Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

MASTER OF SCIENCE IN BIOMATHEMATICS

END OF SEMESTER EXAMINATION

BMA 8305: METHODS & MODELS IN GENOMICS

Date: 13th January, 2023 **Time: 3 Hours**

Instruction: Answer Question one and any other two questions

Question ONE (30 Marks)

a. Differentiate between the following terms as used in genomics:

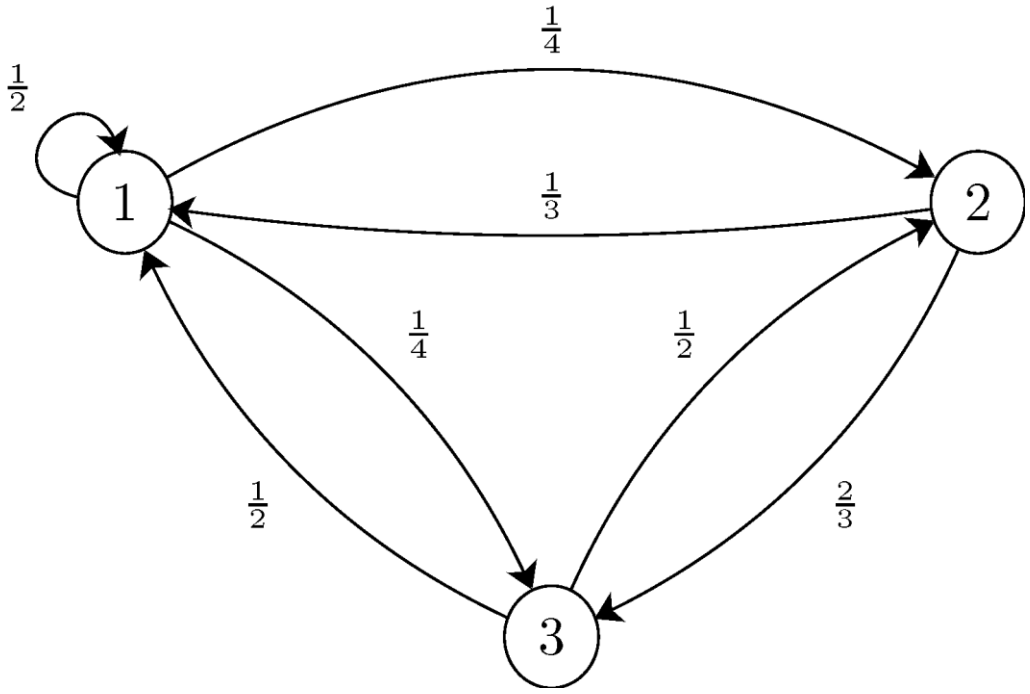
i. mRNA and tRNA (2 marks)

ii. Leading strand and lagging strand (2 marks)

iii. Introns and exons (2 marks)

iv. Sequence homology and Sequence similarity (2 marks)

b. Consider the Markov chain shown in the figure below



i. Is this chain irreducible? (1 mark)

ii. Is this chain aperiodic? (1 mark)

iii. Find the stationary distribution for this chain. (3 marks)

iv. Is the stationary distribution a limiting distribution for the chain? (1 mark)

c. Discuss briefly at least 3 genetic markers used in a genome mapping (6 marks)

d. Discuss at least 3 next generation sequencing technologies (6 marks)

e. Describe genome annotation as used in bioinformatics (4 marks)

Question TWO (15 marks)

a. Discuss the application of biopharma to improve the living standards of people (8 marks)

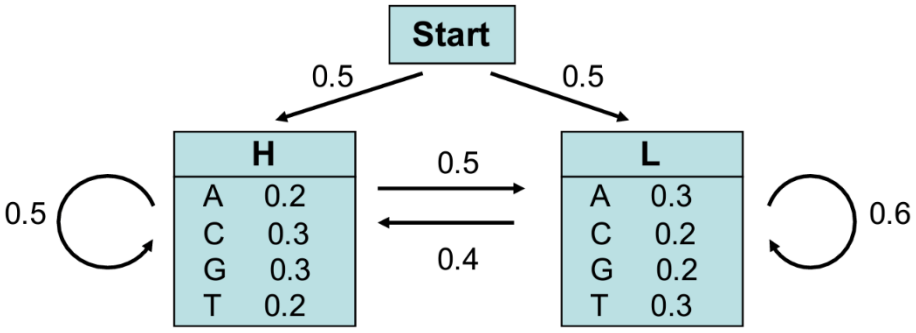
b. Hidden Markov Models (HMMs) are probabilistic models that are used in a wide variety of sequence analysis problems. We define an HMM for K classes of hidden states and T data points. Let the data set be $X= \{x_1, ..., x_T\}$, where each x_i a discrete observed variable. Hidden state variables are $Z = \{z_1, ..., z_T\}$, where each hidden state is $z_t \in \{1,...,K\}$. The transition probabilities are given by a $K \times K$ matrix A, where $a_{kj} = P(z_t = k \mid z_{t-1} = j)$. The initial state variable z_1 is special since it does not have a parent node. Its distribution can be represented by a vector of probabilities π where $P(z_1) = \pi z_1$. Finally, the emission distribution for a hidden state class k is parametrized by $\varphi_{.k}$, where $\varphi_{xk} = P(x_i = x \mid z_i = k)$. Let $\Theta = \{A,\pi,\varphi\}$. If we have a data set $X = \{x_1,...,x_T\}$, write the following expressions in terms of the parameters.

i. Write down the full likelihood of observed and latent variables, $P(X, Z \mid \Theta)$. (4 marks)

ii. Describe the Expectation-Maximization (EM) for estimating A and φ that maximizes the likelihood of the data set $P(X \mid \Theta)$. (3 marks)

Question THREE (15 marks)

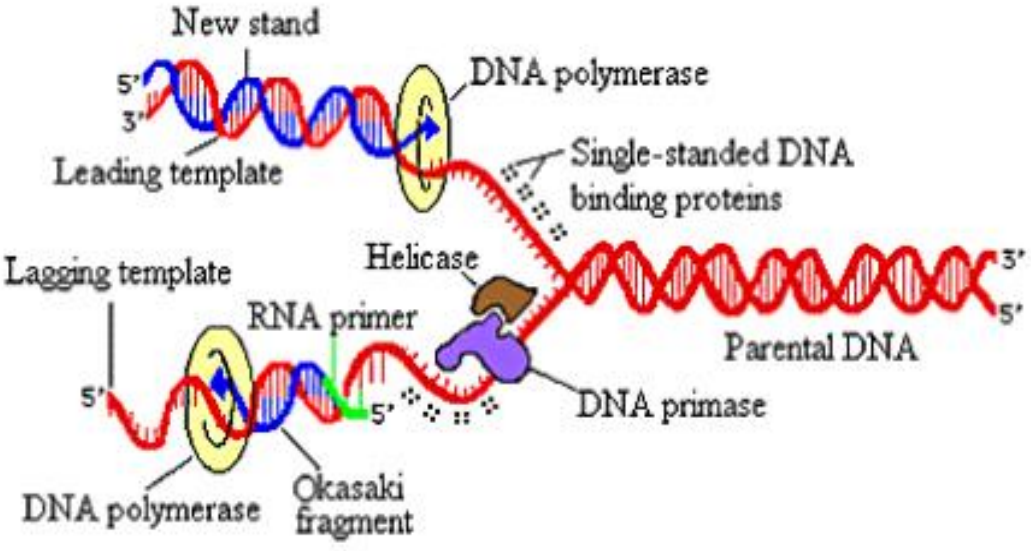
Consider a toy example



Let's consider the following simple HMM. This model is composed of 2 states, H (high GC content) and L (low GC content). We can for example consider that state H characterizes coding DNA while L characterizes non-coding DNA. Use Viterbi algorithm to determine the most probable path for sequence **GGCACTGAA** (15 marks)

Question FOUR (15 marks)

a. With the help of a diagram below,



elaborate the enzymes involved in the process of DNA replication
(10 marks)

b. Give a detailed definition of a hidden Markov model (HMM). Include in your answer a description of the assumptions made by this model.
(5 marks)