TASK 1

Prediction of gene expression from different information (graded)

You are in the group Amoebas consisting of dsumner (dsumner@student.ethz.ch (mailto://dsumner@student.ethz.ch)), lkasak (lkasak@student.ethz.ch (mailto://lkasak@student.ethz.ch)) and tfang (tfang@student.ethz.ch (mailto://tfang@student.ethz.ch)).

1. READ THE TASK DESCRIPTION

2. SUBMIT SOLUTIONS

3. HAND IN FINAL SOLUTION

1. TASK DESCRIPTION

INTRODUCTION

The topic of the first project is the use of epigenetic information to predict gene expression. The goal is to build a simple tool that uses different epigenetic information to accurately predict expression of genes from unseen cell lines.

In particular, in this project you will:

- Explore various epigenetic datasets (histone marks and chromatin accessibility).
- Choose and implement a prediction model that takes various epigenetic datasets as inputs and predicts gene expression.
- Evaluate the performance of your tool.

We have split the work into several packages. There is no need for you to follow this division of work exactly, but it should serve as a guide to which steps are deemed to be important.

Please refer to the handout for project 1 on Moodle for additional details.

The data is available via Polybox (https://polybox.ethz.ch/index.php/s/iY6d8qbMMiy4dQh). Please refer to the course Moodle for the password.

Since Polybox has been having some issues lately, we also provide an alternative GDrive link which contains identical data to the Polybox: https://drive.google.com/drive/folders/1XABEZRTaeUgfd4lLXbf-wXnTg0eHL2Cm?usp=sharing.

SUBMISSION FORMAT

For every data instance in the test set, submission files should contain three columns: (a row index without a column name), *gene_name* and *gex_predicted* where *gex_predicted* should be a double with your predicted expression for that gene.

The file should contain a header and have the following format:

```
,gene_name,gex_predicted
0,CAPN9,0.1
1,ILF2,3.5
...
```

Please zip your submission file (as this will be needed for later projects) before submission and name the ZIP file as LastName_FirstName_Project1.zip. Please ensure that your submission file (within the ZIP) is named gex_predicted.csv. We provide sample code with which you can check that your submission file is correctly zipped and formatted.
Please keep in mind that, as a group, you have a limited number of submissions as stated on the submissions page.

GRADING

We provide you with a dataset for which you have to make prediction or compute other quantities. For each of your submissions, a score is computed. When handing in the task, you need to select which of your submissions will get evaluated and provide a short description of your approach as set out in the project description. This has to be done **individually by each member** of the team. We will then compare your selected submission to our baseline. This project task is graded with either **pass or fail**. To pass the project, you need to achieve a better score than the baseline. Passing the projects is a mandatory part of the course. You will have to pass both Project 1 and Project 2 (unless you choose to do a paper presentation, in which case you only have to pass one of the two projects). In addition, if you achieve performance in the top 10% of all final submissions within one of the mandatory projects, you will receive a bonus (+0.25) to your final course grade (non-cumulative!). You will not be able to see if you have performed better or worse than the mean on the leaderboard; decisions will be communicated via email after the end of each project. The groups receiving a bonus will get the opportunity to present their methodology in a subsequent tutorial. We will consider the code and your solution description for the pass/fail decision. The following **non-binding** guidelines provides you with an idea of what is expected to pass the project. If you hand in a properly-written description, your source code is runnable and reproduces your predictions, and your submission performs better than the baseline, you can expect to have passed the assignment.

⚠ Make sure that you properly hand in the task, otherwise you may obtain zero points for this task.

FREQUENTLY ASKED QUESTIONS

⊙ WHICH PROGRAMMING LANGUAGE AM I SUPPOSED TO USE? WHAT TOOLS AM I ALLOWED TO USE?

You are free to choose any programming language and use any software library. However, **we strongly encourage you to use Python**. You can use publicly available code, but you should specify the source as a comment in your code.

⊙ IN WHAT FORMAT SHOULD I SUBMIT THE CODE?

You can submit it as a single file (you should compress your files into a .zip; jupyter notebook, conda env, README (optional)) having max. size of 1 MB. Please name the ZIP file of your notebook, environment, and README (optional) LastName_FirstName_Project_X_Code.zip.

⊙ IN WHAT FORMAT SHOULD I SUBMIT MY SUBMISSION FILES?

Please submit your submission files as one zip file named LastName_FirstName_ProjectX.zip. The zip should contain all csv files specified in the detailed project description for each respective project.

⊙ WILL YOU CHECK / RUN MY CODE?

We will check your code and compare it with other submissions. We also reserve the right to run your code. Please make sure that your code is runnable and your predictions are reproducible (fix the random seeds, etc.). Provide a readme if necessary (e.g., for installing additional libraries).

⊙ DO I NEED TO PROVIDE A REQUIREMENTS.TXT FILE?

Yes, please provide a requirements.txt file (or equivalent depending on programming language), to ensure that we can reproduce your results as easily as possible.

⊙ SHOULD I INCLUDE THE DATA IN THE SUBMISSION?

No. You can assume the data will be available under the path that you specify in your code.

⊙ CAN YOU HELP ME SOLVE THE TASK? CAN YOU GIVE ME A HINT?

As the tasks are a mandatory part of the class, **we cannot help you solve them**. However, feel free to ask general questions about the course material during or after the exercise sessions.

⊙ CAN YOU GIVE ME A DEADLINE EXTENSION?

⚠ We do not grant any deadline extensions!