Series 4

Genomics and bioinformatics - Week 4 - October 9, 2011

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.

The best alignment is: ....

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.
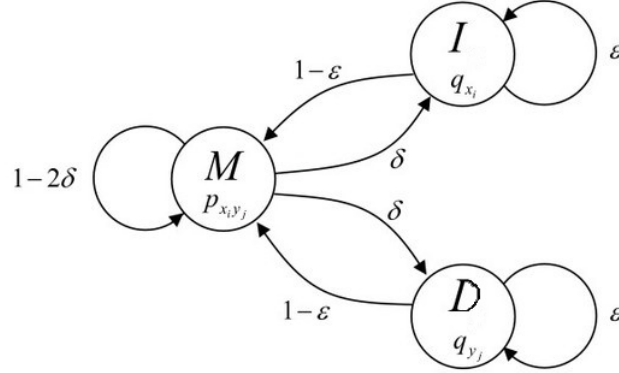

Figure 1: Pair Hidden Markov Model

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

- Three states:

  State M matches one letter from each sequence.

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

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

- Emission probabilities:

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

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

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

- 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,

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

where
S(x,y) = 1 if $x = y$,
S(x,y) = -1 if $x \neq y$,
d = e = -2 for (linear) gap penalty.

The Viterbi algorithm goes through the three steps:

Step 1: Initialization:

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

Step 2: Recursion:

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

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

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

Step 3: Termination:

$$V_E = max(V_M(n,m), V_D(n,m), V_I(n,m))$$

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.
   Hint: at some point, use the formula defining conditional probabilities.

2. Use the algorithm shown above to generate the three matrices for Match(M), Deletion(D) and Insertion(I).

3. Deduce the alignment based on the three matrices.

Matrix V_M:

				4					

Matrix V_D:

Matrix V_I:

				5					

Backtracking matrix:

The possible alignments are: ....