Lecturer: Shlomo Moran
TA: Ilan Gronau

- Exam Duration: 3 hours
- The Exam has 3 questions which sum up to 100 credits, + 5 Bonus Credits. The Maximum possible grade is 100.
- Material: Lectures and Tutorial slides,
  Books: Durbin et al, Setubal et al, Gusfield

<table>
<thead>
<tr>
<th>/33</th>
<th>Question 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>/33+5</td>
<td>Question 2</td>
</tr>
<tr>
<td>/34</td>
<td>Question 3</td>
</tr>
<tr>
<td>/100</td>
<td>Total</td>
</tr>
</tbody>
</table>

Good Luck
Question 1 (33 points)

Two sequences $S, T$ of lengths $n, m$ (respectively) are given over some alphabet $\Sigma$, as well as an additive scoring scheme defined by a function $\sigma : \Sigma \cup \{-\} \times \Sigma \cup \{-\} \rightarrow \mathbb{R}$. In your proofs you may assume correctness of any of the algorithms taught in the lectures or tutorials.

1. (10 pts) Suggest an $O(nm)$ algorithm for constructing a data structure which allows to find in constant ($O(1)$) time the best (global) alignment between each suffix of $S$ and each suffix of $T$.

   - Analyze the time complexity of your algorithm and prove its correctness.
   - Describe the way by which your data structure can be used to answer the desired queries.

2. (23 pts) Denote by $S^i$ the sequence $S$ after deletion of its $i^{th}$ character. Suggest an $O(nm)$ algorithm for finding an optimal alignment between $T$ and each of the reduced sequences $\{S^i\}_{i=1..n}$. Note that your algorithm should return a total of $n$ alignments in a single execution.

   - Analyze the time complexity of your algorithm and prove its correctness (you may assume correctness of the data structure you proposed in (1)).
The Technion Casino introduces the following game: there are two coins, red and blue, each has two possible outcomes: T (tail) and H (head). A round of the game goes as follows: first the blue coin is tossed once. If the outcome is T then the red coin is tossed twice, otherwise the red coin is tossed three times. If the number of times the red coin gave out H is odd then the gambler wins the game (otherwise she loses the game).

**Example:** If the blue coin gave out T and then the red coin gave out TH, then the gambler wins. On the other hand, if the blue coin gave out H and then the red coin gave out HHT, then the gambler loses.

Let $p$ denote the probability the blue coin gives out T and let $q$ denote the probability the red coin gives out T.

1. **(7 pts)** Define the probability space defined by the game:
   - Write down all the "simple events".
   - For each event write its probability (as a function of $p, q$), and its contribution to the various statistics used by the EM algorithm – $N_{BH}, N_{BT}, N_{RH}, N_{RT}$ (B stands for Blue and R stands for Red).

The Technion Faculty suspected that something is not Kosher with the casino, and sent Prof X (X in short) to estimate the fairness of the coins (i.e., the values $p$ and $q$). However X wasn't allowed to enter the casino, and all the information he was able to get is that there was one game and the gambler won that game. We call this event W.

2. **(4 pts)** For given $p$ and $q$, write down the probability of the event W.

3. **(15 pts)** X decided to use the EM algorithm in order to estimate $p$ and $q$ which maximize the likelihood of W. First he did so by starting with $p_1=q_1=0.5$.
   - What is the likelihood of W with these parameters?
   - Describe the calculations done in the E and M steps, and write down the updated parameters $p_2$ and $q_2$ returned after the first round.
   - What happened to the likelihood of W after the first round?

4. **(7 pts)** X decided to run the EM algorithm with different starting parameters: $p_3=q_3=0.1$. With $p_3$, $q_3$ the likelihood of W is 0.6984. X ran a single iteration of the EM algorithm and got updated parameters $p_4$, $q_4$ which give a different likelihood score than $p_3$, $q_3$ (you do not have to calculate the values of $p_4$ and $q_4$).
   - Which of the above parameters ($[p_1, q_1]$ or $[p_2, q_2]$ or $[p_3, q_3]$ or $[p_4, q_4]$) is the best estimate in terms of likelihood? Explain your answer.

5. **(5 pts Bonus)** What values for $p$ and $q$ maximize the likelihood of W? Prove your answer.

   **Hint:** You do not need to make any complex calculations to solve this.
This question deals with the reconstruction of phylogenetic trees from noisy distances.

Let $T = (V,E)$ be a rooted binary phylogenetic tree over $n$ taxa (which are the leaves of $T$), and let $w:E \rightarrow \mathbb{R}$ be its edge-weight function. Let $D$ be the additive distance matrix defined by $T,w$ (i.e. $D(i,j)$ is the distance in $T$ from leaf $i$ to leaf $j$ according to the weights assigned by $w$). Let $r$ denote the root of $T$, and for leaves $i,j$ let $L(i,j)=LCA(i,j)$ be the distance in $T$ from $r$ to the least common ancestor of $i$ and $j$.

1. **(3 pts)** Explain briefly (preferably by a figure) the following formula:

   $$L(i,j) = \frac{1}{2}[D(r,i) + D(r,j) - D(i,j)]$$

2. **(3 pts)** Complete the missing part in the following statement:
   $i$ and $j$ are neighboring leaves in $T$ iff $L(i,j)=\max\{L(x,y):\}$

   In the rest of this question, $m>0$ is a fixed constant satisfying: $w(e) \geq m$ for each edge $e \in E$.

3. **(7 pts)** Show that for every 3 leaves $i,j,k$ in $T$, if $L(i,j) \neq L(i,k)$ then $|L(i,j) - L(i,k)| \geq m$.

   A matrix $L'$ is an $\varepsilon$-approximation of $L$ if for all $i,j$, $|L(i,j) - L'(i,j)| < \varepsilon$.

4. **(7 pts)** Prove that for every 3 leaves $i,j,k$ in $T$, if $L(i,j)<L(i,k)$ and $L'$ is an $m/2$-approximation of $L$, then $L'(i,j)<L'(i,k)$.

5. **(14 pts)** Suggest an efficient algorithm which receives an input matrix $L'$ which is an $m/2$-approximation of $L$ and returns all the cherries (neighboring leaves) of $T$.
   - Analyze the time complexity of your algorithm and prove its correctness.