Deep Unsupervised Representation Learning on Protein Sequences using Variational Autoencoders

Emil Petersen
[empe@di.ku.dk]

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

2019/11/28

Background

Understanding how mutations in amino acid sequences affect protein structure and function is a central challenge in computational biology. Machine learning has proven to be a useful tool in the analysis of protein sequences. One form of machine learning that has been effectively applied to this problem is **representation learning**.

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.

Previous attempts at **protein sequence learning** have shown that fundamental structural features, that capture the function of a given protein, can be learned and represented from raw protein sequences by using Deep learning [1]. Specifically, using a **recurrent neural network** to summarize any protein sequence into a fixed length vector and subsequently averaging, such statistical representations can be discerned from large sets of input data. Additionally, such approaches have been shown to improve performance in nearly all models on downstream tasks [2].

Protein sequences are **comparable to natural language sentences**: both consists of a sequence of symbols. Sentences consists of the letters of the alphabet, while proteins 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 [3] showed that one can 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. If a powerful unsupervised model could be produced, the wealth of this data could be used effectively [4]. Therefore, unsupervised or variants like semi- or self-supervised learning methods could be useful.

Unsupervised representation learning has recently been applied to protein sequences [1] in order to achieve a latent space representation. However, current approaches have extracted such representations, as the hidden internal state of a recurrent layer in the learning network. This makes it difficult to incorporate notions of uncertainty, and other essential properties that we would normally impose on such representations. That is, current approaches are not generative, and so we cannot generate protein sequences corresponding to any point in latent space.

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.

Aims and Method

This project revolves around applying representation learning to protein sequences. 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 such as variational autoencoders and deep neural networks. The results will be compared with current state-of-the-art.

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. The hope is that such a model would produce an overall better representation, with better performance on standard benchmarks.

The latent space defined by such a representation might allow for an understanding of protein sequences, by defining a mapping from sequences to protein properties, such that similar proteins are close together in the space (see figure 1). Previously unseen points could correspond to unseen protein sequences, and properties of these proteins could be inferred from the latent space.

In addition, it will be useful to use representations to explain the behavior of variants of protein that we know, but we need to optimize. For example, the representation is important to help explaining the experimental results of variants screening, trying to map the effect of mutations, especially in case when we do not have a structure.

Learning Objectives

1. Present the theory behind variational autoencoders and deep representation learning.
2. Survey similar approaches (such as [1]) within the field of representation learning on protein sequences and discuss how they relate to the presented theory.
3. Explore the theoretical strengths and weaknesses of model architectures. In addition, discuss the trade-offs between the latent spaces of different models.
4. Design, implement and evaluate representation learning models on protein sequences, using variational autoencoders. Argue for the underlying design and implementation choices and analyze the performance.
5. Discuss how a well-performing representation learning model on protein sequences can be used for exploring new proteins and their properties, and other potential applications, if any.

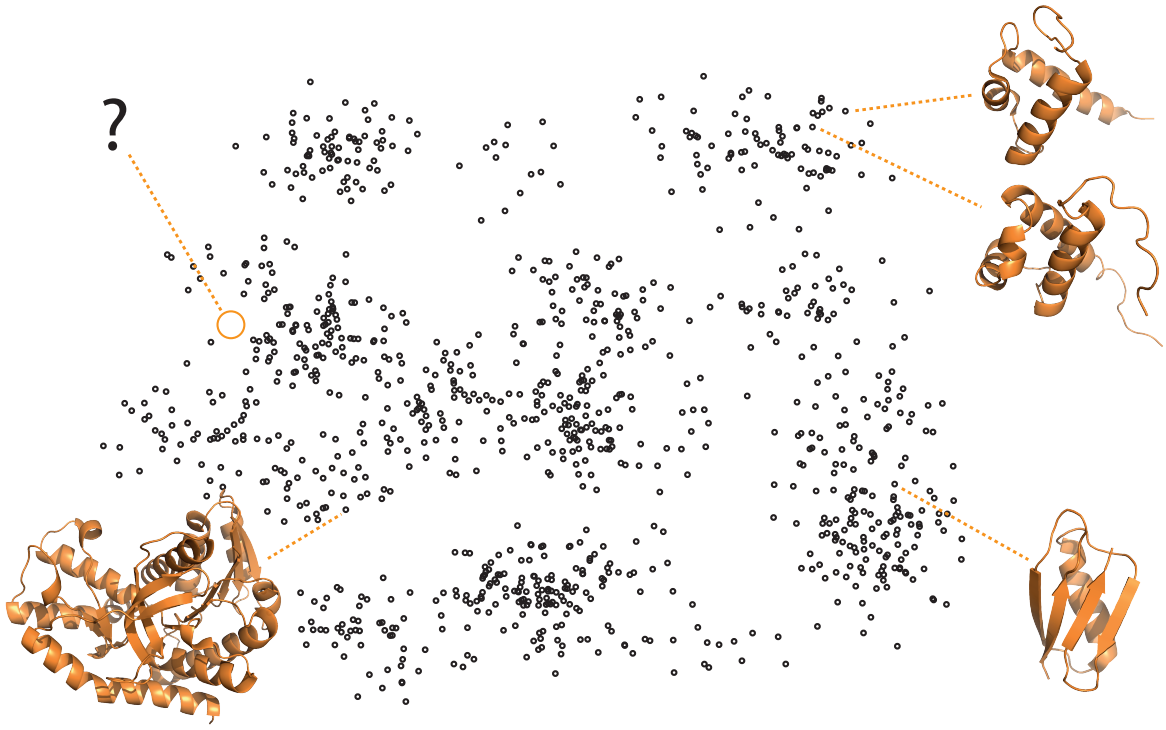

Figure 1: A 2-dimensional view of a protein sequence representation. Similar proteins have representations that are close together. With such a representation, small changes in protein can be explored in the space by looking at points close to the protein. Properties of unknown sequences (represented here by the question mark) could potentially be explored by examining the representation space.

References

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

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

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

[4] 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/.