miRLocator Manual

Machine Learning-based Prediction of Mature MiRNAs within Pre-miRNA Sequences

Version 1.0 September 1st, 2015

Authors: Haibo Cui, Jingjing Zhai, Pan Zhu, Chuang Ma

Contact:

Dr. Chuang Ma: chuangma2006@gmail.com

1 Introduction

Although many computational tools have been developed to identify miRNA-related biological sequences, few can be used to predict mature miRNAs within premiRNA sequences.

Here we present a novel miRNA predictor named miRLocator, which is based on machine learning techniques and sequence and structural features extracted from miRNA:miRNA* dupexes. Prediction models in miRLocator were optimized by considering critical (and often ignored) factors that dramatically affect the prediction accuracy of mature miRNAs. Ten-fold cross-validation on 5854 experimentally validated miRNAs from 19 plant species demonstrated that MiRLocator can be used to accurately locate miRNA from pre-miRNAs.

miRLocator was written in python programming language, and is fast to perform the training and prediction processes.

Welcome to address any comments/suggestions/questions to: chuangma2006@gmail.com.

2 Download

miRLocator can be obtained from: https://github.com/cma2015/miRLocator.

3 Installation

miRLocator (v1.0) was developed based on Linux/Ubuntu OS with the necessity of pre-installing ViennaRNA package, python 2 programming environment and dependent libraries including sklearn, NumPy and SciPy.

(a) <u>ViennaRNA package</u>. miRLocator folds the secondary structure of premiRNAs using miRFold program in ViennaRNA package. For Ubuntu users, the most convenient way to install ViennaRNA package is to use Ubuntu PPA:

\$sudo apt-add-repository ppa:j-4/vienna-rna

\$sudo apt-get update

\$sudo apt-get install vienna-rna

The default installation location of RNAfold is /usr/bin/. More information about the installation can be found at http://www.tbi.univie.ac.at/RNA/.

- (b) Python 2 (version 2.7.6 or newer) and dependent libraries. The source codes of Python 2 can be downloaded from https://www.python.org/. Please note that there are many differences between Python 2 and Python 3 (https://wiki.python.org/moin/Python2orPython3). Errors might occur if you run it in Python 3 programming environment. Two dependent libraries (Numpy and Scipy) are required to perform scientific computation in python, which can be downloaded from: http://www.numpy.org/
- **(c)** <u>Scikit-learn library</u>. Scikit-learn (http://scikit-learn.org/stable/) is a python-based machine learning package for implementing the random forest algorithm. The guide for installation of scikit-learn is presented at: http://scikit-learn.org/stable/install.html/.

(d) <u>miRLocator_v1.0.tar.gz</u>. No need to install miRLocator. First decompress tar.gz file use the command: \$tar -xzvf miRLocator_v1.0.tar.gz. Then, in miRLocator_file_dir/miRLocator fold, you will find two python scripts (source.py and miRLocator.py) and three text files (trainingData.txt, predictionData_Annotated.txt) in miRLocator_file_dir/miRLocator/samples.

4 Implementation

(a) Parameter Setup. Open miRLocator.py with text editor, and define the value of following parameters:

RNAFoldDic: the file directory of RNAfold program

sourceDic: the file directory of miRLocator

<u>resultDic</u>: plase create a file directory named "results" in " sourceDic" for storing input data, intermediate and final prediction results. Read and write rights are needed for this file directory.

<u>cross_validation_flag</u>: a logical parameter indicates whether cross-validation will be performed. The cross-validation results can be used to reveal the prediction performance of miRLocator on training dataset.

<u>trainDataFileName</u>: the file name of training dataset (Figure 1). In this file (e.g., "trainingData .txt" in miRLocator_file_dir/miRLocator/samples), each line represents a miRNA, and contains four or five description items (miRNA identifier, pre-miRNA identifier, miRNA sequence, pre-miRNA sequence, pre-miRNA sequence, pre-miRNA secondary structure[not necessary]) separated with tab key. Please put this training data file into "resultDic" file directory.

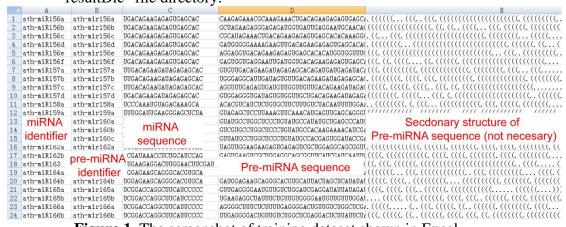


Figure 1. The screenshot of training dataset shown in Excel.

predDataFileName: the name of prediction file (e.g.,"predictionData.txt" in miRLocator_file_dir/miRLocator/samples) in which each line represents a pre-miRNA and contains two or three description items (pre-miRNA identifier, pre-miRNA sequence, pre-miRNA secondary structure [not necessary]). An example of prediction dataset is shown in Figure 2.

	A	В	C
1	ath-mir156a	CAAGAGAAACGCAAAGAAACUGACAGAAGAGAGUGAGCA	. ((((((((((((. ((((((((((((((((
2	ath-mir156b	GCUAGAAGAGGGAGAGAUGGUGAUUGAGGAAUGCAACAC	(((((((((((((((((((((((((((((((((((((((
3	ath-mir156c	CGCAUAGAAACUGACAGAAGAGAGUGAGCACACAAAGGC	. (((((. ((((((((((((((((((((((((((
4	ath-mir156d	GAUGGGGGAAAAGAAGUUGACAGAAGAGAGUGAGCACAC	. (((((((((((((((((((((((((((((((((
5	ath-mir156e	AGGAGGUGACAGAAGAGAGUGAGCACACAUGGUGGUUUU	(((, ((((((((((((((((((((((((((((((((((
6	ath-mir156f	GAGUGGUGAGGAAUUGAUGGUGACAGAAGAGAGUGAGCA	(((, ((, (((,,,,,((,,,,((,,(,((,(,((,(,(,(,(,(,(,
7	ath-mir157a	GUGUUGACAGAAGAUAGAGAGCACAGAUGAUGAGAUACA	(((. (((((((((((((((((((((((((((((((
8	ath-mir157b	UGGGAGGCAUUGAUAGUGUUGACAGAAGAUAGAGAGCAC	. (((((((), (((), (((), (((), ((), ((),
9	ath-mir157c	AGGUUUGAGAGUGAUGUUGGUUGUUGACAGAAGAUAGA((((((, ((((, (((, ((, (, (, (, (, (, (,
10	ath-mir157d	GUGGAGGGUGAUAGUGUGGUUGCUGACAGAAGAUAGAGA	(((((((, ((((, , ((((, , ((((((((((((((
11	ath-mir158a	ACACGUCAUCUCUGUGCUUCUUUGUCUACAAUUUUGGAA	secdonary structure of pre-miRNA
12	ath-mir159a	GUAGAGCUCCUUAAAGUUCAAACAUGAGUUGAGCAGGGU	
13	re-miRNA	GUAUGCCUGGCUCCCUGUAUGCCAUAUGCUGAGCCCAU(sequence (not necessary)
	1-1	GUCGUGCCUGGCUCCCUGUAUGCCACAAGAAAACAUCGA	(((((((, (((((, ((((((, ((, ((, (, (, (,
15	identifier	GUUAUGCCUGGCUCCCUGUAUGCCACGAGUGGAUACCGA	(((((((, ((((((, ((((((, ((((, (, ((((, (,
16	ath-mir162a	pro miDNA coguenco	. ((((((((((((,((((,(((,(((,((,((,((,((,(
17	ath-mir162b	pre-miRNA sequence	((((, ((((((((,(((,(((,((((,(((,((,((,((
18	ath-mir163	ACCCGGUGGAUAAAAUCGAGUUCCAACCUCUUCAACGAC	((, (((, (((((, (((((((((((((((((((((((
19	ath-mir164a	GGGUGAGAAUCUCCAUGUUGGAGAAGCAGGGCACGUGCA	(((((((,(((,(((,(((,((,(((,((,((,((,((,((,((,((,((,((,((,((,())

Figure 2. The screenshot of prediction dataset shown in Excel.

<u>predModelFileDir</u>: the file path of trained prediction model. If this parameter is designated, miRLocator will directly re-load the prediction model; otherwise, the training process will be implemented.

<u>predDataAnnotFileName</u>: the name of file containing annotation information for prediction dataset ("predictionData_Annotated" in miRLocator_file_dir/miRLocator/samples). The format of this file is same with that of training data file.

More information about these parameters is shown in Figure 3.

```
##the directory of RNAfold
                                                                                                                                                             file directory of RNAfold
                                /usr/bin/
##full path of source.py
                                                                                                                                                      file directory of miRLocator ###
sourceDic ="/home/cma/Researches/matureMiRNA/miRLocator/"
##result directory, include trainDataFileName, predDataFileName, predDataAnnotFileName
resultDic = sourceDic + "results/"
need to create a file of
need to create a fi
                                                                                                                                                  need to create a file directory ###
                                                                                                                                                    named results in sourceDic ###
##a logical parameter indicate whether run cross validation on training dataset
cross validation flag = False
##miRNAs and pre-miRNAs for training trainDataFileName = "trainingData.txt" Necessary for training prediction model at the first time###
##pre-miRNAs for prediction
##the full path of trained prediction model. If a full path is given,
##miRLocator will locate it directly, otherwise, miRLocator will run the training program.
predModelFileDir = "" [optional] use the alreadly trained prediction model
##If the file name is defined, miRLocator will evaluate the prediction results
                                                                                                                                                                                                                       ###
                                                                                                                                                                                                                       ###
##based on annotation infomation in this file
```

Figure 3. Important parameters in miRLocator

(b) Dataset Preparation

To train miRLocator, you need to prepare a file including training data (i.e., trainDataFileName). An example is given in miRLocator_file_dir/miRLocator/samples/trainingData.txt.

To perform prediction, you need to prepare a file containing prediction data (i.e., predDataFileName). An example is given in miRLocator_file_dir/miRLocator/samples/predictionData.txt.

(c) Running miRLocator

First, enter into the file directory of miRLocator_v1.0, and prepare training and prediction data.

\$cd miRLocator_file_dir ##run miRLocator in this file fold

\$mkdir results

\$copy ./samples/trainingData.txt ./results ##prepare a training data \$copy ./samples/predictionData.txt ./results ##prepare a prediction

##data[optional]

\$copy ./samples/predictionData_Annotated.txt ./results ##prepare annotation

##info for prediction
##data[optional]

\$python ./miRLocator.py

##run miRLocator

miRLocator finally generates three output files:

trained_prediction_model: The trained prediction model which can be

directly re-loaded at next time running.

miRLocator_predResults.txt: Predicted miRNAs for pre-miRNAs miRLocator_evalResults.txt: The evaluation results of miRLocator at different resolutions.