DOI 10.18699/BGRSSB-2018-029

Siamese neural networks for metagenomics binning

B. Kirillov
Skolkovo Institute of Science and Technology, Moscow, Russia
e-mail: Bogdan.Kirillov@skoltech.ru

Key words: deep learning, few-shot learning, metagenomics, binning

Motivation and Aim: A metagenomic sample usually contains DNA from a lot of different organisms. Each organism has its own genome but only small part of it can be found in pool of sequenced reads and it is hard to identify the source species for each of the resulting reads [1]. This operation is called "binning" and it is one of the major challenges in current metagenomics. It can be done using a reference database or without prior knowledge of taxonomy, in this study I concentrate on the latter case. Most taxonomy-independent binning algorithms rely either only on low-level k-mer features [2] or on additional data, such as DNA methylation [3] or coverage profiles [4]. Ability of deep neural networks to extract high-level features from the sequence may provide an improvement. The research's end goal is to create and evaluate a deep neural network-based solution for binning.
Methods and Algorithms: I use a special neural network architecture, so-called Siamese Networks [5]. The model learns from data how to answer the question "Did that pair of reads come from the same species?" In such a setup the binning problem becomes a binary classification task. To train and test the model I use simulated data and freely accessible dataset of human gut microbiome [6].
Results: Currently, I am in process of hyperparameter selection and optimization. The model shows good generalization ability, also there are indicators that the model is able to identify reads from previously unseen species correctly.
Conclusion: This study provides a novel method of taxonomy-independent metagenomics binning using Deep Learning methods. The trained model, training and read simulation scripts will be available at GitHub.
Acknowledgements: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. Wooley J.C., Godzik A., Friedberg I. (2010) A primer on metagenomics. PLoS Comput. Biol. 6(2):e1000667.
2. Girotto S., Pizzi C., Comin M. (2016) MetaProb: accurate metagenomic reads binning based on probabilistic sequence signatures. Bioinformatics. 32(17):i567-i575.
3. Beaulaurier J. et al. (2018). Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. Nat. Biotechnol. 36(1):61-69.
4. Lin H.H., Liao Y.C. (2016) Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. Sci. Rep. 6:24175.
5. Koch G., Zemel R., Salakhutdinov R. (2015) Siamese neural networks for one-shot image recognition. In: ICML Deep Learning Workshop. Vol. 2.
6. Nielsen H.B. et al. (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat. Biotechnol. 32(8):822.