Swiss Institute of Bioinformatics

INTRODUCTION TO SEQUENCING DATA ANALYSIS

Read alignment

Deepak Tanwar
Frédéric Burdet

April 23-25, 2025
Adapted from previous year courses

elixir SWITZERLAND

Learning Objectives

Understand the concept and purpose of sequence alignment

Distinguish between global and local alignment strategies and their algorithms

Describe how short and long sequencing reads are aligned to reference genomes

Understand indexing strategies (e.g., BWT, suffix arrays) used in fast read alignment tools

In bioinformatics, **alignment** refers to the process of arranging sequences of DNA, RNA, or proteins to identify regions of similarity.

Alignment types

Pairwise alignment

A. **Global Alignment – *Needleman-Wunsch Algorithm***
   - DNA, RNA, or protein sequences of similar length

B. **Local Alignment - *Smith-Waterman Algorithm***
   - Protein or RNA/DNA domains

Multiple Sequence Alignment (MSA)

- Aligns three or more sequences to detect conserved regions.

Read Alignment (e.g. BWA, Bowtie2, minimap2)

- Mapping short reads (DNA/RNA) to a reference genome

Pairwise alignment

Global Alignment

- Aligns **entire sequences** from **end to end**
- Introduces **gaps** to optimize alignment length and score

Local Alignment

- Aligns the best-matching subsequences
- Useful for finding **domains** or **conserved motifs**

Pairwise alignment: Global

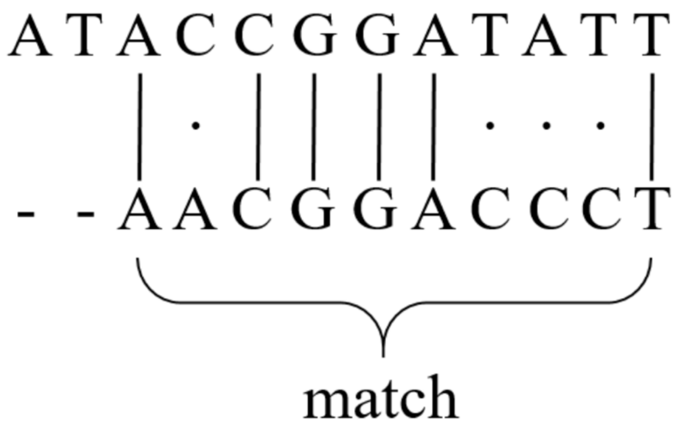
Sequence1 : A T A C C G G A T A T T

Sequence2 : A A C G G A C C C T

Pairwise alignment: Global

Sequence1 : A T A C C G G A T A T T

Sequence2 : A A C G G A C C C T



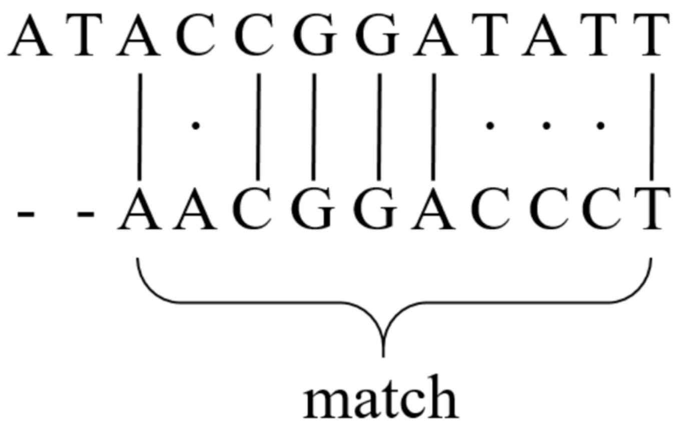
```
A T A C C G G A T A T T
    | . | | | | . . . |
- - A A C G G A C C C T
```

match

Pairwise alignment: Global

Sequence1 : A T A C C G G A T A T T

Sequence2 : A A C G G A C C C T



```
A T A C C G G A T A T T
    | . | | | | . . . |
- - A A C G G A C C C T
```

match

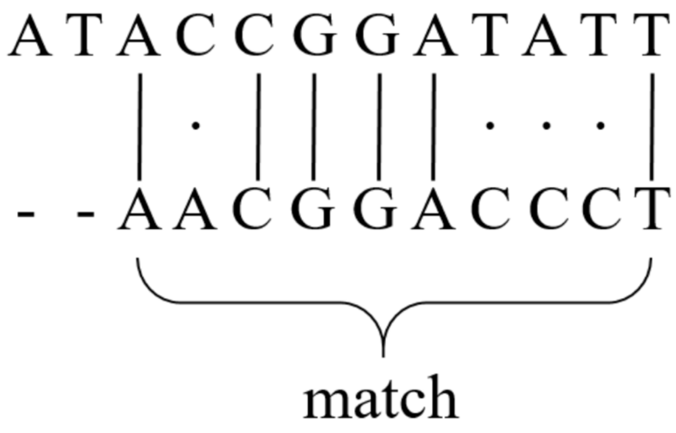
Alignment scoring

Match: +1
Mismatch: -1
Gap penalty: -1 (for each gap)

Pairwise alignment: Global

Sequence1 : A T A C C G G A T A T T

Sequence2 : A A C G G A C C C T



Alignment scoring

Match: +1
Mismatch: -1
Gap penalty: -1 (for each gap)

S1	S2	Score
A	-	-1
T	-	-1
A	A	+1
C	A	-1
C	C	+1
G	G	+1
G	G	+1
A	A	+1
T	C	-1
A	C	-1
T	C	-1
T	T	+1
Total		0

Pairwise alignment: Local

Sequence1 : A T A C C G G A T A T T

Sequence2 : A A C G G A C C C T

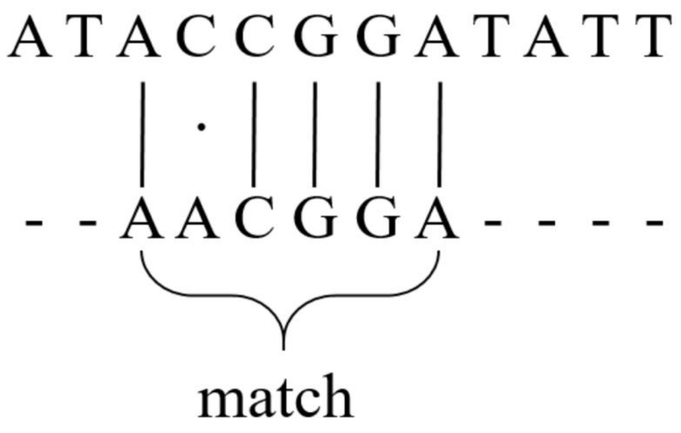
Alignment scoring

Match: +1
Mismatch: -1
Gap penalty: -1 (for each gap)

Pairwise alignment: Local

Sequence1 : A T A C C G G A T A T T

Sequence2 : A A C G G A C C C T



```
A T A C C G G A T A T T
    | . | | | |
- - A A C G G A - - - -
```

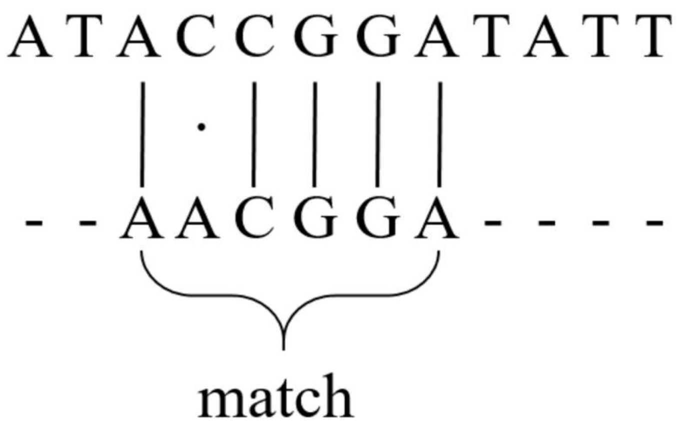
Alignment scoring

Match: +1
Mismatch: -1
Gap penalty: -1 (for each gap)

Pairwise alignment: Local

Sequence1 : A T A C C G G A T A T T

Sequence2 : A A C G G A C C C T



Alignment scoring

Match: +1
Mismatch: -1
Gap penalty: -1 (for each gap)

S1	S2	Score
A	-	0
T	-	0
A	A	+1
C	A	-1
C	C	+1
G	G	+1
G	G	+1
A	A	+1
T	-	0
A	-	0
T	-	0
T	-	0
Total		4

Multiple Sequence Alignment

For **phylogeny**, **function prediction**, or **motif discovery**.

Tool: Clustal Omega, MAFFT Command-line example with Clustal Omega

Multiple Sequence Alignment

For **phylogeny**, **function prediction**, or **motif discovery**.

Tool: Clustal Omega, MAFFT Command-line example with Clustal Omega

A	T	C	G
A	T	G	G
A	C	G	

Multiple Sequence Alignment

For **phylogeny**, **function prediction**, or **motif discovery**.

Tool: Clustal Omega, MAFFT Command-line example with Clustal Omega

A	T	C	G
A	T	G	G
A	C	G	

S1	S2	Score
A	A	+1
A	A	+1
A	A	+1
Total		3

Multiple Sequence Alignment

For **phylogeny**, **function prediction**, or **motif discovery**.

Tool: Clustal Omega, MAFFT Command-line example with Clustal Omega

A	T	C	G
A	T	G	G
A	C	G	

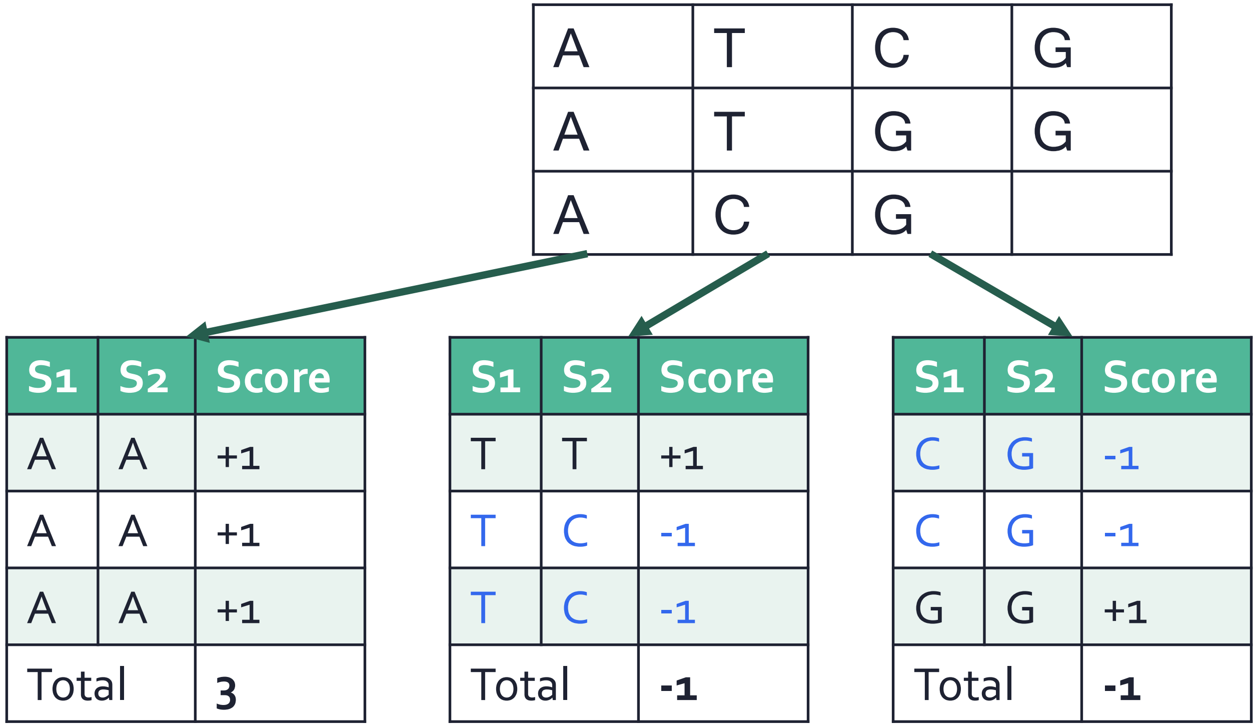
S1	S2	Score
A	A	+1
A	A	+1
A	A	+1
Total		3

S1	S2	Score
T	T	+1
T	C	-1
T	C	-1
Total		**-1**

Multiple Sequence Alignment

For **phylogeny, function prediction**, or **motif discovery**.

Tool: Clustal Omega, MAFFT Command-line example with Clustal Omega



A	T	C	G
A	T	G	G
A	C	G	

S1	S2	Score
A	A	+1
A	A	+1
A	A	+1
Total		3

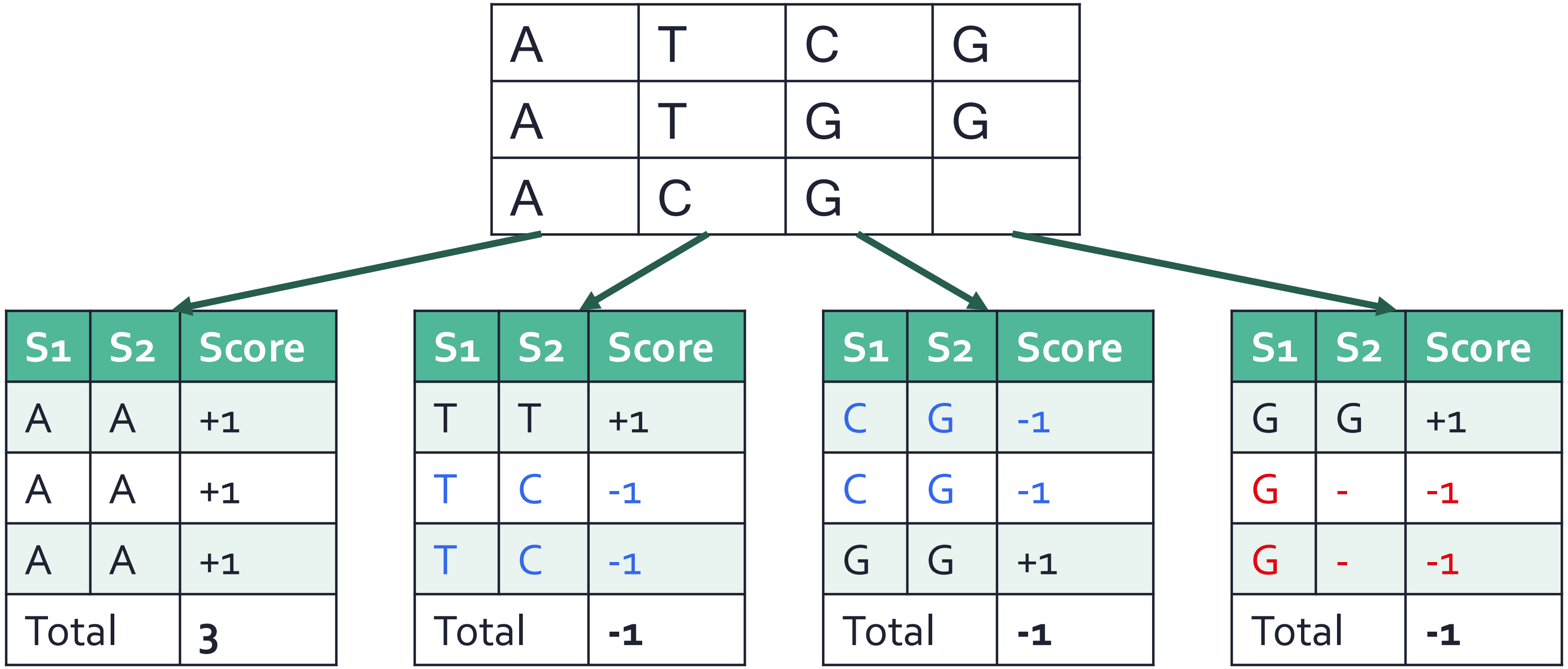
S1	S2	Score
T	T	+1
T	C	-1
T	C	-1
Total		**-1**

S1	S2	Score
C	G	-1
C	G	-1
G	G	+1
Total		**-1**

Multiple Sequence Alignment

For **phylogeny, function prediction**, or **motif discovery**.

Tool: Clustal Omega, MAFFT Command-line example with Clustal Omega



A	T	C	G
A	T	G	G
A	C	G	

S1	S2	Score
A	A	+1
A	A	+1
A	A	+1
Total		3

S1	S2	Score
T	T	+1
T	C	-1
T	C	-1
Total		**-1**

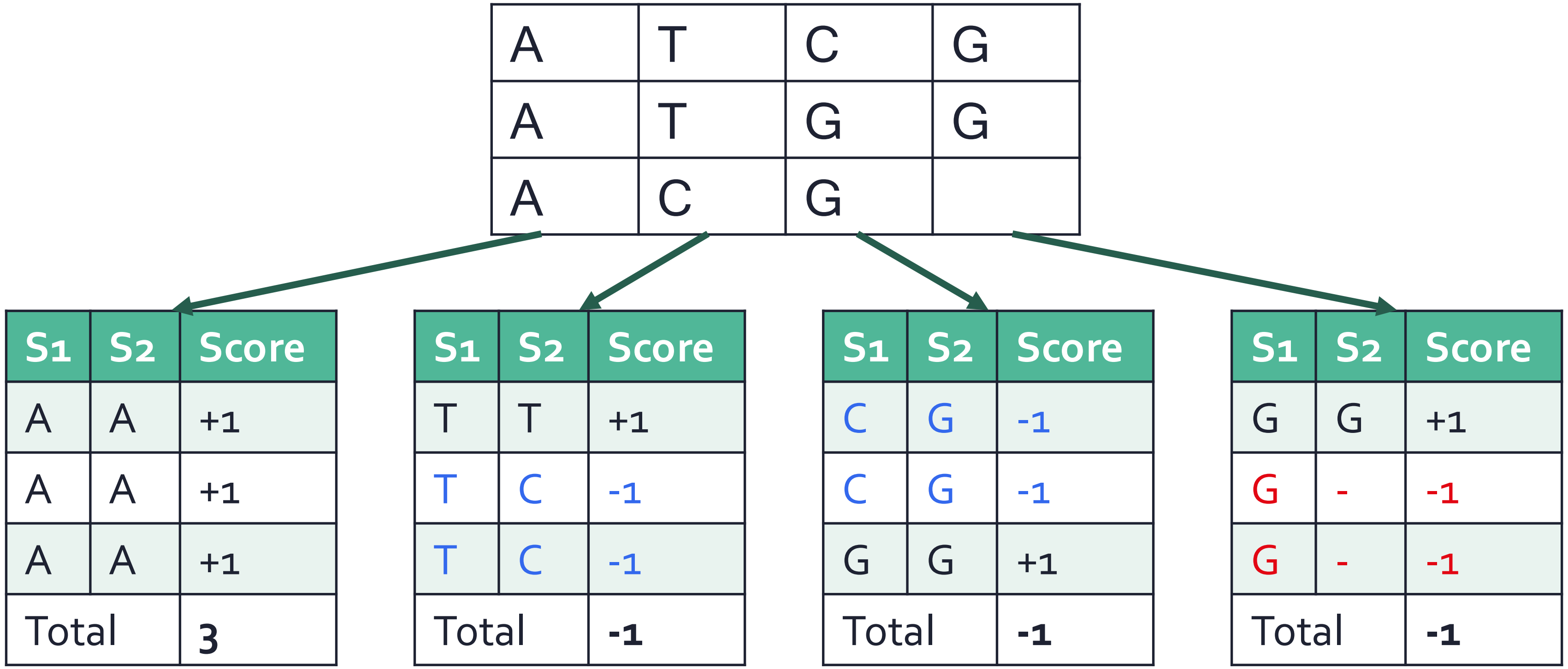
S1	S2	Score
C	G	-1
C	G	-1
G	G	+1
Total		**-1**

S1	S2	Score
G	G	+1
G	-	-1
G	-	-1
Total		**-1**

Multiple Sequence Alignment

For **phylogeny, function prediction**, or **motif discovery**.

Tool: Clustal Omega, MAFFT Command-line example with Clustal Omega



A	T	C	G
A	T	G	G
A	C	G	

S1	S2	Score
A	A	+1
A	A	+1
A	A	+1
Total		3

S1	S2	Score
T	T	+1
T	C	-1
T	C	-1
Total		**-1**

S1	S2	Score
C	G	-1
C	G	-1
G	G	+1
Total		**-1**

S1	S2	Score
G	G	+1
G	-	-1
G	-	-1
Total		**-1**

MSA score: 3 - 1 - 1 - 1 = 0

Read Alignment

Mapping **millions of short reads** to a **reference genome**.

Read Aligners: **BWA**, **Bowtie2, STAR, etc.**

Aim: find **substrings** in **large string**

reads

reference genome

Indexing

Aim: generate a ‘phonebook’ for fast searches

Indexing

Aim: generate a ‘phonebook’ for fast searches

Reference: TAATA$

↑

EOF

Indexing

Aim: generate a ‘phonebook’ for fast searches

Reference: TAATA$

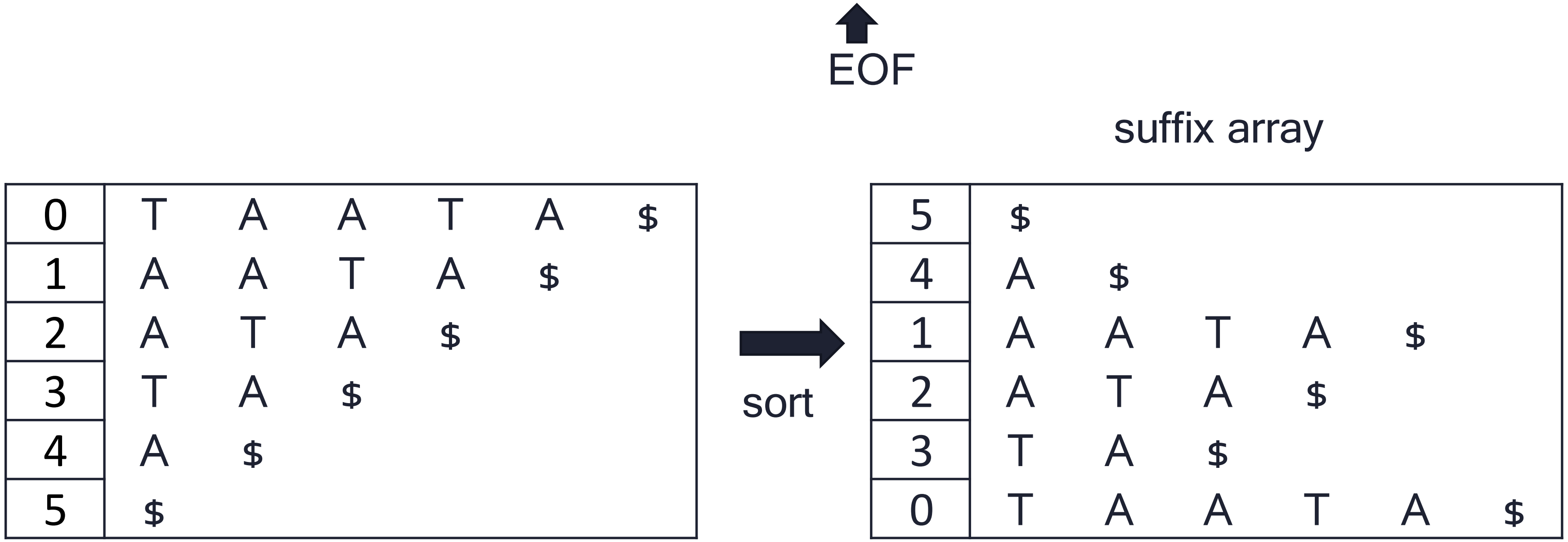
EOF

0	T	A	A	T	A	$
1	A	A	T	A	$	
2	A	T	A	$		
3	T	A	$			
4	A	$				
5	$					

Indexing

Aim: generate a ‘phonebook’ for fast searches

Reference: TAATA$



EOF

0	T	A	A	T	A	$
1	A	A	T	A	$	
2	A	T	A	$		
3	T	A	$			
4	A	$				
5	$					

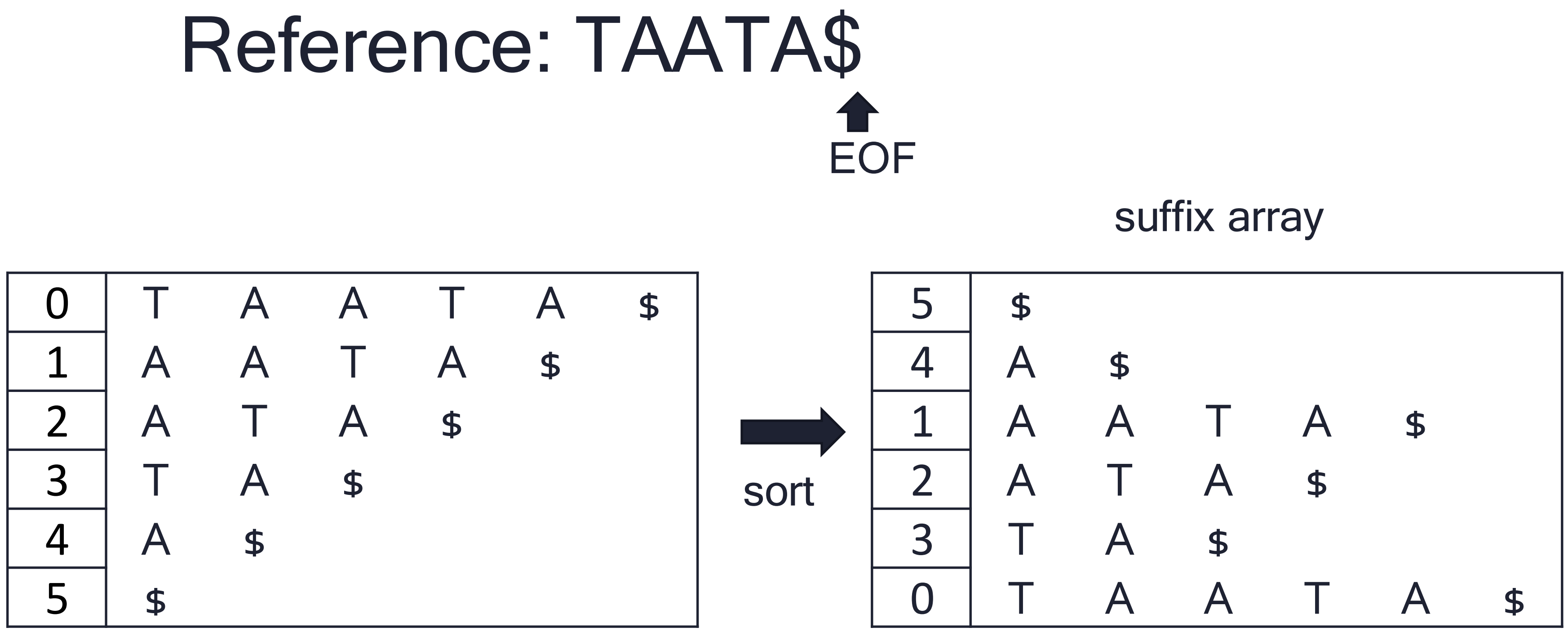
sort

suffix array

5	$					
4	A	$				
1	A	A	T	A	$	
2	A	T	A	$		
3	T	A	$			
0	T	A	A	T	A	$

Indexing

Aim: generate a ‘phonebook’ for fast searches



Reference: TAATA$

EOF

0	T	A	A	T	A	$
1	A	A	T	A	$	
2	A	T	A	$		
3	T	A	$			
4	A	$				
5	$					

sort

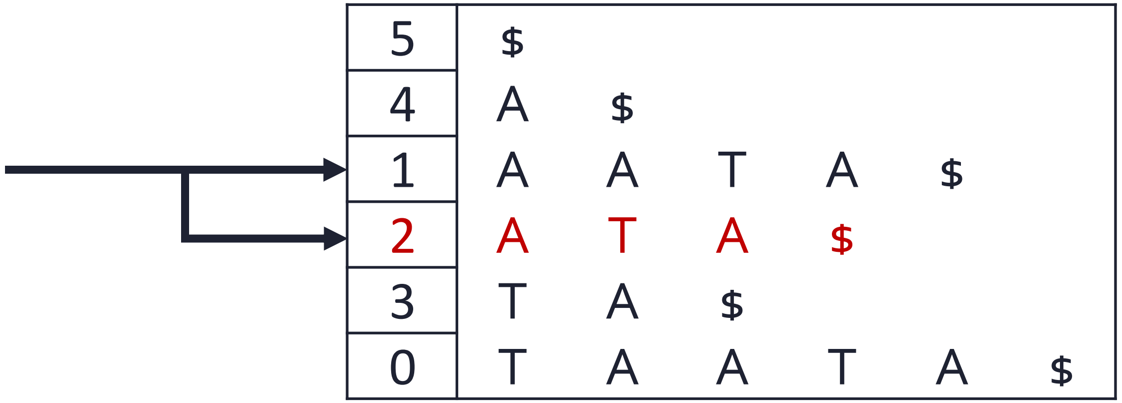
suffix array

5	$					
4	A	$				
1	A	A	T	A	$	
2	A	T	A	$		
3	T	A	$			
0	T	A	A	T	A	$

Querying

Reference: TAATA$

Query: ATA



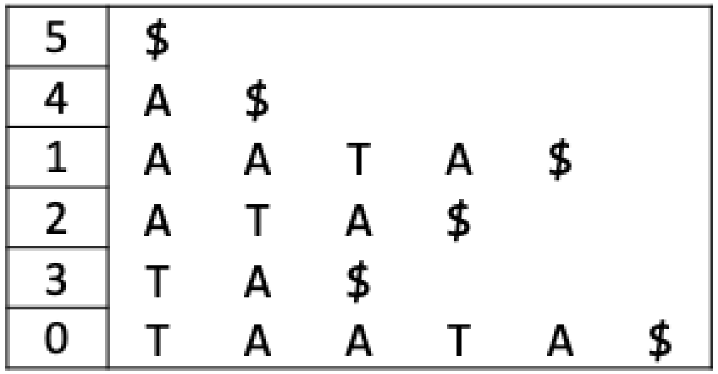
5	$					
4	A	$				
1	A	A	T	A	$	
2	A	T	A	$		
3	T	A	$			
0	T	A	A	T	A	$

Indexing and querying

Suffix array: large, same sequence stored multiple times

BWT: only first and last columns are stored -> still enables fast querying

suffix array



5	$					
4	A	$				
1	A	A	T	A	$	
2	A	T	A	$		
3	T	A	$			
0	T	A	A	T	A	$

Indexing and querying

Suffix array: large, same sequence stored multiple times

BWT: only first and last columns are stored -> still enables fast querying

suffix array

5	$					
4	A	$				
1	A	A	T	A	$	
2	A	T	A	$		
3	T	A	$			
0	T	A	A	T	A	$

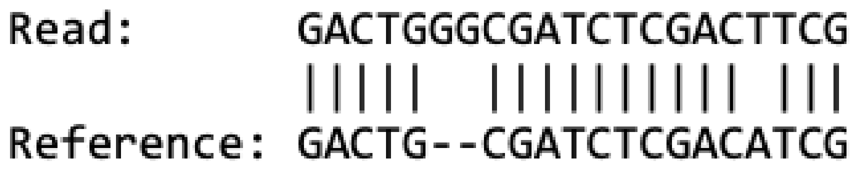
Burrows-**W**heeler **T**ransformation

$	T	A	A	T	A
A	$	T	A	A	T
A	A	T	A	$	T
A	T	A	$	T	A
T	A	$	T	A	A
T	A	A	T	A	$

https://www.youtube.com/watch?v=4WRANhDiSHM&t=1s

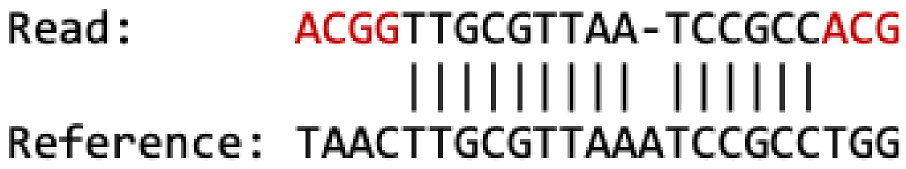
Global vs local

Global (end-to-end)



```
Read:      GACTGGGCGATCTCGACTTCG
           |||||  |||||||||| |||
Reference: GACTG--CGATCTCGACATCG
```

Local (allows for ‘clipping’)



```
Read:      ACGGTTGCGTTAA-TCCGCCACG
               ||||||||| ||||||
Reference: TAACTTGCGTTAAATCCGCCTGG
```

examples from: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#end-to-end-alignment-versus-local-alignment

Software

Basic alignment:

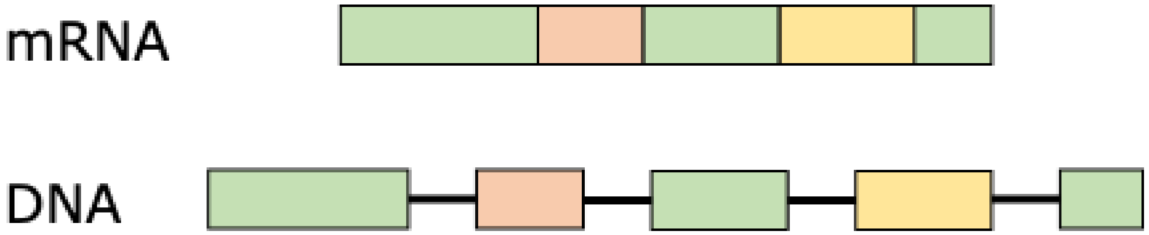
- bowtie2 (BWT; default = global)
- bwa-mem (BWT; default = local)

Splice-aware (RNA-seq):

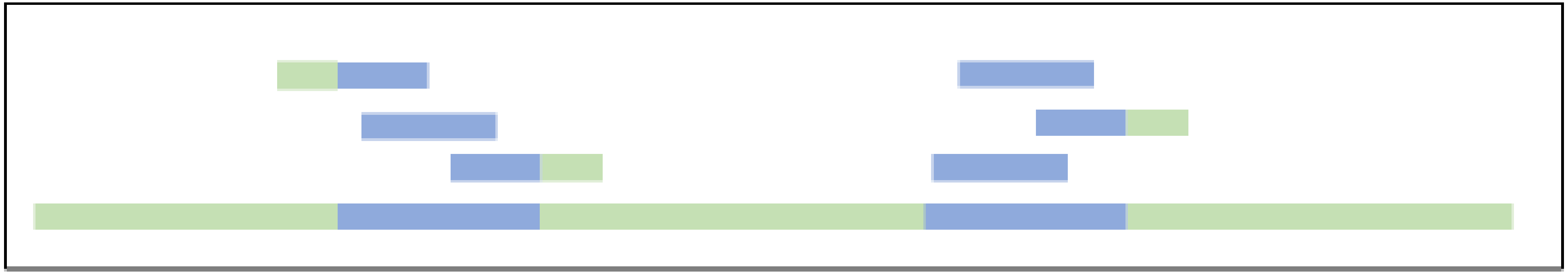
- hisat2
- STAR

Long reads + short reads + splice-aware:

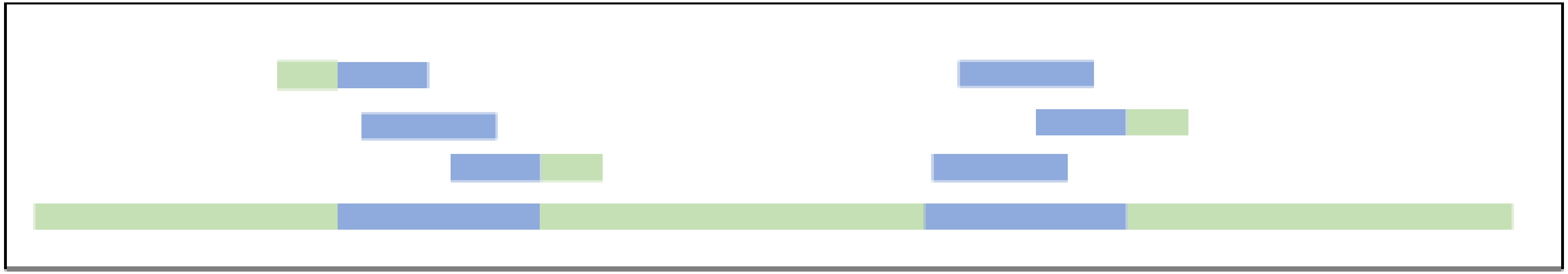
- minimap2



Mapping quality

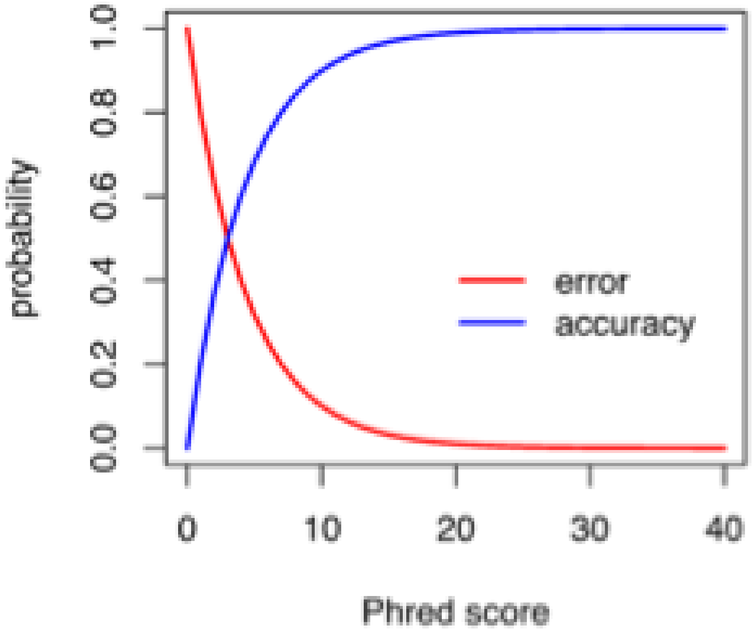


Mapping quality



$$MAPQ$$
$$= -10log_{10}\,\Pr\{mapping\ position\ is\ wrong\}$$

$$-10log_{10}\,(0.01) = 20$$
$$-10log_{10}\,(0.5) = 3$$



Summary

Global alignment aligns full sequences end-to-end

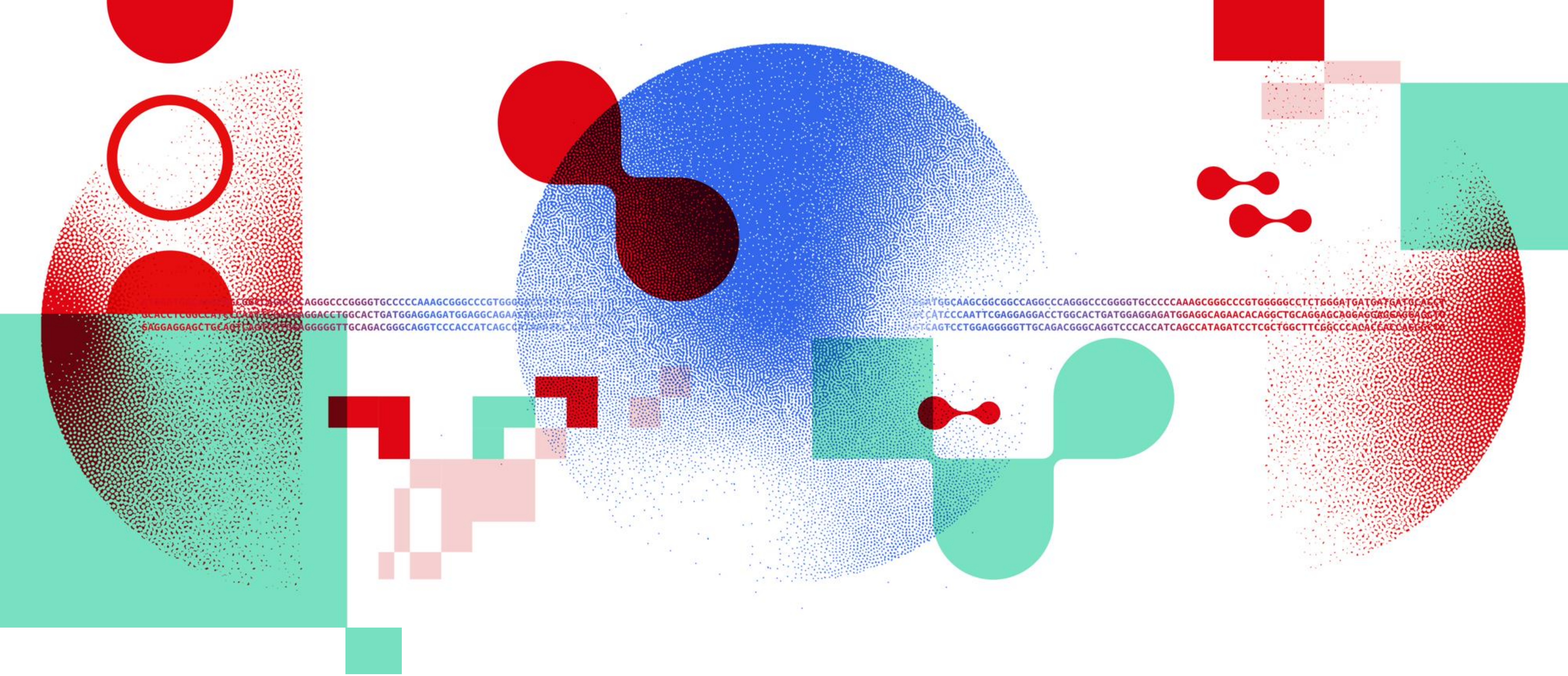
Local alignment is suited for partial matches like conserved domains

MSA aligns multiple sequences to reveal evolutionary or functional patterns

Read alignment maps millions of sequencing reads to a reference genome

Suffix arrays and BWT are used to enable fast searching within large genome indexes

Tools like **BWA**, **Bowtie2**, **STAR**, and **minimap2** support various alignment types, with specific strengths for short reads, long reads, or spliced transcripts


GCCGCCAGGCCAGGGCCCGGGGTGCCCCCAAAGCGGGCCCGTGGG
GGACCTGGCACTGATGGAGGAGATGGAGGCAGAA
GAGGAGGAGCTGCAGT AGGGGGGTTGCAGACGGGCAGGTCCCACCATCAGCC
ATGGCAAGCGGCGGCCAGGCCCAGGGCCCGGGGTGCCCCCAAAGCGGGCCCGTGGGGGCCTCTGGGATGATGATGATGCA
CATCCCAATTCGAGGAGGACCTGGCACTGATGGAGGAGATGGAGGCAGAACACAGGCTGCAGGAGCAGGAGGAGCAGGAGG
AGTCCTGGAGGGGGGTTGCAGACGGGCAGGTCCCACCATCAGCCATAGATCCTCGCTGGCTTCGGCCCACACCACCAGG

Thank you

DATA SCIENTISTS FOR LIFE

sib.swiss

elixir SWITZERLAND

SIB