BIO310 Introduction to Bioinformatics
Computer Lab 1 and Homework 1 Spring 2020

February 11, 2020

"The trouble with the world was," she continued hesitatingly, "that people were still superstitious instead of scientific. He said if everybody would study science more, there wouldn't be all the trouble there was."
"He said science was going to discover the basic secret of life some day," the bartender put in. He scratched his head and frowned. "Didn't I read in the paper the other day where they'd finally found out what it was?"
"I missed that," I murmured.
"I saw that," said Sandra. "About two days ago."
"That's right," said the bartender.
"What is the secret of life?" I asked.
"I forget," said Sandra.
"Protein," the bartender declared. "They found out something about protein."
"Yeah," said Sandra, "that's it."[3]

Instructions:

- We expect you to start working on this assignment in the lab, at the end of the lab you will submit how far you came along. Your overall effort will be graded and it will eventually contribute to your lab grade. This grade will be assigned as a number from 1-5; no effort being 1 and full effort being 5. You are not expected to finish the entire assignment during the lab; you will have a chance to submit the final version till the due date and this will be your homework 1 grade, which is out of 100.
- For the homework submission, submit a PDF document for the answers of the write-up questions, the plots should be appropriately labeled, figures should have captions and should be appropriately cited within the main text. Name your submission as `BIO310-HW1-YourName.pdf` where you substitute in your first and last names into the filename in place of 'YourName' and submit online through SUCourse as a single file. Upload your final report on SuCourse by the due date.
- Upload the code online on SuCourse by the due date. You may code in any programming language you may prefer, but your assistant will only provide help in Python. The code you submit should be in a format that is ready to run. In submitting the code on SuCourse, compress it as a ZIP file with the name `BIO310-HWXcode-YourName.zip` where you substitute in your first and last names into the file name in place of 'YourName' and X with the current homework number.
- If you are considering to submit the homework late, please see the late submission policy in the syllabus.
- Please follow the submission instructions, not adhering the submission standards will lead to point deduction.

Introduction

We have discussed in class that proteins are main actors in the cell; assemblies of proteins make up machines that are involved in nearly all processes that are carried out in cells. It is estimated that the human body has 19,000-20,000 different proteins. Look here to get a brief idea about the types of functions proteins can perform.

Proteins are able to perform these different functions because they have specific shapes. The shapes of proteins arise as a result of the sequence of amino acids that make up proteins. There are exactly 20 amino acids that are used to make up proteins. These amino acids have different chemical properties.

In this lab, we will focus on a single protein: rhodopsin. Rhodopsin belongs to a family of proteins known as the G-Protein Coupled Receptors (GPCRs). The GPCR superfamily has about 900 members [2]. In other words, roughly 4.5 percent of the human coding genome codes for GPCRs. So far, 8 Nobel Prizes have been awarded in the field of signal transduction by G-Proteins, proteins which are associated with GPCRs.

While studying rhodopsin, we will explore a database which is the primary resource that provides information about proteins. You will also implement a program to analyze the amino acid composition of the rhodopsin protein sequence.

Part 1 - UniProt: A Protein Database [15 pts.]

The Universal Protein Resource (UniProt) is a database that provides information on protein sequence and function. UniProt comprises 4 different components:

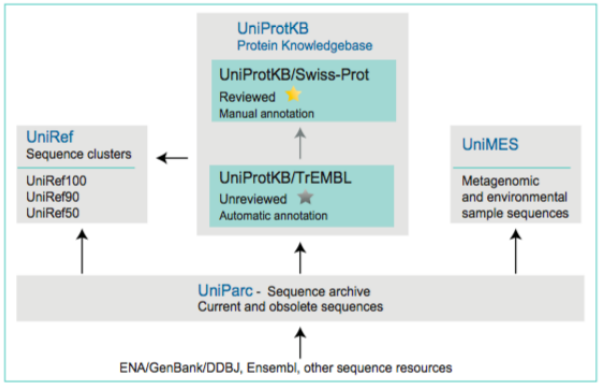

Figure 1: Sources and flow of data for UniProt's component databases

In this lab, we will focus on the UniProt Knowledge Base (UniProtKB). UniProtKB is further subdivided into TrEMBL and SwissProt. The difference between these two is that the former contains unreviewed, automatic annotation while the latter is manually curated. Figures 1 and 2 show some statistics about UniProtKB/TrEMBL. See this for more statistics about UniProtKB/TrEMBL.

You can access UniProt at www.uniprot.org. Search for "rhodopsin" (Note that the search is performed in UniProtKB by default). Play around with the website and answer the following questions:

1. How many results do you find?
2. How many reviewed and unreviewed entries?

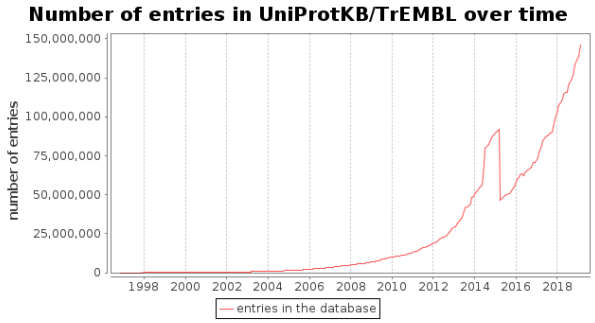

Figure 2: The number of entries in UniProtKB/TrEMBL has increased exponentially over the years.

3. How many reviewed entries are available for Homo sapiens?
4. What is the entry and length of the protein sequence for the reviewed human rhodopsin?
5. Using the entry you found above, answer the following:
   (a) What is the primary function of rhodopsin?
   (b) Which cell type and organelle is rhodopsin found in?
   (c) Name 2 diseases rhodopsin is involved in.

Don't forget to play around with the 3D structure for rhodopsin.

Part 2 - Calculating Amino Acid Frequencies [35 pts.]

1. Go to the Sequence section of rhodopsin in Uniprot. Download the fasta file.
2. Write a program to count the occurrences of amino acids in a given fasta file. Your program should get the fasta file name as a parameter and write the frequencies in a file with the file extension `_aaCount.txt`. For example, for rhodopsin the file should be, `P08100_aaCount.txt`. Use your program to count the frequencies of amino acids in the rhodopsin. There are multiple ways of doing the same thing so feel free to use your favorite packages in your own way. For example, you may use BioPython for reading the file. You may use the information here: http://biopython.org/DIST/docs/tutorial/Tutorial.html. BioPython was developed to read and manipulate biological data. You can use `SeqIO` to read your fasta file or write your own code.
3. What is the most frequent amino acid, what is the least common amino acid in the rhodopsin protein sequence? Plot a bar graph to the frequencies (in percentages) in decreasing order. Properly label your axes and use visible fonts.
4. Go to the https://www.ebi.ac.uk/uniprot/TrEMBLstats and find the most frequent and the least frequent amino acid in the Uniprot. What are they? Inspect the distribution, does rhodopsin's amino acids count follow a similar distribution?

Part 3 - Plotting an Hydropathy Plot [50 pts.]

Analyzing the amino acid sequence can provide some insights into the secondary structure of proteins. In this section, we will see how this is possible for a transmembrane protein. For a transmembrane protein, the membrane-spanning regions must be hydrophobic, since such regions are embedded in the hydrophobic tails of the lipid bilayer. To find such hydrophobic regions in a protein, we examine successive amino acids in the sequence. To measure the hydrophobicity of individual amino acids, a hydropathy score has been defined for each amino acid. A large positive hydropathy score means that the given amino acid is highly hydrophobic. Using the hydropathy scores for individual amino acids and the amino acid sequence for a given protein, a hydropathy plot can indicate regions of a protein which span the membrane [1].

1. Now you will write a program to calculate average hydropathy of the amino-acids along the protein sequence. We'll use the Kyte-Doolittle hydrophobicity scores for amino acids which you can find here. Implement a program `hydropathy_calculator.py`. It should calculate the average hydropathy along the protein sequence and plot a hydropathy plot. To generate a hydropathy plot, the amino acid sequence is scanned in successive segments (called windows) of a given size. For each window, the hydropathy index of the amino acids in that window is averaged to obtain the average hydropathy for that window. Plotting the average hydropathy (y-axis) against the amino acid position in the middle of each window (x-axis) generates the hydropathy plot. Your program inputs are the protein sequence and the window size to be used.
2. Set the window size to 5 and generate a hydropathy plot. Repeat it with the window size 20. Do you see a difference? Please comment.
3. Rhodopsin contains seven transmembrane helices. Go to Uniprot and report the residues where these helices start and end. Now looking at your hydropathy plot with window size set to 20 and the TM positions, do you see anything interesting? Write your observations.

Miscellaneous

- Beautiful Proteins: A blog containing "aesthetically pleasing" protein structures.

References

[1] David L Nelson, Albert L Lehninger, and Michael M Cox. *Lehninger principles of biochemistry*. Macmillan, 2008.

[2] Tetsuji Okada, Oliver P Ernst, Krzysztof Palczewski, and Klaus Peter Hofmann. Activation of rhodopsin: new insights from structural and biochemical studies. *Trends in biochemical sciences*, 26(5):318–324, 2001.

[3] Kurt Vonnegut. *Cat's cradle*. Dial Press, 1998.