Language Models for Protein Feature Prediction Tasks

Dr Daniel Buchan
daniel.buchan@ucl.ac.uk

Background

With the recent success of protein deep learning models such as AlphaFold2 for protein in protein structure prediction there is growing interest in further applications of deep learning in Protein Biochemistry. Of particular interest are protein language models. Biological sequences such as genes, DNA and proteins are easily represented as as sequences of tokens. As a result there is a great deal of interest in protein language models as a means of modelling proteins and protein evolutionary information.

And number of new protein Large Language Models have recently been made available which are capable of encoding a great deal of evolutionary information about protein evolution and protein families. We can use such embeddings to generate new, novel predictors of protein function and structure tasks. One exicting avenue of research is to use such models to traditional Sequence Search methods.

In these projects we would investigate the expressive potential of language embeddings to generate new predictors for protein structure and function.

Project 1

Proteins are complex sequences of amino acids. These sequences encode biologically relevant information about the function and structure of the protein, usually referred to as **Protein Sequence Features**. Common features include; protein domains, active sites, binding sites, signal peptides, disordered regions, metal binding sites and so forth. These sub-sequences have evolutionary shared [degenerate] patterns and as a consequence it has long be recognised that such sequence features can be predicted in protein sequences.

This project would use pre-trained protein LLMs to predict protein sequence features. There is an increasingly rich literature in this field and we would seek to build a novel predictor for one or more protein sequence features. If there is time in the project we would seek to compare the performance of sequence feature predictors built using different available protein LLMs

Project 2

Protein LLMs are typically trained using attention based losses using a masked amino-acid training task. An alternative, natural representation of proteins is as a series of protein domains. Under this view we can regard domains as sequential “words” or tokens along the length of a protein chain. And in turn we can view proteins as a form of psuedo-sentence. This project seeks to build a protein domain based LLM.

In the 2nd phase of the project we would demonstrate that such a ‘Domain LLM’ is capable of predicting protein function as encoded a Gene Ontology terms.

Project 3

Word2vec is a tool developed by google for developing semantic embeddings for words. We have previously used this tool on a corpus of pseudo-sentences that represent proteins to learn and embedding of protein domains. Since then a number of other applications of this method have been been applied to develop protein domain embedding (dom2vec, prot2vec, SPVec).

In this project we would investigate optimising a new word embedding for proteins and using the features to make predictions of protein function. The project would then demonstrate the utility of this embedding by either developing a novel protein predictor or investigating the embedding to demonstrate that it does encode biologically meaningful information

Requirements

Python programming

Introductory Machine Learning or data science (such as COMP088 or COMP081 or equivalent)

Good to know

Introductory Protein biochemistry (https://www.ebi.ac.uk/training/online/courses/protein-classification-intro-ebi-resources/)

Introductory Bioinformatics (such as COMP082 or equivalent)

Knowledge of HPC programming

Pytorch (or equivalent)

Refs and Sample Reading

1. **MSA Transformer**
   Rao, R *et al*
   https://www.biorxiv.org/content/10.1101/2021.02.12.430858v3
2. **Ankh ☥: Optimized Protein Language Model Unlocks General-Purpose Modelling**
   Elnaggar A *et al*
   https://www.biorxiv.org/content/10.1101/2023.01.16.524265v1
3. **Learning a functional grammar of protein domains using natural language word embedding techniques**
   Buchan & Jones
   https://onlinelibrary.wiley.com/doi/10.1002/prot.25842
4. **Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences**
   Rives et al.
   https://www.pnas.org/doi/10.1073/pnas.2016239118
5. **ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning**
   Elnaggae et al.
   https://ieeexplore.ieee.org/document/9477085
6. **Transformer-based deep learning for predicting protein properties in the life sciences**
   eLife 12:e82819.
   https://elifesciences.org/articles/82819
7. **Computational Prediction of Protein Intrinsically Disordered Region Related Interactions and Functions.**
   Genes.2023; 14(2):432. https://doi.org/10.3390/genes14020432https://www.mdpi.com/2073-4425/14/2/432
8. **Protein language-model embeddings for fast, accurate, and alignment-free protein structure prediction**
   Weissenow K *et al*
   https://pubmed.ncbi.nlm.nih.gov/35609601/
9. ***An Analysis of Protein Language Model Embeddings for Fold Prediction***
   Villegas-Morcillo A et al
   https://www.biorxiv.org/content/10.1101/2022.02.07.479394v1.full
10. *https://www.mdpi.com/1999-4893/14/1/28*
11. *https://academic.oup.com/bioinformatics/article/34/15/2642/4951834*