CSCI 5481: Homework 1

Brian Cooper

Homework Overview

- `homework1.py` : Python script for executing the BURST executable (modified from `template.py`)
- `q4.py` : Python code to answ er question 4
- `output.txt` : Output file generated by BURST

`homework1.py` Program Output (stdout)

- Input:

```
python homework1.py -q query.fna -r ref.fna -t taxonomy.txt -c ./burst -o output.txt -V
```

- Output:

```
Command: ./burst -r ref.fna -q query.fna --taxonomy taxonomy.txt -o output.txt

Return value: 0

 --> Assigning taxonomy based on input file taxonomy.txt.
Using up to AVX-128 with 8 threads.
Parsed 130727 queries.
Max query len: 101, avg. divergence: 90.764586 (18.509566 w/o dupes)
Parsed 5000 references.
There are 5000 references and hence 312 clumps (+1)
Average R pack length = 1423.881789
Searching best paths through 35294 unique queries...

Search Progress: [0.32%]
Search Progress: [0.64%]
Search Progress: [2.88%]
Search Progress: [6.07%]
Search Progress: [8.95%]
Search Progress: [11.18%]
Search Progress: [15.02%]
Search Progress: [18.85%]
Search Progress: [20.77%]
Search Progress: [23.00%]
Search Progress: [25.24%]
Search Progress: [27.16%]
Search Progress: [30.67%]
Search Progress: [32.27%]
Search Progress: [33.87%]
Search Progress: [36.10%]
Search Progress: [37.70%]
Search Progress: [39.94%]
Search Progress: [41.85%]
Search Progress: [44.73%]
Search Progress: [47.28%]
Search Progress: [50.48%]
Search Progress: [52.40%]
Search Progress: [55.59%]
Search Progress: [58.15%]
Search Progress: [61.02%]
Search Progress: [63.58%]
Search Progress: [66.45%]
Search Progress: [69.01%]
Search Progress: [71.25%]
Search Progress: [74.12%]
Search Progress: [76.68%]
Search Progress: [78.59%]
Search Progress: [80.83%]
Search Progress: [83.39%]
Search Progress: [85.62%]
Search Progress: [87.86%]
Search Progress: [89.14%]
Search Progress: [91.69%]
Search Progress: [94.25%]
Search Progress: [96.17%]
Search Progress: [98.72%]
Search Progress: [100.00%]
Search complete. Consolidating results...

Alignment time: 2.216912 seconds

No errors!
```

`q4.py` Program Output (stdout)

- Input:

```
python q4.py
```

- Output:

```
Question 4a
===========
Sequences that match database at 97%+ similarity:     34.8436%

Question 4b
===========
Most common determinate species:    s__Faecalibacterium_prausnitzii

Question 4c
===========
Average percent similarity of matches at 97%+ similarity:     98.4779%
```

Question 4 Answers

1. Of the original input query sequences, approximately **34.84%** had a match in the reference database at 97% or above.
2. The most common (determinate) bacterial species in the query set is *Faecalibacterium prausnitzii*.
3. The average percent similarity of the matches among those above 97% is approximately **98.48%**.