SHAKE THE FUTURE.

CENTRALE NANTES

Pairwise sequence alignment

Rokhaya Ba / Sophie Limou

EII - BIOLO TP R

Objectives of pairwise sequence alignment

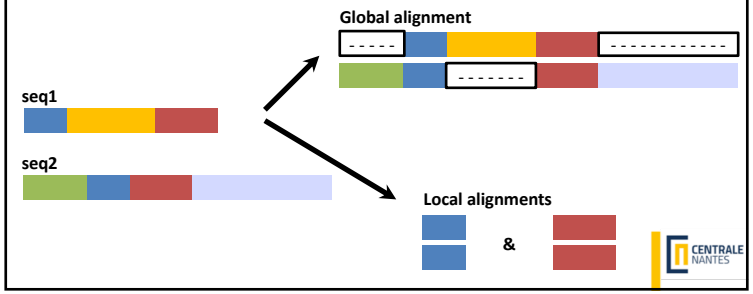
```
Query   VTALWGKVNVD--EVGGEALGRLL
        V  +WGKV  D    G E L RL
Sbjct   VLNVWGKVEADIPGHGQEVLIRLF
```

- % **identity** = % of identical amino acids between two sequences
  = 11/24 = 45%
- % **similarity** = % of identical aa and aa with similar physico-chemical properties (aka conserved substitutions, '+' symbol)
  = (11+1)/24 = 50%

Pairwise alignment = line up two sequences while maximizing the percentage of identity between them

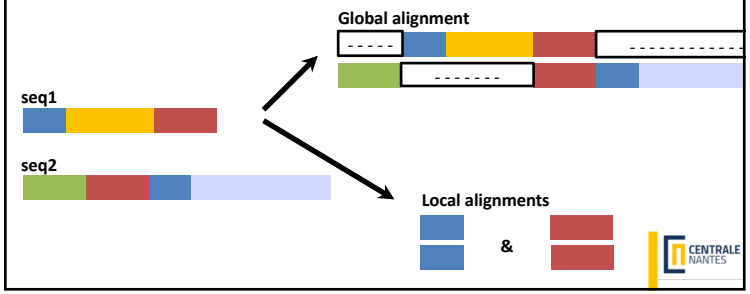
Global vs. local alignment

- **Global alignment** contains the entire sequence/length of each sequence
- **Local alignment** focuses on regions of greatest similarity between 2 sequences



Global vs. local alignment

- **Global alignment** contains the entire sequence/length of each sequence
- **Local alignment** focuses on regions of greatest similarity between 2 sequences



Scoring system

- Attribute a different weight to each amino acid alignment
- Example:
  - +3 for a match
  - -2 for a mismatch
  - -2 for a gap creation
  - -1 for gap extension

```
VTALWGKVNVD--EVGGEALGRLL
V  +WGKV  D    G E L RL
VLNVWGKVEADIPGHGQEVLIRLF
```

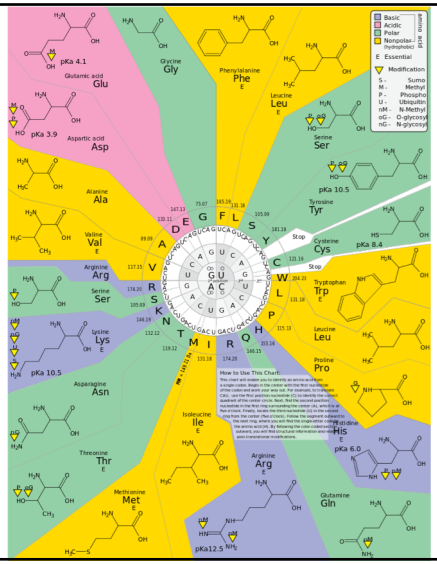
⇨ Score = 11 matches + 10 mismatches + 1 gap of 2nt
= 33 – 22 – 4
= 7

Scoring system – consider amino acid physico-chemical properties

- Attribute a different weight to mismatches based on amino-acid properties (acidic, basic, polar, non polar)
  - ➢ Tolerate same category substitution
  - ➢ Penalize when one amino acid is substituted by an amino acid from a different category

Reminder

Genetic code



Scoring system – the empirical PAM and BLOSUM matrices

- Based on real/observed protein alignments reflecting evolution
- PAM: using alignment of closely related proteins (>85% identity) from superfamilies (global alignment)
- BLOSUM: based on alignment of conserved blocks from distantly related proteins (local alignments)
- Reference matrices = PAM120 and BLOSUM62

BLOSUM90	BLOSUM62	BLOSUM45
PAM30	PAM120	PAM250
Less divergent	⟷	More divergent
Human versus chimpanzee beta globin		Human versus bacterial globins

Take-home messages

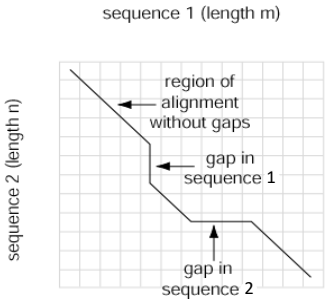
- Genetic code is redundant
- More information in proteins (20 amino acids vs. 4 nucleotides)
- Alignment to maximize identity/similarity score between two sequences
- Difference between global and local alignments
- Possibility to account for shared physico-chemical properties of amino acids in scoring matrices
- Prefer protein alignments but not always: non-coding sequences (intron and intergenic regions)

Bonus !!

- If you are interested by the pairwise alignment algorithm :
- Next are couples of slides explaining the principles of the global alignment algorithm

Global alignment: the Needleman and Wunsch algorithm

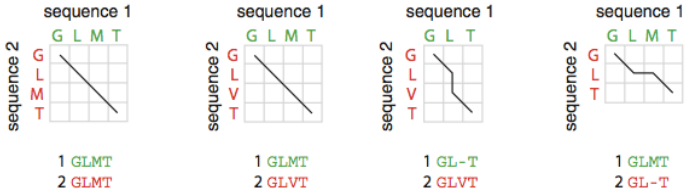
- Example of dynamic programming = find the optimal alignment by extending from optimal subpaths
- Use the diagonal method



- Three steps:
  1) Set the matrix
  2) Score the matrix
  3) Identify the optimal alignment using a trace-back procedure

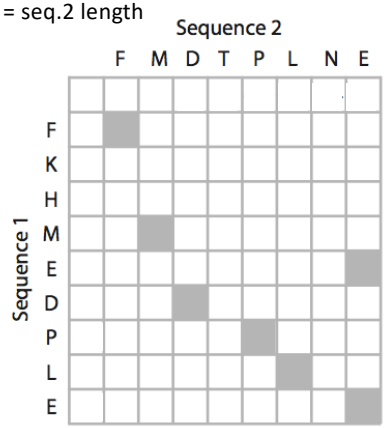
The NW algorithm

- Diagonals represents regions of high similarity.
- Diagonal paths can contain matches and mismatches.
- Vertical and horizontal paths indicate gaps in either sequence.



The NW algorithm – step 1: set-up the matrix

- Matrix of size (m+1) x (n+1)
  where m = seq.1 length and n = seq.2 length
- Grey out the identical cells



The NW algorithm – step 2: score the matrix

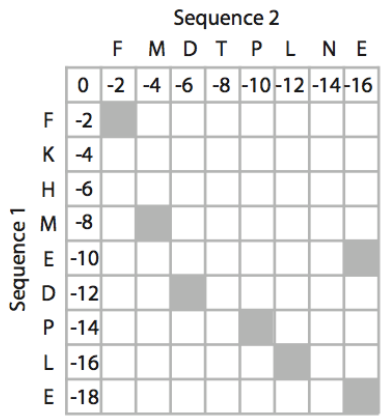
- Define a scoring system
  (you can use a scoring matrix for match/mismatch)

$$\text{Score} = \text{Max} \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - \text{gap penalty} \\ F(i, j-1) - \text{gap penalty} \end{cases}$$

Score (this example) = +1 (match)
-2 (mismatch)
-2 (gap penalty)

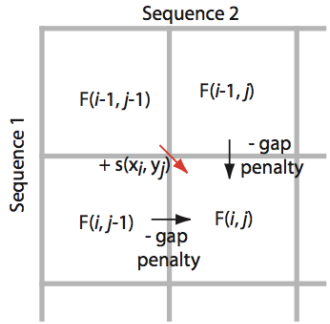
- Fill in the gap penalties in the first row and column

The NW algorithm – step 2: score the matrix

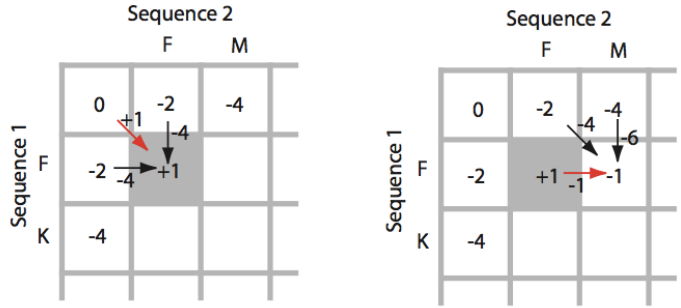


The NW algorithm – step 2: score the matrix

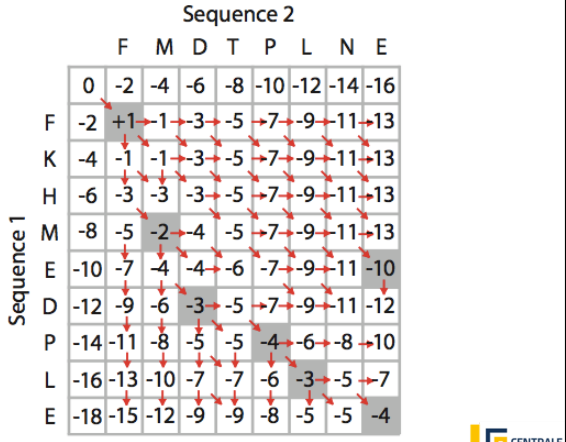
- Fill in the matrix by respecting the scoring system and keeping track of the path giving the best score (red arrow)



The NW algorithm – step 2: score the matrix

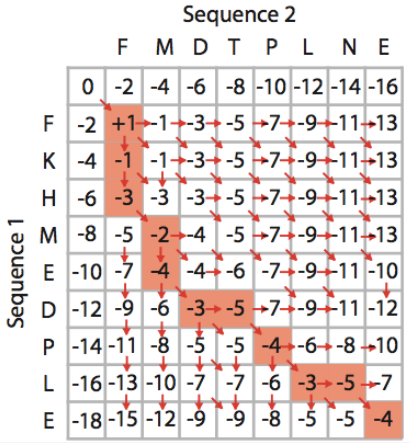


The NW algorithm – step 2: score the matrix



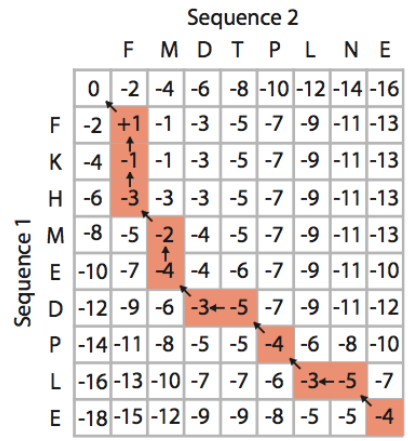
The NW algorithm – step 3: identify the optimal alignment

- Highlight the best path using a trace-back procedure



The NW algorithm – step 3: identify the optimal alignment

- Highlight the best path using a trace-back procedure



The NW algorithm – step 3: identify the optimal alignment

Sequence 2

Sequence 1		F	M	D	T	P	L	N	E
	0	-2	-4	-6	-8	-10	-12	-14	-16
F	-2	+1	-1	-3	-5	-7	-9	-11	-13
K	-4	-1	-1	-3	-5	-7	-9	-11	-13
H	-6	-3	-3	-3	-5	-7	-9	-11	-13
M	-8	-5	-2	-4	-5	-7	-9	-11	-13
E	-10	-7	-4	-4	-6	-7	-9	-11	-10
D	-12	-9	-6	-3	-5	-7	-9	-11	-12
P	-14	-11	-8	-5	-5	-4	-6	-8	-10
L	-16	-13	-10	-7	-7	-6	-3	-5	-7
E	-18	-15	-12	-9	-9	-8	-5	-5	-4

	+1	-1	-3	-2	-4	-3	-5	-4	-3	-5	-4
Sequence 1	F	K	H	M	E	D	-	P	L	-	E
Sequence 2	F	-	-	M	-	D	T	P	L	N	E