Semi-supervised and unsupervised analysis of protein structures

Bioinformatics project, FTE40306 Advanced Machine Learning

Introduction

In the life sciences, high-throughput sequencing has provided an abundance of DNA and RNA measurements. Consequently, we now have genome sequences for a wide range of organisms, as well as the expression levels of genes in various tissues and under different conditions. This increase in data and associated bioinformatics developments have led to a revolution in our understanding of genes and genomes, often fueled by machine learning [1]. However, most functions in a cell are performed by proteins, whose 3D structures determine their actual function, localization, stability, and interactions. While protein sequences are relatively easily measured or derived from gene sequences, protein structures are much harder to measure and work with. Therefore, much research has gone into the development of sequence-based machine learning predictors for protein properties. Such predictors are essential to guide experiments designed to unravel the workings of the living cell and interpret the results. Application of sequence-based machine learning has been successful to some extent, but the approach is born out of the necessity of not having sufficient actual protein structure data available and the challenges involved in working with 3D data. The direct incorporation of structure data in machine learning approaches has indeed been shown to improve predictive power, e.g. [2].

The use of protein structures in computational biological research has thus far been underdeveloped compared to the use of sequences, mainly due to a lack of experimental structures and the challenges of working with 3D data. Recently however, the number and quality of available protein structures has increased (Fig. 1) which has supported the development of deep learning-based protein structure predictors with accuracies in the same range as measurements, such as AlphaFold2, RoseTTAFold, ESMFold etc. AlphaFold2 has been used to create a database that now holds over 200 million predicted protein structures.

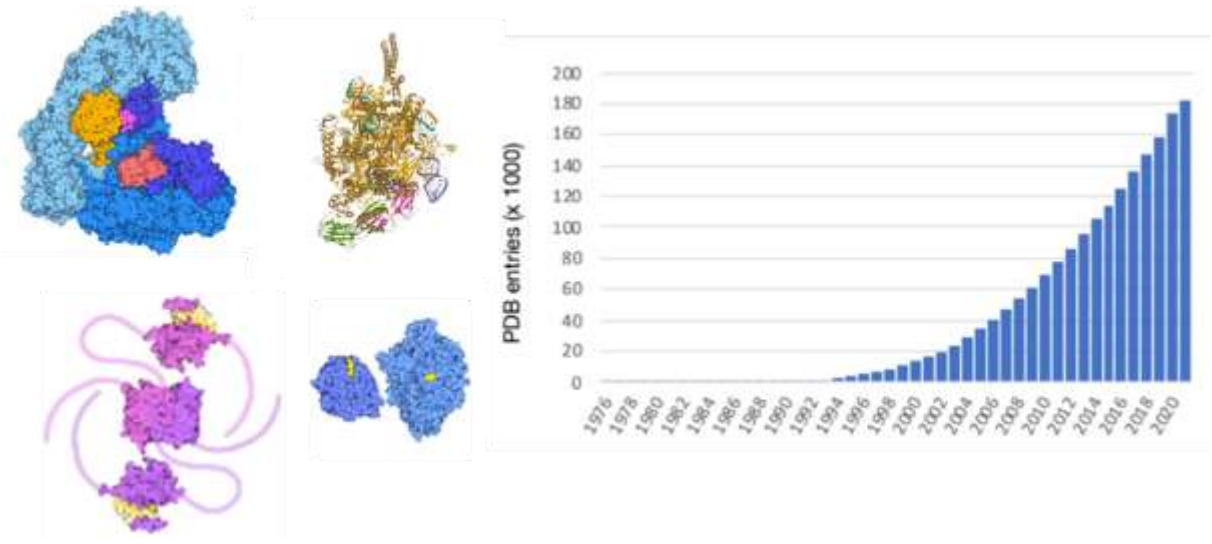

Figure 1. Example protein structures from the protein databank (PDB) at www.rcsb.org (left) and the number of protein structure entries in the PDB (right).

This sudden wide availability of (predicted) protein structures leaves the need for methods that are suited to work with them. A major challenge is how to best represent protein structures as input for machine learning. In particular, there is a huge interest in generating so-called “embeddings” using

unsupervised (or self-supervised) approaches such as autoencoders (Fig. 2). Such approaches have already been applied to DNA and protein sequences, allowing to capitalize on the large amounts of unlabelled data available. In this project, we will investigate different semi- and unsupervised algorithms and explore the development of a generic protein structure embedding method. This would be enormously useful in the context of e.g. protein function prediction.

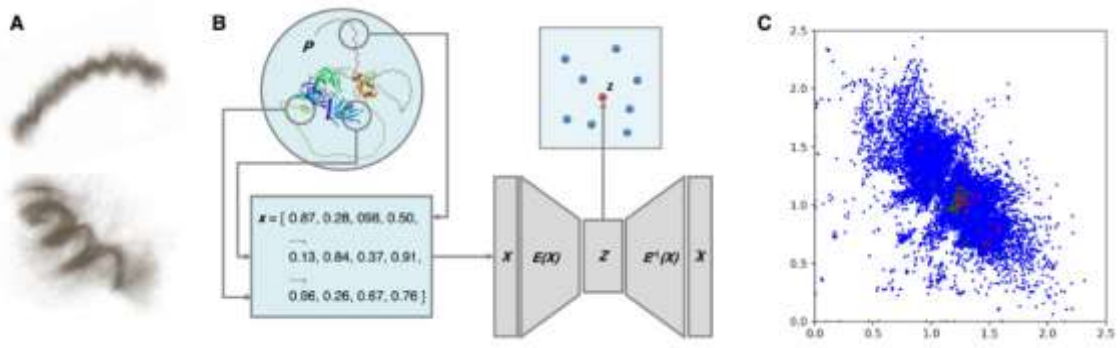

Figure 2. (A) Example of local sub-structures, shape-mers, which can be obtained from structures [3]. (B) Schematic overview of approach: the structure of a protein *P* is represented as a vector of shape-mer counts ***x*** which serves as input to an autoencoder to obtain an embedding ***z*.** (C) An example embedding of ~25,000 protein structures obtained by training a simple autoencoder on shape-mer counts; example functional annotations are indicated in green ("DNA recombination") and red ("ATPase activity").

The assignment

The goal of this project is to investigate different semi-supervised and unsupervised approaches to analyze protein structures at large scale and obtain low-dimensional representations. Such approaches are useful for visualization (e.g. to obtain insight into relations between protein families), to reduce storage, and as pre-training for supervised learning. The initial representation of the proteins that you will use as starting point is based on so-called "shape-mers", count vectors of translation and rotation invariant descriptors of the structure of local protein fragments [3]. Although this representation is less high-dimensional than the original set of 3D coordinates, it is still quite high-dimensional, and the aim of this project is to further reduce the number of dimensions.

In the Advanced Machine Learning course, you learned methods to deal with the above-mentioned challenges, in particular, kernel PCA, t-SNE/UMAP and autoencoders. Your task is to reason about the problem and investigate methods that you think are relevant to solve the assignment. There are some hints to get you started in the notebook **`project_structures.ipynb`**.

Data

Data are provided for training and learning (see the above-mentioned notebook):

- **`pretrain.embedding.edit`** – 9,229 proteins, 1,648 descriptors (shape-mer counts) per protein
- **`pretrain.labels.edit`** – identifiers for the proteins, and information on the sequence length
- **`pretrain.go`** – for the 9,229 proteins, information on whether they are involved in three specific functions (indicated by 0 for no and 1 for yes in columns 2-4). The functions are “localized in the membrane”, “binding to ATP, adenosine 5'-triphosphate (a universally important coenzyme and enzyme regulator)” and “DNA binding”.

Note that the three files contain information on the same set of proteins (and the proteins are ordered in the same way in each of the three files).

Steps

1. Visualize and analyze the contents of the provided datasets. Think of possible filters to apply (e.g. for shape-mers which occur very few times, or a lot of times; also, for very small or very large proteins).
2. Train one or more different autoencoders on this data. Analyze the performance of the methods, and how this depends on settings/hyperparameters.
3. Visualize one of the obtained latent representations (in step 2) using t-SNE or UMAP. This will allow to obtain a 2D representation of the latent representation.
4. Make a supervised prediction model (try both a simple model such as e.g. a decision tree, and a more complex model) for the three gene functions. Compare the performance using the original features with the performance using the newly obtained latent representation(s).
5. **Optional, if time allows**: apply kernel PCA on this data.
6. Write report.

Assessment

The project will be assessed by clarity and quality of both the notebook (50%) and the report (50). The report will be assessed on:

- clarity of text
- overall approach of comparing models
- explanation of methods
- motivation of methods/settings
- experiments and results
- presentation of results
- interpretation of results
- conclusion/discussion

References

[1] https://www.sciencedirect.com/science/article/pii/S2589004220310877
[2] https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008197
[3] https://academic.oup.com/bioinformatics/article/36/Supplement_2/i718/6055902