

RESA Random Forest model

Overview

These scripts use the random forest model (R package of “randomForest” 4.6-12), trained using sliding window approach over RESA profiles, to predict the mRNA stability of user provided sequences based on the weighted importance values given to each sequence features in during the training process.

These scripts are distributed under Mozilla Public License (MPL) version 2.0.

How to train RF model

Please note, you can skip this section of building RF model and use the trained RF model on the complete RESA total data by loading the already trained RF model

"SelectedFeatures_RESATotal_WT_.RData" in the following section for testing directly.

Method

These steps are coded in RFModel_Training.R script.

1. use the RNA-seq data to calculate the average number of reads per each nucleotide at 2hpf and 6hpf
2. per nucleotide, calculate mRNA stability by computing $\log_2(6\text{hpf}/2\text{hpf})$
3. use sliding window (100nt) over the 3'UTRs (for simplicity, we have provided this sliding window example demo data sample as WindowSequences_Demo.csv, this is small part of the sliding window data of RESA total)
4. for each window, the RESA score for the 50th position is used to measure the stability of each window.
5. sliding window data (RESA score, Window sequence) is used for training the RF model, using the selected features from the cross validation procedure

Results

This process creates a trained RF model which is saved as

"Trained_RFmodel_RESATotal_WT_.RData", and can be subsequently used by the following script to predict the stability of each sliding window.

How to use trained RF model to predict sequence

stability

Method

These steps are coded in Trained_RFmodel_Prediction.R script.

1. Add your testing sequence in a csv file and name it "TestingSequence.csv" with (Name,Sequeunce) as shown in "TestingSequence.csv"
2. From the terminal, Go to the directory of the R script and run "Rscript Trained_RFmodel_Prediction.R", this script uses the RF model trained in the previous section to predict the stability using the sequence of each sliding window

Results

Running the script will generate two files namely:

1. "TestingSequence_PredictedStability.csv" which contains the testing sequence along with the predicted stability values in log2 scale (last column)
2. barplot to represent the different stability scores predicted using the trained RF model.

Contacts

For technical questions please contact: antonio.giraldez@yale.edu or marioabdelmessih@gmail.com