Deep neural learning to predict mRNA editing

Jesús Peñuela[1], Michal Zawisza[2], Carlos Herrera[2], Ferran Reverter[2], Esteban Vegas[2], Jordi García[2]

[1]jpenuela@uoc.edu. Open University of Catalonia
[2]mihizawi@gmail.com, carlos.herrera@me.com,freverter@ub.edu,evegas@ub.edu,jordigarcia@ub.edu. Department of Genetics, Microbiology and Statistics, University of Barcelona

RNA-editing is a molecular mechanism that performs chemical changes to specific nucleotides in RNA molecules of eukaryote cells and is one of several posttranscriptional mechanisms that adds versatility to the transcriptome. In metazoa, the most usual form of editing is the change from adenine to inosine (A-to-I) performed by proteins of the ADAR family. Although the mechanisms by which ADAR proteins target specific adenosine the regulation of editing activity is not yet fully understood, it has been shown that this regulation plays a key role in several biological processes. In vertebrates, the regulation of specific editing events is an important factor in changes of gene expression during the development of the neural system. Furthermore, certain brain-expressed proteins present very significant differences in editing across different vertebrate clades.

The initial data are genomic sequences annotated with the secondary structures of the pre-mRNA and with the editing positions that were generated by the information tools developed by M. Zawisza. The data generation process is based on three input files: RNA Editing data extracted from REDIPortal, the fasta file of the reference genome and the gene annotation GTF file. After having all the pre-mRNA sequences, the secondary structures are predicted by LinearFold. At the end of this process, the initial data file is generated, which contains all the pre-mRNA sequences along with the prediction of their secondary structure and the annotations of the edited adenosine positions.

The genomic data files had to be processed properly in order to train the deep learning models. For this purpose, we have used fixed length window sequences, centered on each of the adenosines for a given gene. Window sequences of different lengths were tested, finding that the best results were obtained generating windows of $50 + 1 + 50$ nucleotides.

In this work we have used LSTM neural networks with an attention layer. The attention layer is capable of assigning different weights to different positions in each input sequence, seeking to give more relevance to the positions that are more decisive when classifying the sequence.

We analyzed A-to-I RNA editing in the human genome. In balanced scenarios with 800,000 windows sequences, we obtained 93% Accuracy, 93% F1, 0.854 Kappa and 0.95 AUC. We have done preliminary studies in mice and mackerel, and we have achieved 84% and 72% Accuracy, respectively.

Keywords: RNA Editing, LSTM, Attention mechanism.