BioAI - Pfam Technical Test

As a research engineer, you will have the opportunity to work on biology-related projects. Proteins are ubiquitous in our day-to-day work, hence this technical test aims to introduce you to this type of data.

Problem Setting

The goal of this test is to build a protein classifier: for each protein, you have to assign the corresponding Pfam family (i.e. protein family). You can find more information regarding the Pfam family here.

Data

The dataset to use is hosted on Kaggle: Pfam seed random split - Using Deep learning to Annotate the Protein Universe

Deliverable

Your deliverable should be divided into the following parts:

1. **Dataset Analysis**
   You should explore the statistics and potential issues of the dataset.
2. **Method Explanation**
   You should explain the methods you used. References should be cited.
3. **Experiment Description**
   You should design a set of experiments in order to prove the advantage (if they exist) of the proposed methods.
4. **Result Analysis**
   You should resume your experiment results and analyze them.

The format of your deliverable is up to you (script, Jupyter notebook, pdf report, git repository, etc.)

- If you choose a pdf report, you must also provide the code used for the project.
- If you provide a Jupyter notebook, please re-run all the cells from scratch and save the notebook with the cell outputs.
- If you provide a git repository, please launch from your repository `git bundle create <YOUR_NAME>.bundle --all` and send us the resulting `<YOUR_NAME>.bundle` file.
- If you host your git repository on GitHub/GitLab, please keep it **private**.

You are free to use any machine learning/deep learning framework with the following requirements:

- Python 3.6+
- Easily reproducible on a laptop with 16GB of RAM + 4GB GPU

Evaluation

- You won’t be evaluated on the final performance of your classifier but rather on the methodology you used to tackle this task, so make sure to explain each step.
- We will pay attention to the code quality and the documentation.
- You will be evaluated on your capacity to communicate the results of your work both verbally and in writing to a technical audience.

Compute

In case you need more compute power than locally available on your computer, the following resources provide more compute power for free:

- Google Colab: access to one GPU or one TPU, time limit of 12 hours (kernels are shut down after 12 hours)
- Kaggle notebooks: access to one GPU (NVIDIA P100), time limit of 6 hours

Hope you have fun!

Please feel free to contact us if you have any questions, we’ll be happy to help.