

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 pre-miRNA 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* duplexes. 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) ViennaRNA package. miRLocator folds the secondary structure of pre-miRNAs 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) Scikit-learn library. 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/>.

4 Implementation

RNAFoldDic: the file directory of RNAfold program

resultDic: please 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.

cross-validation will be performed. The cross-validation results can be used to reveal the prediction performance of miRLocator on training dataset.

[illegible]

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.

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.