

Week 02 - Jumping into Python
Assigned September 12th, 2016
Due September 19th, 2016

1) Create a random sequence generator.

Requirements:

- Use 'sys' to get command line arguments
- Create a method to generate the sequence
- Produce multi-FASTA output, with a header as described below
 - Sequence count
 - Sequence GC content
 - Length of sequence
- No outside help through packages

Optional:

- Set a random seed

Example Usage:

```
./random-seq-generator.py TOTAL GC READ_LENGTH
```

Example output:

```
>sequence_001 gc=0.30 length=20
ATGCCATTATATGCCATTAT
>sequence_002 gc=0.25 length=20
ATGCCATTATGAAAATTTC
...
```

2) FASTQ Summary Stats

Files For This Exercise

- [Staphylococcus aureus MiSeq FASTQ](#)

Requirements

- Use 'sys' to get command line arguments
- No outside help through packages
- Output summary statistics of the FASTQ
 - Total Reads, Total Base Pairs
 - Min, Max and Mean read length
 - Read length distribution
 - Min, Max, and Mean quality per read

- Mean quality per read

Optional

- Process text FASTQ and gzipped FASTQ

Example usage:

```
./fastq-stats.py FASTQ
```

Example Output:

```
Total Reads: 120000
Total Base Pairs: 12234984bp
Mean Read (Min, Max): 107.33 (33, 210)
Read Lengths:
    33:44
    ...
    210: 20
Mean Quality (Min, Max) Per Read: 36.67 (13, 38)
Mean Quality Per Base:
    1: 36
    2: 37
    ...
    210: 24
```

3) Extract Proteins From GenBank File

Files For This Exercise:

- [Staphylococcus aureus N315 GenBank](#)

Requirements:

- Use 'sys' to get command line arguments
- No outside help through packages
- Output a multi-FASTA file of translated protein sequences
 - Sequence header should include the locus tag and product

Example Usage:

```
./extract-cds.py GENBANK_FILE
```

Example Output:

```
>SA_RS00145 product=chromosomal replication initiator protein DnaA  
MGDAVL DQYVRTYIVLKLKSKPNKLHQMSKKYVSAKSQAQTLEYLMEQEWFTEEM  
DILKRGRNAKSHTKAKNTDVQTYRKSSAIEAVIGFLYLEKREERLEALLNKIITIVNER  
>... product=...  
MGD...TIV
```