Biopython
Work with biological sequence data in Python

Alexander McFarland

NUIT Research Computing Services

Northwestern University

July 19-23, 2021

Biopython *Work with biological sequence data in Python* is brought to you by NUIT Research Computing Services.

Have a programming or data question about your research? **We're here to help. bit.ly/rcsconsult**

Running the workshop code

The contents of this workshop are available at

https://github.com/agmcfarland/biopython_workshop

Please use look over the contents of the README in github

Two ways to run the code:

1. Open the main github directory in Google Colab and select the notebook (.ipynb) you wish to run **(easy)**
2. Clone the repository, go to the repository directory, open the notebook you wish to run.
   - This assumes you have a way of running a jupyter notebook already set (VScode, JupyterLab (anaconda)

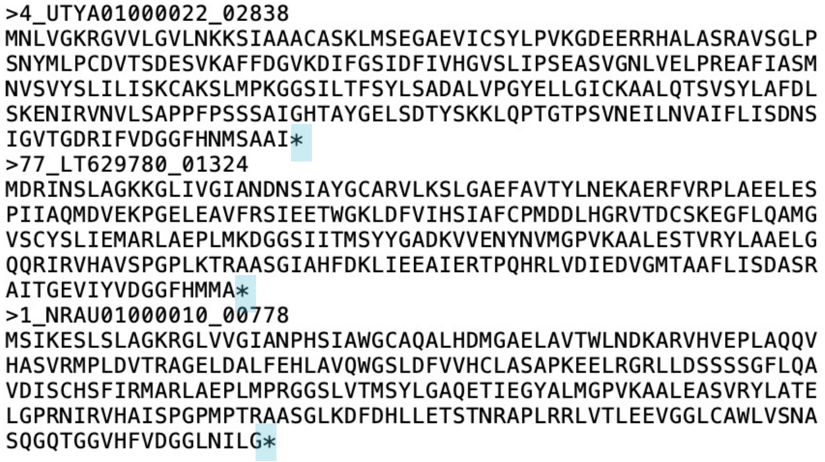
Biopython is designed to work with biological sequences and their formats

- Biological sequences are **DNA, RNA, and amino acids**
- When working with more than a handful of sequence, **manual editing becomes inefficient and time-consuming**
  - Manual work can become error prone and irreproducible
- **Biopython is a Python package** that is built to interact with all types of sequences found in formats used by biologists, including
  - Fasta (.fasta)
  - GenBank file-format (.gbk)
- Biopython can help to make analyses **efficient and reproducible**

By the end of Day 3 you will know how to do all these things!

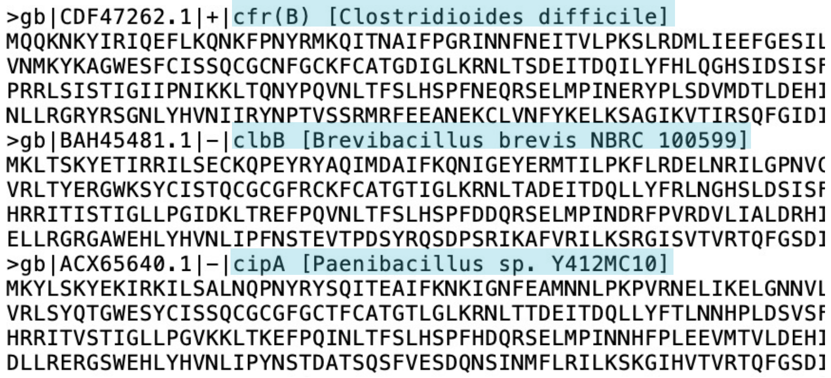
Biopython can help overcome many everyday sequence challenges

Remove the * at the end of each sequence and calculate GC content



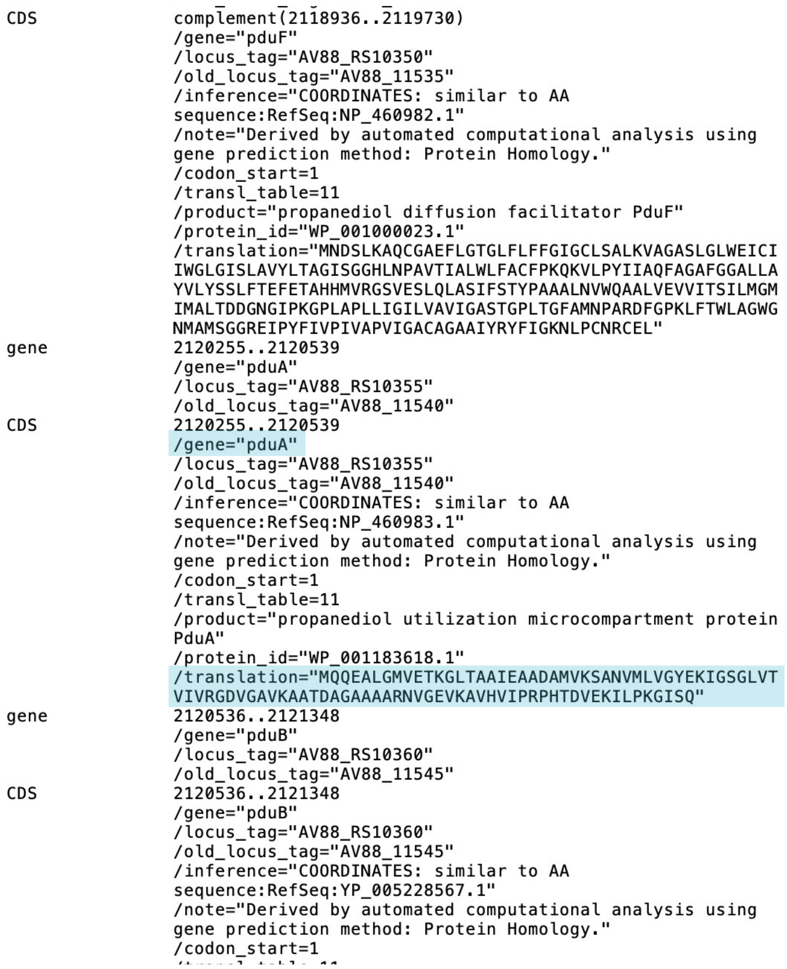
```
>4_UTYA01000022_02838
MNLVGKRGVVLGVLNKKSIAAACASKLMSEGAEVICSYLPVKGDEERRHALASRAVSGLP
SNYMLPCDVTSDESVKAFFDGVKDIFGSIDFIVHGVSLIPSEASVGNLVELPREAFIASM
NVSVYSLILISKCAKSLMPKGGSILTFSYLSADALVPGYELLGICKAALQTSVSYLAFDL
SKENIRVNVLSAPPFPSSSAIGHTAYGELSDTYSKKLQPTGTPSVNEILNVAIFLISDNS
IGVTGDRIFVDGGFHNMSAAI*
>77_LT629780_01324
MDRINSLAGKKGLIVGIANDNSIAYGCARVLKSLGAEFAVTYLNEKAERFVRPLAEELES
PIIAQMDVEKPGELEAVFRSIEETWGKLDFVIHSIAFCPMDDLHGRVTDCSKEGFLQAMG
VSCYSLIEMARLAEPLMKDGGSIITMSYYGADKVVENYNVMGPVKAALESTVRYLAAELG
QQRIRVHAVSPGPLKTRAASGIAHFDKLIEEAIERTPQHRLVDIEDVGMTAAFLISDASR
AITGEVIYVDGGFHMMA*
>1_NRAU01000010_00778
MSIKESLSLAGKRGLVVGIANPHSIAWGCAQALHDMGAELAVTWLNDKARVHVEPLAQQV
HASVRMPLDVTRAGELDALFEHLAVQWGSLDFVVHCLASAPKEELRGRLLDSSSSGFLQA
VDISCHSFIRMARLAEPLMPRGGSLVTMSYLGAQETIEGYALMGPVKAALEASVRYLATE
LGPRNIRVHAISPGPMPTRAASGLKDFDHLLETSTNRAPLRRLVTLEEVGGLCAWLVSNA
SQGQTGGVHFVDGGLNILG*
```

Rename headers to gene name and species, count the number of times a genus is found



```
>gb|CDF47262.1|+|cfr(B) [Clostridioides difficile]
MQQKNKYIRIQEFLKQNKFPNYRMKQITNAIFPGRINNFNEITVLPKSLRDMLIEEFGESIL
VNMKYKAGWESFCISSQCGCNFGCKFCATGDIGLKRNLTSDEITDQILYFHLQGHSIDSISF
PRRLSISTIGIIPNIKKLTQNYPQVNLTFSLHSPFNEQRSELMPINERYPLSDVMDTLDEHI
NLLRGRYRSGNLYHVNIIRYNPTVSSRMRFEEANEKCLVNFYKELKSAGIKVTIRSQFGIDI
>gb|BAH45481.1|-|clbB [Brevibacillus brevis NBRC 100599]
MKLTSKYETIRRILSECKQPEYRYAQIMDAIFKQNIGEYERMTILPKFLRDELNRILGPNVC
VRLTYERGWKSYCISTQCGCGFRCKFCATGTIGLKRNLTADEITDQLLYFRLNGHSLDSISF
HRRITISTIGLLPGIDKLTREFPQVNLTFSLHSPFDDQRSELMPINDRFPVRDVLIALDRHI
ELLRGRGAWEHLYHVNLIPFNSTEVTPDSYRQSDPSRIKAFVRILKSRGISVTVRTQFGSDI
>gb|ACX65640.1|-|cipA [Paenibacillus sp. Y412MC10]
MKYLSKYEKIRKILSALNQPNYRYSQITEAIFKNKIGNFEAMNNLPKPVRNELIKELGNNVL
VRLSYQTGWESYCISSQCGCGFGCTFCATGTLGLKRNLTTDEITDQLLYFTLNNHPLDSVSF
HRRITVSTIGLLPGVKKLTKEFPQINLTFSLHSPFHDQRSELMPINNHFPLEEVMTVLDEHI
DLLRERGSWEHLYHVNLIPYNSTDATSQSFVESDQNSINMFLRILKSKGIHVTVRTQFGSDI
```

Extract specific gene names and associated sequences



```
CDS             complement(2118936..2119730)
                /gene="pduF"
                /locus_tag="AV88_RS10350"
                /old_locus_tag="AV88_11535"
                /inference="COORDINATES: similar to AA
                sequence:RefSeq:NP_460982.1"
                /note="Derived by automated computational analysis using
                gene prediction method: Protein Homology."
                /codon_start=1
                /transl_table=11
                /product="propanediol diffusion facilitator PduF"
                /protein_id="WP_001000023.1"
                /translation="MNDSLKAQCGAEFLGTGLFLFFGIGCLSALKVAGASLGLWEICI
                IWGLGISLAVYLTAGISGGHLNPAVTIALWLFACFPKQKVLPYIIAQFAGAFGGALLA
                YVLYSSLFTEFETAHHMVRGSVESLQLASIFSTYPAAALNVWQAALVEVVITSILMGM
                IMALTDDGNGIPKGPLAPLLIGILVAVIGASTGPLTGFAMNPARDFGPKLFTWLAGWG
                NMAMSGGREIPYFIVPIVAPVIGACAGAAIYRYFIGKNLPCNRCEL"
gene            2120255..2120539
                /gene="pduA"
                /locus_tag="AV88_RS10355"
                /old_locus_tag="AV88_11540"
CDS             2120255..2120539
                /gene="pduA"
                /locus_tag="AV88_RS10355"
                /old_locus_tag="AV88_11540"
                /inference="COORDINATES: similar to AA
                sequence:RefSeq:NP_460983.1"
                /note="Derived by automated computational analysis using
                gene prediction method: Protein Homology."
                /codon_start=1
                /transl_table=11
                /product="propanediol utilization microcompartment protein
                PduA"
                /protein_id="WP_001183618.1"
                /translation="MQQEALGMVETKGLTAAIEAADAMVKSANVMLVGYEKIGSGLVT
                VIVRGDVGAVKAATDAGAAAARNVGEVKAVHVIPRPHTDVEKILPKGISQ"
gene            2120536..2121348
                /gene="pduB"
                /locus_tag="AV88_RS10360"
                /old_locus_tag="AV88_11545"
CDS             2120536..2121348
                /gene="pduB"
                /locus_tag="AV88_RS10360"
                /old_locus_tag="AV88_11545"
                /inference="COORDINATES: similar to AA
                sequence:RefSeq:YP_005228567.1"
                /note="Derived by automated computational analysis using
                gene prediction method: Protein Homology."
                /codon_start=1
                /transl_table=11
```

Biopython applications covered by this workshop

1. **Working with sequences**
   - Remove certain characters (*)
   - Remove first or last 'X' number of nucleotides/amino acids
   - Transcribe/translate DNA sequences
   - Calculate GC content
   - Calculate length of sequences
   - Find motifs
2. **Modifying headers**
   - Make headers more readable
   - Extract information from headers
3. **Extracting and storing sequences**
   - Write only certain sequences to new file
   - Write modified sequences/headers to new file
4. **BLAST-ing against the NCBI database**
   - Set homology thresholds and store results

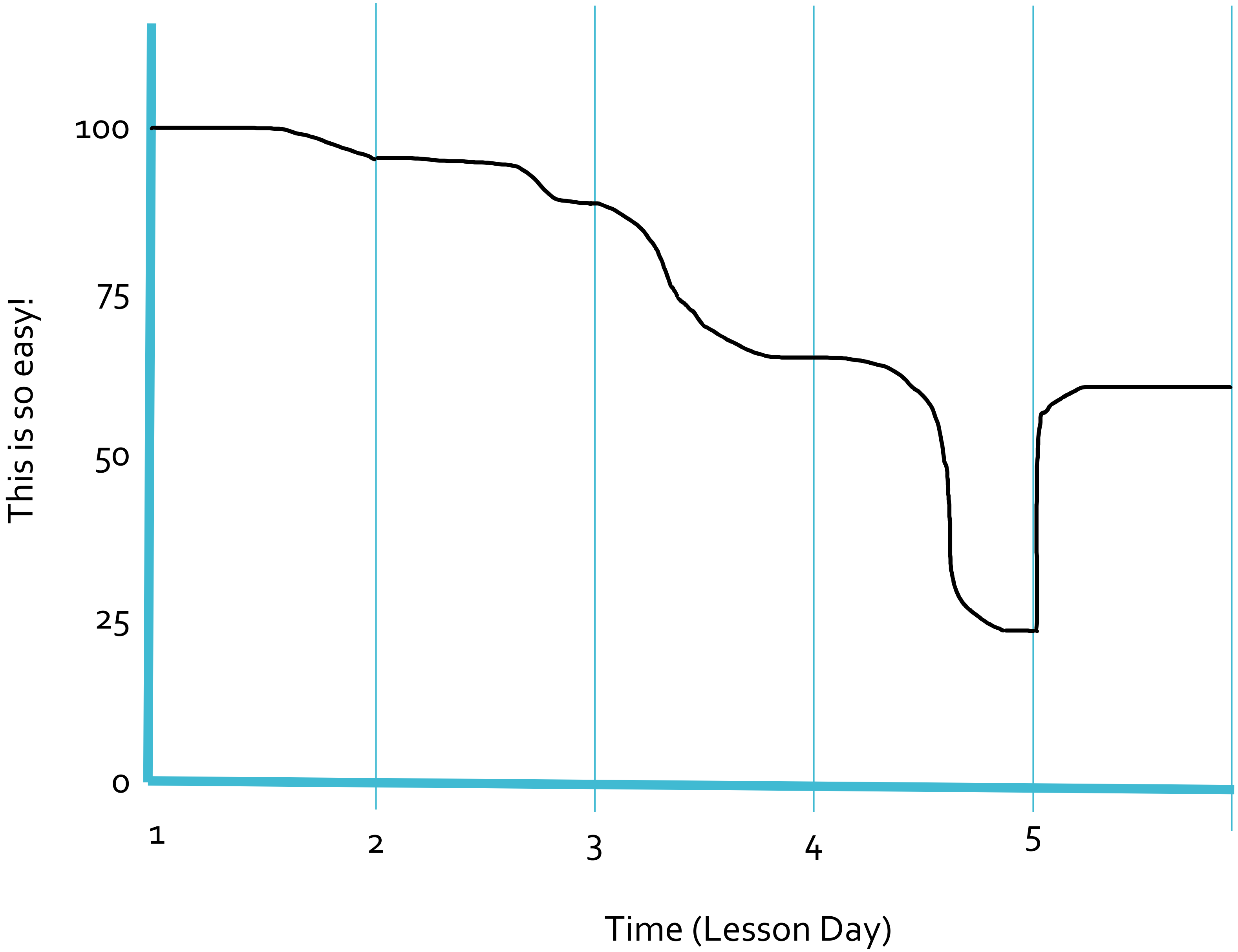
Workshop overview

- **Five days of one-hour sessions**
  - **4-5 pm July 19-23**
- All code is run on a Jupyter notebook
- **When necessary, Python basics will be reviewed** prior to their use with Biopython
- Focus on writing code that can be adapted to workflows
- **Feel free to ask questions/for clarifications in the chat – I will read over them during examples and find the best time to address them!**

Session outline

1. **Monday –** Introduction to strings, Biopython, and Biopython sequences
2. **Tuesday –** Opening, closing, and saving sequence files with Biopython
3. **Wednesday –** More sequence modification and data extraction
4. **Thursday –** Extracting and storing sequence data, working with GenBank files
5. **Friday –** BLAST-ing against the NCBI database

Expected ‘easiness’


This is so easy!
Time (Lesson Day)