Automatic identification of lncRNA transcripts using Artificial Neural Networks

Ana Beatriz Oliveira Villela Silva, Mariana Carmin, Eduardo Jaques Spinosa

Universidade Federal do Paraná

Abstract

A vast number of studies show that only 1.5% of the human genome effectively encodes proteins. Among the remaining 98.5%, a large portion is transcribed as ncRNA (non-coding RNAs), intermediate molecules that participate in the most diverse biological processes. LncRNAs (long non-coding RNAs) are a particularly recently discovered class of ncRNA which contains at least 200 nucleotides in length. Since they can be as long as (or even longer) than mRNAs (messenger RNAs) and also can contain an ORF (Open Reading Frame), differentiating between those two very functionally different types of sequences can be a challenging task. This work proposes a method to discriminate DNA sequences between mRNA and lncRNA transcripts in Human and Mouse genomes using Artificial Neural Networks. In order to train the network, 10000 (human) and 5000 (mouse) sequences from both mRNAs and lncRNAs were manually curated from the GENCODE database. The validation set for both organisms were divided into Test-A and Test-B, with Test-B's difference being the removal of lncRNA sequences that were too similar from mRNAs. A multilayer perceptron was the model adopted to classify the data. In total, nine features represented each sample in the dataset: the number and size of exons, the size of the transcript, the beginning and end of any existing ORFs, the size of the ORF found, a score calculated to represent the quality of the ORF, and the standard deviation based of the number of possible ORFs considering all possible reading frames. A grid search was performed in order to optimize the following hyperparameters: initial learning rate, number and amount of hidden layers, and the maximum number of epochs. The accuracy on both validation datasets was between 82, 36% and 93, 43% for humans and 88, 59% and 94, 18% for mice, with a precision ranging from 91, 77% and 92, 15% for humans and 93, 06% and 94, 13% for mice respectively. The recall achieved varied between 82, 36% and 83, 62% for humans and 88, 59% and 88, 89% for mice. In conclusion, a multilayer perceptron showed to be a valid classification model to discriminate lncRNAs and mRNAs, especially in Test-B, but the overall recall metric of this study can be further improved. Future work can be done using Deep Neural Network approaches in order to achieve even better results.

Funding: