15 March 2025 · CC-BY 4.0

Peer Review

Review of: "METAGENE-1: Metagenomic Foundation Model for Pandemic Monitoring"

Francisco Pascoal[1]

[1] Universidade do Porto, Portugal

Dear authors,

This paper presents a new foundational model based on DNA and RNA sequencing data, obtained from wastewater samples. At this stage, the authors performed pre-training and compared the performance of their model with the most similar models in the literature, showing that the METAGENE-1 outperforms other models in specific tasks.

I believe this paper is relevant and technically sound. However, I have some concerns, which I think can be addressed by correcting and/or clarifying specific segments of the main text. In general, the paper should better explain the concrete aims of the model and the output information that the end-user can reasonably expect. It is also important to better discuss how the model explains the output to the end user and to provide additional detail on the underlying metagenomes. Below, I present comments on specific segments of the text, organized by section.

Abstract

In: “the aim of METAGENE-1 is to capture the full distribution of genomic information present within this wastewater, to aid in tasks relevant to pandemic monitoring and pathogen detection.”

- Considering that the sequence length is very short and that there was no attribution of sequences to genomes (metagenome assembled genomes), I think that the model does not capture the “full distribution of genomic information.” It does, however, capture information derived from multiple genomes.

doi.org/10.32388/YYM9HD

Introduction

In: "Consequently, there have been various efforts to collect wastewater and sequence *metagenomic information*, i.e., information about the diverse collections of organisms and organic material present in these samples"

- Replace "sequece metagenomic information" with "sequence metagenomic DNA";
- Remove "and organic material present in these samples," because metagenomics only provides DNA-level information. This data does not inform about metabolites, for example.

In: "A key motivation for much of this work is the potential to track the prevalence of human pathogens, effectively creating an early warning system for pandemics."

- I suggest removing verbose details, like "much of this work" and "A key motivation," for better readability;

In: "Such models could be fine-tuned for various tasks crucial to pathogen monitoring, including tracking frequencies, trends, and growth of different sequence types"

- Please clarify what you mean by "growth of different sequence types" (changes in abundance? Modifications to the underlying sequence? Something else?)

In: "In this paper, we take an initial step toward developing a metagenomic foundation model by pretraining a model on a large, new dataset sequenced from wastewater."

- Replace "dataset sequenced from wastewater" with "dataset of wastewater metagenomic sequencing" (or equivalent), because the dataset is not sequenced; it is the result of the sequencing process);

In: "This metagenomic dataset, which has never before been used for model training, provides a unique resource for modeling the broad distribution of sequences present in the human microbiome."

- If the objective were to model the broad distribution of sequences present in the human microbiome, why use wastewater-derived metagenomes instead of human gut, skin, etc., microbiome data? I believe that the point to be made, instead, is that we may identify what is happening collectively in a municipality by identifying changes in specific microbial groups or other patterns.
- Additionally, if the data is from wastewater metagenomics, then the model is accounting for human and non-human genomic information; please clarify this in this segment.

In: “Specifically, we pretrain a 7-billion-parameter autoregressive transformer model, which we refer to as METAGENE-1, on a diverse corpus of DNA and RNA sequences comprising over 1.5 trillion base pairs sourced from wastewater samples, which were processed and sequenced using deep metagenomic (next-generation) sequencing[20][11].”

- In this segment and in general, the underlying data used is not very clear because the paper mostly refers to DNA (metagenomics), but it also includes RNA data (metatranscriptomics). It is my understanding that the authors include DNA and RNA in a single idea “metagenomics.” However, this might be confusing because metagenomics usually refers to the DNA-level information. Later on, the authors mention in a footnote that the RNA sequences were transcribed back to DNA, so the model, as far as I understand, only used DNA information, which I think is appropriate. However, I think the authors need to explain this better in the main text.
- I think you can remove “(next-generation),” because it is implied.

In Figure 1:

- It might be better to refer to metagenomics and metatranscriptomics, or sequencing of DNA and RNA, or indicate that there is a step that transcribes RNA back to DNA.

Related work

In “these samples contain potentially tens-of-thousands of species across a wide range of taxonomic ranks, and capture a representative distribution of the full human-adjacent microbiome.”

- Please add references to support the claims:

1. "wastewater contains tens-of-thousands of species across a wide range of taxonomic ranks";
2. that these wastewater samples capture a representative distribution of the human-adjacent microbiome.

Section 3

In: “We pretrain a 7-billion-parameter autoregressive transformer language model, referred to as METAGENE-1, on a novel corpus of diverse metagenomic DNA and RNA sequences comprising over 1.5 trillion base pairs.”

- Please be more specific. How many bp from DNA and how many from RNA?
- Maybe “environmental DNA and RNA” would be better than “metagenomic DNA and RNA” in general because metagenomes refer to DNA.

In: “This dataset is sourced from a diverse set of human wastewater samples, which were processed and sequenced using deep metagenomic (next-generation) sequencing methods.”

- Replace “human wastewater” with “municipal wastewater”;
- How many samples were used for metagenomics (DNA) and metatranscriptomics (RNA)?
- Indicate that this is municipal input wastewater;
- How many different wastewater influents were used?
- What was the sequencing depth per sample? For DNA and RNA.

Section 3.1

In: “material from a broad range of organisms, including bacteria, viruses, cells from human and other eukaryotes, and a diverse array of other species”

- Remove “and a diverse array of other species,” because the previous enumeration already includes all biological groups, except Archaea;
- The wording is a bit confusing because it enumerates biological groups from different taxonomic levels. I suggest simplifying to something like: “DNA from all domains of life, including human DNA.”

In: “we aim to model the complexity of microbial and viral interactions in human-associated environments”

- The authors mention a new aim here, which is to “model the complexity of microbial and viral interactions,” but this was not mentioned before in the objectives of the model; additionally, I don't think this model can identify microbial interactions based on the information available.
- I suggest replacing this aim with more concrete objectives, like pathogen detection.

In: “next-generation sequencing (NGS) or high-throughput sequencing, in which billions of nucleic acid fragments are simultaneously sequenced in a massively parallel manner”

- Even though NGS (with Illumina in this example) sequences billions of nucleic acid fragments, this huge number of nucleotides is associated with the amplification step, in which the same original sequence is multiplied several times. Furthermore, the same sequence might be represented by several reads, all of which are then multiplied several times during the sequencing itself. I don't think it is necessary to include the segment "in which billions of nucleic acid fragments are simultaneously sequenced in a massively parallel manner", because those unaware of the technology's mechanism may get a wrong idea. I think it is better to mention concrete numbers, for example, the number of different sequences, the number of total reads, and their length.

In: "This method produces paired-end reads, where each read consists of two contiguous sequences of base pairs from opposite ends of a DNA or RNA fragment[1]"

- The information in this footnote should be in the main text. As I mentioned previously, the authors seem to conflate DNA and RNA information, even though they represent distinct biological levels;
- So, it should be clear to the reader, at this stage, that the tokenization process will only account for DNA data.

In: "After wastewater samples were collected, the material was filtered and nucleic acids extracted[28][29] before undergoing metagenomic sequencing."

- Please add more methodological information:

1. the number of samples used for metagenomic/metatranscriptomics.

- Additional information that can be provided in Attachments:

1. wastewater sampling protocol;
2. DNA and RNA extraction protocol.

Figures 5 and 7:

- The color labels are difficult to read; consider a different visualization approach or a higher-resolution figure.

In: "In full, the metagenomic dataset for pretraining comprises over 1.5 trillion base pairs."

- In line with previous comments, I would like to see how the authors obtained this number. Probably, answering the previous comments (number of metagenomes, sequencing depth, etc.) will be enough.
- I also want to understand better if this number of nucleotides is derived from unique sequences or if it includes repeated sequences (several reads per sequence).

Section 3.2. Tokenization

In: “We opted for byte-pair encoding (BPE) as our tokenization method, as it satisfies these criteria, and drawing inspiration from its successful application in recent genomic models.”

- Please add references.

In: “The method also has the potential to capture semantic information within a vocabulary of tokens, which can lead to more nuanced representations of genomic data.”

- I think this should be explained in the context of genomics, to not confuse the reader with the meaning of “semantics”. Maybe add some references that explain how semantic concepts employed in LLMs can be translated to genomics.

In: “BPE’s ability to tokenize novel sequences is particularly valuable for modeling diverse metagenomic sequences containing unknown, varied, and possibly novel organisms.”

- This sentence needs clarification. You may have an unknown sequence from a known organism (in this case, the challenge lies in the gene annotation), or you might have an unknown sequence from an unknown origin (but in this case, what reliable information can your model obtain?), or you might have a known sequence (e.g., a known functional gene) that belongs to a previously unknown organism. These situations represent different biological problems. So, I think this sentence can be simplified to avoid confusion. For example, I would suggest something like: “BPE’s ability to tokenize previously unknown sequences is valuable because it allows the model to account for the novelty and diversity of genetic and taxonomic information from metagenomic data.”
- If you are including unknown sequences in the tokenization and pre-training, isn't this a potential source of noise?

In “This vocabulary size strikes a balance between capturing sufficient genomic complexity, maintaining sufficiently long sequence lengths (based on the distribution of token sizes), and allowing for computational efficiency.”

- How did you measure “genomic complexity”?
- I think the context size mentioned in Table 1, which is just 512 bp, is short to capture genomic complexity.

Comments and questions on tokenization:

- What is the frequency of tokens?
- Did you test your results for a different set of tokens?
- Did you consider bi-directionality and reverse complementarity issues in the tokenization of DNA sequences?
- The authors mention that BPE accounts for “different sequence patterns” and “structures,” can you be more specific regarding those "structures"?
- I think it would be important to mention some token examples and provide their size range (e.g., provide the minimum, maximum, and mean length). I see from your GitHub page that you have tokens like “AA” with two nucleotides, but also larger tokens like “ATTTCACCGC.” I think this information is useful for readers.

Section 3.3

In: “For our metagenomic foundation model, we pretrain a 7-billion-parameter autoregressive language model, using a standard dense transformer architecture, similar to the architecture used in popular language models such as the GPT and Llama model families[21][22]. Specifically, we implement a decoder-only style transformer with a causal language modeling objective, where the model aims to predict the next token in a sequence based on the previous tokens.”

- It is worth mentioning examples of foundational models that used this architecture in the context of DNA data because the examples provided are focused on chat applications.

Table 1:

- The context length is small to identify genetic complexity. Many functional genes have thousands of nucleotides. Additionally, very similar sequences might result in very different outcomes. Therefore, why not assemble reads in larger contigs? Could this improve the context length?

Section 4.3.

In: “a curated selection of genomes from multiple species (e.g., fungi, mammalian, invertebrate, bacteria)”

- This sentence should be rewritten because it mentions “multiple species” and then exemplifies non-species level groups: fungi (kingdom level); mammalian (class level); invertebrates (includes several

doi.org/10.32388/YYM9HD

animalia phyla); bacteria (domain level). Thus, instead of mentioning them as examples, simply state that those curated genomes were derived from species from fungi, mammalian, etc.

Section 5.2.

In general:

- I suggest changing the section title to “Human-infecting virus detection benchmark”; and in the remaining text of this section, clarify that you are testing the identification of human-infecting viruses, and not pathogens in general (pathogens can also be bacteria, protists, etc).
- The source of the training (1600) and test (2000) samples is unclear.
- The MCC metric is appropriate, but I would also like to see the underlying contingency table for METAGENE-1 (True positives, False positives, True negatives, False negatives).
- Can you exemplify the output provided by the model when a human-infecting virus is detected? What information about the sequence is made available? How does the model justify to the end user that a specific sequence might be from a pathogen?

In: “(1) Kraken 2[30][5] identifies at least one hit on a *k*-mer associated with a human-infecting virus, and (2) the read aligns with a human-infecting virus genome in GenBank[6].”

- Is this data available? The footnotes refer to Kraken and GenBank in general.
- What viruses were selected? What is the length of the sequences used?

Section 6

- This section mentions valid concerns; however, the specific concerns are raised without referencing studies with concrete evidence of such risks, thus it reads a bit speculative. Please clarify how speculative the mentioned risks are and/or provide concrete references.

Discussion

- Please discuss to what extent the model outputs can be explainable. For example, in human pathogen detection, what makes the algorithm decide that a sequence is pathogenic or not?
- I would like some discussion on the decision to transcribe RNA back to DNA and the methodology to do so (mRNA might undergo splicing, for example). I believe that probably this decision was made to

increase the size of the pre-training data, while not increasing too much complexity.

- I would like more details on the current limitations of this model.
- What are the advantages of possible implementations of this model vs. the traditional approaches (e.g., using this model vs. using BLAST to identify a potential pathogen);

In: “In Figure 7, we show a visualization of (a relatively small subset of) the composition of metagenomic information contained in our pretraining dataset”

- Replace “a relatively small subset of” with the size of the subset.

General questions:

Overall, there were some details that I missed:

- During pre-training, were the sequences labeled? In other words, what information about each sequence was provided?
- Did the pre-training account for the number of reads of each sequence?

I hope my feedback was helpful,

Congratulations on your work,

Francisco Pascoal

Declarations

Potential competing interests: No potential competing interests to declare.