16s rRNA Detection by Using Neural Networks

Semyon Grigorev, Polina Lunina
Saint Petersburg State University
7/9 Universitetskaya nab., St. Petersburg, 199034, Russia
semen.grigorev@jetbrains.com, lunina_polina@mail.ru

The idea that secondary structure of some genomic sequences contains sufficient information which can be used for its detection and classification is widely used in different tools [4, 5, 6, 7]. Real sequences contain huge number of mutations and "noise", so precise methods for secondary structure handling are irrelevant. As a result, probabilistic methods such as probabilistic grammars and covariance models (CMs) are used in this area [1]. For example, CMs are successfully used in the Infernal tool.

Another possible way to deal with "nosy" data is neural networks utilization. There are some solutions for which utilize neural networks for 16s rRNA processing [2, 3] and demonstrate promising results, but more research in this area are required. We propose the way which combines neural networks and context-free grammars. We extract features by using ordinary (not probabilistic) context-free grammar and use dense neural network for features processing.

Let grammar G with start nonterminal S is fixed. Firs step is input sequence parsing. Result is Boolean matrix matrix of features: $M.[i, j] = 1$ iff $S \Rightarrow^*_G w.[i, j]$ where w is the input sequence. The next step is result matrix row-by-row vectorization with "compression": each 32 bits store as unsigned integer. Finally we process vectors by using neural network.

We evaluate proposed approach on 16s rRNA detection. We specify context-free grammars which detects stems with hight more than two pairs and its arbitrary compositions. For network training we use dataset combined from two parts: positive examples are random parts of 16s sequences from Greengenes database, negative examples are random parts of full genes

form NCBI database.All sequences have length 512 symbols, totally up to 310000 sequences. After training current accuracy is 90% for validation set (up to 81000 sequences), and we can conclude that our approach may be useful.

Ongoing experiment is full genome processing: find out all instances of 16s in full genomes. Also we plan to use proposed approach for chimeric sequences filtration and sequences classification. In order to make our approach wore useful for real data processing it is required to investigate possible ways fro composition with other methods and tools. Grammar tuning and detailed performance evaluation and tuning also may be required.

References

[1] Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge: Cambridge University Press.

[2] Sherman D. *Humidor: Microbial Community Classification of the 16S Gene by Training CIGAR Strings with Convolutional Neural Networks.* — 2017.

[3] Higashi S., Hungria M., Brunetto M. *Bacteria classification based on 16S ribosomal gene using artificial neural networks* //Proceedings of the 8th WSEAS International Conference on Computational intelligence, man-machine systems and cybernetics. – 2009. – С. 86–91.

[4] Rivas E, Eddy S.R. *The language of RNA: a formal grammar that includes pseudoknots* // Bioinformatics. — 2000.

[5] Knudsen Bjarne, Hein Jotun. *RNA secondary structure prediction using stochastic context-free grammars and evolutionary history.* //Bioinformatics (Oxford, England).— 1999.— Vol. 15, no. 6.— P. 446–454.

[6] Yuan C. et al. *Reconstructing 16S rRNA genes in metagenomic data* //Bioinformatics. – 2015. – №. 12. – С. 135-143.

[7] Dowell R. D., Eddy S. R. *Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction* //BMC bioinformatics.– 2004.– №. 1.– С. 71.