A novel approach for RNA folding inference based on message-passing graph neural networks

M. Gerard and L. Di Persia
Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH/UNL-CONICET, (3000) SF, ARG

Background:
Nowadays, prediction of secondary structure of RNA is an open challenge. In simple terms, it consists of identifying which nucleotides in the sequence are paired, without considering the backbone. Classical methods based on thermodynamics are typically used for this prediction. More recently, a wide range of deep learning methods have appeared to compete with them, achieving increasingly better results. However, those methods need the training of millions of learnable parameters, which require vast amounts of data, and make those models prone to overfitting.

Results:
Here, we present a novel approach for RNA secondary structure prediction from its sequence. In contrast to other deep learning methods, our proposal uses a small fraction of the learnable parameters used for those neural models. To achieve this goal, our approach first transforms the folding problem into a classification one, where the aim is to classify connections between pairs of nucleotides (nts) as feasible or not. First, the problem is modeled as a graph, where each node describes a particular nucleotide and has an associated set of features, and possible connections are modeled with edges linking them. Then this graph is inverted, turning nodes into arcs and vice versa. On this new graph, where the nodes now represent connections between nucleotides, a message-passing neural model [1] is applied that learns new features that are then used to classify the active connections between nucleotides (nodes on this graph). Preliminary results for this model (~40000 parameters), employing *k*-fold cross-validation (k=5) with archiveII dataset with sequences up to 200 nts, yielded F1 = 0.848 (std: 0.015). In comparison, the state-of-the-art model UFold [2] (8641377 parameters, as stated in [2]) produces F1 = 0.914 (std: 0.036) for the same experiment.

Conclusions:
Our novel approach for predicting RNA secondary structures demonstrates a good performance with minimal parameter usage (~0.4% of those used by UFold), showcasing efficiency in RNA structure prediction. This preliminary result shows our proposal is able to make good predictions using a very small number of parameters. However, it is clear that further work and experiments are needed to improve the performance obtained with this model.

[1] Gilmer, J., Schoenholz, S.S., Riley, P.F., Vinyals, O., Dahl, G.E. (2020). *Message Passing Neural Networks*. In: Schütt, K., Chmiela, S., von Lilienfeld, O., Tkatchenko, A., Tsuda, K., Müller, KR. (eds) Machine Learning Meets Quantum Physics. Lecture Notes in Physics, vol 968. Springer, Cham. https://doi.org/10.1007/978-3-030-40245-7_10

[2]Laiyi Fu, Yingxin Cao, Jie Wu, Qinke Peng, Qing Nie, Xiaohui Xie, *UFold: fast and accurate RNA secondary structure prediction with deep learning*, *Nucleic Acids Research*, Volume 50, Issue 3, 22 February 2022, Page e14, https://doi.org/10.1093/nar/gkab1074