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

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 [4, 5, 6, 7]. 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 [1]. For example, CMs are successfully used in the Infernal tool.Neural networks is another way to deal with "noisy" data. The works [2, 3] utilize neural networks for 16s rRNA processing and demonstrate promising results.

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 M such that $M[i, j] = 1$ iff $S \Rightarrow^*_G w[i, j]$ where w is the input sequence and G is context-free grammar with the start nonterminal S.

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 form positive examples, while the negative exam-

ples are random subsequences of full genes from the NCBI database. 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.

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.

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. – P. 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. – P. 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.– P. 71.