Improving the folding prediction of RNA with deep learning

L.A. Bugnon, L. Di Persia, M. Gerard, A. Edera, J. Raad, S. Prochetto, E. Fenoy, G. Stegmayer and D.H. Milone.

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

Background

The function of noncoding RNAs (ncRNAs) is relevant for numerous biological processes and largely depends on their secondary structure, which determines interactions with partner molecules. However, the determination of RNA structures is a very costly process, which cannot be scaled up efficiently, limiting our ability to functionally characterize such molecules. Computational methods are promising for the prediction of RNA structures, which is speeding up the discovery of function and action mechanisms. Since classical tools strongly rely on hand-crafted thermodynamic features, they show limited capacity for modeling the wide structural diversity of RNAs, including pseudoknots, non-canonical linkages and long sequences. Recently, machine learning techniques have demonstrated being able to capture thermodynamic features in an automated manner.

Results

We have compared several recent methods for secondary structure prediction, including classical and newer ones based on deep learning, using 3023 yeast sequences and a novel benchmark of well-characterized long ncRNA from different species. Classical methods can guarantee a 50%-70% prediction accuracy in all the range of RNA in the dataset. However, their weakness is that their results have stagnated in the last 20 years. Despite being relatively new, machine learning based methods have achieved similar predictive performance to the classical ones for long ncRNAs. We propose a new architecture for secondary structure prediction based on machine learning. The network first encodes information about each nucleotide and its neighbors from which nucleotide-to-nucleotide information is then used to predict the connection matrix. This allows representing any type of connection including pseudoknots. This architecture was evaluated on a curated dataset obtained from RNAs with experimentally verified structures from public repositories. Preliminary results show that our proposal can reach a high performance.

Conclusions

In this work, we identify current limitations of ncRNA folding prediction tools, especially with long ncRNA, and propose new methods to improve them. Using a curated dataset and a benchmark to test model generalization, promising results have been achieved. Computational limitations will be considered in order to process large sequences (>1000nt).

Submission topic

Genomics, Transcriptomics, and Metagenomics.
Machine Learning.