Improving pre-miRNA prediction with complexity measures of the mature and deep learning

Jonathan Raad, Georgina Stegmayer and Diego H. Milone

Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL, CONICET,

Santa Fe, Argentina.

Background:

The miRNAs are small RNA molecules that regulate gene expression in animal and plant cells through post-transcriptional control. They are stored inside precursors of 100 bases long approximately called pre-miRNAs, which have a stem-loop structure. Several experimental methods for detecting pre-miRNAs can be used , such as qPCR, microarray and deep sequencing. However, these techniques present some practical difficulties when evaluating a very large number of candidate sequences in a genome. Due to these technical and practical difficulties, computational methods play an increasingly important role for their prediction . In order to find new candidates for pre-miRNA, many different features sets have been proposed, which mostly describe information of the structure of the pre-miRNA inspired by the action of Drosha. However, the specificity of the subsequent processes impose restrictions on those hairpins that will become mature miRNA. Given that this important information is codified in the mature region, the secondary structure of the precursor by itself might not be sufficient to differentiate a true pre-miRNA from other hairpins.

Results:

In this work, we have developed a new feature for the mature sequences representation based on the Levenshtein distance, which is a string metric for measuring the edit difference between two sequences. Furthermore, this new feature combined with deep learning, has proven to be able to improve significantly the prediction of pre-miRNAs. We have developed a deep neural network (DNN) and train it with and without the new feature using cross validation, and the results obtained indicate that this model was able to improve the separation of classes even in the presence of very high imbalance in the data. Figure 2 shows the classification results (sensibility, precision and F_1) for the new proposed feature and the standard features, with DNN as classifier. The Figure clearly shows how the DNN classifier is capable of maintaining high performance at increasing imbalances, and even increasing both sensibility and precision when the new feature is used. Moreover, it is observed that F_1 is significantly higher for all the imbalances when the new feature is used, increasing from 30% to 80% compared to standard features at the highest imbalance.

Conclusions:

The results showed that the incorporation of mature information with the new Levenshtein distance feature has a high discriminative power, significantly decreasing the number of false positives.

Submission topic: Genomics, Metagenomics and Proteomics

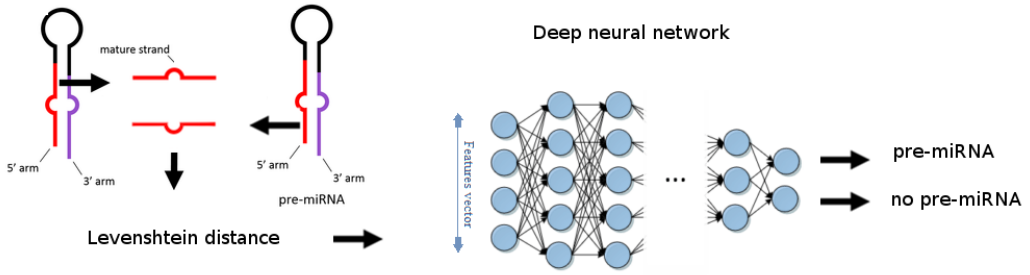
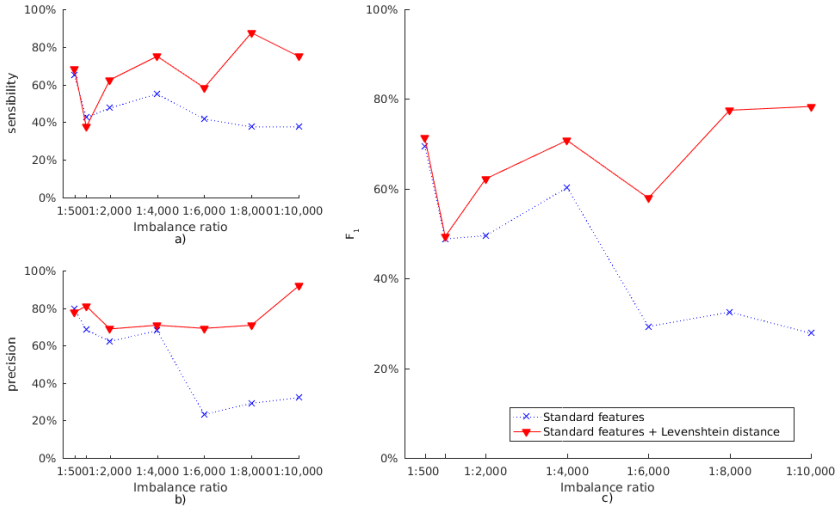


Fig. 1. Levenshtein distance feature extraction

Fig. 2. Results of deep neural networks classifier with standard features (SF) and Levenshtein distance (LD). a) Sensibility;



sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
J. Raad, D. H. Milone & G. Stegmayer; "Improving pre-miRNA prediction with complexity measures of the mature and deep learning"
A2B2C 10th Meeting, 2019.