BBL434 – Bioinformatics **Major Test** (May 10, 2018) **Max Marks = 30**

1. Using the DNA alignment shown below, answer the following questions: **(2 marks)**

```
ATC---TACGTG
ATCTTCTACCTG
ATCAATTACGTC
ATA---TATGTC
ATT---TATGTG
```

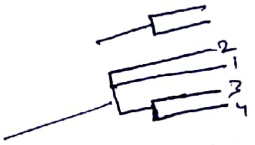
a) In the alignment shown above, which positions are most conserved?

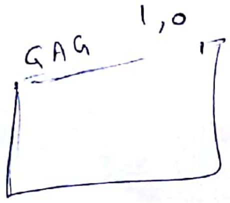
b) If you have to assume that the first position is the first position of a codon (i.e. assume that translation starts from the first position), can you give a biological reason for some positions being more conserved than others?

2. Use the Needleman and Wunsch method to find the distances between the following four sequences, and then use a simple distance matrix method to construct a phylogenetic tree for them. Use a unitary scoring matrix. **(2 marks)**

→ Global

Seq 1 ...GAGTCT...
Seq 2 ...CCGGGT...
Seq 3 ...ATATCA...
Seq 4 ...AGATCA...





3. With a network diagram, show at least 10 different databases that are related to GenBank **(1 mark)**

4. Can you suggest ONE database for the following searches? **(1 mark)**

a) I want to know the annotation of human genome Genbank

b) I want to determine what protein domains my novel protein contains. Uniprot

5. What is the use of the Pfam database and what is the basis of the database? **(2 marks)**

6. If you are provided with a set of homologous sequences (same gene from different species), and you are asked to search in GenBank for finding same gene from more distantly related species, how will you do the database search? **(2 marks)**

↳ Blosum

a) Is it best to use a single sequence as the query?

b) Is it best to use a profile or PSSM in your search?

Explain your answer for both the above situations.

7. Can you illustrate and schematically show how a dotplot is likely to look for two proteins, p1 and p2, whose only significant similarity is a common domainfamily, and where the domain is occurring once in p1 and in three instances in p2? **(2 marks)**

8. You have a set of divergent sequences. Suggest a way by which you can detect weak similarities in these sequences. Can you detect motifs in these sequences?. Describe the methods you would use and what you would expect from them? By chance if you find sequences with E value = 1.0 in your search for the similar sequences, would you consider them or reject them or do so only under certain circumstances? Explain. **(2 marks)**

9.

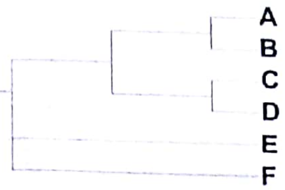
a) Why do we use dynamic programming algorithms for pairwise sequence alignment problems, but not for multiple pairwise alignment?

b) Compare the use of the affine gap penalty with the constant gap penalty. **(2 marks)**

g+ r(l-1)   g

10. Using the guide tree given below, describe the order in which the different sequences could be aligned. **(1 mark)**

(A,B), (C,D) (E),(F)



11. Imagine you are working with an unusual protein. It has very weak similarity to anything in the public databases. To ensure that you keep up to date with the research in the particular area, you "blast" your protein against the nr database at NCBI now and then. When you did your last search, a few months ago, the top hit had E value 0.1. When you do the same search again now, the very same top hit has E value 0.2, even though the score is the same. How come? **(2 marks)**

12. You do a BLAST search to predict the function of a human query protein on two different internet sites that has a BLAST search tool. The alignments of best hits are as follows: **(2 marks)**

```
>gb|AAC60279.1| ubiquitin/ribosomal protein [Gallus gallus]
Length=156
 Score = 47.8 bits (112), E-value = 1e-04
 Identities = 47/95 (49%), Positives = 50/95 (52%), Gaps = 36/95 (37%)
Query  1    IRKETTLHKVLRLWGGAYKDXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXKKKSY  60
            I+KE+TLH VLRL GGA K                                    KKKSY
Sbjct  61   IQKESTLHLVLRLRGGAKK-----------------------------------RKKKSY  85
Query  61   TMPXXXXXXXXXXX-AVLPYYKIDEYGKISRFRRE  94
            T P            AVL YYK+DE GKISR RRE
Sbjct  86   TTPKKNKHKRKKVKLAVLKYYKVDENGKISRLRRE  120
```

```
>sp|P42568|AF9_HUMAN Protein AF-9 (Myeloid/lymphoid or mixed-lineage leukemia associated)
Length=568
 Score = 68.9 bits (167),  E-value = 5e-11
 Identities = 40/52 (76%), Positives = 44/52 (84%), Gaps = 0/52 (0%)
Query  21   SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSKKKSYTMPKKNKHKHKK  72
            SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS   S++ P K   +HK+
Sbjct  154  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSTSFSKPHKLMKEHKE  205
```

a) Which hit is statistically more significant? Explain.

b) What is the reason for the difference between the two BLAST results? Which of the two hits do you think is most likely to be a true homolog? Explain

13.

a) You have determined the genome sequence of a bacterium. How can you use BLAST to identify protein-coding genes in this genome if we only have access to protein sequence databases?

b) A blastp search has not returned any hits at all. Would it be useful to do a PSI-BLAST using the same settings as the original blastp? **(2 marks)**

→ Pribnow ~~box~~

14. Describe the DNA sequence features that can be used to identify a protein coding gene in prokaryotes. **(1 mark)**

↳ Promoters Terminators

15. Discuss the similarities and differences between the prediction of genes and the prediction of regulatory motifs, for example, in terms of the intrinsic difficulties that we face in each approach, in terms of computational techniques to detect signals and variations in the composition of the sequences that must be analyzed, in terms of comparative genomics, etc. **(2 marks)**

16. The number of occurrences of dinucleotides in the genome of Zika virus has been the following: **(2 marks)**

aa	ac	ag	at	ca	cc	cg	ct	ga	gc	gg	gt	ta	tc	tg	tt
1108	720	890	708	901	523	261	555	976	500	787	507	440	497	832	529

Moreover, for this virus, the frequencies of the nucleotides are as follows: a (0.3191430), c (0.2086633), g (0.2580345), t (0.2141593). Let us assume that we are specifically interested in the dinucleotides **tg** and **cg**. Are they over or under-represented? Explain.

0.214 0.258 x

a+t 1

tc 2 1

17. Explain the nature of a Genome-Wide Association Study (GWAS). Describe the principles features of the GWAS studies as discussed in the class for risk reduction in human cardiovascular disease. **(2 marks)**