[BIOINFO 2024]

VertBert: a BERT-based metagenomic contig binning model for viral genomes.

Junho Jung[1], Ho-Jin Gwak[1], Mina Rho[1,2,*]

[1]*Department of Computer Science and Engineering, Hanyang University, Seoul, Korea*

[2]*Department of Biomedical Informatics, Hanyang University, Seoul, Korea*

**Corresponding author: minarho@hanyang.ac.kr*

Motivation: In metagenomics, binning assembled contigs is a useful step to recover the original genome sequences. However, binning without the reference genomes remains challenging. Several learning models have been proposed to enhance the binning accuracies. The methods commonly used two kinds of features: *k*-mer frequencies and read depths of each contig. These approaches, however, have suffered from some limitations. Specifically, read depth varies with the proportion of viruses in each sample, making the model dependent on input samples and limiting the generalizability of read depth as a feature. Additionally, *k*-mer frequencies tend to lack conservation across different regions of the viral genome. To address these limitations, we developed a deep-learning model that more effectively capture the informative context of viral genomes and is independent of the sample context.

Results: In this study, we developed a binary classification model that determines whether two contigs originate from the same genome. A BERT-based model, pre-trained on viral genomes, was used to capture the contextual information of viral genomes. The model was trained using eight double-stranded DNA virus families that infect vertebrate hosts: Alloherpesviridae, Orthoherpesviridae, Polyomaviridae, Papillomaviridae, Iridoviridae, Poxviridae, Adintoviridae, and Adenoviridae. In a species-leave-out test, our model outperformed a *k*-mer-based model by +0.04 in recall while maintaining a precision of 0.95. In addition, we applied a community detection algorithm to bin contigs based on the binary classification results. In terms of binning performance, our model demonstrated higher performance compared to the *k*-mer-based model, achieving improvements of 0.1 in adjusted rand index (ARI) and 0.2 in homogeneity with a minimal decrease of 0.05 in completeness.