Formal Grammars and Neural Networks for RNA Secondary Structure Prediction[⋆]

Polina Lunina[1,2][0000−0002−7172−2647] ✉ and Semyon Grigorev[1,2][0000−0002−7966−0698]

[1] Saint Petersburg State University, 7/9 Universitetskaya nab., St. Petersburg, 199034, Russia
[2] JetBrains Research, Primorskiy prospekt 68-70, Building 1, St. Petersburg 197374, Russia
lunina_polina@mail.ru ✉, semyon.grigorev@jetbrains.com, s.v.grigoriev@spbu.ru

Abstract. RNA secondary structure prediction problem is known to be quite critical in computational genomics, therefore, different tools and algorithms are still competing in this field. In this work, we propose a new approach for secondary structure prediction. We describe the most probable types of stems by a context-free grammar, suchwise, the parsing matrix for some sequence represents all the theoretically possible stems. Then we apply an ensemble of residual neural networks to process such matrices in order to get a contact map of a pure secondary structure. RNA secondary structure databases are not big enough for neural network training, so, we use transfer learning technique. Firstly, we train several base networks with prediction tools output as a reference data, and secondly, we transfer these networks weights to a final composite model, which is trained and evaluated on real world data.

Keywords: CNN · ResNet · Machine Learning · Secondary Structure · Genomic Sequences · Formal Grammars · Parsing.

1 Introduction

Improvement in RNA secondary structure prediction accuracy is one of the key focuses in computational genomics due to its crucial role in functional analysis of RNA molecules. All the diversity of existing secondary structure prediction techniques can be divided into comparative methods that analyse several homologous sequences employing evolutionary approaches [1, 2] and single sequence methods that process one sequence at a time according to some folding constraints, e.g. thermodynamic [3] or statistic [4, 5] rules. One of the challenging parts is pseudoknotted structures processing, because pseudoknots are known to be widely represented in biological data, including functionally important RNA regions, nevertheless, building a model that handles them has always been a non-trivial task.

[⋆] Supported by the Russian Science Foundation grant 18-11-00100

Among other ways, formal grammars can be applied for RNA secondary structure description and some of the algorithms utilize this technique for secondary structure prediction [6, 7]. Due to probabilistic nature of secondary structure formation laws complicated stochastic (probabilistic) grammars are generally used here.

In this work, a new approach to secondary structure prediction which employs the combination of ordinary formal grammars and artificial neural networks is presented. The main ideas were outlined in [8, 9] and this research is conducted to further development of this approach in the context of secondary structure prediction problem. Secondary structure can be formally described as a compositions of stems having different lengths and loop sizes [10], so, we propose to use a simple context-free (not probabilistic) grammar to encode the most common types of stems and search for such stems in the input sequences by matrix-based parsing algorithm. Thereby, the parsing matrix for some sequence will contain the information about whether each subsequence of this sequence can fold to stem or not. This matrix is not yet a representation of a valid secondary structure, because it cannot contain all these stems at once and, besides, there can be more complex elements that are not expressible in terms of our grammar (such as pseudoknots and non-canonical base pairs). Therefore, we propose to process such matrices by neural networks that should filter extra stems and add some missing elements in order to generate a maximal approximation of this sequence secondary structure.

2 Proposed Approach Overview

In this section the brief description of the approach proposed in [8, 9] is provided. Although these works were devoted to RNA sequences classification problems, the main ideas are still relevant in the context of the current research. The proposed approach combines two different techniques: firstly, we use formal grammars for secondary structure features description and secondly, we apply neural networks for these features processing and building a final solution.

Formal Grammars The classical way of formal languages application for secondary structure description is to model the whole structure by means of probabilistic grammars [6, 7]. This approach is known to be quite successful, but it also should be mentioned that building such grammar requires a lot of theoretical and practical difficulties. Therefore, we propose a different way — to encode only stems of secondary structure by simple context-free grammar and leave further processing and probability estimation to neural network.

In figure 1 grammar G_0 that we use in this work as well as in the previous ones is presented. This grammar describes the recursive compositions of stems having height at least 3 (lines **7-12**) and loop size from 2 up to 10 (line **2**). Note that these constants are not mandatory and might be defined experimentally for each task. Also, G_0 allows only conventional base pairs (line **5**) and does not express pseudoknots, because adding the rules for both of these features complicates the grammar in unacceptable for us way, therefore, we expect the neural network to handle them. Also, we consider only stems of height three or more, because including shorter stems would overload the parsing matrix with unnecessary information. So, a sequence folds to a stem (according to the rules we have just defined) iff it is derivable from start nonterminal $s1$ of G_0 (line **1**).

```
1  s1: stem<s0>
2  any_str : any_smb*[2..10]
3  s0: any_str | any_str stem<s0> s0
4  any_smb: A | U | C | G
5  stem1<s>: A s U | G s C | U s A | C s G
6  stem2<s>: stem1< stem1<s> >
7  stem<s>:
8         A stem<s> U
9       | U stem<s> A
10      | C stem<s> G
11      | G stem<s> C
12      | stem1< stem2<s> >
```

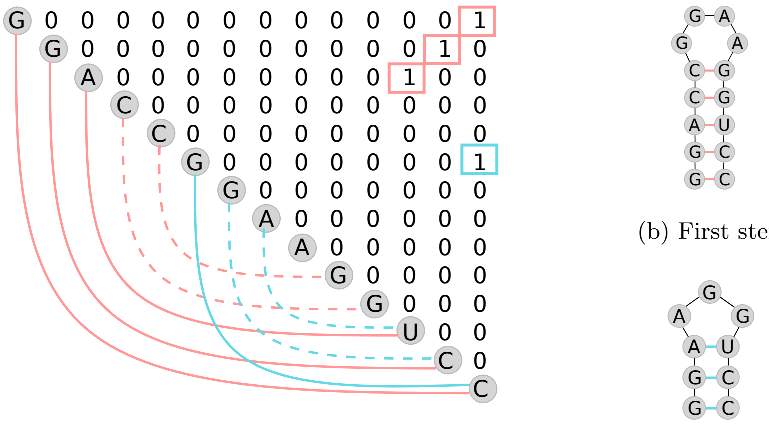
Fig. 1: Context-free grammar G_0 for RNA secondary structure stems description

Having a grammar, we want to find all the subsequences of some given sequence that may fold to stems and this is to be done by parsing algorithm. In all the experiments we use parsing algorithm [11] that is based on matrix operations and demonstrates high performance in practice due to the effective use of GPGPU.

Formally, the result of a matrix-based parsing algorithm for an input string w is an upper-triangular boolean matrix M_P, where $M_P[i,j] = 1$, iff the substring $w[i,j]$ is derivable from grammar start nonterminal. From the practical point of view, this means that parsing matrix contains 1 in a cell iff a correspondent substring folds to a stem according to the rules of a given grammar, so each stem results in a diagonal chain of one-s in the matrix, because if sequence $w_1 ... w_n$ is a stem than it is clear that $w_2 ... w_{n-1}$ is a stem and $w_3 ... w_{n-2}$ is a stem and so on while the stem height is at least 3.

In the figure 2 we provide a parsing result for a short tRNA sequence and show how parsing matrix maps with secondary structure stems. Each 1 in cell describes the stem of height at least 3, so, this sequence contains two subsequences that may fold to stems of the first nesting level. These stems expected hydrogen bonds along with corresponding matrix cells are painted in two different colors. Note that these stems interfere with each other, thereby, real secondary structure cannot contain both of them at the same time.

Neural Networks Parsing matrix is yet a formal construction storing the information about secondary structure features, so, at the final step of our solution



(a) Parsing matrix

(b) First stem

(c) Second stem

Fig. 2: Stems extracted from tRNA sequence

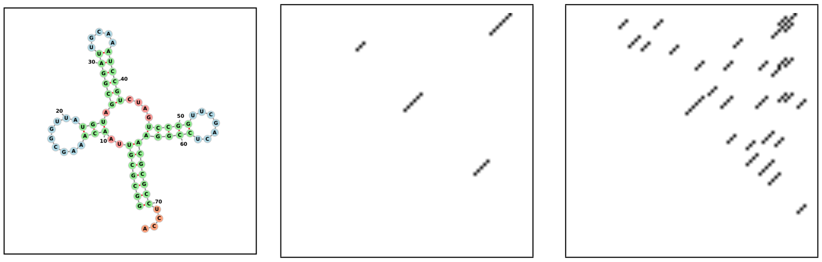
we propose to use artificial neural networks, to process such matrices and correlate them with some real world data. Matrices should be transformed to any suitable format and sent to the input layer of a neural network, constructed for a specific task. For example, in [9] we provide examples of neural networks for small RNA classification problems and show how to process parsing matrices as numerical vectors and black-and-white images.

2.1 Secondary Structure Prediction

In this section we describe all the details of the proposed approach application concerning specifically secondary structure prediction task.

Motivation One of the classical ways of RNA secondary structure formal representation is so-called contact map which for an input string w is a boolean matrix M_C, where $M_C[i,j] = 1$, iff nucleotides in positions i and j form a hydrogen bond (or, to put it simply, a contact) in secondary structure. Consider the discussed earlier parsing matrix for the same sequence w that has 1 in the cell $[i,j]$, iff subsequence $w[i,j]$ folds to a stem. It is clear that the first and the last nucleotides of every stem form a contact, therefore, we can easily transfer between parsing matrix and contact map definitions and view the parsing matrix as a sort of a contact map. Note that if parser finds a stem of height three than we will see only one cell with 1 in matrix, but such stem always wraps a stem of height two which wraps a stem of height one, so, we are always missing two contacts, therefore, after parsing we set $M_P[i-1, j+1] := 1$, $M_P[i-2, j+2] := 1$ if $M_P[i][j] = 1$, $i = 0..size(M_p), j = i..size(M_p)$.

In the figure 3 we provide two-dimensional secondary structure visualization for tRNA sequence (the plot 3a was made by tool [12]) with corresponding contact and parsing matrices and it can be seen clearly that the actual amount of contacts is far less than parsing matrix contains. Moreover, our grammar has certain limitations and cannot, for example, describe pseudoknots and non-canonical base pairs, so, the contacts that are formed by such rules will not be reflected in the parsing matrix.



(a) Visualization (b) Contact matrix (c) Parsing matrix

Fig. 3: RNA secondary structure representations

So, our task is set as follows: we want to process parsing matrices by a neural network in order to achieve a maximal similarity with the expected sequence secondary structure. This network should take parsing matrices as inputs and contact maps as desired outputs for the same set of RNA sequences. For convenience, we transform both matrices to black-and-white images by replacing zero cells with black pixels and one cells with white pixels. Also, we code sequences at the images diagonals by four types of gray pixels in case the nucleotide chains contain some important information about secondary structure shape.

Reference data

ResNet

2.2 Experiments

References

1. I. L. Hofacker and P. F. Stadler, "Automatic detection of conserved base pairing patterns in rna virus genomes," *Computers & chemistry*, vol. 23, no. 3-4, pp. 401–414, 1999.
2. F. Tahi, M. Gouy, and M. Régnier, "Automatic rna secondary structure prediction with a comparative approach," *Computers & chemistry*, vol. 26, no. 5, pp. 521–530, 2002.

3. 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.
4. S. R. Eddy and R. Durbin, "Rna sequence analysis using covariance models," *Nucleic acids research*, vol. 22, no. 11, pp. 2079–2088, 1994.
5. C. B. Do, D. A. Woods, and S. Batzoglou, "Contrafold: Rna secondary structure prediction without physics-based models," *Bioinformatics*, vol. 22, no. 14, pp. e90–e98, 2006.
6. B. Knudsen and J. Hein, "Pfold: Rna secondary structure prediction using stochastic context-free grammars," *Nucleic acids research*, vol. 31, no. 13, pp. 3423–3428, 2003.
7. M. E. Nebel and A. Scheid, "Evaluation of a sophisticated scfg design for rna secondary structure prediction," *Theory in Biosciences*, vol. 130, no. 4, pp. 313–336, 2011.
8. 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 - Volume 3 BIOINFORMATICS: BIOINFORMATICS,*, pp. 234–241, INSTICC, SciTePress, 2019.
9. P. Lunina and S. Grigorev, "On secondary structure analysis by using formal grammars and artificial neural networks," in *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*, pp. 193–203, Springer, 2019.
10. 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: BIOINFORMATICS,*, pp. 302–309, INSTICC, SciTePress, 2019.
11. 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), Association for Computing Machinery, 2018.
12. P. Kerpedjiev, S. Hammer, and I. L. Hofacker, "Forna (force-directed rna): Simple and effective online rna secondary structure diagrams," *Bioinformatics*, vol. 31, no. 20, pp. 3377–3379, 2015.