1. **Title:** Antimicrobial Peptide Discovery Using Computational Models
2. **Names:**

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

**3. Background:** This research focuses on identifying antimicrobial peptides (AMPs) using machine learning techniques. Amps are short protein sequences that exhibit antimicrobial properties and serve as potential alternatives to conventional antibiotics.  
The study leverages computational models to classify peptide sequences as antimicrobial or non-antimicrobial, addressing a critical need in combating antibiotic resistance.

**4. Research Question and Objectives:**

**Research Question:** How can machine learning models optimize the identification and classification of antimicrobial peptides?

**Objectives:**

1. Utilize the ESM-2 Transformer model for feature extraction from protein sequences.
2. Compare the performance of transformer-based models against traditional classifiers to validate their accuracy and robustness.
3. Explore potential correlations between specific amino acid motifs and antimicrobial properties.
4. Evaluate the scalability of the approach for larger datasets and diverse microbial targets.

**5. Data Description:**

* **File Format:**  
  The data is formatted in the FASTA format, commonly used for storing biological sequences such as DNA, RNA, or proteins.  
  Each sequence entry begins with a header line starting with a greater-than symbol (>), followed by a unique identifier or description. The subsequent lines contain the amino acid sequence of the peptide.
* **Content:**  
  The file comprises peptide sequences represented by single-letter amino acid codes (e.g., A, C, D, E).
* **Data Source:**  
  The data originates from the [Antimicrobial-Peptides repository](https://github.com/zswitten/Antimicrobial-Peptides) on GitHub.
* **Data Origin:**  
  The repository aggregates Minimum Inhibitory Concentration (MIC) data for antimicrobial peptides from multiple sources. This includes the GRAMPA database, which contains peptide sequences and their antimicrobial activities against various bacterial strains. Key columns include bacterium, sequence, strain, and MIC value.

**6. Hypothesis:** Using transformer-based models (e.g., ESM-2) for peptide sequence analysis will yield higher accuracy and better generalizability in identifying AMPs compared to traditional classification techniques.

**7. Computational and Statistical Methods:**

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

**8. Execution Plan:**

1. **Data Collection:** Obtain labeled peptide sequences from public databases such as APD, DBAASP, and GRAMPA.
2. **Feature Extraction:** Convert peptide sequences into embeddings using the ESM-2 model.
3. **Model Training:** Train classifiers to distinguish AMPs from non-AMPs using Random Forest and Logistic Regression.
4. **Model Evaluation:** Evaluate the models using metrics such as accuracy, recall, precision, and AUC-ROC.
5. **Application Development:** Develop a Streamlit-based web application for real-time peptide classification.

**9. References:**

* Mizuno, F., Kumagai, M., et al. (2017). Imputation approach for deducing a complete mitogenome sequence... *Journal of Human Genetics*, 62, 631–635. DOI:10.1038/jhg.2017.14.
* [PMC Article on AMP Prediction](https://pmc.ncbi.nlm.nih.gov/articles/PMC9126312)
* [MITOIMP Repository](https://github.com/omics-tools/mitoimp/tree/master)
* [MLpeptide Repository](https://github.com/reymond-group/MLpeptide)
* [Antimicrobial-Peptides Repository](https://github.com/zswitten/Antimicrobial-Peptides)

**10. Bonus:**  
We are a diverse group of males and females, comprising one neuroscientist and computer science engineer, along with three computer science engineers.