110

RAJARATA UNIVERSITY OF SRI LANKA
FACULTY OF APPLIED SCIENCES

B.Sc. (Honors) in Information Technology
Fourth Year - Semester I Examination – January/February 2021

ICT 4307 – Bioinformatics and Computational Biology

Time: THREE (03) hours

- **Answer ALL questions in Part I.**
- **Answer THREE (03) questions in Part II.**
- **This is a closed book examination.**
- **This paper includes SEVEN (07) pages.**
- **Calculators are allowed.**

PART I

Write the English letter of the most appropriate answer from the given choices in your answer book. (5x20 marks)

1. Gene duplication has been found to be one of the major reasons for genome expansion in eukaryotes. In general, what would be the selective advantage of gene duplication?
   a. If one gene copy is nonfunctional, a backup is available
   b. Larger genomes are more resistant to spontaneous mutations
   c. Duplicated genes will make more of the protein product
   d. Gene duplication will lead to new species evolution

2. What would be a likely explanation for the existence of pseudogenes?
   a. Gene duplication
   b. Gene duplication and mutation events
   c. Mutation events
   d. Evolutionary pressure

3. The initial guide tree in progressive alignment is determine by as efficient clustering methods such as ________________.
   a. Back-tracking
   b. Divide-and-Conquer
   c. Both a and b are correct
   d. None of the above is correct

4. The process of finding relative location of genes on a chromosome is called

   a. Gene tracing

   b. Genome walking

   c. Genome mapping

   d. Chromosome mapping

5. ____________ is a graphical control element that presents a hierarchical view of information about phylogenetic analysis.

   a. Diagonal view

   b. Alignment view

   c. Tree view

   d. Domain view

6. ________ is a branch of which study the organismal variation in phenotype as it changes during its life span.

   a. Phenomics

   b. Metabolomics

   c. Transcriptomics

   d. Genomics

7. What is the statement that does NOT apply to the FASTA format?

   a. FASTA format can be used to store multiple sequences in tandem in a unique computer file.

   b. In the FASTA file, the definition line always starts with a greater than (>) symbol, and usually, no constraints restrict its length.

   c. In addition to the plain text file extension (.txt), there is no other file extension for a text file containing FASTA formatted sequences.

   d. FASTA sequence lines could contain line breaks or paragraph marks <¶> at the end of each line.

8. The most commonly used multiple aligners

   a. Align all the sequences at once

   b. Align the sequences 2 by 2 using a progressive approach

   c. Add the sequences 1 after the other to a larger alignment

   d. Add the sequences 3 by 3 using a progressive approach

9. Phylogenetic trees:

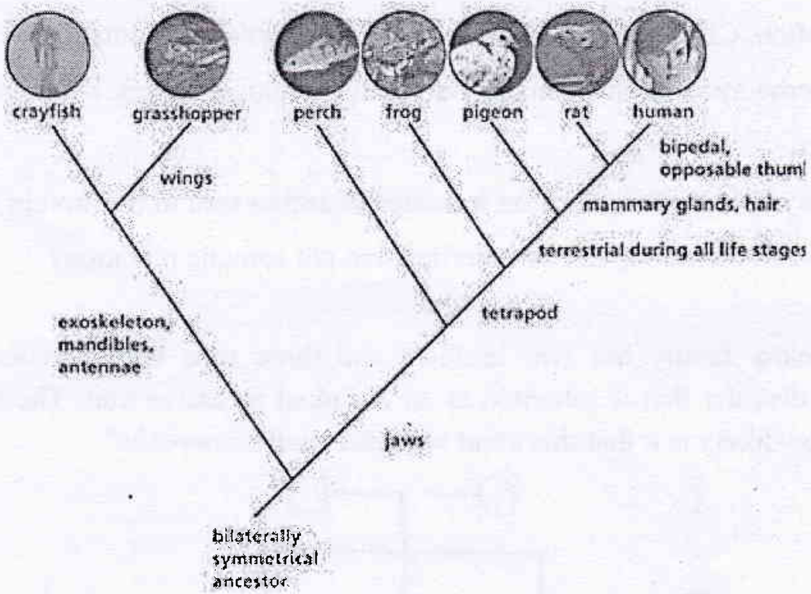
   a. Aim to show phenotype similarity

   b. Show the exact ages of particular species

   c. Aim to show evolutionary histories through common ancestors

   d. Are static never change

10. According to the diagram which trait do only mammals have?



a. Jaws

b. Mammary glands & hair

c. Terrestrial during all life stages

d. Bipedal & opposable thumbs

11. The term sequence _______ in two or more amino acid sequences is similar at the same position of the same amino acid along with similar physiochemical properties such as size, charge, and hydrophobicity.

a. Identity

b. Similarity

c. Homology

d. Xenology

12. Why do cells acquire and synthesis energy?

a. For cellular respiration and photosynthesis

b. Cells need food and carbon dioxide to live

c. To transport carbon dioxide and oxygen

d. To copy DNA and divide it into daughter cells

13. Which is incorrect?

a. Gene is a sequence of nucleic acids in DNA or RNA

b. Genome is the genetic material of an organism which consists of DNA

c. Chromosomes are DNA

d. Humans have 21 pairs of chromosomes excluding the sex chromosome

14. Select the incorrect statement about the mutations

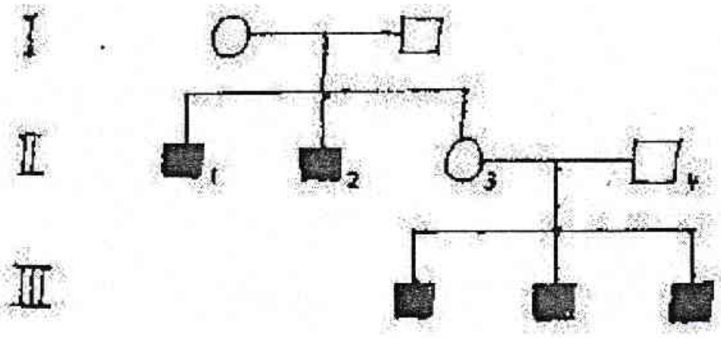
a. Radiation, Chemicals, and Viruses can be the causes of somatic cell mutations

b. Nonsense mutation happens by the change of single nucleic acid in a codon to STOP codon

c. Synonymous mutations cause a change of amino acid in the protein sequence

d. Germ cell mutations can be inherited, but not somatic mutations

15. In the below family has two brothers and three sons with classical hemophilia, a bleeding disorder that is inherited as an X-linked recessive trait. The sister is pregnant again. How likely is it that this child will also have hemophilia?



a. 100% for a son, 50% for a daughter

b. 50% for a son, zero for a daughter

c. 25% for a son, zero for a daughter

d. 50% for both sons and daughters

16. Select the correct statement about inheritance patterns

a. In x-linked recessive inheritance, the sex-chromosome is not involved

b. In codominant inheritance, both alleles from parents influence the genetic trait

c. Autosomal recessive diseases are depending on the sex

d. Mitochondrial inherited disorders are not depending on the sex

17. What are the correct steps of multiple sequence alignment?

a. Individual pairwise alignment, calculating the guide tree, repeat the pairwise alignment for a different set of sequence, get the consensus with the final sequence, complete the alignment

b. Individual pairwise alignment, repeat the pairwise alignment for a different set of sequence, calculating the guide tree, get the consensus with the final sequence, complete the alignment

c. Individual pairwise alignment, calculating the guide tree, get the consensus with the final sequence, repeat the pairwise alignment for the same set of sequence, complete the alignment

d. Individual pairwise alignment, calculating the guide tree, repeat the pairwise alignment for a different set of sequence, get the consensus with the final sequence, complete the alignment

18. What is the main objective of data normalization?

a. Adjust measurements so that they can be to make colorful visual representations

b. Adjust measurements so that they can be easy to evaluate

c. Adjust measurements so that they can be appropriately compared among samples

d. Adjust measurements so that they can be algorithmic and easy to cluster

19. Your company is selling a multivitamin drug in an area where 30 % of the people live in the city and the rest live in the suburbs. Currently, 20 % of the city dwellers use your drug and 10 % of the suburbanites use your drug. You are presented with two new sales strategies the first will increase your market share in the suburbs to 15 %. The second will increase your market share in the city to 25 %. Which strategy should you adopt? What percentage of the people who own your product are city dwellers before your new sales drive?

a. Strategy 1 is better to sand before the new sales drive $\frac{6}{13}$ of the people who use the drug are city dwellers

b. Strategy 2 is better to sand before the new sales drive $\frac{6}{13}$ of the people who use the drug are city dwellers

c. Strategy 1 is better to sand before the new sales drive $\frac{9}{13}$ of the people who use the drug are city dwellers

d. Strategy 2 is better to sand before the new sales drive $\frac{9}{13}$ of the people who use the drug are city dwellers

20. Given the following statistics, what is the probability that a woman has cancer if she has a positive mammogram result?

- One percent of women over 50 have breast cancer.
- Ninety percent of women who have breast cancer test positive on mammograms.
- Eight percent of women will have false positives.

a. 50%

b. 20%

c. 10%

d. 0%

115

PART II

Select THREE (03) questions and answer

1. a) Briefly explain the following. **(25 marks)**

   i. What is Information Theory?

   ii. Basis of Information Theory

   iii. Give 3 examples of fields where you can apply Information Theory

   b) What is the relationship between Information and Biological Systems? **(30 marks)**

   c) Consider a test to detect a disease that 0.1 % of the population has. The test is 99 % effective in detecting an infected person. However, the test gives a false positive result in 0.5 % of cases. If a person tests positive for the disease what is the probability that they actually have it?

   (45 marks)

2. a) Explain the following keywords. **(25 marks)**

   i. K-nearest neighbor

   ii. Principle Component Analysis

   iii. Batch effect

   b) What is Cluster analysis? Explain why do we need cluster analysis with examples?

   (30 marks)

   c) Create the approximate clustering representation and the respective dendrogram that depicts the hierarchical clustering using given six points with the following attributes.

	p1	p2	p3	p4	p5	p6
p1	0.0000	0.2357	0.2218	0.3688	0.3421	0.2347
p2	0.2357	0.0000	0.1483	0.2042	0.1388	0.2540
p3	0.2218	0.1483	0.0000	0.1513	0.2843	0.1100
p4	0.3688	0.2042	0.1513	0.0000	0.2932	0.2216
p5	0.3421	0.1388	0.2843	0.2932	0.0000	0.3921
p6	0.2347	0.2540	0.1100	0.2216	0.3921	0.0000

Table 1: Distance matrix for six points

(45 marks)

3. a) Write short notes about the following keywords. **(25 marks)**

   i. Homology, ortholog, and paralog

   ii. Similarity

   iii. Genome Assembly

   b) What is Sequence alignment? Explain the difference between local and global pairwise alignments?

   (30 marks)

   c) What is a phylogenetic tree? Explain why phylogenetic trees are important?

   (45 marks)

4. a) Write short notes about the following mutation types. **(25 marks)**

i. Point mutation

ii. Frameshift mutation

iii. Trinucleotides repeat mutation

b) What is Dynamic programming? Explain the differences and similarities between the Needleman-Wunsch method and the Smith-Waterman method? **(30 marks)**

c) Use the Needleman-Wunsch method and create the matrix to align the two sequences. After Traceback write the aligned sequences.

Reference seq: GTACGC

Sample seq: GCATGC

(45 marks)

5. a) Write short notes about the following keywords. **(25 marks)**

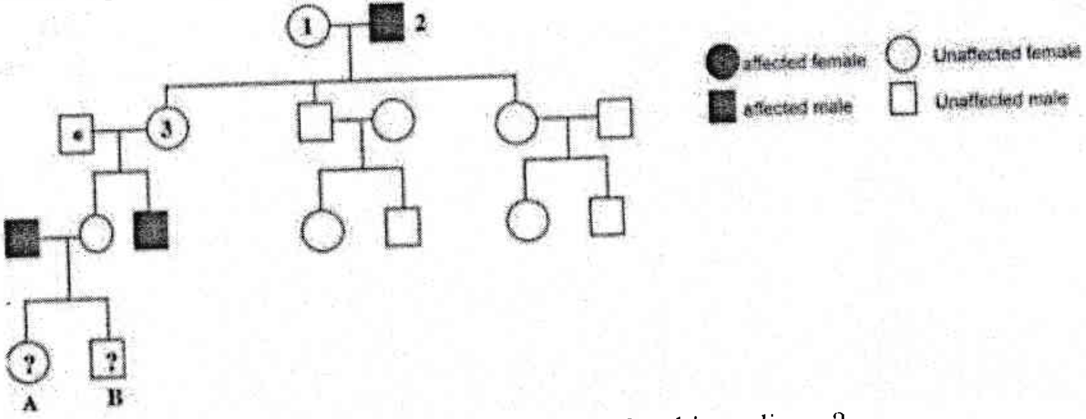
i. Genotype

ii. Phenotype

iii. Allele

b) What are the patterns of inheritances in medical genetics? Explain the difference between Autosomal and X-linked inheritance. **(30 marks)**

c) You are analyzing the following human pedigree. Assume that the individual marked with an asterisk (*) does not carry any allele associated with the affected phenotype and that no other mutation spontaneously occurs. Use "R or XR " for the allele associated with the dominant phenotype, "r or Xr " for the allele associated with the recessive phenotype.



A. What is the most likely mode of inheritance for this pedigree?

B. List all possible genotypes of 1 & 3 individuals in the pedigree.

C. What is the probability of Individual A being affected?

(45 marks)

- END -