Computing for Medicine: Phase 3, Seminar 4 Project

Michelle Craig
Associate Professor, Teaching Stream
mcraig@cs.toronto.edu

Seminar 4 Project

- The project handout is posted:
  - http://c4m.cdf.toronto.edu/cohort2/phase3/
- Two approaches for doing your work:
  - Use the Computer Science Teaching Labs computing network.
  - Use your personal computer.
- Python3 packages to install:
  - `Biopython`
    - http://biopython.org/wiki/Download
    - `pip3 install biopython`

GLOB

Python's `glob` module

- https://docs.python.org/3/library/glob.html
- Used to find files whose names match a given pattern.
- Symbols used:
  - * (matches zero or more characters)
  - ? (matches exactly one character)
  - [] (matches one character contained within the brackets)

Demo

- Example directory contains the following files:
  - a.txt, apple.txt, b.jpg, banana.txt, carrot.txt, carrot.jpg

```
>>> glob.glob('*.txt')
['a.txt', 'apple.txt', 'banana.txt', 'carrot.txt']
>>> glob.glob('*.jpg')
['b.jpg', 'carrot.jpg']
>>> glob.glob('?.txt')
['a.txt']
>>> glob.glob('?.*')
['a.txt', 'b.jpg']
```

Demo (continued)

```
>>> glob.glob('a*')
['a.txt', 'apple.txt']
>>> glob.glob('*a*')
['a.txt', 'apple.txt', 'banana.txt',
'carrot.jpg', 'carrot.txt']
>>> glob.glob('[ab].*')
['a.txt', 'b.jpg']
>>> glob.glob('[bc]*.txt')
['banana.txt', 'carrot.txt']
```

BIOPYTHON

Python's `biopython` module

- The starter code in `calculate_consensus.py` uses the `biopython` module.
- To complete this project, you should read the starter code and aim to understand what that `biopython` code is doing.
- You will need to model part of your solution to `find_mutations.py` on the starter code provided in `calculate_consensus.py`.

Demo

```
filename =
"EBOV_REDC502_MinION_GUI_Conakry_2015-07-13.reads.fasta"

# Open the file containing the input reads
handle = open(filename, "rU")

# Iterate over the input reads and save their sequence in a list
read_sequences = []
for record in SeqIO.parse(handle, "fasta"):
    read_sequences.append(str(record.seq))
```

UPCOMING SEMINARS

Seminar 5: Dr. Michael Brudno

- Tuesday February 27, 2018 6-8pm
- Location: DCS Innovation Lab
- Topic: Rare Disease Data Capture
- http://www.cs.toronto.edu/~brudno/

FEEDBACK