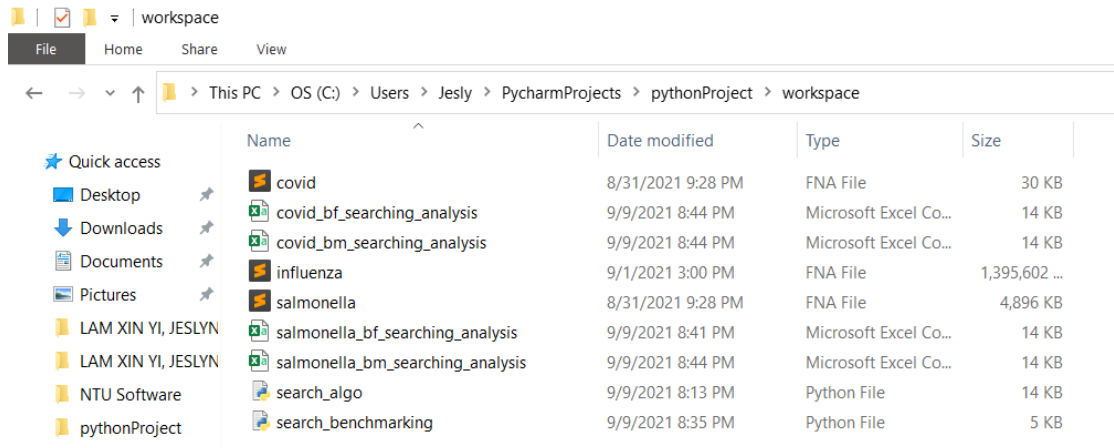


RECOMMENDED SET UP

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.



Name	Date modified	Type	Size
covid	8/31/2021 9:28 PM	FNA File	30 KB
covid_bf_searching_analysis	9/9/2021 8:44 PM	Microsoft Excel Co...	14 KB
covid_bm_searching_analysis	9/9/2021 8:44 PM	Microsoft Excel Co...	14 KB
influenza	9/1/2021 3:00 PM	FNA File	1,395,602 ...
salmonella	8/31/2021 9:28 PM	FNA File	4,896 KB
salmonella_bf_searching_analysis	9/9/2021 8:41 PM	Microsoft Excel Co...	14 KB
salmonella_bm_searching_analysis	9/9/2021 8:44 PM	Microsoft Excel Co...	14 KB
search_algo	9/9/2021 8:13 PM	Python File	14 KB
search_benchmarking	9/9/2021 8:35 PM	Python File	5 KB

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.

```
C:\Windows\System32\cmd.exe
Microsoft Windows [Version 10.0.19043.1165]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Jesly\PycharmProjects\pythonProject>python search_algo.py -h
usage: search_algo.py [-h] -q QUERY -g GENOME -a {all,bf,bm,brute,boyer}

optional arguments:
  -h, --help            show this help message and exit
  -q QUERY, --query QUERY
                        input the query string to look for in genome sequence, for example: CCCAAATTT
  -g GENOME, --genome GENOME
                        input the name of the target fna file, for example: target.fna
  -a {all,bf,bm,brute,boyer}, --algo {all,bf,bm,brute,boyer}
                        indicate which of string searching algorithm to use

C:\Users\Jesly\PycharmProjects\pythonProject>
```

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

```
python search_algo.py -q AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -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  
AAAAAAAAAAAAAABBBBAAAAAAAAAAAA  
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:

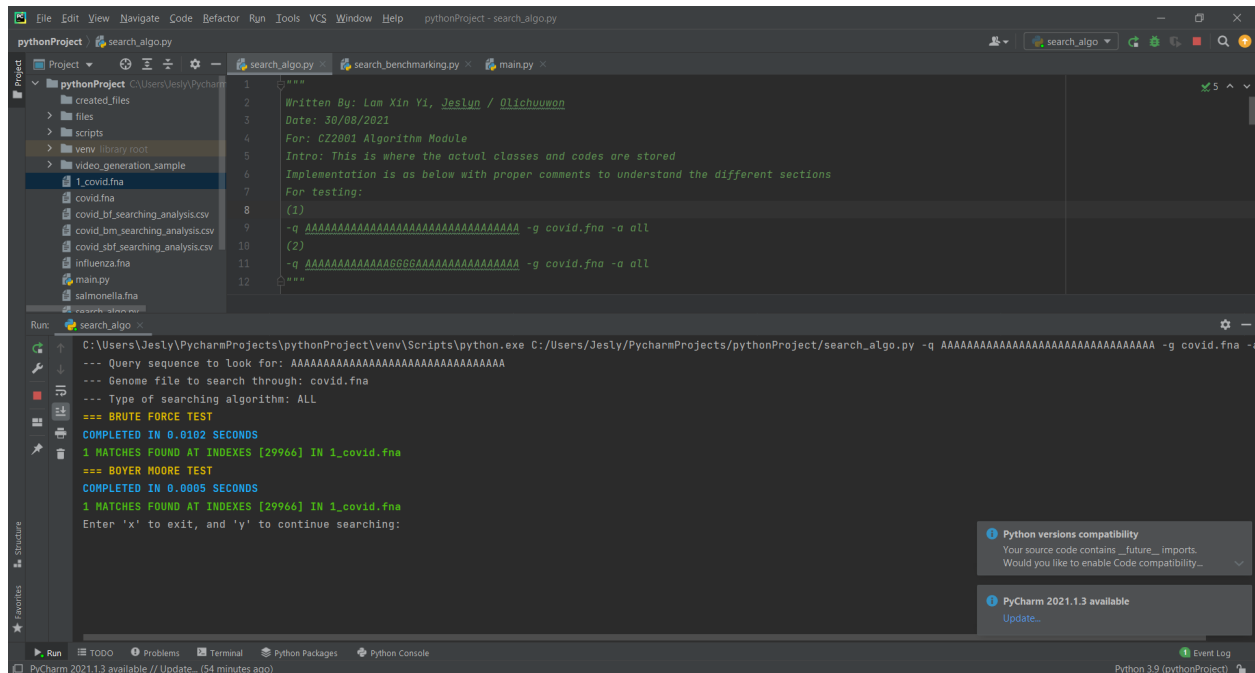
```
covid.fna  
influenza.fna  
salmonella.fna  
target.fna
```

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
```


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

-q AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -g covid.fna -a all

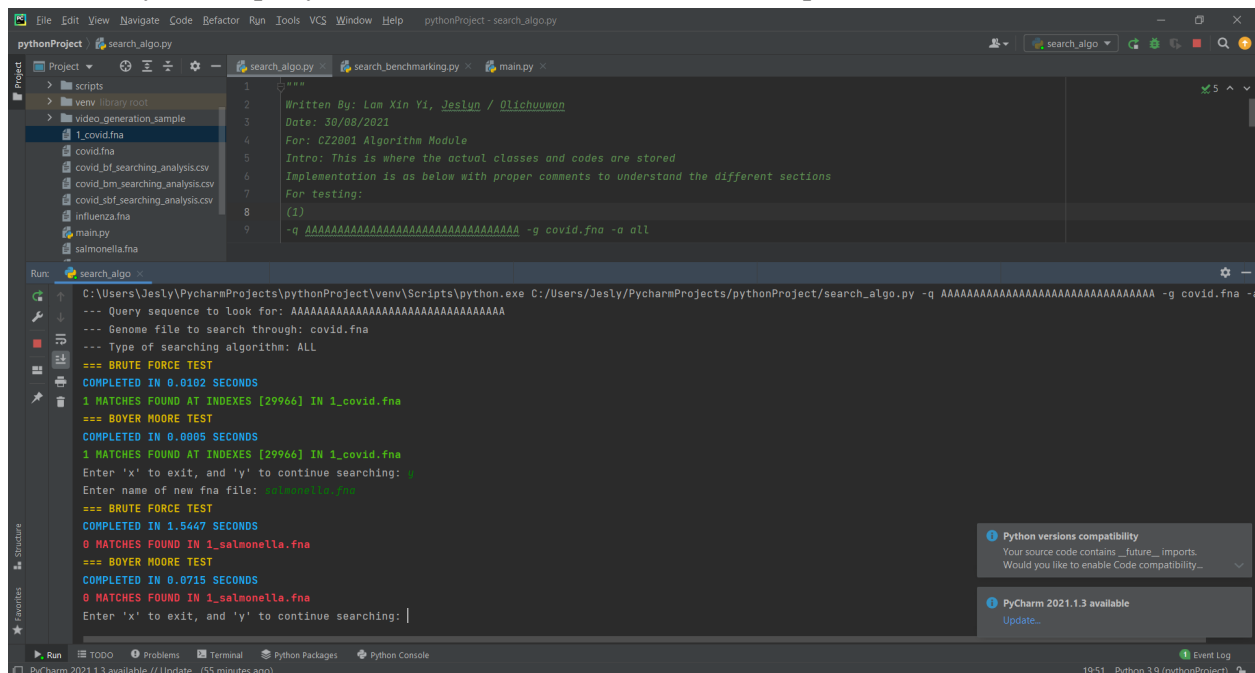


The screenshot shows the PyCharm IDE with the `search_algo.py` file open. The file contains a script for searching sequences in FNA files. The Run console shows the execution of the script with the command `-q AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -g covid.fna -a all`. The output indicates that the search was completed in 0.0102 seconds and found 1 match at index 29966 in `1_covid.fna`.

```
1 """
2 Written By: Lam Xin Yi, Jeslyn / Olichuwon
3 Date: 30/08/2021
4 For: CZ2001 Algorithm Module
5 Intro: This is where the actual classes and codes are stored
6 Implementation is as below with proper comments to understand the different sections
7 For testing:
8 (1)
9 -q AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -g covid.fna -a all
10 (2)
11 -q AAAAAAAAAAAAAAGGGAAAAAAAAAAAAAAAAAAAAA -g covid.fna -a all
12 """
```

```
Run: search_algo
C:\Users\Jesly\PycharmProjects\pythonProject\venv\Scripts\python.exe C:\Users\Jesly\PycharmProjects\pythonProject\search_algo.py -q AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -g covid.fna -a all
--- Query sequence to look for: AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
--- Genome file to search through: covid.fna
--- Type of searching algorithm: ALL
=== BRUTE FORCE TEST
COMPLETED IN 0.0102 SECONDS
1 MATCHES FOUND AT INDEXES [29966] IN 1_covid.fna
=== BOYER MOORE TEST
COMPLETED IN 0.0005 SECONDS
1 MATCHES FOUND AT INDEXES [29966] IN 1_covid.fna
Enter 'x' to exit, and 'y' to continue searching:
```

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

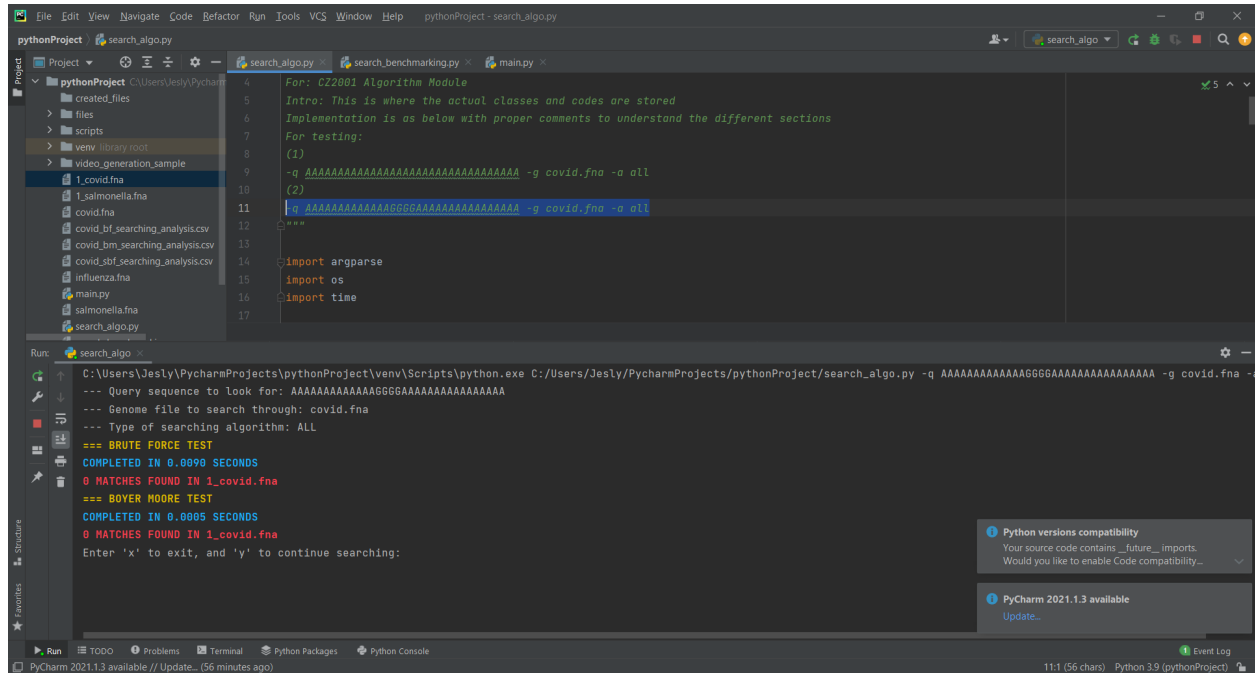


The screenshot shows the PyCharm IDE with the `search_algo.py` file open. The Run console shows the execution of the script with the command `-q AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -g covid.fna -a all`. The output indicates that the search was completed in 0.0102 seconds and found 1 match at index 29966 in `1_covid.fna`. The user then enters 'y' to continue searching and provides the name of a new FNA file, `salmonella.fna`. The script then performs a search on `salmonella.fna`, which is completed in 1.5447 seconds and found 0 matches.

```
Run: search_algo
C:\Users\Jesly\PycharmProjects\pythonProject\venv\Scripts\python.exe C:\Users\Jesly\PycharmProjects\pythonProject\search_algo.py -q AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -g covid.fna -a all
--- Query sequence to look for: AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
--- Genome file to search through: covid.fna
--- Type of searching algorithm: ALL
=== BRUTE FORCE TEST
COMPLETED IN 0.0102 SECONDS
1 MATCHES FOUND AT INDEXES [29966] IN 1_covid.fna
=== BOYER MOORE TEST
COMPLETED IN 0.0005 SECONDS
1 MATCHES FOUND AT INDEXES [29966] IN 1_covid.fna
Enter 'x' to exit, and 'y' to continue searching: y
Enter name of new fna file: salmonella.fna
=== BRUTE FORCE TEST
COMPLETED IN 1.5447 SECONDS
0 MATCHES FOUND IN 1_salmonella.fna
=== BOYER MOORE TEST
COMPLETED IN 0.0715 SECONDS
0 MATCHES FOUND IN 1_salmonella.fna
Enter 'x' to exit, and 'y' to continue searching: |
```

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

-q AAAAAAAAAAAAAAGGGGAAAAAAAAAAAAAAAA -g covid.fna -a all



The screenshot shows the PyCharm IDE interface. The top pane displays the source code of `search_algo.py`. The bottom pane shows the output of running the script. The code defines a function `search` that takes a query sequence, a genome file, and a search algorithm type. It performs a brute force search and a Boyer-Moore search. The output shows that no matches were found for the query `AAAAAAAAAAAAAGGGGAAAAAAAAAAAAAAAA` in the genome file `1_covid.fna`.

```
4 For: C22001 Algorithm Module
5 Intro: This is where the actual classes and codes are stored
6 Implementation is as below with proper comments to understand the different sections
7 For testing:
8 (1)
9 -q AAAAAAAAAAAAAAGGGGAAAAAAAAAAAAAAAA -g covid.fna -a all
10 (2)
11 -q AAAAAAAAAAAAAAGGGGAAAAAAAAAAAAAAAA -g covid.fna -a all
12
13
14 import argparse
15 import os
16 import time
17
```

Run: search_algo.py

```
C:\Users\Jesly\PycharmProjects\pythonProject\venv\Scripts\python.exe C:\Users\Jesly\PycharmProjects\pythonProject\search_algo.py -q AAAAAAAAAAAAAAGGGGAAAAAAAAAAAAAAAA -g covid.fna -a all
--- Query sequence to look for: AAAAAAAAAAAAAAGGGGAAAAAAAAAAAAAAAA
--- Genome file to search through: covid.fna
--- Type of searching algorithm: ALL
=== BRUTE FORCE TEST
COMPLETED IN 0.0090 SECONDS
0 MATCHES FOUND IN 1_covid.fna
=== BOYER MOORE TEST
COMPLETED IN 0.0005 SECONDS
0 MATCHES FOUND IN 1_covid.fna
Enter 'x' to exit, and 'y' to continue searching:
```

Python versions compatibility
Your source code contains `future` imports.
Would you like to enable Code compatibility...

PyCharm 2021.1.3 available
Update...

PyCharm 2021.1.3 available // Update... (56 minutes ago) 11:1 (56 chars) Python 3.9 (pythonProject)

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: AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
--- 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: