RESEARCH

Open Access

DNASimCLR: a contrastive learning-based deep learning approach for gene sequence data classification

Minghao Yang[1,2], Zehua Wang[2], Zizhuo Yan[2], Wenxiang Wang[2], Qian Zhu[1] and Changlong Jin[1*]

*Correspondence: cljin@sdu.edu.cn

[1] Shandong University, Weihai, People's Republic of China
[2] Beijing Research Institute of Automation for Machinery Industry, Beijing, People's Republic of China

Abstract

Background: The rapid advancements in deep neural network models have significantly enhanced the ability to extract features from microbial sequence data, which is critical for addressing biological challenges. However, the scarcity and complexity of labeled microbial data pose substantial difficulties for supervised learning approaches. To address these issues, we propose DNASimCLR, an unsupervised framework designed for efficient gene sequence data feature extraction.

Results: DNASimCLR leverages convolutional neural networks and the SimCLR framework, based on contrastive learning, to extract intricate features from diverse microbial gene sequences. Pre-training was conducted on two classic large scale unlabelled datasets encompassing metagenomes and viral gene sequences. Subsequent classification tasks were performed by fine-tuning the pretrained model using the previously acquired model. Our experiments demonstrate that DNASimCLR is at least comparable to state-of-the-art techniques for gene sequence classification. For convolutional neural network-based approaches, DNASimCLR surpasses the latest existing methods, clearly establishing its superiority over the state-of-the-art CNN-based feature extraction techniques. Furthermore, the model exhibits superior performance across diverse tasks in analyzing biological sequence data, showcasing its robust adaptability.

Conclusions: DNASimCLR represents a robust and database-agnostic solution for gene sequence classification. Its versatility allows it to perform well in scenarios involving novel or previously unseen gene sequences, making it a valuable tool for diverse applications in genomics.

Keywords: Biological sequence data, Representation learning, Contrastive learning, SimCLR, Convolutional neural networks, Sequence classification

Background

Gene sequence classification

The rapid advancement in high-throughput sequencing technologies has revolutionized the study of microorganisms, shifting away from reliance solely on cultured cells or virus strains to direct sampling from unknown environmental sources [4]. In the realm of medical disease research, the significance of microorganisms in numerous diseases is evident [32]. However, processing genetic data from microorganisms collected within the human body presents challenges due to the presence of unknown components resulting from direct environmental sampling. The first thing we need to do is to make a judgment on the source of the samples [15]. Consequently, the classification of short gene sequences becomes a basic task [29]. Furthermore, in infectious disease virus research, swift identification of pathogen types holds paramount importance for subsequent treatments [33]. Therefore, the classification of microbial gene sequence emerges as a pivotal field of study.

Traditionally, microbial gene sequencing classification relied on a homology-based approach—searching for similar DNA/RNA sequences within databases. Methods such as BLAST [2], BLAT [20], BLASTX [2], Diamond [7], BWA [23], BOWTIE [22], and others have demonstrated high accuracy. However, considerable limitations arise as numerous gene sequences cannot be classified due to poor matches with all gene types in the database. This often stems from missing data in genomic databases that is, the genetic sequences of many of these species are missing. Consequently, homology-based approaches often ineffective when dealing with new species. Additionally, the slow data processing speed of homology-based methods severely restricts their utility [27].

Recently, diverse machine learning-based approaches, including deep learning, have emerged to address these challenges. Unlike traditional methods relying on existing databases, machine learning techniques learn mathematical functions by training on available databases to accomplish predictive tasks. Meanwhile Deep learning also holds significant research value in the representation learning of microbial gene sequence data [14]. The exploration of deep learning, is gaining momentum in handling microbial gene sequence data [12]. Antonino Fiannaca et al., 2018 [3] proposed a 16S short-read sequence classification technique based on k-mer representation and deep learning architecture, which accordingly generated a model of each taxonomic unit, validated it as an effective method for bacterial sequence classification, and could be integrated into commonly used metagenomic analysis tools to successfully classify SG and AMP data. Mateo Roja-Carulla et al., 2019 [25] proposed that GeNet is a method for Shotgun metagenomic classification from original DNA sequences, using hierarchical structures between tags for training. It shows competitive accuracy and good recall rates, and requires fewer memory resources. The representation of GeNet learning is practical for biological tasks, enabling pathogen detection accuracy of more than 90%. Qiaoxing Liang et al., 2020 [24] proposed DeepMicrobes, a deep learning-based framework that overcomes the limitations of new species taxonomic in metagenome studies, has superior species and genus identification accuracy, and has demonstrated competitiveness in abundance estimation, helping to explore the role of unknown metagenome species. Meryem Altin Karagoz et al. 2021 [1] proposed a deep learning method based on k-mer representation, which combined with relative abundance index (RAI) to classify metagenomic fragments, showing that metagenomic data generated under different sequencing platforms is competitive. For the first time, the RAI score is used as a spectral representation in a deep learning algorithm, showing improved performance for data sets with multiple parameter ranges. In the field of natural and natural processing models, Florian Mock et al. 2022 [13] proposed BERTax, a neural network using natural language processing, precisely classifies DNA sequence superkingdoms and phyla without relying on representative relatives in databases. It matches or exceeds existing methods of species classification, especially when dealing with new species. Combining BERTax with databases further improves prediction quality, expanding accurate classification across diverse genomic sequences and enhancing overall information acquisition.

In addition to metagenomic applications, deep learning models have also been applied to the field of virus sequences. Tampuu A, et al. at 2019 [30] introduce ViraMiner, a novel deep learning method, to identify diverse viruses in human biospecimens, overcoming the challenge of detecting unknown or highly divergent viruses. Using Convolutional Neural Networks on raw metagenomic contigs from 19 experiments, ViraMiner significantly outperforms other machine learning methods, achieving a high accuracy of 0.923 area under the ROC curve with 300 bp contigs. It is the first model capable of detecting viral sequences within raw metagenomic data, providing insights into "unknown" sequences and enhancing our understanding of infectious diseases. Jie Ren, et al. at 2020 [28] introduce DeepVirFinder, a reference free machine learning method, excels in identifying viral sequences in metagenomic data, surpassing traditional methods. Trained on extensive pre-2015 data and enriched with additional viral sequences, it outperforms VirFinder. In colorectal carcinoma patient samples, it detected 51,138 viral sequences within 175 bins, showing potential for non-invasive CRC diagnosis. Jakub M. Bartoszewicz et al. 2021 [5] uses deep neural networks to reliably predict whether a virus can directly infect humans and has developed interpretative tools and novel nucleotide resolution correlates graph methods that can be used to detect regions of interest in novel pathogens, such as SARS-CoV-2 coronavirus. In addition, in the field of proteins, Wang Liu-Wei et al. proposed DeepViral in 2021 [31], a deep learning method for predicting protein–protein interactions (PPI) between humans and viruses. However, these methods typically rely on labeled data for model training, which becomes challenging due to the scarcity of microbial data labels, leading to complexities in feature extraction. Additionally, achieving a model with broad applicability proves to be difficult.

Contrastive learning

Currently, contrastive learning stands out as a promising direction in the field of machine learning, particularly in the realm of unsupervised feature extraction. The fundamental concept of contrastive learning involves training the network's feature extraction capability by contrasting similar and dissimilar data points in the feature space (Fig. 1). The vector representations of similar data obtained through the encoder are as close as possible, while the vector representations of dissimilar data are as distinct as possible. This approach has proven its efficacy in various domains, including computer vision, signal processing, and natural language processing, delivering promising performance [21]. Several noteworthy studies have emerged in the field of contrastive learning, such as SimCLR v1/v2 [8, 9], MoCo v1/v2/v3 [10, 11, 18], and BYOL [16], achieving state-of-the-art performance across multiple domains. SimCLR, MoCo, and BYOL represent three significant methods for unsupervised feature extraction in computational technology. SimCLR emphasizes data augmentation and contrastive loss to learn more useful feature representations through contrastive learning. MoCo employs momentum contrast to learn from unlabelled data, utilizing momentum updates to construct a contrast set. BYOL is a self-supervised learning approach encouraging the network to predict

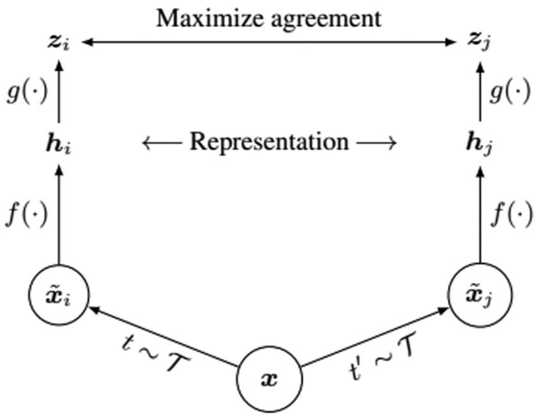


Fig. 1 A concise statement for contrastive learning [8]. Two independent data enhancement operations (t ∼ T and t′ ∼ T) are applied to the same input data, resulting in two associated data representations. An gated embedding vectors encoder network f(·) and a feedforward neural network g(·) are trained to maximize agreement using a contrastive loss. After the pre-training, we throw away the feedforward neural network g(·) and use the encoder network f(·) to complete the follow-up work

its augmented versions for learning visual feature representations. These methods train models to distinguish between different data points from a large pool of unlabelled data to derive the final feature extraction model, significantly enriching the training methods of unsupervised learning and enabling the application of various complex neural network models to large-scale unlabelled data. Given its principle of contrasting different data, this method can learn rich and distinct representations, showcasing broad prospects for the application of contrastive learning to various types of data [6].

In summary, current research in contrastive learning demonstrates the effectiveness of training feature extraction networks based on contrasting different data. Many contrastive learning models have achieved excellent results in their respective domains [26], proving their ability to efficiently derive a powerful feature extraction model from unlabelled data. However, despite its success in other fields of machine learning, including computer vision and natural language processing [17], contrastive learning has not been widely applied in microbiome bioinformatics research. While it holds immense potential, as demonstrated in various domains, its adoption remains relatively limited in the context of microbial genomics and metagenomics analysis. Most studies in microbiome bioinformatics primarily focus on traditional supervised and unsupervised learning techniques, leaving untapped potential for contrastive learning to advance microbiome bioinformatics research.

Our research contributions

To address the aforementioned challenges, this paper introduces the DNASimCLR framework, a deep learning method based on contrastive learning for the feature extraction of microbial sequence data. Unlike other approaches, we leverage unlabelled data for pre-training to enhance feature extraction. Our methodology involves two key steps: initial pre-training using unlabelled gene sequence data, followed by fine-tuning the resulting network for classification during the training phase.

In terms of data processing, we employ one-hot coding to represent DNA sequences. Based on SimCLR framework, with convolutional neural network serving as the feature extraction module. To assess the performance of our classification method, we conducted tests on a microbial gene database from various sources. Applying our method, we performed taxonomic classification and short-sequence virus host prediction on read sequences of varying lengths (250 bp, 500 bp, 1000 bp, 1300 bp, and 10,000 bp), achieving a remarkable classification accuracy of 99%. Our contributions include:

(1) Pioneering the application of contrastive learning to the feature extraction of microbial gene sequences, along with the development of a data processing method that extends contrastive learning to genetic data, overcoming limitations observed in the original SimCLR approach designed for image data.
(2) Establishing a high-performance gene sequence classifier, substantially enhancing the effectiveness of existing deep learning methods.
(3) The division of our method into pre-training and classification phases facilitates easy adaptation to other genomics problems, such as gene function and metagenomic clustering. This adaptability underscores the versatility and broad applicability of the proposed DNASimCLR framework in advancing genomics research.

Methods

The workflow of DNASimCLR consists of two primary stages: Pretraining Phase of Contrastive Learning and Fine-tuning Phase of Classification Networks (Fig. 2). In pre-training stage, we transform the unlabeled original DNA gene sequence data into a machine-learning-compatible format using the One-Hot encoding method. The

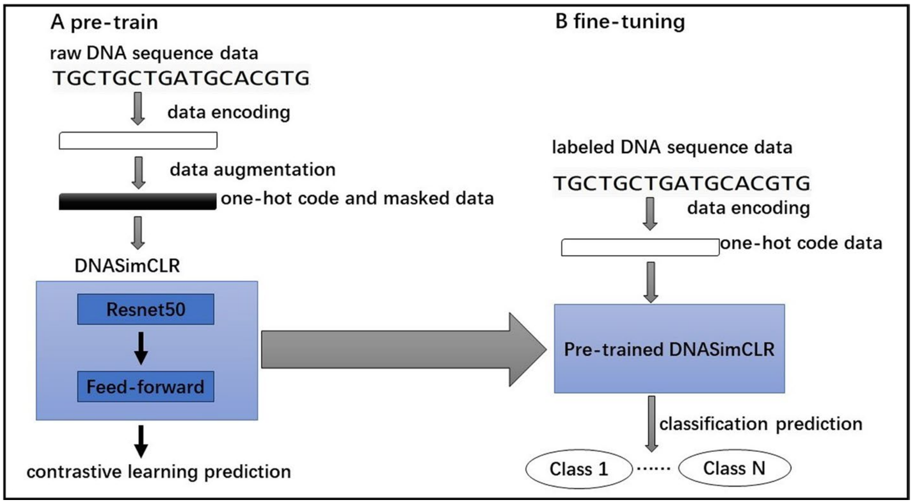


Fig. 2 Overview of the DNASimCLR framework. (A) Tokenization Data And Pretraining Model: The DNA sequences undergo preprocessing via one-hot encoding, converting sequences into a digital feature matrix, followed by a random mask application for data augmentation. Subsequently, the SimCLR model is pretrained using the masked data to derive pretrained model. (B) Training Classification Network: The feature extraction model acquired from the previous steps is employed for classification training. Ultimately, this process yields a classification network capable of determining the categorical classification of DNA sequences

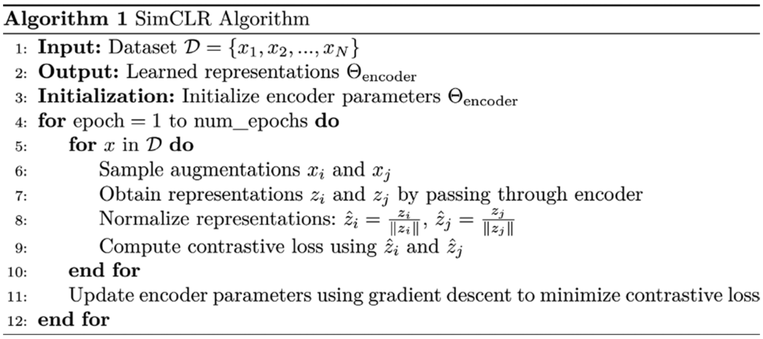
One-Hot encoded data undergoes random masking to generate the training dataset during the pre-training phase. In this stage, we employ the SimCLR framework model to obtain vector representations of unlabeled sequences. This process embeds the gene sequences into a fixed-dimensional high-dimensional space through contrastive learning. In fine-tuning stage, with the feature extraction model obtained from the pre-training phase, annotated data is encoded using the One-Hot encoding method without masking operations. We proceed with the training for classification prediction, aiming to derive a classification network equipped with a classification function.

Pretraining phase of contrastive learning

We chose SimCLR as the contrastive learning method for this study, our approach embraces the core principles of the SimCLR model while making adjustments to its implementation and data augmentation methods tailored for DNA data. The fundamental concept behind the SimCLR (Contrastive Learning for Unsupervised Visual Representation) framework is to train the feature vectors of similar samples to be as close as possible and those of dissimilar samples to be as distant as possible through comparative learning (Fig. 3). This approach facilitates the extraction of more effective feature representations. Through contrastive learning applied to unlabelled data, SimCLR generates high-quality feature vectors that densely represent the data space. These vectors prove valuable for various visual tasks.

In the processing of each DNA sequence, we segment it into fixed lengths and perform one-hot encoding. To further enhance the data for training after employing the SimCLR model, we adopt a data augmentation by masking the encoding augmentation by masking the encoding with a probability of p. During each batch reading, 30% of the bases in the input sequence are randomly masked, effectively substituting "0, 0, 0, 0" in place of the original one-hot encoding (Fig. 4). In this study, the masked encoding sequence j of the original sequence i is treated as a related representation of the same sample and is selected as a positive sample pair for SimCLR (Fig. 5), this meticulous data processing strategy is employed to prepare the data adequately for the pretraining phase, aligning with the unique characteristics of DNA data.

After completing the one-hot encoding and random masking of the original data, A neural network encoder generates a representation vector from enhanced data example. The neural network encoder in the framework is replaceable. A smaller network



```
Algorithm 1 SimCLR Algorithm
Input: Dataset $\mathcal{D} = \{x_1, x_2, ..., x_N\}$
Output: Learned representations $\Theta_{\text{encoder}}$
Initialization: Initialize encoder parameters $\Theta_{\text{encoder}}$
for epoch = 1 to num_epochs do
    for $x$ in $\mathcal{D}$ do
        Sample augmentations $x_i$ and $x_j$
        Obtain representations $z_i$ and $z_j$ by passing through encoder
        Normalize representations: $\hat{z}_i = \frac{z_i}{\|z_i\|}$, $\hat{z}_j = \frac{z_j}{\|z_j\|}$
        Compute contrastive loss using $\hat{z}_i$ and $\hat{z}_j$
    end for
    Update encoder parameters using gradient descent to minimize contrastive loss
end for
```

Fig. 3 Pseudocode for SimCLR algorithm

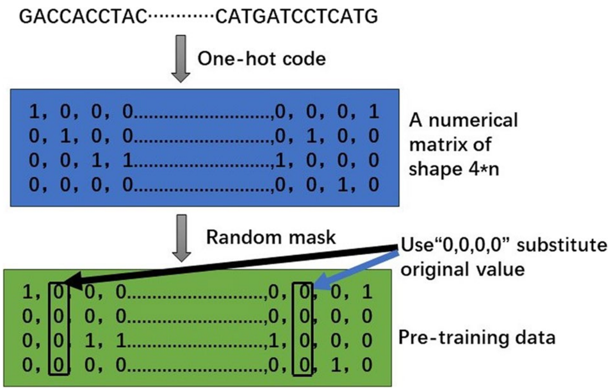


Fig. 4 Data processing flow. The original gene sequence data is one-hot encoded and converted into the image form that can be processed by convolutional neural network (“1,0,0,0” is used to express base G, “0,1,0,0” to express base A, “0,0,1,0” to express base C, and “0,0,0,1” to express base T.). Then 30% of them are randomly selected and masked with “0, 0, 0, 0”

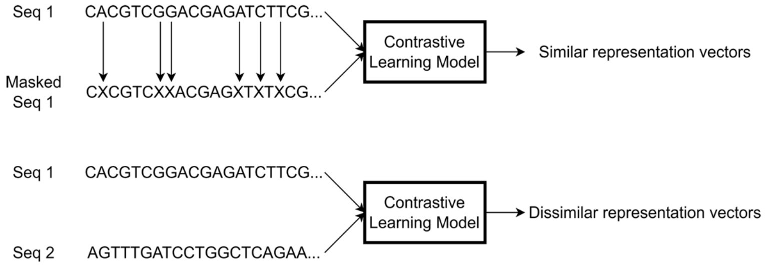


Fig. 5 The Contrastive Learning Framework in This Study. The core concept is to train a feature extraction model where feature vectors obtained from masked original DNA sequences through the contrastive learning model should be maximally similar; meanwhile, feature vectors from different DNA sequences through the same contrastive learning model should be maximally distinct

projection head maps the representation vector into the hidden space of contrastive learn. The loss function of the contrastive learning task directly impacts the feature extraction capability of the contrastive learning model. Assuming that the view x_i and the view x_j are a positive pair in the given set, the contrast prediction task is to find xj when given x_i. We take a batch at random with N samples, and each sample will produce 2 views, so we get 2N views. We don’t explicitly take negative samples directly. When given a positive pair, then you can form a negative pair with 2(N-1) samples. Then the loss function for a positive pair of examples (i, j) is defined as

$$\ell_{i,j} = -\log \frac{\exp\left(\sim\left(z_i, z_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{P}_{[k\neq i]} \exp\left(\sim\left(z_i, z_k\right)/\tau\right)}$$

where $\mathbb{P}_{[\not k=i]}$ is an indicator function evaluating to 1 if $[\not k=i]$ and ff denotes a temperature parameter. The final loss is computed across all positive pairs, both (i, j) and (j, i), in a mini-batch.

The architecture and pooling operation of Convolutional Neural Networks (CNNs) are pivotal in extracting features from input data, demonstrating proven efficacy in various computer vision domains, including image and video processing. In our DNASimCLR framework, we adopt the ResNet-50 architecture as the encoder. ResNet-50, introduced by Microsoft Research in 2016 [19], represents a significant leap in deep CNN structures. As part of the ResNet (Residual Network) family, this model addresses challenges encountered in traditional CNN training, such as gradient disappearance or explosion. Distinguished by its depth and large-scale training on the ImageNet dataset, ResNet-50 has showcased outstanding performance in subsequent evaluations, reshaping industry perceptions of CNN networks. A key innovation of ResNet-50 lies in its utilization of residual connections to mitigate the issue of gradient disappearance during network training. This approach facilitates information propagation across network layers, substantially alleviating the impact of gradient disappearance and enabling effective training of deep and intricate neural networks. Despite its considerable depth, the model's stability during training is enhanced by the inclusion of residual connections, resulting in superior performance in subsequent tasks. The ResNet-50 architecture has exerted a significant influence on the evolution of deep learning in computer vision, finding widespread application in image classification, object recognition, and video processing. Its success underscores the importance of thoughtful architectural design in addressing training challenges, contributing to broader advancements in the field.

We conducted a pre-training phase for our model, spanning 100 epochs on all unlabeled data with a batch size of 64. To effectively manage GPU running memory, we implemented a cumulative gradient strategy. Specifically, after a certain number of pre-training steps, we calculated and updated the network parameters based on the accumulated gradient. Subsequently, we cleared the gradient to proceed to the next loop (with the default initial cumulative gradient step number set at 36). For optimization, we employed the Adam optimizer with a dynamic learning rate following a trapezoidal cycle. The initial learning rate was set at 5e-2. To prevent overfitting and conserve computing resources as the model approached convergence, we implemented an early stop strategy with 10 ~ 50 epochs. Our model was trained using single-precision floating-point operations on a machine equipped with a single NVIDIA 3070Ti GPU. After pre-training on a large unlabeled corpus, we successfully obtained a gene sequences feature extraction model poised for downstream tasks.

Fine-tuning phase of classification networks

During the model's pre-training phase, masked data is employed for training. However, in the subsequent fine-tuning training phase, we abstain from using data masks. Instead, we leverage the SimCLR model previously trained with unlabeled data and integrate a simple fully connected classification neural network for fine-tuning the classification network. The architecture of the classification network comprises multiple layers of deep neural networks (DNNs). Each layer includes a fully connected layer, a ReLU activation function layer, a batch normalization layer, and a dropout layer. The dataset is then split into training and test sets. Following the fine-tuning training, the final classification network is obtained, ready for deployment in down-stream tasks.

Results and discussion

Datasets and evaluation criteria

To evaluate the effectiveness of our approach, we employed benchmark datasets obtained from the NCBI database, accessible at https://www.ncbi.nlm.nih.gov/. The first dataset, proposed by [1], is a low-complexity metagenomic dataset explicitly designed for classification. Its primary objective is the identification of relevant pathogens within samples. Authentic sequencing data was obtained from the NCBI database, representing bacterial genera such as Bacteroides, Klebsiella, Yersinia, Mycobacterium, Clostridium, and Escherichia. The corresponding NCBI SRA accession numbers for these genera are ERR1898312, ERR1474981, SRR5117441, SRR5277601, and SRR5344355, respectively. This dataset is referred to as the lr-WGS dataset. To enhance the complexity of problem-solving, we conducted tests on short reads of varying lengths, specifically 500, 1000, and 10,000 base pairs. These tests aim to assess the adaptability and performance of our approach across different data complexities and lengths.

The second dataset, denoted as the sr-16S dataset, employed in this study encompasses 16S rRNA gene regions derived from metagenome data simulations as introduced by Fiannaca et al. [3]. These simulations involve short-read sequencing data depicting full-length, high-quality 16S rRNA gene sequences sourced from the RDP database. The dataset is represented as short sequences of 250 bp and 1300 bp in length. Within this dataset, there are 100 genera within the Proteobacteria phylum, and each genus is represented by 10 species. This dataset is designed to provide a diverse and comprehensive set of sequences for evaluating the performance and adaptability of our approach across different conditions. The last dataset, referred to as the virus-host dataset and proposed by [5], serves the purpose of virus-host prediction. It comprises meticulously curated virus genomes and annotations, offering a rich source of information covering various virus species. The database encompasses both DNA and RNA viruses, with RNA sequences encoded as DNA. The dataset is represented as short sequences of 250 bp in length. Table 1 provides an overview of the dataset's attributes.

In this study, accuracy indicators were utilized to evaluate the performance of our deep learning model in the classification of microbial gene sequences. The comprehensive nature of this virus-host dataset allows for a thorough assessment of the model's ability to predict virus-host interactions across diverse genomic sequences.

The Performance of DNASimCLR on benchmark datasets

In this article, we evaluated our method, DNASimCLR, using a baseline dataset. The evaluation involved a comparative analysis with methods highlighted in the most recent CNN-based metagenomic reading classification tool [1]. Following the experimental framework outlined earlier, we initiated the process by pre-training the model on our

Table 1 Properties of the Datasets

Dataset	Read length	Number of class	Number of samples
lr-WGS	500 bp	6	4,725
	1000 bp	6	4,161
	10,000 bp	6	418
sr-16S	250 bp	100	28,224
	1300 bp	100	1,000
virus-host	250 bp	2	10,774

baseline dataset. Upon obtaining the pre-trained model, we labeled the baseline dataset and proceeded with the classification training of DNA fragments using the same network. The comparative performance between our method and the metagenomic reading classification tool [1] on lr-WGS and sr-16S datasets is summarized in Tables 2 and 3 (Fig. 6).

Additionally, applying a similar approach, we established a classification network on the virus-host dataset, achieving notably high prediction accuracy that surpasses the classification accuracy reported in the original paper of the latest CNN-based viral gene sequence reading host prediction tool [5]. The comparison between our method and the method mentioned in the paper [5] and paper [34] on the virus-host dataset, as illustrated in Table 4, underscores the superior performance of our model. These results highlight the effectiveness and robustness of the DNASimCLR framework in diverse genomics classification tasks.

Furthermore, to evaluate the feature extraction capabilities of our model, DNASimCLR, we performed feature extraction on the lr-WGS dataset after pretraining. Initially,

Table 2 DNASimCLR, CNN-RAI and RAIphy accuracy (%) results for lr-WGS Metagenomic taxonomic classification datasets

Method	**500bp_Acc**	**1000bp_Acc**	**10000bp_Acc**
CNN-RAI [1]	80.84	87.38	90.64
RAIphy [3]	67.44	75.96	95.00
DNASimCLR	**84.42**	**89.39**	**95.14**

Table 3 DNASimCLR, CNN-RAI and RAIphy accuracy (%) results for sr-16S Metagenomic taxonomic classification datasets

Method	**250bp_Acc**	**1300bp_Acc**
CNN-RAI [1]	76.47	98.44
RAIphy [3]	83.00	97.00
DNASimCLR	**94.00**	**99.99**

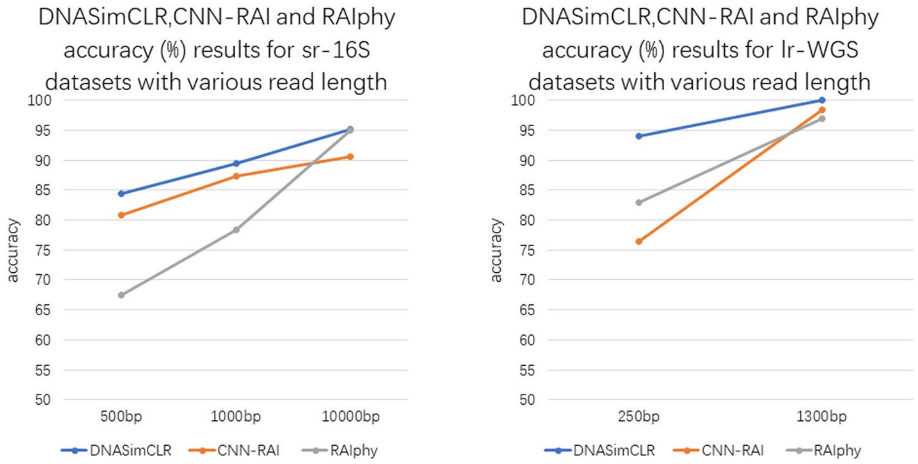


Fig. 6 Model accuracy results for sr-WGS and lr-WGS data with various read lengths

Table 4 Comparison of DNASimCLR with other tools in virus-host dataset (Acc. (Accuracy): The overall correctness of predictions. Prec. (Precision): The accuracy of positive predictions. Rec. (Recall): The proportion of actual positives correctly predicted. Spec. (Specificity): The proportion of actual negatives correctly predicted.)

Method	Acc	Prec	Rec	Spec
CNN [5]	87.8	89.9	85.2	90.4
LSTM [5]	84.7	86.0	82.8	86.5
k-NN [34]	75.5	76.3	73.9	77.1
BLAST [34]	78.4	98.3	79.2	77.6
VirusBERTHP [34]	99.1	98.8	99.3	98.8
DNASimCLR	**99.9**	**99.8**	**99.7**	**99.8**

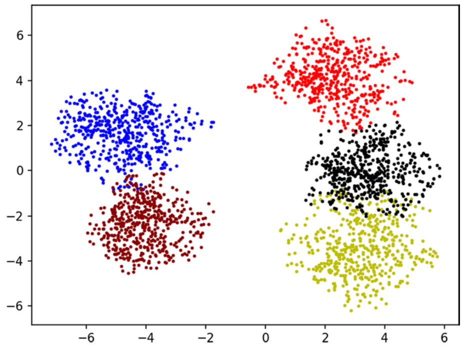


Fig. 7 Effect of linear discriminant analysis dimensionality reduction on lr-WGS data set.We performed a random selection of 500 sample points from each genus present in the dataset. Subsequently, our pre-trained model was applied to generate 128-dimensional feature vectors for each selected sample. Following this, we labeled individual vectors and reduced their dimensions to two using Linear Discriminant Analysis (LDA) for visualization purposes. The resulting visual representation indicates the feature extraction capability of the DNASimCLR model post pre-training

we obtained the eigenvector of the original data, and then employed linear discriminant analysis (LDA) to reduce the dimensionality to 2 dimensions for visualization purposes. As depicted in Fig. 7, it is evident from the visualization that our model demonstrates a remarkable capacity for feature extraction following pre-training. This visualization provides insights into the distribution and separability of features extracted by our model, validating its effectiveness in capturing meaningful representations from the input data.

Conclusions

Even the most comprehensive microbial gene databases currently available exhibit challenges with missing data and labels, significantly limiting the effectiveness of numerous supervised deep learning methods. Addressing this incompleteness is an urgent challenge that demands immediate attention. In this paper, we propose a neural network feature extraction method based on contrastive learning to address the issue of representation learning for microbial gene sequence data. The method involves two key steps: first, pre-training on unlabelled gene sequence data, and then using labelled data for fine-tuning to obtain classification networks for downstream tasks. To process the data, one-hot coding is employed to encode the DNA sequence, and the SimCLR framework is utilized to complete the pre-training model, with RESNET50 selected as the feature extraction module. In terms of performance evaluation, the study tests genomic databases from different sources. For metagenomic segment classification and virus host prediction, the proposed method demonstrates significant advantages over NN-based models on short sequence data, achieving significantly improved accuracy. The contributions of this study are multifaceted. Firstly, contrastive learning is applied to the representation learning of microbial gene sequence data for the first time. A novel data processing method for gene data is developed, overcoming the limitation that the SimCLR method is traditionally applicable only to image data. This expansion broadens the application field of contrastive learning. Secondly, the microbial gene sequence data classifier proposed in this study exhibits a substantial improvement in performance, opening new opportunities for the development of convolutional neural network methods in processing biological data. Additionally, due to the separation of the pre-training stage and the classification stage, the method can be easily applied to other genomics problems, such as protein function prediction and new virus detection. In conclusion, DNASimCLR represents an advanced exploration of microbial gene sequence feature extraction utilizing a self-supervised learning model. This approach holds the potential to introduce innovative concepts in the field of bioinformatics, providing a pathway to derive biological sequence features through convolutional neural networks.

Author contributions
All authors conceptualized the study and reviewed the manuscript. Minghao Yang developed the methodology, conducted data analysis, created visualizations, Took the lead in writing the original draft, review and editing of the manuscript. Changlong Jin optimized the methodology and participated in writing the original draft. Zehua Wang, Zizhuo Yan and Wenxiang Wang reviewed and edited the draft. Changlong Jin and Qian Zhu administrated the project and supervised.

Funding
Not applicable.

Availability of data and materials
Sequence data for contrastive learning are publicly accessible at NCBI Sequence Read Archive (SRA) with the primary accession codes ERR1898312, ERR1474981, SRR5117441, SRR5277601, SRR5344355 at https://www.ncbi.nlm.nih.gov/, the sr-16S dataset can be accessed at https://github.com/IcarPA-TBlab/MetagenomicDC/tree/master/data and the virus-host dataset is available at https://www.genome.jp/virushostdb/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 29 February 2024 Accepted: 9 October 2024
Published online: 14 October 2024

References

1. Altın KM, Nalbantogl OU. Taxonomic classification of metage-nomic sequences from relative abundance index profiles using deep learning. Biomed Signal Process Control. 2021;67:102539.
2. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
3. Antonino F, Laura LP, Massimo LR, Giosuè LB, Giovanni R, Riccardo R, Salvatore G, Alfonso U. Deep learning models for bacteria taxonomic classification of metagenomic data. BMC Bioinform. 2018;19(Suppl 7):61–76.
4. Baird SJE. The impact of high-throughput sequencing technology on speciation research: maintaining perspective. J Evolut Biol. 2017;30(8):14820–70.

5. Bartoszewicz JM, Seidel A, Renard BY. Interpretable detection of novel human viruses from genome sequencing data. NAR Genom Bioinform. 2021;3(1):lqab004.
6. Benjamin E., Tianjun Z., Ruslan S., Sergey L. Contrastive Learning as Goal-Conditioned Reinforcement Learning. Conference on Neural Information Processing Systems (2022)
7. Byrd AL, Perez-Rogers JF, Manimaran S, Castro-Nallar E, Toma I, McCaffrey T, Siegel M, Benson G, Crandall KA, Johnson WE. Clinical pathoscope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. BMC Bioinform. 2014;15:262.
8. Chen T., Simon K., Mohammad N., Geoffrey H. A simple framework for contrastive learning of visual representations. International Conference on Machine Learning. PMLR. 2020.
9. Chen T, Simon K, Kevin S, Mohammad N, Geoffrey H. Big self-supervised models are strong semi-supervised learners. Conf Neural Inform Process Syst. 2020;33:22243–55.
10. Chen X, Fan H, Girshick R, He K. Improved baselines with momentum contrastive learning. Learning. 2020;9:04297.
11. Chen X, Xie S, He K. An empirical study of training self-supervised vision transformers. IEEE Int Conf Comput Vis. 2021;57:9620–9.
12. Florian M, Adrian V, Emanuel B, Manja M. Vidhop, viral host prediction with deep learning. Bioinformatics. 2021;37(3):318–25.
13. Florian M., Fleming K., Anton K., Sebastian B., Manja M.: BERTax: Taxonomic Classification of DNA Sequences with Deep Neural Networks (2021)
14. Gökcen E, Žiga A, Julien G, Fabian JT. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20(7):389–403.
15. Gargi C, Sangeeta N, Supratim B, Joel F, Anthonia O, Pratyoosh S, et al. Microbiome systems biology advancements for natural well-being. Sci Total Environ. 2022;838(Pt 2):155915.
16. Grill J, Strub F, Altché F, Tallec C, Richemond PH, Buchatskaya E, et al. Bootstrap your own latent - a new approach to self-supervised learning. Conf Neural Inform Process Syst. 2020;33:21271–84.
17. Haifeng L, Jun C, Jiawei Z, Qinyao L, Silu H, Xuyin W. Augmentation-free graph contrastive learning of invariant-discriminative representations. IEEE Trans Neural Netw Learn Syst. 2023;4:1–11.
18. Kaiming H, Haoqi F, Yuxin W, Saining X, Ross G. Momentum contrast for unsupervised visual representation learning. Comput Vis Patt Recogn. 2020;2020(1):9726–35.
19. Kaiming H, Xiangyu Z, Shaoqing R, Jian S. Deep residual learning for image recognition. Proc IEEE Comput Soc Conf Comput Vis Patt Recogn. 2016;03385(1):770–8.
20. Kent WJ. Blat-the blast-like alignment tool. Genome Res. 2002;12:656–64.
21. Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, Isola P, et al. Supervised contrastive learning. Conf Neural Inform Process Syst. 2020;11362:18661–73.
22. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. Genome Biol. 2009;10:R25.
23. Li H, Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform. Bioinformatics. 2010;26:589–95.
24. Liang Q, Wang PB, Liu Y, Zou B, Lai W. Deepmicrobes: taxo-nomic classification for metagenomics with deep learning. NAR Genom Bioinform. 2020;2(1):lqaa009.
25. Mateo R, Ilya OT, Guillermo L, Nicholas Y, Ruth L, Bernhard S. GeNet: deep representations for metagenomics. Comput Res Repos. 2019;15:537795–13.
26. Mengru C, Chao H, Lianghao X, Wei W, Yong X, Ronghua L. Heterogeneous graph contrastive learning for recommendation. WSDM. 2023;95:544–52.
27. Quince C, Walker AW, Simpson JT, et al. Shotgun metagenomics, from sampling to analysis. Nat Biotechnol. 2017;35(9):833.
28. Ren J, Song K, Deng C, Ahlgren NA, Fuhrman JA, Li Y, Xie X, Poplin R, Sun F. Identifying viruses from metagenomic data by deep learning. Quantit Biol. 2020;8:64–77.
29. Stanton RA, Vlachos N, Laufer HA. GAMMA: a tool for the rapid identification, classification and annotation of translated gene matches from sequencing data. Bioinformatics. 2022;38(2):546–8.
30. Tampuu A, Bzhalava Z, Dillner J, Vicente R. ViraMiner: deep learning on raw DNA sequences for identifying viral genomes in human samples. PLoS ONE. 2019;14:e0222271.
31. Wang L, Şenay K, Jun C, Nicholas JD, Jesper T, Robert H. Deepviral: prediction of novel virus-host interactions from protein sequences and infectious disease phenotypes. Intell Syst Mol Biol. 2021;37(17):2722–9.
32. Wei M, Lu Z, Pan Z, Chuanbo H, Jianwei L, Bin G, Jichun Y, Wei K, Xuezhong Z, Qinghua C. An analysis of human microbe-disease associations. Brief Bioinform. 2017;18(1):85–97.
33. Xiaoyuan Y, Kai M, Yuxia Z, Lihong Q, Wu A, Youling W. Establishment and Application of Rapid Diagnosis for Reverse Transcription-Quantitative PCR of Newly Emerging GooseOrigin Nephrotic Astrovirus in China. mSphere. 2018;3(6):18.
34. Yunzhan W, Jin Y, Yunpeng C. VirusBERTHP: Improved Virus Host Prediction Via Attention-based Pre-trained Model Using Viral Genomic Sequences. IEEE Int Conf Bioinform Biomed. 2023;3:678–83.

Publisher's Note