Student Assignment

BioInformatics Final Project

Your final project will combine many of the elements that we have discussed throughout the semester. For this project, you will choose a protein of interest to you: this can be any protein you wish. Almost everything we do could work on proteins from any organism, however some analyses relate specifically to human pathology, so only select a human protein for this work. If you like cancer research, consider a protein involved in DNA damage surveillance or repair. If you like immunology, consider a protein in or on a favorite cell type – perhaps a T cell signaling protein.

Background, Protein Identification

Start by identifying your protein and providing some background information on it. Use the following resources to identify your protein and provide some background information

I. **NCBI Gene**. Obtain the AA sequences for two proteins of their choice using the NCBI's gene database.
   - i. Protein Common Name
   - ii. Host species (*Genus species*)
   - iii. NCBI Gene ID/Official Symbol
   - iv. Short Description
     - i. Write a simple description distilled from the NCBI Gene Summary and/or Wikipedia.
   - v. Expression Bar Chart
     - i. Click "See more" to view a bar chart containing data on where in the body's tissues the gene is expressed (as determined by RNA sequencing). Save and include this bar chart as the deliverable for this step.

II. Universal Protein Research Knowledgebase (UniProtKB)
   - 6. UniProt Entry Number
     - i. Follow the UniProt link in the Resources then search for the protein using the NCBI Gene ID
       - a. Carefully select the result that best matches the gene and organism of interest by clicking on the blue entry number.
       - b. This page will be used later to gather further details about the protein.

III. RCSB Protein Data Bank (PDB)
   - 7. RCSB PDB Solved Structure Identifier
     - i. Follow the RCSB PDB link in the Resources and search for the protein by either the common name or the NCBI Gene ID, making sure to select the organism of interest on the left.
     - ii. You must ensure that your chosen protein has an existing solved structure in this data bank in order to do a mutational analysis in later parts of this exercise.

IV. NCBI GenBank

8. AA Protein Sequence
    i. From the NCBI Gene page, go to the “Genomic regions, transcripts, and products” section and then click “GenBank” on the right. Scroll down to the first Coding Sequence “CDS” section and look directly after “/translation=” for the full protein sequence.
    ii. Sequence needs to be in FASTA Format consisting of ‘>’ followed by a simple name, a return, and then the sequence in one continuous line of text. See “FASTA Formatting” link in Resources.

Signal Sequence and Transmembrane Regions

Use the following resources to identify various regions of your protein

V. SignalP 6.0 Server

9. Predicted Signal Sequence.
    i. Following the link in the Resources for SignalP 6.0, paste the AA sequence copied from step 8, and select the following options: Organism: Other; Output format: Long Output; and Model Mode: Slow.
    ii. The results here will predict if a signal sequence is present on the protein, directing that ribosome towards the ER for synthesis.
    iii. If this signal sequence is present, copy and paste the FASTA sequence here, making the predicted signal sequence blue. If no signal sequence is present, indicate so.
10. Predicted Cut Site
    i. If a signal peptide is predicted, the cut site will be reported by SignalP as well.
11. Signal Sequence Probability Plot
    i. SignalP will also create a graphic representation showing the AAs predicted to be in this signal peptide region. Attach this plot here.

VI. Expasy ProtScale Web Tool

12. Kyte-Doolittle Hydropathy Plot
    i. Follow the ProtScale link in the Resources to open the web tool. Making sure to exclude the predicted signal sequence (if present),copy and paste the remaining chain of AAs from Step 9 into the indicated box.
    ii. Select the “Hydropath./Kyte & Doolittle” plot and leave all other options as the default then hit submit. Attach the resulting hydropathy plot as the deliverable for this step.

VII. UniProt Knowledgebase

13. Comparison of Predicted vs. Experimental Results
    1. Copy and paste the FASTA sequence from Step 8 here with the predicted signal sequence in blue for further editing.
    2. Return to the UniProt protein page and select the “Subcellular Location” section on the left and scroll down to the “Features” subheading.
    3. Here the experimentally determined extracellular, transmembrane, and intracellular regions of the protein can be found.
    4. On the FASTA sequence here, underline the extracellular region, make the transmembrane region red, and the intracellular region green.
    5. Note any discrepancies between the predicted and the actual regions of the protein.

3-Dimensional Modeling and Printing

Use the following resources to go from an AA sequence to a 3-D printed model

VIII. Swiss-Model Protein Structure Homology-Modeling Server

Swiss-Model is a fully automated server that references a database of proteins whose structure has been solved via X-Ray Crystallography. It then uses these solved structures as templates and combines them to generate a predictive model for new proteins.
Either the entire protein or just the extracellular or intracellular portions of the protein may be modeled, whichever is of greater interest.

14. 3-D Model of the Protein (.pdb file)
    1. To generate this model, follow the Swiss-Model link provided in the resources, paste the desired portion of the AA sequence from Step 17 into the first box, give the model a project title, and click “Build Model”.
    2. Once the model is done rendering, change the view on the right from “Cartoon” to “Surface”.
    3. Attach a screenshot of this entire page here. This view should show the template Swiss-Model used to construct the model (left) and the final image of the 3-D rendered protein (right).
    4. To download the .pdb file, click the “Model 01” dropdown menu and select the first “PDB Format” option.

IX. UCSF ChimeraX Molecular Visualization Program

Swiss-Model creates protein database (.pdb) files, however, most 3-D printers need the file in a stereolithography (.stl) format.

15. Printable 3-D Model of the Protein (.stl file)
    1. Use ChimeraX to convert the Swiss Model file into an .stl file, readable by most 3D printers. (note: ChimeraX only on Mac/iOS computers, Jmol and PyMol are similar programs which run on Mac, Windows, and Linux computers)

ii. This program is only available for mac iOS however alternative software such as Jmol or PyMol can be used if using a Windows- or Linux-based machine (links in Resources).

iii. Once this program is installed and open, go to "File" then "Open" and open the .pdb file from Swiss-Model for the protein.

iv. Once the file is open, select "Molecule Display" from the top banner then select "Electrostatic" in the "Coloring" section at the top.

v. Attach a clear screenshot of this protein here. Go to the "Right Mouse" section in the top banner and use the "Movement" tools to gain a better view of the protein.

vi. Next, go to "File" then "Save..." then select "STL (3D Printing)" for the type of file. Save this file for the next step.

X. **3D Printing**. Print your protein using a 3D printer. Download and convert your model into an .stl file and get it 3D printed. You will be turning in your models, so if you would like your own, be sure to make extra copies.

16. Printed 3-D Model

i. Using a 3-D printing software that is paired with the available 3-D printer, upload the .stl file and then print the protein.

a. Johnson County Public Library Makerspaces will print one free item per week per member. They will help you set up the machine and do everything you need to.

b. Alternatively, you can have your protein printed elsewhere, one local source is https://docs.google.com/forms/d/e/1FAIpQLSeiC7-fDuh7W66xgWirgLEvDHh0P2y7Ajcc8-aFu4OltI_AUw/viewform.

Experimental Analysis of Mutations

The goal for this portion of the assignment is to better visualize and understand how different mutations can affect proteins in different ways and it will require critical thinking along with some trial and error. For these next deliverables, use what is known of the properties of AAs (size, charge, hydrophobicity) to ascertain one tolerated and one non-tolerated mutation in the protein.

XI. RCSB PDB

The tools being used can only perform mutational analysis on a portion of a protein if that portion has a solved structure in the RCSB Protein Data Bank. Therefore, for Steps 17-23, use the FASTA format given by this data bank.

17. Solved RCSB FASTA Sequence

i. Return to the RCSB PDB page for the protein, click the "Display Files" button on the right of the "Structure Summary" page, and then click "FASTA Sequence".

ii. It is of note that the crystallized structure may be a fraction of the total protein. This sequence may also be expressed with an artificial start codon or other tags of AAs used for isolation of the recombinant protein. These AAs will not be found within the original sequence and should be ignored.

iii. Paste the AA sequence from Step 13 here then bold the portion of the sequence that corresponds to the FASTA sequence from the RCSB PDB.XII.

XII. Sorting Intolerant From Tolerant (SIFT) Sequence Tool

The SIFT Sequence tool is an AA substitution tool that predicts whether any given AA change will alter the structure and function of a particular protein based on sequence homology and the physical and electrochemical properties of the AAs themselves.

18. Tolerated Mutation Example

i. Follow the SIFT link in the Resources. Under "Single Protein Tools" on the right, select the "SIFT Sequence" tool.

ii. Paste the AA sequence excluding the signal sequence in FASTA format (including the name) copied from Step 17 into the first box.

iii. In the second box, enter an AA mutation that is predicted to be **Tolerated**. The format is to have X#Y where X is the original AA, # is the position of the mutation, and Y is the new AA. When determining the AA position, for this resource specifically, start counting from the first AA in the submitted sequence.

iv. It is essential that the AA residue picked is within the bold portion of the sequence.

v. Leave all other parameters on this page the same and click "Submit".

vi. If this AA substitution was predicted to be tolerated, copy and paste the resulting text here.

vii. If this AA substitution was reported as "affecting the protein's function" and is therefore non-tolerated, evaluate why that might be and try again.

19. Reason For Being Tolerated

i. Propose an explanation for why this mutation would be tolerated based on the structural and electrochemical properties of the AAs involved.

20. Non-Tolerated Mutation Example

i. Repeat Step 18, but with the intent of finding an AA substitution that would be **Non-tolerated** and would affect the protein's function.

21. Reason For Being Non-Tolerated

i. Propose an explanation for why this mutation would be non-tolerated based on the structural and electrochemical properties of the AAs involved.

XIII. Site-Directed Mutator (SDM) Server

SDM is a web server used to predict the effects of various mutations on a specific protein's structure and function. A wild-type AA structure along with a proposed AA substitution are input and a stability score is calculated.

22. Tolerated Mutation Comparison
    - i. Follow the SDM link in the Resources.
    - ii. In the "Single Mutation" section enter the following:
        - a. First text box: the PDB Identifier from Step 7
        - b. Second: Enter the AA mutation that was predicted to be **Tolerated** by SIFT. The format is the same, however, when determining the AA position, for this resource specifically, start counting from the first AA in the Step 17 sequence.
        - c. Third: For most cases, entering "A" here will work
    - iii. Click the green "Run SDM" button
    - iv. The results of this prediction will be on the top left under "Predicted pseudo ΔΔG:" This will reflect how this specific mutation will the form and possibly function of the protein.
    - v. Use the interactive prediction tool on the right to examine how the substituted AA physically fits into the folded protein.
    - vi. Capture a screenshot of this entire page and attach it here along with a statement comparing these results to those previously obtained from SIFT.

23. Non-Tolerated Mutation Comparison
    - i. Repeat Step 22, but with the mutation that was predicted by SIFT to be **Non-Tolerated**.

XIV. PinSnps Tool for Human Protein Networks and SNPs

While the two previous mutational analysis were done using random, user-defined mutations, in the final portion of this assignment, use the PinSnps tool to investigate what actual, clinical mutations exist and appreciate their corresponding health consequences. Note that Single Nucleotide Polymorphisms (SNPs) are commonly observed mutations that may result in a change to the amino acid called for at a specific location. Many known SNPs are associated with pathologies.

24. Pathological Mutation ID with Reference
    - i. Follow the PinSnps link in the Resources and clear out the default entries in the first three boxes.
    - ii. Copy and paste the UniProt Entry Number from Step 6 into the first box and click "Query".

iii. Shown below are any published SNPs or SNVs (Single Nucleotide Variants) that are known, sorted into five sections. Expand any of these sections containing identified mutations by clicking the plus button on the right.

iv. If possible, identify one of these known mutations that is associated with a predicted non-benign/"possibly damaging" outcome.

v. Follow the link for that mutation by clicking on the "Mutation ID"

vi. Record the Mutation ID here.

vii. Scroll down to the References sections and also include the Reference Title, Author, year, and Journal of any one of the sources where this mutation was described.

Deliverable

Use the outline above to create your deliverable. Remove the instruction text from each section and replace it with your answer or a screenshot image of your result. Turn in a 3D model of your protein.

Resources

HHMI BioInteractive	www.biointeractive.org
NCBI Gene Database	https://www.ncbi.nlm.nih.gov/gene/
UniProt	https://www.uniprot.org/
RCSB Protein Database	https://www.rcsb.org/
FASTA formatting:	https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp
SignalP 6.0	https://services.healthtech.dtu.dk/service.php?SignalP
Swiss Model	https://swissmodel.expasy.org/ https://swissmodel.expasy.org/docs/help
ChimeraX	https://www.cgl.ucsf.edu/chimera/
Jmol	http://jmol.sourceforge.net/
PyMol	https://pymol.org/pymol.html?
JoCo Library Makerspace	https://www.jocolibrary.org/makerspace
SIFT	https://sift.bii.a-star.edu.sg/
SDM	http://marid.bioc.cam.ac.uk/sdm2/prediction
PinSnps	https://fraternalilab.kcl.ac.uk/PinSnps/faces/index.xhtml