Sequence-Based Toxicity Prediction of Antimicrobial Peptides Using Hybrid Machine Learning and Alignment-Guided Preprocessing for Drug Safety Assessment

Prahalika Nivarthi*, Sahil Nambiar*, Agastya Nambiar*
* School of Computing, Department of CSE
Computer Science (Artificial Intelligence)
Amrita Vishwa Vidyapeetham
bl.sc.u4aie24036@bl.students.amrita.edu, bl.sc.u4aie24045@bl.students.amrita.edu, bl.sc.u4aie24006@bl.students.amrita.edu

***Abstract*—The antimicrobial peptides (AMPs) are the novel alternatives to the traditional antibiotics since they are broad-spectrum agents that develop less resistance. Nevertheless, some AMPs have cytotoxic or hemolytic properties with regard to human cells that also restrict their area of application. Safe development of AMP toxicity through early-stage computational screening is consequently inevitable. The present research paper is an attempt to present a sequence-based hybrid deep learning and machine learning paradigm to predict the toxicity of antimicrobial peptides. Redundancy removal and sequence filtering followed by analysis of alignment processing were methods to curate a given AMP dataset to provide a biologically-integrated, overfitting-reduced dataset. The methods of global and local alignment were used to confirm the similarity of the sequences and determine conserved toxicity-related motifs. After that, toxicity classification was performed by a hybrid feature representation of deep learning-derived features and physicochemical features. The last system produces toxicity report and establishes drug suitability of candidate peptides. The approach suggested allows conducting scalable and interpretable early-stage toxicity analysis without structural data.**

***Index Terms*—Antimicrobial peptide, toxicity prediction, hybrid machine learning, sequence alignment, drug safety.**

I. Introduction

The fast increase in the antimicrobial resistance has accelerated efforts in the discovery of alternative therapeutic agents. Antimicrobial peptides (AMPs) are short bioactive molecules, which possess anti-bacterial, antifungal, and antiviral properties. Although they have a therapeutic potential, several AMPs have a toxicity profile on human cells, including hemolytic and cytotoxic activities, that restrict their clinical translation.

The conventional system of toxicity testing is based on lab experiments which are both expensive and time consuming. The computation methods based on machine learning enable the prediction of toxicity in relation to amino acid sequence without any structural data. The paper introduces a well-organized AMP toxicity prediction preprocessing and modeling pipeline based on redundancy elimination, sequence filtering, such as alignment analysis, and hybrid machine learning.

II. Materials and Methods

A. Dataset Description

The database only includes those antimicrobial peptide sequences that have been experimentally verified and retrieved in curated AMP databases. These records include a peptide identifier and amino acid sequence and toxicity information (toxic or non-toxic to human cells).

B. Redundancy Removal

Biological databases usually possess quite similar sequences. In order to avoid overfitting, sequence identity clustering was used to eliminate redundancy.

Sequence identity is computed as:

$$\text{Identity} = \frac{N_{match}}{L_{alignment}} \times 100 \tag{1}$$

where N_{match} represents the number of identical residues and $L_{alignment}$ represents the alignment length.

Sequences sharing greater than 80% identity were clustered, and one representative sequence per cluster was retained.

C. Data Filtering

Additional filtering steps were applied:

- **Length Filtering:** Peptides outside the typical AMP range (10–60 amino acids) were removed.
- **Character Validation:** Sequences containing ambiguous amino acids were excluded.
- **Class Balancing:** Toxic and non-toxic samples were balanced.

D. Global Alignment

The NeedlemanWunsch algorithm was used to measure global alignment. Scoring function can be defined as:

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + S(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases} \tag{2}$$

where $S(x_i, y_j)$ denotes substitution score and d represents gap penalty.

Global alignment validates overall sequence similarity.

E. Local Alignment

The Smith-Waterman algorithm was used as local alignment:

$$H(i,j) = \max \begin{cases} 0 \\ H(i-1, j-1) + S(x_i, y_j) \\ H(i-1, j) - d \\ H(i, j-1) - d \end{cases} \tag{3}$$

Local alignment: The local programs recognize conserved shorter sub sequences related to toxicity related motifs.

F. Hybrid Feature Extraction

Two categories of features were extracted:

- **Handcrafted Features:**
  - Amino acid composition
  - Net charge
  - Hydrophobicity
  - Molecular weight
  - Sequence length
- **Deep Learning Features:**
  - One-hot encoded sequences
  - 1D Convolutional Neural Network embeddings

The obtained learned deep features were also stacked with handmade descriptors in order to create a hybrid feature vector.

G. Toxicity Classification and Drug Suitability

The hybrid feature set was trained on a supervised classifier that was used to predict toxicity. The final output includes:

- Toxicity classification (Toxic / Non-toxic)
- Confidence score
- Drug appropriateness decision

Peptides that are toxic are not drug-appropriate, and the non-toxic peptides are labeled drug-appropriate.

III. Results and Discussion

The Removal of redundancy caused the decrease of sequence similarity and the inhibition of overfitting. The analysis of alignment demonstrated that the toxic peptides have known cationic and hydrophobic motifs that are conserved across all, which indicates the biologic validity. The hybrid model utilizes interpretable physicochemical descriptors with deep learned sequence representations, which is much more resilient to single-feature methods.

The system proves that toxicity can be predicted successfully with the help of sequence information only.

IV. Conclusion

The research paper contains a systematic model on prediction of antimicrobial peptide toxicity based on sequence analysis. Incorporated in the proposed system are redundancy removal, filtering, global and local alignment, and a hybrid machine learning modeling that ensures biological integrity and computational robustness.

The structural information is not necessary, and the developed pipeline allows conducting toxicity screening and assessment of drug appropriateness at an early stage. Future research can include transformer-based systems in architecture and multi-class predictiveness of toxicity.