ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Special Issue 7, October 2015**

# Protein Function Prediction Using Machine Learning Techniques

Sinchana H K[1], Hemalatha N[2]

M.Sc(ST), AIMIT, St. Aloysius College (Autonomous), Mangalore, Karnataka, India[1]

Assistant Professor, AIMIT, St. Aloysius College (Autonomous), Mangalore, Karnataka, India[2]

**ABSTRACT:** Machine learning is a field of computer science that has evolved from the study of computational learning and recognition of pattern theory in artificial intelligence. There are various techniques used in machine learning, like Support Vector Machines, Artificial Neural Network, Decision Tree, Cluster Analysis, Bayesian Network, Random Forest, Hierarchical clustering etc. These are the techniques used in bioinformatics to solve different type of problems. Protein function prediction is an important and challenging field in Bioinformatics. Protein function prediction is techniques that are used by researchers to assign biological roles to proteins. Proteins are categorized on the basis of amino acid patterns. In this work we propose to study the papers which have used Machine learning Techniques to predict the function of a protein using different feature extraction methods.

## I. INTRODUCTION

Machine learning is a process that enables the computers to learn from experience, learn by example, and learn by analogy. Machine learning focuses on prediction, based on known properties learned from the training data. Machine learning means different things to different people, and there is no set of algorithms that must be learned upon the agreement. Machine Learning techniques are suitable for applications in bioinformatics because the subjects can be easily adapted to a new environment. Machine learning is about learning to do better in the future based on what was experienced in the past. Machine learning is a core subarea of artificial intelligence. It focuses on prediction, based on properties which is known and learned from data set which is trained. Techniques used for machine learning are Support Vector Machine, K Nearest Neighbor, Decision Tree, and Association Learning.

Proteins are formed from a group of 20 amino acids and the function of a protein is closely related to the structure. Based on functions there are different proteins which help in catalysis, transport and information [1]. As compared to biological experiments, using the computational approaches are much cheaper and cost-effective in protein function predictions. Protein function plays a main role in the understanding of the comprehension of living organisms in complex machinery. Based on sequence similarity there are many developed methods for protein function prediction.

Assigning of biological or biochemical roles of proteins are called protein function prediction [2]. Proteins molecular function is referred as protein function, such as: gene regulation, Transport of materials, catalysis of biochemical reactions (enzymes), among others. A common evolutionary origin between sequences can be identified by searching sequence similarity on new/query protein and non-protein.
In this paper an attempt has been made to review different papers on proteins functions that are predicted using different Machine learning techniques.

## II. LITERATURE SURVEY

Juliana S *et-al.*, in their paper proposed that protein function prediction was one of the most challenging problems in the post-genomic era [3]. The number of newly identified proteins has been exponentially increasing with the advances of the high-throughput techniques. The functional characterization of new proteins was not in the same proportion. To fill that gap, in the literature many number of computational methods has proposed. To the newly discovered proteins, the known functions have explored to homology relationships from early approaches. When new protein was different from old one then approaches tend to fail. In order to get more relevant methods appropriate computational program is

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Special Issue 7, October 2015**

required. Regarding those points, that review provides a comprehensible description of machine learning approaches that are currently applied to protein function prediction problems. They started by defining several problems enrolled in understanding the aspects of the protein function and it explains how machine learning is implemented to problems. Sometimes difficulty occurred due to vigorous evolvement and to the protein function inference, the systematical framework was exposed. Machine learning methods in functional proteomics has to be provided for classification with recent attachments and highlighted most representative contributions.

Huang *et-al.*, in their paper proposed a novel scoring card method to estimate solubility scores of dipeptides and amino acid residues for predicting solubility of proteins and analyzing the tendency of physicochemical properties [4]. The proposed method problems have dipeptide composition features and protein function problems were easily adapted to dipeptide propensities. The scoring card method with solubility scoring matrix performed well in predicting solubility, compared with existing methods using complementary features that associated well with solubility. The results approving with the literature reports reveal that the solubility scoring matrices are effective. Since the proposed scoring card method is effective for generating solubility scoring matrix to predict protein solubility, their future work was where amino acid and dipeptide plays important role and to investigate protein function prediction to generate different kinds of scoring matrices.

Xiong *et-al.*, in their paper came up with a new method which combines to increase the performance of protein-protein interaction based on collective classification [5]. According to their method, protein function prediction is divided into two phases: First, by adding number of edges that are from the information of protein sequence is enriched from protein-protein interaction. The added edges are called as implicit edges, and the existing ones are called explicit edges. Second, a collective classification algorithm is employed on the new network to predict protein function. Their key idea is to improve the performance of prediction by adding number of edges that are computed by PPI network to enrich protein information interaction. They conducted extensive experiments on two real, publicly available protein-protein interaction datasets. Compared to four existing protein function prediction approaches, their method performed better in many situations, which shows that adding implicit edges can indeed improve the prediction performance. The experimental results demonstrate that their method outperforms the existing approaches does not work well in sparsely-labeled networks due to protein-protein interaction information inadequacy. Experimental results also validate the robustness to the labeled proteins for the approach in protein-protein interaction networks.

Qingyao *et-al.*, have proposed that to solve the problem in classifications from protein-protein networks for interrelated proteins, Markov chain based Collective Classification algorithm was used. To intensify the performance of protein-protein interaction network data, the algorithm will focus how to use label and unlabeled data. They aimed the classifiers to make separate predictions, using two distinct Markov chain classifiers with regard to relational features from relational information and attribute features from protein data [6]. The algorithm combines the results of both the classifiers to compute the ranks of labels to indicate the instance to labelled groups which was repeatedly used by ICA framework for the performance in labelled data of protein function prediction from protein-protein interaction networks by machine learning models. In the limited labeled data, protein-protein interactions showed that ICAN is better than ICA methods from given data. For the purpose of studying, the protein function prediction from protein-protein interaction networks is the valuable tool. For collective classifications the other techniques of semi-supervised learning for protein-protein interaction network, they will consider it in future study.

Amit Bhola *et-al.*, proposed that protein function prediction is an important and challenging field in Bioinformatics. Based on many machine learning techniques, by using the sequence derived properties it has been proposed to the functions to predict the protein [7]. In their paper 857 many types of machine learning techniques are used to derive the features such as correlation, transaction, composition, amino acid composition. To select the approximate number of features their paper used various techniques like Gain Ratio, One R attribute, ReliefF The comparative analysis of result shows that the random forest based method with reliefF provide the overall accuracy of 89.20% and Matthews's correlation coefficient (MCC) 0.87% that is better to others.

## III. MACHINE LEARNING TECHNIQUES

### A. Support Vector Machine

Support vector machines are supervised learning models which has learning algorithms associated that can analyze data and recognize the patterns, used for classification and regression analysis. Given a set of training examples, each

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Special Issue 7, October 2015**

marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary classifier. Based on separation of different class labels, SVM constructs the hyper planes by performing classification. SVM can handle many categorical and continuous variables and supports classification and regression. Maximum separation of hyper plane is constructed and higher-dimensional space can be mapped into input samples. SVM assumes that better accuracy prediction is implied when there is separation between the classes. The largest margin is determined by a set of observations that are the most difficult classifying training points, so-called support vectors [8].

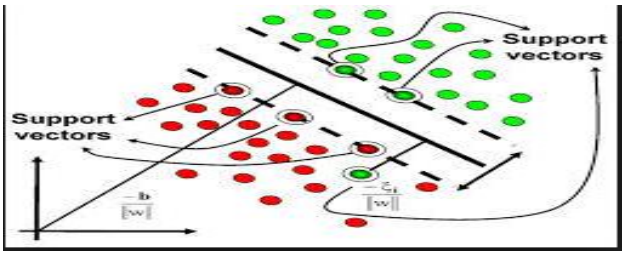
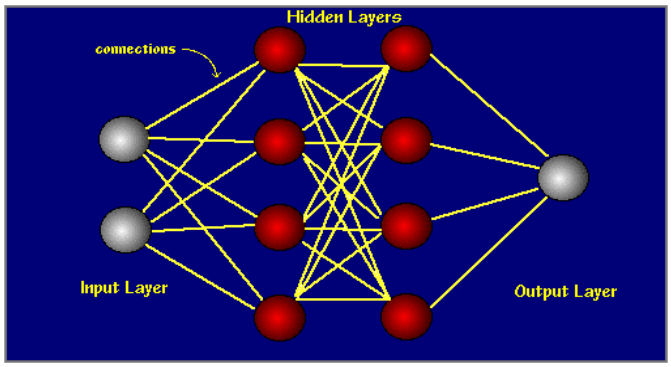



Fig 1:Support Vector

B. Artificial Neural Network

Artificial neural network (ANN) is artificial neurons where it is interconnected to each other. The general function is when a particular input pattern is given neural network produces output pattern. Similar to protein folding process, ANN is suitable for empirical approach to protein function prediction.




C. Self-Organizing Maps

SOM is unsupervised learning and its goal is develop system with classification rules. SOM learning assumes that data classification is related to data structure or topological structure. SOM explores output of classification rules. There are common reserved motifs in a sequence with similar functions and structure within a protein. Therefore unsupervised learning approaches can be used to discover the patterns constructed by motifs.

D. *K*-Nearest Neighbour

K-Nearest Neighbor is used for regression and classification and it is also a non-parametric method. The input consists of examples of k closest training and output depends whether regression and classification is used for k-NN. It is the simplest of the machine learning algorithms. KNN can be regarded as one of the most important factors of the model that can strongly influence the quality of predictions. k-Nearest Neighbors (KNN) is a memory-based model defined by a set of objects known as examples. Each example consists of a data case having a set of independent values labeled by a set of dependent outcomes.

## IV. CONCLUSION AND FUTURE SCOPE

In this paper we had studied many papers which had worked on protein function prediction. Also have studied different machine learning techniques which are used for creating the models for function prediction. In our future work we

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Special Issue 7, October 2015**

propose to use these machine learning techniques to predict toxicity in plants and select the one with best accuracy for creating a prediction tool.

## REFERENCES

[1] Altschul S, Gish W, Miller W, Myers E and Lipman. D, 'Basic Local Alignment Search Tool', Journal Molecular Biology, Vol. 215, pp. 403-410, 1990

[2] Pearson W, 'Rapid and sensitive sequence comparison with FASTP and FASTA Methods', Enzymol, Vol 183, pp. 63-98, 1985

[3] Boser B, Guyon I and Vapnik V, ' A training algorithm for optimal margin classifiers', In: Proceedings of the fifth annual ACM workshop on Computational learning theory, ACM Press, pp. 144-152, 1992.

[4] S Bernardes and Juliana., 'A review of protein function prediction under machine learning perspective', Recent patents on biotechnology Vol. 7, Issue 2, pp 122-141,2013.

[5] Huang, Hui-Ling, et al., 'Prediction and analysis of protein solubility using a novel scoring card method with dipeptide composition.',BMC Bioinformatics, Vol 13, Issue 17, Suppl 3 ,2012.

[6] Wu, Qingyao, 'Collective prediction of protein functions from protein-protein interaction networks.', BMC bioinformatics, Vol. 15, Issue 2, Suppl 9, 2014.

[7] Bhola, Amit, Sanjeev Kumar Yadav, and Arvind Kumar Tiwari, 'Machine Learning based Approach for protein Function Prediction using Sequence Derived Properties.', International Journal of Computer Applications, Vol 105 pp 12 ,2014.

[8] Lan, Liang, et al., 'MS-kNN: protein function prediction by integrating multiple data sources', BMC bioinformatics, Vol. 14, Issue 3, Suppl 8,2013.