BBL 434 Bioinformatics
Minor Test 2, Mar 24, 2018
Max Marks = 20

15/20

(Write your answers as clearly and precisely as possible in the space provided)

1. In a genome with average GC-content 40%, what is the expected frequency of occurrence of the tetranucleotide AACG? (1 mark)

$\frac{1}{4} \times \frac{1}{4} \times 0.4 = 0.025$

GC → 40%
AT → 60%

A → 30%, T → 30%, G → 20%, C → 20%

$\frac{3 \times 3 \times 2 \times 2}{10^4} = 0.0036$

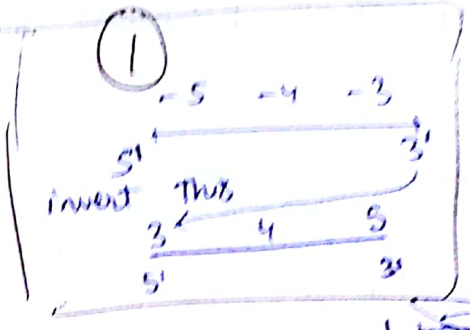
2. Two species are found to share a cluster of 8 genes, but the genes are in different orders in the two species. The orders are represented by signed permutations as given below: (2 marks)

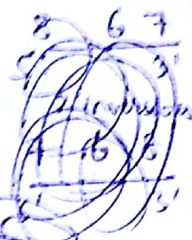
Species X 1, 2, 3, 4, 5, 6, 7, 8
Species Y 1, 2, -5, -4, -3, 8, 6, 7

(B)

Choose one of the following correct answers and explain why.

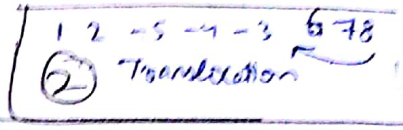
The transformation between the two gene orders
(A) cannot be achieved by inversions alone.
(B) can be achieved by one translocation and one inversion.
(C) can be achieved by three inversions.
(D) requires six separate genome rearrangement events.





1 + 2 = 3 inversions

one inversion for inverting
one translocation for translocating



3. If two distantly related protein sequences of roughly equal length are to be aligned, which one of the following would work best and why? (2 marks)

(A) PAM250 matrix and default gap penalty
(B) PAM250 matrix, default start gap penalty and 10 times less of gap continuation penalty
(C) BLOSUM62 matrix with high gap penalty
(D) BLOSUM62 matrix and default gap penalty

We need high gap penalty to align distantly related proteins. Small penalty would lead to poor alignment.

distant seq.
→ higher PAM
Lower BLOSUM
We need flexibility in alignment
∴ less gap penalty
∴ PAM 250, BLOSUM 62
(C) as high BLOSUM has specially designed for distantly related seq.

4. A DNA sequence of length x (at least) is needed to be in order for it not to be found by chance more than once in the human genome. The value of x is (1 mark)

$4^x = 3.3$ billion

x = 16.

5. Which one of the following terms describes SNP's that result in an amino acid change in the protein? Why? (2 marks)

(A) synonymous change in the non-coding region
(B) synonymous change in the coding region
(C) non-synonymous change in the non-coding region
(D) non-synonymous change in the coding region

Since there is change of A.A. → non-synonymous and should be in coding region.
Synonymous change would not alter A.A.

6. You have been asked to use bioinformatics tools to annotate the putative function of a protein using its amino acid sequence information. Which one of the following methods will be most appropriate for this task? Why? **(2 marks)**

(A) use the BLAST program to compare sequence of your protein with entries in PDB
(B) use the BLAST program to compare sequence of your protein with entries in nr database of NCBI
(C) get the corresponding DNA sequence from GenBank and calculate GC content
(D) predict its secondary structure

Since we are interested in function only, ∴ use local alignment tool BLAST, Checking in PDB will not help in annotation of function so we have to search in nr DB of NCBI. Also it helps in finding novel gene by this method.

---

7. Describe these terms very important while performing annotation (a) ORF (b) UTR (c) mRNA processing **(1 mark)**

ORF: Open reading frame Codon which is translated

UTR: Un-translated region. Region of DNA which is not translated

mRNA processing: Processing of mRNA to remove introns.

5' cap
3' poly A

---

8. We discussed in the class that most SNPs in the human genome do not have an effect on phenotype; ie., they are selectively neutral. Why do you expect this to be true? **(1 mark)**

(i) SNP might occur in region of DNA which does not code for protein or phenotypical characters.

(ii) Due to redundancy in protein coding, change in nucleotide still forms same protein.

---

9. We discussed a paper that reported genome-wide association (GWA) study to find SNPs associated with increased risk to cardiac myopathy. If you are now given an entire dataset of say, 100,000 SNPs from 10,000 individuals to find SNPs with increased susceptibility to the same clinical condition, how will you conduct the GWA study? Will you have all the SNPs as one large group or split into two groups and study? Explain **(2 marks)**

Split into 2 grps and plot Manhattan plot. From the plot, we will get information about association in different chromosomes, higher is the peak, higher is the association for the clinical condition. Each dot represents SNP

---

10. What are ESTs? How is this helpful in the closure phase of genome assembly? **(1 mark)**

EST: Expressed sequence tag - These are sub-sequence of cDNA and are helpful in assembly of short cDNA seq. to generate full length. assembly. How ?

---

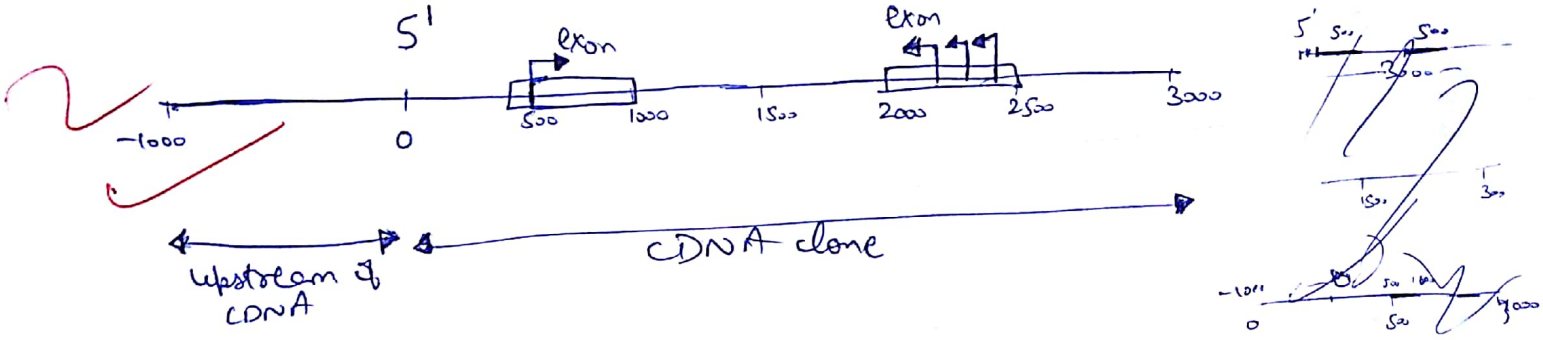
11. When I summarized the findings of the *Nature* paper that described the draft sequence of the human genome, I told you that the paper had indicated that there only ~ 20,000 – 25,000 genes in humans. However, I also told you that it is likely that there are more than 25,000 different proteins made in humans. How can you explain this? **(1 mark)**

Human in Eukaryote and in the DNA there are large number of introns (non-coding). During mRNA processing, splicing event occurs due to and phenomena of alternate splicing one exon combines with other in different combinations. This alternate splicing leads to many more no of proteins.

12. My research lab in the department (DAILAB) is actively conducting research projects on integrated drug screening for stress, aging and cancer intervention with prime focus on elucidation of functional mechanisms of natural drugs. We are studying a gene called "Gene X" that is involved in the regulation of secondary metabolite biosynthesis in an Indian medicinal plant called *Ashwagandha*. The protein encoded by Gene X appears to be a good candidate for a possible "switch" protein that determines whether a metabolite A or B is formed in the metabolic pathway. Now, we have managed to obtain a partial cDNA sequence (3000 nucleotides long) as well as a clone containing the corresponding genomic DNA sequence (4000 bp long). The genomic clone includes exactly 1000 additional nucleotides upstream (on the 5' side) of the 5' end of the cDNA clone sequence. When the Gene X sequence was BLASTed against the existing database, it was found that Gene X has ~ 90% sequence identity with part of a larger protein (Ash-P) encoded by a gene previously cloned and sequenced in *Ashwagandha*. The function of the *Ashwagandha* protein (Ash-P) is not known, but mutations in the Ash-P gene result in the metabolite A formation rather than metabolite B. The Gene X appears to have at least two exons (each about 500 bp). There is a candidate for a translational start signal (ATG) within one exon (at position 1500 on the genomic clone) and a cluster of 3 potential translational STOP codons within the other exon (beginning at position 3000 on the genomic clone).

Question 12-A (2 marks)

i) Draw a line to represent the map of the cDNA clone
ii) Label the 5' end of the cDNA
iii) Label the positions of the translational START and STOP codons
iv) Draw two boxes to represent the two exons mentioned above (at their approximate locations)



Question 12-B (1 mark)

Would you use any of the gene prediction methods to predict the exact intron/exon boundaries in the Gene X? Explain.

Physio-chemical method can be used. It will help in estimating the binding positions of different proteins on DNA. Hence, we can approximate boundary of exon/intron.

Question 12-C (1 mark)

You are now asked to check the genomic DNA sequence between 1-1000 for potential additional exon(s) not present in the cDNA clone. You find a region in the genomic DNA sequence (between positions 500 and 800) that is strongly predicted to contain another ORF. What are the criteria you would assume/choose to make such a prediction?

To find ORF upstream, we will have to make different frames like frame +1, frame +2, frame +3, frame -1, frame -2, frame -6. From these 6 frames will will run along the sequence finding the start codon in 500-800 region upstream of cDNA. If we find a start codon and correspondingly stop codon downstream to that then we can say that there is ORF.

0.5