# The way to Needleman-Wunsch

#### The goal

- Given two sequences A and B:
  - 1. compute the alignment score
  - 2. build the alignment representation

#### • Example:

G-ATTACAAAA GCAT-GCAAAA

#### O.Longest common sub sequence

- Given two sequences find the longest common subsequence
- Example
  - Seq1 CATISNOTADOG
  - Seq2 NOTABENE
  - Result NOTA

#### Naïve approach

- Input
  - Seq1 CATISNOTADOG
  - Seq2 NOTABENE
- Step1:

Create all the subsequences of Seq2: N, NO, NOT ... OT, OTA ... NE, E

• Step2

For each of them check if it is also a subsequence of Seq1 Keep the longest found

#### Towards improvements (1)

- 1. Build binary matrix of matches
- 2. Count the long streak alongside diagonals

	С	Α	Т	I	S	N	0	Т	Α	D	0	G
N	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	Q	1	0	0	0	1	0
Т	0	0	1	0	0	0	8	1	0	0	0	0
Α	0	0	0	0	0	0	0	0	1	0	0	0
В	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	1	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0

Need to scan the diagonals

0

00

010

• • •

1111000

#### Towards improvements (2)

- 1. Accumulate number of mismatches
- 2. Break the streak with 0 in case of a mismatch

	С	Α	T	I	S	N	0	T	Α	D	0	G
N	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	8	2	0	0	0	1	0
Т	0	0	1	0	0	0	8	3	0	0	0	0
Α	0	0	0	0	0	0	0	0	4	0	0	0
В	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	1	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0

Need to scan only for the maximum value

#### Towards improvements (3)

- 1. Allow mismatches
- 2. No longer "common subsequence" but rather "common similar subsequence"

	С	Α	Т	I	S	N	0	Т	Α	D	0	G
N	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	2	0	0	0	1	0
T	0	0	1	0	0	0	0	3	0	0	0	0
A	0	1	0	0	0	0	0	0	4	0	0	0
F	0	0	0	0	0	0	0	0	0	3	0	0
0	0	0	0	0	0	0	1	0	0	0	4	0
G	0	0	0	0	0	0	0	0	0	0	0	5

Need to scan only for the maximum value

We need to introduce the mismatch score

Still only diagonal operations

#### Limitations of diagonal operations

- Example:
  - Seq1 CATISNOTADOG
  - Seq2 NOTDOG
- Would be so nice to have a gap introduction operation:

```
NOTADOG
NOT DOG
```

#### Towards improvements (4)

Allow gap introduction operation

	С	Α	Т	I	S	N	0	Т	Α	D	0	G
N	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	2	1	0	0	0	0
Т	0	0	1	0	0	0	1	3	2	1	0	0
D	0	0	0	0	0	0	0	2	2	3	2	1
0	0	0	0	0	0	0	1	1	1	2	4	3
G	0	0	0	0	0	0	0	0	0	1	3	5

Need to scan only for the maximum value

We need to introduce mismatch score and gap introduction score

Now we have horizontal operations

#### Last steps towards NW: Initialization

Allows recursive computation for the whole matrix

		С	Α	Т	I	S	N	0	Т	Α	D	0	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12
N	-1												
0	-2												
Т	-3												
D	-4												
0	-5												
G	-6												

# Alignment

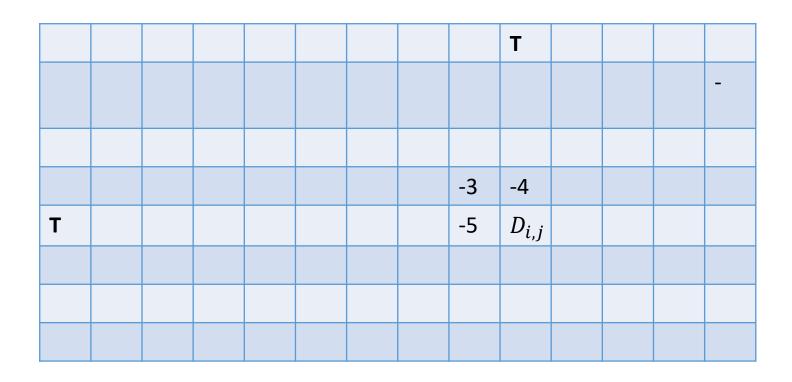
		С	Α	Т	I	S	N	0	Т	Α	D	0	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12
N	-1						-4						
0	-2							-3					
Т	-3								-2	-3			
D	-4										-2		
0	-5											-1	
G	-6												0

CATISNOTADOG

NOT DOG

#### Recursion

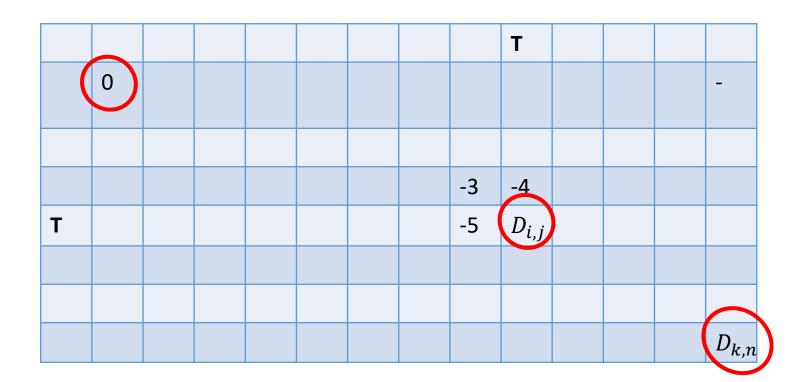
Three values to check



$$D_{ij} = \max \begin{cases} D_{i-1,j-1} + w(a_i, b_j) \\ D_{i,j-1} + w(-, b_j) \\ D_{i-1,} + w(a_i, -) \end{cases}$$

#### Traceback

Three values to check



$$D_{ij} = \max \begin{cases} D_{i-1,j-1} + w(a_i, b_j) \\ D_{i,j-1} + w(-, b_j) \\ D_{i-1,} + w(a_i, -) \end{cases}$$

### PAM250 BLOSSUM62 for proteins

С	12		_																			
S	0	2																				
Т	-2	1	3																			
Р	ო	1	0	6																		
Α	-2	1	1	1	2																	
G	-ე	1	0	-1	1	5																
N	-4	1	0	-1	0	0	2															
D	-5	0	0	-1	1	2	2	4														
Ε	-5	0	0	٦-	0	0	1	3	4													
Q	-5	-1	-1	0	0	-1	1	2	2	4												
Н	-3	-1	-1	0	-1	-2	2	1	1	3	6											
R	-4	0	-1	0	-2	ფ	0	-1	-1	1	2	6										
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5									
М	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6								
-	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5							
L	-6	-3	-2	უ	-2	-4	-3	-4	-3	-2	-2	უ	-3	4	2	6						
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4					
F	-4	-3	-ე	-5	-4	-5	-4	-6	-5	-5	-2	4	-5	0	1	2	-1	9				
Υ	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10			
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17		
В	-4	0	0	-1	0	0	2	3	2	1	1	-1	1	-2	-2	-3	-2	-5	-3	-5	2	
Z	-5	0	-1	0	0	-1	1	3	3	3	2	0	0	-2	-2	უ	-2	-5	-4	-6	2	3
	С	S	T	Р	Α	G	N	D	Ε	Q	Н	R	K	М	Ι	L	V	F	Υ	W	В	Z

#### Important notes

- 1. Gap penalty: make sure you are penalizing the gaps (there is no double negation)
- 2. Make sure that columns/rows of your matrix are independent objects (test if you can change a singe value in your matrix)
- 3. Keep track of the indexes. Make sure you are not exceeding the matrix limits while iterating over the cells
- 4. Test your code on toy examples which can be easily checked by eye
- 5. Document your code
- 6. Submit even an incomplete assignment if you are stuck; mark the steps which were implemented
- 7. Do not be scared to ask us for help if you are stuck
- 8. Work on your own.

## Good luck with your exercise