Homework 6

Working with protein data in R

Translate DNA, annotate your protein with GO terms, predict protein structure or pull protein structure from available databases.

1. Install the following R packages:   
   UniprotR   
   protti

You may also need to install GenomicAlignments from BiocManager in order to get UniprotR to load correctly:  
BiocManager::install("GenomicAlignments")

If possible, install this package as well. It’s for visualizing protein structures, but we have other ways of doing that if it won’t install:   
r3dmol

More info on UniprotR:

<https://github.com/Proteomicslab57357/UniprotR>

1. Note: you may see some stuff about B-factors as you work through this exercise. These are the ‘atomic displacement factor’, which is related to the energy required to fold a protein, and can be used to identify the flexibility of atoms, side chains, or even whole regions of a protein.
2. Choose one of your sequences from the last homework, that you translated into an amino acid sequence. You will likely need to write the amino acid sequence to a .fasta file, using this function in Biostrings: writeXStringSet()

Search for it in UniProt using the protein BLAST: <https://www.uniprot.org/blast>

Choose the top 5 matches and download the UniProt accession numbers to a file in your GitHub folder.

If for some reason you get no matches, use the following two accession numbers to move forward in the exercise:

"P0A799" "P08839"

1. Read this file into R using the appropriate function. Likely one of these:

read.csv()  
read.txt()

1. Format your list of accession numbers into a character string
2. Get the Gene Ontology (GO) terms using this function: GetProteinGOInfo()  
   and save them to a variable
3. Plot your results using PlotGoInfo()
4. There is a ‘handy visualization for publications’ code line in the UniprotR GitHub page. Use this to save a plot of your GO terms to your GitHub repository
5. What are some interesting GO terms for your gene?

**The GO terms for the amoA gene are membrane and monooxygenase activity**

**GO terms for P0A799 (Phosphoglycerate kinase) are ATP binding, Phosphoglycerate kinase activity, and gluconeogenessis**

1. Use GetPathology\_Biotech() and Get.diseases() to find information on any diseases or pathologies associated with your gene. List an example for each below:

N.A

1. Now we are going to access structural information using the protti package. Use fetch\_uniprot() to get information on the UniProt Accessions for your samples. That function will give you a dataframe of information on these accessions. One column, called ‘xref\_pdb’ is a list of the Protein DataBase (pdb) IDs for these samples, if they exist.
2. Use fetch\_pdb() to pull any available structural information from the Protein DataBase. If none are available from your data, use these two: "1ZMR", "2HWG"

More information is here:

<https://cran.r-project.org/web/packages/protti/vignettes/protein_structure_workflow.html>

1. Use fetch\_alphafold\_prediction() to get information on any available 3D structures for your gene. Here you can either try to use the r3dmol package to visualize the 3D structure of the protein, or go to the Alpha Fold website and type in the Alpha Fold accession number, which will bring up the 3D structure:

<https://alphafold.ebi.ac.uk/>

Save a photo of the 3D structure to your GitHub folder

1. Push your data to GitHub!