A Quick Review of Latest Developments in the Application of Advances in Natural Language Processing (NLP) to Protein Sequence and Structure Modelling

CS593A (Fall 2020) Term Paper

Somadina Mbadiwe

Project Advisor: Dr. Vasudevan Jagannathan

Lane Department of Computer Science and Electrical Engineering
West Virginia University
Morgantown WV, USA

December 7, 2020

Contents

1 Introduction

A great diversity of cells exist in nature, but they all have some common features. All cells have a life cycle: they are born, eat, replicate, and die. During a cell's life cycle, it has to make a lot of important decisions . However, cells do not have brains so these decisions are made possible through complex networks of chemical reactions, called pathways, that synthesize new materials, break other materials down for spare parts, or signal that it is time to eat or die [15].

All life on this planet depends on three types of molecule: DNA (deoxyribonucleic acid), RNA (ribonucleic acid), and proteins. Roughly speaking, a cell's DNA holds a vast library describing how the cell works. RNA acts to transfer certain short pieces of this library to different places in the cell, at which point those smaller volumes of information are used as templates to synthesize proteins. Proteins form enzymes that perform biochemical reactions, send signals to other cells, form the body's major components, and otherwise perform the actual work of the cell [13]. This relationship between DNA, RNA and protein is popularly known as *the central dogma in molecular biology* or, informally, *the central dogma of life* [7].

Proteins are responsible for almost all biological processes critical to life. Hemoglobin carries oxygen to your cells; insulin regulates your blood glucose levels; and rhodopsin helps you see. It even extends past life itself. Proteins have been used in industrial settings to break down plastic waste and create laundry detergents. The rest of the article reviews the recent advances in studying proteins.

2 Background

The DNA is a long molecule consisting of four types of bases: *adenine* (A), *thymine* (T), *guanine* (G), and *cytosine* (C). RNA also has four bases, three of which are the same with DNA bases. Thymine from DNA is replaced with *uracil* (U) [15]. Thus DNA data is typically a sequence of strings constructed using only four letters: A, C, G, and T. For RNA, the letters are A, C, G and U.

Proteins are also represented as sequence of strings but its sequences are constructed using 25 letters: 20 characters for the standard amino acids, 2 for the non-standard amino acids selenocysteine and pyrrolysine, 2 for ambiguous amino acids, and 1 for when the amino acid is unknown [12]. This model of representation (a sequence of discrete symbols) enables us to treat these molecules as a *language* and allows us to model them using machine learning architectures developed for natural language. Protein sequences, for instance, are in many ways comparable to text: discrete symbols (amino acids), dictionary of symbols (similar to characters) and access to large databases of unlabelled sequences (similar to news publications or Wikipedia).

Beyond its encoding as a sequence $(x_1, ..., x_L)$, a protein has a 3D molecular structure. The different levels of protein structure include primary (amino acid sequence), secondary (local features), and tertiary (global features) [13].

3 Datasets

There are generally two kinds of datasets for studying proteins. One is unlabelled data containing protein sequences usually. The size is typically in the tens oto hundreds of millions. Such datasets are typically used for training protein-based language models. The other kind are labelled data curated with particular task(s) in mind. See Section 5.3 for a description of tasks usually studied in protein science. These labelled data are typically small - in the thousands or tens of thousands. They can be used purely as test dataset to evaluate a model; or split into the usual train / validation / test splits and used to train or finetune a model. Below, we discuss some of the datasets in use.

3.1 Protein Data Bank

Protein Data Bank (PDB) was established in 1971 as the first open-access, molecular data resource in biology [5]. Today, it has become a core data resource essential for understanding the functional roles that macromolecules play in biology and medicine. In fact, most scientific journals will not publish new macromolecular structures unless the authors deposit to the PDB the 3D atomic coordinates comprising the structural model plus experimental data used to derive the structures and associated metadata; and many governmental and non-governmental research funding agencies also require PDB deposition of unpublished macromolecular structure data [1].

The main challenge of using data from PDB, from a machine learning perspective, is that it comes as raw protein structures, requiring post-processing before they are usable by machine learning frameworks.

3.2 UniRef

The UniRef databases (UniProt Reference Clusters) [24] provide clustered sets of sequences from the UniProt Knowledgebase and selected UniParc records to obtain complete coverage of sequence space at several resolutions (100%, 90% and 50% identity) while hiding redundant sequences. The UniRef100 database combines identical sequences and subfragments from any source organism into a single UniRef entry (i.e. cluster) with 216 million protein sequences (80 billion amino acids) and requiring 150GB of disk space. UniRef90 and UniRef50 are built by clustering UniRef100 sequences at the 90% or 50% sequence identity levels. UniRef a powerful alternative to native sequence databases for similarity searches and in using those searches in functional annotation.

3.3 BFD

BFD is a dataset that merged all protein sequences available in UniProt [8] and proteins translated from multiple metagenomic sequencing projects, making it the largestcollection of protein sequences available at the time of writing. The original BFD set contained several copies of identical sequences; only one of those was kept, resulting in a subset with 2.1 billion protein sequences (with >393 billion amino acids requiring 527GB of disk space as text).

3.4 Pfam

Pfam [9] is a database of thirty-one million protein domains used extensively in bioinformatics. It can be used as pretraining corpus for a protein-based LM. Sequences in Pfam are clustered into evolutionarily-related groups called *families*. The latest version (as of December 2020) is 33.1 and contains 18259 entries (families).

3.5 ProteinNet

ProteinNet [3] is a standardized data set for machine learning of protein structure. It provides protein sequences, structures (secondary and tertiary), multiple sequence alignments (MSAs), position-specific scoring matrices (PSSMs), and standardized training / validation / test splits. ProteinNet builds on the biennial CASP assessments, which carry out blind predictions of recently solved but publicly unavailable protein structures, to provide test sets that push the frontiers of computational methodology. It is organized as a series of data sets, spanning CASP 7 through 12, to provide a range of data set sizes that enable assessment of new methods in relatively data poor and data rich regimes.

ProteinNet integrates sequence, structure, and evolutionary information in programmatically accessible file formats tailored for machine learning frameworks like Tensorflow and PyTorch. Two file formats are avalable: text-based and TFRecords.

3.6 NetSurfP-2.0

NetSurfP-2.0 dataset is a labeled dataset prepared and released as part of developing the the NetSurfP-2.0 LM [16]. (See Section 4.4 for a description of the model). The dataset contains 10,837 sequences and is publicly available.

4 Protein-based language models

To make it easier adaptiing solutions from the NLP world, a common approach to training protein-based LMs is to treat a protein sequence as a *document* and amino acids as *words* [**?**]. The tokens therefore are word-based tokens. This approach essentially makes it possible to take existing LMs from the NLP domain and train them on protein sequences. Several existing protein-based LMs follow this approach and have seen some successes on various downstream tasks. We discuss some existing LMs in the following subsections.

4.1 ProtVec

ProtVec [4] is a model based on *word2vec* [18]. ProtVec assumes that every token or word consists of three consecutive residues (amino acid 3-mers). During training, each protein sequence is split into overlapping 3-mers and the skip-gram version of word2vec is used to predict adjacent 3-mers, given the 3-mer at the center. After training, protein sequences can be split into overlapping 3-mers which are mapped onto a 100-dimensional latent space.

4.2 UniRef

UniRef [2], published in October 2019, is a model based on multiplicative LSTM (mLSTM) [17] and trained on UniRef50 dataset[23]

4.3 SeqVec

SeqVec [11], published in December 2019, is a model based on EMLo [19] which uses a deep, bi-directional LSTM model to create word representations. SeqVec learned to model a probability distribution over a protein sequence. The sum over this probability distribution constituted a very informative input vector for any machine learning task trying to predict protein features. It also picked up context-dependent protein motifs without explicitly explaining what these motifs are relevant for.

4.4 NetSurfP-2.0

NetSurP-2.0 [16], published in February 2019, is a successor to NetSurfP-1.0 (released in 1999). In addition to using protein sequences only, like most transformer-based models, the model exploits protein sequence profiles built from the dataset.

As described by the authors, the model was implemented using the Keras library. The input layer of the model consists of the one-hot (sparse) encoded sequences (20 features) plus the full HMM (hidden markov model) profiles from HH-suite (30 features in total, comprising 20 features for the amino acid profile, 7 features for state transition probabilities, and 3 features for local alignment diversity), giving a total of 50 input features. This input is then connected to two Convolutional Neural Network (CNN) layers, consisting of 32 filters each with size 129 and 257, respectively. The CNN output is concatenated with the initial 50 input features and connected to two bidirectional long short-term memory (LSTM) layers with 1024 nodes.

4.5 ProtTrans

ProtTrans [10], published in July 2020, is a transformer-based LM. The project explored the limits of up-scaling language models trained on proteins as well as protein sequence databases used for training, and compared the effects of auto-regressive and auto-encoding pre-training upon the success of the subsequent supervised training.

The ProtTrans project published several LMs:

- ProtTXL: Transformer-XL model trained on UniRef100 data.
- ProtTXL-BFD: Transformer-XL model and trained on BFD-100 data.
- ProtBert: Bert model trained on UniRef100 data.
- ProtBertTXL-BFD: Bert model trained on BFD-100 data.
- ProtAlbert: Albert (version: xxlarge v2) model trained on UniRef100 data.
- ProtXLNet: XLNet model trained on UniRef100 data.

From their experiments, they found that the LMs learnt rudimentary information about how proteins are formed, shaped, and function. They also found that unlike NLP where uni-directional models (auto-regressive; e.g. TransformerXL) perform on par with bi-directional models (auto-encoding; e.g. Albert), bi-directional models far outperform uni-directional models on protein data.

4.6 Evolutionary Scale Modeling (ESM)

ESM [21], published in August 2020, use unsupervised learning to traina deep contextual language model on 86 billion amino acids across 250 million protein sequences spanning evolutionary diversity.

5 Performance Evaluations

5.1 Perplexity

Perplexity is the standard evaluation metric of a language model. It is defined in the context of a language model as the exponential of a model's (crossentropy) loss averaged per token. In the case of the Transformer (with base-2 assumed) this is $2^{\mathcal{L}_{MLM}}$, where $\mathcal{L}_{MLM}$ is negative log likelihood of the true amino acid x_i given the masked sequence $x_{/M}$ and is given by

$$\mathcal{L}_{MLM} = \mathbb{E}_{x \sim X} \mathbb{E}_M \sum_{i \in M} -\log p(x_i | x_{/M}) \tag{1}$$

Perplexity describes the mean uncertainty of the model among its set of options for every prediction: ranging from 1 for an ideal model to 25 (the number of unique amino acid tokens in the data) for a completely random prediction [21]. Lower values are better.

5.2 CASP

Critical Assessment of protein Structure Prediction (CASP) [14] is perhaps the most well-known protein modeling benchmark which focuses on structure modeling. Once every two years the CASP assessment is held. During this competition structure predictors from across the globe are presented with protein sequences whose structures have been recently solved but which have not yet been made publicly available. The predictors make blind predictions of these structures, which are then assessed for their accuracy. The CASP structures thus provide a standardized benchmark for how well prediction methods perform at a given moment in time. Note that the test set consists of new experimentally validated structures which are held under embargo until the competition ends. This prevents information leakage and overfitting to the test set.

There have been fourteen CASP experiments so far. The first (CASP1) was held in 1994. The fourteenth (CASP14) was concluded in November 2020, where AlphaFold 2, a model developed by DeepMind, bested all other models in solving the protein-folding problem by achieves a median score of 92.4 GDT overall across all targets [6]. GDT is acronym for *Global Distance Test* and it ranges from 0-100. GDT is, roughly speaking, the percentage of amino acid residues within a threshold distance

from the correct position. It is interesting first version of AlphaFold also bested all the submissions in the CASP13 held in 2018 [22].

5.3 TAPE

Tasks Assessing Protein Embeddings (TAPE) [20] is the first attempt at systematically evaluating semi-supervised learning on protein sequences. TAPE includes a set of five biologically relevant donstream supervised tasks that evaluate the performance of learned protein embeddings across diverse aspects of protein understanding. Those tasks were chosen to highlight three major areas of protein biology where self-supervision can facilitate scientific advances: structure prediction, evolutionary understanding, and protein engineering. The five tasks are listed below:

5.3.1 Task 1: Secondary Structure (SS) Prediction

This is a sequence-to-sequence task where each input amino acid x_i is mapped to a three-class label $y_i \in \{Helix, Strand, Other\}$. It is classified as a Structure Prediction Task. SS prediction tests the degree to which models learn local structure. SS is an important feature for understanding the function of a protein, especially if theprotein of interest is not evolutionarily related to proteins with known structure. The metric being measured here is accuracy.

5.3.2 Task 2: Contact Prediction (Structure Prediction Task)

This is a pairwise amino acid task, where each pair (x_i, x_j) of input amino acids from sequence x is mapped to a label $y_{ij} \in \{0, 1\}$, where the label denotes whether the amino acids are "in contact" or not. It is classified as a Structure Prediction Task. Accurate contact maps provide powerful global information; e.g., they facilitate robust modeling of full 3D protein structure. Of particular interest are medium- and long-range contacts, which may be as few as twelve sequence positions apart, or as many as hundreds apart. The metric being measured here is precision of the $L/5$ most likely contacts for medium- and long-range contacts.

5.3.3 Task 3: Remote Homology Detection

This is a sequence classification task where each input amino acid x_i is mapped to a three-class label $y_i \in \{1, ..., 1195\}$, representing different possible protein folds. It is classified as an Evolutionary Understanding Task. Detection of remote homologs is of great interest in microbiology and medicine; e.g., fordetection of emerging antibiotic resistant genes and discovery of new CAS enzymes. The metric being measured here is accuracy.

5.3.4 Task 4: Fluorescence Landscape Prediction

This is a regression task where each input protein x is mapped to a label $y \in \mathcal{R}$, corresponding to the log-fluorescence intensity of x. It is classified as a Protein Engineering Task. This task is important because for a protein of length L, the number of possible sequencesmmutations away is $O(L^m)$, a

prohibitively large space for exhaustive search via experiment, even if m is modest. Moreover, due to epistasis (second- and higher-order interactions between mutations at different positions), greedy optimization approaches are unlikely to succeed. Accurate computational predictions could allow significantly more efficient exploration of the landscape, resulting in better optima. This task tests the model's ability to distinguish betweenvery similar inputs, as well as its ability to generalize to unseen combinations of mutations. The metric being measured here is Spearman's ρ (rank correlation coefficient).

5.3.5 Task 5: Stability Landscape Prediction (Protein Engineering Task)

This is a regression task where each input protein x is mapped to a label $y \in \mathcal{R}$, measur-ing the most extreme circumstances in which proteinxmaintains its fold above a concentrationthreshold (a proxy for intrinsic stability). It is classified as a Protein Engineering Task.

6 Interpretability

One of the main challenges of using machine learning models is interpretability. Treating ML models as black box may suffice for certain low-risk / low impact applications, but in healthcare where models can drive decisions on treatment options, drug manufacturing, etc., a black-box approach is simply unacceptable. In contrast to NLP, which seeks to automate a capability that humans already possess — understanding natural language — protein modeling seeks to shed light on biological processes that are not yet fully understood. By analyzing the differences between the model's representations and our current understanding of proteins, we may be able to gain insights that fuel scientific discovery.

7 Conclusion

This is a very active area of research. In fact, much of the advancements in applying NLP to protein sequences have happened within the past one year. Currently, alignment-based methods still outperform deep-learning-based methods for certain tasks [20] further highlighting huge opportunities for improvement. This work is by no means a comprehensive review of the state of the art in using machine learning approaches to study proteins. However, we hope it provides enough overview and insight into some of the methods, tools, techniques and datasets that can be used as a steping stone to launch deeper into this emerging but exciting area of research.

References

[1] Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic acids research*, 47(D1):D520–D528, 2019.

[2] Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

[3] Mohammed AlQuraishi. ProteinNet: a standardized data set for machine learning of protein structure. *BMC Bioinformatics*, 20(1):311, Jun 2019.

[4] Ehsaneddin Asgari and Mohammad RK Mofrad. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PloS one*, 10(11):e0141287, 2015.

[5] FC Bernstein, TF Koetzle, GJB Williams, EJ Meyer Jr, MD Brice, JR Rodgets, O Shimanouchi Kennard, and M. Tasumi. The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol*, 112:535–542, 1977.

[6] Ewen Callaway. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. https://www.nature.com/articles/d41586-020-03348-4, 2020. Accessed: 2020-12-04.

[7] Matthew Cobb. 60 years ago, francis crick changed the logic of biology. *PLOS Biology*, 15(9):1–8, 09 2017.

[8] The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212, 10 2014.

[9] Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, et al. The pfam protein families database in 2019. *Nucleic acids research*, 47(D1):D427–D432, 2019.

[10] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *bioRxiv*, 2020.

[11] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics*, 20(1):723, 2019.

[12] IUPAC-IUB. Recommendations on nomenclature and symbolism for amino acids and peptides. *Pure Appl. Chem*, 56:595–623, 1984.

[13] R. Jiang, X. Zhang, and M.Q. Zhang. *Basics of Bioinformatics: Lecture Notes of the Graduate Summer School on Bioinformatics of China*, volume 9783642389511. 11 2013.

[14] Moult John, Fidelis Krzysztof, Kryshtafovych Andriy, Schwede Torsten, and Tramontano Anna. Critical assessment of methods of protein structure prediction (casp)—round xii. *Proteins: Structure, Function, and Bioinformatics*, 86(7-15):10–1002, 2018.

[15] Neil C. Jones and Pavel A. Pevzner. *An introduction to bioinformatics algorithms*. MIT Press, 2004.

[16] Michael Schantz Klausen, Martin Closter Jespersen, Henrik Nielsen, Kamilla Kjærgaard Jensen, Vanessa Isabell Jurtz, Casper Kaae Sønderby, Morten Otto Alexander Sommer, Ole Winther, Morten Nielsen, Bent Petersen, et al. Netsurfp-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, 87(6):520–527, 2019.

[17] Ben Krause, Liang Lu, Iain Murray, and Steve Renals. Multiplicative lstm for sequence modelling, 2017.

[18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[19] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[20] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with TAPE. In *Advances in Neural Information Processing Systems*, 2019.

[21] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*, 2020.

[22] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[23] Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

[24] Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, Cathy H. Wu, and the UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 11 2014.