Learning the Language of Life

Jose Juan Almagro Armenteros[1], Alexander Rosenberg Johansen[2], Ole Winther[2] and Henrik Nielsen[1]

[1]Section for Bioinformatics, Department of Health Technology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark and [2]Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark.

What determines how a protein looks, how it works, and where it carries out its function? Within the field of bioinformatics, many methods have been developed to predict the structure, function, and location of proteins based on their amino acid sequences. However, these prediction methods could be much better if we had an understanding of the language of proteins.

In the field of Natural Language Processing (NLP), methods for making machines "understand" human languages are developing rapidly these years. Tasks such as automated translation and text classification are being handled by deep learning methods, e.g. at companies like Google or Facebook. A natural choice would be to apply methods from cutting-edge NLP to the language of proteins, i.e. their amino acid sequences, and indeed, both Google [1] and Facebook [2] have made initial efforts at protein understanding.

We have also started our own foray into this field, using a recurrent Long Short-Term Memory neural network to build language models for protein sequences, i.e. models that predict an amino acid given its context in the sequence [3]. From this work, we have learned that the language of proteins has *dialects* – the predictability of amino acids depends on the origin (domain of life) of the sequences used for training and testing. As an example, bacterial proteins seem to be generally more predictable than eukaryotic proteins.

Our long-term goal is to use the internal states of trained protein language models as representations of proteins for various prediction tasks, including the structure, function, and location of proteins. In this way, the vast amounts of unannotated protein sequence data could be brought to use in such prediction tasks, which otherwise depend upon severely limited amounts of experimentally annotated sequences. It has already been shown that such a representation is able to improve predictive performance of secondary structure and subcellular location [4], but that study was, like the Facebook and Google attempts, done without taking dialects into account.

Another way to use a trained protein language model is to ask it to generate novel sequences without homology to any known proteins, but with biological properties similar to those in its training set. By generating novel proteins using our language model and a simple background model, we have e.g. shown that proteins from the language model have a realistic proportion of predicted signal peptides in contrast to the background model, which generates almost no signal peptides. This shows that a neural network representing an understanding of proteins in general can have tangible technological implications.

Bileschi, Maxwell L.; Belanger, David; Bryant, Drew; Sanderson, Theo; Carter, Brandon; Sculley, D.; DePristo, Mark A.; Colwell, Lucy J. 2019. Using Deep Learning to Annotate the Protein Universe. *BioRxiv*. https://doi.org/10.1101/626507.

Rives, A.; Goyal, S.; Meier, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. 2019. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *BioRxiv*. https://doi.org/10.1101/622803.

Almagro Armenteros, Jose J.; Johansen, Alexander R.; Winther, Ole; Nielsen, Henrik 2020. Language Modelling for Biological Sequences – Curated Datasets and Baselines. *BioRxiv*, https://doi.org/10.1101/2020.03.09.983585.

Heinzinger, Michael; Elnaggar, Ahmed; Wang, Yu; Dallago, Christian; Nechaev, Dmitrii; Matthes, Florian; Rost, Burkhard 2019. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* 20. https://doi.org/10.1186/s12859-019-3220-8.