OXFORD

Briefings in Bioinformatics, 2025, **26(3)**, bbaf224

https://doi.org/10.1093/bib/bbaf224

Problem Solving Protocol

VirNucPro: an identifier for the identification of viral short sequences using six-frame translation and large language models

Jing Li[1,2], Jia Mi[1], Wei Lin[2], Fengjuan Tian[2], Jing Wan[1,*], Jingyang Gao[1,*], Yigang Tong[2,*]

[1]The College of Information Science and Technology, Beijing University of Chemical Technology, No. 15 North Third Ring East Road, Chaoyang District, Beijing 100029, China

[2]The College of Life Science and Technology, Beijing University of Chemical Technology, No. 15 North Third Ring East Road, Chaoyang District, Beijing 100029, China

[*]Corresponding authors. Jing Wan, The College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China. E-mail: wanj@mail.buct.edu.cn; Jingyang Gao, The College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China. E-mail: gaojy@mail.buct.edu.cn; Yigang Tong, The College of Life Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China. E-mail: tongyigang@mail.buct.edu.cn

Abstract

Viruses are ubiquitous in nature, yet our understanding of them remains limited. High-throughput sequencing technology facilitates the unbiased revelation of genetic composition in samples; however, viral sequences typically make up a small proportion of the entire sequencing data, making it challenging to accurately identify the few or fragmented viral sequences present in a sample. The limited features and information provided by short sequences result in insufficient resolution of viral sequences by existing models. Therefore, we propose a new model, VirNucPro, for short viral sequence identification. Based on a six-frame translation strategy and large language models, we combine nucleotide and amino acid sequence information to enhance feature extraction for short sequences, achieving high accuracy in identifying short viral sequences. Ablation experiments compared the contributions of nucleotide and amino acid sequence features to the model, confirming that the introduced amino acid features significantly contribute to the classification results. Our model outperforms others, such as GCNFrame, DeepVirFinder, DETIRE, and Virtifier, which have demonstrated good performance in identifying short viral sequences of 300 and 500 bp. Our model demonstrates excellent performance on carefully created real-world datasets. Additionally, it can scan for prophage regions within long bacterial fragments, offering a wide range of applications. The codes are available at: https://github.com/Li-Jing-1997/VirNucPro.

Keywords: VirNucPro; viral short sequences identifier; six-frame translation; large language models

Introduction

Viruses, as microorganisms that depend on other species for survival, are ubiquitous in nature. However, our understanding of viruses is limited, perhaps only representing one-thousandth of their true scale. Characterizing viruses forms the foundation for the future development of virology [1, 2]. Metagenomic method, which can objectively reveal microbial genomes within samples, have overcome the limitations of genome acquisition. This advancement aids in understanding the scale and diversity of the virosphere, contributing to the development of virology [3]. However, in metagenomic sequence data, the proportion of viral data is relatively low due to the small genome size of viruses and the low abundance of inactive viruses in samples, which makes detection challenging [4, 5]. For example, in a study of viruses in alpine permafrost, the sample with the lowest viral proportion contained only 230 viral reads, accounting for less than 0.01% of the total [6]. Typically, viral reads account for less than 5% of the total metagenomic sequence data, and in cases of mixed infections, less abundant viruses are often missed, possibly due to the low read count or for reporting purposes [7, 8]. At common sequencing depths, viral data may be insufficient to cover complete viral genomes, making the assembly and identification of viral sequences challenging [9]. Fragmented viral sequence segments are a more common occurrence. Therefore, more accurate strategies for identifying short viral sequences are essential.

The method of identifying pathogen information based on homologous sequence alignment provides extremely useful functionality due to its high accuracy. However, homologous sequence alignment heavily relies on sequences from reference databases, which only gradually expands the known virus repertoire [10]. This approach has limited effectiveness in identifying the highly diverse viral 'dark matter.' Some artificial intelligence methods, such as VirGrapher [11], VirSorter2 [12], and VIBRANT [13], have shown good results in identifying whether long sequences are viral sequences. However, a study has indicated that the accuracy of tool significantly decreases for sequences below 3 kb [14]. The lack of rich feature information in short sequences adds difficulty to sequence identification. A comparative study of virus identification tools has shown that K-mer-based tool DeepVirFinder [15] and BLAST-based tools MetaPhinder excel at identifying viruses

Received: November 19, 2024. **Revised:** March 26, 2025. **Accepted:** April 26, 2025

from short (< 3 kb) viral genome fragments [7]. In addition, several tools for identifying short sequences have been developed, such as Virtifier, which is based on long short-term memory networks [16], and DETIRE, which is built on a hybrid deep learning model [17]. However, the most models only consider nucleic acid sequences, which limits their ability to capture richer features, especially in the case of shorter sequences. In recent years, the development of large language models has provided more accurate and effective text feature extraction methods for many fields, as well as added available tools for sequence identification. Recently, LucaProt, a method for identifying viral RdRp amino acid sequences based on protein language models, has demonstrated the effectiveness of large models in viral identification tasks and confirmed the excellent performance of structural information extracted from protein sequences for viral recognition tasks [18].

In conclusion, in sequencing data with low pathogen abundance, the recovered viral genomes are usually shorter, and the available feature information is limited. To improve identification accuracy and enhance the ability to uncover the viral ‘dark matter’ in the real world, additional methods for acquiring feature information need to be incorporated. Compared to the four constituent units (A, G, C, T) in nucleotide sequences, natural amino acid sequences are building with 20 constituent units, providing richer feature information. In this study, based on a six-frame translation strategy, we supplemented the corresponding amino acid sequence information based on nucleotide sequences, and used large language models based on nucleotide (DNABERT_S [19]) and amino acid (ESM2 [20]) sequences to extract sufficient features, addressing the issue of limited available features in short sequences. We proposed a new model using nucleotide-amino acid features for identifying short viral sequences. By incorporating large language models for both nucleic acid and amino acid sequences, the feature information provided by short sequences is enriched, thereby enhancing the recognition performance of fragmented viral short sequences across various sequencing datasets.

Materials and methods

Data preparation and preprocessing

We constructed training and test sets for short sequences of 300 bp and 500 bp in length. We constructed the positive dataset composed of viral RefSeq genomes, and the negative dataset that includes genomes from archaea, bacteria, fungi, vertebrates, plants, invertebrates, and protozoa. 18,723 viral RefSeq sequences were download as positive sequences from NCBI (https://ftp.ncbi.nlm.nih.gov/refseq/release/) as of July 11, 2024. And their corresponding gbff annotation files were downloaded for extracting translation information. All sequences were uniformly fragmented, and the resulting fragments were translated in all six reading frames. Identify sequences that can be translated without termination and the resulting amino acid sequences can be matched to those annotated in the gbff files. Retain the nucleotide sequences that meet the requirements as samples. This process resulted in 1,020,000 fragments of 300 bp length and 460,000 fragments of 500 bp length, as well as their corresponding amino acid sequences, forming the positive dataset. Using the same approach, we randomly and proportionally downloaded and extracted 1,020,000 fragments of 300 bp length and 460,000 fragments of 500 bp length, along with their corresponding amino acid sequences, from the remaining seven classes in the negative dataset. Ten percent of the samples were then randomly selected from both positive and negative datasets to form the test set.

Model building

VirNucPro uses nucleotide sequences and their corresponding amino acid sequences from the training dataset as input, embedding with large language models (Fig. 1A). Specifically, the nucleotide sequences are processed using the nucleotide language model DNABERT_S, resulting in a 768-dimensional feature vector, while the amino acid sequences are processed with the protein language model ESM2-3B to obtain a 2560-dimensional feature vector. DNABERT-S is a foundation model built upon DNABERT-2, leveraging the BERT architecture and pretrained on a vast, multi-species genomic dataset. Specifically designed to generate DNA embeddings, DNABERT-S can effectively cluster and differentiate the genomes of various species in the embedding space, enhancing its utility for a wide range of genomic classification tasks [19]. ESM2 is a state-of-the-art protein language model built upon the Evolutionary Scale Modeling (ESM) framework [20]. Trained on a massive dataset of 250 million protein sequences, ESM captures predictive representations of proteins’ biochemical and biological properties, including their functions. These two pretrained biological language models helps incorporate rich prior knowledge into our viral classification task. After concatenating all embedded features, a two-layer MLP was used to create a binary classification model for identifying viral and non-viral sequences. During the training process of this study, the MLP’s hidden layer consists of 512 neurons, and the final layer contains two neurons, reflecting the prediction of whether the input sequence is viral or non-viral. The output prediction is made through the softmax function, generating the binary classification scores for viral and non-viral sequences. The optimizer is defined using Stochastic Gradient Descent (SGD) with a learning rate of 0.0002 and a momentum of 0.9. A learning rate scheduler, StepLR, is implemented to decrease the learning rate by a factor of 0.85 every 10 steps. The DataLoader is configured to load the training dataset in mini-batches of size 32, with shuffling enabled for better training dynamics. In order to prevent overfitting, the model also incorporates a dropout layer with a dropout rate of 0.5 and we adopted a five-epoch early stopping mechanism based on the test set. The models for the 300 bp and 500 bp sequences were trained for 109 and 85 epochs, respectively, at which points the loss no longer showed improvement, indicating that training was complete.

For sequence prediction, VirNucPro takes nucleotide sequences as input (Fig. 1B). Upon inputting the nucleotide sequences, they are screened for translatable regions using the six-frame translation method based on standard genetic code to identify potential coding sequence (CDS) areas. The nucleotide translation order, direction, and the resulting amino acid sequences are retained. The input nucleotide and amino acid sequences are then processed through the model, where DNABERT_S and ESM-2 are used for embedding. After embedding, the features are classified using a trained MLP layer. For sequences that can form multiple potential coding sequences through different reading frames, the viral and non-viral prediction scores for each translation are computed. The sequence’s final prediction is determined by the highest score among all the predictions, ensuring that the model retains the predictions most closely aligned with the real-world translation products learned during training.

Evaluation criteria

We evaluate the model’s performance using several key metrics derived from true positives (TPs, examples correctly labeled as positive), false positives (FPs, examples incorrectly labeled as

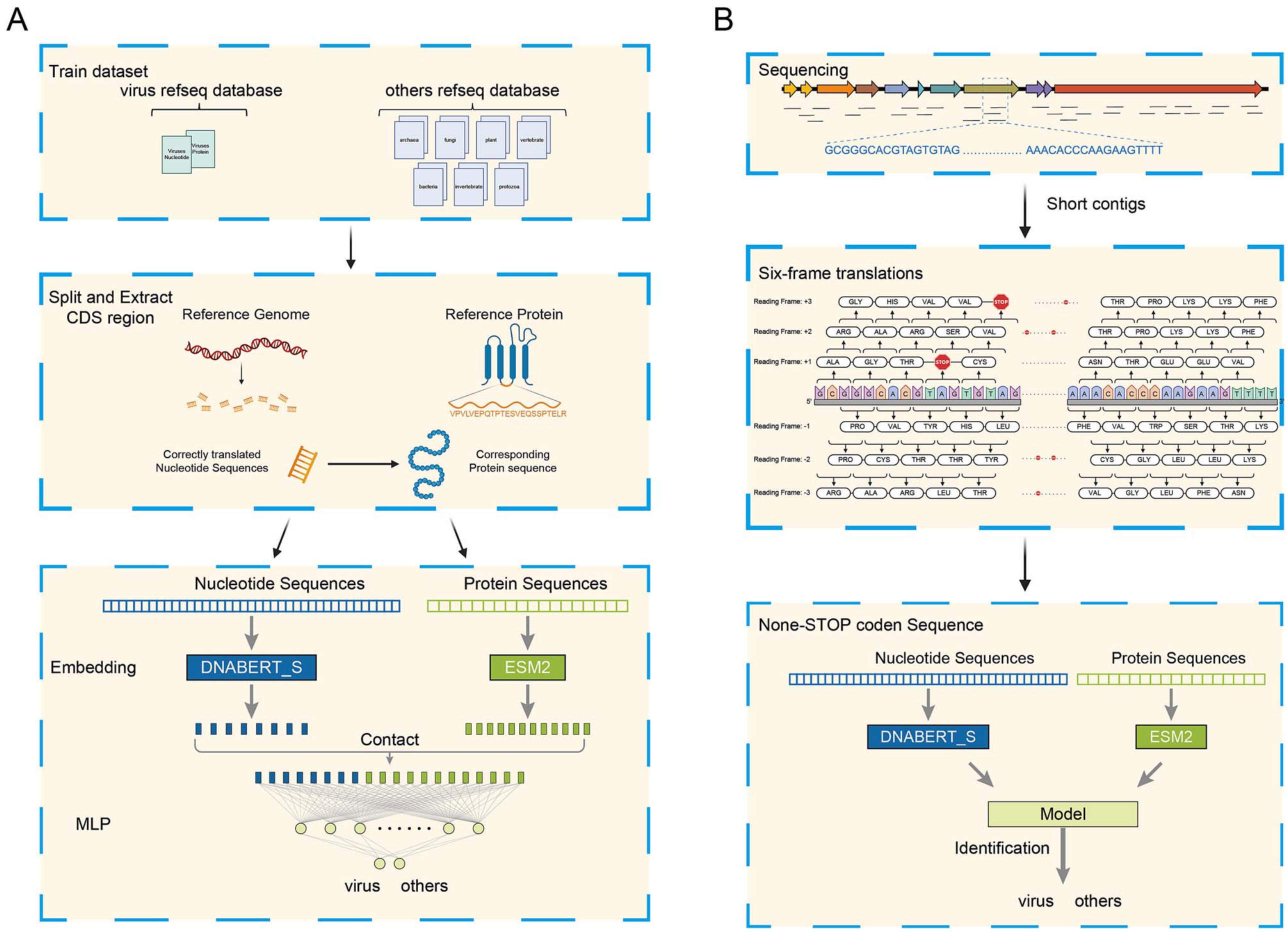
A
Train dataset
virus refseq database
Viruses Nucleotide
Viruses Protein
others refseq database
archaea
fungi
plant
vertebrate
bacteria
invertebrate
protozoa
Split and Extract CDS region
Reference Genome
Reference Protein
VPVLVEPQTPTESVEQSSPTELR
Correctly translated Nucleotide Sequences
Corresponding Protein sequence
Nucleotide Sequences
Protein Sequences
Embedding
DNABERT_S
ESM2
Contact
MLP
virus
others
B
Sequencing
GCGGGCACGTAGTGTAG .... AAACACCCAAGAAGTTTT
Short contigs
Six-frame translations
Reading Frame: +3 GLY HIS VAL VAL STOP THR PRO LYS LYS PHE
Reading Frame: +2 ARG ALA ARG SER VAL THR PRO LYS LYS PHE
Reading Frame: +1 ALA GLY THR STOP CYS ASN THR GLU GLU VAL
5' G C G G G C A C G T A G T G T A G A A A C A C C C A A G A A G T T T T 3'
Reading Frame: -1 PRO VAL TYR HIS LEU PHE VAL TRP SER THR LYS
Reading Frame: -2 PRO CYS THR THR TYR CYS GLY LEU LEU LYS
Reading Frame: -3 ARG ALA ARG LEU THR VAL GLY LEU PHE ASN
None-STOP coden Sequence
Nucleotide Sequences
Protein Sequences
DNABERT_S
ESM2
Model
Identification
virus
others

Figure 1. VirNucPro schematic diagram. (A) VirNucPro training data screening and construction process. (B) VirNucPro workflow for predicting short sequences.

positive), true negatives (TNs, examples correctly labeled as negative), and false negatives (FNs, examples incorrectly labeled as negative). These metrics include the true positive rate (TPR), false positive rate (FPR), precision, recall, and F1 score. Based on the TPR and FPR, a receiver operating characteristic (ROC) curve can be plotted to assess the performance of the proposed model. The ROC curve is created by plotting the relationship between the TPR and the FPR. The area under the ROC curve (AUROC) is then used to evaluate prediction performance, with a higher AUROC value indicating better model performance.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Recall(TPR) = \frac{TP}{TP + FN} \tag{2}$$

$$FPR = \frac{FP}{FP + TN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

Results

The feasibility of only identifying CDS regions

The genomes of viruses are typically small, and to provide normal survival functions, viruses accomplish complex life activities through densely encoded amino acids. We investigated the proportion of CDS in the total genomes of all species in the dataset (Fig. 2). Compared to other species, the CDS of viruses occupies the highest proportion in complete genomes, followed by archaea and bacteria. Therefore, we consider the identification of viral CDS sequences as a reliable criterion for defining the presence of viruses.

VirNucPro performance on the testing dataset

To assess the outstanding performance of VirNucPro, we compared it with commonly used tools for virus identification, including DeepVirFinder [15], and tools designed for short sequence prediction, such as DETIRE [17] and Virtifier [16], as well as a novel genomic encoding framework, GCNFrame [21], using the same training and testing datasets. Using the same training and testing datasets. The accuracy, precision, recall, F1 score, and AUROC score of VirNucPro on sequence lengths of 300 bp and 500 bp were superior to those of the other tools (Fig. 3).

The rejection ability of VirNucPro

Additionally, we examined the performance of VirNucPro in rejecting short sequences from other species. The model performed best in rejecting vertebrate sequences, with rejection rates of 96.18% and 98.64% for sequences of 300 bp and 500 bp, respectively. This was followed by plant sequences, with rejection rates of 94.70% and 98.07%. The model performed the worst on bacteria, rejecting 87.47% and 95.05% of short sequences at

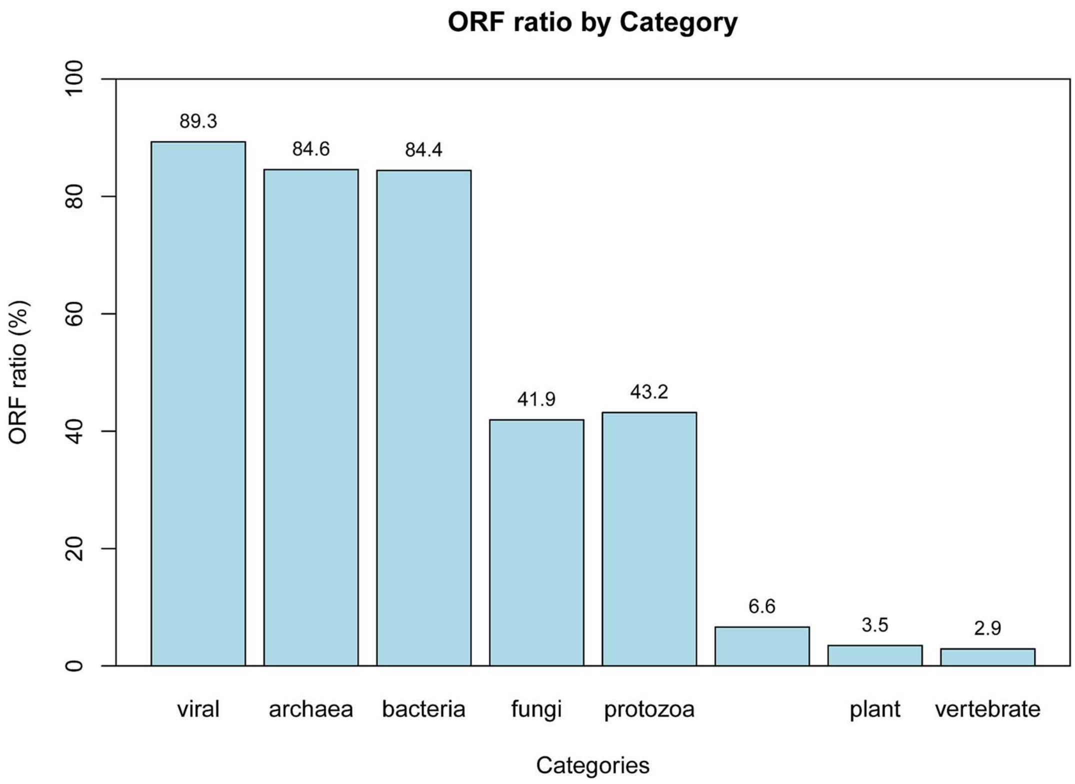

Figure 2. The proportion of CDS in different species in the data used in this study.

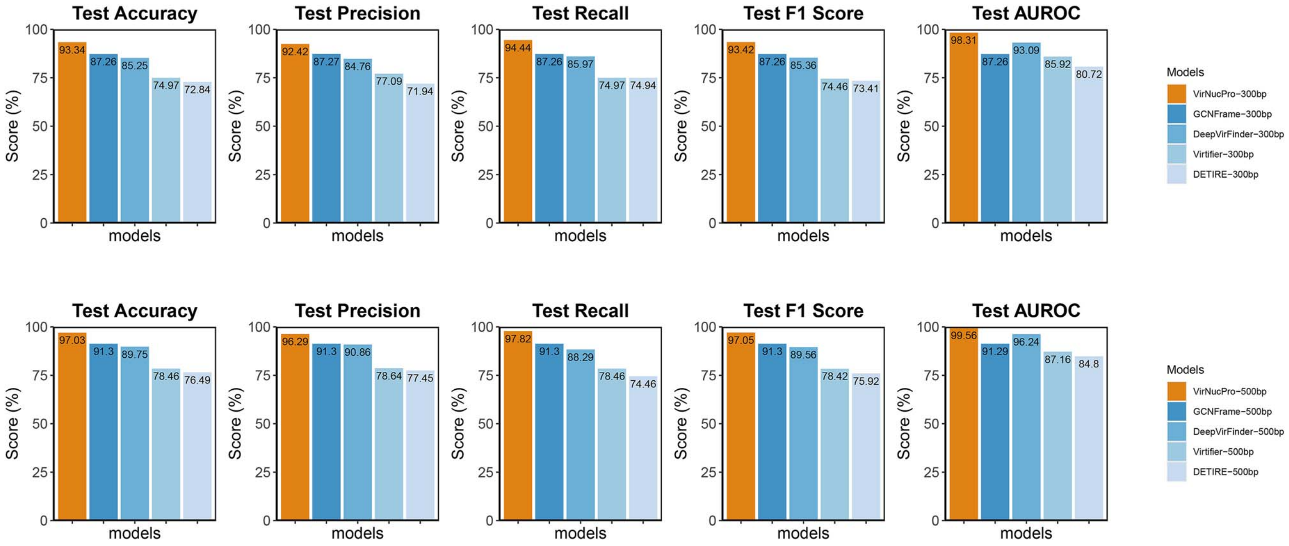

Figure 3. Performance of VirNucPro on 300 bp and 500 bp sequence lengths.

300 bp and 500 bp, respectively (Fig. 4). A possible explanation is that bacteria, in their long-term host-pathogen arms race with bacteriophages, undergo complex processes of gene integration, exchange, and other forms of element transfer [22, 23]. After fragmentation, bacterial sequences that contain mobile elements of viral origin exhibit more viral characteristics, causing them to be recognized as viral sequences by VirNucPro, which leads to a decrease in identification accuracy.

Ablation study

To compared the contributions of nucleotide and amino acid sequence features to the model, we trained classification models using only the amino acid features extracted from ESM2 and only the nucleotide features extracted from DNABERT_S. We found that both DNABERT_S, based on nucleotide sequences, and ESM2, based on amino acid sequences affect the identification performance. Features of ESM2 from amino acid sequences provided more assistance in predicting viral sequences, with classification models using only amino acid features achieving AUROCs of 97.01 and 99.06 at 300 bp and 500 bp lengths, respectively (Fig. 5). This indicates that the additional amino acid feature information improves the model's ability to recognize viral sequences compared to models using only nucleotide features.

Moreover, using DNABERT_S to extract nucleotide embeddings showed limited performance in classifying viral and non-viral sequences. A similar trend was observed with ViraLM [24], a model

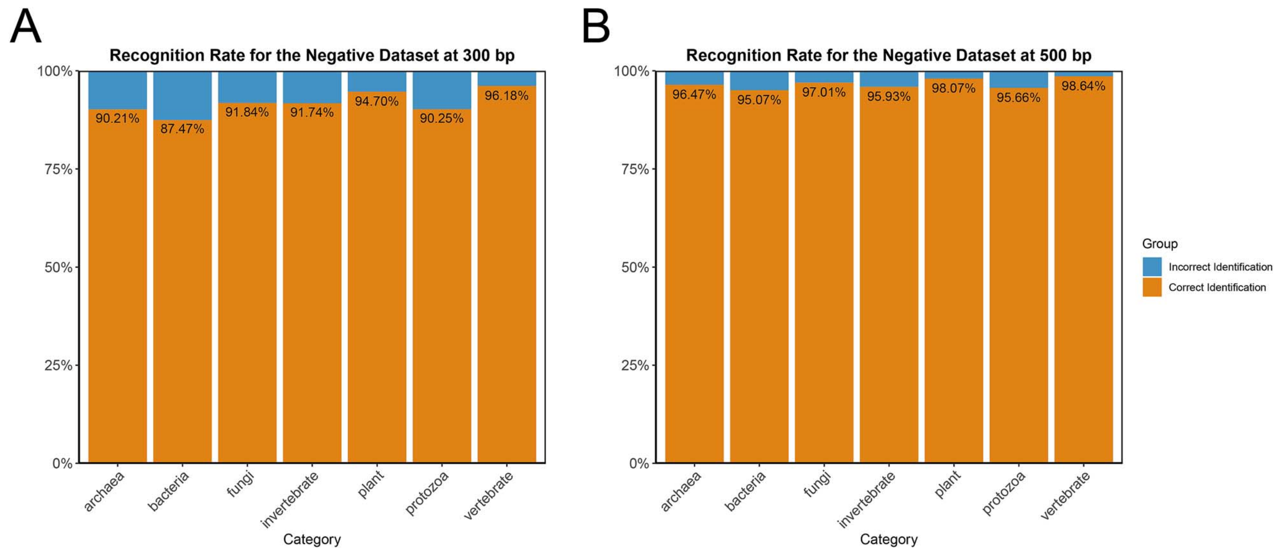

Figure 4. Performance of VirNucPro on rejecting other species.

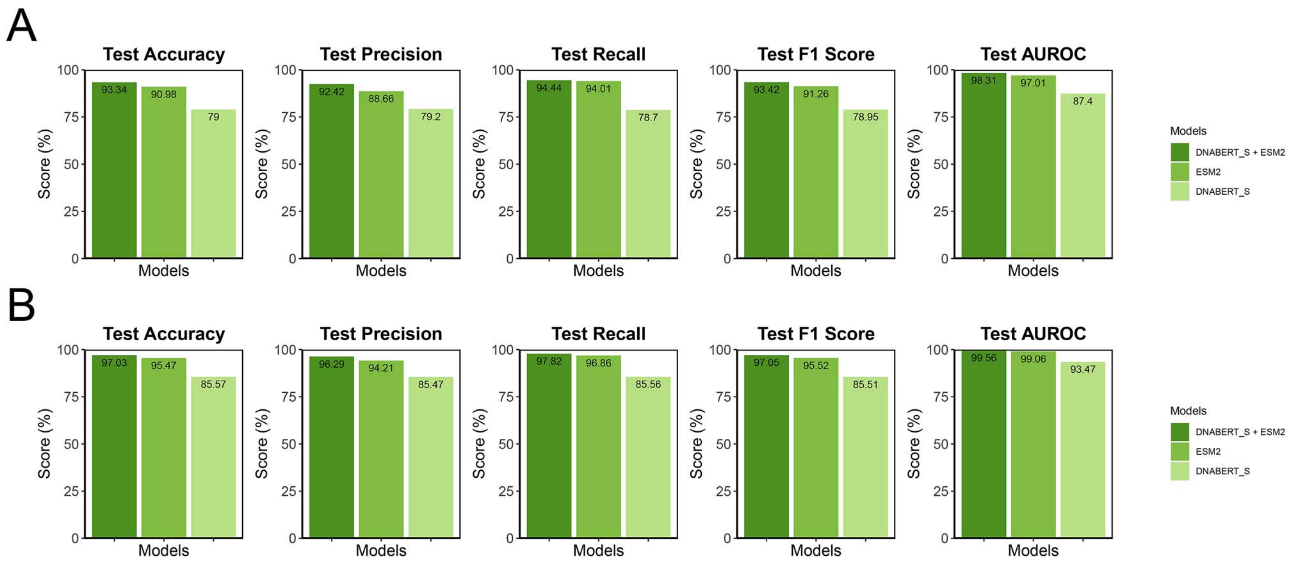

Figure 5. Ablation experiments with 300 bp and 500 bp lengths.

fine-tuned on DNABERT_S, which showed limited performance at 500 bp, but the model's classification ability improved as the sequence length increased. The varying impact of nucleotide and amino acid models on classification performance may be related to the different dimensions of information they provide. Amino acids directly correspond to the structure and function of microorganisms, and their sequences are generally more conserved in evolution compared to nucleotide sequences, which tend to exhibit larger variations. For many organisms, the conservation of amino acid sequences better reflects inter-species relationships and functional similarities than nucleotide sequences, providing stronger classification signals. Additionally, when using large models to embed sequence features, amino acid representations, with 20 distinct units, are more complex than nucleotide representations, which have only four units. This complexity allows amino acid sequences to capture more dimensional information during large-scale data learning, contributing to improved model performance.

VirNucPro's performance in viral identification on real-world data

We utilized a carefully created set of real-world metagenomic data as a test dataset to evaluate VirNucPro's ability to identify viruses in complex samples [7]. This sequencing data was physically separated at the sample level into viral and non-viral groups, generating eight paired viral and microbial datasets for each of the three biomes. After assembly, only contigs longer than 1500 bp were retained for tool evaluation, ensuring high complexity and authenticity. We incorporated our tool into this benchmarking study for comparison. In our tools, all contigs were fragmented into 300 bp and 500 bp sequences, and each fragment was classified as viral or non-viral. A relatively strict criterion was applied to define a sequence as viral—at least 70% of its detected fragments needed to be predicted as viral. A more stringent criterion would further reduce the model's false rejection rate (FRR), while also decreasing the number of viral contigs predicted by the model. Under this condition, VirNucPro outperformed

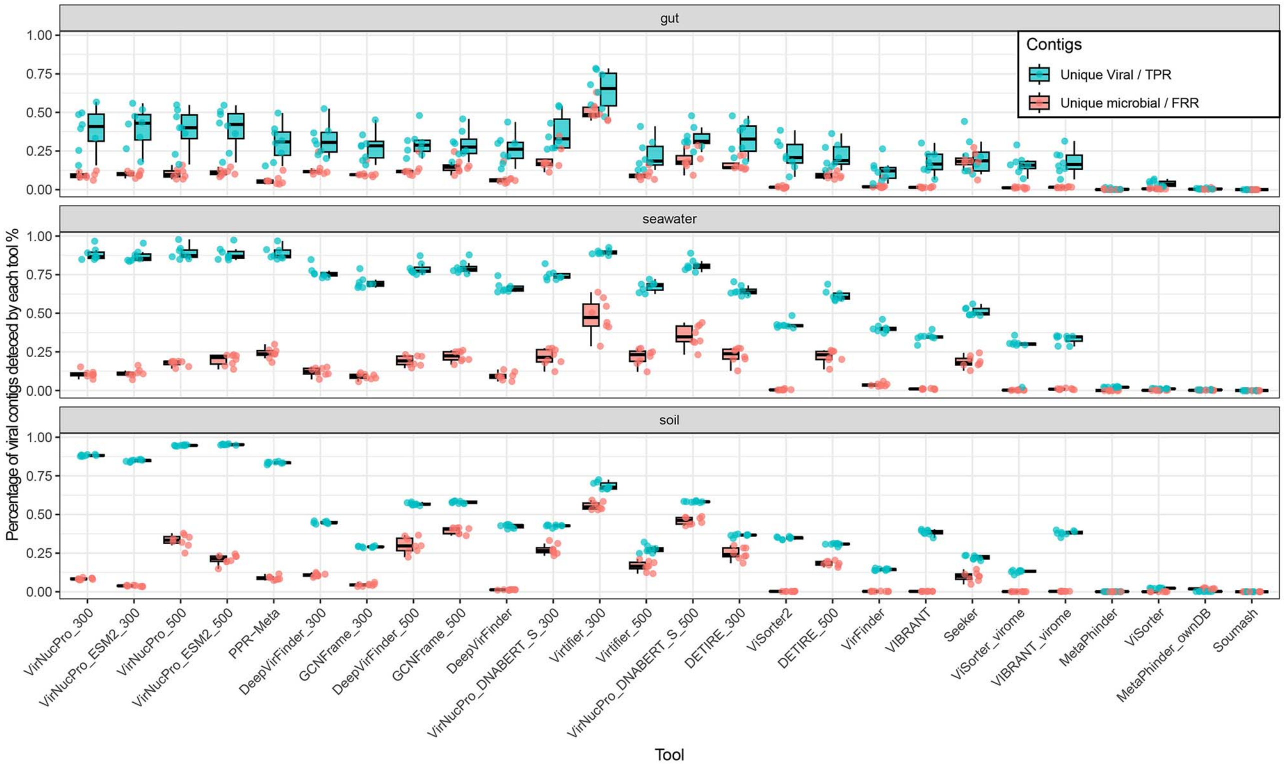

Figure 6. The percentage of contigs identified as viral in viral datasets (true positive rate, blue) and microbial datasets (false positive rate, red), ranked by the average difference in detection rates across eight paired viral and microbial datasets in gut, seawater, and soil biomes.

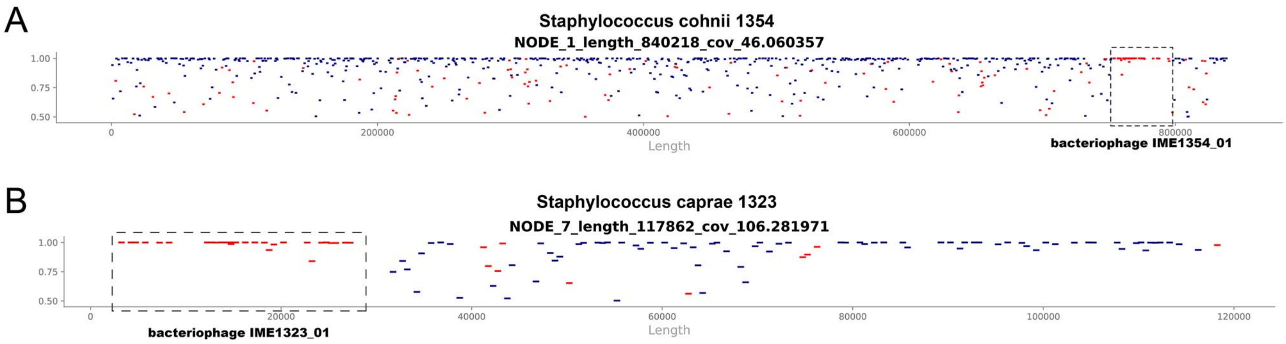

Figure 7. VirNucPro identifies prophages effects in bacterial genomes.

PPR-Meta, the best-performing method from prior evaluations, at both 300 bp and 500 bp sequence lengths (Fig. 6).

The gut dataset may have been affected by processing challenges, which resulted in insufficient separation between viral and microbial populations, leading to reduced accuracy in identification. However, VirNucPro was still able to identify a greater number of viral contigs with a lower FRR. In the higher-quality datasets of seawater and soil, VirNucPro performed best at 300 bp length, identifying more viral contigs in a low FRR. At 500 bp length, VirNucPro identified even more viral sequences compared to the 300 bp length, but the FRR slightly increased, particularly in the soil dataset. In this study, DeepVirFinder and GCNFrame, which were trained on short sequence fragments, exhibited a similar trend. This may be because, at the same total length, shorter fragments provide more units for classification, thereby increasing the basis for decision-making. Longer fragment segmentation may cause some short coding sequences to be masked by surrounding stop codons, potentially affecting classification performance.

Additionally, we compared the models trained in the ablation study using only amino acid sequences or only nucleotide sequences. Consistent with the ablation study results, the amino acid-only model demonstrated strong classification performance on the test data. At 300 bp, its performance was second only to the complete VirNucPro model. At 500 bp, its performance in the seawater dataset was slightly lower than that of the complete VirNucPro, while in the soil dataset, it identified more viral sequences than complete VirNucPro, but with a significantly lower FRR.

Overall, VirNucPro outperformed all other tools when handling complex datasets, particularly with the 300 bp model, which effectively balanced sensitivity and specificity in viral identification. Its strong performance can be attributed to the amino acid sequence features introduced through six-frame translation. In some cases, these amino acid features alone were sufficient to achieve high classification accuracy.

VirNucPro's ability to identify phages in bacterial contigs

We demonstrated the potential application of VirNucPro in specific scenarios. VirNucPro can fragment long sequences and identify coding sequences to detect the presence of viral sequences. Bacterial prophages can insert their genomes into host sequences and engage in complex gene exchange with the host during evolution, making them more challenging to identify. We validated VirNucPro's ability to scan for viral sequences within long fragments using a 500 bp segmentation length. Two phage sequences previously confirmed in studies were tested [25, 26]. In this section, we retained the classification results and scores for all complete translation products from the six-frame translation, with viral and non-viral identification results represented in red and blue, respectively, showed in Fig. 7 to display the virus scanning results across long sequences. The results indicate that VirNucPro can recognize prophages within long contigs as multiple contiguous viral segments, which stand out within bacterial genomes. This demonstrates VirNucPro's ability to identify viral segments within long sequences.

Discussions and conclusions

Here, we propose a new model, VirNucPro, for the identification of short viral sequences. VirNucPro utilizes a six-frame translation strategy to extract fully translated amino acid sequences from short nucleic acid sequences. Using two pre-trained large language models, DNABERT_S and ESM2, it extracts features from both nucleic acid and amino acid sequences, providing the model with sufficient feature information and increasing the accuracy of identification. Our ablation experiment results show that amino acid sequence features contribute the most to identifying viral sequences, demonstrating that supplementing amino acid sequence information is essential and effective for short sequence identification.

To be confidently applied to environmental data, virus identification tools must be trained on representative sequences from the microbial communities under study, ensuring that the tool has encountered enough sequence space to correctly classify viral sequences [14]. Compared to other viral classification models, we used more complex negative sample training and testing, which posed additional challenges for the model, it also broadens the model's applicability. In the test set containing all species, the model performed the poorest at rejecting bacterial sequences at 300 bp, but this improved significantly when the sequence length was increased to 500 bp. Our model has broad application potential: beyond identifying viral contigs in complex data types, it can also scan long sequences in fragments to detect viral sequences inserted into host genomes (e.g., prophages), thereby expanding the model's applicability.

Our proposed model still has certain limitations. VirNucPro utilizes all translationally possible amino acids, covering all potential coding sequences in the genome. However, the six-frame translation strategy itself may introduce non-authentic translation products, and some of these products may not naturally occur in the environment [27]. In addition, many microorganisms use non-standard codon tables, which may affect the accuracy of VirNucPro's identification. The results of VirNucPro may be limited by the accuracy of translation prediction. Future advancements in methods for predicting authentic translation products from short sequences are expected to further improve the identification performance of VirNucPro. Furthermore, distinguishing viral-like mobile elements from true viral sequences within hosts based on short sequences remains a significant challenge. Increasing the length of sequences to provide more evidence for identification seems to be one of the currently effective approaches. In the future, improvements may be made to the model's classification of these two similar data types based on more carefully curated datasets and more accurate annotations.

In conclusion, we present a new model, VirNucPro, for identifying short sequences located in viral CDS regions. The model uses a six-frame translation strategy to obtain the corresponding amino acid sequences of the input sequences, providing richer feature information and accurately identifying short viral sequences within diverse host species. VirNucPro outperforms models such as DeepVirFinder, DETIRE, and Virtifier, which have shown good performance in short viral sequence identification, on datasets of both 300 bp and 500 bp. Our research contributes to the rapid identification and exploration of viral information in metagenomic sequencing, improving the resolution of viral sequence identification.

Key Points

- VirNucPro employs a six-frame translation strategy to combine short nucleotide sequences with their corresponding amino acid sequences, enriching the data available for analysis.
- VirNucPro leverages large language models to extract rich features from both nucleotide and amino acid sequences, enhancing the model's ability to capture complex characteristics.
- VirNucPro addresses the limitations of feature scarcity in short sequences, significantly improving the accuracy and resolution of viral sequence identification.

Conflict of interest: None declared.

Funding

This work is supported in the Capital Health Research and Development of Special under grant (2024-1G-4421).

Data availability

The codes are available at: https://github.com/Li-Jing-1997/VirNucPro

References

1. Holmes EC, Krammer F, Goodrum FD. Virology-the next fifty years. *Cell* 2024;**187**:5128–45. https://doi.org/10.1016/j.cell.2024.07.025
2. Zhang YZ, Shi M, Holmes EC. Using metagenomics to characterize an expanding Virosphere. *Cell* 2018;**172**:1168–72. https://doi.org/10.1016/j.cell.2018.02.043
3. Tian FJ, Li J, Liu WL. *et al.* Virome in healthy pangolins reveals compatibility with multiple potentially zoonotic viruses. *Zool Res* 2022;**43**:977–88. https://doi.org/10.24272/j.issn.2095-8137.2022.246
4. Lind AL, Pollard KS. Accurate and sensitive detection of microbial eukaryotes from whole metagenome shotgun sequencing. *Microbiome* 2021;**9**:58. https://doi.org/10.1186/s40168-021-01015-y

5. Benler S, Yutin N, Antipov D. *et al.* Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome* 2021;**9**:78. https://doi.org/10.1186/s40168-021-01017-w
6. Wen Q, Yin X, Moming A. *et al.* Viral communities locked in high elevation permafrost up to 100 m in depth on the Tibetan plateau. *Sci Total Environ* 2024;**932**:172829. https://doi.org/10.1016/j.scitotenv.2024.172829
7. Wu LY, Wijesekara Y, Piedade GJ. *et al.* Benchmarking bioinformatic virus identification tools using real-world metagenomic data across biomes. *Genome Biol* 2024;**25**:97. https://doi.org/10.1186/s13059-024-03236-4
8. de Vries JJC, Brown JR, Fischer N. *et al.* Benchmark of thirteen bioinformatic pipelines for metagenomic virus diagnostics using datasets from clinical samples. *J Clin Virol* 2021;**141**:104908. https://doi.org/10.1016/j.jcv.2021.104908
9. Orłowska A, Iwan E, Smreczak M. *et al.* Evaluation of direct metagenomics and target enriched approaches for high-throughput sequencing of field rabies viruses. *J Vet Res* 2019;**63**:471–9. https://doi.org/10.2478/jvetres-2019-0067
10. Ho SFS, Wheeler NE, Millard AD. *et al.* Gauge your phage: Benchmarking of bacteriophage identification tools in metagenomic sequencing data. *Microbiome* 2023;**11**:84. https://doi.org/10.1186/s40168-023-01533-x
11. Miao Y, Sun Z, Ma C. *et al.* VirGrapher: A graph-based viral identifier for long sequences from metagenomes. *Brief Bioinform* 2024;**25**:bbae036. https://doi.org/10.1093/bib/bbae036
12. Guo J, Bolduc B, Zayed AA. *et al.* VirSorter2: A multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* 2021;**9**:37. https://doi.org/10.1186/s40168-020-00990-y
13. Kieft K, Zhou Z, Anantharaman K. VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* 2020;**8**:90. https://doi.org/10.1186/s40168-020-00867-0
14. Hegarty B, Riddell VJ, Bastien E. *et al.* Benchmarking informatics approaches for virus discovery: Caution is needed when combining in silico identification methods. *mSystems* 2024;**9**:e0110523. https://doi.org/10.1128/msystems.01105-23
15. Ren J, Song K, Deng C. *et al.* Identifying viruses from metagenomic data using deep learning. *Quant Biol* 2020;**8**:64–77. https://doi.org/10.1007/s40484-019-0187-4
16. Miao Y, Liu F, Hou T. *et al.* Virtifier: A deep learning-based identifier for viral sequences from metagenomes. *Bioinformatics* 2022;**38**:1216–22. https://doi.org/10.1093/bioinformatics/btab845
17. Miao Y, Bian J, Dong G. *et al.* DETIRE: A hybrid deep learning model for identifying viral sequences from metagenomes. *Front Microbiol* 2023;**14**:1169791. https://doi.org/10.3389/fmicb.2023.1169791
18. Hou X, He Y, Fang P. *et al.* Using artificial intelligence to document the hidden RNA virosphere. *Cell* 2024;**187**:6929–6942.e16. https://doi.org/10.1016/j.cell.2024.09.027
19. Zhou Z, Wu W, Ho H. *et al.* DNABERT-S: Pioneering species differentiation with species-aware DNA Embeddings. *arXiv* 2024; arXiv:2402.08777v3.
20. Lin Z, Akin H, Rao R. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30. https://doi.org/10.1126/science.ade2574
21. Wang RH, Ng YK, Zhang X. *et al.* Coding genomes with gapped pattern graph convolutional network. *Bioinformatics* 2024;**40**:btae188. https://doi.org/10.1093/bioinformatics/btae188
22. Jiang Y, Wang Y, Che L. *et al.* GutMetaNet: An integrated database for exploring horizontal gene transfer and functional redundancy in the human gut microbiome. *Nucleic Acids Res* 2024;**53**:D772–82. https://doi.org/10.1093/nar/gkae1007
23. Sargen MR, Helaine S. A prophage competition element protects salmonella from lysis. *Cell Host Microbe* 2024;**32**:2063–2079.e8. https://doi.org/10.1016/j.chom.2024.10.012
24. Peng C, Shang J, Guan J. *et al.* ViraLM: Empowering virus discovery through the genome foundation model. *Bioinformatics* 2024;**40**:btae704. https://doi.org/10.1093/bioinformatics/btae704
25. Tian F, Li J, Li F. *et al.* Characteristics and genome analysis of a novel bacteriophage IME1323_01, the first temperate bacteriophage induced from Staphylococcus caprae. *Virus Res* 2021;**305**:198569. https://doi.org/10.1016/j.virusres.2021.198569
26. Tian F, Li J, Li L. *et al.* Molecular dissection of the first Staphylococcus cohnii temperate phage IME1354_01. *Virus Res* 2022;**318**:198812. https://doi.org/10.1016/j.virusres.2022.198812
27. Omasits U, Varadarajan AR, Schmid M. *et al.* An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics. *Genome Res* 2017;**27**:2083–95. https://doi.org/10.1101/gr.218255.116

Briefings in Bioinformatics, 2025, **26(3)**, bbaf224
https://doi.org/10.1093/bib/bbaf224
Problem Solving Protocol