Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 5
MMTCS 2024

Research on Sequence Clustering and Alignment in DNA Storage Based on the K-Means Model

Guanglin Yang, Yinting Zhu *

School of Mathematics and Information Science, Nanchang Normal University, Nanchang, China, 330022

* Corresponding Author Email: 18679460830@163.com

Abstract. The technology of DNA storage uses artificially synthesized deoxynucleotide chains to store information, ensuring precise and error-free reading. Compared with traditional electronic information storage, it has advantages in terms of capacity, density, and energy consumption. In this study, a K-Means clustering model is constructed with the aim of accurately clustering DNA sequences after DNA storage sequencing. To objectively evaluate the effectiveness of the model, clustering results are compared in detail with the correct DNA sequences. Experimental data show that when processing 100,000 DNA storage sequencing sequences, the accuracy of the model exceeds 90%, and the entire clustering process only takes 10 seconds. This result fully demonstrates the important role of the K-Means model in restoring original information sequences in DNA storage and provides a solid theoretical and practical foundation for future research.

Keywords: K-Means clustering; DNA storage; photobiology.

1. Introduction

DNA storage technology refers to the technology in which information such as documents, images, and audio is stored and completely read using artificially synthesized deoxynucleotide chains. Compared with traditional electronic information storage, DNA storage has the advantages of large capacity, high density, and low energy consumption [1-3]. In an era where data storage demands are growing exponentially, DNA is expected to become a potential storage medium to replace traditional storage devices. However, some errors are inevitably introduced during the DNA storage process, and error correction and accurate restoration of information stored in DNA are currently the main challenges faced by DNA storage [4-8].

In the 1970s, foreign scholars proposed the idea of using different states of DNA as a way of representing information. In recent years, significant progress in the information storage field regarding DNA storage performance has been made with the enthusiastic participation of many research institutions and companies such as Microsoft, the European Bioinformatics Institute, Columbia University, Washington University, Harvard University, and the UK's Cambridge Consultants [10-12]. In 2017, researchers such as Shipman at Harvard Medical School successfully incorporated DNA sequences encoded with black-and-white images and short video images into the genome of E. coli using the CRISPR-Cas gene-editing system in DNA shearing technology, achieving rapid replication of information [13]. In 2018, Organick and others further accomplished the capability to encode and store 35 different files among over 1.3 million DNA oligonucleotides, and each file could be independently and accurately retrieved through random access. In June of the same year, the Pentagon announced the use of innovative DNA storage strategies to protect large amounts of sensitive information about citizens, and the Defense Advanced Research Projects Agency successfully developed a desktop device capable of writing data onto DNA and other artificial molecular structures [14].

By contrast, domestic research in the field of DNA data storage is still in its infancy. Currently, universities such as Huazhong University of Science and Technology, Tianjin University, and the National University of Defense Technology have formed research teams to conduct relevant studies.

Systematic research on DNA data storage in China is gradually being carried out, with DNA data storage technology listed as a subproject in the 2018 “Synthetic Biology” initiative.

Based on the existing issues in current research, the application of the K-Means model in sequence clustering analysis in DNA storage is explored in this paper. The K-Means clustering model is constructed to accurately cluster DNA sequences after DNA storage sequencing. To objectively evaluate the model, clustering results are compared in detail with the correct DNA sequences, and the clustering speed of the model is tested.

2. Establishment and solution of K-Means model

2.1. Sequential analysis

2.1.1. Sequence error

In DNA storage technology, information needs to undergo many processes, including transmission, DNA synthesis, PCR replication, and DNA sequencing. When information is transmitted through the channel, some channel noises such as white noise are quite significant and can easily cause sequence errors. Sequence errors exist in four types, as shown in Figure 1, 2, and 3. Among them, the breakage error occurs when one of the strands of the DNA double helix breaks during the process of replication, forming one or more temporary single-strand breaks. It is stipulated that if the sequence length is reduced by 25% or more compared to the original target length, a breakage is considered to have occurred.

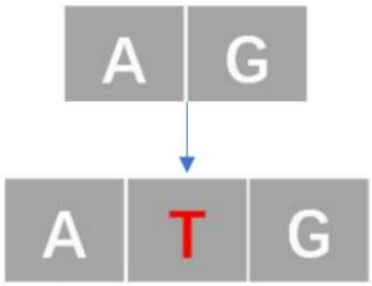

Figure 1. Adds an error type diagram

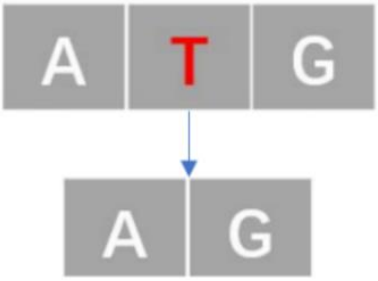

Figure 2. Delete error type diagram

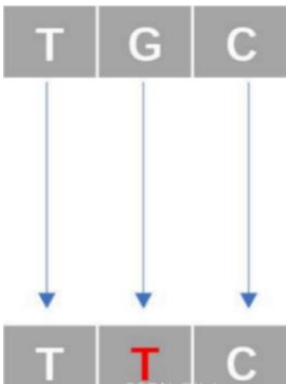

Figure 3. Replacing error types

2.1.2. Copy number analysis

During DNA sequencing, a large number of copies of the original sequences (the sequences storing the correct DNA information) are made. In this paper, the K-Means model is employed to cluster the sequences by comparing the copied sequences with the original sequences. In the original sequences, each sequence is assigned a sequence number. The copied sequences obtained from sequencing also have a sequence number corresponding to the original sequence from which they were copied. To verify the accuracy of the K-Means model in subsequent steps, the sequence numbers of the original and copied sequences are compared, and the number of copied original sequences after sequencing is counted. A bar chart visualizing the number of copied sequences after sequencing is drawn. As shown in Figure 4, the sequence with sequence number 25 has the fewest copies, approximately 90, while the sequence with sequence number 34 has the most copies, exceeding 140. The number of copies for other sequences is relatively uniform, mostly ranging between 100 and 140.

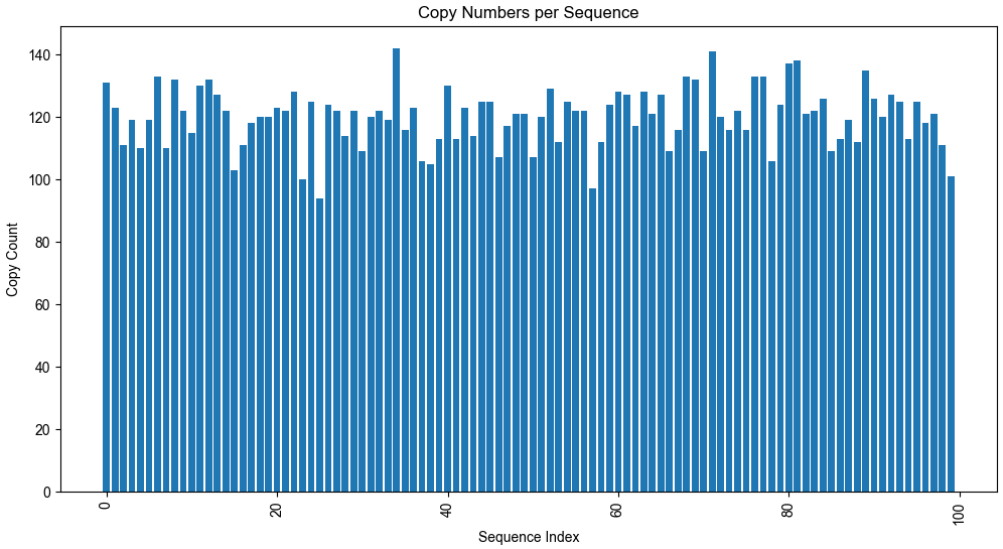

Figure 4. Copy number bar chart of each sequence

2.2. Establishment of K-Means clustering model

2.2.1. K-Means model theory

Clustering algorithms are known as algorithms that divide data samples into several categories based on their intrinsic relationships, without any data labels, utilizing the characteristic that samples within the same category are highly similar, while samples between different categories are less similar. Among these algorithms, K-Means clustering (or K-Means) is considered the most fundamental clustering algorithm. Due to its strong local search ability, high clustering efficiency in large data sample spaces, fast convergence, and ease of implementation [8-9], it has been widely applied in the field of image processing.

Its basic concept involves finding K clusters through iterations, such that the loss function corresponding to the clustering results is minimized. The loss function describes the compactness among the cluster center points. The smaller the value of the loss function, the higher the similarity among the samples within the same cluster, thus indicating better clustering performance. The loss function is usually defined as:

$$J(c,u)=\sum_{i=1}^{M}\left\|x_i-u_{c_i}\right\|^2 \tag{1}$$

In the formula, J (c, u) represents the sum of squared errors between the center points of each sample and the sequence to which it belongs. x_i, c_i, u_{ci}, and M represent the i-th sample, the sequence to which x_i belongs, the corresponding center points of the sequence, and the total number of samples.

2.2.2. K-Means algorithm steps

The core of the K-Means algorithm lies in dividing the given dataset into K clusters based on the number of clusters, and then determining through iterative looping whether the clustering results meet the stopping conditions. The algorithm's process can be specifically divided into four steps:

(1) Data preprocessing. The data is primarily standardized and outliers are filtered.

(2) Randomly select K centers, denoted as u1(0), u2 (0), u3(0), ..., Uk(0).

(3) Define loss function:

$$J(c,u) = \sum_{i=1}^{M} \left\| x_i - u_{c_i} \right\|^2 \tag{2}$$

(4) Set the number of iterations t and repeat the following process to make the loss function J converge:

Evaluate the distance from each sample to the cluster center and assign each sample x_i to the sequence belonging to the nearest center:

$$c_i^t <- \arg\min_k \left\| x_i - u_{c_i} \right\|^2 \tag{3}$$

If the loss condition is not met or the iteration stop condition is not reached, recalculate the center point k for each category:

$$u_k^{(t+1)} <- \arg\min_u \sum_{i:c_j^t=k}^{b} \left\| x_i - u_{c_i} \right\|^2 \tag{4}$$

3. Discussion of results

3.1. Model accuracy

Clustering is a method by which similar data points are grouped. Sequencing reads are encoded using k-mer to transform them into word frequency matrices, which are then clustered using the K-Means algorithm. The number of clusters is set to the number of original sequences. To evaluate the accuracy of the K-Means model, it is assumed here that the sequence indices are unknown post-sequencing. The original sequences corresponding to the replicated sequences are labeled through the clustering method. Clustering is based on sequence similarity and compared with the original sequence indices to evaluate the model's accuracy. From Figure 5, which shows the spatial distribution of sequences in K-Means clustering, it can be observed that similar sequences are categorized into the same group and are represented by the same color. There are 100 types of original sequences, so the number of sequence types here is also 100. The quantities of sequences after clustering with the K-Means model are calculated and visualized, as shown in Table.1 and Figure 6.

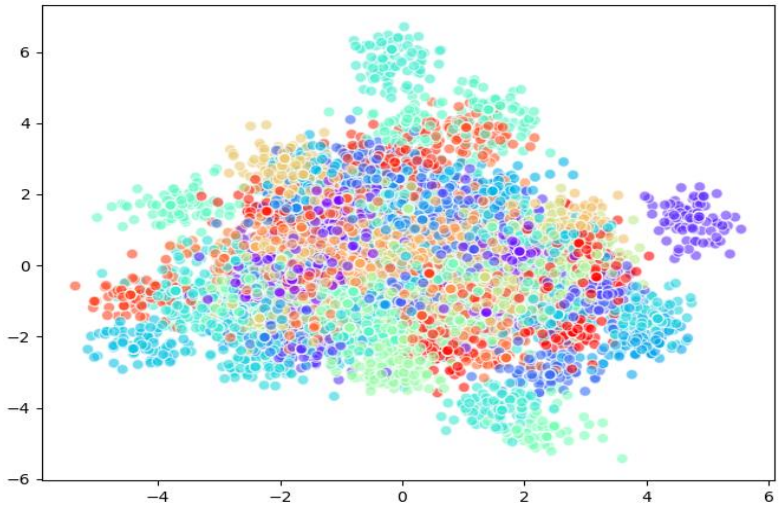

Figure 5. Spatial distribution of K-Means clustering sequences

Table 1. Sample Size of K-Means Clustering for Each Sequence

Sequence	Sample Size
0	111
1	113
2	107
…	…
98	113
99	109

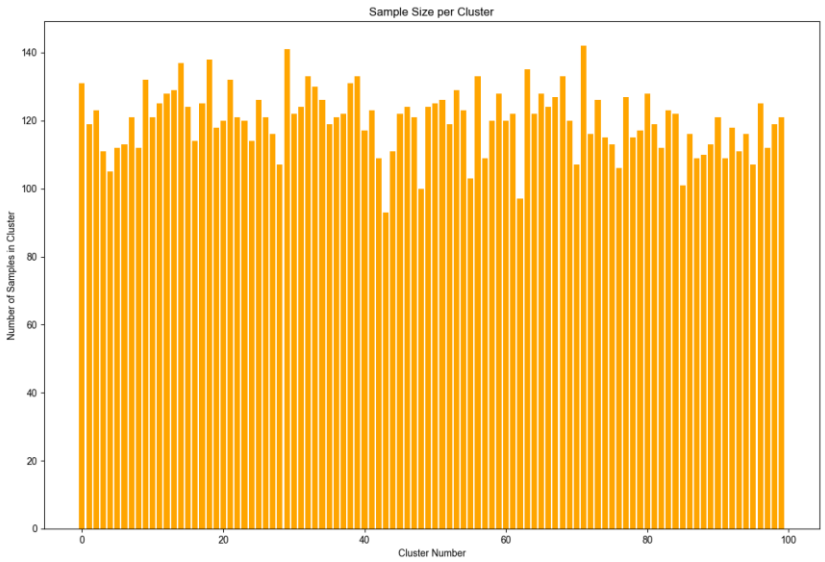

Figure 6. Bar chart of sample size for K-Means clustering sequences

By closely comparing Table.1, Figure 6, and Figure 4, it can be observed that the sample quantities of each sequence are essentially similar, with a calculated similarity of over 90%. The preliminary evaluation indicates that the accuracy of the K-Means model exceeds 90%.

Sequence purity is an important metric for assessing the performance of a clustering model, used to evaluate the model's ability to cluster similar samples into the same sequence. High sequence purity indicates that the model accurately segregates similar samples, whereas low purity may suggest shortcomings in the clustering process. To further evaluate the model's accuracy, the purity of each sequence is obtained, with the results shown in Table.2 and Figure 7. From the analysis of Table.2 and Figure 7, it can be observed that the purity of each sequence clustered by the K-Means model

post-sequencing is close to 1. This result strongly indicates that the model effectively assigns the sequenced DNA reads to their original sequences, thereby confirming the high accuracy of the designed model again.

Table 2. Summary of Purity of Each Sequence

Cluster	Most_Common_ID	Purity
0	73	1
1	44	1
2	53	1
…	…	…
97	1	1
98	24	1
99	66	1

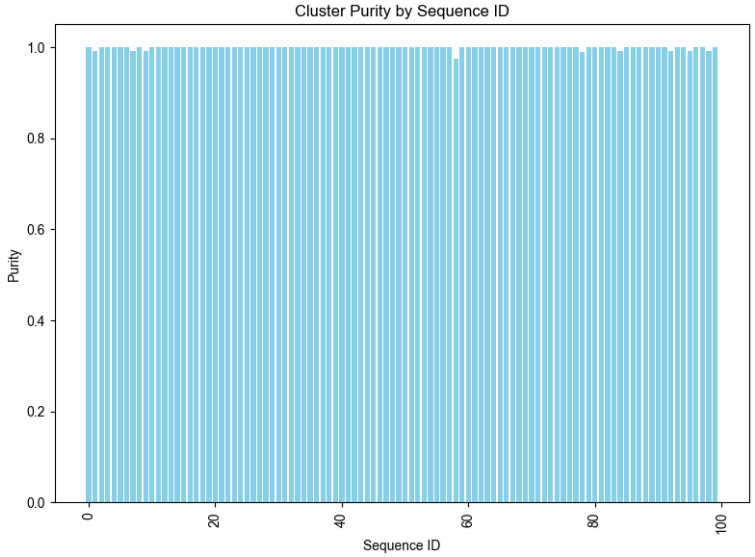

Figure 7. Purity bar chart of each sequence

3.2. Clustering speed

To evaluate the efficiency of the K-Means clustering algorithm in handling data, the time taken by the model to perform the clustering task was precisely measured. The specific method involved starting a timer at the beginning of the clustering process and stopping it when the clustering was completed, thereby recording the total time elapsed for the entire clustering operation.

According to the experimental records, the K-Means algorithm took only 0.18 minutes, or approximately 10.8 seconds, to complete the clustering task. This result indicates that the K-Means algorithm has a quite fast processing speed, enabling it to efficiently cluster large amounts of data in a very short time. The time complexity of the K-Means clustering algorithm is O (Kn), where K is the number of clusters and n is the size of the dataset. In practical applications, since DNA sequences are very long, n is usually quite large. However, due to the initialization and iterative optimization strategies employed by the K-Means clustering algorithm, its actual clustering speed is typically lower than the theoretical time complexity.

4. Conclusion

In this study, the potential application of K-Means clustering technology in the field of DNA data storage and restoration was thoroughly investigated. A K-Means clustering model was built, aimed at accurately clustering DNA sequences post-sequencing of DNA storage. To objectively evaluate the model's effectiveness, the clustering results were compared in detail with the correct DNA

sequences. The experimental data indicated that when faced with the task of processing 100,000 DNA storage sequencing reads, the model achieved an accuracy rate of over 90%, with the entire clustering process taking only 10.8 seconds. Not only is an efficient and accurate data restoration method provided, but the rapid processing speed also makes it possible to handle larger-scale bioinformatics data, suggesting a broad application prospect in future biological computing and data storage technologies.

Acknowledgments

This work was financially supported by the Nanchang Normal University School-level Scientific Research Project 23XJZR03.

References

[1] Zhirnov V, Zadegan R M, Sandhu G S, et al. Nucleic acid memory [J]. Nature Materials, 2016, 15 (4): 366.

[2] Panda D, Molla K A, Baig M J, et al. DNA as a digital information storage device: hope or hype? [J] 3 Biotech, 2018, 8 (5): 239.

[3] Extance A. How DNA could store all the world's data [J]. Nature, 2016, 537 (7618).

[4] Bar-Lev D, Orr I, Sabary O, et al. Deep DNA storage: Scalable and robust DNA storage via coding theory and deep learning [J]. arxiv preprint arxiv: 2109.00031, 2021.

[5] Clermont D, Santoni S, Saker S, et al. Assessment of DNA encapsulation, a new room-temperature DNA storage method [J]. Biopreservation and biobanking, 2014, 12 (3): 176-183.

[6] Howlett S E, Castillo H S, Gioeni L J, et al. Evaluation of DNAstable™ for DNA storage at ambient temperature [J]. Forensic Sci Int Genet, 2014, 8 (1): 170-178.

[7] Doricchi A, Platnich C M, Gimpel A, et al. Emerging approaches to DNA data storage: Challenges and prospects [J]. ACS nano, 2022, 16 (11): 17552-17571.

[8] Wang S, Mao X, Wang F, et al. Data Storage Using DNA[J]. Advanced Materials, 2024, 36 (6): 2307499.

[9] Mayer C, Mcinroy G R, Murat P, et al. An Epigenetics-Inspired DNA-Based Data Storage System [J]. Angew Chem Int Ed Engl, 2016, 128 (37): 11310-11314.

[10] Goldman N, Bertone P, Chen S, et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. [J]. Nature, 2013, 494 (7435): 77-80.

[11] Church G M, Gao Y, Kosuri S. Next-generation digital information storage in DNA [J]. Science, 2012, 337 (6102): 1628-1628.

[12] Erlich Y, Zielinski D.DNA fountain enables a robust and efficient storage architecture [J]. Science, 2017, 355 (6328): 950-954.

[13] Shipman S L, Nivala J, Macklis J D, et al. CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria [J]. Nature, 2017, 547 (7663): 345-349.

[14] Organick L, Ang S D, Chen Y J, et al. Random access in large-scale DNA data storage [J]. Nature biotechnology, 2018, 36 (3): 242-248.