NotificationNumber:**569/2024**
Notificationdate:**25/10/2024**

Name of the Scholar: ShabanamKhatoon
Name of the Supervisor: Dr.Suraiya Jabin(Professor)
Name of the Department/ Faculty: Department of Computer Science/ Faculty of Sciences
Topic of the Research : Function Prediction of Hypothetical Proteins Using Machine Learning Techniques
Keywords: Hypothetical proteins, Functionprediction, Molecularfunction, Deeplearning, Motif, Physicochemical feature, Sequence based feature, Annotation based feature.

Findings

Protein function prediction (PFP) is the most researched and challenging subject among computational biologists. The great majority of known proteins have yet to be experimentally characterised, and there is a large disconnect between their structures and functions. A publicly available dataset is generated towards function prediction of proteins of bacterial phyla. This dataset is based on four types of features: sequence based, physicochemical based, annotation based, and sub-sequence based. Each sample contains the following columns named as Entry, Entry name, Sequence, Sequence based Features, Physicochemical Features, Annotation based features, Subsequence based features amounting to a total count of 9890 features which is the most exhaustive set of features till date. For subsequence based features, we extracted motifs from Prosite database for all protein sequences, and these motifs were counted for eachsequence. We created a dataset 2ofpathogenic unreviewed proteins from 9 bacterialphyla, each containing 9890 features, similar to the train/test dataset 1 of reviewed proteins but withouttarget labels, to predict their functions.

We prepared Train dataset using reviewed (Swiss-Prot) protein sequences and two types of Test datasets; one taken from reviewed proteins only (25% of full data), and another one using unreviewed protein sequences (TrEMBL) of bacterial phyla from UniProtKB. Protein function prediction being multi-class and multi-label problem, we initially incorporated almost all GO terms (4000 approximately) for the collected bacterial protein sequences (approximately 323,719). To cope up with this situation, frequent data mining algorithm was applied over theMF domain's GO terms in the whole web-scrapped dataset to reduce the number of class labels (GO terms) and preserving large proportion of the data at the same time. As an outcome, we retrieved a total of 17 GO termsas the most frequent ones in our dataset of 323,719 reviewed bacterial protein sequences. Thus, the initial dataset was worked upon to extract the proteins associated with these 17 GO terms that came out to be 171,212 proteins. Due to the fact that multiple GO terms (MF) may be associated with a protein, these 17 frequent GO terms were present incombinationwithother GOterms inthe extracteddataset. So, after consideringallthe unique GO terms occurring in combination with these most frequent 17 terms, total GO terms came out to be 1739 with evidence codes. This step ensured that the Train dataset of reviewed proteins for implementation of proposed model is possessing good number of positive and negative samples for eachof the target GOterm. A train dataset created in such a wayhas many advantages, themostimportantbeing it isbalancedwithrespecttopositiveandnegativesamples belonging to specific target label/GO term and covering most of the GO terms belonging to MF domain for bacterial organisms and thus capable of capturing all hidden patterns in the dataset with respect to input-output mapping. We used this dataset for training of various machine

learning models out of which deep neural network (DNN) was doing the best. We experimented with different combinations of features for training of multiple deep neural network models. DNN trained on feature combination of sequence based features, annotation based features, and motifcount based features performed the best with F1 measure of0.7567 over the test dataset of reviewed bacterial proteins. We noticed that motif count based features in combination withother 3 groups did better. We trained various DNN models with combination ofdifferent feature groups and selected top most performing five models to create an ensemble which gave an F1 score of 0.7912 on reviewed bacterial proteins, thereby increasing the performance of best individual model. We then used our ensemble to predict function of hypothetical bacterial proteins for which there's no significant homology modelling is available.

It will be interesting to extend this model for other organisms of proteins to see effectiveness of the model designed. This work is a novel effort towards function prediction of bacterial proteins as compared to other works inthe literature. The success ofthe proposed work advocates design of protein function prediction systems for a dedicated organism over large train dataset using deep learning. The proposed work is a refreshing computer science solution on the famous biologicalproblem of protein function prediction and we hope to further improve it in future.