Paolo Federico – s212975

Predicting Phage-Host Interaction with Language Models

Background: Recent advancements have highlighted the efficacy of language models over traditional supervised machine learning approaches in the realm of protein structure prediction. Traditional methods rely on evolutionary information sourced from multiple sequence alignment, a process which can be computationally intensive. Conversely, language models have the capacity to derive information directly from amino acid sequences.

Objective: This research aims to explore the capabilities of language models in predicting protein-protein interactions. The primary focus will be on assessing the potential interactions between the proteins of a phage and its bacterial host. Successfully predicting such interactions can be indicative of a virus's ability to infect the microorganism.

Methods:

1. **Embedding Creation:** We will generate embeddings derived from the association of receptors found on the host's outer membrane and the receptor-binding proteins of the virus.
2. **Protein Interaction Predictor Training:** These embeddings will serve as the primary input for training the interaction predictor. To facilitate effective training, the dataset will be augmented with negative examples to represent non-interactions.

The whole process will be employed in two distinct scenarios:

a. Utilizing proteins that have been experimentally validated.
b. Sourcing phage-host protein pairings from bacterial genomes that contain prophages. Here, the putative bacterial proteins and corresponding viral proteins will be respectively extracted from the genome and the prophage.