**Title: Antimicrobial Peptide Discovery Using Computational Models**

**Names:**

* Yousef Abugosh : 315177444, yousef.abu-gosh@mail.huji.ac.il, yousefak
* Areen Mansour : 212900211, areen.mansour@mail.huji.ac.il, areen0507
* Nosiba Othman : 212378111, nosiba.otman@mail.huji.ac.il, nosiba
* Adan Mhameed : 214310666, adan.mhameed2@mail.huji.ac.il, adan.mhameed2

**Background:**

This research focuses on identifying **antimicrobial peptides (AMPs)** using machine learning techniques.  
AMPs are short protein sequences that exhibit **antimicrobial properties** and are potential alternatives to conventional antibiotics.  
The study leverages **computational models** to classify peptide sequences as antimicrobial or non-antimicrobial.

**Reference Article:**

📄 **Title:** AI-driven Antimicrobial Discovery  
🔗 <https://www.nature.com/articles/s41586-020-1990-9>  
**Summary:**  
This study demonstrates how **artificial intelligence** can be used to discover new antimicrobial peptides by analyzing **protein sequences** and predicting their activity against bacteria.

**Research Question and Objectives:**

**Research Question:**

How can machine learning models optimize the identification and classification of antimicrobial peptides?

**Objectives:**

* Utilize the **ESM-2 Transformer model** for feature extraction from protein sequences.

**Data Description:**

**Data Includes:**

* A labeled dataset of peptide sequences classified as **AMPs** or **non-AMPs**.
* Data format: **CSV files** containing peptide sequences and labels.

**Data Sources:**

📂 **DBAASP - Database of Antimicrobial Peptides** 🔗 <https://dbaasp.org/>  
📂 **Kaggle Antimicrobial Peptide Dataset** 🔗 <https://www.kaggle.com/datasets/ashishreddy08/antimicrobial-peptide-amps-dataset>

**Hypothesis:**

Using **transformer-based models (such as ESM-2)** for peptide sequence analysis will result in **higher accuracy** in identifying AMPs compared to traditional classification techniques.

**Computational and Statistical Methods:**

* **Model:** Transformer-based classification using **ESM-2**.
* **Feature Extraction:** Representing peptide sequences as **vector embeddings**.
* **Machine Learning:** Training a **Random Forest / Logistic Regression model** for AMP classification.
* **Validation Metrics:** Accuracy, Precision, and AUC-ROC.
* **Software Tools:** Python, TensorFlow/PyTorch, Biopython, and Streamlit for web app development.

**Execution Plan:**

1️ **Data Collection:** Obtain labeled peptide sequences from public databases.  
2️ **Feature Extraction:** Convert peptide sequences into embeddings using **ESM-2**.  
3️ **Model Training:** Train a classifier to distinguish AMPs from non-AMPs.  
4️ **Model Evaluation:** Measure **accuracy, and AUC-ROC**.

**References:**

📄 **AI-driven Antimicrobial Discovery** – Nature (2020) 🔗 <https://www.nature.com/articles/s41586-020-1990-9>  
📂 **ESM-2 Transformer for Protein Analysis** 🔗 <https://github.com/facebookresearch/esm>  
📂 **Antimicrobial Peptide Classification** 🔗 <https://github.com/itsmeh/antimicrobial-peptides>