BIOL550 - Assignment #2 (5 questions, 45 points)

Always read through assignment/project/exam instructions carefully; they are provided for a reason, and a lot of time can be saved by taking an extra second to read through them!

In this assignment, you will be provided with five (5) different biology-related files. For each file, you will be tasked with designing a regular expression that returns desired lines. Please feel free to utilize Regexer (https://regexr.com/) to aid in the design of your regular expressions. A 'desired output' file containing the correctly extracted lines is provided for each input file.

To check the validity and correctness of your regular expression, you can perform the following command lines:

```
## This extracts the lines that match your regular expression
grep -P 'regular expression' Q#_input.txt > Q#_output.txt
```

followed by:

```
## This checks if the files have any differences
diff -s Q#_output.txt Q#_desired_output.txt
```

***What to do*:**

1. Download the A2_F23_files.tar.gz from Blackboard to your local computer.
2. Upload A2_F23_files.tar.gz from your local computer to your ~/Assignment/Assignment_2 folder on the class server
3. Untar the A2_F23_files.tar.gz file archive and work from your account on the class server.
4. Write your final regular expression for each question in a single text file named assignment2_[lastname]_[firstname].txt. **Do not** use a word processor to save your regular expression, use a code editor instead, to prevent corrupting your regular expressions!
5. Submit the text file containing your regular expression to Blackboard.

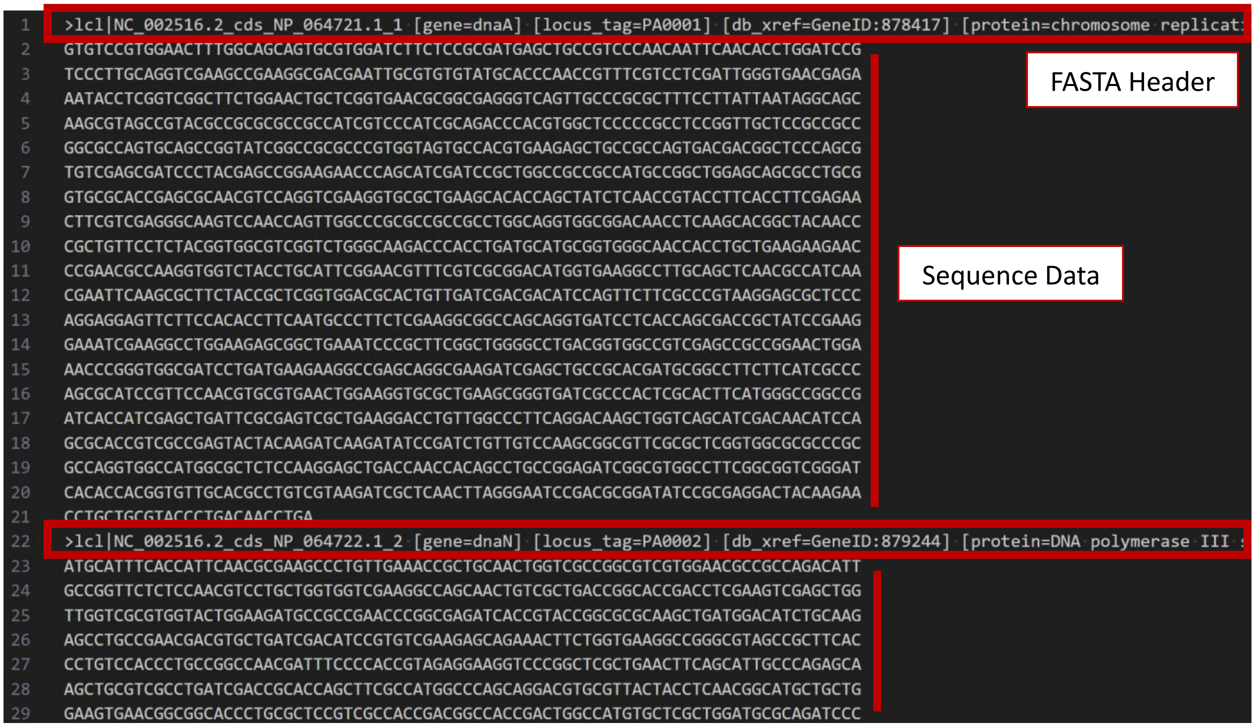
```
## Example of format
# Question 1
^\d+\w{3}\s+$
# Question 2
\d\d\d\w+\s*
```

Please reach out for any clarification needed.

Question 1 (5 points)

FASTA (for FAST-All), has become one of the standards in which sequence information, either nucleotide or amino-acid, is stored. The format is made up of two components: the header and the sequence. The header component is a single line that starts with a '>' and is followed by a string of characters containing the sequence ID and potentially some other metadata. The sequence component has variable number of lines (it depends on the length of the sequence and text-wrap formatting used).

1. Create a regular expression that returns only the **FASTA header** lines of the file Q1_fasta.fna.

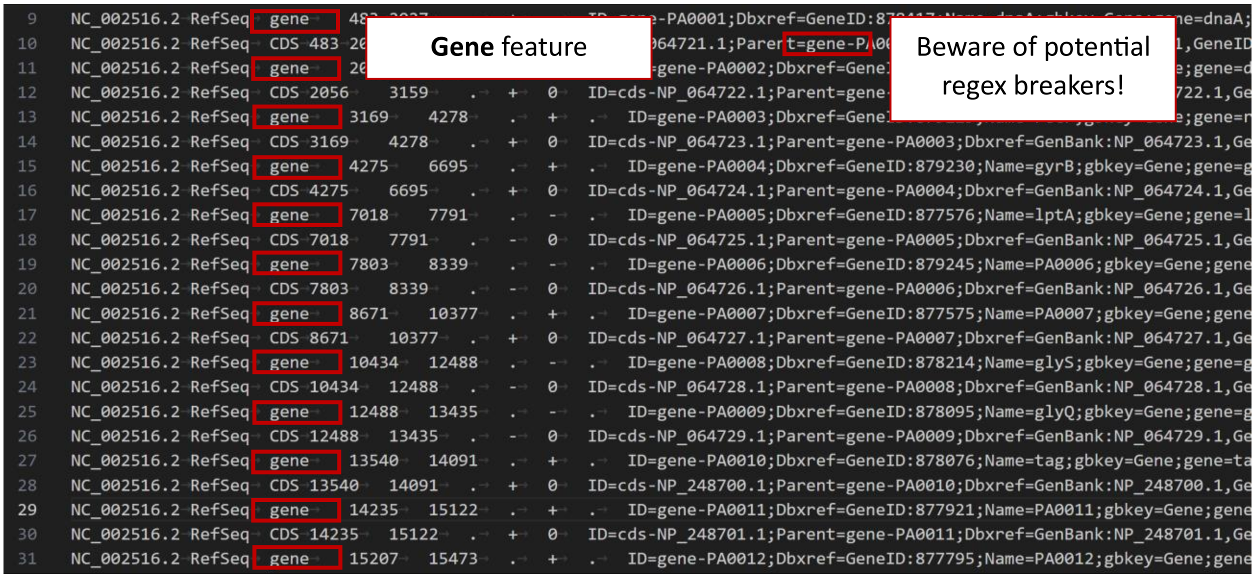


```
>lcl|NC_002516.2_cds_NP_064721.1_1 [gene=dnaA] [locus_tag=PA0001] [db_xref=GeneID:878417] [protein=chromosome replicat
GTGTCCGTGGAACTTTGGCAGCAGTGCGTGGATCTTCTCCGCGATGAGCTGCCGTCCCAACAATTCAACACCTGGATCCG
TCCCTTGCAGGTCGAAGCCGAAGGCGACGAATTGCGTGTGTATGCACCCAACCGTTTCGTCCTCGATTGGGTGAACGAGA
AATACCTCGGTCGGCTTCTGGAACTGCTCGGTGAACGCGGCGAGGGTCAGTTGCCCGCGCTTTCCTTATTAATAGGCAGC
AAGCGTAGCCGTACGCCGCGCGCCGCCATCGTCCCATCGCAGACCCACGTGGCTCCCCCGCCTCCGGTTGCTCCGCCGCC
GGCGCCAGTGCAGCCGGTATCGGCCGCGCCCGTGGTAGTGCCACGTGAAGAGCTGCCGCCAGTGACGACGGCTCCCAGCG
TGTCGAGCGATCCCTACGAGCCGGAAGAACCCAGCATCGATCCGCTGGCCGCCGCCATGCCGGCTGGAGCAGCGCCTGCG
GTGCGCACCGAGCGCAACGTCCAGGTCGAAGGTGCGCTGAAGCACACCAGCTATCTCAACCGTACCTTCACCTTCGAGAA
CTTCGTCGAGGGCAAGTCCAACCAGTTGGCCCGCGCCGCCGCCTGGCAGGTGGCGGACAACCTCAAGCACGGCTACAACC
CGCTGTTCCTCTACGGTGGCGTCGGTCTGGGCAAGACCCACCTGATGCATGCGGTGGGCAACCACCTGCTGAAGAAGAAC
CCGAACGCCAAGGTGGTCTACCTGCATTCGGAACGTTTCGTCGCGGACATGGTGAAGGCCTTGCAGCTCAACGCCATCAA
CGAATTCAAGCGCTTCTACCGCTCGGTGGACGCACTGTTGATCGACGACATCCAGTTCTTCGCCCGTAAGGAGCGCTCCC
AGGAGGAGTTCTTCCACACCTTCAATGCCCTTCTCGAAGGCGGCCAGCAGGTGATCCTCACCAGCGACCGCTATCCGAAG
GAAATCGAAGGCCTGGAAGAGCGGCTGAAATCCCGCTTCGGCTGGGGCCTGACGGTGGCCGTCGAGCCGCCGGAACTGGA
AACCCGGGTGGCGATCCTGATGAAGAAGGCCGAGCAGGCGAAGATCGAGCTGCCGCACGATGCGGCCTTCTTCATCGCCC
AGCGCATCCGTTCCAACGTGCGTGAACTGGAAGGTGCGCTGAAGCGGGTGATCGCCCACTCGCACTTCATGGGCCGGCCG
ATCACCATCGAGCTGATTCGCGAGTCGCTGAAGGACCTGTTGGCCCTTCAGGACAAGCTGGTCAGCATCGACAACATCCA
GCGCACCGTCGCCGAGTACTACAAGATCAAGATATCCGATCTGTTGTCCAAGCGGCGTTCGCGCTCGGTGGCGCGCCCGC
GCCAGGTGGCCATGGCGCTCTCCAAGGAGCTGACCAACCACAGCCTGCCGGAGATCGGCGTGGCCTTCGGCGGTCGGGAT
CACACCACGGTGTTGCACGCCTGTCGTAAGATCGCTCAACTTAGGGAATCCGACGCGGATATCCGCGAGGACTACAAGAA
CCTGCTGCGTACCCTGACAACCTGA
>lcl|NC_002516.2_cds_NP_064722.1_2 [gene=dnaN] [locus_tag=PA0002] [db_xref=GeneID:879244] [protein=DNA polymerase III
ATGCATTTCACCATTCAACGCGAAGCCCTGTTGAAACCGCTGCAACTGGTCGCCGGCGTCGTGGAACGCCGCCAGACATT
GCCGGTTCTCTCCAACGTCCTGCTGGTGGTCGAAGGCCAGCAACTGTCGCTGACCGGCACCGACCTCGAAGTCGAGCTGG
TTGGTCGCGTGGTACTGGAAGATGCCGCCGAACCCGGCGAGATCACCGTACCGGCGCGCAAGCTGATGGACATCTGCAAG
AGCCTGCCGAACGACGTGCTGATCGACATCCGTGTCGAAGAGCAGAAACTTCTGGTGAAGGCCGGGCGTAGCCGCTTCAC
CCTGTCCACCCTGCCGGCCAACGATTTCCCCACCGTAGAGGAAGGTCCCGGCTCGCTGAACTTCAGCATTGCCCAGAGCA
AGCTGCGTCGCCTGATCGACCGCACCAGCTTCGCCATGGCCCAGCAGGACGTGCGTTACTACCTCAACGGCATGCTGCTG
GAAGTGAACGGCGGCACCCTGCGCTCCGTCGCCACCGACGGCCACCGACTGGCCATGTGCTCGCTGGATGCGCAGATCCC
```

Question 2 (5 points)

Genome annotations can be stored in various file formats; for this question, the annotations are in gene feature format (GFF). The GFF format stores gene location, source database, feature type, as well as several additional pieces of information needed to completely define genomic annotations, with one feature entry per line.

1. Create a regular expression that returns only lines with the **gene feature-type** from Q2_annotations.gff.



```
NC_002516.2 RefSeq  gene    48                                -PA0001;Dbxref=GeneID:87                            e=dnaA;
NC_002516.2 RefSeq  CDS 483 20                               064721.1;Parent=gene-PA00                            1,GeneID
NC_002516.2 RefSeq  gene    20                               gene-PA0002;Dbxref=GeneI                             e;gene=d
NC_002516.2 RefSeq  CDS 2056    3159    .   +   0   ID=cds-NP_064722.1;Parent=gene-                           722.1,Ge
NC_002516.2 RefSeq  gene    3169    4278    .   +   .   ID=gene-PA0003;Dbxref=GeneI                           e;gene=r
NC_002516.2 RefSeq  CDS 3169    4278    .   +   0   ID=cds-NP_064723.1;Parent=gene-PA0003;Dbxref=GenBank:NP_064723.1,Ge
NC_002516.2 RefSeq  gene    4275    6695    .   +   .   ID=gene-PA0004;Dbxref=GeneID:879230;Name=gyrB;gbkey=Gene;gene=g
NC_002516.2 RefSeq  CDS 4275    6695    .   +   0   ID=cds-NP_064724.1;Parent=gene-PA0004;Dbxref=GenBank:NP_064724.1,Ge
NC_002516.2 RefSeq  gene    7018    7791    .   -   .   ID=gene-PA0005;Dbxref=GeneID:877576;Name=lptA;gbkey=Gene;gene=l
NC_002516.2 RefSeq  CDS 7018    7791    .   -   0   ID=cds-NP_064725.1;Parent=gene-PA0005;Dbxref=GenBank:NP_064725.1,Ge
NC_002516.2 RefSeq  gene    7803    8339    .   -   .   ID=gene-PA0006;Dbxref=GeneID:879245;Name=PA0006;gbkey=Gene;gene
NC_002516.2 RefSeq  CDS 7803    8339    .   -   0   ID=cds-NP_064726.1;Parent=gene-PA0006;Dbxref=GenBank:NP_064726.1,Ge
NC_002516.2 RefSeq  gene    8671    10377   .   +   .   ID=gene-PA0007;Dbxref=GeneID:877575;Name=PA0007;gbkey=Gene;gene
NC_002516.2 RefSeq  CDS 8671    10377   .   +   0   ID=cds-NP_064727.1;Parent=gene-PA0007;Dbxref=GenBank:NP_064727.1,Ge
NC_002516.2 RefSeq  gene    10434   12488   .   -   .   ID=gene-PA0008;Dbxref=GeneID:878214;Name=glyS;gbkey=Gene;gene=g
NC_002516.2 RefSeq  CDS 10434   12488   .   -   0   ID=cds-NP_064728.1;Parent=gene-PA0008;Dbxref=GenBank:NP_064728.1,Ge
NC_002516.2 RefSeq  gene    12488   13435   .   -   .   ID=gene-PA0009;Dbxref=GeneID:878095;Name=glyQ;gbkey=Gene;gene=g
NC_002516.2 RefSeq  CDS 12488   13435   .   -   0   ID=cds-NP_064729.1;Parent=gene-PA0009;Dbxref=GenBank:NP_064729.1,Ge
NC_002516.2 RefSeq  gene    13540   14091   .   +   .   ID=gene-PA0010;Dbxref=GeneID:878076;Name=tag;gbkey=Gene;gene=ta
NC_002516.2 RefSeq  CDS 13540   14091   .   +   0   ID=cds-NP_248700.1;Parent=gene-PA0010;Dbxref=GenBank:NP_248700.1,Ge
NC_002516.2 RefSeq  gene    14235   15122   .   +   .   ID=gene-PA0011;Dbxref=GeneID:877921;Name=PA0011;gbkey=Gene;gene
NC_002516.2 RefSeq  CDS 14235   15122   .   +   0   ID=cds-NP_248701.1;Parent=gene-PA0011;Dbxref=GenBank:NP_248701.1,Ge
NC_002516.2 RefSeq  gene    15207   15473   .   +   .   ID=gene-PA0012;Dbxref=GeneID:877795;Name=PA0012;gbkey=Gene;gene
```

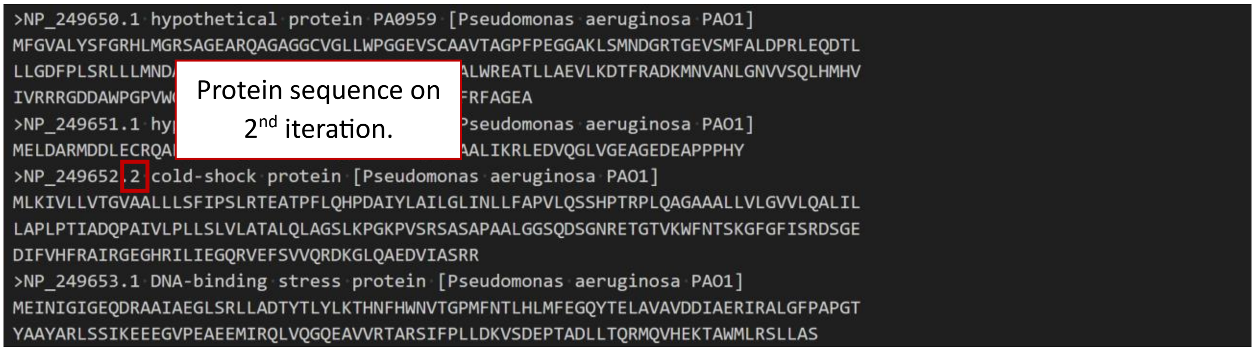
Gene feature

Beware of potential regex breakers!

Question 3 (10 points)

Sometimes newer evidence comes forward that may result in the need to update the sequences present in FASTA files. When these changes are made, the version number of the protein sequence is modified to delineate past analysis from future analysis.

1. Create a regular expression the grabs FASTA headers for proteins whose sequence is not on its first iteration (**iteration ≥ 2**) from Q3_protein.faa.



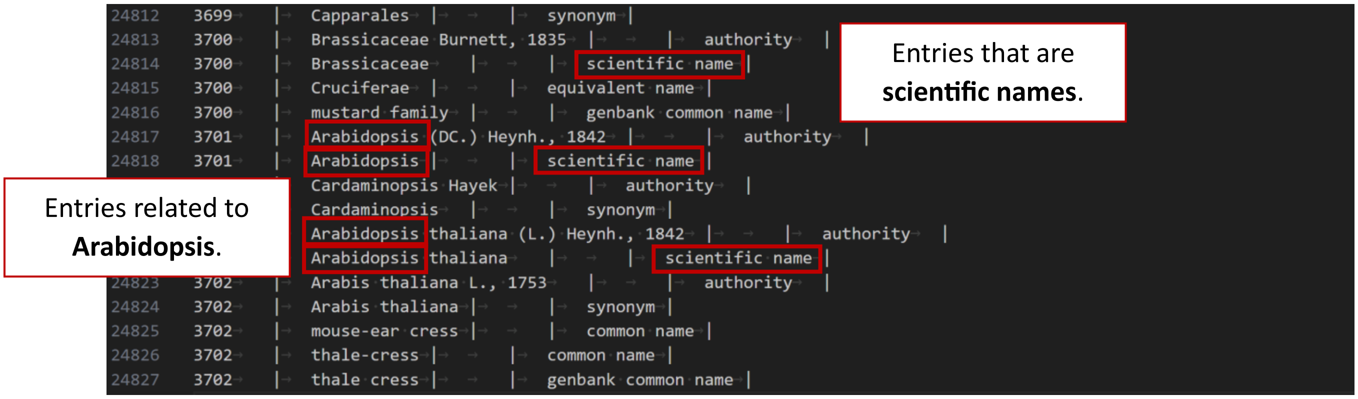
```
>NP_249650.1 hypothetical protein PA0959 [Pseudomonas aeruginosa PAO1]
MFGVALYSFGRHLMGRSAGEARQAGAGGCVGLLWPGGEVSCAAVTAGPFPEGGAKLSMNDGRTGEVSMFALDPRLEQDTL
LLGDFPLSRLLLMND                            ALWREATLLAEVLKDTFRADKMNVANLGNVVSQLHMHV
IVRRRGDDAWPGPVW                            FRFAGEA
>NP_249651.1 hy                             Pseudomonas aeruginosa PAO1]
MELDARMDDLECRQA                            AALIKRLEDVQGLVGEAGEDEAPPPHY
>NP_249652.2 cold-shock protein [Pseudomonas aeruginosa PAO1]
MLKIVLLVTGVAALLLSFIPSLRTEATPFLQHPDAIYLAILGLINLLFAPVLQSSHPTRPLQAGAAALLVLGVVLQALIL
LAPLPTIADQPAIVLPLLSLVLATALQLAGSLKPGKPVSRSASAPAALGGSQDSGNRETGTVKWFNTSKGFGFISRDSGE
DIFVHFRAIRGEGHRILIEGQRVEFSVVQRDKGLQAEDVIASRR
>NP_249653.1 DNA-binding stress protein [Pseudomonas aeruginosa PAO1]
MEINIGIGEQDRAAIAEGLSRLLADTYTLYLKTHNFHWNVTGPMFNTLHLMFEGQYTELAVAVDDIAERIRALGFPAPGT
YAAYARLSSIKEEEGVPEAEEMIRQLVQGQEAVVRTARSIFPLLDKVSDEPTADLLTQRMQVHEKTAWMLRSLLAS
```

Protein sequence on 2nd iteration.

Question 4 (**10** **points)**

A dump file (.dmp), named because they are typically created by "dumping" a bunch of data into a file, can be used to store a variety of information. It's important to note that just because the data is "dumped" does not mean there isn't a defined structure to the file! In the NCBI names.dmp, the data stored is used to connect the taxonomic ID of an organism to its more descriptive name.

1. Create a regular expression that returns all lines that have a **scientific name** related to **Arabidopsis** in Q4_names.dmp.



```
3699    |   Capparales  |       |   synonym |
3700    |   Brassicaceae Burnett, 1835  |       |   authority   |
3700    |   Brassicaceae    |       |   scientific name |
3700    |   Cruciferae  |       |   equivalent name |
3700    |   mustard family  |       |   genbank common name |
3701    |   Arabidopsis (DC.) Heynh., 1842  |       |   authority   |
3701    |   Arabidopsis |       |   scientific name |
            Cardaminopsis Hayek |       |   authority   |
            Cardaminopsis   |       |   synonym |
            Arabidopsis thaliana (L.) Heynh., 1842  |       |   authority   |
            Arabidopsis thaliana    |       |   scientific name |
3702    |   Arabis thaliana L., 1753    |       |   authority   |
3702    |   Arabis thaliana |       |   synonym |
3702    |   mouse-ear cress |       |   common name |
3702    |   thale-cress |       |   common name |
3702    |   thale cress |       |   genbank common name |
```

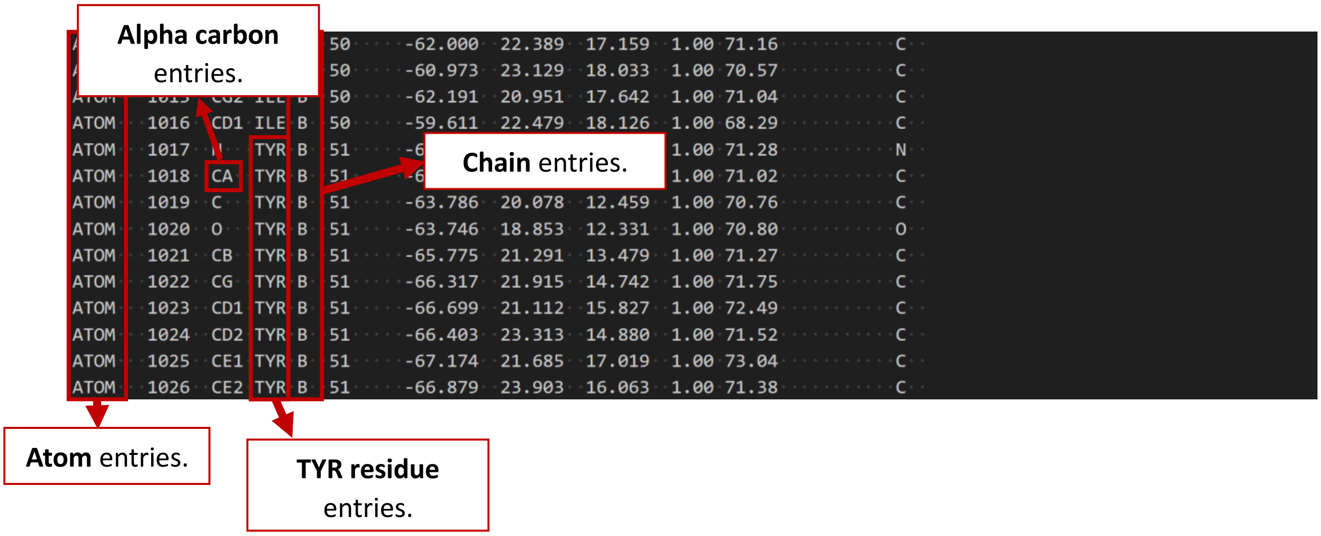
Entries that are **scientific names**.

Entries related to **Arabidopsis**.

Question 5 (15 points)

The PDB format (.pdb) contains information related to the orientation of biological molecules in 3D space, most commonly proteins (the P in PDB stands for proteins!). In its barebones form, the PDB file contains 3D coordinates for almost each atom in the structure. Additionally, PDB files will almost always contain a metadata block at the top of the file, describing what the molecules in the structure are, how they were obtained, and the individuals who helped obtain it. When it comes to text parsing, the PDB format is one of the most interesting, as the file does not delimit data by the traditional method (tab, comma, semicolon), rather each column is assigned to a particular piece of data depending on what kind of line is being read (For more details: https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/tutorials/pdbintro.html)!

1. Grab all lines that denote a protein (**ATOM**) alpha carbon (**CA**) in a tyrosine residue (**TYR**) for chains **B**, **D**, or **E** in Q5_structure.pdb.



```
A                           50     -62.000  22.389  17.159  1.00 71.16           C
A                           50     -60.973  23.129  18.033  1.00 70.57           C
                            50     -62.191  20.951  17.642  1.00 71.04           C
ATOM   1016  CD1 ILE B  50     -59.611  22.479  18.126  1.00 68.29           C
ATOM   1017       TYR B  51     -6                      1.00 71.28           N
ATOM   1018  CA  TYR B  51     -6                      1.00 71.02           C
ATOM   1019  C   TYR B  51     -63.786  20.078  12.459  1.00 70.76           C
ATOM   1020  O   TYR B  51     -63.746  18.853  12.331  1.00 70.80           O
ATOM   1021  CB  TYR B  51     -65.775  21.291  13.479  1.00 71.27           C
ATOM   1022  CG  TYR B  51     -66.317  21.915  14.742  1.00 71.75           C
ATOM   1023  CD1 TYR B  51     -66.699  21.112  15.827  1.00 72.49           C
ATOM   1024  CD2 TYR B  51     -66.403  23.313  14.880  1.00 71.52           C
ATOM   1025  CE1 TYR B  51     -67.174  21.685  17.019  1.00 73.04           C
ATOM   1026  CE2 TYR B  51     -66.879  23.903  16.063  1.00 71.38           C
```