Series 4

Genomics and bioinformatics - Week 4

October 11, 2011

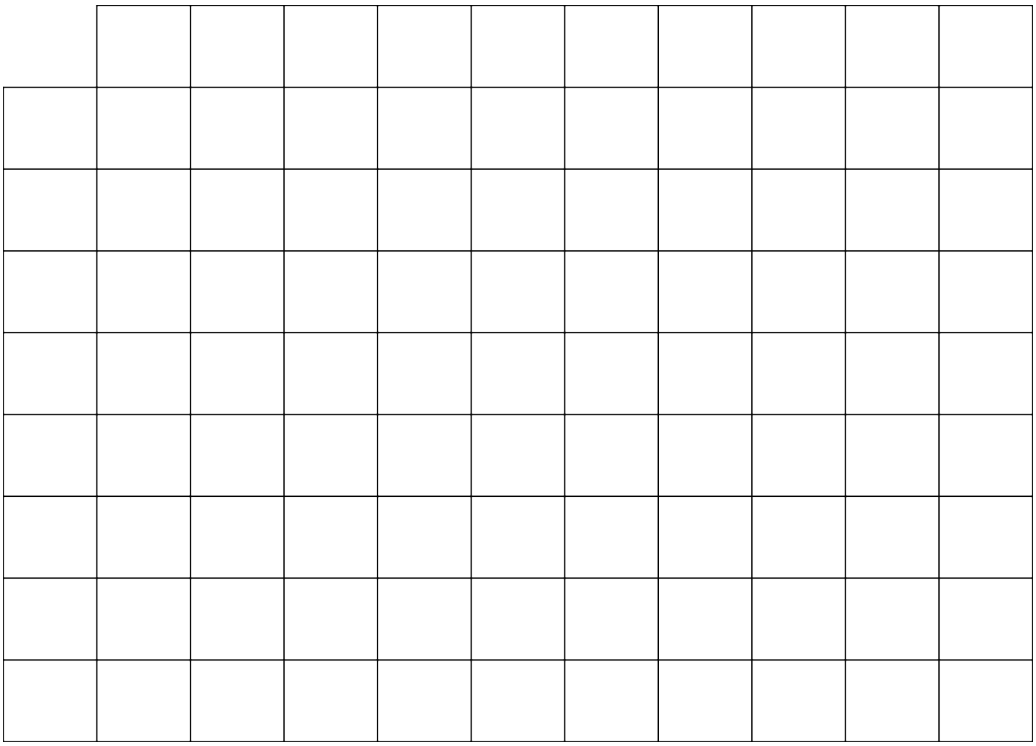
1 Sequence alignment

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`.



Solution:

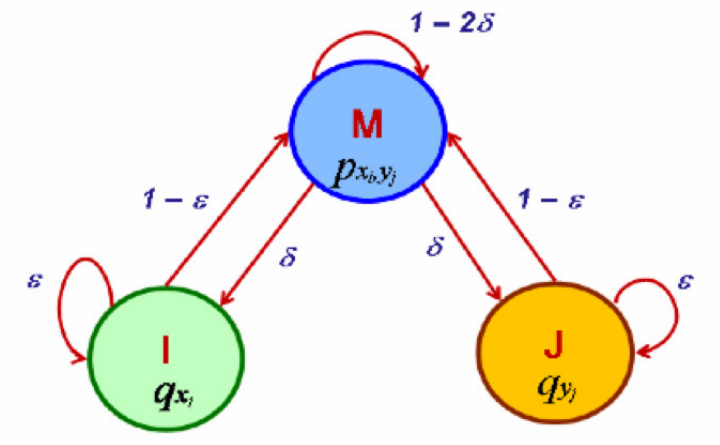


Figure 1: Pair Hidden Markov Model

2 Pair HMM

In this exercise, we will construct a pair Hiden Markov Model for the same sequences as in the first exercise and align them using the path with maximum probabiity.

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 during the exercise that the Viterbi algorithm is actually similar to the Needleman-Wunsch algorthm.

The algorithm goes through the three steps
Step1: Intialization
Step2: Recursion
Step3: Termination

A general pair HMM is shown in Figure 1. Pair HMM consists of the following parameters,

Three states: M, I, J .
State M matches one letter from each sequence
State I inserts a gap in the second sequence
State J inserts a gap in the first sequence

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

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