INTRODUCTION

RNA-Binding-Proteins play important roles in many biological processes like:
- Gene Regulation
- Protein Synthesis
- Sequence encoding during both transcription and post-transcription

Prediction of RNA-Binding-Proteins is hence important. Experimental processes of protein prediction are time-consuming and tedious. Hence comes the need of computational procedures for RNA-Binding-Protein prediction.

OBJECTIVES

- Create a Computational approach for prediction of RNA-Binding-Proteins
- Identify the RNA-Binding-Proteins from sequence information only
- Present a machine learning technique based on a comprehensive set of encoded features.
- Compare the results with the state-of-the-art methods.
- Finally, develop a predictor that can be applied for RNA-Binding-Protein prediction from sequence only.

METHODS

Training Set: a non-redundant set of RNA-Binding-Proteins (RBPs) and Non-RNA-Binding-Proteins (NRBPs).

For C-T-D, the Amino Acids were divided into groups where amino acids with similar characteristics fall under a group (Figure 4). The composition refers to composition of each amino acid group in the sequence. Similarly, transition refers to change of amino acids from one group to other as we go linearly through the sequence. Distribution refers to how one amino acid group is distributed throughout the protein sequence. For instance, where does the first element of the group fall in the protein sequence, similarly 25 percent, 50 percent, 75 percent and all the element’s position in the sequence chain is calculated for the transition.

FEATURE ENCODING OF SEQUENCES

The dataset collected was converted into fasta files, individual fasta files representing a single protein. These individual fasta files were used as an input to feature encoding.

| PSSM-DDT | 1800 |
| PSSM-DDT | 1000 |
| PSSM-DDT | 800 |
| Hydrophobicity | 21 |
| Polarity | 21 |
| Vander Waals Volume | 21 |
| Polarizability | 21 |
| Predicted Secondary Structure | 21 |
| Predicted Solvent Accessibility | 13 |

Figure 1: Features used to encode the protein sequences.

The proposed prediction technique outperforms state-of-the-art approach significantly with an accuracy around ninety percent. As for machine learning, stacking was used to predict RNA-Binding-Proteins from Non-RNA-Binding-Proteins.

RESULTS

Performance of our sequence-based RNA Binding predictor was measured using 10-fold cross-validation. Sensitivity (SN), Specificity (SP), and Accuracy (ACC) was calculated. RNA-Binding-Protein was successfully predicted with an overall ACC of 89.58 percent, with a SP of 90.41 percent and a SN of 89.58 percent.

Evident from the result comparison, our predictor outperformed the most recent RNA-Binding-Protein Predictor published.

DISCUSSIONS

From performance evaluation and feature ranking processes, we could see the evolutionary features play the most significant role in the prediction of RNA-Binding-Proteins. Our feature set consisted of more than half of its features from evolutionary features (PSSM).

Similarly, for PSSM features we apply distance transformations namely, PSSM-DDT, PSSM-DDT, PSSM-EDT. These take the pssm matrix generated from psi-blast as an input and gave multidimensional vectors based on the distance between amino acids in the protein sequence.

PSSM-DDT measures the occurrence probabilities of a pair of same amino acids separated by a distance d in a protein from the PSSM profile. PSSM-DDT measures the occurrence probabilities of pairs of different amino acids separated by a distance d in a protein from the PSSM profile. PSSM-DDT (Evolutionary Distance Transformation) measures non-co-occurrence probability for two amino acids separated by a certain distance d in a protein from the PSSM profile.

Acknowledgements

We gratefully acknowledge the Louisiana Board of Regents through the Board of Regents Support Fund, LEQSF (2016-19)-RD-B-07.