

Series 4

Genomics and bioinformatics - Week 5 - October 11, 2011

1 Sequence alignment: Needleman-Wunsch

The Needleman-Wunsch algorithm uses a method called “dynamic programming” - a very general programming technique. It involves three main steps:

1. Initialization
2. Scoring (matrix fill)
3. Alignment (backtracking)

In the first exercise of this session you will manually perform a global alignment of two sequences based on the following scoring scheme: *Match*: +1, *Mismatch*: -1, *Gap*: -2 .

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 of generating the alignment and the corresponding path are calculated by a dynamic programming algorithm which is called the Viterbi Algorithm. You will see through the exercise that the Viterbi algorithm is actually similar to the Needleman-Wunsch algorithm.

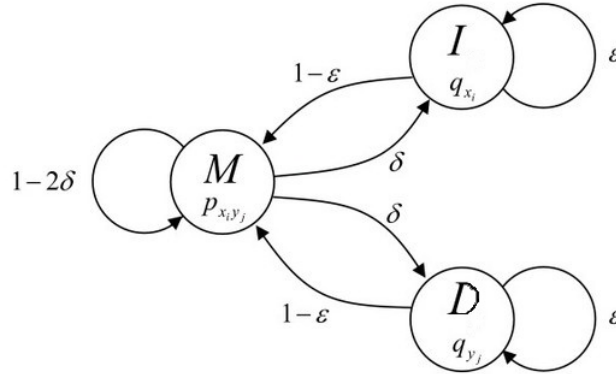


Figure 1: Pair Hidden Markov Model

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 second sequence.

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

Emission probabilities:

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

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

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

Transition probabilities:

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,

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

where

$S(x, y) = 1$ for match,

$S(x, y) = -1$ for mismatch,

$d = e = -2$ for gap penalty.

The Viterbi algorithm goes through the three steps:

Step 1: Initialization:

$$V_M(0, 0) := 0$$

$$V_D(0, 0) := -\infty$$

$$V_I(0, 0) := -\infty$$

$$V_*(-1, j) = V_*(i, -1) := -\infty \quad (* \text{ accounts for either M, D or I})$$

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.
2. Use the algorithm as shown above to generate the three matrices for Match(M), Deletion(D) and Insertion(I).
3. Deduce the alignment based on the three matrices.

Solution:

Matrix V_M :

Matrix V_D :

Matrix V_I :

Backtracking matrix:

The possible alignments are: