Ab initio prediction of pri-miRNAs based on structural and sequence motifs

Renato Cordeiro Ferreira[1], Alan Durham[1]

1 IME

Abstract

MicroRNAs (miRNAs) are a category of small non-coding RNAs that help to regulate the translation process within the cell. They are originated from a long type of transcript called primary miRNAs (pri-miRNAs), which present a distinctive hairpin loop secondary structure and have a set of conserved motifs. Different proteins use these characteristics to distinguish pri-miRNAs from other similar molecules, so that they can generate the mature miRNAs from them. The aim of this project is to explore these patterns to create an ab initio pri-miRNA predictor. The first step to achieve this goal was to create a simple proof-of-concept classifier that used regular expressions and sequence alignment to select candidate pri-miRNAs. The program was tested on 467,100 segments of size 200 nucleotides (obtained with a sliding window of 100 nucleotides) from the human chromosome 21. It filtered a total of 29 sequences that matched the profile, 6 of which presented high similarity (alignment with e-value less than $10e^{-5}$ against the miRBase database) with sequences annotated in other human chromosomes. This result shows the potential of using these signals to identify likely candidates. The next step will be to implement a full probabilistic model, such as a Context-Sensitive Hidden Markov Model (csHMM), to identify pri-miRNAs. A csHMM will be able to describe the long-range dependencies between positions in the hairpin loop, besides encoding the distribution of nucleotides obtained from real training examples. This way, we expect to create an automatic way to find candidate miRNAs that have not been experimentally observed yet.

Funding: CAPES