Novel microRNA discovery from genome-wide data: a computational pipeline with unsupervised machine learning

G. Stegmayer[1], C. Yones[1], L. Kamenetzky[2], N. Macchiaroli[2], M. Perez[2], M.C. Rosenzvit[2], D.H. Milone[1]

[1] Institute for Signals, Systems, Computational Intelligence (sinc(i)), FICH-UNL, CONICET, ARG.
[2] Instituto de Investigaciones en Microbiología y Parasitología Médica, IMPAM-UBA,CONICET, ARG.

Background:

There are several challenges related to the computational prediction of novel microRNAs (miRNAs), especially from genome-wide data and non-model organisms. First of all, many pre-processing steps on the raw data must be done to cut it into sequences, which involve the selection and use of a variety of software packages written in different programming languages, with many different possible configurations and parameters, most of the time unclear and very difficult to set by the final user. After that, each sequence must be analyzed one by one to classify it as possible candidate to pre-miRNA. The classical way of doing this has been training a binary supervised classifier with well-known pre-miRNAs (for example, extracted from miRBase) and artificially defining the no-pre-miRNA class, which is very difficult. Thus, a single, complete, and simple procedure for unsupervised pre-miRNA prediction from genome-wide data is of high interest today.

Results:

We have developed an integrated pipeline (Fig. 1) of just 5 simple steps, that starts from genome-wide data and can be applied for model and non-model organisms. First, an intelligent pre-processing cuts the genome into overlapped windows of nucleotides (sequences) with greater length than the mean pre-miRNA length of the species under analysis (or a phylogenetically related one, if well-known are not available). The secondary structure is predicted using RNAfold and pre-miRNA properties are verified, such as folding into stem-loops with a fixed value of minimum free energy. Multi-loop segments are split and duplicated stem-loops are deleted. If available, known RNA can be filtered as well. The remaining sequences go through a feature extraction process that can calculate all published features up to date. The resulting feature vectors are used for training an unsupervised machine learning model named miRNA-SOM, a deep architecture of several nested SOMs (Self organizing maps) that requires only positive labelled examples (the well-known pre-miRNAs). It clusters all the unlabelled sequences with the pre-miRNAs. miRNA-SOM allows for the quick identification of the best candidates to pre-miRNAs as those sequences clustered together with known precursors at the last level of the deep model.

We have performed a benchmarking test with the *Caenorhabditis elegans* full genome. The pipeline was applied obtaining 1,739,124 sequences. From miRBase v17, 200 well-known pre-miRNAs of *C. elegans* were used as positive labelled samples. After training, the unsupervised prediction model has been tested with the pre-miRNAs more recently added to miRBase v18-21 and absent in v17. In this test, 44 out of 48 have been identified as positive, resulting in a model sensitivity of 92%. The proposed pipeline was also applied for *Echinococcus multilocularis* and *Taenia solium* parasites genomes, allowing the effective discovery of 11 and 7 novel pre-miRNAs, respectively. These novel candidates have been even validated afterwards with “wet” experiments and RNAseq data.

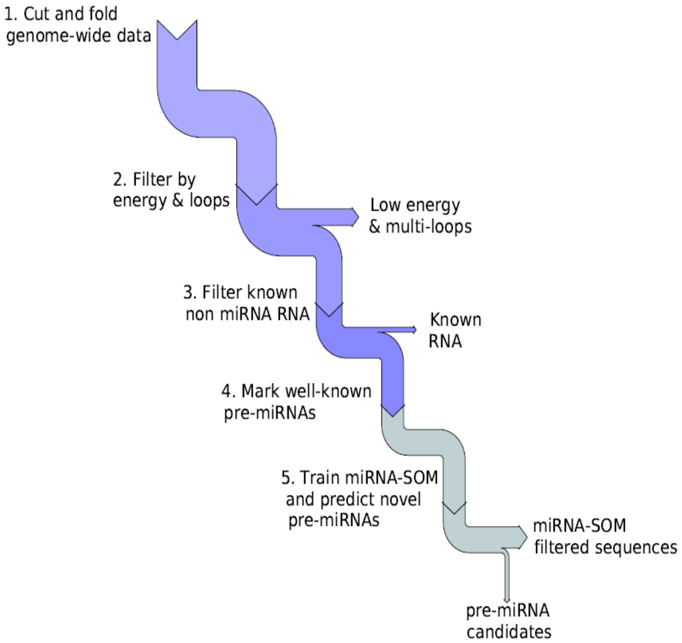
1. Cut and fold genome-wide data
2. Filter by energy & loops
Low energy & multi-loops
3. Filter known non miRNA RNA
Known RNA
4. Mark well-known pre-miRNAs
5. Train miRNA-SOM and predict novel pre-miRNAs
miRNA-SOM filtered sequences
pre-miRNA candidates

Figure 1: Computational pipeline for novel pre-miRNA prediction from genome-wide data

Conclusions:

We have described a pipeline that, receiving input genome-wide data and a set of well-known pre-miRNAs of a given organism, can automatically cut the genome into sequences, extract features and train an unsupervised machine learning model for novel pre-miRNAs prediction. It is based on the clustering of unlabelled sequences and well-known miRNA precursors for the organism under study. Novel pre-miRNAs have been effectively discovered with this methodology, which can help in the design of “wet” experiments that otherwise would be impossible to address.

Supported by: CONICET, MinCyT, UNL