Secondary Structure Prediction by Combination of Formal Grammars and Neural Networks

Semyon Grigorev[1], Dmitry Kutlenkov[1], Polina Lunina[1]
[1]*Saint Petersburg State University, JetBrains Reserach, St. Petersburg, Russia*
E-mail: *semyon.grigorev@jetbrains.com, kutlenkov.dmitri@gmail.com, lunina_polina@mail.ru*

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.

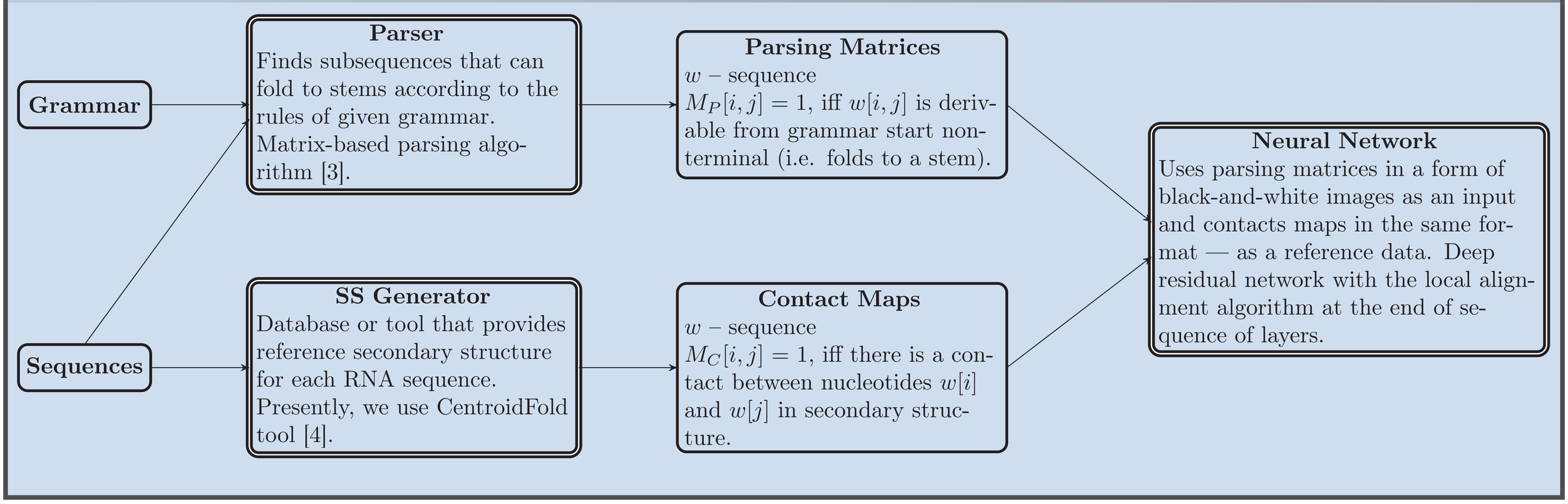
An approach for sequences secondary structure analysis by combination of formal grammars and neural networks was proposed in [1, 2]. The main idea here is to encode stems of secondary structure by means of context-free grammar, extract them by parsing algorithm and then process the parsing provided data by neural network to solve some given task. In this work, we apply this approach to RNA secondary structure prediction.

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 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 using neural network to handle them in order to generate an actual secondary structure.

Solution Overview



Results

For experimental research we took sequences from RnaCentral [5] 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.

Future Research

- Improvement of prediction accuracy.
- Experiments on structures with pseudoknots and corresponding adaptation of the alignment algorithm.
- Building a model that predicts secondary structure for sequences of any length.
- More accurate choice of the reference data source.

Acknowledgments

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

Information

All materials are available on GitHub:
...
...

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] Semyon Grigorev and Polina Lunina. Improved architecture of artificial neural network for secondary structure analysis. *BMC Bioinformatics*, 20(S17), 2019.

[3] 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.

[4] 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.

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