Semester: 2021-2.
Programming for Biologists, Biology, EAFIT
Microproject I - gbfviewer V0.1
Associated evaluations:
22% *Code Design, execution, completeness, testing strategy, and quality of code*
8% *Presentation and documentation, code in github repository, demo of a solution*

PURPOSE

To get confidence with the activity of programming on the very basics using the Python language. To acquire a minimum knowledge on the usage of imperative languages control structures, like conditional sentences, loops, and processing of text files.

GENBANK FORMAT FILE QUERY TOOL

The genbank data format file tool will be a command line computational program that allows a user to make queries against a folder containing a set of GBFF format files.

TECHNICAL CONSIDERATIONS

- Files do not require to be related. As an example, an specific folder can contain a set of files or archaeas, prokaryotes and eukaryotes, although this concrete practice will be tested with a set of organisms of your preference (you must get at least 50 files in the same folder). The teacher will use gbff files for archaeas available in NCBI, more than 370+ files.
- Gbff files must be in a **data** folder as argument (option) for the command to operate correctly.
- Gbff files should be intact and unmodifiable after download.
- New files can be added to the data folder.
- The language for the development of the command line **gbfviewer** tool is Python.
- Code should be managed in a github repository.
- Recommended development environments are ***Visual Studio Code*** and ***Spider***. You can use any other tool to generate the command line tool.

SPECIFICATIONS.

COMMAND.

- The command line tool name is **biosql.** A user use the command line to write the name of the command, like **$>python3 *gbfviewer.py options***

ID OPTIONS

- **-data**=*"relative or absolute path to data folder"* ***(required)***
- ***-id="name:****organism name* | ***-acc****:accession code* | **protein***:match protein_id"*

QUERY OPTIONS

- One of
  - ***files**, **totals**, **header**, **dnaseq**, **proteinlist**, **proteinseq***

DESCRIPTION OF DATA ORIGIN OPTIONS

The *data options* define the origin of data. For this version, only a folder is possible.
data=*"relative or absolute path to data folder" .*
Is the relative or absolute path to the system directory when the GBFF files are located. This argument is always required in order to work with different directories. **(required)**.
Validation: If no data source is given, an error message is emitted.

DESCRIPTION OF ID OPTIONS

The *id options* are arguments used to identify an organism or set of organisms to be queried. Id options are optional: if no id options are present, all the records in the **data** folder are selected.
For concrete records, use one of 3 possible id options:

- ***id=****name:organism name*
- ***id=****acc:accession code*
- ***id=****"prot:match protein_id . (use with caution!)*

Used for the identification of a unique file or concrete group of files.
VERY IMPORTANT: Be aware some GBFF files can have multiple records! The record separation is done using a line containing only 2 slashes (//).
Validation:

- If no id option is given, the query is applied to all files.
- If a bad id is given, a message is emitted. **(optional)**

DESCRIPTION OF QUERY OPTIONS

In the most simple form of explanation, query options are the information we want to see. The information must be presented for the record or group of records selected using the following options:

files

List the name of the files that match the id section, including the first line of the header.
If you use a unique identifier like a complete accession id, only a filename is shown as output.
Example #1:

```
$> gbffviewer data=./gbff id=acc:NC_014222 files <enter>
GCF_000006175.1_ASM617v2_genomic.gbff:
LOCUS       NC_014222            1936387 bp    DNA     circular CON 16-DEC-2020
$>

```

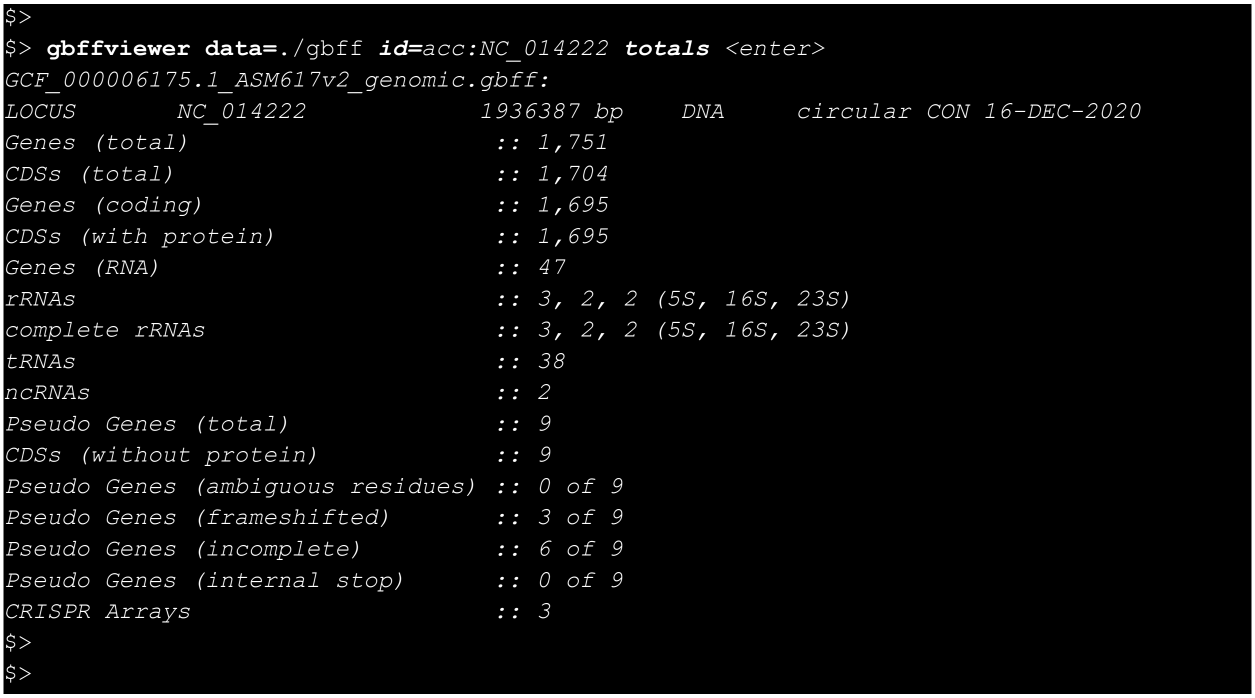
Example #2:

```
$> gbffviewer data=./gbff id=acc:NC_0144 files <enter>
GCF_000145295.1_ASM14529v1_genomic.gbff:
LOCUS       NC_014408            1634695 bp    DNA     circular CON 02-NOV-2020

GCF_000145295.1_ASM14529v1_genomic.gbff:
LOCUS       NC_014409               4440 bp    DNA     circular CON 02-NOV-2020
$>
```

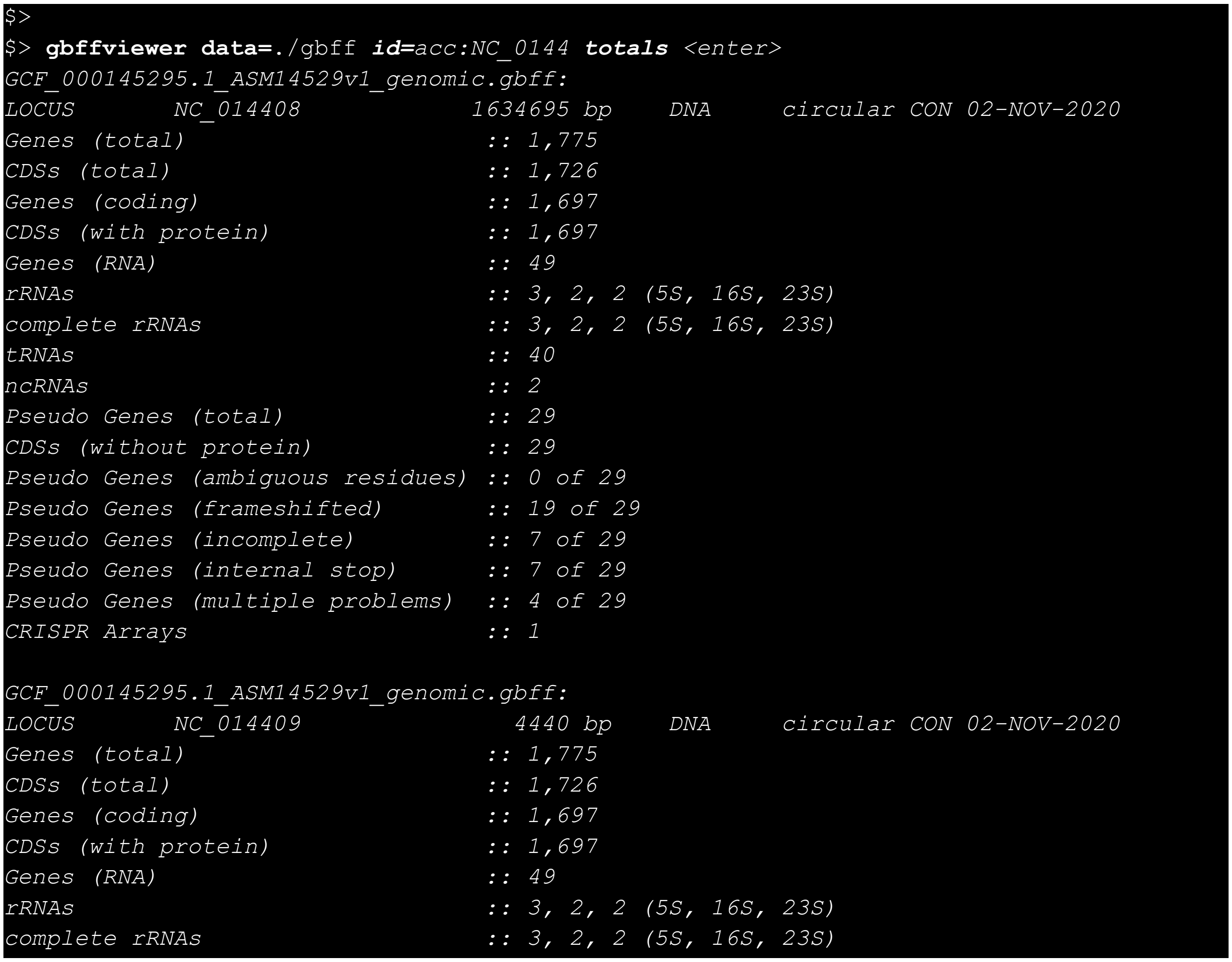
totals

Print the locus header and the numeric information between the line *##Genome-Annotation-Data-START##* and the line *##Genome-Annotation-Data-END##*.
Example #1:



```
$>
$> gbffviewer data=./gbff id=acc:NC_014222 totals <enter>
GCF_000006175.1_ASM617v2_genomic.gbff:
LOCUS       NC_014222            1936387 bp    DNA     circular CON 16-DEC-2020
Genes (total)                     :: 1,751
CDSs (total)                      :: 1,704
Genes (coding)                    :: 1,695
CDSs (with protein)               :: 1,695
Genes (RNA)                       :: 47
rRNAs                             :: 3, 2, 2 (5S, 16S, 23S)
complete rRNAs                    :: 3, 2, 2 (5S, 16S, 23S)
tRNAs                             :: 38
ncRNAs                            :: 2
Pseudo Genes (total)              :: 9
CDSs (without protein)            :: 9
Pseudo Genes (ambiguous residues) :: 0 of 9
Pseudo Genes (frameshifted)       :: 3 of 9
Pseudo Genes (incomplete)         :: 6 of 9
Pseudo Genes (internal stop)      :: 0 of 9
CRISPR Arrays                     :: 3
$>
$>
```

Example #2:



```
$>
$> gbffviewer data=./gbff id=acc:NC_0144 totals <enter>
GCF_000145295.1_ASM14529v1_genomic.gbff:
LOCUS       NC_014408            1634695 bp    DNA     circular CON 02-NOV-2020
Genes (total)                     :: 1,775
CDSs (total)                      :: 1,726
Genes (coding)                    :: 1,697
CDSs (with protein)               :: 1,697
Genes (RNA)                       :: 49
rRNAs                             :: 3, 2, 2 (5S, 16S, 23S)
complete rRNAs                    :: 3, 2, 2 (5S, 16S, 23S)
tRNAs                             :: 40
ncRNAs                            :: 2
Pseudo Genes (total)              :: 29
CDSs (without protein)            :: 29
Pseudo Genes (ambiguous residues) :: 0 of 29
Pseudo Genes (frameshifted)       :: 19 of 29
Pseudo Genes (incomplete)         :: 7 of 29
Pseudo Genes (internal stop)      :: 7 of 29
Pseudo Genes (multiple problems)  :: 4 of 29
CRISPR Arrays                     :: 1

GCF_000145295.1_ASM14529v1_genomic.gbff:
LOCUS       NC_014409               4440 bp    DNA     circular CON 02-NOV-2020
Genes (total)                     :: 1,775
CDSs (total)                      :: 1,726
Genes (coding)                    :: 1,697
CDSs (with protein)               :: 1,697
Genes (RNA)                       :: 49
rRNAs                             :: 3, 2, 2 (5S, 16S, 23S)
complete rRNAs                    :: 3, 2, 2 (5S, 16S, 23S)
```

```
tRNAs                            :: 40
ncRNAs                           :: 2
Pseudo Genes (total)             :: 29
CDSs (without protein)           :: 29
Pseudo Genes (ambiguous residues) :: 0 of 29
Pseudo Genes (frameshifted)      :: 19 of 29
Pseudo Genes (incomplete)        :: 7 of 29
Pseudo Genes (internal stop)     :: 7 of 29
Pseudo Genes (multiple problems)  :: 4 of 29
CRISPR Arrays                    :: 1

$>
```

header

Print the header of the files, corresponding to the fields *LOCUS, DEFINITION, ACCESSION, VERSION, DBLINK, KEYWORDS, SOURCE*, and *ORGANISM*.

Example #1:

```
$>
$> gbffviewer data=./gbff id=acc:NC_014222 header <enter>
GCF_000006175.1_ASM617v2_genomic.gbff:
LOCUS       NC_014222            1936387 bp    DNA     circular CON 16-DEC-2020
DEFINITION  Methanococcus voltae A3, complete sequence.
ACCESSION   NC_014222 NZ_ABHB01000000 NZ_ABHB01000001 NZ_ABHB01000002
            NZ_ABHB01000003 NZ_ABHB01000004 NZ_ABHB01000005
VERSION     NC_014222.1
DBLINK      BioProject: PRJNA224116
            BioSample: SAMN00000040
            Assembly: GCF_000006175.1
KEYWORDS    RefSeq.
SOURCE      Methanococcus voltae A3
  ORGANISM  Methanococcus voltae A3
            Archaea; Euryarchaeota; Methanomada group; Methanococci;
            Methanococcales; Methanococcaceae; Methanococcus.
$>
$>
```

Example #2:

```
$> gbffviewer data=./gbff id=acc:NC_0144 header <enter>
GCF_000145295.1_ASM14529v1_genomic.gbff:
LOCUS       NC_014408            1634695 bp    DNA     circular CON 02-NOV-2020
DEFINITION  Methanothermobacter marburgensis str. Marburg, complete sequence.
ACCESSION   NC_014408
VERSION     NC_014408.1
DBLINK      BioProject: PRJNA224116
            BioSample: SAMN02603260
            Assembly: GCF_000145295.1
KEYWORDS    RefSeq.
SOURCE      Methanothermobacter marburgensis str. Marburg
  ORGANISM  Methanothermobacter marburgensis str. Marburg
            Archaea; Euryarchaeota; Methanomada group; Methanobacteria;
            Methanobacteriales; Methanobacteriaceae; Methanothermobacter.

GCF_000145295.1_ASM14529v1_genomic.gbff:
```

```
LOCUS       NC_014409               4440 bp    DNA     circular CON 02-NOV-2020
DEFINITION  Methanothermobacter marburgensis str. Marburg plasmid pMTBMA4,
            complete sequence.
ACCESSION   NC_014409
VERSION     NC_014409.1
DBLINK      BioProject: PRJNA224116
            BioSample: SAMN02603260
            Assembly: GCF_000145295.1
KEYWORDS    RefSeq.
SOURCE      Methanothermobacter marburgensis str. Marburg
  ORGANISM  Methanothermobacter marburgensis str. Marburg
            Archaea; Euryarchaeota; Methanomada group; Methanobacteria;
            Methanobacteriales; Methanobacteriaceae; Methanothermobacter.
$>
$>
```

proteinseq=XXXXXXXXX

Print the locus header and the formatted protein sequence, corresponding to the translation field of the CDS for a protein.

Example #1:

```
$>
$> gbffviewer data=./gbff id=acc:NC_014222 proteinseq=WP_157209417.1 <enter>
LOCUS       NC_014408            1634695 bp    DNA     circular CON 02-NOV-2020
GCF_000006175.1_ASM617v2_genomic.gbff:
MKTLKSKYKVYKTSKYLTKKDINNIIEKDYDEIIMPQSIYKLLN
EKNKSSMEKLRLCGIIVKTTDNVGRPKKITKYDKDKIKELLVDGKSVRKTAEIMDMKK
TTVWENIKDCMNEIKIEKFRKMIYEYKELLIMQERYGSYVESLFLELDIYINNEDMEN
ALEILNKIIIYVKSEDKKD
$>
$>
```

proteinlist

Print the locus header and the product list of the organism. The protein list contains the protein code and the product. Be aware that for a lot of organisms, this output could be very large.

Example #1:

```
$>
$> gbffviewer data=./gbff id=acc:NC_014222 proteinlist <enter>
GCF_000006175.1_ASM617v2_genomic.gbff:
LOCUS       NC_014408            1634695 bp    DNA     circular CON 02-NOV-2020
WP_013179449.1 tRNA-Glu
WP_013179450.1 tRNA-Arg
WP_013179749.1 DNA-directed RNA polymerase
WP_013179447.1 DNA-directed RNA polymerase subunit E
WP_013179452.1 DUF359 domain-containing protein
WP_013179449.1 30S ribosomal protein S24e
WP_013179450.1 30S ribosomal protein S27ae
WP_013178449.1 hypothetical protein
WP_157209383.1 tyrosine-type recombinase/integrase
...
...
$>
$>
```

dnaseq

Print the locus header and the formatted dna sequence of the organisms, corresponding to the field *CONTIG*. Be aware that for a lot of organisms, this output could be very large. Try to filter properly.
Example #1:

```
$>
$> gbffviewer data=./gbff id=acc:NC_014222 dnaseq <enter>
GCF_000006175.1_ASM617v2_genomic.gbff:
LOCUS       NC_014408            1634695 bp    DNA     circular CON 02-NOV-2020
        1 gaaggagttg ataagatgag tggcaaccac aaccacccac cccacagaac cacccctata
       61 gaggtgatgt gcaatccacc cacaaaggag gtgaactaca atgaatagta tgtcaaagaa
      121 ggaatataaa cattttgagc gatcaataac cctacacatt aaacacgctg attttgaatt
      181 agtaggccag aaatacaggg gcagagacct ctacgagtac ttgttcatca aggggagtaa
      241 accaatccac acagccacag gtaaaaccag caacctctac gagatcatat cagaggataa
      301 aggccctgat gaagccctca aaatcatagg ggacacattc acagaggatg atataaacat
      361 cctcctcagg ggagggttcc atgatgataa caagacccct gagggtgtcc tcgagttaat
      421 ccagcacatc ctcctcgcag gtgaggtgct ccacccagga ggggatgtta ttgaacccca
      481 atcattcaag gactaccctg agaagataca agcctacgct gaccagttaa tcaatgatga
      541 cagcatcgac atcctcgaat ccatcaccag agtcataggg gaggcccatt atggtgatga
      601 gaaggccgtg aaattactcc tattatccat cggcacccta ttcctcaggg acaccccacc
      661 agtccaccag gcactgaggg gctccacagg ttcaggtaaa acagaccttg tattgaagac
      721 agtactcgca gtcccagaga ggtatgtgca catcctcaga tcagcatcac ctaaatactt
      781 attctacgca tcagagactg gcatcctccg tgaagactac aatatcttcg tatttgatga
      841 tattgagttg aatgatgaga tcatagcgat ctccaagacc atcacagata acatcctccc
      901 agagaaggaa caccacaccg tgaaggatca ggaggccctg aaactggaga tcccaggtga
      961 gggcctggcg atattcacaa gggcgaggga catccatgat aatgaactca atgac


...
...
$>
$>
```