1. **The basics**
   a) Define the term "silent mutation"

   Silent mutations are DNA mutations that do not result in a change to the amino acid sequence of a protein

   b) What is the "central dogma" of molecular biology?

   Biological information flows in only one direction, from DNA to RNA to proteins. Replication (DNA) -> Transcription (DNA-RNA) -> Translation (RNA->protein)

   c) Identify the longest open reading frame in the following DNA sequence and translate it into an amino-acid sequence (note: translation table provided at the end of the exam)

   TGCATATGTAGTCAGACGGTGACCGCTTGCAGGCTAAGCGACG

   ATG TCA GAC GGT GAG ACG CTT GCG GGC TAA
   M S D G E T L A G

2. **Sequence alignment**

   a) Describe the recurrence and location of answer for global alignment between two sequences.

   Recurrence:

   \[ OPT(i, j) = \min \begin{cases} 
   \text{cost}(a_i, b_j) + OPT(i - 1, j - 1) & \text{match } a_i, b_j \\
   \text{gap} + OPT(i - 1, j) & a_i \text{ is not matched} \\
   \text{gap} + OPT(i, j - 1) & b_j \text{ is not matched} 
\end{cases} \]

   Cost of the optimal alignment between \( a_1...a_i \) and \( b_1...b_j \)

   Written in terms of the costs of smaller problems

   **Bottom rightmost cell**
b) Perform a global **multiple** sequence alignment on the following sequences and report the alignment and Sum-of-Pairs score. Use Seq1 as Sc in both (center of star tree). **MATCH** = +1, **MISMATCH** = -1, **GAP** = -1.

Seq1: AGT
Seq2: ACT
Seq3: AGAT

\[
\begin{align*}
\text{Seq1: AGT} \\
\text{Seq2: ACT} \\
\text{Seq3: AGAT}
\end{align*}
\]

\[
\begin{align*}
\text{AG-T} \\
\text{AC-T} \\
\text{AGAT} \\
\text{S(-,-) = 0 (if not we double count)}
\end{align*}
\]

\[
+1, +1, +1, -1, +1, -1, 0, -1, -1, +1, +1, +1 = 3
\]
3. Genome assembly

a) The Lander-Waterman model describes the expected number of contigs (N) in a genome project as a function of the genome length G, read length L, depth of coverage c, and the overlap between sequences o. Without remembering the exact formula, sketch the rough shape of the dependency between N and c, assuming G, L, and o are fixed.
4. Genome alignment

Briefly describe what is depicted in the dot plot below:

A large scale inversion event between genome A & genome B.
4. Data structures: Suffix trees
a) Given the following string, construct a suffix tree of ATGTAG

![Suffix tree diagram]

a) Label the path of the string GTAG in the above suffix tree. Give the time complexity of finding a query of length ‘n’.

\[O(n)\]
## Translation table

### Ter - stop codon

<table>
<thead>
<tr>
<th>Codon</th>
<th>Amino Acid</th>
<th>Codon</th>
<th>Amino Acid</th>
<th>Codon</th>
<th>Amino Acid</th>
<th>Codon</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTT</td>
<td>F Phe</td>
<td>TCT</td>
<td>S Ser</td>
<td>TAT</td>
<td>Y Tyr</td>
<td>TGT</td>
<td>C Cys</td>
</tr>
<tr>
<td>TTC</td>
<td>F Phe</td>
<td>TCC</td>
<td>S Ser</td>
<td>TAC</td>
<td>Y Tyr</td>
<td>TGC</td>
<td>C Cys</td>
</tr>
<tr>
<td>TTA</td>
<td>L Leu</td>
<td>TCA</td>
<td>S Ser</td>
<td>TAA</td>
<td>* Ter</td>
<td>TGA</td>
<td>* Ter</td>
</tr>
<tr>
<td>TTG</td>
<td>L Leu</td>
<td>TCG</td>
<td>S Ser</td>
<td>TAG</td>
<td>* Ter</td>
<td>TGG</td>
<td>W Trp</td>
</tr>
<tr>
<td>CTT</td>
<td>L Leu</td>
<td>CCT</td>
<td>P Pro</td>
<td>CAT</td>
<td>H His</td>
<td>CGT</td>
<td>R Arg</td>
</tr>
<tr>
<td>CTC</td>
<td>L Leu</td>
<td>CCC</td>
<td>P Pro</td>
<td>CAC</td>
<td>H His</td>
<td>CGC</td>
<td>R Arg</td>
</tr>
<tr>
<td>CTA</td>
<td>L Leu</td>
<td>CCA</td>
<td>P Pro</td>
<td>CAA</td>
<td>Q Gln</td>
<td>CGA</td>
<td>R Arg</td>
</tr>
<tr>
<td>CTG</td>
<td>L Leu</td>
<td>CGG</td>
<td>P Pro</td>
<td>CAG</td>
<td>Q Gln</td>
<td>CGG</td>
<td>R Arg</td>
</tr>
<tr>
<td>ATT</td>
<td>I Ile</td>
<td>ACT</td>
<td>T Thr</td>
<td>AAT</td>
<td>N Asn</td>
<td>AGT</td>
<td>S Ser</td>
</tr>
<tr>
<td>ATC</td>
<td>I Ile</td>
<td>ACC</td>
<td>T Thr</td>
<td>AAC</td>
<td>N Asn</td>
<td>AGC</td>
<td>S Ser</td>
</tr>
<tr>
<td>ATA</td>
<td>I Ile</td>
<td>ACA</td>
<td>T Thr</td>
<td>AAA</td>
<td>K Lys</td>
<td>AGA</td>
<td>R Arg</td>
</tr>
<tr>
<td>ATG</td>
<td>M Met</td>
<td>ACG</td>
<td>T Thr</td>
<td>AAG</td>
<td>K Lys</td>
<td>AGG</td>
<td>R Arg</td>
</tr>
<tr>
<td>GTT</td>
<td>V Val</td>
<td>GCT</td>
<td>A Ala</td>
<td>GAT</td>
<td>D Asp</td>
<td>GGT</td>
<td>G Gly</td>
</tr>
<tr>
<td>GTC</td>
<td>V Val</td>
<td>GCC</td>
<td>A Ala</td>
<td>GAC</td>
<td>D Asp</td>
<td>GCC</td>
<td>G Gly</td>
</tr>
<tr>
<td>GTA</td>
<td>V Val</td>
<td>GCA</td>
<td>A Ala</td>
<td>GAA</td>
<td>E Glu</td>
<td>GGA</td>
<td>G Gly</td>
</tr>
<tr>
<td>GTG</td>
<td>V Val</td>
<td>GCG</td>
<td>A Ala</td>
<td>GAG</td>
<td>E Glu</td>
<td>GGG</td>
<td>G Gly</td>
</tr>
</tbody>
</table>