Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MACHINE LEARNING CONFIGURATIONS FOR HUMAN PROTEIN CLASSIFICATION USING SDFES

SUNNY SHARMA [1], AMRITPAL SINGH [2], GURVINDER SINGH [3], RAJINDER SINGH [4]

[1, 2] Research Scholar, DCS, Guru Nanak Dev University, Amritsar-143001, Punjab, India

[3] Professor and Dean, DCS, Guru Nanak Dev University, Amritsar-143001, Punjab, India

[4] Professor, DCS, Guru Nanak Dev University, Amritsar-143001, Punjab, India
E-mail: [1]sunnysharma05@yahoo.co.in

ABSTRACT

The identification of target proteins for diseased condition yields the development of the disease detection recommender system and drug discovery processes whose reticence can demolish the pathogen. The testing of this drug discovery is done through clinical and in addition through pre-clinical observations first on the creatures then on people. Thereafter the discovered drug is ready for public use. But if the drug discovery testing phase does not show the suitable consequences, then the entire task must be repeated. This repetitive clinical as well as the preclinical experimentation task is very cumbersome. But keeping in view the importance of the disease detection and drug discovery phase in protein identification as well as in the protein classification process this task must be done by researchers. The advancements in computational biology reveal the importance of computational prediction of protein function or to identify the target on the basis of protein sequence extracted features. To accurately predict the human protein functionalities. lots of approaches are incorporated but this is a very cumbersome task due to the large and versatile nature of the domain. The present work will help to do this job through computational prediction. This paper involves the development of a model which use associative rule mining to extract the sequence derived features at a single platform (SDFES-Sequence derived feature extraction server) from the given human protein sequence and then critically analyzed with machine learning (ML) approaches under the aegis of data analysis tool WEKA. The new sequence derived features are identified and incorporated in the data set, and the scopes of ML approaches were examined for effective prediction. The important configuration incorporation and their configured comparison of approaches are completed to accomplish higher accuracy. In addition to comparative analysis, the limitation of ML approach is discussed along with its remedies by changing the configurations. The proposed work will assist to derive the sequence extracted feature together at a single place and further predict the class or function of the protein which leads to the innovation in drug discovery and disease detection recommender systems.

Keywords: *Protein, Machine Learning, WEKA, Random Forest, Decision Tree.*

1. INTRODUCTION

Protein function prediction, protein classification, disease detection and drug discovery as well as their recommendation systems, are immense areas with the colossal amount of information. Lots of research activities are happening for protein classification and protein function prediction, but the learning about its right perception is quite low, so there is a need to explore this vast domain. The machine learning (ML) approaches gives promising results to vast as well as to not so clearly characterized zones of research, this encourages using the power of ML to explore such a vast domain and boost the present understanding of protein. The survey of 65 papers on ML approach shows that ML approaches are extensively used for human protein function predictions and it is the prominent area where ML got some challenges to show its supremacy [1]. The association of rule mining in data mining is a frequently used technique, which provides significant valuable rules or patterns. This association of rule mining extracts the possibility of the co-occurrence of common features in a data collection, this also encourages to use the power of rule mining to explore such a vast and not so clearly characterized zones of research and find the occurrence of common features while boosting up the present knowledge of protein. On the other

hand, there is very much reliable, vibrant and the white box decision tree [2] [3] technique of machine learning that helps to predict protein class, and do the classification of protein with the help of rule mining. Its step by step approach encourages computational experts to use this technique even without much information of the concerned area; with having nodes and edges this technique portrays different functionalities at various levels of the tree [4]. It clearly characterizes the issue structure and its elucidations progressively in the hierarchical way which is considerably less demanding to fathom. This hierarchical approach guides the input parameters to achieve its goals [1] [5]. So distinguishing challenges and the conceivable answers for protein classification problems is the key concentration of the investigation.

Human protein classification is an important research area because of its implication on various key research areas like disease detection, drug discovery, crop hybridization, etc.

The motivation for present research is to facilitate the sequence derived feature extraction process which in turn improves the accuracy of classification and prediction of human protein.

Second motivation is the use of machine learning approach for the process which is already applied in existing literature. But this research showcase that machine learning approach too have some pitfalls.

So to overcome those issues various configuration were tried and varying accuracy results of individual classes were obtained. This shows that for tasks like drug discovery, when the interest in identifying a particular class is more, these configuration are very useful.

1.1 Associative Rule Mining

In data mining association rule mining has been extensively studied & it is a frequently used technique, which produces significant valuable rules or patterns. Association is a rule mining which extracts the possibility of the co-occurrence of items or features in a collection. Association rules or the production rules convey the relationships between co-occurring features or items. Among the association the strong association rules favors the strong relation between data items & the weak association rules are considered least related items. To measure the interest, the support and confidence are two listed measures. To extract associated variables, the values for support and confidence are extracted to the production system. This grabbed the important relationship between variables only if satisfied minimum support and confidence. To find out all the frequently occurring items in the repository is a cumbersome because to find out the entire item from the repository with all combinations is tough job. The set of probable item sets is said to be the power set over all the items & has size $2n-1$. As the data items increases, the size of combination increases exponentially. The possible efficient search can be performed using anti- monotonicity [35].

1.2 Protein and Protein Function Information

Proteins are the complex foundations for all the living creatures on earth. Under the aegis of protein, the body's tissues and organs perform functionalities, structuring and controlling the body's parts. The building blocks of proteins are amino acids, they assist or anchored to each other and perform functionalities like store information, send information to body organs as well as structuring in the living cell. It can be portrayed as a string of 20 diverse amino-acids (AA) and they are anchored to make a protein of the living cell. These20 distinct kinds of amino acids are recorded as Serine, Alanine, Proline, Arginine, Valine Asparagine,Threonine, Glumatic, Aspartic. Glycine, Tryptophan, Histidine, Cysteine. Isoleucine, Phenylalanine, Tyrosine, Glutamine. Leucine, and Methionine, Lysine.

Based on functionalities of proteins, they are categorized as Transport, storage, motor, structural, receptor, signaling, hormones, Antibodies, enzymes etc. Fundamentally the broad chain of amino acids is framed with peptide bond which is the bridge bond of amino corrosive structures with another amino acid corrosive structure. The progression of such peptide bonds is known as polypeptide bonds which is actually responsible for the correct functionality of every protein and depicting the interesting 3D structure [5].

1.2 Protein Features Extraction from Sequence

Protein classes having different protein sequences and these sequences constitute with various features which are actually responsible for functionalities of body organs and tissues. The sequence-based features are useful to anticipate protein class and also very helpful to build automatic protein class predictor classifier. The features can be extracted from protein sequence by utilizing different online bio-informatics servers or tools, which are freely available online such as TMHMM Sererv.2.0 [6], SignalP4.1 Server [7], NetNGlyc1.0 Server [8], ProtParam [9], PSORT