Programming for Bioinformatics | BIOL7200

Week 13 Exercise

November 22, 2022

Assume that the user gives you correct inputs all the time. **Your script will be graded on the output produced and not on how the errors are handled.**

For this assignment, the module list is "`multiprocessing`", "`sys`", "`re`", "`os`", "`subprocess`", and "`argparse`". Do not use `input()` for any input.

Instructions for submission

- **This assignment is also due Monday, Dec 6th, 2022 at 11:59pm. Late submissions will not be graded**
- Name your script as `parallel_ani.py`
- Your code should run as `./parallel_ani.py -o <Output file> [-t <Number of threads>] fasta_file1 fasta_file2 fasta_file3…`
- **DO NOT HARDCODE** any file name!
- Use `#!/usr/bin/env python3` as your shebang
- **Your script should finish within 5min for 10 input genomes and 5 threads, partial credit will be awarded up to 10 min of run time.**

Another common task in bioinformatics – running the same task for different inputs. The use case here will be computing average nucleotide identity (ANI) for each pair (all-against-all pairs) of input fasta file. The fasta files are microbial genome sequences. The ANI is being calculated using MUMmer's dnadiff. Install the MUMmer package yourself from here: https://mummer4.github.io/. Assume that dnadiff is in our environment PATH.

Your objective is simple – you are given a set of files, say A.fasta, B.fasta and C.fasta. You must calculate the pairwise distance between them and print them in a matrix format as follows (Each column is tab-separated):

	A.fasta	B.fasta	C.fasta
A.fasta	100		
B.fasta		100	
C.fasta			100

The threads argument is key. The threads specify how many parallel instances of pairwise ANI calculations should be performed. So, if the user says -t 3, launch 3 ANI computations (obviously for different pairs) simultaneously. Your program should finish ~3 times faster for -t 3 when compared to a single thread (-t 1).

Helpful notes:

1. You will have to run dnadiff like this:

   ```
   dnadiff -p <unique prefix> file1.fasta file2.fasta
   ```

   E.g.,

   ```
   dnadiff -p output123 genome1.fasta genome2.fasta
   ```

   This will create a bunch of files starting with output123

2. The file with extension .report will have the ANI (in the 19th line – both numeric columns will have the same values).