

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,

```

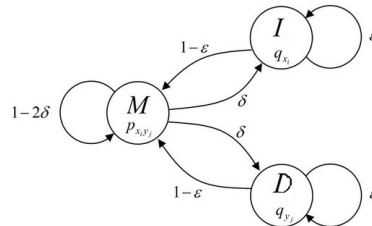
G A A T T C A G
M M M M D M D M
G A A T T C A G
|   |   |   |   |
G G A T - C - G

```

2 Pair HMM

2.1 Finding the Emission and Transition Probabilities

The corresponding HMM is given by the following figure .



If we consider all probability values with respect to a random model in log-odds

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

where $p(x)$ is the probability of choosing one nucleotide at random. So, $p(x) = 1/4$.

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

$p(x, y) = p(x|y) p(y)$ where $p(x|y)$ is the probability of x conditioned on y . Therefore,

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

So, we can find the emission probability values with this following equation

$$p(x|y) = 2^{S(x,y)} p(x)$$

The emission probability matrix is given by

Table 1: Emission Probability Matrix

–	A	C	G	T
A	1/2	1/8	1/8	1/8
C	1/8	1/2	1/8	1/8
G	1/8	1/8	1/2	1/8
T	1/8	1/8	1/8	1/2

and $q(x, y) = q(y) = 1/4$ and the Transition Probabilities are $\delta = 1/4$ and $\epsilon = \delta$.




2.2 Constructing the Three Matrices (VM, VD and VI)

Then we can construct the three matrices for VM, VD and VI by the Viterbi algorithm.

Matrix V_M :

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

Matrix V_D :

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

Matrix V_I :

	-----	G	A	A	T	T	C	A	G
-----	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
G	-2	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
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

2.3 Deducing the Alignments

The final matrix is formed by taking maximum value from the three elements of each matrix.

Alignment by backtracking:

	----	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

There are two possible optimal alignments.

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