

Improved Architecture of Artificial Neural Network for Secondary Structure Analysis

Polina Lunina¹, Semyon Grigorev¹

¹Saint Petersburg State University, JetBrains Reserach, St. Petersburg, Russia **E-mail:** lunina polina@mail.ru, semyon.grigorev@jetbrains.com



Motivation

An approach for biological sequences processing by combination of formal grammars and neural networks is proposed in the work [1]. While classical way is to model secondary structure of the full sequence by using grammar, the proposed approach utilizes it only for primitive secondary structure features description. These features can be extracted by parsing algorithm and processed by using artificial neural network. It is shown that this approach is applicable for real-world data processing and some questions are formulated for future research. In this work we provide answers to some of them.

Results

We use the proposed improvements for tRNA sequences analysis tasks: classification of tRNA into 2 classes: eukaryotes and prokaryotes (EP) and 4 classes: archaea, bacteria, plants and fungi (ABFP). We use sequences from databases [2, 3]. Results for both image- and vector-based classifiers are presented in the table, where base model means network which handles parsing result and extended model handles sequences and is based on the corresponding base model.

Classifier	EP		ABFP	
Approach	Vectors	Images	Vectors	Images
Base model accuracy	94.1%	96.2%	86.7%	93.3%
Extended model accuracy	97.5%	97.8%	96.2%	95.7%
Total samples (train:valid:test)	20000:5000:10000		8000:1000:3000	

Questions

Is it possible to use convolutional neural networks for parsing result processing? The result of parsing algorithm is a set of upper triangular boolean matrices. The original idea is to vectorize these matrices row by row and use DNNs for these vectors processing. Matrices can be also treated as bitmaps, where the false bits of matrix correspond to white pixels and the true bits to black ones. To handle these images we use network with a small number of convolutional layers followed by linearization and then the same structure as for vectorized data (dense and dropout layers with batch normalization).

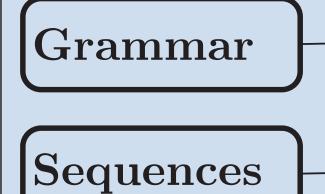
Is it possible to move parsing to network training step? Parsing is the most time-consuming operation of the proposed solution. We solve this problem by using two-staged learning. At the first step, we prepare a neural network (vector- or image-based) for our task which takes parsed data as an input. After that we extend trained network with a number of input layers that should convert the original nucleotide sequence into parsing result. This way we create a network which can handle sequences, not parsing result. So, parsing is required only for training the base network.

Future Research

- Chimeric sequences filtration
- Secondary structure prediction
- Proteins functions prediction

Solution Overview

Neural Networks



Parser
Parser extracts secondary structure features described by grammar. We use a matrix-based parsing algorithm [4].

Matrices

Parsing result for sequence w and nonterminal N is an uppertriangular boolean matrix M_N , where $M_N[i,j]=1$, iff w[i,j-1] is derivable from N.

Vectors

We drop out the bottom left triangle and vectorize the rest of matrix row by row. It requires the equal length of the input sequences, therefore we propose to either cut sequences or add some empty symbol till the definite length.

Images

We represent the false bits of matrix as white pixels and the true bits as black ones to get a bitmap image. This approach makes it possible to process sequences with different lengths since the images may be easily transformed to the same size.

dropout 0.05 Conv2D → dropout 0.3 5, (3, 3) 512 8194 normalization normalization flatten 128 512 relu dropout 0.3 normalization normalization dense dropout 0.9 relu relu dropout 0.75 1024 1024 dropout 0.75 normalization normalization 2048 relu relu normalization 64 sigmoid sigmoid dropout 0.9 dropout 0.9 dense 2048 normalization 512 sigmoid sigmoid normalization relu 1024 dense normalization

3028

normalization

relu

dropout 0.9

Neural Networks

N1 is a model for images classification, N2 is a model that classifies vectorized data and N0 is the extension block which converts the original nucleotide sequence into a set of features which can be handled by using weights of N1 or N2.

Acknowledgments

dropout 0.9

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

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: BIOINFORMATICS*, pages 234–241. INSTICC, SciTePress, 2019.
- 2] Genomic tRNA Database. http://gtrnadb.ucsc.edu/. Last accessed 05.06.2019.
- [3] tRNADB-CE. http://trna.ie.niigata-u.ac.jp/cgi-bin/trnadb/index.cgi. Last accessed 05.06.2019.
- [4] 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)*, page 5. ACM, 2018.