AMRITA VISHWA VIDYAPEETHAM

Computational Framework for Understanding Microbial Pathogenesis and Antimicrobial Resistance (AMR)

Molecular Biology and Basic Cellular Physiology (24AIM112)
Ethics, Innovative Research, Businesses & IPR (24AIM115)

RAGUL U. - CB.AI.U4AIM24036
RAMKUMAR R. - CB.AI.U4AIM24033
SHWETHA P. - CB.AI.U4AIM24042
PRAGALYA M. - CB.AI.U4AIM24032

Introduction

- Antimicrobial resistance (AMR) is the ability of microorganisms such as bacteria, viruses, and fungi to resist the effects of drugs that once killed or inhibited them, making infections harder to treat.
- A key factor driving AMR is the action of virulence proteins, which enhance a pathogen's ability to invade, damage host tissues, and evade immune system.
- Prediction of AMR and protein virulence enables early detection of high-risk pathogens, supports targeted treatments, and strengthens efforts to combat antibiotic resistance and infectious disease outbreaks

Objective

- Predict Antimicrobial Resistance (AMR)
- Prediction of Virulence factor in Protein
- Apply Machine Learning for Predictions
- UI for easy use
- Provide a Scalable and Cost-Effective Approach

AMRITA VISHWA VIDYAPEETHAM

LITERATURE REVIEW

TITLE	KEY FINDINGS
1.Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning	• Demonstrated SVM,CNN models can effectively predict AMR with label encoding, one-hot encoding and frequency matrix chaos game representation (FCGR encoding) on whole-genome sequencing data,identify mutations that are associated with AMR for each antibiotic.
2.Machine learning-enabled prediction of antimicrobial resistance in foodborne pathogens	• Application of ML combined with WGS and spectroscopy techniques to identify and predict AMR in foodborne pathogens which is crucial to assure food safety.
3.Towards routine employment of computational tools for antimicrobial resistance determination via high-throughput sequencing	• It highlights advancements in bioinformatics pipelines, database improvements, and automation to enhance AMR prediction. • The study emphasizes the need for standardization and routine implementation to improve clinical and public health responses.

TITLE	KEY FINDINGS
4.Balancing the risks and benefits of antibiotic use in a globalized world: the ethics of antimicrobial resistance	• Addressing AMR's ethical dimensions includes fair resource allocation, environmental impact, and conflicts of interest in antibiotic development. • Equitable access to antibiotics and stakeholder collaboration in stewardship are crucial to ensure public interest and responsible use. • Balancing risks and benefits demands innovative, global strategies to preserve antibiotic efficacy for future generations.
5.Ethics and antibiotic resistance	• The ethical challenges posed by antibiotic resistance, emphasizing the severe and unevenly distributed health consequences. • It critiques common frameworks like patient responsibility, the tragedy of the commons, and antibiotic stewardship, highlighting their limitations

COMPUTATIONAL ASPECT

Data collected

- Gene sequence
- Protein sequence

Sequence	Description	DNA Sequence	Class
Entry ID	Description	Sequence	Class

Sources

- NCBI (National Centre for Biotechnology Information) for gene sequences
- UniProt for protein sequences

AMR Prediction using CNN (K-mer + TF-IDF)

In [6]:
```
new_sequence_kmers = " ".join(get_kmers("ATGCGTACGTAGCTAGC"))
new_sequence_tfidf = vectorizer.transform([new_sequence_kmers]).toarray().reshape(1, -1, 1)
cnn_prediction = model_cnn.predict(new_sequence_tfidf)

print("CNN Prediction:", cnn_prediction)
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 126ms/step
CNN Prediction: [[0.95005983]]
```

In [10]:
```
# Example new gene sequence
new_sequence = "GGGGGCCGCCTTCGCCACCGGTATTCCTCCAGATCTCTACGCATTTCACCGCTACACCTGGAATTCTACCCC

# Make prediction
predicted_class, confidence = predict_amr_cnn(new_sequence, model_cnn, vectorizer)

# Print results
print(f"Predicted Class: {'AMR (1)' if predicted_class == 1 else 'Non-AMR (0)'}")
print(f"Prediction Confidence: {confidence:.4f}")
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 66ms/step
Predicted Class: AMR (1)
Prediction Confidence: 0.9501
```

Protein virulence prediction

MODEL USED	ACCURACY	FEATURE OF THE MODEL
ProtBert + Random Forest	92%	Learns relationship between sequences
CNN + BiLSTM	93%	Captures the structural patterns

```
In [51]:
# Test on a new sequence
new_seq = "MGGRWSKSSIVGWPAIRERIRRTDPAADGVGAVSRDLEKHGAITSSNTRGTNADCAWLEAQEESEEVGFPVRPQVPLRPMTYKGALDLSHFLKEKGG"
new_seq_encoded = one_hot_encode(new_seq).reshape(1, 200, 20)

# Predict
prob = model.predict(new_seq_encoded)[0][0]
pred_class = 1 if prob > 0.5 else 0

print(f"Predicted Class: {'Virulent (1)' if pred_class == 1 else 'Non-Virulent (0)'}")
print(f"Prediction Confidence: {prob:.4f}")

1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 71ms/step
Predicted Class: Virulent (1)
Prediction Confidence: 0.8128
```

Fig. Output obtained from the final model (CNN+BiLSTM)

XG-BOOST model trained for Virulent Prediction

```
              0       0.87      0.77      0.82       171
              1       0.85      0.93      0.89       254

       accuracy                           0.86       425
      macro avg       0.86      0.85      0.85       425
   weighted avg       0.86      0.86      0.86       425

📄 Model saved as: D:/2ND SEM/delete/NEW/mar6 - proteins/xgboost_model.txt

🔍 Enter a new protein sequence: python -u "d:\2ND SEM\delete\NEW\mar6 - proteins\protein pred user.py"

📌 Prediction Result: Virulent
PS D:\2ND SEM\AMR_2nd>
```

```
✅ Model Accuracy: 0.8612

🔹 Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.77      0.82       171
           1       0.85      0.93      0.89       254

    accuracy                           0.86       425
   macro avg       0.86      0.85      0.85       425
weighted avg       0.86      0.86      0.86       425

📄 Model saved as: D:/2ND SEM/delete/NEW/mar6 - proteins/xgboost_model.txt

🔍 Enter a new protein sequence: MAVMAPRTLLLLLSGALALTQTWAGSHSMRYFFTSVSRPGRGEPRFIAVGYVDDTQFVRFDSDAASQRMEPRAPWIEQEGPEYWDQE
TRNVKAQSQTDRVDLGTLRGYYNQSEAGSHTIQIMYGCDVGSDGRFLRGYRQDAYDGKDYIALNEDLRSWTAADMAAQITKRKWEAAHEAEQLRAYLDGTCVEWLRRYLENGKETLQRTD
PPKTHMTHHPISDHEATLRCWALGFYPAEITLTWQRDGEDQTQDTELVETRPAGDGTFQKWAAVVVPSGEEQRYTCHVQHEGLPKPLTLRWELSSQPTIPIVGIIAGLVLLGAVITGAVV
AAVMWRRKSSDRKGGSYTQAASSDSAQGSDVSLTACKV

📌 Prediction Result: Non-Virulent

D:\2ND SEM\delete\NEW\mar6 - proteins>
```

AMRITA VISHWA VIDYAPEETHAM

Problem faced

Class imbalance in dataset

- Extracting gene and protein sequences from NCBI (AMR) & UniProt (Virulence) was time-consuming and complex .
- The dataset had a skewed distribution (more AMR/Virulent sequences than non-AMR/non-virulent ones)
    - 1319 -Virulence protein sequence
    - 807-Non Virulence protein sequences
    - 480-AMR gene sequence
    - 124- Non AMR gene sequences

Solution

- This issue was addressed by replacing the past dataset with a balanced dataset with:
    - 2,76,666 -Virulence protein sequence
    - 2,76,666-Non Virulence protein sequences
    - 527-AMR gene sequence
    - 527- Non AMR gene sequences

FINAL UPDATED WORK AND RESULTS

Protein Virulence Prediction Result

✅ Accuracy: 84.34%

Classification Report:

	precision	recall	f1-score	support
0	0.80	0.92	0.86	55335
1	0.91	0.76	0.83	55334
accuracy			0.84	110669
macro avg	0.85	0.84	0.84	110669
weighted avg	0.85	0.84	0.84	110669

The Dataset contains 5,53,341 entries.
Evenly split between
2,76,666 Virulence Protein
2,76,675 non Virulence Protein

AMR Prediction Result

```
Test Accuracy: 0.8957

Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.97      0.90       106
           1       0.97      0.82      0.89       105

    accuracy                           0.90       211
   macro avg       0.91      0.90      0.90       211
weighted avg       0.90      0.90      0.90       211
```

The Dataset contains 1,054 entries.

Evenly split between

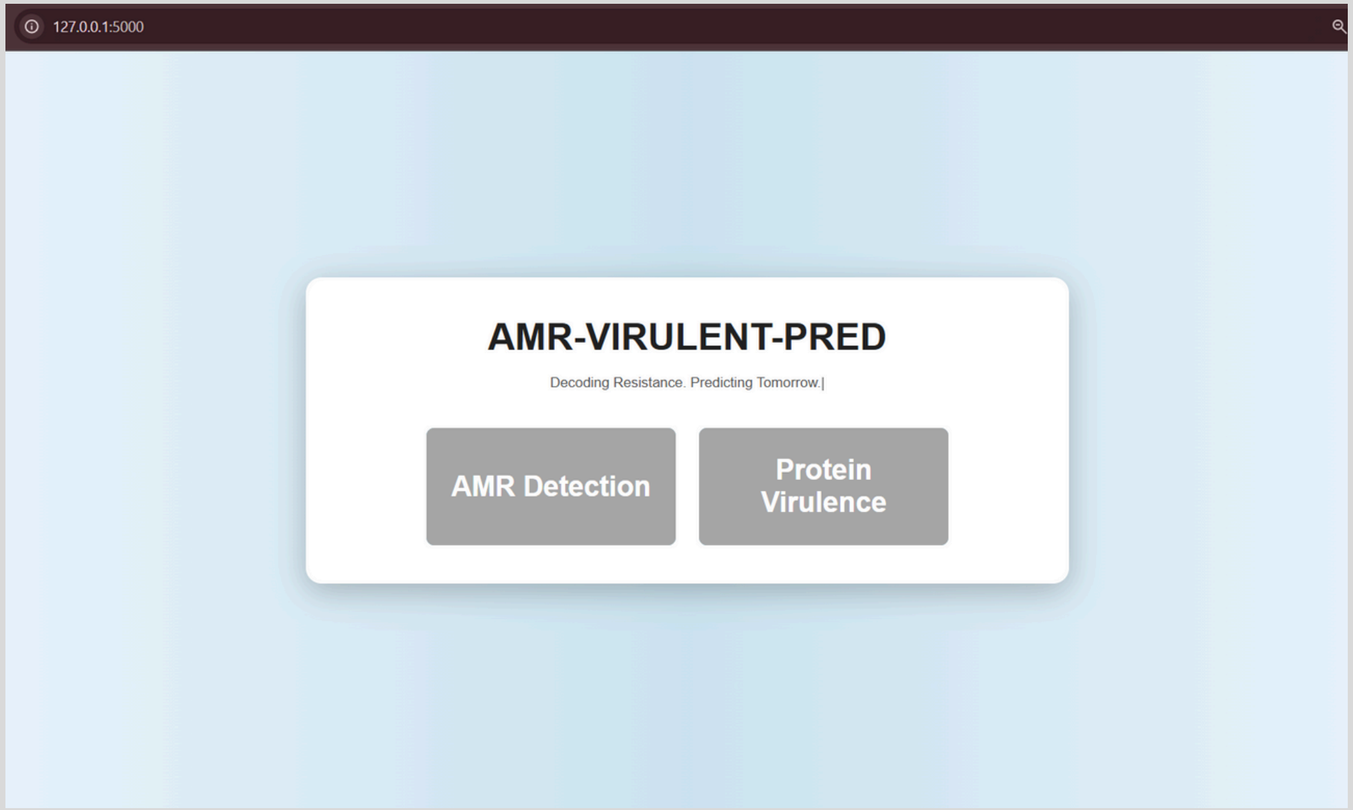
- 527 AMR (antimicrobial-resistant)
- 527 non-AMR samples.

User - Interface

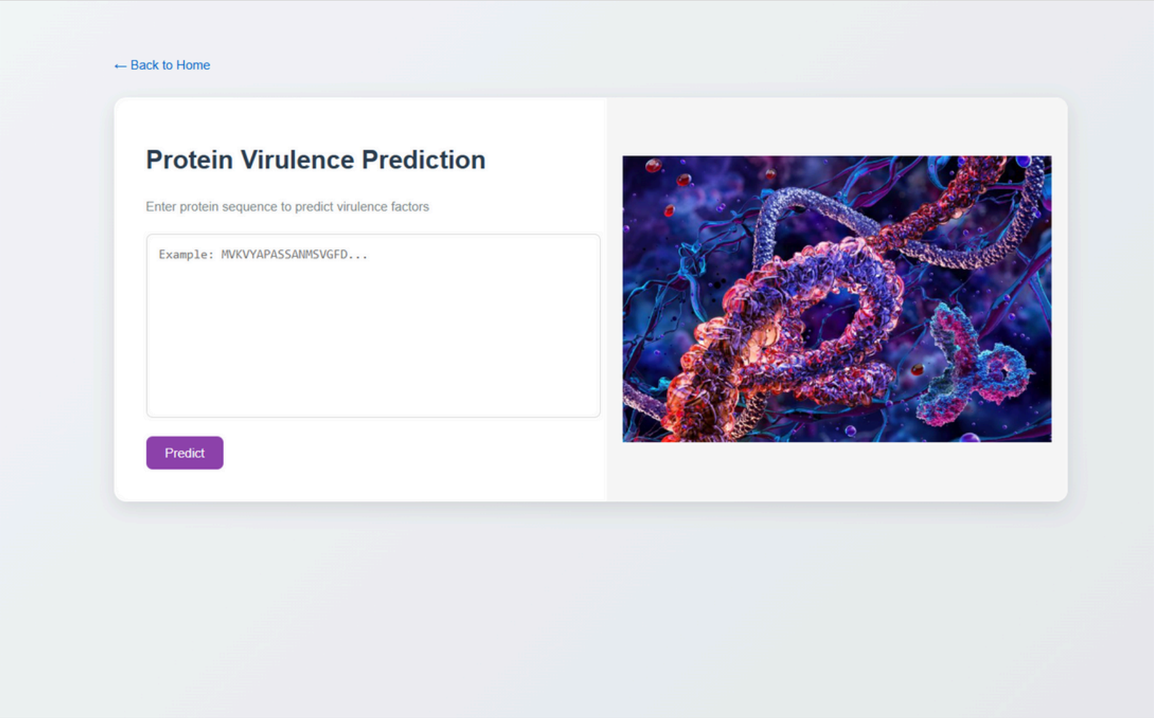
AMR-Virulent-Pred

Two Features :

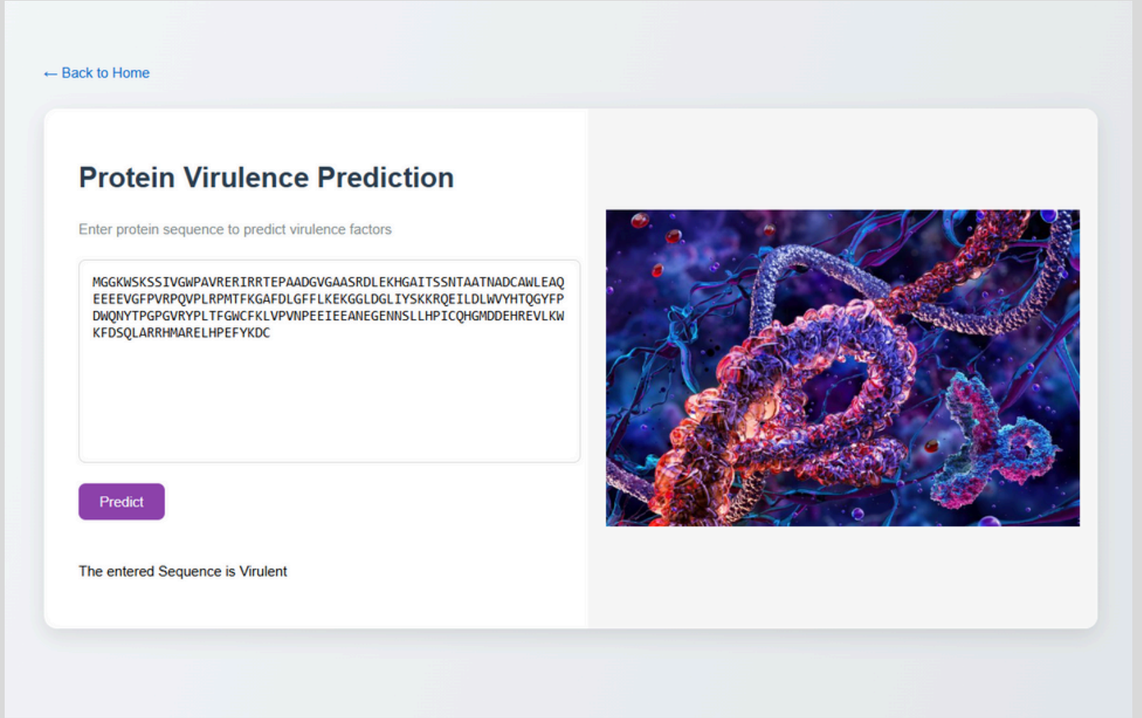
- AMR Detection.
- Protein Virulence



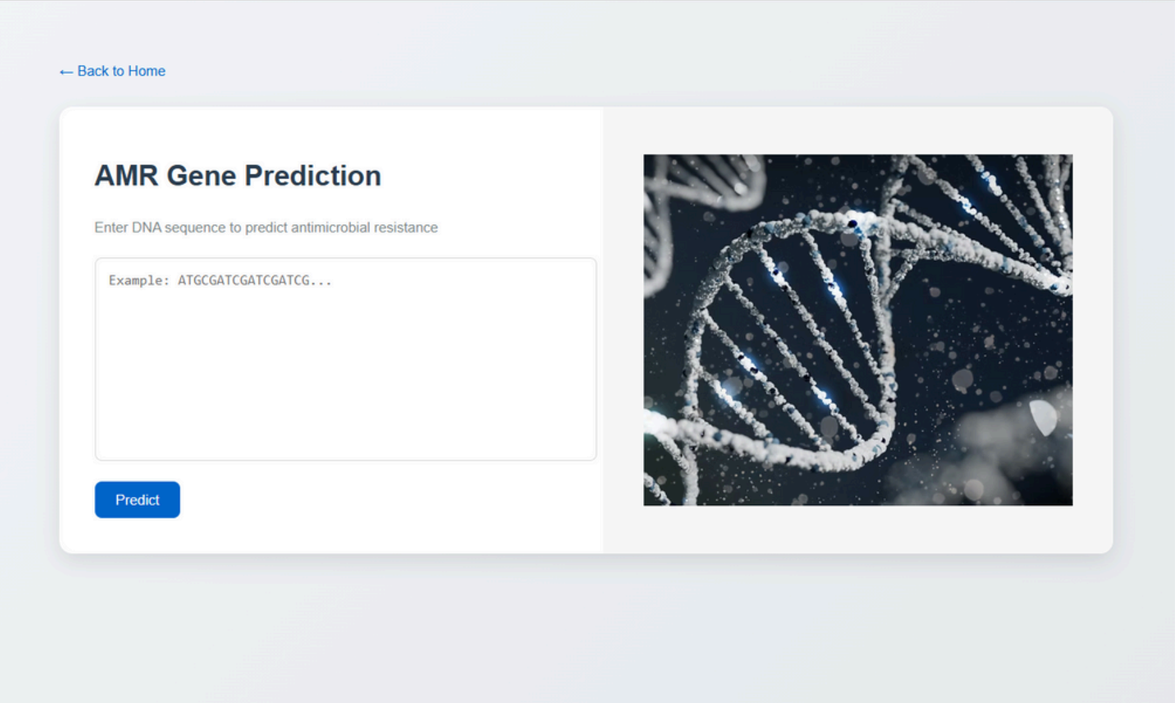
Landing page



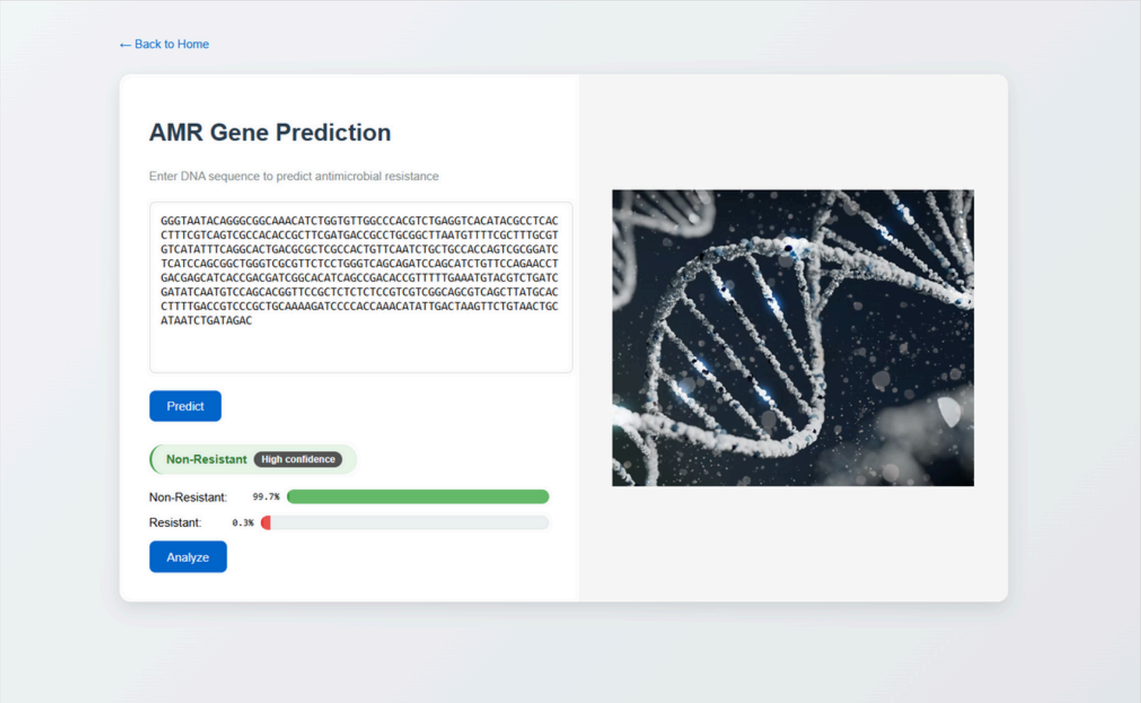
Protein Virulence Prediction Page



Sample run : Protein sequence entered and the correct output is being displayed



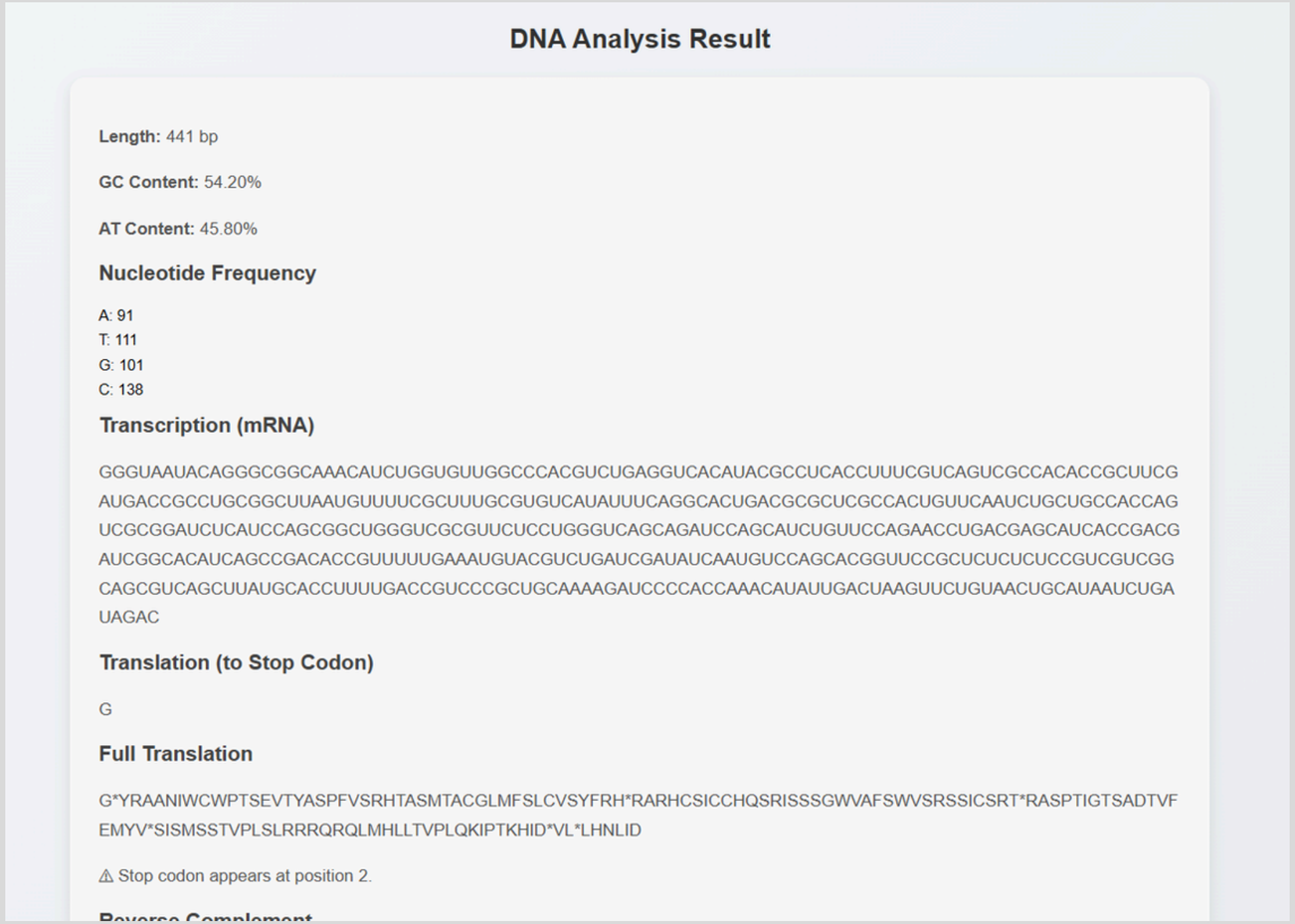
AMR Gene Detection page



<u>Sample run</u> : Gene sequence entered and the output with the prediction percentage is being displayed

DNA Analysis Result

- Length
- GC Content
- AT Content
- Nucleotide Frequency
- Transcription (mRNA)
- Full Translation
- Reverse Complement
- Codon Usage



Sample DNA Analysis Report

Machine learning for predicting antimicrobial resistance in critical and high-priority pathogens (Carlos, Daniel, Sergio)

Published 25 Feb 2025

They found that Gradient Boosted Decision Trees (GBDT), Random Forest, and XGBoost were the top-performing ML models.

assessing 688,107 patients and 1,710,867 antimicrobial susceptibility tests. GBDT, Random Forest, and XGBoost were the top-performing ML models for predicting antibiotic resistance in CHPP infections. GBDT exhibited the highest AuROC values compared to Logistic Regression (LR), with a mean value of 0.80 (range 0.77–0.90) and 0.68 (range 0.50–0.83), respectively. Similarly, Random Forest generally showed better AuROC values compared to LR (mean value 0.75, range 0.58–0.98 versus mean value 0.71, range 0.61–0.83). However, some predictors selected by these algorithms align with those suggested by LR.

Case Study

PATENT NO.: US 2012/0196309 A1
TITLE: METHODS AND KITS FOR DETECTION OF ANTIBOTC RESISTANCE
DATE OF PATENT: Aug. 2, 2012

- Detects antibiotic resistance using mass spectrometry.
- Includes a kit with reagents for rapid testing.
- Combines phenotypic and chemical analysis in a unified platform.
- Applicable to multiple sample types: colonies, blood, body fluids.

Artificial intelligence in predicting pathogenic microorganisms' antimicrobial resistance: challenges, progress, and prospects

Author : Yan Li, Xiaoyan Cui, Xiaoyan Yang, Guangqia Liu ,Juan Zhang
Published : 01 November 2024

Objective :

1. This paper reviews the latest advancements in AI and ML for predicting antimicrobial resistance in pathogenic microorganisms.
2. Highlights the main AI and ML models used in resistance prediction,
3. Finally the paper discuss about the new perspectives and solutions for research into microbial resistance through algorithm optimization, dataset expansion, and interdisciplinary collaboration.

Algorithms used :

SVM (Support Vector Machine)
RF (Random Forest)
CNN (ConvolutionalNeural Network)
KNN (KNearest Neighbors)
NB (Naive Bayes)

Antimicrobial Resistance (AMR) and Multidrug Resistance (MDR): Overview of current approaches, consortia and intellectual property issues

Overuse of antibiotics:

Evidence shows that in low-income and middle-income countries (LMICs), antibiotic use is increasing with rising incomes, high rates of hospitalization, and high prevalence of hospital infections.[44] However,

Equality in distribution:

Generally, innovators and generics are commercially incentivized to sell high volumes of product. This cannot work in the case of AMR/MDR as the goal is to provide access to only those patients who absolutely need the state-of-the-art treatment.[61] Low sales generally lead to an unsustainable business model, but high levels of sales would result in overconsumption and contribute to high levels of resistance.

Profit from Sales

Numerous experts have proposed antibiotic business models that reinforce conservation efforts by completely severing a developer's ROI from sales volume and price. This concept is known as "de-linkage" and is beneficial for three key reasons. Firstly, it provides developers with a concrete ROI that is extraneous to the market. Secondly, it removes the motivation for developers to overmarket their antibiotic.

Case Study

PATENT: GLYCOMIMETICS TO INHIBIT PATHOGEN-HOST INTERACTIONS (US 9,605,014 B2)

Date of Patent: Mar. 28, 2017

- The patent US9605014B2 describes glycomimetic compounds designed to inhibit pathogen-host interactions
- A glycomimetic as described herein may be used to impregnate filters, masks, and clothing or any combination thereof in prophylactic strategies to reduce transmission and inhibit the binding of any of a variety of pathogens to their target.

Case Study

PATENT: GLYCOMIMETICS TO INHIBIT PATHOGEN-HOST INTERACTIONS (US 9,605,014 B2)

In addition to administration to a subject for therapeutic use, a glycomimetic as described herein may be used to impregnate filters, masks, and clothing or any combination thereof in prophylactic strategies to reduce transmission and infection.

performing a computational simulation to predict the binding energies of glycomimetic-pathogen complex or glycomimetic scaffold-pathogen complex;

performing computational substituent remodelling of the glycomimetic or glycomimetic scaffold to further improve the affinity of the molecule for the pathogen, or to modify certain properties of the molecule including, but not limited

Large scale application of glycomimetics in ->everyday materials could result in unintended environmental effects like resistance development in pathogens

-> while computational methods reduce need for human trails,eventual human testing requires proper data protection and informed consent,raising privacy concerns

Case Study

TITLE:DEVICE FOR DETERMINING ANTIMICROBIAL SUSCEPTABILITY OF A MICROORGANISM

PATENT NUMBER: US 12 203 125 B2

DATE OF PATENT: Jan. 21,2025

- The invention provides methods for determining antimicrobial susceptibility.
- The device was tested under various growth conditions using different antibiotics to evaluate its performance.
- It helps in efficient identification of the most effective antibiotic against a specific microorganism.

PAPER 1 - Ethics and antibiotic resistance

- The ethical challenges posed by antibiotic resistance, emphasizing the severe and unevenly distributed health consequences. It critiques common frameworks like patient responsibility, the tragedy of the commons, and antibiotic stewardship, highlighting their limitations

Responsibility and Accountability

Areas of agreement: Ethical analyses have focused on the moral responsibilities of patients to complete antibiotic courses, resistance as a tragedy of the commons and attempts to limit use through antibiotic stewardship.

to the problem.[3] Making and implementing policy to address drug resistance involves balancing multiple ethical values, as well as multiple types of harms and benefits.[2,3]

problems related to resistant bacteria involve multiple species and sectors, and at this broader level, ABR may be best characterized as a One Health problem.[9,10]

Public Health vs. Individual Freedom

The recognition that antibiotic use increases acquired ABR, and thereby may reduce the availability of effective antibiotic therapy for all, led to the formulation that antibiotic use is analogous to one type of collective action problem known as the tragedy of the commons.[23–30] Standard commons

infection[32,33]; (ii) therapeutic uncertainty, as not all patients with a bacterial infection benefit from antibiotic therapy[34]; (ii) microbiological uncertainty, microbiological uncertainty, as pathogens are not uniformly susceptible to all available antibiotic options (and this information may not be available

Conflicts of Interest in Healthcare

ations between individuals' antibiotic consumption and colonization or infection with resistant isolates can be identified even when associations at the ecological ('commons') level are absent.[41,42] The implication of this observation is that the 'commons' necessarily places in opposition. When attempting to resolve this assumed conflict, and in concordance with dominant professional norms, clinicians often prioritize their immediate patient over the interests of other, distant and/or future patients.[32,33,44] In

involves risk, these risks are concentrated among users), and patients who use disproportionately large amounts of antibiotics are not 'free riding' on

Although inescapable within the 'commons' formulation, this conflict of interest can disappear when the limitations of this formulation are considered.

Moral hazard arises when prescribers' incentives do not align with those of the other two parties, and qualitative research has identified these adverse drivers of antibiotic prescribing as active in multiple medical contexts.[50] Chief among these appear to

need for other policies to address inappropriate antibiotic use.

Moreover, stewardship resources are often concentrated in hospitals, where some of the more overt harms from resistant bacterial infections often

Autonomy and Individual Rights

risks. Unlike infections with respiratory viruses that last days or weeks, resistant bacteria are often carried asymptomatically for years, meaning that cumulative infringements on carriers' lives may be even more significant. Being identified as a carrier can result in restricted access to healthcare and mental health issues due to stigmatization,[60,61] but also infringements on privacy, freedom of movement and free choice of occupation.[12] Determining the con-

promote the spread of resistant organisms (and resistance traits between organisms) as much of the environment in poor communities is contaminated with resistant bacteria.[9] Without addressing these factors, stewardship efforts to reduce antibiotic use among the global poor (and, for that matter, in livestock and agriculture) will remain largely futile.

Above and beyond altruistic motivations for high-income countries to help address these social deter-

Justice and Fairness

the global human population. Even before questions of whether people from such communities can access resistant organisms. The recent COVID-19 pandemic has highlighted the ethical salience of public health measures for asymptomatic infection, which some-

'courses' of antibiotics appear excessive); (ii) the minimally effective antibiotic spectrum for specific antibiotics and/or a healthcare provider to diagnose the disease and supervise access to antibiotics comes the need for basic public health measures (see Fig. 1).

of 'access'.[62] While high-income countries struggle with the development of appropriate policies to curb inappropriate use of antibiotics available in abundance, hundreds of thousands of people die in low-income countries every year for want of access to antibiotics.[63] Yet this contrast hides more

and health policy. From an ethical perspective, policy and clinical decisions should be based on value judgements informed by sound evidence. Where this evidence is lacking there is an ethical imperative for more relevant scientific research and public health surveillance. Policy should also focus on harm reduc-

Much of this disease burden is concentrated in low- and middle-income countries (LMICs) where surveillance for resistance is often incomplete.[2]

and freedom (or liberty).[8] Before embarking on a tour of the ethical issues related to acquired ABR, it is important to review the biology and

References

[1]de Nies, L., Lopes, S., Busi, S.B. et al. PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. Microbiome 9, 49 (2021). https://doi.org/10.1186/s40168-020-00993-9

[2]Ren Y, Chakraborty T, Doijad S, et al. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. Bioinformatics. 2022;38(2):325-334. doi:10.1093/bioinformatics/btab681

[3]Sanabria, A.M., Janice, J., Hjerde, E. et al. Shotgun-metagenomics based prediction of antibiotic resistance and virulence determinants in Staphylococcus aureus from periprosthetic tissue on blood culture bottles. Sci Rep 11, 20848 (2021). https://doi.org/10.1038/s41598-021-00383-7

[4] Gao Y, Li H, Zhao C, Li S, Yin G and Wang H (2024) Machine learning and feature extraction for rapid antimicrobial resistance prediction of Acinetobacter baumannii from whole-genome sequencing data. Front. Microbiol. 14:1320312. doi: 10.3389/fmicb.2023.1320312

[5]Yitong Liu, Xin Cao, Jiani Li et al. Advancing virulence factor prediction using protein language models, 29 July 2024, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-4664562/v1]

[6]Qian J, Jin P, Yang Y, Ma N, Yang Z, Zhang X. Protein function annotation and virulence factor identification of Klebsiella pneumoniae genome by multiple machine learning models. Microb Pathog. 2024;193:106727. doi:10.1016/j.micpath.2024.106727