Alignment and blast: basic principles and limitations to keep in mind for downstream analyses

David Dylus

ESCMID Postgraduate technical workshop

10/09/2019

Outline

- Motivation
  - Simple alignment
  - Complexity
- Querying a sequence against a database (BLAST)
  - Local vs Global alignment
  - How does it work?
  - What does the output mean?
- Mapping a sequence against a genome
  - Burrows Wheeler Transform
  - Tool Performance
  - Things to consider
- Multiple Sequence alignment
  - Basic Idea
  - Which tool is good for which purpose
- Conclusion

Why do you need alignment?

- Determine the expression of genes
  - Count the number of genes mapped to specific regions in the genome
- Find SNPs/SNVs in genomes
- Determine homology of genes
- Find the distance between two sequences
- Multiple sequence alignment for phylogenetic analysis

Simple alignment

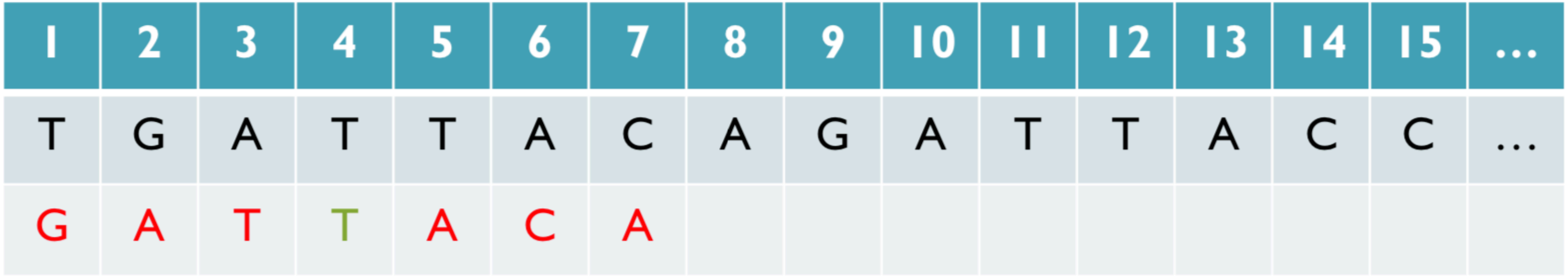
- Where is `GATTACA (len = 7)`?

```
T G A T T A C A G A T T A C C (len = 15)
```

1.) Align at the beginning
2.) Compare each nucleotide with each nucleotide and count the number of matches
3.) Move one position along the genome and go back to 2.)

Simple alignment

- Where is GATTACA?



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...
T	G	A	T	T	A	C	A	G	A	T	T	A	C	C	...
G	A	T	T	A	C	A									

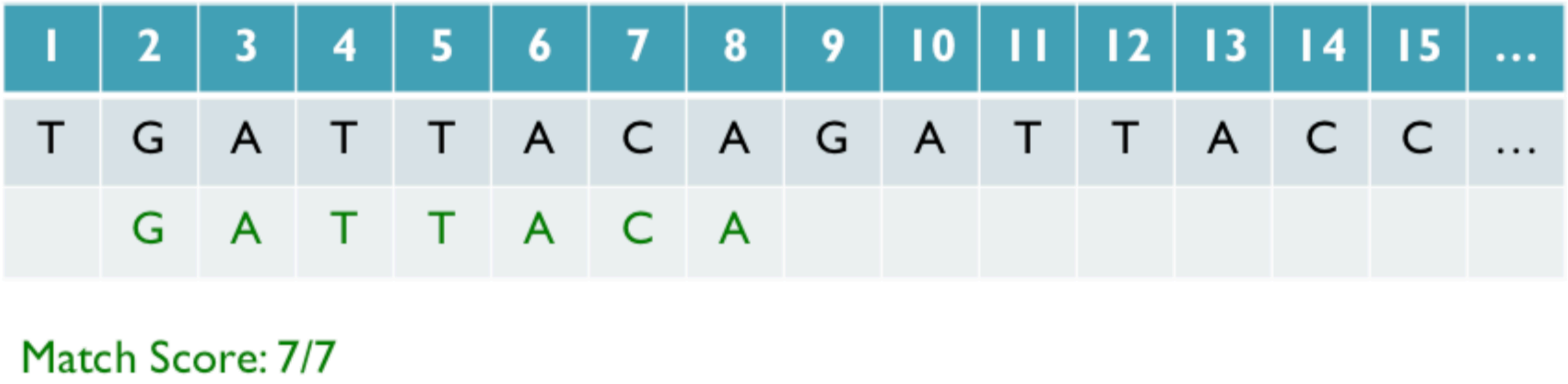
Match Score: 1/7

First iteration = 7 comparisons
Total = 7

http://schatzlab.cshl.edu/teaching/2010/Lecture%202%20-%20Sequence%20Alignment.pdf

Simple alignment

- Where is GATTACA?



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	...
T	G	A	T	T	A	C	A	G	A	T	T	A	C	C	...
	G	A	T	T	A	C	A								

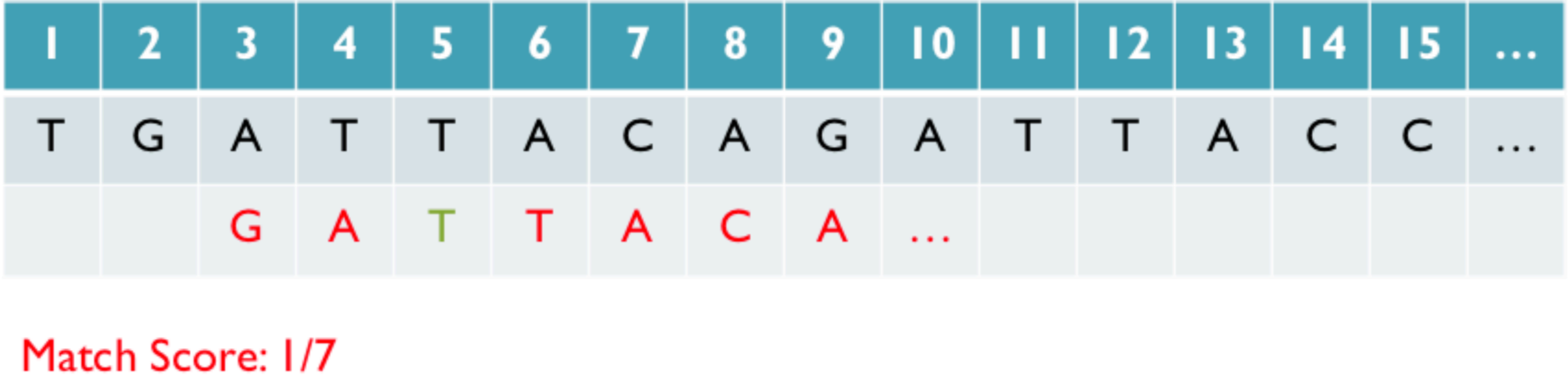
Match Score: 7/7

Second iteration = 7 comparisons
Total = 14

http://schatzlab.cshl.edu/teaching/2010/Lecture%202%20-%20Sequence%20Alignment.pdf

Simple alignment

- Where is GATTACA?



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	…
T	G	A	T	T	A	C	A	G	A	T	T	A	C	C	…
		G	A	T	T	A	C	A	…						

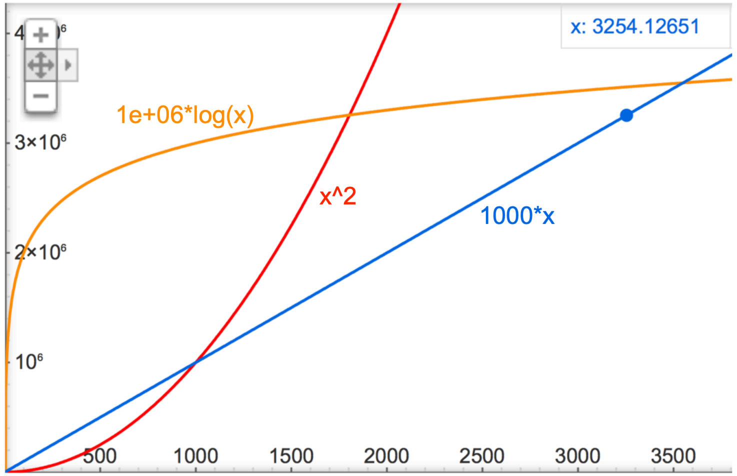
Match Score: 1/7

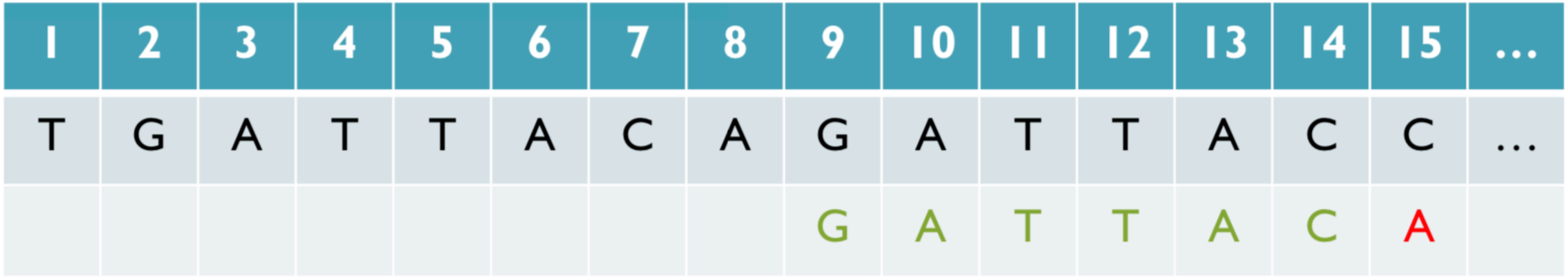
Third iteration = 7 comparisons
Total = 21

http://schatzlab.cshl.edu/teaching/2010/Lecture%202%20-%20Sequence%20Alignment.pdf

Simple alignment

- Where is `GATTACA`?





1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	…
T	G	A	T	T	A	C	A	G	A	T	T	A	C	C	…
								G	A	T	T	A	C	A	

Match Score: 6/7 <- We may be very interested in these imperfect matches
Especially if there are no perfect end-to-end matches

9th iteration = 7 comparisons
Total = 63

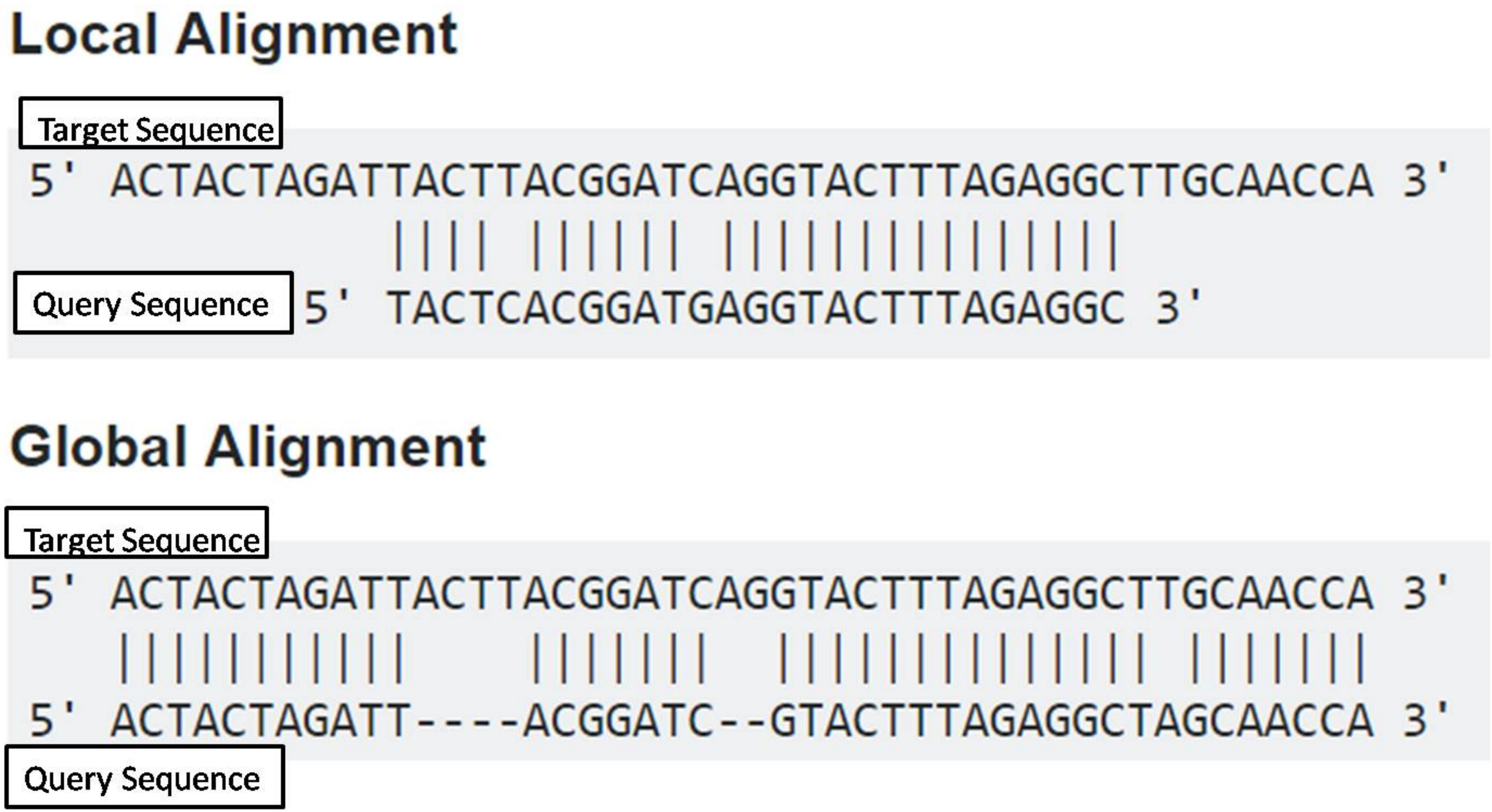
num_queries * (len_query * (len_genome - len_query + 1))

http://schatzlab.cshl.edu/teaching/2010/Lecture%202%20-%20Sequence%20Alignment.pdf

What do we see?

- Brute force will take a long time for many queries
- Indels will be characterized as mismatches and alignment might be scored wrongly
- If we want to find alignments with small number of mismatches we need to store all of these, therefore not only runtime problem but also memory

Alignment



Local Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| |||||||||||||||
```

Query Sequence

```
          5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Global Alignment

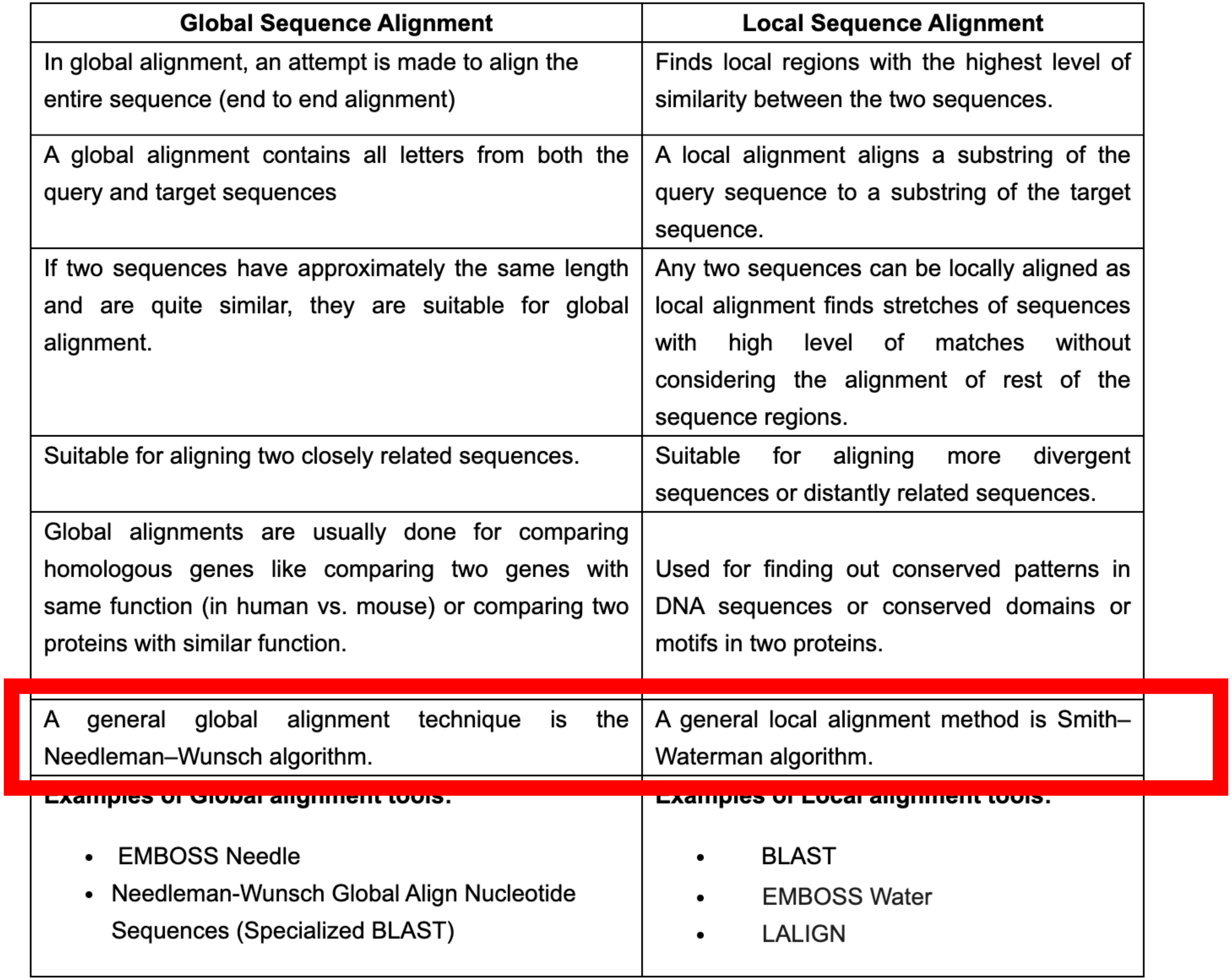
Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence

https://www.majordifferences.com/2016/05/difference-between-global-and-local.html

Alignment



Global Sequence Alignment	Local Sequence Alignment
In global alignment, an attempt is made to align the entire sequence (end to end alignment)	Finds local regions with the highest level of similarity between the two sequences.
A global alignment contains all letters from both the query and target sequences	A local alignment aligns a substring of the query sequence to a substring of the target sequence.
If two sequences have approximately the same length and are quite similar, they are suitable for global alignment.	Any two sequences can be locally aligned as local alignment finds stretches of sequences with high level of matches without considering the alignment of rest of the sequence regions.
Suitable for aligning two closely related sequences.	Suitable for aligning more divergent sequences or distantly related sequences.
Global alignments are usually done for comparing homologous genes like comparing two genes with same function (in human vs. mouse) or comparing two proteins with similar function.	Used for finding out conserved patterns in DNA sequences or conserved domains or motifs in two proteins.
A general global alignment technique is the Needleman–Wunsch algorithm.	A general local alignment method is Smith–Waterman algorithm.
Examples of Global alignment tools: • EMBOSS Needle • Needleman-Wunsch Global Align Nucleotide Sequences (Specialized BLAST)	**Examples of Local alignment tools:** • BLAST • EMBOSS Water • LALIGN

https://www.majordifferences.com/2016/05/difference-between-global-and-local.html

BLAST

Build database with all 3 letter words part of query

N K C K T P Q G Q R L V N Q W N K

20 Amino acids

3 letter words

Word	Position
NKC	1
KCK	2
CKT	3
...	
PQG	6
...	

BLAST

Use these 3 letter words to find high scoring neighbors by comparing it to all 20^3 possible words -> First SEED

N K C K T P Q G Q R L V N Q W N K

P Q G 24 = 8+8+6

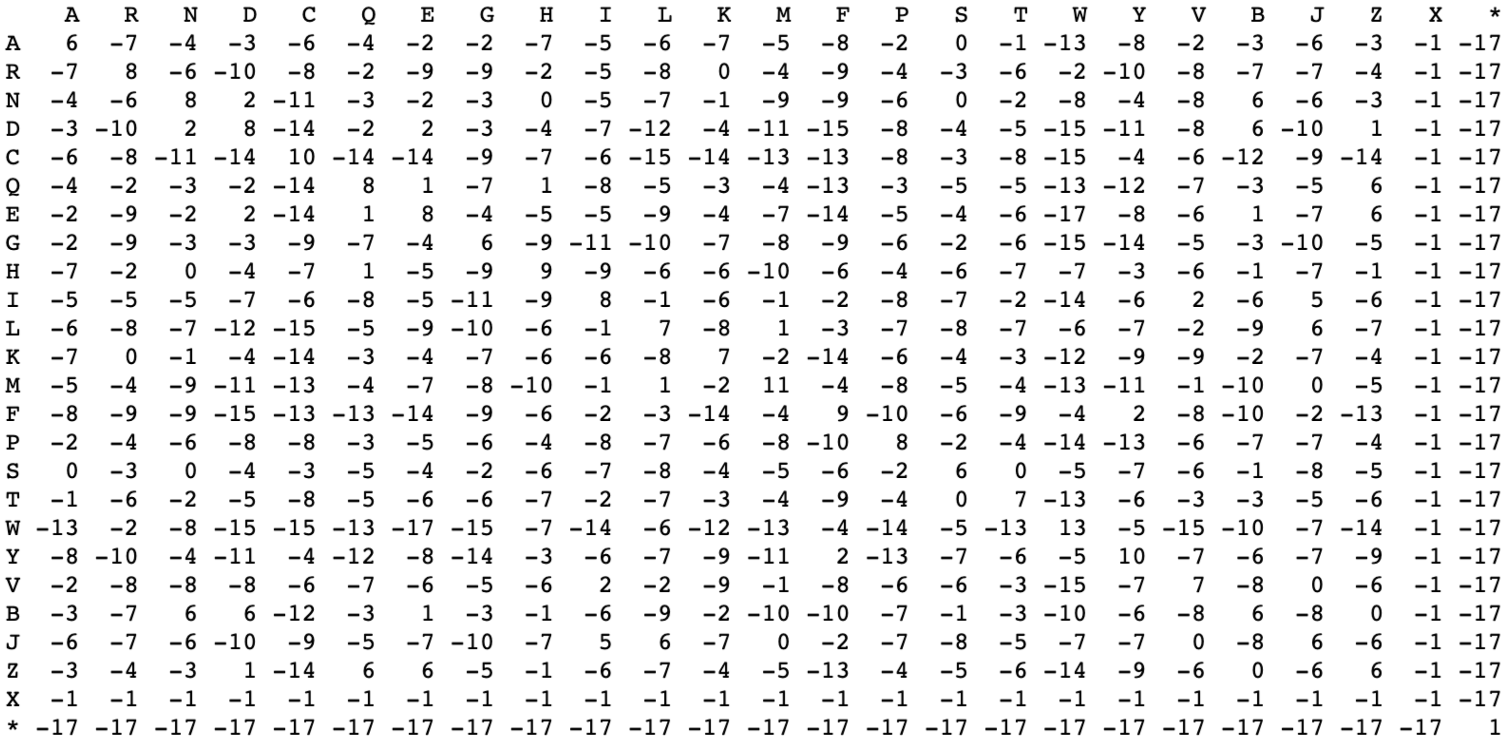
P E G 15

P R G 12

•

•

•

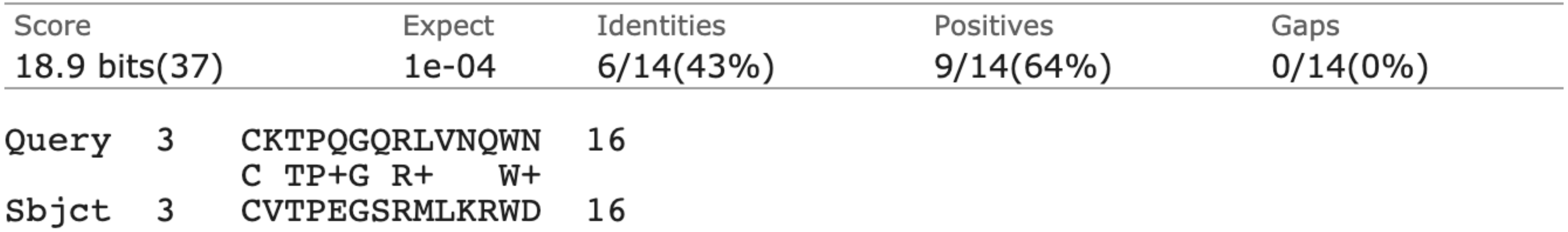


	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	J	Z	X	*
A	6	-7	-4	-3	-6	-4	-2	-2	-7	-5	-6	-7	-5	-8	-2	0	-1	-13	-8	-2	-3	-6	-3	-1	-17
R	-7	8	-6	-10	-8	-2	-9	-9	-2	-5	-8	0	-4	-9	-4	-3	-6	-2	-10	-8	-7	-7	-4	-1	-17
N	-4	-6	8	2	-11	-3	-2	-3	0	-5	-7	-1	-9	-9	-6	0	-2	-8	-4	-8	6	-6	-3	-1	-17
D	-3	-10	2	8	-14	-2	2	-3	-4	-7	-12	-4	-11	-15	-8	-4	-5	-15	-11	-8	6	-10	1	-1	-17
C	-6	-8	-11	-14	10	-14	-14	-9	-7	-6	-15	-14	-13	-13	-8	-3	-8	-15	-4	-6	-12	-9	-14	-1	-17
Q	-4	-2	-3	-2	-14	8	1	-7	1	-8	-5	-3	-4	-13	-3	-5	-5	-13	-12	-7	-3	-5	6	-1	-17
E	-2	-9	-2	2	-14	1	8	-4	-5	-5	-9	-4	-7	-14	-5	-4	-6	-17	-8	-6	1	-7	6	-1	-17
G	-2	-9	-3	-3	-9	-7	-4	6	-9	-11	-10	-7	-8	-9	-6	-2	-6	-15	-14	-5	-3	-10	-5	-1	-17
H	-7	-2	0	-4	-7	1	-5	-9	9	-9	-6	-6	-10	-6	-4	-6	-7	-7	-3	-6	-1	-7	-1	-1	-17
I	-5	-5	-5	-7	-6	-8	-5	-11	-9	8	-1	-6	-1	-2	-8	-7	-2	-14	-6	2	-6	5	-6	-1	-17
L	-6	-8	-7	-12	-15	-5	-9	-10	-6	-1	7	-8	1	-3	-7	-8	-7	-6	-7	-2	-9	6	-7	-1	-17
K	-7	0	-1	-4	-14	-3	-4	-7	-6	-6	-8	7	-2	-14	-6	-4	-3	-12	-9	-9	-2	-7	-4	-1	-17
M	-5	-4	-9	-11	-13	-4	-7	-8	-10	-1	1	-2	11	-4	-8	-5	-4	-13	-11	-1	-10	0	-5	-1	-17
F	-8	-9	-9	-15	-13	-13	-14	-9	-6	-2	-3	-14	-4	9	-10	-6	-9	-4	2	-8	-10	-2	-13	-1	-17
P	-2	-4	-6	-8	-8	-3	-5	-6	-4	-8	-7	-6	-8	-10	8	-2	-4	-14	-13	-6	-7	-7	-4	-1	-17
S	0	-3	0	-4	-3	-5	-4	-2	-6	-7	-8	-4	-5	-6	-2	6	0	-5	-7	-6	-1	-8	-5	-1	-17
T	-1	-6	-2	-5	-8	-5	-6	-6	-7	-2	-7	-3	-4	-9	-4	0	7	-13	-6	-3	-3	-5	-6	-1	-17
W	-13	-2	-8	-15	-15	-13	-17	-15	-7	-14	-6	-12	-13	-4	-14	-5	-13	13	-5	-15	-10	-7	-14	-1	-17
Y	-8	-10	-4	-11	-4	-12	-8	-14	-3	-6	-7	-9	-11	2	-13	-7	-6	-5	10	-7	-6	-7	-9	-1	-17
V	-2	-8	-8	-8	-6	-7	-6	-5	-6	2	-2	-9	-1	-8	-6	-6	-3	-15	-7	7	-8	0	-6	-1	-17
B	-3	-7	6	6	-12	-3	1	-3	-1	-6	-9	-2	-10	-10	-7	-1	-3	-10	-6	-8	6	-8	0	-1	-17
J	-6	-7	-6	-10	-9	-5	-7	-10	-7	5	6	-7	0	-2	-7	-8	-5	-7	-7	0	-8	6	-6	-1	-17
Z	-3	-4	-3	1	-14	6	6	-5	-1	-6	-7	-4	-5	-13	-4	-5	-6	-14	-9	-6	0	-6	6	-1	-17
X	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-17
*	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	-17	1

BLAST

Once a HSP is found extend the seed using dynamic programming until the score drops below a specific threshold and report the output

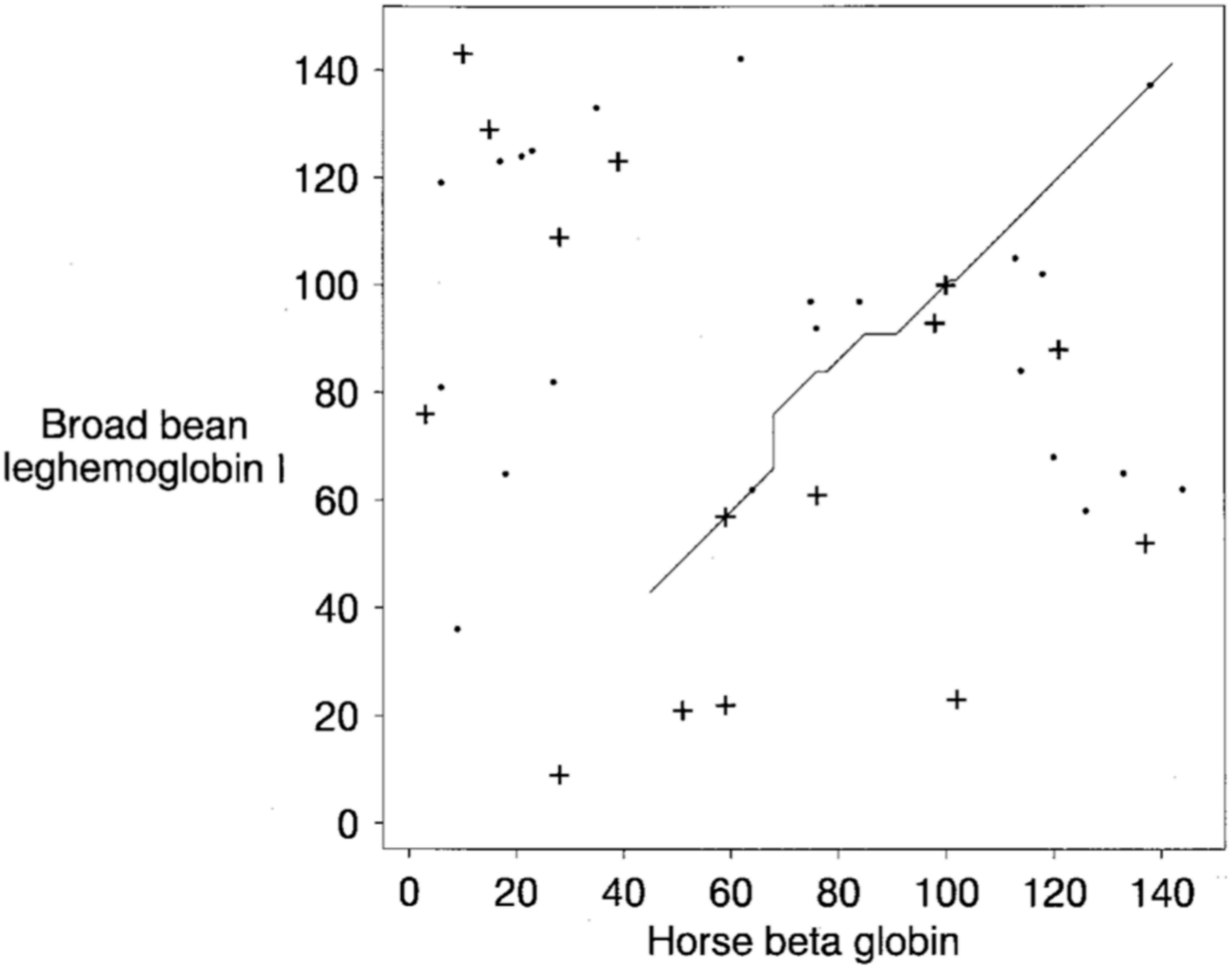
```
Query:  N   K   C   K  T  P  Q  G  Q   R  L   V   N   Q   W   N   K
DB:     D   S   C   V  T  P  E  G  S   R  M   L   K   R   W   D   S

        2  -4  10  -9  7  8  1  6  -5  8  1  -2  -1  -2  13   2  -4
```



Score	Expect	Identities	Positives	Gaps
18.9 bits(37)	1e-04	6/14(43%)	9/14(64%)	0/14(0%)

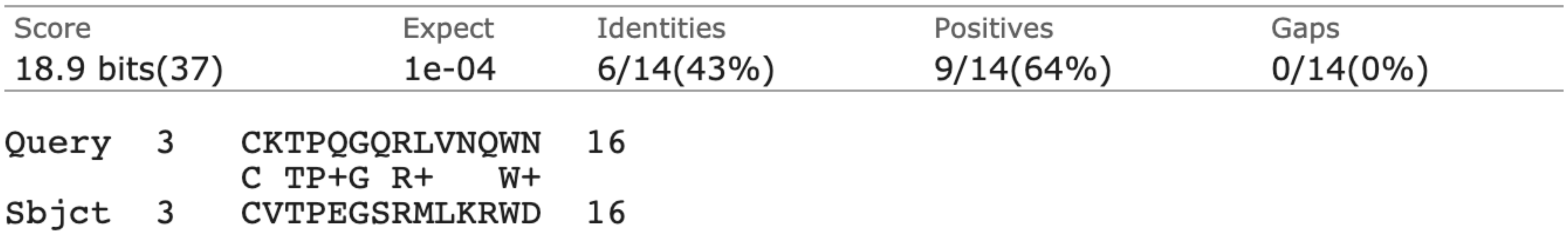
```
Query  3    CKTPQGQRLVNQWN   16
            C TP+G R+    W+
Sbjct  3    CVTPEGSRMLKRWD   16
```

BLAST


Broad bean
leghemoglobin I
140
120
100
80
60
40
20
0
0
20
40
60
80
100
120
140
Horse beta globin

BLAST Stats

- Because we are using a matrix with different values for the alignment between two residues score and length of the alignment are not comparable
- E-value depends on database size and especially when using custom database can be misleading



Score	Expect	Identities	Positives	Gaps
18.9 bits(37)	1e-04	6/14(43%)	9/14(64%)	0/14(0%)

```
Query  3    CKTPQGQRLVNQWN  16
            C TP+G R+   W+
Sbjct  3    CVTPEGSRMLKRWD  16
```

BLAST

Program	Query Sequence	Target Sequence
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide, six-frame translation	Protein
TBLASTN	Protein	Nucleotide, six-frame translation
TBLASTX	Nucleotide, six-frame translation	Nucleotide, six-frame translation

BLAST - Summary

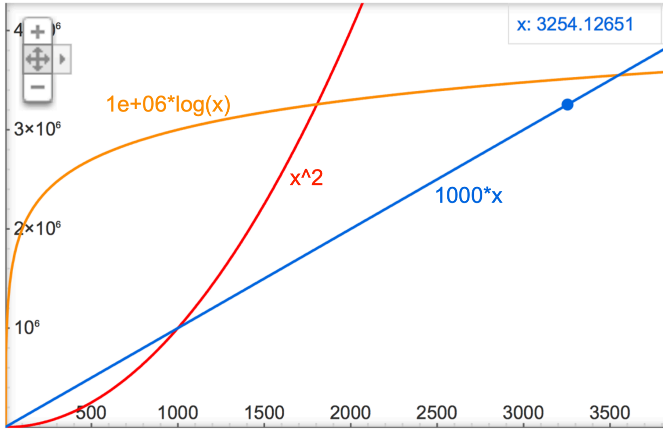
- Searches for high-scoring segment pairs (HSPs)
    - Look for high scoring words of length W
    - Compile list L of all W-mers that score >T with some word in query sequence
    - Scan database for words in L
    - When some word found: Extend alignment
    - When score drops more than X below hitherto best score stop extension
    - Report all words with large score S
- Results in all plausible alignments between two sequences

Questions BLAST

- Does BLAST give you the best possible alignment?

- Is the e-value a good choice to measure how well aligned it is?

- Would you use blast to align reads to a genome?



BLAST is great but SLOW

O(Number Queries * Length Query * DB Size)

Typical:

Align 1.000.000.000 reads with 100bp length to genome with 3.000.000 bp (Bacteria)

Alignment Problem: reads against Genome

- Detection of SNV
- Detection of SNPs
- Detection of presence
- Expression of genes / regions

How to align sequence reads to a reference?

Seeding: for each position, find longest exact match covering the position

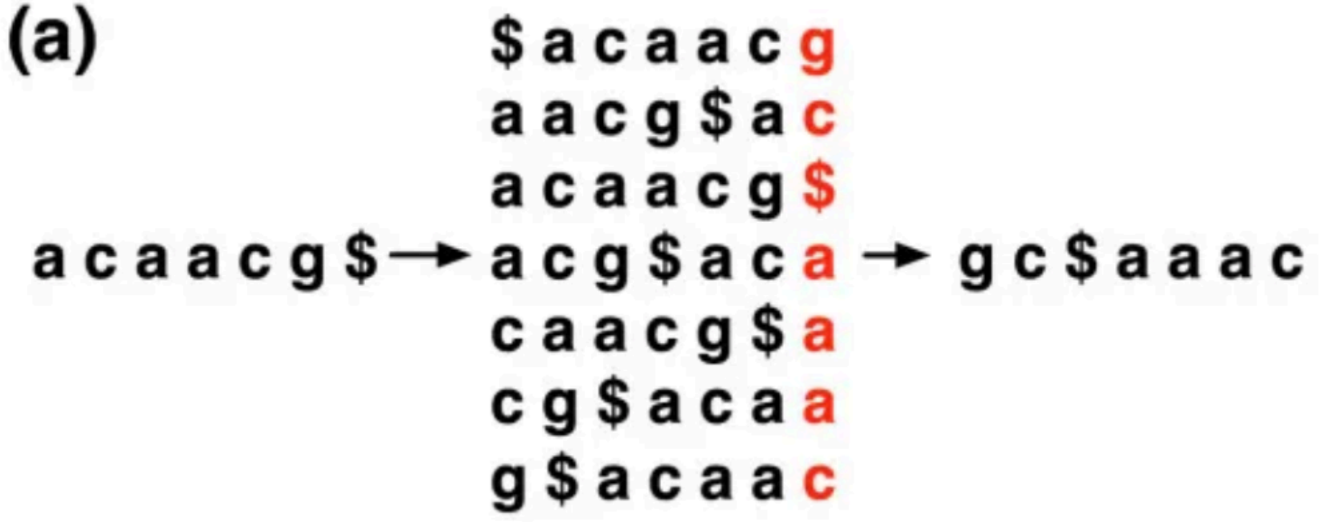
Index genome with Burrows-Wheeler Transform

Slide kindly provided by Aitana Lebrand

Align reads against genome

Burrows Wheel Transform (BWT):

- Computational trick on how data can be stored and searched
- Write all permutations of string and sort alphabetically ($ first)



Current Pos	Nucleotide	Old Pos
1	g_1	7
2	c_1	5
3	$	1
4	a_1	2
5	a_2	3
6	a_3	4
7	c_2	6

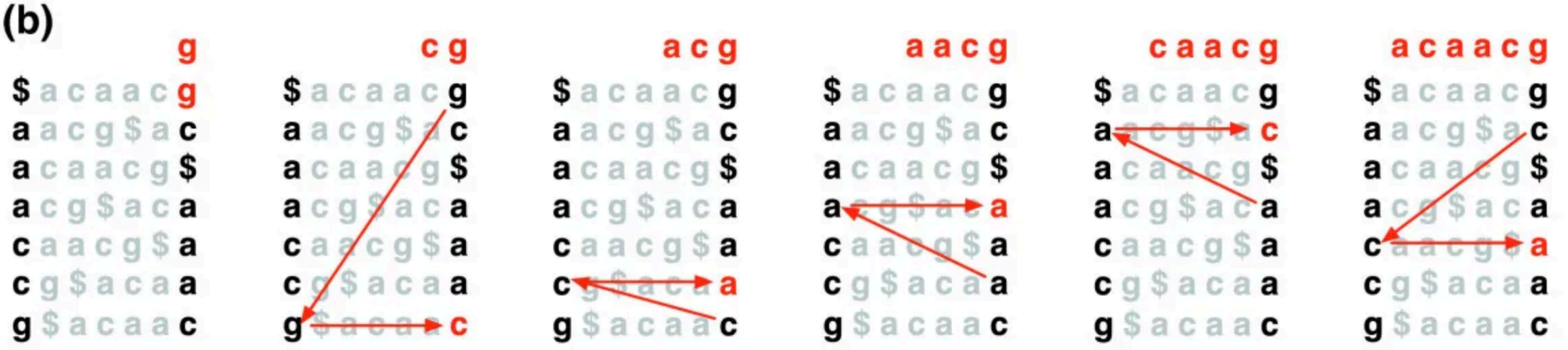
Taken from Langmead et al, 2010, Genome Biology

Align reads against genome

Current Pos	Nucleotide	Old Pos
1	g_1	7
2	c_1	5
3	$	1
4	a_1	2
5	a_2	3
6	a_3	4
7	c_2	6

Burrows Wheel Transform (BWT):

- From last column alone we are able to reconstruct the whole genome



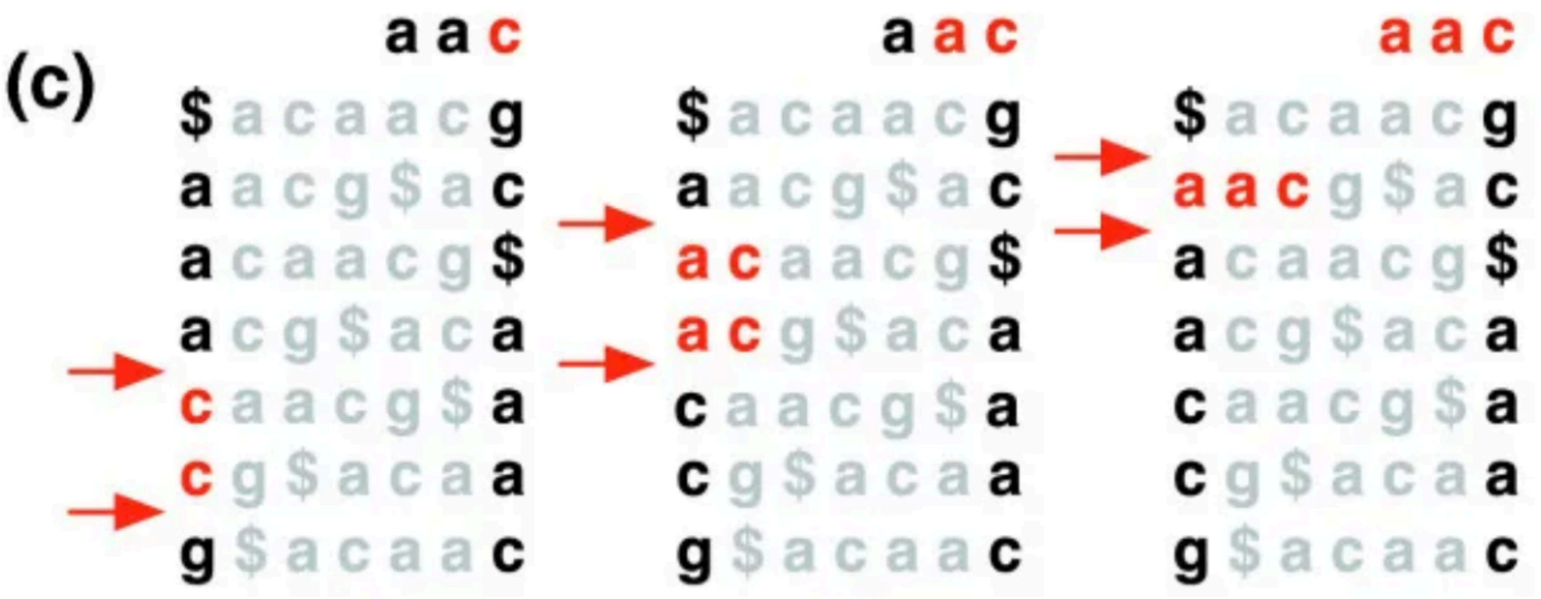
```
g          cg         acg        aacg       caacg      acaacg
$acaacg    $acaacg    $acaacg    $acaacg    $acaacg    $acaacg
aacg$ac    aacg$ac    aacg$ac    aacg$ac    aacg$ac    aacg$ac
acaacg$    acaacg$    acaacg$    acaacg$    acaacg$    acaacg$
acg$aca    acg$aca    acg$aca    acg$aca    acg$aca    acg$aca
caacg$a    caacg$a    caacg$a    caacg$a    caacg$a    caacg$a
cg$acaa    cg$acaa    cg$acaa    cg$acaa    cg$acaa    cg$acaa
g$acaac    g$acaac    g$acaac    g$acaac    g$acaac    g$acaac
```

Taken from Langmead et al, 2010, Genome Biology

Align reads against genome

Burrows Wheel Transform (BWT):

- Exact matching can be performed in O(len query seq) time
- We know that a [2:4], c [5:6], g [7]

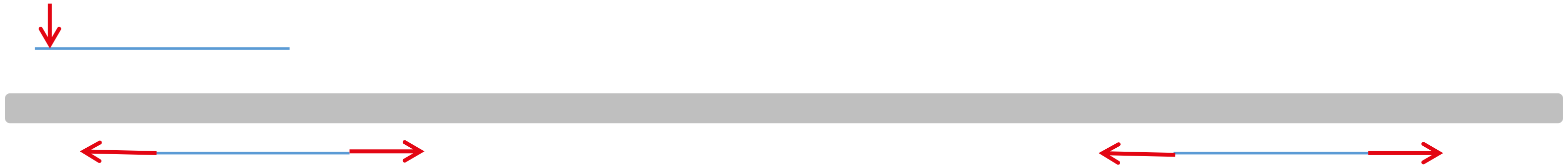


Current Pos	Nucleotide	Old Pos
1	g_1	7
2	c_1	5
3	$	1
4	a_1	2
5	a_2	3
6	a_3	4
7	c_2	6

Taken from Langmead et al, 2010, Genome Biology

How to align sequence reads to a reference?

Extend seed (allow both local and end-to-end alignments with inexact matching)



Famous mappers

BWA (Li and Durbin 2009)

Bowtie2 (Langemead et al. 2009)

Slide kindly provided by Aitana Lebrand

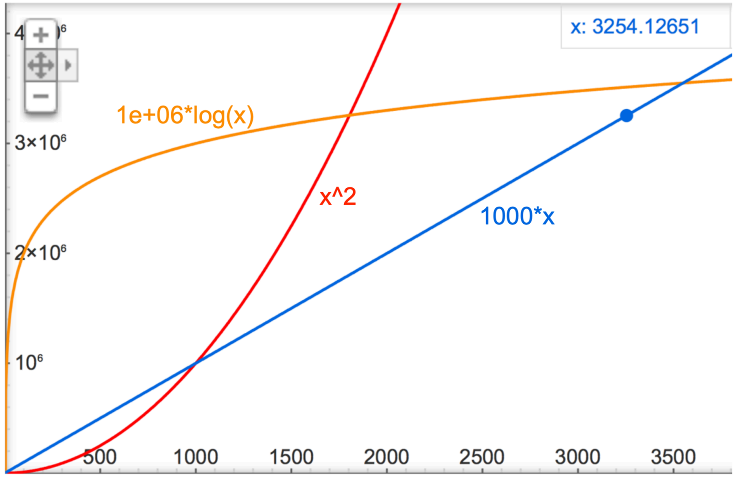
Align reads against genome

Burrows Wheel Transform (BWT):

- Inexact matching uses extend trick where at positions that not overlap quality of base call is considered and possible substitutions

Taken from Langmead et al, 2010, Genome Biology

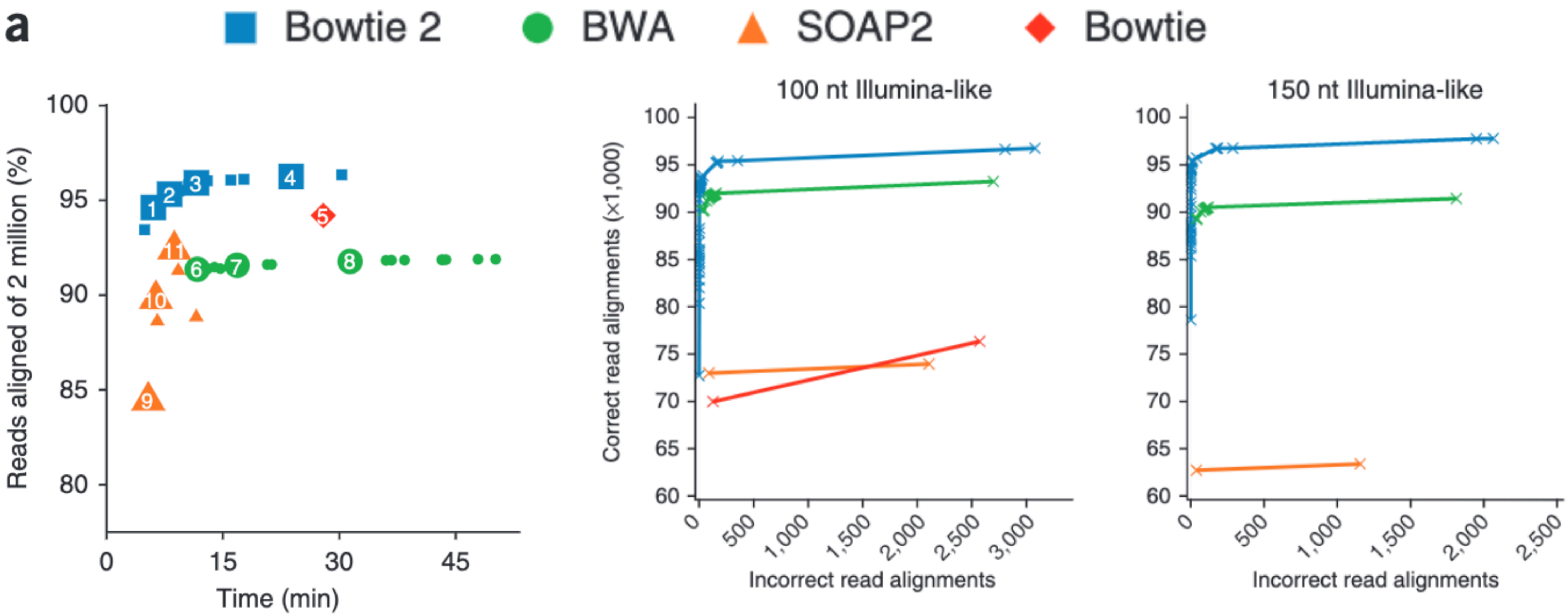
Align reads against genome



Burrows Wheel Transform (BWT):

- Transformation possible in O(n)
- Reconstruct Genome from BWT(Genome) in time O(|genome|)
- Search for all exact occurrences of read in time O(|read|)
- BWT(Genome) is easier to compress than Genome

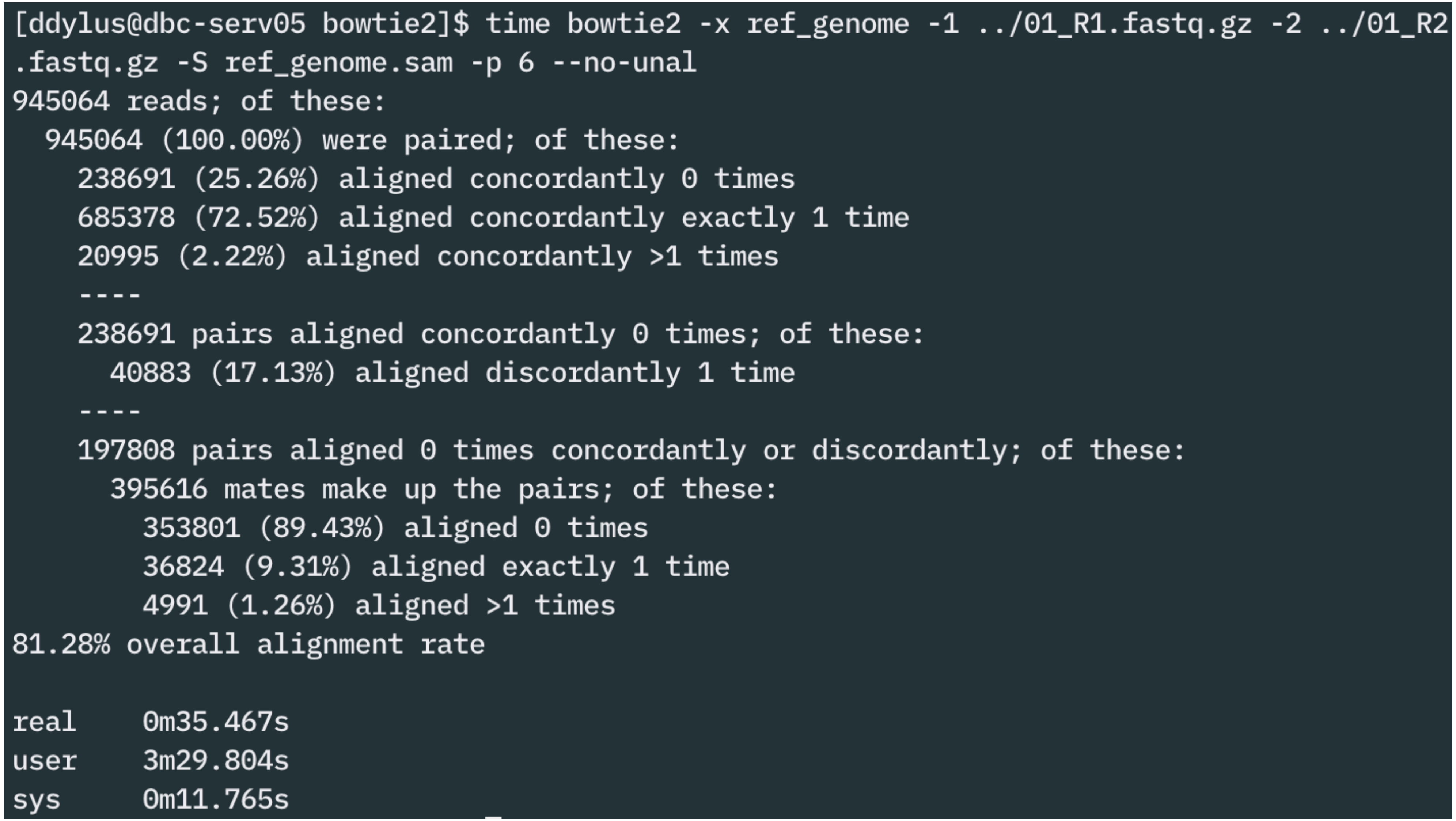
Comparison of Tools



Be aware that this benchmark was done for Bowtie2 and other benchmarks might be different

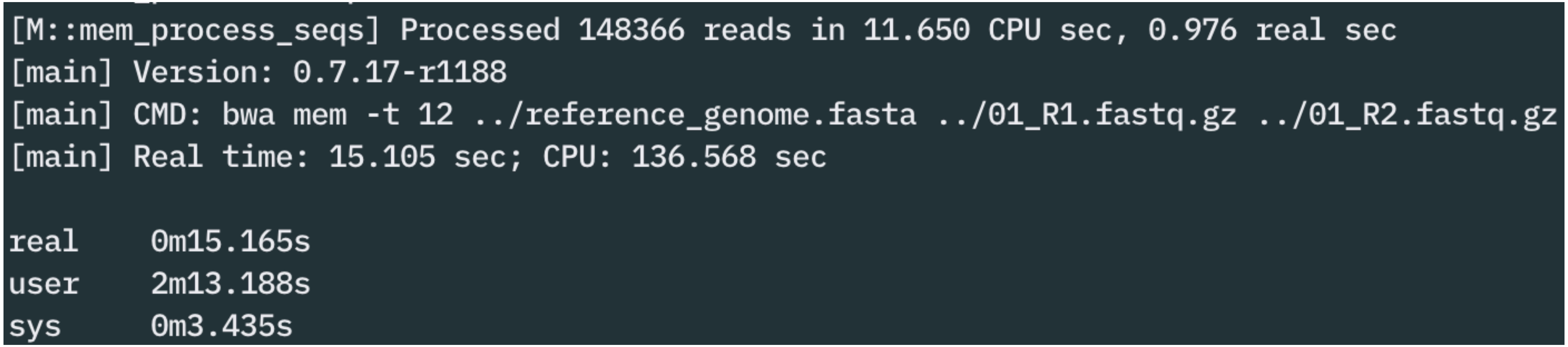
Taken from Langmead et al, 2012, Nature Methods

Bowtie 2 typical run



```
[ddylus@dbc-serv05 bowtie2]$ time bowtie2 -x ref_genome -1 ../01_R1.fastq.gz -2 ../01_R2
.fastq.gz -S ref_genome.sam -p 6 --no-unal
945064 reads; of these:
  945064 (100.00%) were paired; of these:
    238691 (25.26%) aligned concordantly 0 times
    685378 (72.52%) aligned concordantly exactly 1 time
    20995 (2.22%) aligned concordantly >1 times
    ----
    238691 pairs aligned concordantly 0 times; of these:
      40883 (17.13%) aligned discordantly 1 time
    ----
    197808 pairs aligned 0 times concordantly or discordantly; of these:
      395616 mates make up the pairs; of these:
        353801 (89.43%) aligned 0 times
        36824 (9.31%) aligned exactly 1 time
        4991 (1.26%) aligned >1 times
81.28% overall alignment rate

real	0m35.467s
user	3m29.804s
sys	0m11.765s
```

BWA typical runs



```
[M::mem_process_seqs] Processed 148366 reads in 11.650 CPU sec, 0.976 real sec
[main] Version: 0.7.17-r1188
[main] CMD: bwa mem -t 12 ../reference_genome.fasta ../01_R1.fastq.gz ../01_R2.fastq.gz
[main] Real time: 15.105 sec; CPU: 136.568 sec

real	0m15.165s
user	2m13.188s
sys	0m3.435s
```

Aligning the whole read… or part of it

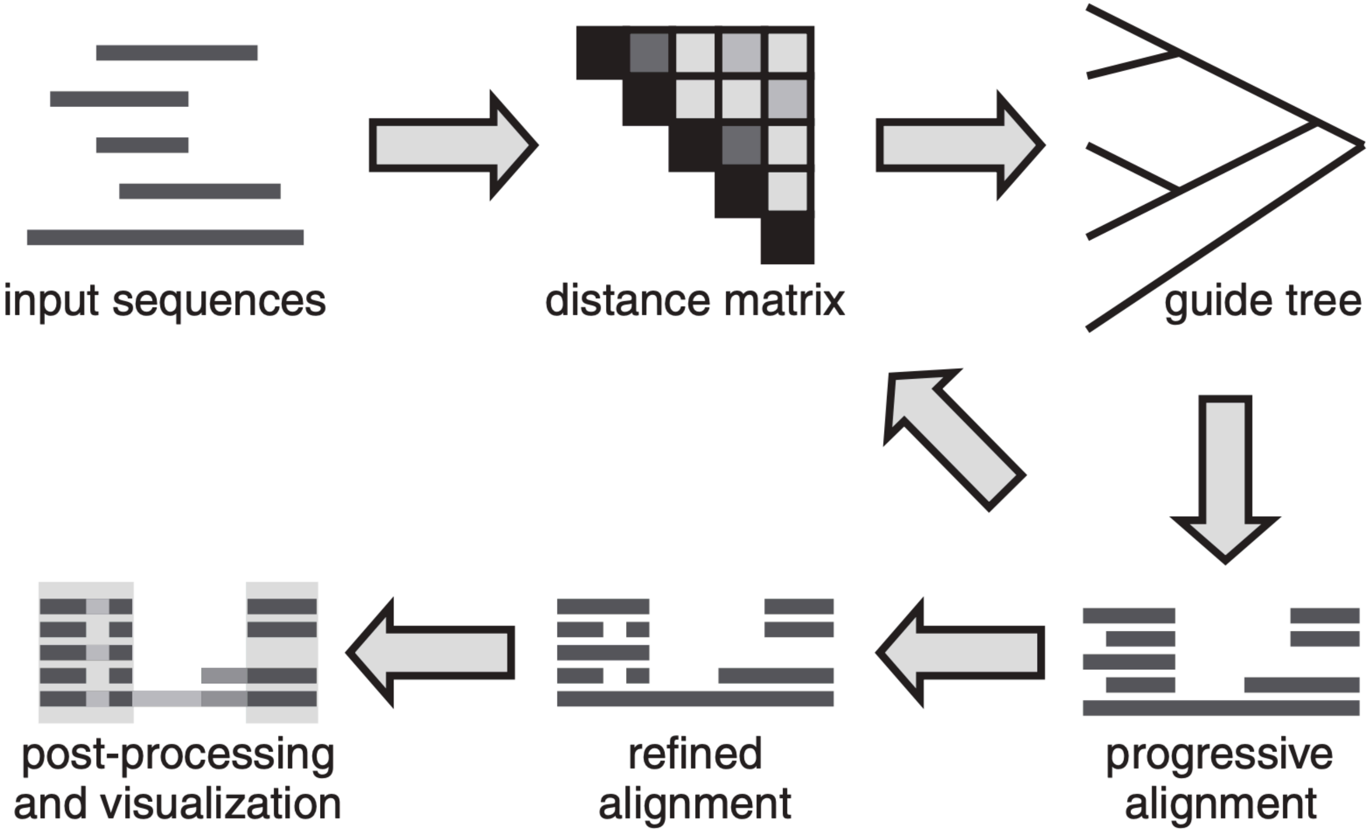
- Real life datasets are often not perfect:
- poor base call qualities, sequencing errors, insertions or deletions, structural variation, contaminants or adapter…

→ **Despite adapter/quality trimming and allowing for mismatches/indels, some parts of the reads may still consist of sequences not present in the genome**

- **Hard-clipping** (removed from BAM) **vs. soft-clipping** (still part of BAM; not used for SVN calling, depth calculation, but can be useful to look into translocations, deletions…)
- **! Neither soft nor hard clipped regions are displayed in a viewer!**

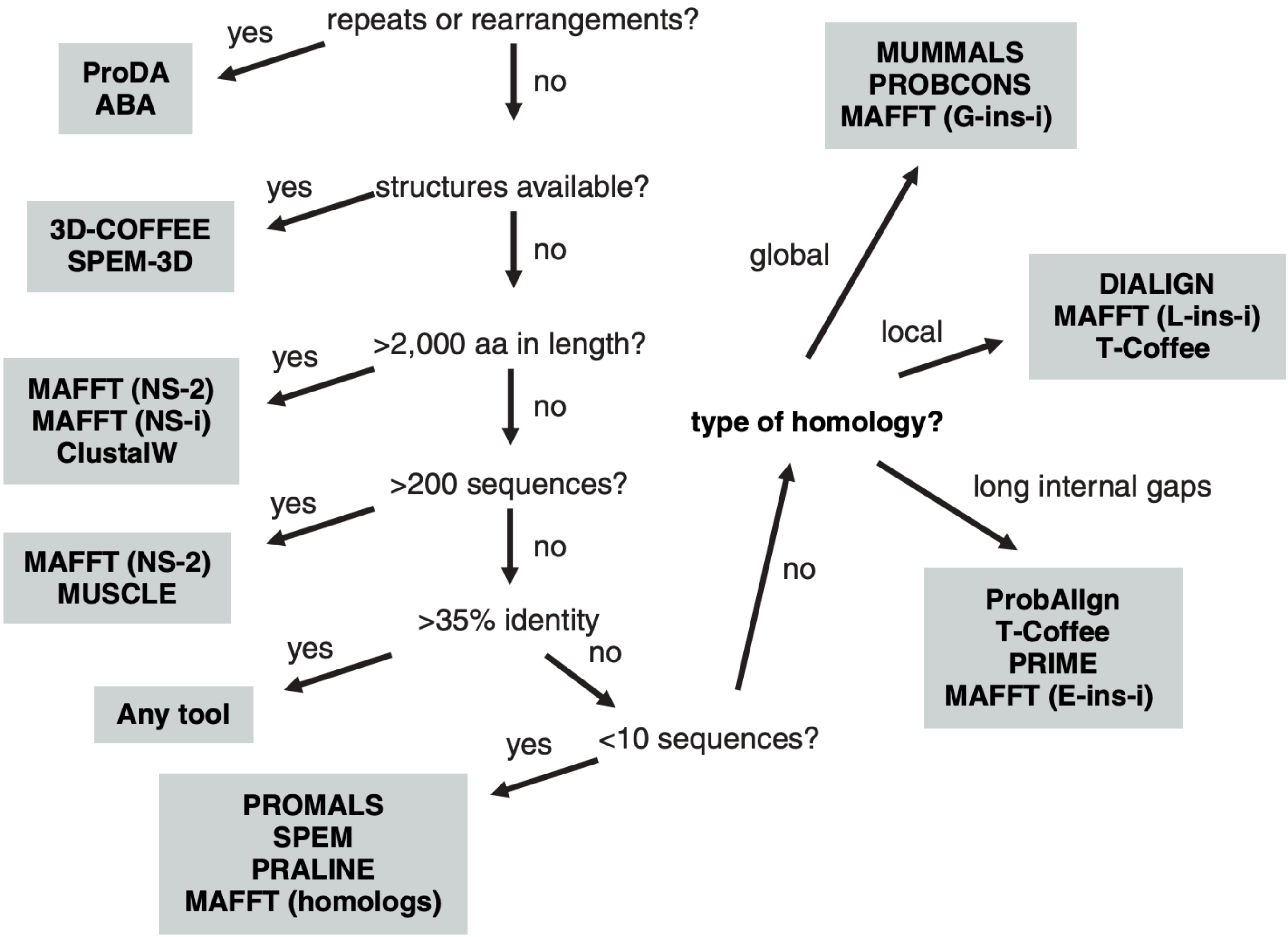
Slide kindly provided by Aitana Lebrand

What about Multiple Sequence Alignment



http://ai.stanford.edu/~chuongdo/papers/alignment_review.pdf

What about Multiple Sequence Alignment



http://ai.stanford.edu/~chuongdo/papers/alignment_review.pdf

Take home messages

- Be careful when comparing a blast result when using different search DBs
- Blast does not guarantee the most optimal alignment between you query and the obtained sequence from the DB
- Be aware that depending on read alignment tool you might end up with differences (for instance different SNPs)
- Do not just trust the MSA algorithm that is readily presented to you but decide based on your data

Source for this talk

Computational Biology:

- https://ocw.mit.edu/courses/biology/7-91j-foundations-of-computational-and-systems-biology-spring-2014/video-lectures/

Bowtie and BWA:

- http://merenlab.org/2015/06/23/comparing-different-mapping-software/

BWT:

- https://www.youtube.com/watch?v=4n7NPk5lwbI

Alignment:

- https://www.youtube.com/watch?v=hpb-mH-yjLc&list=PL2mpR0RYFQsBiCWVJSvVAO3OJ2t7DzoHA

Blast:

- https://www.ndsu.edu/pubweb/~mcclean/plsc411/Blast-explanation-lecture-and-overhead.pdf
- https://developer.ibm.com/articles/j-seqalign/
- http://csc.columbusstate.edu/carroll/7840/private/papers/BasicLocalAlignmentSearchTool-BLAST.pdf
- https://www.youtube.com/watch?v=SAweFv8I8ow
- http://web.math.ku.dk/~richard/courses/binf_project/Stinus-BLAST.pdf

References

MSA Benchmark

https://arxiv.org/pdf/1211.2160.pdf

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2995051/

https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz552/5530966

Mapper benchmark:

https://www.ecseq.com/support/benchmark.html

https://www.biostars.org/p/125020/