INF 503 - Homework 1

Jadon Fowler

To build the homework1 executable:

```
make
```

Problem 1A

```
./homework1 problem1A /common/contrib/classroom/inf503/hw_dataset.fa
```

- *How many unique sequence fragments are in each of the 14 datasets?*
- *How many total sequence fragments are in each dataset (i.e. when you consider copy numbers)?*

Dataset 1 has 3153996 unique fragments and 4000000 total fragments.
Dataset 2 has 2612547 unique fragments and 4000000 total fragments.
Dataset 3 has 2651696 unique fragments and 4000000 total fragments.
Dataset 4 has 3425236 unique fragments and 4000000 total fragments.
Dataset 5 has 3317288 unique fragments and 4000000 total fragments.
Dataset 6 has 3142908 unique fragments and 3735552 total fragments.
Dataset 7 has 3272592 unique fragments and 4000000 total fragments.
Dataset 8 has 3318272 unique fragments and 4000000 total fragments.
Dataset 9 has 3250754 unique fragments and 4000000 total fragments.
Dataset 10 has 3497973 unique fragments and 4000000 total fragments.
Dataset 11 has 3564950 unique fragments and 4000000 total fragments.
Dataset 12 has 3246447 unique fragments and 4000000 total fragments.
Dataset 13 has 3252996 unique fragments and 4000000 total fragments.
Dataset 14 has 3196619 unique fragments and 4000000 total fragments.

I don’t know why dataset 6 has a different number of total fragments… very suspicious…

Problem 1B

`./homework1 problem1B /common/contrib/classroom/inf503/hw_dataset.fa`

- *What is the 'big O' notation of your search (linear / quadratic / cubic / etc)?*

O(n)

- *How long does it take (in seconds) to search for all fragments of dataset 1 within dataset 2? Please note that depending on the efficiency of your algorithm, this step may take a long time. First estimate the total time using 1,000, 10,000, and 100,000 queries – if total time estimate is greater than 24 CPU hours, provide estimate rather than exact number.*

```
...1% (8967/896609) done in 166.876s...
...7% (62763/896609) done in 1167.38s...
...13% (116560/896609) done in 2167.89s...
...19% (170356/896609) done in 3168.51s...
...25% (224153/896609) done in 4169.01s...
...31% (277949/896609) done in 5169.44s...
...37% (331746/896609) done in 6121.42s...
...43% (385542/896609) done in 6977.81s…
srun: Force Terminated job 50234090
srun: Job step aborted: Waiting up to 32 seconds for job step to
finish.
slurmstepd: error: *** STEP 50234090.0 ON cn69 CANCELLED AT
2022-02-15T17:11:54 DUE TO TIME LIMIT ***
srun: error: cn69: task 0: Terminated
```

^ This is my run with a sample of that data. The job was killed because it took too long.

There are 3,153,996 unique fragments in dataset 1 and 2,612,547 unique fragments in dataset 2. My example run took about 4000 seconds to check 200,000 fragments. Checking all of dataset 1 would take about 63,000s or 17.5 hours.

- *How many sequence fragments in dataset 1 are also in dataset 2?*

Estimate is > 1000

Problem 1C

`./homework1 problem1C /common/contrib/classroom/inf503/hw_dataset.fa`

- *What is the 'big O' notation of your search (linear / quadratic / cubic / etc)?*

O(log n)

- *How long (in seconds) does it take to search for 1000 queries? How about 10,000 or 100,000? Does the time increase make sense? Explain the differences (if any) when compared to search times obtained as part of 1B.*

6 seconds! We are comparing far fewer strings than part 1B because the pre-sorted lists let us skip huge sections when searching.

- *How many sequence fragments in dataset 1 are also in dataset 2?*

```
...sorting the first dataset...
...sorting the second dataset...
...1% (31540/3153996) done in 4.72368s...
...7% (220780/3153996) done in 4.79562s...
...13% (410020/3153996) done in 4.86618s...
...19% (599260/3153996) done in 4.94327s...
...25% (788499/3153996) done in 5.01565s...
...31% (977739/3153996) done in 5.08701s...
...37% (1166979/3153996) done in 5.15538s...
...43% (1356219/3153996) done in 5.21859s...
...49% (1545459/3153996) done in 5.287s...
...55% (1734698/3153996) done in 5.35412s...
...61% (1923938/3153996) done in 5.41904s...
...67% (2113178/3153996) done in 5.48564s...
...73% (2302418/3153996) done in 5.55445s...
...79% (2491657/3153996) done in 5.63219s...
...85% (2680897/3153996) done in 5.7035s...
...91% (2870137/3153996) done in 5.77027s...
...97% (3059377/3153996) done in 5.84022s...
```

There are 1386 fragments from dataset 1 in dataset 2.
Searching took 5.87697s.