58

RAJARATA UNIVERSITY OF SRI LANKA
FACULTY OF APPLIED SCIENCES, MIHINTALE

B.Sc. Four Year Degree in Information and Communication Technology
Forth Year – Semester I Examination – Oct/Nov 2015

ICT 4207 – BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

Answer **THREE** questions only Time allowed: 2 Hours

The use of a non-programmable electronic calculator is permitted.

1. Consider aligning short DNA sequence with a genome with huge number of base pairs. A student uses following algorithm to identify exact matching. If three are matching segments and return index of the first matching segment of DNA of the genome:

Length of DNA sequence A= m Length of Genome B = n

```
J=1;
FOR j=1..(n-m)+1
SegmentEqual=TRUE;
i ←1;
FirstValue=j
        WHILE  SegmentEqual AND i<m        DO
                IF NOT (A(i)= B(j)) Then
                        SegmentEqual=FALSE;
                End If
                i← i+1;
                j← j+1;
        END WHILE
IF (Segment Equal=TRUE) return FirstValue;
END FOR
```

59

(i) Mention Time Complexity of above algorithm based on m and n for the Worst Case.

[15 marks]

(ii) Same genome may be used for thousands of queries for different DNA segments .Based on this, propose a suitable indexing technique to represent the genome. Explain your indexing method using following DNA sequence as the genome:

AATCGGTCAG$

[20 marks]

(iii) Provide an efficient searching algorithm that can be used with above indexing technique mentioned in section (ii).

[15 marks]

(iv) Apply searching algorithm mentioned in section (iii) to search GTC sequence in Index created in section (ii).

[35 marks]

(v) Compare efficiency of this algorithm with the method discussed in section (i) for 3000,000,000 bp length genome with 2000bp DNA sequence.

[15 marks]

2. (i) a. Provide an equation that can be used to estimate uncovered bases in genome assembly

b. Following details regarding a genome assembly are given:

Genome size- 4,000,000,000 bp

Number of reads-4000000

Length of read-2000 bp

Calculate an estimation for number of uncovered bases in the genome.

[20 marks]

(iI) Explain Transitively- Inferable-Edges using a suitable example.

[20 marks]

60

(iii) Apply Overlap-Layout-Consensus assemble technique for following fragments of DNA.

S1	TTATCGGTTGA	S6	CGGTTGATGTTA
S2	TGTTAACATGTACGGCTGA	S7	GGCTGAAGTCC
S3	AGTCCGATAGGCTG	S8	GATAGGCTGGCTAATTTA
S4	GCTAATTTAGCGCTACGT	S9	GCGCTACGTGCATA
S5	GCATACCC	S10	TGTTAACATGTA

Table 1

[60 marks]

3. (i) Explain one advantage of De Bruijn graph assembly over Overlap-Layout-Consensus assemble technique.

[15 marks]

(ii) Discuss following properties of node of graph with suitable examples.

a. balanced

b. semi-balanced

c. connected

Library
Faculty of Applied Science
Rajarata University of Sri Lanka
Mihintale.

[15 marks]

(iii) What are the conditions needed to be satisfied by a directed connected graph to be Eulerian?

[10 marks]

(iv) Consider following DNA sequences as segments of one DNA string:

TGTTAACA TGTACGGC AACATGTA

a. Represent above sequences with De Bruijn Graph of 4-mers nodes (Edge represent 5 mer).

b. Apply De Bruijn Graph Assembly method to get the original DNA string.

[60 marks]

4. (i) Discuss three problems of Hidden Markov Model (HMM) using a suitable example.

[30 marks]

(ii) Provide an algorithm to calculate probability of happening specified sequence of observations when λ (π,A,B) is given.

[20 marks]

(iii) Apply HMM to predict most probable sequence for states of nucleotides (Intron or Exon) when observed DNA sequence is ATCC. λ is given in table 2 ,3 and 4.

Category	Probability
Intron	.995
Exon	.005

Table 2

	Intron	Exon
Intron	.99	0.01
Exon	.01	.99

Table 3

	A	T	C	G
Intron	.2	.2	.3	.3
Exon	.3	.3	.2	.2

Table 4

[50 marks]