15/20

BBL434 – Bioinformatics
Minor Test 1, Feb 05, 2017
Max Marks = 20

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

---

1. BLAST search of a query protein sequence against **nr** database of NCBI gave you 4 hits with the E-values as mentioned below. Which one of these hits would be most similar to your query sequence? Why? **(2 marks)**
(a) 1e-230 (b) 9e-22 (c) 0.003 (d) 0.0000001

(a) will represent hit most similar to query seq. Because lower the value of 'E' higher is the result significant ie. there are very less chances of matching to a random seq. in database than others.

2. One particular sequence alignment program is using a non-standard scoring matrix instead of BLOSUM62. All the substitution scores on the diagonal of this matrix are +2 and scores at all other positions of this matrix are -1. In other words, identical substitutions are given a score +2 and non-identical substitutions are given a score -1. It aligns GFIRIGKTYL and GFVKDGRTYL over their entire length without any gaps. What would be the score of this alignment and percentage identity? **(2 marks)**

```
G F I R I G K T Y L
| |   |   | | |
G F V K D G R T Y L
```

% Identity = $\frac{6}{10} \times 100 = 60\%$

Score = +2 +2 −1 −1 −1 +2 −1 +2 +2 +2
= 8

---

3. (a) Use the appropriate dynamic programming algorithm to **LOCALLY** align the sequences ACGGTTG and ACGTTG. Use +2 for a match, -2 for a mismatch and -1 for a gap. (b) Show **ALL** optimal alignments. (c) What is the name of the algorithm? (d) What is the score of the/all optimal alignment(s)? **(5 marks)**

	-	A	C	G	G	T	T	G
-	0	0	0	0	0	0	0	0
A	0	2	←1	←0	←0	←0	←0	←0
C	0	↑1	↖4	←3	←2	←1	←0	←0
G	0	↑0	↑3	↖6	↖5	←4	←3	↖2
T	0	0	↑2	↑5	↖4	↖7	↖6	↖5
T	0	0	↑1	↑4	↖3	↖6	↖9	←8
G	0	0	↑0	↖3	↖6	←5	↑8	↖11

(b) (i) A C G G T T G
A C − G T T G
Score = 11

(ii) A C G G T T G
A C G − T T G
Score = 11

(c) Name - Smith – Waterman Algorithm

(d) optimal score = +11

---

4. The conserved sequence motif for the family of AMP binding proteins is [LIVMFY]-X(2)-[STG]-[STAG]-G-[ST]-[STEI]-[SG]-X-[PASLIVM]-[KR]. You are given sequence of a 34 amino acid stretch starting from the first residue of the motif. Which one of the following proteins is likely to have AMP binding function ? **(1 mark)**

a) LIVMFYNGSTGSTAGGSTSTEISGAPASLIVMKR
b) MAGTAGSEGYIRHHCSCDGSYPFDVITVNGKTYL
c) LIVMFYNGSTGSTAGGSTSTEISGAPASLIVMDE
d) LSSTAYTTSALKAAAAAAAAAAAARRRRRRRRRR

Rule out other 3

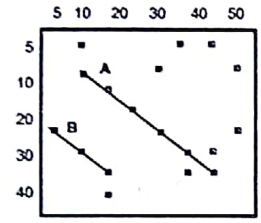
5. Which one of the following computational methods would you choose for identifying a conserved sequence **motif** in a family of AMP binding proteins ? Why? Why would the other three be not suitable? **(2 marks)**

a) BLAST alignment of any two AMP binding proteins
b) Global alignment of any two AMP binding proteins
c) Multiple Sequence Alignment of a representative set of AMP binding proteins — by comparing large set of Blog seq. we can predict more accurately what seq. it is
d) BLAST alignment of nucleotide sequences of any two AMP binding proteins

2

- Will not choose (b) because it will also give us region having less similarity whereas we are interested in a particular region only
- will not choose (d) because there is redundancy in A.A coding and chances of random mutation in DNA are more than AA seq.
- Will not choose (a) bcoz from only 2 seq. we can't accurately predict conserved seq

6. Dot matrix analysis of the amino acid sequences of lambda phage cI (horizontal sequence) and phage P22 c2 (vertical sequence) repressors is shown below. Which does line A and line B indicate? **(1 mark)**



A → lambda phage cI
B → phage p22 c2

no answer ?

A represents similarity in 2 seq
B rep. repeats

7. What is the total score for alignment of sequence A with sequence B if you use the amino acid substitution matrix given below? **(1 mark)**

= 2 ✗

Sequence A	Tyr	Cys	Asp	Ala
Sequence B	Phe	Met	Glu	Gly
BLOSUM62 matrix value	3	-1	2	0

8. Suppose the BLAST search returned 100 hits. Of these, 17 were false positives and we knew that there were 165 sequences in the database that should have returned a hit with our sequence. How many false negatives were there, and what is the sensitivity and selectivity of BLAST in this instance? **(2 marks)**

0.5

False +ve = 17
True +ve = 83
False -ve = 165 - 83 = 82

Sensitivity = $\frac{TP}{TP+FP} = \frac{83}{83+17} = 0.83$ ✗

Selectivity = $\frac{TP}{TP+FN} = \frac{83}{83+82} = 0.503$

9. A sample genetic code is given below:

Amino Acid	Pro	Val	Gly	His	Asp	Tyr	Thr	Lys
Codon	CCN	GUN	GGN	CAY	GAY	UAY	CAN	AAR

1.5

An amino acid substitution matrix needs to be constructed for sequence alignment and analysis from evolutionary studies. Which amino acids would be considered more similar and why? **(2 marks)**

His & Thr. are more similar, since they have 2/3 nucleotides same. Reason not 100% true

10. You are interested to identify distant relatives of uncharacterized proteins and to get insights into the functions of this family of proteins. Which one of the following tools can reliably establish an evolutionary link between two proteins and align them even if they share little sequence similarity? Explain. **(2 marks)**

2

a) BLAST (b) PSI-BLAST (c) Needlman-Wunch algorithm (d) FASTA

PSI-BLAST would be a better tool, as we are interested in finding insights for functions ∴ local alignment is better than global. Also in PSI-BLAST we can iteratively ~~generate~~ create new profile until we get significant alignment, which is not possible in BLAST.