St. Jude Children's Research Hospital

St. Jude BioHackathon

Title
ML dashboard for real-time model building using natural language processing on biological sequence inputs

Category
GUI Tool Development

Challenge
ML dashboard for interactive real-time model building on biological sequence inputs

Natural language processing is a highly promising field in artificial intelligence. Efforts are currently underway to extend these language-specific models to the biological domain. Biological sequences (DNA, RNA, protein sequence) can benefit from such models but the current software landscape is scarce. Our goal is to build an interactive dashboard based on Python, that will implement some of these language models on biological sequence data. Users will be able to provide any biological sequences as input (e.g., from a multiple sequence alignment) and then open the interactive dashboard where they can interactively select the language model, define the parameters, and start calculations. Depending on the model selected (and this is something that can be prioritized for the purpose of the hackathon), the user can in real-time change parameters and have the results updated on the screen.

The only solution that currently exists, to my knowledge, is BioSeq-BLM (https://academic.oup.com/nar/article/49/22/e129/6377401). However, the server interface is quite archaic, web-based, expects users to know all input parameters, and is not at all interactive.
Our goal for the hackathon is to select only a subset of the BioSeq-BLM models (perhaps 2 or 3), which we think is sensible given the timeline, and develop a prototype. A prototype for this project would entail a working dashboard that accepts a biological sequence as input, processes it using standard Python packages on the backend, and provides a few applications to interactively explore the results (e.g., two or three interactive plots).

Because building the backend is an involved task, we think users should aim to code the initial solution on Jupyter notebooks and then use programs such as voila (https://github.com/voila-dashboards/voila) to serve the dashboard. This would also provide participants with the unique experience of learning about interactive data exploration, which I assume is not that common in hackathon challenges.

Benefit
We currently have a working 3D sequence clustering dashboard webserver working in our group. We would like to add the capability of making selections on these clusters, and then provide the above Python-based dashboard to users to interactively explore the different models. We also think that solutions supported by BioSeq-BLM are useful to computational proteomics and genomics and would find beneficial use by other groups here at St. Jude as well.

Helpful Tools, Packages, or Software
Jupyter Notebook, numpy, scipy, scikit-learn, BioSeq-BLM (the downloaded standalone package), bokeh/plotly.

Test Data
BioSeq-BLM can be used as a reference solution to test and compare.