Master Thesis Project Description (1200 words)

This projects revolves around applying representation learning to protein sequence modelling. Previous attemps have applied language models to protein sequences (source), making latent representations of proteins, reaching both levels of performance comparable to the currently best known, and faster computation. A recent study (source: TAPE article) emphasizes the importance of representations in the performance on downstream tasks.

Such a representation allows for comparison between sequences directly in their latent space representation, instead of current similarity measures, such as pairwise sequence alignments (?). Thus a good representation of protein sequences is important.

We wish to examine the performance of representations and their implications for protein informatics by making an unsupervised representation using tools from machine learning and language modelling such as variational autoencoders and deep neural networks. The results will be compared with current state-of-the-art performance on downstream tasks.

Representations are important not only for protein informatics, but also language modelling, as the protein domain is one of the richest in terms of data and structure.

Initially, we wish to inspect the unsupervised UniRep model ([1]) which compresses protein sequences into a fixed length vector, representing the sequence in a latent space. One key aspect is that the representations created by UniRep corresponds to the internal state of the LSTM used for training. In this project, we wish to make a proper representation instead, meaning that the representation itself is the output, not the internal state.

Finally, current work on representations of proteins might carry potential toward a general representation of proteins, implying that general unseen protein sequences can be analyzed with such a representation with respect to functionality or structure, or at least infer commonalities from their representations. We wish to explore such implications.

Learning from protein sequences is also desirable in terms of data, as it would be able to utilize the growing number of protein sequence data that is being aquired. That is, there is a lot of unlabelled data and little labeled.

In addition, this field of research, being the understanding and formulation of protein sequence space, is comparatively new, allowing for many different ventures to be explored. This project is one step in that direction.

References

[1] Ethan C Alley et al. “Unified rational protein engineering with sequence-only deep representation learning”. In: *bioRxiv* (2019), p. 589333.