CSCI 5481 Homework 1

Christopher White

September 9, 2018

Homework 1: BURST

PRINT PYTHON FILE

```
%cat exercise01.py
```

```
# Output file for CSCI 5481 Fall 2018 Exercise 1
# Christopher White
# September 19, 2018
#
# Usage:
# exercise01.py -h
import sys, os
import argparse
from subprocess import Popen, PIPE

def make_arg_parser():
    parser = argparse.ArgumentParser(prog='exercise01.py',
                          # version="%prog 1.0",
                          formatter_class=argparse.ArgumentDefaultsHelpFormatter)
    parser.add_argument("-q","--query",
                        default=argparse.SUPPRESS,
                        required=True,
                        help="Path to query fasta [required]")
    parser.add_argument("-r","--ref",
                        default=argparse.SUPPRESS,
                        required=True,
                        help="Path to reference fasta [required]")
    parser.add_argument("-t","--taxonomy",
                        default=None,
                        required=True,
```

```
                    help="Path to taxonomy file [required]")
    parser.add_argument("-o","--output",
                    default=None,
                    required=True,
                    help="Path to output file [required]")
    parser.add_argument("-c","--command",
                    default='./burst',
                    help="Path to BURST command")
    parser.add_argument("-V","--verbose",
                    action="store_true",
                    help="Verbose output")
    return parser

# Runs BURST to search query sequences against reference sequences
def run_burst(query, ref, taxonomy, output, burst_cmd='./burst', verbose=False):
    """thread worker function"""

    cmd = burst_cmd + ' -q ' + query + ' -r ' + ref + ' -t ' + taxonomy + ' -o ' + output

    return run_command(cmd, verbose=verbose)

# runs the given command and returns return value and output
def run_command(cmd, verbose=False):
    if verbose:
        print(cmd)

    proc = Popen(cmd,shell=True,universal_newlines=True,stdout=PIPE,stderr=PIPE)

    stdout, stderr = proc.communicate('Running command')
    return_val = proc.returncode
    return str(return_val), stdout, stderr


if __name__ == '__main__':
    parser = make_arg_parser()
    args = parser.parse_args()

    return_value, stdout, stderr = run_burst(args.query, args.ref, args.taxonomy, args.outpu

    print('\nReturn Value: '+return_value)
    print('\nSTDOUT: ...\n'+stdout)
    print('\nSTDERR: '+stderr)
    print('---')
```

RUN COMMAND LINE

```
!python exercise01.py -q query.fna -r ref.fna -t taxonomy.txt -o output.txt -c ./burst -V
./burst -q query.fna -r ref.fna -t taxonomy.txt -o output.txt

Return Value: 0

STDOUT: ...
 --> Setting threads to 0
Using up to AVX-128 with 0 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: [0.96%]
Search Progress: [1.28%]
Search Progress: [1.60%]
Search Progress: [1.92%]
Search Progress: [2.24%]
Search Progress: [2.56%]
Search Progress: [2.88%]
Search Progress: [3.19%]
Search Progress: [3.51%]
Search Progress: [3.83%]
Search Progress: [4.15%]
Search Progress: [4.47%]
Search Progress: [4.79%]
Search Progress: [5.11%]
Search Progress: [5.43%]
Search Progress: [5.75%]
Search Progress: [6.07%]
Search Progress: [6.39%]
Search Progress: [6.71%]
Search Progress: [7.03%]
Search Progress: [7.35%]
Search Progress: [7.67%]
Search Progress: [7.99%]
Search Progress: [8.31%]
Search Progress: [8.63%]
Search Progress: [8.95%]
```

```
Search Progress: [9.27%]
Search Progress: [9.58%]
Search Progress: [9.90%]
Search Progress: [10.22%]
Search Progress: [10.54%]
Search Progress: [10.86%]
Search Progress: [11.18%]
Search Progress: [11.50%]
Search Progress: [11.82%]
Search Progress: [12.14%]
Search Progress: [12.46%]
Search Progress: [12.78%]
Search Progress: [13.10%]
Search Progress: [13.42%]
Search Progress: [13.74%]
Search Progress: [14.06%]
Search Progress: [14.38%]
Search Progress: [14.70%]
Search Progress: [15.02%]
Search Progress: [15.34%]
Search Progress: [15.65%]
Search Progress: [15.97%]
Search Progress: [16.29%]
Search Progress: [16.61%]
Search Progress: [16.93%]
Search Progress: [17.25%]
Search Progress: [17.57%]
Search Progress: [17.89%]
Search Progress: [18.21%]
Search Progress: [18.53%]
Search Progress: [18.85%]
Search Progress: [19.17%]
Search Progress: [19.49%]
Search Progress: [19.81%]
Search Progress: [20.13%]
Search Progress: [20.45%]
Search Progress: [20.77%]
Search Progress: [21.09%]
Search Progress: [21.41%]
Search Progress: [21.73%]
Search Progress: [22.04%]
Search Progress: [22.36%]
Search Progress: [22.68%]
Search Progress: [23.00%]
Search Progress: [23.32%]
Search Progress: [23.64%]
```

```
Search Progress: [23.96%]
Search Progress: [24.28%]
Search Progress: [24.60%]
Search Progress: [24.92%]
Search Progress: [25.24%]
Search Progress: [25.56%]
Search Progress: [25.88%]
Search Progress: [26.20%]
Search Progress: [26.52%]
Search Progress: [26.84%]
Search Progress: [27.16%]
Search Progress: [27.48%]
Search Progress: [27.80%]
Search Progress: [28.12%]
Search Progress: [28.43%]
Search Progress: [28.75%]
Search Progress: [29.07%]
Search Progress: [29.39%]
Search Progress: [29.71%]
Search Progress: [30.03%]
Search Progress: [30.35%]
Search Progress: [30.67%]
Search Progress: [30.99%]
Search Progress: [31.31%]
Search Progress: [31.63%]
Search Progress: [31.95%]
Search Progress: [32.27%]
Search Progress: [32.59%]
Search Progress: [32.91%]
Search Progress: [33.23%]
Search Progress: [33.55%]
Search Progress: [33.87%]
Search Progress: [34.19%]
Search Progress: [34.50%]
Search Progress: [34.82%]
Search Progress: [35.14%]
Search Progress: [35.46%]
Search Progress: [35.78%]
Search Progress: [36.10%]
Search Progress: [36.42%]
Search Progress: [36.74%]
Search Progress: [37.06%]
Search Progress: [37.38%]
Search Progress: [37.70%]
Search Progress: [38.02%]
Search Progress: [38.34%]
```

```
Search Progress: [38.66%]
Search Progress: [38.98%]
Search Progress: [39.30%]
Search Progress: [39.62%]
Search Progress: [39.94%]
Search Progress: [40.26%]
Search Progress: [40.58%]
Search Progress: [40.89%]
Search Progress: [41.21%]
Search Progress: [41.53%]
Search Progress: [41.85%]
Search Progress: [42.17%]
Search Progress: [42.49%]
Search Progress: [42.81%]
Search Progress: [43.13%]
Search Progress: [43.45%]
Search Progress: [43.77%]
Search Progress: [44.09%]
Search Progress: [44.41%]
Search Progress: [44.73%]
Search Progress: [45.05%]
Search Progress: [45.37%]
Search Progress: [45.69%]
Search Progress: [46.01%]
Search Progress: [46.33%]
Search Progress: [46.65%]
Search Progress: [46.96%]
Search Progress: [47.28%]
Search Progress: [47.60%]
Search Progress: [47.92%]
Search Progress: [48.24%]
Search Progress: [48.56%]
Search Progress: [48.88%]
Search Progress: [49.20%]
Search Progress: [49.52%]
Search Progress: [49.84%]
Search Progress: [50.16%]
Search Progress: [50.48%]
Search Progress: [50.80%]
Search Progress: [51.12%]
Search Progress: [51.44%]
Search Progress: [51.76%]
Search Progress: [52.08%]
Search Progress: [52.40%]
Search Progress: [52.72%]
Search Progress: [53.04%]
```

```
Search Progress: [53.35%]
Search Progress: [53.67%]
Search Progress: [53.99%]
Search Progress: [54.31%]
Search Progress: [54.63%]
Search Progress: [54.95%]
Search Progress: [55.27%]
Search Progress: [55.59%]
Search Progress: [55.91%]
Search Progress: [56.23%]
Search Progress: [56.55%]
Search Progress: [56.87%]
Search Progress: [57.19%]
Search Progress: [57.51%]
Search Progress: [57.83%]
Search Progress: [58.15%]
Search Progress: [58.47%]
Search Progress: [58.79%]
Search Progress: [59.11%]
Search Progress: [59.42%]
Search Progress: [59.74%]
Search Progress: [60.06%]
Search Progress: [60.38%]
Search Progress: [60.70%]
Search Progress: [61.02%]
Search Progress: [61.34%]
Search Progress: [61.66%]
Search Progress: [61.98%]
Search Progress: [62.30%]
Search Progress: [62.62%]
Search Progress: [62.94%]
Search Progress: [63.26%]
Search Progress: [63.58%]
Search Progress: [63.90%]
Search Progress: [64.22%]
Search Progress: [64.54%]
Search Progress: [64.86%]
Search Progress: [65.18%]
Search Progress: [65.50%]
Search Progress: [65.81%]
Search Progress: [66.13%]
Search Progress: [66.45%]
Search Progress: [66.77%]
Search Progress: [67.09%]
Search Progress: [67.41%]
Search Progress: [67.73%]
```

```
Search Progress: [68.05%]
Search Progress: [68.37%]
Search Progress: [68.69%]
Search Progress: [69.01%]
Search Progress: [69.33%]
Search Progress: [69.65%]
Search Progress: [69.97%]
Search Progress: [70.29%]
Search Progress: [70.61%]
Search Progress: [70.93%]
Search Progress: [71.25%]
Search Progress: [71.57%]
Search Progress: [71.88%]
Search Progress: [72.20%]
Search Progress: [72.52%]
Search Progress: [72.84%]
Search Progress: [73.16%]
Search Progress: [73.48%]
Search Progress: [73.80%]
Search Progress: [74.12%]
Search Progress: [74.44%]
Search Progress: [74.76%]
Search Progress: [75.08%]
Search Progress: [75.40%]
Search Progress: [75.72%]
Search Progress: [76.04%]
Search Progress: [76.36%]
Search Progress: [76.68%]
Search Progress: [77.00%]
Search Progress: [77.32%]
Search Progress: [77.64%]
Search Progress: [77.96%]
Search Progress: [78.27%]
Search Progress: [78.59%]
Search Progress: [78.91%]
Search Progress: [79.23%]
Search Progress: [79.55%]
Search Progress: [79.87%]
Search Progress: [80.19%]
Search Progress: [80.51%]
Search Progress: [80.83%]
Search Progress: [81.15%]
Search Progress: [81.47%]
Search Progress: [81.79%]
Search Progress: [82.11%]
Search Progress: [82.43%]
```

```
Search Progress: [82.75%]
Search Progress: [83.07%]
Search Progress: [83.39%]
Search Progress: [83.71%]
Search Progress: [84.03%]
Search Progress: [84.35%]
Search Progress: [84.66%]
Search Progress: [84.98%]
Search Progress: [85.30%]
Search Progress: [85.62%]
Search Progress: [85.94%]
Search Progress: [86.26%]
Search Progress: [86.58%]
Search Progress: [86.90%]
Search Progress: [87.22%]
Search Progress: [87.54%]
Search Progress: [87.86%]
Search Progress: [88.18%]
Search Progress: [88.50%]
Search Progress: [88.82%]
Search Progress: [89.14%]
Search Progress: [89.46%]
Search Progress: [89.78%]
Search Progress: [90.10%]
Search Progress: [90.42%]
Search Progress: [90.73%]
Search Progress: [91.05%]
Search Progress: [91.37%]
Search Progress: [91.69%]
Search Progress: [92.01%]
Search Progress: [92.33%]
Search Progress: [92.65%]
Search Progress: [92.97%]
Search Progress: [93.29%]
Search Progress: [93.61%]
Search Progress: [93.93%]
Search Progress: [94.25%]
Search Progress: [94.57%]
Search Progress: [94.89%]
Search Progress: [95.21%]
Search Progress: [95.53%]
Search Progress: [95.85%]
Search Progress: [96.17%]
Search Progress: [96.49%]
Search Progress: [96.81%]
Search Progress: [97.12%]
```

```
Search Progress: [97.44%]
Search Progress: [97.76%]
Search Progress: [98.08%]
Search Progress: [98.40%]
Search Progress: [98.72%]
Search Progress: [99.04%]
Search Progress: [99.36%]
Search Progress: [99.68%]
Search Progress: [100.00%]
Search Progress: [100.00%]
Search complete. Consolidating results...

Alignment time: 7.511785 seconds


STDERR:
---
```

PROCESS OUTPUT FILE

```
import pandas as pd
import numpy as np

df = pd.read_csv('output.txt', header=None, delim_whitespace=True,
                 names=['qseqid','sseqid','pident','length','mismatch','gapopen','qstart','qe

df.head(5)
```

qseqid

sseqid

pident

length

mismatch

gapopen

qstart

qend

sstart

send

evalue

bitscore

0

h146M.1.418838_256197075

851865

97.029701

101

3

0

1

101

308

409

3

0

1

h165M.1.418394_282943524

851865

97.029701

101

3

0

1

101

308

409

3

1

2

Amz5chldF2.418405_806016025

851865

97.826088

92

2

0

1

92

308

400

2

0

3

Amz4chldM2.418774_769559065

851865

97.029701

101

3

0

1

101

308

409

3

0

4

h257M.1.418454_646157849

851865

96.969696

99

3

0

1

99

308

407

3

0

Question 1: What fraction of the input query sequences had a match in the database at 97% or above?

```
df[df['pident']>=97].count()['qseqid'], df['pident'].count()
```

(45550, 46011)

There are 45,550 at 97% or above out of 46,011 entries.

Question 2: What is the most common bacterial species in the query set?

```
df['qseqid'].value_counts()
```

k278A.2.418424_651300242 1 h9M.1.418588_808479870 1 Amz6chldM.418668_793073478
1 TS109.418691_885856118 1 Amz6teen.418569_1029400378 1 TS129.418618_381258008
1 USchp33ChildB.418578_213517173 1 TS7.418860_392489734 1 TS195.418848_408859884
1 h95M.1.418831_743636910 1 TS129.418618_398711882 1 h264M.1.418377_251539154
1 h147M.1.418531_262271267 1 TS111.418684_365510845 1 USchp33ChildB.418578_234887031
1 h9M.1.418588_777264590 1 Amz29adlt.418370_1021060745 1 k278A.2.418424_647844107
1 USchp25Child.418345_206657105 1 TS109.418691_921630454 1 Amz29adlt.418370_1017418414
1 TS111.418684_379500016 1 TS1.418828_346501103 1 USchp18Mom.418783_228226310
1 Amz6chldM.418668_767329927 1 TS1.418828_355543881 1 h273M.1.418507_296460623
1 h68M.1.418773_188195634 1 TS109.418691_947299040 1 h279M.1.418530_648185250
1 .. USchp4Mom.418666_192812823 1 Amz5chldF1.418757_1012389222 1
Amz30adlt.418837_766998154 1 TS193.418750_406059473 1 TS4.418810_367555296
1 TS25.418407_916630067 1 Amz4adltF.418711_801824048 1 h165M.1.418394_292960889
1 Amz4chldM2.418774_768826770 1 h101M.1.418586_633813835 1 Amz4chldM2.418774_816398505
1 h146M.1.418838_251654143 1 TS195.418848_376963069 1 USchp36Mom.418718_375555754
1 Amz4adltF.418711_819346976 1 h146M.1.418838_270865891 1 TS193.418750_357094037
1 USchp3Mom.418727_892595014 1 Amz30adlt.418837_743246919 1
TS193.418750_363283526 1 Amz5chldF2.418405_762273143 1 TS193.418750_364868631
1 TS129.418618_375016390 1 TS195.418848_410389318 1 TS195.418848_359383225
1 USchp1Mom.418814_911228524 1 TS111.418684_348606674 1 h146M.1.418838_255619300
1 h257M.1.418454_584706830 1 h235M.1.418489_623997400 1 Name: qseqid,
Length: 46011, dtype: int64

The answer is that all of the id's are unique.

Question 3: What is the average percent similarity of the matches?

```
np.average(df['pident'])
```

98.46259236202215

The average of the percent similarity is 98.4626%.