Series 4

Genomics and bioinformatics - Week 4

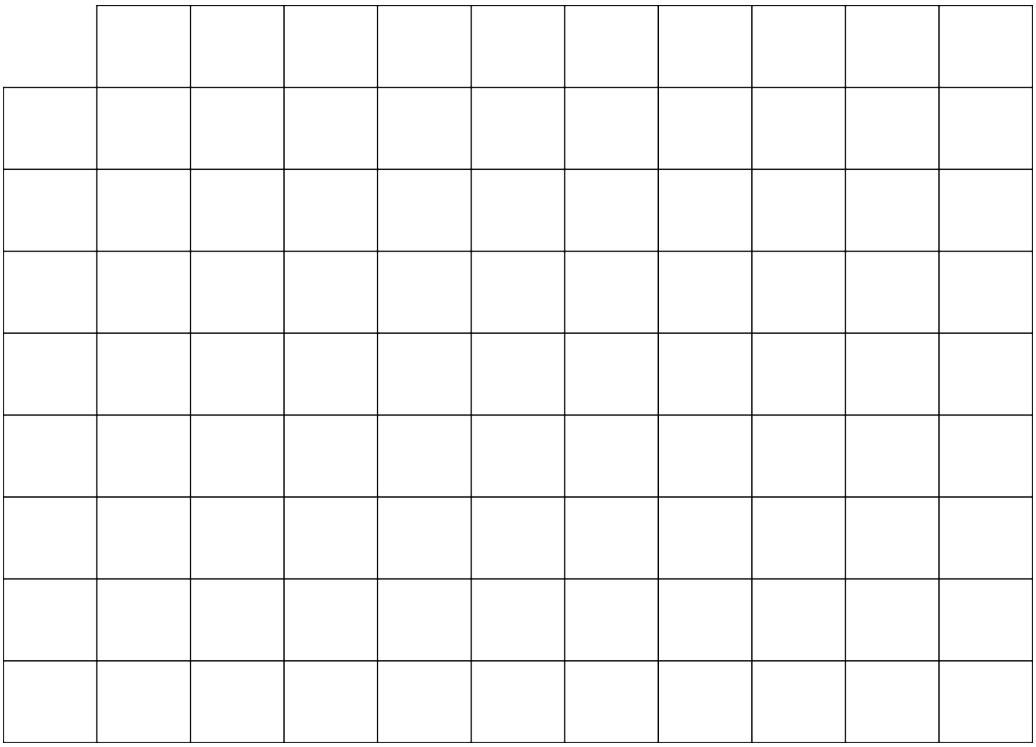
October 11, 2011

1 Sequence alignment

The Needlman-Wunsch algorithm uses a method called "dynamic programming". This is 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: `GAATTCAGA`
Sequence 2: `GGATCGA`.



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.

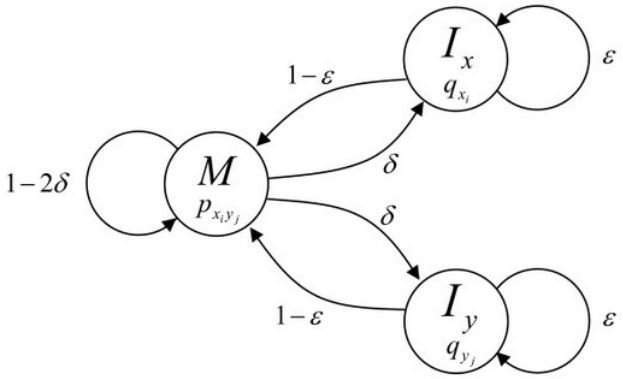


Figure 1: Pair Hidden Markov Model

The Pair HMM consists of the following parameters(Figure)

Three states: M, I, D
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

So, three states will be denoted by M, I and D in the algorithm written below.

Emission probabilities: p(x, y), q(x) and q(y), where
p(x,y) = probability of emitting a pair of characters [x,y]
q(x) = probability of emitting a pair of character [x,_]
q(y) = probability of emitting a pair of character [_,y]

Transition probabilities:
δ = probability of opening a gap
ε = probability of extending a gap

The algorithm goes through the three steps:

Step 1: Initialization

$$VM(0,0) = 0; VD(0,0) = -\infty; VI(0,0) = -\infty; V*(-1,j) = V*(i,-1) = -\infty;$$

Step 2: Recursion

$$VM(i,j) = S(x_i, y_j) + max \begin{cases} VM(i-1, j-1) \\ VD(i-1, j-1) \\ VI(i-1, j-1). \end{cases}$$

$$VD(i,j) = max \begin{cases} VM(i-1, j) - d \\ VD(i-1, j) - e. \end{cases}$$

$$VI(i,j) = max \begin{cases} VM(i, j-1) - d \\ VI(i, j-1) - e. \end{cases}$$

Step 3: Termination

$$VE = max(VM(n,m), VD(n,m), VI(n,m))$$

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

$$S(x,y) = log \frac{p(x,y)}{p(x)\ p(y)}$$

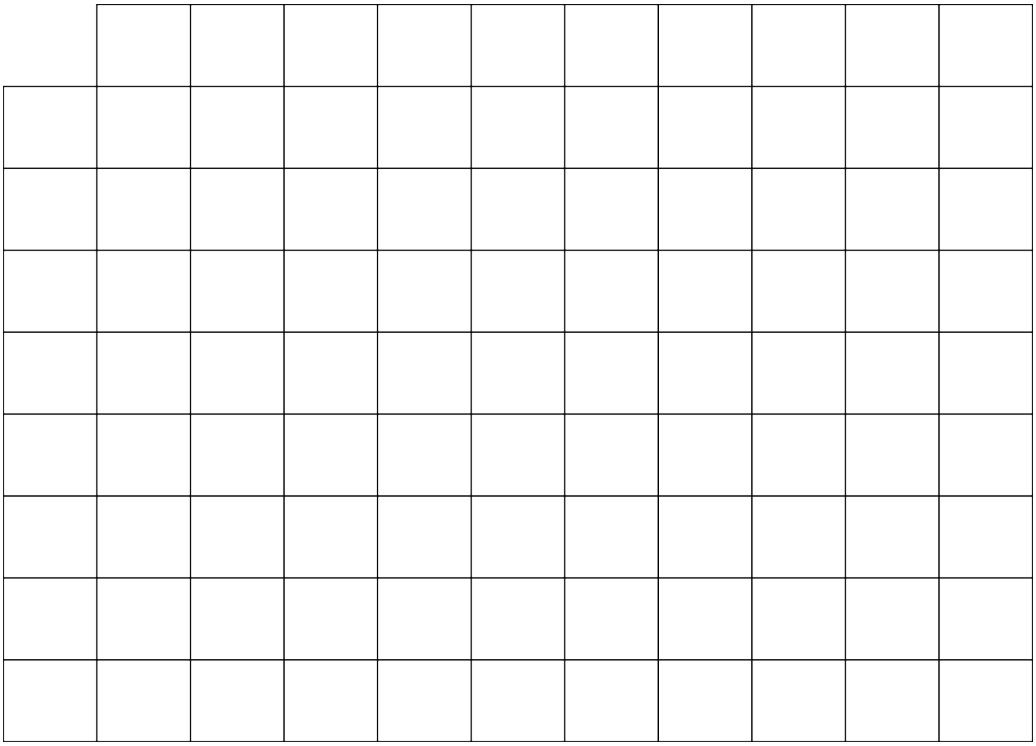
$$d = -log(\delta)$$

where
S(x,y) = 1 for match,
S(x,y) = -1 for mismatch,
d = -2 for gap penalty.

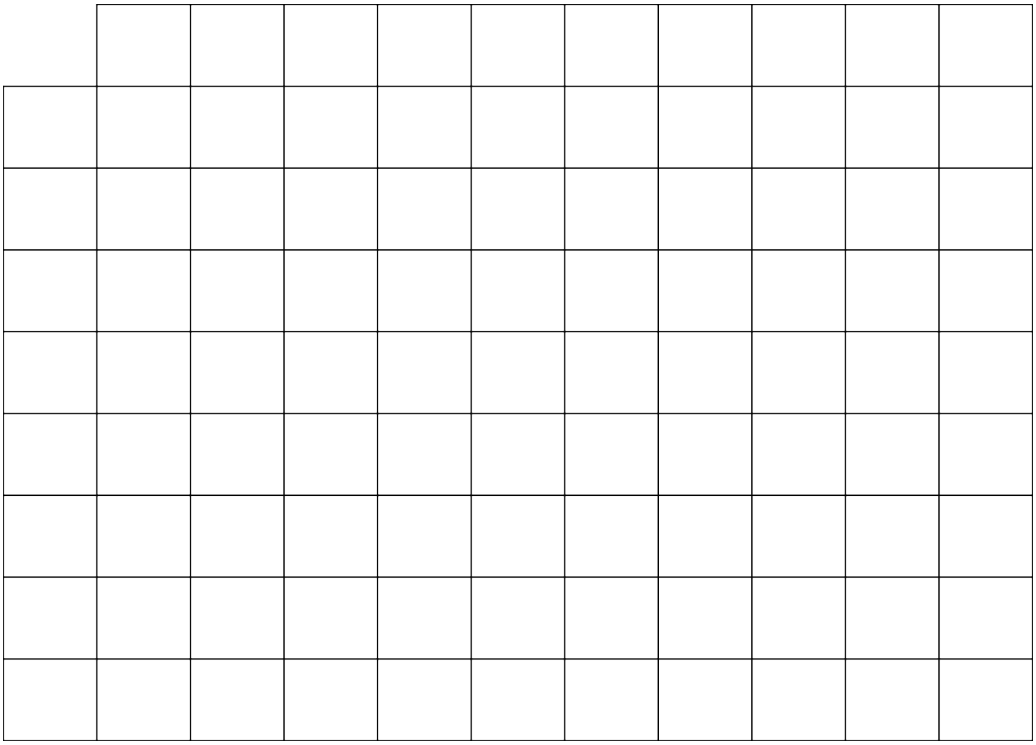
1. Deduce the emission probability matrix and the transition probabilities for the HMM
2. Use the algorithm as shown above to generate the three matrices for Match(M), Delete(D) and Insertion(I)
3. Deduce the alignment based on the three matrices.

Solution:

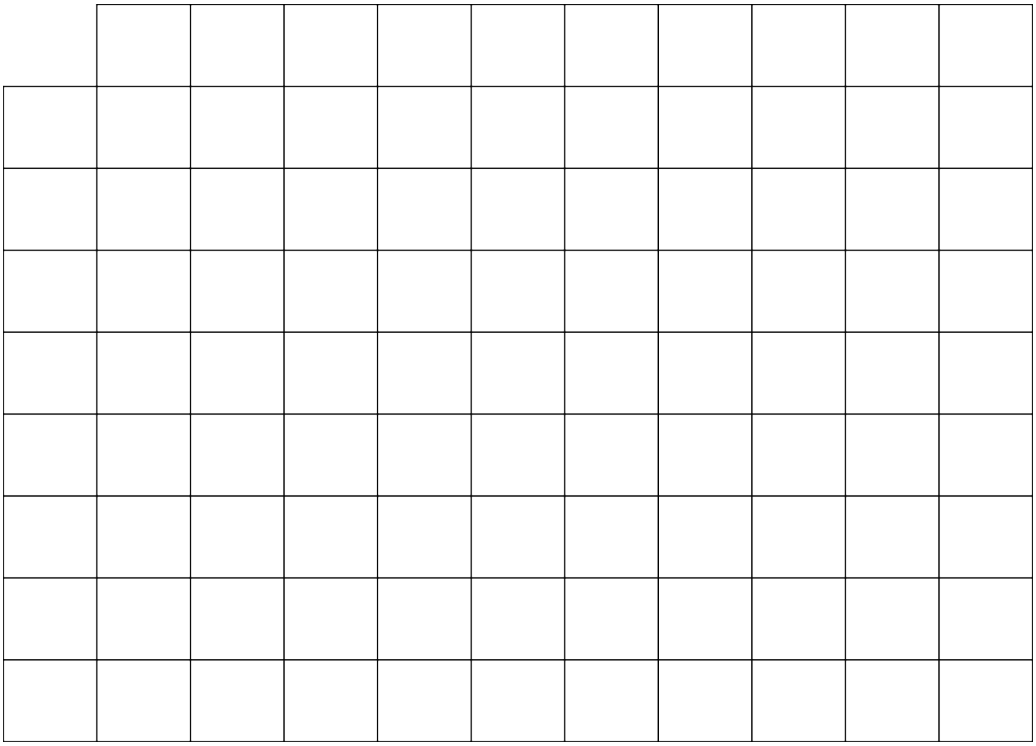
Matrix 1:



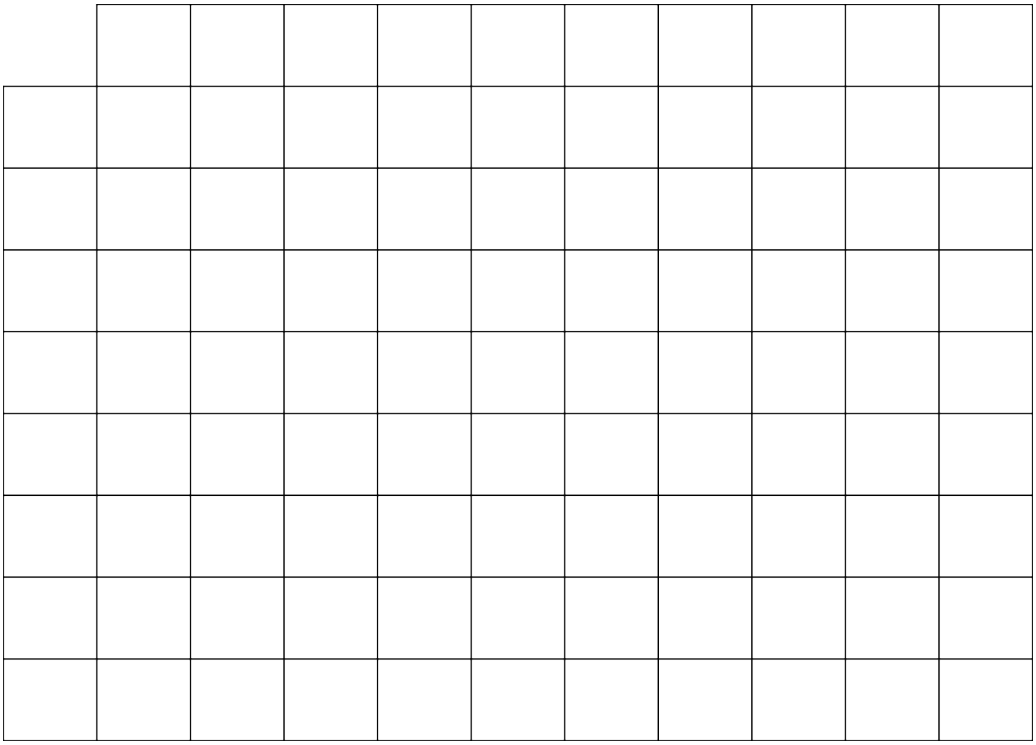
Matrix 2:



Matrix 3:



Backtracking matrix:



The possible alignments are: .....