Computational Molecular Biology and Bioinformatics

EvoGradient

Malay Bhattacharyya

Associate Professor

Machine Intelligence Unit
Indian Statistical Institute, Kolkata

October, 2025

Background

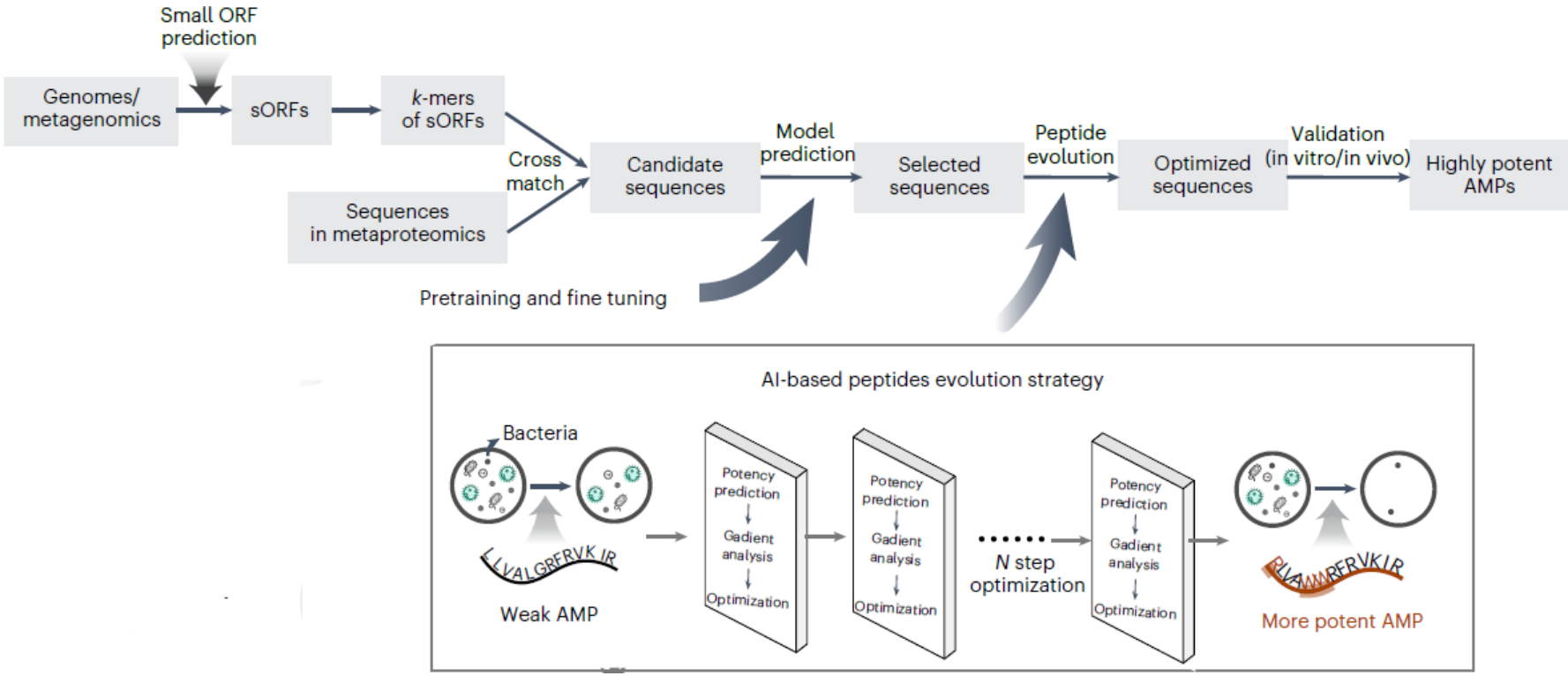
Antimicrobial peptides (AMPs) are small peptides known to inhibit the growth of various microorganisms, including bacteria and fungi. Found in a diverse range of organisms such as plants, animals, humans and microorganisms, natural AMPs display a broad spectrum of specificity and mechanisms of action.

AMPs are effective against pathogens resistant to conventional antibiotics. Therefore, they are considered as potential solutions to the public health concerns posed by pandrug-resistant microbes and the diminishing efficacy of current antibiotics.

What is EvoGradient?

EvoGradient is an explainable deep learning model that predicts the potency of antimicrobial peptides (AMPs) and virtually modifies peptide sequences to produce more potent AMPs akin to in silico directed evolution [1].

The method


Small ORF prediction
Genomes/ metagenomics
sORFs
k-mers of sORFs
Cross match
Sequences in metaproteomics
Candidate sequences
Model prediction
Selected sequences
Peptide evolution
Optimized sequences
Validation (in vitro/in vivo)
Highly potent AMPs
Pretraining and fine tuning
AI-based peptides evolution strategy
Bacteria
Weak AMP
Potency prediction
Gadient analysis
Optimization
N step optimization
More potent AMP

Dataset preparation

- **The initial dataset:** It comprises peptide sequences labelled as either 1 or 0, indicating their classification as AMPs or non-AMPs. It was accumulated from 5 publicly available AMP databases (namely APD, LAMP1, LAMP2, BAGEL4 and dbAMP) and Uniprot (for Non-AMPs) by excluding duplicated peptides and those longer than 100 AAs. This dataset is used for the pre-training and testing (90:10) of the peptide model aiming a classification task.
- **The extended dataset:** It includes peptides labeled with minimum inhibitory concentration (MIC) values, a quantification measurement to assess the ability of peptides to kill bacteria. It was collected from the publicly available Grampa dataset [2]. This dataset is used for further quantitative adaptive tuning of the peptide model through training and testing (85:15) aiming a regression task.

The pre-training phase

Before feeding the peptide sequences into the model, each of the 20 AAs was initially encoded as a unique one-hot vector. A subsequent zero-padding ensures that all peptide sequences are aligned to a uniform length of 100 AAs. After this pre-processing, the input shape becomes batch size $\times$ 100 (uniform length) $\times$ 20 (AA embeddings).

The model is pre-trained on this large, prep-processed and balanced dataset. The main objective of the pre-training phase is to identify AMPs from the input peptides, essentially performing a binary classification task. The cross-entropy loss is used as the initial loss function for training. To further minimize the occurrence of false positive samples and enhance model precision, a conservative loss function is designed to impose penalties on overly confident predictions generated by the model.

The pre-training phase

- **Cross-entropy loss:**

$$L(y_c, t_c) = -\frac{1}{n}\sum_{i=1}^{n}(t_{c,i}\log(y_{c,i}) + (1 - t_{c,i})\log(1 - y_{c,i})),$$

where $L(y_c, t_c)$ represents the cross-entropy loss. $y_{c,i}$ denotes the predicted probabilities of the positive class AMP of sample i, and $t_{c,i}$ represents the true classification labels of sample i, with 0 representing non-AMP and 1 denoting AMP, n is the number of samples.

- **Conservative loss:**

$$L_{con}(y_c, t_c) = L(y_c, t_c) + \lambda\sum_{i=1}^{n}(\mathrm{ReLU}(y_{c,i} - t_{c,i}))^2,$$

where the hyperparameter λ scales the penalty degree, allowing control over the influence of the penalty term.

The adaptive tuning phase

The adaptive tuning stage integrates a continuous MIC regression task into the discrete binary classification pre-training framework. With this, we can capture both strong and weak antimicrobial potency, ensuring that the output value accurately represented the level of antimicrobial activity exhibited.

The objective during the adaptive tuning stage becomes the minimization of MSE loss as follows:

$$L_r(y_r, t_r) = \frac{1}{n}\sum_{i=1}^{n}(t_{r,i} - y_{r,i})^2.$$

We further apply a learning rate ($l_r = 0.001$) to slowly update the entire neural network (base model).

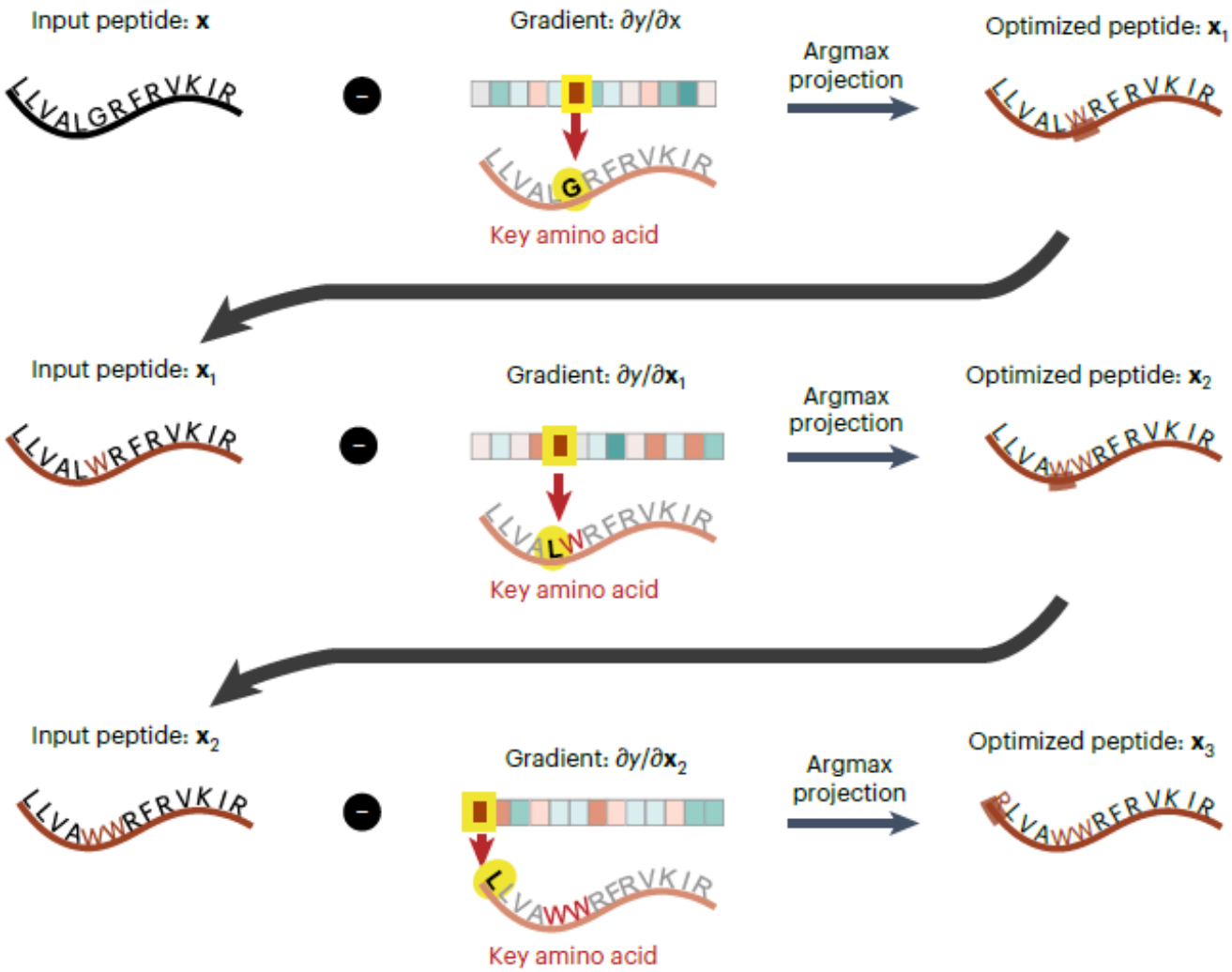
The model

EvoGradient is an iterative gradient descent approach combined with a projection operation to discover a locally optimal sequence near the original peptide. Through forward propagation, we could obtain a predicted MIC value for the original peptide. We then employed a back-propagation algorithm to compute the gradient of the predicted MIC with respect to the input sequence. The difference between the scaled gradient and the input vector (x) is given by

$$x' = x - \alpha \frac{\partial y_r}{\partial x},$$

where α is a hyperparameter that determines the magnitude of each step taken during optimization. It is set as 0.01, 0.0005, 0.005 and 0.001 for CNN, Transformer, Attention and LSTM, respectively. This iterative process is performed until half of the original sequence was altered or a local optimum is achieved.

The model


Input peptide: x
Gradient: ∂y/∂x
Argmax projection
Optimized peptide: x1
LLVALGRFRVKIR
LLVALWRFRVKIR
G
Key amino acid
Input peptide: x1
Gradient: ∂y/∂x1
Argmax projection
Optimized peptide: x2
LLVALWRFRVKIR
LLVAWWRFRVKIR
L
Key amino acid
Input peptide: x2
Gradient: ∂y/∂x2
Argmax projection
Optimized peptide: x3
LLVAWWRFRVKIR
LLVAWWRFRVKIR
L
Key amino acid

References

1. Wang, B., Lin, P., Zhong, Y., Tan, X., Shen, Y., Huang, Y., Jin, K., Zhang, Y., Zhan, Y., Shen, D. and Wang, M., Explainable deep learning and virtual evolution identifies antimicrobial peptides with activity against multidrug-resistant human pathogens. Nature Microbiology, 10(2):332-347, 2025.
2. Witten, J. and Witten, Z., 2019. Deep learning regression model for antimicrobial peptide design. BioRxiv, p.692681.