Solutions - Series 3

Genomics and bioinformatics - Week 4

October 9, 2012

1 Global alignment

1.1 The table

Using the given scoring matrix M, one deduces the following rules to compute the 3 intermediate scores in each cell:

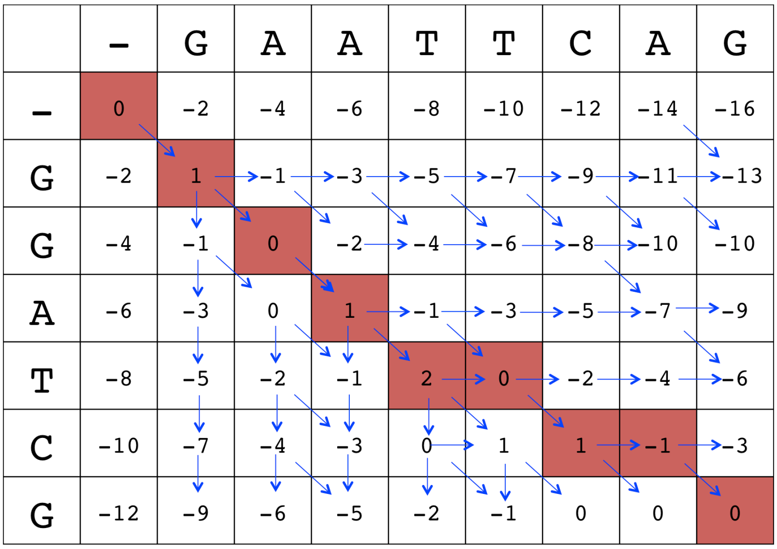
- Upper neighbour cell score - 2
- Left neighbour cell score - 2
- Upper-left neighbour cell score $\begin{cases} +1 \text{ if nucleotides are identical} \\ -1 \text{ if nucleotides are different} \end{cases}$

The highest score is then retained and the corresponding arrow is labelled. Here is the resulting table:

	–	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

1.2 Backtracking

One then applies the traceback process to the obtained table to deduce the best alignment between the two reads. The traceback begins with the bottom-right cell and is completed when the top-left cell of table is reached. Note that several traceback paths are possible and thus in general there may be more than one optimum alignment. The best alignments are given by the paths that have the maximum score. A traceback path 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

1.3 Alignment

An optimal alignment is easily obtained by using the following arrow rules:

Left/Right = Deletion, Up/Down = Insertion, and Diagonal = Match.

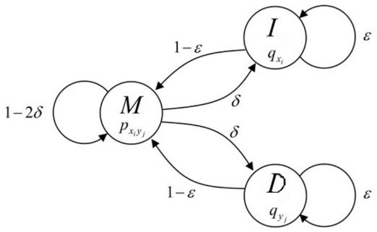
An optimal alignment is

```
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 HMM

2.1 Finding the Emission and Transition Probabilities

The corresponding HMM is described in the following figure:



First, from $d = -log_2(\delta)$ and $d = -2$ we deduce $\delta = 2^{-2} = \frac{1}{4} = \varepsilon$, giving the transition probabilities. Then we consider all probability values with respect to a random model in log-odds, i.e.:

$$S(x, y) = log_2 \frac{p(x, y)}{p(x)\, p(y)},$$

with $S(x, y) = 1$ for match, $S(x, y) = -1$ for mismatch, and $p(x)$ is the probability of choosing one nucleotide at random: $p(x) = p(y) = 1/4$.

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

$$S(x, y) = log_2 \frac{p(x|y)}{p(x)}$$

.

So we can find the emission probability values with the following equation:

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

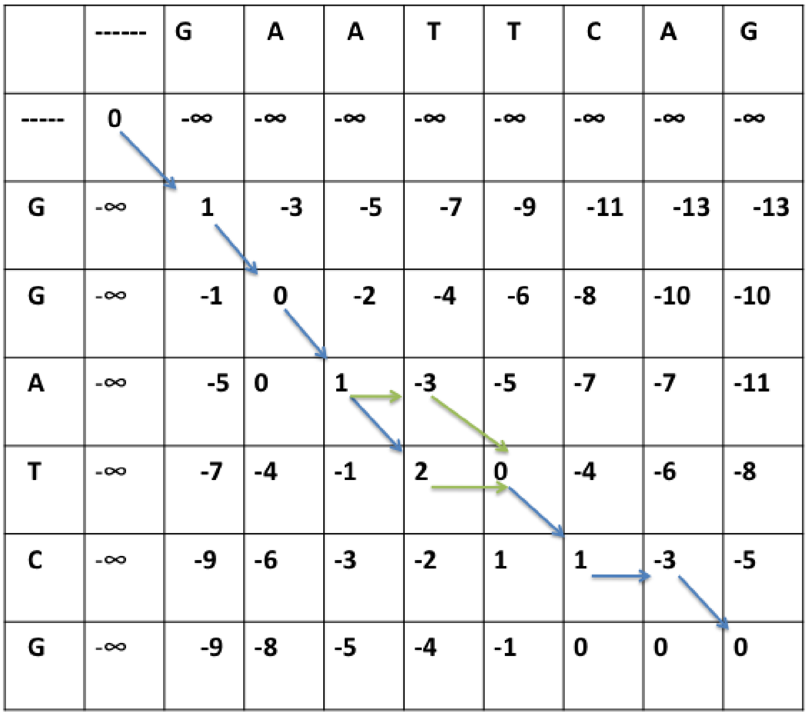
Finally, the emission probability matrix is given by

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

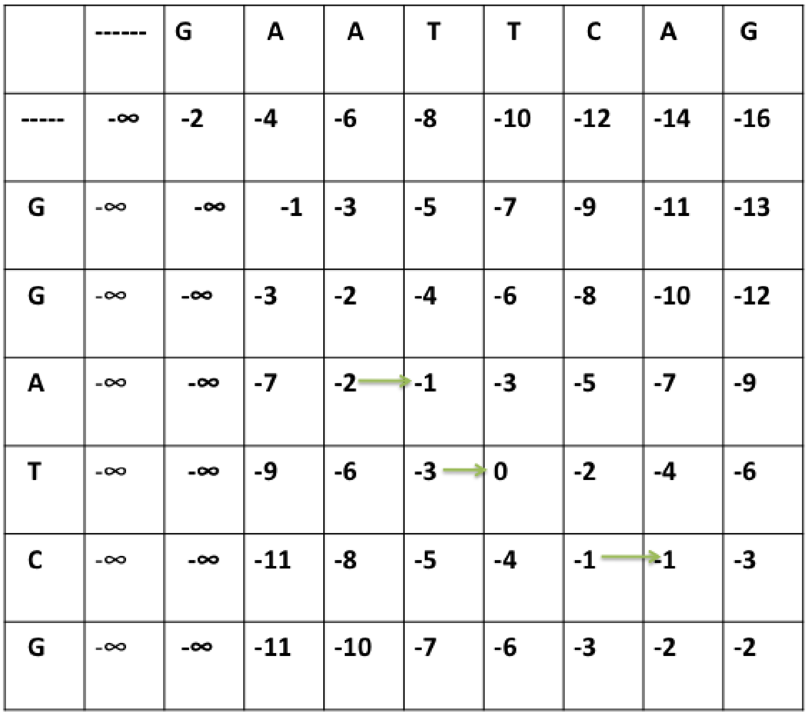
2.2 Constructing the Three Matrices (VM,VD and VI)

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



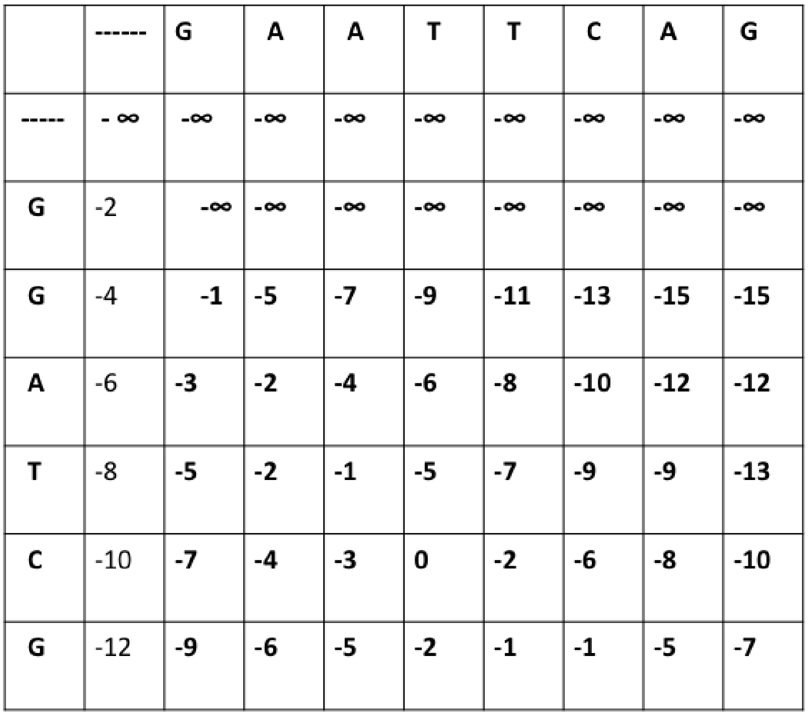
	------	G	A	A	T	T	C	A	G
-----	0	-∞	-∞	-∞	-∞	-∞	-∞	-∞	-∞
G	-∞	1	-3	-5	-7	-9	-11	-13	-13
G	-∞	-1	0	-2	-4	-6	-8	-10	-10
A	-∞	-5	0	1	-3	-5	-7	-7	-11
T	-∞	-7	-4	-1	2	0	-4	-6	-8
C	-∞	-9	-6	-3	-2	1	1	-3	-5
G	-∞	-9	-8	-5	-4	-1	0	0	0

Matrix V_D:



	------	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

Matrix V_I:

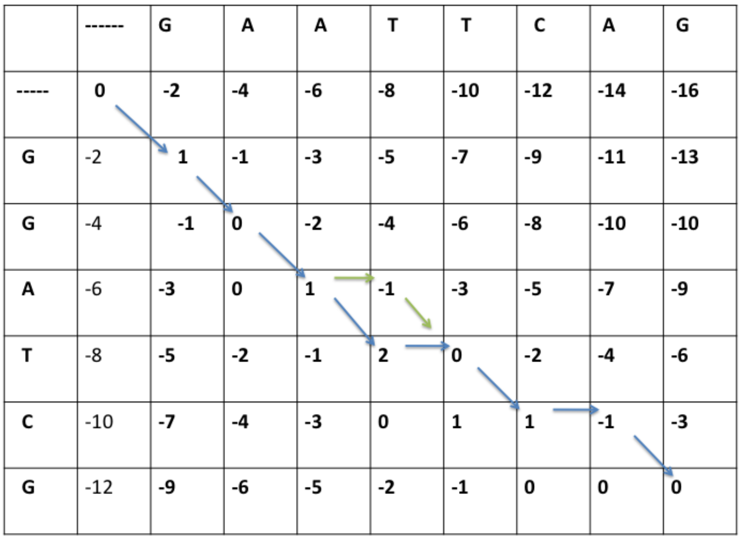


	------	G	A	A	T	T	C	A	G
-----	- ∞	-∞	-∞	-∞	-∞	-∞	-∞	-∞	-∞
G	-2	-∞	-∞	-∞	-∞	-∞	-∞	-∞	-∞
G	-4	-1	-5	-7	-9	-11	-13	-15	-15
A	-6	-3	-2	-4	-6	-8	-10	-12	-12
T	-8	-5	-2	-1	-5	-7	-9	-9	-13
C	-10	-7	-4	-3	0	-2	-6	-8	-10
G	-12	-9	-6	-5	-2	-1	-1	-5	-7

2.3 Deducing the Alignments

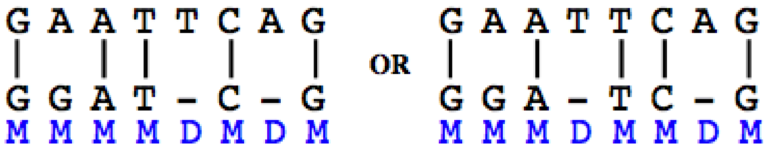
The final matrix is formed by taking maximum value from the three elements of each matrix.

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
```