This document includes a walkthrough of the program, echoing the comments found throughout. I have summarised the issues which I faced during this assignment in a preface ‘Challenges’. Please type python<ICA1\_python\_script.py in a bash shell to run the code once unzipped.

## Challenges (line numbers for ease of navigation)

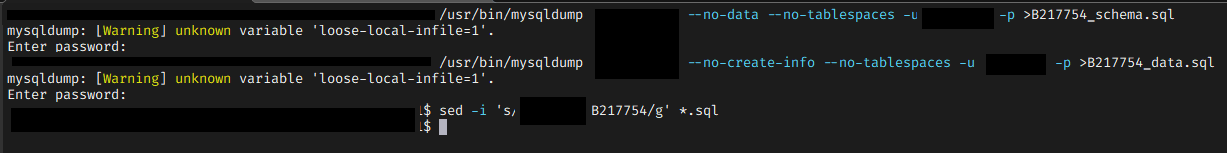
* 42 I was unable to loop this code through the list of gene symbols given earlier in the code, as the following error would be given:

*200~File "/usr/lib/python3.8/threading.py", line 848, in start*

*raise RuntimeError("threads can only be started*

* 71 +153+175 This code should transfer the pandas dataFrame to the SQL table made earlier, however, I was not able to do this and as a result, work with true data ceases here.
* 179 For the purposes of this assignment, I have manually entered data from the three primary biological databases into three dummy SQL tables so that I can demonstrate how I would have manipulated such databases into a summary table. The lines of code 182-307 exist purely to demonstrate this.

I attempted to run the code below as instructed, but am unsure if this has had the desired effect?



## Code walkthrough

This code should use APIs to retrieve data from 3 primary biological databases and store it in a MySQL table which can be queried.

7 Making a connection to my personal sql db and opening the cursor to execute commands. CODE TAKEN FROM BD2\_Relational\_DataBases\_2per.pdf, author: Simon Tomlinson.

18 Setting list of gene names

24 ########## Entrez Pubmed #######

25 This section retrieves the PMID of published materials which are obtained when searching for the gene symbol.

26 modified from BD5\_BiologicalDatabases, author Simon Tomlinson. Previously modified from https://entrezpy.readthedocs.io/en/master/functions/efetch\_func.html

once")

42 I was unable to loop this code through the list of gene symbols given earlier in the code, as the following error would be given:

43 200~File "/usr/lib/python3.8/threading.py", line 848, in start

44 raise RuntimeError("threads can only be started

71 At this point, the data should be converted into a pandas df as seen in QuickGO section. See thiat section for issues on data conversion.

73 ################QUICKGO###################

74 QuickGO, geneproduct from https://www.ebi.ac.uk/QuickGO/api/index.html#!/gene\_products/geneProductSearchUsingGET

77 Installing the required SQL package SQLAlchemy modified from https://www.activestate.com/resources/quick-reads/how-to-install-python-packages-using-a-script/

81 implement pip as a subprocess:

84 Creating the an SQL tables for the query entries to be uplaoded to (before combination into one table)

106 code taken from https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.to\_sql.html

108 setting variables to be concatenated as part of url construction

116 Looping through search term gene names to get hits from GO gene products

118 construct the request URL for the API to use

128 Convert the Json into a Python object and print the type to make manipulation easier

153 This code should transfer the pandas dataFrame to the SQL table made earlier, however, I was not able to do this and as a result, work with true data ceases here.

156 ############### Interpro ###############

157 modified from https://gustavo-salazar.github.io/ProteinFamiliesTalks/InterProAPI.html?fbclid=IwAR29\_rfILGLKmzDCz7iUW-CMzLBI0pEeQPeF0b6fu8bN814EEsoaOGvUJL8#/25

175 Similarly to above, I could not find a way to translate the resulting information into an sql database from the pandas DataFrame.

179 For the purposes of this assignment, I have manually entered data from the three primary biological databases into three dummy SQL tables so that I can demonstrate how I would have manipulated such databases into a summary table. The lines of code 182-307 exist purely to demonstrate this.

308 The following block of code creates the summary table which integrates the information from three tables, relying on the gene symbol and accession numbers. The code was modified from the information at BD2\_answers\_SQL, author= Simon Tomlinson.

327 Closing the connection to my personal sql database to keep things tidy