PlantMirP-rice: an efficient and program for rice pre-miRNA prediction

Version 1.0.0

June 13, 2020

Author: Yuangen Yao

Contact: yygen89@163.com

Department of Physics, College of Science, Huazhong Agricultural University, Wuhan, Hubei 430070, China

1. Statement

Implementation: RiceMirP, a random forest-based classifier, is implemented in Perl (v5.24.1) and R (v3.2.2), with the recommended versions in parentheses. The Random Forest algorithm is implemented by the randomForest R package (v4.6–14).

Availability: The local package of riceMirP is freely available to the academic community at https://github.com/yygen89/riceMirP. For non-profit users, you can copy, distribute and use the software for your scientific studies. Our software is not free for commercial usage.

Usage: Our software is designed in an easy-to-use manner. We invite you to read the manual before using the software.

Updating: Based on users' suggestions and advices, we will update software routinely. Therefore, your feedback is greatly important for our future updating. Please do not hesitate to contact with us if you have any concerns.

Citation: If the software has been helpful for your work, we wish you could cite the article.

Acknowledgements: This work was supported by the National Natural Science Foundation of China (Grant nos. 31601071, 11675060).

2. Introduction

Rice microRNAs (miRNAs) are important post-transcriptional regulation factors and play vital roles in many biological processes, such as growth, development, and stress resistance. Identification of these molecules is the basis of dissecting their regulatory functions. Current various machine learning techniques have been developed to identify precursor miRNAs (pre-miRNAs). However, no tool is implemented specifically for rice pre-miRNAs. This study aims at improving prediction performance of rice pre-miRNAs by constructing novel features with high discriminatory power, and training model with species-specific data. PlantMirP-rice (hereinafter called riceMirP for short), a stand-alone random forest-based miRNA prediction tool, achieves a promising accuracy of 93.48% based on independent (unseen) rice data. Comparisons with other competitive pre-miRNA prediction methods demonstrate that plantMirP-rice performs surpassingly to existing tools for rice and other plant pre-miRNA classification.

3. Dependencies

Before running software, Perl (v5.24.1), R (v3.2.2) and randomForest R package (v4.6–14), with the recommended versions in parentheses, are required to preinstall in local machines.

Perl: The installation package of Perl language can be freely obtained from CPAN (Comprehensive Perl Archive Network) through link: https://www.perl.org/cpan.html. Then, follow the installation instructions to finish installation.

R: The installation package of R language can be freely obtained from CRAN (Comprehensive R Archive Network) through link: https://cran.r-project.org/. Then, follow the installation instructions to finish installation.

randomForest R package: The R package of randomForest can be freely obtained through link: https://cran.r-project.org/web/packages/randomForest/index.html. Then, follow the installation instructions to finish installation.

4. Installation

RiceMirP is designed in an easy-to-use manner, and is an installation-free software. Unpack the local package of riceMirP in your specified directory. Go to the specified directory, you can directly run riceMirP.

5. Usage of riceMirP

Input: RiceMirP requires three FASTA-formatted input files: the first two files containing nucleotide sequences of positive and negative training samples, respectively, and the last one containing nucleotide sequences of testing samples. Please make sure that the identifier for each sequence in FASTA-formatted files is unique. More details about FASTA format is obtained at http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml. All necessary data (including data used in this study) are contained in the local package of riceMirP.

Command: perl riceMirP.pl -P train_positive.fa -N train_negative.fa -T test.fa

Output: The output file of riceMirP includes three-column contents of the identifier, predicted label (positive or negative) and corresponding score for each testing sample. Please make sure that the identifier for each sequence in FASTA-formatted files is unique. Generally, higher score indicates that the testing sample is more likely to be predicted label.

Application scenarios: RiceMirP can be applied to sequence classification problem: for a given sequence, what is the likelihood of this given sequence to be positive. For miRNA prediction from small RNA sequencing data, riceMirP, in conjunction with another kind of miRNA biogenesis-based approaches, can further reduce the false positive rate.

6. Data

The training and testing datasets, and some datasets from other approaches are included in the local package of riceMirP.

(1). train_positive.fa & test_positive.fa

Pre-miRNA sequences of rice were downloaded from the miRBase database (Release 22.1). After removing sequences containing non-AGCU characters, 604 known pre-miRNAs of rice were collected as positive dataset. Then, 422 known pre-miRNAs are randomly selected as positive training dataset, and the other remaining 182 known pre-miRNAs are collected as independent positive testing dataset.

(2). train_negative.fa & test_negative.fa

Pseudo pre-miRNAs we selected from cDNAs. The CDS sequences of rice were obtained from PlantGDB database (http://www.plantgdb.org/download/Download/xGDB/OsGDB/Osativa_193_cds.fa.bz2), and then fragmented into non-overlapped segments under a constraint condition that the length distribution of extracted segments was identical with that of known plant pre-miRNAs. Further, they should satisfy some criteria. The criteria are determined by observing real plant pre-miRNAs. The criteria for selecting the pseud pre-miRNA are: minimum of 14 base pairings in the hairpins and maximum of −9.7 kcal/mol free energy of secondary structures (including GU wobble pairs). Finally, 502 and 216 pseudo pre-miRNAs were randomly selected as negative training dataset and testing dataset, respectively.

(3). datasets_from_miPlantPreMat

The training and testing datasets of miPlantPreMat, which included in the software of miPlantPreMat, were downloaded from website: https://github.com/kobe-liudong/miPlantPreMat.

(4). datasets_from_PlantMiRNAPred

The training and testing datasets of PlantMiRNAPred were downloaded from web site of PlantMiRNAPred: http://nclab.hit.edu.cn/PlantMiRNAPred/.