# Title

Semyon Grigorev, Polina Lunina
Saint Petersburg State University
7/9 Universitetskaya nab., St. Petersburg, 199034, Russia
semyon.grigorev@jetbrains.com, lunina_polina@mail.ru

Algorithms that can efficiently perform sequences classification and subsequences detection have recently become a focus in bioinformatics and many of them utilize the idea about considering these sequences secondary structure [?, ?, ?, ?]. 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 proposed in the work [?]. While classical way is to model secondary structure of the full sequence by using grammar, the proposed approach utilizes grammar only for primitive secondary structure features description. These features can be extracted by parsing algorithm and processe by using artificial neural network. It is shouwn that this approach is applicabe for real-world data processing, and some questions are formulated for future research. In this work we answer 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 an input string and fixed nonterminal is an upper triangular boolean matrix. Presently, we came up with two possible ways of these matrices representation. The first one is to drop out the bottom left triangle, vectorize the rest of matrix row by row and transform it to the vector. It requires the equal length of the input sequences, therefore we propose to either 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 of matrix as white pixels and the true bits as black ones. This approach makes it possible to process sequences with different length since the images could be easily resized to the same size. To handle these images we use network with small number of convolutional layers, linearization and dense laysers with same structure as for vectorized data.

The second qiestion is whether it is possible to move parsing network trainig step. This question is important because parsing is the most time-consuming operation.

We solve this problem by using two staged learning. At the first step, we prepare a neural network for our task (vector or image based) which takes parsing data as an input. After that we extend trained network with a number of input layers which should convert original nucleotide sequence into parsing result. This way create a netwok which can handle sequences, not parsing result. So, parsing is required only for network training.

We use proposed improvements to create networks for tRNA sequences analysis problems: classification of tRNA into 2 classes (eukaryotes and prokaryotes) and 4 classes (archaea, bacteria, plants, and fungi). We train networks on !!!! sequences from !!! database and test it on !!!!. Results for both image- and vector-based classifiers are presented in the table **??**. Base model meens network which handle parsing resut, extended wodel handles sequences and is based on appropriate base model.

| | | Base model accuracy | Extended model accuracy |
|---|---|---|---|
| Eukaryotik/prokaryotik classifier | Vector-based | 94.1% | 97.5% |
| | Image-based | 96.2% | 97.8% |
| Archaeal/bacterial/fungi/plant classifier | Vector-based | 86.7% | 96.2% |
| | Image-based | 93.3% | 95.7% |

Table 1: Test results

# References

[1] Rivas E, Eddy S.R. *The language of RNA: a formal grammar that includes pseudoknots* // Bioinformatics. — 2000.

[2] Knudsen Bjarne, Hein Jotun. *RNA secondary structure prediction using stochastic context-free grammars and evolutionary history.* //Bioinformatics (Oxford, England).— 1999.— Vol. 15, no. 6.— P. 446–454.

[3] Yuan C. et al. *Reconstructing 16S rRNA genes in metagenomic data* //Bioinformatics. – 2015. – №. 12. – P. 135-143.

[4] Dowell R. D., Eddy S. R. *Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction* //BMC bioinformatics.– 2004.– №. 1.– P. 71.

[5] Grigorev S., Lunina P. *The Composition of Dense Neural Networks and Formal Grammars for Secondary Structure Analysis*