HervEVO: Accurate Classification and Generation of Human Endogenous Retroviruses with DNA Foundation Model

Jieli Zhou[1*] Duo Du[2*] Hongyi Xin[1†] Lei Liu[2†]
[1]Shanghai Jiao Tong University, Shanghai, China
[2]Fudan University, Shanghai, China
[*]Both authors contributed equally to this research.
[†]Corresponding authors: **hongyi.xin@sjtu.edu.cn** and **liulei_sibs@163.com**

ABSTRACT

Human Endogenous Retroviruses (HERVs) make up approximately 10% of the human genome and are linked to a variety of diseases. Accurate classification of HERVs, determination of their coding statuses, and generation of new HERV sequences are essential for advancing research in therapeutics and vaccines. In this study, we present HervEVO, a novel long-context DNA foundation model framework that accurately classifies HERV sequences with an overall AUROC over 92% and a PRAUC over 93%. Specifically, HervEVO achieves 97% AUROC for non-coding HERV sequences and 91% AUROC for coding HERV sequences. Additionally, HervEVO is finetuned to generate high-fidelity coding HERV sequences, which can enhance classification model training and support vaccine development. This generative system not only opens new avenues for medical advancements involving HERVs but also deepens our understanding of their functions.

1 INTRODUCTION

HERV sequences, remnants of ancient viral insertions into the human genome, are implicated in various complex diseases, including cancer [2]. A recent study [1] highlighted the capabilities of DNA language models such as DNABert [3], to classify HERV sequences. However, these DNA language models have several shortcomings, such as restricted input size (512bp), lack of sequence generation capability, and inaccurate classification results. This study tackles these challenges by building on a long-context DNA foundation model, EVO [4], to generate high-fidelity coding HERV sequences and improve classification results through model distillation.

2 HUMAN ENDOGENOUS RETROVIRUS DATASET AND PROBLEM DEFINITION

The HERV sequence dataset, sourced from a recent study [1], was filtered to exclude sequences longer than 8K bp due to HervEVO's input size limitation and the average sequence length of 1200 bp. This filtering retained 99.32% of the original dataset, preserving 774,797 out of 780,095 sequences. This dataset was initially created using Bedtools[1] to generate the necessary regions. Subsequently, the corresponding sequences were extracted based on these regions' coordinates, and gene annotation information was utilized to compile the final dataset.

The classification task is then framed as a 4-class imbalanced classification problem: coding non-Herv (18,454 or 2.38%), coding Herv (16,015 or 2.06%), non-coding non-Herv (341,869 or 44.12%), and non-coding Herv (398,516 or 51.43%). With over 95% of samples being non-coding, the classification task is extremely imbalanced and challenging. The generation task is to generate high-quality coding Herv sequences.

[1]https://bedtools.readthedocs.io/en/latest/

3 HERVEVO SYSTEM DESIGN AND PRELIMINARY RESULTS

We developed HervEVO based on EVO, a long-context DNA foundation model based on the StripedHyena architecture [4]. EVO has 7 billion parameters and a context window of 8K. However, since it is only pre-trained on prokaryotic and phage genomes, EVO has poor performance on human genomic tasks [4]. In this work, we extend EVO to both classify and generate Herv sequences.

First, we developed a classification branch with EVO embeddings (using mean and last token embeddings) and K-mer representations (denoted as HervEVO-mean and HervEVO-last). As shown in Table 1, the overall classification results as measured in AUROC and PRAUC are much better than previous models, while other metrics being comparable. We further investigated the classification results of sequences in coding and non-coding regions separately, and as shown in Table 2, the model outperforms the coding region in the non-coding region, which may be closely related to the number of sequences in the dataset (95.56% non-coding vs 4.44% coding).

Motivated by this observation, we finetune EVO using next token prediction supervision on the 16,015 coding HERV sequences to augment the classification dataset. Finally, the augmented dataset is further utilized for model distillation to refine the classifier part. We also analyze the generated HERV sequences using bioinformatics tools like LTR Finder [5] to ensure its generation quality.

Table 1: Model Performance on the HERV 4-Class Classification task, HerEVO has better overall classification performances compared to other DNA language models.

Method	Accuracy	F1	Precision	Recall	AUROC	PRAUC
HervEVO-mean	0.8788	0.8715	0.8796	0.8788	**0.9192**	**0.9308**
HervEVO-last	0.8781	0.8713	0.8802	0.8781	**0.9202**	**0.9324**
DNABert	0.8880	0.8895	0.8989	0.8880	0.8322	0.8455
HumanGPT2	0.8954	0.8956	0.8985	0.8954	0.8369	0.8515

Table 2: Comparison of HervEVO's performance in detecting HERV sequences in coding and non-coding regions.

Region	% of data	AUROC	PRAUC
Coding	4.44%	0.911	0.899
Non-coding	95.56%	**0.972**	**0.977**

REFERENCES

[1] Duo Du, Fan Zhong, and Lei Liu. 2023. Enhancing Recognition and Interpretation of Functional Phenotypic Sequences through Fine-Tuning Pre-Trained Genomic Models. *bioRxiv* (2023), 2023–12.

[2] Johan Jakobsson and Michelle Vincendeau. 2022. SnapShot: Human endogenous retroviruses. *Cell* 185, 2 (2022), 400–400.

[3] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* 37, 15 (2021), 2112–2120.

[4] Eric Nguyen, Michael Poli, Matthew G Durrant, Armin W Thomas, Brian Kang, Jeremy Sullivan, Madelena Y Ng, Ashley Lewis, Aman Patel, Aaron Lou, et al. 2024. Sequence modeling and design from molecular to genome scale with Evo. *bioRxiv* (2024), 2024–02.

[5] Zhao Xu and Hao Wang. 2007. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research* 35, suppl_2 (2007), W265–W268.