MicroRNA prediction from genome-wide data with deep learning: a novel approach based on convolutional residual networks

C. Yones, L.A. Bugnon, J. Raad, D.H. Milone and G. Stegmayer

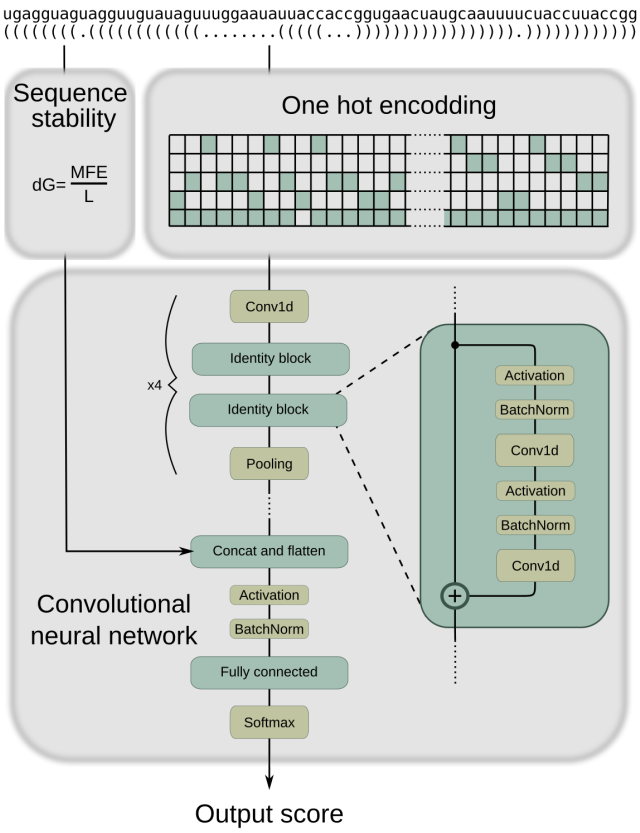
Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH/UNL-CONICET

Background

The relevance and importance of miRNAs has been widely acknowledged by the community given their fundamental role in gene regulation, by promoting or inhibiting certain diseases and infections. The computational prediction of novel miRNAs involves identifying pieces of a genome with the highest chance of being miRNA precursors (pre-miRNAs). Although several machine learning (ML) classifiers can be used for this task, with high classification performance, most predictors depend heavily on the selection of the features used to represent the sequences under analysis. Moreover, just very few pre-miRNAs are known in any genome, in comparison to the hundreds of thousands of potential candidates to novel miRNAs, which results in very high class imbalance problem to be tackled by a ML model. In fact, most of the state-of-the-art methods cannot work with raw genome-wide data as input, and apply some kind of artificial class balancing, which makes them unsuited for finding novel pre-miRNAs in a real context.

Results

Here we present mirDNN, a nucleotide-level convolutional neural network (CNN) model for pre-miRNAs discovery (see figure). By using one-hot encoding and padding, all genome sequences are converted into numeric matrices. Then, identity blocks composed of convolution operations can automatically extract features from the sequences. Finally, a fully connected network generates a prediction for each test sequence (candidate pre-miRNA, or not). To tackle the imbalance problem during training, a combination of over-sampling and a loss function specially designed for imbalance learning is used. In addition, we propose a novel training and validation scheme, which is more realistic because it takes into account the existence of homologous miRNAs among species .



MirDNN was evaluated on 5 model species and the area under the precision recall curve (AUPRC) was measured. The results showed that mirDNN outperformed state-of-the-art methods, with AUCPRs above 80%. This means more than ten times the precision of traditional machine learning methods. Also, mirDNN was compared with other methods based on sequence alignment to show the convenience of using an automatically “learned” sequence distance instead of generic alignment distances.

Conclusions

This work provided a novel and very effective approach for dealing with the computational prediction and discovery of di-novo pre-miRNAs: convolutional deep residual neural networks.