

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 20th, 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 amino acid sequences.

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 [3].

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 K12 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 not 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 [2]. In detail, it will probably see a protT5 embedding module followed by a random forest classifier.

Lastly, an additional volume of protein sequences will be obtained *E. coli* genomic sequences containing prophages. Infection from these phages is taken 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 lipopolysaccharide
- 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

REFERENCES

1. Elnaggar et al. – DOI: 10.1109/TPAMI.2021.3095381
2. Gonzales et al. – DOI: 10.1371/journal.pone.0289030
3. Linial et al. – DOI: 10.1093/bioinformatics/btac020
4. Maffei et al. – DOI: 10.1371/journal.pbio.3001424

PROJECT TIMELINE

Period: 28/08/2023 – 28/01/2024

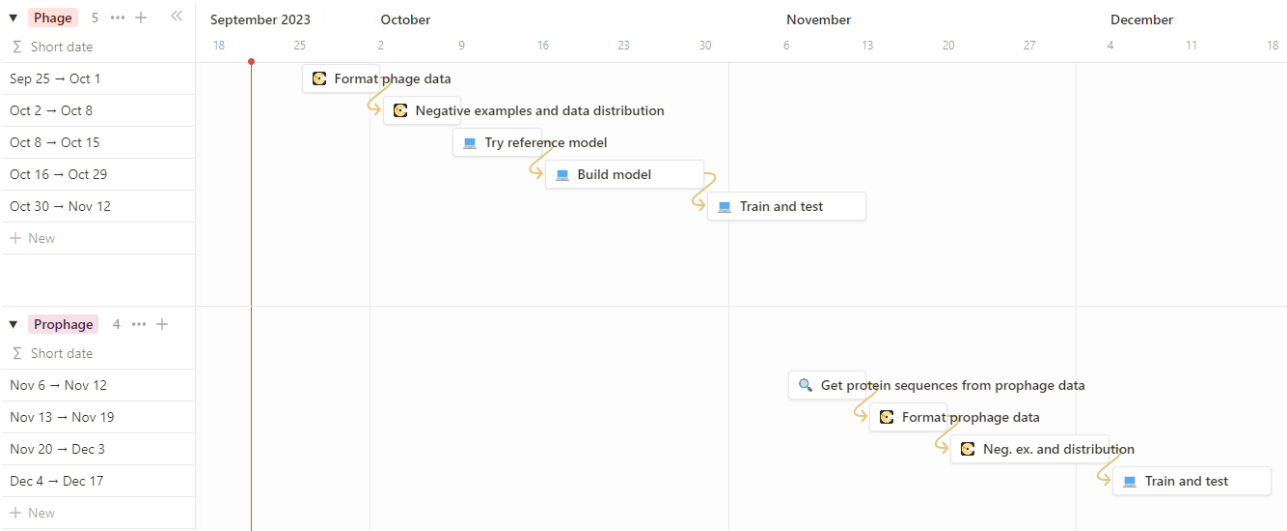


Figure 1 - Data handling and model development for phages and prophages

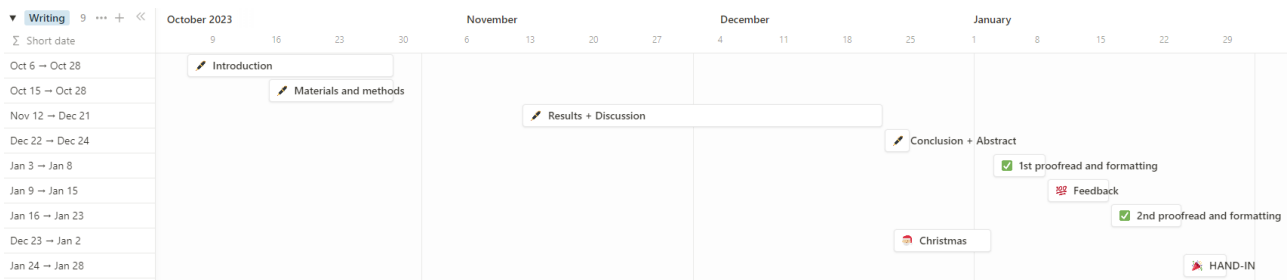


Figure 2 - Project report writing