Application and Comparison of Machine Learning and Database-Based Methods in Taxonomic Classification of High-Throughput Sequencing Data

Qinzhong Tian\textsuperscript{1,2}, Pinglu Zhang\textsuperscript{1,2}, Yixiao Zhai\textsuperscript{1,2}, Yansu Wang\textsuperscript{1,2,*}, Quan Zou\textsuperscript{1,2,*}

1. Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China
2. Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 324003, Zhejiang, China.

*Corresponding author: zouquan@nclab.net; wangyansu@uestc.edu.cn

Abstract

Motivation: The advent of high-throughput sequencing technologies has not only revolutionized the field of bioinformatics but has also heightened the demand for efficient taxonomic classification. Despite technological advancements, efficiently processing and analysing the deluge of sequencing data for precise taxonomic classification remains a formidable challenge. Existing classification approaches primarily fall into two categories, database-based methods and machine learning methods, each presenting its own set of challenges and advantages. On this basis, the aim of our study was to conduct a comparative analysis between these two methods while also investigating the merits of integrating multiple database-based methods.

Results: Through an in-depth comparative study, we evaluated the performance of both methodological categories in taxonomic classification by utilizing simulated datasets. Our analysis revealed that database-based methods excel in classification accuracy when backed by a rich and comprehensive reference database. Conversely, while machine learning methods show superior performance in scenarios where reference sequences are sparse or lacking, they generally show inferior performance compared to database methods under most conditions. Moreover, our study confirms that integrating multiple database-based methods does, in fact, enhance classification accuracy. These findings shed new light on the taxonomic classification of high-throughput sequencing data and bear substantial implications for the future development of computational biology.

Availability and Implementation: For those interested in further exploring our methods, the source code of this study is publicly available on https://github.com/LoadStar822/Genome-Classifier-Performance-Evaluator.
Additionally, a dedicated webpage showcasing our collected database, datasets, and various classification software can be found at:

http://lab.malab.cn/~tqz/project/taxonomic/.

**Keywords**: Taxonomic classification, Machine learning, Metagenomics, Comparison, Database

1. **Significance Statement**

In the burgeoning field of bioinformatics, efficiently classifying species from vast amounts of sequencing data remains a central challenge. Our research conducted an exhaustive comparison between database-based and machine learning methods for this task, unveiling their individual strengths and weaknesses. By integrating these approaches, we provide a novel strategy that enhances taxonomic classification, promising significant advances in biodiversity research and other related biological disciplines.

2. **Introduction**

As high-throughput sequencing technology continues to evolve, taxonomic classification has emerged as a cornerstone in the field of bioinformatics (Hassemer et al. 2019). Accurate taxonomic classification is not merely an academic concern; it is crucial for uncovering biological diversity, conducting disease-associated metagenome studies, and undertaking environmental surveillance (Uyaguari-Diaz et al. 2016; Nooij et al. 2018; Kim et al. 2017). Yet, efficiently processing and analyzing the sheer volume of sequencing data for precise taxonomic classification remains a daunting task (Gardiner et al. 2021).

In the landscape of existing taxonomic classification techniques, the dominant paradigms are database-based and machine learning methods. Database-based approaches primarily focus on aligning sequencing reads against a known reference genome database to determine taxonomic
classification (Bonin et al. 2023). These methods require substantial hard drive space to store the database and a significant amount of memory to load the database for use (Wright et al. 2023). In contrast, machine learning techniques endeavor to classify species by discerning patterns within the training dataset and thereby making predictions (Liang et al. 2020). Typically, these methods only need a much smaller model, making them more efficient in terms of storage and memory usage.

Each of these methodologies comes with its own set of merits and drawbacks. While database-based methods can achieve higher classification precision when the reference database is extensive and complete, they are constrained by the quality and scope of the reference database itself (Parks et al. 2020). Machine learning approaches, on the other hand, offer advantages when reference sequences are sparse or lacking, as they can extrapolate the existence of unknown species from the training data; however, their performance is often limited by the representativeness and volume of the training data (Mock et al. 2022).

In light of these complexities, our study sets out to not only compare database-based methods with machine learning approaches, but also to investigate the feasibility of integrating multiple database-based techniques. This multifaceted research aims to rigorously assess the performance and suitability of these methods using simulated datasets. Ultimately, our objective is to explore avenues for optimizing and amalgamating these techniques, with the goal of elevating the accuracy and efficiency of future taxonomic classification efforts.

2.1 Database-Based taxonomic classification method

Database-based taxonomic classification methods hold a pivotal position in bioinformatics and constitute a major trend in contemporary classification strategies (Ames et al. 2013). Utilizing reference databases, these methods determine the taxonomic affiliations of unknown sequences; their accuracy and efficiency are strongly influenced by both the comparison strategies employed and the quality of the databases themselves (Portik et al. 2022). With the surge in
high-throughput sequencing data, the task of processing and analyzing large
datasets for accurate classification has become increasingly challenging (Yang et
al. 2020; Johnson et al. 2019). To tackle this, database-based methods deploy
sophisticated algorithms and computations, serving as a potent solution. These
methods are broadly categorized into three primary types: methods based on
Alignment, methods based on Markers, and methods based on k-mer. Each
category exhibits its own unique features, possesses applicability across diverse
scenarios, and thus manifests the adaptability and richness of the field.

2.1.1 Alignment-Based Methods

Alignment-based taxonomic classification methods are pivotal in
bioinformatics, facilitating species identification by aligning unknown sequences
against known sequences in reference databases. These methods are applied
extensively across various domains, including evolutionary biology and
metagenomics. With technological advancements, sophisticated sequence
alignment techniques have been developed to enhance efficiency and
accuracy (Furstenau et al. 2022; Gao et al. 2017). These advanced methodologies
offer more precise measurements of sequence similarity, thereby improving
classification accuracy. However, the computational complexity of these methods
remains relatively high, particularly for large-scale datasets, necessitating
significant computational resources. Alignment-based methods have wide
applications in bioinformatics, and ongoing technological improvements are
expected to further enhance their efficiency and applicability for research
purposes (Zhang et al. 2024; Eisenhofer & Weyrich 2019).
2.1.2 Marker-Based Methods

In microbiomics and bioinformatics, marker-based classification methodologies leverage highly conserved marker genes or proteins within specific species or groups for species identification (Hugenholtz et al. 2021; Tovo et al. 2020). These approaches achieve precise species determination by comparing marker gene sequences from unknown samples with known sequences in reference databases, exemplified by the widespread use of the 16S rRNA gene in bacterial classification (Lan et al. 2016; Vicente Dos Santos & Tixier 2017). While these marker-based classification methods have gained recognition in the scientific community, the increasing volume and diversity of sequencing data necessitate further research and innovation to address challenges in identifying unknown or rare microbes and managing the surge in data storage and retrieval demands (Blanco-Míguez et al. 2023).

2.1.3 K-mer-based method

In bioinformatics, the taxonomic classification method based on k-mer is a vital classification technique, which complements the methods based on Alignment and Marker. In this approach, k-mer refers to a DNA sequence fragment with a length of k. Such fragments serve as feature units in species’ genomic data. By analyzing the occurrence frequency of k-mer in diverse species’ genomes, taxonomic classification is achieved (Shaw & Yu 2022). The merits of this method are its efficiency, universality, and expandability. As the count of k-mers is finite, structures like hash tables can be utilized for storing and searching k-mers, enabling swift taxonomic classification (Breitwieser et al. 2018). Additionally, the k-mer approach isn’t dependent on particular genes or genomic structures, allowing its application to any species’ genomic data. Kraken is a notable application example of the k-mer based classification method (Wood & Salzberg 2014). At the core of Kraken is a database containing k-mers and their lowest common ancestor (LCA). It quickly assigns a classification to each k-mer by matching it with the most specific node in the taxonomic tree associated with
the given k-mer. This efficient approach is also utilized by some other traditional k-mer based classification software, which similarly employs the LCA for taxonomy resolution (Menzel et al. 2016; Piro et al. 2020).

Nonetheless, the k-mer method has its restrictions. Initially, this technique presumes that every genome region equally contributes to taxonomic classification. Yet, diverse gene regions might have varied contributions (Han & Cho 2019). Next, the k-mer approach is exceptionally sensitive to sequence inaccuracies; slight errors can result in numerous k-mer mistakes (Vinje et al. 2015). Ultimately, the k-mer method can’t address intricate biological events like gene recombination or horizontal gene transfers (Dubinkina et al. 2016).

Overall, the k-mer-based taxonomic classification technique is potent and adaptable. In combination with the Alignment and Marker-based methods, it offers a holistic and precise taxonomic classification solution. Yet, this method presents its challenges and limitations. It demands additional research and enhancements to cater to varying application contexts and requirements. For a comprehensive understanding of this procedure and potential optimization tactics, we present a flowchart (Figure 1) illustrating the conventional k-mer classification process from database establishment to DNA sequence input and results, along with unique optimization techniques used by certain software.

To address the limitations of the k-mer taxonomic classification method, a variety of improvement strategies have been proposed by researchers. These are aimed at enhancing both its accuracy and efficiency. Significantly, these improvement strategies are tailored not just for the unique features of the k-mer method itself but also consider the synergistic effects when used in conjunction with methods based on Alignment and Marker genes. Below, we outline some primary directions for improvement.

Opting for specific data structures serves as one effective strategy. For example, Kraken2 has managed to increase its query speed while simultaneously reducing memory consumption by adopting compact hash tables and utilizing storage structures like minimizers (Lu & Salzberg 2020). Similarly, KrakenUniq
employs the HyperLogLog (HLL) data structure to estimate the unique number of k-mers in each classification unit, yielding significant benefits in memory efficiency, calculation speed, and accuracy (Breitwieser et al. 2018). Centrifuge leverages the FM-index data structure, based on the Burrows-Wheeler transform (BWT), enabling quick searches for k-mers of any length and thus facilitating efficient sequence categorization (Kim et al. 2016). In the case of MegaBLAST, its core feature lies in the strategic organization of database indices, coupled with the deployment of the "seed search strategy." This allows for the identification of initial matching substrings within both queries and databases (Morgulis et al. 2008). This arrangement encompasses compressed sequence information and positional data for k-mers, while also integrating unique offset and prefix management strategies. Such a setup enables MegaBLAST to efficiently process k-mers, ensuring that the identified seeds adhere to the seed qualification criteria. This innovative data structure furnishes MetaOthello with notable advantages in terms of both memory and computation, facilitating quick and accurate metagenome read classifications (Liu et al. 2018).

**Optimize k-mer selection.** For instance, CLARK’s hallmark feature is its employment of target-specific or distinguishing k-mers for categorization (Ounit et al. 2015). By constructing the target sequence’s k-spectrum and eliminating shared k-mers, CLARK opts for k-mers that uniquely and unambiguously represent each target for classification. This focused use of selective k-mers significantly enhances the precision and speed of classification. Similarly, TaxMaps’ core strategy aims to optimize k-mer selection and management through a unique compression algorithm, thereby achieving efficient species categorization (Corvelo et al. 2018). This compression strategy not only eradicates database redundancy but also streamlines the categorization process by carrying out lowest common ancestor (LCA) pre-allocation and k-mer collapsing, thereby facilitating approximate searches and improving query efficiency. Kaiju introduces a complementary method to this k-mer optimization by focusing on protein-level sequence comparison for classification (Menzel et al.
2016). Utilizing the Burrows-Wheeler transform for rapid and precise string matching, Kaiju surpasses traditional k-mer-based approaches by identifying maximum exact matches (MEMs) at the protein level, thereby enhancing recall especially for underrepresented genera in the database. This capability allows Kaiju to classify more reads with higher accuracy. Additionally, Kaiju’s greedy heuristic search mode, which permits a certain number of amino acid substitutions at the match termini, further boosts recall, making it a robust tool for sequence classification at the protein level.

**Incorporating intricate algorithms.** For example, Kraken achieves high precision and rapid classification rates by synergistically merging the k-mer approach with the LCA strategy (Wood & Salzberg 2014). k-SLAM employs k-mers for initial alignment but goes a step further by additionally integrating Smith-Waterman pairwise alignment and pseudo-assembly techniques to further boost accuracy (Ainsworth et al. 2017). These additional computational layers make k-SLAM’s algorithm more intricate compared to conventional k-mer-based methods. MMseqs2, on the other hand, takes a more comprehensive approach. It doesn’t solely rely on k-mers for matching; instead, it employs a complex strategy that involves first identifying target sequences with two sequential analogous k-mer matches on a single diagonal and then leveraging the Burrows-Wheeler transform (BWT) for exact string alignments (Steinegger & Söding 2017). These multi-step tactics make MMseqs2’s approach both more intricate and more efficient compared to traditional k-mer-centric methods. Additionally, MMseqs2 facilitates iterative profile-to-sequence and sequence-to-profile queries, adding yet another layer to its algorithmic complexity. PathSeq further contributes to this field by employing a unique classification method that begins with the removal of low-quality, low-complexity host-derived reads (Walker et al. 2018). It utilizes a Bloom filter for rapid k-mer searches to detect short sequences in the host reference sequences, effectively eliminating a significant portion of host reads before sequence alignment. The remaining non-host reads are then mapped to a microbial genome reference library using the BWA-MEM aligner,
with paired reads required to map to the same microorganism to enhance specificity. These steps, including parallel data processing, improved low-complexity sequence filtering algorithms, and efficient alignment tools, greatly accelerate analysis speed and efficiency, setting PathSeq apart from traditional methods.

**Unique k-mer approach.** For example, CLARK-S and DIAMOND, both leveraging spaced seeds, exemplify this innovation. CLARK-S employs a method called spaced k-mers, which allows for mismatches at predefined positions, thereby enhancing the tool's recall without compromising specificity (Ounit & Lonardi 2016). This approach facilitates more flexible and accurate sequence matching, especially useful in complex metagenomic datasets. Similarly, DIAMOND introduces a dual indexing and spaced seed strategy for protein sequence alignment, significantly improving upon traditional continuous seed methods. By identifying all seeds and their positions in both query and reference sequences and then linearly traversing the lists to find all matching seeds, DIAMOND enhances data locality and reduces memory bandwidth requirements. Its use of longer seeds, with only a subset of positions being utilized, strikes a balance between speed and recall, enabling faster alignments without sacrificing accuracy (Buchfink et al. 2015).

These improved strategies showcase the diversity and flexibility of k-mer taxonomic classification methods. By strategically combining diverse technologies and strategies, the precision and performance of k-mer methods can be advanced to new heights. Nevertheless, these enhancements come with new challenges and concerns that demand continued investigation and rigorous probing.

In the face of intricate and varied genomic data, k-mer taxonomic classification techniques have evolved into an essential instrument for researchers. Their high efficiency and generality have led to their extensive application in diverse settings and fields, ranging from environmental genomics to metagenomics, epidemiology, and origin tracking, to name a few (Van Etten et al. 2023). More
importantly, k-mer methods do not rely on specific genes or genomic structures, which is particularly beneficial for genomes without detailed annotations or those that are highly heterogeneous.

2.2 Machine learning-based taxonomic classification methods

In some cases, traditional classification methods can be inefficient and may struggle to recognize new species, particularly when tasked with processing massive amounts of data (Borba et al. 2021). Transitioning from this challenge, machine learning has recently emerged as a potent computational tool across various scientific and engineering domains, notably including taxonomic classification. Machine learning introduces a novel method for identifying and categorizing unknown biological entities, particularly in the fields of viral genomics and microbiomics (Bartlett et al. 2022). Employing a range of sophisticated algorithms and statistical methodologies, machine learning can adeptly distill features from intricate biological data, thereby leading to precise categorizations (Garcia et al. 2021).

In the realm of taxonomic classification, both traditional machine learning and deep learning techniques have showcased promise. Specifically, algorithms such as VirusTaxo, VirFinder, and IDTAXA in the traditional machine learning category, are complemented by others like DeepVirFinder, CHEER, and DeepMicrobes in the deep learning domain (Liang et al. 2020; Murali et al. 2018; Raju et al. 2022; Ren et al. 2017, 2020; Shang & Sun 2021). Each of these methods brings their own strengths and constraints concerning classification precision and recall, but together they collectively furnish a robust instrument for probing into biological diversity and virus taxonomy.

2.2.1 Traditional machine learning-based methods

In the domain of taxonomic classification, traditional machine learning
Methodologies play a pivotal role, surpassing deep learning techniques in terms of interpretability and computational efficiency, while often matching their performance levels (Alam & Chowdhury 2020). These methods encompass a variety of algorithms, including Naive Bayes, logistic regression, tree-based techniques, and sequence analysis, each offering distinct advantages in addressing taxonomic challenges (Song & Wang 2023).

Specifically, QIIME 2 is renowned for its q2-feature-classifier plugin, which leverages a suite of machine learning classifiers from scikit-learn to redefine taxonomic classification of marker-gene sequences into a structured document classification framework (Bokulich et al. 2018; Bolyen et al. 2019). This plugin’s adaptability allows for sophisticated customization, enabling the integration of diverse feature extraction and transformation methods, alongside the selection of an optimal machine learning algorithm. At the heart of this plugin is a method that utilizes k-mer counts from reference sequences to train a multinomial Naive Bayes classifier, a technique that notably enhances its capabilities by employing the HashingVectorizer for feature extraction. This advancement extends the k-mer length analysis from the traditional 8-mers used by the RDP Classifier to up to 32-mers, significantly surpassing the conventional classifier’s limitations (Wang et al. 2007). Additionally, this method fine-tunes class weights, allowing the classifier to move beyond uniform class prior assumptions and more accurately reflect the natural distribution of taxa in samples, thereby underscoring the balance between interpretability and computational efficiency that traditional machine learning methods, particularly Naive Bayes, bring to the field of taxonomic classification (Song & Wang 2023).

IDTAXA combines decision tree methods with bootstrap sampling to form an effective tree-based classification system. It builds decision trees to segment data into binary decisions, capturing complex relationships (Murali et al. 2018). Nodes in the tree represent dataset features, chosen for their class separation ability, which helps in segmenting the feature space. Bootstrap sampling generates multiple subsets of the original data to create an ensemble of trees, improving
accuracy and generalizability. The model is trained by selecting features and
splits iteratively, and tested on separate datasets, including simulated ones, to
evaluate performance. This approach allows IDTAXA to accurately classify
complex biological data by leveraging the ensemble of trees. However, tree-based
methods may confront challenges such as lengthy training times and recall to
parameter selection.

2.2.2 Deep Learning-based methods

As the demands for complex, detailed, and accurate classification tasks
continue to grow within the biological sciences, deep learning, with its
Convolutional Neural Networks (CNN), Long Short-Term Memory Networks
(LSTM), and other architectures, has emerged as a crucial component of the
solution repertoire (Yang et al. 2022). These methods are adept at automatically
extracting features from raw data and capturing complex nonlinearities,
significantly enhancing classification accuracy and generalizability compared to
traditional machine learning approaches. Deep learning’s application spans
genome and protein sequence analysis, as well as specific tasks such as RNA virus
classification, despite its computationally intensive training process and the need
for meticulous hyperparameter tuning, its strengths in automatic feature
learning and complex model construction establish it as a cornerstone in
advanced bioinformatics tools.

BERTax is a novel method for DNA sequence classification that relies on deep
learning and employs the BERT architecture commonly used in natural language
processing (Mock et al. 2022). What sets this method apart is its ability to
operate independently of similar sequences in databases, which makes it
outstanding for the classification of new species. By segmenting DNA sequences
into k-mers as inputs, BERTax trains on these segments to classify sequences
across diverse taxonomic levels, ranging from kingdoms to genera. Nonetheless,
BERTax may show reduced predictive power at lower taxonomic levels, such as at the genus level.

3. Results

3.1 Comparison of Classifier Performance

In the current research, we have thoroughly compared the performance of different types of classifiers, aiming to highlight performance disparities between database-based classification algorithms and machine learning classifiers across diverse biological taxonomic levels. Our particular attention was given to the database approaches at the species and genus levels because these taxonomic categories often present challenges and practical relevance in microbial identification (McIntyre et al. 2017; Sczyrba et al. 2017) (refer to Figure 2). In contrast, machine learning approaches, we particularly examined their efficacy at the genus, phylum, and superkingdom taxonomic levels (refer to Figure 3). The selection was made for two primary reasons: Firstly, machine learning-based classifiers are still evolving, and their effectiveness in fine-grained classification in microbiology, like at the species level, is yet to be improved (Liang et al. 2020; Mock et al. 2022). Secondly, given the capability of machine learning methods in dealing with large and high-dimensional data sets, we anticipate they may excel at higher taxonomic levels like phylum and superkingdom. Such a multi-tiered comparison will assist us in fully grasping the strengths and weaknesses of the different approaches.

At the species-level (refer to Figure 2a), methods based on databases demonstrate notable disparities in their performance (Sczyrba et al. 2017). Numerous software suites misclassify organisms into incorrect subspecies, severely compromising their accuracy and leading to subpar overall performance. Yet, at the genus level (refer to Figure 2b), continuously updated
methods based on databases have garnered exceptionally good results. Despite this, at the superkingdom level where they are purportedly strong (see Figure 3c), machine learning classification methods still fall short of competing with database-based approaches. However, in scenarios where species are absent from databases and not included in the training set, machine learning classification methods show a significant advantage in accuracy. A box plot (see Figure 4) now visually compares the classification precision, recall, accuracy, and false positive rate (FPR) of machine learning and database-based methods under these circumstances. Although database-based methods demonstrate exceptionally high recall, machine learning approaches, such as Bertax, generally outperform in precision, accuracy, and FPR. Notably, Bertax ranks first in overall performance metrics. This highlights the significant potential of machine learning methods in classifying species not represented in databases, showing their capability to effectively handle cases where species are unknown or new to the system. In addition, we conducted experiments with seven different sensitivity settings of Diamond, which demonstrated that adjusting sensitivity has a minimal impact on the results (see supplementary).

The Merge method, which consolidates the results from eight database-reliant software tools, stands out for achieving the highest accuracy levels at both species and genus levels in the comprehensive test dataset, as shown in Figure 2. This indicates the method’s effectiveness in integrating diverse software outputs to enhance classification precision for species well-represented within databases. However, its precision significantly diminishes when applied to sequences from species not included in the databases, as illustrated in Figure 4, where it falls below the median precision among the evaluated tools. This downturn in effectiveness can be primarily attributed to the fact that the majority of the individual software tools in the Merge ensemble also misclassify these unknown sequences. Since Merge methodologically selects the consensus or majority classification from its constituent tools, its precision is inherently limited by the accuracy of these individual tools. Consequently, when these tools collectively
misclassify an unknown sequence, the Merge method is predisposed to adopt these incorrect classifications, leading to reduced precision. This outcome highlights a fundamental limitation of the Merge approach: its dependency on the accuracy of individual database-based tools, which, in the absence of database representation for certain species, often leads to the propagation of classification errors. This scenario underscores the challenge of relying on consensus-based methods for classifying species absent from reference databases, revealing a critical area for improvement in the design and application of integrated classification strategies.

Regarding actual runtime and memory usage, a more nuanced understanding can be gained by examining Figures 5a, 5b, and 6. In Figure 5a, it is evident that machine learning classification methods often take longer in most cases. Figure 5b further corroborates that these machine learning methods tend to consume noticeably less memory compared to database-based methods. This dichotomy can be attributed to the inherent characteristics of machine learning algorithms, which often invest more time in model training, thus efficiently conserving computational resources and lowering memory usage during runtime. Consequently, in scenarios where computational resources are constrained but time is not a limiting factor, machine learning methods could offer significant advantages.

It is noteworthy that while Qiime2 appears to operate with relatively shorter runtimes as depicted in Figure 5a, the actual time consumed for data processing is significantly longer and is not reflected in the runtime calculations. On the flip side, classification methods based on databases generally show shorter runtimes (Figure 5a) but consume significantly more memory (Figure 5b). The primary reason for this is the necessity to load substantial reference databases into memory to facilitate quick searches, thereby significantly increasing memory consumption. In settings where time is of the essence but resources are abundant, these database-based methods may be more advantageous.

However, database-based classification methods also offer options to reduce
memory consumption. For instance, Kraken2 allows for a reduction in memory
usage by downsizing the database, with official low-memory versions of the
database available, albeit at the expense of a reduced number of identifiable
species. KrakenUniq introduces a technique of loading only a small portion of the
database into memory at a time for processing, thus diminishing memory
requirements. Clark has developed a lightweight variant, Clark-l, which reduces
memory demands to 4% of the original Clark requirements by compromising on
execution speed and recall. Diamond permits the adjustment of block size and
the number of seed index chunks processed, enabling a balance between memory
usage and performance. MMseqs2 reduces memory demands by segmenting the
database.

Despite these solutions offering relief in memory-constrained environments, it
is imperative to acknowledge that reducing memory consumption invariably
impacts other performance metrics, such as classification speed, accuracy, and
the breadth of identifiable species. The selection of the appropriate tool and
configuration depends not only on memory limitations but also on the specific
requirements and objectives of data processing. Therefore, users must navigate a
trade-off among memory usage, execution time, and classification capabilities,
opting for the method that best suits their particular scenario.

For a more holistic understanding, Figure 6 provides an integrated view by
plotting the log-transformed runtime against the log-transformed memory usage
for each classifier. The size and color of each bubble in Figure 6 additionally
represent the precision and recall metrics, respectively, providing a multi-faceted
evaluation of classifier performance at the genus level. This comprehensive view
aids in better understanding the trade-offs involved and could be instrumental in
the informed selection of an appropriate method based on the specific needs of a
given study.
3.2 Evaluation of Performance in Results Integration

The strategy of direct integration of classification outcomes, employed in this study, led to a substantial enhancement in the performance of taxonomic classification. By leveraging the strengths of each classification method, we achieved the best results across nearly all evaluation criteria. However, our focus on precision, particularly evident in the Merge(P) strategy, led to exceptionally high precision while slightly reducing recall and accuracy. Notably, the performance of Merge(P) at the species level is very close to that of CLARK, yet it underperforms compared to the Merge strategy at the species level and performs similarly at the genus level.

By using the integration strategy, we not only achieved more comprehensive and precise species identification but also succeeded in significantly minimizing the classification errors that could occur using a single method. Within this strategy, Merge focuses on achieving a balance between precision and recall, resulting in robust overall classification results. In contrast, Merge(P) specifically aims to maximize precision, which leads to exceptionally high precision rates but slightly lowers recall and accuracy. This contrast between the two strategies illustrates the trade-offs involved in optimizing different aspects of taxonomic classification. Despite these trade-offs, neither strategy significantly detracts from the overall performance of the integrated outcomes, demonstrating the efficacy of employing tailored approaches depending on specific classification goals.

It is worth noting that this strategy of direct result integration aims not to optimize or modify individual classification methods but rather relies on the original outcomes of each respective method; this ensures that the integration strategy maintains the unique strengths of each method while simultaneously circumventing any uncertainties that could be introduced through optimization or adjustments.

Furthermore, under this integration strategy, the classification results for some species exhibited such marked improvements that it became evident this could
be attributed to the synergistic impact of multiple methods. This synergy allows for the accurate categorization of species that are hard to classify using a single approach.

In summary, this strategy of direct result integration consistently surpassed the performance of any individual method on most evaluation criteria, thereby offering a robust and effective strategy for taxonomic classification in high-throughput sequencing data. This clearly affirms the tremendous potential of employing multiple strategies in parallel and integrating results.

4. Discussion

In this research, we have conducted a comprehensive analysis comparing database-driven classification approaches and machine learning classification approaches. A strategy for integrating various methods to enhance the accuracy of taxonomic classification was also introduced. Through this comparative lens, experimental outcomes reveal that both types of methods have specific advantages and limitations that manifest across different levels of biological taxonomy.

Classification methods relying on databases exhibit notable disparities in performance, particularly at the species and genus levels. This may stem from the incorrect categorization of organisms into the wrong subspecies, consequently lowering the overall accuracy of classification (Martínez-Porchas et al. 2016). However, at the genus level, continuously updated software employing database-based methods have yielded exceptionally favorable results. These findings not only validate the efficacy of database-driven methods but also spotlight the challenges they could encounter in species-level classification (Nørskov-Lauritsen 2014).

On the other hand, machine learning methods show their strengths in specific scenarios, such as when databases are incomplete or when classifying species
not included in the training set. Despite these advantages, they generally lag behind in terms of actual classification accuracy when benchmarked against database-based methods (Shang & Sun 2021). However, it is important to note that the merits of machine learning approaches in managing large-scale and high-dimensional data should not be overlooked.

From an operational standpoint, database-based methods often exhibit a runtime advantage but require significantly more memory than machine learning methods. This can be attributed to the necessity for these database-based methods to load extensive reference databases into memory for quick retrieval. Conversely, the significant time investment during the training phase for machine learning methods leads to effective savings in computational resources, thereby reducing memory consumption during actual operation.

Having assessed the strengths and weaknesses of both methods, we devised a new strategy for integrated classification results. The strategy employs a weighted voting mechanism to finalize the classification of each species, thereby effectively exploiting the strengths of each classification method. The experimental findings indicate that this strategy significantly enhances performance in taxonomic classification, leading to more comprehensive and accurate species identification.

It is noteworthy that tools such as FlexTaxD and MetaMeta also aim to enhance classification accuracy by integrating various methods, offering potential value in taxonomic classification. FlexTaxD improves sequence read classification accuracy and precision by modifying and amalgamating both official and custom taxonomic databases (Sundell et al. 2021). This approach is particularly applicable to studies requiring highly customized databases for identifying specific biological communities, thereby addressing complex diversity issues in taxonomy. It enhances the ability to recognize diverse biological information, proving particularly valuable in areas such as bacterial population structure analysis. Meanwhile, MetaMeta provides a comprehensive perspective on species classification by integrating the results of different metagenomic analysis
tools (Piro et al. 2017). This is especially effective in handling large-scale metagenomic data, combining the strengths of various analytical tools to offer significant insights into the discovery of new species and understanding the structure and function of microbial communities. However, despite their advantages in improving classification accuracy, both methods support a limited range of classification tools and employ integration techniques different from ours, and they are no longer being updated. Therefore, while these methods have their merits, our proposed integrated approach may be superior in terms of simplicity and effectiveness, although it too has its limitations.

First, the existing merging strategy is somewhat limited, focusing mainly on boosting recall, possibly at the cost of precision. We plan to explore additional merging strategies to enhance the precision of the classification results. Second, while machine learning methods offer certain benefits in classifying unknown species, our current integrated strategy does not yet incorporate results from machine learning approaches. As a result, adding outcomes from machine learning methods to the integration strategy is likely to boost the classification performance for unidentified species. Last, the limited number of species in the dataset used for this experiment, along with the less-than-optimal GPU performance, may affect the experiment's reliability and general applicability.

Overall, our study suggests that employing a multi-strategy parallel approach and integrating results holds significant potential in the realm of taxonomic classification in high-throughput sequencing data. While each approach has its own set of strengths and weaknesses, we find that by employing an integrated strategy, we can capitalize on the advantages of each to enhance the accuracy of taxonomic classification. In future research, we aim to further refine the integration strategies and improve current classification methods to heighten the accuracy and reliability of species identification.
5. Materials and Methods

5.1 Dataset Selection

In this study, we employed the ART-Illumina simulator to generate a comprehensive collection of 240 simulated datasets, encompassing a total of 113,282,142 sequences, to assess and compare the efficacy of various species classification methodologies (Huang et al. 2012). ART-Illumina was selected for its capability to accurately mimic the sequencing processes of Illumina platforms, such as the HiSeq 2500 and HiSeq 2000, including their characteristic error patterns. The simulated datasets encompassed 30 species validated in the NCBI and SILVA databases, representing a broad spectrum of bacteria and fungi prevalent within microbial communities (Rinke et al. 2021). It is important to note that approximately 10% of the data, corresponding to three species, were intentionally included from outside the databases to simulate real-world scenarios where not all species present in a sample are known or cataloged. The selection of these species reflects their widespread distribution across different ecosystems and the significant variance in their genomic features and ecological traits, laying the groundwork for constructing a challenging microbial community model.

A focal point of our simulation process was the application of read error models, a crucial element in evaluating the accuracy of species classification tools. The error models in ART-Illumina, derived from real sequencing data, replicate the specific error rates and types associated with the HiSeq 2500 and HiSeq 2000 platforms, such as base substitutions, insertions, and deletions. These simulated error types account for the randomness and complexity of errors encountered in actual sequencing workflows; for instance, base substitution errors may arise from imperfections in the sequencing instruments’ optical systems, whereas insertions and deletions could result from slippage during the PCR amplification process.
Regarding simulated read lengths and coverage, we established parameters ranging from 100 to 150 base pairs for read lengths and 10 to 30x for coverage, mirroring common settings in actual sequencing experiments based on the performance characteristics of the HiSeq 2500 and HiSeq 2000 platforms. Additionally, to reflect the variability in fragment lengths during library preparation, average fragment lengths were set between 200 to 400, with standard deviations (10, 25, 50) adjusted to simulate this variability. Such parameterization not only reflects the conditions of real sequencing experiments but also enhances the diversity of the simulated datasets, aiding in a comprehensive assessment of the performance of species classification tools under varied conditions.

Through this highly customized simulation process, we were able to generate simulated datasets that are both realistic and embody diverse challenges, providing a solid foundation for evaluating the accuracy, robustness, and applicability of species classification tools in practical scenarios. This approach ensures that the simulated datasets accurately reflect the myriad complexities potentially encountered in real sequencing data, thereby enabling a thorough and effective evaluation of species classification tools.

5.2 Evaluation Methods

In this study, we evaluated the performance of species classifiers using four core metrics: accuracy, precision, recall, and a modified false positive rate (FPR) (Saito & Rehmsmeier 2015). Accuracy directly reflects the proportion of species correctly identified by the classifier. Precision focuses on the classifier's ability to correctly label samples as a specific species, while recall measures the proportion of all relevant species samples that the classifier identifies (Ye et al. 2019). Given our simulated dataset, where each sequence has a target species, resulting in a true negative count of zero, we redefine FPR as the proportion of sequences incorrectly classified to non-target taxa relative to the total number of sequences. This adjustment is depicted in the confusion matrix.
(Table 2), which illustrates the distribution of true positives (TP), false positives (FP), and false negatives (FN) in our classification results. These metrics allow us to comprehensively evaluate the classifier's performance, ensuring its reliability and accuracy in species identification (McIntyre et al. 2017).

For database-dependent methods, the default operation mode seeks hits that meet the software-defined identity and alignment criteria. Sequences that do not find matches meeting these thresholds are classified as unknown. On the other hand, machine learning methods classify sequences based on whether their confidence scores are above a certain confidence threshold, again defined by the software. Although some tools allow adjusting this threshold, not every software provides such flexibility. To maintain consistency and fairness in our evaluation, we used the default settings of each classification tool, ensuring a standardized approach to assessing the performance of various methods.

5.3 Applied Taxonomic Classification Methods

For this experiment, we have chosen a range of taxonomic classification methods that are commonly used in practical applications for our assessment. These approaches can be categorized as either database-based or machine-learning-based methods, as illustrated in Table 1. To maintain the experiment's fairness and accuracy, all classifiers operated in single-thread mode using default settings, aimed at realizing the best performance for each method, and were run on a single server equipped with a Tesla K80 GPU. Moreover, while all the database methods utilized NCBI's public reference database, we modified it by removing certain species to allow machine learning methods a chance to showcase their strengths (Nasko et al. 2018).

5.4 Result Integration

To enhance the reliability and accuracy of species classification, we adopted an integration strategy that consolidates the results from multiple classification methods. The eight selected methods—Kraken2, KrakenUniq, Centrifuge, CLARK,
CLARK-S, k-SLAM, MegaBlast, and Kaiju—were chosen for their rapid processing capabilities and their proficiency in generating definitive classification results for individual sequences. Despite their varied focal points, a common feature among these methods is their reliance on sequence-level precision over phylogenetic tree structures or direct abundance outputs.

In the process of integrating these classification outcomes, we devised a weighted voting system (as illustrated in Figure 7), where each software’s result for every sequence (typically in the form of a taxonomic identifier, or taxid) is voted on based on assigned weights, with the highest-weighted outcome being selected. The allocation of weights was based on tests conducted on a separate simulated dataset, distinct from the one used in the current experiment. Although this weight distribution underwent preliminary testing and optimization, we acknowledge that it does not represent an optimal solution but rather serves as an exploratory attempt. Hence, in our description, we specifically emphasize that the weight assignment was based on "preliminary settings derived from testing on an alternate simulated dataset, indicating a provisional and exploratory nature."

Furthermore, we encourage users to customize weights according to the specific characteristics of their data and research objectives. This flexibility is crucial for accommodating diverse research contexts and aims.

Our integration strategy explores two distinct paths: one leaning towards high recall, by excluding results that were not successfully classified and retaining only those unanimously agreed upon by the methods; and another leaning towards high precision, which considers all results, including unsuccessful classifications. Notably, when the aggregated weight of the 'unknown' category is high, this approach classifies the corresponding sequence as unknown. By embracing the diversity of these strategies and providing researchers with the ability to customize weights, we aim to optimize the overall effectiveness of species classification, leveraging the strengths of each method to their fullest.
6. Data Availability

All data supporting the findings of this study are available within the article and its supplementary materials. The data sets generated and analyzed during the current study are available in the Zenodo repository at https://doi.org/10.5281/zenodo.10666087.

The data analysis scripts and code used in this study are available in the GitHub repository at https://github.com/LoadStar822/Genome-Classifier-Performance-Evaluator.

Acknowledgment

The work was supported by the National Natural Science Foundation of China (No. 62373080, No. 62131004, No. 62102269).

References


Han G-B, Cho D-H. 2019. Genome classification improvements based on k-mer intervals in


Figure 1  Flowchart of Taxonomic Classification Based on k-mer and Software Enhancement Strategies. Blue arrows illustrate the conventional k-mer classification pathway, encompassing the formulation of databases and the incorporation of DNA sequences leading to the classification outcome. Green arrows highlight the unique optimization techniques implemented by specific software, along with the software names. Panel A describes the preprocessing of the genetic sequences to be classified, including Quality Check and Sequence Trimming. Panel B showcases the construction of reference databases by species classification software, detailing unique optimization methods employed by three different software tools during database construction. Panel C discusses the querying of input sequences within the reference databases, listing the querying methods of three software tools. Panel D presents the output of classification results, which are hierarchical classifications ranging from domain to species.
Figure 2 Performance evaluation of database-based classification methods. This boxplot illustrates the Accuracy, precision, recall, and FPR values for various database-based classification approaches on a specific dataset. The "Merge" method is referred to as the result integration strategy discussed in this paper. "Merge(P)" represents our high-precision integration strategy. Accuracy represents the proportion of samples correctly identified by the classifier, precision measures the classifier's ability to correctly label a sample as a specific category, recall is the proportion of all relevant category samples that are identified by the classifier, and the false positive rate indicates the frequency at which non-target samples are incorrectly labeled as target categories. The methods are sorted by the median Accuracy to facilitate comparison.
Figure 3 Performance Comparison of Classification Methods Based on Machine Learning.

This figure presents three boxplots, each representing the performance metrics - accuracy, precision, recall, and false positive rate (FPR) - across different taxonomic levels: genus, phylum, and superkingdom.

Figure 4 shows a boxplot comparing the performance of machine learning (ML) and database-based (DB) methods at the genus level on data not included in training databases, showcasing precision, recall, accuracy, and FPR. The boxplots are sorted by the average accuracy, with different colors distinguishing the metrics for each method.
Figure 5 Running Resource Consumption of Classification Software

(A) Comparison of Total Running Time among Classifiers

(B) Comparison of Total Memory Usage among Classifiers
Figure 6 Comprehensive Performance Comparison of Various Microbiome Classifiers at the Genus Level. The x-axis and y-axis represent the log-transformed total running time in minutes and total memory usage in gigabytes (Gb), respectively. The size of each bubble indicates the precision of the classifier, while the color indicates its recall rate.
Figure 7 The Process of Integration Using the Weighted Voting Method. The figure shows how the final classification outcome is chosen from among various classification methods using weighted voting. Each classification method’s weight (W) is founded on performance indicators like accuracy, recall rate, and the like. The classification outcomes with high recall and precision are represented using two pie charts. The final classification outcomes are then tabulated below the figure.
## Tables

<table>
<thead>
<tr>
<th>Type</th>
<th>Classifier</th>
<th>Synopsis</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database-based</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kraken2</td>
<td>Employs a compact hash table for minimizers to LCAs mapping and stores only minimizers, reducing memory needs</td>
<td>(Lu &amp; Salzberg 2020)</td>
</tr>
<tr>
<td></td>
<td>KrakenUniq</td>
<td>Combines k-mer classification with unique k-mer counts using HyperLogLog, supports hierarchical searches, strain/plasmid detection, and integrates extensive viral resources</td>
<td>(Breitwieser et al. 2018)</td>
</tr>
<tr>
<td></td>
<td>Centrifuge</td>
<td>Based on Burrows-Wheeler transform and FM index for genome sequence storage and search, reduces index through compression and redundancy removal, supports variable-length k-mer search</td>
<td>(Kim et al. 2016)</td>
</tr>
<tr>
<td></td>
<td>CLARK</td>
<td>Multilevel short metagenomic read classification, scalable on multi-core, low operational requirements, provides confidence scores, supports BAC, transcript, and centromeric region inference</td>
<td>(Ounit et al. 2015)</td>
</tr>
<tr>
<td></td>
<td>CLARK-S</td>
<td>Allows limited mismatches in k-mers to improve recall, introduces discriminative spaced k-mers for precision</td>
<td>(Ounit &amp; Lonardi 2016)</td>
</tr>
<tr>
<td></td>
<td>k-SLAM</td>
<td>Conducts full sequence alignment for gene and variant identification, differing from rapid classifiers that rely solely on k-mers; incorporates pseudo-assembly to enhance accuracy, notably for species with high intra-genus homology</td>
<td>(Ainsworth et al. 2017)</td>
</tr>
<tr>
<td></td>
<td>MegaBLAST</td>
<td>Employs genome-database-derived indexing for seed finding, utilizes WindowMasker for soft-masking of repeats, with performance gradually declining in cases of long or highly matched queries</td>
<td>(Morgulis et al. 2008)</td>
</tr>
<tr>
<td></td>
<td>PathSeq</td>
<td>Subtracts human sequences, aligns microbial sequences to detect known/unknown pathogens and resident microorganisms; supports any dataset size in parallel computing and cloud environments</td>
<td>(Walker et al. 2018)</td>
</tr>
<tr>
<td></td>
<td>taxMaps</td>
<td>Maps short-read data taxonomically, reduces redundancy and improves query performance with database compression, uses GEM for non-exact match mapping, achieving similar accuracy to BLASTN with reduced time</td>
<td>(Corvelo et al. 2018)</td>
</tr>
<tr>
<td></td>
<td>DIAMOND</td>
<td>Supports ultra-fast large-scale protein searches suitable for extensive datasets; utilizes double indexing and multiple spaced seeds for enhanced specificity, optimized by high-performance and cloud computing</td>
<td>(Buchfink et al. 2015)</td>
</tr>
<tr>
<td></td>
<td>Kaiju</td>
<td>Translates metagenomic reads into reading frames for maximum exact matches in a protein database, utilizing Burrows-Wheeler transform for rapid string matching, and employs sparse FM-index and adjustable suffix arrays to minimize memory usage</td>
<td>(Menzel et al. 2016)</td>
</tr>
<tr>
<td></td>
<td>MMseqs2</td>
<td>Parallelized for iterative profile-to-sequence and sequence-to-profile searches, capable of handling</td>
<td>(Steinegger &amp; Söding)</td>
</tr>
</tbody>
</table>
large datasets, offers clustering to reduce database redundancy, saving storage and computational resources (2017).

<table>
<thead>
<tr>
<th>Tool</th>
<th>Description</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bertax</td>
<td>Uses deep neural network and natural language processing to classify DNA sequences' superkingdom and phylum without database relatives; excels with novel organisms and can classify any genome region (Mock et al. 2022)</td>
<td></td>
</tr>
<tr>
<td>IDTAXA</td>
<td>Uses machine learning to reduce overclassification and enhance accuracy, fewer overclassifications with missing data, suitable for varied sequence lengths, auto-corrects errors in reference taxonomies, part of R's DECIPHER package (Murali et al. 2018)</td>
<td></td>
</tr>
<tr>
<td>Machine Learning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QIIME 2</td>
<td>Integrates various feature classifiers and databases for species annotation, supports complex statistical analyses for biodiversity exploration, offers data visualization tools for result presentation, and has a plugin architecture for easy extension of new features (Bolyen et al. 2019)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Confusion matrix for DNA sequence classification task using simulated data where each sequence has a target species, resulting in zero true negatives (TN). The table categorizes DNA sequences as true positives (TP), false positives (FP), and false negatives (FN) based on their actual and predicted classifications.

<table>
<thead>
<tr>
<th>Actual Target Species (P)</th>
<th>Predicted as Target Species (PP)</th>
<th>Predicted as Non-Target Species (PN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive (TP): DNA sequence correctly classified as its species</td>
<td>False Negative (FN): DNA sequence failed to be classified</td>
<td></td>
</tr>
<tr>
<td>Actual Non-Target Species (N)</td>
<td>False Positive (FP): DNA sequence misclassified as another species</td>
<td>True Negative (TN): DNA sequence correctly classified as non-target species</td>
</tr>
</tbody>
</table>