International Journal for Modern Trends in Science and Technology, 6(11): 63-66, 2020

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST061111
Available online at: http://www.ijmtst.com/vol6issue11.html

Function Prediction of Human Proteins

Prashansa Roy[1] | Bhavesh Tanawala[1]| Hetal Gaudani[2]

[1]Computer Engineering, Birla vishvakarma Mahavidyalaya
[2]Computer Engineering, G H Patel Collage of Engineering & Technology

To Cite this Article

Prashansa Roy, Bhavesh Tanawala and Hetal Gaudani, "Function Prediction of Human Proteins", *International Journal for Modern Trends in Science and Technology*, 6(11): 63-66, 2020.

Article Info

Received on 13-October-2020, Revised on 04-November-2020, Accepted on 07-November-2020, Published on 10-November-2020.

ABSTRACT

Due to the rapidness in research, accumulation of biological data is happening at an overwhelming rate. Advanced computation techniques are required to gather the useful information from this enormous amount of protein data such that the knowledge is practically useful and easily interpretable. For instance, drug discoverers need biological or computational methods to predict the functions of proteins, responsible for different sort of diseases in human body. Since traditional biological methods were time consuming and comparatively expensive, various computational methods have been introduced in the respective research areas. In this project, we have tried to generate machine learning models that predict the protein function of unknown proteins and analyze their performance to get a model with highest accuracy. Protein function's sequence annotations such as Amino Acid modifications, Molecule Processing and other structural features like Active Site, Beta strand, Chain, etc. along with it even protein mass and length are considered for prediction of protein functions. To further improve the accuracy feature selection has been performed. According to the enzyme nomenclature scheme the protein are classified into 6 groups. This enzyme classes is nothing but the crystalize reactions of proteins and shows the functions of it.

KEYWORDS: *Machine Learning, Function Prediction, Proteins, Data Mining, Protein Family.*

I. INTRODUCTION

Every human cell contains proteins and knowing the functions of it, is help to the drug discoverers to find and detectvarious deceases. Basically, Proteins are made up of four levels of structure. Primary, Secondary, Tertiary, and Quaternary structure.Proteins are constructed from a set of 20 amino acids.**Primary Structure** describes the unique order in which amino acids are linked together to form a protein.**Secondary Structure** refers to the coiling or folding of a polypeptide chain that gives the protein its 3-D shape. There are two types of secondary structures observed in proteins. One type is the **alpha (α) helix** structure. This structure resembles a coiled spring and is secured by hydrogen bonding in the polypeptide chain. The second type of secondary structure in proteins is the **beta (β) pleated sheet.** This structure appears to be folded or pleated and is held together by hydrogen bonding between polypeptide units of the folded chain that lie adjacent to one another.**Tertiary Structure** refers to the comprehensive 3-D structure of the polypeptide chain of a protein. There are several types of bonds and forces that hold a protein in its tertiary structure. **Quaternary Structure** refers to the structure of a protein macromolecule formed by interactions between multiple polypeptide chains. Each polypeptide chain is referred to as a subunit. These sequence of the amino acids and structures of the proteins helps the most to assign the biological roles of the proteins. Protein Function

prediction includes its Molecular Functions, Biological Process, and Cellular components. From that we can predict the Homology based function prediction, structure based function prediction, and Genomic context based function prediction. Also, functions can predict based on types of the various proteins. In this research, we have classified the function into the 6 groups based on the enzyme classes [1].

PROTEIN FUNCTION PREDICTION

Proteins are classified in various categories: Structural Proteins, Enzymes, Hormones, Respiratory Proteins, Transport Proteins, Contract, Storage, Toxins. As we know proteins are responsible for a number of different functions, which are like: Molecule Processing, Amino Acid modification, Active Site, Beta Strand, and Chain.This Functions can be predicted on the basis of enzyme commission classification. Based on this we can classify protein functions into 6 different classes: EC1, EC2, EC3, EC4, EC5, and EC6. This EC1, EC2.., EC6 are the Enzyme Commission number (EC number). Which is a numerical classification scheme for enzymes, based on the chemical reactions they catalyze.As a system of enzyme nomenclature, every EC number is associated with a recommended name for the respective enzyme.

In EC number the first number shows to which of the six main divisions (classes) the enzyme belongs, the second figure indicates the subclass, the third figure gives the sub-subclass,and the fourth figure is the serial number of the enzyme in its sub-subclass. In research, it's defined as:

EC1: Hydrolases: Water to cleave chemical bonds,
EC2: Isomerases: Convert Molecule,
EC3: Ligases: Ligation of DNA,
EC4: Lyases: Joining of Specified Molecules,
EC5: Oxydoreducates: Transfer of electrons,
EC6: Transferases: Transfer of group of atoms.

Our main motivation is, how Machine learning Algorithms can help to predict the functions of any unknown proteins. Proteins are responsible for the so many functions in every cell, so finding the working role of it can help in many other deceases [1].

RELATED WORK

In paper [1] authors have used various machine learning algorithms for function prediction of human proteins like C5.0, Neural Net, SVM, Bayesian Network and CHAID Algorithms. In this, Decision Tree Based C5.0 Algorithm gave best accurate result around 90.86%. In paper [2] Identification of protein functions using a machine learning approach based on sequence derived properties they have used, machine learning approach based solely on protein sequence properties. In predicting the functions of 11 different proteins a high performance around 94.23% was achieved. In paper [3] protein function via graph kernel, authors have used kernel methods and support vector machine and obtained the accuracy around 72.33%. In paper [4] Global protein function prediction from protein-protein interaction networks, they have used the support vector machine and obtained 79% accuracy. In paper [5] Machine learning classifiers for human protein function prediction, authors have used C5.0 algorithm and for 25 molecular sequences and 21 SDFs. the tree obtained by using C5 algorithm gives the best accuracy of 83% for the 25 SDFs. In paper [6] Deep GO: predicting protein function from sequence and interaction using deep ontology aware classifier have used deep learning, cross-species protein-protein interaction network, for evaluating this method using standards established by the computational assessment of the function annotation and demonstrate a significant improvement over baseline methods such as BLAST in particular for predicting for cellular locations. In paper [7] the physical characteristics of human proteins in different biological functions, authors have used multitask deep neural networks, for developing the multi task deep neural network architect to tackle the multi label problem in protein function prediction. In paper [8], new deep learning approach predicts protein structure from amino acid sequence, so accuracies and efficiently predicting protein folding has been a holy grail for the field, this approach. In paper [9] data mining frame work for protein function prediction authors have used SVM, for this the large input dimensionally that can be extracted from a sequence and secondly in dealing with small sample size. In paper [10] Deepred: automated protein function prediction with multi task feed-forward deep neural networks , authors have used Deepred method which describes the Deepred method for predicting GO term based protein functions using stack of feed forward multi task deep neural network. In paper [11] the authors have done SVM based methods for prediction of proteins by using protein sequence

and amino acid. Also have developed a classifiers for rRNA-, DNA and RNA which are responsible for controlling many cells in process. The accuracies were around 84%, 78% and 72% accordingly.

COMPARISION TABLE

Author Name	Method Name	Database	Remarks
Arun Vikram	Machine Learning Algorithms	UniportKB based manual Dataset	Model is giving the best accuracy of 90.86 with C5.0 Algorithm.
Shuzlina Abdul Rahman	SVM	UniportKB based Manual Dataset	The large input dimensionali-t y that can be extracted from a sequence and secondly in dealing with Small sample sizes
Mohammed Al Quraishi	Deep Learning	UniportKB based manual Dataset	Accurately and efficiently predicting protein folding has been a holy grail for the field, and it is hope and Expectation that this approach.
Ahmet Sureyya Rifaioglu 1	DEEPred Method	UniportKB based manual Dataset	Described the DEEPred method for predicting GO term based protein functions using a stack Of feed-forward multi-task deep neural networks.

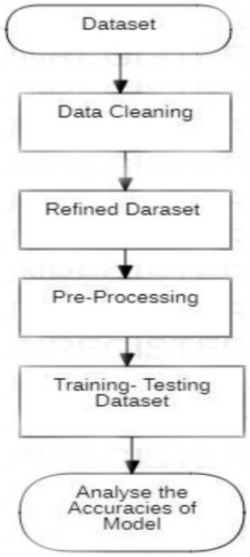
[Table 1: Comparison table of different methods]

II. METHOD & METHODOLOGY

This prediction required a layered workflow to reach the proper results. In this survey, tried to understand machine learning models that predict the molecule perform of unknown proteins and analysed their performance to obtain a model with highest accuracy. The list of genes appreciate Human Protein Dataset alongside Mass and Length is obtained from Uniprot. Then the Sequence Annotations corresponding to every of the genes is obtained. Hence the dataset is formed.

The main focus of this research was to enhance the classification accuracy while integrating the different sequence annotations like molecule processing, amino acid modifications and other structural related features. With an increase in the number of protein databanks, the need of an automated function prediction approach has emerged in order to handle the big data. For a given unknown protein sequence, predicting the functional class can be done on basis of different sequence annotations extracted, thanks to UniprotKB.

WORKING MODEL



[Figure 1: Flowchart of System]

DATASET & COLLECTION

Data collection is the process of gathering information of variable of interest so that we can reach to the desire output. Here is taken around 20430 protein dataset from the UniportKB. UniportKB is a freely accessible database of protein sequence and functional information. It contains the large amount of information about biological function of proteins derived from the research literature.

DATA CLEANING

Data cleaning is process of detecting and correcting the corrupt or inaccurate records from the dataset and refers to identifying incomplete, incorrect part of data and then replacing and modifying the data. UniportKB handles so many details about Proteins, which all are not required for Protein function

prediction. So dataset is contains the manually selected values.

REDEFINED DATASET

Here dataset contains Protein Name, Gene Name Along with their Mass, Length and sequence and their annotation score to Predict Functions.

PRE-PROCESSING

Data Pre-processing involves transforming raw data into an understandable format. Real world data is often incomplete, inconsistent and lacking in certain behaviour trends. And it might contains many errors. Data Pre-processing prepares raw data for further processing.

Understanding of dataset and Pre-processing with classifying the dataset into EC1, EC2, EC3, EC4, EC5 and EC6.

Divided dataset into two parts: X- contains Protein Details, Y- Contains accordingly EC1, EC2, EC3, EC4, EC5 and EC6.

TRAINING AND TESTING OF DATASET

The Training dataset is used to make sure the machine recognize the patterns in the data. The cross validation data is used to ensure better accuracy and efficiency of algorithm used to train the machine. And Test data is used to see how well the machine can predict new answers based on its training.

Will split Dataset into two parts for Training and Testing and check and compare the accuracies.

ANALYSE THE RESULT

Appling the Various Machine Learning Algorithm to find the best accuracy. Here the used algorithms are decision tree based C5.0 algorithm, SVM and Metrix based approach.

III. TOOLS AND TECHNOLOGY

Jupyter Note book Python Programming editor is used for completing the work at various stages of research and development. Different Machine Learning Libraries were used for programing like Numpy, Pandas, Matplotlib, Sklearn, Seaborn, GUI python Library: Tkinter, Web-based Development: Python-Flask.

IV. CONCLUSIONS

Protein is one of the important component in human body and Protein Function Prediction its self a challenging task for medical science. There are so many other ways to predict the functions of the proteins but predicting the functions from the enzyme classes is one of the best way to assign the roles to unknown proteins. With the help of the Machine learning algorithm we can classify and predict the functions of the proteins. After using various algorithms like decision tree, SVM, Metrix based algorithms, the C5.0 algorithm is giving the best accuracy around 90.86%. Which is highest to compare with any other algorithms.

REFERENCES

[1] Ahmad, S., Gromiha, M. M., & Sarai, A. (2004). Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics, 20(4), pp. 477–486.

[2] Ahmad, S., & Sarai, A. (2005). PSSM-based prediction of DNA binding sites in proteins. *BMC bioinformatics*, 6(1), 33.

[3] Kuznetsov, I.B., Gou, Z., Li, R., & Hwang, S. (2006). Using evolutionary and structural information to predict DNA- binding sites on DNA- binding proteins. PROTEINS: Structure, Function, and Bioinformatics, 64(1), pp. 19–27.

[4] Li, T., Li, Q. Z., Liu, S., Fan, G. L., Zuo, Y. C., & Peng, Y. (2013). PreDNA: accurate prediction of DNA-binding sites in proteins by integrating sequence and geometric structure information. Bioinformatics, 29 (6), pp. 678–685.

[5] U. Consortium, "Uniprot: the universal protein knowledgebase," Nucleic acids research, Vol. 45, No. D1, pp. D158–D169, 2016.

[6] Webb, E.C. (1992). Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes (No. Ed. 6). Academic Press.

[7] Gaudet, Pascale, *et al.* "neXtProt: organizing protein knowledge in the context of human proteome projects." Journal of proteome research 12.1 (2012), pp. 293–298.

[8] Bruxella, J. Mary Dallfin, S. Sadhana, and S. Geetha. "Categorization of Data Mining Tools Based on Their Types. " International Journal of Computer Science and Mobile Computing 3.3 (2014), pp. 445–452.

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

[10] Li, Tao, Chengliang Zhang, and MitsunoriOgihara. "A comparative study of feature selec-tion and multiclass classification methods for tissue classification based on gene expression." Bioinformatics 20.15 (2004), pp. 2429–2437.

[11] X. Yu, J. Cao, Y. Cai, T. Shi, and Y. Li (2006). "Predicting rrna-, rna-, and dna-binding proteins from primary structure with support vector machines," Journal of Theoretical Biology, Vol. 240, No. 2, pp. 175–184.

[12] R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin,D.A. Natale, C.O'Donovan, N. Redaschi, and L.S. Yeh, "UniProt:the Universal Protein knowledgebase," Nucleic Acids Research, Vol. 32, No. 1, pp. D115-D119, Jan. 2004.Electronic copy