Secondary Structure Prediction by Combination of Formal Grammars and Neural Networks*

Semyon Grigorev, Dmitry Kutlenkov, Polina Lunina

Saint Petersburg State University

7/9 Universitetskaya nab., St. Petersburg, 199034, Russia

JetBrains Research

Primorskiy prospekt 68-70, Building 1, St. Petersburg, 197374, Russia

semyon.grigorev@jetbrains.com, kutlenkov.dmitri@gmail.com, lunina_polina@mail.ru

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. Comparative methods of secondary structure prediction analyse several homologous sequences employing evolutionary approaches, while single sequence methods process one sequence at a time, and the most popular approach here is to find the minimum free energy structure [1, 2]. Also, secondary structure can be theoretically described by means of formal grammars [3, 4].

An approach for sequences secondary structure analysis by combination of formal grammars and neural networks was proposed in [5, 6]. In this work, we investigate the possibilities of this approach for RNA secondary structure prediction. Secondary structure can be described as composition of stems having different heights and loop sizes [7]. We use context-free grammar G_0 from [5, 6] to encode the most common kinds of stems and parsing algorithm [8] to find the subsequences of sequence that should fold to such stems.Note that this grammar describes only the classical base pairs and

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

cannot express pseudoknots. The result of a matrix-based parsing algorithm for string w is a boolean matrix M, where $M[i, j] = 1$ iff the substring $w[i, j-1]$ is derivable from grammar start nonterminal, i.e. folds to a stem. Suchwise we get a representation of all the theoretically possible stems in terms of G_0, but the real secondary structure cannot contain all of them at once and, besides, there are more complex elements that are not expressible in given grammar. Therefore, parsing matrices require further processing and we propose to use neural network to handle them in order to generate an actual secondary structure.

For experimental research we took sequences from RnaCentral [9] database and as a reference data for network we used contact matrices generated by CentroidFold tool [2], where $[i, j]$ element of matrix is equal to 1 iff there is a connection between nucleotides i and j in secondary structure. We transformed parsing matrices and contact maps to black-and-white images, where white pixel in position $[i, j]$ corresponds to 1 in matrix and black — to 0. These images were used for training the generative neural network which takes 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 sequence of layers.

We trained models on several datasets with fixed sequences length interval. Trained models were estimated by precision, recall and F1 score metrics calculated for numbers of correctly and incorrectly guessed contacts for each image. Results for different sequences lengths for models with and without alignment are presented in the table 1. It is shown that the smaller the window size, the more accurate the model. Also, alignment significantly improves precision of neural networks due to removing the contacts that break the secondary structure.

Sequence length	Alignment	Precision	Recall	F1 score
90	×	67%	75%	68%
	✓	80%	66%	70%
88-90	×	66%	78%	69%
	✓	81%	62%	68%
50-90	×	60%	72%	63%

Table 1: Test results for all trained models

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

References

[1] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster, "Fast folding and comparison of rna secondary structures," *Monatshefte für Chemie/Chemical Monthly*, vol. 125, no. 2, pp. 167–188, 1994.

[2] M. Hamada, H. Kiryu, K. Sato, T. Mituyama, and K. Asai, "Prediction of rna secondary structure using generalized centroid estimators," *Bioinformatics*, vol. 25, no. 4, pp. 465–473, 2009.

[3] R. D. Dowell and S. R. Eddy, "Evaluation of several lightweight stochastic context-free grammars for rna secondary structure prediction," *BMC bioinformatics*, vol. 5, no. 1, p. 71, 2004.

[4] B. Knudsen and J. Hein, "Rna secondary structure prediction using stochastic context-free grammars and evolutionary history.," *Bioinformatics (Oxford, England)*, vol. 15, no. 6, pp. 446–454, 1999.

[5] S. Grigorev and P. Lunina, "The composition of dense neural networks and formal grammars for secondary structure analysis," in *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies*, vol. 3, pp. 234–241, 2019.

[6] S. Grigorev and P. Lunina, "Improved architecture of artificial neural network for secondary structure analysis," *BMC Bioinformatics*, vol. 20, no. S17, 2019.

[7] M. Quadrini., E. Merelli., and R. Piergallini., "Loop grammars to identify rna structural patterns," in *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 3: BIOINFORMATICS,*, pp. 302–309, INSTICC, SciTePress, 2019.

[8] R. Azimov and S. Grigorev, "Context-free path querying by matrix multiplication," in *Proceedings of the 1st ACM SIGMOD Joint International Workshop on Graph Data Management Experiences & Systems (GRADES) and Network Data Analytics (NDA)*, GRADES-NDA '18, (New York, NY, USA), pp. 5:1–5:10, ACM, 2018.

[9] T. R. Consortium, "RNAcentral: a hub of information for non-coding RNA sequences," *Nucleic Acids Research*, vol. 47, pp. D221–D229, 11 2018.