Series 4

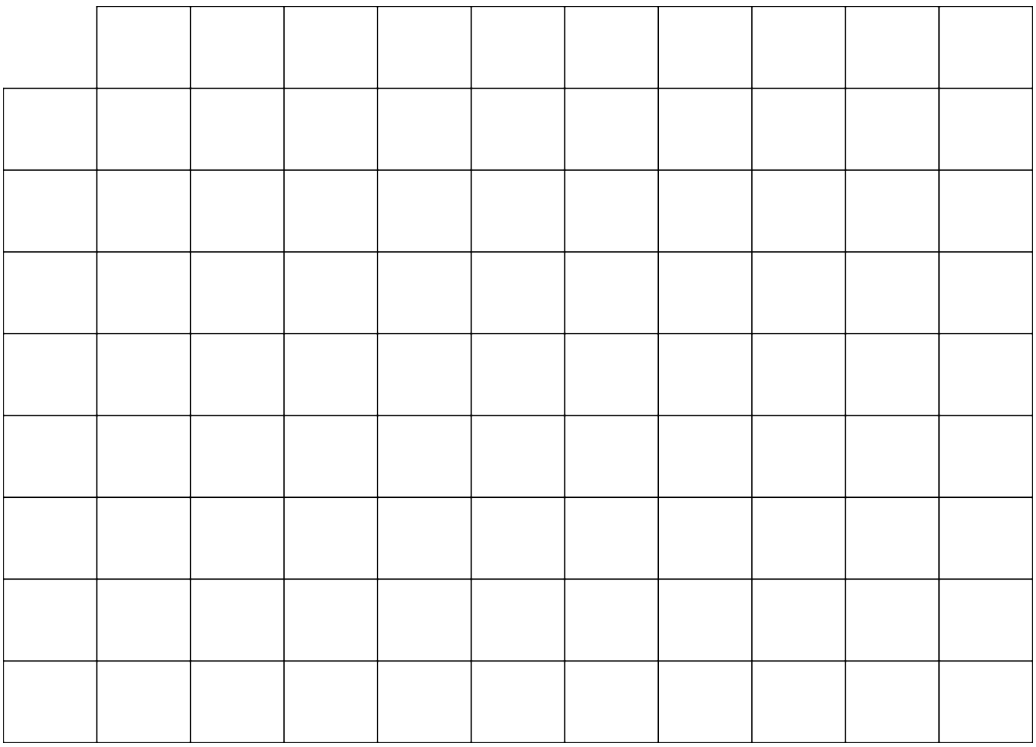
Genomics and bioinformatics - Week 4

1 Sequence alignment: Needleman-Wunsch

In this exercise you will manually perform a global alignment of two sequences using the Needleman-Wunsch algorithm and based on the following scoring scheme: for X, Y in $\{A, T, G, C, -\}$,

$$M(X,Y) = \begin{cases} +1 & \text{if} \quad X = Y \\ -2 & \text{if} \quad X = - \quad \text{or} \quad Y = - \\ -1 & \text{otherwise} \end{cases}$$

Sequence 1: `GAATTCAG`
Sequence 2: `GGATCG`.



Find the best alignment and its score. Is it unique ?

2 Pair Hidden Markov Model

In this exercise, we will construct a pair Hidden Markov Model for the same sequences as in the first exercise and align them using the path with maximum probability. The maximum probability path and the corresponding alignment are calculated by an algorithm called the Viterbi Algorithm. You will see through the exercise that the Viterbi algorithm is actually equivalent to the Needleman-Wunsch algorithm.

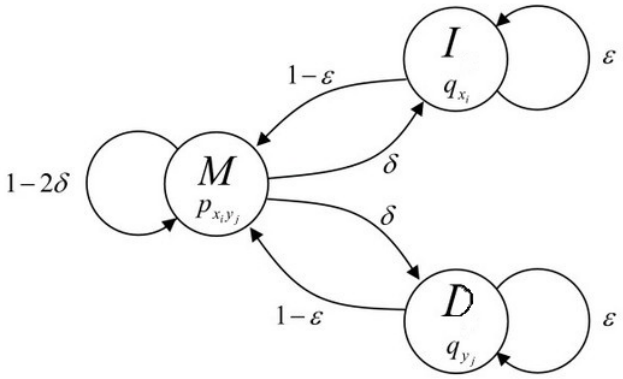


Figure 1: Pair Hidden Markov Model

The Pair HMM consists of the following parameters (see fig. 1):

- Three states:

  State M matches one letter from each sequence.

  State D (Deletion) inserts a gap to the second sequence.

  State I (Insertion) inserts a gap to the first sequence.

- Emission probabilities:

  $p(x, y)$ = probability of emitting a pair of characters $[x, y]$ in state M ($x, y \in \{A, T, G, C\}$).

  $p(x) = p(x, -)$ = probability of emitting a pair of characters $[x, -]$ in state D.

  $p(y) = p(-, y)$ = probability of emitting a pair of characters $[-, y]$ in state I.

- Transition probabilities (affine gaps case):

  δ = probability of opening a gap

  ε = probability of extending a gap

To make correspondence to the Needleman-Wunsch algorithm with the scores given in Exercise 1,

$$S(x, y) = \lambda \cdot log_2 \frac{p(x, y)}{p(x)\ p(y)} \quad , \quad d = log_2(\delta) \quad ,$$

$S(x, y) = 1$ if $x = y$,
$S(x, y) = -1$ if $x \neq y$,
$d = e = -2$ for (linear) gap penalty,
$\lambda = \frac{1}{log_2(3)}$ (just for scaling).

The Viterbi algorithm goes through the three steps:

Step 1: Initialization:

$V_M(0,0) := 0$
$V_I(0,0) := -\infty$
$V_D(0,0) := -\infty$
$V_*(-1,j) = V_*(i,-1) := -\infty$ ($_*$ accounts for either M, D or I)

Step 2: Recursion:

$$V_M(i,j) = S(x_i, y_j) + \max \begin{cases} V_M(i-1,j-1) \\ V_D(i-1,j-1) \\ V_I(i-1,j-1) \end{cases}$$

$$V_I(i,j) = \max \begin{cases} V_M(i-1,j) + d \\ V_I(i-1,j) + e \end{cases}$$

$$V_D(i,j) = \max \begin{cases} V_M(i,j-1) + d \\ V_D(i,j-1) + e \end{cases}$$

Step 3: Termination:

$$V_E(n,m) = \max(V_M(n,m), V_D(n,m), V_I(n,m)), \quad \text{for all } m,n$$

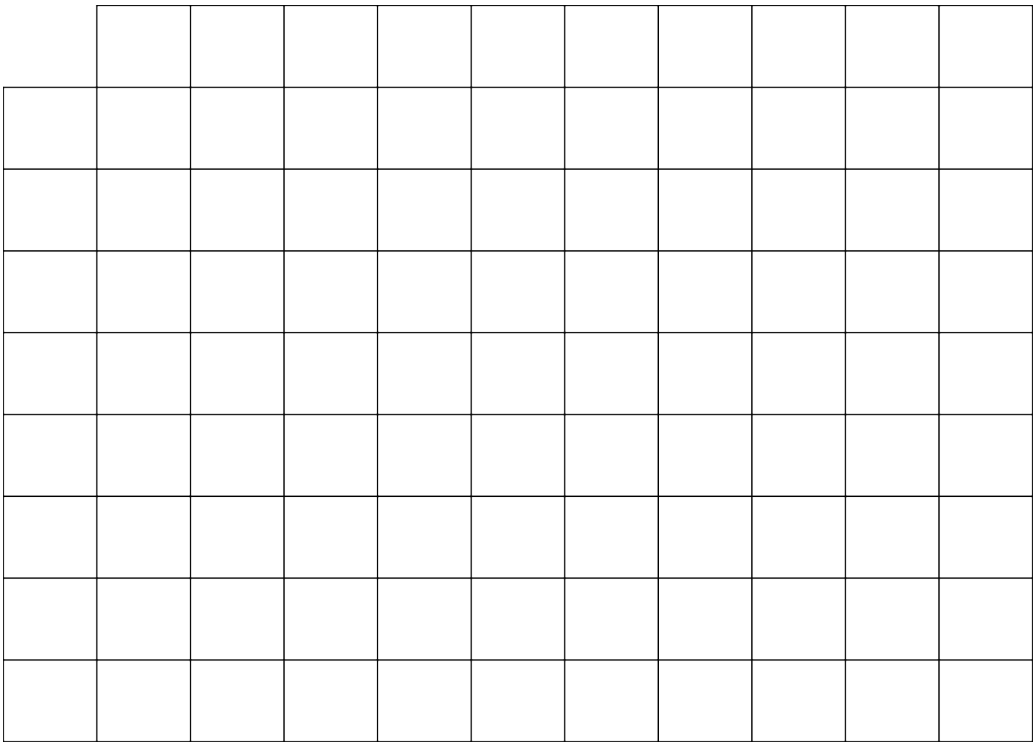
Questions:

1. Deduce the emission probability matrix and the transition probabilities of the HMM. *Note*: suppose that independent emissions of A,T,G,C are equiprobable (1/4).
2. Use the algorithm shown above to generate V_M, V_D, V_I, V_E.
3. Deduce the alignment based on these matrices.

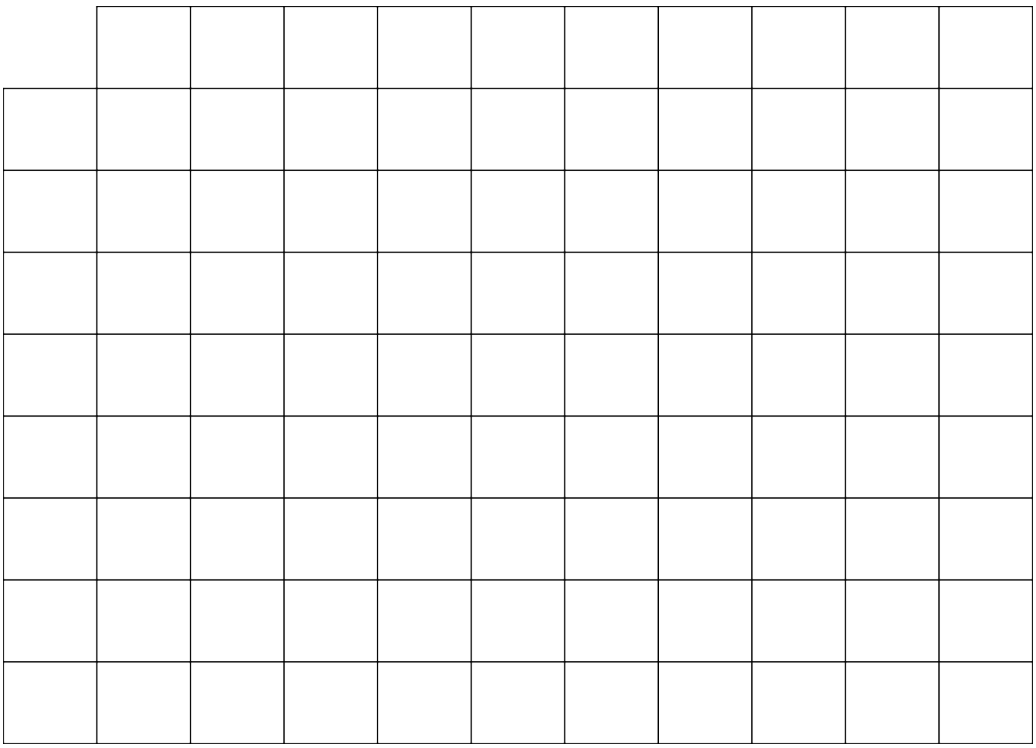
Matrix V_M:

Matrix V_D:

Matrix V_I:



Backtracking matrix V_E:



The possible alignments are: .....