Solutions - Series 4

Genomics and bioinformatics - Week 4

October 11, 2011

1 Sequence alignment

Match: `+1`, *Mismatch:* `-1`, *Gap:* `-2` Sequence 1: `GAATTCAGA` Sequence 2: `GGATCGA`.

1.1 Initialization

Create a matrix with `m + 1` columns and `n + 1` rows where `m` and `n` correspond to the sizes of Sequences 1 and 2, respectively.

1.2 Scoring

Using the given scoring scheme, at each cell, 3 scores are calculated:

- Upper neighbor score + Gap cost
- Left neighbor score + Gap cost
- Upper-left neighbor score + Match score (if nucleotides match), *OR* Upper-left neighbor score + Mismatch cost (if nucleotides do not match)

The highest score is retained and the arrow is labelled. Here is the resulting scoring matrix:

	–	G	A	A	T	T	C	A	G
–	0	-2	-4	-6	-8	-10	-12	-14	-16
G	-2	1	-1	-3	-5	-7	-9	-11	-13
G	-4	-1	0	-2	-4	-6	-8	-10	-10
A	-6	-3	0	1	-1	-3	-5	-7	-9
T	-8	-5	-2	-1	2	0	-2	-4	-6
C	-10	-7	-4	-3	0	1	1	-1	-3
G	-12	-9	-6	-5	-2	-1	0	0	0

1.3 Backtracking

The process of deduction of the best alignment from the score matrix is known as traceback. The traceback begins with the last cell to be filled, i.e. the bottom-right cell, and is completed when the first, i.e. the top-left cell of the matrix is reached. Several traceback paths are possible. The solution resulting in the best final score is selected to deduce the optimum alignment. It is possible to have more than one optimum alignment. A traceback path for the scoring matrix generated in Step 2 is highlighted below.

	–	G	A	A	T	T	C	A	G
–	0	-2	-4	-6	-8	-10	-12	-14	-16
G	-2	1	-1	-3	-5	-7	-9	-11	-13
G	-4	-1	0	-2	-4	-6	-8	-10	-10
A	-6	-3	0	1	-1	-3	-5	-7	-9
T	-8	-5	-2	-1	2	0	-2	-4	-6
C	-10	-7	-4	-3	0	1	1	-1	-3
G	-12	-9	-6	-5	-2	-1	0	0	0

1.4 Alignment

After backtracking, the optimal alignment is easy to recover using the following rule:
Left = Deletion , *Up* = Insertion, and *Diagonal* = Match
The optimum alignment for the two sequences `GAATTCAGA` and `GGATCGA` is,

```
G A A T T C A G
M M M M D M D M
G A A T T C A G
|   | |   |   |
G G A T - C - G
```

2 Pair HMM

The corresponding HMM is given by figure **??**.
where $p(x, y) = 0.125$ and $p(x, y) = 0.04$ for mismatch
$p(x, \_) = q(\_, y) = 0.25$
If we consider all probability values with respect to a random model in log-odds

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

$S(x, y) = 1$ for match and $s(x, y) = -1$ for mismatch
The we can construct the three matrices for VM, VX and VY by the Viterbi algorithm.
There are two possible optimal alignments.

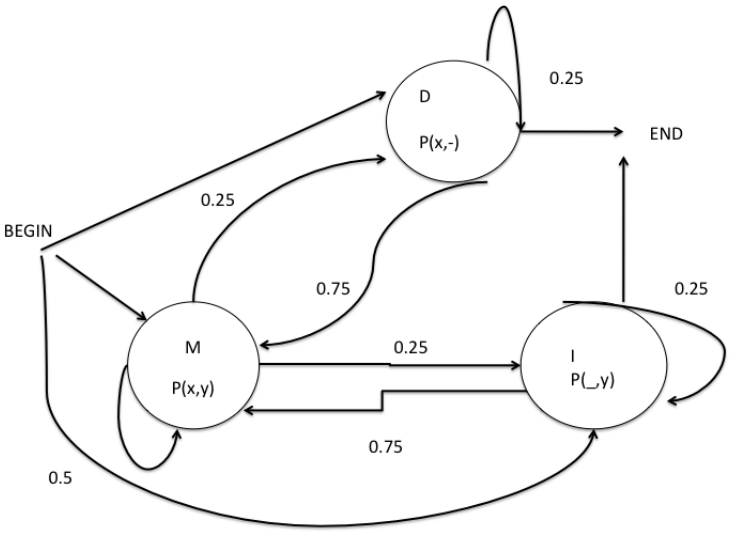

Figure 1: The HMM model for sequence alignment

$$\log V^M(i,j) = \max\begin{cases}\log V^M(i-1,j-1)+s(x_i,y_j)\\ \log V^I(i-1,j-1)+s(x_i,y_j)\\ \log V^J(i-1,j-1)+s(x_i,y_j)\end{cases}$$

$$\log V^I(i,j) = \max\begin{cases}\log V^M(i-1,j)-d\\ \log V^I(i-1,j)-e\end{cases}$$

$$\log V^J(i,j) = \max\begin{cases}\log V^M(i,j-1)-d\\ \log V^J(i,j-1)-e\end{cases}$$

Figure 2: Viterbi Algorithm

	------	G	A	A	T	T	C	A	G
-----	0	-∞	-∞	-∞	-∞	-∞	-∞	-∞	-∞
G	-∞	1	-3	-5	-7	-9	-11	-13	-13
G	-∞	-1	0	-2	-4	-6	-8	-10	-10
A	-∞	-5	0	1	-3	-5	-7	-7	-11
T	-∞	-7	-4	-1	2	0	-4	-6	-8
C	-∞	-9	-6	-3	-2	1	1	-3	-5
G	-∞	-9	-8	-5	-4	-1	0	0	0

Figure 3: The matrix for the match state

	------	G	A	A	T	T	C	A	G
-----	-∞	-2	-4	-6	-8	-10	-12	-14	-16
G	-∞	-∞	-1	-3	-5	-7	-9	-11	-13
G	-∞	-∞	-3	-2	-4	-6	-8	-10	-12
A	-∞	-∞	-7	-2	-1	-3	-5	-7	-9
T	-∞	-∞	-9	-6	-3	0	-2	-4	-6
C	-∞	-∞	-11	-8	-5	-4	-1	-1	-3
G	-∞	-∞	-11	-10	-7	-6	-3	-2	-2

Figure 4: Matrix for deletion state

Figure 5: Matrix for insertion state

	------	G	A	A	T	T	C	A	G
-----	- ∞	-∞	-∞	-∞	-∞	-∞	-∞	-∞	-∞
G	-2	-∞	-∞	-∞	-∞	-∞	-∞	-∞	-∞
G	-4	-1	-5	-7	-9	-11	-13	-15	-15
A	-6	-3	-2	-4	-6	-8	-10	-12	-12
T	-8	-5	-2	-1	-5	-7	-9	-9	-13
C	-10	-7	-4	-3	0	-2	-6	-8	-10
G	-12	-9	-6	-5	-2	-1	-1	-5	-7

	------	G	A	A	T	T	C	A	G
-----	0	-2	-4	-6	-8	-10	-12	-14	-16
G	-2	1	-1	-3	-5	-7	-9	-11	-13
G	-4	-1	0	-2	-4	-6	-8	-10	-10
A	-6	-3	0	1	-1	-3	-5	-7	-9
T	-8	-5	-2	-1	2	0	-2	-4	-6
C	-10	-7	-4	-3	0	1	1	-1	-3
G	-12	-9	-6	-5	-2	-1	0	0	0

Figure 6: Alignment by backtracking

```
G A A T T C A G      G A A T T C A G
| | | | | | |  OR  | | | | | | |
G G A T - C - G      G G A - T C - G
M M M M D M D M      M M M D M M D M
```

Figure 7: The alignments