Indraprastha Institute of Information Technology Delhi (IIITD)

Department of Computational Biotechnology

BIO213 – Introduction to Quantitative Biology

Quiz-2 (April 04, 2024)

Total time: 1 hour Total marks: 40

Question 1. What are the four major factors that add to the complexity of fragment assembly problem? **(2 marks)**

Base call errors, chimeric fragments, contamination with vector, Unknown orientation, Repeated regions, Lack of coverage (any four – 0.5 mark each)

Question 2. Fill in the blanks in the following statements. **(10 marks)**

a. Hamiltonian path is a walk that visits each node (node/edge) once.

b. The overall height of a logo position reflects how conserved the position is, and the height of each letter in a position reflects the relative frequency of the residue in the alignment (height/width/amplitude/relative frequency).

c. An invariant column in a multiple sequence alignment will have less entropy (less/high).

d. The maximum height of the entire stack of residues in case of DNA sequence logo is 2 bits. (or $\log_2 4$)

e. Viterbi algorithm is used to find the most likely sequence of states in an HMM model.

f. Even if many pseudocounts are added in comparison to real sequence counts, the amino acid frequencies will not have any effect or influence. False (True or False)

g. Motifs are formed by simple combination of secondary structures.

h. The two most commonly used methods for DNA fingerprinting include RFLP (restriction fragment length polymorphism) and PCR (polymerase chain reaction). [Short forms will also do]

(1 mark for each correctly filled blank)

Question 3. What does this representation mean - <M-x-[AC]-x(2)-V-{ED}? **(2 mark)**

a. C terminal-Met-any-[Ala or Cys]-any-any-Val-{any but Glu or Asp}

b. N terminal-Met-[any but not Ala and Cys]-any-any-Val-{Glu or Asp}

c. N terminal-Met-[Ala or Cys]-any-any-Val-{any but Glu or Asp}

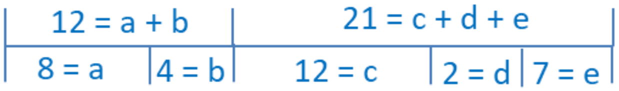
d. N terminal-Met-any-[Ala or Cys]-any-any-Val-{any but Glu or Asp}

Question 4. Given the following results of double digest experiments, try to find a solution. **(6 marks)**

Enzyme A: 12, 21
Enzyme B: 7, 4, 10, 12
Enzyme A+B: 8, 4, 12, 2, 7

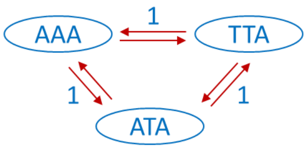
A+B =	8	4	12	2	7
	a	b	c	d	e

A	B
12 = a + b	7 = e
21 = c + d + e	4 = b
	10 = a + d
	12 = c



There can be other answers as well. So, check accordingly.

Question 5. Construct an overlap graph for F = (AAA, TTA, ATA). Find a shortest common superstring for this collection. **(6 marks)**



```
TTA
 AAA     → TTAAATA
   ATA

TTA
 ATA     → TTATAAA
   AAA
```

3 marks for correct graph; 3 marks for the correct superstring (one is also acceptable); if complementary fragments have been used to increase overlaps that is also acceptable. Can have multiple answers, so check accordingly.

Question 6. Which of the fragments is wrongly placed in the following alignment layout for finding the target sequence, provided ε = 0.4? **(5 marks)**

```
     GCAGGCAACTT
  CGGAA  CAACATCA
 TCGTCAGG   GTCCATT
--------------------
 TCGGCAGGCAACTTCATT
```

d_s (GCAGGCAACTT, TCGGCAGGCAACTTCATT) = 11
ε| f | = 0.4 X 11 = 4.4
Here, d_s (f, S) > | f |, therefore incorrect placement

d_s (CGG**A**A, TCGGCAGGCAACTTCATT) = 1
ε| f | = 0.4 X 5 = 2.0
Here, d_s (f, S) < ε| f |, therefore correct placement

d_s (CAAC**A**TCA, TCGGCAGGCAACTTCATT) = 1
ε| f | = 0.4 X 8 = 3.2
Here, d_s (f, S) < | f |, therefore correct placement

d_s (TCG**T**CAGG, TCGGCAGGCAACTTCATT) = 1
ε| f | = 0.4 X 8 = 3.2
Here, d_s (f, S) < | f |, therefore correct placement

d_s (**G**T**C**CATT, TCGGCAGGCAACTTCATT) = 2
ε| f | = 0.4 X 7 = 2.8
Here, d_s (f, S) < | f |, therefore correct placement

Question 7. Which of the following regular expressions would be matched by sequence MAVDEKC? **(2 marks)**

a. {MG}-A-x(2)-[DE]-[ILK]-x
b. M-A-x-[DE]-[ILK]-x
c. {HS}-A-x(2)-[DE]-[ILK]-x
d. M-A-x-[DE]-{LK}-C

Question 8. Calculate the contribution of each nucleotide for making the sequence logo using the column information as given below: **(4 marks)**

G
A
A
G
G
A
A
C
A
A

Give the value for: Entropy, Information index and contribution.

Entropy: -(0.3*$\log_2 0.3$ + 0.6*$\log_2 0.6$ + 0.1*$\log_2 0.1$) = 1.298 (1.5 marks)

Information: $\log_2 4$ – Entropy = 2 – 1.298 = 0.702 (1 mark)

Contribution: i) G = 0.3*Information = 0.3*0.702 = 0.2106 (0.5 mark)
ii) A = 0.6*0.702 = 0.4212 (0.5 mark)
iii) C = 0.1*0.702 = 0.0702 (0.5 mark)

(Answers in log are also acceptable)

Question 9. What is iterated in PSI-BLAST? **(1 mark)**

a. PSSM
b. Z-score
c. E-value
d. BLOSUM62

Question 10. List one advantage of MULTICONTIG over RECONSTRUCTION model of fragment assembly. **(2 marks)**

MULTICONTIG allows reconstruction of multiple contigs, which is not possible is RECONSTRUCTION. RECONSTRUCTION is designed to return the shortest common superstring.