JTE HCMUTE
ISSN: 2615-9740

JOURNAL OF TECHNICAL EDUCATION SCIENCE
Ho Chi Minh City University of Technology and Education
Website: https://jte.edu.vn
Email: jte@hcmute.edu.vn

Using Deep Learning for the Taxonomic Classification of Microbial Sequences

Manh Hung Hoang[1], Vu Hoang[2], Van Vinh Le[2*]
[1]*HCMC Industry and Trade College, Vietnam*
[2]*HCMC University of Technology and Education, Vietnam*

[*]Corresponding author. *Email: vinhlv@hcmute.edu.vn*

ARTICLE INFO

Received: 15/01/2024
Revised: 01/02/2024
Accepted: 21/02/2024
Published: 28/02/2024

KEYWORDS

Taxonomic classification;
Deep learning;
DNA;
Metagenomic;
k-mer embedding.

ABSTRACT

Microbes are common creatures and play a crucial role in our world. Thus, the understanding of microbial communities brings benefits to human lives. Because the material samples of microbes contain sequences belonging to different organisms, an important task in analyzing processes is to classify the sequences into groups of different species or closely related organisms, called metagenomic classification. Many classification approaches were proposed to analyze the metagenomic data. However, due to the complexity of microbial samples, the accuracy performance of those methods still remains a challenge. This study applies an effective deep learning framework for the classification of microbial sequences. The proposed architecture combines a sequence embedding layer with other layers of a bidirectional Long Short-Term Memory, Seft-attention, and Dropout mechanisms for feature learning. Experimental results demonstrate the strength of the proposed method on datasets of real metagenomes.

Doi: https://doi.org/10.54644/jte.2024.1521

1. Introduction

Metagenomics is the field of studying microbial communities. Different from traditional approaches, genetic materials are collected directly from the environment and put into the sequencing process without culturing and separating in laboratories. With the development of sequencing technologies, biologists are currently able to analyze a huge amount of data from various environments such as soils, the human body, and ocean water [1]. The field brings an opportunity to deeply understand microbial worlds and how they affect human lives.

In a metagenomic project, data are generated from DNA samples by sequencing machines. Many tasks are applied to analyze the sequenced data. One of the most important tasks is the taxonomic classification which separates DNA sequences into groups of species or related organisms, and determines the organisms in the groups. Results of the application allow scientists to continue studying each individual species or apply DNA assembling to build single genomes for further analysis. Taxonomic classification approaches for metagenomics can be divided into two groups of homology based and composition based methods.

Homology based approaches such as MEGAN [2], [3], Qmatey [4], and MTSv [5] use the similarity information between sequences as major features for the classification process. MEGAN firstly finds matched sequences to reference databases using alignment tools, *e.g.* BLAST [6], or DIAMOND [7]. The approach then determines relationships between sequences with known organisms using a Lowest Common Ancestor (LCA) algorithm. Qmatey, and MTSv try to boost the speeds of the searching task by using other techniques such as MegaBLAST [8], or Smith-Waterman algorithm with seed matches, respectively. However, the full alignment methods are still very time consuming and hard in working with a huge amount of data in real metagenomic projects.

Dealing with the challenge of the full alignment algorithms in computational performance, some methods, *e.g.* Kraken [9], Kraken 2 [10], CLARK [11], and K2Mem [12], are based on exact *k*-mer matching techniques. Instead of doing a comparison of sequences with their whole lengths, the approaches extract *k*-mers from them and perform *k*-mers alignments. Kraken is one of the initial

approaches which applies the fast alignment technique. After sequences are compared to each other through k-mer matching, they are classified by the LCA algorithm. Kraken 2 is an improvement of Kraken in which it reduces memory storage significantly compared to its previous version. CLARK and K2Mem aim to increase speed and the accuracy of the classification process using discriminative *k*-mers. The two approaches do not store and perform all *k*-mers, they remove any common *k*-mers. The remaining ones are target-specific *k*-mers and represent genomic regions with uniquely characteristic.

Some metagenomic classification methods utilize composition features which are extracted from sequences to classify them. While NBC [13] applies the naïve Bayes classifier for its classification process, TAC-ELM [14] uses an Extreme Learning Machine technique to work with GC-content and *k*-mer frequency features. Another composition based approach, TACOA [15] separates and assigns long sequences using *k*-nearest neighbor algorithm. One of the strengths of those methods is that they are very fast. However, they often return low accuracy and are very sensitive to sequencing errors or variants in sequences.

Recent composition based approaches apply the strength of deep learning techniques to achieve better classification quality. DeepMicrobes [16] uses a deep neural network with two strategies for encoding DNA sequences including one-hot encoding and *k*-mer embedding. BERTax [17] and LP-MeTaxa [18] are other metagenomic classification approaches derived from models for natural language processing. BERTax is based on the BERT model which employs a self-attention mechanism. The method determines parts of inputs which are relevant to each other autonomously. LP-MeTaxa uses Word2vec, a powerful training technique for word embedding. In this case, it considers word2vec vectors to be similar to nucleotides concatenation. Another deep learning based algorithm, MetaTransformer [19], also uses self-attention models in its analysis tasks. Besides, the approach applies sparse embedding gradients and mixed-precision computation strategies in training tasks.

This study utilizes deep learning techniques to classify microbial sequences of metagenomic samples. The proposed method applies different mechanisms to discover meaningful features of data such as Biditional Long-Short Term Memory (BiLSTM), self-attention, and dropout technique. Furthermore, a *k*-mer embedding technique is used to consider deeper relationships between sequences belonging to the same organisms. The strengths of the approach are demonstrated on datasets generated from real metagenomes.

2. Methods

2.1. General process

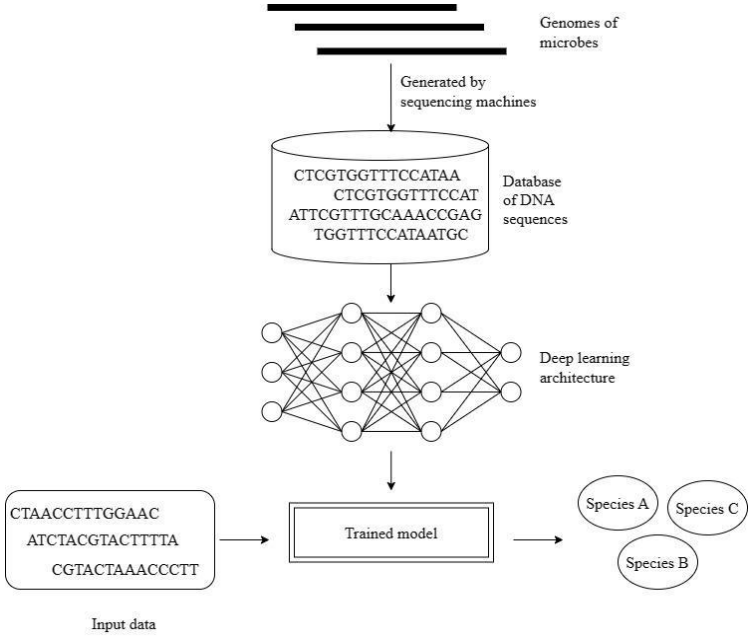

Figure 1. *Training and classifying process*

The analyzing process of the proposed approach to classify microbial sequences contains two phases of training and classifying, presented in figure 1. In order to train the model, genomes of microbes are sequenced from environmental materials. The genomes contain a series of DNA (Deoxyribonucleic Acid) data including: A (Adenine), C (Cytosine), G (Guanine), T (Thymine). After being generated by a sequencing machine, each genome is a set of DNA sequences with short length (*e.g.* from 100 bp to 400 bp with Illumina sequencing machine). The trained model is used for classifying input sequences. It labels each sequence with name or taxonomy identification of microbial species.

2.2. Model architecture

The architecture of the model used in this work (presented in figure 2) contains different layers. It firstly generates feature matrices using *k*-mer embedding technique to generate feature matrices, and then using Bidirectional LSTM, Self-attention, Dropout and other fully connected layers to learn features. The model also includes three fully connected layers which use ReLU activations as linear transformation between them.

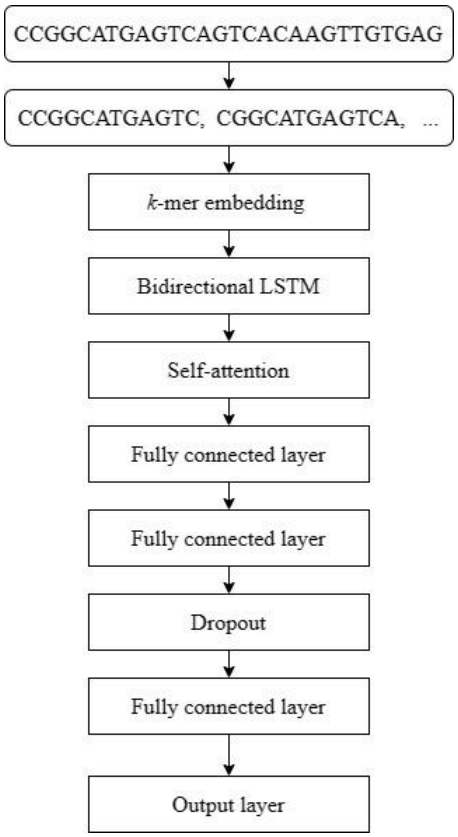

Figure 2. *Model architecture of the proposed approach*

2.2.1. k-mer embeddings

The *k*-mer embedding method aims to map DNA short sequences to real-number vectors. While traditional *k*-mer based approaches simply calculate the frequency vector of *k*-mer without considering the co-occurrence relationship of *k*-mers, the embedding technique utilizes the meaningful information as classification features. With the technique, a feature matrix is constructed to be a representation of each sequence.

Given a sequence of length *L*, using a sliding window to slide along the sequence with a step of one nucleotide, and extract subsequence with the length of *k*. In total, we have *L*-*k*+1 *k*-mers for each sequence. For instance, with a sequence of 80bp, there are 69 *k*-mers generated with a fit length of 12.

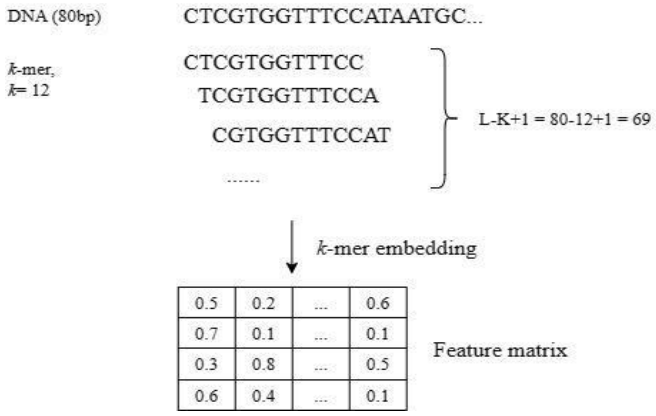

Figure 3. *k-mer embeddings technique*

In order to generate a feature matrix, each *k*-mer is indexed with a positive integer according to its lexicographical order in a *k*-mer dictionary. The learning process of embedding representations utilizes the information of *k*-mer occurrences. Each *k*-mer is represented as an embedding vector. The collection of these vectors forms the feature matrix.

2.2.2. Bidirectional Long Short-Term Memory

This study employs the Bidirectional Long Short-Term Memory (BiLSTM) layer (Figure 4). The process ensures that any neural network has sequential information in both backward (from the future to the past) and forward (from the past to the future) directions. This architecture offers several advantages in practical applications, as every component of the input sequence obtains information from both the past and the future. Therefore, BiLSTM helps generate a more meaningful output by combining LSTM layers from both directions. The BiLSTM phase executes a bidirectional LSTM network on the input to capture long-term dependencies, ultimately yielding a fixed-length feature vector as the result.

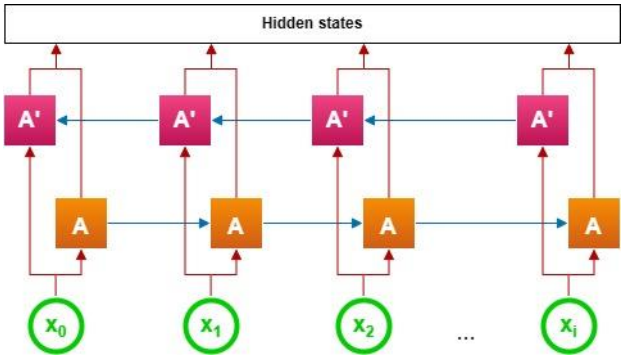

Figure 4. *Bidirectional LSTM*

2.2.3. Self-Attention and Dropout

In this model, a self-attention mechanism is applied to capture contextual relationships within sequences of data. The technique allows each element in a sequence to weigh the importance of every other element. It enables the model to focus on different parts of the input when making predictions. The attention layer proves effective in handling long-range dependencies.

In addition to improving quality performance, the proposed approach utilizes a dropout technique. With the technique, neurons are randomly selected and omitted during training, including both the forward and backward propagation processes. When the neurons are excluded randomly, their contribution to activation of other neurons is temporarily reduced, and any weight updates do not apply to these omitted neurons. This layer aims to mitigate the risk of overfitting. The dropping out of some neurons during training forces other neurons to participate more to compensate for the missing ones. As

a result, the network could learn more robust features from the neurons, and become less reliant on specific neurons.

2.3. Evaluation metrics

In order to evaluate the performance of the classification approaches, this work uses two metrics *precision* and *recall* which are defined as follows.

$$precision = \frac{No.\ sequences\ classified\ correctly}{No.\ sequences\ classified}$$

$$recall = \frac{No.\ sequences\ classified\ correctly}{No.\ sequences}$$

Besides, an additional metric named F-measure which emphasizes comprehensively on precision and recall to fully reflect the performances of a classification method.

$$F\text{-}measure = \frac{2 * precision * recall}{precision + recall}$$

3. Experiments and Results

3.1. Training the model

In order to train the proposed model, we use 8.265.856 sequences which are generated from 25 real microbial genomes downloaded from the National Center for Biotechnology Information (NCBI) database. These sequences are generated by MetaSim tool [20] following the Illumina error profile with length of 100 bp. A dataset of 918.428 sequences which are not included in the training set is used to test the model.

In this case, *k*-mer is set to a length of 12. Adam optimizer is used to control the network training where it updates and calculates the network parameters. Learning and Dropout rates are set to 0.001, and 0.1, respectively.

3.2. Results

3.2.1. Classification performance

The proposed method is compared with DeepMicrobes on the tested dataset. Results of the two methods on the test dataset are presented in table 1. It can be seen on the table that the proposed method archives better both precision and recall value. Totally, the proposed method returns higher *F-measure* than DeepMicrobes.

Table 1. *Performance of DeepMicrobes and Proposed method on test data*.

Metrics	DeepMicrobes	Proposed method
Precision	93.48%	94.63%
Recall	90.3%	92.54%
F-measure	91.86%	93.57%

3.2.2. The effect of k-mer length

In order to evaluate the effect of *k* value to the performance of the proposed approach, this experiment performs a test with a dataset of 1.112.554 sequences from 5 microbial genomes. We use 10-fold cross-validation technique to validate the method with different values of *k* from 8 to 12. The data is divided into 10 groups. The model is trained 10 times. Each training time uses 9 groups as training data and the remaining one used as validation data. Average value of the all testing results is calculated for evaluating the model.

The bar chart in figure 5 show that the proposed approach gets better accuracy when the length of *k*-mer increases. However, the increase of *k* value also requires more computing resources because the number of *k*-mers rises significantly. Therefore, the length of *k*-mers used depends on the performance of computing system.

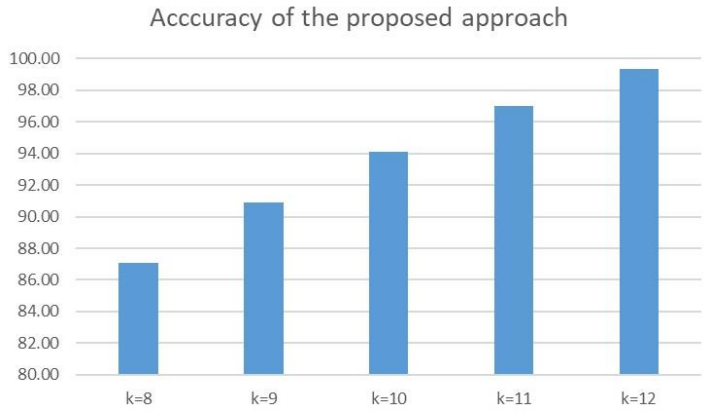

Figure 5. *Accuracy of the proposed approach with different value of k*

4. Conclusions

The complexity of microbial communities requires a deep analysis process to decipher meaningful features from genomic data. This work demonstrates the strength of deep learning techniques in accurately classifying and characterizing microbial species. With the usages of BiLSTM, Self-attention, combined with the dropout technique, the proposed model gains a better quality performance compared to an available method.

The model applied in this study is only trained and tested with the datasets containing organisms at the same species level. In future works, it should be evaluated for other levels such as Genus, Family or higher ones. Additionally, exploring various other variants of deep learning models is also a future research direction.

Acknowledgments

The authors would like to thank the Faculty of Information Technology, HCMC University of Technology and Education for providing facilities for this study. The applications presented in this paper were tested in the High Performance Computing lab of the faculty.

Conflict of Interest

The authors declare no conflict of interest.

REFERENCES

[1] C. Simon and R. Daniel, "Metagenomic analyses: past and future trends," *Applied and Environmental Microbiology*, vol. 77, no. 4, pp. 1153-1161, 2011.
[2] D. H. Huson, *et al.*, "MEGAN analysis of metagenomic data," *Genome Research*, vol. 17, no. 3, pp. 377-386, 2007.
[3] C. Bağcı, S. Patz, and D. H. Huson, "DIAMOND+ MEGAN: fast and easy taxonomic and functional analysis of short and long microbiome sequences," *Current Protocols*, vol. 1, no. 1, pp. e59, 2021.
[4] T. N. Furstenau *et al.*, "MTSv: rapid alignment-based taxonomic classification and high-confidence metagenomic analysis," *Peer J.*, vol. 10, no. 3, pp. e14292, 2022.
[5] A. K. Adams *et al.*, "Qmatey: an automated pipeline for fast exact matching-based alignment and strain-level taxonomic binning and profiling of metagenomes," *Briefings in Bioinformatics*, vol. 24, no. 2, pp. bbad351, 2023.
[6] T. Madden, "The BLAST sequence analysis tool," *The NCBI Handbook*, vol. 2, no. 5, pp. 425-436, 2013.
[7] B. Buchfink, C. Xie, and D. H. Huson, "Fast and sensitive protein alignment using DIAMOND," *Nature Methods*, vol. 12, no. 1, pp. 59-60, 2015.
[8] Y. Chen *et al.*, "High speed BLASTN: an accelerated MegaBLAST search tool," *Nucleic Acids Research*, vol. 43, no. 16, pp. 7762-7768, 2015.
[9] D. E. Wood and S. L. Salzberg, "Kraken: ultrafast metagenomic sequence classification using exact alignments," *Genome Biology*, vol. 15, no. 3, pp. 1-12, 2014.
[10] D. E. Wood, J. Lu, and B. Langmead, "Improved metagenomic analysis with Kraken 2," *Genome Biology*, vol. 20, pp. 1-13, 2019.
[11] R. Ounit *et al.*, "CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers," *BMC Genomics*, vol. 16, no. 1, pp. 1-13, 2015.
[12] D. Storato and M. Comin, "K2mem: discovering discriminative k-mers from sequencing data for metagenomic reads classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 1, pp. 220-229, 2021.
[13] G. L. Rosen, E. R. Reichenberger, and A. M. Rosenfeld, "NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads," *Bioinformatics*, vol. 27, no. 1, pp. 127-129, 2011.
[14] Z. Rasheed and H. Rangwala, "TAC-ELM: Metagenomic Taxonomic Classification with Extreme Learning Machines," *BICoB*, 2011.

[15] N. N. Diaz *et al.*, "TACOA–Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach," *BMC Bioinformatics*, vol. 10, pp. 1-16, 2009.
[16] Q. Liang *et al.*, "DeepMicrobes: taxonomic classification for metagenomics with deep learning," *NAR Genomics and Bioinformatics*, vol. 2, no. 1, pp. qaa009, 2020.
[17] F. Mock *et al.*, "BERTax: taxonomic classification of DNA sequences with Deep Neural Networks," *BioRxiv*, vol. 07, 2021.
[18] B. Matougui *et al.*, "NLP-MeTaxa: A Natural Language Processing Approach for Metagenomic Taxonomic Binning Based on Deep Learning," *Current Bioinformatics*, vol. 16, no. 7, pp. 992-1003, 2021.
[19] A. Wichmann *et al.*, "MetaTransformer: deep metagenomic sequencing read classification using self-attention models," *NAR Genomics and Bioinformatics*, vol. 5, no. 3, pp. lqad082, 2023.
[20] D. C. Richter *et al.*, "MetaSim - a sequencing simulator for genomics and metagenomics," *PLoS ONE*, vol. 3, no. 10, pp. e3373, 2008.

Hoang Manh Hung received the bachelor's degree in Information Technology from University of Information Technology in 2013, and MSc degree in Computer Science from Ho Chi Minh City University of Technology and Education in 2023. Currently, he is working at the faculty of Information Technology, HCMC Industry and Trade College, Vietnam.
Email: manhhung@hitu.edu.vn

Hoang Vu received the bachelor's degree in Information Technology from Ho Chi Minh City University of Technology in 2006, and MSc degree in Computer Science from Ho Chi Minh City University of Technology and Education in 2021. Currently, he is studying at the faculty of Information Technology, HCMC University of Technology and Education, Vietnam. Email: hvu267@gmail.com

Le Van Vinh received the bachelor's degree in Information Technology, and MSc degree in Computer Science from University of Science (Vietnam National University, Ho Chi Minh City) in 2005 and 2009, respectively. He received the PhD degree in Computer Science from HCMC University of Technology (Vietnam National University, Ho Chi Minh City) in 2017. Currently, he is working at the faculty of Information Technology, HCMC University of Technology and Education, Vietnam. His research interests include bioinformatics, high-performance computing, data science, and deep learning. Email: vinhlv@hcmute.edu.vn. ORCID: https://orcid.org/0000-0001-5218-3089