Secondary Structure Prediction by Combination of Formal Grammars and Neural Networks

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

[1] *Saint Petersburg State University, JetBrains Reserach, St. Petersburg, Russia*

E-mail: *lunina_polina@mail.ru, kutlenkov.dmitri@gmail.com, semyon.grigorev@jetbrains.com*

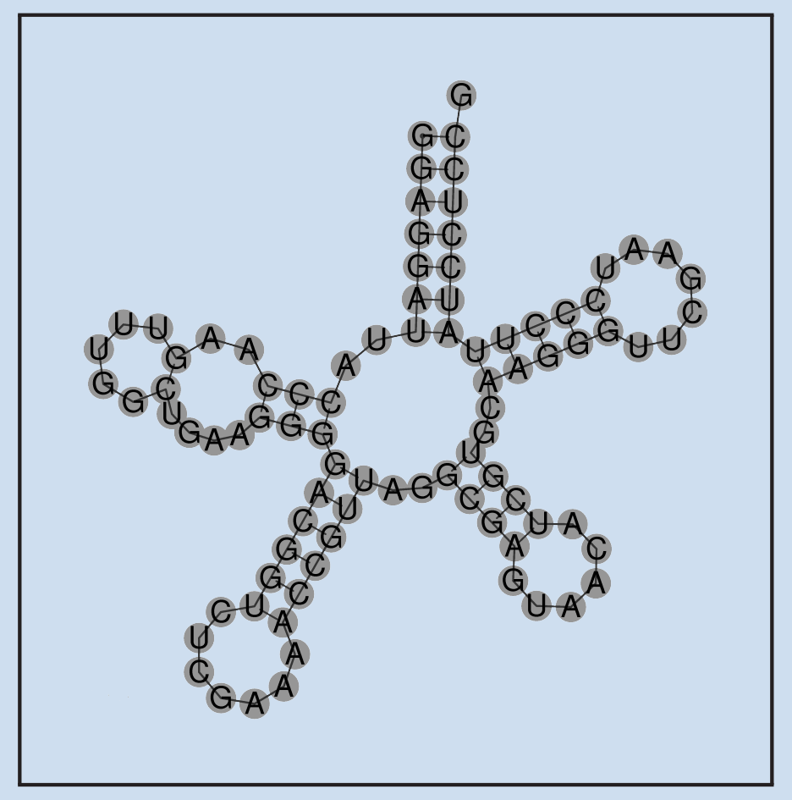
Introduction

Secondary structure is known to have a crucial impact on the RNA molecule functioning, therefore, development of the algorithms for secondary structure modeling and prediction is a fundamental task in computational genomics. An approach for sequences secondary structure analysis by combination of formal grammars and neural networks was proposed in [1]. We encode stems of secondary structure by means of context-free grammar, extract them by parsing algorithm and then process the parsing provided data by neural network. In this work, we apply this approach to RNA secondary structure prediction problem.

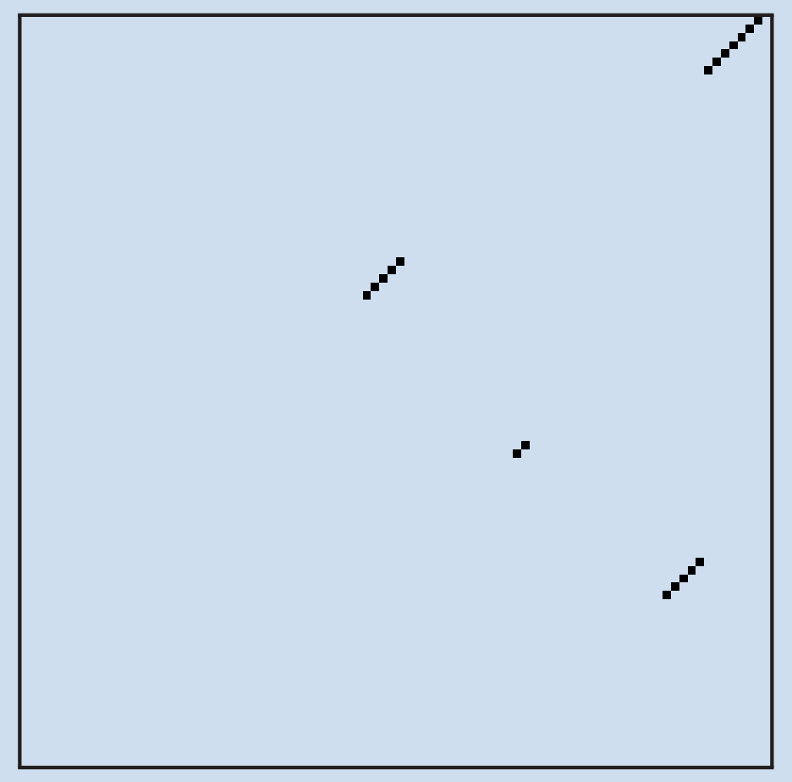
Research Motivation

Secondary structure can be described as composition of stems having different heights and loop sizes. We use context-free grammar to encode the most common kinds of stems and parsing algorithm to find the subsequences of sequence that should fold to such stems. Parsing matrix represents all the theoretically possible stems in some sequence in terms of grammar, but the real secondary structure is more complex than that. Therefore, parsing matrices require further processing and we propose using neural network to handle them in order to generate an actual secondary structure.

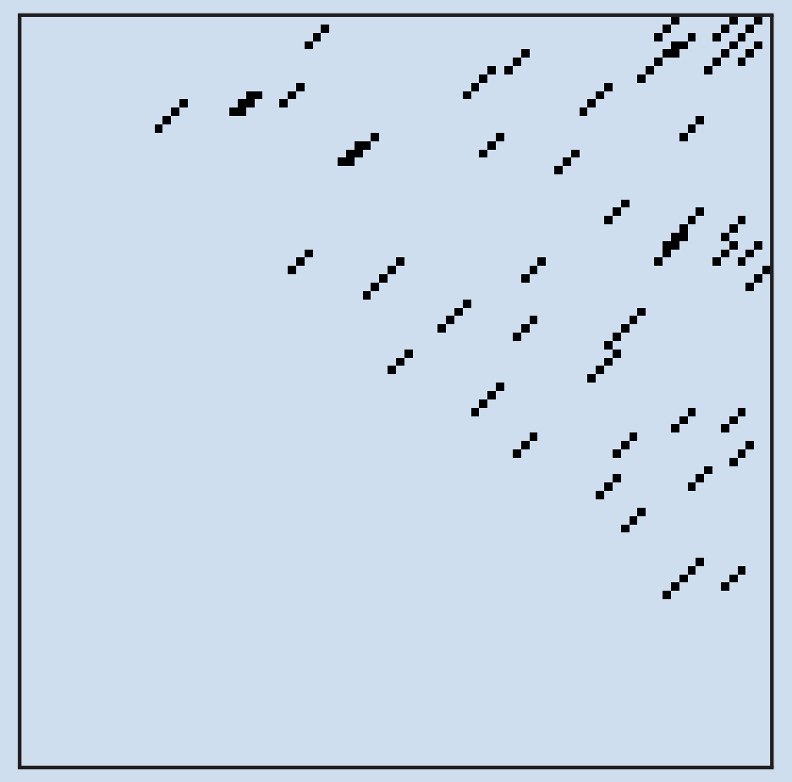
Example



Secondary structure



Contact map



Parsing matrix

Metrics

Consider TW (true white), TB (true black), FW (false white) and FB (false black) as amounts of correctly and incorrectly predicted pixels of each color for all images.

- $Precision = \frac{TW}{TW+FW}$
- $Recall = \frac{TW}{TW+FB}$
- $F1\ score = 2 * \frac{Precision*Recall}{Precision+Recall}$

Results

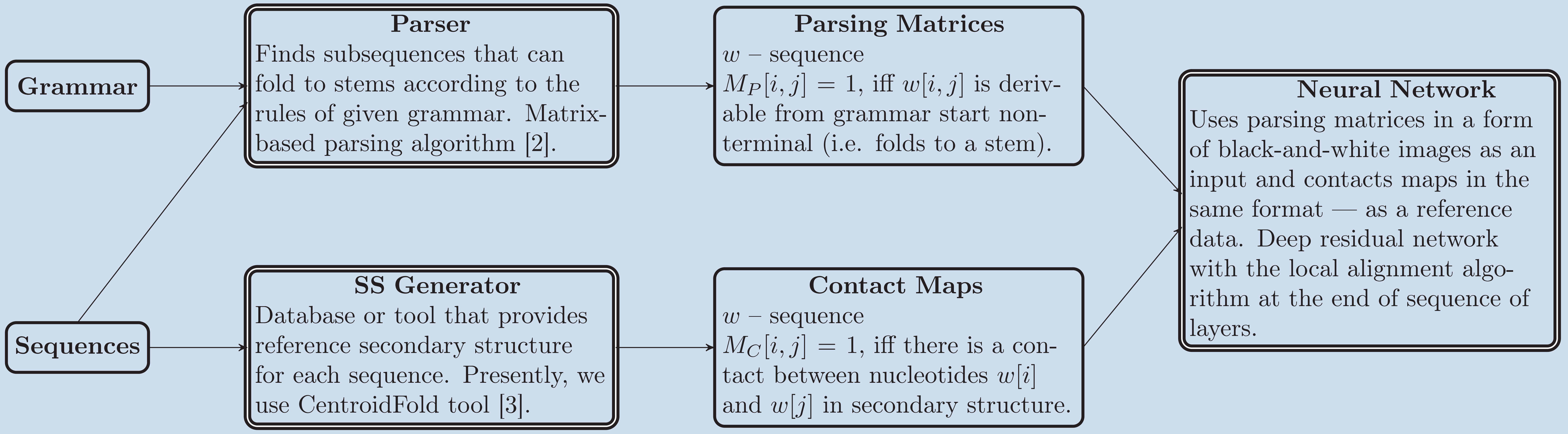
We took sequences from RnaCentral [4] database with 70%:10%:20% split and trained models on several datasets with fixed sequences length interval with and without alignment.

Length	Samples	Alignment	Precision	Recall	F1 score
90	26511	×	67%	75%	68%
		✓	80%	66%	70%
88-90	77976	×	66%	78%	69%
		✓	81%	62%	68%
50-90	141835	×	60%	72%	63%
		✓	71%	61%	63%

We can make the following conclusions.

- The smaller the window size, the more accurate the model.
- Alignment significantly improves precision of neural networks due to removing the contacts that break the secondary structure.
- From the other hand, it decreases recall, probably because it also removes a part of necessary information.
- So, our approach is applicable to secondary structure analysis problem, but further research is required.

Solution Overview



Grammar

Sequences

Parser
Finds subsequences that can fold to stems according to the rules of given grammar. Matrix-based parsing algorithm [2].

SS Generator
Database or tool that provides reference secondary structure for each sequence. Presently, we use CentroidFold tool [3].

Parsing Matrices
w – sequence
$M_P[i, j] = 1$, iff $w[i, j]$ is derivable from grammar start nonterminal (i.e. folds to a stem).

Contact Maps
w – sequence
$M_C[i, j] = 1$, iff there is a contact between nucleotides $w[i]$ and $w[j]$ in secondary structure.

Neural Network
Uses parsing matrices in a form of black-and-white images as an input and contacts maps in the same format — as a reference data. Deep residual network with the local alignment algorithm at the end of sequence of layers.

Future Research

- Improvement of models quality and performance.
- More accurate tuning of the models hyperparameters and alignment technique usage.
- Experiments on structures with pseudoknots and corresponding adaptation of the alignment algorithm.
- Building a model that predicts secondary structure for sequences of an arbitrary length.
- More accurate choice of the reference data source.
- More detailed testing and comparison with another tools.

References

[1] Semyon Grigorev and Polina 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*, volume 3, pages 234–241, 2019.

[2] Rustam Azimov and Semyon 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, pages 5:1–5:10, New York, NY, USA, 2018. ACM.

[3] Michiaki Hamada, Hisanori Kiryu, Kengo Sato, Toutai Mituyama, and Kiyoshi Asai. Prediction of rna secondary structure using generalized centroid estimators. *Bioinformatics*, 25(4):465–473, 2009.

[4] The RNAcentral Consortium. RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Research*, 47(D1):D221–D229, 11 2018.

Information

https://github.com/LuninaPolina/SecondaryStructureAnalyzer.
https://github.com/SacredArrow/Secondary_structure_public.

Acknowledgments

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