LPI-EnEDT: An Ensemble Framework with Extra Tree and Decision Tree Classifiers for Imbalanced lncRNA-protein Interaction Data Classification

Lihong Peng (plhhnu@163.com)
Hunan University of Technology https://orcid.org/0000-0002-2321-3901
RuYa Yuan
Hunan University of Technology
Ling Shen
Hunan University of Technology
Pengfei Gao
Hunan University of Technology
Liqian Zhou
Hunan University of Technology

Research

Keywords: lncRNA-protein interaction, Ensemble, Class imbalance

Posted Date: May 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-480398/v1

Abstract

Background: Long noncoding RNAs (lncRNAs) have dense linkages with various biological processes. Identifying interacting lncRNA-protein pairs contributes to understand the functions and mechanisms of lncRNAs. Wet experiments are costly and time-consuming. Most computational methods failed to observe the imbalanced characterize of lncRNA-protein interaction (LPI) data. More importantly, they were measured based on a unique dataset, which produced the prediction bias.

Results: In this study, we develop an Ensemble framework (LPI-EnEDT) with Extra tree and Decision Tree classifiers to implement imbalanced LPI data classification. First, five LPI datasets are arranged. Second, lncRNAs and proteins are separately characterized based on Pyfeat and BioTriangle and concatenated as a vector to represent each lncRNA-protein pairs. Finally, an ensemble framework with Extra tree and decision tree classifiers is developed to classify unlabeled lncRNA-protein pairs. The comparative experiments demonstrate that LPI-EnEDT outperforms four classical LPI prediction methods (LPI-BLS, LPI-CatBoost, LPI-SKF, and PLIPCOM) under cross validations on lncRNAs, proteins, and LPIs. The average AUC values on the five datasets are 0.8480, 0,7078, and 0.9066 under the three cross validations, respectively. The average AUPRs are 0.8175, 0.7265, and 0.8882, respectively. Case analyses suggest that there are underlying associations between HOTTIP and Q9Y6M1, NRON and Q15717.

Conclusions: Fusing diverse biological features of lncRNAs and proteins and exploiting an ensemble learning model with Extra tree and decision tree classifiers, this work focus on imbalanced LPI data classification as well as interaction information inference for a new lncRNA (or protein).

Full Text

Due to technical limitations, full-text HTML conversion of this manuscript could not be completed. However, the manuscript can be downloaded and accessed as a PDF.

Figures

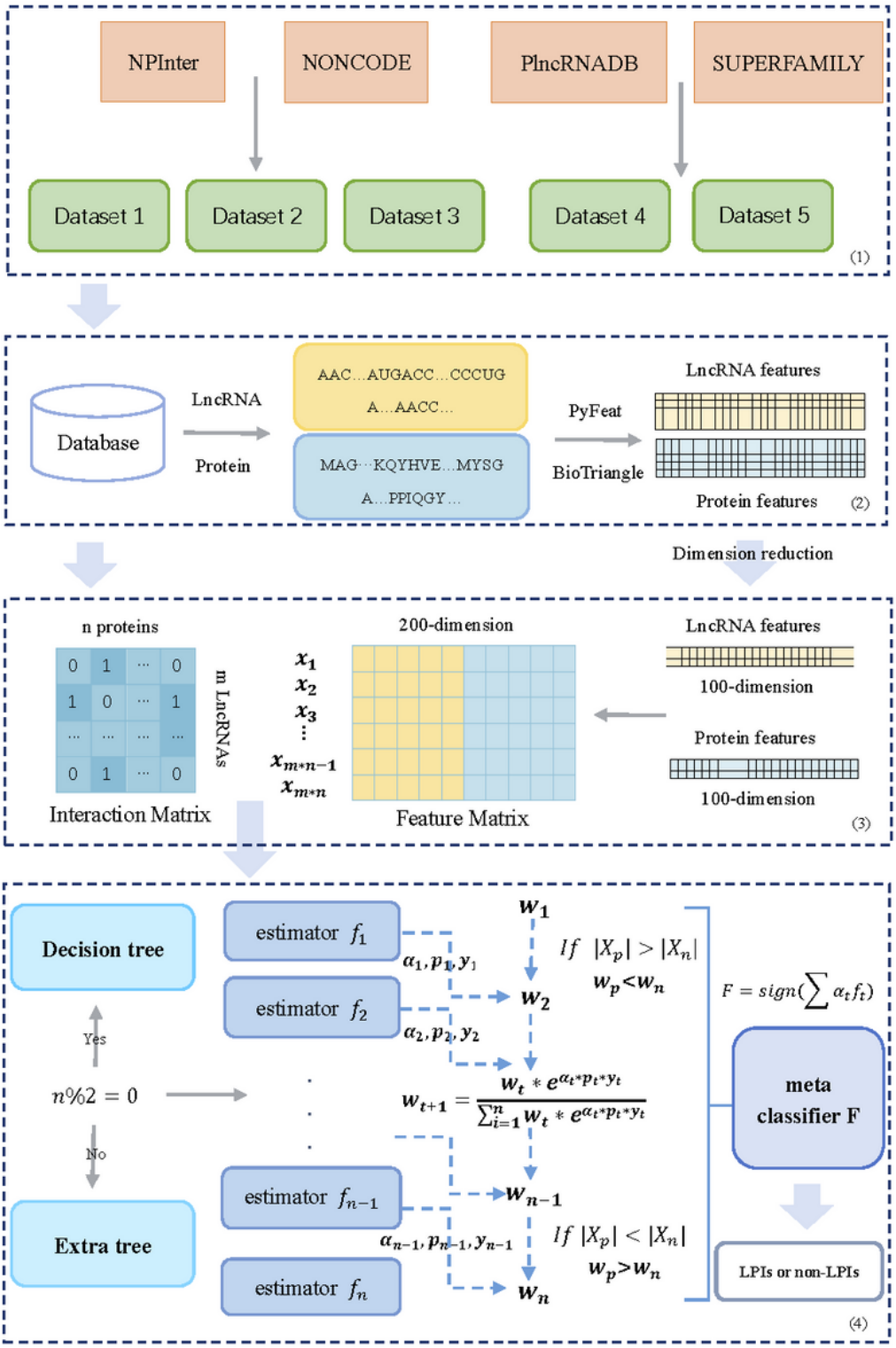

Figure 1

The Pipeline of the LPI-EnEDT framework. (1) Dataset arrangement. (2) Feature description. (3) Feature selection. (4) LPI classification.

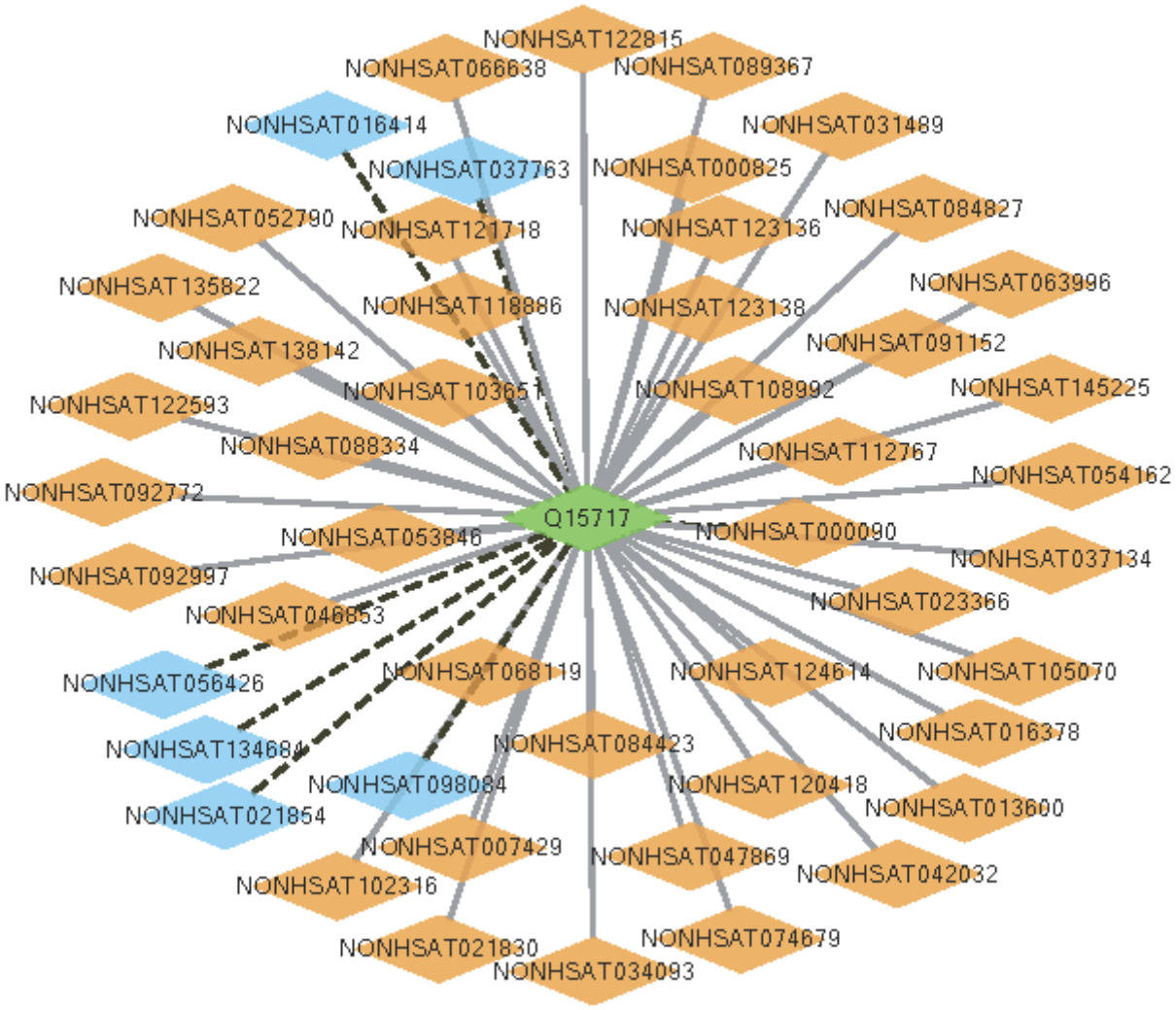

Figure 2

The predicted top 50 LPIs on Dataset 1.

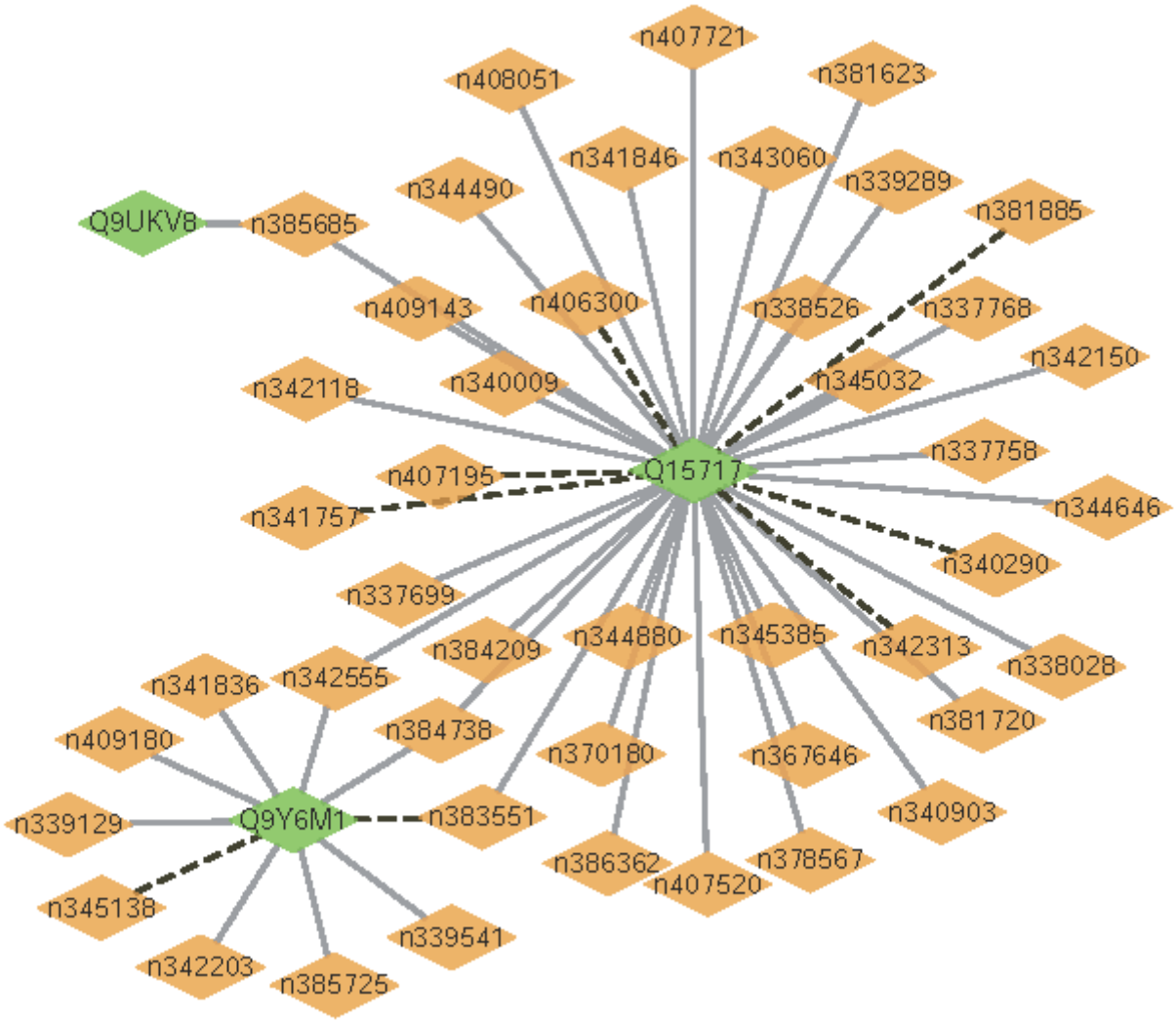

Figure 3

The predicted top 50 LPIs on Dataset 2.

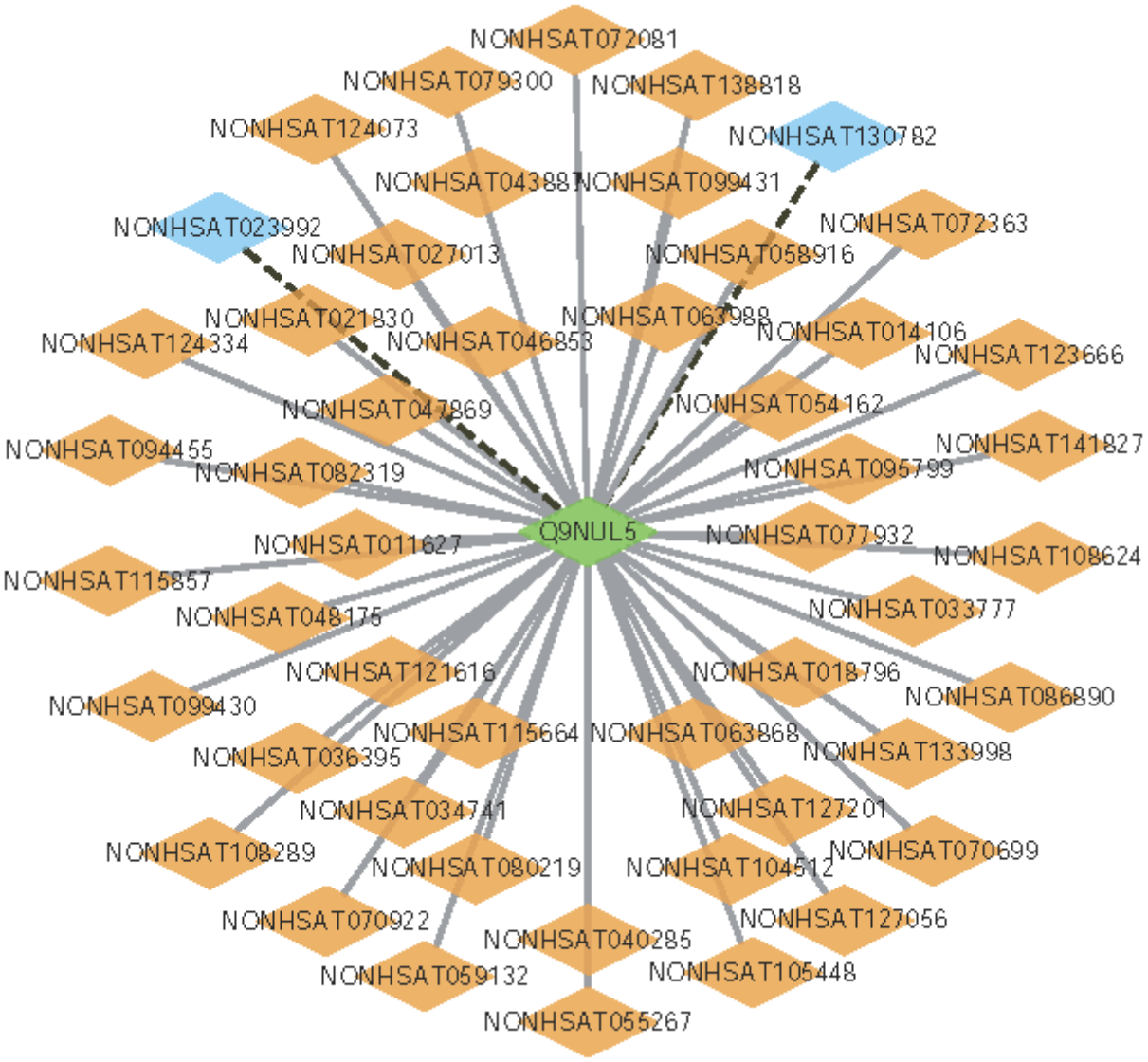

Figure 4

The predicted top 50 LPIs on Dataset 3.

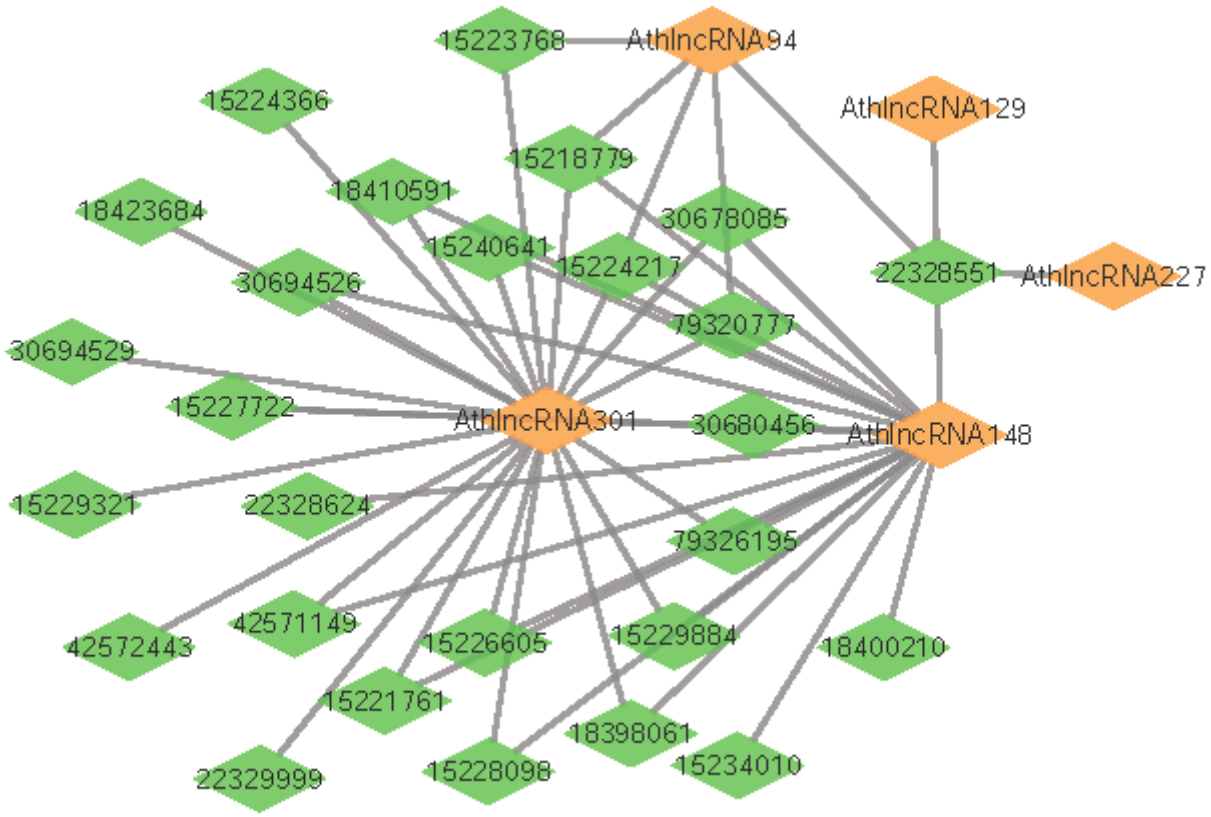

Figure 5

The predicted top 50 LPIs on Dataset 4.

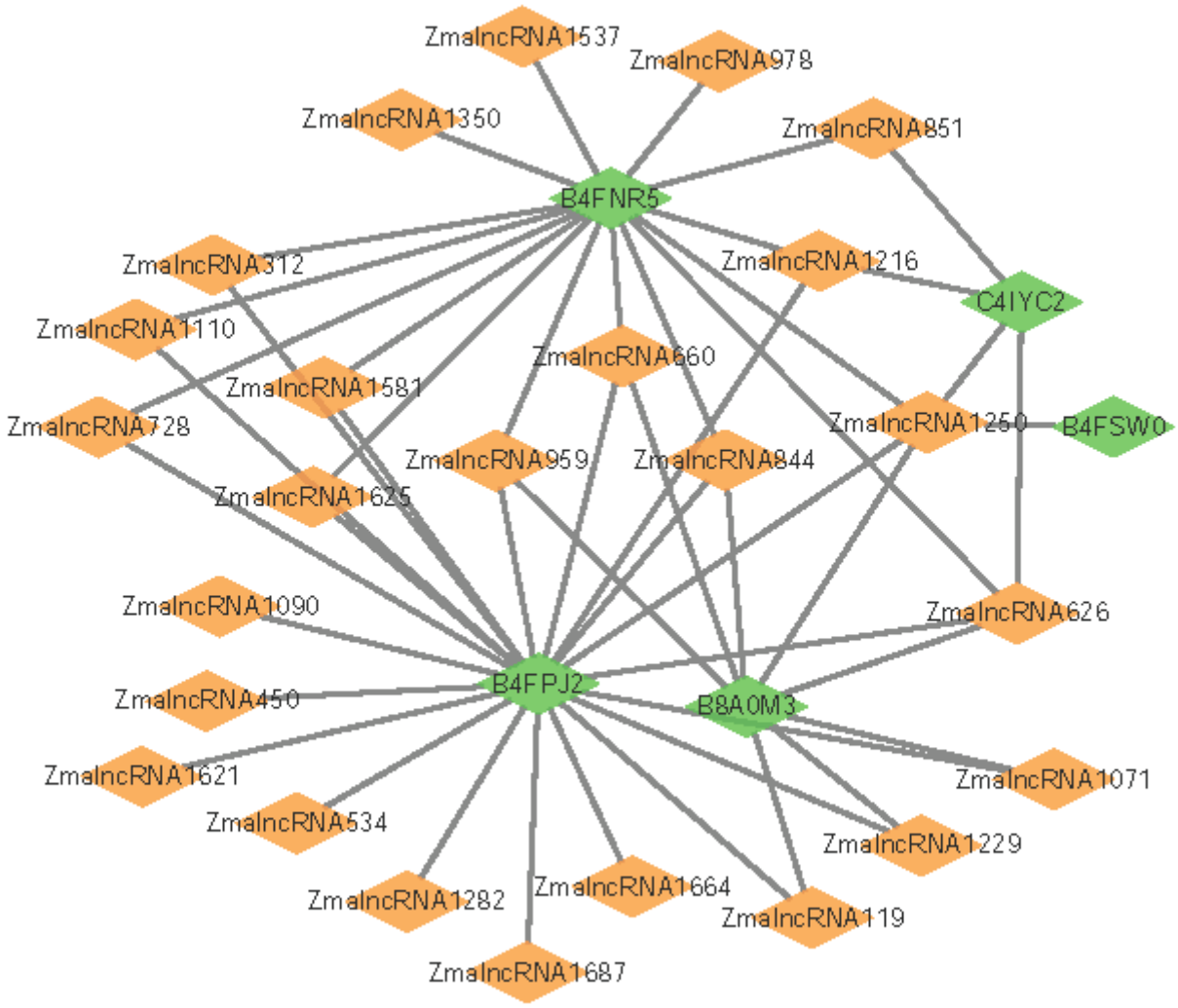

Figure 6

The predicted top 50 LPIs on Dataset 5.