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

**Joon-Su Kim1, and Dong-Keun Kim2[[1]](#footnote-2)**

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 is suggested. Many studies have used high-cost deep learning such as CNN and RNN to increase the accuracy of biological classification with sequence datasets. Instead of using that, the analysis algorithm preprocess the gene dataset, so that it allows meaningful results in a short time with low-cost supervised or unsupervised learning models. In addition, several thousand or more sequences may be stably preprocessed with limited resources. The COI DNA genetic dataset for each mammal, bird, reptile, fish, and amphibian in the form of a FASTA file was used to preprocess. The new algorithm used ‘k-mers slicing’ with GC-content and ‘k-mers embedding’ to reduce the deviation in genetic samples and those size. Compared to the commonly used alignment-based sequence analysis programs, it showed significantly faster processing speeds. In addition, the accuracy of machine learning for species classification was higher than when learned with the programs.

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

**I. INTRODUCTION**

The commercialization of Sanger sequencing, a method of DNA sequencing that analyzes DNA (or RNA) sequences by Sanger, has made the importance of DNA sequencing widely known throughout biology and contributed to the development of biology. However, Sanger sequencing has a fatal drawback that the length of sequence that could be decoded at once was short. This led to the human genome project being carried out over a long period of 13 years and caused a high cost of collecting DNA information. Since then, the NGS, which produces and analyzes a large amount of parallel data through computers at high speed, has been created, enabling biological classification using computers.

Methods of processing to extract features from DNA information in biological classification can be sorted alignment-based sequence analysis and alignment-free sequence analysis. Alignment-based sequence analysis is a method of sorting the positions of bases on DNA with a complex sequence similarity measurement algorithm. In contrast, alignment-free sequence analysis is an algorithm that converts a sequence into a string and extracts features for each sequence. Alignment-free sequence analysis can be analyzed at a low cost, but its accuracy is low. Alignment-based sequence analysis is highly accurate but requires high costs for analysis [1].

In this paper, we propose a new alignment-free sequencing analysis algorithm that can solve the low accuracy problem, which is a disadvantage of the existing alignment-free sequencing analysis algorithm.

**II. SYSTEM MODEL AND METHODS**

As CO1(Cytochrome c oxidase subunit I) is being used as a genetic marker for species classification, it can effectively act on species-level classification [2]. As a result, CO1 sequence FASTA files were collected at the National Center of Biotechnology Information (NCBI) for mammals, birds, reptiles, fish, and amphibia. This was composed of multi-FASTA and analyzed.

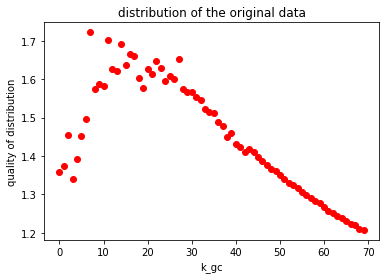
***A. K-mers Embedding and K-mers Slicing with GC-content***

Positional information between sequences is known as important information when analyzing sequence similarity. In this study, a technique called k-mers was applied. K-mers converts a sequence into unique substrings containing positional information as long as k. The analysis was conducted using a new method called k-mers slicing with GC-content in advance as well as simply k-mers embedding. The GC-content can be a primary factor shaping amino acid, so the richness of the GC-content may suggest large amounts of protein genes [3]. Thus, GC-content was used to improve the performance of sequencing analysis.

K-mers slicing with GC-content is a method of compressing capacity, leaving only major information on existing datasets of vast sequence lengths and samples. Through this method, only strings with the highest GC-content of the substrings were adopted to extract the main sequence data replacing the existing massive-length sequence. This allows the elimination of mutations in the sequences and the distinct separation of species, contributing to improving the accuracy of biological clustering or classification. Accordingly, it is possible to improve the algorithm processing speed and analyze considering the distribution of the sequences length.

Here, it is important to set an appropriate k value because the k value refers to the amount of position information of the sequences. According to Fig 1, if the k value is set to be small, information loss occurs, or if the k value is set to be large, information deviation occurs, thus the algorithm performance is degraded. Therefore, I propose the equation that allows us to find an appropriate k value of k.

(1)



**Fig. 1.** The clustering level of amphibia CO1 DNA datasets(quality of distribution) according to the change in the k value of k-mers sliding with GC content(k\_gc). There is a specific section in which the clustering level is maximized according to the change in the k value.

***B. Evaluating the Algorithm***

For each mammals, birds, fish, amphibia, and reptiles sequencing data used in this study, the training datasets was constructed by applying the alignment-based sequence analysis method and the proposed alignment-free sequence analysis method. The accuracy of the supervised learning model SVM (Support Vector Machine) and the unsupervised learning model K-means was measured and compared. To overcome the curse of dimensionality problem, the training datasets was reduced to two dimensions through linear discriminant analysis (LDA) dimension reduction technique and t-sne (t-distributed Stochastic Neighbor Embedding) dimension reduction technique, followed by classification and clustering.

**III. RESULTS**

***A. Optimal K value setting on K-mers Slicing***

The difference in clustering levels between setting the optimal k value in k-mers slicing and not was visually compared through the distribution of data in the two-dimensional plane. The side to which the k value derived through the linear equation is applied generally shows a higher level of clustering.

***C. Result of the Algorithm***

The proposed algorithm compared to MUSCLE, an alignment-based sequence analysis program commonly used in preprocessing COI DNA data, showed higher accuracy. According to Table 1, when comparing the average accuracy of species classification for the five classes, the SVM model achieved 99.12%, about 1.5% higher than before, and 90.7%, about 1.2% higher than K-means model. It also showed excellent results in terms of processing speed. In general, alignment-based sequence analysis is difficult to calculate the time complexity, and the processing speed differs significantly from the changes in the sequence length and number of samples. In contrast, the proposed algorithm can calculate the time complexity, and the difference in processing speed is relatively small [1]. In particular, the reptile datasets with large data sizes were processed 92 times faster than MUSCLE.

**Table 1.** Units Accuracy of MUSCLE and the new alignment-free sequence analysis algorithm

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **SVM with MUSCLE** | **K-means with MUSCLE** | **SVM with**  **the algorithm** | **K-means with the algorithm** |
| Mammals | 97.4% | 87.5% | 99.4% | 94.1% |
| Birds | 98.6% | 92.2% | 99.5% | 87.4% |
| Fish | 99.5% | 92.2% | 100% | 93.0% |
| Reptiles | 100% | 99.9% | 100% | 95.6% |
| Amphibia | 92.5% | 76.0% | 96.7% | 83.4% |

**IV. DISCUSSION AND CONCLUSIONS**

In this work, we were able to overcome the disadvantage of low accuracy, while maintaining the advantages of alignment-free sequence analysis, which was used to process large amounts of DNA sequence datasets. It is expected that using the proposed algorithm will enable quick and accurate judgment in a situation where more sequence analysis is needed in the future to solve problems such as COVID-19 or a decrease in the number of endangered species due to climate change.

**REFERENCES**

[1] Zielezinski, Andrzej, et al. “Alignment-Free Sequence Comparison: Benefits, Applications and Tools.” *Genome Biology*, vol. 18, no. 1, 2017.

[2] Hebert, Paul D., et al. “Biological Identifications through DNA Barcodes.” *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 270, no. 1512, pp. 313–321, 2003.

[3] Du, Meng-Ze, et al. “The GC Content as a Main Factor Shaping the Amino Acid Usage during Bacterial Evolution Process.” *Frontiers in Microbiology*, vol. 9, 2018.

1. Corresponding author [↑](#footnote-ref-2)