Leveraging Large Language Models for Metagenomic Analysis

M.S. Refahi, B.A. Sokhansanj and G.L. Rosen

Department of Electrical & Computer Engineering, Drexel University, Philadelphia, PA, USA
{sr3622, bas44, glr26}@drexel.edu

***Abstract*— Analyzing sequencing data from microbiome experiments is challenging, since samples can contain tens of thousands of unique taxa (and their genes) and populations of millions of cells. Reducing the dimensionality of metagenomic data is a crucial step in improving the interpretability of complex genetic information, as metagenomic datasets typically encompass a wide range of genetic diversity and variations.**

In this study, we implement RoBERTa, a state-of-the-art large language model, and pre-train it on relatively large genomic datasets to obtain a model that can be used to generate embeddings that can help simplify complex metagenomic data sets. The pre-training process enables RoBERTa to capture the inherent characteristics and patterns present in the genomic sequences. We then evaluate the effectiveness of embeddings generated using the pre-trained RoBERTa model in downstream tasks, with a particular focus on taxonomic classification. To assess whether our method can be generalizable, we conduct extensive downstream analysis on three distinct datasets: 16s rRNA, 28s rRNA, and ITS. By utilizing datasets containing 16S rRNA exclusive to bacteria and eukaryotic mitochondria, as well as datasets containing 28S rRNA and ITS specific to eukaryotes (such as fungi), we were able to assess the performance of RoBERTa embeddings across diverse genomic regions. We tune the RoBERTa model through hyperparameter optimization on each dataset. Our results demonstrate that RoBERTa embeddings exhibit promising results in taxonomic classification compared to conventional methods.

Keywords*— *DNA sequence analysis, metagenomics, taxonomic classification, microbiome, natural language processing.

I. Introduction

Metagenomics data offers valuable insights into the composition and functionality of microbial communities across diverse environments. However, analyzing such data presents significant challenges due to the intricate nature and genomic diversity of microorganisms.

By leveraging the advances in Natural Language Processing (NLP), researchers have made remarkable progress in training machine-learning models to generate vector representations for word sequences, offering promising solutions for analyzing genomics and metagenomics data [1]. This advancement, known as representation learning, has proven to be a powerful tool in extracting meaningful information from data and has opened up exciting possibilities in the field of biological research. Leveraging the power of these models, we aim to overcome the challenges posed by the genomic diversity of bacteria and extract informative features for any downstream analysis tasks like taxonomic classification.

In biological language processing (BLP), the utilization of k-mer representation has gained prominence as a method for working with lower-dimensional representations. Also leveraging the frequency of k-mers, has demonstrated success representation method in various BLP applications [2] . Researchers have employed various techniques, including Word2Vec [3], LSTM [4], GloVe [5] and Transformers [6, 7], to obtain representation embedding vectors for biological k-mer sequences. Transformers based Language models like BERT and RoBERTa offer several advantages over RNNs like LSTM and static word embeddings like Word2Vec. First, models like BERT and RoBERTa provide contextualized word representations, capturing the meaning of words in relation to their surrounding context. This contextual understanding enhances their performance in tasks such as sentiment analysis and machine translation. Additionally, language models excel at sentence-level understanding by considering the entire sentence, making them suitable for tasks like natural language inference and text classification. While,Word2Vec relies on word co-occurrence statistics, limiting its scope to word-level similarities.

In recent years, language models such as BERT, ALBERT, and XLNet have been widely employed to explore the potential of Transformer models in studying protein structures and their properties [8]. These models leverage attention mechanisms to unveil the three-dimensional structure of proteins, revealing spatial relationships between amino acids that may be distant in the sequence but close in space. ProteinBERT [6], trained on a diverse dataset of 106 million proteins, is an efficient tool for training predictors and analyzing protein properties. ProteinBert embedding is valuable for limited labeled data scenarios, providing insights into protein characteristics. ProteinBert has also been fine-tuned using contrastive-learning for other protein downstream tasks [9].

DNA sequences provide insights into control mechanisms for gene expression, potential of cellular processes, and the impact of genetic variations on gene regulation. DNA sequences are composed of individual nucleotides, much like letters in a sentence. There are various methods used to extract larger units of sequence, akin to words. One of the long-used techniques for effectively extracting meaningful sub-words from a sequence of nucleotides is k-mer tokenization. DNABERT [10] is highlighted as a variant of the BERT model that has been specifically trained on human genome data using a wide variety of k-mer tokenizations. This specialized training empowers DNABERT to effectively capture genomic language patterns and encode biological information, making it suitable for various downstream tasks.Furthermore, the paper "BioAutoMATED" [11] introduces a flexible framework for analyzing multi-omics data using different model architectures for specific downstream tasks. The authors demonstrate that DNABERT outperforms their own method in predicting ribosome-binding sites with a specific learning rate. This showcases the effectiveness of using a large language model for these tasks.Considering downstream applications involving bacteria genomes, it is crucial to prioritize the use of a pre-trained model specifically trained on bacteria data rather than one trained on human data. This choice is motivated by the computational demands and the phylogenetic closeness of the bacteria data [12].

In this study, we chose RoBERTa for its superior performance and efficiency in language understanding tasks, particularly on large datasets. RoBERTa, an advanced variant of BERT, employs enhanced transformer architectures and is trained using token prediction, resulting in improved language understanding and representation of intricate patterns and relationships within the data.

The implementation challenges were mainly tied to unique intricacies of prokaryotic genomes. Unlike eukaryotes, prokaryotic genomes exhibit compact structures with a high gene density and diverse operon arrangements. Adapting RoBERTa to effectively understand and accurately represent prokaryotic genomic language posed a significant challenge. We needed to tailor the model architecture and hyperparameters to address these specific challenges, ensuring optimal performance in downstream tasks like taxonomic classification.

II. Methods

II-A. DNA Language Model Training

The initial step in training any type of language model involves tokenization. Tokenization is especially important when working with DNA sequences since there are no specific tokens or whitespace in DNA sequences. There are various methods available for tokenizing DNA sequences, and one commonly used approach is segmenting them into k-mers with a one-base overlap. In this study, we experimented with different k-mer sizes ranging from 3-mers to 6-mers in order to validate our approach. To enhance analysis of prokaryotic genomes, we conducted experiments using customized RoBERTa [13] models. We adopted the main architecture and transformers from the RoBERTa model while exploring various modifications to optimize its performance for genomics sequences. We focused on the masked language modeling (MLM) task, where a random sample of tokens in the input sequence is replaced with the special token "[MASK]". We varied the embedding size, testing values of 512, 768, 1024, and 2048, to assess their impact on model performance. Additionally, we investigated different input lengths, choosing a length of 200bp due to the short-read nature of popular Illumina sequencing. Furthermore, we examined other important parameters, such as the number of attention heads and the number of forward layers. By varying these parameters, we aimed to evaluate their influence on model effectiveness and identify the best configuration for metagenomic analysis. Through these experiments, we sought to determine the most suitable configuration of the RoBERTa model for metagenomic data, with a focus on improving the accuracy and interpretability of downstream tasks. The results of our analysis and model comparisons will be presented in the following sections.

II-B. Downstream and Pre-Train Datasets

We used one bacterial dataset and two fungal datasets to evaluate our models. The inclusion of fungal datasets allowed us to assess how well our embeddings generalized to previously unseen data, providing valuable insight into the models' performance. Each dataset was split into 90% for training and 10% for testing in order to analyze downstream tasks at different taxonomic levels. For both sets, we attempted to maintain a balanced distribution of samples across taxonomic groups in order to prevent potential biases. Additionally, we filtered datasets before splitting, considering only classes with more than 10 genera.

II-B1. Pretraining-Dataset

We collected a comprehensive dataset of 33,902 complete prokaryotic genomes from the National Center for Biotechnology Information (NCBI) for pretraining our model. These genomes exhibit an average length of 3.4 megabases (Mb), with a wide range spanning from 0.11 Mb to 13 Mb. The dataset encompasses a diverse distribution of genomes, representing various taxonomic groups and genetic characteristics. Two datasets were extracted from the dataset of this prokaryotic genomes for pretraining. The first dataset comprised 7 million genomic 200bp fragments, while the second dataset consisted of 51 million 200bp fragments. These datasets were obtained from the same distribution of genomic fragments but varied in the number of fragments extracted from each genome.

II-B2. Bacterial-16s Dataset

The DairyDB [14] dataset consists of 10,612 full-length 16S rRNA sequences derived from microbial species present in dairy products. To facilitate further analysis, we extracted 80,227 fragments of 200bp from these sequences. This fragmented dataset encompasses sequences from 42 distinct phyla, 197 different orders, and 1069 different genera.

II-B3. Fungi-ITS Dataset

This dataset was obtained from the Fungi RefSeq ITS project [15]. This dataset consists of 15,551 sequences derived from the non-gene coding region of the genome. We extracted 50,068 fragments, each containing 200bp, from the sequences for further analysis. This dataset showcases sequences from 6 different phyla, 235 different orders, and 516 families.

II-B4. Fungi-28s Dataset

We obtained the curated 28s rRNA sequences from the mothur project [16]. The dataset comprises 8,506 unique sequences extracted from the 28S rRNA of fungi. In addition, we extracted 42,766 fragments, each consisting of 200 bp, from these sequences. Our dataset exhibits a broad taxonomic level, encompassing sequences from 8 different phyla, 105 orders, and 293 families.

III. Results & Discussion

We extensively analyzed and optimized 44 different models, each representing a unique combination of parameter settings. To ensure reliable and accurate results, we conducted the training process with consistent parameters except for the specific parameter being investigated. The training was performed on a high-performance infrastructure consisting of 4 GPUs, each with 32GB of memory. The duration of training varied for each model, typically lasting between 12 and 40 hours per epoch, depending on the complexity of the architecture. We evaluated our system using two metrics: the pre-training loss function and accuracy for taxonomic classification tasks at various levels.

III-A. Kmer Size

Figure 1 illustrates the evaluation of different k-mer sizes.Increasing the k-mer size led to a consistent increase in loss, indicating the model's ability to capture longer sub-sequences and higher computational complexity. This complexity can be attributed to the larger vocabulary and the presence of rare words during training. Additionally, it is noteworthy that increasing the k-mer size from 3 to 6 led to improvements across all 9 tasks. Moreover, the model trained with a k-mer size of 6 achieved the highest accuracy (refer to Figure 1). This significant difference in accuracy compared to other models can be attributed to the fact that each 6-mer represents two codons, which correspond to amino acids. We explored k-mer sizes up to 6 due to the potential increase in model complexity and computational resources beyond that. In Figure 2, we also have demonstrated the benefits of different k-mer sizes by examining Zipf's law on a pre-trained dataset. This analysis revealed that the 6-mer exhibits both high frequency and a relatively low occurrence of rare words.

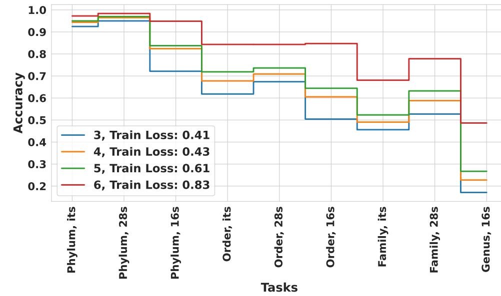


Figure 1. Comparison of Model Accuracies for Different Tasks with Respect to K-mer Size

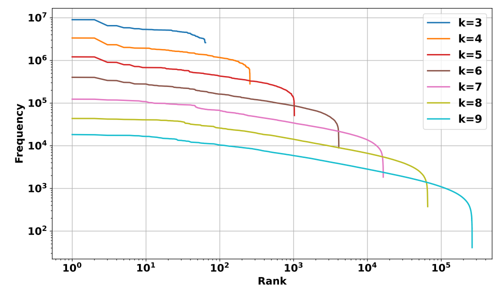


Figure 2. Zipf's Law of the pre-training dataset of 33k genomes for Different k-mer Lengths

III-B. Embedding Size

The impact of the embedding size on model performance was investigated, and the findings are presented in Figure 3(a). Increasing the embedding size, which depends on the vocabulary size (function of k-mer size), can improve model performance by capturing more complex relationships in the data, leading to a decrease in loss. Increasing the embedding size can improve the model's performance, but beyond a certain threshold (e.g., 2048), further increases may lead to difficulties in training and underfitting. Interestingly in our case, the best results were achieved when employing embeddings from the model with an embedding size of 2048 for downstream tasks (refer to Figure 3(a)). The higher-dimensional embeddings demonstrated superior performance across all tasks, suggesting that they effectively captured the diverse distribution of the data and improved classification accuracy. To conduct further analysis, it is crucial to preserve this diverse representation, which is beneficial for taxonomic classification. By reducing the model size, we trained our final model that effectively fit the genomics dataset.

III-C. Positional Embedding Type

Positional embeddings are vectors that represent token positions in a sequence, either uniquely for specific positions (absolute) or in relation to other tokens (relative). Figure 3(b) shows the impact of positional embedding type on model performance. Positional embeddings are a significant parameter that impacts both training time and model performance. Through multiple evaluations, we consistently observed superior performance of the

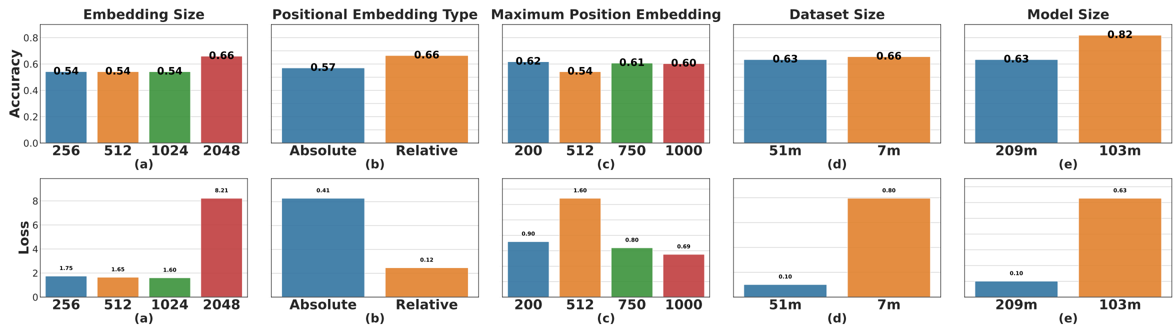


Figure 3. Comparison of Average Accuracy and Training Loss for Different Models. Here we represent the two performance metrics for each optimized parameter. The top row of plots shows the average accuracy of 9 different downstream tasks for each model.The bottom row of plots displays the pre-training loss value for each model.

relative positional embedding model compared to the absolute positional embedding model in downstream tasks.In our view, relative positional embeddings are preferable because relative information within tokens (k-mers) is more critical than absolute positional information within a sequence.

III-D. Maximum Positional Embedding Size

Figure 3(c) summarizes the results of varying the maximum positional embedding size on model performance. The maximum positional embedding size is a parameter that determines the length of the sequence for which the model can effectively capture positional information. In the original BERT model, the default value for this parameter is 512. Our experiments reveal that the choice of maximum positional embedding size significantly influenced the model's performance. As we explored a range of values from 200-1000, our results showed that setting the maximum positional embedding size to 200 resulted in the best average performance (see Figure 3).

III-E. Dataset Size

In Figure 3(d), we analyzed two datasets: the 51 million dataset and the 7 million dataset. Both datasets share the same distribution of genomes, but we intentionally included fewer sequences in the 7M dataset compared to the 51M dataset. Importantly, the frequency proportion of each k-mer is nearly identical in both datasets. Although the training loss was lower for the 51 million dataset, the performance on downstream tasks was actually better for the 7 million dataset (refer to Figure 3(d). This indicates that overfitting might have happened with the larger dataset, causing a decrease in performance when working with new, unseen data.

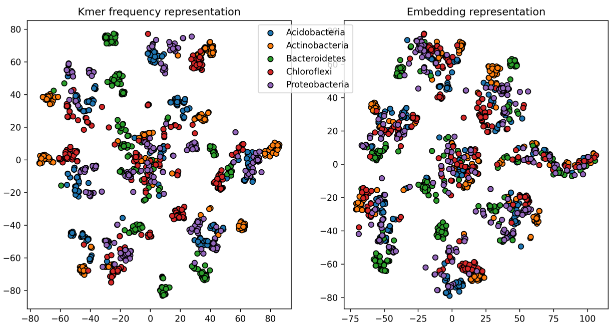
III-F. Model Size

The model size significantly depends on two parameters: the feedforward dimension and the number of feedforward layers. We extensively investigated these two parameters multiple times while keeping the remaining parameters constant. In Figure 3(e), we compare two models: one with 209 million parameters and the other with 103 million parameters. Surprisingly, we observed that a larger model size does not necessarily result in better performance. In fact, in some tasks, the larger model even yielded lower accuracy results.

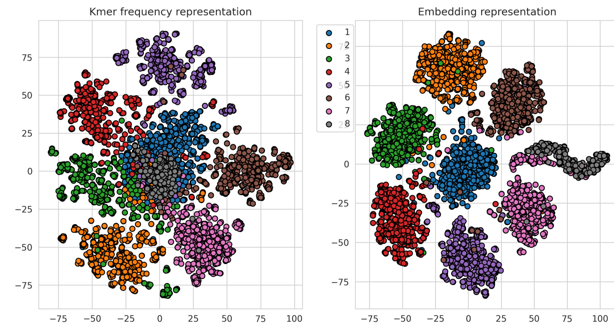
III-G. Optimized Model Results

We conducted t-SNE visualization to gain insights into the distribution and separability of phyla within the embeddings. In the t-SNE plots, we colored the data points by phylum and focused on the top five most highly represented phyla. Figures 4a, 4c, and 4d display the embedding results for the 16S, ITS, and 28S datasets, respectively. From the visualizations, we observed distinct clusters and patterns for both the pre-trained model embeddings and the k-mer frequency representation. Notably, the pre-trained model embeddings exhibited tighter and more well-defined clusters compared to the k-mer frequency representation. This difference was particularly evident in the ITS and 28s datasets. The improved cluster separation in the pre-trained model embeddings suggests that the model captures more nuanced relationships and discriminative features between the phyla, leading to enhanced representation quality in the embedding space. Interestingly, when examining the results for the 16s dataset (Figure 4b), we observed an intriguing pattern: the embeddings were organized based on the position of each fragment in the sequence. Fragments with similar positions clustered together, indicating that our embedding captures positional information effectively, which is not evident in the k-mer representation. This discovery has inspired further exploration of the pre-trained model's potential in capturing information related to variable regions or functional characteristics, which could be valuable for downstream tasks.

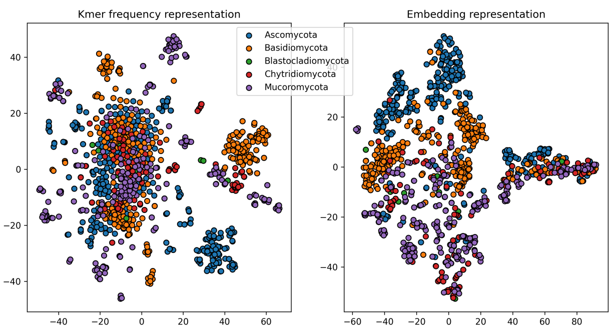
We conducted taxonomic classification experiments comparing our model's embedding representation, k-mer representation, and the DNABERT model's embedding on three downstream datasets at different taxonomic levels. All evaluations were performed using 10-fold cross-validation, and the reported results are the



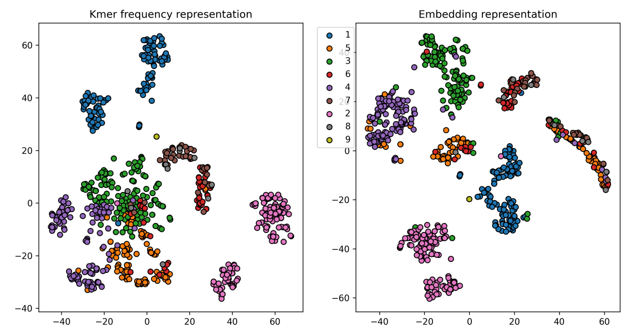
(a) TSNE plots: Top five phyla in 16s dataset



(b) TSNE plot: Fragments positions in 16s dataset



(c) TSNE plots: Top five phyla in ITS dataset



(d) TSNE plot: Fragments positions in 28s dataset

Figure 4. Unveiling Phylum Distribution and Positional Insights: RoBERTa Model Embeddings(right) vs. 6-mer Frequency(left)

average accuracies. Figures 5,6 and 7 display accuracy results for various tasks across different datasets. The combination of 6-mer representation and Random Forest (RF) consistently outperformed the other models. The 6-mer representation demonstrated its effectiveness by capturing local sequence information and effectively discriminating between different taxa. RF's ability to handle high-dimensional data, capture complex relationships, and handle noisy data contributed to its superior performance. It is worth noting that the 6-mer representation achieved a high-dimensional vector size of 4096, which may have contributed to its success. However, we observed a decrease in performance when reducing the k-mer size to 3-mers, indicating under performance of this representation. Therefore, there appears to be a trade-off between representation dimensionality and classification accuracy.

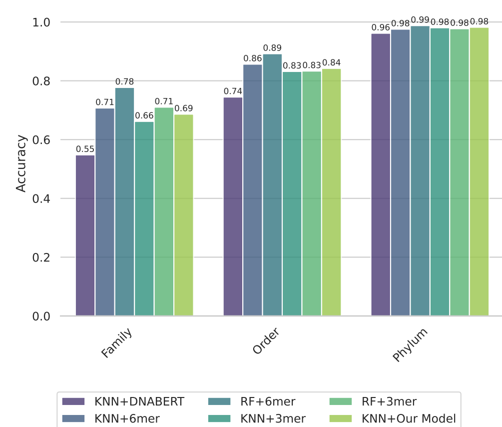


Figure 5. Taxonomic Classification Accuracy at Different Taxonomic Levels (Phylum, Order, Genus) for Different Models Using the 28s rRNA Dataset (10-fold Cross-Validation).

IV. Conclusions

We have demonstrated the effectiveness of our optimized RoBerta model for taxonomic classification across diverse downstream datasets and genes. The model's strong performance showcases its versatility and applicability in various biological contexts, even within the limitation of sequence length, which currently restricts its usage for sequences exceeding $\sim$200 nucleotides—posing a potential challenge for tasks like SNP detection that require analysis of longer genomic sequences. For future work, our optimized RoBerta model opens up possibilities for further downstream analysis. For example, it should be explored for prediction of gene functions. This provides valuable insights into the functional capabilities of different organisms and uncovers novel biological pathways. Moreover, our model holds promise for metagenomic studies, where it can accurately classify and identify the taxonomic

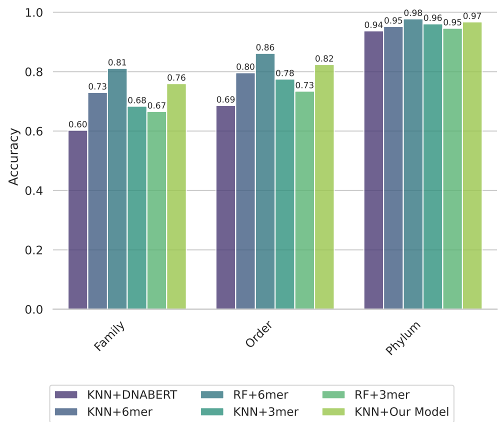


Figure 6. Taxonomic Classification Accuracy at Different Taxonomic Levels (Phylum, Order, Genus) for Different Models Using the ITS Dataset (10-fold Cross-Validation).

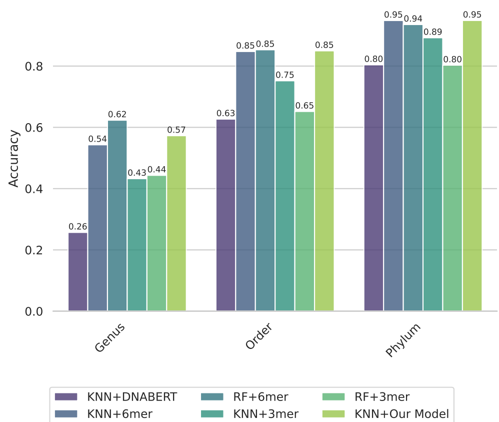


Figure 7. Taxonomic Classification Accuracy at Different Taxonomic Levels (Phylum, Order, Genus) for Different Models Using the 16S rRNA Dataset (10-fold Cross-Validation).

composition of microbial communities. This information contributes to a deeper understanding of ecosystem dynamics, species interactions, and potential functional roles within these communities.

Supplementary Information

The source code and relevant data are available at https://github.com/EESI/MetaBERTa.

Acknowledgements

This work is supported by National Science Foundation under Grant Numbers #1936791, #1919691 and #2107108. We thank the University Research Computing Facility for their paid services.Additionally, we extend our appreciation to Dr. James Brown for his valuable advice in enhancing the manuscript.

References

[1] N. S. Detlefsen, S. Hauberg, and W. Boomsma, "Learning meaningful representations of protein sequences," *Nature communications*, vol. 13, no. 1, pp. 1–12, 2022.

[2] S. Solis-Reyes, M. Avino, A. Poon, and L. Kari, "An open-source k-mer based machine learning tool for fast and accurate subtyping of hiv-1 genomes," *PloS one*, vol. 13, no. 11, p. e0206409, 2018.

[3] S. Woloszynek, Z. Zhao, J. Chen, and G. L. Rosen, "16s rrna sequence embeddings: Meaningful numeric feature representations of nucleotide sequences that are convenient for downstream analyses," *PLoS computational biology*, vol. 15, no. 2, p. e1006721, 2019.

[4] Z. Zhao, S. Woloszynek, F. Agbavor, J. C. Mell, B. A. Sokhansanj, and G. L. Rosen, "Learning, visualizing and exploring 16s rrna structure using an attention-based deep neural network," *PLoS computational biology*, vol. 17, no. 9, p. e1009345, 2021.

[5] C. A. Tataru and M. M. David, "Decoding the language of microbiomes using word-embedding techniques, and applications in inflammatory bowel disease," *PLoS computational biology*, vol. 16, no. 5, p. e1007859, 2020.

[6] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, and M. Linial, "Proteinbert: a universal deep-learning model of protein sequence and function," *Bioinformatics*, vol. 38, no. 8, pp. 2102–2110, 2022.

[7] B. A. Sokhansanj, Z. Zhao, and G. L. Rosen, "Interpretable and predictive deep neural network modeling of the sars-cov-2 spike protein sequence to predict covid-19 disease severity," *Biology*, vol. 11, no. 12, p. 1786, 2022.

[8] J. Vig, A. Madani, L. R. Varshney, C. Xiong, R. Socher, and N. F. Rajani, "Bertology meets biology: interpreting attention in protein language models," *arXiv preprint arXiv:2006.15222*, 2020.

[9] M. Heinzinger, M. Littmann, I. Sillitoe, N. Bordin, C. Orengo, and B. Rost, "Contrastive learning on protein embeddings enlightens midnight zone," *NAR genomics and bioinformatics*, vol. 4, no. 2, p. lqac043, 2022.

[10] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome," *Bioinformatics*, vol. 37, no. 15, pp. 2112–2120, 2021.

[11] J. A. Valeri, L. R. Soenksen, K. M. Collins, P. Ramesh, G. Cai, R. Powers, N. M. Angenent-Mari, D. M. Camacho, F. Wong, T. K. Lu *et al.*, "Bioautomated: An end-to-end automated machine learning tool for explanation and design of biological sequences," *Cell Systems*, vol. 14, no. 6, pp. 525–542, 2023.

[12] Y.-z. Zhang, Z. Bai, and S. Imoto, "Dysfunctional analysis of the pre-training model on nucleotide sequences and the evaluation of different k-mer embeddings," *bioRxiv*, pp. 2022–12, 2022.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[14] M. Meola, E. Rifa, N. Shani, C. Delbès, H. Berthoud, and C. Chassard, "Dairydb: a manually curated reference database for improved taxonomy annotation of 16s rrna gene sequences from dairy products," *BMC genomics*, vol. 20, no. 1, pp. 1–16, 2019.

[15] C. L. Schoch, B. Robbertse, V. Robert, D. Vu, G. Cardinali, L. Irinyi, W. Meyer, R. H. Nilsson, K. Hughes, A. N. Miller *et al.*, "Finding needles in haystacks: linking scientific names, reference specimens and molecular data for fungi," *Database*, vol. 2014, p. bau061, 2014.

[16] P. D. Schloss and et al., "Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Applied and Environmental Microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009. (available at: *https://aem.asm.org/content/75/23/7537*).