Week 02 - Jumping Into Python

Robert Petit
IBS796 - Python
9/12/2016

Bioinformatics and Data Formats

- Many different file formats in Bioinformatics
- Sequences
  - FASTA, FASTQ, FAST5
- Gene Annotations
  - GFF3, GenBank, ASN
- Variants
  - VCF
- Alignments
  - SAM, BAM, FASTA
- Many, many, many more!

A few guidelines for Parsing

- Step 1: Get an example and look through it
- Step 2: Determine the pattern
  - Is it tab delimited? Are there repeated elements?
- Step 3: Code a parser
- Step 4: Debug

This Week

- You will be dealing with:

    1. FASTA
    2. FASTQ
    3. Genbank

FASTA

```
>gi|410687891|ref|NC_019227.1| Bacillus thuringiensis serovar rongseni plasmid pBMB2062-56, complete sequence
ATGCAAGTTTATTTGGATAGGCTAATGATTAAGTATAAAGATGTAACAGAGAAACAATTTAGTGATGTTTTAACTAAAATATCGTCAAAGCAGATTTTTTTACCGAATACA
CCTATTAGGTCAGAACATGGGACGTCTGTTAGAGATTATCATAGAGTTATACATATTGGATATGGTGAAGGTGCAGTTTATATAGGGTGGAAACATAATTCGGAAAAGGAA
AAAGATAGCTATGATATGAAAGTTGATTTTAACCCTTCTAAATTTGAAAATAACGAGTTGCAAAAAGATAGTTATGAAAAAGTGTTTGAAACCGTTTTTCATACGTTAAAT
GCAGTTTTGAAGTCTAATAAGCGAGTGGTTTATGGTATGGATATTGCTTTTGATATAGAGCGTCATATGAGTGATATTGTGTCTTATAGTAAAACGGGAAAGCAACAGGAT
AGACATAAAGGAACTGTTTATTATGGAAATAGAAATAAAGATGGATATTTGAAGATATATGATAAGAAAAAGGAGTTATATAATCATTTTAAAAGAATGATAGAAGAAGAG
AATTTGACTCGTATTGAGTATAGTTGGAGAGACTCTGACGGTGTAGTGGTAGACGAAATAAGGAAGAGTCCTCCGTTTAGTATTGATGAATCTTATACATTCTCGATTTTT
AATTTGAATAATGTTAAAGGGGCATTAAAAGCTTGTTTGATTTGTTATTCTAATGGAACTATGGATATGAAAGAGTTCCCTCGTAGAACTAAAGAGAGTATAAAAAAAGCC
CTTGAAGAAATGGATCACTTGGCGGTGGACCCCATTCTACAGGACTGTTGGTTATCTATATTAGAAAATATTAAGAACTATACTCGTTTATGATATTAGCGTGTGCTTCTC
TGTGTGTCAAGAGGGTGTCAATATGATGCTCTCTTTTTGTTTTTCTTAACTTGTTTATATTAATAGCGGATAGAGTCCCACTTTTACATTGTTCTGGTGTATCTTAGTGTT
GATATTGTGTTTAAACTGATGTTATATTTATGTAGTACGATATACAAGAGGTGATTAGATGAGTGAAATGGTTCGTGTTAATACACGTATCAGTAAAAAGTTAAATGATTG
TTGGACGAGTATAGCAAAGAAAGTGGTGTACCGAAAAGCACTTTAGTTCATTTAGCTTTAGAGAATTATGTGAATCAAAAGGTTATGTTGGAACAAATGCCAAAGATGCAA
CAAATGTTGAGTATGATGTTTGAAAATGTAACGCAGCAACAATTGAATCAAAAAGGGAATATGTTTGAGTTGAAGTAACGGTTATGTTTCGAAAATGTAGTCTTATTGATT
AGGAGATGCACATCGATAAACTAAATGCGTAGTTGGTGTGGCTGAAGTTTGCCCGCCACCTACTCATTTAGAATATCCGTGCATGGGTCCCTGAACAATTAGGAAACGGCT
TTTGTAATCGTATGAGTTGTAATTATTATTCACTTGGGTGAACGTTGTATGACTGAAAGGAAACCCATGGCAGTTAACTGTGTAAATGGGCACTTATTTATTAAGTGTATT
CGACGGAGCCGGAAGTAGACGACTTAAGGGAGCGACATGAATGAAGTGAAGTGCAGACAGAAGGTAAGGTGACGGAGTGGACTCTCAAAGGACACGACGGCGCGTGAACTG
ATGGCAAACGAAATGGAATGAGTGAGAAGAGCGTACGGTCGGCGAGGTAACGGAGGTGTAGGAGCAGATTGATAGAAAGTGAGGGTAACAATTTGAAACTGACAGAAAGAC
AATTGAATGATTTGAAAAGAATTAGCGAATTACGTGTAAAGTTGTTTGGAGTTCCTGGTGAAAGTGTAGTTGATCCAGAGAATGTTGAGTTTTTATTGGATAATGCTATTA
GTTCTTATTTAGGGCAATTAGAAATTTTTGAAGTCACGATAGAGATTGAACAGTATAATTCAATGTGTGGGTAAATTGTAGAAATGTGGCGAAGACATTTTCGGACATTCT
AATAGCCGAAAATCGTGTACAAAATGACATGTTTAATAAAAAATAAGGAGCGGGATAGATTTT
```

FASTA Format

```
>SOME SEQUENCE INFORMATION 001
ATCATTGACTGATGCTGATGCTAGTCGTAGTCAGTACGTTACTGCATG
>SOME SEQUENCE INFORMATION 002
ACTGATCGTACGTAGCTAGCTAGCTAGCTACTGACTGACTGACTTCTT
```

One of the most used and basic data formats

- New entries always start with ‘>’ symbol
- The line following is the sequence
  - Sequence can be split across multiple lines

FASTQ

```
@HWI-700819F:355:HLW5VADXX:1:1101:1889:2189 1:N:0:AGGCAGAACTCTCTAT
GTTCACTAGTTATGAAACCAAGTGAAATTACACCATTAACAACAATACGTGTTTTTGAATTAATGGAAGAAGTTGGTTTCCCTAAAGGAACAATTAATCTTATTCTAGGTGCAGGTTCTGAAGTT
+
BBBFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFFIIIIIIIIIIIIIIIIIIIFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFF
@HWI-700819F:355:HLW5VADXX:1:1101:1849:2242 1:N:0:AGGCAGAACTCTCTAT
TTCTATTCTTGTATCTTTCAAATACATGCTTTGTGTTATGAGTTTCGCGTTTAATTCGTTCTGAATCTTTTTTAATAAAGGTGTTAAACAGAACTTCTCAAATACAAGCGTGTTCTTTTCCAAAT
+
BBBFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIFFIIIIIFIIIIIIIIIIIIIIIIIFFIIIIIIIIIIIIFFFFFFFFFBBFFFFFFFFFFFBFFFFFFBBFFBBBFFBBFFFFFFFFFFBFF
@HWI-700819F:355:HLW5VADXX:1:1101:2098:2105 1:N:0:AGGCAGAACTCTCTAT
ATAATAAGCATTCAATATATCATGCGTATCAAAGTGACTTACTGTTACAACACAATCAGATATACCACGAATAGCTAATATTGCATTTTCAATTTCATCAAGTTCAATACGGTACCCGTTAACTT
+
BBBFFFFFFFFFFIFIIIIIIIIIIIIFFFFIIIIBFIIIIIIIIFIIIIIIIIIIIIFIFFFIIIIIIIIIIIIIFFFFFFFFFFFFFFFFFFFFFFFFFFFBFBFFFFFFFFFBBFFFFBFFFFBF
```

FASTQ Format

```
@HWI-700819F:355:HLW5VADXX:1:1101:1889:2189 1:N:0:AGGCAGAACTCTCTAT
GTTCACTAGTTATGAAACCAAGTGAAATTACACCATTAACAACAATACGTGTTTTTGAATTAATGGAAGAAGTTGGTTT
+
BBBFFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFFFIIIIIIIIIIIIIIIIIIIFFFFFFF
```

Pattern of 4 elements

1. Sequence Header
   a. Always starts with the ‘@’ symbol
2. Sequence
   a. Same length as the quality
3. “Plus” line, separates sequence and quality scores
   a. Always starts with the ‘+’ symbol
4. Quality Scores
   a. Same length as the sequence

FASTQ Gotchas

There are multiple versions, some have/are:

- Different offsets for quality scores
  - Phred+33, Phred+64, Solexa+64
- Broken up into 4 line entries
- Sequences/quality scores split into multiple lines
- ‘@’ and ‘+’ symbols are used in quality scores

GenBank

- A data format to store sequences and annotations.
  - Genes, CDS, RNAs, IS elements, etc...

- NCBI Sample Record Link
  - You might need this for homework!

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
```

Python for this Week

You may (or may not!) use:

- Strings, Lists, Dictionaries
- Built-in Functions
- String Methods
- List Functions (Slicing)
- open()
- print()
- format()
- random()
- len()
- min()
- max()
- startswith()
- rstrip()
- split()
- ord()

Python Methods (Functions)

```
def some_name(a, b, c, ..., z):
    Do something...
    ...
    ...
    ...
    return something
```

Python Method Examples

```
>>> def add(a, b):
...     return a + b
...
>>> def lower_case(string):
...     return string.lower()
...
>>> def power(base, exponent):
...     return base ** exponent
...
>>> add(2, 4)
6
>>> power(2,4)
16
>>> lower_case('ATGC')
'atgc'
```

Opening and Writing Files in Python

```
# File Operations
with open("your_file", 'r') as fh:
    for line in fh:
        # DO SOMETHING WITH LINES

# Write File (replace existing)
with open("your_file.txt", 'w') as fh:
    fh.write("super cool data\n")

# Append To File
with open('your_file.txt", 'a') as fh:
    fh.write("more super cool data\n")
```

For Loops in Python

```
# Loops
for i in my_list:
    print(i)

for i in my_string:
    print(i)

for key, value in my_dict.items():
    print(key, value)
```

Joining Strings in Python

```
# String append in Python
# This is a no no in Python
my_string = "A"
for i in "ATGCATCGC":
    my_string = my_string + i

# Instead
my_list = []
for i in "ATGCATCGC":
    my_list.append(i)
my_string = ''.join(my_list)
```

Home Work

Random Sequence Generator

- Generate multi-FASTA of random sequences

FASTQ Parser With Stats

- Parse FASTQ
- Output read length distribution, per read mean quality, per base mean quality

Extract CDS From GenBank File

- Parse GenBank
- Output FASTA of translated CDS features