

Dense Neural Networks for Secondary Structure Analysis of Biological Sequences

Semyon Grigorev^{1,2}, Polina Lunina^{1,2}

¹*St. Petersburg State University, 7/9 Universitetskaya nab., St.Petersburg, Russia*

²*JetBrains Research, Universitetskaya emb., 7-9-11/5A, St.Petersburg, Russia*
s.v.grigoriev@spbu.ru, Semen.Grigorev@jetbrains.com, lunina_polina@mail.ru

Keywords: Dense Neural Network, DNN, Machine Learning, Secondary Structure, Formal Grammar, Parsing

Abstract: Abstract is very abstract.

1 INTRODUCTION

Accurate sequences classification and subsequences detection are an open problems in different areas of bioinformatics, such as genomics and proteomics. Challenge here is a high variability of sequences which one want to mark as a same class. For some type of sequences its secondary structure is a !!! and this fact may be used for !!!.

For example, algorithms that can efficiently and accurately identify and classify bacterial taxonomic hierarchy have become a focus in computational genetics. The idea that secondary structure of genomic sequences is sufficient for solving the detection and classification problems lies at the heart of many tools (Rivas and Eddy, 2000; Knudsen and Hein, 1999; Yuan et al., 2015; Dowell and Eddy, 2004). The secondary structure can be specified in terms of formal grammars. The sequences obtained from the real bacteria usually contain a huge number of mutations and “noise” which renders precise methods impractical. Probabilistic grammars and covariance models (CMs) are a way to take the noise into account (Durbin et al., 1998). For example, CMs are successfully used in the Infernal tool (Nawrocki and Eddy, 2013). Neural networks is another way to deal with “noisy” data. The works (Sherman, 2017; Higashi et al., 2009) utilize neural networks for 16s rRNA processing and demonstrate promising results.

In this work we propose the way to combine for-

mal grammars and neural networks for secondary structure features processing. The key idea is does not try to model full (sub)sequence by grammar, but create grammar which describes features of secondary structure and use neural network for these features processing.

2 PROPOSED SOLUTION

We combine neural networks and ordinary context-free grammars to detect genomic sequences. We extract features by using the ordinary (not probabilistic) context-free grammar and use the dense neural network for features processing. Features can be extracted by any parsing algorithm and then presented as a boolean matrix but we choose parsing algorithm based on matrix multiplication.

2.1 Context-Free Grammars

It is a well-known fact that secondary structure of sequence may be approximated by using formal grammars. There is number of works that utilize this fact for !!!

The !!! is to use probabilistic grammars. We use ordinary (not probabilistic) gramars. Our goal is not to model secondary structure of whall sequence (wich required probabilistic grammars), but describe

```

s1: stem<s0> any

any_str : any*[2..10]

s0: any_str | any_str stem<s0> s0

any: A | U | C | G

stem1<s>: A s U | G s C | U s A | C s G

stem2<s>: stem1< stem1<s> >

stem<s>:
    A stem<s> U
    | U stem<s> A
    | C stem<s> G
    | G stem<s> C
    | stem1< stem2<s> >
}

```

Figure 1: Context-free grammar for RNA secondary structure features extraction

features of secondary structure, such as stems, loops, pseudoknots and its composition. The set of feature types is limited by class of the grammar which we use. For example, pseudoknots can not be expressed by using conjunctive (Devi and Arumugam, 2017; Zier-Vogel and Domaratzki, 2013; Okhotin, 2001) or multiple context-free (Seki et al., 1991; Riechert et al., 2016).

The context-free grammar which we use in our experiments is presented in figure 1. More details on it. Four letters in the alphabet (). Only classical base pairs. s_1 is a start nonterminal. metarules as a feature of language. Stems as an example of metarules usage.

For example, one can vary length of unfoldable sequence $a_{0_7} : any*[0..10]$ or $a_{0_7} : any*[1..8]$. Also one can increase (or decrease for some reason) the minimal height of stem, or add pseudoknots description in the grammar.

2.2 Parsing Algorithm

Parsing is a feature extraction, so undirected parsing: we want to find all derivable substrings of given string for all nonterminals, not to check derivability of given string.

CYK — as a classical well-known algorithm.

Matrices.

Valiant (Valiant, 1975) — subcubic algorithm based on matrix multiplication.

Rustam (Azimov and Grigorev, 2018) — generalization for graph. Theoretical time complexity is !!! than complexity of the Valiant's algorithm, but in

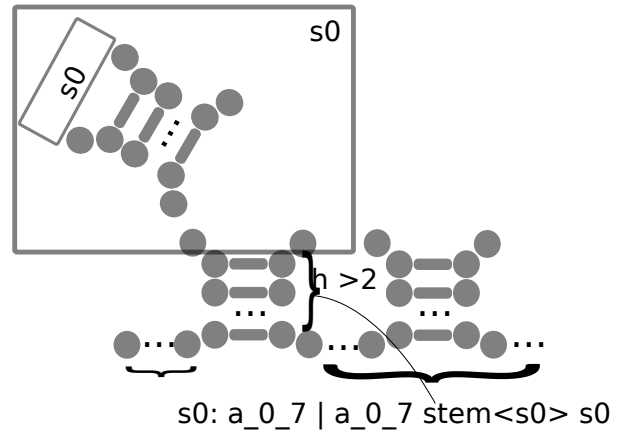


Figure 2: !!!

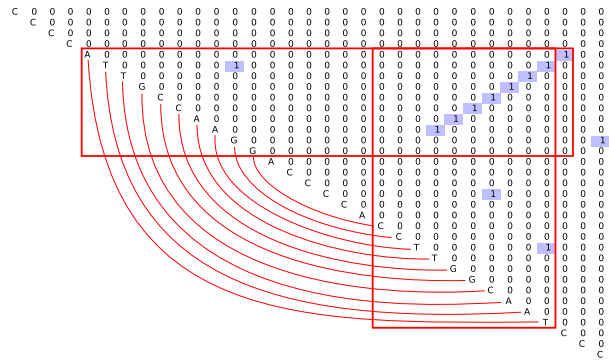


Figure 3: !!!!!

practice these algorithm avoid machinery on submatrices manipulation and demonstrate better performance with simple implementation.

Sparse matrices, boolean, GPGPU, etc.

Matrix-based approach can be generalized to conjunctive and even boolean grammars (Okhotin, 2014), as far as to multiple context-free grammars (Cohen and Gildea, 2016), which can provide a base for more expressive features descriptions handling.

2.3 Matrices

The result of parsing is a set of square boolean matrices. Each matrix M_N contains information of all substrings which can be derived from nonterminal N . In other words, $M_N[i, j] = 1$ iff $N \Rightarrow_G^* w[i, j]$ where w is the input sequence and G is context-free grammar, and N is a nonterminal.

Detailed description

One matrix for each nonterminal. We can select nonterminals of interest.

Empty triangle can be omitted. In order to handle matrices by using neural networks we vectorize it. Row by row. Compression to int or byte. It is the

reason why we want to try boolean networks.

2.4 Neural Networks

One of possible choice for classificaton. Classical scenario is to provide features vectors and try to classify them which means that network can select important features for each class. In our case ht fact that “ $w[i, j]$ is derivable from nonterminal N ” which is encoded in the matrix is a feature. So, matrix is a set (or vector) of features.

We use dense neural network because locality is broken during vectorisation and any convolutons is inapplicable.

Current architecture and its motivation and explanation. Huge dropout and batch normalization.

2.5 Examples

Here we provide more examples of matrices and point out some observations about it.

As mentioned above, pseudoknots can not be expressed in terms of context-free grammars. But Blue and red If newural network is powerful enough to !!! that these two fwetures should occurre simultaneously, then thay can detect pseudocnots.

The second example is a matrix for real tRNA.

Result of folding by tool!!!

Matrix contains all these variants. Blue and red boxes. And some additional.

Nontrivial compositions may be detected. May we use NNs for secondary structure prediction?

3 EVALUATION

Two cases. 16s rRNA and tRNA.

We evaluate the proposed approach for 16s rRNA detection. We specify context-free grammars which detect stems with the hight of more than two pairs and their arbitrary compositions. For network training we use dataset consisting of two parts: random subsequences of 16s rRNA sequences from the Green Genes database (DeSantis et al., 2006) form positive examples, while the negative examples are random subsequences of full genes from the NCBI database (Geer et al., 2010). All sequences have the length of 512 symbols, totally up to 310000 sequences. After training, current accuracy is 90% for validation set (up to 81000 sequences), thus we conclude that our approach is applicable.

Description of evaluation on tRNA classification. Procariot — eucarion. 500000 sequences. Upper bound of length. Results!!!

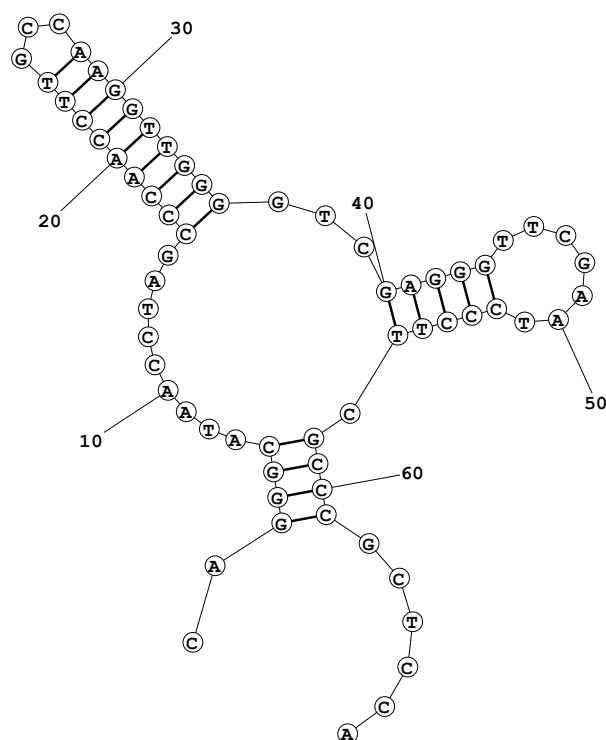


Figure 4: !!!!!

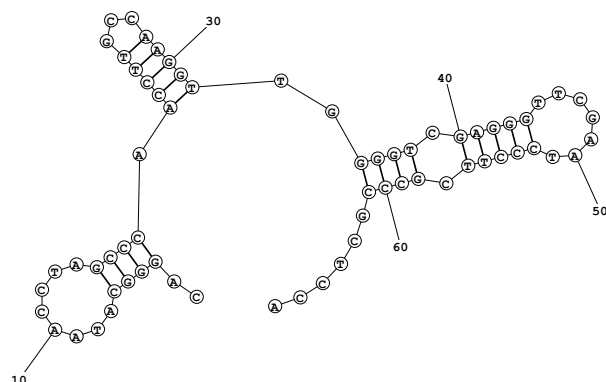


Figure 5: !!!!!

4 FUTURE WORK

The presented is a work in progress. The ongoing experiment is finding all instances of 16s rRNA in full genomes. Also we plan to use the proposed approach for the filtration of chimeric sequences and the classification. Composition of our approach with other methods and tools as well as grammar tuning and detailed performance evaluation may improve the applicability for the real data processing.

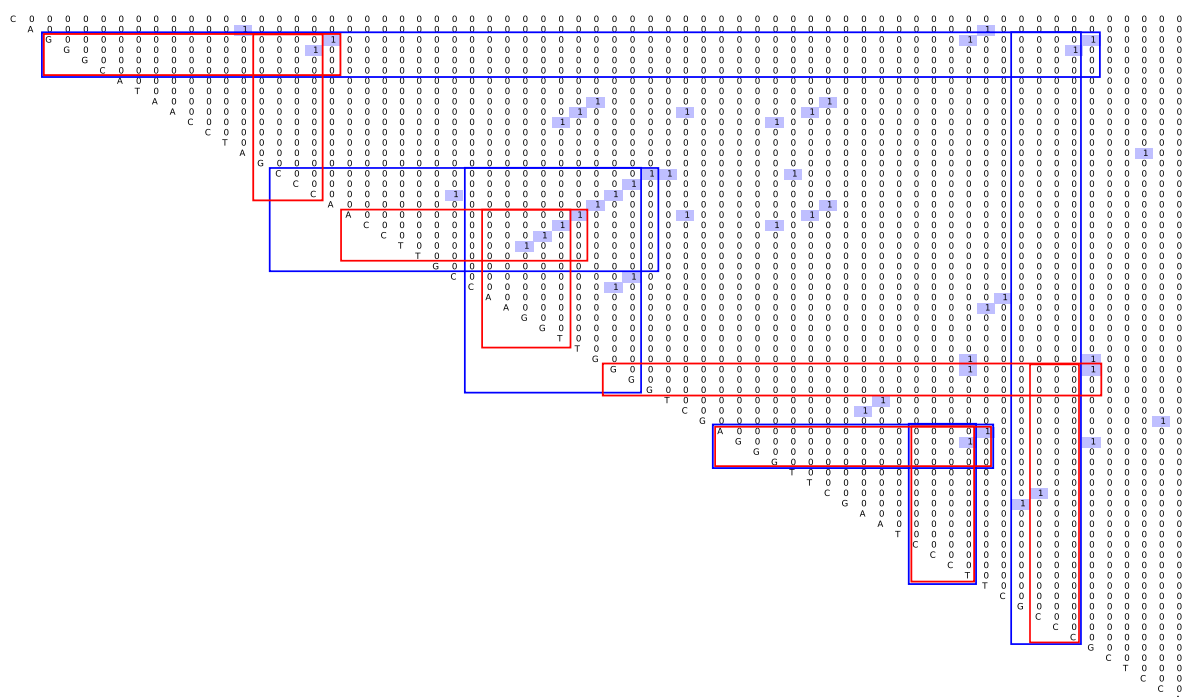


Figure 6: !!!!!

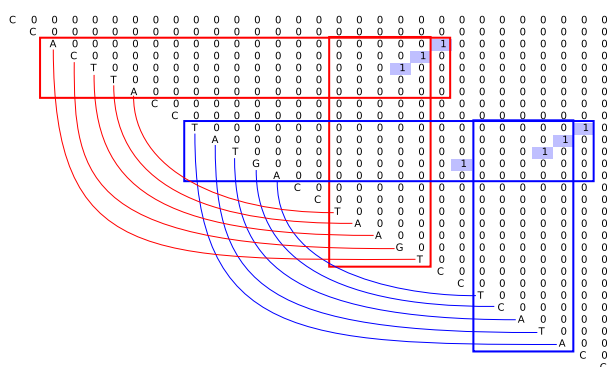


Figure 7: Parsing result for pseudoknot

5 DISCUSSION

Protenomics (Witold Dyrka) (Dyrka et al., 2018) More complex grammar: more symbols in alphabet, more complex features. More powerful languages required. One of the possible crucial problem is functionally equivalence sequences with different length in proteomics.

Different lengths. Is a problem. How can we normalize input?

Construct network which can handle sequences, not parsing data. It may help to create an embedding. It may be done by the next way.

1. Build and train the network which handle vectorized matrices.

2. Extend this network with head which should convert sequence to !!!
3. Train. Weights of first network is fixed.
4. For concrete problem we can tune weights for full network after second trained to appropriate quality.

Other types of NNs. Binary, convolutional (try to process matrix as a picture). Pictures: problem with size, typical matrix size is big.

Problems with data: how to create balanced set for training. Datasets (like GreenGenes) contains huge number of samples for some well-studied organisms and very small number of samples for other.

Huge amount of experiments in different directions. Plans should be discussed with community.

ACKNOWLEDGEMENTS

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

REFERENCES

- Azimov, R. and Grigorev, S. (2018). 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. ACM.
- Cohen, S. B. and Gildea, D. (2016). Parsing linear context-free rewriting systems with fast matrix multiplication. *Computational Linguistics*, 42(3):421–455.
- DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, 72(7):5069–5072.
- Devi, K. K. and Arumugam, S. (2017). Probabilistic conjunctive grammar. In *Theoretical Computer Science and Discrete Mathematics*, pages 119–127. Springer International Publishing.
- Dowell, R. D. and Eddy, S. R. (2004). Evaluation of several lightweight stochastic context-free grammars for rna secondary structure prediction. *BMC bioinformatics*, 5(1):71.
- Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.
- Dyrka, W., Coste, F., and Talibert, J. (2018). Estimating probabilistic context-free grammars for proteins using contact map constraints. *CoRR*, abs/1805.08630.
- Geer, L. Y., Marchler-Bauer, A., Geer, R. C., Han, L., He, J., He, S., Liu, C., Shi, W., and Bryant, S. H. (2010). The NCBI BioSystems database. *Nucleic Acids Res.*, 38(Database issue):D492–496.
- Higashi, S., Hungria, M., and Brunetto, M. (2009). Bacteria classification based on 16s ribosomal gene using artificial neural networks. In *Proceedings of the 8th WSEAS International Conference on Computational intelligence, man-machine systems and cybernetics*, pages 86–91.
- Knudsen, B. and Hein, J. (1999). Rna secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics (Oxford, England)*, 15(6):446–454.
- Nawrocki, E. P. and Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22):2933–2935.
- Okhotin, A. (2001). Conjunctive grammars. *J. Autom. Lang. Comb.*, 6(4):519–535.
- Okhotin, A. (2014). Parsing by matrix multiplication generalized to boolean grammars. *Theoretical Computer Science*, 516:101 – 120.
- Riechert, M., Höner zu Siederdissen, C., and Stadler, P. F. (2016). Algebraic dynamic programming for multiple context-free grammars. *Theor. Comput. Sci.*, 639(C):91–109.
- Rivas, E. and Eddy, S. R. (2000). The language of rna: a formal grammar that includes pseudoknots. *Bioinformatics*, 16(4):334–340.
- Seki, H., Matsumura, T., Fujii, M., and Kasami, T. (1991). On multiple context-free grammars. *Theoretical Computer Science*, 88(2):191 – 229.
- Sherman, D. (2017). Humidor: Microbial community classification of the 16s gene by training cigar strings with convolutional neural networks.
- Valiant, L. G. (1975). General context-free recognition in less than cubic time. *J. Comput. Syst. Sci.*, 10(2):308–315.
- Yuan, C., Lei, J., Cole, J., and Sun, Y. (2015). Reconstructing 16s rna genes in metagenomic data. *Bioinformatics*, 31(12):i35–i43.
- Zier-Vogel, R. and Domaratzki, M. (2013). Rna pseudoknot prediction through stochastic conjunctive grammars. *Computability in Europe 2013. Informal Proceedings*, pages 80–89.