Framework for Independent Study

Background of this project

In this stage, our ultimate goal is to train a high-quality neural network for aptamer-protein interaction prediction. Based on our literature review, we have already found some interesting work that applies BERT, transformer, and CNN models. These models are good at known proteins but fail on unseen proteins. We would like to explore whether continue learning (CL) is effective for improving these models, and we can prepare to train our own transformer-based models.

To begin, let's create some datasets.

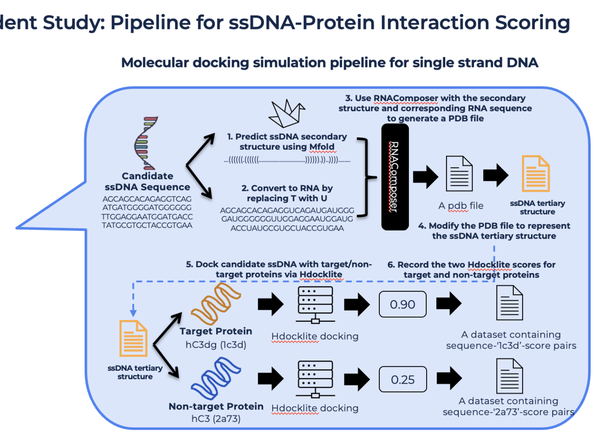
Materials for CL:
https://neptune.ai/blog/continual-learning-methods-and-application
https://paperswithcode.com/task/continual-learning
https://arxiv.org/abs/1904.07734

WK3-WK4: successfully run the pipeline and give a try to generate two binding scores for one ssDNA sequence

Date: 9/13-9/22

Tasks

T0: read the background and pipeline illustration, and understand what's your todos.



T1: build the hdock docker on your windows laptop

1. download "docker desktop" on your windows laptop (hdock docker can not be used on macbook m1/m2, so try to find one with windows/linux)
2. follow the tutorial to install hdock docker on your laptop: https://hub.docker.com/r/pegi3s/hdock-builder

T2: read all the code I provide in the "pipeline" folder and try one new sequence you randomly create to make the pipeline run!

0. install required packages

Try different functions to see whether they can successfully work on your laptop and install all the packages that you need. You may need to change the "chrome_driver_path" in "rna_downloader.py". Try your own sequence by replacing "aptamer_sequence" in "grader.py". Finally, you should execute "python grader.py" in the command line.

```python
# example
grader = Grader()
aptamer_sequence = "TCACCGCCGTTTTAA"  # your aptamer sequence
grader.test_one_sequence(aptamer_sequence)
```

```
PS C:\CodeProjects\github.com\JianJinglin\multiple-thread-hdock> python grader.py
Testing aptamer sequence: TCACCGCCGTTTTAA
Submit:
>rna
UCACCGCCGUUUUAA
..((....)).....
====================
==========Attempt 1 for RNAComposer==========

DevTools listening on ws://127.0.0.1:8461/devtools/browser/5edb2911-f487-4687-a104-ea0047771d0
==========Submitting sequence to RNAComposer==========
[28956:1132:0912/204848.683:ERROR:device_event_log_impl.cc(196)] [20:48:48.683] USB: usb_service_win.cc:105 SetupDiGetDeviceProperty({{A45C254E-DF1C-4EFD-8020-67D146A850E0}, 6}) failed: µ¥Ì¿1o8¨oàâc|ám¢é (0x490)
Created TensorFlow Lite XNNPACK delegate for CPU.
```

This is the screenshot of the output you should see when you successfully execute "python grader.py" in the command line.

```
Converted ssDNA file saved as: ./Aptamer_pdb/ssUESAFYTJ.pdb
Starting hdock: ssUESAFYTJ-C3dg.out
Starting hdock: ssUESAFYTJ-2a73.out
 This docking calculation will take about    425.74  seconds
 This docking calculation will take about    681.11  seconds
```

This is the signal that indicates you have successfully executed "hdock".

When the process finishes, you will see "Finished docking ..." and you will find two "hdock.out" files in your folder. This means you have completed this task.

2. use a text editor to open these two -Hdock.out files and check the scores.

3. use "output_get_excel.py" to extract scores and create one excel file.

4. manually reorganize the two pairs and put them into two excel files and upload it to google drive. (you can also write code to do this)

you can check these two samples and add your results at the bottom of these two files.

WK5: create two datasets for target protein and non-target protein

Date: 9/23-9/29

T3: massively repeat the process to generate more than 1000 pairs for each proteins

1. generate a list of ssDNA, and massively put into the pipeline you already successfully run.

2. get a banch of new pairs (aptamer, protein, hdock_lite score)

Jinglin's todo

1. create a github repo and a google drive shared folder — https://drive.google.com/drive/folders/1eraY7pbQy4Uo33GSwMmUAgf2pTyNLxYS?usp=sharing

2. create a jupyternotebook in wk3-4