Final Exam

Short Answer Questions

1. How can **BWT** (Burrows-Wheeler Transform) be used for short read alignment?

2. Explain what **DNA splicing** is and its advantages.

3. If the genome is identical across all cells in an organism, why do cells differ in type and function? Why don't they transform into each other?

4. Describe two differences between **Illumina** and **Ion Torrent** sequencing technologies.

5. Assume the function of a specific protein leads to a disease. You aim to design a drug to prevent the disease, but you cannot inhibit the gene expression producing this protein (i.e., the protein will still be produced). What would your approach be for designing the drug?

Dynamic Programming and Sequence Alignment

6. Find the shortest circular sequence that includes all the substrings below:

Substrings: TACGT, TAATAA, GTAC, CGTAC, TAAC, AACTG, CTGT, TGTAA

7. Given a sequence of numbers $\mathbf{x_1, x_2, ..., x_n}$, an increasing subsequence of length **k** is defined as $\mathbf{x_{i1}, x_{i2}, ..., x_{ik}}$ such that $\mathbf{i_1 < i_2 < ... < i_k}$ and $\mathbf{x_{i1} < x_{i2} < ... < x_{ik}}$.

a) Propose a dynamic programming algorithm to find the longest increasing subsequence in a given sequence.

b) A two-increasing subsequence is a sequence that can be divided into two increasing subsequences. For example, the sequence **2,1,6,5,7,9** can be divided into **1,5,7,9** and **2,6.** Provide a counterexample where a greedy algorithm, which removes the longest increasing subsequence first and then finds the next one, fails to yield the correct solution.

Heuristic MSA Algorithm

8. At a specific step in a heuristic **MSA algorithm**, the following two alignments are given. Align these two sequences:

Alignment 1:

```
- GGAT - -
AGC - TAC
- - CAGAC
```

Alignment 2:

```
TGAC
- CAC
```

Note: In cases of equal scores, prioritize the alignment with the higher total score.

HMM (Hidden Markov Model)

9. Suppose a specific family of proteins starts with an **alpha-helix** structure, which either transitions to an irregular structure (**coil**) after a few amino acids or directly continues as a **beta-sheet** structure. Based on statistical analysis, we know the following probabilities:

- The probability of transitioning directly from alpha-helix to beta-sheet: **0.1**
- If an amino acid is in the alpha-helix region, the next amino acid has a **0.5** probability of being in the same region.
- If an amino acid is in the coil region, the next amino acid has a **0.2** probability of being in the same region.
- The probabilities of all amino acids occurring in the coil region are equal (**20 amino acids**).
- In the alpha-helix and beta-sheet regions, the probabilities of **P, A** are **0.1**, and **N, I** are **0.2.** The probabilities of other amino acids are equal.

a) Draw a model representing this scenario (only include edges with non-zero weights).

b) A newly discovered protein from this family has an unknown structure. Using your model, determine the most probable structure of the protein with the sequence **NPALSIL** using the **Viterbi algorithm.**

---

10. For the given HMM and DNA sequence below, calculate the probability that the fourth base is in the coding region:

Sequence: CTTAG

Probability Analysis: Casino Problem

11. Two coins are available:

- Coin 1: **80% heads**, **20% tails**
- Coin 2: **55% heads**, **45% tails**

After 100 flips with one of these coins (without switching), we observe **69 heads** and **31 tails.**

Someone claims that since **31** is closer to **20**, Coin 1 was likely used. What is your opinion on this claim?

Phylogenetic Tree Construction

12. The following distance matrix is provided for six species:

	A	**B**	**C**	**D**	**E**	**F**
A	0	5	7	9	5	8
B	5	0	4	10	6	9
C	7	4	0	7	9	8
D	9	10	7	0	6	11
E	5	6	9	6	0	8
F	8	9	8	11	8	0

Using the **Neighbor Joining** algorithm, construct an unrooted tree from this matrix.

---

13. For the following four species, the sequences are given as follows:

Sequences:
A: TAGGCATAGTAGT
B: TAGGCCCTCCTAG
C: TAGGCGCTCCTC
D: TAGGCCGCTCTG

a) Compute the Hamming distance matrix for the species based on their sequences.

b) Using the matrix from part (a), construct an unrooted tree using the **Neighbor Joining** algorithm.

c) Root the tree obtained in part (b) by identifying the species most distantly related to the others.

Motif Discovery

14. The **Planted Motif Problem** involves identifying a consensus motif of a specified length (**l**) from a set of sequences.

Input: A set of sequences (**S**) of equal length (**n**), and integers **l** and **d.**

Objective: Find a motif **m** such that:

- **m** has a Hamming distance of at most **d** with at least one substring of length **l** in each sequence.

Propose an algorithm to solve this problem.