Latent Space Representations for Evolutionary Diversity of Protein Families

Gagana B

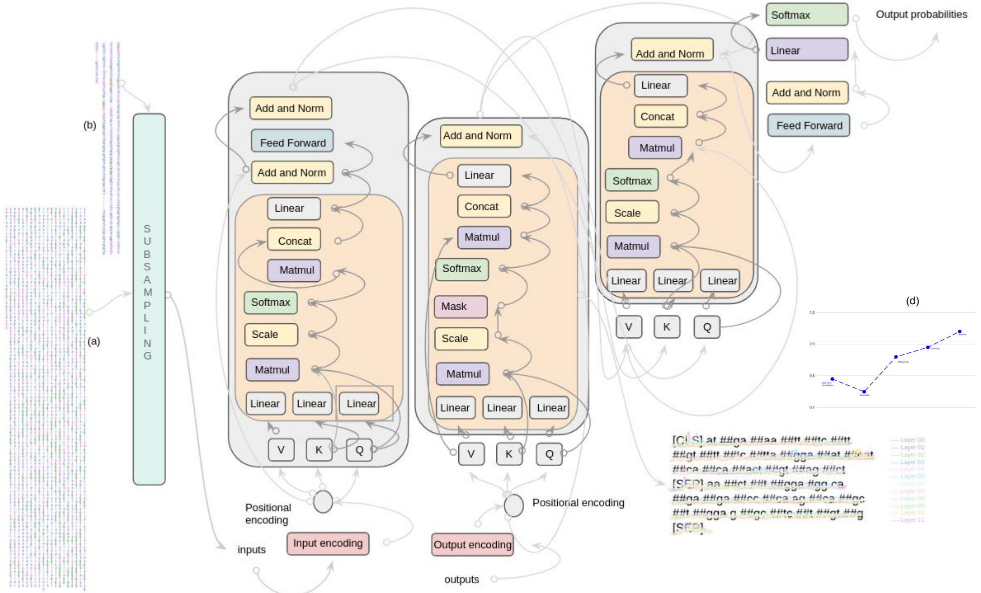
Introduction

Molecular biology, specifically protein folding and heuristic structural prediction problems, have been revolutionised by computational methods (such as gradient descent algorithms {2} and transformer based architectures (like BERT{1} and it's variants{10})) which proffer interesting insights on multi-scale representations{6}, cross-modality embeddings{7}, rotation invariant shape-mers{8}, simultaneous inclusion of protein backbone{4} and extension to euclidean vector spaces{9}. But the generalisability of the behavior of these systems and their respective representations to two levels of evolutionary diversity: (a) multitude of protein compositions that varies across interaction mechanisms (host-microbiome{3}) (b) single cell analyses and how they scale to sub-populations and populations of cells remains obscure. We propose to tackle the same by applying similar methods to various interaction settings, and understanding how single cell analysis methods generalise to capture heterogeneity: patterns within cells to patterns which within sub-populations of cells. While repetitions across scale in different organisms perfectly embodies the recursiveness that connectionist models excel at, the underlying influencing factors (protein localisation sites, cell types and cell organisations that affect recruitment, inter and intra cellular communication etc.) pose an adverse challenge to both basic and applied artificial intelligence. The vast amounts of unannotated data along with the capability for empirical verification and the potential of comparing radically differing techniques under a unifying set of problems could help uncover fundamental strengths and weaknesses of NLP and computer vision approaches{5}.

Methodology

Interaction Mechanisms: The input protein sequences are extracted and further subsampled for a bi-fold advantage where the incoming subsequence is representative of the binding sites which can be further indexed with relative ease within the architecture.

Single cell to subpopulation generalisations: Similar methods are applied to capture patterns and signatures to understand "generalisation quotient" which is a function of applicability and accuracy across nucleoplasm, nuclear membrane, nucleoli, nucleoli fibrillar center, nuclear speckles, nuclear bodies, endoplasmic reticulum, golgi apparatus, intermediate filaments, actin filaments, microtubules, mitotic spindle, centrosome, plasma membrane, mitochondria, aggresome, cytosol, vesicles and punctate cytosolic patterns.



Detailed MicroBERT Architecture to capture interaction mechanisms. In this figure, the pulmonary host surfactant protein interactions with binding sites of ORF8 SARS-COV-2 variants is shown. (a) Color coded input sequence of A1 variant AD' 6A (SFTPA1) mRNA protein; (b) Color coded input sequence of ORF8 sequence of SARS-COV-2; (c) (left) Visualisation of the per-layer multi-attention head activation on the tokenized surfactant substring and (right) corresponding layer-wise legend. (d) Graph of generalisation quotient where x-axis refers to latent representation based model score and y-axis refers to average performance measured as focal loss.

Acknowledgements

The author would like to acknowledge: Anil B Murthy, V Krishnamurthy and Nikhil Kasukurthi.

References

1. Vig et al, "BERTology meets biology: Interpreting attention in protein language models", arXiv: 2006.15222, 2020
2. Senior et al, "Improved protein structure prediction using potentials from deep learning", Nature, 2020
3. B, "MicroBERT: A BERT based framework to map host-microbiome protein-protein interactions", MIT Microbiome Symposium, 2021
4. King et al, "SidechainNet: An all-atom protein structure dataset for machine learning", Machine Learning for Structural Biology Workshop - NeurIPS, 2020
5. Hunter, "Artificial intelligence and molecular biology", AAAI Symposium, 1990
6. Rives et al, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences", biorXiv, 2020
7. You et al, "Cross-modality protein embedding for compound-protein affinity and contact prediction", biorXiv, 2020
8. Durairaj et al, "Fast and adaptive protein structure representation for machine learning", Machine Learning for Structural Biology Workshop - NeurIPS, 2020
9. Jing et al, "Learning from protein structure with geometric vector perceptrons", arXiv: 2009.01411, 2020
10. Filipavicius et al, "Pre-training Protein Language models with label-agnostic binding pairs enhances performance in downstream tasks", arXiv: 2012.03084, 2020
11. Lundberg et al, Human Protein Atlas: Single Cell Classification, hpa-768768, 2021.

ICLR