Algorithms for Bioinformatics - lect 4

Francesco Penasa

March 9, 2020

1 Global sequence alignment

https://en.wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm

1.1 Exercise

1. We have got two sequences s_1 and s_2 and the following weights $match = 1$ $mismatch = -1$ $gap = -2$;
2. Put the sequences in a matrix like in table 1;
3. Init the first row and the first column like in table 2
4. Starting from the 0 on the top-left (with $i = 1$ and $j = 1$) find the max between the following equations:
   (a) $if\ M[i+1][label] == M[label][j+1]\ then\ M[i][j] + match$
   (b) $if\ M[i+1][label] \neq M[label][j+1]\ then\ M[i][j] + mismatch$
   (c) $M[i][j] + gap$
5. Repeat row to row
6. Since it is global the best alignment will result in the max score at the bottom right of the matrix.

Table 1: Init table $s_1 = GCATGCU$ $s_2 = GATTACA$

	-	G	C	A	T	G	C	U
-								
G								
A								
T								
T								
A								
C								
A								

Table 2: Init first row and first column $s_1 = GCATGCU$ $s_2 = GATTACA$

	-	G	C	A	T	G	C	U
-	0	-2	-4	-6	-8	-10	-12	-14
G	-2							
A	-4							
T	-6							
T	-8							
A	-10							
C	-12							
A	-14							

Table 3: First row iteration $s_1 = GCATGCU$ $s_2 = GATTACA$

	-	G	C	A	T	G	C	U
-	0	-2	-4	-6	-8	-10	-12	-14
G	-2	$\nwarrow 1$	$\leftarrow -1$	$\leftarrow -3$	$\leftarrow -5$	$\leftarrow\nwarrow -7$	$\leftarrow -9$	$\leftarrow -11$
A	-4							
T	-6							
T	-8							
A	-10							
C	-12							
A	-14							

Table 4: Second row iteration $s_1 = GCATGCU$ $s_2 = GATTACA$

	-	G	C	A	T	G	C	U
-	0	-2	-4	-6	-8	-10	-12	-14
G	-2	$\nwarrow 1$	$\leftarrow -1$	$\leftarrow -3$	$\leftarrow -5$	$\leftarrow\nwarrow -7$	$\leftarrow -9$	$\leftarrow -11$
A	-4	$\nwarrow\uparrow -1$	$\nwarrow 0$	$\nwarrow 0$	$\leftarrow -2$	$\leftarrow -4$	$\leftarrow -6$	$\leftarrow -8$
T	-6							
T	-8							
A	-10							
C	-12							
A	-14							

2 Local sequence alignment

`https://en.wikipedia.org/wiki/Smith%E2%80%93Waterman_algorithm`
`https://www.youtube.com/watch?v=QphFHG9tmOY`
As Needleman Wunsch for global alignment we use a table. There are only three differences:

1. The table initialization is done with all 0 as in table 5
2. We have gap penality to incentivize not starting the alignment.
3. We search for the max number in the table, we don't look only at the last cell.

Table 5: Smith Waterson Init first row and first column $s_1 = GCATGCU$ $s_2 = GATTACA$

	-	G	C	A	T	G	C	U
-	0	0	0	0	0	0	0	0
G	0							
A	0							
T	0							
T	0							
A	0							
C	0							
A	0							

3 Substitution Matrices

3.1 PAM

https://en.wikipedia.org/wiki/Point_accepted_mutation

3.2 BLOSUM

https://en.wikipedia.org/wiki/BLOSUM