Chapter 5: Sequence alignment

Answers.

1. Aligning AGGWCTTY and GWCTTY with: \( s(i,j) = 1 \), \( s(i,j) = 0 \), no gap penalty.

\[
\begin{array}{cccccccc}
A & G & G & W & C & T & T & Y \\
G & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
w & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\
c & 0 & 0 & 1 & 2 & 2 & 2 & 2 \\
t & 0 & 0 & 1 & 2 & 2 & 3 & 3 \\
y & 0 & 0 & 1 & 2 & 3 & 4 & 5 \\
\end{array}
\]

\# of alignments with score 5: \(14\)

2) Same as in 1, but now a gap penalty of 1 (i.e. \(-1\))

Assume gaps at the beginning count.

\[
\begin{array}{cccccccc}
A & G & G & W & C & T & T & Y \\
G & 0 & 0 & 0 & -1 & -1 & -1 & -1 \\
w & -1 & 0 & 0 & 1 & -1 & -1 & -1 \\
c & -1 & -1 & 0 & 0 & 0 & 2 & 0 & 0 \\
t & -1 & -1 & -1 & 0 & 0 & 3 & 2 & 1 \\
y & -1 & -1 & -1 & -1 & 0 & 1 & 3 & 3 \\
\end{array}
\]
3. The 3 sequence alignments are:

\[
\begin{align*}
\text{THESE SENTENCES ALIGN} & \quad \text{--- NICELY} \\
\text{THE SEQUENCE} & \quad \text{--- ALIGNED NICELY} \quad \text{Score: 180}
\end{align*}
\]

\[
\begin{align*}
\text{THESE SENTENCES ALIGN} & \quad \text{--- NICELY} \\
\text{THE-SEQ-ENCE} & \quad \text{--- ALIGNED NICELY} \quad \text{Score: 190}
\end{align*}
\]

\[
\begin{align*}
\text{THESE SENTENCES ALIGN} & \quad \text{--- NICELY} \\
\text{THE-SEQ-ENCE} & \quad \text{--- ALIGNED NICELY} \quad \text{Score: 190}
\end{align*}
\]

All 3 alignments have the same score, 190.

4. When we account for gaps:

Alignment 1:
\[
\text{Score} = 190 - 5 - 2 \times 4 - 5 - 2 \times 2 = 168
\]

Alignment 2:
\[
\text{Score} = 190 - 5 - 2 - 5 - 2 \times 2 - 5 - 2 - 5 - 2 \times 2 = 15
\]

Alignment 3:
\[
\text{Score} = 190 - 5 - 2 \times 2 - 5 - 2 - 5 - 2 - 5 - 2 \times 2 = 13
\]

Alignment 1 is optimal.

5. As alignment B and C have similar gaps, we need to change the scores of individual matches = answer D.
6. The size of the alphabet matters: you do better with proteins (20 letters) than with DNA (4 letters).

Answer: C

7.8. ATGCA 3  CTCCCA 2.4  GAAAA 2.6
     CTCCCA 2.9  CTCCCA 3.6
     ATGCA 3.3  CTGCA

Most dissimilar: CTCCCA with GAAAA
Most similar: CTCCCA with CTGCA

9. PAT250, just like BLAST, reflects the properties of amino acids: answer C

10. W and C are very different: W is aromatic, bulky, C is small, involved in disulfide bridges — answer B

11. All 5 amino acids (I, L, T, F, V) are hydrophobic (non polar) — answer A

12. P-value = E-value = $10^{-4}$, answer C

P-value: probability of a random match with same score.