Manual

RNAplonc: A tool for identication of plant long non-coding RNAs.

Pipeline

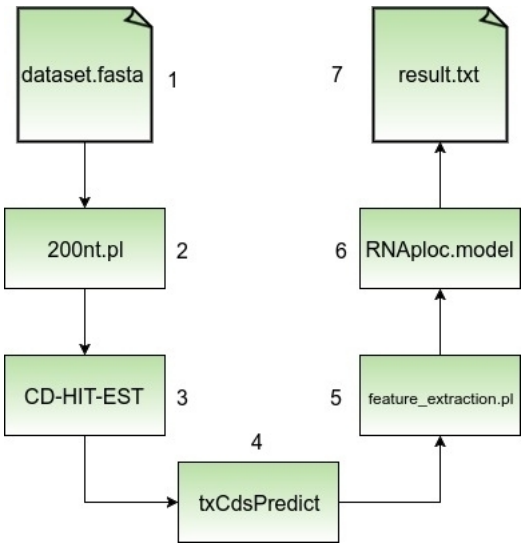

Figura 1:

Before to start, it is necessary to install CD-HIT-EST and txCdsPredict. After, it just unzips the RNAplonc.zip.

CD-HIT-EST is for taking similar sequences, if you want the whole result not to do this step. This step is optional, since the CD-hit takes the strings with redundancy.

PS: path is the location of the RNAplonc folder.

1 - dataset.fasta

Input of data: The input sequences must be in fasta format, according to the model below:

>12345xyz this is a nice sequence of the Foo gene
atgccatgataggactatttatttttctcactaccatcacccncacttaaagcatgggcggatttacta
>12345xyz this is a nice sequence of the Foo gene
atgccatgataggactatttatttttctcactaccatcacccncacttaaagcatgggcggatttacta

2 - 200nt.pl

The 200nt.pl le, compacted in RNAplonc.zip in the download section,will be executed to remove sequences smaller than 200 nucleotides.

Open the command terminal (Ctrl + Alt+ t)

Command to execute: perl path/200nt.pl path/le.fasta

OBS: path = Path of the le

The output le will have the same name as the input le with _ at the end, eg. dataset_.fasta

3 – CD-HIT-EST (optional)

The CD-HIT-EST program is presented in the download section.Your installation is explained in the install section. Its execution is a little slow, due to it removes all sequences with similarity of 80%.

It will run with the le resulting from step 2.

Command to execute: cd path

./cd-hit-est -i dataset_.fasta -o result.fasta -c 0.8
-i = Name of the output le from step 2
-o = Output le name
-c = Percentage cut used of 80% similarity

The output le will have the name you put after the -o

4 - txCdsPredict

The txCdsPredict program is presented in the download sectionYour installation is explained in the install section.

The le resulting from step 3 (result.fasta) will be entered in the txCdsPredict

Command to execute: cd path/kentUtils/src/hg/txCds/txCdsPredict/

./txCdsPredict result.fasta result.cds

result.fasta = Name of the output le from step 3 - CD-HIT-EST
result.cds = Output le name

5 - feature_extraction.pl

The entry in this step 5 will be the resulting les of steps 3 will be the results les of steps 3 - CD-HIT-EST and 4 - txCdsPredict

Command to execute: perl feature_extraction.pl result.fasta result.cds >result.ar

result.fasta = Name ofthe output le from step 3 - CD-HIT-EST
result.cds = Name of the output le from step 4 - txCdsPredict
The output le will have the extension .ar

6 - RNAplonc.model

The entry in this step 6 is the resulting le from step 5 - feature_extraction.pl

Command to execute: java -cp weka.jar weka.classifiers.trees.REPTree -l RNAplonc.model -T result.arff -p 0 >resultado_end.txt

The le weka.jar is in the download section in the compressed le RNAplonc.zip

7 - Filter result

Stars by using the output file from step 6 (resultado_end.txt), this novel script will help filter the output results for the final user.

* withouth filter optional parameters:

>python3 FilterResults.py -c result.cds -r resultado_end.txt -o resultFinal.txt

FilterResults.py = script to filter the lncRNA/mRNA results
-c = result.cds = Name of the output file from step 4 - txCdsPredict
-r = resultado_end.txt = Name of the output file from step 6
-o = resultFinal.txt = result display (result display)

Filtering optional parameters:

>python3 FilterResults.py -c result.cds -r resultado_end.txt -o resultFinal.txt -p 0.5 -t 1

-p = 0.5* = Filter the output percentage, float valeu between 0 and 1
-t = 1 = Filter the output by type: 1- lncRNA , 2-mRNA

* According to WEKA official website (https://waikato.github.io/weka-wiki/making_predictions/), prediction value is the probability to belong to that class. >=0.5 will be considered lncRNAs. User is free to choose their threshold. However, close to 1 you will get less False Positive results.

Any questions please contact us.