Class: Biol/CS123A **Name:** Zayd Hammoudeh

Assignment: #3

Problem: #1 / |

Problem: #2 / |

Problem: #3 / |

Problem: #4 / |

Grade: Total / |

Problem #1

Note: The page numbers below are for my studying later on and can be ignored during grading.

Chapter #2 (Continued):

1.1) In general, where do large hydrophobic groups of amino acids cluster, and where do most polar groups cluster in a protein?

Hydrophobic amino acids tend to be buried within the protein surrounded by other hydrophobic amino acids. The polar, hydrophilic amino acids like to be surrounded by water molecules, which they can interact with; they are often buried inside the protein with another oppositely charged hydrophilic residue that it interacts with. (See page 29)

1.2) What is a protein domain? How many amino acids does a typical domain have? What is the core of each domain composed of?

A **protein domain** is a conserved part of a given protein sequence that can evolve, function, and exist independently of the rest of the protein chain. (Wikipedia entry on protein domain) The protein chain folds into these discrete structural units. A protein domain can be from 50 to around 350 amino acids in length. The core of each domain is mainly composed of tightly packed α-helices, β-sheets, or a mixture of both. (See page 41)

1.3) Is the adult hemoglobin an example of a tetrameric quaternary structure? Explain.

A **tetramer** is a protein consisting of four monomer subunits. Tetrameric hemoglobin has four units, specifically two α and two β units. (See page 43)

1.4) According to the summary of Chapter #2, what is one of the main aims of bioinformatics?

One of the main aims of bioinformatics is to predict and analyze the structure of proteins and the relationship of the structure to the function. (See page 43)

1.5) According to the summary of chapter #2, what should be done to the protein sequence under study in order to obtain more information it and to perform accurate predictions?

The protein should be aligned with other proteins to find homologs. (See page 43)

Chapter #3:

1.6) 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?

Many of the more complex database structures depend on specific, often expensive software whereas the flat files can be read and analyzed by many alternate programs according to the user's preferences. (See page 49)

1.7) What do SQL, HTML, and XHTML stand for?

SQL – Structured Query Language (See page 49)
HTML – Hypertext Markup Language (See page 50)
XHTML – Extended Hypertext Markup Language (See page 50)

1.8) What is meant by "annotation," and what are some of the features it can include?

Entries in the major protein and nucleotide sequence databases have large amount of relevant non-sequence information. This additional information is referred to as the **annotation** and can include links to related entries in other databases, interpretation of the data, and relevant research citations. (See page 53)

1.9) What is meant by "Gene Ontology"?

An **ontology** is a set of field-specific descriptors that enable the sharing of the same concepts and definitions for specific terms. One of the most common ontologies is the gene ontology, which provides a controlled vocabulary for genes. (Ontology definition page 743) **Gene Ontology** is a collaborative project across many laboratories to provide a controlled vocabulary that describes gene and gene-associated information (but not gene byproducts) for all organisms. (See page 54)

1.10) Which journal reports new and updated databases at the beginning of every year?

Nucleic Acid Research (See page 55)

1.11) What are three types of DNA sequences are stored in databases containing information about nucleic acid sequences?

There are three types of DNA sequences stored in databases. They are: raw genomic sequences, cDNA, and expressed sequence tags (ESTs). **Raw genomic sequence** data represents the chromosomal DNA and includes non-coding regions, introns, control regions, and exons. **cDNA** (abbreviation for complementary DNA) is the result of reverse transcription from RNA to DNA. Since these samples are synthesized from RNA, they do not including anything beyond the coding sequence (e.g. introns, control sequences, etc. are excluded). An **expressed sequence tag** (EST) is a partial CDNA sequence that is generally around 300 nucleotides in length. (See page 56).

1.12) What kinds of databases are DIP and pSTIING? What is the main difference between them?

DIP stands for "Database of Interaction Proteins"; it contains information only on protein-protein interactions. It employs rigorous criteria for evaluating the reliability of each interaction.

pSTIING stands for "protein Signaling, Transcriptional Interaction, and Inflammation Networks Gateway"; in addition to protein-protein interaction, it also integrate protein-anything else interactions as well as transcriptional associations. At its simplest level, both are protein interaction databases. (See pages 58-59)

1.13) What is a "non-redundant database"?

A **non-redundant database** has no duplicate entries. (See definition on page 743) This can also mean that in two entries in a database, there is not duplicate information. However, the definition of redundancy does differ from database to database.

1.14) When and why are genes and proteins labeled hypothetical?

A gene is marked **hypothetical** in a database when it is identified in a nucleotide sequence purely through computational methods, and there is no experimental data available yet to support the predicted gene. This distinction is required because while the gene may be correctly predicted, it is also possible no gene exists at all or that there where errors in prediction that would lead to a different amino acid sequence. This enables the user of the database to show sufficient caution when encountering these entries. (See page 65)

1.15) How are the sequences in the Swiss-Prot protein sequence database curated?

Swiss-Prot does not use computer based annotation. Instead, it is done manually by specialists to produce high-quality annotations. (See page 65)

Chapter #4:

1.16) The identification of similar sequences has many applications. Name some of them explaining the importance of that application.

Uncharacterized genomic DNA sequences can be compared with known sequences in databases to determine whether the DNA sequence is likely to contain or be part of a protein coding gene.

Protein function is entirely determined by its amino acid sequence, which drives its unique 3-dimensional shape. By comparing an amino acid sequence with unknown function to other known amino acid sequences, it is often possible to predict the protein's structure and in turns its behavior. (See page 72)

Sequence alignment can indicate homology between sequences. While genes may be similar in that they have some degree of match, that does not imply they are homologous and have common evolutionary origins. Sequence comparison methods and the scoring systems used take these factors into account and allow a researcher to discriminate between fortuitously good alignments and real evolutionary relationships.

1.17) What are pseudogenes? Do they all arise from gene duplication? What is the estimated number of pseudogenes? Give some examples of pseudo genes.

Pseudogenes are sequences in genomic DNA that are similar to known coding-genes but do not produce a functional protein. Pseudogenes are assumed to arise from gene duplication when one of the gene copies undergoes a mutation that either prevents its transcription or leads to a non-functional protein. Since the pseudogene sequence is no longer under selection pressure to retain protein function, it will generally accumulate further mutations at a higher rate than a functional gene. There are up to 20,000 pseudogenes in the human genome. In one case, RNA from a transcribed pseudogene regulates the expression of the corresponding functional gene. (See page 73). Examples of pseudo genes is the gene that codes the enzyme L-gulono-γ-lactone oxidase, deactivation of the caspase 12 gene, and the Bovine Seminal Ribonuclease which was a pseudogene for 20 million years until it was resurrected through mutation (Pseudogene Wikipedia article).

1.18) What is meant by convergent evolution? What is meant by divergent evolution?

Convergent evolution is when organs, proteins, and DNA sequence that are **unrelated in their evolutionary origin** acquire the same structure or function. An example of convergent evolution is wings in bats and insects, which share no common ancestral structure. Convergent evolution does not generally produce highly similar sequences of any great length.

Divergent evolution produces different structures or sequences from a **common ancestor**. (See page 75) It is the process by which a single ancestor or ancestral gene is modified over time into two or more descendants that have an increasing degree of dissimilarity as the time since they diverged increases. It is also called **adaptive evolution**. (See definition on page 737).

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

First, there are only 4 unique bases in DNA compared to 20 unique amino acids. Hence, one amino acid position contains much more information than one character in a nucleic acid sequence. Second, the genetic code is redundant in that there are two or more different codons for most amino acids. This means slight, insignificant variations in the nucleic acid sequence are filtered out when looking at an amino acid sequence. Third, the structure of a protein is fully derived from its amino acid sequence. Hence, the importance of maintaining protein function leads to protein amino acid sequences changing less over time than nucleic acid sequences.

It is necessary to compare DNA sequences when looking for promoters or other regulatory sequences that are not present in the amino acid sequence. DNA sequence comparison is also used when looking at gene identification. (See pages 75-76)

1.20) When working with dot plots, they often suffer from background noise. How do we get rid of this noise? Explain.

Background noise can be eliminated by adjusting the window size (i.e. number of characters considered per comparison) and the stringency (number of identical matches in the comparison window to be considered a match). Adjusting these settings acts as a filter by requiring that a comparison achieves some minimum identity score to be considered a match. When this is done, only diagonals will survive the filter removing spurious noise. (See pages 77-78)

Problem #2

Using the RCSB Protein Database.

1. **For the 2DN2 hemoglobin molecule, where are the 6-GLU molecules located on chains B and D? Are they on the outside or inside?**

They are located on the outside. Figure 1 is captured from **RCSB Protein Workshop**.

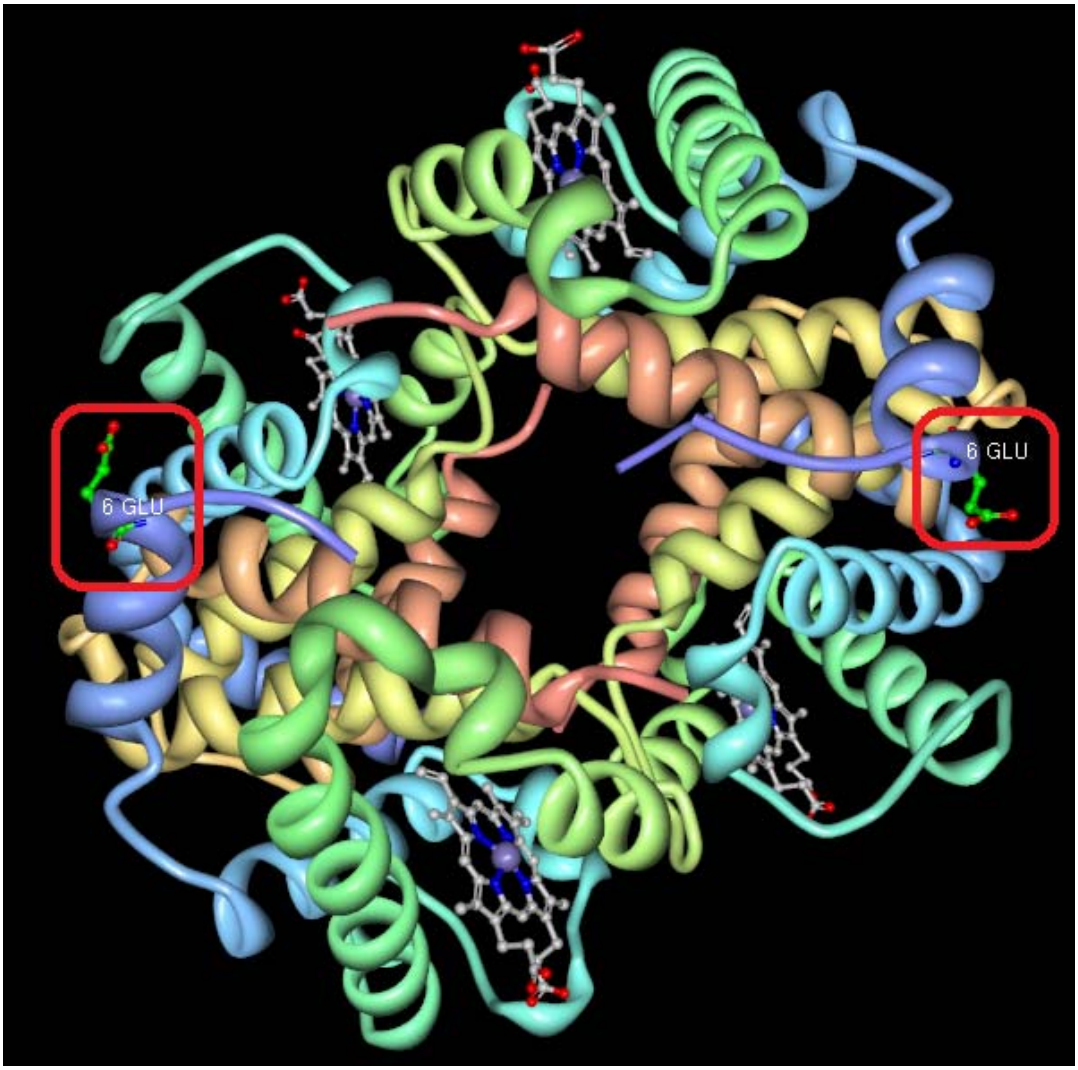

Figure 1 – 6-GLU Amino Acid in Hemoglobin Protein

2. **In the 2HBS hemoglobin molecule, what residue is now found on all 4 beta chains of the two hemoglobins?**

In place of the glutamate in position 6, there are now valine amino acids. There are also 153 HEM molecules on the hemoglobins where previously there were previously 142 HEM and 147 HEM molecules on the hemoglobin's alpha and beta chains.

3. **In a screenshot, show the association of the 6-VAL from one hemoglobin's beta chain to the 85 PHE and 88 LEU in another hemoglobin's beta chain.**

Figure 2 shows the two hemoglobin molecules interacting with the valine (6 VAL) amino acid of one hemoglobin adjacent to the phenylalanine (85 PHE) and leucine (88 LEU) amino acids of another hemoglobin molecule. It is circled in red.

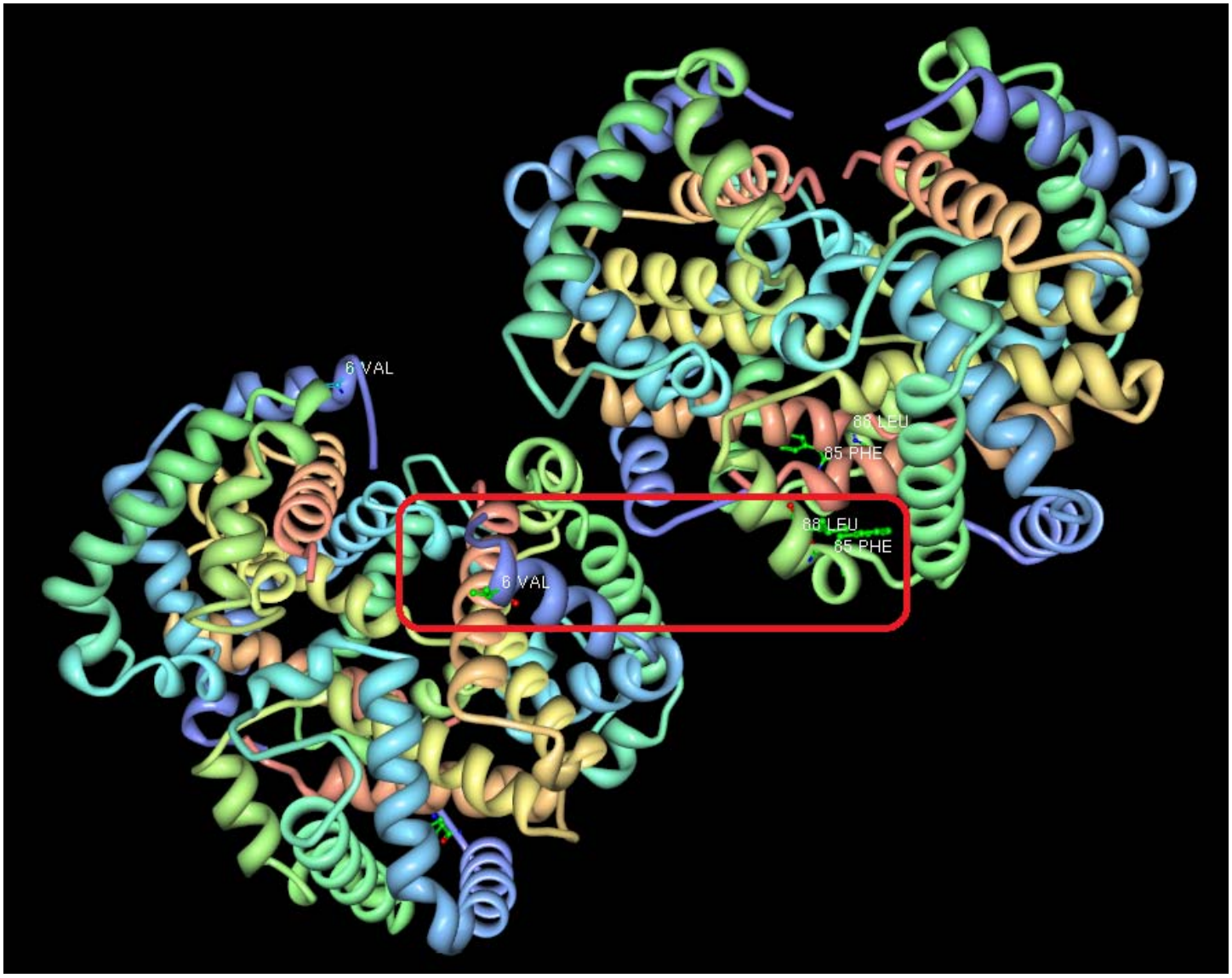

Figure 2 – Interaction Between two Hemoglobin Molecules

Problem #3

Given the nucleic acid sequence below, answer the following questions using BLASTN and GenBank:

GCTGGATCCACTGGAGCAGGCAAGACTTCACTTCTAATGGTGATTATGGGAGAACTGGAG
CCTTCAGAGGGTAAAATTAAGCACAGTGGAAGAATTTCATTCTGTTCTCAGTTTTCCTGG
ATTATGCCTGGCACCATTAAAGAAAATATCATCTTTGGTGTTTCCTATGATGAATATAGA
TACAGAAGCGTCATCAAAGCATGCCAACTAGAAGAGGACATCTCCAAGTTTGCAGAGAAA
GACAATATAGTTCTTGGAGAAGGTGGAATCACACTGAGTGGAGGTCAACGAGCAAGAATT

a) To what organism does the selected sequence most probably belong?

Homo sapiens (Human)

b) Which type of sequence is it? (DNA, RNA)

It is mRNA.

Choose a human, full length mRNA which significantly aligns with the given sequence.

c) What is its accession number?

XM_006715842

d) What is the mRNA's GI number?

578813888

e) On what chromosome is the mRNA sequence located?

Chromosome 7

f) What part of the sequence actually encodes for a protein?

The coding sequence codes for the gene. It is nucleotides 1 through 4767 (note the codon frame is 1 – codon_start=1).

g) What is the protein's function?

Cystic fibrosis transmembrane conductance regulator in humans. This gene encodes a member of the ATP-binding cassette (ABC) transporter superfamily. ABC proteins transport various molecules across extra- and intra-cellular membranes.

Using the mRNA record for the organism *Sus scrofa* with accession number NM_001104950.1. Answer the following questions.

h) What organism is that?

Sus scrofa is a pig.

i) Explain why translations ends with the specific amino acids "VQETRL".

The GenBank record includes the complete gene, which is 4449 bases in length. It also says the gene in the record uses the first codon (i.e. first fame). Hence, the last 21 bases (7 codons) in the record form the last seven codons; these bases are shown in Table 1. For each codon, the associated amino acid is listed. Note that the sequence is VQETRL followed by the stop codon.

Codon #	DNA Sequence	mRNA Sequence	Amino Acid
1476	GTG	GUG	Val (V)
1477	CAA	CAA	Gln (Q)
1478	GAA	GAA	Glu (E)
1479	ACA	ACA	Thr (T)
1480	AGA	AGA	Arg (R)
1481	CTT	CUU	Leu (L)
1482	TAG	UAG	Stop

Table 1 - Amino Acid Assignment for *Sus Scrofa* Gene

Problem #4

The paper I selected as the reference is: "Prevalence of various mutations in beta thalassaemia and its association with hematological parameters" by Khattak *et. al*. Below is a table showing the first 31 codons in the beta thalassemia protein.

Codon #	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Codon	AUG	GUG	CAU	CUG	ACU	CCU	GAG	GAG
Codon #	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Codon	AAG	UCU	GCC	GUU	ACU	GCC	CUG	UGG
Codon #	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**
Codon	GCC	AAG	GUG	AAC	GUG	GAU	GAA	GUU
Codon #	**24**	**25**	**26**	**27**	**28**	**29**	**30**	
Codon	GGU	GGU	GAG	GCC	CUG	GGC	AG**X**G	

Table 2 – First 31 Codons in Beta Thalassemia Protein

Between the two guanine molecules in codon #30, there is an intron (shown as a red **X** in Table 2). Khattak et. al. mention two possible mutations in codon #30; they are: G→C and G→A. Depending on which of the two guanine molecules in codon #30 (AGG) is affected by these mutations, the effect on the resulting amino acid sequence could be essentially the same or very different. If the first guanine molecule is mutated, then the possible resulting codons for these two mutations is: ACx and AAx. In both cases, the first intron would be affected since the intron splice sequence (AG-GU) is compromised. This would cause the protein to be elongated and possibly experience a frame shift. In contrast, if the second guanine molecule was mutated, then the second exon would be either partially or entirely truncated since the splice termination sequence (AG-G) is compromised.

Another possible mutation is in codon #15 (UGG) where a guanine molecule is mutated to an adenine. This will result in the codon being changed to UGA or UAG, which are both stop codons. This is an example of a **nonsense mutation** since the protein is truncated prematurely.