Python Libraries

Gzip, CSV

- https://docs.python.org/3/library/gzip.html
- https://docs.python.org/3/library/csv.html
- https://daler.github.io/pybedtools/ - PyBed tools
- http://biopython.org/wiki/GFF_Parsing

BioPython - a library of modules for bioinformatics

BioPython Tutorial

Modules for Sequence data, BLAST parsing, Multiple alignments

Already installed on biocluster

To installed on own computer you control use Python tool 'pip'

```
$ pip3 install biopython
```

Simple BioPython

```
import Bio
from Bio.Seq import Seq
my_seq = Seq("ATGAGTACACTAGGGTAA")

print(my_seq)

rc = my_seq.reverse_complement()
pep = my_seq.translate()
print("revcom is", rc)
print(pep)
```

Parsing sequence files

```
more ../data/E3Q6S8.fasta
>tr|E3Q6S8|E3Q6S8_COLGM RNAse P Rpr2/Rpp21/SNM1 subunit domain-containing protein OS
MAKPKSESLPNRHAYTRVSYLHQAAAYLATVQSPTSDSTTNSSQPGHAPHAVDHERCLET
NETVARRFVSDIRAVSLKAQIRPSPSLKQMMCKYCDSLLVEGKTCSTTVENASKGGKKPW
ADVMVTKCKTCGNVKRFPVSAPRQKRRPFREQKAVEGQDTTPAVSEMSTGAD
```

```python
import sys
#import Bio
from Bio import SeqIO
from Bio.Seq import Seq

# seqfile
filename = sys.argv[1]
for seq_record in SeqIO.parse( filename , "fasta"):
    print(seq_record.id)
    print(repr(seq_record.seq))
    print(seq_record.seq)
    print(len(seq_record))
```

```
tr|E3Q6S8|E3Q6S8_COLGM
Seq('MAKPKSESLPNRHAYTRVSYLHQAAAYLATVQSPTSDSTTNSSQPGHAPHAVDH...GAD',
SingleLetterAlphabet())
MAKPKSESLPNRHAYTRVSYLHQAAAYLATVQSPTSDSTTNSSQPGHAPHAVDHERCLETNETVARRFVSDIRAVSLKAQIRPS
172
```

GenBank files: another sequence format

```
LOCUS       AJ240084                1905 bp    DNA     linear   PRI 03-FEB-2000
DEFINITION  Homo sapiens TRIM gene, promoter.
ACCESSION   AJ240084
VERSION     AJ240084.1  GI:6911579
KEYWORDS    T-cell receptor interacting molecule; TRIM gene.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
      Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
      Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
      Catarrhini; Hominidae; Homo.
REFERENCE   1
AUTHORS   Hubener,C., Mincheva,A., Lichter,P., Schraven,B. and Bruyns,E.
TITLE     Genomic organization and chromosomal localization of the human gene
          encoding the T-cell receptor-interacting molecule (TRIM)
JOURNAL   Immunogenetics 51 (2), 154-158 (2000)
  PUBMED   10663578
REFERENCE   2  (bases 1 to 1905)
  AUTHORS   Huebener,C.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-MAY-1999) Huebener C., Immunomodulation Laboratory,
            Institute for Immunology, University of Heidelberg, Im Neuenheimer
            Feld 305, Heidelberg, 69120, GERMANY
```

GenBank file

```
FEATURES             Location/Qualifiers
     source          1..1905
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /clone_lib="RPCI1,3-5 Human PAC library"
     gene            1..1902
                     /gene="TRIM"
     regulatory      1..1746
                     /regulatory_class="promoter"
                     /gene="TRIM"
     5'UTR           1747..1902
                     /gene="TRIM"
```

GenBank file

```
ORIGIN
   1 ccaaaaattt ccagtcctga aacccttttct ctttccaatg tcctctgtaa gctcgagttg
  61 tgggcatcta ctttgcccat attccaaggt cttgcttagg taacctctgt agtcctttct
 121 tgagcctagg acttctactt ttcttaccag ttaccctctt tcaggaccaa agctcaactc
 181 ctcaaggcca taactaggcc ctctcctctc aaactgattt atcaggtgcc cgaatcttcc
 241 tgaatgtctg ggattcaact tttcagcagt cttcctccct acgttccatc taattctaag
 301 atgaaacctt ctgattcttt gttgtcctct gatccctaca tgaacctgag gctgctgttc
 361 cctgaagtct tgttctgtca gcatccaggc ctgcttcata aaacctgtca ctctgctaat
 421 ggttagcggc tgaacaaaga gtcctctggc caaataagtt tagaaaaact ctgataaaaa
 481 tattatttgg gtttcctttt cgcaggactt acctaagcct ttaatatgca tctacggagg
 541 taaaaataaa gctatatatt ttttccaaag atatttgttg aagaaacatt tgtcttctgc
 601 gtttcttaaa ggccgagtgt tctatggaac atactttaaa aaacccttttt aaagaagctt
 661 agaccagaga atctccaagg tctctttcag ttttacagcc tctgagtcaa cgattcacca
 721 aaaaatattt tgggggggaag tgattgaagt ggaaaaattt gttagtgttt agccagcttt
 781 gtccaaagga taagatgcac tgtattttgc ttactaggga gttattttct ataatggaag
 841 acaaagaaag cacaagacac ccatggtttt gtttgttcaa tcactgagag taagtctcaa
 901 ttattgagac ttacgatgtg ccggtgtgct taattctagt tatgaaattt taataatgaa
 961 taatatagat tctattcctt atatgagttt ccaaaagcat tgtccagaac atctatatta
1021 aaatatctta tcatatacaa tatatgtaat ttaaaatgca ctcagaaaat ctgcttgtta
1081 aaatgcagat tctagtgctt caccctaaat agtctaattt agacgggccc aggattttaa
1141 actagcatct tatagcatac ttatgtacac caacatgtaa gaactgctgc tattaagatt
1201 ctgggatggt ggttgagaac aggagcttgt tgtcaggtgg ctctagattg gacagagaaa
1261 ctcatactga taaggtgagg attgtcagga aataaggcag gcatctagcc tcgcattaag
1321 atgaggtata gaaggcaact gatacatact aagtgctcaa aaaatattaa ctccctgtcc
1381 tccatcatgg ctcaagaaaa tacaacagct gagcacaccc acgggttgct tactatttac
```

Now parse GenBank

```
import sys
#import Bio
from Bio import SeqIO
from Bio.Seq import Seq

# seqfile
filename = sys.argv[1]
for seq_record in SeqIO.parse( filename , "genbank"):
    print(seq_record.id)
    print(repr(seq_record.seq))
    print(seq_record.seq)
    print(len(seq_record))
```

```
python bp_parse_gbk.py ../data/AJ240084_TRIM.gbk
AJ240084.1
Seq('CCAAAAATTTCCAGTCCTGAAACCCTTTCTCTTTCCAATGTCCTCTGTAAGCTC...ATG',
IUPACAmbiguousDNA())
CCAAAAATTTCCAGTCCTGAAACCCTTTCTCTTTCCAATGTCCTCTGTAAGCTCGAGTTGTGGGCATCTACTTTGCCCATATTC
1905
```

Parse features from GenBank file

See documentation on SeqIO here and the tutorial

```
#!/usr/bin/env python3
import sys
import Bio
from Bio import SeqIO
from Bio.Seq import Seq
filename = sys.argv[1]
for seq_record in SeqIO.parse( filename , "genbank"):
print(seq_record.id)
for feature in seq_record.features:
    print("\t",feature.type,feature.location)
    print("\t",feature.type,feature.location.start, feature.location.end, f
```

An even simpler fasta parser to dictionary

```
from Bio import SeqIO
handle = open("example.fasta", "rU")
record_dict = SeqIO.to_dict(SeqIO.parse(handle, "fasta"))
handle.close()
print record_dict["gi:12345678"] #use any record ID
```

Convert file formats

From GenBank to Fasta

```
from Bio import SeqIO

input_handle = open("cor6_6.gb", "rt")
output_handle = open("cor6_6.fasta", "w")

sequences = SeqIO.parse(input_handle, "genbank")
count = SeqIO.write(sequences, output_handle, "fasta")

output_handle.close()
input_handle.close()
```

Or even more simply

```
from Bio import SeqIO
count = SeqIO.convert("cor6_6.gb", "genbank", "cor6_6.fasta", "fasta")
print("Converted %i records" % count)
```

Other BioPython modules

- Pairwise Alignment Parsing (BLAST, FASTA, HMMER)
- Multiple Alignment Parsing
- Database access (local, fast indexed files; Remote databases via Web)
- Some Graphics drawing support

GFF parsing

http://biopython.org/wiki/GFF_Parsing