

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

0.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

	C	A	T	I	S	N	O	T	A	D	O	G
N	0	0	0	0	0	1	0	0	0	0	0	0
O	0	0	0	0	0	0	1	0	0	0	1	0
T	0	0	1	0	0	0	0	1	0	0	0	0
A	0	0	0	0	0	0	0	0	1	0	0	0
B	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

0 0

0 1 0

...

1 1 1 1 0 0 0

Towards improvements (2)

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

	C	A	T	I	S	N	O	T	A	D	O	G
N	0	0	0	0	0	1	0	0	0	0	0	0
O	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	0	0	0	0	0	0	0	4	0	0	0
B	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”

	C	A	T	I	S	N	O	T	A	D	O	G
N	0	0	0	0	0	1	0	0	0	0	0	0
O	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
O	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_DOOG

Towards improvements (4)

- Allow gap introduction operation

	C	A	T	I	S	N	O	T	A	D	O	G
N	0	0	0	0	0	1	0	0	0	0	0	0
O	0	0	0	0	0	0	2	1	0	0	0	0
T	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
O	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

Alignment

		C	A	T	I	S	N	O	T	A	D	O	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12
N	-1						-4						
O	-2							-3					
T	-3								-2	-3			
D	-4										-2		
O	-5											-1	
G	-6												0

CATISNOTADOG

_____NOT__DOG

Recursion

- Three values to check

									T				
													-
								-3	-4				
T								-5	$D_{i,j}$				

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

Traceback

- Three values to check

									T				
0													-
								-3	-4				
T								-5	$D_{i,j}$				
													$D_{k,n}$

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

PAM250 BLOSSUM62 for proteins

[illegible]

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 single 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