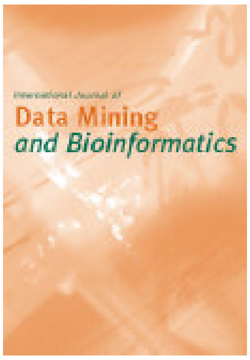


International Journal of
Data Mining
and Bioinformatics

International Journal of Data Mining and Bioinformatics

ISSN online: 1748-5681 - ISSN print: 1748-5673
https://www.inderscience.com/ijdmb

Research on cloud storage biological data deduplication method based on Simhash algorithm

Haijuan Du

DOI: 10.1504/IJDMB.2023.10058134

Article History:

Received:	28 December 2022
Last revised:	10 February 2023
Accepted:	19 June 2023
Published online:	17 October 2023

Research on cloud storage biological data deduplication method based on Simhash algorithm

Haijuan Du

School of Information Science and Engineering,
Jiaxing University,
Jia'xing, 314001, China
Email: hjuan_du@126.com

Abstract: Aiming at the problems of high duplication error, low data similarity accuracy and poor throughput in cloud storage biological data deduplication, a cloud storage biological data deduplication method based on Simhash algorithm is designed. First, we analyse the cloud storage mode, characteristics and advantages of biological data, and determine the distribution rule of biological data in cloud storage. Then, K-nearest neighbour algorithm and Bayesian algorithm are used to extract the features of cloud storage biological data. Finally, Simhash algorithm is introduced to map the data into digital signatures that are longer than specialty to the maximum extent; digital signatures of different dimensions after cloud storage biological data mapping is set, the similarity of biological data signature bit values is determined, and duplication removal is completed. The results show that the proposed method has lower error and is feasible.

Keywords: Simhash algorithm; cloud storage; biological data; deduplication method; K-nearest neighbour algorithm; Bayesian algorithm; digital signature.

Reference to this paper should be made as follows: Du, H. (2023) 'Research on cloud storage biological data deduplication method based on Simhash algorithm', *Int. J. Data Mining and Bioinformatics*, Vol. 27, No. 4, pp.252–266.

Biographical notes: Haijuan Du received her Master's degree from the School of Software, Liaoning University of Engineering and Technology in 2008, and currently a Lecturer in the School of Information Science and Engineering, Jiaxing University. Her research interests include algorithms, data processing and modern educational technology.

1 Introduction

With the continuous development of science and technology, the scale and quality of biological data have also made a qualitative leap. Biological data is a kind of special data produced by the cross cutting of various disciplines in bioinformatics (Wen et al., 2021). With the first evolution of biological science and technology as well as instruments and equipment, the amount of biological data also shows a straight upward trend (Yang et al., 2021). Biological data contains a lot of important biological characteristics and relationships and other important information. It is the key data support for human to fight against diseases. Its mining and research is the key to promote social development. However, due to the interference of gene dimensions, variable data sources and duplicate data in biological data, many biological data cannot be widely used in medical research at present (Kosvyra et al., 2021). Therefore, the research on duplicate data removal methods in biological data has become one of the hot issues in this field.

Yin et al. (2020) proposed a MUSE framework for multi-tiering and SLA driven data deduplication for cloud storage systems. First, by quantifying the performance/space cost combination at different levels, the Dedup SLA can be used as a perfect service quality agreement between service providers and customers. MUSE uses multi-tier deduplication technology to consolidate multiple combinations of deduplication into multiple tiers with different 'deduplication strengths'. Finally, a mechanism called dynamic deduplication reconciliation (DDR) is implemented to adjust the deduplication behaviour at runtime. The results show that MUSE provides higher quality data deduplication services for cloud storage systems that support data deduplication, but the throughput of the storage system after data deduplication is low. Acharya et al. (2020) proposed a diversified biological data deduplication method for multi-view feature selection. In this method, feature selection and marker gene detection are designed as a multi-view and multi-objective clustering problem. On this point, a gene selection method based on unsupervised multi-view multi-objective clustering is proposed. Three important biological data resources (gene ontology, protein interaction data, and protein sequence) and gene expression values are jointly used to design two different views. UMVMO select aims to reduce the gene space to minimise the damage to the classification efficiency of samples, and determines relevant and non redundant gene markers from the three biological data expression benchmark datasets, realising the effective deduplication of biological data. Although this method has high accuracy in calculating similar data, the throughput of the storage system after deduplication is not high and the utilisation rate is low. Berger et al. (2021) proposed a method to compare the data of deduplication biological database by Levenshtein distance and sequence. The method first reviews the history of dynamic programming algorithms used to calculate Levenshtein distance and sequence alignment. Then the heuristic method used in BLAST, the most widely used software in bioinformatics, is a program that searches DNA and protein databases to find evolutionary related similarities. This paper summarises how Levenshtein distance as a mathematical formula can be used to optimise similarity search in biological context. Focusing on the low entropy and fractional dimension of the biological database, the duplicate data is removed according to the research on similarity, so as to realise the research on deduplication of biological data. Although this method has high efficiency in removing duplicate data, its error and accuracy in calculating similar data are low.

In order to solve the above problems of high deduplication error, low data similarity calculation accuracy and throughput, and improve the effectiveness of biological data elimination methods, this paper designs a cloud storage biological data deduplication method based on Simhash algorithm. Through the introduction of Simhash algorithm, the biological data in cloud storage is effectively deduplicated, and after a variety of pre-processing methods to determine the biological data, so as to achieve this research.

The key steps of this study are as follows:

Step 1 Analyse the biological data cloud storage mode, its characteristics and advantages, and determine the distribution rule of biological data in cloud storage, lay a foundation for the subsequent biological data planning and feature classification and extraction, and improve the accuracy of biological data feature extraction.

Step 2 On the basis of the above, K-nearest neighbour algorithm is used to determine the same type of data within the minimum distance range of biological data, and then Bayesian algorithm is used to classify various types of biological feature data to speed up the subsequent feature extraction. Then use the feature subset optimisation selection algorithm to realise the feature extraction of cloud storage biological data, laying the foundation for the next data similarity determination, so as to improve the accuracy of the final data deduplication.

Step 3 After the data feature extraction is completed, Simhash algorithm is introduced to map the data to the maximum extent into the digital signatures that are longer than the specialty, set the digital signatures of different dimensions after the cloud storage biological data mapping. The similarity of the bit value of the biological data signature is determined with the help of approximation algorithm to ensure the accuracy of the duplicate data obtained and reduce the duplication error. Finally, the cloud storage biological data duplication is completed based on Simhash algorithm.

Step 4 Experimental analyses: Set the experimental parameters, and take the methods in Acharya et al. (2020) and Berger et al. (2021) as the comparison methods. Under the same experimental parameters and environmental settings as the proposed methods, the cloud storage biological data deduplication error, data similarity calculation accuracy, and the throughput of the storage system after deduplication are tested respectively, and the corresponding results are analysed.

2 Research on cloud storage analysis and feature extraction of biological data

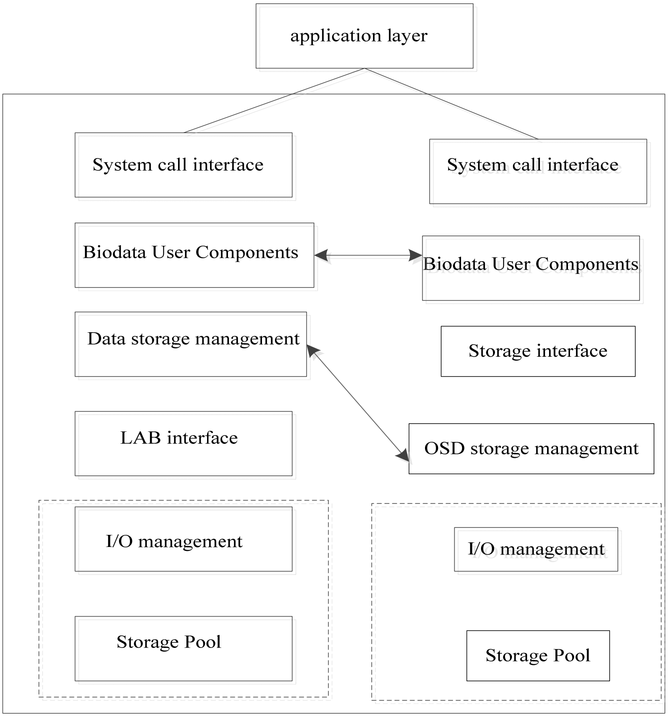
In order to realise the deduplication research of biological data stored in the cloud, it is necessary to analyse the characteristics and advantages of biological data cloud storage. On the basis of this analysis, the characteristics of cloud storage biological data are targeted to be extracted, and the same type of data within the minimum distance range of biological data is determined by the K-nearest neighbour algorithm, and then the Bayesian algorithm is used to classify each type of biological data to speed up the subsequent feature extraction. Then the feature subset optimisation selection algorithm is used to realise the feature extraction of cloud storage biological data, laying the foundation for the next data similarity determination, so as to provide more accurate data support for the subsequent data deduplication and improve the accuracy of the final data deduplication.

2.1 Analysis of biological data cloud storage characteristics

Cloud storage is a storage mode based on cloud computing, which mainly aims at the storage of massive data. In the biological data cloud storage, it provides powerful conditions for the storage of massive biological data. This storage mode requires low cost in storing biological data, and has a good expansion mode. It is an actual storage mode connected to the internet (Saraswathi and Malarvizhi, 2021), providing users with a more transparent storage convenience. This storage mode has unlimited expansion capacity, which can support the continuous expansion of massive data. It also has the advantage of parallel computing, which improves the capacity of massive computing of biological data.

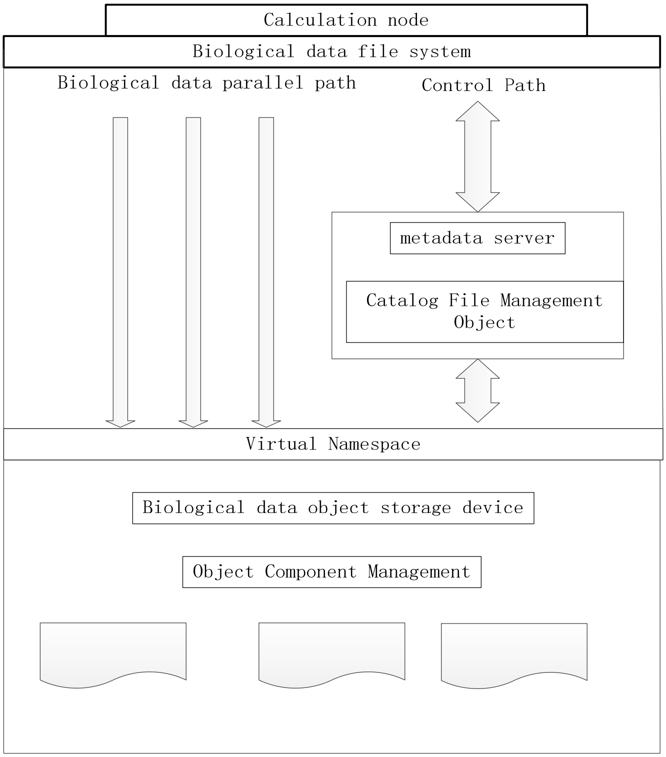
In the biological data cloud storage, massive unstructured data is stored in data files through key keys, and the access interfaces of biological data are linked, effectively integrating the storage advantages of this mode. Abstract data in biological data can be safely stored in local area policies through this storage mode, or directly stored through high-performance scalable switching network structure (Wang et al., 2022). The coupling degree of this storage mode is relatively loose, and it has the advantages of cross platform operability, etc. The biological data cloud storage mode is shown in Figure 1.

Figure 1 Schematic diagram of biological data cloud storage mode



In the biological data cloud storage, the basic unit of storage objects and the core of storage are based on the biological data file system. The storage of biological data is completed in the metadata server and OSD communication (Vélez-Pereira et al., 2021), and the effective sharing of biological data is achieved through network links. The structure of object storage system in biological data cloud storage is shown in Figure 2.

Figure 2 Schematic diagram of object storage system structure in biological data cloud storage


Calculation node
Biological data file system
Biological data parallel path
Control Path
metadata server
Catalog File Management Object
Virtual Namespace
Biological data object storage device
Object Component Management

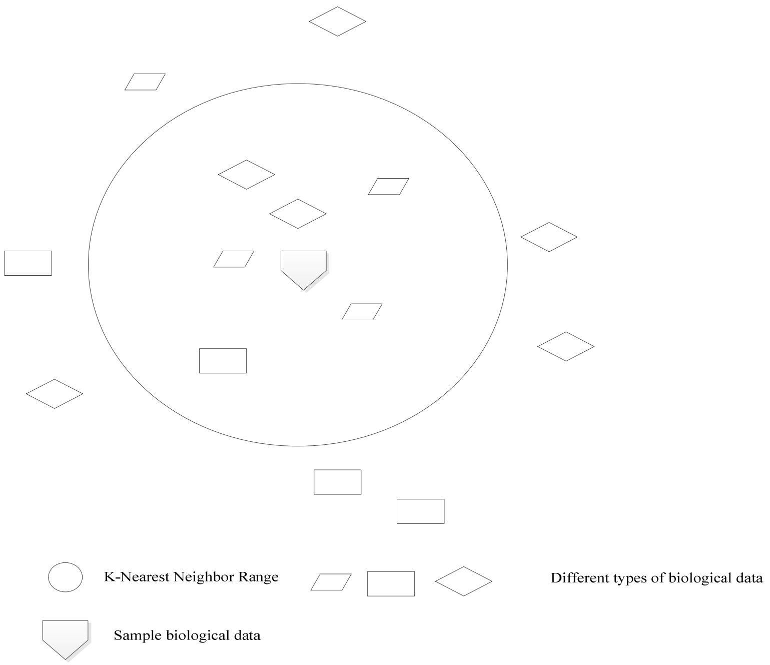
In the biological data cloud storage system, data is a data symbol with a unique mark between files and blocks. By storing these objects and studying the access of biological data, the process of biological data cloud storage is simplified, and its adaptability across data centres is realised (Kanehisa et al., 2021). Therefore, it is determined that biological data in cloud storage has the distribution law of data source dispersion and difficulty in concentration.

In the analysis of cloud storage of biological data, the characteristics and advantages of cloud storage mode of biological data are analysed. It has low cost, good expansion mode, and can realise the effective sharing of biological data and the adaptability of cloud storage to determine the distribution law of biological data in cloud storage. Thus, it lays a foundation for the following biological data planning and feature classification and extraction, which can accelerate the speed of biological data planning and improve the accuracy of biological data feature extraction to a certain extent.

2.2 Research on cloud storage biological data feature extraction

Based on the characteristics of cloud storage of biological data analysed above, in order to achieve effective deduplication of cloud storage biological data, the characteristics of biological data in cloud storage are extracted in this chapter, and the next step of deduplication research is conducted according to the extracted characteristics. There are many kinds of biological data, and the criticality is inconsistent. Therefore, in order to describe the biological data information more clearly, the bioinformatics database is used as the initial data source, and the biological data information is encoded into global information and local information. The composition of different data in biological data is different (Babitsch et al., 2021). Therefore, in the removal of local and all biological information data, the K-nearest neighbour algorithm is used to first determine the known data within the minimum distance range of a biological data sample. If the data is more consistent with the set sample biological data, the data within the range will be regarded as data of the same kind. The process of determination is shown in Figure 3.

Figure 3 Schematic diagram of biological data category K-nearest neighbour determination process



According to Figure 3, calculate the minimum distance from the sample biological data to determine the biological data within this range, and the result is:

$$d_{\min} = \sqrt{\frac{(a+b)^2(a-b)}{\left[(a-b)(a+b)\right]^2}} \tag{1}$$

In formula (1), $d_{\min}$ represents the minimum distance value between the sample biological data, and a and b represent the biological data types within this range, respectively.

On the basis of consistent type data obtained according to formula (1), in order to extract more accurate biological data features, Bayesian algorithm is introduced to determine the probability result that biological data belongs to the same type (Ke et al., 2021), namely:

$$Q(a \mid b) = \frac{Q(b \mid a)Q(a)}{Q(b)} \tag{2}$$

In formula (2), $Q(a|b)$ represents the probability that the biological data belongs to the same type, $Q(a)$ and $Q(b)$ represent the prior probability result, and $Q(b|a)$ represents the posterior probability result.

According to the determined probability results of occurrence of the same type, all biological data are effectively analysed and realised through Bayesian classifier. The classified results are as follows:

$$h_i = \arg\max Q(a) \prod_{i=1}^{f} Q(a \mid b) \tag{3}$$

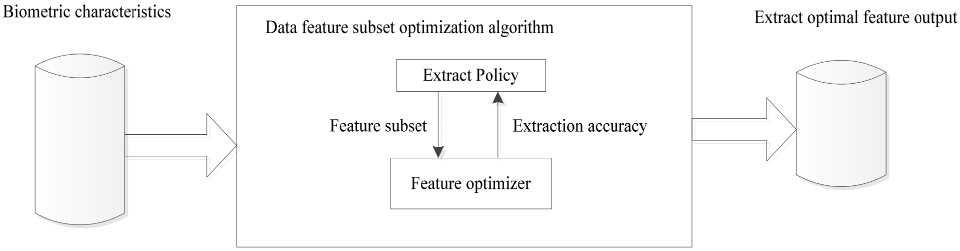
In formula (3), f represents the biological data attribute, $h_i(x)$ represents the Bayesian classifier description, and argmax represents the set function.

According to the classified biological data, the characteristics of each type of biological data are extracted. In this paper, the feature extraction of biological data is realised through the feature subset optimisation selection algorithm. The formula for extracting biological data features is:

$$\varphi_i = \mu \sum_{i=1}^{n} \prod_{i=1}^{f} Q(a \mid b) \tag{4}$$

In formula (4), φ_i represents the feature extraction results of biological data, μ represents the feature extraction optimisation classifier, and n represents the number of biological data types. The process of biological data features extracted using this algorithm is shown in Figure 4.

Figure 4 Schematic diagram of optimisation extraction process of biological data feature subset


Biometric characteristics
Data feature subset optimization algorithm
Extract optimal feature output
Extract Policy
Feature subset
Extraction accuracy
Feature optimizer

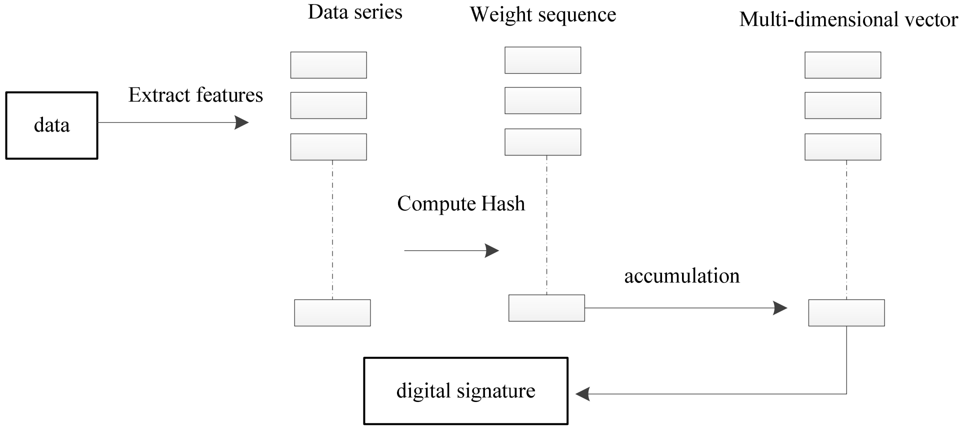
In the feature extraction of cloud storage biological data, K-nearest neighbour algorithm is used to determine the same type of data within the minimum distance range of biological data, Bayesian algorithm is used to determine the probability that biological data belongs to the same type, Bayesian classifier is used to classify different types of biological feature data, and feature extraction of cloud storage biological data is realised through feature subset optimisation selection algorithm, laying a foundation for subsequent duplication research.

3 Design of cloud storage biological data deduplication method based on Simhash algorithm

Based on the above determination of the distribution of biological data in cloud storage and the completion of biometric data classification and feature extraction, in order to achieve effective data deduplication of biological data and solve the problems of high deduplication error, low precision calculation of data similarity and poor throughput in deduplication of biological data in cloud storage, this paper introduces Simhash algorithm to design a deduplication method. First, map the data to a longer digital signature than the professional to the greatest extent, and set the digital signature of different dimensions after the cloud storage biological data is mapped, then determine the similarity of the biological data signature bit value with the help of approximation algorithm, and finally complete the cloud storage biological data deduplication based on Simhash algorithm.

Simhash algorithm is a key algorithm used for data near similarity detection in recent years. The budget estimation method is formed on the basis of hash function to avoid similar collision in data duplication. This algorithm maps data to the maximum extent into digital signatures longer than specialty, which can avoid conflicts between biological data in duplication (Moon et al., 2019). By combining the approximation algorithm to determine the similarity of the biological data signature bit value, the accuracy of the duplicate data obtained is guaranteed, and the deduplication error is further reduced. Moreover, the algorithm can automatically reduce the dimension of the data after mapping to the adjacent data area, reduce the noise carried by the data, and improve the precision of deduplication. Compared with the existing methods, this method is a data processing method with great advantages. Therefore, this algorithm is used in the design of biological data deduplication in cloud storage. The algorithm is basically as shown in Figure 5.

Figure 5 Schematic diagram of basic principle of Simhash algorithm



In this paper, the main implementation process of cloud storage biological data deduplication with Simhash algorithm is as follows:

Step 1 Determine the distance of digital signatures between cloud storage biological data. According to the above determination of cloud storage biological data characteristics, the difference between biological data signatures is calculated by Hamming distance (Wang et al., 2021). Hamming distance is to determine the signature distance of its biological data by controlling the change of coding through biological data transmission in cloud storage (Steffen et al., 2020). This algorithm maps two biological data of the same length on the bit, and the number of different bits is the distance. The calculation formula is:

$$D(x, y) = \sum_{i=1}^{n} x_i \oplus y_i \tag{5}$$

In formula (5), $D(x, y)$ represents the signature distance between the biological data, $\oplus$ representing the symbol of the difference value calculation, and x_i and y_i represent the different signature sequences of the biological data.

Step 2 Map the distance data that determines the digital signature to the Hash algorithm. By mapping the biological data distance obtained above, determine the degree of difference between the original biological data (Palgen et al., 2022). The dataset determined in this process is:

$$(X, Y) = \left[(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\right] \tag{6}$$

The data difference is determined by local hash, namely:

$$D(x, y) \le \sigma_i \rightarrow p\left(f(x)\right) \tag{7}$$

In formula (7), σ_i represents the difference biological data measurement function, p represents the hash coefficient, and $f(x)$ represents the mapping of the digital signature distance data.

Step 3 Set digital signatures of different dimensions after cloud storage biological data mapping. Set any random variable in the biological data to determine the multi-dimensional reference vector of the biological data. The expression formula is:

$$R = \left\{refr_1, refr_2, \dots, refr_n\right\} \tag{8}$$

In formula (8), R represents the arbitrary dataset of the biological data, and $refr_1$, $refr_2$, …, $refr_n$ represents the multi-dimensional reference vector composition of the biological data.

Update the multi-dimensional reference vector of biological data in formula (8) (Chen et al., 1971), and determine the new result as follows:

$$R' = \left\{refr_1', refr_2', \dots, refr_n'\right\} \tag{9}$$

Step 4 Determine the bit values of digital signatures corresponding to different dimensions in cloud storage biological data. By building a biological data signature and setting a circular order table in the library, determine that the digital signature bit value (Qin et al., 2020) of the biological data in the library is expressed as:

$$b_i = \begin{cases} 1 & vec\ refr_i > 0 \\ 0 & vec\ refr_i \leq 0 \end{cases} \tag{10}$$

In formula (10), b_i represents the digital signature bit value of the biological data.

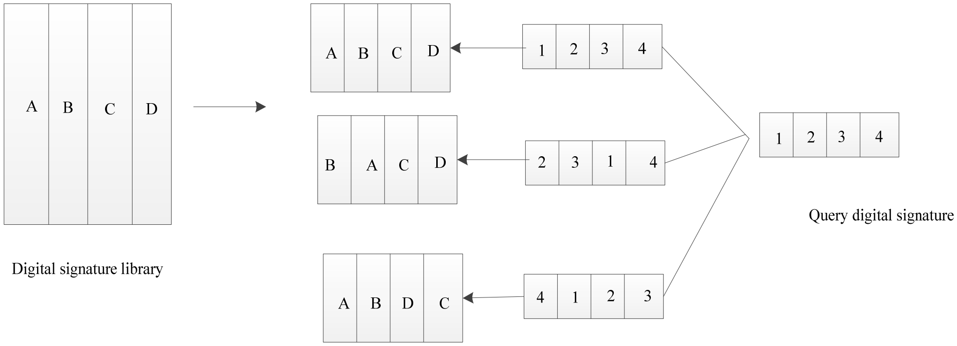
Step 5 Determine the similar value of the digital signature bit value of the biological data. Determine the duplicate data in the biological data by determining the similar value of the digital signature bit value of the biological data, and implement it with the aid of an approximate algorithm (Li, 2021). The result of determining the biological data containing duplicates is:

$$E_i(x) = \frac{1}{n}\sum\nolimits_{i=1}^{n} b_i \frac{\sigma_i}{p} \int h \tag{11}$$

In formula (11), $E_i(x)$ represents the description of the repeated biological data results, and h represents the proximity detection scale factor.

Step 6 Complete the research on cloud storage biological data deduplication. According to the similar value of the digital signature bit value of the above determined object data, determine the range of duplicate data to be removed, and remove duplicate biological data within this range (Sheela and Janet, 2021). The search process of proximity biological data in the deduplication is shown in Figure 6.

Figure 6 Schematic diagram of the search process for biological data of proximity in deduplication



Finally, the result of cloud storage biological data deduplication based on Simhash algorithm is:

$$W_i(x) = \frac{1}{n}\sum\nolimits_{i=1}^{n} b_i E_i(x) \int h \tag{12}$$

In formula (12), $W_i(x)$ represents the result of storing biological data weight.

In the cloud storage biological data deduplication based on Simhash algorithm, Hamming distance is used to determine the distance of the minimum digital signature of biological data, and the distance data that determines the digital signature is mapped to the Hash algorithm. The digital signatures of different dimensions after cloud storage biological data mapping are set to determine the similarity of biological data signature bit values. The cloud storage biological data deduplication is implemented based on Simhash algorithm.

4 Experimental analysis

4.1 Experimental scheme

In order to verify the feasibility of the proposed method, experimental tests were conducted. In the experimental test, 10,000 pieces of data from the biological data professional database were selected as the objects of this study. The specific experimental parameters are shown in Table 1.

Table 1 Experimental parameters

Parameter	*Data*
Quantity of biological data/bar	10,000
Repeat the biological data/bars	3,000
Biological data noise/dB	[–2, 2]
Program implementation software	Simulation software
Denweight similarity detection error	<1

Carry out experimental analysis according to the above set experimental parameters, and take the method in Acharya et al. (2020) and the method in Berger et al. (2021) as the comparison method to test under the same experimental parameter setting as the proposed method. The cloud storage biological data deduplication error, data similarity calculation accuracy, and the throughput of the storage system after deduplication are taken as performance test indicators. In order to ensure the accuracy of the experimental research, the experimental tests are carried out in a unified experimental environment.

4.2 Analysis of experimental results

First, under the same experimental environment, the proposed method, Acharya et al. (2020) method and Berger et al. (2021) method were tested by cloud storage biological data deduplication error experiment. In this experiment, 5,000 pieces of data were randomly selected from 10,000 pieces of data in the professional biological data database as the sample biological data for error removal analysis. The results are shown in Table 2.

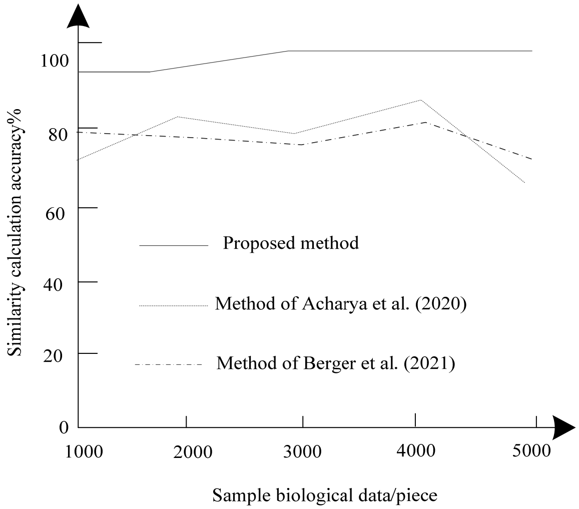
Table 2 Analysis of biological data removal error results (%)

Sample biological data/bar	*Proposed method*	*Acharya et al. (2020) method*	*Berger et al. (2021) method*
1,000	0.25	1.21	1.35
2,000	0.28	1.35	1.45
3,000	0.28	1.45	1.55
4,000	0.29	1.46	1.58
5,000	0.29	1.52	1.69

From the analysis of the experimental results in Table 2, it can be seen that the error results of the proposed method, Acharya et al. (2020) method and Berger et al. (2021) method for sample data deduplication have some changes, and their deduplication errors increase with the increase of the number of sample biological data. When the sample biological data is 1,000, the weight removal error result of the proposed method is 0.25%, while the weight removal error result of the method in Acharya et al. (2020) and the method in Berger et al. (2021) are 1.21% and 1.35% respectively; when the sample biological data is 3,000, the weight removal error result of the proposed method is, while the weight removal error result of the method in Acharya et al. (2020) and the method in Berger et al. (2021) are 1.45% and 1.55% respectively; when the sample biological data is 5,000, the weight removal error result of the proposed method is 0.29%, while the weight removal error result of the method in Acharya et al. (2020) and the method in Berger et al. (2021) are 1.52% and 1.69% respectively. Through the comparison of the error results, the error of the proposed method changes slightly with the increase of the number of biological data in the sample, which shows that the error of the proposed method is low and has good application performance. This is because this method can automatically reduce the dimension of the data, reduce the noise carried by the data, and improve the precision of deduplication.

Then, the proposed method, Acharya et al. (2020) method and Berger et al. (2021) method were tested for the accuracy of data similarity calculation in sample biological data deduplication. In the experiment, 5,000 biological data samples randomly selected in the above experiment were still taken as the research object, and the results of data similarity calculation accuracy of the three methods are shown in Figure 7.

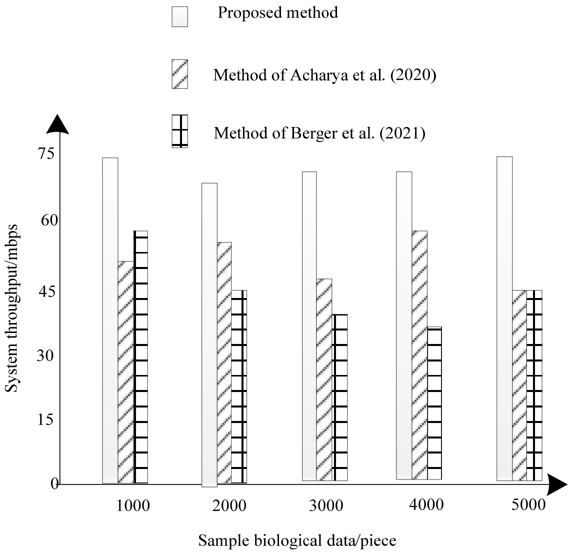
Figure 7 Analysis of data similarity calculation precision results in biological data deduplication



It can be seen from the analysis of Figure 7 that with the constant change of sample biological data volume, the data similarity calculation precision curves of the proposed method, Acharya et al. (2020) method and Berger et al. (2021) method show different changes. Among them, the data similarity calculation accuracy curve of the proposed method shows a steady upward trend with the increase of the sample biological data volume, and the data similarity calculation accuracy is not less than 95%; the similarity calculation accuracy of the method in Acharya et al. (2020) is about 80%, and the similarity calculation accuracy of the method in Berger et al. (2021) is below 80%. Compared with the results of the three methods, the proposed method has higher accuracy of similarity calculation and more reliable deduplication. This is because the method classifies and extracts features of biological data before deduplication, and determines the similarity of the signature bit value of biological data in combination with the approximation algorithm to ensure the accuracy of the duplicate data obtained.

Finally, the proposed method, Acharya et al. (2020) method and Berger et al. (2021) method are tested for the throughput change of the storage system after deduplication of sample storage biological data. The results are shown in Figure 8.

Figure 8 Throughput change analysis of the storage system after deduplication



It can be seen from the analysis of Figure 8 that there are some differences in the throughput of the storage system after the proposed method, the method in Acharya et al. (2020) and the method in Berger et al. (2021) are used to deduplicate the sample storage biological data. Among them, the proposed method has a high throughput after data deduplication, which remains above 70%, while the throughput of the other two methods is below 55%. This is because the proposed method analyses the change characteristics of biological data in cloud storage before removal, and effectively improves the problems in storage according to this characteristic. Based on the above test results, it can be concluded that the proposed method is more applicable to the deduplication of biological data.

5 Conclusions

Aiming at the problems of high duplication error, low data similarity accuracy and poor throughput in cloud storage biological data deduplication, a new cloud storage biological data deduplication method based on Simhash algorithm is designed. By analysing the characteristics and advantages of cloud storage, determine the distribution rule of biological data in cloud storage; then, K-nearest neighbour algorithm is used to determine the same type of data within the minimum distance range of biological data, and Bayesian algorithm is used to classify different types of biological feature data to achieve cloud storage biological data feature extraction; finally, Simhash algorithm is introduced to map the data to a digital signature that is superior to others to the maximum extent, avoiding conflicts between biological data in duplication removal, setting digital signatures of different dimensions after cloud storage biological data mapping, determining the similarity of biological data signature bit values, and completing cloud storage biological data duplication removal based on Simhash algorithm. The test results show that the deduplication error of the proposed method is relatively low, the lowest is about 0.25%, the data similarity calculation accuracy is always higher than 95% when removing biological data from cloud storage, and the throughput of biological data storage after deduplication is improved. The method has certain feasibility and reliable application.

References

Acharya, S., Cui, L. and Pan, Y. (2020) ‘Multi-view feature selection for identifying gene markers: a diversified biological data driven approach’, *BMC Bioinformatics*, Vol. 21, No. 18, pp.483–490.

Babitsch, D., Berger, E. and Sundermann, A. (2021) ‘Linking environmental with biological data: low sampling frequencies of chemical pollutants and nutrients in rivers reduce the reliability of model results’, *Science of the Total Environment*, Vol. 77, No. 2, p.145498.

Berger, B., Waterman, M.S. and Yu, Y.W. (2021) ‘Levenshtein distance, sequence comparison and biological database search’, *IEEE Transactions on Information Theory*, Vol. 67, No. 6, pp.3287–3294.

Chen, G., Chen, G., Wu, D. et al. (2021) ‘An improved Simhash algorithm based malicious mirror website detection method’, *Journal of Physics Conference Series*, Vol. 1971, No. 1, pp.1206–1210.

Kanehisa, M., Sato, Y. and Kawashima, M. (2021) ‘KEGG mapping tools for uncovering hidden features in biological data’, *Protein Science*, Vol. 21, No. 3, pp.19–24.

Ke, P.F., Xiong, D.S., Li, J.H. et al. (2021) ‘An integrated machine learning framework for a discriminative analysis of schizophrenia using multi-biological data’, *Scientific Reports*, Vol. 11, No. 1, pp.3354–3356.

Kosvyra, A., Ntzioni, E. and Chouvarda, I. (2021) ‘Network analysis with biological data of cancer patients: a scoping review’, *Journal of Biomedical Informatics*, Vol. 120, No. 5, pp.103873–103878.

Li, W. (2021) ‘Fast retrieval algorithm of English translation core words based on Simhash’, *Science of the Total Environment*, Vol. 11, No. 2, pp.14–19, Springer, Cham.

Moon, K.R., Dijk, D.V., Wang, Z. et al. (2019) ‘Visualizing structure and transitions in high-dimensional biological data’, *Nature Biotechnology*, Vols. 14/54, No. 1, p.38, p.781.

Palgen, J.L., Perrillat-Mercerot, A., Ceres, N. et al. (2022) ‘Integration of heterogeneous biological data in multiscale mechanistic model calibration: application to lung adenocarcinoma’, *Acta Biotheoretica*, Vol. 70, No. 3, pp.1–24.

Qin, J., Cao, Y., Xiang, X. et al. (2020) ‘An encrypted image retrieval method based on Simhash in cloud computing’, *Computer, Materials and Continuum*, English, Vol. 16, No. 4, p.1532.

Saraswathi, S.S. and Malarvizhi, N. (2021) 'Block level time variant dynamic encryption algorithm for improved cloud security and de-duplication using block level topical similarity', *International Journal of Advanced Intelligence Paradigms*, Vol. 3, No. 4, p.19.

Sheela, J. and Janet, B. (2021) 'Caviar-sunflower optimization algorithm-based deep learning classifier for multi-document summarization', *The Computer Journal*, Vol. 13, No. 7, pp.1478–1481.

Steffen, P., Wu, J., Hariharan, S. et al. (2020) 'OmixLitMiner: a bioinformatics tool for prioritizing biological leads from 'Omics data using literature retrieval and data mining'', *International Journal of Molecular Sciences*, Vol. 21, No. 4, pp.412–417.

Vélez-Pereira, A.M., Linares, C.D., Canela, M. et al. (2021) 'Spatial distribution of fungi from the analysis of aerobiological data with a gamma function', *Aerobiologia*, Vol. 3, No. 16, pp.1–17.

Wang, L., Tan, Y., Yang, X. et al. (2022) 'Review on predicting pairwise relationships between human microbes, drugs and diseases: from biological data to computational models', *Briefings in Bioinformatics*, Vol. 41, No. 3, p.3.

Wang, L., Wang, Y. and Wen, D. (2021) 'Tunable biological nonvolatile multilevel data storage devices', *Physical Chemistry Chemical Physics*, Vol. 23, No. 10, pp.321–328.

Wen, Z., Yan, C., Duan, G. et al. (2021) 'A survey on predicting microbe-disease associations: biological data and computational methods', *Briefings in Bioinformatics*, Vol. 22, No. 3, pp.157-1–157-20.

Yang, C., Cronin, M., Arvidson, K. et al. (2021) 'COSMOS next generation – a public knowledge base leveraging chemical and biological data to support the regulatory assessment of chemicals', *Computational Toxicology*, Vol. 19, No. 1, p.100175, Amsterdam, Netherlands.

Yin, J., Tang, Y., Deng, S. et al. (2020) 'MUSE: a multi-tierd and SLA-driven deduplication framework for cloud storage systems', *IEEE Transactions on Computers*, Vol. 70, No. 5, pp.759–774.