A long-context language model for the generation of bacteriophage genomes

Bin Shao[1,2*]

[1] Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA

[2] Present address: Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA

*Correspondence should be addressed to Bin Shao (shaobin@broadinstitute.org)

Abstract: Generative pre-trained transformers (GPTs) have revolutionized the field of natural language processing. Inspired by this success, we develop a long-context generative model for genomes. Our multiscale transformer model was pre-trained on unannotated bacteriophage genomes with byte-level tokenization. It generates *de novo* sequences up to 96K with functional genomic structure, including regulatory elements and novel proteins with phage-related functions. Our work paves the way for the *de novo* design of the whole genome.

Large pre-trained language models have drastically transformed the natural language processing (NLP) field[1,2]. Drawing on the similarity of natural language and genome sequences, genomic language models have been developed. These models were trained on large scale genomic datasets, and they effectively predict regulatory elements, uncover co-regulation patterns in protein and identify genome-wide variant effects[3–7]. However, it remains an open question whether language models can be tailored to generate genome-scale sequence with functional structure, which holds potential for the rational design of the whole genome.

Most of the current models used masked language modeling like BERT (Bidirectional Encoder Representations from Transformers)[1] which is not ideal for tasks that involve generating new content. In addition, current models face technical constraints such as short context size and aggregation of sequences in k-mer tokenization. These limitations hinder their ability to efficiently learn from genome scale data while maintaining the resolution needed for precise design of functional elements.

In this work, we developed a long-context language model that can generate *de novo* sequence with functional genomic structure. Our model draws inspiration from the GPT model[2], which is renowned for its proficiency in generating long and coherent texts. We utilized a multiscale transformer structure[8] that enables us to train the model on unannotated bacteriophage genomes up to 96K bp at the single nucleotide-level. The trained model generates sequences that share similar genomic structure with the natural bacteriophage genomes. We found functional promoter and ribosome binding sites (RBS) in the 5' untranslated regions (5'UTR) of the predicted genes. In addition, the proteins from the generated sequences are predicted to be structurally plausible and span a wide variety of functional families. We believe our model is a timely advance that paves the way for DNA design at the genome scale. The model is available from GitHub: https://github.com/lingxusb/megaDNA

To construct the training dataset, we collected bacteriophage genomes with high confidence from three sources including NCBI genebank, Metagenomic Gut Virus (MGV) catalogue[9] and Gut Phage Database (GPD)[10] (Supplementary Fig. 1). After data cleaning, we constructed a dataset with 99.7K bacteriophage genomes up to 96K bp, which was used to pre-train our model (Methods). The training data was byte-level tokenized. We employed the multi-scale transformer structure with long-context length from Yu et al.[8] and our model is named megaDNA, correspondingly. In model inference, we generated a total of 1,024 sequences longer than 1K bp. Then geNomad[11] was used for functional annotation of the generated sequences. Among all these sequences, 607 of them have a virus score larger than zero and we focused our analysis on them. Their mean sequence length is 43K bp and the mean number of

predicted genes is 67, both are similar to the training dataset (Mean length: 48K bp, mean number of predicted genes: 68). The gene length distribution is close to that of the training dataset (Fig. 1b, average gene length: 558 bp vs 646 bp), while the predicted gene numbers show wider spread (Fig. 1c). The median virus score of these generated sequences is 0.84 and the maximum virus score is 0.97, comparable to the virus scores for natural bacteriophage genomes which range from 0.70 to 0.98 (Fig. 1c). 223/607 (37%) of the generated sequences are predicted to be Caudoviricetes by geNomad (Fig. 1d). As a comparison, 98% of the training dataset was classified as Caudoviricetes.

We then examined the 5'UTR region of the annotated genes in the generated sequences to see whether they contain functional regulatory elements like promoters and RBS to initiate transcription and translation. We chose generated sequence #87 for further analysis due to its high predicted virus score (0.96) and relatively small size (28K bp). Using a machine learning tool (Promoter Calculator)[12], we identified the -35 box and -10 box of the promoter within the 5'UTR of the predicted gene. Notably, they closely aligned with the established consensus motifs: TTGACA and TATAAT (Fig. 1f). Prior to the start codon of the predicted phage stabilization protein, we observed a region enriched in adenine (A) and guanine (G) nucleotides, which is a motif characteristic of functional ribosome binding sites (Fig. 1f). Analyzing all 5'UTR sequences of the predicted genes from this sequence, we observed significantly higher mean promoter activity compared to random sequences of the same length. (Fig. 1g). Intriguingly, the proportion of A and G nucleotides peaked around 10 bp upstream of the start codon, close to the optimal position for RBS to drive translation initiation (Fig. 1h). This trend of A/G enrichment is also consistent across all the generated sequences (Supplementary Fig. 2). In short, our generated sequences harbor functional regulatory sequences that would enable expression of the predicted genes.

Among the annotated genes in our generated sequences, 343 of them were predicted to match geNomad markers. We employed ESMfold[13] to predict structures for these genes and calculated the average predicted local distance difference test (pLDDT) score. This score reflects the confidence of ESMfold on the predicted structures. The median pLDDT score for these proteins is higher than the pLDDT of random peptide sequence of the same length (28 vs 18). We also randomly sampled 10K annotated genes from the generated sequences and found a high pLDDT score for them (median value of 36, Supplementary Fig. 3), suggesting that these generated proteins are more likely to adopt a stable conformation. We further used deepFRI[14] for functional annotation of all generated proteins and we only retained proteins with high scores (> 0.5). Our analysis reveals several large protein families with functional roles, including the transporter activity and the structural molecule activity (Fig. 1j).

Interestingly, we identified several proteins with DNA-binding activity, and the predicted structure resembles the canonical helix-turn-helix (HTH) domain in this protein family (Supplementary Fig. 4).

To the best of our knowledge, our work presents the first long-context generative model for genomic sequences. Our language model effectively learns the high-order genomic grammar via a single step of self-supervised training on unannotated whole genomes. The generated sequences match the length of natural bacteriophage genomes and display functional genomic architecture. With further scaling up, we envision that the generative genomic models will pave the way for the *de novo* design of the genome sequence, which offers opportunities for breakthroughs in medicine, agriculture, and environmental science. This field also faces ongoing challenges in ethical considerations, biosafety, and regulatory frameworks, which are critical for the responsible advancement of generative modeling in synthetic biology.

Figures

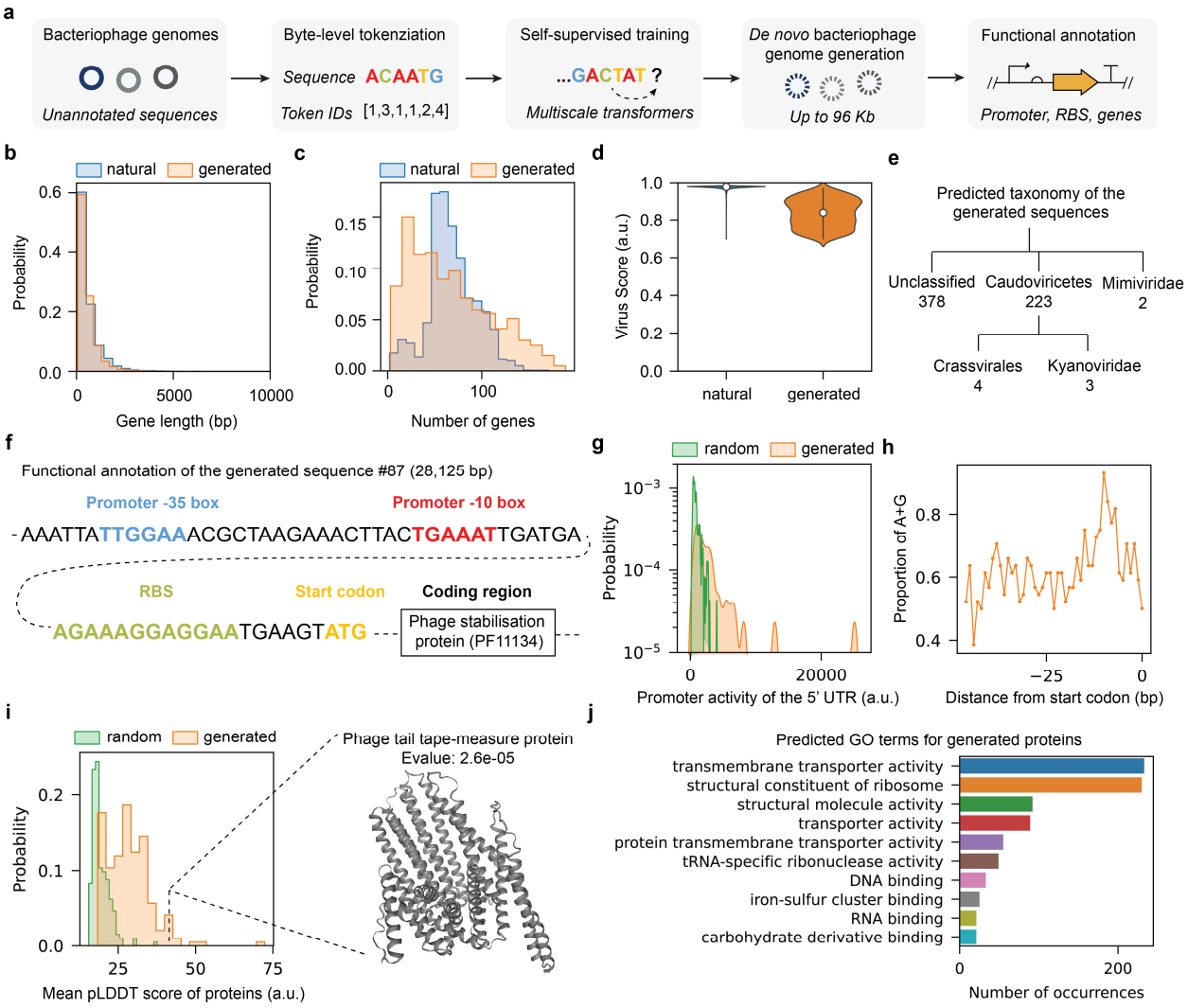


Figure 1. Language model generates sequences with functional genomic structures.

a) the workflow schematic. **b**) comparison of gene length distributions between predicted genes in generated sequences (n = 40,399) and a randomly sampled subset of genes from the training dataset (n = 10,000). **c**) distributions of the number of predicted genes for the generated sequences (n = 607) and the training dataset (n = 99,673). **d**) comparison of the predicted virus scores for generated sequences and the training dataset. **e**) predicted taxonomy for the generated sequences. Only taxonomies with > 1 sequence are shown. **f**) functional annotation of a selected sequence fragment (generated sequence #87). **g**) predicted promoter activity for all the 5'UTRs in the generated sequence #87 (n = 44), along with the promoter activity of the random sequences with the same length. **h**) proportions of adenine (A) and guanine (G) nucleotides preceding the start codon of all the predicted genes in the generated

sequence #87. **i**) mean predicted pLDDT scores for proteins with geNomad markers from generated sequences (sample size: n = 343; median value: 28) against random peptide sequences of the same lengths (sample size: n = 343; median value: 18). A sample generated protein is shown on the right. **j**) top 10 predicted functions of proteins derived from the generated sequences.

References

1. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

2. Brown, T. *et al.* Language models are few-shot learners. *Adv Neural Inf Process Syst* **33**, 1877–1901 (2020).

3. Ji, Y., Zhou, Z., Liu, H. & Davuluri, R. V. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* **37**, 2112–2120 (2021).

4. Dalla-Torre, H. *et al.* The nucleotide transformer: Building and evaluating robust foundation models for human genomics. *bioRxiv* 2021–2023 (2023).

5. Benegas, G., Batra, S. S. & Song, Y. S. DNA language models are powerful predictors of genome-wide variant effects. *Proceedings of the National Academy of Sciences* **120**, e2311219120 (2023).

6. Hwang, Y., Cornman, A. L., Kellogg, E. H., Ovchinnikov, S. & Girguis, P. R. Genomic language model predicts protein co-regulation and function. *bioRxiv* 2023–2024 (2023).

7. Nguyen, E. *et al.* Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *arXiv preprint arXiv:2306.15794* (2023).

8. Yu, L. *et al.* Megabyte: Predicting million-byte sequences with multiscale transformers. *arXiv preprint arXiv:2305.07185* (2023).

9. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat Microbiol* **6**, 960–970 (2021).

10. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098–1109 (2021).

11. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nat Biotechnol* 1–10 (2023).

12. LaFleur, T. L., Hossain, A. & Salis, H. M. Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria. *Nat Commun* **13**, 5159 (2022).

13. Lin, Z. *et al.* Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv* **2022**, 500902 (2022).

14. Gligorijević, V. *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat Commun* **12**, 3168 (2021).

Methods

Training dataset

Our training dataset was curated from three sources. Firstly, we downloaded all the complete virus genomes from NCBI genebank, retaining only those with "phage" in the organism's name. Secondly, the phage genomes from MGV were downloaded, and we only included genomes with a completeness score larger than 95% and classified under the order Caudovirales. Our third source was GPD, and we kept all the genomes with a completeness score above 0.95. Following the initial collection, we undertook an additional round of filtering. We used geNomad to predict the taxonomy of these genomes and then deleted all the genomes whose predicted host is not a unicellular organism. All genomes smaller than 96K bp were used to construct the final training dataset.

Model training and inference

Our megaDNA model utilized a three-layer transformer structure[8]. Each layer had a depth of 8 and progressively fewer dimensions (512, 256, 196), capturing local-to-global information. The sequence length for three layers is 128, 64, 16. The model contains 145M parameters in total. We assigned numerical tokens (1, 2, 3, and 4) to the nucleotides A, T, C, and G, respectively. For model training, we used a batch size of 1 and set the learning rate at 0.0002. The learning rate was progressively increased during the initial 50,000 steps as part of a warmup schedule. We utilized the Adam optimizer and applied gradient clipping with a norm of 0.5 to prevent gradient explosion.

We generated sequences from the trained model using a predefined set of parameters. Specifically, we adjusted the temperature to 0.95 to ensure a balance between variety and coherence in the sequences and kept the filter threshold at 0.0 to avoid limiting the range of token probabilities. For model training and inference, we utilized Nvidia's A100 GPU (40GB) and 3090 Ti GPU (24GB) and used the PyTorch version 2.1.1 software package.

Analysis of the generated sequence

geNomad[11] was used for sequence annotation of all generated sequences. The 100 base pair regions preceding the start codon of each predicted gene was designated as the 5'UTR. We employed the

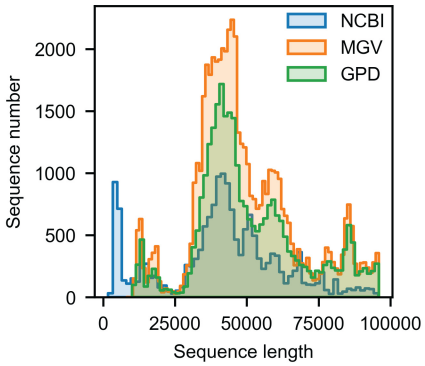
Promoter Calculator[12] to find the promoters in these regions. Only the promoter with the highest predicted activity was annotated. For protein structure prediction, we used the pretrained ESMfold model v1[13]. The chunk size of the model was set to be 64 for proteins longer than 700 AA and 128 for shorter proteins. We limited our structure calculation to proteins less than 1000 AA in length. Function prediction for these proteins was carried out using the default deepFRI model[14], as available on GitHub (https://github.com/flatironinstitute/DeepFRI). We used a score cutoff of 0.5 which was reported to be significant in the original publication.

Data availability

The bacteriophage genomes were downloaded from public databases including NCBI genebank (ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/), MGV (https://portal.nersc.gov/MGV), and GPD (https://www.sanger.ac.uk/data/gut-phage-database/).

Code availability

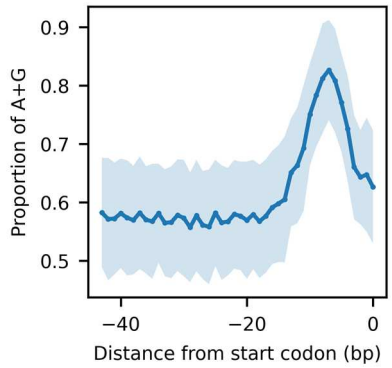
Our trained model and codes for model inference are available from GitHub:
https://github.com/lingxusb/megaDNA



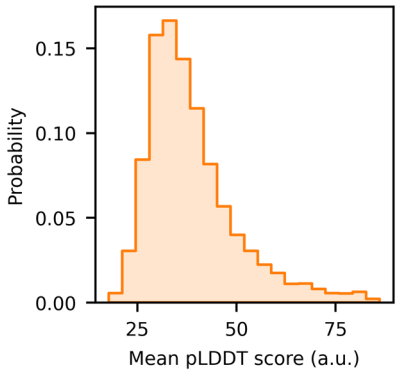
191

192 **<u>Supplementary Figure 1:</u> Distribution of genome sizes of three data sources.** Distributions of genome

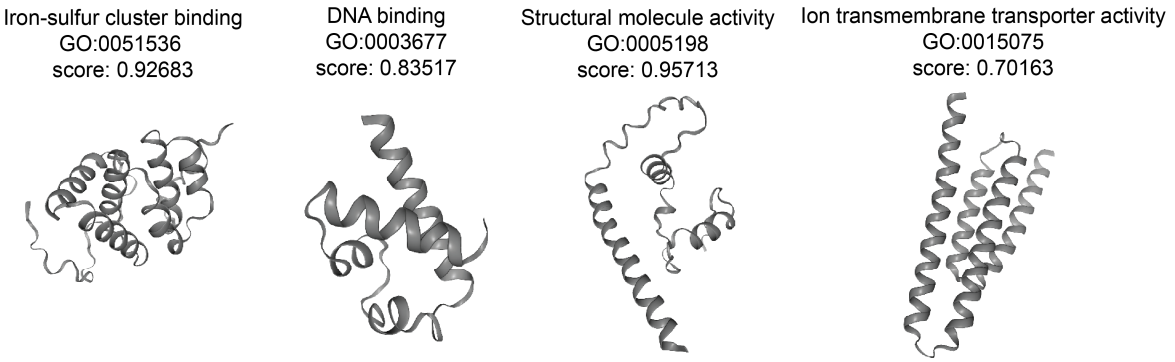
193 sizes within the training dataset: NCBI (n = 16,609), MGV (n = 53,032) and GPD (n = 30,032).



<u>Supplementary Figure 2:</u> Proportions of adenine (A) and guanine (G) nucleotides preceding the start codon for all the generated sequences. Blue line denotes the mean A+G nucleotides proportion profile for all the generated sequences (n = 607). The shaded region represents the standard derivation of all profiles.



Supplementary Figure 3: Mean pLDDT score for proteins derived from the generated sequences. The distribution from a randomly sampled subset of the generated proteins is shown (sample size: n = 10,000; median value: 36).



203

204 **<u>Supplementary Figure 4:</u> Representative proteins from the generated sequence with predicted**

205 **functions and structures.** The protein structures were predicted using ESMfold[13] and the functions were

206 annotated using deepFRI[14]. Predicted scores and GO terms from deepFRI are shown.