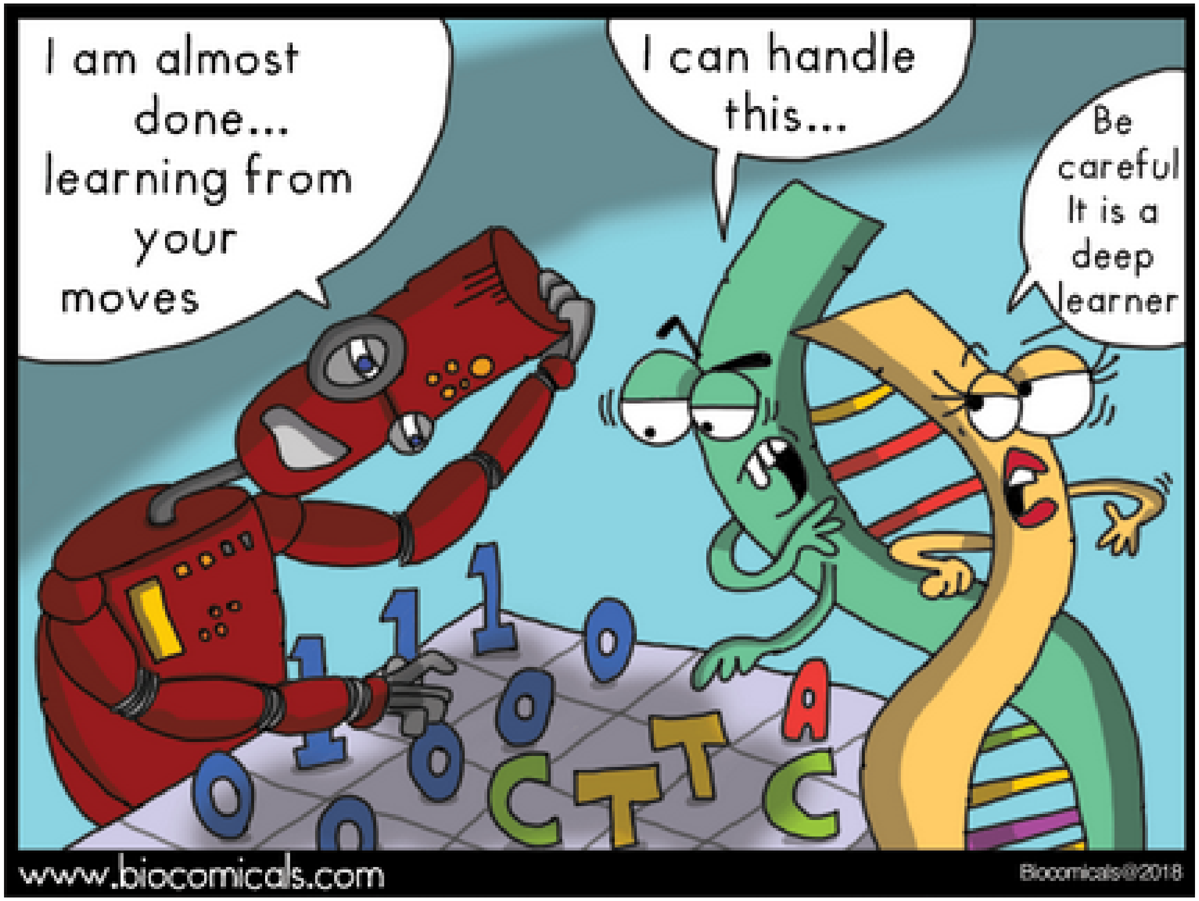
CS123A Bioinformatics Module 3 – Week 7 – Presentation 2

Leonard Wesley

Computer Science Dept

San Jose State Univ



Agenda

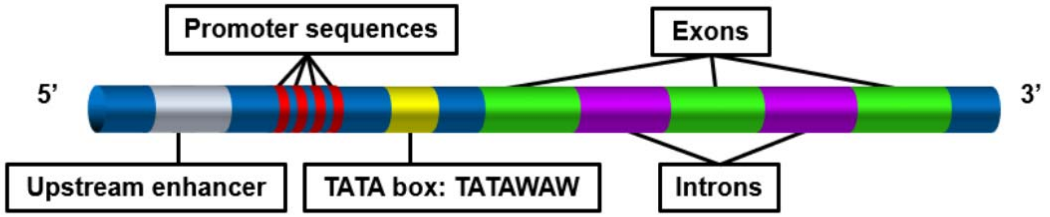
- Midterm Study Guide

- Review of Needleman-Wunsch global alignment algorithm

Variations Of Local & Global Alignment Algorithms

- Variations/modifications of local & global alignment algorithms can be made to fit the biological situation at hand.

- For example,



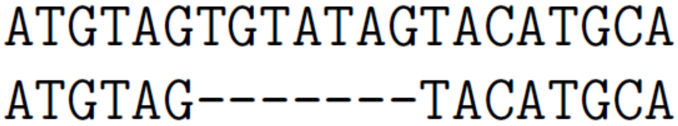
- Might need to align sequences such that gaps are not inserted in the TATAA box area of a reference sequence.

- Various ways to accomplish this: (1) Initial insertion & gap extension penalties; (2) Scoring pre-gapped sub-sequences (in-class exercise); (3) possibly more ways where some are private/proprietary.

Initial Insertion & Gap Extension Penalties (a.k.a. Affine Gap Penalties)

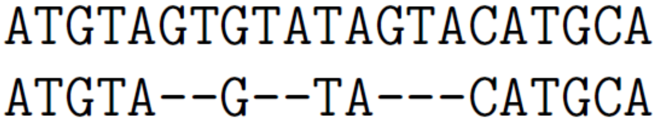
- $g(n_{gap})$ = $- I - E(n - 1)$

- Affine gaps favor the alignment:



```
ATGTAGTGTATAGTACATGCA
ATGTAG-------TACATGCA
```

- Over the alignment

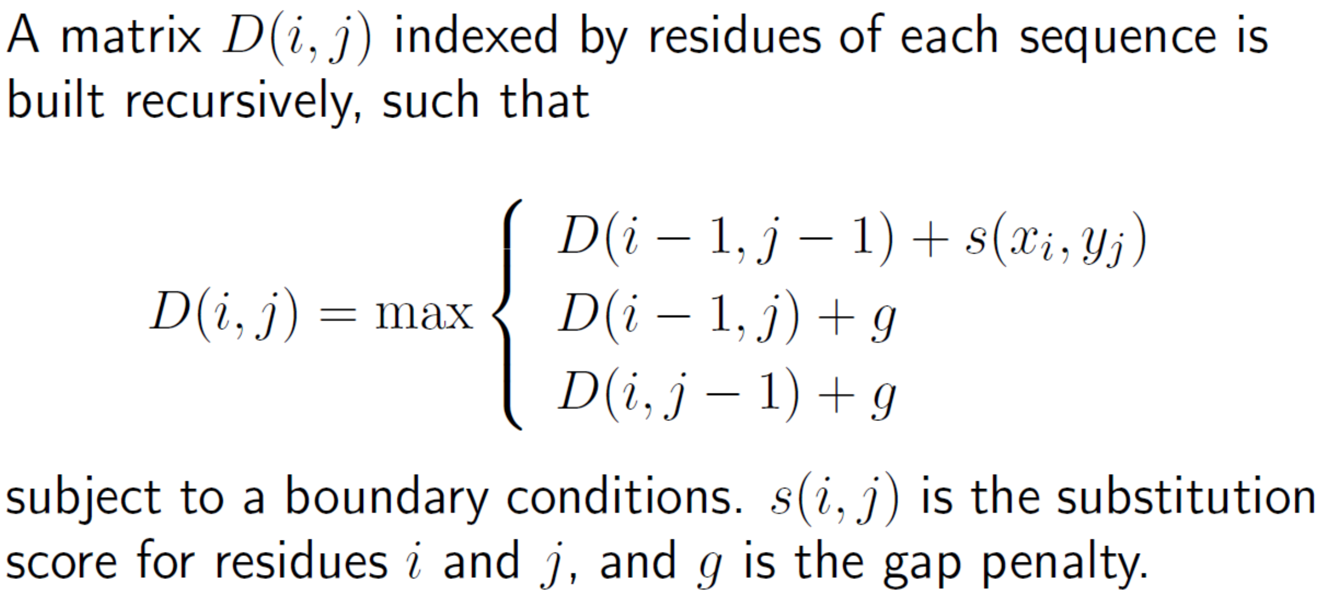


```
ATGTAGTGTATAGTACATGCA
ATGTA--G--TA---CATGCA
```

Classic Needleman-Wunsch Alignment

Example Alignment Using BLOSUM62 Matrix

- Align the amino acid residues SEND and AND using



A matrix $D(i,j)$ indexed by residues of each sequence is built recursively, such that

$$D(i,j) = \max \begin{cases} D(i-1,j-1) + s(x_i, y_j) \\ D(i-1,j) + g \\ D(i,j-1) + g \end{cases}$$

subject to a boundary conditions. $s(i,j)$ is the substitution score for residues i and j, and g is the gap penalty.

The Scoring Matrix …

The cells of the score matrix are labelled $C(i,j)$ where $i = 1, 2, ..., N$ and $j = 1, 2, ..., M$

Notice no gaps pre-inserted … yet

		S	E	N	D
	C(1,1)	C(1,2)	C(1,3)	C(1,4)	C(1,5)
A	C(2,1)	C(2,2)	C(2,3)	C(2,4)	C(2,5)
N	C(3,1)	C(3,2)	C(3,3)	C(3,4)	C(3,5)
D	C(4,1)	C(4,2)	C(4,3)	C(4,4)	C(4,5)

	S	E	N	D

The Scoring Matrix … *(cont.)*

- The first row and the first column of the score and traceback matrices are filled during the initialization.

		S	E	N	D
	0	−10	−20	−30	−40
A	−10				
N	−20				
D	−30				

	S	E	N	D
done	left	left	left	left
up				
up				
up				

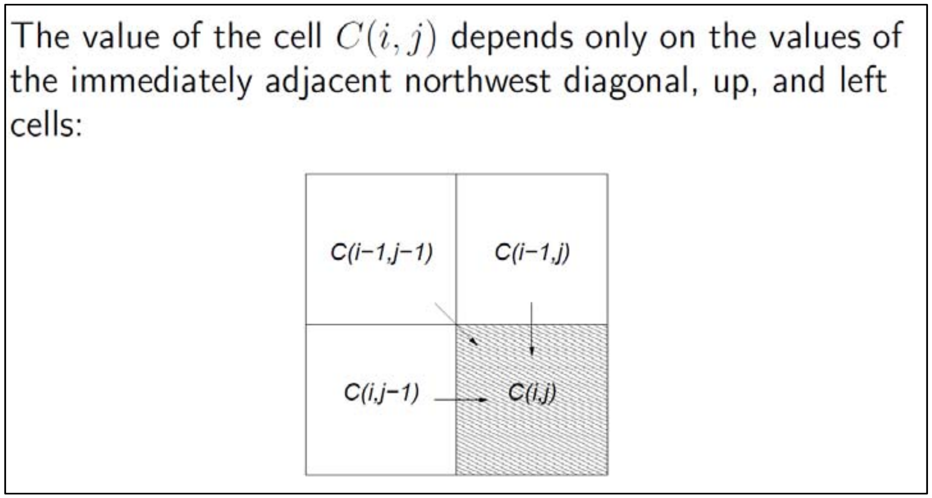
The Scoring Matrix … *(cont.)*

The score matrix cells are filled by row starting from the cell $C(2,2)$

The score of any cell $C(i,j)$ is the maximum of:

$$\begin{aligned} q_{diag} &= C(i-1,j-1) \; + \; S(i,j) \\ q_{up} &= C(i-1,j) \; + \; g \\ q_{left} &= C(i,j-1) \; + \; g \end{aligned}$$

where $S(i,j)$ is the substitution score for letters i and j, and g is the gap penalty.



The value of the cell $C(i,j)$ depends only on the values of the immediately adjacent northwest diagonal, up, and left cells:

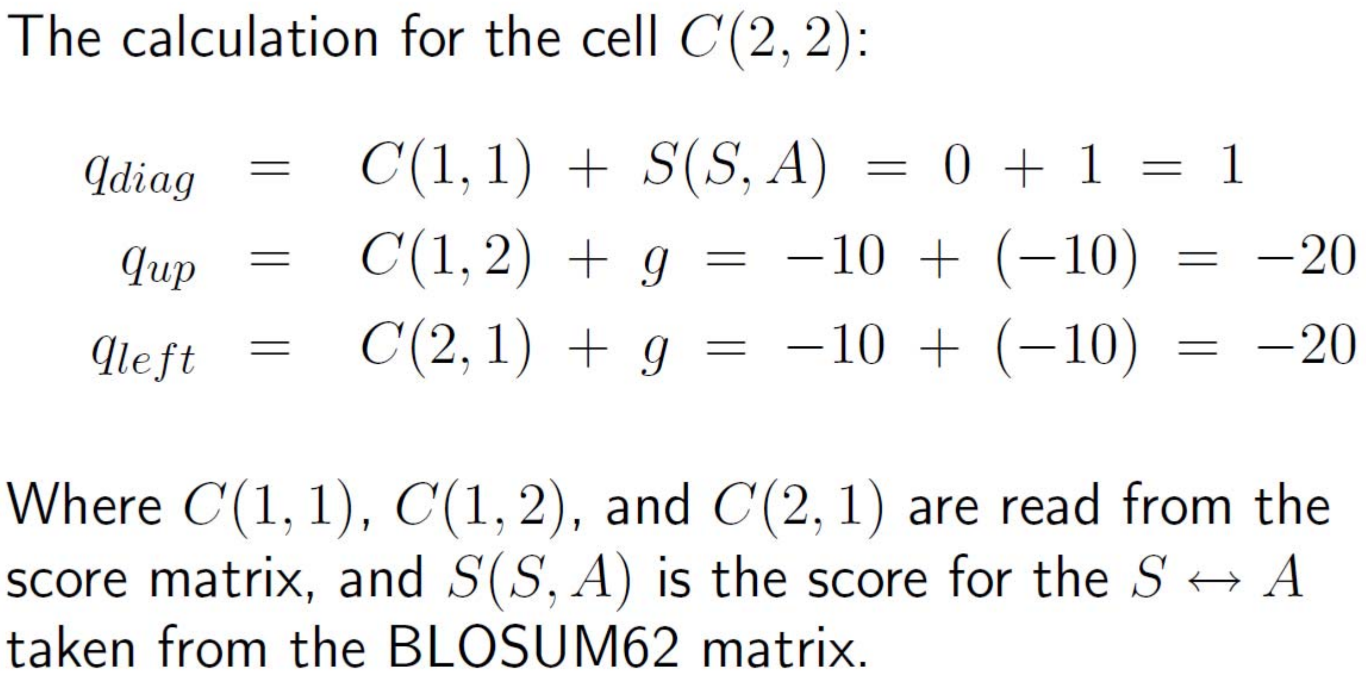
The Scoring Matrix *(cont.)*

The first step is to calculate the value of $C(2,2)$:

		S	E	N	D
	0	−10	−20	−30	−40
A	−10	**?**			
N	−20				
D	−30				

	S	E	N	D
done	left	left	left	left
up	**?**			
up				
up				

The Value of C(2,2)



The calculation for the cell $C(2,2)$:

$$\begin{aligned} q_{diag} &= C(1,1) + S(S,A) = 0 + 1 = 1 \\ q_{up} &= C(1,2) + g = -10 + (-10) = -20 \\ q_{left} &= C(2,1) + g = -10 + (-10) = -20 \end{aligned}$$

Where $C(1,1)$, $C(1,2)$, and $C(2,1)$ are read from the score matrix, and $S(S,A)$ is the score for the $S \leftrightarrow A$ taken from the BLOSUM62 matrix.

Filling The Score And Traceback Matrix

For the score matrix $C(2,2) = q_{diag}$ which is 1. The corresponding cell of the traceback matrix is "diag":

		S	E	N	D
	0	−10	−20	−30	−40
A	−10	**1**			
N	−20				
D	−30				

		S	E	N	D
	done	left	left	left	left
	up	**diag**			
	up				
	up				

The Process Is Recursive … C(3,2)

		S	E	N	D
	0	−10	−20	−30	−40
A	−10	**1**	**?**		
N	−20				
D	−30				

	S	E	N	D
done	left	left	left	left
up	**diag**	**?**		
up				
up				

Calculation for C(3,2)

The calculation for the cell $C(2,3)$

$$\begin{aligned} q_{diag} &= C(1,2) + S(E,A) = -10 + -1 = -11 \\ q_{up} &= C(1,3) + g = -20 + (-10) = -30 \\ q_{left} &= C(2,2) + g = 1 + (-10) = -9 \end{aligned}$$

Thus $C(3,2) = -9$ and the corresponding cell of the traceback matrix is "left".

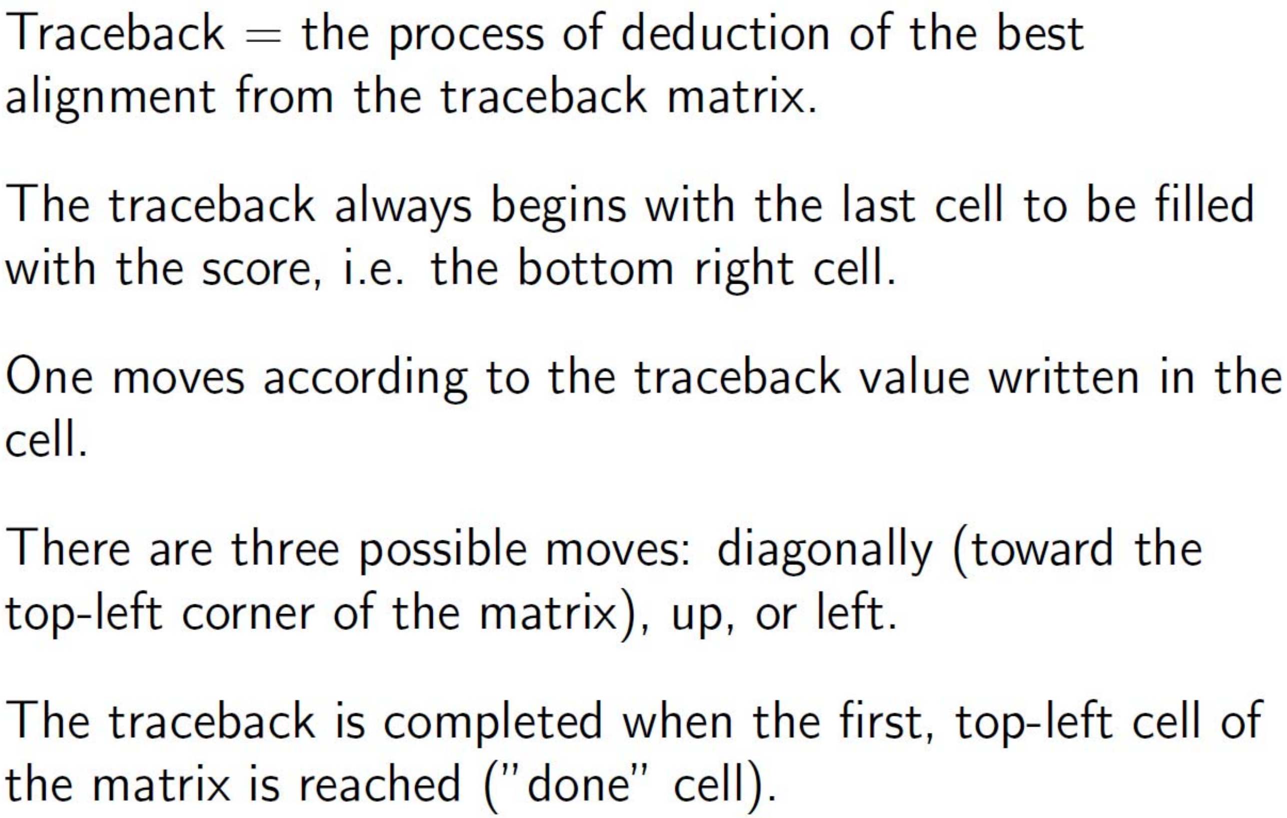
The Final Score & Traceback Matrices

After all cells are filled, the score and traceback matrices are:

		S	E	N	D
	0	-10	-20	-30	-40
A	-10	1	-9	-19	-29
N	-20	-9	1	-3	-13
D	-30	-19	-7	2	3

	S	E	N	D
done	left	left	left	left
up	diag	left	left	left
up	diag	diag	diag	left
up	up	diag	diag	diag

Traceback Process



Traceback = the process of deduction of the best alignment from the traceback matrix.

The traceback always begins with the last cell to be filled with the score, i.e. the bottom right cell.

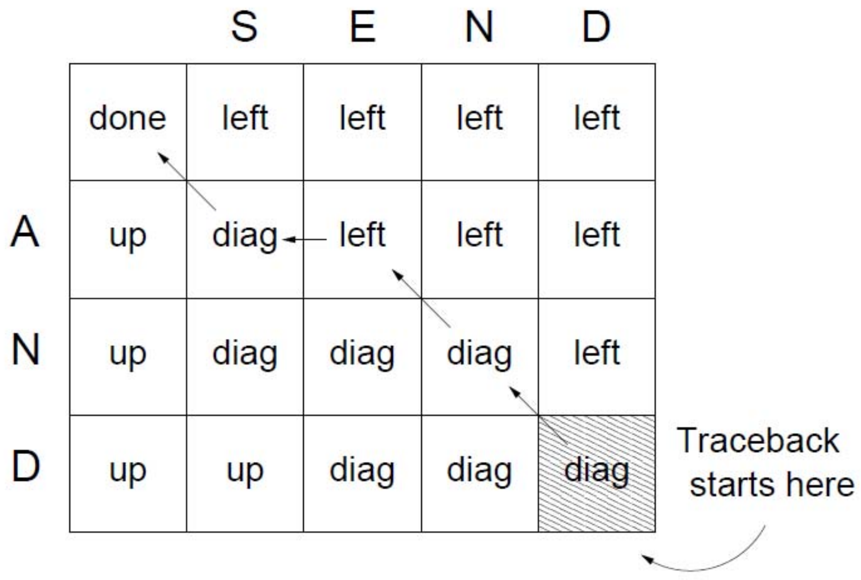
One moves according to the traceback value written in the cell.

There are three possible moves: diagonally (toward the top-left corner of the matrix), up, or left.

The traceback is completed when the first, top-left cell of the matrix is reached (”done” cell).

Traceback Path

The traceback performed on the completed traceback matrix:



		S	E	N	D
	done	left	left	left	left
A	up	diag	left	left	left
N	up	diag	diag	diag	left
D	up	up	diag	diag	diag

Traceback starts here

The Best Alignment

The alignment is deduced from the values of cells along the traceback path, by taking into account the values of the cell in the traceback matrix:

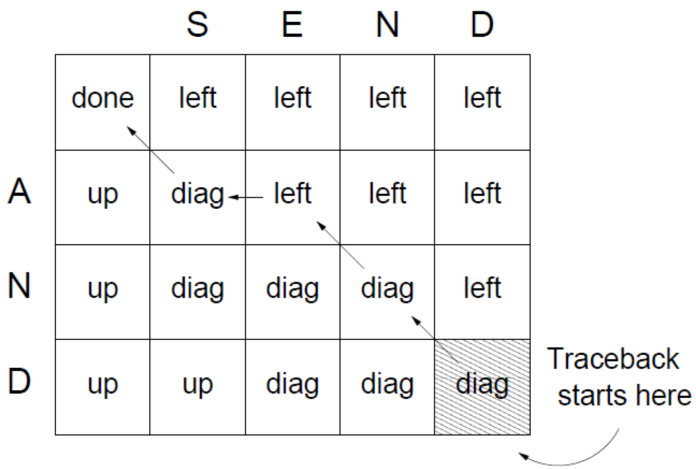
- *diag* – the letters from two sequences are aligned
- *left* – a gap is introduced in the left sequence
- *up* – a gap is introduced in the top sequence

Sequences are aligned backwards.

Traceback Step 1

The first cell from the traceback path is "diag" implying that the corresponding letters are aligned:

```
D
D
```



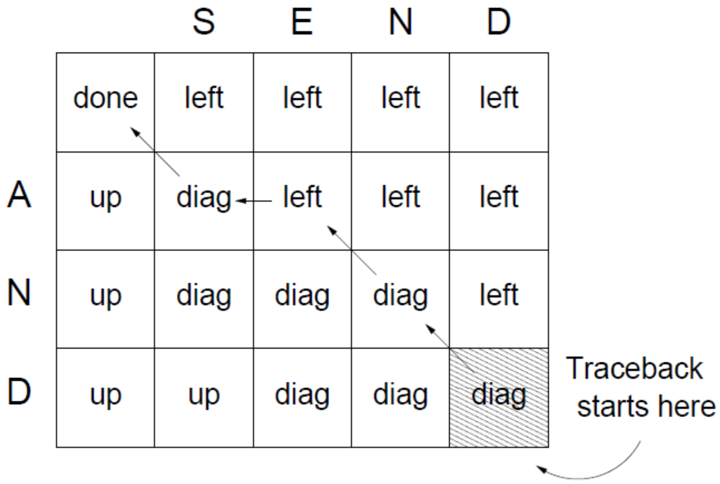
		S	E	N	D
	done	left	left	left	left
A	up	diag	left	left	left
N	up	diag	diag	diag	left
D	up	up	diag	diag	diag

Traceback starts here

Traceback Step 2

The second cell from the traceback path is also "diag" implying that the corresponding letters are aligned:

```
ND
ND
```



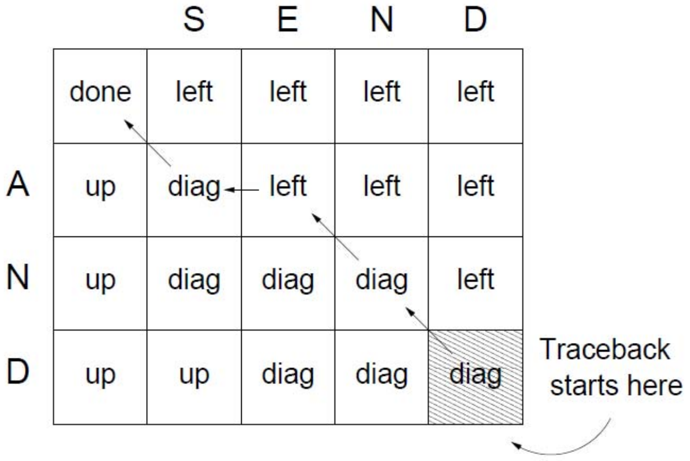
		S	E	N	D
	done	left	left	left	left
A	up	diag	left	left	left
N	up	diag	diag	diag	left
D	up	up	diag	diag	▨ diag

Traceback starts here

Traceback Step 3

The third cell from the traceback path is "left" implying the gap in the left sequence (i.e. we stay on the letter A from the left sequence):

```
END
-ND
```



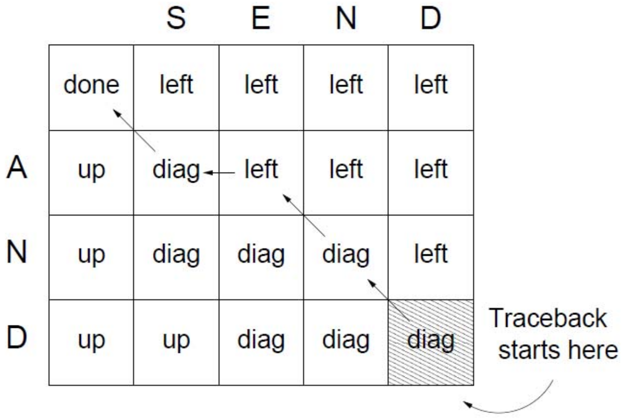
	S	E	N	D
	done	left	left	left	left
A	up	diag	left	left	left
N	up	diag	diag	diag	left
D	up	up	diag	diag	diag

Traceback starts here

Traceback Step 4

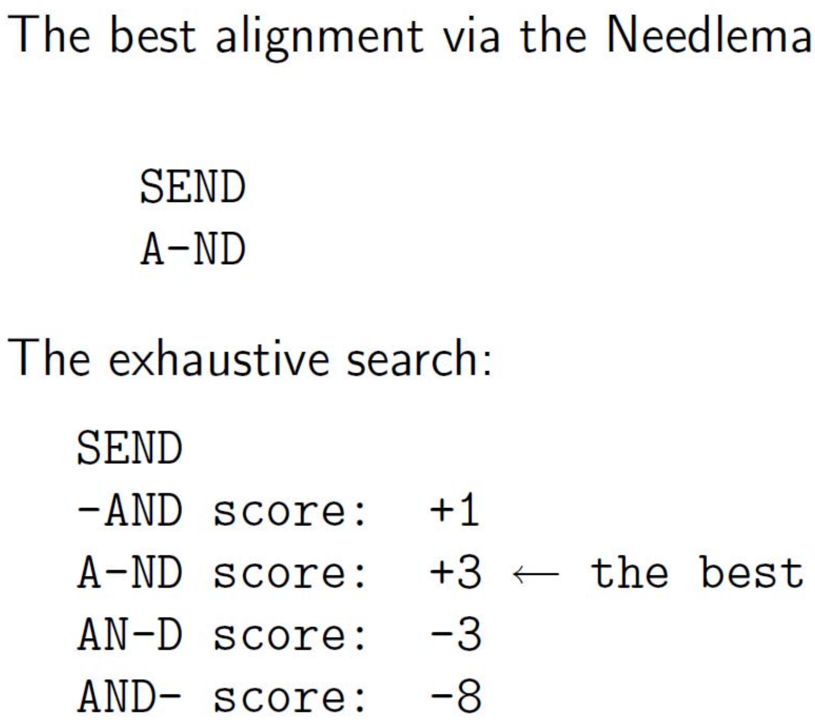
The fourth cell from the traceback path is also ”diag” implying that the corresponding letters are aligned. We consider the letter `A` again, this time it is aligned with `S`:

```
SEND
A-ND
```



	S	E	N	D
	done	left	left	left
A	up	diag	left	left	left
N	up	diag	diag	diag	left
D	up	up	diag	diag	diag

Best Alignment



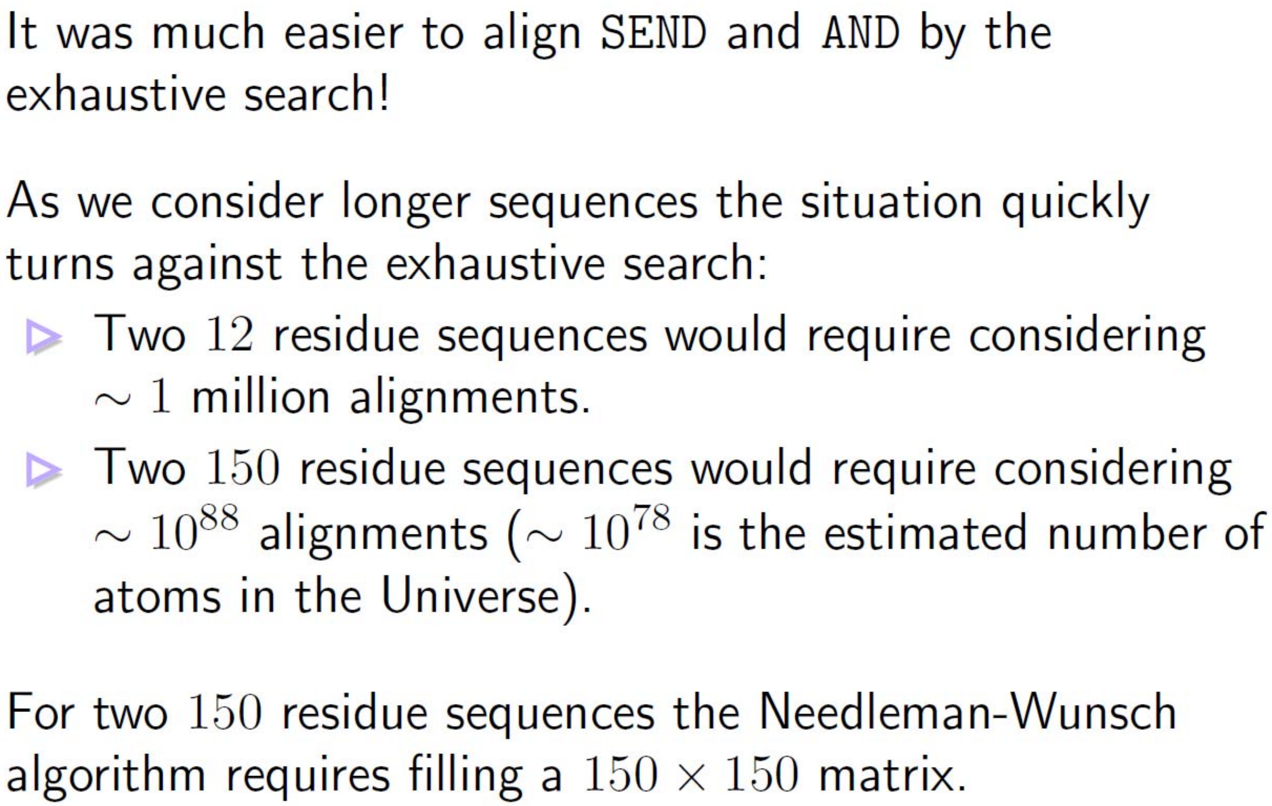
The best alignment via the Needleman-Wunsch algorithm:

```
SEND
A-ND
```

The exhaustive search:

```
SEND
-AND score:  +1
A-ND score:  +3 ← the best
AN-D score:  -3
AND- score:  -8
```

Some Observations



It was much easier to align `SEND` and `AND` by the exhaustive search!

As we consider longer sequences the situation quickly turns against the exhaustive search:

- Two 12 residue sequences would require considering ~ 1 million alignments.
- Two 150 residue sequences would require considering $\sim 10^{88}$ alignments ($\sim 10^{78}$ is the estimated number of atoms in the Universe).

For two 150 residue sequences the Needleman-Wunsch algorithm requires filling a 150×150 matrix.

Summary

The alignment is over the entire length of two sequences: the traceback starts from the lower right corner of the traceback matrix, and completes in the upper left cell of the matrix.

The Needleman-Wunsch algorithm works in the same way regardless of the length or complexity of sequences, and *guarantees* to find the best alignment.

The Needleman-Wunsch algorithm is appropriate for finding the best alignment of two sequences which are (*i*) of the similar length; (*ii*) similar across their entire lengths.

In-Class Exercise

- Globally align TODAY and DAYTIME

- For the midterm be able to compute the score for
  - Pre-gapped sequences, e.g., I S_LAND
    WAYLAND
  - Or an non pre-gapped sequence, e.g., ISLAND
    WAYLAND

- Initial gap and gap extension penalty alignment questions will not be on the midterm