FACULTY OF SCIENCE
UNIVERSITY OF COPENHAGEN

Deep Unsupervised Representation Learning on Protein Sequences using Variational Autoencoders

Master's Thesis

Emil Petersen
[empe@di.ku.dk]

Victor Nordam Suadicani
[swl460@alumni.ku.dk]

2019/11/20

Supervisors: Wouter Krogh Boomsma

Master Thesis Project Descriptions

Representation learning is the task of training a machine to produce a suitable representation of features for the desired input, as opposed to manually engineering the representation. Good representations are important in order to learn useful properties of the given data.

This project revolves around applying representation learning to protein sequences. Previous attempts 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 [1] emphasizes the importance of representations in the performance on downstream tasks.

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

Protein sequences are comparable to natural language sentences, in the sense that both consists of a sequence of symbols. In the case of sentences, the symbols are the letters of the alphabet, and in the case of proteins, the symbols are amino acids. Thus one might look at the results of natural language sentence representations in order to learn about protein sequence representation. A study by Google Brain [2] showed that it is possible to represent entire sentences as single vectors in a latent space, allowing interpolation between sentences. Additionally, the latent space inherently maps common properties of sentences, such as style and topic. It is possible that a similar representation for proteins may be able to map properties of proteins, such as secondary structure.

Unsupervised learning is particularly suited for protein machine learning, since there is an overweight of unlabeled data in comparison to labeled data. If a powerful unsupervised model could be produced, the wealth of this data could be used effectively [3]. Therefore, unsupervised or variants like semi- or self-supervised learning methods could be especially useful.

Initially, we wish to inspect the unsupervised UniRep model [4] 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. The hope is that such a model would produce an overall better representation, with better performance on standard benchmarks (?).

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.

References

[1] Roshan Rao et al. “Evaluating Protein Transfer Learning with TAPE”. In: *arXiv preprint arXiv:1906.08230* (2019).

[2] Samuel R Bowman et al. “Generating sentences from a continuous space”. In: *arXiv preprint arXiv:1511.06349* (2015).

[3] Mohammed AlQuraishi. “The Future of Protein Science will not be Supervised”. In: (2015). URL: https://moalquraishi.wordpress.com/2019/04/01/the-future-of-protein-science-will-not-be-supervised/.

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