Homework Two

Please hand in the solutions to the following problems on Thursday, October 3, 2013. **Hand in a hard copy (required)** and a CD (optional) or USB key (optional) containing your solutions.

Problem 1 Problems from the textbook – Chapter One and Chapter Two

1) a) What are promoters?
   b) Bacterial promoters typically occur immediately before the position of what site?
2) What is one of the main problems in finding promoters in DNA sequences?
3) a) Is the prokaryote terminator sequence more variable than the promoter sequence?
   b) Is the prokaryote terminator sequence usually included in genome annotations?
4) a) What are activators and repressors?
   b) Why are they of critical biological importance?
5) a) What is the most important core promoter sequence in genes transcribed by RNA polymerase II?
   b) Where is it located?
6) What is the major difference between eukaryotes and prokaryotes in terms of their transcription and translation processes?
7) a) What is the role of the spliceosome?
   b) What does it consist of?
8) a) What is meant by "alternative splicing"?
   b) Is it quite common in the genes of humans and other mammals?

9) a) What is the Shine-Dalgarno sequence?
   b) What is its typical consensus sequence?
   c) Where does it occur?
10) What is an operon?
11) How do viruses replicate?
12) Give an example of an unusual feature that is found in some viral genomes but not in cellular genomes.
13) What are plasmids?
14) On what process does the fate of a mutation (to be lost or to be retained) depend?

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

-------------------------------- End of Chapter One ---------------------------------

16) What is one of the challenges facing bioinformatics?

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

18) a) What is meant by α-helix?
    b) What is meant by β-sheet?

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

20) a) What are globular proteins?
    b) What are fibrous proteins?

Problem 2

A single nucleotide addition followed by a single nucleotide deletion approximately 20 bp apart in the DNA causes a change in the protein sequence from sequence a to sequence b:

a) His – Thr – Glu – Asp – Trp – Leu – His – Gln – Asp

b) His – Asp – Arg – Gly – Leu – Ala – Thr – Ser – Asp

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

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

Problem 3

In this problem we will explore some current bioinformatics tools. We will use different databases and other Internet resources to learn about the PAX6 gene in different organisms.

A) Using BLAST at NCBI for finding orthologs

We are going to use BLAST at NCBI to find a human ortholog of the zebrafish Pax6 protein

- Go to the main page of NCBI: http://www.ncbi.nlm.nih.gov
- Click on "BLAST" under "Popular Resources" on the right hand side of the page

- In the new page, click on "protein blast" under "Basic BLAST"

We are now ready to enter the pax6 protein of the zebrafish.

- Retrieve the pax6 protein of the zebrafish by copying it from www.cs.sjsu.edu/faculty/khuri/CS123A/zebrafish_pax6_protein.fasta
- Paste the sequence in the BLAST window under "Enter Query Sequence"

The pax6 sequence can now be compared to various datasets of protein sequences in various databases

- Scroll down and choose "UniProtKB/Swiss-Prot(swissprot)" from the dropdown window next to "Database"
- Start typing "Homo sapiens" in the "Organism" window (to limit the search to human proteins) and choose "Homo sapiens (taxid:9606)"
- Make sure that "blastp (protein-protein BLAST)" is chosen under "Program Selection"
- Choose "Show results in a new window"
- Click on the blue "BLAST" button to start the search

The graphic representation under "Show Conserved Domains" in the new page shows that two conserved domains have been detected in the Pax6 protein: PAX domain (**pa**ired bo**x** domain) and homeodomain (or homeobox domain).

The graphical view (under "Color key for alignment scores") shows an overview of the results where the human sequences detected in SwissProt by the BLAST search (the "hits") are aligned with the zebrafish Pax6 protein (represented as a red scale bar next to "Query"). The "Color key for alignement scores" shows the degree of similarity between the Query sequence and the results. Below the graphical overview, the detailed list of the sequences producing significant Alignments is given.

In the "Descriptions" section (under "Sequences producing significant alignments:"), you can examine the database matches in more details. Each database sequence has an identifier string, an accession number shown as a blue link. For example: P26367.2 is an *accession number*:

- Click on P26367.2 (under "Accession) and you will be taken to the database entry whose accession number is P26367.2. Note that it is the human Pax6.

1) When was the last time this record was updated?

- Go back to the BLAST output page.

To the right of the accession number you will find a one-line, short description of the protein, and the Total score that shows the level of similarity to the QUERY sequence and the E value assigned to each "hit". The Total score and E value are special statistics that measure the degree of similarity between two sequences. Basically, the higher the Total score, the greater the similarity between the two sequences. The lower the E value, or the closer it is to zero, the more "significant" the match is.

Below the list of hits, the individual "Alignments" for each hit are shown. For each alignment, the query sequence ("Query" – the zebrafish protein) is shown at the top and the hit ("Sbjct" – the human protein returned by BLAST) underneath it, with the position of the amino acids indicated on the right and left of the alignment.

Go to the first alignment and answer the following questions:

2) a) Which protein in the human dataset is the closest to the zebrafish Pax6?
   b) How long is this protein?
3) What is the degree of similarity between the query and the hit?
4) What is the probability that the similarity between the query and the hit occurs only by chance?
5) In the first alignment, what do you think the stretches "---" represent?
6) Look at the second and third most relevant hits. How similar are they to the zebrafish Pax6 protein sequence?

The human sequence most similar to our QUERY is the protein Pax6. It has the highest Total score and the lowest E value in the list of hits. It is the human ORTHOLOG of the zebrafish Pax6 protein. Its accession number in the SwissProt database is P26367.2. Let us study the information available about it in several relevant biological databases.

B) The SwissProt Database

By going to http://www.uniprot.org/, you would have accessed the "Swiss-Prot Protein knowledgebase" database hosted by the "Swiss Institute of Bioinformatics". Note that the Protein Knowledgebase database is one of several UniProt (Universal Protein Resource) databases.

7) What is the mission of UniProt?

- Type the accession number “P26367” in the “Query” field at the top of and click on “Search”.

The result of your search is a page for the human Pax6 protein.
The page contains information grouped in categories [Name and origin], [Protein attributes], [General annotation (Comments)], [Ontologies], [Binary interactions], [Alternative products], [Sequence annotation (Features)], [Sequences], [References], [Web resources], [Cross-references], [Entry information], and [Relevant documents] easily identified by light blue headers.
Scroll up and down the page to study the different categories of information available and answer the following questions.

8) In which tissues is the protein found?
9) How many diseases are described in relation with defects in the Pax6 protein? Which organs are affected by mutations in the PAX6 gene?
10) What is the function of Pax6?

11) Scroll down to the "References" section of the SwissProt Report. How many bibliographic references are quoted in this entry? Which paper describes the evolutionary conservation of PAX6 gene?

C) Studying the architecture of proteins with SMART

Let us use SMART to study the 3-Dimensional structure of Pax6.

- Go to SMART at http://smart.embl-heidelberg.de/

SMART (Simple Modular Architecture Research Tool) is based on the principle that proteins are modular in nature, i.e. they contain functional modules (or domains) that are detectable because they are conserved between species. SMART allows the identification of protein domains and the analysis of domain architectures. More than 500 domain families found in signaling, extra-cellular and chromatin-associated proteins are detectable. These domains are extensively annotated with respect to phylogenic distributions, functional class, TERTIARY STRUCTURES and functionally important residues.

- Click on the blue “SMART MODE:” under “Normal mode”
- Type “P26367” in the ”Sequence ID or ACC” window
- Click on “Sequence SMART”

The resulting page, “Domains within *Homo sapiens* protein PAX6_HUMAN (P26367)”, shows the two conserved domains detected by SMART: the PAX domain and the HOX domain. The figure also shows low complexity regions (LCRs) as pink bars.

Study the SMART result page and answer the following questions.

12) Name the two conserved domains found in PAX6 and write down their start and end positions.
13) Is the function of the paired box domain known?
14) Are paired box genes found in plants? In fungi?
15) What is the function of the HOX domain?