CSC 483 - Needleman-Wunsch Example

doingg@union.edu

2026-01-21

Intructions for the Needleman-Wunsch Alignment Algorithm

Given a mismatch scoring scheme $s(x_i, y_j)$ where x_i and y_i are nucleotides (nt) and g is a gap penalty, the optimal aligmnet score of two nucleotide sequences X and Y, indexed by i and j respectively, is :

$$S(i,j) = \max \begin{cases} S(i-1, j-1) + s(x_i, y_j) & \text{(Match or Mismatch)} \\ S(i-1, j) + g & \text{(Gap in } Y\text{)} \\ S(i, j-1) + g & \text{(Gap in } X\text{)} \end{cases}$$

For example, setting $g = -2$ and

$$s(x_i, y_j) = \begin{cases} 1 & \text{if Match} \\ -1 & \text{if Mismatch} \end{cases}$$

fill each square as:

$$S(i,j) = \max \begin{cases} \begin{cases} S(i-1, j-1) + 1 & \text{(Match)} \\ S(i-1, j-1) - 1 & \text{(Mismatch)} \end{cases} \\ S(i-1, j) - 2 & \text{(Gap in } Y\text{)} \\ S(i, j-1) - 2 & \text{(Gap in } X\text{)} \end{cases}$$

Dynamic Programming Implementation

1. In the upper left corner, initiate with 0, then, move across columns and down rows. Fill in each square with:
    - the maximum value given the values at the previous indices
    - the previous indices from which the value builds (diagonal, left or up arrows)

	Seq X		**A**	**G**	**T**
Seq Y		$i = 0$	$i = 1$	$i = 2$	$i = 3$
	$j = 0$	0			
A	$j = 1$				
A	$j = 2$				
G	$j = 3$				
C	$j = 4$				final score

2. Once all squares are filled, **the bottom right corner has your best possible score**.

Example

Step 1. Fill in scores for all i and j

For $i = 1, j = 0$:

$$S(i,j) = \max \begin{cases} \begin{cases} S(0, j-1) + 1 = NA \text{ (indices cannot be < 0)} & \text{(Match)} \\ S(0, j-1) - 1 = NA \text{ (indices cannot be < 0)} & \text{(Mismatch)} \end{cases} \\ S(0,0) - 2 = S(0,0) - 2 = -2 & \text{(Gap in } Y\text{)} \\ S(1, j-1) - 2 = NA \text{ (indices cannot be < 0)} & \text{(Gap in } X\text{)} \end{cases}$$

$S(1,0) = -2$, from $S(0,0) = S(i-1, j)$, so mark with a left arrow

	Seq X		**A**	**G**	**T**
Seq Y		$i = 0$	$i = 1$	$i = 2$	$i = 3$
	$j = 0$	0	← -2		
A	$j = 1$				
A	$j = 2$				
G	$j = 3$				
C	$j = 4$				

Next, for $i = 2, j = 0$:

$$S(i,j) = \text{max} \begin{cases} \begin{cases} S(1,j-1) + 1 = NA \text{ (indices cannot be < 0)} & \text{(Match)} \\ S(1,j-1) - 1 = NA \text{ (indices cannot be < 0)} & \text{(Mismatch)} \end{cases} \\ S(1,0) - 2 = -2 - 2 = -4 & \text{(Gap in } Y\text{)} \\ S(2,j-1) - 2 = NA \text{ (indices cannot be < 0)} & \text{(Gap in } X\text{)} \end{cases}$$

$S(2,0) = -4$, from $S(1,0) = S(i-1,j)$, so mark with a left arrow

	Seq X		**A**	**G**	**T**
Seq Y		$i = 0$	$i = 1$	$i = 2$	$i = 3$
	$j = 0$	0	← -2	← -4	
A	$j = 1$				
A	$j = 2$				
G	$j = 3$				
C	$j = 4$				

Note, that $i = 3, j = 0$ follows similarly, as above. Give it a try, I've filled it in below.

Next, for $i = 0, j = 1$:

$$S(i,j) = \text{max} \begin{cases} \begin{cases} S(i-1,0) + 1 = NA \text{ (indices cannot be < 0)} & \text{(Match)} \\ S(i-1,0) - 1 = NA \text{ (indices cannot be < 0)} & \text{(Mismatch)} \end{cases} \\ S(i-1,1) - 2 = NA \text{ (indices cannot be < 0)} & \text{(Gap in } Y\text{)} \\ S(0,0) - 2 = 0 - 2 = -2 & \text{(Gap in } X\text{)} \end{cases}$$

$S(0,1) = -2$ derived from $S(0,0) = S(i,j-1)$(up arrow)

	Seq X		**A**	**G**	**T**
Seq X		$i = 0$	$i = 1$	$i = 2$	$i = 3$
	$j = 0$	0	← -2	← -4	← -6
A	$j = 1$	↑ -2			
A	$j = 2$				
G	$j = 3$				
C	$j = 4$				

The rest of the column ($i = 0$) follows similarly.

To fill in the center squares, we start with $i = 1, j = 1$:

$$S(i,j) = \text{max} \begin{cases} \begin{cases} S(0,0) + 1 = 0 + 1 = 1 & \text{(Match)} \\ S(0,0) - 1 = NA \text{ (not a mismatch, A=A)} & \text{(Mismatch)} \end{cases} \\ S(0,1) - 2 = -2 - 2 = -4 & \text{(Gap in } Y\text{)} \\ S(1,0) - 2 = -2 - 2 = -4 & \text{(Gap in } X\text{)} \end{cases}$$

$S(1,1) = 1$ from $S(0,0) = S(i-1, j-1)$, so mark with a diagonal arrow

	Seq X		**A**	**G**	**T**
Seq Y		$i = 0$	$i = 1$	$i = 2$	$i = 3$
	$j = 0$	0	← -2	← -4	← -6
A	$j = 1$	↑ -2	↖ 1		
A	$j = 2$	↑ -4			
G	$j = 3$	↑ -6			
C	$j = 4$	↑ -8			

Continuing, for $i = 2, j = 1$:

$$S(i,j) = \text{max} \begin{cases} \begin{cases} S(1,0) + 1 = NA \text{ (not a match, G!=A)} & \text{(Match)} \\ S(1,0) - 1 = -2 - 1 = -3 & \text{(Mismatch)} \end{cases} \\ S(1,1) - 2 = 1 - 2 = -1 & \text{(Gap in } Y\text{)} \\ S(2,0) - 2 = -4 - 2 = -6 & \text{(Gap in } X\text{)} \end{cases}$$

$S(2,1) = -1$ derived from $S(1,1) = S(i-1, j)$(left arrow)

	Seq X		**A**	**G**	**T**
Seq Y		$i = 0$	$i = 1$	$i = 2$	$i = 3$
	$j = 0$	0	← -2	← -4	← -6
A	$j = 1$	↑ -2	↖ 1	← -1	
A	$j = 2$	↑ -4			
G	$j = 3$	↑ -6			
C	$j = 4$	↑ -8			

Something to look out for, continuing, for $i = 1, j = 2$:

$$S(i,j) = \text{max} \begin{cases} \begin{cases} S(0,1) + 1 = -2 + 1 = -1 & \text{(Match)} \\ S(0,1) - 1 = NA & \text{(Mismatch)} \end{cases} \\ S(0,2) - 2 = -4 - 2 = -6 & \text{(Gap in } Y\text{)} \\ S(1,1) - 2 = 1 - 2 = -1 & \text{(Gap in } X\text{)} \end{cases}$$

$S(1,2) = -1$ could be from $S(0,1) = S(i-1, j-1)$ (diagonal arrow)…

OR….could be from $S(1,1) = S(i, j-1)$ (up arrow).

	Seq X		**A**	**G**	**T**
Seq Y		$i = 0$	$i = 1$	$i = 2$	$i = 3$
	$j = 0$	0	← -2	← -4	← -6
A	$j = 1$	↑ -2	↖ 1	← -1	
A	$j = 2$	↑ -4	↑↖ -1		
G	$j = 3$	↑ -6			
C	$j = 4$	↑ -8			

Step 2. Find the final score in the bottom, right cell

$$S(X,Y) = -1$$

	Seq X		**A**	**G**	**T**
Seq Y		$i = 0$	$i = 1$	$i = 2$	$i = 3$
	$j = 0$	0	← -2	← -4	← -6
A	$j = 1$	↑ -2	↖ 1	← -1	← -3
A	$j = 2$	↑ -4	↑↖ -1	↖ 0	←↖ -2
G	$j = 3$	↑ -6	↑ -3	↖ 0	↖ -1
C	$j = 4$	↑ -8	↑ -5	↑ -2	↖ -1

Step 3. Backtrace following the arrows for the optimal alignment

3. To figure out the alignment that leads to that best possible score, back-trace from the bottom right to the upper left, following the arrows of highest score attribution. Starting with the last nucleotide (3' end), pre-pend the nucleotides as you go.
    - if diagonal arrow, record both nt at those indices (they may or may not match)

- if left arrow, record the nt from the top sequence (Seq X, i indices) and a gap in the bottom sequence (Seq Y, j indices)
- if up arrow, record the nt from the bottom sequence (Seq Y, j indices) and a gap in the top sequence (Seq X, i indices)

Continuing the example

Starting in the bottom right cell $i = 3, j = 4$, given the diagonal arrow, we begin our backward aligment by recording both nucleotides at these indices:

```
Seq X:   T
         |
Seq Y:   C
```

This brings us to $i = 2, j = 3$, with another diagonal arrow so both nt are added again:

```
Seq X:   GT
         ||
Seq Y:   GC
Match:   *
```

Then to $i = 1, j = 2$, with a diagonal arrow **and** a left arrow, so we have two options:

Option 1, diagonal arrow, add both nt:

```
Seq X:   AGT
         |||
Seq Y:   AGC
Match:   **
```

Option 2, up arrow, add a gap to Seq2 (i indices):

```
Seq X:   –GT
         |||
Seq Y:   AGC
Match:    *
```

Finishing out the trace, we can choose an option (maybe follow the arrow that points to the higher score) or we can follow both options:

Option 1, now we are at $i = 0, j = 1$ with an up arrow so add a gap to Seq2:

```
Seq X:   –AGT
         ||||
Seq Y:   AAGC
Match:    **
```

Option 2, now we are at $i = 1, j = 1$ with a diagonal arrow so both nt get added:

Seq X: A–GT
| | | |
Seq Y: AAGC
Match: * *

Note that both alignments produces a score of -1 and both have 2 mismatches. If we broke this tie by following the path that always chooses the higher score, we would end with Option 2 since $[S(1,1) = 1] > [S(0,1) = -2]$

You can see this using the online tool by setting your sequences to

Seq1 (Y): AAGC
Seq2 (X): AGT

https://bioboot.github.io/bimm143_W20/class-material/nw/

If you "Clear Path" and manually trace Option 1, however, you will get the same score.

Practice

Try this again with new sequences (any length), check your results using the online tool.

	Seq X									
Seq Y		$i = 0$	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$
	$j = 0$									
	$j = 1$									
	$j = 2$									
	$j = 3$									
	$j = 4$									
	$j = 5$									
	$j = 6$									
	$j = 7$									
	$j = 8$									