Accurate Identification of RNA-binding Proteins (AIRBP) Using Machine Learning Techniques

Avdesh Mishra, Reecha Khanal, Md Tamjidul Hoque
email: amishra2@uno.edu, rkhanal@uno.edu, thoque@uno.edu
Computer Science Department, University of New Orleans, LA, USA

OBJECTIVES

- Create Computational approach to efficient identification of RNA-Binding-Proteins (RBPs) from sequence information alone.
- Extract features like Evolutionary information, physio-chemical properties and disordered properties from a protein sequence.
- Use advanced evolutionary algorithm to get a feature-subset.
- Use advanced machine learning techniques like stacking to train a robust model.
- Finally, develop a predictor that can be efficiently applied for RBP prediction.

INTRODUCTION

RNA-Binding Proteins (RBPs) play important roles in many biological processes like Gene Regulation and Protein Synthesis. They have also been linked to some critical diseases like:
- cancer, neuro-degenerative diseases
- immunological disorders
- muscular atrophies and metabolic disorders

Therefore, prediction of RNA-Binding Proteins is very important. Effective Computational methods for RBP prediction also can help assist existing experimental techniques.

FEATURE ENCODING

Feature vector for each protein sequence was derived from the PSSM profile, Physiochemical Properties (PP), Residue-wise Contact Energy Matrix (RCEM) and Molecular Recognition Features (MoRFs). Each of the 10 properties along with their encoding mechanism is described in Figure 3.

Machine Learning Model

AIRBP was developed using stacking, an ensemble based machine learning approach, that collects information from multiple models in different phases and combines them to form a new model. Sets of stacking framework tested are:
- i) SF1: RDF, XGBoost, LogReg, KNN in base-level and XGBoost in meta-level,
- ii) SF2: Bagging, XGBoost, LogReg, KNN in base-level and XGBoost in meta-level and
- iii) SF3: ET, XGBoost, LogReg, KNN in base-level and XGBoost in meta-level.

RESULTS

AIRBP outperforms top performing RBPPred by 3.26%, 6.75% and 9.53% in accuracy, F1-score and MCC respectively, on a benchmark dataset. F1-score and MCC are two widely used measures for imbalanced dataset. Moreover, AIRBP outperforms RBPPred by 4.76%, 3.58% and 9.11% in accuracy, F1-score, and MCC, respectively. Hence, AIRBP can be used for efficiently predicting RBPs and providing valuable insights for treating critical diseases.

CONCLUSIONS

AIRBP outperforms the best-performing state-of-the-art approach, RBPPred, significantly on both benchmark dataset and independent test datasets.

REFERENCES


ACKNOWLEDGEMENTS

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