QGen: Qatari Language Modeling for Protein Sequence Generation

Project Description:

Generating proteins with desired properties is one of the most complex yet impactful problems in biology. Protein engineering research has grown over the past 50 years and yielded remarkable outcomes including the development of new enzymes, therapies, and sensors. The raw amino acid sequence encodes a protein, and during synthesis, this chain of amino acids folds in ways that exhibit local (secondary) and global (tertiary) structure. These structural properties then directly determine a unique function, which is of ultimate interest to protein engineers. Hence, our aim is to develop a two-step pipeline. In the first step, a model takes a protein sequence as input and generates potential scaffolds along with desired properties. This model would focus specifically on understanding the correlations between protein sequences and structural, or functional properties of molecular scaffolds, allowing for a more precise interpretation of biological and chemical interactions. In the second step, a subsequent model utilizes the generated properties and scaffolds to produce a complete drug molecule. By separating these tasks into two distinct models, each model can learn more focused and specialized correlations, leading to improved performance. This modular approach could potentially outperform a single end-to-end model that directly maps a protein sequence to a drug molecule, offering enhanced flexibility, better targeted drug designs, and more refined optimization of therapeutic properties. A Qatari dataset will be utilized to fine-tune the language model.

Duties/Activities:
The Interns will build a protein sequence repository. Then, they utilize generative artificial intelligence models to create embeddings followed by investigating protein-to-target interactions. In addition, we will build a dashboard that accepts partial protein sequence and returns full protein sequence with desired properties.

Required Skills:
Python programming skills is required. Understand the basics of bioinformatics, deep leaning models, and previous experience in building dashboards is a plus.

Learning Opportunities:
The intern will have opportunity to work with scientists and software engineers on promising research problems. In addition, it is great opportunity to participate in solving real-world healthcare problems.

Expected Team Size: *3-4*

Mentors

Name: Dr. Mohamed Elshrif email: melshrif@hbku.edu.qa