**Improved Architecture of Artificial Neural Network for Secondary Structure Analysis**

Semyon Grigorev1, 2\*, Polina Lunina1, 2

1Saint Petersburg State University, St. Petersburg, 199034, Russia

2JetBrains Research, St. Petersburg, 197374, Russia

\*semyon.grigorev@jetbrains.com

The idea about using secondary structure while solving different sequences analysis problems is considered in many works (1-4). One of the classical ways of describing secondary structure is formal grammars.

An approach for biological sequences processing by combination of formal grammars and neural networks is proposed in the work (5). While classical way is to model secondary structure of the full sequence by using grammar, the proposed approach utilizes it only for primitive secondary structure features description. These features can be extracted by parsing algorithm and processed by neural network. It is shown that this approach is applicable for real-world data processing and several questions are formulated for future research. In this work we provide answers to some of them.

The first question is whether it is possible to use convolutional neural networks for parsing result processing. The result of matrix-based parsing algorithm for some string and fixed nonterminal is an upper-triangular boolean matrix. The first way to represent this matrix is to drop out the bottom left triangle and vectorize the rest of matrix row by row. The vector can be handled by using dense neural network. It requires the equal length of the input sequences, therefore, we propose to cut sequences or add some special symbol till the definite length. The second way is to represent the matrix as an image: the false bits as white pixels and the true bits as black ones, so, we can process sequences with different length and resize images afterwards. To handle these images we use network with a small number of convolutional layers. After linearization we use the same network as for vectors handling.

The second question is whether it is possible to move parsing to network training step, because parsing is the most time-consuming operation of our solution. We solve this problem by using two-staged learning. Firstly, we prepare a network which takes parsed data as an input. After that we extend trained network with a number of input layers that convert original nucleotide sequence into parsing result. This network handles sequences, so, we require parsing only for training the first network.

We use these improvements for tRNA sequences analysis problems: classification into two classes (eukaryotes, prokaryotes) and four classes (archaea, bacteria, plants, fungi). Accuracy for image- and vector based classifiers is up to 90% and accuracy for extended networks is up to 95% on the test dataset.

**Acknowledgements**

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

**References**

1. Rivas E, Eddy SR. The language of RNA: a formal grammar that includes pseudoknots. Bioinformatics. 2000; 16(4): 334-340.

2. Knudsen B, Hein J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics (Oxford, England). 1999; 15(6): 446–454.

3. Yuan C, Lei J, Cole J, Sun Y. Reconstructing 16S rRNA genes in metagenomic data. Bioinformatics. 2015; 12: 135-143.

4. Dowell RD, Eddy SR. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. BMC bioinformatics. 2004; 5(1): 71.

5. Grigorev S, Lunina P. The Composition of Dense Neural Networks and Formal Grammars for Secondary Structure Analysis. Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 3: BIOINFORMATICS. 2019; 234-241.