RESEARCH ARTICLE

Accurate annotation method for food microbial communities based on K-mer fragmentation strategy

Wei Du*

Department of Ecological Environmental Protection, Linyi Vocational University of Science and Technology, Linyi, Shandong, China.

Received: January 28, 2026; **accepted:** May 16, 2026.

Accurately annotating food microbial communities is crucial to food quality and safety control. However, existing methods have problems such as high false positives for highly similar sequences, gene copy number interference with abundance calculations, and low retrieval efficiency. This research constructed an accurate annotation scheme for microbial communities suitable for food habitats to improve annotation accuracy and efficiency. The study first conducted quality control on the 16SrRNA sequences of food microbial communities, determined the optimal fragmentation parameters through the K-mer repetition rate model, constructed a reference database with a conflict-free hash index, and corrected the gene copy number. The global comparison optimization, dual-mode consensus annotation, and parallel computing were integrated to optimize performance. The results showed that the precision and recall were both above 0.91, the data processing efficiency was stable at around 0.88, and the lowest retrieval time was 100 ms when the data volume was 5,000. The number of species detected in actual food samples was 7.7 species/sample higher than that of the traditional culture method, and the number of low-abundance species detected was 2.3 times that of the traditional culture method. The annotation method could quickly and accurately detect low-abundance probiotics in yogurt, trace pathogenic bacteria in ham, and toxin-producing molds in bread. This research provided direct technical support to food companies to shorten the microbial testing cycle, provided early warning of safety risks, and optimized fermentation processes.

Keywords: food microbial community; K-mer fragmentation strategy; accurate annotation; hash index; copy number correction.

***Corresponding author:** Wei Du, Department of Ecological Environmental Protection, Linyi Vocational University of Science and Technology, Linyi, Shandong 276000, China. Email: duwei0426@163.com.

Introduction

The composition of food microbial communities is directly related to food quality, safety, and nutritional value. Accurate analyzing its species composition and abundance is the core direction of food microbiology research. It is of great significance to the optimization of food processing technology, prevention and control of corruption risks, and discovery of functional strains [1, 2]. Currently, the popularization of metagenomic sequencing technology has provided massive sequence data for the research of food microbial communities. However, efficient processing and accurate annotation of largescale sequencing data still face multiple challenges [3, 4]. Traditional microbial classification methods based on sequence comparison are prone to false positives when dealing with highly similar sequences in food habitats, and data storage takes up too much memory and has low retrieval efficiency, making

it difficult to adapt to large food microbial samples and significant differences in habitats. In addition, most existing annotation algorithms are not optimized for the unique distribution characteristics of conserved and variable regions of 16SrRNA sequences in food microbial communities. Moreover, the interference of gene copy number differences in species abundance calculations has not been effectively eliminated, resulting in a low detection rate of low-abundance functional microorganisms such as probiotics and pathogenic bacteria, making it difficult to meet the food industry's needs for accurate analysis of microbial communities [5]. K-mer is a nucleotide fragment with a fixed length of k bases. The annotation algorithm based on K-mer technology uses K-mer as the data base and integrates strategies such as hash index and copy number correction to efficiently classify and accurately annotate food microbial communities [6].

Aiming at the memory usage in largescale sequence indexing, Campanelli *et al*. proposed a lossless compression data structure for coloring De Bruyne diagrams. This method mapped the k-substring to its color set, which contained all reference sequence identifiers of the k-substring. The results showed that, by exploiting the inherent repetitiveness of color collections to extract and store repeated patterns only once, space efficiency could be significantly improved, allowing index sizes to be reduced by an order of magnitude in some cases with less impact on query efficiency [7]. In response to the challenges faced by biological data storage, management and transmission, Cai *et al*. classified the algorithms developed in the past fifteen years and analyzed them based on their key technologies and compression optimization strategies. Although these algorithms could theoretically improve data transmission efficiency and save storage costs, their practical applications should be modified to improve support for common data sets and reduce matching time overhead [8]. To solve the sampling difficulties and lack of interdisciplinary research in the impact of deep-sea volcanic ash deposition on the abyss environment, Bai *et al*. combined geology and microbiology to explore the sediments affected by rhyolite volcanic ash in the Kermadec Trench. The results showed that the viruses in these sediments did not directly affect prokaryotes-mediated iron cycling by carrying iron-related accessory genes, but regulated iron cycling of iron-cycling prokaryotes [9]. To classify natural genes and disease-related genes in bioinformatics, Juneja *et al*. proposed a DNA sequence classification method based on machine learning algorithms to analyze the DNA sequence by converting it into subsequences of a specific length, identifying the gene category to which it belongs. The results showed that this method could effectively identify gene family categories including G protein-coupled receptors, tyrosine kinases, synthetases, *etc*. [10].

Existing research has explored technical problems in different fields but has not clearly verified the performance of the annotation algorithm in more special food habitats and complex microbial communities and has not deeply explored the adaptability of the algorithm under lower sequencing depth data. To solve the above challenges, this research aimed to accurately annotate food microbial communities, focusing on the core analysis object of 16SrRNA sequence. A K-mer based food microbial community accurate annotation algorithm with multi-strategy consensus annotation (K-mer-CA) was proposed to efficiently classify and accurately annotate food microbial communities in different habitats. This study constructed a K-mer repetition rate statistical model to optimize fragmentation parameters and combined base binary coding and conflict-free hash index to build a food-specific amplicon reference database. Copy number correction technology was introduced to eliminate the interference of gene copy number differences on abundance calculations, improving data storage and retrieval efficiency and abundance accuracy. The global alignment optimization and dual-mode consensus annotation were integrated to accurately classify highly similar sequences in

food habitats, balancing annotation accuracy and habitat adaptability. The parallel computing architecture and multi-threaded computing mechanism were also integrated to improve the analysis efficiency of largescale food microbial samples through parallel processing of data sharding.

Materials and methods

K-mer feature optimization extraction and database construction of food microbial communities

National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database (https://www.ncbi.nlm.nih.gov/sra) was used for this research. A total of 120 sets of raw metagenomic 16S rRNA paired-end sequencing data of food microorganisms were retrieved from the database with a total data volume of 1,200 G. The data collection period was from January 2025 to June 2025, covering three typical food habitats of fermented foods for 40 sets including 15 sets of yogurt, 10 sets of pickles, and 15 sets of fermented meat products; fresh foods for 40 sets including 20 sets each of fresh fruits and vegetables; and processed foods for 40 sets including 20 sets each of canned foods and baked foods. All sample sequences were obtained through Illumina Nova Seq6000 (Illumina, San Diego, CA, USA) with paired-end 150 bp, the single sample sequencing data volume of 10 G, and the proportion of Q30 bases ≥ 90%. The study employed Intel Xeon Gold 6230 processor, 256 GB of RAM, and a 2 TB solid-state drive for storage under Ubuntu 20.04 operating system (https://ubuntu.com/20-04), Python 3.8 environment (https://www.python.org/) along with bioinformatics analysis toolkits including BioPython (https://biopython.org/) and Pandas (https://pandas.pydata.org/). The number of parallel computing threads was set to 16. The multi-dimensional screening and quality control on the 16SrRNA sequences of food microbial communities were performed. After extracting 16SrRNA sequences from the raw metagenomic sequencing data of food samples, sequence length normalization was performed to remove sequences less than 1,000 bp in length. Low quality bases were then filtered based on the Phred quality value. Bases with a quality value Q < 20 in a single sequence were eliminated when they accounted for more than 15%. The cluster database at high identity with tolerance (CD-HIT) tool (https://www.bioinformatics.org/cd-hit/) was used to deduplicate redundant sequences with 100% sequence similarity, and the sequence with the longest length and the best base quality in each redundant sequence was retained as the representative sequence [11, 12]. This study used the sliding window method to fragment representative sequences into K-mers of different lengths. A K-mer repetition rate statistical model was constructed to determine the optimal k value. The K-mer repetition rate was calculated as follows.

$$R_k = \frac{\sum_{i=1}^{n}(C_{i,k}-1)}{N_k} \quad (1)$$

where R_k was the K-mer repetition rate when the length was k, reflecting the interference of repeated K-mers in the sequence on feature extraction at this k value. n was the number of different K-mer types in a single sequence. $C_{i,k}$ was the number of occurrences of the i-th K-mer in the sequence. N_k was the total number of K-mers in a single sequence. The study calculated the average R_k of all representative sequences under different k values, and selected the k value with the smallest average R_k and covering the specific differences in food microbial sequences as the optimal fragmentation length. The average K-mer repetition rate of various food sample sequences presented a decline followed by an increase with the change of k value. The k value corresponding to the minimum point was the optimal K-mer length for that type of food. The optimal K-mer lengths for different food types had slight differences, but the overall fluctuation range was relatively small. The study finally took the average of the optimal k values of various foods as a unified fragmentation parameter to ensure adaptability to different

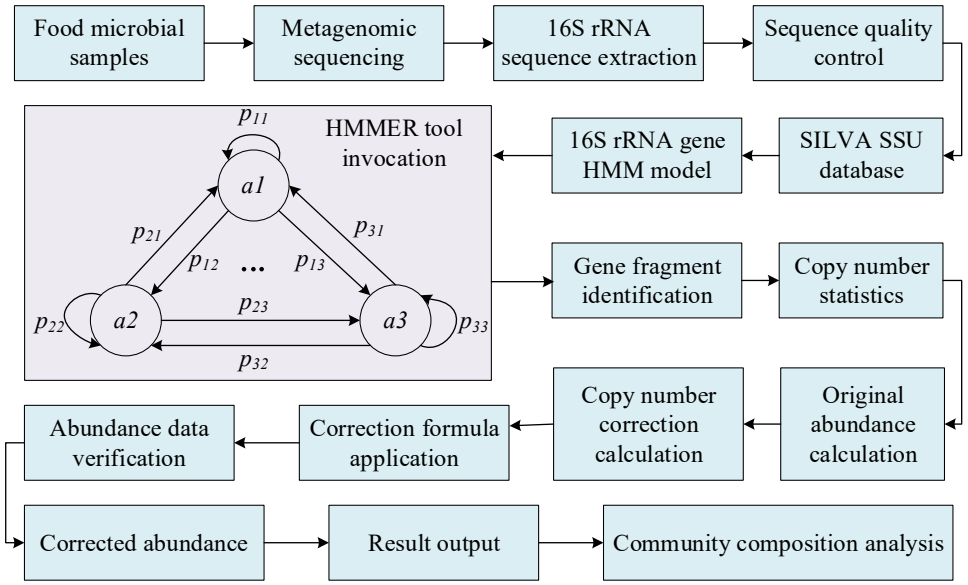

Figure 1. Relative abundance copy number correction process of food microbial species.

food habitats. To eliminate the impact of gene copy number differences on the abundance calculation of food microbial communities, the study introduced copy number correction technology. Through the pre-trained 16SrRNA gene hidden Markov model (HMM), the 16SrRNA gene fragments were identified from the whole genome sequence of food microbial communities [13]. The copy number C_g of the 16SrRNA gene in each genome sequence was counted, and the relative abundance of the species was corrected based on the copy number as shown below.

$$A_{corrected} = \frac{A_{original}}{C_g} \quad (2)$$

where $A_{corrected}$ was the corrected relative abundance of food microbial species. $A_{original}$ was the relative abundance of species calculated based on original sequencing data. C_g was the copy number of the 16SrRNA gene in the genome of the species, which was determined by the number of gene fragments identified by the HMMER tool (https://www.ebi.ac.uk/Tools/hmmer/). The relative abundance copy number correction process of food microbial species used multi-link quality control and precise calculations throughout the entire process to eliminate the interference of gene copy number differences on abundance assessment and ensure the accuracy of community composition analysis (Figure 1). The corrected data could more truly reflect the actual community composition of microorganisms in food samples. To efficiently store and quickly retrieve database sequences, the conflict-free hash index was constructed. Considering 16SrRNA sequence containing A, T, C, G four bases, the study converted each base into a 2-digit binary code with A = 00, T = 11, C = 01, G = 10 [14, 15]. The K-mer sequence was converted into an integer format through base code conversion shown below.

$$Int(K) = \sum_{i=0}^{k-1} B(K_i) \times 4^{k-1-i} \quad (3)$$

where $Int(K)$ was the integer after conversing the K-mer sequence. K_i was the i -th base in the K-mer sequence. $B(K_i)$ was the binary code value corresponding to the base (A = 0, T = 3, C = 1, G = 2). 4^{k-1-i} was the weight coefficient of the i -th base, which decreased from left to right with the base position. Through this conversion, the character K-mer sequence could be converted into numerical data to facilitate hash index

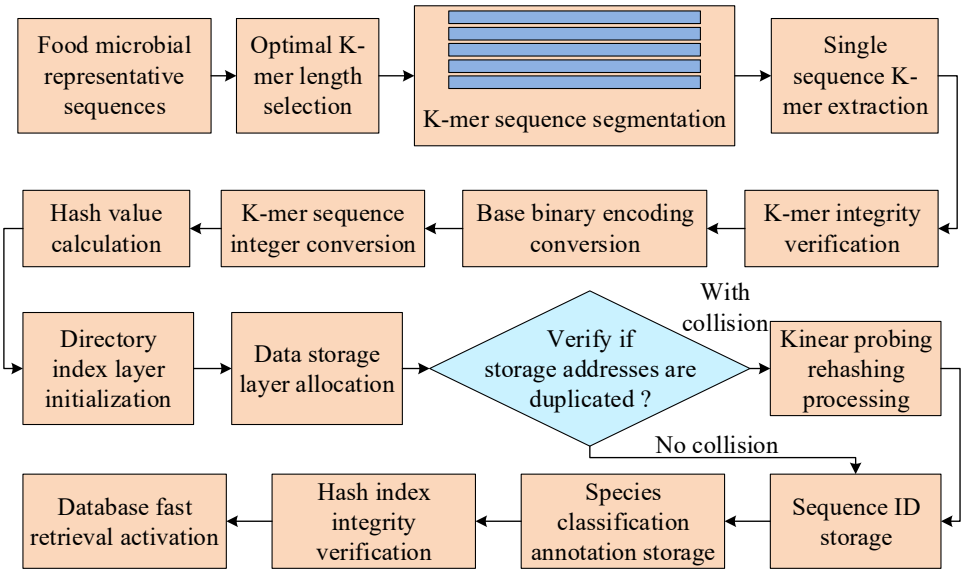

Figure 2. Construction process of K-mer conflict-free hash index for food microbial communities.

storage. Based on the converted K-mer integers, a segmented hash table was constructed as a database index structure, which consisted of a directory index layer and a data storage layer. The directory index layer stored the starting address of the hash table corresponding to each K-mer integer, while the data storage layer stored the food microorganism sequence ID, species classification annotation, and K-mer position information in the sequence corresponding to the K-mer in the order of the starting address [16, 17]. To avoid hash conflicts, the linear detection and then hashing method were used with the conflict resolution as shown below.

$$Addr = (Hash(Int(K)) + d) \bmod M \quad (4)$$

where $Addr$ was the final storage address of K-mer in the hash table. $Hash(\cdot)$ was the hash function. d was the detection increment, and the initial value was 0. M was the hash table length, which was set based on 1.2 times the total number of K-mers in the database to ensure that the hash table load factor was in the optimal range of 0.7 - 0.8. The construction process of the hash index included that the food microorganism representative sequence underwent optimal K-mer fragmentation and integrity verification to obtain a qualified K-mer sequence. Through base encoding conversion and integer conversion, the character K-mer was converted into numerical data for hash calculation (Figure 2). The integrity and accuracy of the constructed food microbial amplicon reference database were verified. The Simpson's diversity index was used to evaluate the coverage of food microbial species in the database as shown below.

$$D = 1 - \sum_{i=1}^{S} P_i^2 \quad (5)$$

where D was Simpson's diversity index with a value ranging from 0 to 1. The closer this value was to 1, the higher the species diversity in the database. S was the total number of food microbial species in the database. P was the sequence proportion of the i -th species in the database.

Construction of accurate annotation algorithm for food microbial communities integrating multi-strategy consensus annotations

The K-mer-CA integrating multi-strategy consensus annotation was constructed. By integrating global comparison optimization, dual-mode consensus annotation and habitat

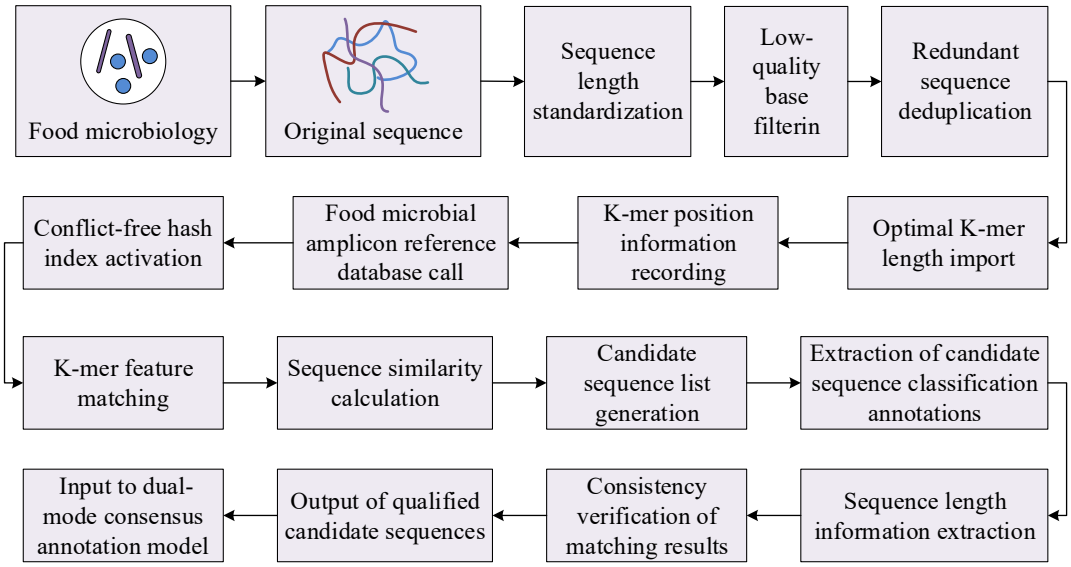

Figure 3. Process of global alignment optimization algorithm for food microbial sequences.

adaptation adjustment, food microbial communities were efficiently classified and accurately annotated. Food microbial sequences were first divided into several K-mer fragments according to the optimal K-mer length, and similar sequences were quickly matched in the reference database through conflict-free hash indexes. The number Q_s of matching K-mers between each candidate sequence and the sequence to be classified was counted. The sum of K-mer position differences G_s was calculated as follows.

$$G_s = \sum_{i=1}^{Q_s} | Pos_{i,query} - Pos_{i,ref} | \quad (6)$$

where G_s was the sum of the K-mer position differences between a single candidate reference sequence and the sequence to be classified. The smaller the value, the closer the K-mer distribution of the two was. $Pos_{i,query}$ was the starting position of the i-th matching K-mer in the sequence to be classified. $Pos_{i,ref}$ was the starting position of the i-th matching K-mer in the candidate reference sequence. Q_s was the number of matching K-mers between the two. The study screened candidate sequences by setting a similarity threshold. The similarity was calculated as follows.

$$S = \frac{Q_s \times L_{ref}}{(L_{query-k+1}) \times (G_s + 1)} \quad (7)$$

where S was the similarity between the candidate reference sequence and the sequence to be classified with the value range of 0 - 1. The larger the value, the higher the similarity between the two. L_{ref} was the length of the candidate reference sequence. L_{query} was the length of the sequence to be classified. Only candidate sequences with similarity $S \geq 0.85$ were retained and entered into the subsequent annotation process to ensure the biological relevance of the candidate sequences and the sequences to be classified. The process of the global comparison optimization algorithm included that the original sequence underwent three-step preprocessing of length normalization, low-quality filtering, and redundant deduplication to ensure sequence quality. The optimal K-mer length parameters were imported, the sequence was fragmented, and the position information was recorded. By calling the reference database and conflict-free hash index, K-mer feature matching was quickly completed, and the number of matches, the sum of position differences, and sequence similarity were simultaneously calculated. After filtering by

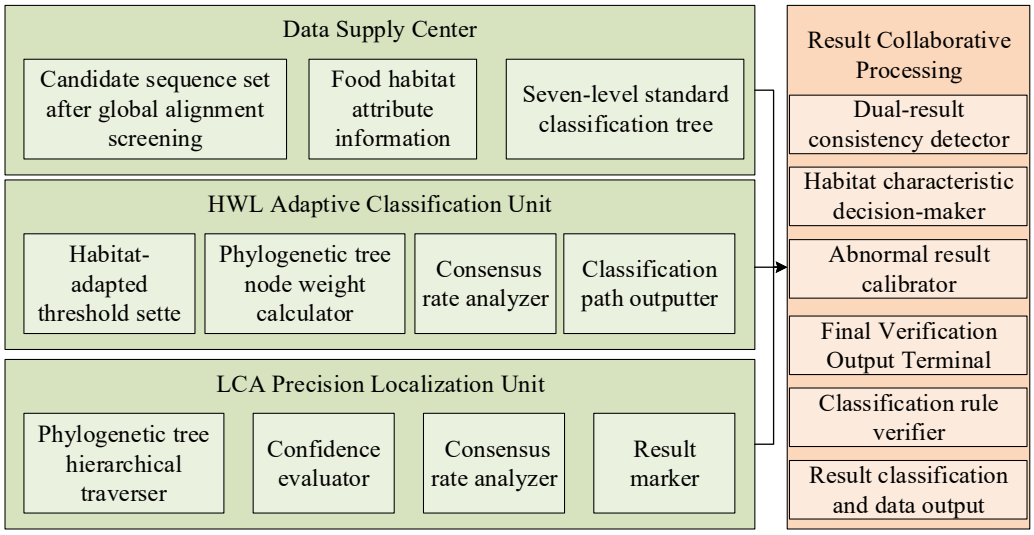

Figure 4. Dual-mode consensus classification annotation model synergy mechanism.

the similarity threshold, qualified candidate sequences were retained, and their classification annotations and length information were extracted. The qualified results were output after consistency verification (Figure 3). To improve the accuracy and adaptability of classification annotation, a dual-mode consensus classification annotation model was constructed including the highest weight leaf node (HWL) adaptive threshold optimization strategy and the lowest common ancestor (LCA) precise positioning method [18, 19]. For the HWL strategy, the study first constructed a food microorganism classification tree based on the classification annotations of candidate sequences. The weight of each node and the node consensus rate were calculated to evaluate the credibility of the classification results as follows.

$$CR = \frac{W_{node}}{W_{total}} \tag{8}$$

where CR was the consensus rate of a certain node in the classification tree, reflecting the support of the node at the corresponding classification level. W_{node} was the weight of the node. W_{total} was the total weight of all nodes at the current classification level. The study introduces an adaptive threshold mechanism, which no longer used a fixed threshold, but dynamically adjusted the consensus rate threshold CR_{th} based on the type of food habitat. For fermented foods, $CR_{th} = 0.6$ was set because the composition of the microbial community was relatively stable. For fresh food, $CR_{th} = 0.5$ was set because the microbial community fluctuates greatly. The HWL strategy started traversing from the bottom of the classification tree. If there was a node of $CR \geq CR_{th}$ at a certain level, the path from the node to the root node was used as the classification result. If there was no node that met the threshold at all levels, the root node annotation was output. To quantify the reliability of LCA classification, the study defined LCA confidence $Conf_{LCA}$ as below.

$$Conf_{LCA} = \frac{L_{final}}{L_{\max}} \tag{9}$$

where $Conf_{LCA}$ was the confidence level of the LCA classification result with the value range of 0 - 1. The larger the value, the more reliable the classification result was. L_{final} was the hierarchical depth where the final classification result was located. L_{max} was the maximum level depth of the classification tree being fixed to 7 corresponding to the "species" level. Only the classification results of $Conf_{LCA} \geq 0.9$ were

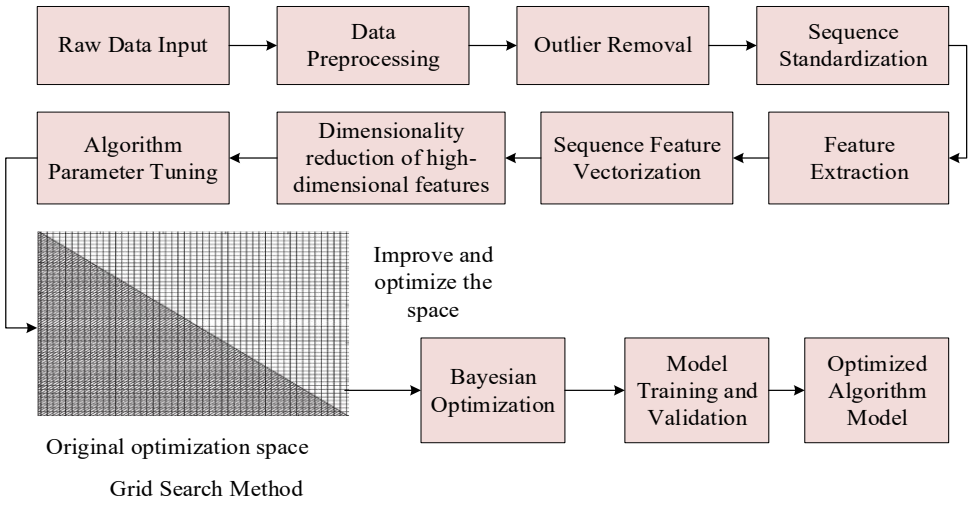

Figure 5. Performance optimization process of food microbial classification algorithm.

retained. Otherwise, they were marked as "undetermined" to prompt subsequent manual verification. The synergistic mechanism of the dual-mode consensus classification annotation model was that, through the parallel calculation and complementary results of HWL and LCA strategies, a balance between accuracy and coverage of food microbial classification was achieved. The data supply center provided unified and standard basic data for the entire system to ensure the accuracy of subsequent operations. The HWL adaptive classification unit and the LCA precise positioning unit worked in parallel and independently based on the same data source. The former filtered classification results using habitat dynamic thresholds and node consistency rates, while the latter obtained high-confidence results through classification tree traversal and common node localization. The results of both were simultaneously transmitted to the result collaborative processing center and were passed to the final verification output after consistency detection, habitat adaptation decision-making, and abnormal calibration (Figure 4). To adapt to the analysis needs of largescale food microbial samples, the research further optimized the algorithm performance and integrated parallel computing architecture and multi-threaded efficient computing mechanism. The operation process after K-mer-CA performance optimization was that the original data sequentially passed through the data preprocessing module including outlier removal and sequence standardization, feature extraction unit including implementing sequence feature vectorization and high-dimensional feature dimensionality reduction, K-mer-CA algorithm parameter tuning using grid search method and Bayesian optimization, model training and verification steps before outputting the optimized algorithm model and performance report (Figure 5). To evaluate the predictive performance of K-mer-CA, all 120 sets of metagenomic data from food samples were randomly divided into a training set and a test set in a 8:2 ratio. The training set contained 96 samples, which were used for learning and optimizing the model parameters. The test set contained 24 samples, which were used to verify the predictive accuracy of the model. The comparison results of predicted values and true values were all calculated based on the test set data.

Actual food sample collection and traditional detection method comparison

To verify the annotation reliability of K-mer-CA algorithm, four typical food samples were selected for microbial community analysis including plain yogurt (Inner Mongolia Yili

Industrial Group Co., Ltd., Hohhot, Inner Mongolia Autonomous Region, China), fresh lettuce (Shanghai Qingmei Green Food (Group) Co., Ltd., Shanghai, China), low-temperature ham (Henan Shuanghui Investment Development Co., Ltd., Luohe, Henan, China), and vacuum-packed bread (Taoli Bread Co., Ltd., Shenyang, Liaoning, China). 25 g of each food sample was mixed with 225 mL of sterile physiological saline for homogenization treatment followed by the preparation of 10-fold gradient dilution solutions. The diluted solution was then spread on LB and PDA culture medium plates (Bailingwei Technology Co., Ltd., Beijing, China) for bacterial and fungal counting, respectively, after incubating at 37°C for 48 hours. Based on the characteristics of colony morphology, color, and size, the colonies were preliminarily classified, and single colonies were picked for Gram staining and microscopic observation. Bacterial species identification was carried out by using 16S rRNA gene polymerase chain reaction (PCR) and Sanger sequencing method. The PCR reaction was conducted using Mastermix 16S Complete reagent kit (Molzym GmbH & Co. KG, Bremen, Germany), containing universal primers for the V3/V4 variable region of 16S rRNA gene. The PCR reaction included 20 µL of 2.5× Mastermix, 2 µL of template DNA, and nuclease free water up to 50 µL. The PCR program was 95°C for 5 minutes followed by 35 cycles of 95°C for 30 seconds, 58°C for 40 seconds, 72°C for 2 minutes, and 72°C for for 5 minutes. The PCR products were verified by 1% agarose gel electrophoresis before Sanger sequencing by Beijing Heuluahua Genetech Technology Co., Ltd. (Beijing, China). The sequencing results were compared with the reference database through NCBI BLAST (https://blast.ncbi.nlm.nih.gov/) to determine the bacterial species. To verify the performance advantages of K-mer-CA proposed in this study, three algorithms commonly used in the field of food microbial annotation were selected for comparative experiments including sequence alignment-based microbial classification algorithm (SAMCA) [20], naive Bayesian microbial annotation algorithm (NBMA) [21], and deep learning-based microbial community classification algorithm (DL-MCCA) [22]. The precision and recall rates of different algorithms were determined by comparing the classification results with the actual species labels. The species annotation results output by the algorithm were taken as the predicted values, while the database standard annotations were regarded as the true values. The precision was calculated by the ratio of the number of correctly predicted samples to the total number of samples. The recall was calculated by the ratio of the number of correctly predicted positive samples to the total number of actual positive samples.

Results and discussion

Performance evaluation of K-mer-CA

The results demonstrated that the precision of K-mer-CA quickly rose to 0.92 when the number of iterations reached about 100 and remained stable, showing that the algorithm could quickly converge to high precision in early iterations (Figure 6a). The recall of K-mer-CA also rose rapidly to 0.91 when the number of iterations reached about 100 and remained stable, showing that the algorithm also performed well in recall (Figure 6b). The results highlighted the core advantages and practical value of K-mer-CA in accurately and efficiently annotating food microbial communities. The relationship between the predicted values of different algorithms and the true values showed that there was a high consistency between the predicted values and the true values. The data points clustered closely around the ideal fit line, demonstrating the algorithm's high predictive accuracy (Figure 7a). The consistency between the predicted value of DL-MCCA and the true value was slightly lower (Figure 7b), while the NBMA and SAMCA data points were concentrated in low predicted value areas but scattered in high predicted value areas, showing poor consistency between predicted value and true values (Figures 7c, 7d). The results indicated that K-mer-CA performed excellently in prediction performance, and the deviation between the predicted value and the true value

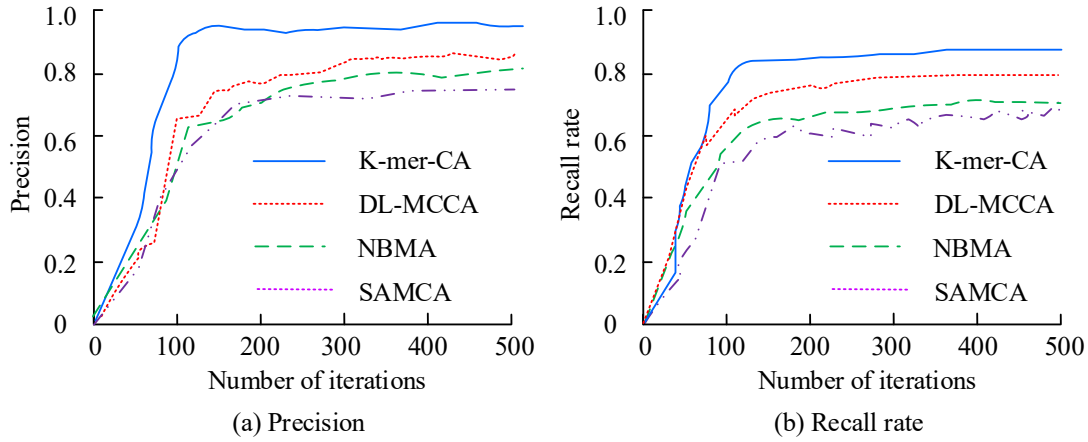

Figure 6. Precision and recall rates of different algorithms.

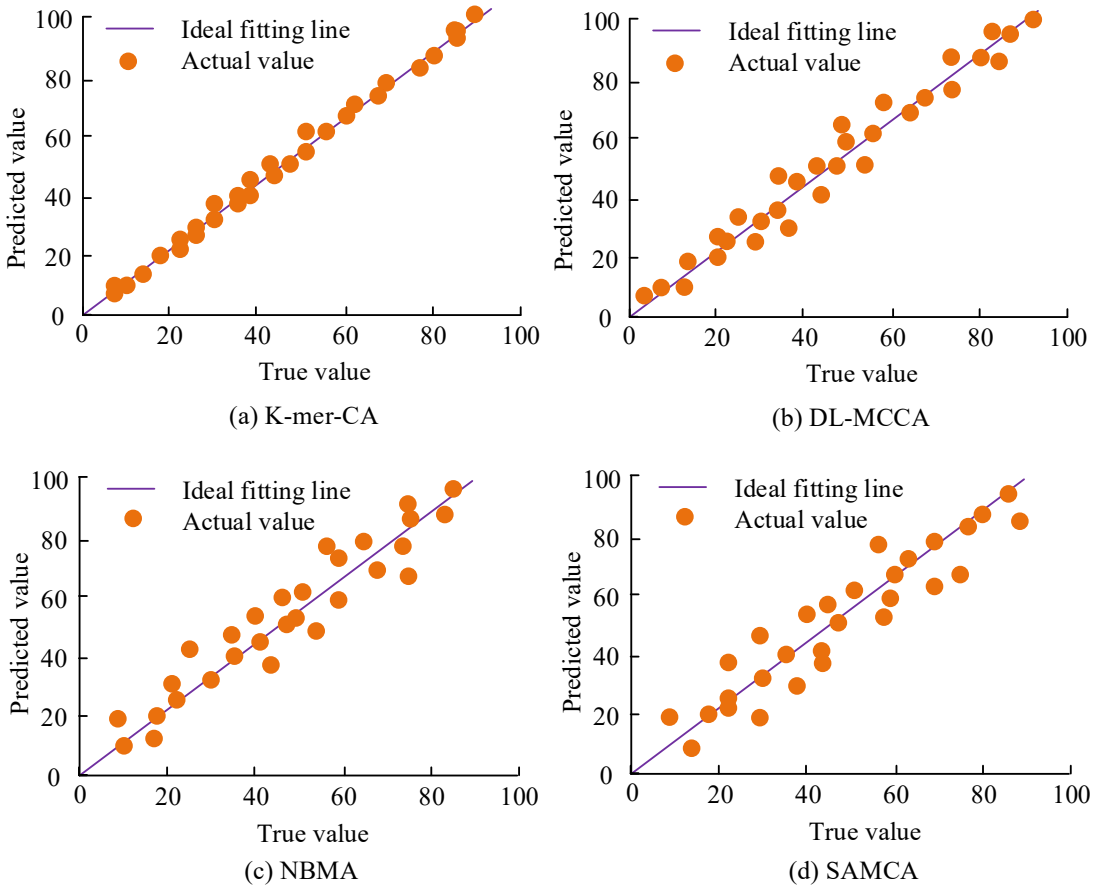

Figure 7. The relationship between the predicted values and the true values.

was small. The data processing efficiency and retrieval time of different algorithms demonstrated that K-mer-CA data processing efficiency increased rapidly, reaching more than 0.85 when the number of iterations was about 50, and finally stabilized at around 0.88 (Figure 8a). K-mer-CA maintained the lowest retrieval time of 100 ms when the data volume was 5,000,

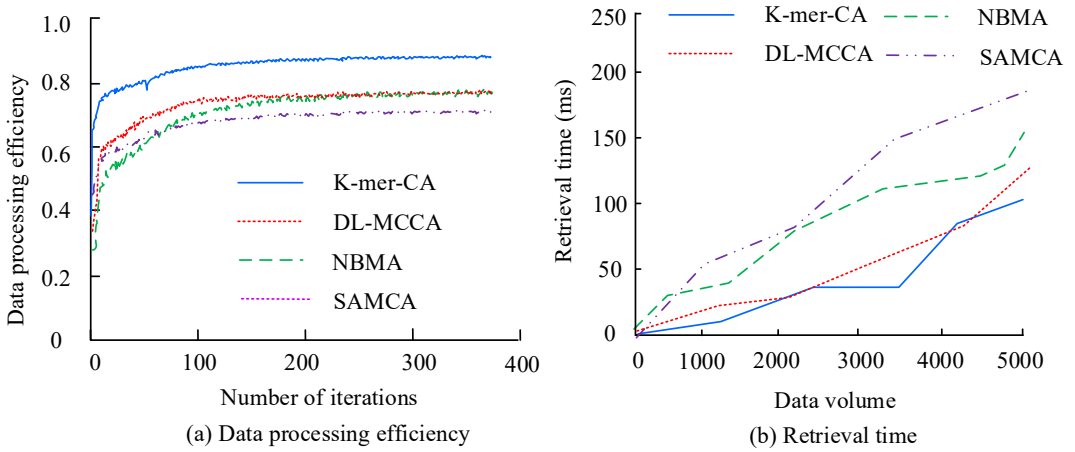

Figure 8. Data processing efficiency and retrieval time of different algorithms.

Table 1. Comparison of microbial community between K-mer-CA and traditional culture method.

Food samples	Methods	Dominant microorganisms (relative abundance)	Detected functional microorganisms	Total number of species detected	Detection volume of low-abundance species
Plain yogurt	K-mer-CA	*Bifidobacterium lactis* (32.5%) *Streptococcus thermophilus* (28.3%)	*Lactobacillus casei* (1.2%)	15	6
	Traditional cultivation	*Bifidobacterium lactis* (31.8%) *Streptococcus thermophilus* (27.9%)	No	9	2
Fresh lettuce	K-mer-CA	*Pseudomonas* (18.6%) *Enterobacteriaceae* (12.3%)	*Pseudomonas fluorescens* (2.1%) *Escherichia coli* (1.5%)	22	8
	Traditional cultivation	*Pseudomonas genus* (17.9%) *Enterobacteriaceae* family (11.8%)	*Pseudomonas fluorescents* (2.0%) *Escherichia coli* (1.4%)	12	3
Low-temperature ham	K-mer-CA	*Staphylococcus* (25.8%) *Lactobacillus* (19.2%)	*Listeria* (0.3%)	16	5
	Traditional cultivation	*Staphylococcus* (25.1%) *Lactobacillus* (18.8%)	*Listeria* (0.3%)	10	2
Vacuum-packed bread	K-mer-CA	Yeast (8.6%) Mold (3.2%)	*Aspergillus flavus* (0.8%)	19	7
	Traditional cultivation	Yeast (8.2%) Mold (3.0%)	No	11	3

showing excellent retrieval performance (Figure 8b). K-mer-CA had dual advantages in data processing efficiency and retrieval speed and was more suitable for scenarios that required higher processing efficiency and retrieval speed. K-mers-CA performed better in all indicators, achieving rapid convergence and maintaining a stable high recognition accuracy. It also had shorter search time and higher processing efficiency, indicating that this algorithm had significant advantages when dealing with highly similar food microbial sequences, large amounts of data, and complex habitats. It could effectively improve the accuracy and operational efficiency of microbial community annotation, which was suitable for the rapid analysis of large-scale food microbial samples.

Analysis of the application effect of K-mer-CA on actual food samples

The comparison between the analysis results of the microbial community composition of four types of food samples by K-mer-CA and the traditional culture method demonstrated that K-mer-CA could accurately annotate dominant microorganisms in all types of samples, and the relative abundance deviations of dominant species from traditional culture methods were less than 2%. In addition, the total number of microbial species detected by K-mer-CA for the

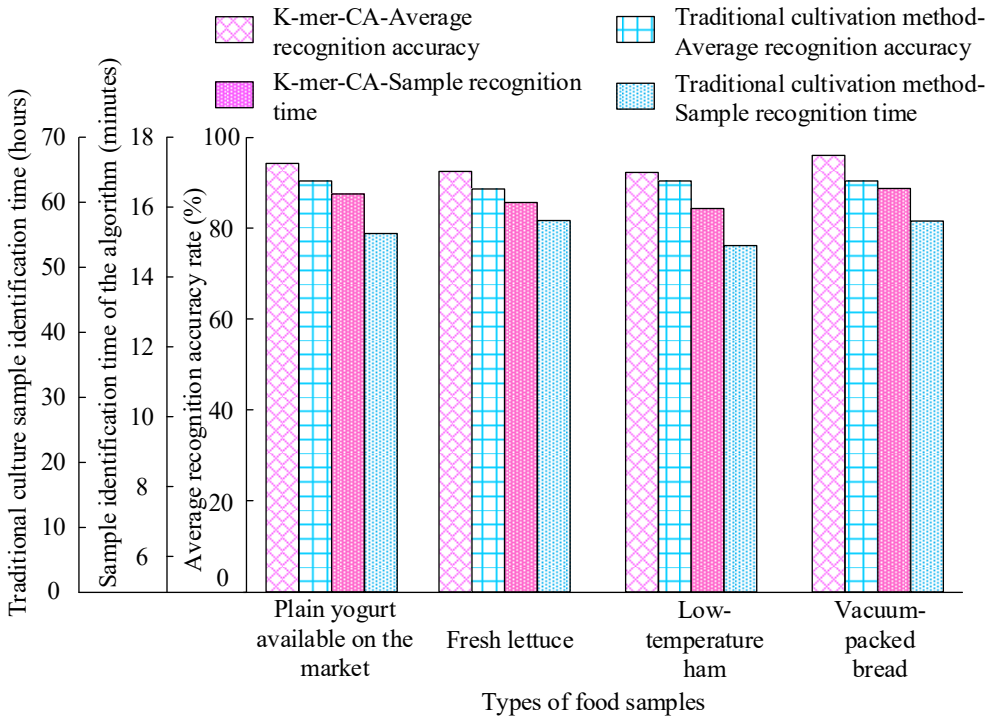

Figure 9. Recognition accuracy and efficiency of K-mer-CA and traditional culture method.

four types of samples was higher than that of the traditional culture method. The average detection amount reached 18.2 species/sample, while the traditional culture method was only 10.5 species/sample. Especially in detecting low-abundance species, the detection amount of the proposed algorithm was 2.3 times that of the traditional culture method, reflecting stronger species coverage (Table 1). The recognition accuracy and efficiency of K-mer-CA and the traditional culture method showed that K-mer-CA algorithm achieved an average recognition accuracy rate of approximately 95% for the commercial plain yogurt samples, which was higher than 89% of the traditional culture method. The sample recognition time of K-mer-CA was approximately 17 minutes, significantly shorter than the time required by the traditional culture method. For the fresh lettuce samples, the K-mer-CA average recognition accuracy was about 92%, which was higher than the 87% of the traditional culture method with a recognition time of approximately 16 minutes compared to 58 hours in the traditional method. For the low-temperature ham samples, the average recognition accuracy was about 93%, higher than the 88% of the traditional culture method with a recognition time of approximately 15.5 minutes. For the vacuum-packed bread samples, the average recognition accuracy was about 96%, higher than the 89% of the traditional culture method with a recognition time of approximately 16.1 minutes (Figure 9). The accuracy of K-mer-CA was superior to that of the traditional method in all samples, and the recognition time was significantly shortened. The results indicated that K-mer-CA had higher accuracy and better efficiency in identifying microbial communities in actual food samples and had good application potential. The application effect of K-mer-CA in practical scenarios was superior to that of traditional cultivation methods. This proposed method could not only detect more microbial species but also significantly enhance the identification ability of low abundance species. Moreover, it outperformed traditional methods in identification accuracy and detection time. K-mer-CA could be applied to various food

environments such as fermented foods, fresh foods, and processed foods, which enabled rapid, accurate, and high-throughput identification of microorganisms, providing practical technical support for food quality safety monitoring, risk warning, and production process optimization.

References

1. Kautsar A. 2025. DNA sequence classification using machine learning models based on k-mer features. J Comput Digit Bus. 4(2):100-105.
2. Uddin M, Islam MK, Hassan MR, Jahan F, Baek JH. 2023. A fast and efficient algorithm for DNA sequence similarity identification. Complex Intell Syst. 9(2):1265-1280.
3. Van Etten J, Stephens TG, Bhattacharya D. 2023. A k-mer-based approach for phylogenetic classification of taxa in environmental genomic data. Syst Biol. 72(5):1101-1118.
4. Shaw J, Yu YW. 2022. Theory of local k-mer selection with applications to long-read alignment. Bioinformatics. 38(20):4659-4669.
5. Jha M, Gupta R, Saxena R. 2023. Fast and precise prediction of non-coding RNAs (ncRNAs) using sequence alignment and k-mer counting. Int J Inf Technol. 15(2):577-585.
6. Arslan H. 2023. A k-mer based metaheuristic approach for detecting COVID-19 variants. Dicle Univ Eng Fac Eng J. 14(1):17-26.
7. Campanelli A, Pibiri GE, Fan J, Patro R. 2024. Where the patterns are: Repetition-aware compression for colored de Bruijn graphs. J Comput Biol. 31(10):1022-1044.
8. Cai J, Hu C, Wang H, Shen Z. 2024. A survey on gene sequence compression algorithms based on reference sequences. Front Data Comput. 6(4):59-76.
9. Bai S, Wang Z, Guo Y, Xu H, Li J, Peng X. 2025. Iron is an important influence of volcanic ash input on the evolution of deep-sea ecosystems. Microbiol Spectr. 13(9):e00715-e00725.
10. Juneja S, Dhankhar A, Juneja A, Bali S. 2022. An approach to DNA sequence classification through machine learning: DNA sequencing, k-mer counting, thresholding, sequence analysis. Int J Reliab Qual E-Healthc. 11(2):1-15.
11. Liu S, Koslicki D. 2022. CMash: Fast, multi-resolution estimation of k-mer-based Jaccard and containment indices. Bioinformatics. 38(1):28-35.
12. Jenike KM, Campos-Domínguez L, Boddé M, Cerca J, Hodson CN, Schatz MC, *et al*. 2025. K-mer approaches for biodiversity genomics. Genome Res. 35(2):219-230.
13. Theepalakshmi P, Srinivasulu Reddy U. 2024. Finding the transcription factor binding locations using novel algorithm segmentation to filtration (S2F). J Ambient Intell Humaniz Comput. 15(9):3347-3358.
14. Dao FY, Lv H, Zhang ZY, Lin H. 2022. BDselect: A package for k-mer selection based on the binomial distribution. Curr Bioinform. 17(3):238-244.
15. Svetlicic E, Alarcon LA, Karlsson R, Jers C, Mijakovic I. 2025. Mass spectrometry-based analysis of surface proteins in *Staphylococcus aureus* clinical strains: Identification of promising k-mer targets for diagnostics. J Proteome Res. 24(9):4575-4585.
16. You H, Yu L, Tian S, Ma X, Xing Y, Song J, *et al*. 2022. Anti-cancer peptide recognition based on grouped sequence and spatial dimension integrated networks. Interdiscip Sci Comput Life Sci. 14(1):196-208.
17. Jiang L, Sun N, Zhang Y, Yu X, Liu X. 2023. Bioactive peptide recognition based on NLP pre-train algorithm. IEEE/ACM Trans Comput Biol Bioinform. 20(6):3809-3819.
18. Djukova AP, Djukova EV. 2024. Recognition of specific genome regions using machine learning methods. Pattern Recognit Image Anal. 34(4):924-929.
19. Simon K, Vicent M, Addah K, Bamutura D, Atwiine B, Nanjebe D, *et al*. 2023. Comparison of deep learning techniques in detection of sickle cell disease. Artificial Intelligence and Applications. 1(4):252-259.
20. Helal M, El-Gindy H, Gaeta B, Sinchenko V. 2023. High performance multiple sequence alignment algorithms for comparison of microbial genomes. arXiv:2312.02994.
21. Zhao H, Zhang S, Qin H, Liu X, Ma D, Han X, *et al*. 2024. DSNetax: A deep learning species annotation method based on a deep-shallow parallel framework. Brief Bioinform. 25(3):bbae157.
22. Remya S, Anjali T. 2023. An intelligent and optimal deep-learning approach in sensor-based networks for detecting microbes. IEEE Sens J. 23(17):19367-19381.