ISSN 2047-3338

Machine Learning Classifiers for Human Protein Function Prediction

Manpreet Singh, Dr. Gurvinder Singh and Dr. Karanjeet Singh Kahlon

Abstract—**Protein Function Prediction is important for its numerous applications in the field of drug discovery, gene ontology and its role in the complex structure formation. HPRD (Human Protein Reference Database) database contains the protein sequences for different protein classes. The sequences are accessed from HPRD and then different online tools are used to extract SDFs (Sequence Derived Features) for each protein sequence. In the present paper, two types of databases are considered: one containing the continuous values of SDFs and other containing the discrete values having range of SDFs for particular sequence. The paper summarizes the different decision tree methodologies applied for the problem of Human Protein Function Prediction. The accuracies for different approaches are presented for the specific databases. A new approach based on C5 decision tree is also presented and the results are compared with the existing approaches.**

Index Terms— **Machine Learning, Decision Tree, HPRD and Protein Function Prediction**

I. Machine Learning

MACHINE learning is a way of automatically improving, of using “training” data to build or alter a model which can later be used to make predictions for new unseen data.

Machine learning algorithms can be broadly classified into two categories: Supervised and Unsupervised. Supervised algorithm is first trained from the given set of data whose classes (outcomes) are already known. Based on the training, the algorithm builds the profile of classes which are used to predict the classes of unknown data. Decision trees and neural networks fall under this category. Unsupervised algorithms do not undergo training, and are usually used when the classes are not known in advance. Clustering is an example of unsupervised classification. In this case, data is clustered together based on their similarity and groups are formed which act as classes. Unknown data is classified by assignment to the closest matching cluster, and is assumed to have characteristics similar to the other data in the cluster [1].

II. Proteins and Protein Function

Proteins are the main building blocks of our body and the all the living beings on earth. They are necessary for the structure, function, and regulation of the body’s tissues and organs. Proteins are constituted of multiple smaller units called amino acids. There are 20 different types of amino acids that are attached together in a form of long chain called protein.

Based on the different molecular functions, proteins can be divided into various classes. Protein class prediction plays important part in the drug discovery process because proteins are the common drug targets. The drug discovery process is a complex and labor intensive problem and hence it is time consuming and expensive. Bioinformatics promises to reduce the labour and time associated with this process, allowing drugs to be developed faster and at a lower cost [2].

III. Background

In the recent years, numerous techniques have been developed for protein function prediction. The quality of function prediction models developed so far depends upon the accuracy of the prediction model. The support vector machines and neural network are considered as black box as the intermediate computational results are invisible [3]-[6]. Black box model do not provide any biological significance of the genomic process [7]. However, black box model can be transformed into white box model [8]-[11] and these techniques are yet to be explored for function prediction models. On the other hands, white box models like decision trees and rule sets are widely used in this field [12]-[15]. HPF prediction can be done based on the similarity score of the unknown sequence and the existing database. The class of the sequence giving the highest score with the unknown sequence can be designated to the input sequence. But the potentially similar sequence can’t guarantee the similar function [16]-[19]. Due to these reasons, HPF prediction based on SDFs is preferred. The present paper discusses about the existing decision tree methodologies and also introduces a new approach based on C5 algorithm for HPF prediction.

IV. Data Source

The amino acid sequences are accessed from Human

Manpreet Singh is with Guru Nanak Dev Engineering College, Ludhiana, Punjab, INDIA (phone: +91-161-2490339; e-mail: mpreet78@ yahoo.com).

Dr. Gurvinder Singh is with Guru Nanak Dev University, Amritsar, Punjab, INDIA (e-mail: gsbawa71@yahoo.com).

Dr. Karanjeet Singh Kahlon is with Guru Nanak Dev University, Amritsar, Punjab, INDIA (e-mail: karanvkahlon@yahoo.com).

Protein Reference Database (HPRD) [20]. The HPRD provides the information regarding the domain architecture, post-translational modifications, interaction networks and disease association for each protein in the human proteome. It includes approximately 163 classes of protein functions. The database provides information about protein function under the heading 'molecular class' covering all the major protein function categories [21].

Some of the HPRD classes are: Defensin (Def), Heat Shock Protein (HSP), Voltage Gated Channel (VGC), Cell Surface Receptor (CSR), DNA Repair Protein (DRP), Aminopeptidase (Ami), Decarboxylase (Dec), G-Protein (GP), RNA Binding Protein (RBP) and Transport/Cargo Protein (T/CP).

V. Sequence Derived Features

Sequence derived features (SDFs) are various parameters based on the amino acid sequence which are used to predict human protein function (HPF). Sequence derived features are very important in protein prediction as these are the input to the HPF predictor as labeled vector. The different sequence derived features are: Number of amino acids, Molecular weight, pI, Number of negative ions, Number of positive ions, Extinction coefficients 1 and 2, Instability index, Aliphatic index, Gravy, T, S, Ser, Thr, Tyr, Mean, D, Probability, ExpAA, Number of helices(PredHel) and ProbN.

Following are the various bioinformatics Tools for obtaining sequence derived features (SDFs):

NetNGlyc server is used for the prediction of N-Glycosylation sites in human proteins using artificial neural networks.

PSORT server predicts of protein localization sites in cells. It analyzes the input sequence by applying the stored rules for various sequence features of known protein sorting signals. Finally, it reports the possibility for the input protein to be localized at each candidate site with additional information.

TMHMM server is a program for predicting transmembrane helices based on a hidden Markov model.

NetOGlyc server predicts the O-GalNAc (mucin type) glycosylation sites in mammalian proteins.

Signal-P server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks and hidden Markov models.

ExPASy ProtParam server computes various physico-chemical properties of protein like isoelectric point, extinction coefficient [21]-[24].

VI. Machine Learning Classifiers For HPF

In the following section, the different decision tree methodologies are discussed used for HPF prediction.

A. C4.5 Decision Tree

C4.5 was proposed by Quinlan (1993) which uses gain ratio as its splitting criteria. C 4.5 can handle pruning, missing value and numeric values thus is known as successor of ID3 algorithm. It is used to find missing gain values using corrected gain ratio from a given training set. This also follows greedy approach in finding the best attribute. The Decision Tree C4.5 involves entropy calculation. Entropy is the expected information based on the partitioning into subsets by an attribute. The smaller the entropy value, the greater is the purity of the subset partitions. The information gain measure is used to select the test attribute at each node in the tree [25]. The attribute with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node.

B. Decision Tree based on Uncertainty Measure

The new prediction technique based on uncertainty measure was developed which gives the greater depth in tree as compared to C4.5. The technique does not encode the information in terms of bits. The attribute with the least uncertainty measure is chosen as the best attribute by this prediction technique during decision tree creation [21].

C. See5

Quinlan developed See5 tool based on C5 algorithm which mainly emphasizes on rule-based classifiers in which each rule can be separately examined and validated, without having to consider it as a whole [26].

D. C5 Algorithm

C5 uses the concept of maximum gain to find out the best attribute which is considered as root in making of decision tree. The best attribute at a node is the attribute having maximum Gain among all the possible attributes that can be used at the node. Thus this attribute is considered as the root node of decision tree and rest nodes are been calculated using same criteria and hence we finally reach to a decision tree which is simple and small in size and occupies less memory space.

An C5 algorithm was developed considering the discrete values of sequence derived features. The SDF values are in the form of ranges. The range of each SDF is fixed manually by comprehensively studying the variation. The advantage of discrete data is that it reduces the overall complexity of the database. Some SDF values are having minor variations as compared to the other values which vary at the extreme end for the sequences in the different classes. Thus the significant variation can be easily recognized and the minute variation can be neglected by choosing the database with discrete values. The algorithm is shown in Figure 1 and the splitting criterion followed is shown in Figure 2.

VII. RESULTS

For 25 molecular sequences and 21 SDFs the tree obtained by using C5 algorithm is shown in Figure 3.

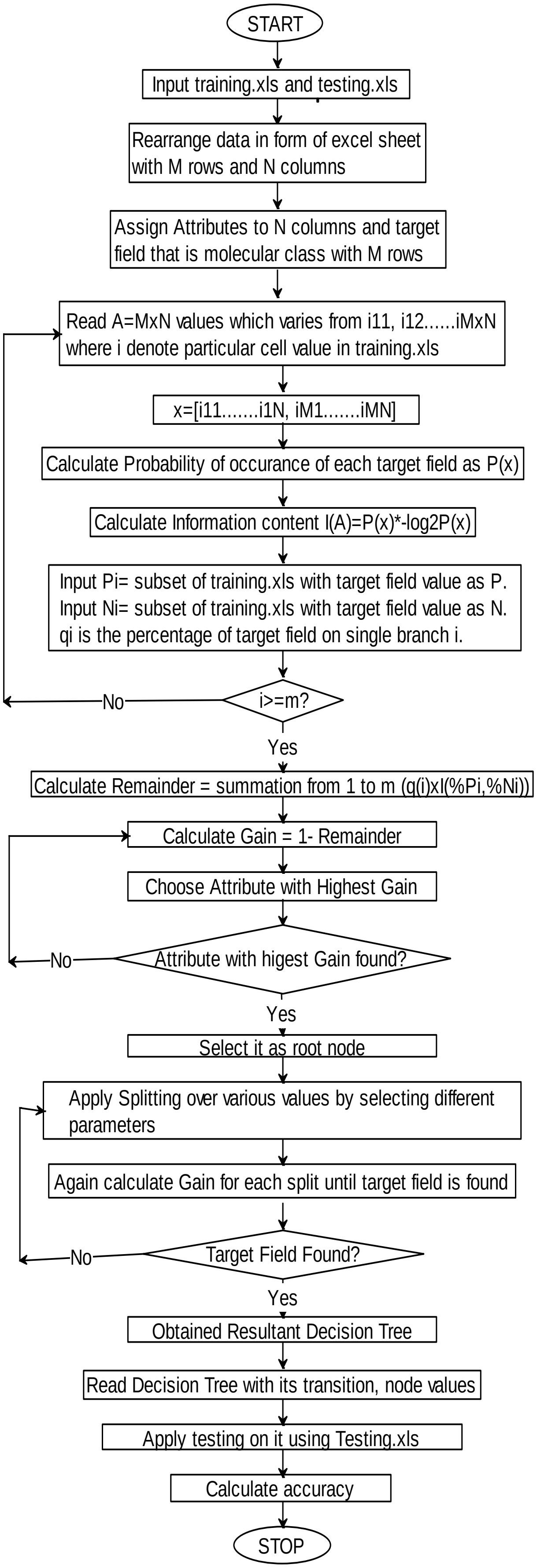
START
Input training.xls and testing.xls
Rearrange data in form of excel sheet with M rows and N columns
Assign Attributes to N columns and target field that is molecular class with M rows
Read A=MxN values which varies from i11, i12......iMxN where i denote particular cell value in training.xls
x=[i11.......i1N, iM1.......iMN]
Calculate Probability of occurance of each target field as P(x)
Calculate Information content I(A)=P(x)*-log2P(x)
Input Pi= subset of training.xls with target field value as P.
Input Ni= subset of training.xls with target field value as N.
qi is the percentage of target field on single branch i.
i>=m?
No
Yes
Calculate Remainder = summation from 1 to m (q(i)xI(%Pi,%Ni))
Calculate Gain = 1- Remainder
Choose Attribute with Highest Gain
Attribute with higest Gain found?
No
Yes
Select it as root node
Apply Splitting over various values by selecting different parameters
Again calculate Gain for each split until target field is found
Target Field Found?
No
Yes
Obtained Resultant Decision Tree
Read Decision Tree with its transition, node values
Apply testing on it using Testing.xls
Calculate accuracy
STOP

Fig. 1. C5 Algorithm

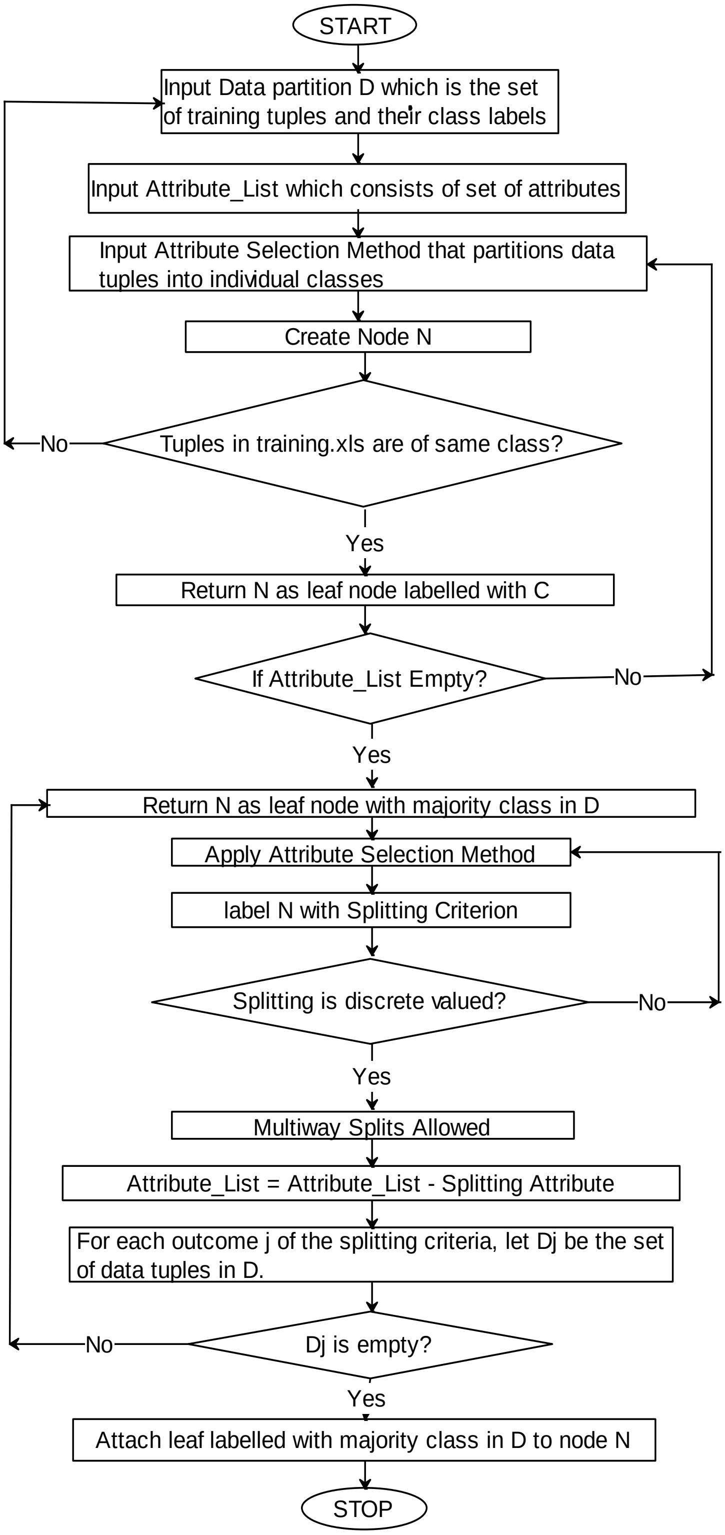
START
Input Data partition D which is the set of training tuples and their class labels
Input Attribute_List which consists of set of attributes
Input Attribute Selection Method that partitions data tuples into individual classes
Create Node N
Tuples in training.xls are of same class?
No
Yes
Return N as leaf node labelled with C
If Attribute_List Empty?
No
Yes
Return N as leaf node with majority class in D
Apply Attribute Selection Method
label N with Splitting Criterion
Splitting is discrete valued?
No
Yes
Multiway Splits Allowed
Attribute_List = Attribute_List - Splitting Attribute
For each outcome j of the splitting criteria, let Dj be the set of data tuples in D.
Dj is empty?
No
Yes
Attach leaf labelled with majority class in D to node N
STOP

Fig. 2. Splitting Criteria in C5

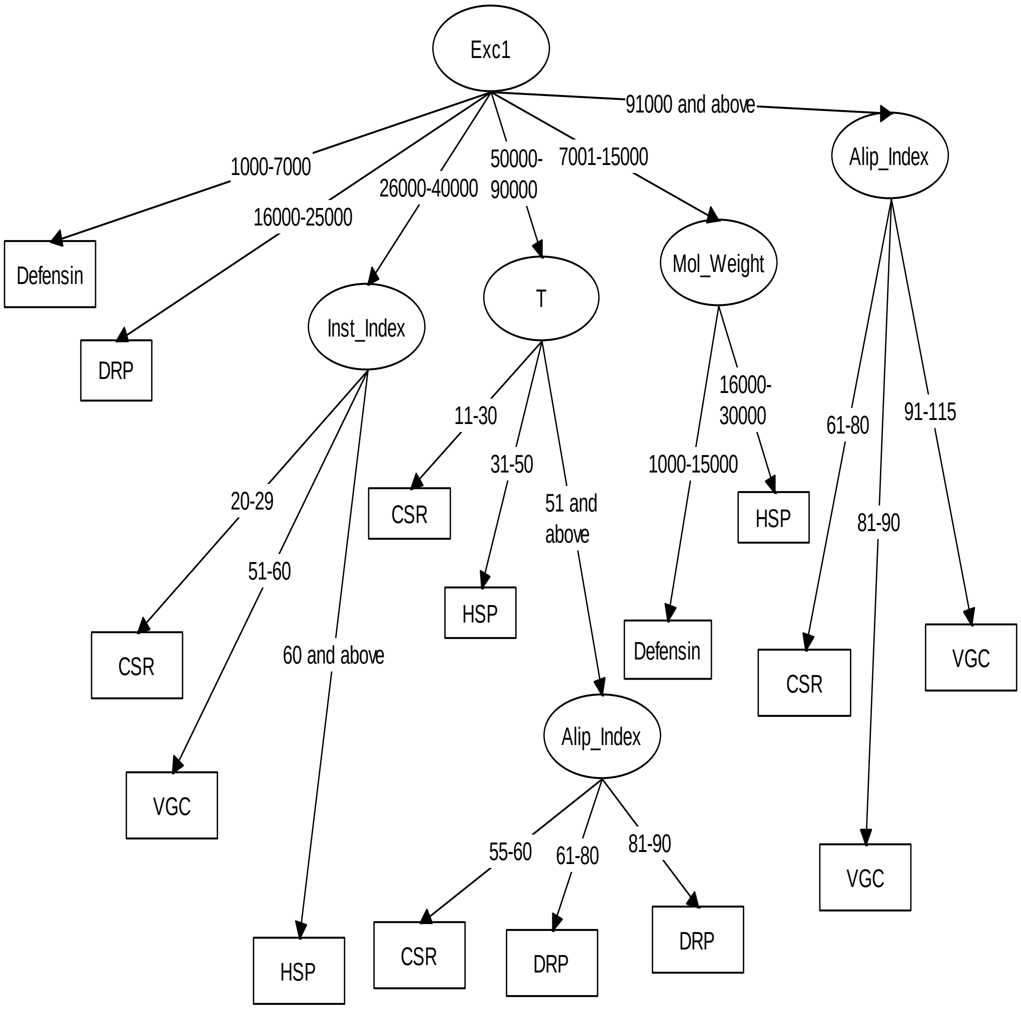

Fig. 3. C5 Decision Tree for HPF Prediction

VIII. Conclusion

The See5 tool is based on the continuous data values. The decision tree was constructed based on training data of 55 sequences and test data of 29 sequences. Continuous data involving 21 SDFs was given as input. Different advanced options and combinations are tried out of which decision tree powered with boosting and winnowing give the maximum accuracy of 30% for the data under consideration. If the continuous data set for 25 sequences is taken then the accuracy comes out to be 64% with the same technique [27]. If the discrete values are taken to reduce the complexitity of data set, C4.5 algorithm gives the accuracy of 44% and algorithm based on uncertainty measure gives the accuracy of 72% with 25 sequence data having the 17 SDFs [21]. In the same scenario, C5 algorithm gives the accuracy of 83% with 25 SDFs.

References

[1] A. Clare, "Machine learning and data mining for yeast functional genomics", Ph.D. Thesis, Department of Computer Science University of Wales, 2003.

[2] D. Krane and M. Raymer, *Fundamental Concepts of Bioinformatics*, Pearson Education, 2006, pp 155-178.

[3] L.J. Jensen, R. Gupta, H.H. Staerfeldt, and S. Brunak, “Prediction of Human Protein Function According to Gene Ontology Categories”, *Bioinformatics*, vol. 19, no. 5, 2003, pp. 635-642.

[4] K. Tu, H. Yu, Z. Guo, and X. Li, “Learnability-Based Further Prediction of Gene Functions in Gene Ontology”, *Genomics*, vol. 84, 2004, pp. 922-928.

[5] A. Vinayagam, R. Konig, J. Moormann, F. Schubert, R. Eils, K.-H. Glatting, and S. Suhai, “Applying Support Vector Machines for Gene Ontology Based Gene Function Prediction”, *BMC Bioinformatics*, vol. 5, no. 116, 2004.

[6] W.R. Weinert and H.S. Lopes, “Neural Networks for Protein Classification”, *Applied Bioinformatics*, vol. 3, no. 1, 2004, pp. 41-48.

[7] B. Mirkin and O. Ritter, "A Feature-Based Approach to Discrimination and Prediction of Protein Folding Groups," *Genomics and Proteomics: Functional and Computational Aspects*, Kluwer Academic/Plenum Publishers, 2000, pp. 157-177.

[8] H. Jacobson, "Rule Extraction from Recurrent Neural Networks:A Taxonomy and Review", *Neural Computation*, vol. 17, 2005,pp. 1223-1263.

[9] A.B. Tickle, R. Andrews, M. Golea, and J. Diederich, "The Truth Will Come to Light: Directions and Challenges in Extracting Knowledge Embedded within Trained Artificial Neural Networks", *IEEE Trans. Neural Networks*, vol. 9, no. 6, 1998, pp. 1057-1068.

[10] G. Fung, S. Sandilya, and R.B. Rao, "Rule Extraction from Linear Support Vector Machines", *Proc. ACM SIGKDD '05*, 2005, pp. 32-40.

[11] H. Nunez, C. Angulo, and A. Catala, "Rule Extraction from Support Vector Machines", *Proc. European Symp. Artificial Neural Networks (ESANN '02)*, 2002, pp. 107-202.

[12] A. Clare, A. Karwath, H. Ougham, and R.D. King, "Functional Bioinformatics for Arabidopsis thaliana", *Bioinformatics*, vol. 22, no. 9, pp. 1130-1136, 2006.

[13] J. He, H.-J. Hu, R. Harrison, P.C. Tai, and Y. Pan, "Transmembrane Segments Prediction and Understanding Using Support Vector Machine and Decision Tree", *Expert Systems with Applications*, vol. 30, 2006, pp. 64-72.

[14] G.L. Pappa, A.J. Baines, and A.A. Freitas, "Predicting Post-Synaptic Activity in Proteins with Data Mining", *Bioinformatics*, vol. 21, no. Suppl. 2, 2005, pp. ii19-ii25.

[15] M. Singh, P.S. Sandhu, H. Singh "Decision Tree Classifier for Human Protein Function Prediction", *Proceedings of International Conference on Advanced Computing and Communications, ADCOM 2006*, 20-23 Dec., 2006, pp. 564-568.

[16] I. Friedberg, "Automated Protein Function Prediction—The Genomic Challenge", *Briefings in Bioinformatics*, vol. 7, no. 3, 2006, pp. 225-242.

[17] J.A. Gerlt and P.C. Babbitt, "Can Sequence Determine Function," *Genome Biology*, vol. 1, no. 5, 2000.

[18] U. Syed and G. Yona, "Using a Mixture of Probabilistic Decision Trees for Direct Prediction of Protein Function", *Proceedings of Seventh Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, 2003.

[19] D. Szafron, P. Lu, R. Greiner, D.S. Wishart, B. Poulin, R. Eisner, Z. Lu, B. Poulin, R. Eisner, J. Anvik, and C. Macdonell, "Proteome Analyst—Transparent High-Throughput Protein Annotation:Function, Localization and Custom Predictors", *Proceedings of ICML Workshop Bioinformatics*, 2003.

[20] Human Protein Reference Database (HPRD) http://www.hprd.org/moleculeClass

[21] M. Singh, P. Singh and P.K. Wadhwa "Human Protein Function Prediction using Decision Tree Induction", *International Journal of Computer Science and Network Security*, Vol. 7, No. 4, 2007, pp. 92-98.

[22] L. Jensen, "Prediction of Protein Function from Sequence Derived Protein Features", Ph.D. thesis, Technical University of Denmark, 2002.

[23] L. Jensen, M. Skovgaard and S. Brunak "Prediction of Novel Archaeal Enzymes from Sequence Derived Features", Protein Science, Vol. 11, 2002, pp. 2894-2898.

[24] L.J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, C. Kesmir, H. Nielsen, H.H. Staerfeldt, K. Rapacki, C. Workman, C.A.F. Andersen, S. Knudsen, A. Krogh, A. Valencia and S. Brunak "Prediction of Human Protein Function from Post-Translational Modifications and Localization Features", *Journal of Molecular Biology*, Vol. 319(5), 2002 pp. 1257-1265.

[25] J. Han, and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2004.

[26] http://rulequest.com/see5-info.html.

[27] M. Singh, G. Singh and S. Sharma, Human Protein Function Prediction from Sequence Derived Features using See5, *IJSER*, 3(7) (2012), July-2012.