PathoVF: A Dual-Approach Tool for Predicting Pathogenic and Virulence Factors Using Multi-Modal DNA Encoding and User-Friendly Interface

Abstract

The development of machine learning and deep learning tools for pathogenic detection has revolutionized the ability to analyze biological and environmental samples. Despite significant advancements, current tools are limited by their inability to process both long and short read DNA sequences, reliance on simplistic DNA encoding methods, lack of functionality for predicting pathogen properties such as virulence factors and non-user-friendly interfaces. Here, we introduce PathoVF, a robust dual-approach tool that addresses these challenges by incorporating multi-modal DNA encoding strategies and a user-friendly interface. PathoVF is based on both long and short DNA reads, utilizes various DNA encoding methods including an ensemble approach, and provides a novel classification system for virulence factors. Our tool features a straightforward web server interface that requires no programming skills, allowing for broad usability among biological researchers. The long-read model, based on a machine learning algorithm using random forests, achieves 91% accuracy on our test dataset and 86% accuracy on external metagenome-assembled genomes (MAG) without re-training, demonstrating its effectiveness and efficiency.

Introduction

The detection and characterization of pathogenic organisms are crucial in understanding infectious diseases and improving public health interventions. Modern advancements in bioinformatics with the use of machine learning and deep learning technology have greatly enhanced the capabilities for pathogen detection in both clinical and environmental settings. However, the application of these tools is often limited by challenges such as the need for specific types of DNA read lengths, simplistic and insufficient DNA encoding strategies, and the complexity of operational requirements. Additionally, many tools do not support the prediction of critical pathogen characteristics, such as virulence factors, which are essential for understanding the pathogenicity and potential impact of organisms.

From these challenges, we developed PathoVF, a tool designed to overcome the limitations of existing pathogenic identification tools. PathoVF integrates both long and short DNA read analysis capabilities and employs a variety of sophisticated DNA encoding methods. Moreover, it extends its functionality to predict virulence factors, offering insights into the pathogenic potential of detected organisms. Importantly, PathoVF is developed with a focus on user accessibility, featuring a straightforward web-based interface that requires no programming skills, thus democratizing the use of advanced bioinformatics tools. This paper details the development, functionality, and validation of PathoVF, demonstrating its effectiveness through rigorous testing and its potential to transform the landscape of pathogen detection and analysis.

Methods

Dataset Collection

For our study, we utilized the BacRefSeq dataset collected by DCiPatho. This dataset comprises 32,927 complete bacterial genomes, sourced from the Reference Sequence Database provided by the National Center for Biotechnology Information (NCBI) as of June 2022. Each genome was labeled as pathogenic or nonpathogenic based on comprehensive data curation from multiple sources. To enhance the dataset's usability for our analyses, we downloaded the genomes and manually removed any duplicated entries, resulting in 29,110 unique genomes. For virulence factor classification, we collected 27,982 virulence bacterial pathogens complete protein sequences from VFDB.

Data Processing-Genome Processing

The genomes were initially segmented into approximately 1.1 billion short reads, each 150 base pairs (bps) long with an overlapping region of 40 bps to ensure continuity and accuracy in assembly. From this extensive collection, we implemented a random selection process to manage the dataset efficiently. Specifically, we extracted 1,000 representative short reads from each complete genome, culminating in a streamlined compilation of around 2 million short reads in total. This reduction balanced computational feasibility with the integrity of our dataset for subsequent analyses.

Data Processing-External Test Data

Additionally, we incorporated external test data assessed by DCiPatho, which included metagenome-assembled genomes (MAGs) from the gut microbiome of hospitalized adults. These MAGs were pre-filtered using CheckM to ensure high quality (completeness > 90% and contamination < 5%).

Encoding Techniques

For long read data, we focused on k-mer frequency analysis with k-values of 3, 4, and 5 due to the extensive length of each sequence, which makes traditional encoding methods like one-hot encoding impractical. In contrast, the encoding for short read data was more diverse, including methods such as GC Content, Shannon Entropy, Nucleotide Frequency, Binary Encoding, Fourier Transform, Kmer Frequencies, Dinucleotide Properties, Positional Nucleotide Frequencies, Mismatch Profile, and Nucleotide Auto Covariance. These methods were chosen to capture a broader spectrum of genetic information and to suit the varied nature of short DNA reads.

Method	Description
GC Content	Measures the percentage of guanine (G) and cytosine (C) bases in a DNA sequence, important for understanding different properties of GC-rich regions.
Shannon Entropy	Quantifies the uncertainty or randomness in the nucleotide distribution of a DNA sequence, with higher entropy indicating more complexity.
Nucleotide Frequency	Calculates the proportions of each type of nucleotide (A, T, C, G) in a DNA sequence, aiding in understanding the basic composition.
Binary Encoding	Represents DNA sequences as binary data, encoding each nucleotide type as a unique binary string for computational models.
Fourier Transform	Applies Fourier transform techniques to identify periodic patterns and hidden structures in DNA sequences.
Kmer Frequencies	Counts the occurrences of kmers (subsequences of length k) to identify common motifs and regulatory elements in DNA sequences.
Dinucleotide Properties	Studies properties of dinucleotide pairs, focusing on their chemical and physical characteristics that can influence DNA structure.
Positional Nucleotide Frequencies	Analyzes the frequency of each nucleotide type at specific positions within a DNA sequence, revealing positional biases.

Kmer Type 1	Focuses on the analysis of specific types of kmers according to predefined criteria, potentially involving complex patterns.
Mismatch Profile	Examines the distribution and frequency of mismatches between two DNA sequences, useful in comparative genomics and mutation analysis.
Nucleotide Covariance	Analyzes the covariance of nucleotides at different positions to identify correlations indicating structured regions in DNA.
PSSM	A statistical representation that scores the likelihood of each nucleotide at each position based on known biological data, used in motif analysis.

Machine Learning Algorithms

Our approach compared several machine learning algorithms to determine the most effective models for our dual analysis focus. We employed Logistic Regression, Random Forest, Support Vector Machines, K-Nearest Neighbors, Decision Trees, Gaussian Naive Bayes, and XGBoost. Each algorithm was trained on a split of 90% of our data, with the remaining 10% used for testing, ensuring balanced representation and rigorous evaluation of both pathogenic and nonpathogenic samples.

Web-based development

This server architecture is distinguished by its simplicity and efficiency. Given that our application does not require the storage of temporary or user-specific data, we decided not to implement a backend server. This decision not only streamlines the system, reducing potential overhead and maintenance concerns, but also enhances the application's performance, rendering it as a lightweight and powerful tool for genomic analysis in identification of pathogen and virulence factor.

Result

Model Performance on Controlled Datasets

Our analysis began with testing the models on controlled datasets comprising both long and short DNA reads. The Random Forest model demonstrated exceptional performance, achieving an accuracy of 91% on long-read sequences and 86% accuracy on external MAG datasets. This was

significantly higher than other models, which underscored its robustness in handling complex genomic data. For short reads, the models incorporating advanced DNA encoding techniques, such as Fourier Transform and Nucleotide Auto Covariance, also showed promising results, providing deeper insights into sequence characteristics.

External Validation with MAGs

The external validation on MAGs was crucial in testing the practical applicability of PathoVF. Here, the Random Forest model again stood out, with precision, recall, and F1 scores that outperformed other models, particularly in distinguishing between pathogenic and nonpathogenic strains. The decision tree and K-Nearest Neighbors models also performed well, indicating that ensemble and instance-based learning approaches are beneficial in genomic analysis.

Model	Accuracy	Precision	Recall	F1 Score
decision_tree	0.729483	0.814922	0.729483	0.751661
gaussian_naive_bayes	0.437690	0.736746	0.437690	0.461832
knn	0.782675	0.848930	0.782675	0.799237
logistic_regression	0.215805	0.831703	0.215805	0.078759
svm	0.586626	0.737981	0.586626	0.623654
random_forest	0.860182	0.886265	0.860182	0.867364
xgboost	0.471125	0.771562	0.471125	0.496371

Short Read Comparisons

In our analysis of short read data, we evaluated the performance of several DNA encoding methods using a standard machine learning model. The results, summarized in the table below, indicate that none of the encoding techniques achieved satisfactory performance levels, with all metrics generally falling below expectations. Here are the key findings:

Overall Performance

The highest accuracy observed was 0.51095 for the Fourier Transform method, which, like most other methods, performed only marginally better than random guessing. The precision, recall, and F1 scores for all methods were also generally low, indicating a lack of predictive reliability across the board.

Comparison of Encoding Methods

Methods like GC Content and Shannon Entropy, which are relatively simple, provided slightly better results compared to more complex methods like Binary Encoding and Kmer Frequencies, yet still did not reach a level of performance that would be considered effective.

Advanced methods such as Fourier Transform and Nucleotide Auto Covariance showed some potential by achieving the highest scores among the tested methods but still fell short of achieving practical applicability.

The Binary Encoding method, despite being one of the more complex approaches with a high feature dimensionality, did not translate into better performance, indicating possible overfitting or inefficiency in capturing useful predictive signals from the data.

Implications of Short Read Performance

The suboptimal performance of the encoding methods on short read data suggests significant challenges in the model's ability to discern meaningful patterns from such highly fragmented information. This outcome highlights the need for further refinement in our approach, potentially through the development of more sophisticated machine learning algorithms, optimization of encoding techniques, or a deeper integration of contextual genomic data that might enhance predictive accuracy.

Model	Accuracy	Precision	Recall	F1 Score	Feature_dim
GC_Content	0.48285	0.479593	0.48285	0.461358	1.0
Shannon_Entropy	0.51010	0.510105	0.51010	0.510041	1.0
Nucleotide_Frequency	0.49030	0.490177	0.49030	0.488702	4.0
Binary_Encoding	0.50135	0.501391	0.50135	0.497682	100.0
Fourier_Transform	0.51095	0.511191	0.51095	0.508305	10.0
Kmer_Frequencies	0.48170	0.480755	0.48170	0.475258	64.0
Dinucleotide_Properties	0.49320	0.492851	0.49320	0.486947	16.0
Positional_Nucleotide_Frequencies	0.50245	0.502467	0.50245	0.501569	40.0
Kmer_Type_1	0.49215	0.491760	0.49215	0.486072	16.0
Mismatch_Profile	0.50000	0.250000	0.50000	0.333333	1.0
Nucleotide_Auto_Covariance	0.50670	0.506706	0.50670	0.506589	1.0
PSSM	0.48095	0.479833	0.48095	0.473665	64.0

These results will inform our future efforts to improve the tool's efficacy, particularly through focusing on enhancing the data preprocessing and encoding strategies, which are crucial for handling the complex nature of genomic data.