# Assignment #3

*The goal of this project is to create a script that analyzes various features of a bacterial genome.*

**You are given the following types files to analyze:**

Sequence file is a FASTA file containing the DNA sequence of a bacterial species. The DNA is organized into 1 or more chromosomes.

Annotation file is a text file containing tab-delimited data for each gene. There should be a header line, containing the following five columns: <GeneName><Chromosome><Strand><Start><Stop>. Each line will contain information for a single gene. Assume the coordinate system is 1 based.

**Your script should take the following arguments:**

*Positional arguments*

Sequence file – required, must be a string

Annotation file – required, must be a string

*Optional arguments*

Codon analysis flag – optional, should not take a value

Gene sequence flag – optional, should take 1 or more gene names to return the sequence

**Your script should do the following things:**

A. Using **argparse**, take in all the above arguments and store them appropriately into a single object.

B. Read in and perform error checking on the sequence and annotation file.

   For the sequence file, use the **pyfaidx** module to read in the data. Verify that:

   1. The file exists
   2. It is proper fasta format
   3. All nucleotides are A,C,G, or T (uppercase or lowercase are allowed)

   For the annotation file, you should use pandas to read in the data. Verify that:

   1. The file exists
   2. It contains five columns
   3. The headers of the columns are named: GeneName, Chromosome, Strand, Start, Stop
   4. None of the genes have the same name
   5. Strand equals '+' or '-'
   6. Start is less than stop
   7. The length of the gene is divisible by 3

   If *any* of these conditions are violated, the program should print an informative statement of all of the violations and quit the program.

C. If **no optional arguments** are given, your script should report: name, length, number of genes, and GC content for each of the chromosomes.

D. If the codon analysis option is used, you should report that calculates the amino acid and codon usage for the entire genome (i.e. how often each amino acid is used within all of the proteins and how often each codon is used for a given amino acid):

   A 5.5% - GCA: 23%; GCC – 37%; GCG – 21%; GCT – 19%

E. If the gene sequence option is used, you should print on the protein sequence for each of the genes that are requested in FASTA format.

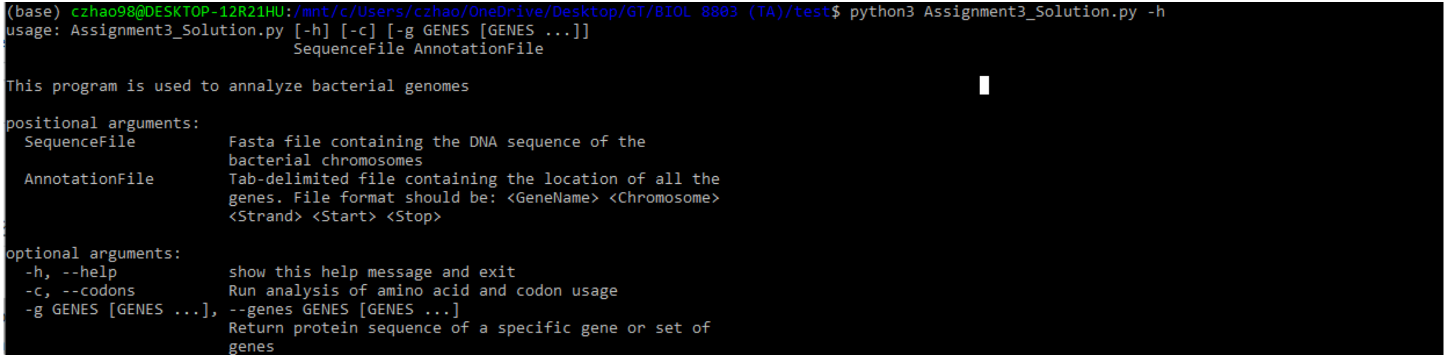
We've provided a template script for you to use. Some example outputs for this script are given below:

# OUTPUTS

## General Usage

### Help documentation:

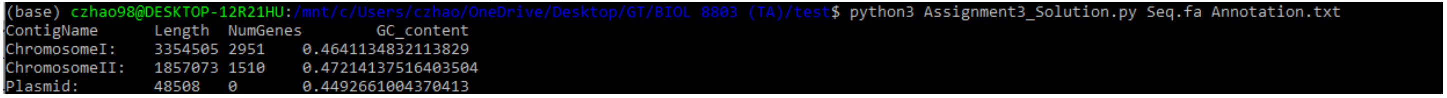
<user>$ python3 Assignment3_Solution.py -h



```
(base) czhao98@DESKTOP-12R21HU:/mnt/c/Users/czhao/OneDrive/Desktop/GT/BIOL 8803 (TA)/test$ python3 Assignment3_Solution.py -h
usage: Assignment3_Solution.py [-h] [-c] [-g GENES [GENES ...]]
                               SequenceFile AnnotationFile

This program is used to annalyze bacterial genomes

positional arguments:
  SequenceFile          Fasta file containing the DNA sequence of the
                        bacterial chromosomes
  AnnotationFile        Tab-delimited file containing the location of all the
                        genes. File format should be: <GeneName> <Chromosome>
                        <Strand> <Start> <Stop>

optional arguments:
  -h, --help            show this help message and exit
  -c, --codons          Run analysis of amino acid and codon usage
  -g GENES [GENES ...], --genes GENES [GENES ...]
                        Return protein sequence of a specific gene or set of
                        genes
```

### Base case:

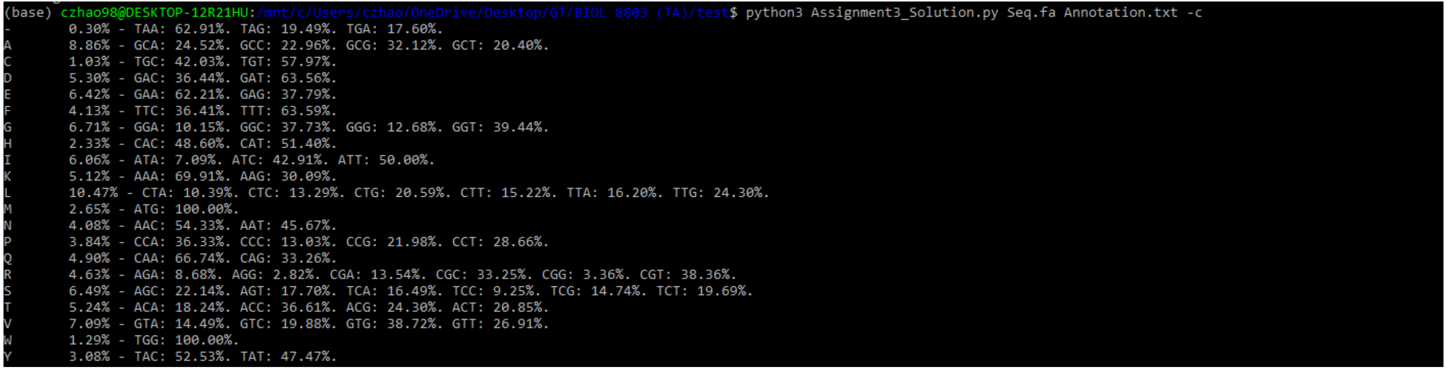
<user>$ python3 Assignment3_Solution.py Seq.fa Annotation.txt



```
(base) czhao98@DESKTOP-12R21HU:/mnt/c/Users/czhao/OneDrive/Desktop/GT/BIOL 8803 (TA)/test$ python3 Assignment3_Solution.py Seq.fa Annotation.txt
ContigName      Length  NumGenes        GC_content
ChromosomeI:    3354505 2951    0.4641134832113829
ChromosomeII:   1857073 1510    0.47214137516403504
Plasmid:        48508   0       0.4492661004370413
```

### Codon flag:

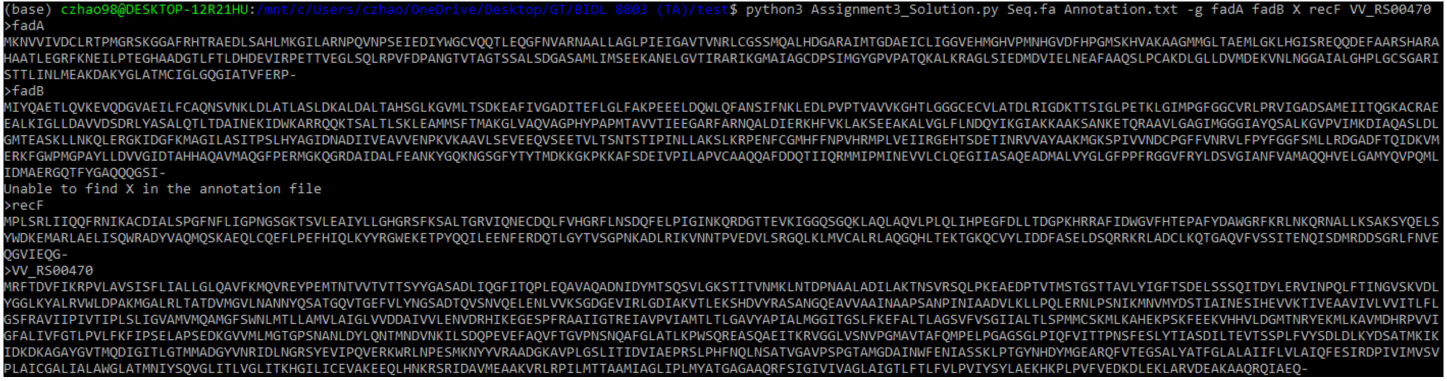
<user>$ python3 Assignment3_Solution.py Seq.fa Annotation.txt -c



```
(base) czhao98@DESKTOP-12R21HU:/mnt/c/Users/czhao/OneDrive/Desktop/GT/BIOL 8803 (TA)/test$ python3 Assignment3_Solution.py Seq.fa Annotation.txt -c
-       0.30% - TAA: 62.91%. TAG: 19.49%. TGA: 17.60%.
A       8.86% - GCA: 24.52%. GCC: 22.96%. GCG: 32.12%. GCT: 20.40%.
C       1.03% - TGC: 42.03%. TGT: 57.97%.
D       5.30% - GAC: 36.44%. GAT: 63.56%.
E       6.42% - GAA: 62.21%. GAG: 37.79%.
F       4.13% - TTC: 36.41%. TTT: 63.59%.
G       6.71% - GGA: 10.15%. GGC: 37.73%. GGG: 12.68%. GGT: 39.44%.
H       2.33% - CAC: 48.60%. CAT: 51.40%.
I       6.06% - ATA: 7.09%. ATC: 42.91%. ATT: 50.00%.
K       5.12% - AAA: 69.91%. AAG: 30.09%.
L       10.47% - CTA: 10.39%. CTC: 13.29%. CTG: 20.59%. CTT: 15.22%. TTA: 16.20%. TTG: 24.30%.
M       2.65% - ATG: 100.00%.
N       4.08% - AAC: 54.33%. AAT: 45.67%.
P       3.84% - CCA: 36.33%. CCC: 13.03%. CCG: 21.98%. CCT: 28.66%.
Q       4.90% - CAA: 66.74%. CAG: 33.26%.
R       4.63% - AGA: 8.68%. AGG: 2.82%. CGA: 13.54%. CGC: 33.25%. CGG: 3.36%. CGT: 38.36%.
S       6.49% - AGC: 22.14%. AGT: 17.70%. TCA: 16.49%. TCC: 9.25%. TCG: 14.74%. TCT: 19.69%.
T       5.24% - ACA: 18.24%. ACC: 36.61%. ACG: 24.30%. ACT: 20.85%.
V       7.09% - GTA: 14.49%. GTC: 19.88%. GTG: 38.72%. GTT: 26.91%.
W       1.29% - TGG: 100.00%.
Y       3.08% - TAC: 52.53%. TAT: 47.47%.
```

### Gene flag:

<user>$ python3 Assignment3_Solution.py Seq.fa Annotation.txt -g fadA fadB X recF VV_RS00470

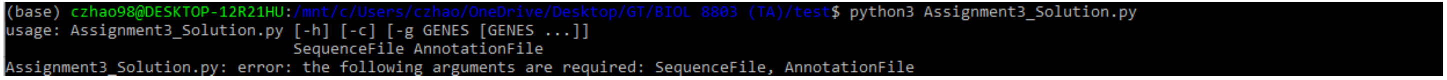


```
(base) czhao98@DESKTOP-12R21HU:/mnt/c/Users/czhao/OneDrive/Desktop/GT/BIOL 8803 (TA)/test$ python3 Assignment3_Solution.py Seq.fa Annotation.txt -g fadA fadB X recF VV_RS00470
>fadA
MKNVVIVDCLRTPMGRSKGGAFRHTRAEDLSAHLMKGILARNPQVNPSEIEDIYWGCVQQTLEQGFNVARNAALLAGLPIEIGAVTVNRLCGSSMQALHDGARAIMTGDAEICLIGGVEHMGHVPMNHGVDFHPGMSKHVAKAAGMMGLTAEMLGKLHGISREQQDEFAARSHARA
HAATLEGRFKNEILPTEGHAADGTLFTLDHDEVIRPETTVEGLSQLRPVFDPANGTVTAGTSSALSDGASAMLIMSEEKANELGVTIRARIKGMAIAGCDPSIMGYGPVPATQKALKRAGLSIEDMDVIELNEAFAAQSLPCAKDLGLLDVMDEKVNLNGGAIALGHPLGCSGARI
STTLINLMEAKDAKYGLATMCIGLGQGIATVFERP-
>fadB
MIYQAETLQVKEVQDGVAEILFCAQNSVNKLDLATLASLDKALDALTAHSGLKGVMLTSDKEAFIVGADITEFLGLFAKPEEELDQWLQFANSIFNKLEDLPVPTVAVVKGHTLGGGCECVLATDLRIGDKTTSIGLPETKLGIMPGFGGCVRLPRVIGADSAMEIITQGKACRAE
EALKIGLLDAVVDSDRLYASALQTLTDAINEKIDWKARRQQKTSALTLSKLEAMMSFTMAKGLVAQVAGPHYPAPMTAVVTIEEGARFARNQALDIERKHFVKLAKSEEAKALVGLFLNDQYIKGIAKKAAKSANKETQRAAVLGAGIMGGGIAYQSALKGVPVIMKDIAQASLDL
GMTEASKLLNKQLERGKIDGFKMAGILASITPSLHYAGIDNADIIVEAVVENPKVKAAVLSEVEEQVSEETVLTSNTSTIPINLLAKSLKRPENFCGMHFFNPVHRMPLVEIIRGEHTSDETINRVVAYAAKMGKSPIVVNDCPGFFVNRVLFPYFGGFSMLLRDGADFTQIDKVM
ERKFGWPMGPAYLLDVVGIDTAHHAQAVMAQGFPERMGKQGRDAIDALFEANKYGQKNGSGFYTYTMDKKGKPKKAFSDEIVPILAPVCAAQQAFDDQTIIQRMMIPMINEVVLCLQEGIIASAQEADMALVYGLGFPPFRGGVFRYLDSVGIANFVAMAQQHVELGAMYQVPQML
IDMAERGQTFYGAQQQGSI-
Unable to find X in the annotation file
>recF
MPLSRLIIQQFRNIKACDIALSPGFNFLIGPNGSGKTSVLEAIYLLGHGRSFKSALTGRVIQNECDQLFVHGRFLNSDQFELPIGINKQRDGTTEVKIGGQSGQKLAQLAQVLPLQLIHPEGFDLLTDGPKHRRAFIDWGVFHTEPAFYDAWGRFKRLNKQRNALLKSAKSYQELS
YWDKEMARLAELISQWRADYVAQMQSKAEQLCQEFLPEFHIQLKYYRGWEKETPYQQILEENFERDQTLGYTVSGPNKADLRIKVNNTPVEDVLSRGQLKLMVCALRLAQGQHLTEKTGKQCVYLIDDFASELDSQRRKRLADCLKQTGAQVFVSSITENQISDMRDDSGRLFNVE
QGVIEQG-
>VV_RS00470
MRFTDVFIKRPVLAVSISFLIALLGLQAVFKMQVREYPEMTNTVVTVTTSYYGASADLIQGFITQPLEQAVAQADNIDYMTSQSVLGKSTITVNMKLNTDPNAALADILAKTNSVRSQLPKEAEDPTVTMSTGSTTAVLYIGFTSDELSSSQITDYLERVINPQLFTINGVSKVDL
YGGLKYALRVWLDPAKMGALRLTATDVMGVLNANNYQSATGQVTGEFVLYNGSADTQVSNVQELENLVVKSGDGEVIRLGDIAKVTLEKSHDVYRASANGQEAVVAAINAAPSANPINIAADVLKLLPQLERNLPSNIKMNVMYDSTIAINESIHEVVKTIVEAAVIVLVVITLFL
GSFRAVIIPIVTIPLSLIGVAMVMQAMGFSWNLMTLLAMVLAIGLVVDDAIVVLENVDRHIKEGESPFRAAIIGTREIAVPVIAMTLTLGAVYAPIALMGGITGSLFKEFALTLAGSVFVSGIIALTLSPMMCSKMLKAHEKPSKFEEKVHHVLDGMTNRYEKMLKAVMDHRPVVI
GFALIVFGTLPVLFKFIPSELAPSEDKGVVMLMGTGPSNANLDYLQNTMNDVNKILSDQPEVEFAQVFTGVPNSNQAFGLATLKPWSQREASQAEITKRVGGLVSNVPGMAVTAFQMPELPGAGSGLPIQFVITTPNSFESLYTIASDILTEVTSSPLFVYSDLDLKYDSATMKIK
IDKDKAGAYGVTMQDIGITLGTMMADGYVNRIDLNGRSYEVIPQVERKWRLNPESMKNYYVRAADGKAVPLGSLITIDVIAEPRSLPHFNQLNSATVGAVPSPGTAMGDAINWFENIASSKLPTGYNHDYMGEARQFVTEGSALYATFGLALAIIFLVLAIQFESIRDPIVIMVSV
PLAICGALIALAWGLATMNIYSQVGLITLVGLITKHGILICEVAKEEQLHNKRSRIDAVMEAAKVRLRPILMTTAAMIAGLIPLMYATGAGAAQRFSIGIVIVAGLAIGTLFTLFVLPVIYSYLAEKHKPLPVFVEDKDLEKLARVDEAKAAQRQIAEQ-
```

## Error Handling

### Missing command inputs:

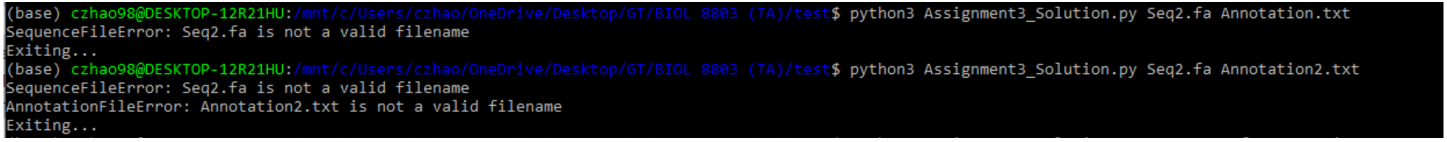
```
<user>$ python3 Assignment3_Solution.py
```



```
(base) czhao98@DESKTOP-12R21HU:/mnt/c/Users/czhao/OneDrive/Desktop/GT/BIOL 8803 (TA)/test$ python3 Assignment3_Solution.py
usage: Assignment3_Solution.py [-h] [-c] [-g GENES [GENES ...]]
                               SequenceFile AnnotationFile
Assignment3_Solution.py: error: the following arguments are required: SequenceFile, AnnotationFile
```

### Missing files:

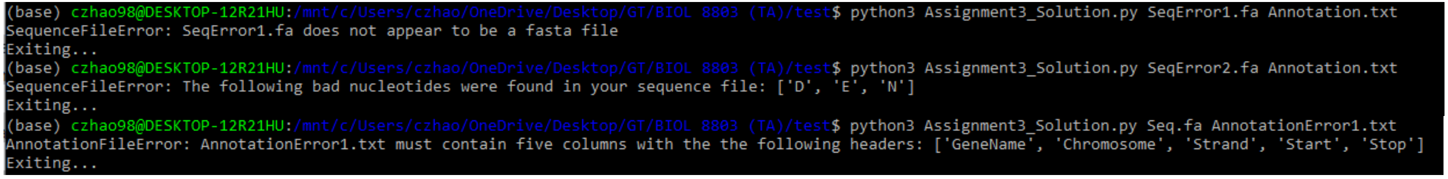
```
<user>$ python3 Assignment3_Solution.py Seq2.fa Annotation.txt
<user>$ python3 Assignment3_Solution.py Seq2.fa Annotation2.txt
```



```
(base) czhao98@DESKTOP-12R21HU:/mnt/c/Users/czhao/OneDrive/Desktop/GT/BIOL 8803 (TA)/test$ python3 Assignment3_Solution.py Seq2.fa Annotation.txt
SequenceFileError: Seq2.fa is not a valid filename
Exiting...
(base) czhao98@DESKTOP-12R21HU:/mnt/c/Users/czhao/OneDrive/Desktop/GT/BIOL 8803 (TA)/test$ python3 Assignment3_Solution.py Seq2.fa Annotation2.txt
SequenceFileError: Seq2.fa is not a valid filename
AnnotationFileError: Annotation2.txt is not a valid filename
Exiting...
```

### Bad input files:

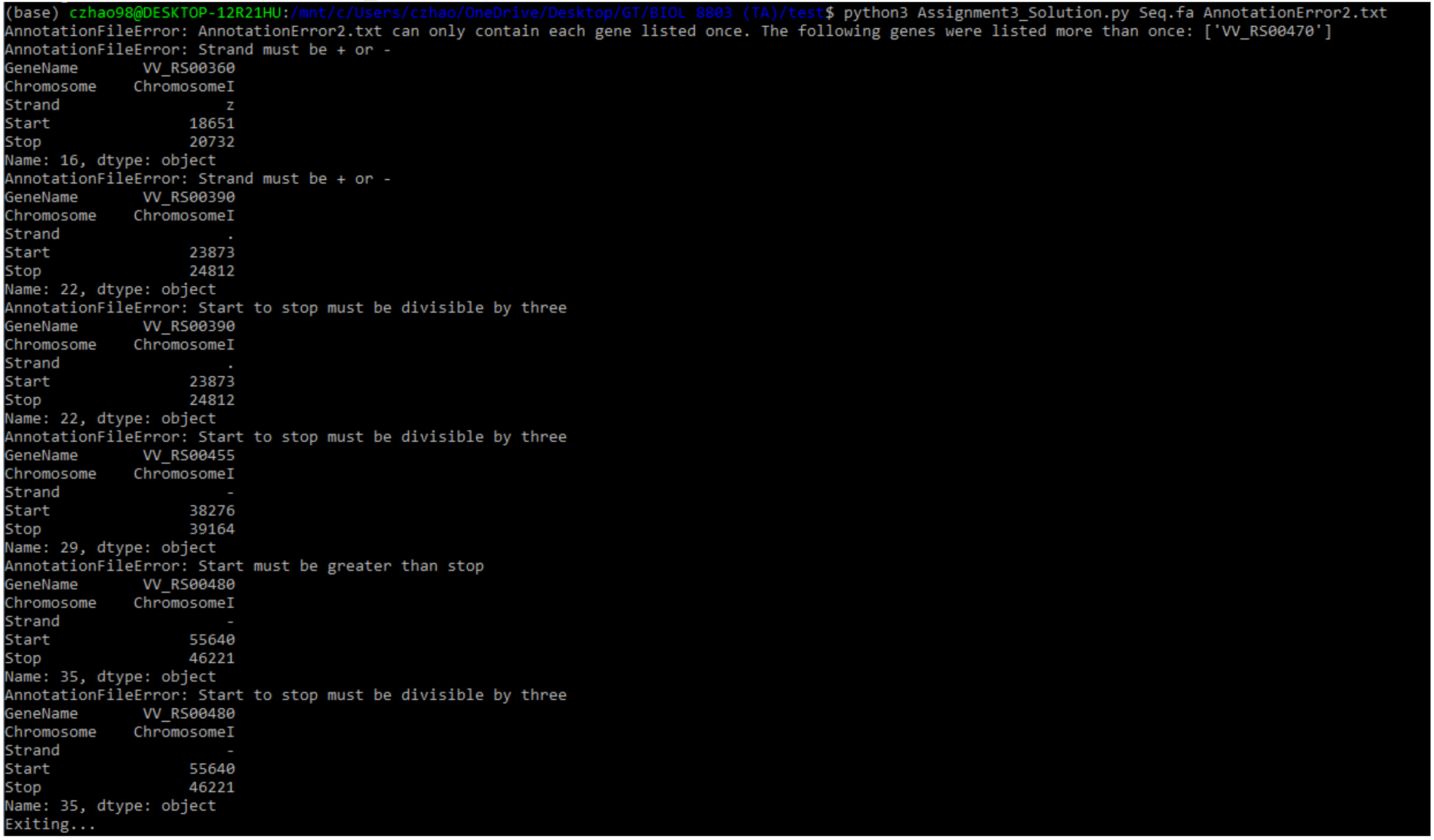
```
<user>$ python3 Assignment3_Solution.py SeqError1.fa Annotation.txt
<user>$ python3 Assignment3_Solution.py SeqError2.fa Annotation.txt
<user>$ python3 Assignment3_Solution.py Seq.fa AnnotationError1.txt
```



```
(base) czhao98@DESKTOP-12R21HU:/mnt/c/Users/czhao/OneDrive/Desktop/GT/BIOL 8803 (TA)/test$ python3 Assignment3_Solution.py SeqError1.fa Annotation.txt
SequenceFileError: SeqError1.fa does not appear to be a fasta file
Exiting...
(base) czhao98@DESKTOP-12R21HU:/mnt/c/Users/czhao/OneDrive/Desktop/GT/BIOL 8803 (TA)/test$ python3 Assignment3_Solution.py SeqError2.fa Annotation.txt
SequenceFileError: The following bad nucleotides were found in your sequence file: ['D', 'E', 'N']
Exiting...
(base) czhao98@DESKTOP-12R21HU:/mnt/c/Users/czhao/OneDrive/Desktop/GT/BIOL 8803 (TA)/test$ python3 Assignment3_Solution.py Seq.fa AnnotationError1.txt
AnnotationFileError: AnnotationError1.txt must contain five columns with the the following headers: ['GeneName', 'Chromosome', 'Strand', 'Start', 'Stop']
Exiting...
```

### Bad input files (duplicate gene):

```
<user>$ python3 Assignment3_Solution.py Seq.fa AnnotationError2.txt
```



```
(base) czhao98@DESKTOP-12R21HU:/mnt/c/Users/czhao/OneDrive/Desktop/GT/BIOL 8803 (TA)/test$ python3 Assignment3_Solution.py Seq.fa AnnotationError2.txt
AnnotationFileError: AnnotationError2.txt can only contain each gene listed once. The following genes were listed more than once: ['VV_RS00470']
AnnotationFileError: Strand must be + or -
GeneName       VV_RS00360
Chromosome    ChromosomeI
Strand                  z
Start               18651
Stop                20732
Name: 16, dtype: object
AnnotationFileError: Strand must be + or -
GeneName       VV_RS00390
Chromosome    ChromosomeI
Strand                  .
Start               23873
Stop                24812
Name: 22, dtype: object
AnnotationFileError: Start to stop must be divisible by three
GeneName       VV_RS00390
Chromosome    ChromosomeI
Strand                  .
Start               23873
Stop                24812
Name: 22, dtype: object
AnnotationFileError: Start to stop must be divisible by three
GeneName       VV_RS00455
Chromosome    ChromosomeI
Strand                  -
Start               38276
Stop                39164
Name: 29, dtype: object
AnnotationFileError: Start must be greater than stop
GeneName       VV_RS00480
Chromosome    ChromosomeI
Strand                  -
Start               55640
Stop                46221
Name: 35, dtype: object
AnnotationFileError: Start to stop must be divisible by three
GeneName       VV_RS00480
Chromosome    ChromosomeI
Strand                  -
Start               55640
Stop                46221
Name: 35, dtype: object
Exiting...
```