

Python Packages

Python 3

To use Python 3 on the cluster you can use a module that has it as the default.

```
$ module load miniconda3
```

This will load an environment where python3 is the python installed. You can check that this works by doing `python -V` and see the version is 3.XX

You can setup your own conda environment

This will allow you to install python packages as well as any other pkgs you might want to.

Do this one-time

```
mkdir ~/bigdata/.conda
ln -s ~/bigdata/.conda ~/.conda
conda create -y -n gen220 python=3
source activate gen220
conda config --add channels defaults
conda config --add channels bioconda
conda config --add channels conda-forge
conda install biopython bcbio-gff
```

Python Libraries

Python has many built in packages already installed and a whole host of contributions that span everything from plotting data, interfacing with other datatypes.

- <https://docs.python.org/3/library/gzip.html>
- <https://docs.python.org/3/library/csv.html>
- PyBed tools (processing BED files) - <https://daler.github.io/pybedtools/>
- BioPython GFF parsing
- There is a nice list of available libraries <https://wiki.python.org/moin/UsefulModules>

Gzip

Can open a file that is compressed or write to a file already compressed. This can save space for large files or when you get data from a resource without having to decompress it.

```
import gzip
with gzip.open(file,"rt") as fh:
    for line in fh:
        print("The first line from uncompressed")
        break
```

URL / Web requests directly

```
# this is a URL at the UniProt database to get a protein sequence based on
# accession number
import urllib.request
url="https://www.uniprot.org/uniprot/P10127.fasta"
seqdata = urllib.request.urlopen(url)
for line in seqdata:
    linestrip = line.decode('UTF-8').strip()
    print(linestrip)
```

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'

```
$ conda create -y -n gen220 # only need to do this once
$ source activate gen220
$ conda install biopython # if you forget to do the line above it will fail
```

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 /bigdata/gen220/share/data/E3Q6S8.fasta
>tr|E3Q6S8|E3Q6S8_COLGM RNase P Rpr2/Rpp21/SNM1 subunit domain-containing protein OS=Colletotrichum
MAKPKSESLPNRHAYTRVSYLHQAAYLATVQSPTSDSTTNSSQPGHAPHAVDHERCLET
NETVARRFVSDIRAVSLKAQIRPSPSLKQMMCKYCDSSLVEGKTCSTTVENASKGGKKPW
ADVMVTKCKTCGNVKRFPVSAPRQKRRPFREQKAVEGQDTPAVSEMSTGAD
```

To process this file:

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

# seqfile
filename = "/bigdata/gen220/shared/data/E3Q6S8.fasta"
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))
```

This will output:

```
tr|E3Q6S8|E3Q6S8_COLGM
Seq('MAKPKSESLPNRHAYTRVSYLHQAAYLATVQSPTSDSTTNSSQPGHAPHAVDH...GAD',
SingleLetterAlphabet())
MAKPKSESLPNRHAYTRVSYLHQAAYLATVQSPTSDSTTNSSQPGHAPHAVDHERCLETNETVARRFVSDIRAVSLKAQIRPSPSLKQMMCKYCDSSLVEGKTCSTTVENASKGGKKPW
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

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"

ORIGIN
 1 ccaaaaattt ccagtcctga aaccctttct ctttccaatg tcctctgtaa gctcgagttg
 61 tgggcatcta ctttgcccat attccaaggt cttgcttagg taacctctgt agtcctttct
 121 tgagcctagg acttctactt ttcttaccag ttacctctt tcaggaccaa agctcaactc
 181 ctcaaggcca taactaggcc ctctcctctc aaactgattt atcagggtgcc cgaatcttcc
 241 tgaatgtctg ggattcaact tttcagcagt cttcctccct acgttccatc taattctaag
 301 atgaaacctt ctgattcttt gttgtcctct gatccctaca tgaacctgag gctgctgttc
 361 cctgaagtct tgttctgtca gcatccaggc ctgcttcata aaacctgtca ctctgctaata
 421 ggtagcggc tgaacaaaga gtctctgtgc caataagtt tagaaaaact ctgataaaaa
 481 tattatttgg gtttcctttt cgcaggactt acctaagcct ttaatatgca tctacggagg
 541 taaaaataaa gctatatatt ttttccaaag atatttggtg aagaaacatt tgtcttctgc
 601 gtttcttaaa ggccgagtgt tctatggaac atactttaaa aaacctttt aaagaagctt
 661 agaccagaga atctccaagg tctctttcag ttttacagcc tctgagtcaa cgattcacca
 721 aaaaatattt tggggggaag tgattgaagt ggaaaaattt gttagtgttt agccagcttt
 781 gtccaaagga taagatgcac tgtattttgc ttactaggga gttattttct ataattggaag
 841 acaaagaaag cacaagacac ccatggtttt gtttgttcaa tctactgagag taagtctcaa
 901 ttattgagac ttacgatgtg cgggtgtgct taattctagt tatgaaattt taataatgaa
 961 taatatagat tctattcctt atatgagttt caaaagcat tgtccagaac atctatatta
 1021 aaatatctta tcatatacaa tatatgtaat ttaaaatgca ctcagaaaaat ctgcttgtaa
 1081 aaatgcagat tctagtgtt caccctaaat agtctaattt agacggggcc aggattttaa
 1141 actagcatct tatagcatac ttatgtacac caacatgtaa gaactgctgc tattaagatt
 1201 ctgggatggt gggttgagaac aggagcttgt tgtcaggtgg ctctagattg gacagagaaa
 1261 ctcatctga taagtgagg attgtcagga aataaggcag gcatctagcc tcgcattaag
 1321 atgaggtata gaaggcaact gatacact aagtgtcaa aaaaatttaa ctccctgtcc
 1381 tccatcatgg ctcaagaaaa tacaacagct gagcacaccc acgggttgct tactatttac

```

1441 ttatcagttt agtgtatctt attttgtttc catgtgaatt tacttgtgaa gagatgactg
1501 gattctctcc agagatagga agatccctcc tggtttaatt cctaccttta tttatttatt
1561 tttcaattag actcaggtat tgataaaaat tcaaagtca gattacaaag gtgtgtggga
1621 tttttcttcc cacgttacac aatttaagtc gactgttttc agatcaaaac tcaagacaac
1681 tccttcacca catttcctgt ttgtaactga aacaaagtac acacaaaaga ttttaagaaa
1741 cagaagagaa aagaatccga ggcacagata aagataagtt ttactgtcat gctgctttta
1801 acataacaga gcaacatcac ctaggaaaaa agttttagg aggattttta atccatatat
1861 ttgtcttatg gctagataaa gatttctccg aaaaaaagaa gcatg
//

```

Now parse GenBank

To Parse the genbank file - it is the same code! Just change the format.

```

import sys
import Bio
from Bio import SeqIO
from Bio.Seq import Seq

# seqfile
filename = "/bigdata/gen220/shared/data/AJ240084.gbk"
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))

```

Produces

```

python bp_parse_gbk.py ../data/AJ240084_TRIM.gbk
AJ240084.1
Seq('CCAAAAATTCCAGTCCTGAAACCCCTTTCTCTTTCCAATGTCCTCTGTAAGCTC...ATG',
IUPACAmbiguousDNA())
CCAAAAATTCCAGTCCTGAAACCCCTTTCTCTTTCCAATGTCCTCTGTAAGCTCGAGTTGTGGGCATCTACTTTGCCCATATTCCAAGGTC
1905

```

urlcolor: blue