Secondary structure prediction by combination of formal grammars and neural networks

Semyon Grigorev[1, 2*], Dmitry Kutlenkov[1, 2], Polina Lunina[1, 2]

[1]Saint Petersburg State University, St. Petersburg, 199034, Russia
[2]JetBrains Research, St. Petersburg, 197374, Russia

*semyon.grigorev@jetbrains.com

Secondary structure is known to have a crucial impact on RNA molecules functioning, therefore, development of algorithms for secondary structure modeling and prediction is a fundamental task in computational genomics. Among other methods, secondary structure can be theoretically described by means of formal grammars [1, 2].

An approach for sequences secondary structure analysis by combination of formal grammars and neural networks was proposed in [3, 4]. In this work, we apply this approach to RNA secondary structure prediction. Secondary structure can be described as composition of stems having different heights and loop sizes [5]. We use context-free grammar from [3] to encode the most common kinds of stems and parsing algorithm [6] to find such stems in sequences. Note that this grammar describes only the classical base pairs and cannot express pseudoknots. The result of a matrix-based parsing algorithm for some sequence is a boolean matrix that represents all the theoretically possible stems in terms of grammar, but the real secondary structure cannot contain all of them at once and, besides, there can be more complex, not expressible in given grammar elements. Therefore, parsing matrices require further processing and we propose to use a neural network to handle them in order to generate an actual secondary structure.

For experimental research we took sequences from RnaCentral [7] database and as reference data for network training we used the output of CentroidFold tool [8]: contact matrices that represent connections between nucleotides in secondary structure. We transformed parsing matrices and contact maps to black-and-white images. These images were used for training the generative neural network which takes a parsing-provided image as an input and transforms it to the maximal approximation of the considered contact map. We applied deep residual networks with the local alignment algorithm at the end of the sequence of layers.

We trained models with and without alignment on several datasets with fixed sequence length interval and estimated them by precision, recall and F1 score metrics calculated for numbers of correctly and incorrectly guessed contacts for each image. All models showed F1 score up to 70% and we discovered that the smaller the window size, the more accurate the model, moreover, alignment significantly improves precision of neural networks due to removing the contacts that break the secondary structure.

To conclude, the set of experiments confirmed that the proposed approach is applicable to secondary structure prediction problem and further research is required.

Acknowledgments
The research was supported by the Russian Science Foundation grant 18-11-00100 and a grant from JetBrains Research.

References

1. Dowell RD, Eddy SR. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. BMC Bioinformatics. 2004;5(1):71.
2. Knudsen B, Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics. 1999 Jun 1;15(6):446-54.
3. Grigorev S, Lunina P. The composition of dense neural networks and formal grammars for secondary structure analysis. In De Maria E, Gamboa H, Fred A, editors, BIOINFORMATICS 2019 - 10th International Conference on Bioinformatics Models, Methods and Algorithms, Proceedings; Part of 12th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2019. SciTePress. 2019. p. 234-241

4. Grigorev S, Lunina P. Improved Architecture of Artificial Neural Network for Secondary Structure Analysis. BMC Bioinformatics. 2019;20(S17). P2..
5. Quadrini M, Merelli E, Piergallini R. Loop Grammars to Identify RNA Structural Patterns. In: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies [Internet]. SCITEPRESS - Science and Technology Publications; 2019.
6. Azimov R, Grigorev S. Context-free path querying by matrix multiplication. In Bhattacharya A, Fletcher G, Roy S, Arora A, Larriba Pey JL, West R, editors, Proceedings of the 1st ACM SIGMOD Joint International Workshop on Graph Data Management Experiences and Systems (GRADES) and Network Data Analytics (NDA), GRADES-NDA 2018. Association for Computing Machinery. 2018. a5. (Proceedings of the 1st ACM SIGMOD Joint International Workshop on Graph Data Management Experiences and Systems (GRADES) and Network Data Analytics (NDA), GRADES-NDA 2018).
7. Sweeney BA, Petrov AI, Burkov B, Finn RD, Bateman A, Szymanski M, et al. RNAcentral: a hub of information for non-coding RNA sequences. Nucleic Acids Research. 2019 Jan 8;47(D1):D221-D229.
8. Hamada M, Kiryu H, Sato K, Mituyama T, Asai K. Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics. 2009 Feb 15;25(4):465-73.