Machine Learning Configurations for Enhanced Human Protein Function Prediction Accuracy

Amritpal Singh, Sunny Sharma, Gurvinder Singh and Rajinder Singh

Abstract Molecular class prediction of a protein is highly relevant for conducting research in domains of disease-detection and drug discovery process. Numerous approaches are incorporated to increase the accuracy of Human protein Function (HPF) prediction task, but it is highly challenging due to wide and versatile nature of this domain. This research is focused on sequence derived attributes/features (SDF) approach for HPF prediction and critically analyzed with the WEKA data analysis tool. New SDFs were identified and included in the training dataset from the Human protein reference database, enhanced as in number of sequences and the related features for deriving the relation with various protein classes. A range of Machine Learning approaches were analyzed for prediction effectiveness and a comprehensive comparison is carried out to achieve higher classification accuracy. The Machine Learning approach is also analyzed for its limitation on application of broad spectrum data domain and remedies for the limitation were also explored by changing the configuration of data sets and prediction classes.

Keywords Bagging · Bayes Net · C5 · Decision tree · HPF · IBK · J48 · Logistic approach · PART · Random forest · SDF · Weka

1 Introduction

Protein classification is a vast domain with enormous amount of data available for research and analysis yet the knowledge about its correct perception is very limited. On the other hand Machine learning (ML) provides promising answers to not-so-clearly defined areas of research. Thus, it's a powerful tool to explore the possibilities of the enhancement of the current understanding of protein.

A. Singh · S. Sharma (✉) · G. Singh · R. Singh
Guru Nanak Dev University, Amritsar, India
e-mail: sunnysharma05@yahoo.co.in

A. Singh
e-mail: amritsinghmand@gmail.com

A. K. Luhach et al. (eds.), *Smart Computational Strategies: Theoretical and Practical Aspects*, https://doi.org/10.1007/978-981-13-6295-8_4

Decision tree [1, 2] based prediction approach of machine learning is very clear and reliable for protein classification. Being a white-box approach it clearly illustrates the sequence of computations involved at each and every stage. This plus point enables its usage by computational experts even without much knowledge of the concerned domain. Similarly, a domain expert is empowered for examining the toute followed by an expert of computation. So the gap between technical knowledge and domain expertise. Nodes and edges indicates various utilities at the different stages of computations in a Decision tree [3]. A decision tree neatly depicts the results required or outputs of various possibilities of outcome. It clearly defines the problem structure and its interpretations in a hierarchical way which is much easier to comprehend. As the model has a unique ability of considering different initial parameters and reaching a goal [4, 5]. However, recent advancements suggests that the prediction of Protein-Function is a domain where ML faces some challenges. A thorough collection of almost 65 papers in 'A Survey of Computational Methods for Protein Function Prediction' helped arrive on this aspect of ML applicability on HPF prediction [6]. So identifying the challenges and the possible solutions to overcome such situations is also the key focus of the study.

2 Protein and Protein Function Information

Proteins are the large and complex building blocks for all the life forms on this planet. They play a defining role in functionality and the structuring of organs and tissues of the body of an organism and also perform regulatory functions in it. Several diminutive components form protein. These diminutives are referred as Amino-Acid (AA) groupings. This arrangement helps in various tasks of a life cell. Proteins can be assigned to various categories based on functions, these groups for each protein are listed here as: Transport proteins, Enzymes, Hormones, Motor proteins, Immunoglobulin or Antibodies, Receptors, Storage proteins, Structural proteins, Signaling proteins. There are various diminutive units, 20 in number, listed as: Alanine, Arginine, Asparagine, Aspartic acid, Cysteine, Glumatic acid, Glycine, Glutamine, Histidine, Leucine, Lysine, Isoleucine, Methionine, Serine, Phenylalanine, Proline, Threonine, Tyrosine, Tryptophan and Valine. In the prime form the protein sequence it can be characterized as a string of Twenty AAs and they are chained to make a protein [7]. Amino part (–NH2) and the carboxylic part (–COOH) of the nearby AA is joined to generate a bond (peptide) as:

$$\begin{array}{cccc} -\mathrm{NH2} & + & -\mathrm{COOH} & == -\mathrm{CONH} \\ \text{A}-\text{group} & & \text{C}-\text{group} & \text{Resultant bond} \\ (\text{AA1}) & & (\text{AA2}) & (\text{Peptide}) \end{array}$$

Basically AA forms a Peptide bond with the other AA and an extensive grouping of AA's is formed. A series of such peptide bonds is known as a polypeptide. The chain of AA's describes the unique 3D structure and the exact function of each protein [8].