Amanda Qi
acqi

Project Proposal: Generative Design of Anti-MRSA Peptides

Biomedical Problem and Motivation

Methicillin-resistant Staphylococcus aureus (MRSA) is a major global health concern responsible for life-threatening infections that are increasingly untreatable with conventional antibiotics. The slow pace of traditional antibiotic discovery and the rapid emergence of resistant strains highlight the need for new therapeutic approaches.

Antimicrobial peptides (AMPs) are short, naturally occurring proteins that kill bacteria by disrupting cell membranes or interfering with essential enzymes. They are less prone to resistance development than small-molecule antibiotics, which makes them promising candidates for new antibacterial therapies. However, experimental discovery of new AMPs is expensive and limited in scope.

Recent advances in generative AI allow for the computational exploration of peptide sequence space to propose novel candidates with antibacterial potential. This project aims to use generative protein language models to design new peptide sequences that share biophysical features of known anti-MRSA AMPs. The goal is to assess whether current sequence-generation models can meaningfully produce candidates that resemble known antibacterial peptides.

State of the Art

Large-scale protein language models such as ProGen2 can generate realistic protein-like sequences by learning statistical patterns from large biological datasets. Predictive tools such as AMP-scanner v3, AI4AMP, and AMPEP30 use deep learning to estimate antimicrobial potential directly from amino acid sequences. Structural predictors like ESMFold and AlphaFold 3 can identify secondary structures such as amphipathic alpha-helices that are common in functional AMPs.

Previous studies have identified specific features that make AMPs effective. Huan et al. (2020) found that most antibacterial peptides have a positive charge, an amphipathic alpha-helical structure, and moderate hydrophobicity, which together allow selective interaction with bacterial membranes. Arif et al. (2024) developed iMRSAPred, a model for predicting anti-MRSA peptide activity, and reported that successful peptides tend to have net charges between +1 and +5, balanced hydrophobicity, and aromatic residues such as phenylalanine.

These findings inform this project’s design strategy. Instead of retraining the generative model, I will guide sequence generation using biological context and post-generation filtering to bias results toward AMP-like properties.

Proposed Approach

1. Data Collection and Reference Analysis
   - Retrieve experimentally validated anti-MRSA peptide sequences from public databases such as DBAASP and APD3.
   - Analyze their key physicochemical properties, including charge, hydrophobicity, and amino acid composition, to establish reference ranges.
   - The dataset will serve as a reference for evaluation rather than for model fine-tuning.
2. Sequence Generation
   - Use ProGen2 in inference mode to generate peptide sequences between 10 and 50 amino acids.
   - Apply prefix conditioning such as “antimicrobial peptide | sequence:” to encourage AMP-like outputs.
   - Filter generated sequences based on charge and hydrophobicity values that fall within the range observed in real anti-MRSA peptides.
3. Activity Screening
   - Evaluate generated peptides using AMP-scanner v3 and AI4AMP to estimate antimicrobial activity.
   - Rank and retain the top candidates that score highly for predicted antibacterial probability.
4. Structural and Stability Evaluation
   - Use ESMFold to predict the secondary structure and calculate fold confidence (mean pLDDT).
   - Identify sequences predicted to form amphipathic alpha-helices that are associated with strong antibacterial activity.
5. Feature and Novelty Analysis
   - Calculate each peptide’s net charge, hydrophobic moment, molecular weight, and sequence length.
   - Compare these distributions with those of the reference anti-MRSA peptides.
   - Quantify novelty by computing sequence similarity and embedding distances to ensure that generated sequences are distinct from known ones.

Evaluation Plan

- Predicted Activity Rate: Percentage of generated peptides predicted to be active by AMP-scanner v3 or AI4AMP.

- Feature Alignment: Proportion of generated peptides whose charge and hydrophobicity values match those of experimentally validated AMPs.
- Structural Confidence: Average pLDDT score from ESMFold and visual inspection for alpha-helical formation.
- Novelty: Sequence similarity analysis using BLAST and embedding distance metrics to identify new sequence patterns.
- Justification: The activity and structure evaluation methods are consistent with established AMP research, including Arif et al. (2024) and Huan et al. (2020), ensuring a validated and literature-supported workflow.

Resources and Compute Needs

- Models and Tools: ProGen2 (Hugging Face), AMP-scanner v3, AI4AMP, and ESMFold (Google Colab or PSC Bridges-2).
- Hardware: 1 GPU (V100 or A100) with approximately 30 GPU-hours required for sequence generation and structure prediction.
- Storage: Up to 30 GB for sequences, model outputs, and predicted structures.
- Software Environment: Python 3.10 or higher with BioPython, Pandas, and Matplotlib for data analysis and visualization.

Timeline and Milestones

Week	Task
1	Collect and analyze experimental anti-MRSA peptide dataset.
2	Generate new sequences using ProGen2 with prefix conditioning.
3	Screen generated peptides for predicted activity with AMP-scanner v3 and AI4AMP.
4	Predict secondary structures and folding confidence using ESMFold.
5	Analyze physicochemical features, assess novelty, and visualize trends.
6	Prepare final report, visualizations, and poster presentation.

References

1. Arif M. et al. (2024). iMRSAPred: Improved Prediction of Anti-MRSA Peptides Using Physicochemical and Pairwise Contact-Energy Properties of Amino Acids. ACS Omega, 9(3), 3106–3118. https://doi.org/10.1021/acsomega.3c08303
2. Huan Y. et al. (2020). Antimicrobial Peptides: Classification, Design, Application and Research Progress. Frontiers in Microbiology, 11, 582779. https://pmc.ncbi.nlm.nih.gov/articles/PMC9476499/