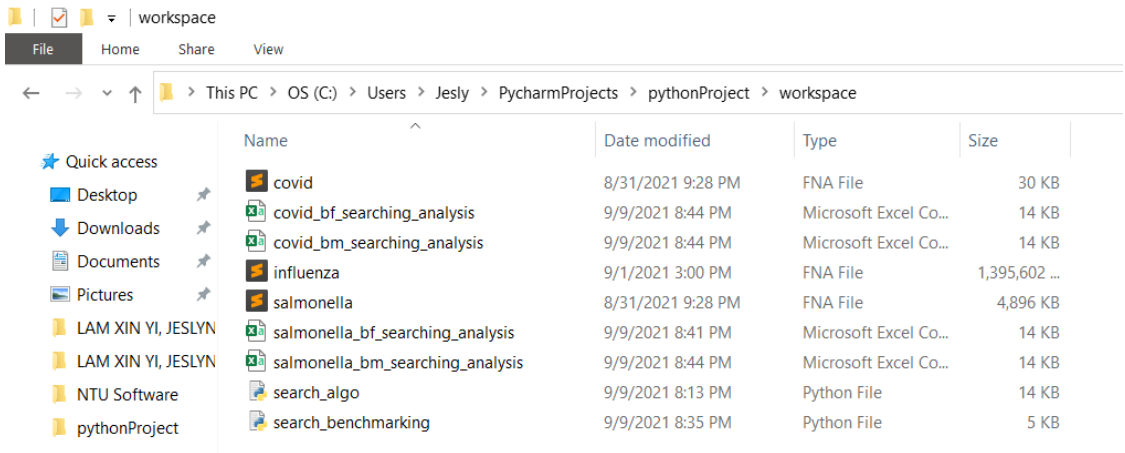
RECOMMENDED SET UP

It would be recommended that you run the python files in a folder created just for this project as the codes tend to create and generate many files, be it due to the chunking of fna files or the generation of the csv files.

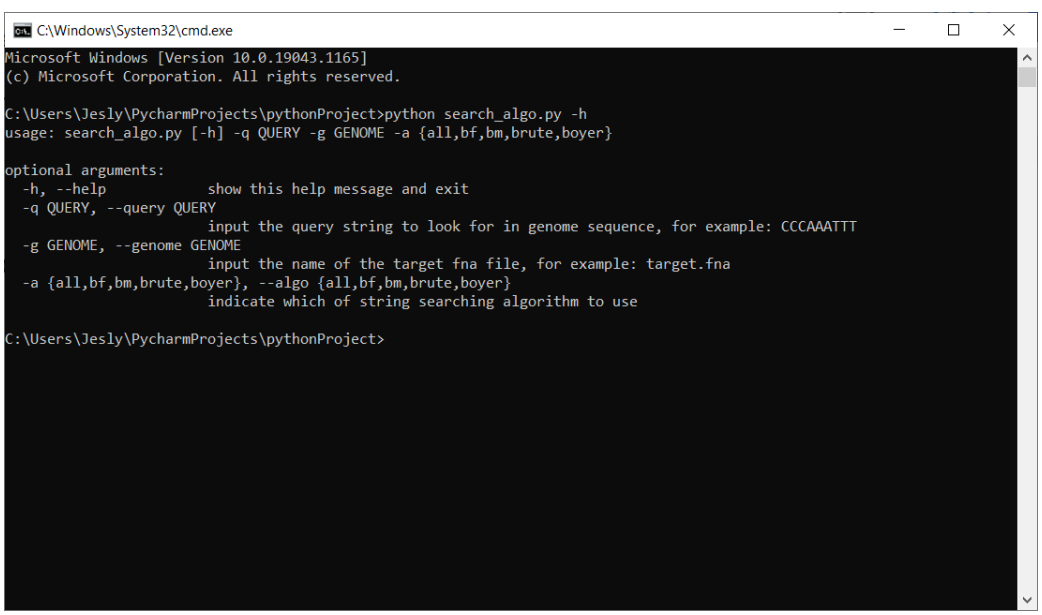


TO CHECK THE RESULTS

Do make use of the new file generated to look for index matches generated in the results. The file to be referred to would be the file appended with "new". Example: For the file "covid.fna", the file with the joined dna sequence would be named "new_covid.fna". This would contain the matched indexes.

PERFORM_SEARCH.PY

To perform a search it can be run through the cmd line and the -h flag would guide you on the extensions to be used.



For example, a search on the covid fna,
for the string “AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA”,
with all of the algorithms implemented can be done like this:

```
python search_algo.py -q AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -g covid.fna -a all
```

In blue, the -q flag represents the query flag where you would put the string that you are querying:

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAABBBBAAAAAAAAAAAAAAA
AAACCCGGGTTT
ACGTACGTACGT

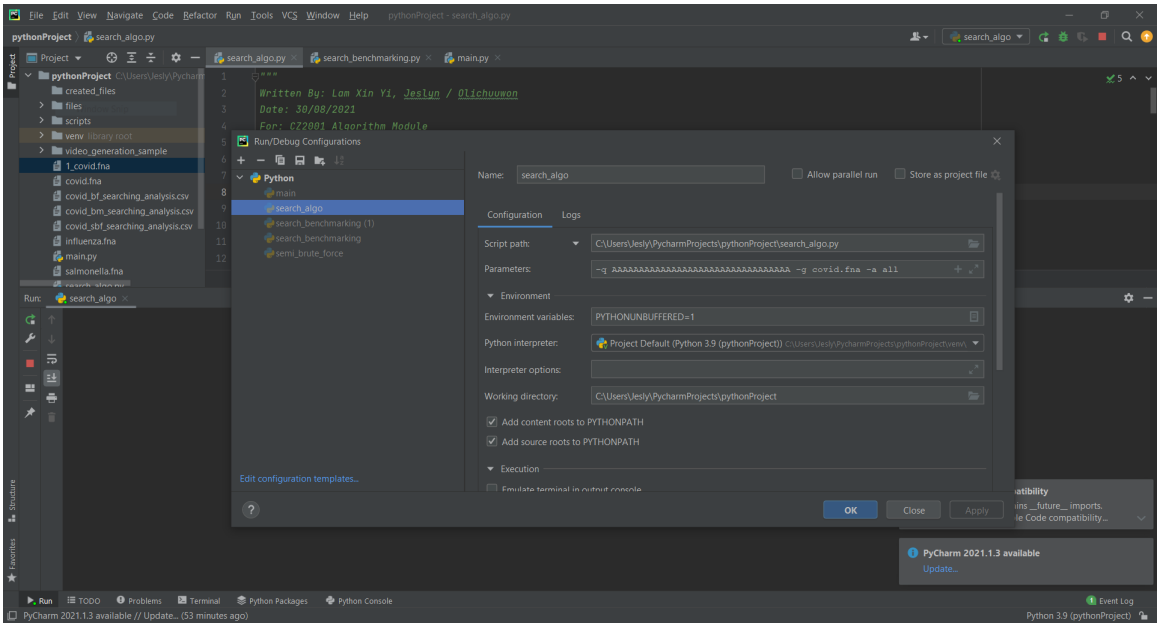
In purple, the -g flag represents the genome flag where you would input the name of the fna file that you would want to search through:

covid.fna
influenza.fna
salmonella.fna
target.fna

In red, the -a flag represents the algorithm flag where you would input the type of algorithm that you would want to test, there is also the “all” flag where you could test on all the algorithms that we have:

all - to test all algorithms
bf - only the brute force algorithm
brute - only the brute force algorithm
bm - only the boyer moore algorithm
boyer - only the boyer moore algorithm
kmp - only the knuth morris algorithm
kmp - only the knuth morris algorithm

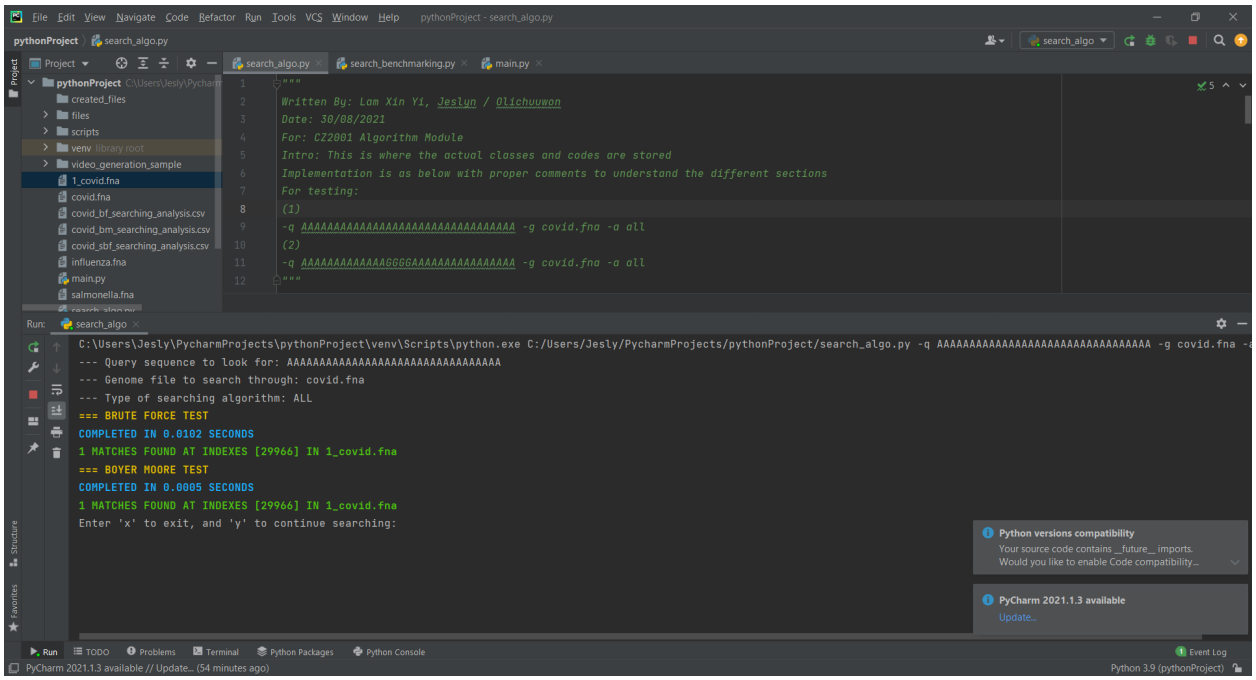
For running in PyCharm, you can directly include arguments in the “Parameters” field.



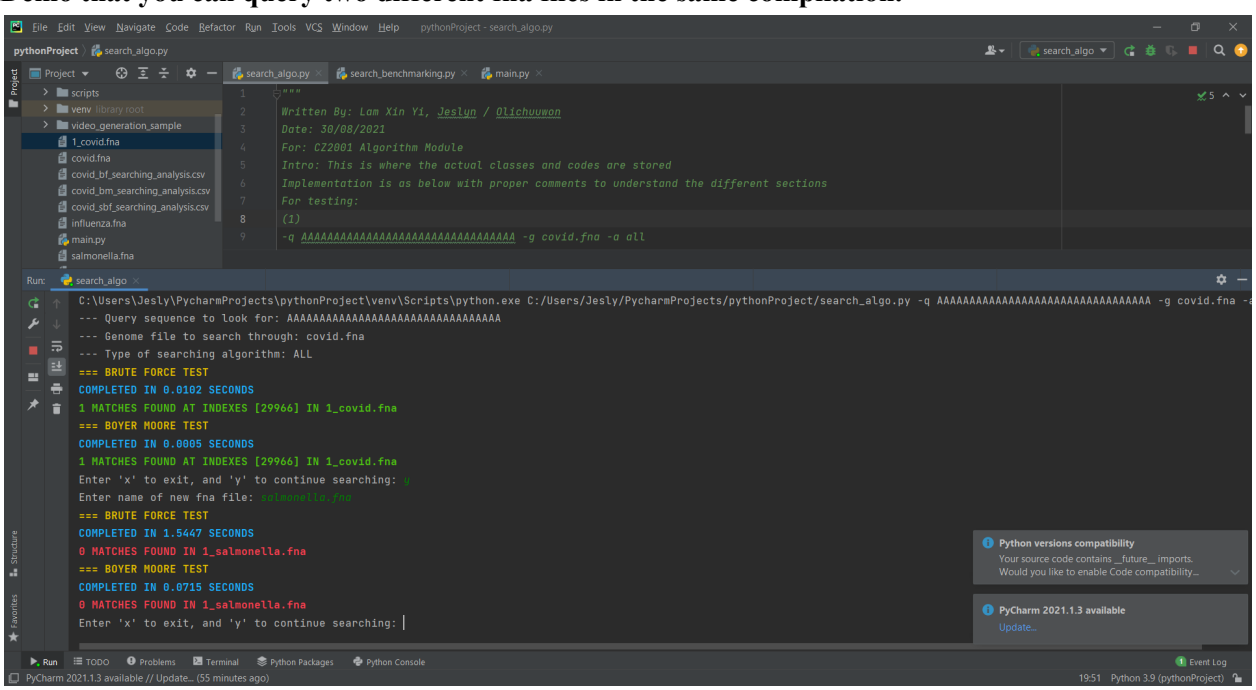
Parameters: -q AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -g covid.fna -a all

Sample output of the search with the input labeled as (1).

`-q AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -g covid.fna -a all`

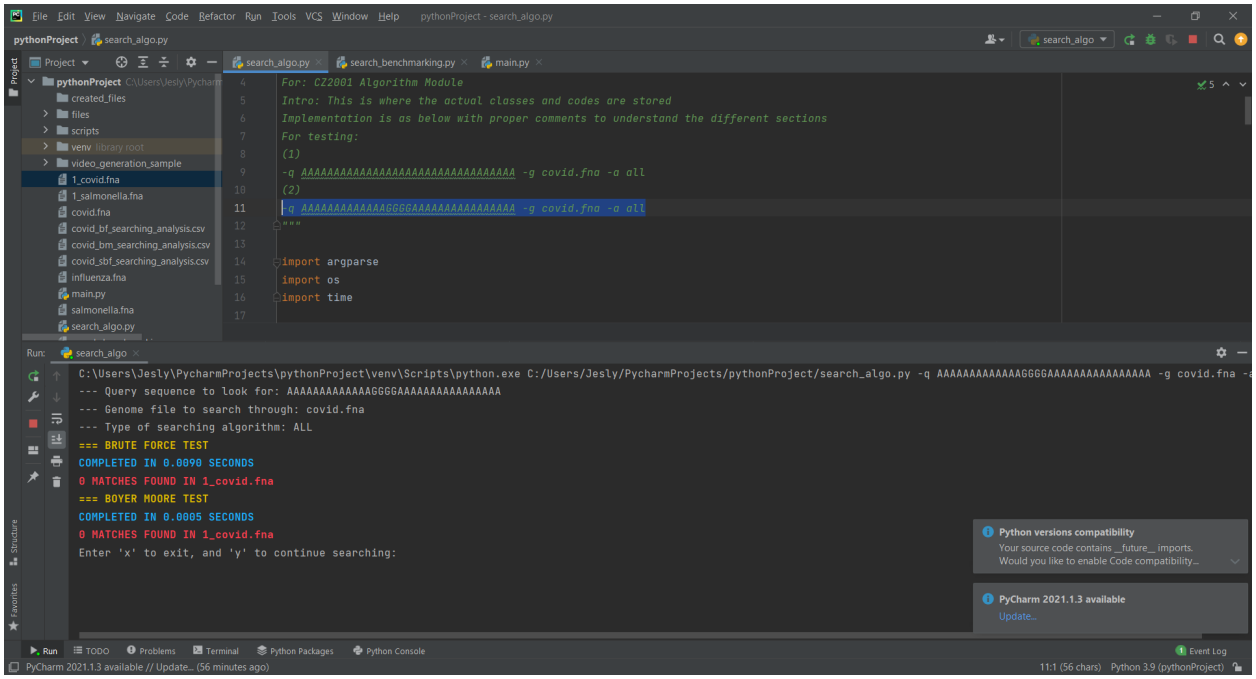


Demo that you can query two different fna files in the same compilation.



Demo of when there are no matches, it is clearly indicated, with query labeled (2).

`-q AAAAAAAAAAAAAGGGGAAAAAAAAAAAAAAAA -g covid.fna -a all`



Quick demo to show that chunking works - CHUNK_SIZE = 1000

```
--- Query sequence to look for: AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
--- Genome file to search through: covid.fna
--- Type of searching algorithm: ALL
=== BRUTE FORCE TEST
COMPLETED IN 0.0002651000 SECONDS
COMPLETED IN 0.0002794000 SECONDS
COMPLETED IN 0.0002762000 SECONDS
COMPLETED IN 0.0002767000 SECONDS
COMPLETED IN 0.0002817000 SECONDS
COMPLETED IN 0.0002757000 SECONDS
COMPLETED IN 0.0002809000 SECONDS
COMPLETED IN 0.0002679000 SECONDS
COMPLETED IN 0.0002784000 SECONDS
COMPLETED IN 0.0002631000 SECONDS
COMPLETED IN 0.0002682000 SECONDS
COMPLETED IN 0.0005218000 SECONDS
COMPLETED IN 0.0002782000 SECONDS
COMPLETED IN 0.0002723000 SECONDS
COMPLETED IN 0.0002725000 SECONDS
COMPLETED IN 0.0002738000 SECONDS
COMPLETED IN 0.0002706000 SECONDS
COMPLETED IN 0.0002727000 SECONDS
COMPLETED IN 0.0002752000 SECONDS
COMPLETED IN 0.0002684000 SECONDS
COMPLETED IN 0.0002777000 SECONDS
COMPLETED IN 0.0002764000 SECONDS
COMPLETED IN 0.0002721000 SECONDS
COMPLETED IN 0.0002689000 SECONDS
COMPLETED IN 0.0002764000 SECONDS
COMPLETED IN 0.0002731000 SECONDS
COMPLETED IN 0.0002694000 SECONDS
COMPLETED IN 0.0002693000 SECONDS
COMPLETED IN 0.0002713000 SECONDS
COMPLETED IN 0.0002783000 SECONDS
COMPLETED IN 0.0001023000 SECONDS
1 MATCHES FOUND AT INDEXES [335] IN 31_covid.fna
TOTAL OF 1 MATCHES FOUND AT INDEXES [29965] IN covid.fna
=== KNUTH MORRIS TEST
COMPLETED IN 0.0002148000 SECONDS
COMPLETED IN 0.0002310000 SECONDS
COMPLETED IN 0.0002278000 SECONDS
COMPLETED IN 0.0002293000 SECONDS
COMPLETED IN 0.0002329000 SECONDS
```

```
COMPLETED IN 0.0002279000 SECONDS
COMPLETED IN 0.0002327000 SECONDS
COMPLETED IN 0.0002221000 SECONDS
COMPLETED IN 0.0002294000 SECONDS
COMPLETED IN 0.0002165000 SECONDS
COMPLETED IN 0.0002218000 SECONDS
COMPLETED IN 0.0002167000 SECONDS
COMPLETED IN 0.0002279000 SECONDS
COMPLETED IN 0.0003599000 SECONDS
COMPLETED IN 0.0009201000 SECONDS
COMPLETED IN 0.0002379000 SECONDS
COMPLETED IN 0.0002310000 SECONDS
COMPLETED IN 0.0002302000 SECONDS
COMPLETED IN 0.0002308000 SECONDS
COMPLETED IN 0.0002275000 SECONDS
COMPLETED IN 0.0002353000 SECONDS
COMPLETED IN 0.0002365000 SECONDS
COMPLETED IN 0.0002302000 SECONDS
COMPLETED IN 0.0002282000 SECONDS
COMPLETED IN 0.0002350000 SECONDS
COMPLETED IN 0.0002313000 SECONDS
COMPLETED IN 0.0002376000 SECONDS
COMPLETED IN 0.0002226000 SECONDS
COMPLETED IN 0.0002250000 SECONDS
COMPLETED IN 0.0002322000 SECONDS
COMPLETED IN 0.0000868000 SECONDS
1 MATCHES FOUND AT INDEXES [335] IN 31_covid.fna
TOTAL OF 1 MATCHES FOUND AT INDEXES [29965] IN covid.fna
=== BOYER MOORE TEST
COMPLETED IN 0.0000225000 SECONDS
COMPLETED IN 0.0000182000 SECONDS
COMPLETED IN 0.0000186000 SECONDS
COMPLETED IN 0.0000168000 SECONDS
COMPLETED IN 0.0000163000 SECONDS
COMPLETED IN 0.0000167000 SECONDS
COMPLETED IN 0.0000162000 SECONDS
COMPLETED IN 0.0000170000 SECONDS
COMPLETED IN 0.0000169000 SECONDS
COMPLETED IN 0.0000161000 SECONDS
COMPLETED IN 0.0000161000 SECONDS
COMPLETED IN 0.0000161000 SECONDS
COMPLETED IN 0.0000169000 SECONDS
COMPLETED IN 0.0000173000 SECONDS
COMPLETED IN 0.0000162000 SECONDS
```

COMPLETED IN 0.0000174000 SECONDS
COMPLETED IN 0.0000163000 SECONDS
COMPLETED IN 0.0000166000 SECONDS
COMPLETED IN 0.0000167000 SECONDS
COMPLETED IN 0.0000170000 SECONDS
COMPLETED IN 0.0000175000 SECONDS
COMPLETED IN 0.0000169000 SECONDS
COMPLETED IN 0.0000160000 SECONDS
COMPLETED IN 0.0000171000 SECONDS
COMPLETED IN 0.0000174000 SECONDS
COMPLETED IN 0.0000161000 SECONDS
COMPLETED IN 0.0000166000 SECONDS
COMPLETED IN 0.0000178000 SECONDS
COMPLETED IN 0.0000160000 SECONDS
COMPLETED IN 0.0000168000 SECONDS
COMPLETED IN 0.0000148000 SECONDS
1 MATCHES FOUND AT INDEXES [335] IN 31_covid.fna
TOTAL OF 1 MATCHES FOUND AT INDEXES [29965] IN covid.fna
Enter 'x' to exit, and 'y' to continue searching:

Default chunk size - CHUNK_SIZE = 100000
--- Query sequence to look for: AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
--- Genome file to search through: covid.fna
--- Type of searching algorithm: ALL
=== BRUTE FORCE TEST
COMPLETED IN 0.0082989000 SECONDS
1 MATCHES FOUND AT INDEXES [29965] IN 1_covid.fna
TOTAL OF 1 MATCHES FOUND AT INDEXES [29965] IN covid.fna
=== KNUTH MORRIS TEST
COMPLETED IN 0.0070104000 SECONDS
1 MATCHES FOUND AT INDEXES [29965] IN 1_covid.fna
TOTAL OF 1 MATCHES FOUND AT INDEXES [29965] IN covid.fna
=== BOYER MOORE TEST
COMPLETED IN 0.0004375000 SECONDS
1 MATCHES FOUND AT INDEXES [29965] IN 1_covid.fna
TOTAL OF 1 MATCHES FOUND AT INDEXES [29965] IN covid.fna
Enter 'x' to exit, and 'y' to continue searching: