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

Assignment: #2

Problem: #1 / |

Problem: #2 / |

Problem: #3 / |

Grade: / |

Problem #1

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

Chapter #1 (Continued):

1.1) What are promoters? In bacteria, promoters typically occur immediately before the position of what site?

A **promoter** is a control region in DNA at which RNA polymerase binds to initiate transcription. RNA polymerase binds more tightly to these regions of DNA than other regions. Bacterial promoters typically occur immediately before the position of the **transcription start site** (TSS). (See page 15)

1.2) What is one of the main problems in finding promoters in DNA sequences?

Motifs are characteristic short sequences in DNA/RNA. Gene to gene, promoter motifs vary somewhat. What is more, outside the set of conserved motifs, promoter sequences vary considerably. (See page 16)

1.3) Is the prokaryote terminator sequence more variable than the promoter sequence? Is the prokaryote terminator sequence usually included in genome annotations?

In prokaryotes, the terminator sequence is more variable than the promoter. It is not usually included in genome annotations. (See page 16)

1.4) What are activators and repressors? Why are they of biological importance?

Proteins that improve the correct binding of RNA polymerase are called **activators** while those that block the promoter sites and inhibit the expression of a gene are called **repressors**. These additional repressor and activator proteins have a profound influence on whether, and when, transcription actually occurs. (See page 16)

1.5) What is the most important core promoter sequence in genes transcribed by RNA polymerase II? Where is it located?

The most important core promoter is called the **TATA Box** which is characterized by the TATA nucleotide sequence. It is about 25 nucleotides upstream (i.e. before) the start of transcription. (See page 17)

1.6) What is the major difference between eukaryotes and prokaryotes in terms of their transcription and translation processes?

The major difference between eukaryotes and prokaryotes in terms of their transcription and translation is that eukaryote mRNA transcripts are substantially modified before translation. The first modification occurs while the transcription is in process and involves the addition of a modified guanosine nucleotide to the 5' end of the transcript in a process called **RNA capping**. The second step involves the cleavage of the mRNA after a Cytosine-Adenine (CA) sequence and adding of approximately 200 adenosine nucleotides at the 3' end in a process called **polyadenylation**. (See page 18)

1.7) What is the role of the spliceosome and what does it consist of?

The spliceosome is complex consisting of small nuclear RNA (snRNA) molecules and proteins. It contains the enzymatic activity needed for **RNA splicing** where the mRNA sequence is cleaved, introns removed, and exons rejoined in the mRNA transcript. (See page 18)

1.8) What is meant by "alternative splicing"? Is it quite common in the genes of humans and other mammals?

In **alternative splicing**, some exons and/or parts of exons are excluded in RNA splicing. This allows different versions of a protein to be produced from the same gene. It is quite common in humans and mammals and is thought to be one of the means by which a relatively small number of genes present in the genome can specify a much greater number of proteins. (See page 19)

1.9) What is the Shine-Dalgarno sequence? What is its consensus sequence, and where does it occur?

In bacterial DNA, the **Shine Dalgarno sequence** is a short sequence at the 5' end of mRNA that indicates the ribosome binding site. The typical consensus sequence is AGGAGGU and occurs a few bases upstream of the AUG translation start codon. (See page 19)

1.10) What is an operon?

Functionally related protein coding sequences can be clustered together into **operons**. Each operon is transcribed into a single mRNA transcript; the separate, individual proteins are then translated separately from this one long mRNA molecule. Operons are rarely found in eukaryotes. (See page 19)

1.11) How do viruses replicate?

A virus has a very small genome that encodes the proteins that make up the virus structure. Viruses replicate by "hijacking" a living cell's biochemical machinery for replicating DNA and synthesizing proteins. Viruses may have either DNA or RNA genomes. (See page 21)

1.12) Give an example of an unusual feature that is found in some viral genomes but not in cellular genomes.

Overlapping genes (see page 21)

1.13) What are plasmids?

A **plasmid** is extrachomosomal DNA found in bacteria. These small circular DNA can be transmitted from bacterium to bacterium. Plasmids are often associated with genes for drug resistance. (See page 21)

1.14) On what process does the fate of a mutation (to be lost or to be retained) depend?

Whether a mutation is lost or retain is dependent on the process of natural selection. Beneficial and benign mutations can be passed on to subsequent generations while deleterious mutations are generally not. (See page 23)

1.15) What can be said about the general statements made in chapter one?

Chapter #1 consisted primarily of general statements regarding DNA sequences, protein encoding, and cell signaling. However, there are exceptions to these general statements. For example, the codon UGA while usually a stop codon can also code for the amino acid selenocysteine. (See page 23)

Chapter #2:

1.16) What is one of the challenges facing bioinformatics?

The structure adopted by a protein chain, and in turn its function, is determined entirely by its amino acid sequence. However, the rules that govern how a given amino acid sequence folds are not yet known. Thus, it is not yet possible to fully predict the folded structure of a protein from an amino acid sequence alone. This remains one of the most important challenges facing bioinformatics. (See page 27)

1.17) What are some of the physical and chemical properties that are used to classify proteins into groups? Are these groups overlapping?

The functional properties of proteins are almost entirely due to the **side chain** of the amino acids. Each type of amino acid has specific chemical physical properties that are conferred due to the structure and chemical properties of the side chain. Amino acids can be classified into overlapping groups that share physical and chemical properties. Examples of classifying chemical and physical properties are:

- Size

- Electrical charge (hydrophobic and hydrophilic)
- Acidic and Basic
- Polar and Nonpolar

(See page 28-30)

1.18) What is meant by α-helix and β-sheet?

An **α-helix** is a type of stable right-handed helix. They are formed due to energetically favorable hydrogen bonding between atoms of the backbone of the protein chain. A helix is classified as a secondary structure in proteins. Another type of secondary structure is a β-strand, which is an extended strand of amino acids in a protein. These β-strands can align and bond with each other. A set of β-strands bonded together side by side form a **β-sheet**. (See page 33-34)

1.19) What are homologous proteins? When comparing proteins, where are most amino acids that change during evolution found?

Proteins that share a common ancestor are **homologous**. Most amino acids that change during evolution are found in regions that are not structurally or functionally important; this includes in many of the loops (or variable) regions. (See page 38)

1.20) What are globular and fibrous proteins?

Globular proteins are those that are roughly spherical when folded up. Globular proteins or proteins composed of multiple globular units perform most cellular functions. Fibrous proteins are rod- or wire-like in shape. Examples of fibrous proteins include the keratin of wool and hair, and the silk protein. (See page 41)

Problem #2

A single nucleotide addition followed by a single nucleotide deletion approximately 20 bp apart in the DNA caused a change in the protein sequence from sequence A to sequence B.

Sequence A: His (H) – **Thr (T)** – Glu (E) – **Asp (D)** – Trp (W) – **Leu (L)** – His (H) – **Gln (Q)** – Asp (D)
Sequence B: His (H) – **Asp (D)** – Arg (R) – **Gly (G)** – Leu (L) – **Ala (A)** – Thr (T) – **Ser (S)** – Asp (D)

1) **Which nucleotide has been added, and which nucleotide has been deleted?**

The amino acids in the first and last amino acids (i.e. Histidine and Aspartic Acid respectively) are in the same location for both sequences. Hence the nucleotide addition and deletion occurred in the codons immediately following and preceding those locations (i.e. in the second and eighth amino acids) respectively.

Addition: In sequence A, the second amino acid is Threonine (Thr), which corresponds to codons ACU, ACC, ACA, and ACG. In sequence B, the second amino acid is Aspartic Acid (Asp), which corresponds to codons GAU and GAC. For this change to be possible, there needs to be an insertion of a Guanine (G) nucleotide between the third and fourth nucleotides (i.e. in the fourth position) in sequence A.

Deletion: In sequence A, the eighth (second to last) amino acid is Glutamine (Gln), which corresponds to codons CAA and CAG. Moreover, in sequence A, the seventh (third to last) amino acid is Histidine, which corresponds to codons CAU and CAC. Since there was an insertion in sequence A before the eighth codon and the deletion is in the eighth codon, then the last nucleotide of the seventh amino acid in sequence A (i.e. either U or C) forms the first nucleotide in the eighth codon in sequence B.

In sequence B, the eighth amino acid is Serine (Ser), which corresponds to codons AGU, AGC, UCU, UCC, UCA, and UCG. As explained previously, the first nucleotide in sequence B's is from the last nucleotide in sequence A's seventh codon. Hence the eighth codon in sequence B must begin with a U since none of the codons for Serine begin with a C. Using the information from sequence A, valid combinations for the eighth codon in sequence B are:

- UAA (Stop) and UAG (Stop) if the 22nd nucleotide in sequence A (or 23rd nucleotide in sequence B including the insertion) is deleted.
- UCA (Serine) and UCG (Serine) if the 23rd nucleotide in sequence A (or 24th nucleotide in sequence B including the insertion) is deleted.
- UCA (Serine) if the 24th nucleotide in nucleotide in sequence A (or 25th nucleotide in sequence B including the insertion) is deleted.

Hence the deletion was either in the 23rd (A) or 24th (A or G) nucleotide of sequence A (i.e. the 24th or 25th nucleotide in sequence B).

2) **What are the original and the new mRNA sequences?**

The genetic code is **degenerate**; hence in most cases (including this one), it is not possible to unambiguously deduce a nucleic acid sequence from a protein sequence. Table 1 lists all the possible nucleotide/codon combinations for each amino acid sequence (A and B). Table 2 is a simplification of Table 1 that combines information across the two sequences to eliminate impossible combinations; Table 2 is the possible amino nucleotide chains for sequences A and B. Note for both tables that if there are multiple amino acid triplets/codons available for any amino acid, all combinations are listed in separate rows.

Set of Possible Codons for Sequence A without Simplification Using Sequence B:

His (H)	Thr (T)	Glu (E)	Asp (D)	Trp (W)	Leu (L)	His (H)	Gln (Q)	Asp (D)
CAU	ACU	GAA	GAU	UGG	UUA	CAU	CAA	GAU
					UUG			
	ACC				CUU			
CAC	ACA	GAG	GAC		CUC	CAC	CAG	GAC
					CUA			
	ACG				CUG			

Set of Possible Codons for Sequence B without Simplification Using Sequence A:

His (H)	Asp (D)	Arg (R)	Gly (G)	Leu (L)	Ala (A)	Thr (T)	Ser (S)	Asp (D)
CAU	GAU	AGA	GGU	UUA	GCU	ACU	AGU	GAU
		AGG		UUC			AGC	
		CGU	GGC	CUU	GCC	ACC	UCU	
CAC	GAC	CGC	GGA	CUC	GCA	ACA	UCC	GAC
		CGA		CUA			UCA	
		CGG	GGG	CUG	GCG	ACG	UCG	

Table 1 – mRNA Nucleotide Sequences A and B Without Simplification

Sequence A Maximally Simplified:

His (H)	Thr (T)	Glu (E)	Asp (D)	Trp (W)	Leu (L)	His (H)	Gln (Q)	Asp (D)
CAU	ACC	GAG	GAU	UGG	CUA	CAU	CAA	GAU
CAC	ACA		GAC				CAG	GAC

Sequence B Maximally Simplified:

His (H)	Asp (D)	Arg (R)	Gly (G)	Leu (L)	Ala (A)	Thr (T)	Ser (S)	Asp (D)
CAU	GAC	AGA	GGA	UUG	GCU	ACA	UCA	GAU
CAC		CGA		CUG			UCG	GAC

Table 2 – mRNA Nucleotide Sequences A and B Maximally Simplified

Problem #3

An ortholog are genes in different species that evolved from a common ancestral gene. A homolog is a gene related to a second gene by descent from a common ancestral sequence.

Part A:

1) When was the Homo sapien Pax-6 gene (accession# P26367) last updated?

The sequence was last updated on July 15, 1999, and the annotation was last updated on October 1, 2014.

2a) Which protein in the human dataset is closest to the zebrafish Pax6?

It is the Pax-6 protein. Its full name is Aniridia type II protein and is also referred to as Oculorhombin.

2b) How long is this protein?

It is 422 amino acids in length.

3) What is the degree of similarity between the query and the hit?

The sequences are 96% similar (404 identities out of 422).

4) What is the probability that the similar between the query and the hit occurs only by chance?

The E-value is 0% (minimum value, or underflow). Hence, the probability the similarity is due to chance is 0 (i.e. underflow). This means the genes are orthologs.

5) In the first alignment, what do you think the stretches "---" represent?

These are gaps in the protein sequence. In this case, they are three amino acids that are in the subject protein that are not in the query protein. There are four total gaps shown for this gene pairing.

6) Look at the second and third most relevant hits. How similar are they to the zebrafish Pax6 protein sequences?

Pax-3 is 50.95% similar (134 identities out of 263) while Pax-4 is 50.90% similar (141 identities out of 277).

Part B:

7) What is the mission of UniProt?

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality, and freely accessible resource of protein sequence and functional information.

8) In which tissues is the protein (P26367) found?

The tissue specificity of this protein is: fetal eye, brain, spinal cord, and olfactory epithelium.

9) **How many diseases are described in relation with defects in the Pax6 protein? Which organs are affected by the mutations in the Pax6 gene?**

Eight diseases are described in the heading "Involvement in disease heading." They are:

a. *Aniridia (AN)* – A complete absence of the iris or extreme iris hypoplasia (under/incomplete development). Aniridia is also associated with macular and optic nerve hypoplasia, cataract, corneal changes, nystagmus (involuntary eye movement), low visual acuity, and to a secondary extent glaucoma. All of these occur in the human eye.

b. *Peter's Anomaly (PETAN)* – Affects the eye's cornea and iris.

c. *Foveal Hypoplasia 1 (FVH1)* – Affects the eye.

d. *Keratisis Hereditary (KERH)* – Affects the eye's cornea and its stroma.

e. *Coloboma of Iris Choroid and Retina (COI)* – Leads to eye deformities.

f. *Coloboma of Optic Nerve (COLON)* – Affects the eye

g. *Bilateral Optic Nerve Hypoplasia (BONH)* – Causes defects in the eye as well as brain defects, cerebral anomalies, pituitary dysfunction, and structural abnormalities in the pituitary.

h. *Aniridia, Cerebrallar Ataxia, and Mental Deficiency (ACAMD)* – Leads to a partially rudimentary iris, cerebellar impairment, and mental retardation.

Organs affected by defects in this gene include the eye, optic nerve, brain (BONH and ACAMD), and pituitary (BONH).

Nine diseases as listed under Orphanet (with some being duplicates from above). They are:

a. *Aniridia*
b. *Autosomal Dominant Keratitis*
c. *Foveal Hypoplasia*
d. *Isolated Aniridia*
e. *Isolated Optic Nerve Hypoplasia*
f. *Morning Glory Syndrome*
g. *Ocular Coloboma*
h. *Peter's Anomaly*
i. *WAGR Syndrome*

10) **What is the function of the PAX6 gene?**

This gene is a transcription factor involved in the development of the human eye, nose, central nervous system, and pancreas. It is required for the differentiation of pancreatic islet alpha cells. It also regulates the specification of the ventral neuron subtypes.

11) **How many bibliographic references are quoted in this entry? Which paper describes the evolutionary conservation of the Pax6 gene?**

37 references are cited. The paper "*Genomic structure, evolutionary conservation and aniridia mutations in the human PAX6 gene*" discusses this gene's evolution conservation.

Part C:

12) Name the two conserved domains found in PAX6, and write down their start and end positions.

The two conserved domains are:

a. PAX (Paired Box Domain) which starts at position 4 and continues to position 128.
b. HOX (Homeodomain) which starts at position 210 and continues to position 272.

13) Is the function of the paired box domain known?

No, and no description of domain is provided.

14) Are paired box genes found in plants and/or in fungi?

No this gene is found only in animals, primarily mammals as well as the zebrafish.

15) What is the function of the HOX domain?

HOX is a DNA binding factor that is involved in the transcriptional regulation of key development processes.