**Predicting Phage-Host Interaction with Language Models**

Project plan

**Danmarks Tekniske Universitet – DTU  
Department of Health Technology, Bioinformatics**

**Paolo Federico – s212975  
Bioinformatics and Systems Biology**

**Henrik Nielsen (Assoc. Prof. – Supervisor)  
Nicholas Taylor (Assoc. Prof. – Co-supervisor)  
Simon Rasmussen (Assoc. Prof. – Co-supervisor)  
Felix Teufel (PhD – Co-supervisor)**

**September 26th, 2023**

# Introduction

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 single amino acid sequences. Most noticeably, such models can exploit the information encoded in the vast number of unlabeled sequences used for their pre-training. On top of that, language models allow protein-specific predictions, as opposed to other methods which rely on family-averaged alignments.

The information encoded from neural networks are finding a growing number of applications, from protein structure estimation *[1]* to host prediction from a viral protein *[2]*. State of the art results make use of embeddings, which are protein information encoded via transformers architecture like protBERT *[1]*.

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

An in-depth investigation of the phages infecting *Escherichia coli* was conducted by *Maffei et al [4].* The data they collected concerns the K-12 MG1655 ΔRM strain and will be the foundation for training and testing this project’s neural model.

The first task requires the labelling of virus-host protein pairs as interacting or not. Such information will be inferred from the occurred infection and knowledge about receptor–receptor binding protein interaction.

Secondly, negative examples of expectedly non-interacting proteins might be added to the dataset in order to reinforce the model understanding of correct protein interaction. On top of that, the protein-class distribution will be assessed to avoid any bias.

The next step contemplates the model realization. The architecture will be based on the work of Gonzales et al, where the host genus is computed from a viral receptor-binding protein embedding [2]. On the other side, the here proposed project’s aim concerns the compatibility prediction of a virus-host protein pair. In detail, the architecture will probably see a protT5 embedding module [1] followed by a random forest classifier.

Lastly, an additional volume of protein sequences will be obtained from ***E. coli*** genomic sequences containing prophages. Infection from these phages is taken heuristically for granted. Further data processing and model training will follow.

# Side projects

The subject is exciting and open to various possibilities. Hereby are listed some possible developments:

* Consider phage interaction with the lipopolysaccharides
* Improve biological knowledge increasing embeddings interpretability - e.g. projection on lower dimensions
* Move the focus on other phylogenetic elements
* Estimate confidence on predictions via evidential learning
* Use protein sets as input to estimate infection instead of protein-protein interactions

# Project timeline

Immagine che contiene testo, schermata

Descrizione generata automaticamenteImmagine che contiene testo, Carattere, numero, linea

Descrizione generata automaticamentePeriod: 28/08/2023 – 28/01/2024

Figure 2 - Project report writing

Figure 1 - Data handling and model development for phages and prophages

# References

1. Elnaggar et al. – DOI: 10.1109/TPAMI.2021.3095381
2. Gonzales et al. – DOI: 10.1371/journal.pone.0289030
3. Maffei et al. – DOI: [10.1371/journal.pbio.3001424](https://doi.org/10.1371/journal.pbio.3001424)