ISSN 0973-2063 (online) 0973-8894 (print)

Bioinformation 20(9): 986-989 (2024)

BIOINFORMATION
Discovery at the interface of physical and biological sciences

BIOMEDICAL INFORMATICS

www.bioinformation.net

Volume 20(9)

Research Article

Received September 1, 2024; Revised September 30, 2024; Accepted September 30, 2024, Published September 30, 2024

DOI: 10.6026/973206300200986

BIOINFORMATION 2022 Impact Factor (2023 release) is 1.9.

Declaration on Publication Ethics:
The author's state that they adhere with COPE guidelines on publishing ethics as described elsewhere at https://publicationethics.org/. The authors also undertake that they are not associated with any other third party (governmental or non-governmental agencies) linking with any form of unethical issues connecting to this publication. The authors also declare that they are not withholding any information that is misleading to the publisher in regard to this article.

Declaration on official E-mail:
The corresponding author declares that lifetime official e-mail from their institution is not available for all authors

Comments from readers:
Articles published in BIOINFORMATION are open for relevant post publication comments and criticisms, which will be published immediately linking to the original article without open access charges. Comments should be concise, coherent and critical in less than 1000 words.

Disclaimer:
The views and opinions expressed are those of the author(s) and do not reflect the views or opinions of Bioinformation and (or) its publisher Biomedical Informatics. Biomedical Informatics remains neutral and allows authors to specify their address and affiliation details including territory where required. Bioinformation provides a platform for scholarly communication of data and information to create knowledge in the Biological/Biomedical domain.

Edited by P Kangueane
Citation: Sangar *et al.* Bioinformation 20(9): 986-989 (2024)

Species annotation using a *k-mer* based KNN model

Srushti Sangar[1], Prathamesh Kolage[2] & Pritee Chunarkar-Patil[3,*]

[1]Department of Bioinformatics, Rajiv Gandhi Institute of IT and Biotechnology, Bharati Vidyapeeth (Deemed to be University), Pune, Maharashtra, India; *Corresponding author

Institute URL:
https://www.bvuniversity.edu.in/rgitbt/

Author contacts:
Srushti Sangar-E-mail: srushti.sangar-rgitbt@bvp.edu.in; Phone: +91 9325122998
Prathamesh Kolage-E-mail: prathamesh.kolage-rgitbt@bvp.edu.in; Phone: +91 9370215432
Pritee Chunarkar-Patil-E-mail: preeti.chunarkar@bharatividyapeeth.edu; Phone: +91 9730038142

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 20(9): 986-989 (2024)

Abstract:
Bacterial identification is a critical process in microbiology, clinical diagnostics, environmental monitoring, and food safety. Machine learning holds great promise for improving bacterial identification by increasing accuracy, speed, and scalability. However, challenges such as data dependency, model interpretability, and computational demands must be addressed to fully realize it's potential. *k-mer* based bacterial identification algorithm is an attempt to address these issues. Sequence matching is completed using the KNN technique. This included feature extraction, dataset preparation, classifier training, and label prediction based on k-mer frequency distribution similarity. The algorithm's performance has been cross-checked through accuracy assessment metrics such as F1 score and precision with an impressive 93% accuracy rate.

Availability: The tool is available at https://github.com/prathamesh21575/K-mer-based-Bacterial-Identification-KNN-Approach

Keywords: k-mer, bacterial identification, sequence comparison, KNN classification & bio pytho

Background:
The annotation and identification of genomic regulatory elements, such as enhancers, splice sites, transcription start sites, and promoters, as well as the classification of various phenotypes, are just a few of the genomics and bioinformatics issues that have been extensively addressed by machine learning algorithms **[1]**. Utilizing the 16S rRNA gene as a pattern-based computational tool for taxonomy classification from the phylum Firmicutes down to the genus Bacillus, DNA Barcode Identification (DNA Bar ID) has recently generated a lot of interest in the use of k-mer-based methods to predict the phenotypic features of bacteria. The researchers mapped the patterns into several hyperactive variable areas of the 16S rRNA gene, with V3–V4 being one of the most highly variable regions. The produced signatures displayed good sensitivity and specificity when compared **[2]**. K-mer-based models may be easier to interpret if techniques for decreasing redundancy and collecting genomic context are employed. Evaluation of prediction models constructed with sparse machine learning techniques, such as decision trees or lasso-regression, can be more difficult when bacterial genomes are represented using k-mers. This is because, although there might be more linked features that are equally predictive, these algorithms usually select a random subset of the connected features **[3]**. Bio Seq-Analysis is a robust platform for biological sequence analysis that leverages machine-learning techniques **[4]**. It streamlines feature extraction, predictor development, and performance evaluation, automates prediction creation while allowing users to contribute benchmark datasets, and outperforms some state-of-the-art methods in sequence analysis tasks. Virus sequences may now be identified from prokaryotic metagenomics data thanks to tools such as VirFinder, which improve and supplement gene-based methods for viral sequence categorization. These methods provide an effective way to find new viruses that might not share gene sequences by utilizing virus-specific k-mer patterns. The efficient and dependable classification of viral sequences over a broad variety of host domains and phyla is made possible by the use of k-mer patterns in place of gene-centric approaches for viral sequence identification **[5]**. K-mers are brief segments of a predetermined length (k) that are taken out of DNA sequences **[6]**. Many methods of sequence classification have been proposed, with the goal of improving BLAST accuracy with machine learning and sequence matching algorithms. The MEGAN program searches a sequence (using BLAST) against many databases **[7]**. The lowest common ancestor (LCA) of the best matches discovered in each database is assigned to the sequence. To achieve higher accuracy than BLAST on its own, PhymmBL **[8, 9]** combines BLAST results with scores derived from interpolated Markov models. The Naïve Bayes Classifier (NBC) applies a Bayesian rule to the distributions of k-mers inside a genome **[10]**. Therefore, it is of interest to describe a *k-mer* based model using KNN.

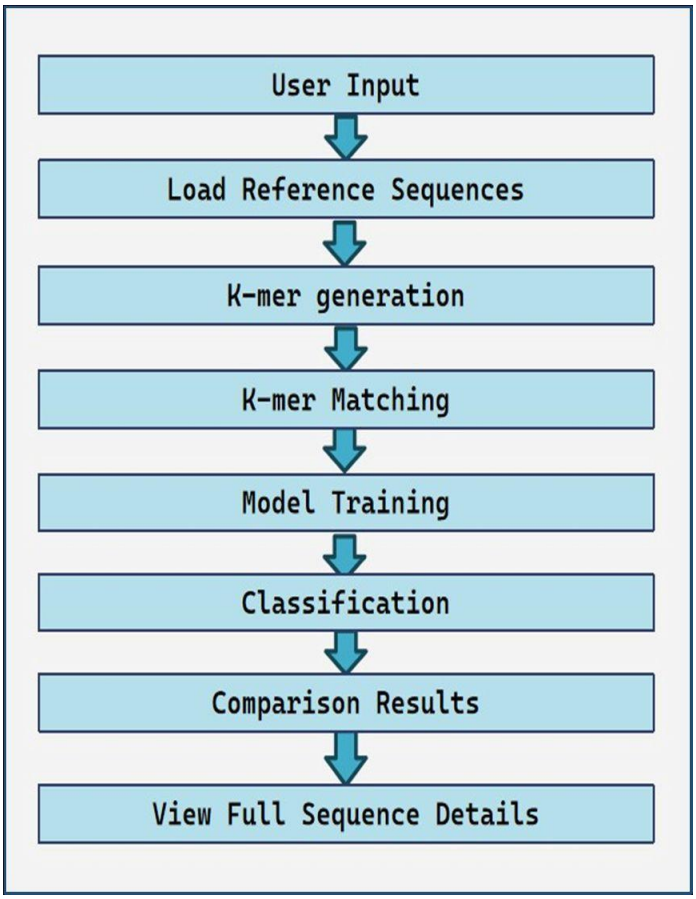


Figure 1: Flowchart of workflow of the tool

Materials and Methodology:

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 20(9): 986-989 (2024)

Through the integration of KNN algorithm and k-mer analysis, the tool facilitates fast and precise DNA sequence comparison for the identification of bacteria. This method increases the efficacy and efficiency of sequence matching in a variety of applications by utilizing both the natural properties of DNA sequences and the capabilities of machine learning. This program created to using the scikit-learn Tkinter, and bio-python libraries. The modular design allows for easy maintenance, scalability, and future enhancements.

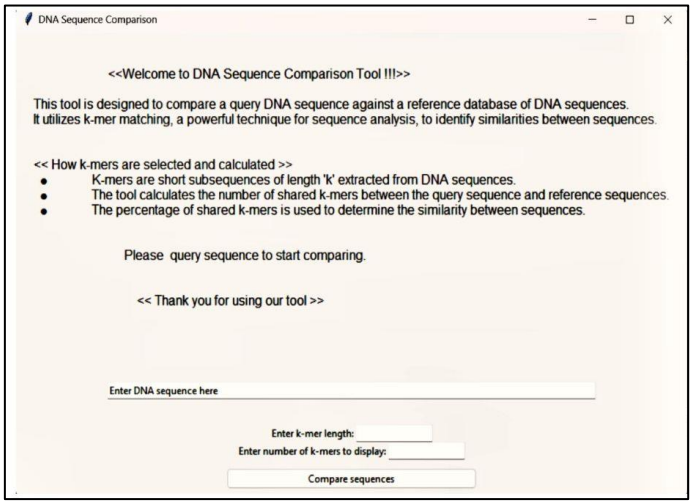


Figure 2: GUI of Tool

Data gathering and pre-processing:
DNA sequences were collected from various sources in order to build our reference database. These sequences span many organisms and genetic regions to ensure robustness in sequence comparison. We used the Biopython package to process FASTA files in an effective manner.

Sequence comparison algorithm:
Our methodology is based on k-mer matching, a popular bioinformatics technique for sequence analysis. K-mers are short sub sequences of length 'k' that are extracted from DNA sequences. For each query sequence, we calculate the number of shared k-mers between the query and reference sequences.

K-mer calculation:
A technique has been developed to extract k-mers from DNA sequences with efficiency. The function returns every possible k-mer of length 'k' given a sequence. These k-mers serve as the basis for the comparison between the reference and query sequences.

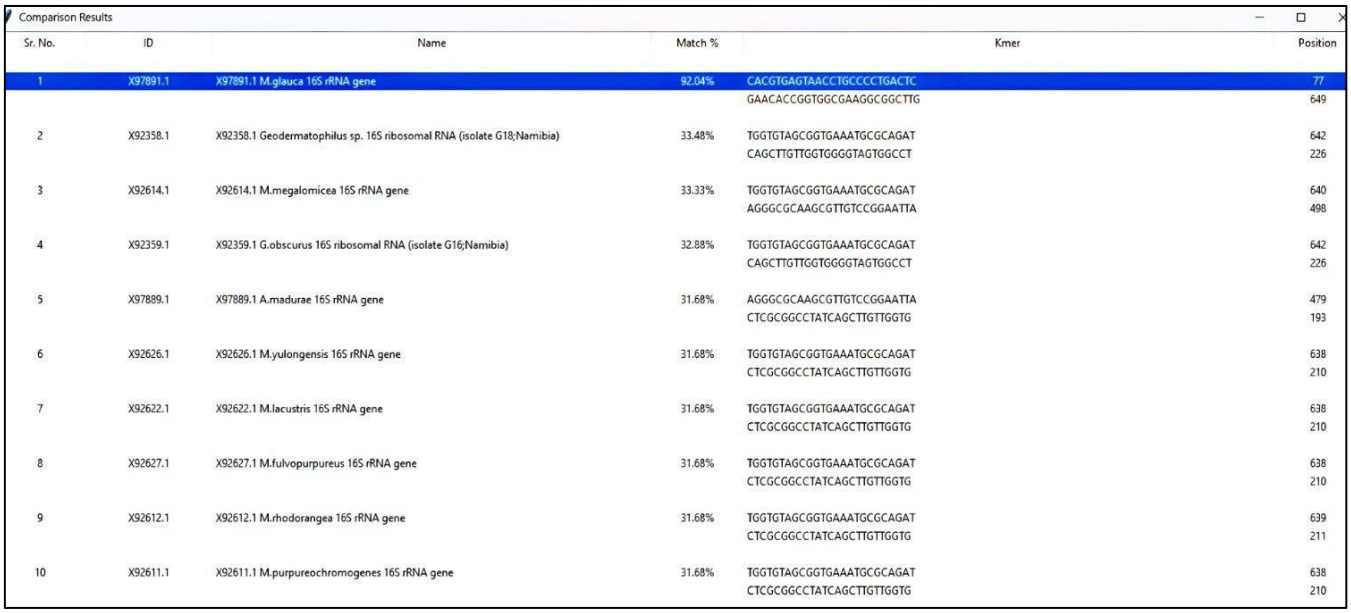
Similarity calculation:
The degree of similarity between each reference sequence and the query sequence was measured by the fraction of shared k-mers. This percentage indicates how similar the sequence is to one another. Greater percentages imply greater similarity.

User interface design:
An easy-to-use graphical user interface created with the Python Tkinter module. Users can input their query sequences, start the comparison process, and adjust settings like the length of the k-mer and the number of k-mers to display with only one click. The complete workflow of the tool is shown in Figure 1.

Results:
The KNN method was selected for sequence matching in our model due to its resilience, simplicity, and efficiency in classification tasks. KNN is non-parametric and depends only on feature vector similarity, in contrast to parametric approaches that assume certain aspects of data distributions. This makes it especially appropriate for this study objective, which is to categorize query sequences according to how similar they are to reference sequences. *k-mer*-based prediction model provides a strong and effective method for identifying bacteria through the integration of KNN algorithm and k-mer analysis. This tool offers a flexible platform for quick and precise DNA sequence comparison. Users are able to change the number of matching k-mers that will be shown in the results as well as the k-mer length (k). By customizing the analysis parameters to the unique properties of their data and research goals, researchers can improve the relevance and usefulness of their findings. The homepage of the tools is shown (Figure 2). The tool will take sequence from the user and show the best 10 hits from the database (Figure 3). User can access the sequence through cross-link of accession number (Figure 4). The tool shows the percent match and position of the matched k-mer. The astounding 93% accuracy rate of our system was validated by metrics like precision and F1 score. Especially, considering the increasing application of machine learning in genomics and bioinformatics, our k-mer based prediction model stands out for its computational efficiency compared to traditional alignment approaches. Our model swiftly and precisely detects bacterial species using k-mer matching from DNA sequences. Python modules like Biopython and collections power it.

Discussion:
Current study significantly enhances our understanding of k-mer-based prediction models in bioinformatics and genomics. It demonstrates that k-mer matching is an effective method for classifying DNA sequences, achieving a 93% accuracy rate in this prediction model. Unlike popular programs like Kraken and Mash, which focus on taxonomic classification or k-mer counting, current study approach prioritizes direct sequence comparison, offering researchers a novel paradigm **[3, 11]**. The model's flexibility allows users to create custom reference databases, making the analysis more relevant to specific research questions. This adaptability sets this method apart from technologies that rely on large, pre-built databases, enabling its application to a broader range of datasets. Our program also employs exact k-mer matching, which is crucial for accurate genomic research, providing detailed assessments of sequence similarities, including specific matching k-mers and their positions **[12, 13].** Additionally, the user-friendly interface, developed with Tkinter, enhances accessibility and allows for interactive visualization of results, promoting the wider

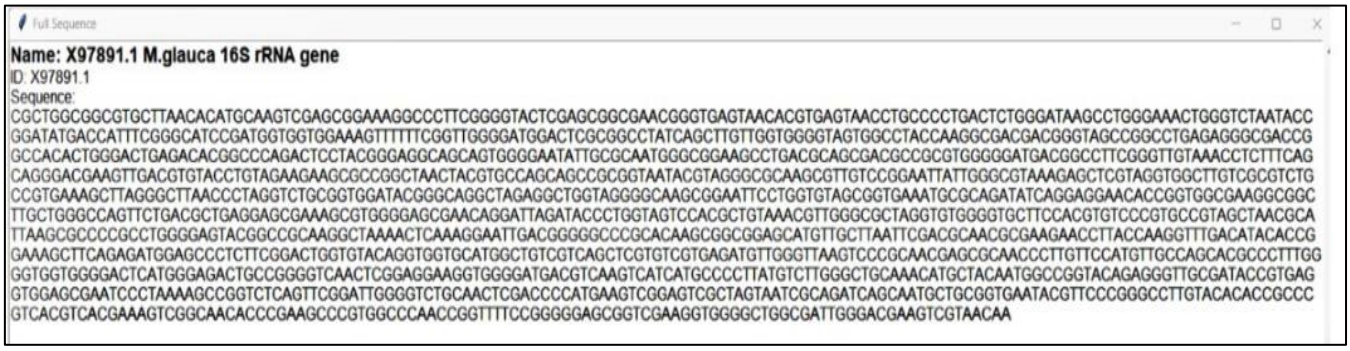
adoption of advanced genetic analysis techniques and democratizing the use of bioinformatics tools.



Comparison Results

Sr. No.	ID	Name	Match %	Kmer	Position
1	X97891.1	X97891.1 M.glauca 16S rRNA gene	92.04%	CACGTGAGTAACCTGCCCCTGACTC	77
				GAACACCGGTGGCGAAGGCGGCTTG	649
2	X92358.1	X92358.1 Geodermatophilus sp. 16S ribosomal RNA (isolate G18;Namibia)	33.48%	TGGTGTAGCGGTGAAATGCGCAGAT	642
				CAGCTTGTTGGTGGGGTAGTGGCCT	226
3	X92614.1	X92614.1 M.megalomicea 16S rRNA gene	33.33%	TGGTGTAGCGGTGAAATGCGCAGAT	640
				AGGGCGCAAGCGTTGTCCGGAATTA	498
4	X92359.1	X92359.1 G.obscurus 16S ribosomal RNA (isolate G16;Namibia)	32.88%	TGGTGTAGCGGTGAAATGCGCAGAT	642
				CAGCTTGTTGGTGGGGTAGTGGCCT	226
5	X97889.1	X97889.1 A.madurae 16S rRNA gene	31.68%	AGGGCGCAAGCGTTGTCCGGAATTA	479
				CTCGCGGCCTATCAGCTTGTTGGTG	193
6	X92626.1	X92626.1 M.yulongensis 16S rRNA gene	31.68%	TGGTGTAGCGGTGAAATGCGCAGAT	638
				CTCGCGGCCTATCAGCTTGTTGGTG	210
7	X92622.1	X92622.1 M.lacustris 16S rRNA gene	31.68%	TGGTGTAGCGGTGAAATGCGCAGAT	638
				CTCGCGGCCTATCAGCTTGTTGGTG	210
8	X92627.1	X92627.1 M.fulvopurpureus 16S rRNA gene	31.68%	TGGTGTAGCGGTGAAATGCGCAGAT	638
				CTCGCGGCCTATCAGCTTGTTGGTG	210
9	X92612.1	X92612.1 M.rhodorangea 16S rRNA gene	31.68%	TGGTGTAGCGGTGAAATGCGCAGAT	639
				CTCGCGGCCTATCAGCTTGTTGGTG	211
10	X92611.1	X92611.1 M.purpureochromogenes 16S rRNA gene	31.68%	TGGTGTAGCGGTGAAATGCGCAGAT	638
				CTCGCGGCCTATCAGCTTGTTGGTG	210

Figure 3: Results showing top 10 hits



Full Sequence

Name: X97891.1 M.glauca 16S rRNA gene

ID: X97891.1

Sequence:

CGCTGGCGGCGTGCTTAACACATGCAAGTCGAGCGGAAAGGCCCTTCGGGGTACTCGAGCGGCGAACGGGTGAGTAACACGTGAGTAACCTGCCCCTGACTCTGGGATAAGCCTGGGAAACTGGGTCTAATACC
GGATATGACCATTTCGGGCATCCGATGGTGGTGGAAAGTTTTTTCGGTTGGGGATGGACTCGCGGCCTATCAGCTTGTTGGTGGGGTAGTGGCCTACCAAGGCGACGACGGGTAGCCGGCCTGAGAGGGCGACCG
GCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGCGCAATGGGCGGAAGCCTGACGCAGCGACGCCGCGTGGGGGATGACGGCCTTCGGGTTGTAAACCTCTTTCAG
CAGGGACGAAGTTGACGTGTACCTGTAGAAGAAGCGCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGGCGCAAGCGTTGTCCGGAATTATTGGGCGTAAAGAGCTCGTAGGTGGCTTGTCGCGTCTG
CCGTGAAAGCTTAGGGCTTAACCCTAGGTCTGCGGTGGATACGGGCAGGCTAGAGGCTGGTAGGGGCAAGCGGAATTCCTGGTGTAGCGGTGAAATGCGCAGATATCAGGAGGAACACCGGTGGCGAAGGCGGC
TTGCTGGGCCAGTTCTGACGCTGAGGAGCGAAAGCGTGGGGAGCGAACAGGATTAGATACCCTGGTAGTCCACGCTGTAAACGTTGGGCGCTAGGTGTGGGGTGCTTCCACGTGTCCCGTGCCGTAGCTAACGCA
TTAAGCGCCCCGCCTGGGGAGTACGGCCGCAAGGCTAAAACTCAAAGGAATTGACGGGGGCCCGCACAAGCGGCGGAGCATGTTGCTTAATTCGACGCAACGCGAAGAACCTTACCAAGGTTTGACATACACCG
GAAAGCTTCAGAGATGGAGCCCTCTTCGGACTGGTGTACAGGTGGTGCATGGCTGTCGTCAGCTCGTGTCGTGAGATGTTGGGTTAAGTCCCGCAACGAGCGCAACCCTTGTTCCATGTTGCCAGCACGCCCTTTGG
GGTGGTGGGGACTCATGGGAGACTGCCGGGGTCAACTCGGAGGAAGGTGGGGATGACGTCAAGTCATCATGCCCCTTATGTCTTGGGCTGCAAACATGCTACAATGGCCGGTACAGAGGGTTGCGATACCGTGAG
GTGGAGCGAATCCCTAAAAGCCGGTCTCAGTTCGGATTGGGGTCTGCAACTCGACCCCATGAAGTCGGAGTCGCTAGTAATCGCAGATCAGCAATGCTGCGGTGAATACGTTCCCGGGCCTTGTACACACCGCCC
GTCACGTCACGAAAGTCGGCAACACCCGAAGCCCGTGGCCCAACCGGTTTTCCGGGGGAGCGGTCGAAGGTGGGGCTGGCGATTGGGACGAAGTCGTAACAA

Figure 4: Sequence for one of the hits is displayed

Conclusion:

The k-mer-based prediction model represents a significant advancement in sequence comparison techniques, opening new possibilities for its application in various biological contexts **[14]**. The tool is having 93% accuracy, 96% positive predictive value, 91% sensitivity, 96% specificity and 90% negative predictive value (NPV). This study addresses current challenges and provides practical solutions for researchers, contributing valuable insights to genomics.

References:

[1] Libbrecht MW & Nobel SW, *Nat Rev Genet.* 2015 **16:**321. [PMID: 25948244].
[2] More RP & Purohit HJ, *Journal of Computational Biology* 2016 **23:**651. [PMID: 27104769].
[3] Jaillard M *et al. Giga Science* 2020 **9:**giaa110. [PMID: 33068113].
[4] Liu B, *Briefings in Bioinformatics* 2019 **20:**1280. [PMID: 29272359].
[5] Bussi Y *et al. PLoS One* 2021 **16:** e0258693. [PMID: 34648558].
[6] https://en.wikipedia.org/wiki/K-mer
[7] Huson DH *et al. Genome Res.* 2007 **17:**377. [PMID: 17255551]
[8] Brady A & Salzberg SL, *Nat Methods* 2009 **6:**673. [PMID: 19648916].
[9] Brady A & Salzberg S, *Nat Methods* 2011 **8:**367 [PMID: 21527926].
[10] Rosen G *et al. Advances in Bioinformatics* 2008 **2008**:205969. [PMID: 19956701].
[11] Erki A *et al. PLoS Comput Biol* 2018 **14**:e1006434. [PMID: 30346947]
[12] Lorenzi C *et al. Genome Biol* 2020 **21**:261. [PMID: 33050927]
[13] Aylward AJ *et al. Bioinformatics* 2023 **39**:btad621. [PMID: 37846049]
[14] Karikari B *et al.* Genes (Basel) 2023 **14**:1439. [PMID: 37510343]