Generating new protein with specific task with LLMs

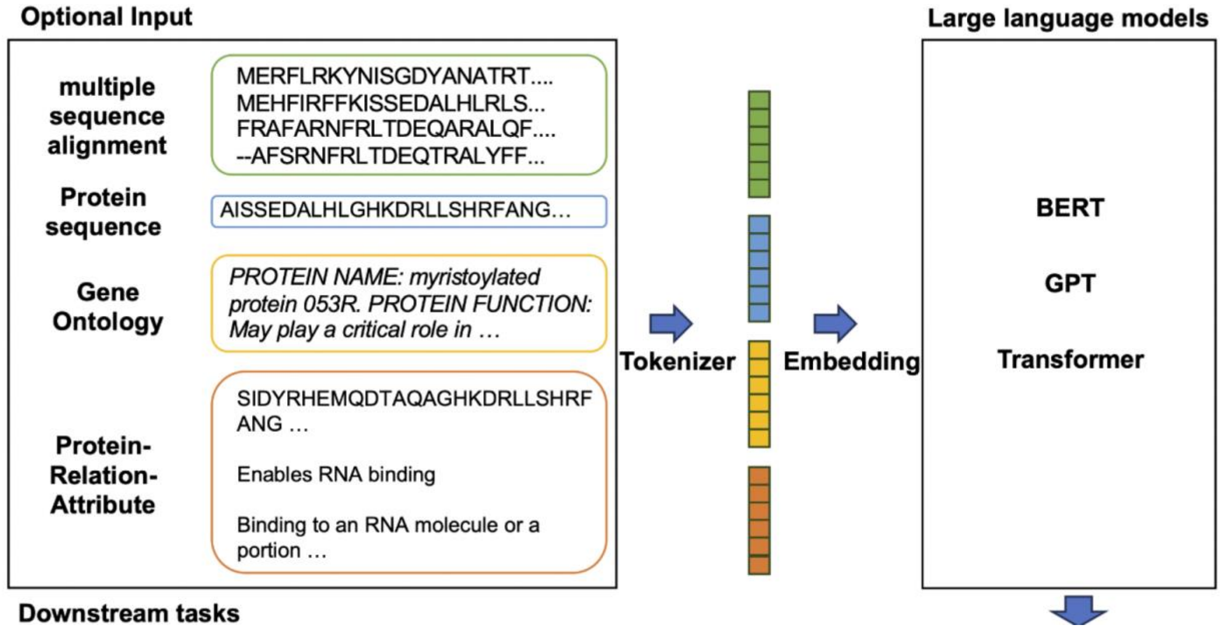
Halima Bensmail
Qatar Computing Research Institute

Aim: Generation of protein has broad application prospects in fields such as drug design and protein engineering By using methods such as machine learning or deep learning, protein sequences can be generated. The generated sequences are hoped to have good foldability so that they can form stable three-dimensional structures. Moreover, the desired proteins are expected to exhibit specific functional properties, including enzyme activity and antibody binding capability. The advancement of large language models and the integration of conditional models have significantly propelled the progress in the field of protein generation.

The model, referred to as ProGen, incorporates UniprotKB Keywords as conditional tags in 2020. These tags encompass a vocabulary consisting of various categories, including 'biological process', 'cellular component', and 'molecular function'. In total, the conditional tags encompass over 1,100 distinct terms. When assessing protein sequences generated by ProGen using metrics for sequence similarity, secondary structure accuracy, and conformational energy, they exhibit desired structural properties. In 2022, inspired by the remarkable achievements of generative Transformer-based language models like the GPT-x series, the development of ProtGPT2 emerged. Notably, the proteins generated by ProtGPT2 exhibit amino acid propensities following the principles of natural ones. Assessments involving disorder and secondary structure prediction reveal that a substantial majority (88%) of ProtGPT2-generated proteins possess globular characteristics, aligning with the attributes found in natural sequences. Employing AlphaFold on ProtGPT2 sequences yields well-folded non-idealized structures, encompassing the presence of extensive loops, and the emergence of previously unseen topologies that are absent from current structure databases. It appears that ProtGPT2 has acquired the language specific to proteins.

The proposed work will focus on doing an assessment of these two algorithms ProGen and ProtGPT2.

We also will use LLMs for protein function prediction using protein sequence as input and fine tune the model for several protein tasks such as homology prediction, and secondary structure prediction, protein solubility and protein crystallization and compare that to SPRoBERTa.


Optional Input
multiple sequence alignment
MERFLRKYNISGDYANATRT....
MEHFIRFFKISSEDALHLRLS...
FRAFARNFRLTDEQARALQF....
--AFSRNFRLTDEQTRALYFF...
Protein sequence
AISSEDALHLGHKDRLLSHRFANG…
Gene Ontology
PROTEIN NAME: myristoylated protein 053R. PROTEIN FUNCTION: May play a critical role in …
Protein-Relation-Attribute
SIDYRHEMQDTAQAGHKDRLLSHRF ANG …
Enables RNA binding
Binding to an RNA molecule or a portion …
Tokenizer
Embedding
Large language models
BERT
GPT
Transformer
Downstream tasks

Fine-tuning step

Protein crystallization protein solubility protein-protein interaction

Deliverable:

a- Learn how to have access to protein database

b- Download data and process it using similarity assessment and other task

c- Train LLMs model on protein sequences.

d- Use the multi-head model and fine tune the model to answer the tasks:

   a. Crystallization

   b. Solubility

Student: Knowledge in programming with any language.
Student is willing to learn about sequences data such as protein sequence or DNA sequence
Student is willing to learn about language models

References:

[1] Steinegger, M. and J. Sding, *Clustering huge protein sequence sets in linear time.* Nature communications, 2018. **9**(1): p. 2542.

[2]. Steinegger, M., M. Mirdita, and J. Sding, *Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold.* Nature methods, 2019. **16**(7): p. 603-606.

[3]. Strokach, A. and P.M. Kim, *Deep generative modeling for protein design.* Current opinion in structural biology, 2022. **72**: p. 226-236.

[4]. Ferruz, N. and B. Hcker, *Controllable protein design with language models.* Nature Machine Intelligence, 2022. **4**(6): p. 521-532.

[5]. Madani, A., et al., *Progen: Language modeling for protein generation.* arXiv preprint arXiv:2004.03497, 2020.