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

A general pair HMM is shown in Figure 1. It consists of the following parameters:

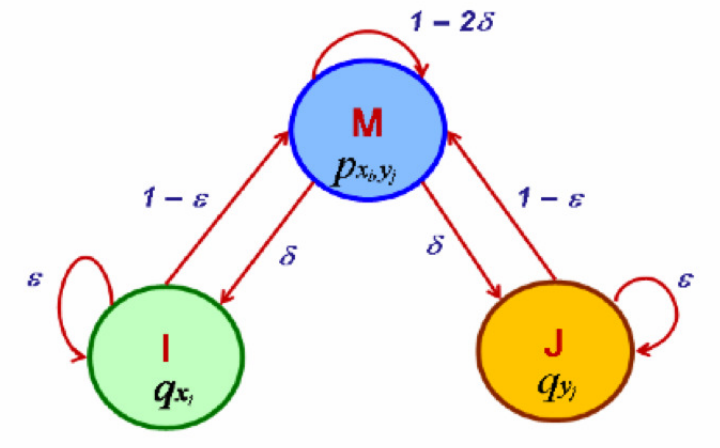

Figure 1: Pair Hidden Markov Model

- Three states: M, I, D.
  State M matches one letter from each sequence
  State I inserts a gap in the second sequence
  State D 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 = probability of 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

Find the maximum probability path, and deduce the best possible alignment.