SincFold: a new tool for RNA folding based on deep learning

L.A. Bugnon[1], L. Di Persia[1], M. Gerard[1], J. Raad[1], S. Prochetto[1,2], E. Fenoy[1], U. Chorostecki[3], F. Ariel[2], G. Stegmayer[1], D.H. Milone[1]

1 Research Institute for Signals, Systems and Computational Intelligence, sinc(i), FICH-UNL, CONICET, Ciudad Universitaria UNL, 3000 Santa Fe, Argentina.

2 Instituto de Agrobiotecnología del Litoral, CONICET-UNL, CCT-Santa Fe, Ruta Nacional N° 168 Km 0, s/n, Paraje el Pozo, 3000 Santa Fe, Argentina.

3 Faculty of Medicine and Health Sciences, Universitat Internacional de Catalunya, 08195 Sant Cugat del Vallès, Barcelona, Spain.

Background: Non-coding RNAs fold into well-defined secondary structures, which defines their functions. However, its computational prediction from a raw sequence is a long-standing unsolved problem, without significant changes in performance in the last decades. Traditional RNA secondary structure prediction algorithms are based on thermodynamic models and dynamic programming for free energy minimization. More recently, deep learning methods have appeared.

Results: We present sincFold, an end-to-end deep learning approach that predicts the contact matrix among nucleotides in a sequence using only the RNA sequence as input. The model is based on hierarchical 1D-2D residual neural networks that can learn short- and long-range interaction patterns. Extensive experiments on several benchmark datasets were conducted, comparing sincFold against classical methods and recent deep learning models. Results show that sincFold can outperform state-of-the-art methods on all datasets assessed. The source code is available at https://github.com/sinc-lab/sincFold and an online demo is provided at https://huggingface.co/spaces/lbugnon/sincFold

Conclusions: SincFold has proven to be better suited to identify structures that might defy traditional modeling. Even when there is a small number of examples of some RNA families to learn from (around hundreds of sequences), sincFold can learn the structures with high performance. As more examples are available, the generalization capability of the model is significantly better than other methods. Results also show that sincFold, thanks to its capability for capturing a wide range distances in interactions, is significantly better than all other methods for the secondary structure prediction in longer ncRNA sequences.