Homework Three

Please hand in the solutions to the following problems on Thursday, October 21, 2014. Please remember to attach the assignment cover sheet to your solution.

Problem 1 Problems from the textbook – Chapters Two, Three and Four

1) In general, where do large hydrophobic groups of amino acids cluster and where do most polar groups cluster in a protein?
2) a) What is a protein domain?
   b) How many amino acids does a typical domain have?
   c) What is the core of each domain mainly composed of?
3) Is adult hemoglobin an example of a tetrameric quaternary structure? Explain.
4) According to the summary of Chapter Two, what is one of the main aims of bioinformatics?
5) According to the summary of Chapter Two, what should be done to the protein sequence under study in order to obtain more information about it and to perform accurate predictions?

---------- End of Chapter Two ----------

6) Relational databases are more sophisticated than flat file structures. Yet, flat files are still being used, especially for data distributing purposes. Why are flat files still being used?
7) What do SQL, HTML, and XHTML stand for?
8) a) What is meant by “annotation”?
   b) What are some of the features it can include?
9) What is meant by “Gene Ontology”?
10) Which journal reports new and updated databases in the beginning of every year?
11) What are the three types of DNA sequences stored in databases containing information about nucleic acids?
12) a) What kind of databases are DIP and pSTIING?
   b) What is the main difference between them?
13) What is a “nonredundant database”?
14) When and why are genes and proteins labeled “hypothetical”?
15) How are the sequences in the Swiss-Prot protein sequence database curated?

---------- End of Chapter Three ----------

16) The identification of similar sequences has many applications. Name

some of them explaining the importance of that application.

17) a) What are pseudogenes?
b) Do they all arise from gene duplication?
c) What is the estimated number of pseudogenes in the human genome?
d) Give some examples of pseudogenes.

18) a) What is meant by convergent evolution?
b) How does it differ from divergent evolution?

19) a) Why is it easier to detect homology when comparing protein sequences than when comparing nucleic acid sequences?
b) When is it necessary to compare DNA sequences?

20) When working with dotplots, they often suffer from background noise. How do we get rid of that noise? Explain.

Problem 2

We have studied in class the mutation in the hemoglobin sequence that causes sickle cell anemia. Let us now look at where that mutation is in the hemoglobin quaternary structure. For proteins, there are 4 levels of structure.

- The first, **primary structure** is composed of the amino acid sequence.
- The **secondary structure** explains how the different amino acids next to each other in the sequence are organized.
- The **tertiary structure** is the folded 3-D structure of the protein that allows it to perform its functions.
- The **quaternary structure** is the total protein structure that is made when all the subunits of the protein are in place.

A) Visualizing Proteins with Protein Workshop at PDB

You will use a molecular modeling program called Protein Workshop which is 3D program for visualizing proteins. It is part of the Protein Data Bank (PDB) that contains structures of many proteins.

- Go to www.rcsb.org
- To access the hemoglobin structure of a normal protein, type “2DN2” in the “Search” window at the top of the page and click on the blue key (with the magnifying lens) to search the database.
- From the new page, click on "Protein Workshop" on the right hand side.
- If you are not already in “RCSB PDB Protein Workshop 4.1.0 (powdered by the MBT): 2DN2”, then click on "Launch RCSB - Protein Workshop" and then on "Run" from the window that pops up.
- From the window with title: “RCSB PDB Protein Workshop 4.1.0 (powdered by the MBT): 2DN2”, click on the “Visibility” button.

We are going to use the Protein Workbench to look at the structure of the wild-type hemoglobin first, and later, at the sickled cell hemoglobin.

B) Structure of Normal Hemoglobin

You should now have a ribbon structure of the normal, wild-type (not mutated) hemoglobin in the Protein Workshop window. This is a type of visualization in the Protein Workshop package. Note that the four hemes are represented ball and stick style.

- Try grabbing (keep the left key pushed down) and rotating the molecule with the cursor.
- Click on the small "▷" sign next to "Chain B" to view the amino acids of the beta globin protein (chain B) with their positions.
- Scroll to "6 GLU" and click on it. This will put a small mark on the graph. You might have to rotate the molecule with the mouse to be able to see the position of the glutamate on the graph.
- Click on the small "▷" sign next to "Chain D" this time, to view the amino acids of the beta globin protein (chain D) with their positions.
- Scroll to "6 GLU" and click on it. This will put a small mark on the graph. You might have to rotate the molecule with the mouse to be able to see the position of the glutamate on the graph.

Let us display the word glutamate in the sixth position of the normal beta globin chains.

- Choose "Labels" from "Select your tool".
- Choose "Label by residues" from "Change the tool's options, if necessary."
- Click once more on "6 GLU" on the B and D chains to see the positions of glutamate on the hemoglobin.

1) Where are the "6 GLU" residues in relation to the rest of the hemoblogin: inside or outside?

- Choose "exit" from the drop-down window one gets by clicking on "File" (upper left-hand corner) of the Protein Workshop window.

In the next section we study the structure of the sickle cell hemoglobin to see how it differs from the normal hemoglobin.

C) Structure of Sickle Cell Hemoglobin

To access and view the hemoglobin molecule with the sickle cell mutation:

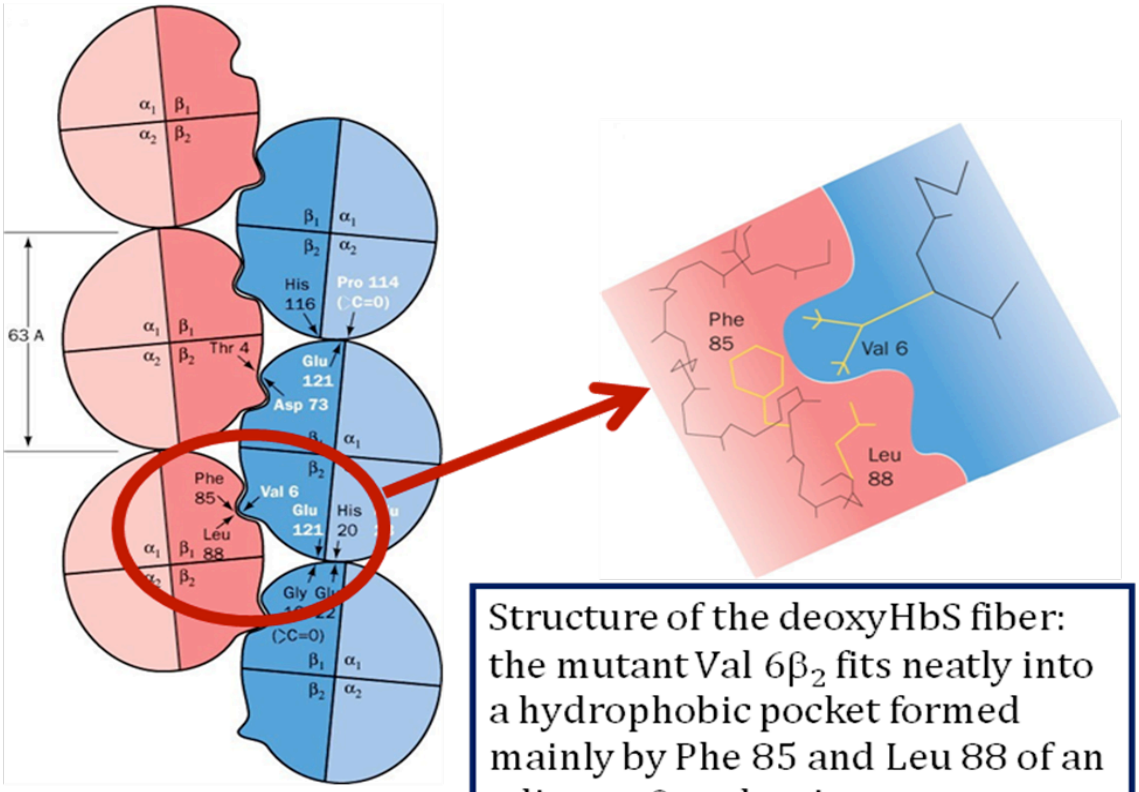
- Go to www.rcsb.org
- Type “2HBS” in the “Search” window at the top of the page and click on the blue key (with the magnifying lens) to search the database.
- Click on the arrow to the left of “Biological Assembly 1” to get “Assymmetric Unit” (which will display 2 sickle-cell hemoglobins).
- Click on "Protein Workshop" on the right hand side.
- If you are not already in “RCSB PDB Protein Workshop 4.1.0 (powdered by the MBT): 2HB2”, then click on "Launch RCSB - Protein Workshop" and then on "Run" from the window that pops up.
- From the window with title: “RCSB PDB Protein Workshop 4.1.0 (powdered by the MBT): 2HB2”, click on the “Visibility” button.

Notice that there are actually 2 hemoglobin tetramers. A tetramer is a protein made up of four subunits. We will see why we see 2 hemoglobins pretty soon.
2) What residue is now found on all 4 beta chains of the two hemoglobins?

In the next section we study how and why the two hemoglobins are "attached".

D) Sickled Hemoglobins Attach to Other Sickled Hemoglobins

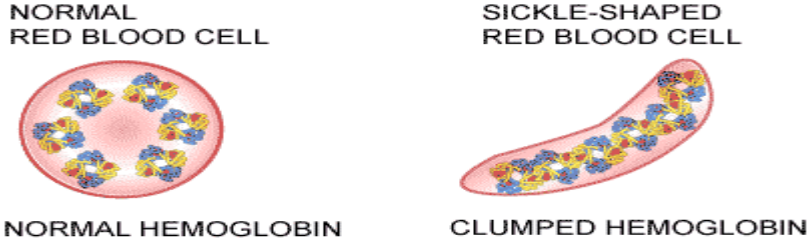
We know that valine is a hydrophobic ("afraid of or does not like water") amino acid, and that it is located on the outside of the hemoglobin tetramer, surrounded by water. Hydrophobic amino acids try to associate with each other. In order to get away from the water molecules, valine-6 interacts with hydrophobic amino acids, such as phenylalanines (85 PHE) and leucines (88 LEU), of neighboring hemoglobins. This causes the association of 2 individual hemoglobin molecules as can be seen in the following figure (Biochemistry by Voet and Voet, 4th Edition, 2010).



Structure of the deoxyHbS fiber: the mutant Val $6\beta_2$ fits neatly into a hydrophobic pocket formed mainly by Phe 85 and Leu 88 of an adjacent β_1 subunit.

The mutated valine-6 residues just keep adding on more hemoglobin molecules as they try to stabilize their structure. The amino acids in a protein form associations with each other to form fibers.
As the fibers form, they cause the shape of the red blood cell to become sickle shaped. The long fibers push the cell membrane out of shape, causing the characteristic shape of the red blood cells in the disease. These cells can no longer move normally through the blood vessels, so normal delivery of oxygen to the body is interrupted. This is what causes the disease, sickle cell anemia.

3) Use "Visibility", "Labels", "Label by residues", and any other feature you deem necessary to show how one 6 VAL from a beta globin chain from one hemoglobin is associated with 85 PHE and 88 LEU of the beta globin chain of the other hemoglobin. Give a screen shot of the figure you create.



Problem 3

Consider the following sequence:

```
GCTGGATCCACTGGAGCAGGCAAGACTTCACTTCTAATGGTGATTATGGGAGAACTGGAG
CCTTCAGAGGGTAAAATTAAGCACAGTGGAAGAATTTCATTCTGTTCTCAGTTTTCCTGG
ATTATGCCTGGCACCATTAAAGAAAATATCATCTTTGGTGTTTCCTATGATGAATATAGA
TACAGAAGCGTCATCAAAGCATGCCAACTAGAAGAGGACATCTCCAAGTTTGCAGAGAAA
GACAATATAGTTCTTGGAGAAGGTGGAATCACACTGAGTGGAGGTCAACGAGCAAGAATT
```

- Go to the NCBI website at http://www.ncbi.nlm.nih.gov

You will perform a BLASTN search of the given sequence against the NCBI non-redundant database.

- Click on “BLAST”
- Paste the sequence
- Choose “Others (nr etc.):” from “Database”
- Choose “Somewhat similar sequences (blastn)” from “Optimize for”
- Click on “BLAST”

a) To what organism does the selected sequence most probably belong?
b) What type of sequence is it? (DNA, RNA)

From the BLASTN output, choose a human, full-length mRNA which significantly aligns with the given sequence.

c) What is the mRNA's accession number?
d) What is the mRNA's GI number?
e) On what chromosome is the mRNA sequence located?
f) What part of the sequence actually encodes for a protein?
g) What is the protein's function?

One of the hits you obtain is with the organism Sus scrofa with NM_001104950.1.
h) What organism is that?
i) Explain why translation ends with the specific amino acids: “VQETRL”. Show all your work.

Problem 4

Choose only **one** of the following 4 problems (A, B, C, or D) to answer the following questions about the **two mutations** you choose:

1) locate the mutation on the sequence
2) explain what consequences on the protein it might have

A) "Changes in the Epidemiology of Thalassemia in North America: A New Minority Disease" by Elliott P. Vichinsky, et al. was published in "Pediatrics" in 2005 (issue 1, pages 1098-4275). The article can be found at: http://pediatrics.aappublications.org/cgi/content/full/peds.2005-0843v1
Choose two mutations from the upper half of Table 1 (mutations of the β globin gene associated with the phenotypes in the β-thalassemias in North America) that were not covered in Hands-On 8.

B) "Prevalence of various mutations in beta thalassaemia and its association with haematological parameters" by Khattak et al. was published in "Journal Pakistan Medical Association" in January 2012.
Choose two mutations from Table 1 (Frequencies of β thalassaemia mutations and their distribution in various ethnic groups) that were not covered in Hands-On 7.

C) "Beta-thalassemia" by Galanello and Origa was published in "Orphanet of Rare Diseases" in 2010.
Choose two mutations from Table 1 (Common types of beta-thalassemia: severity and ethnic distribution) that were not covered in Hands-On 7.

D) "Beta-thalassemia" by Cao et al. is part of GeneReviews™ - NCBI Bookshelf at NCBI: http://www.ncbi.nlm.nih.gov/books/NBK1426/
Choose two mutations from Table 5 (Mild and Silent HBB Mutations Causing β-Thalassemia) that were not covered in Hands-On 7.