International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-3)

Research Article

March
2015

# Feature based Protein Function Prediction by Using Random Forest

**Vatan Singh Azad**
Department of Computer Science and Engineering
Uttarakhand Technical University,
Dehradun, U.K., India

***Abstract-***

***In this paper we had discuss the prediction of protein function by using the Random Forest method on WEKA tool. Proteins are main building blocks of our Life. Proteins are essential parts of our life and participate in virtually every process within a cell. Protein function prediction methods are those used in bioinformatics to assign biological or biochemical roles to proteins. Here we have extracted 41 sequence-derived features of enzymes from freely available online tool. In this paper we have achieved the overall accuracy about72.5% by using Random Forest. Our Datasets have taken from PDB.***

***Keywords- enzymes function, classification, machine learning, protein sequence features.***

## I. INTRODUCTION

Proteins are main building blocks of our Life. They are responsible for catalyzing and regulating biochemical reactions, transporting molecules, and they form the basis of structures such as skin, hair, and tendon. The shape of protein is specified by its amino acid sequence. There are 20 different kinds of amino acid and each amino acid is identified by its side chain which determines the properties of amino acid. Formation of protein passes through different levels of structure. The primary structure of a protein is simply the linear arrangement, or sequence, of the amino acid residues that compose it. Secondary protein structure occurs when sequence of amino acid are linked by hydrogen bonds.

In light of the key biological role of enzymes, the Enzyme Commission (EC) of the International Union of Biochemistry and Molecular Biology (IUBMB) has created a hierarchical classification scheme based on the functional role of enzymes [3]. Each enzyme is designated an EC number. The six main classes are: (1) Oxidoreductases, (2) Transferases, (3) Hydrolases, (4) Lyases, (5) Isomerases, (6) Ligases.

There are three prominent approaches that have been widely experimented with: firstly, using sequence similarity between enzymes belonging to same functional class and secondly protein structure comparison [1] [2]. These methods have been considered inefficient since enzymes belonging to same functional class are not necessarily similar in sequence and structure [5] [4]. The third approach involves representing enzymes using sequence and structure driven features or sequence motifs that do not use similarity.

Studies that propose methods from the third category of approaches are found in [9][8][6][7]. References [9] and [8] established that Support Vector Machine (SVMs) is useful for protein function classification showing accuracy from 84% to 96%. In this work, proteins were classified into categories like RNA-binding, homodimer, drug absorption, drug delivery etc. using sequence derived features like amino acids composition, hydrophobicity, polarizability and secondary structure. Reference [12] uses features to represent subtle distinctions in local regions of sequence along with attributes as used in [9]. It applies SVM to predict the main class and report accuracy in the range of 66.02% to 90.78%. Reference [15] is a recent work that uses artificial neural networks to predict enzymes and non-enzymes. An interesting part of this work is set of 41 sequence- derived features that have been extracted from PROTPARAM and EMBOSS PEPSTST tool [14].

In this paper, we represent a new approach to predict enzyme function class using random forest. Random Forest is an ensemble classification and regression approach which is consider unsurpassable in accuracy among current data mining algorithm [11]. Random Forest algorithms have been applied extensively in prediction, probability estimation, information retrieval and until recently in bioinformatics [13]. Using an unique set of sequence features extracted with aid of online tool [14] [10] our model gives an overall accuracy of 72.5%. Feature selection and analysis of important features is also presented in this paper.

## II. DATA DESCRIPTION

The protein raw data set used in this paper is obtained from PDB. In the data set 517 protein enzymes taken from PDB are classified according to EC number and Enzyme name. We have taken 41 features extracted from PROTPARAM. It is an online tool that computes values for 36 sequence features [10]. However, it provides for feature values such as number of negatively or positively charged residues, number of carbon, hydrogen, oxygen and sulphur atoms, GRAVY, theoectical-pI and aliphatic index. The use of these features is reasoned and motivated in previous works [16] [17]. From our experiments, we find that a union of the features of ProtParam delivers better accuracy in comparison to using only one of the two feature sets. Data preparation and all manipulation have been done using Microsoft Excel. There are two tables given in TABLE 1 shows features extracted of sequences and TABLE 2 shows the proteins according to class and the total set of each class taken.

Table 1.Features description of Data sets

| Features | Description |
|---|---|
| Structure Molecular Weight | Structure molecular weight of atom |
| Residue Count | Residue count of an atom |
| Sequence | Primary structure of protein |
| Molecular Weight | Molecular weight of atom |
| Atom Count | No. of atoms |
| Amino Acid | Percentage of amino acid composition |
| Carbon | Number of carbon(C) |
| Hydrogen | Number of hydrogen(H) |
| Nitrogen | Number of nitrogen(N) |
| Oxygen | Number of oxygen(0) |
| Sulphur | Number of sulphur(S) |
| No. of Atoms | Total no. atom |
| Negcharged residue | Number of negatively charged residue |
| Pos charged residue | Number of positively charged residue |
| Theoretical Pi | Theoretical Pi |
| GRAVY | Grand Average of Hydropathicity |
| Aliphatic Index | Aliphatic index |
| Instability Index | Instability index |
| Ext. Coefficient | Extinction coefficients based on the assumption Extoe ALCy that all cysteine residues appear as half cystines |
| Classes | 6 enzyme classes |

Table2. Data description of six enzymes

| **EC.NO** | **Class(Enzymes)** | **Function** | **Total Set** |
|---|---|---|---|
| 1 | Oxidoreductases | Catalyze the reduction oxidation reactions. | 63 |
| 2 | Transferases | Transfer a functional grouping and a donor group to a receptor | 86 |
| 3 | Hydrolases | Hydrolases<br>Catalyze hydrolysis, the breaking of links and structures by the action of water. | 86 |
| 4 | Lyases | Enzymes which catalyze the cleavage of C-C, C-O and C-N links. | 93 |
| 5 | Isomerases | Catalyze the isomerization reactions of simple molecules. | 82 |
| 6 | Ligases | Formation of links by condensation of substances. | 99 |

## III. METHODOLOGY

In this paper we use WEKA tool to do our calculation [24]. WEKA, is a widely used open source in machine learning was used to carry out all the experiment [20]. We used here 10 fold cross validation in our data sets for its manipulation. Data access from the excel file and then we convert the data compatible to read in WEKA and after that we perform the calculation

### 3.1. RANDOM FOREST

Random forests are an ensemble learning method for classification (and regression) that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. Random forest is a classification algorithm developed by Leo Breiman that uses an ensemble of decision trees [11]. The term came from random decision forests that was first proposed by Tin Kam Ho ofBell Labs in 1995. The method combines Breiman's "bagging" idea and the random selection of features, introduced independently by Ho and Amit and Geman in order to construct a collection of decision trees with controlled variance [25]. Each tree is constructed by a bootstrap sample from the data, and at every split it uses a candidate set of features selected from a random set. Thus, random forest uses both bagging and random variable selection for tree building. Once the forest is formed, test instances are percolated down each tree and trees make their respective class prediction. Random Forest has several characteristics that make it well suited for enzyme function classification: a) It runs efficiently on large data sets with many features and does not require for data to be normalized. b) It can handle missing values. c) Because many trees are built and each tree is effectively an independent model, the model builder tends not to over-fit to the training dataset. d) Incorporates interactions among predictor variables. The error rate of a random forest depends on the correlation between any two trees and the strength of each tree. Random Forest can be used to rank the importance of variables in a regression or classification problem in a natural way.

We assume that the user knows about the construction of single classification trees. Random Forest grows many classification trees. To classify a new object from an input vector, put the input vector down each of the tree in the forest. Each tree gives a classification, and we say the tree “votes” for that class. The forest chooses the classification having the most votes (over all the trees in the forest). Forest error rate depends on two things: a) the correlation between any two trees in the forest. Increasing the correlation increases the error rate. b) the strength of each individual tree in the forest. A tree with a low error rate is strong classifier. Increasing the strength of the individual trees decrease the forest error rate [18].

## IV. RESULT and DISCUSSION

In this paper we use 10 fold cross validation method to measure the performance of the RANDOM FOREST. In comparison to other, RANDOM FOREST demonstrates better accuracy result despite a larger data set and wider distribution of test data instances. Below table shows the result for a 10 fold cross validation experiment with Random Forest.

Table 3. 10 fold cross validation results.

| CLASS | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC-Area |
|---|---|---|---|---|---|---|
| 1 | 0.641 | 0.015 | 0.854 | 0.641 | 0.732 | 0.955 |
| 2 | 0.92 | 0.026 | 0.879 | 0.92 | 0.899 | 0.979 |
| 3 | 0.828 | 0.028 | 0.857 | 0.828 | 0.842 | 0.963 |
| 4 | 0.684 | 0.093 | 0.625 | 0.684 | 0.653 | 0.914 |
| 5 | 0.602 | 0.072 | 0.617 | .602 | 0.61 | 0.899 |
| 6 | 0.66 | 0.101 | 0.611 | 0.66 | 0.635 | 0.898 |
| Weighted AVERAGE | 0.725 | 0.059 | 0.731 | 0.725 | 0.726 | 0.933 |

Here we observed that the overall accuracy of Random Forest is 72.5%. The overall Precision and Recall values are 73.1% and 72.5% respectively. Here TP represents true positive rate and FP represent false positive rate.

## V. CONCLUSION

In this paper, we have presented an alternative approach to represent enzyme protein sequences using 41 different features. Random Forest correctly achieved an overall accuracy of around 72.5% on a widely distributed and reasonably large datasets. In future the performance can be increased by integrating more than one classifier together and by integrating multiple data sets together. So in future we can increase the overall accuracy also.

## REFERENCES

[1] Shah and L. Hunter, "Predicting enzyme function from sequence: a systematic appraisal," In the proceeding of ISMB, pp. 276-283, 1999.

[2] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller et al., "Gapped BLAST and PSIBLAST: a new generation of protein database search programs," Nucleic Acids Research, vol. 35, pp. 3389-3402, 1997.

[3] Enzyme-Nomenclature, Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB), Academic Press, NY (http:Hwww.chem.qmul.ac.uk/iubmb/enzyme/).

[4] Umar Syed and Golan Yona, "Enzyme function prediction with interpretable models," Computational Systems Biology, pp. 1-33, 2007.

[5] Xiangyun Wang, Diane Schroeder, Drena Dobbs, and Vasant Honavar, "Automated data-driven discovery of motif-based protein function classifiers," Information Sciences, vol. 155, pp. 1-18, 2003.

[6] B.J. Lee, H. G. Lee, J. Y. Lee, K. H. Ryu, “Classification of Enzyme Function from Protein Sequence based on Feature Representation,” IEEE Xplore, vol. 10, pp. 741-747, October 2007.

[7] U. Syed and G. Yona, "Using a mixture of probabilistic decision trees for direct prediction of protein function," RECOMB, pp. 289-300, 2003.

[8] L.Y. Han, C.Z. Cai, Z.L. Ji, Z.W. Cao, J. Cui, and Y.Z. Chen,"Predicting functional family of novel enzymes irrespective of sequence similarity," Nucleic Acids Research, vol. 32, pp. 6437-6444, 2004.

[9] C.Z. Cai, W.L. Wang, L.Z. Sun, and Y.Z. Chen, "Protein function classification via support vector machine approach," Mathematical Biosciences, vol. 185, pp. 111-122, 2003.

[10] E. Gasteiger, C. Hoogland, et al., Protein Identification and Analysis Tools on the ExPASy Server, John M. Walker (ed): The Proteomics Protocols Handbook, pp. 571-607, Humana Press, 2005.

[11] L. Breiman, “Random Forests”, Machine Learning, vol. 45, pp. 5-32, 2001.

[12] B.J. Lee, H. G. Lee, K. H. Ryu, “Design of a Novel Protein Feature and Enzyme Function Classification,” Proceedings of the 2008 IEEE 8th International Conference on Computer and Information Technology Workshops, vol. 00, pp. 450-455, 2008.

[13] J. Zhang, M. Zulkernine, "A Hybrid Network Intrusion Detection Technique Using Random Forests," ares,pp.262-269, First International Conference on Availability, Reliability and Security (ARES'06), 2006

[14] P. Rice, I. Longden, A. Bleasby, “Emboss: the European Molecular Biology Open Software Suite,” Trends in Genetics, vol. 16, pp. 276-282, June 2000.

[15] ] P.K. Naik, V.S. Mishra, M. Gupta, K. Jaiswal, “Prediction of enzymes and non-enzymes from protein sequences based on sequence derived features and PSSM matrix using artifical neural network,” Bioinformation, vol. 2, pp. 107-112, December 2007.

[16] P. D. Dobson and A. J. Doig, “Predicting Enzyme Class from Protein Structure without Alignments,” JMB, vol. 345, pp. 187-199, 2005.

[17] Lars Juhl Jensen, Marie Skovgaard and Soren Brunak, “Prediction of novel archael enzymes from sequence-derived features”, Protein Science, vol. 3, pp. 2894-2898, 2002.

[18] http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.html

[19] V. N. Vapnik, “Statistical Leaning Theory,” Wiley-Interscience, New York 1998.

[20] Ian H., Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005, (http://www.cs.waikato.ac.nz/ml/weka/).

[21] Ross Quinlan, "C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[22] Oskar Markovic and Stefan Janecek, “Pectin degrading glycoside hydrolases of family 28: sequence-structural features, specifities and evolution,” vol. 14, no. 9, pp. 615-631, September 2001.

[23] Graham Williams, Rattle: A graphical user interface for data mining in R using GTK. R package version 2.4.10, (2008) http://rattle.togaware.com/

[24] www.cs.waikato.ac.nz/ml/weka/

[25] http://www.whrc.org/education/indonesia/pdf/DecisionTrees_RandomForest_v2.pdf