**DNA Alignment-free Sequence Analysis Algorithm with GC-contents and K-mers Preprocessing**

**Joon-Su Kim1, and Dong-Keun Kim2\***

1Department of Human-Centered Artificial Intelligence, Sangmyung University, Seoul 03016, Korea

2Department of Human-Centered Artificial Intelligence, Sangmyung University, Seoul, 03016, Korea

*[*Email *: nagabuti13@naver.com]*

**Abstract**

In this study, to improve the shortcomings of existing alignment-free sequence analysis and alignment-based sequence analysis, the new DNA alignment-free sequence analysis algorithm with high processing speed and gene mutation resistance compared to popular alignment-based sequence analysis programs is suggested. Now, many studies have shown a trend to increase the accuracy of biological classification by preprocessed sequence datasets with high-cost deep learning such as CNN and RNN. Alternatively, the gene dataset preprocessed with the analysis algorithm allows meaningful results to be achieved in a short time with only low-cost supervised or unsupervised learning models. In addition, several thousand or more sequences may be stably preprocessed at once with limited resources.

For the study, I obtained a COI DNA genetic dataset for each mammal, bird, reptile, fish, and amphibian from the National Center for Biotechnology Information (NCBI) in the form of a FASTA file. It was converted into a single multi-FASTA metagenome file and preprocessed with the new algorithm. A differentiated k-mers technique named ‘k-mers slicing’ was applied to reduce the deviation in the amount of information in genetic samples caused by sequencer errors or mutations. The k value of k-mers, which is an important variable in sequence analysis, was set as the optimal value in the direction of maximizing the biological classification accuracy in consideration of the genetic distribution of the samples. After that, only the fragments with the highest GC-content among the sliced gene fragments were taken for removing outliers and reducing the data size. Finally, I applied ‘k-mers embedding’ to the data and constructed the training datasets for sequence analysis. Species classification was performed with the K-means unsupervised learning model and the SVM supervised learning model.

Compared to MUSCLE, MEGA11, and ClusterW, commonly used alignment-based sequence analysis programs in preprocessing COI DNA data, the suggested algorithm showed significantly faster processing speeds. Especially, about 300 samples of the 9 species of reptile CO1 DNA dataset was processed 92 times faster than MUSCLE. In addition, the accuracy of machine learning for species classification was higher than when learned with the programs. Comparing the average accuracy for the five classes, when applying the supervised learning model, we achieved 99.12%, about 1.5% higher than before and when applying the unsupervised learning model, achieved 90.7%, about 1.2% higher than before.

Index Terms: CO1, GC-content, K-mers embedding, K-mers slicing

* Corresponding Author Information
  + Name : Dong-Keun Kim
  + Email : dkim@smu.ac.kr
  + Phone : +82-2-2287-5431
* Paper Target (Please check one.)

|  |  |  |  |
| --- | --- | --- | --- |
| SCOPUS-indexed Journal | JICCE  (Journal of Information and Communication Convergence Engineering) | JKIICE  (Journal of Korea Institute of Information and Communication Engineering) | ICFICE2023 Proceeding only |
|  |  |  | O |