A Lightweight Web Application for Genome-Based Drug Repurposing Using Machine Learning Models

Ahmed AlShahab[1], Dr. Vaishali A. Chavan[2]

[1,2] Department of Computer Science, Dr. G. Y. Pathrikar College of Computer Science and IT, MGM University

Abstract—**Drug repurposing reduces the cost, time, and risk associated with conventional drug discovery by identifying new therapeutic uses for existing drugs. With the rapid growth of genomic data, machine learning models provide effective tools for genome similarity analysis and drug target identification. This paper presents a lightweight web-based application for genome- based drug repurposing that predicts similar viral genomes using a Random Forest machine learning model and recommends associated antiviral drug molecules. The system accepts viral genome FASTA protein sequences, validates input data, performs real-time prediction via a FastAPI backend, and presents results through a ReactJS frontend. Approved and experimental drug references are linked to PubChem for validation. Experimental evaluation demonstrates accurate genome similarity prediction with low latency, highlighting the feasibility of deploying ma- chine learning-driven bioinformatics solutions as accessible web applications.**

Index Terms—**Drug Repurposing, Genomics, Machine Learn- ing, Random Forest, FASTA, Web Application, Bioinformatics**

I. INTRODUCTION

Drug discovery is traditionally a costly and time-intensive process with high attrition rates during clinical trials. Drug repurposing, also known as drug repositioning, provides an effective alternative by identifying new therapeutic applica- tions for existing drugs, thereby reducing development cost and time.

Advances in genome sequencing technologies have enabled large-scale analysis of viral and disease-associated genomes. Genome-based similarity analysis allows researchers to infer functional relationships between pathogens and known therapeutic agents. However, the high dimensionality of genomic data poses significant computational challenges.

Machine learning (ML) techniques have shown strong per- formance in modeling complex biological data, including genome classification and similarity detection. Despite these advances, most existing approaches focus on offline analysis or lack deployable platforms for real-time usage. This work pro- poses a lightweight, end-to-end web application that integrates machine learning-based genome similarity prediction with antiviral drug recommendation, making advanced bioinformatics analysis accessible to a wider research community.

II. RELATED WORK

Genome-based drug repurposing has been widely studied using computational and machine learning techniques. Early approaches relied on sequence alignment, molecular docking, and protein–protein interaction networks to identify potential drug targets. While effective, these methods are computation- ally expensive and often require expert intervention.

Machine learning-based approaches have gained popularity due to their ability to learn patterns directly from genomic data. Supervised learning algorithms such as Support Vector Machines, Decision Trees, K-Nearest Neighbors, and Random Forests have been applied to classify genomes and predict drug–target associations. These methods demonstrate strong predictive performance when appropriate feature encoding techniques are employed.

AlShahab and Chavan proposed a genome-based drug re- purposing framework using FASTA protein sequence data and classical machine learning algorithms. Their experimental evaluation showed that the Random Forest classifier achieved superior performance, with an accuracy of approximately 98.79%, highlighting its robustness and suitability for genomic similarity prediction. This prior work provides the method- ological foundation for the machine learning model adopted in the present study.

Recent studies have explored deep learning techniques, including convolutional neural networks and graph neural networks, to capture complex relationships in genomic and molecular data. Although these models achieve high accuracy, they require substantial computational resources and are diffi- cult to deploy in lightweight environments.

Several web-based bioinformatics platforms have been pro- posed to improve accessibility; however, many focus on visu- alization rather than real-time prediction and drug recommen- dation. The proposed work differentiates itself by combining a proven Random Forest model with a lightweight, deployable web architecture that supports real-time genome analysis and direct drug repurposing recommendations.

III. SYSTEM ARCHITECTURE

The proposed system follows a client–server architecture composed of three layers:

- Frontend Layer: Implemented using ReactJS, providing genome sequence input, validation feedback, and visual- ization of results.
- Backend Layer: Developed using FastAPI to handle RESTful POST requests, preprocessing, and machine learning inference.
- Machine Learning Layer: Utilizes a trained Random Forest model to predict genome similarity and retrieve associated antiviral drugs.

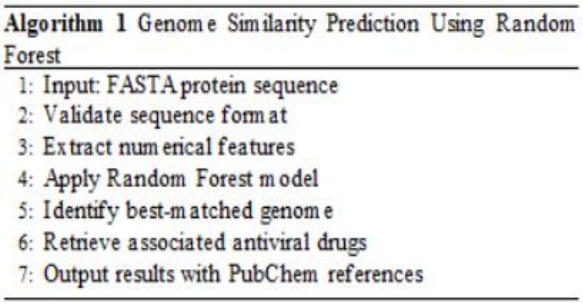
The system is deployed as a web application and is acces- sible at https://dr.dualsysco.com.

IV. DATASET AND FEATURE ENCODING

Viral genome protein sequences in FASTA format are used as input. Sequences are preprocessed to remove noise and converted into numerical feature vectors using sequence- based encoding techniques. These features capture biologically relevant information required for machine learning-based similarity prediction. FASTA protein sequences were downloaded from the NCBI protein database [4].

V. MACHINE LEARNING MODEL

A Random Forest classifier is employed due to its robust- ness, ability to handle high-dimensional data, and resistance to overfitting. The model consists of an ensemble of decision trees trained on bootstrapped samples of the dataset. Genome similarity is inferred based on majority voting across trees.



Algorithm 1 Genome Similarity Prediction Using Random Forest

```
Input: FASTA protein sequence
Validate sequence format
Extract numerical features
Apply Random Forest model
Identify best-matched genome
Retrieve associated antiviral drugs
Output results with PubChem references
```

VI. WORKFLOW

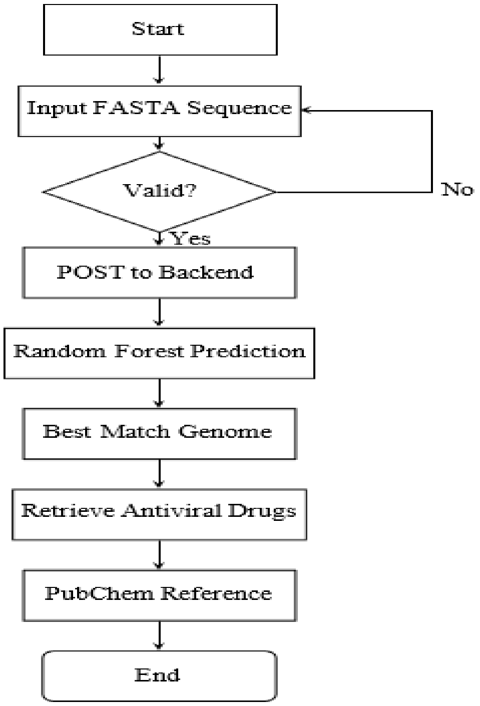


Fig. 1. Workflow of the proposed genome-based drug repurposing system

VII. RESULTS AND DISCUSSION

Experimental evaluation indicates that the Random For- est model is capable of identifying genome similarity with stable performance and low response time. The lightweight web-based deployment enables real-time prediction and drug recommendation, making the system suitable for exploratory research and decision support in genome-based drug repurposing.

A. Example Prediction Case Study

To demonstrate the practical functionality of the proposed system, an example prediction result generated through the web application is presented. A viral protein sequence in FASTA format was provided as input by the user and pro- cessed after successful validation.

1) Input Genome Sequence: The following protein se- quence corresponds to *Human Metapneumovirus (HMPV)* and was obtained from the National Center for Biotechnology Information (NCBI) protein database [4].

```
MSLQGIHLSDLSYKHAILKESQYTIKRDVGTTTA
VTPSSLQQEITLLCGEILYAKHADYK
YAAEIGIQYISTALGSERVQQILRNSGSEVQVVL
TRTYSLGKIKNNKGEDLQMLDIHGVE
KSWVEEIDKEARKTMATLLKESSGNIPQNQRPS
APDTPIILLCVGALIFTKLASTIEVGL
ETTVRRANRVLSDALKRYPRMDIPKIARSFYDL
FEQKVYHRSLFIEYGKALGSSSTGSKA
ESLFVNIFMQAYGAGQTMLRWGVIARSSNNIML
GHVSVQAELKQVTEVYDLVREMGPESG
LLHLRQSPKAGLLSLANCPNFASVVLGNASGLG
IIGMYRGRVPNTELFSAAESYAKSLKE
SNKINFSSLGLTDEEKEAAEHFLNVSDDSQNDY
E
```

2) Prediction Output: After clicking the *Process Sequence* option, the validated sequence was sent to the backend via an HTTP POST request. The Random Forest model analyzed the input and produced the following results:

- Predicted Closest Genome Match: Rotavirus
- Similarity Confidence Score: 66.00%

Based on the identified genome similarity, the system sug- gested antiviral drug molecules associated with the predicted viral family.

3) Suggested Drug Molecules: The recommended com- pounds included:

- Nitazoxanide ($C_{12}H_9N_3O_5S$, 157.10 g/mol), an experi- mental broad-spectrum antiviral agent.
- Lactoferrin ($C_{141}H_{226}N_{46}O_{29}S_3$, 3125.8 g/mol), an ex- perimental compound under investigation for antiviral and immunomodulatory effects.

Each suggested compound is accompanied by a direct reference to the PubChem database, allowing researchers to further examine chemical properties, biological activity, and related studies.

4) Discussion: This example illustrates the system's ability to process an unseen viral protein sequence, identify the most similar genome, and recommend relevant antiviral drug candidates. The results are intended to assist researchers in preliminary drug repurposing analysis and hypothesis gen- eration, rather than serve as a final or clinically validated therapeutic conclusion.

VIII. LIMITATION

Despite the promising performance of the proposed system, several limitations should be acknowledged.

First, the Random Forest model has been trained on a comprehensive but domain-specific dataset of viral protein se- quences, including Chickenpox, Chikungunya, Dengue, Ebola, Rotavirus, and Zika viruses. The training dataset comprises a total of 246,650 variant genome sequences across these diseases. While this dataset provides strong coverage for the selected viral families, certain viral diseases were intentionally excluded during training for testing and validation purposes. Consequently, the model's predictive capability is currently limited to the viral genome families included in the training dataset. Future work will focus on expanding the database to incorporate additional viral genome families as more genomic data become available.

Second, the proposed web application is designed primarily as a research-oriented tool and has been developed specifically to support academic research, including doctoral-level studies. The system is not intended for direct use by the general public or for clinical decision-making by non-experts. Appropriate domain knowledge is required to interpret the results respon- sibly.

Finally, the predicted genome similarity and associated an- tiviral drug recommendations are intended to assist and guide computational drug repurposing research. The results should not be interpreted as a final or definitive therapeutic verdict. Instead, the suggested drug molecules represent computational hypotheses that require further experimental, molecular, and clinical validation before any real-world application.

IX. CONCLUSION

This paper presented a lightweight web application for genome-based drug repurposing using a Random Forest ma- chine learning model. By integrating genome similarity pre- diction with a modern web architecture, the system provides an accessible and scalable solution for computational drug repurposing. Future work will focus on deep learning integration and multi-omics data analysis.

X. ABBREVIATIONS

ML	Machine Learning
RF	Random Forest
FASTA	Fast-All Sequence Alignment
API	Application Programming
Interface REST	Representational State Transfer
HTTP	Hypertext Transfer Protocol
UI	User Interface
JSON	JavaScript Object Notation
PubChem	Public Chemical Database (NCBI)
PhD	Doctor of Philosophy

REFERENCES

[1] T. T. Ashburn and K. B. Thor, “Drug repositioning: Identifying and de- veloping new uses for existing drugs,” *Nature Reviews Drug Discovery*, 2004.

[2] S. Pushpakom *et al.*, “Drug repurposing: Progress, challenges and recommendations,” *Nature Reviews Drug Discovery*, 2019.

[3] A. AlShahab and V. A. Chavan, “Genome-Based Drug Repurposing: Identifying Potential Targets Using FASTA Sequences and Machine Learning,” *International Journal of Scientific Advances in Technology (IJSAT)*, 2024, DOI: pending.

[4] National Center for Biotechnology Information (NCBI), “Protein Database,” Available: https://www.ncbi.nlm.nih.gov/protein/