TLRs - pathogen nucleic acid specificity prediction

To understand the dynamics of pathogen-specific host responses, we identified some key sequence, expression and function related features of nucleic acid sensing Toll-like receptor proteins of the host. Our findings suggest that such host-specific features are directly related to the strand (single or double) specificity of nucleic acid from pathogens. Therefore, we developed a model to predict the pathogen nucleic acid strand specificity of TLRs as follows.

RFC-LOO (Random Forest Classifier - Leave One Out) model

This repository includes the source code for the RFC-LOO model. It can be used to predict the nucleic acid strand specificity of pathogens sensed by novel TLRs.

The model is trained on 27 features of 129 TLR proteins of 16 species that includes sequence network evolutionary-based features, gene expression-based features and functional annotation-based features. All the features are mentioned in the following steps and their estimations are defined in the method section of publication (under review).

- ➔ ***Sequence, network and evolutionary features***
  - ◆ GC content, gene length and protein lengths
  - ◆ Protein domain and sequence length
  - ◆ Evolutionary gene age
  - ◆ Sequence-predicted solvent accessibility and disorder content
  - ◆ Sequence-predicted nucleic acid binding sites prediction
- ➔ ***Gene expression features***

- ◆ Tissue-specificity and average gene expression level

➔ ***Function annotation features***

- ◆ Gene ontologies

Overall performance measures, each class performance measures, ROC curve and misclassified prediction of the RFC-LOO model along with strand specificity prediction for novel and blind set TLRs are generated.

Running the model

Conditions

The model was developed in python version 3 and above. It is necessary to install the sklearn, matplotlib, seanorn and joblib libraries to run the RFC-LOO model.

Steps

1. Extract all the features and create a tab separated file as described in the format provided in the “prediction_input.txt” file.

2. Run the python script using the following command
   $python3 specificity_prediction.py <input_file>

   Here, input_file is “prediction_input.txt” which have above mentioned feature values for novel and blind set TLRs.

3. Predicted specificity for novel TLRs is shown on terminal.