CLASSIFICATION OF ENZYMES USING MACHINE LEARNING BASED APPROACHES: A REVIEW

Sanjeev Kumar Yadav[1] and Arvind Kumar Tiwari[2]

[1]Department of CSE, KIT, Varanasi, U.P., India
[2]Department of CSE, IIT (BHU), Varanasi, India

ABSTRACT

Enzymes play an important role in metabolism that helps in catalyzing bio-chemical reactions. A computational method is required to predict the function of enzymes. Many feature selection technique have been used in this paper by examining many previous research paper. This paper presents supervised machine learning approach to predict the functional classes and subclass of enzymes based on set of 857 sequence derived features. It uses seven sequence derived properties including amino acid composition, dipeptide composition, correlation feature, composition, transition, distribution and pseudo amino acid composition .Support vector machine recursive Feature elimination (SVRRFE) is used to select the optimal number of features. The Random Forest has been used to construct a three level model with optimal number of features selected by SVMRFE, where top level distinguish a query protein as an enzyme or non-enzyme, second level predicts the enzyme functional class and the third layer predict the sub functional class. The proposed model reported overall accuracy of 100%, precision of 100% and MCC value of 1.00 for the first level, whereas accuracy of 90.1%,precision of 90.5% and MCC value of 0.88 for second level and accuracy of 88.0%, precision of 88.7% and MCC value of 0.87 for the third level.

KEYWORDS

Enzymes, Feature extraction, Feature selection, Classification, SVMRFE, Random forest

1. INTRODUCTION

Enzyme function prediction is a very challenging task in Bioinformatics. It is the most essential molecule in our life. Enzyme is responsible for catalysis of biochemical reaction for metabolism, structuring the organs and for maintenance of cellular component. The knowledge of functionality of enzyme is very crucial to develop new approaches in biological process. The enzyme function prediction based on experiments requires a large experimental and human effort to analyze single gene or enzyme. This drawback is removed by a number of experimental procedures containing high-throughput that have been invented to investigate the methods which is used for function prediction. These procedures generate a variety of data, such as enzyme sequences, enzyme structures, enzyme-enzyme interaction network and gene expression data. There are many databases that maintain these data, such as, DIP [1], SWISS-PROT [2], NCBI [3], PDB [4] and STRING [5]. The functions of the enzymes are categorized by the activity of the enzymes. These functions describes the activity at the molecular level such as catalysis and biological process. It describes the border function which is carried out by assemblies of molecular function such as cellular component and metabolic pathways that describes the

DOI : 10.5121/mlaij.2015.2404

component of a cell in which the enzyme perform functions.[6].There are various computational methods that are used to predict these enzyme functions. The similarity based methods that use the structure of a enzyme and the method identifies the enzyme with most similar structure by using structural alignment techniques. The feature based methods finds the features of the amino acid sequence, enzyme-enzyme interactions and structure of the enzyme. It uses these features to find the function of the enzyme. In this approach, the features are extracted from the individual enzyme hence these features are more important since they are defined by the knowledge of the enzyme function and factors that affects the enzyme function. The computational intelligence based methods such as Support Vector Machine, Decision Tree, Random Forest, Neural Network, Naive Bayesian Classifier and k Nearest Neighbor are used to identify the most appropriate functional class of enzyme from amino acid sequence, structure, enzyme-enzyme interaction and gene expression data

The amino acid sequence of an enzyme is also called as primary structure of enzyme. It plays an important role in enzyme function prediction. Homology based methods are used for enzyme function prediction from amino acid sequence. The global and the local sequence alignment [7, 8, 9] and sequence motifs [10, 11] are used for enzyme function prediction. The BLAST [7] is used for comparison of the amino acid sequences that optimizes the maximal segment pair score. FASTA [12] is also used for the comparison of amino acid sequences. The sequence profile based methods such as gapped BLAST and Position-Specific Iterated BLAST (PSI-BLAST)[8] uses the position-specific score matrix to search the enzyme databases, which provides a high sensitivity for detecting remote homologs. The enzymes that are diverged from a common ancestral gene might have same function, but no detectable sequence similarity [13]. Therefore the sequence similarity-based approaches may not be always useful for enzyme function prediction.

The enzymes structure is more conserved than sequences. When a sequence-based function prediction cannot achieve high accuracy then the three-dimensional structures of enzymes is used for enzyme function prediction. The structure of the enzyme determines various functional features such as active site residues, cellular location, overall fold and their conformation in enzymes, interactions with ligands and other enzyme. On the basis of this information, we classify the structure based enzyme function prediction. In paper [14] the fold information depends on global and local structure alignment algorithms. The global and local similarities between enzymes indicate the functional similarities and it is useful for inferring functions of enzymes. The author of paper [15] developed a molecular binding site prediction method that integrates structure based methods with sequence conservation estimates to identify enzyme surface cavities, ligand binding packets, catalytic sites, individual ligand binding residues, and drug binding pockets. The paper [16] developed a binding sites database that gives known enzyme-ligand binding sites and allows fast retrieval of other binding sites with similar structure that are independent of sequence and fold similarity. The catalytic site structure is highly conserved between distantly related enzymes. Hence, when sequence and overall structure based enzyme function method fails, templates representing the catalytic sites are used for enzyme function prediction. Research paper [17] presents a library of structural templates representing catalytic sites. Detection of similar local geometries of functionally important residue implies similar functions even in distantly related enzymes. This approach is useful for prediction of enzyme function. The analysis of enzyme structure has given valuable information for enzyme function prediction. The structural properties based enzyme function prediction is more useful for single static structure but it is not useful in dynamic structure. Structural dynamics can enhance the function prediction. To find the binding sites for enzyme function prediction, the author of paper [18] used the molecular dynamics simulation along with the structure based function prediction algorithm. The major limitations of this method is the availability of the high resolution structural data of enzymes.

Sequence and structure-based methods uses homology relationships among enzymes for enzyme function prediction. When sequence based homology is failed, then the structure based homology is used to predict the enzyme function. But these methods have several problems in enzyme function prediction due to the availability of adequate data of homologous enzymes and it may contain a different function. This method fails when these homology relationships cannot be established for target enzymes which is described in paper [19].Structure based enzyme function prediction has been restricted due to the availability of a limited number of and folds and structures in the databases. Sequence based enzyme function prediction is a great challenge for those enzymes that has low sequence similarity or no sequence similarity to enzymes of known function. Due to this, computational intelligence techniques have been useful in enzyme function prediction by using sequence derived properties that are independent of sequence similarity. It has a great potential for low and non-homologous enzyme [20].

The advancement of high throughput technologies that produces large amount of high throughput data such as enzyme-enzyme interaction and gene expression data that are useful in enzyme function prediction. Gene expression measurement provides which genes are active under certain condition and produces enzyme to perform a given function under such condition. It is expected that co-expressed genes perform similar cellular function. Various computational intelligence techniques have been used to annotate unknown gene that co-express with known genes. Enzyme performs a specific function by interacting with another enzyme. So enzyme-enzyme interaction network provides a valuable data that are useful in enzyme function prediction. The usefulness of these technique has been discussed in several recent studies, hence it is important to have a deep knowledge about these computational intelligence techniques used in enzyme function prediction.

2. DATA PREPARATION STEPS

This section describes how computational intelligence techniques are used in enzyme function prediction. The general steps used in computational intelligence techniques are as follows.

2.1 Data Acquisition

Machine learning needs two things to work properly, i.e. data and model. While acquiring the data, make sure that enough features are populated to train correctly the learning model.

2.2 Data Pre-Processing

Raw data is highly susceptible to noise, inconsistency and missing values. The quality of the data affects the data mining result. In order to improve the quality of data and results, raw data is pre-processed. It improves the efficiency of the mining process. Data pre-processing is the most critical steps in data mining process. It deals with preparation and transformation of the initial dataset. Data pre-processing methods are mainly divided into four categories:

Data Cleaning: Data cleaning is used to fill in missing values, identifies outliers, smooth noisy data, and correct data inconsistencies.

Data Integration: Data integration combines the data from multiple sources and form a data store. Metadata, data conflict detection and correlation analysis contribute towards smooth data integration.

Data Transformation: Data transformation routines transform data into appropriate forms for mining. The attribute data may be normalized so that it can fall between ranges within 0 to 1.0.

Data Reduction: Data reduction technique such as dimension reduction, discretization, data compression, data cube aggregation and numerosity reduction can be used to obtain a reduced representation of the data while minimizing the loss of information content.

2.3 Feature Extraction

In machine learning, pattern recognition or feature extraction starts from an initial set of data and builds features intended to be informative, facilitating the subsequent learning and generalization steps, non-redundant. In some cases it leads to better human interpretations. Feature extraction is closely related to dimensionality reduction.

When input data to an algorithm is too large that it cannot be processed and it is suspected to be redundant e.g. measurement is same in both meters and feet. The repetitiveness of the images presented as pixels. It can be transformed into a reduced set of features (also known as features vector). This process is called as feature extraction. The features that are extracted are expected to contain relevant information from the input data. Due to this, the desired task can be performed by using the reduced representation instead of complete initial data.

2.4 Feature Selection

In machine learning, feature selection is also known as variable selection, attribute selection or variable subset selection. It is the process of selecting a subset of the relevant features to use in model construction. The main assumption while using the feature selection technique is that data contains many redundant and irrelevant features. Redundant features are those that provide no information than the currently selected features, and irrelevant features provide no useful information in any context.

Feature selection techniques are different from feature extraction technique. Feature extraction technique creates new features from functions of original features. Feature selection technique returns a subset of features. It is often used in domains where there are so many features and comparatively only few samples.

2.5 Enzyme Function Prediction

Enzyme function prediction is done by applying various classification techniques along with that various feature selection techniques is used. The steps (Figure 1) are as follows:

1. Preparation of the training datasets in specific format for each computational intelligence techniques by using feature extraction from the input datasets.
2. Select the features by using feature selection techniques that affects the particular class of input data.
3. Design and develop computational intelligence techniques to predict the function of enzyme.
4. Using appropriate parameters and input data train the Computational intelligence techniques and construct a prediction model.
5. Validation of prediction model using test data to evaluate the performance of the model.

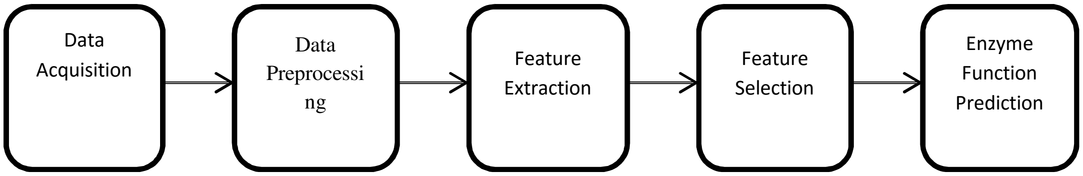

Figure 1. Steps in enzyme function prediction

3. COMPUTATIONAL INTELLIGENCE TECHNIQUES

This section shows an overview of various computational intelligence techniques used for enzyme function prediction, including feature selection techniques such as filter and wrapper method, classification techniques such as artificial neural network, K-nearest-neighbor, Decision Tree, Random Forests, Naive Bayes, support vector machine.

3.1. Features Selection Techniques

Feature selection is the process of selecting a best subset of features, among all the features that are useful for the learning algorithms. The goals of the feature selection is to avoid over fitting, increase the overall accuracy and improve the prediction performance.

3.1.1. Filter method

Filter method calculates the relevance score of features by using the essential properties of data and then low scoring features are removed. This method evaluates features in isolation without considering the correlation between features, but it is useful for large high dimensional datasets. Filter methods select variables regardless of the model. The method is based on general features like correlation with variables to predict. This method suppresses the least interesting variables. Other variables will be part of the model classification, and regression is used to classify or data prediction. Filter methods are generally effective in computation time and it is robust to overfitting. Filter methods selects redundant variables as they do not consider the relationships between variables. Therefore, the method is mainly used as pre-process method.

3.1.2. Wrapper method

Wrapper method uses the classifier for searching the subset of features. It uses the backward elimination process to remove the irrelevant features from subset of features**.** In this method the

rank of the features is calculated recursively and low rank features are removed from the result. It interacts feature subset and classifier so predict the future dependencies. It has a higher over-fitting risk than the filter method. Wrapper method evaluates subsets of variables that allows unlike the filter approach to detect the possible interactions between the variables.
The two main disadvantages are:

- Increasing overfitting risk when number of observations is insufficient.
- The significant computation time when number of variables is too large.

3.2. Artificial Neural Network

Artificial neural networks are inspired by the concept of biological nervous system. ANNs are the collection of computing elements (neurons) that may be connected in several ways. In ANNs the effect of the synapses is represented by the connection weight that modulates the input signal. The architecture of the ANNs is fully connected, a three layered (input layer, hidden layer and output layer) structure of nodes where information flows from input layer to output layer through the hidden layer. The ANNs are capable of linear and nonlinear classification. The artificial neural network learns by adjusting the weights in accordance with the learning algorithms. ANN is capable of processing and analyzing large complex datasets, containing non-linear relationships. There are various types of artificial neural network architecture that are used for enzyme function prediction such as perceptron, multi-layer perceptron (MLP), radial basis function networks and Kohonen self-organizing maps.

3.3. Support Vector Machine

The Support Vector Machine is based on the statistical learning theory [21]. It is capable of resolving linear and non-linear classification problems. The idea of the classification by using SVM is to separate the examples through the linear decision surface. It is used to maximize the margin of separation between classes to be classified. The SVM works by mapping the data with a high-dimensional feature space so that the data points can be categorized, even when the data are not linearly separable. A separator between the categories is found, and then data are transformed in such a way so that the separator could be drawn as a hyperplane. The characteristic of new data is used to predict the group to which a new record should belong. After transformation of data, the boundary between the two categories can be defined by a hyperplane. The mathematical function used for transformation is known as kernel function. SVM supports the Polynomial, Linear,Radial basis function (RBF) and Sigmoid kernel types. When there is a straightforward linear separation then linear function is used otherwise we use Radial basis function (RBF), Polynomial and sigmoid kernel function. Besides the separating line between the categories, SVM also finds marginal lines that define the space between these two categories. Data points that lie on margins are known as support vectors.

3.4. K Nearest-Neighbor

The kNN classifiers are based on finding the *k* nearest neighbor, and taking a majority of vote among the classes of these *k* neighbors, to assign a class for the given query [22]. kNN is more efficient for a large datasets and it is robust when processing noisy data, but high computation cost, reduces its speed. In pattern recognition, the *k* Nearest Neighbor algorithm is a non-parametric method that is used for regression and classification .In both case, the input consists of *k* closest training examples in the feature space. The output depends on whether the *k*-NN is used for regression or classification. kNN is a type of instance-based learning, or called as lazy learning, where function is the only approximated locally and all the computations is deferred until classification. The *k*-NN algorithm is the easiest among all machine learning algorithms.

3.5. Decision Trees

Decision tree is a branch-test-based classifier. C4.5 [23] is a decision tree classifier. By using the knowledge of the training data it creates a decision tree that is used to classify test data. In the decision tree classifier every branch represents a set of classes and leaf represents a particular class. A decision node identifies a test on single attribute value with one branch and its subsequent classes represent as class outcomes.

Decision tree builds the classification or regression model in the form of tree like structure. It breaks down the dataset into smaller and smaller subsets. The associated decision tree are incrementally developed at the same time. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g. Sunny, Overcast, Rainy). Leaf nodes (e.g., Play) represents the classification or decision. The top most decision node in the tree that corresponds to the best predictor is called as root node. The decision trees can handle both numerical and categorical data.

3.6. Random forests

Random Forests is an ensemble classifier of randomly generated decision trees [24]. In this, multiple trees are constructed by using the training datasets. Each tree has access to only randomly sampled subset of the attributes of the training data. In this method, each individual tree predicts a separate class and the majority of the class predicted among the trees is used to predict the class of the test data. It performs better in comparison with the single tree classifier such as CART [25] and C4.5 [23].

Random forest is a classification algorithm. It uses an ensemble of classification trees. Each of the classification trees is built by using a boot-strap sample of the data. At each node of the tree, a set of features is selected from a random subset of the entire feature set and it is used to calculate the feature with highest information. This technique performs very well when compared to other classifiers, including SVMs, neural networks, etc. Random forest uses both bagging and selecting random variable for tree building. Each tree classifies instances by voting for a particular class, once the forest is formed. The class that receives the maximum votes is chosen as final classification. This classifier has various characteristics that is well suited for the enzyme function classification:

(a) It does not require for data to be normalized and can run efficiently on very large datasets.
(b) It can easily handle the missing values.

3.7. Naive Bayes classifier

Naive Bayes classifier is a statistical which is based on Bayes theorem. It calculates the probability of each training data for each class. The class of the test data assigns by using the inverse probability. It assumes that the entire variables are independent, so only mean and variance are required to predict the class. The main advantage of this classifier is that it requires only a small amount of the training data to estimate the mean and variance of data that are used to predict the class.

In machine learning,the Naive Bayes classifiers are family of a simple probabilistic classifiers based on applying Bayes' theorem with strong independence assumptions between features. This classifier are highly scalable, requires number of parameters linear in the number of variables (predictors /features) in learning problem. The maximum-likelihood training can be done by evaluating the closed form expression which takes linear time, rather than by expensive iterative approximation that is used for many other types of classifiers.

4. COMPUTATIONAL INTELLIGENCE TECHNIQUES IN THE PREDICTION OF ENZYME FUNCTION/ FAMILY

Enzyme catalyzes bio-chemical reactions and plays very important role in metabolic pathways. A small fraction of residue may involve in the catalytic reactions, these catalytic residues are the most important part of enzyme. The knowledge about the catalytic residue is necessary for understanding enzyme function and catalytic mechanism. The author of paper [26] developed an artificial neural network based model. The model is used for the classification of Enzyme from sequence by using the sequence similarity and some other sequence derived features such as co-translational and post translational modification, physical and chemical properties and secondary structures. The paper [27] used the nearest neighbor method along with the functional domain composition of enzyme to predict the enzyme family classes. [28] proposed a Bayesian based approach for the enzyme classification with structure derived properties of an enzyme, [29] used support vector machine based methods by using the feature vector from enzyme functional domain composition. [30] proposed a support vector machine based method with the amphiphilic pseudo amino acid composition.[31] proposed a fuzzy kNN based method with the amphiphilic pseudo-amino acid composition that includes both the features such as function related features and sequence order related features. The author of paper [32] used optimized evidence-theoretic k nearest neighbor classifier with the functional domain composition and PSSM. They also constructed a top-down three layer model where top layer classifies a query enzyme sequence as enzyme or non-enzyme, second layer predicts the main functional class and the bottom layer predicts the sub-function class. The paper [33] proposed a K-nearest neighbor (KNN) based method for the prediction of Enzyme family with amino acid composition. The author of paper [34] proposed an ANN based method for the enzyme sequence classification by using the sequence motif. For the enzyme function classification, the paper [35] proposed a SVM-Port, a SVM based tool by using the amino acid sequence. For the prediction of enzyme structural classes, author of the paper [36] proposed Fusion based approach for the prediction of enzyme structural classes of dual layer support vector machine with the pseudo amino acid composition that contain the information related to the sequential order of enzyme and the distribution of hydrophobic amino acid along with the chain of amino acid sequences. The author of paper [37] proposed the fuzzy K nearest neighbor classifier that is based on pseudo amino acid composition by using the approximate entropy and hydrophobicity pattern of amino acid sequence [38] and paper [39] used the support vector machine based method with pseudo amino acid composition along with conjoint triad features (CTF) to represent the enzyme sequences as not only the composition of amino acid, but also the neighbor relationship in the sequence for the prediction of subfamily and function of Enzymes respectively. In paper [40] an integrated method of support vector machine was proposed with discrete wavelet transform for the classification of enzyme family by using the hydrophobicity of amino acid from pseudo amino acid composition. The paper [41] proposed a Random Forest based method that is used to predict the functional class and sub-class of enzyme based on sequence-derived features. A top-down three layer model is constructed where the top layer classifies a query enzyme sequence as enzyme or non-enzyme, second layer predicts the main functional class and the bottom layer predicts the sub-function class. The paper [42] presented N-to-1 Neural Network for prediction of the Enzyme by using the amino acid sequences. In the paper [43] random forest based method was proposed for predicting the enzyme functions, with a set of specificity determining the residues. For the classification of secretory and non-secretory enzymes. The paper [44] proposed a SVM based method with PSI-BLAST by using the sequence similarity, amino acid composition, dipeptide composition and physiochemical properties. The author of the paper [45] presented a SVM and Random forest based methods for the prediction of enzyme function by using the sequence derived properties. In the paper [46] a SVM based approach by using the features extracted from global structure based on fragment libraries was proposed.

The author of the paper [47] and [48] presented SVM that is useful for protein function classification with accuracy of 84-96%. This paper uses protein classes such as RNA binding, drug absorption, drug delivery, homodimer, etc. using feature vectors like amino acid composition, polarizability, hydrophobicity and secondary structure. It proves that classification using machine learning approach and sequence features can be useful for protein function prediction. In the paper [49] that used support vector machine to predict the enzyme main functional class and reports accuracy around 66-90%.The paper [50] classify enzymes by employing a C4.5 classifier on 36 features drawn from protein sequence to build classification model and achieve precision and recall in the range of 86-92%.

The author of paper [51] proposed a method that assign the function from the structure of protein using EC number. The paper used one-class versus one-class SVM to predict the protein function. He found the accuracy between 35-60%. Below Table 1 describes the various methods and their accuracy.

Table 1. Summary of Computational Intelligence techniques for prediction of enzyme function/family

Author	**Computational Method**	**Enzyme Function/ Family**	**Performance**	**Datasets**
(Cai, Y. D.*et al.,* 2005)	kNN	Family	Accuracy: 85%	Functional domain composition
(Zhou, X. B.*et al.,* 2007)	SVM	Family	Accuracy: 80.87%.	Amphiphilic pseudo amino acid composition
(Huang, W. L*et al.,* 2007)	kNN	Family	Accuracy : 76.6%,	Amphiphilic pseudo-amino acid composition
(Qiu JD *et al.,* 2010)	SVM with DWT	Family	Accuracy: 91.9.	Pseudo amino acid composition
(WangYC *et al.,* 2010)	SVM	Family	MCC: 0. 92 and Accuracy: 93%	Pseudo amino acid composition with (CTF)

(Borro *et al.,* 2006)	Bayesian Classifier	Function	Accuracy: 45%.	Structural properties
(Lu L *et al.,* 2007)	SVM	Function	Accuracy :91.32%	Functional domain composition
(Shen, H. B. *et al.,* 2007)	OET-kNN	Function	Overall accuracy: 91.3%, 93.7% and 98.3% for the 1st, 2nd and 3rd level	Functional domain composition and PSSM
(Nasibov, E.*et al* .,2009)	k-NN	Function	Accuracy: 99%	Amino acid composition
(WangYC *et al.,* 2011)	SVM	Function	Accuracy: 81% to 98% and MCC: 0.82 to 0.98	Pseudo amino acid composition with (CTF
(Kumar, C.*et al.,* 2012)	Random Forest	Function	Overall accuracy: 94.87%, 87.7% and 84.25% for the 1st, 2nd and 3rd level.	Sequence-derived features
(Volpato *et al.,* 2013)	N-to-1 Neural Network	Function	Overall accuracy: 96%, Specificity: 80% and FP rates: 7%.	Amino acid sequences
(Nagao C, *et al.,* 2014)	Random forest	Function	Precision: 0.98 and Recall: 0.89	Set of specificity determining residues
(Cai *et al.,* 2003)	SVM	Enzyme function	Accuracy: 69.1–99.6%.	Amino acid sequence
(Yadav *et al.,* 2012)	SVM	Enzyme function	Accuracy: 95.25%	Structural features based on fragment libraries.
(Lee *et al.,* 2009)	SVM and Random Forest	Enzyme function	Accuracy: 71.29- 99.53% by SVM and 94- 99.31% by random forest	Sequence derived properties
(Chen C *et al.,* 2006)	SVM	Enzyme structural class	Sensitivity: 85.6% and Specificity: 86.1%.	Pseudo amino acid composition
(ZhangTL *et al.,* 2008)	Fuzzy kNN	Enzyme structural class	Accuracy: 56.9%	Pseudo amino acid composition, approximate hydrophobicity and entropy
(Cai *et al.,* 2003)	SVM	Protein function	Accuracy: 69.1–99.6%.	Amino acid sequence
(Han L.Y et al.,2004)	SVM	Protein function	Accuracy: 84-96%	AAC, Polarizability, Hydrophobicity
(Lee Bum Ju, 2004)	SVM	Function	Accuracy:66-90%	Sequence Derived Feature
(Lee Bum Ju et al,2007)	C4.5	Function	Precision and Recall:86-92%	Amino Acid Composition
(Paul Dobson, et al., 2005)	SVM	Protein Function	Accuracy:35-60%	Amino Acid Composition

5. CASE STUDY FOR ENZYME FUNCTION PREDICTION USING MACHINE LEARNING BASED APPROACHES

5.1 Datasets

The sequence of enzymes have been collected from the enzyme repository of UNIPROT [52] database. Uniprot is a universal enzyme resource, and also a central repository of enzyme data. It is created by combining the Swiss-Prot, TrEMBL and PIR-PSD databases. In this paper, 2647 enzyme sequences and 700 non-enzymes has been selected. For level 1, a total of 2647 number of enzyme sequences has been selected. For level 2 and level 3, we have selected the optimized features by testing on various numbers of features (50, 100, 150, 200, 250, 300, 350, and 400) .The best result is obtained with 300 features. SVM-RFE has been used to perform feature selection. The data and their description used in the analysis is mentioned below in table 1. WEKA [53] is a widely used machine learning open source tool that has been used in this paper to analyze the data and carry out result.

5.2 Methodology

5.2.1 FEATURE EXTRACTION

In machine learning, research has proved that in order to build an accurately predicting model, class size should be balanced. Therefore, the number of sequences was kept balanced while extracting from every main class. Each main class has many sub-class, hence SVMRFE was used to extract and well distribute the sequence in each sub-class. Here the data used was obtained after removal of all identical sequences present in Enzyme (each main class) and Non Enzyme. The distribution of sequences across all enzyme functional main and sub-class are described in Table 2. The table represents the sequences after removing all the identical sequences. PROFEAT [54]: an online tool is used in order to extract sequence-derived features. It generates a list of sequence-derived features. This tool generates values for features such as amino acid composition, Moran autocorrelation, dipeptide composition, transition, composition, distribution and pseudo amino acid composition.

5.3 Feature Selection

5.3.1 SVMRFE

The Support Vector Machine Recursive Feature Elimination technique (SVM-RFE) [55] algorithm is a wrapper based feature selection method that generates the ranking of features by using backward feature elimination technique. SVM-RFE was originally proposed to perform gene selection for cancer classification problem. The key idea is to eliminate redundant data and gives better and more compact data subsets. The features are eliminated according to a specific criteria related to support for their discrimination function. This is a weight based method, where at each step the coefficients of weight vector of a linear SVM are used as the feature ranking criterion.

The SVM-RFE algorithm has four major steps:

1. Train an SVM on training set.
2. The features are ordered using the weights of the resulting classifier.
3. The smallest weight features are eliminated.
4. Repeat the same process with the training set restricted to the remaining features.

Table 2. Data Description

Families	Enzyme classes	EC sub-classes	No.Seq.	No. Seq.	No. Seq.
Enzymes	Oxidoreductases	1.1	99	533	2647
		1.2	88		
		1.3	97		
		1.4	88		
		1.5	98		
		1.6	63		
	Transferases	2.1	98	553	
		2.2	91		
		2.3	72		
		2.4	97		
		2.5	97		
		2.6	98		
	Hydrolases	3.1	98	420	
		3.2	56		
		3.3	90		
		3.4	99		
		3.5	93		
		3.6	82		
	Lyases	4.1	89	330	
		4.2	97		
		4.3	61		
		4.4	83		
	Isomerases	5.1	90	445	
		5.2	99		
		5.3	62		
		5.4	97		
		5.5	98		
	Ligases	6.1	59	365	
		6.2	73		
		6.3	40		
		6.5	98		
		6.6	95		
Non					700

5.4 Classification of Enzyme Functional Classes and Subclasses

5.4.1 Proposed Model

In this paper, a three tier model is used to predict enzyme function class and subclass. This model consists of three layers: the first layer of the model classifies enzymes and non-enzymes; second layer predicts the main functional class of enzymes and the third layer predict their sub-class of enzymes. The three layer classifier is built by using Random Forest with the best 300 number of features extracted using SVMRFE to achieve highest accuracy. A flowchart of this model with optimized feature technique at each level is illustrated in Figure 2.

The figure shows the different components of the three tier model. In this three tier model, first level classifies enzymes and non-enzymes. The model has been trained using random forest with parameter value mtry = 25 and ntree = 500. The second level classifies enzyme into their main function class, and the third level classifies enzymes whose main class is predicted at level 2, in their sub-classes. In Level 3, six classifiers are used, each for the corresponding main class. Level 3 classifier is built using random forest where parameter values similar to level 2, i.e., mtry = 7 and ntree = 500. The values correspond to minimum OOB error rate that is obtained by using this classifier.

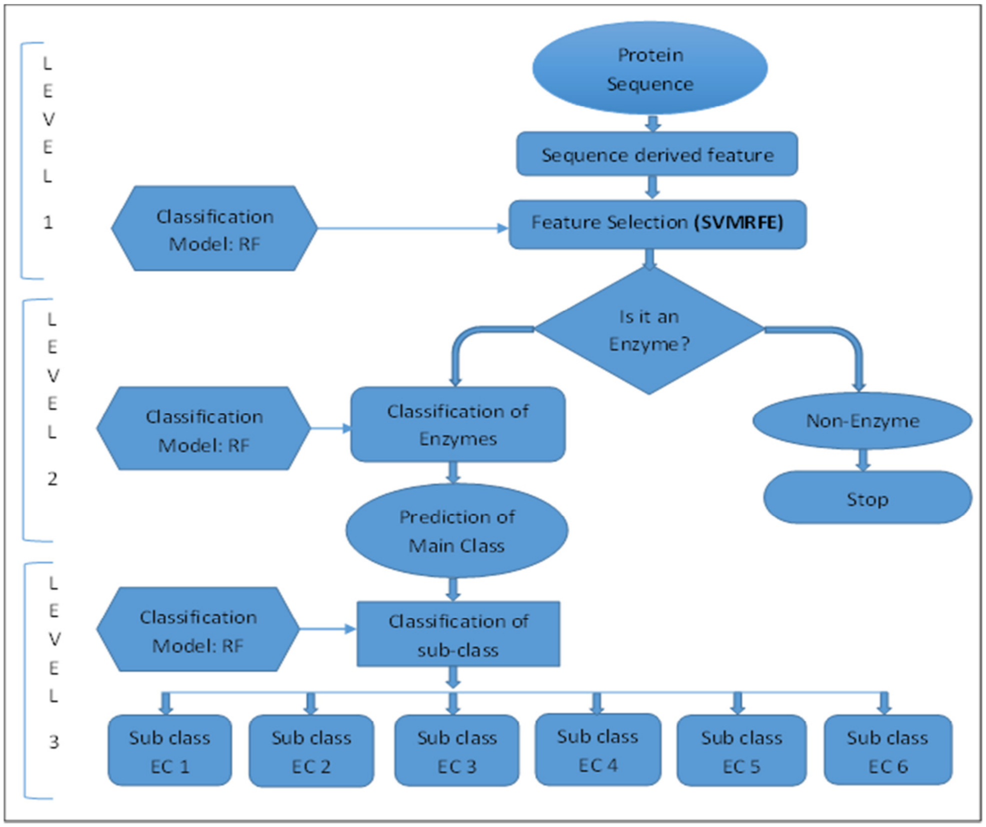

Figure 2. Three tier model to predict enzyme functional classes and sub-classes

5.5. Performance Evaluation of Classification

The performance of different classifiers is measured by using the the quantity of True positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). Here TP (True Positive) is the number of positive instances that are classified as positive, FP (False Positive) is the number of Negative instances that are classified as positive, TN (True Negative) is the number of Negative instances that are classified as Negative and FN (False Negative) is the number of positive instances that are classified as Negative.

Accuracy, Precision and Matthew correlation coefficient (MCC) are defined as follows:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$$

$$Precision = \frac{TP}{TP+FP}$$

MCC is a balanced measure that considers both true and false positives and negatives. The MCC can be obtained as:

$$MCC = \frac{(\mathrm{TP})(\mathrm{TN}) - (\mathrm{FP})(\mathrm{FN})}{\sqrt{[\mathrm{TP} + \mathrm{FP}][\mathrm{TP} + \mathrm{FN}][\mathrm{TN} + \mathrm{FP}][\mathrm{TN} + \mathrm{FN}]}}$$

6 RESULT AND DISCUSSION

Initially, research was carried out with many different classifiers to identify the best classifier for dataset. In this paper we performed tenfold cross-validation experiments between LibSVM, Naive Bayes, k-Nearest Neighbor and Random Forest. The experiment was performed on all level for the given model, i.e. to differentiate enzymes and non-enzymes and to predict the main and sub-class of the enzymes. Figure 3 illustrates the True Positive (TP) Rate for four different classifiers, out of which random forest out performed all other remaining classifiers by providing the highest accuracy, precision and MCC.

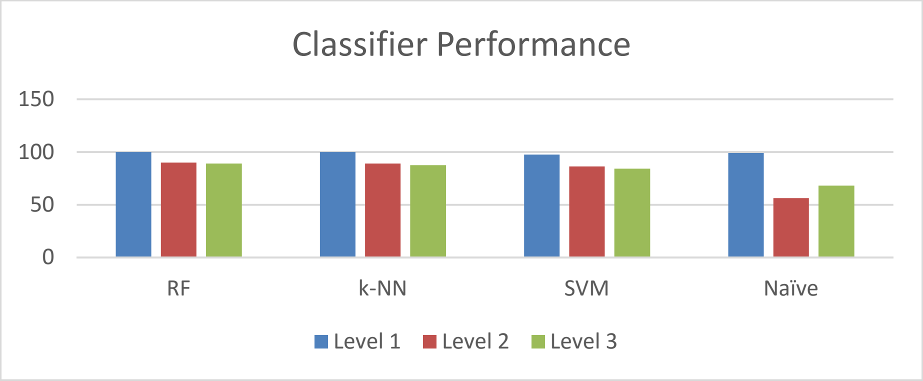

Figure 3. Performance evaluation of different classifiers

Table 3. Result Analysis for the prediction of enzymes and non-enzymes

Enzyme Type	Sequence	True Positive	Precision	Accurac y	MCC
Enzyme	2647	2647	100	100	1.00
Non-Enzyme	700	700	100	100	1.00
Overall	**3347**	**3347**	**100**	**100**	**1.00**

6.1 Classification of Enzyme | Non-Enzyme

Level 1 classifies enzyme sequence from non-enzyme sequence. In this paper, tenfold cross validation is performed on the dataset.2647 enzyme sequence and 700 non-enzyme sequences are selected. The least OOB error is obtained when the two parameter values are ntree = 500 and mtry = 25. We anchor these parameter values for first level classifier. Table 3 shows the results obtained from tenfold cross-validation experiment. The overall accuracy achieved is 100%, which is quiet favorable when we compare it to other articles [18, 19], like neural network which shows the accuracy of approximately 75% and 91.3%.

6.2 Classification of Enzyme Main Class

Using a dataset of 2647 enzyme sequence, level 2 classifies enzyme sequences out of six main functional classes. We again tested the two parameters for different values (mtry and ntree). The summarized classification result that is obtained for level 2 is shown in Table 4. The accuracy obtained for this level was 90.1%, where 2384 enzymes being correctly classified into their main functional class out of a total of 2647 instances. The result was obtained by testing different set of feature(50, 100, 150, 200, 250, 300, 350, and 400) by using SVM-RFE from original set, and the best result was achieved by using 300 features. The comparative analysis of the result shows that the Random Forest along with SVMRFE provides the overall accuracy of 90.1% at level 2 i.e. for the prediction of enzyme functional enzyme class.

Table 4. Result analysis for the classification of enzyme functional classes

EC Class	**Total Enzymes**	**True Positive**	**False Positive**	**Precision**	**Accuracy**	**MCC**
1	533	486	47	90.7	91.2	0.88
2	553	503	50	82.5	91.0	0.82
3	420	375	45	91.2	89.3	0.88
4	330	278	22	100.0	84.2	0.90
5	446	398	48	93.6	89.2	0.89
6	365	344	21	88.9	94.2	0.90
Overall	**2647**	**2384**	**233**	**90.5**	**90.1**	**0.88**

6.3 Classification of Enzyme Subclass

In Level 3 of this model, we classify enzymes whose main class has been predicted, into their corresponding subclass. In this level, there are six random forest classifiers, each to predict the sub class for enzymes under the main class. Here, we used the same parameter values as we used in second level for all six classifiers of level 3, i.e., ntree = 500 and mtry = 7. It is because we did not find any big difference in OOB error even after varying values of mtry between 5 and 25. The result for each sub class is show below in Table 5. This result was also achieved by testing different feature sets shown in level 2 by using SVM-RFE from original set, and the best result was achieved using 300 features.

Table 5. Result analysis for the classification of enzyme sub-functional classes

Classes	Sub-	No.	Precision	Accuracy	MCC
Oxidoreductases	1.1	99	95.9	93.9	0.93
	1.2	88	100.0	95.5	0.97
	1.3	97	93.1	96.9	0.93
	1.4	88	95.1	88.6	0.90
	1.5	98	76.0	96.9	0.82
	1.6	63	97.7	68.3	0.79
	Overall	**533**	**92.5**	**91.4**	**0.90**
Transferases	2.1	98	84.7	95.9	0.87
	2.2	91	97.8	98.9	0.98
	2.3	72	100.0	84.7	0.91
	2.4	97	87.1	90.7	0.86
	2.5	97	98.9	91.8	0.94
	2.6	98	94.9	94.9	0.93
	Overall	**553**	**93.6**	**93.1**	**0.91**
Hydrolases	3.1	56	97.9	83.9	0.89
	3.2	90	93.5	96.7	0.96
	3.4	99	86.0	92.9	0.83
	3.5	93	96.6	92.5	0.93
	3.6	82	90.4	91.5	0.88
	Overall	**420**	**92.4**	**92.1**	**0.90**
Lyases	4.1	89	98.9	98.9	0.98
	4.2	97	96.9	97.9	0.96
	4.3	61	100.0	93.4	96.0
	4.4	83	95.3	98.8	0.96
	Overall	**330**	**97.6**	**97.6**	**0.96**
Isomerases	5.1	90	94.6	96.7	0.94
	5.2	99	95.0	97.0	0.94
	5.3	62	100.0	95.2	0.97
	5.4	97	93.1	96.9	0.93
	5.5	98	98.9	93.9	0.95
	Overall	**446**	**96.1**	**96.0**	**0.94**
Ligases	6.1	59	100	100.0	1.00
	6.2	73	97.3	98.6	0.97
	6.3	40	100.0	87.5	0.92
	6.5	98	93.3	99.0	0.94
	6.6	95	95.7	93.7	0.92
	Overall	**365**	**96.5**	**96.4**	**0.95**
Complete Overall		**2647**	**88.7**	**88.0**	**0.87**

7 CONCLUSIONS

In this paper, we presented the state of art comprehensive review based on the computational intelligence technique used for the prediction of functional class and subclass of enzyme.
The summary of the result obtained by various researchers available in literature to predict the enzyme functional class and subclass is also presented.

The case study including the computational study of various machine learning based approaches were presented. Here in this paper, it is observed that Random Forest with SVMRFE based feature selection may be useful for the prediction of enzyme functional class and subclass.
Enzyme function classification is a challenging problem to accurately predict enzyme mechanisms, but by using a different set of features extracted from enzyme sequence and classifier Random Forest, we have demonstrated a three tier model to accurately predict enzyme functional classes. 300 features have been extracted by using feature selection technique SVMRFE. We highlighted different existing tools can be re-used to address interesting problems in Bioinformatics. The results show that Random Forest classifier is useful for classifying multi-class problems like enzyme function classification. The RF classifier achieved a high accuracy on a large enzyme dataset. Further, our analysis suggests that RF with SVMRFE could improve the result by correctly predicting different functional classes of enzymes at each level.

REFERENCES

1. I. Xenarios, L. Salw´ınski, X. J. Duan, P. Higney, S.-M. Kim,and D. Eisenberg, “DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions,”*Nucleic Acids Research*, vol. 30, no. 1, pp. 303–305, 2002.
2. B. Boeckmann, A. Bairoch, R. Apweiler et al., “The SWISSPROTprotein knowledgebase and its supplement TrEMBL in2003,” *Nucleic Acids Research*, vol. 31, no. 1, pp. 365–370, 2003.
3. R. Edgar, M. Domrachev, and A. E. Lash, “Gene expressionomnibus: NCBI gene expression and hybridization array data repository,” *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.
4. H. M. Berman, J. Westbrook, Z. Feng et al., “The protein databank,” *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000
5. D. Szklarczyk, A. Franceschini, M. Kuhn et al., “The STRINGdatabase in 2011: functional interaction networks of proteins,globally integrated and scored,” *Nucleic Acids Research*, vol. 39,supplement 1, pp. D561–D568, 2011.
6. Bork, Peer, Eugene V. Koonin, Predicting functions from protein sequences—where are the bottlenecks?, *Nature genetics* 18, no. 4: 313-318, 1998.
7. Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, David J. Lipman, Basic local alignment search tool, *Journal of molecular biology* 215, no. 3: 403-410, 1990.
8. Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, David J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic acids research* 25, no. 17: 3389-3402, 1997.
9. Pearson, W. R., Effective protein sequence comparison, *Methods Enzymol*., 266, 227–258, 1996.
10. Bairoch, Amos, Philipp Bucher, Kay Hofmann, The PROSITE database, its status in 1995, *Nucleic Acids Research* 24, no. 1: 189-196, 1996.
11. Attwood, T. K., M. E. Beck, A. J. Bleasby, D. J. Parry-Smith, PRINTS- a database of protein motif fingerprint, *Nucleic acids research* 22, no. 17: 3590, 1994.
12. Pearson, W. R., Lipman, D. J., Improved tools for biological sequence comparison, *Proc Natl Acad Sci USA*, 85, 2444-2448, 1998.

13. Benner, Steven A., Stephen G. Chamberlin, David A. Liberles, Sridhar Govindarajan, Lukas Knecht, Functional inferences from reconstructed evolutionary biology involving rectified databases–an evolutionarily grounded approach to functional genomics, *Research in microbiology* 151, no. 2: 97-106, 2000.
14. Harrison, A., Pearl, F., Sillitoe, I., Slidel, T., Mott, R., Thornton, J., Orengo, C., Recognizing the fold of a protein structure, *Bioinformatics*, 19(14), 1748-1759, 2003.
15. Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M., Funkhouser, T. A., Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure, *PLoS computational biology*, 5(12), e1000585, 2009.
16. Gold, ND., Jackson, RM., Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships, *J Mol Biol*., 3, 355(5), 1112-24, 2006.
17. Torrance, J. W., Bartlett, G. J., Porter, C. T., Thornton, J. M., Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families, *Journal of molecular biology*, 347(3), 565-581, 2005.
18. Glazer, DS. Radmer, RJ. Altman, RB. Improving structure-based function prediction using molecular dynamics, *Structure*, 17, 919–929, 2009.
19. Whisstock, JC. Lesk, AM., Prediction of protein function from protein sequence and structure, *Q Rev Biophys*, 36, 307-40, 2003.
20. Han, L., Cui, J., Lin, H., Ji, Z., Cao, Z., Li, Y., Chen, Y., Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity, *Proteomics*, 6(14), 4023-4037, 2006.
21. Cortes, C.,Vapnik, V., Support-vector networks, *Machine learning*, 20(3), 273-297, 1995.
22. Johnson, R. A. Wichern, Applied multivariate statistical analysis, Edition, New Jersey: *Prentice-Hall, Inc,* 1982.
23. Quinlan, J. R., C4. 5: programs for machine learning (Vol. 1). *Morgan kaufmann,* 1993.
24. Breiman, Leo, Random forests, *Machine learning* 45, no. 1: 5-32, 2001.
25. Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A., Classification and regression trees, *CRC press*, 1984.
26. Jensen, L. J., Skovgaard, M., Brunak, S., Prediction of novel archaeal enzymes from sequence-derived features, *Protein Sci*., 3, 2894–2898, 2002.
27. Cai, Y. D., Chou, K. C. ,Using functional domain composition to predict enzyme family classes, *Journal of proteome research*, 4(1), 109-111, 2005.
28. Borro, Luiz C., Stanley RM Oliveira, Michel EB Yamagishi, Adaulto L. Mancini, José G. Jardine, Ivan Mazoni, E. H. D. Santos, Roberto H. Higa, Paula R. Kuser, Goran Neshich, Predicting enzyme class from protein structure using Bayesian classification, *Genet. Mol. Res* 5, no. 1: 193-202, 2006.
29. Lu, L., Qian, Z., Cai, Y. D., Li, Y., ECS: An automatic enzyme classifier based on functional domain composition, *Comput Biol Chem*., 31 (3), 226-232, 2007.
30. Zhou, X. B., Chen, C., Li, Z. C., Zou, X. Y., Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes, Journal of theoretical biology, 248(3), 546-551, 2007.
31. Huang, W. L., Chen, H. M., Hwang, S. F., & Ho, S. Y., Accurate prediction of enzyme subfamily class using an adaptive fuzzy k-nearest neighbor method, *Biosystems*, 90(2), 405-413, 2007.
32. Shen, H. B., Chou, K. C. EzyPred: a top–down approach for predicting enzyme functional classes and subclasses, *Biochemical and Biophysical Research Communications*, 364(1), 53-59, 2007.
33. Nasibov, E., & Kandemir-Cavas, C., Efficiency analysis of KNN and minimum distance-based classifiers in enzyme family prediction, *Computational biology and chemistry*, 33(6), 461-464, 2009.
34. Blekas, Konstantinos, Dimitrios I. Fotiadis, Aristidis Likas, Motif-based protein sequence classification using neural networks, *Journal of Computational Biology* 12, no. 1: 64-82, 2005.
35. Cai, C. Z., L. Y. Han, Zhi Liang Ji, X. Chen, Yu Zong Chen, SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence, *Nucleic acids research* 31, no. 13: 3692-3697, 2003.

36. Chen, C., Tian, Y. X., Zou, X. Y., Cai, P. X., Mo, J. Y., Using pseudo-amino acid composition and support vector machine to predict protein structural class, *Journal of Theoretical Biology*, 243(3), 444-448, 2006.
37. Zhang, T. L., Ding, Y. S., Chou, K. C., Prediction protein structural classes with pseudo-amino acid composition: Approximate entropy and hydrophobicity pattern, *J Theor Biol*, 250 (1), 186-193, 2008.
38. Wang, Y. C., Wang, X. B., Yang, Z. X., Deng, N. Y., Prediction of enzyme subfamily class via pseudo amino acid composition by incorporating the conjoint triad feature, *Protein Pept. Lett.* 17, 1441–1449, 2010.
39. Wang, Y. C., Wang, Y., Yang, Z. X., Deng, N. Y., Support vector machine prediction of enzyme function with conjoint triad feature and hierarchical context, *BMC systems biology*, 5(Suppl 1), S6, 2011.
40. Qiu, J. D., Huang, J. H., Shi, S. P., Liang, R. P., 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*., 17(6), 715-22, 2010.
41. Kumar, C., Choudhary, A., A top-down approach to classify enzyme functional classes and sub-classes using random forest, *EURASIP Journal on Bioinformatics and Systems Biology*, (1), 1-14, 2012.
42. Volpato, V., Adelfio, A., Pollastri, G., Accurate prediction of protein enzymatic class by N-to-1 Neural Networks, *BMC Bioinformatics*, 14(Suppl 1), S11, 2013.
43. Nagao C, Nagano N, Mizuguchi K., Prediction of Detailed Enzyme Functions and Identification of Specificity Determining Residues by Random Forests, *PLoS ONE* 9(1): e84623, 2014.
44. Garg, A., Raghava, G. P.S., A Machine Learning Based Method for the Prediction of Secretory Proteins Using Amino Acid Composition, Their Order and Similarity-Search, In *Silico Biol*, 8, 129-140, 2008.
45. Lee, B. J., Shin, M. S., Oh, Y. J., Oh, H. S., Ryu, K. H., Identification of protein functions using a machine-learning approach based on sequence-derived properties, *Proteome science*, 7(1), 27, 2009.
46. Yadav, A., Jayaraman, V. K., Structure based function prediction of proteins using fragment library frequency vectors, *Bioinformation,* 8(19), 953, 2012.
47. CZ Cai, WL Wang, LZ Sun, YZ Chen, Protein function classification via support vector machine approach. Math Biosci. 185, 111–122, 2003
48. LY Han, CZ Cai, ZL Ji, ZW Cao, J Cui, YZ Chen, Predicting functional family of novel enzymes irrespective of sequence similarity. Nucleic Acids Res. 32, 6437–6444, 2004.
49. BJ Lee, HG Lee, KH Ryu, Design of a novel protein feature and enzyme function classification, in Proceedings of the 2008 IEEE 8th International Conference on Computer and Information Technology Workshops, pp. 450–455, Sydney, 2008.
50. BJ Lee, HG Lee, JY Lee, KH Ryu, Classification of enzyme function from enzyme sequence based on feature representation. IEEE Xplore. 10, 741–747, 2007.
51. Paul D. Dobson and Andrew J. Doig, "Predicting Enzyme Class from Enzyme Structure without Alignments," JMB, vol. 345, pp. 187-199, 2005.
52. UniProt Consortium. "The universal enzyme resource (UniProt)." Nucleic acids research 36.suppl 1: D190-D195, 2008.
53. Frank, Eibe, et al. "Data mining in bioinformatics using Weka." Bioinformatics 20.15, 2004.
54. Li, Ze-Rong, et al. "PROFEAT: a web server for computing structural and physicochemical features of enzymes and peptides from amino acid sequence." Nucleic Acids Research 34.suppl 2, W32-W37: 2006
55. Samb, Mouhamadou Lamine, et al. "A Novel RFE-SVM-based feature selection approach for classification." International Journal of Advanced Sciencce and Technology 43, 2012

Authors

Mr. Sanjeev Kumar Yadav is currently pursuing M.Tech. (Master of Technology) in CSE (Computer Science and Engineering) from Kashi Institute of Technology, Varanasi, India affiliated to Uttar Pradesh Technical University (UPTU), Lucknow. His research area includes machine learning computational approaches in the field of bioinformatics. He has published a paper in IJCA (International Journal of Computer Applications) and also has a conference paper in IEEE Xplorer.

Dr. Arvind Kumar Tiwari received the BE degree in CSE from CCS university, Meerut, M.Tech. in CSE from UPTU, Lucknow and Ph.D. in CSE from I.I.T. (BHU), Varanasi, India. He has worked as Professor and Vice Principal in GGS College of Modern Technology, Kharar, Punjab, India. His research interests include computational biology and pattern recognition.