THE UNIVERSITY *of* EDINBURGH
School of Biological Sciences

Assignment 2]
[Bioinformatics Programming and System Management]

Student Exam Number: B193818

ASSIGNMENT 2

GitHub link: https://github.com/B193818-2021/Assignment2

Zipfile code: BPSM1234

USER MANUAL FOR ORDINARY USER

1. To start the pipeline, user just run the python script.

```
python pipeline.py
```

2. First, the pipeline will ask for a directory to save all of the outputs, user need to type in a path for a directory, if the workspace aready existed, the pipeline will ask the user weather to overwrite it or input anothter directory.

```
Please input a dir for workspace:workspace(Enter)
```

3. Then, the pipeline will create the workspace for you, and all your output will place in this directory. When the workspace has been settled down, the pipeline will ask for a protein family and taxonomic group user want to analysis.

```
Please input a protein family: glucose-6-phosphatase(Enter)
Please input a taxonomic group:birds(Enter)
```

4. Here, we will use **glucose-6-phosphatase** in **birds** as a example. When the pipeline recieve these two paramaters, it will search proteins in NCBI Protein database and report hit counts. If the pipeline can't find enough priteins for downstream analysis, the pipeline will terminate! Othterwise, the pipeline will ask user how many sequences you want to download.

```
Seaching for glucose-6-phosphatase protein in birds ...
Got 883 hits
Note: download and analysis too many sequences will take a long
time.
Please input a number of sequences you want to download and
analysis(Enter directly to use all):500(Enter)
Using maximun of 500 sequences
Protein sequences saved in "workspace/protein-sequences.fa".
```

5. According to the user's request, the pipeline will download only the request sequences to avoid using too many times and disk space. After that, the path of saved FASTA file of protein sequences will be reported on the screen. Then, the pipeline will count how many species these proteins come from, user can determinde weather they wanna to continue the analysis or terminate.

```
These sequences come from 89 species. Continue?(yes):yes(Enter)
Aligning sequences ...
Alignment result saved in: "workspace/protein-sequences-
align.fa"
```

6. If the sequences are satisfied, the pipeline will process in multi-sequence alignment and report the path of alignment result.

```
Alignment result saved in: "workspace/protein-sequences-
align.fa"
Note: analysis conservation for too many sequences will take a
long time.
Please input a number of sequences you want to continue
analysis(Enter directly to use all):100(Enter)
Analysis most similar 100 sequences
```

7. Then, the pipeline will ask for the number of sequence user want to plot conservation and extract the most similar sequences to plot conservation and report the path of result.

```
Plot conservation of a sequence alignment
Plotcon output image saved in: "workspace/plotcon.1.svg"
```

8. After that, the pipeline will scan motifs from PROSITE database and report the path of result.

```
Please input the window size:4(enter)
Scaning motif ...
Found 8 motif:
LECTIN_LEGUME_BETA
ATP_GTP_A
THIOL_PROTEASE_HIS
TYR_PHOSPHO_SITE_2
AMIDATION
TYR_PHOSPHO_SITE_1
RGD
TUBULIN_B_AUTOREG
Motif list saved to: workspace/protein-sequences-motifs.txt
```

9. Finally, the pipeline will count codon frequency for users.

```
Counting codon frequencies ...
Calculate the composition of unique words in sequences
Codon frequencies saved to: workspace/protein-sequences-
codon.comp
```

10. That is all the analysis procedure.

USER MANUAL FOR PYTHON DEVELOPER

There are 3 parts of this pipeline:

1. Preprocess

In this part, the pipeline will ask for workspace, protein family and taxonomic group. Each input is place in a indefinitely while loop, only break until the pipeline get the valid input from user.

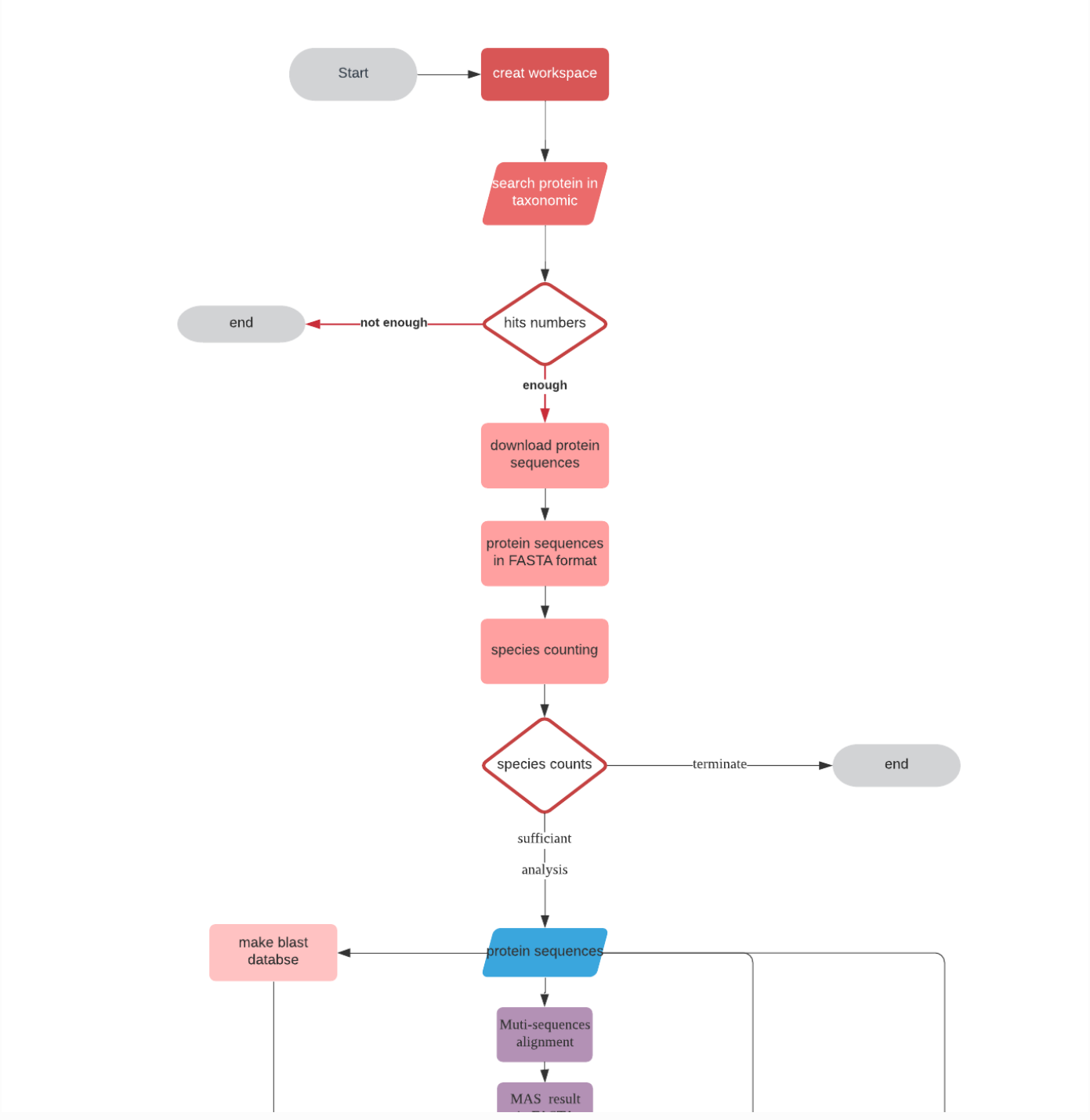
2. Fetch Sequences

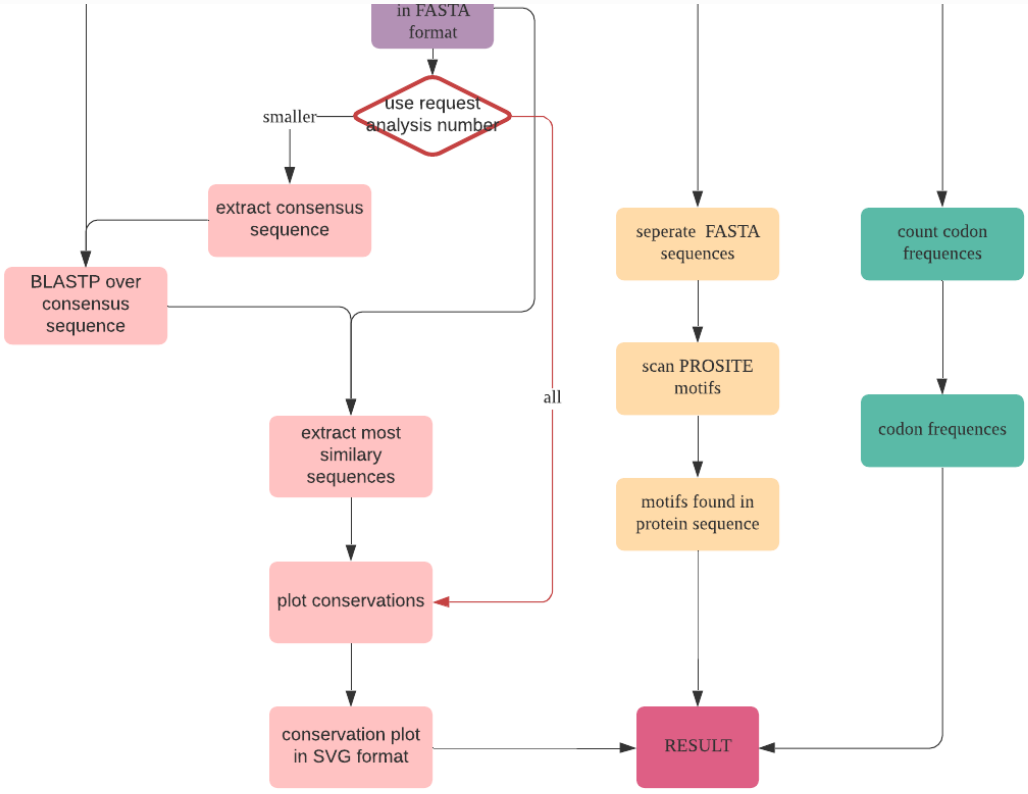
In this part, the pipeline will search and fetch protein sequences fron NCBI Protein database. In this process, the pipeline will only download sequences according to user's request to save running time and disk space.

3. Analysis

Conservation plot, PROSITE motif scan and Codon counts will apply to downloaded sequences in this process. When there are too many sequences, the pipeline will analysis sequences similarity and extract most similarly sequence to save running time according to the user's request.

THIS IS A FLOW CHART FOR USERS




in FASTA format
use request analysis number
smaller
all
extract consensus sequence
BLASTP over consensus sequence
extract most similary sequences
plot conservations
conservation plot in SVG format
seperate FASTA sequences
scan PROSITE motifs
motifs found in protein sequence
count codon frequences
codon frequences
RESULT