Sequence alignment

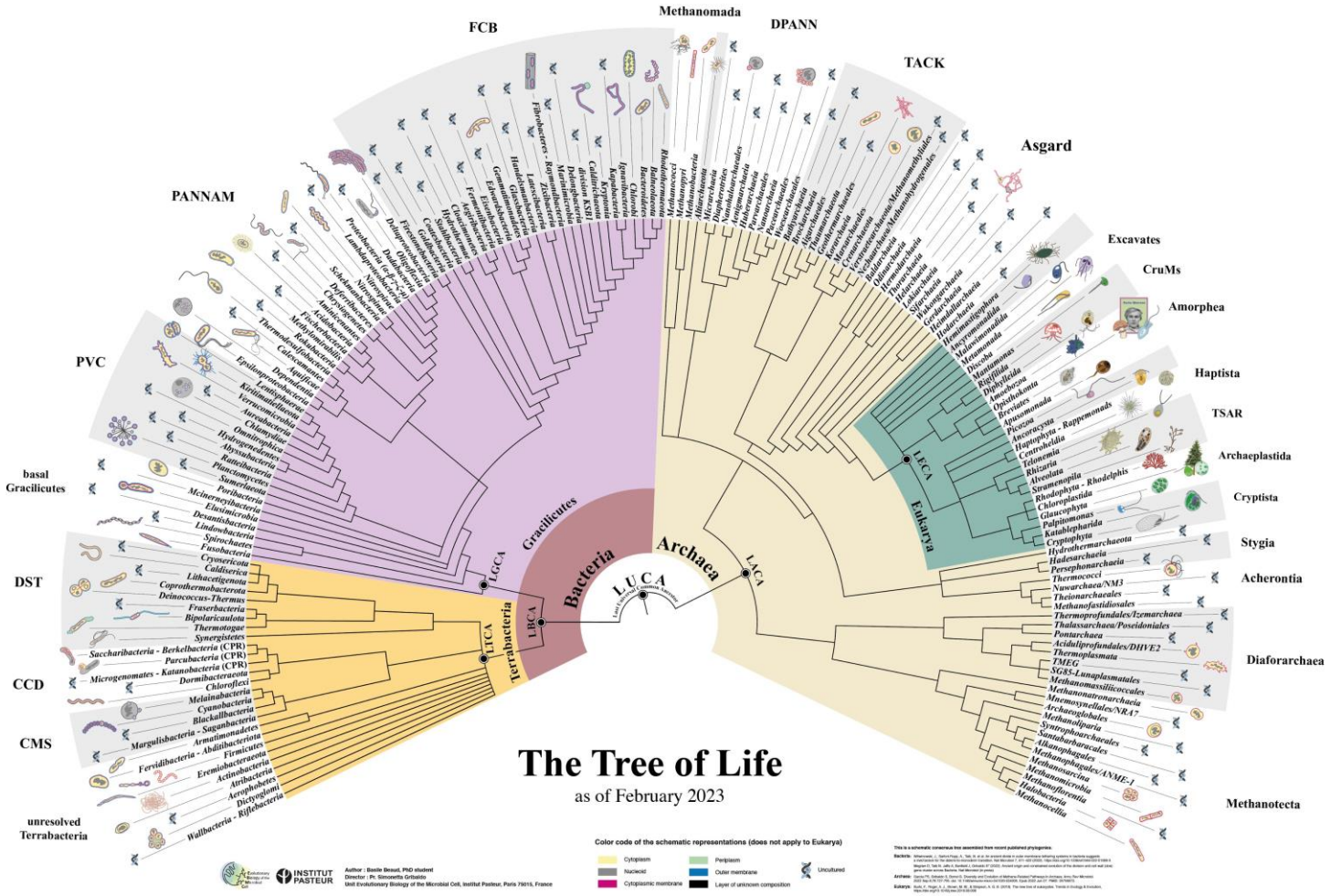
TRANSCRIPTOMICS

BIGNAUD AMAURY
26/09/2023

Why do you want to align sequences in biology ?

Why do you want to align sequences in biology ?

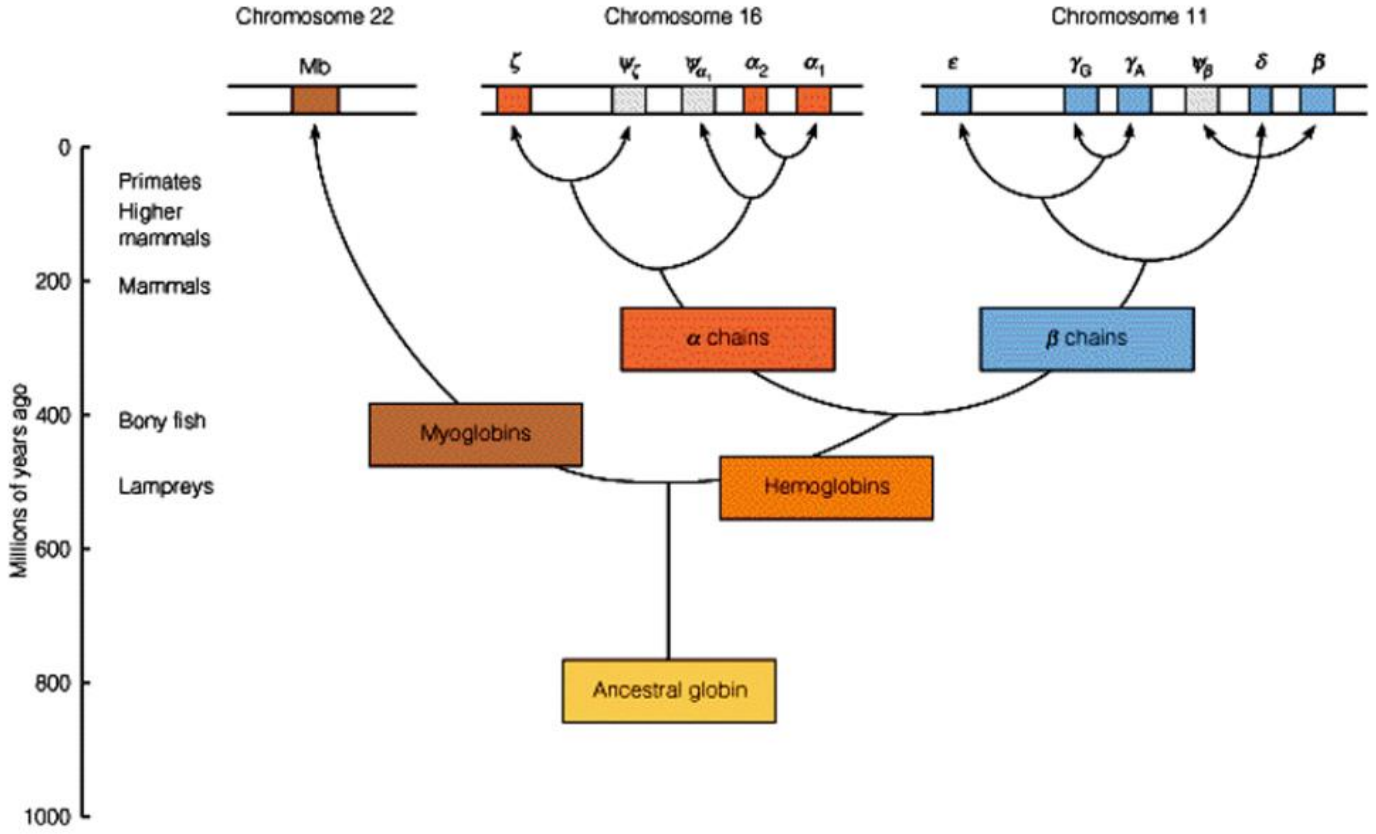
- **Homology** in sequences:
  - Find a common ancestor and build **phylogenetic** tree.



Why do you want to align sequences in biology ?

- **Homology** in sequences:
  - Find a common ancestor and build **phylogenetic** tree.
  - Infer **structure/function** of a protein based on sequence similarities.

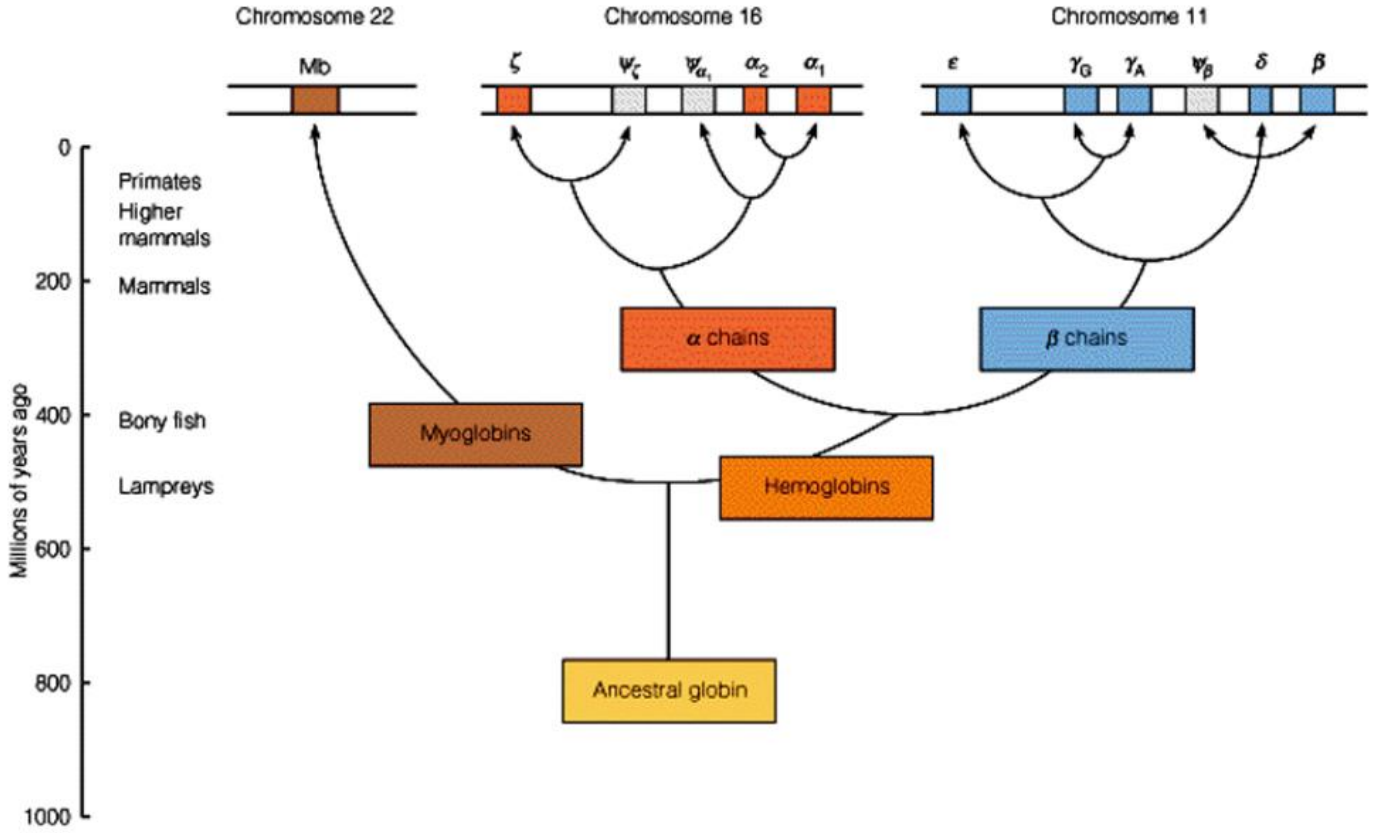
Two homologous genes from the same species (duplication event) are **paralogous** (globin α and β) and two genes homologous from different species are **orthologous** (human globin α and mouse globin α).



Why do you want to align sequences in biology ?

- **Homology** in sequences:
  - Find a common ancestor and build **phylogenetic** tree.
  - Infer **structure/function** of a protein based on sequence similarities.

  Two homologous genes from the same species (duplication event) are **paralogous** (globin α and β) and two genes homologous from different species are **orthologous** (human globin α and mouse globin α).



- **Sequencing processing**.

Task definition

- **Given:**

  – A pair of sequences (DNA or protein).
  – A method for scoring a candidate alignment.

- **To Do:**

  – Determine the correspondences between substrings in the sequences such that the similarity score is maximized.

Sequence variations

Sequences may have diverged through various types of mutations:

- **Substitutions** (A**C**GA → A**G**GA)
- **Insertions** (ACGA → AC**CGGA**GA)
- **Deletions** (A**CGGA**GA → AGA)

The latter two will result in **gaps** or **indels** in alignments.

Issues on sequence alignment

- The sequences we're comparing probably differ in length.
- There may be only a relatively small region in the sequences that match.
- We want to allow partial matches - some nucleotides/amino acid pairs are more substitutable than others.
- Variable length regions may have been inserted/deleted from the reference sequence.

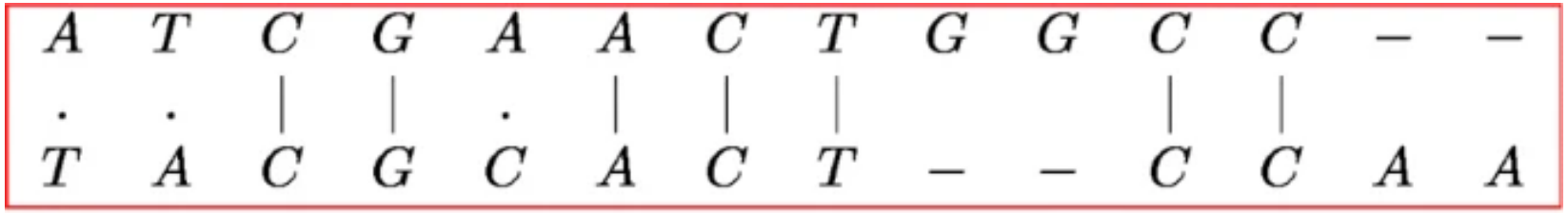
Problematics

- Question 1: What do we want to align ?
- Question 2: How do we "**score**" an alignment ?
- Question 3: How do we find the "**best**" alignment ?

What do we want to align ?

Ahmed *et al.*, *BMC Bioinformatics*, 2019

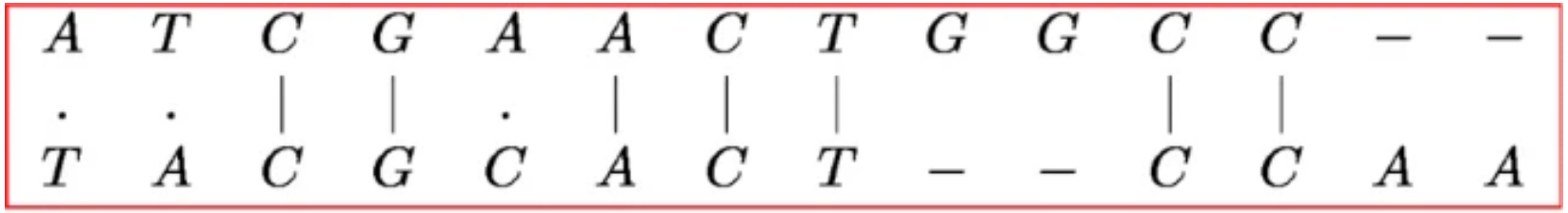
- **Global alignment**: Find best match of both sequences in their entirety



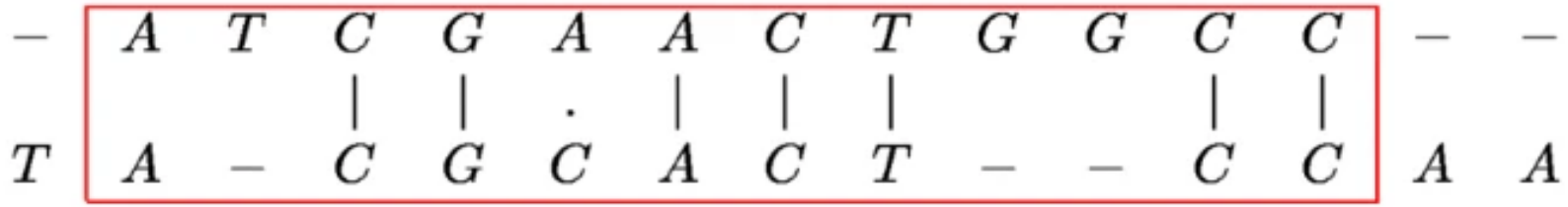
What do we want to align ?

Ahmed *et al.*, *BMC Bioinformatics*, 2019

- **Global alignment**: Find best match of both sequences in their entirety



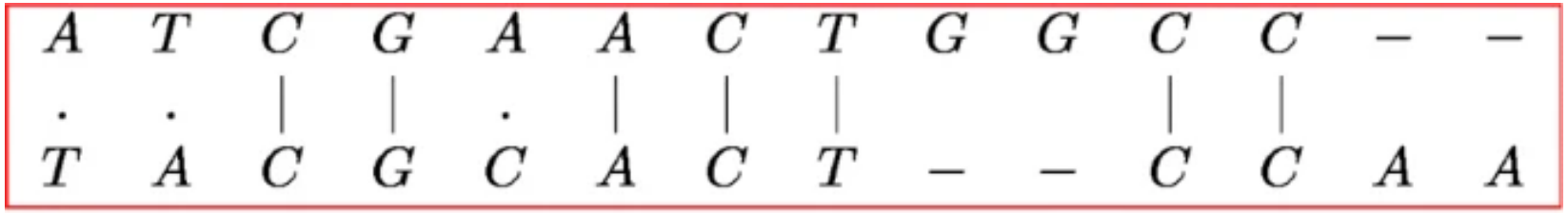
- **Semi-global alignment**: Find best match without penalizing gaps on the ends of the alignment.



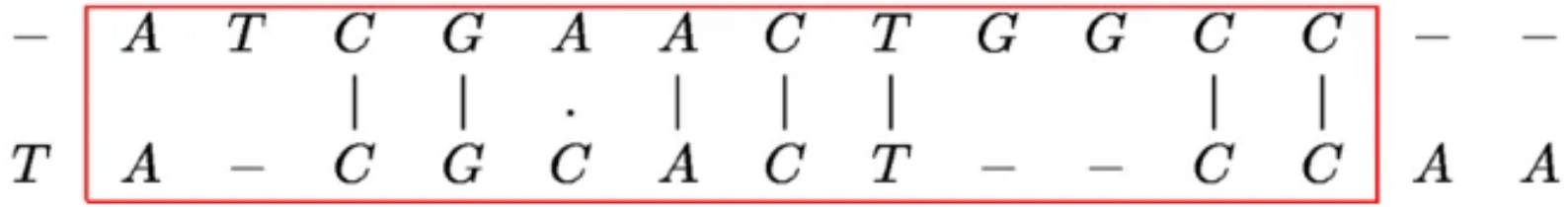
What do we want to align ?

Ahmed *et al.*, *BMC Bioinformatics*, 2019

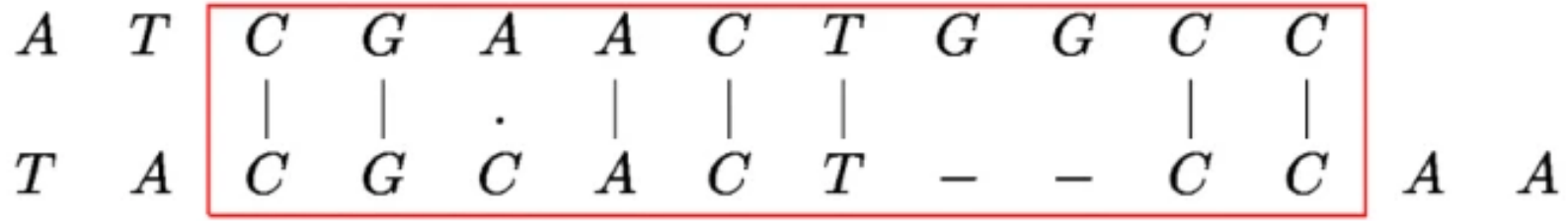
- **Global alignment**: Find best match of both sequences in their entirety



- **Semi-global alignment**: Find best match without penalizing gaps on the ends of the alignment.



- **Local alignment**: Find best subsequence match.



How do we “**score**” an alignment ?

How do we “**score**” an alignment ?

Gap penalty function

w(k) indicates cost of a gap of length k. Here we will take the example of a linear cost.

$$w(k) = g\ x\ k$$

Substitution matrix

s(a, b) indicates score of aligning character a with character b.

-	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

How do we "**score**" an alignment ?

Gap penalty function
w(k) indicates cost of a gap of length k. Here we will take the example of a linear cost.

$$w(k) = g\ x\ k$$

Substitution matrix
s(a, b) indicates score of aligning character a with character b.

-	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

The score of an alignment is the sum of the scores for pairs of aligned characters plus the scores for gaps.

Example: given the following alignment:

ATCGAACTGGCC - -
TACG - - -T- - CCAA

We would score it by:

s(A, T) + s(T, A) + s(C, C) + s(G, G) + 3g + s(T, T) + 2g + s(C, C) + s(C, C) + 2g = ??

How do we "**score**" an alignment ?

Gap penalty function
w(k) indicates cost of a gap of length k. Here we will take the example of a linear cost.

$$w(k) = g\ x\ k$$

Substitution matrix
s(a, b) indicates score of aligning character a with character b.

	-	A	G	C	T
A	10	-1	-3	-4	
G	-1	7	-5	-3	
C	-3	-5	9	0	
T	-4	-3	0	8	

The score of an alignment is the sum of the scores for pairs of aligned characters plus the scores for gaps.

Example: given the following alignment:

ATCGAACTGGCC - -
TACG - - -T- - CCAA

We would score it by:

s(A, T) + s(T, A) + s(C, C) + s(G, G) + 3g + s(T, T) + 2g + s(C, C) + s(C, C) + 2g = ??

-4 + -4 + 9 + 7 + -9 + 8 + -6 + 9 + 9 + -6 = 13

How do we find the "**best**" alignment ?

Simplest approach: compute & score all possible alignments, but there are:

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2}$$

possible global alignments for 2 sequences of length 100 will have $\sim 10^{77}$ possible alignments.

Dynamic programming: solve an instance of a problem by taking advantage of solutions for subparts of the problem.

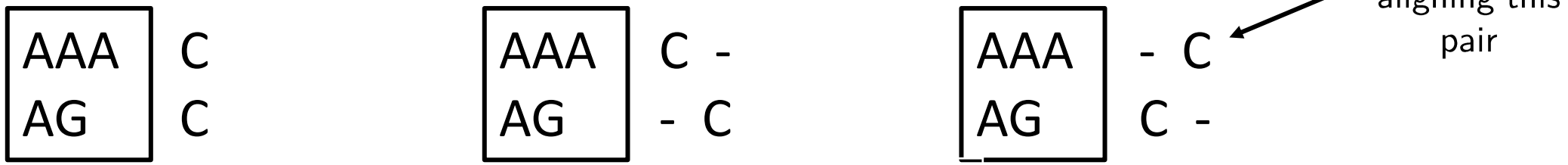
Reduce problem of best alignment of two sequences to best alignment of all prefixes of the sequences.

Avoid recalculating the scores already considered.

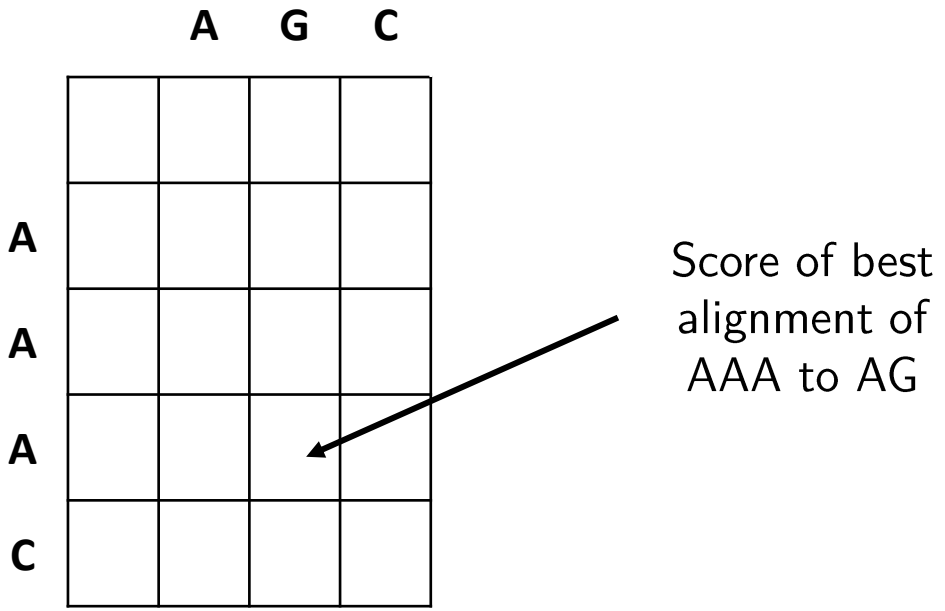
- example: Fibonacci sequence 1, 1, 2, 3, 5, 8, 13, 21, 34...

Needleman & Wunsch algorithm

- Consider last step in computing alignment of AAAC with AGC:
  - Match between the two next nucleotides.
  - Insertion in the first sequence.
  - Deletion in the first sequence.



- Construct an (n+1) x (m+1) matrix F.
- F(i, j) = Score of the best alignment of x[1…i] with y[1…j].



Initializing the matrix

		A	G	C
	0			
A				
A				
A				
C				

Suppose we choose the following scoring scheme:

$$s(x_i, y_i) = \begin{cases} +1 \; when \; x_i = y_i \\ -1 \; when \; x_i \neq y_i \end{cases}$$

$$w(k) = -2k$$

Initializing the matrix

		A	G	C
	0 →	-2 →	-4 →	-6
A	-2			
A	-4			
A	-6			
C	-8			

Suppose we choose the following scoring scheme:

$$s(x_i, y_i) = \begin{cases} +1 \; when \; x_i = y_i \\ -1 \; when \; x_i \neq y_i \end{cases}$$

$$w(k) = -2k$$

Completing the matrix

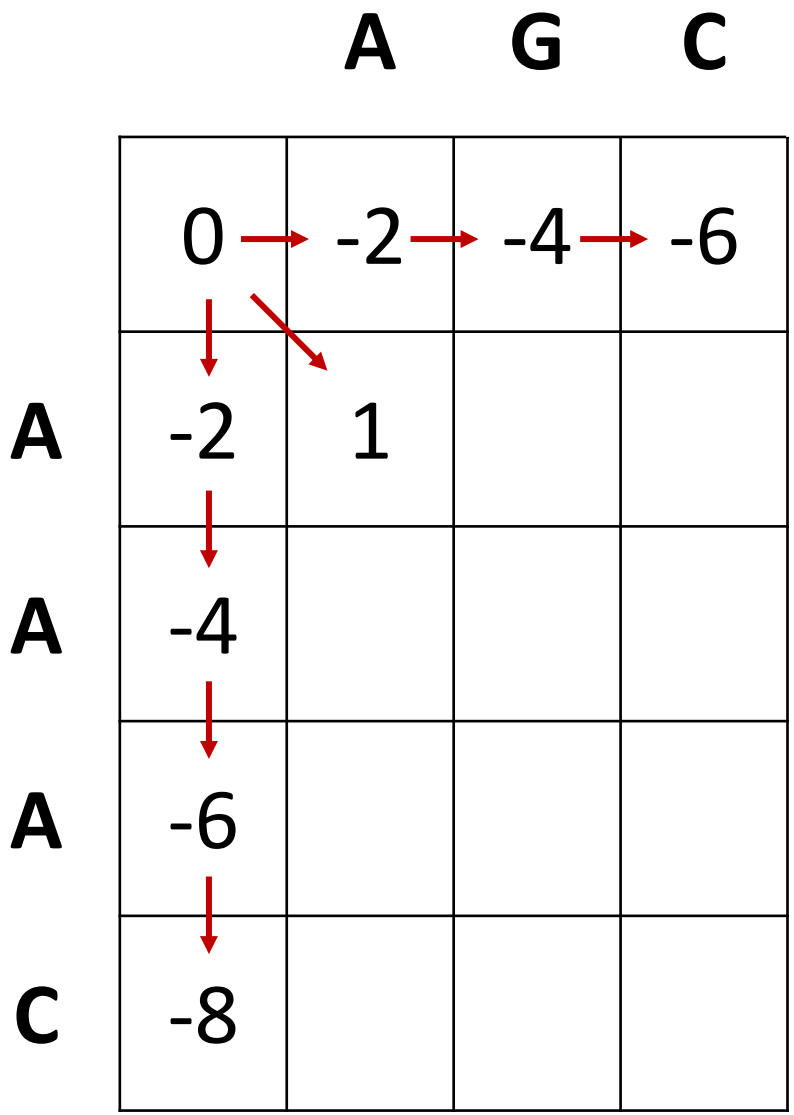
		A	G	C
	0 →	-2 →	-4 →	-6
A	-2			
A	-4			
A	-6			
C	-8			

Suppose we choose the following scoring scheme:

$$s(x_i, y_i) = \begin{cases} +1 \; when \; x_i = y_i \\ -1 \; when \; x_i \neq y_i \end{cases}$$

$$w(k) = -2k$$

Completing the matrix

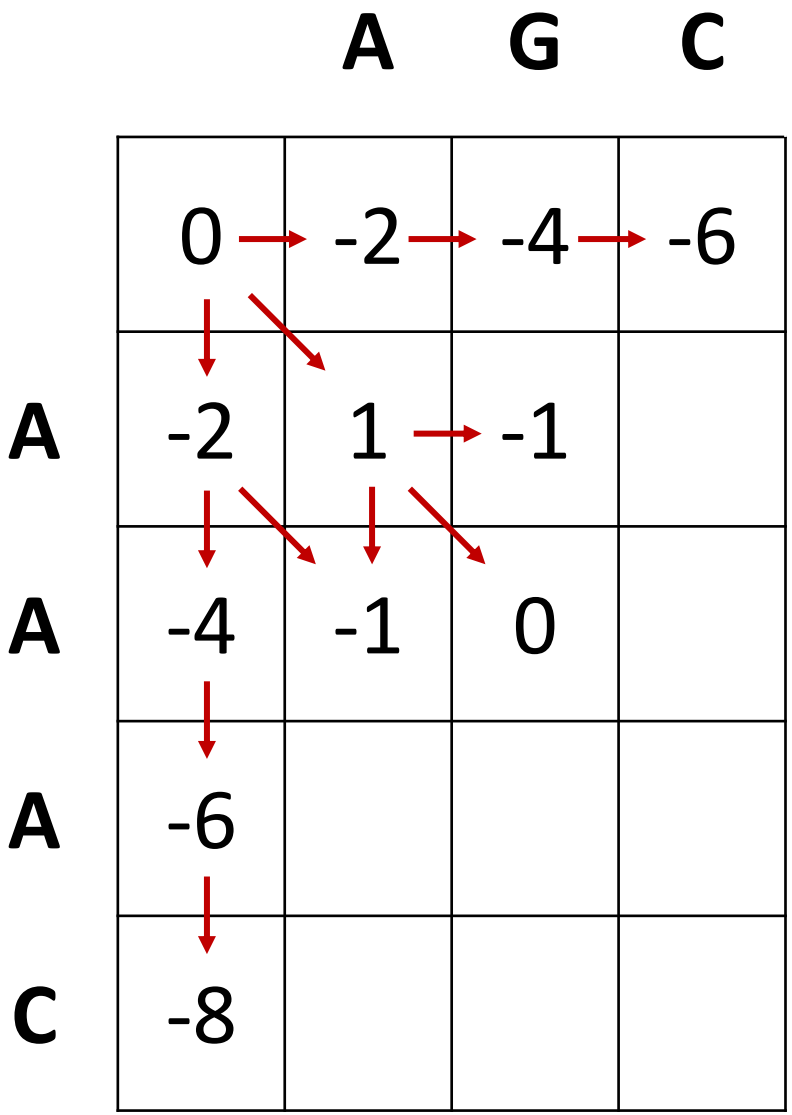


Suppose we choose the following scoring scheme:

$$s(x_i, y_i) = \begin{cases} +1 \; when \; x_i = y_i \\ -1 \; when \; x_i \neq y_i \end{cases}$$

$$w(k) = -2k$$

Completing the matrix

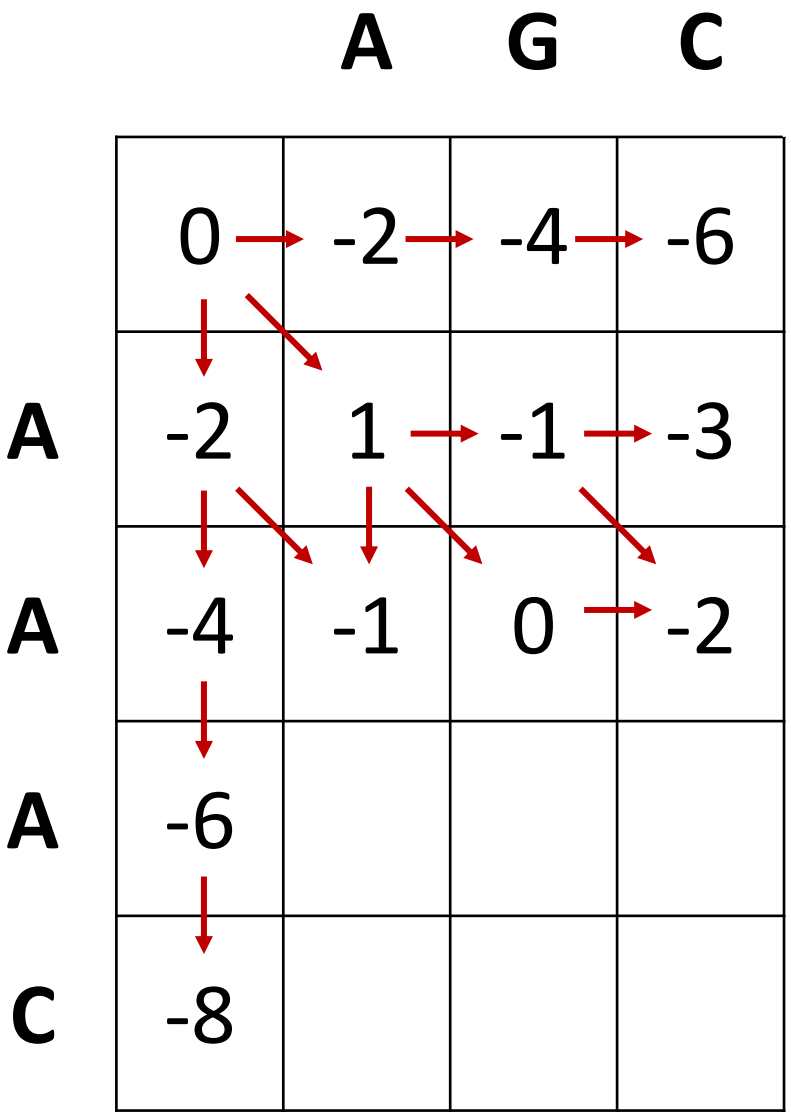


Suppose we choose the following scoring scheme:

$$s(x_i, y_i) = \begin{cases} +1 \; when \; x_i = y_i \\ -1 \; when \; x_i \neq y_i \end{cases}$$

$$w(k) = -2k$$

Completing the matrix

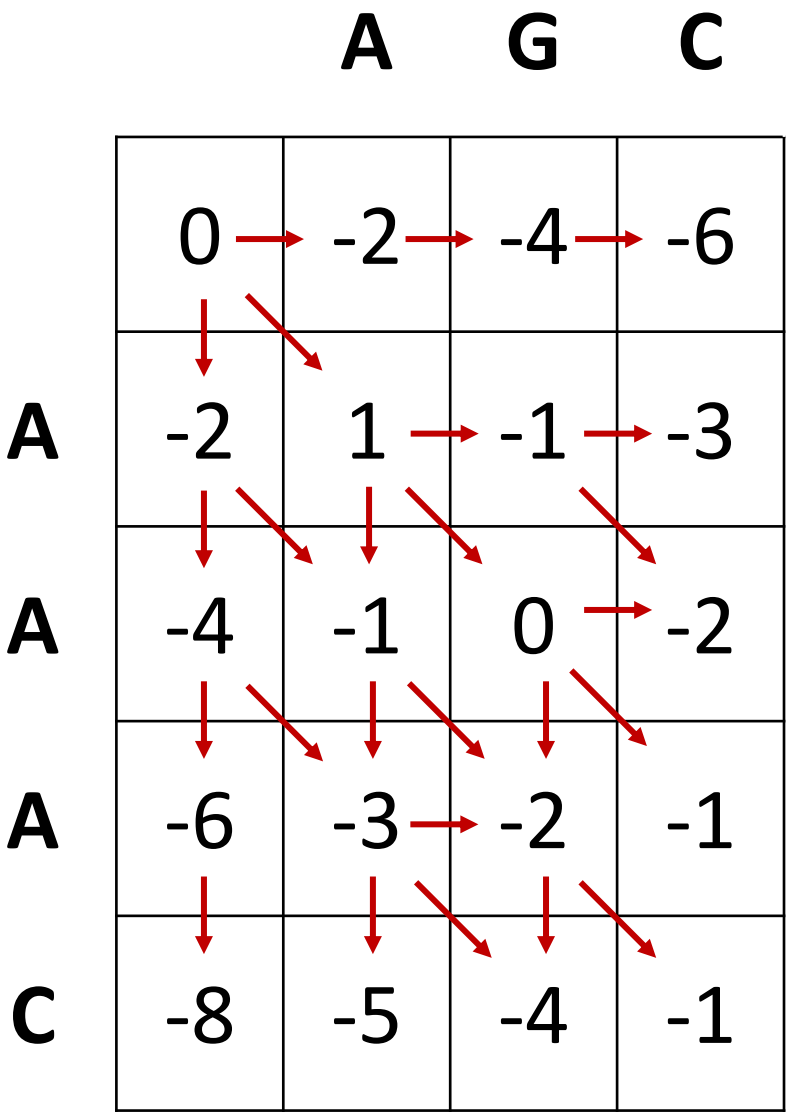


Suppose we choose the following scoring scheme:

$$s(x_i, y_i) = \begin{cases} +1 \; when \; x_i = y_i \\ -1 \; when \; x_i \neq y_i \end{cases}$$

$$w(k) = -2k$$

Completing the matrix

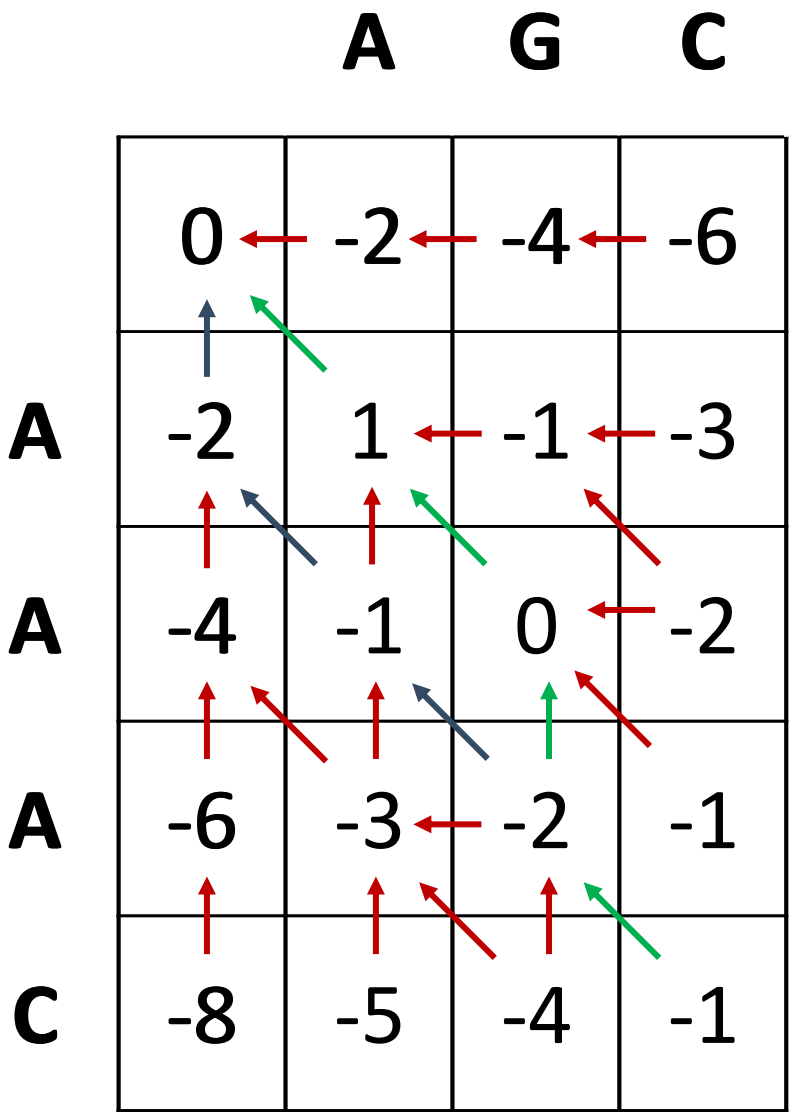


Suppose we choose the following scoring scheme:

$$s(x_i, y_i) = \begin{cases} +1 \; when \; x_i = y_i \\ -1 \; when \; x_i \neq y_i \end{cases}$$

$$w(k) = -2k$$

Solving the matrix



Optimal alignment (highroad):

x: A A A C

y: A G - C

Optimal alignment (lowroad):

x: A A A C

y: - A G C

Local alignment

- So far we have discussed global alignment, where we are looking for best match between sequences from one end to the other.
- More commonly, we will want a local alignment, the best match between subsequences of x and y.

Local alignment

- So far we have discussed global alignment, where we are looking for best match between sequences from one end to the other.
- More commonly, we will want a local alignment, the best match between subsequences of x and y.

Needleman-Wunsch algorithm:

Pairwise **global** alignment only. Needleman and Wunsch, *Journal of Molecular Biology*, 1970.

Smith-Waterman algorithm:

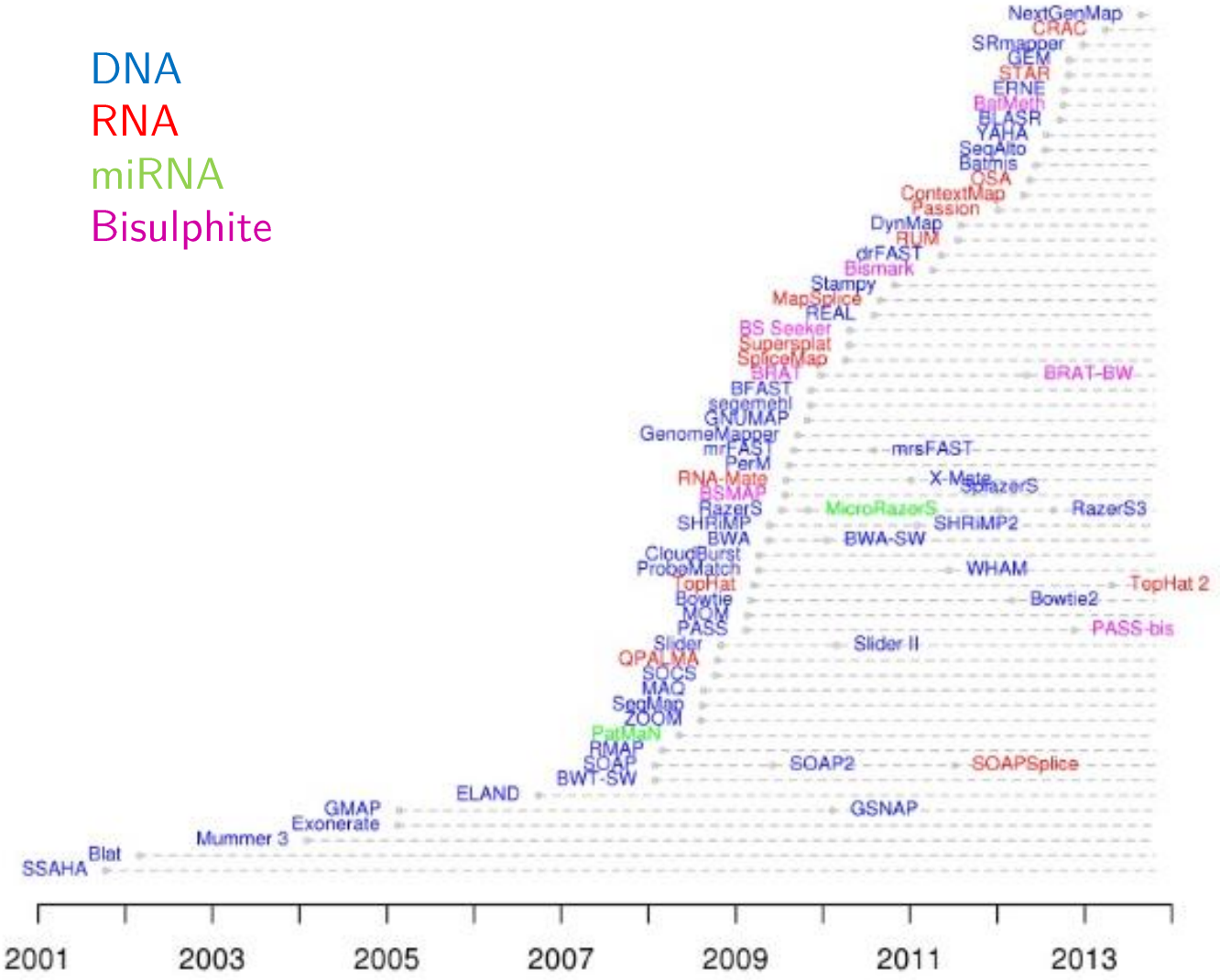
Pairwise, **local** (*or global*) alignment. Smith and Waterman, *Journal of Molecular Biology*, 1981.

BLAST algorithm:

Pairwise **heuristic** local alignment. (Most cited algorithm in bioinformatics). McGinnis and Madden, *Nucleic Acid Research*, 2004.

Sequencing reads aligner

DNA
RNA
miRNA
Bisulphite

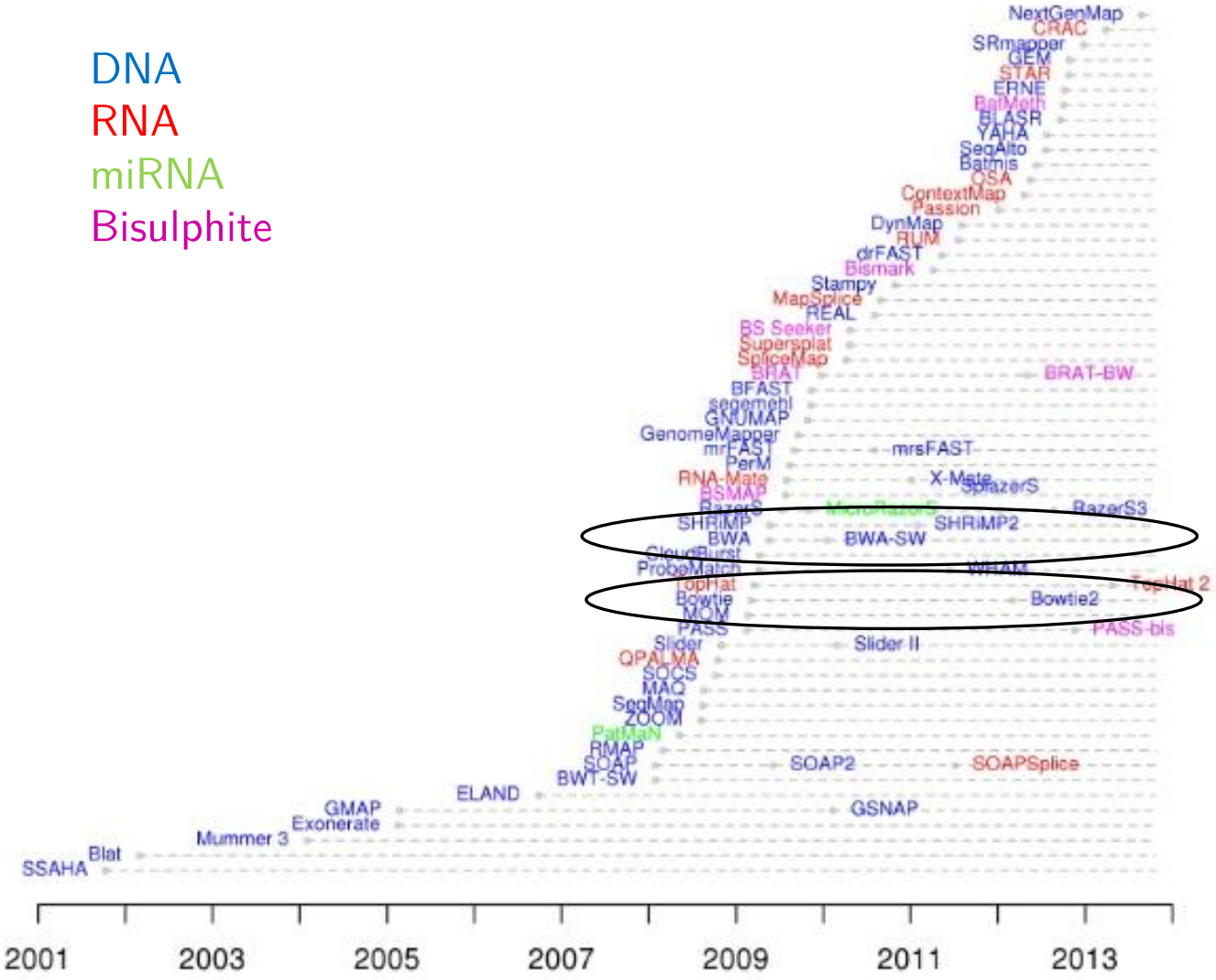


~ 100 aligner algorithms now

Fonseca *et al.*, *Bioinformatics*, 2012

Sequencing reads aligner

DNA
RNA
miRNA
Bisulphite



~ 100 aligner algorithms now

Fonseca *et al.*, *Bioinformatics*, 2012

The main used aligners principle (bowtie/bwa)

- To get a fast alignment on a large genome (3.10^9 for the human genome), the reference is indexed based on Burrows-Wheeler algorithm. The idea is to be able to load the whole genome inside the RAM of your computer to get a very fast access to the sequence.
- You use a seed (part of your reads, usually around ~20bp) which will be mapped to these indexed subsequence.
- Once you map the seed to some positions you try to extend the alignments with the local genome and keep the best match(es).

Bowtie2: Langmead & Salzberg, *Nature methods*, 2012.

BWA: Li & Durbin, *Bioinformatics*, 2009