CBMFW4761 Computational Genomics

Deep Learning Model for Characterizing Sequence Reads

Sachin Kadyan, Abhishek Iyer and William Kindschuh

Abstract

Motivation: Tremendous amounts of data are generated by NGS. A crucial step in characterizing this data is to map the reads to reference genomes. Depending on the sample vast portions of the data may go uncharacterized if the reference genomes are unavailable. This data may contain significant signals that are lost due to the lack of methods to characterize data without reference assemblies. Another limitation of the current methods is that vast amounts of memory is required to load the references to characterize the reads from sequencing data.
Results: With minimal fine-tuning of a deep learning based DNA language model we were able to achieve performance in sequence classification approaching the performance of standard mapping algorithms. The performance of our model was most comparable to the performance of mapping when the input sequences were mutated at a higher rate (0.1 snps/bp).
Availability: The code is available freely at https://github.com/abhishake07/unmapped_reads.git

Contact: sk4835@columbia.edu, ahi2112@columbia.edu, wfk2109@cumc.columbia.edu
Supplementary Information: Supplementary_info.pdf

1. Introduction

Over the past decades the cost of next generation sequencing has dramatically gone down. This has been a trigger for extensive genome analysis by different consortiums (Wetterstrand, 2020; Mardis, 2011). Some with a focus on sequencing the genomes of healthy humans (1000 genomes), others in tumor samples (TCGA, ICGC) and yet others in the space of metagenomics and microbiomes. A majority of current sequencing technologies rely on a key step called mapping, before further analysis of any kind is performed.

Oftentimes, especially in the metagenomic and microbiome whole genome sequencing spaces, reference genomes are not easily available and hence the majority of reads go uncharacterized (Sangiovanni et.al., 2019; Laine et.al., 2019; Zhu et.al., 2019). Additionally, tumor samples often contain a mixture of tumor cells, normal human cells, microbes and viruses present in the tumor microenvironment. While most of the reads belong to the human genome and map very well, ~2-5% of the reads go uncharacterized (Laine et.al., 2019; Tae et.al., 2014). Recent studies have shown the importance of these reads. Kostic et. al. recently published a study where upon close inspection of unmapped reads in tumor WGS samples, they found an association between fusobacterium and colorectal cancer (Kostic et.al. 2012). Another study by Park et. al. on exploring human microbiome data found that the unmapped reads actually belonged to a completely novel organism called crAssphage (Park et.al., 2020). Laine et. al. have stressed on the importance of characterizing reads that do not map to any genome and were able to identify new pathogenic bacteria that affect songbirds after analyzing their genome (Laine et.al., 2019).

Outside unmapped reads of human or tumor samples, a lot of reads in metagenomic samples (60-70%) go uncharacterized (Zhu et.al., 2019). Recently, there have been some methods being developed and more effort is being put into characterizing reads. A popular technique being used right now for such analysis is Pathseq by the Broad institute. This method uses a multi mapping step to characterize reads in a sequencing experiment (Kostic et.al. 2011). Their main focus is on characterizing reads in human sequencing samples through subtractive mapping. While their approach is highly effective, it is limited as it requires reference genomes to characterize the reads. Further, the first step of their approach is essentially to map to the human genome and then to microbial genomes. This approach is well suited for human sequencing samples as the majority of the reads are of human origin, but would create an unnecessary bias for metagenomic and microbiome samples. Another challenge is that the reference databases required for mapping reads to micro-organisms are huge and often demand a huge memory requirement (Ye et.al. 2019) making such analysis difficult on personal computing systems.

Mapping algorithms are extremely robust and can accurately identify where reads map in a reference genome. However, while often very effective, mapping does not scale well when trying to characterize reads for which reference genomes are unavailable (Ye et.al. 2019). Here we would like to present a deep learning approach to characterize reads in any sequencing experiment. We chose a deep learning based language model known as DNABert as it is reference free and could potentially characterize reads from novel organisms as well, which is an inherent limitation of mapping. There have been some studies that have shown for specific applications where deep learning and machine learning models were used to characterize reads (Deneke et.al. 2017, Tampuu et.al. 2019). We show that with minimal hyperparameter adjustment and fine-tuning of the DNABert model we can accurately classify human and bacterial reads.

Compared to the traditional approach of Path-seq which involves multiple mapping steps, our model takes sequence reads as input and is able to classify between human and bacterial reads. Due to computational limitations, we were unable to directly compare our results with Pathseq, but built a similar approach that we could test our model against. See the methods section below for further details on our model and mapping algorithm implementations.

2. Methods

The data and the methods used for the same are explained below.

2.1 Data Procurement

References for human and bacterial genomes were downloaded from the RefSeq database (NCBI). The data was downloaded from the FTP portal. The GCF_000001405.25_GRCh37.p13 version of the human genome

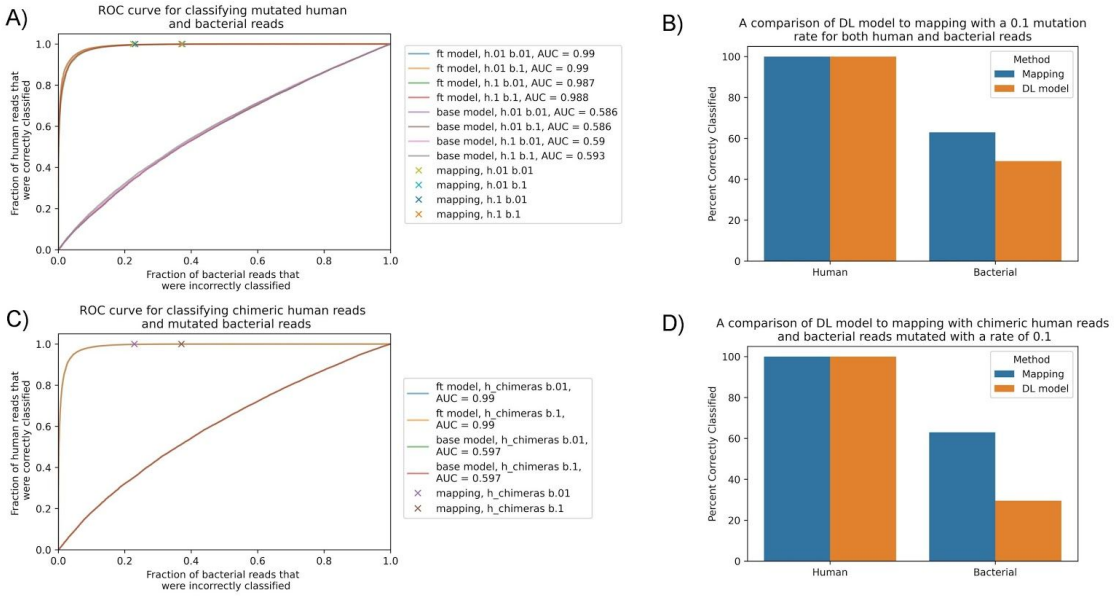

Fig. 1. Performance of mapping and DL model on classifying human and bacterial reads. A) ROC curves for our fine-tuned model as well as the base DNABert model for classifying human and bacterial reads at mutation rates of 0.01 and 0.1 snps/bp (four combinations total). Performance of mapping shown as individual points. B) A comparison of the performance of our fine-tuned DL model to our mapping algorithm at classifying human and bacterial reads mutated at a rate of 0.1 snps/bp. C) ROC curves for our fine-tuned model as well as the base DNABert model for classifying chimeric human

was chosen for all the experiments. Representative bacterial genomes were downloaded from Refseq. (https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/). Data for only organisms that had complete genomes were downloaded (2847 genomes).

2.2 Generation of synthetic reads

Synthetic reads for human and bacterial genomes were created using NEAT-gen reads (Stephens et.al. 2016). For creating mutation datasets we used the mutation rate parameter of the tool to create datasets with mutation of 0.01 and 0.1 snps/bp. Default parameters were used for generating data with coverage of 1 for human genomes and 10 for bacterial genomes. For the generation of chimeric human reads we first generated unmutated human reads using NEAT-gen reads and then randomly merged 50bp substrings from each read in order to generate chimeras.

2.3 Mapping of synthetic reads

GCF_000001405.25_GRCh37.p13 was chosen as the reference human genome. We used BWA (bwtsw) to first index the reference genome and all the test sets were mapped to the human genome with default parameters using BWA mem (Li and Durbin, 2009). Reads that were mapped were tested for accuracy metrics (Supp. Methods 2). Unmapped reads were then classified using Kraken2 (Wood et.al. 2019). Minikraken2 database was used for the analysis as there were memory limitations in using the entire kraken database. The results of this step were again tested for accuracy metrics (Supp. Methods 2).

2.4 Deep Learning model training

In order to build a deep learning based model that could accurately identify reads as either human or bacterial we decided to take advantage of previous work in sequence prediction using deep learning. While many of the earliest methods for sequence prediction or sequence classification using deep learning used convolutional or recurrent neural networks, we chose to base our model on the newer transformer architecture. In 2020 the first transformer based deep learning language model trained on DNA, known as DNABert, was published by Ji et al (Ji et al. 2020). DNABert was trained for masked language modeling on the human genome in which sequences were tokenized into overlapping 6-mers and 15% of tokens in each sequence were randomly masked and the model was trained to predict the identities of the masked tokens. In their paper Ji et al. showed that they could achieve state of the art results on several sequence prediction and classification tasks by then further training (fine-tuning) this pretrained model for specific tasks.

We fine-tuned the DNABert model for the classification of human and bacterial reads by training the model for binary prediction on a dataset of 1 million unmutated human and 1 million unmutated bacterial reads from 6 phyla commonly found in human microbiomes (Supp Method 1). All sequences used for model training and testing were length 100bp. After some limited experimentation with the learning rate and batch size we chose 2e-5 as the learning rate and 32 as our training batch size. Fine-tuning was performed on one NVIDIA 1060 GPU over roughly 30 hours.

2.5 Generation of test sets for both the mapping algorithm and deep learning model

We chose to test the performance of our model's classification compared to the classifications of our mapping algorithm using the following sets of tests: 1) 90% human reads mutated at rates of 0.01 and 0.1 snps/bp and 10% bacterial reads mutated at rates of 0.01 and 0.1 (four combinations total), 2) 90% chimeric human reads and 10% bacterial reads mutated at rates of 0.01 and 0.1 snps/bp. All test datasets contained 200,000 reads total. For each test a corresponding fastq was made that was then used as input for our mapping algorithm. Since our deep learning model does not have a mechanism by which to interpret quality scores, we set the quality scores for all test set reads used for mapping to be the highest possible score. In order to make our tests more challenging we also only used mutated bacterial reads from phyla which did not contribute any reads to our training data.

3. Results

The ROC curve was plotted to compare the proposed model's performance with that of the standard mapping + kraken2 approach. These characteristics are presented in Figure 1 (A) for mutated human and bacterial reads, and Figure 1 (C) for chimeric human and mutated bacterial reads. The figures present the characteristics of the untrained

model (base) and the final trained model. There are separate curves for the four different combinations of mutation rates for the human (0.01 and 0.1 snps/bp) and bacterial reads (0.01 and 0.1 snps/bp). The performance of the mapping-based pipeline has been projected on the plots as multiple single points.

Our proposed deep-learning based model was able to match or approach the performance of a standard mapping-based pipeline in almost all scenarios. The bar charts in Figure 1 (B) and 1 (D) display the percentage of correctly classified human and bacterial reads for the non-chimeric and chimeric reads respectively. For the non-chimeric reads the DL model was able to match the performance of mapping on human reads (both 99.98%) and came close to it on bacterial reads (48.90% vs 62.98%). On chimeric reads, the DL model did not come close to the performance of mapping on bacterial reads (29.53% vs 62.49%) but was able to match it on human reads (both 100%). Still, mapping does seem to incorrectly classify a larger number of bacterial reads at higher mutation rates compared to the DL model.

4. Discussion

Based on the experiments conducted and the results obtained, it can be concluded that the proposed model is able to classify bacterial reads from human reads with high accuracy, despite the fact that the model was tested on reads from phyla on which it was not trained. The high accuracy is sustained when mutations are introduced into the sequences, at high frequencies resembling a rapidly mutating bacterial population. It can be seen in Figure 1 (A) that raising the mutation rate from 0.01 to 0.1 significantly increases the fraction of incorrectly classified bacterial reads by mapping, but this does not appear to be case in the performance of our deep learning model, which is evident in the degree to which the ROCs overlap regardless of the mutation rate.

In comparison to existing methods, the model affords the advantage of having a drastically smaller memory footprint, and hence eliminates the requirement of huge amounts of system memory required in a mapping-based sequence analysis pipeline. The model also does not require pre-curated genome references for training. A small amount of sequenced reads of the target population(s) are sufficient to reach high identification accuracy. Consequently, the memory requirements do not scale higher with respect to the dimension of taxonomy that is used (phyla/taxon/organism).

Still, due to the proposed model being based on the computational approaches of deep learning, it suffers many of the same limitations. The training time is high when compared to a simple indexing procedure on an existing database. Due to the use of a binary classifier as the output node of the model (as of now), it is also currently limited to classifying each read as either human or bacterial.

Sequencing reads contain additional information such as quality scores, which do not currently find a use in the deep learning model. We have also not been able to perform sub-kingdom level classification for bacterial reads, hence the performance of the model in those tasks is unclear.

Mapping-based pipelines are effective after fulfilling the requirement of good reference databases. These techniques would fail when they encounter reads that do not map to the existing reference database (the reason for non-conformity of the read notwithstanding). Deep-learning based models based on their inherent superiority at pattern recognition can serve as potential solutions to such problems.

4.1 Future Prospects

There are many ways that the current model can be extended. To incorporate identification of additional genomes, the deep learning model can be extended to include multi-label classification. Identification of sub-kingdom taxa for bacterial reads is possible through additional relevant data and training pipeline. It would also be worth determining how well a model such as the one described here can distinguish viral from human and bacterial reads.

We believe that Deep Learning based sequence classification models have significant potential in the study of metagenomics as they can be used to identify and characterize unmapped reads in that domain. While the ideal reference for the classification of metagenomic reads via a mapping algorithm may be both enormous and unattainable, it may be possible to build a useful model or family of models for the classification of metagenomic reads using only a small subset of the metagenomic reads space as training data for large language models.

Acknowledgements

We acknowledge the support of Prof Itsik Pe'er, and other course staff for providing us the opportunity to undertake this finding and project, and for their guidance throughout the project and the course.

References

Deneke, C., Rentzsch, R. & Renard, B. (2017) PaPrBaG: A machine learning approach for the detection of novel pathogens from NGS data. Sci Rep 7, 39194.

Heng Li, Richard Durbin (2009) Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*, Volume 25, Issue 14, 754–1760

Ji, Y., Zhou, Z., Liu, H., & Davuluri, R. V. (2020). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *bioRxiv*.

Kostic, A., Ojesina, A., Pedamallu, C. *et al.* (2011) PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat Biotechnol* **29,** 393–396.

Laine, V.N., Gossmann, T.I., van Oers, K. *et al.* (2019) Exploring the unmapped DNA and RNA reads in a songbird genome. *BMC Genomics* **20,** 19

Mardis E. (2011) A decade's perspective on DNA sequencing technology.*Nature*, 470: 198-203.

Park G, Ng T, Freeland AL, et al. (2020) CrAssphage as a Novel Tool to Detect Human Fecal Contamination on Environmental Surfaces and Hands. *Emerging Infectious Diseases*. 26(8):1731-1739.

Sangiovanni, M., Granata, I., Thind, A. *et al.* (2019) From trash to treasure: detecting unexpected contamination in unmapped NGS data. *BMC Bioinformatics* **20,** 168.

Simon, H. Ye, et al. (2019) "Benchmarking metagenomics tools for taxonomic classification." *Cell* 178.4 779-794.

Stephens, Zachary D., et al. (2016) "Simulating next-generation sequencing datasets from empirical mutation and sequencing models." *PloS one* 11.11.

Tampuu A, Bzhalava Z, Dillner J, Vicente R (2019) ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples. PLoS ONE 14(9)

Wetterstrand KA. (2020) DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)

Zhu, Z., Ren, J., Michail, S. *et al.* (2019) MicroPro: using metagenomic unmapped reads to provide insights into human microbiota and disease associations. *Genome Biol* **20,** 154.