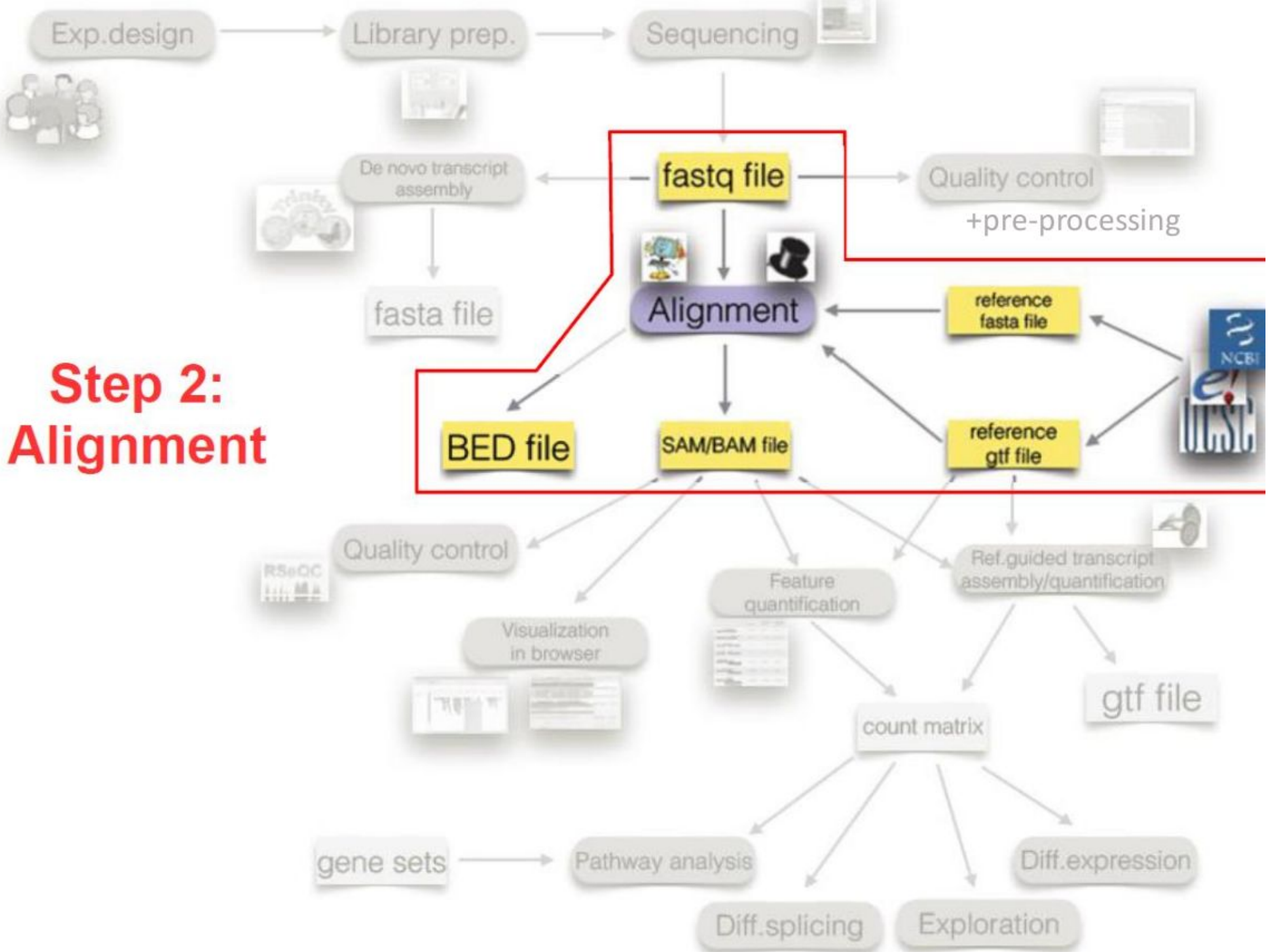
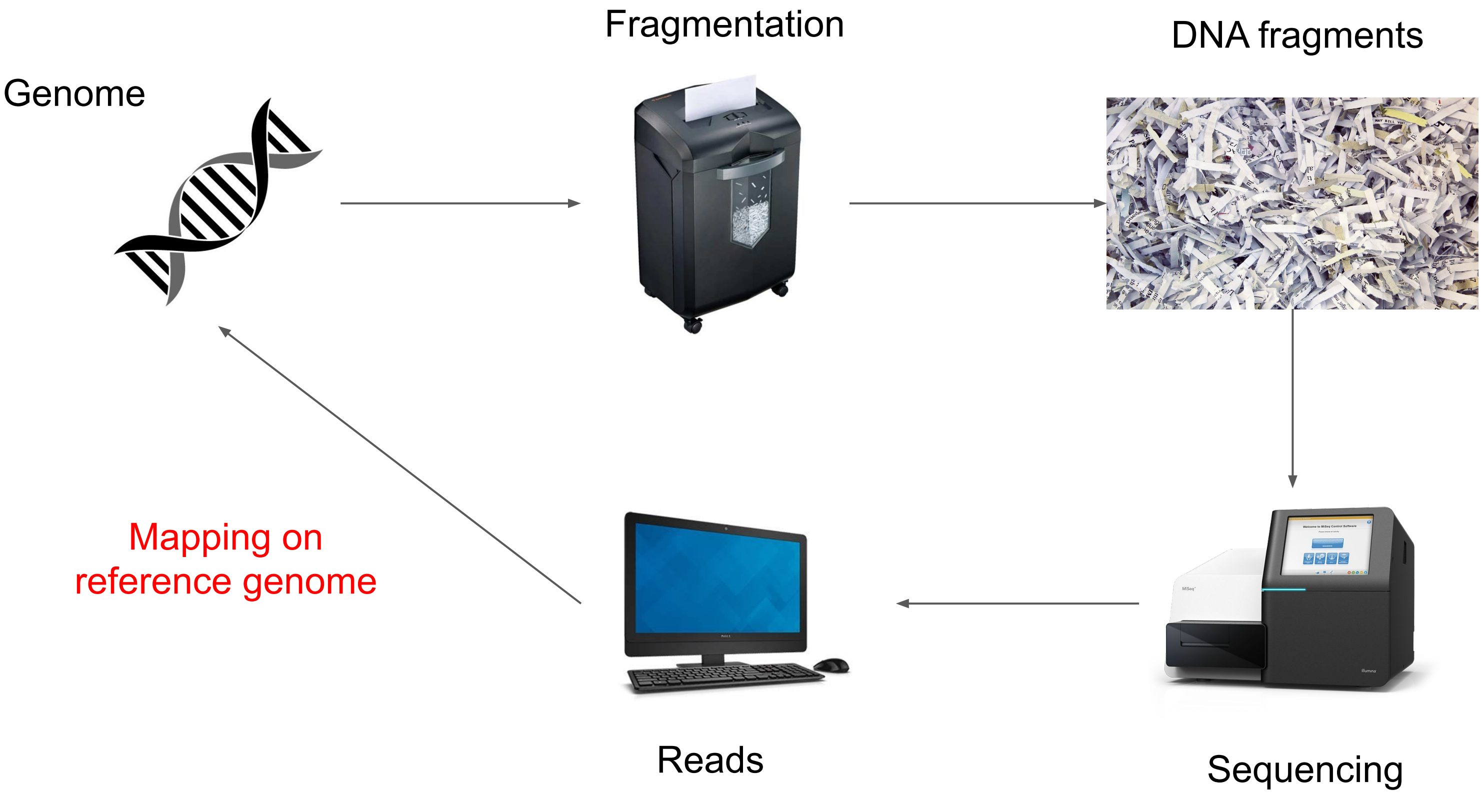
Alignment and mapping

Ing. Stanislav Smatana

Step 2: Alignment



The Main Goal



Genome

Fragmentation

DNA fragments

Sequencing

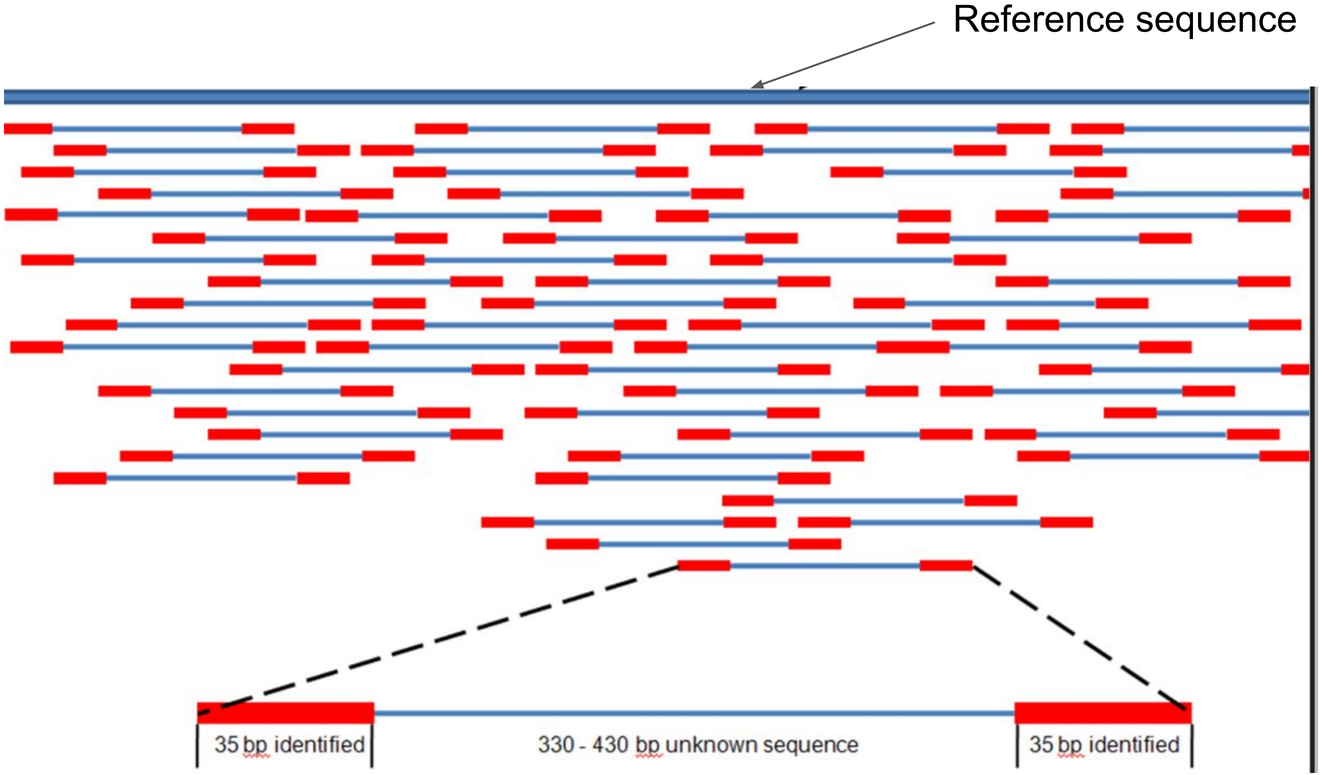
Reads

Mapping on reference genome

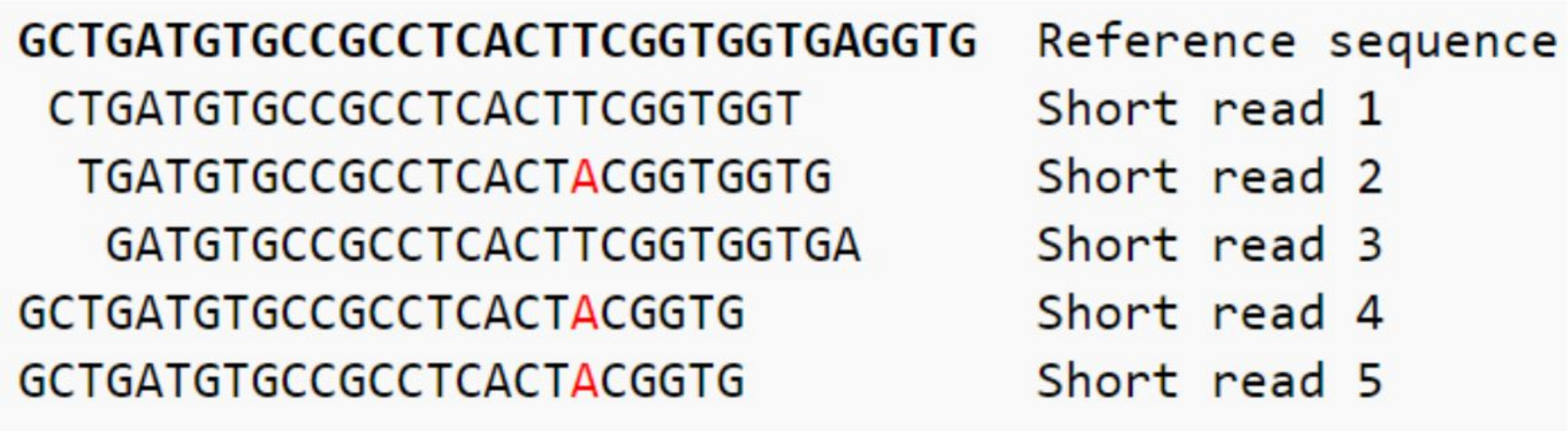
The Main Goal

- Mapping is the essential step in **re-sequencing**
- This means we try to explore something with **known reference sequence**
- We can also construct the reference but let's keep this for another time
- In theory it is quite simple – take a read, compare it with the reference and **find the correct place**

The main goal



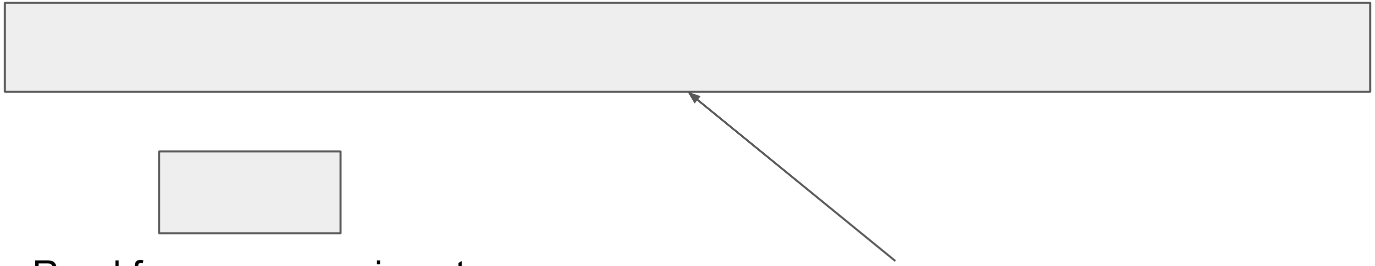
The Main Goal (SNP/SNV)



```
GCTGATGTGCCGCCTCACTTCGGTGGTGAGGTG  Reference sequence
 CTGATGTGCCGCCTCACTTCGGTGGT        Short read 1
  TGATGTGCCGCCTCACTACGGTGGTG       Short read 2
   GATGTGCCGCCTCACTTCGGTGGTGA      Short read 3
GCTGATGTGCCGCCTCACTACGGTG          Short read 4
GCTGATGTGCCGCCTCACTACGGTG          Short read 5
```

It’s a sequence alignment problem

For simplicity, let’s first focus on single (non-paired) reads



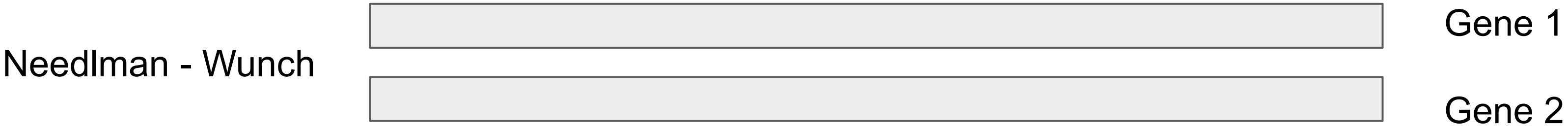
Read from our experiment

Reference genome

We need to **align** reads from sequencing experiment to their corresponding place on reference genome sequence

Sequence Alignment

Global alignment



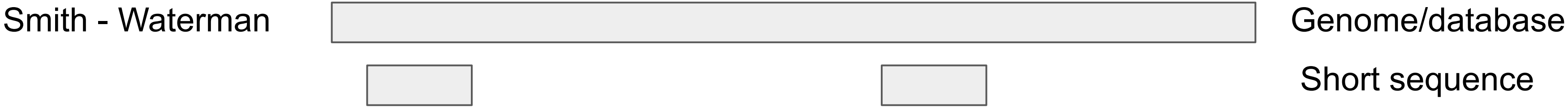
Needlman - Wunch

Gene 1

Gene 2

We want to align two sequences to the same length in order to illuminate evolutionary relationship between them.

Local alignment



Smith - Waterman

Genome/database

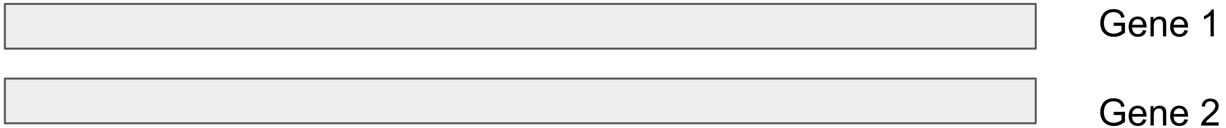
Short sequence

We want to find occurences of shorter sequence in much longer sequence.

Sequence Alignment

Global alignment

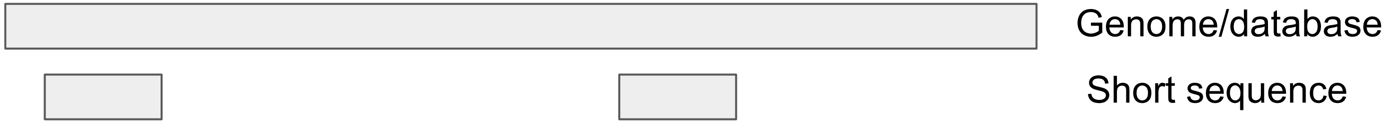
Needlman - Wunch



We want to align two sequences to the same length in order to illuminate evolutionary relationship between them.

Local alignment

Smith - Waterman

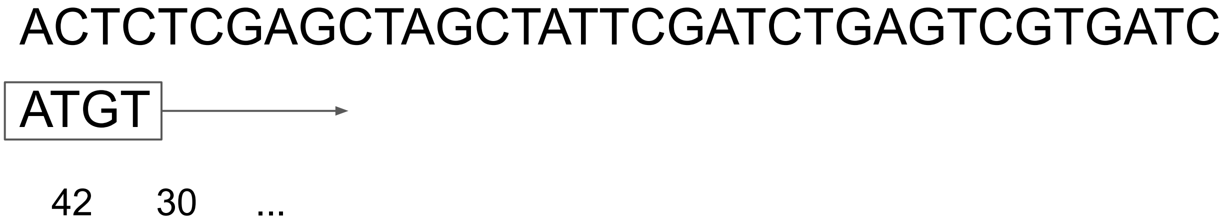


We want to find occurences of shorter sequence in much longer sequence.

We need this !

Naive approach to local alignment

- Compare query to subject string at every position and calculate score
- Correct alignment is at position with the highest score



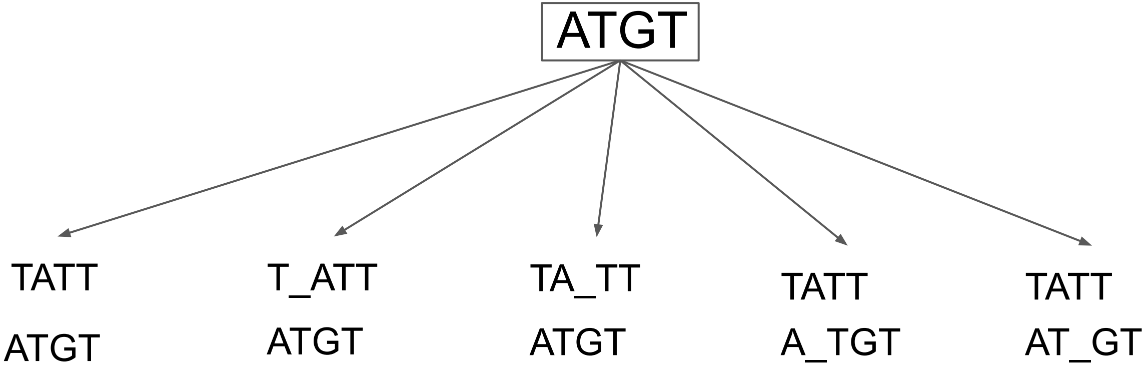
ACTCTCGAGCTAGCTATTCGATCTGAGTCGTGATC

ATGT

42 30 ...

Indels Complicate Things

ACTCTCGAGCTAGCTATTCGATCTGAGTCGTGATC



ATGT

```
TATT     T_ATT    TA_TT    TATT     TATT
ATGT     ATGT     ATGT     A_TGT    AT_GT
```

Much more work !

Dealing with indels - **global** alignment

TGCTGTACTG
TATACCA

→

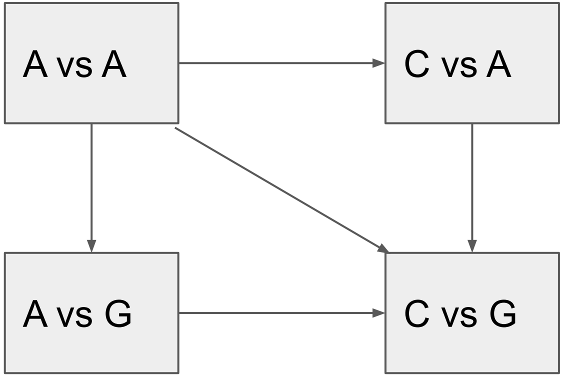
TGCTGTACTG
T_A__TACCA

We want to align two sequences to the same length by inserting gaps in order to illuminate evolutionary relationship between them.

Graph representation of the problem

match = 1
mismatch = 0
gap = -1

AC vs AG



A vs A

C vs A

A vs G

C vs G

AC vs -AG

-AC vs -AG

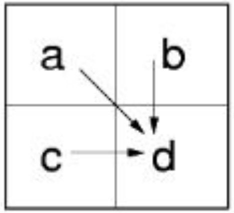
-AC vs AG

-AC vs -AG

Needlman - Wunch algorithm

TGCTGTACTG
T_A__TACCA

		T	G	C	T	G	T	A	C	T	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
T	-1										
A	-2										
T	-3										
A	-4										
C	-5										
C	-6										
A	-7										



d=max(a + match, b - gap, c - gap)

TGCTGTACTG
T_A__TACCA

Needlman - Wunch algorithm

		T	G	C	T	G	T	A	C	T	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
T	-1	1									
A	-2										
T	-3										
A	-4										
C	-5										
C	-6										
A	-7										

TGCTGTACTG

T_A__TACCA

Needlman - Wunch algorithm

		T	G	C	T	G	T	A	C	T	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
T	-1	1	0								
A	-2										
T	-3										
A	-4										
C	-5										
C	-6										
A	-7										

TGCTGTACTG

T_A__TACCA

Needlman - Wunch algorithm

		T	G	C	T	G	T	A	C	T	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
T	-1	1	0	-1							
A	-2										
T	-3										
A	-4										
C	-5										
C	-6										
A	-7										

TGCTGTACTG

T_A__TACCA

Needlman - Wunch algorithm

		T	G	C	T	G	T	A	C	T	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
T	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8
A	-2	0	1	0	-1	-2	-3	-3	-4	-5	-6
T	-3	-1	0	1	1	0	-1	-2	-3	-3	-4
A	-4	-2	-1	0	1	1	0	0	-1	-2	-3
C	-5	-3	-2	0	0	1	1	0	1	0	-1
C	-6	-4	-3	-1	0	0	1	1	1	1	0
A	-7	-5	-4	-2	-1	0	0	2	1	1	1

TGCTGTACTG

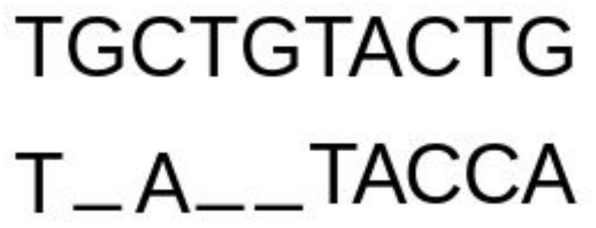
T_A__TACCA

Needlman - Wunch algorithm

		T	G	C	T	G	T	A	C	T	G
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
T	-1	**1**	**0**	-1	-2	-3	-4	-5	-6	-7	-8
A	-2	0	1	**0**	**-1**	**-2**	-3	-3	-4	-5	-6
T	-3	-1	0	1	1	0	**-1**	-2	-3	-3	-4
A	-4	-2	-1	0	1	1	0	**0**	-1	-2	-3
C	-5	-3	-2	0	0	1	1	0	**1**	0	-1
C	-6	-4	-3	-1	0	0	1	1	1	**1**	0
A	-7	-5	-4	-2	-1	0	0	2	1	1	**1**

Needlman - Wunch summary

Given scoring parameters, the algorithm **guarantees** to find all optimal alignments between the two sequences



TGCTGTACTG

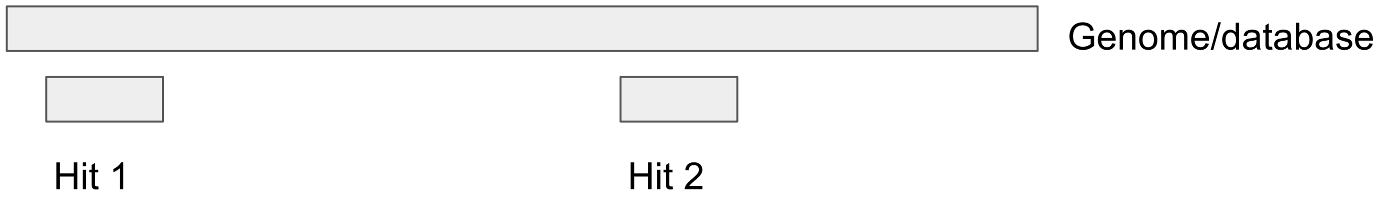
T_A__TACCA

Alignment score = 4x match - 3x gap
Alignment score = 1

Local alignment with indels (Smith - Waterman)

- This can be solved by modification of Needlman - Wunch algorithm
  - First row and first column of the matrix are initialized to zeros
  - Mismatch must have negative score (e.g. -1)
  - If score goes below zero, it is saturated to zero
  - Backtracking from all cells with maximum score
- This modification is called **Smith - Waterman** algorithm
- This algorithm **is guaranteed** to find all occurrences of the shorter sequence in the longer sequence

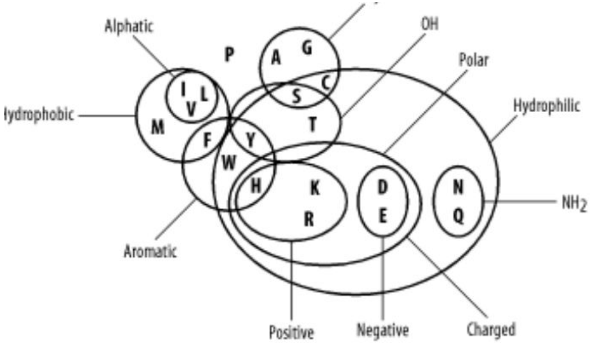
We want to find occurences of shorter sequence in much longer sequence.



Tuning alignment - scoring matrices

- Proteins - matrices
  - Blosum (empiric)
  - PAM (based more on theory)
- Nucleotides
  - Typically only match and mismatch score
- Gaps
  - Gap opening penalty
  - Gap extension penalty

You can create your own scoring matrix based on your domain knowledge !



Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3																
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4
	Ala	**Arg**	**Asn**	**Asp**	**Cys**	**Gln**	**Glu**	**Gly**	**His**	**Ile**	**Leu**	**Lys**	**Met**	**Phe**	**Pro**	**Ser**	**Thr**	**Trp**	**Tyr**	**Val**

Is raw score enough ?

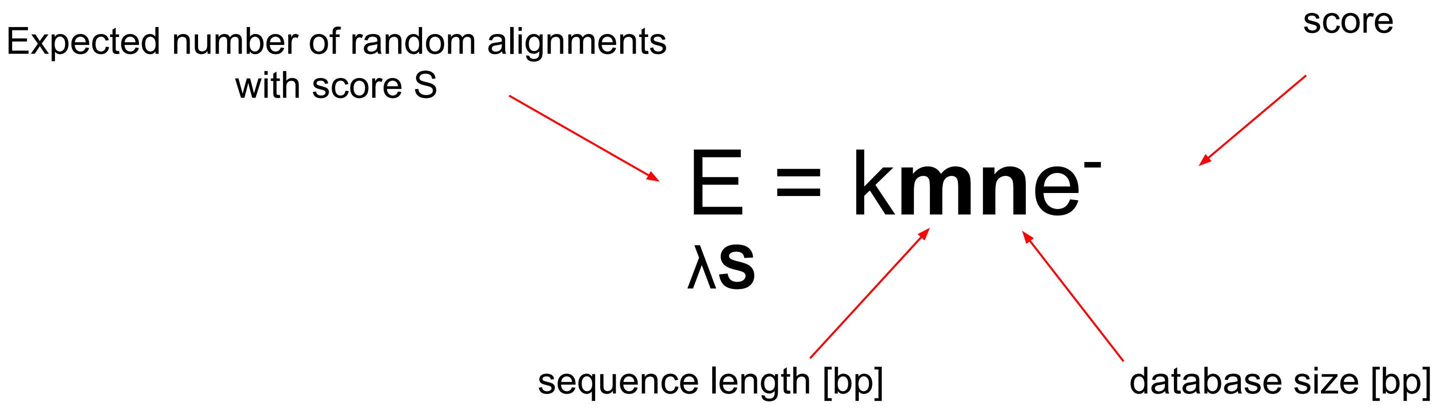
We have aligned sequence X to the database Z and the alignment score is 42. Yay !

Is raw score enough ?

We have aligned sequence X to the database Z and the alignment score is 42. Yay !

Happy ?

Karlin-Altschul alignment statistics (E-value)



Expected number of random alignments with score S

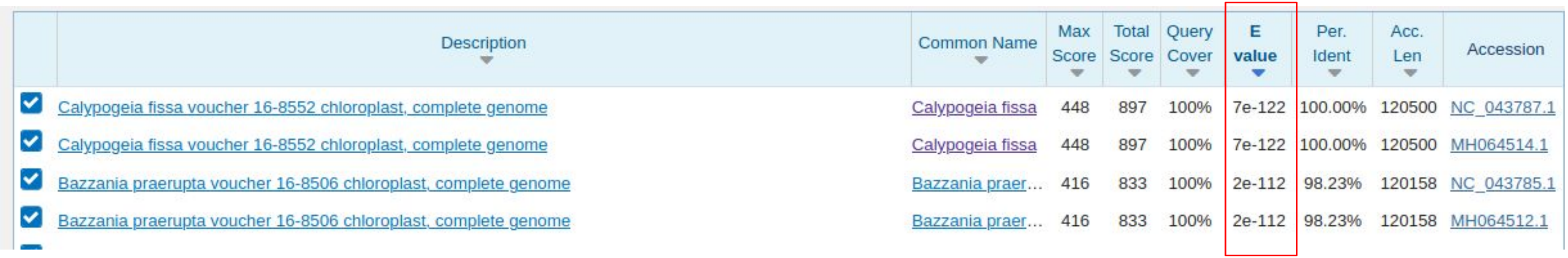
score

$$E = kmne^{-}$$

$$\lambda S$$

sequence length [bp]

database size [bp]



	Description	Common Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Calypogeia fissa voucher 16-8552 chloroplast, complete genome	Calypogeia fissa	448	897	100%	7e-122	100.00%	120500	NC_043787.1
☑	Calypogeia fissa voucher 16-8552 chloroplast, complete genome	Calypogeia fissa	448	897	100%	7e-122	100.00%	120500	MH064514.1
☑	Bazzania praerupta voucher 16-8506 chloroplast, complete genome	Bazzania praer...	416	833	100%	2e-112	98.23%	120158	NC_043785.1
☑	Bazzania praerupta voucher 16-8506 chloroplast, complete genome	Bazzania praer...	416	833	100%	2e-112	98.23%	120158	MH064512.1

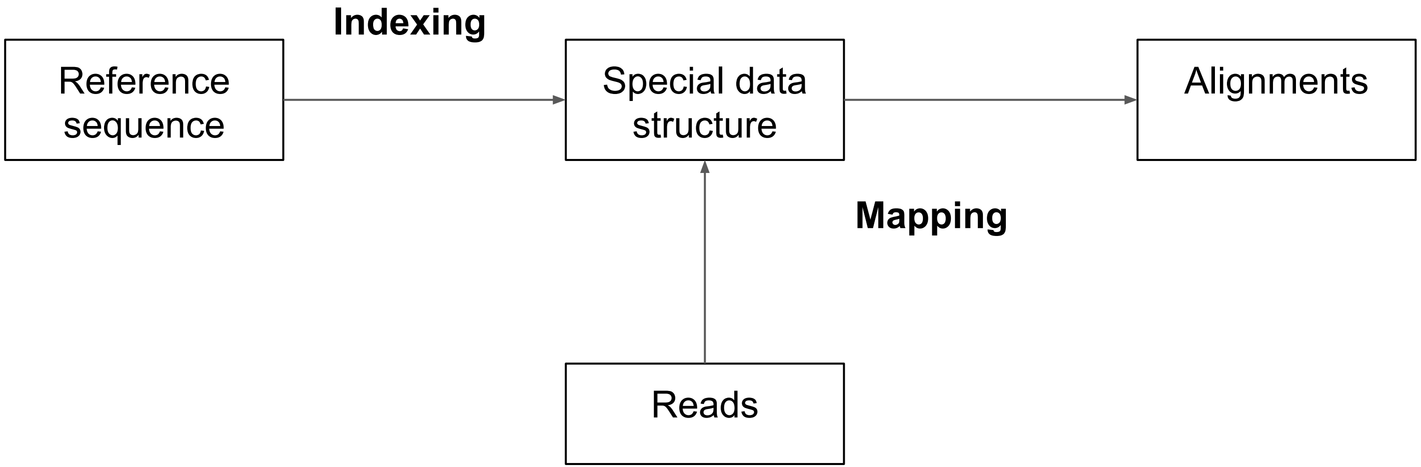
The problem with Smith - Waterman algorithm

Number of nucleotides	Time needed for computation
100	0.2 ms
1,000	0.02 s
10,000	2 s
100,000	3 m
1,000,000	5 h
10,000,000	23 days
100,000,000	6.5 years
1,000,000,000	**650 years**

Calculated by Ing. Tomáš Martínek, PhD. from BUT FIT. Single Xeon 3Ghz CPU.

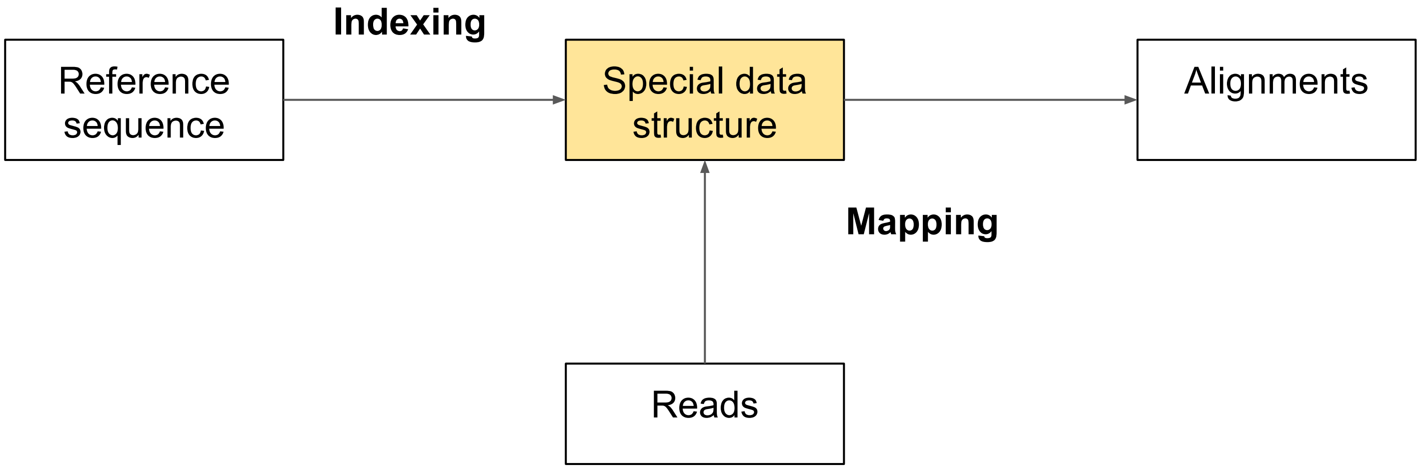
Indexing - Making local alignment faster

Idea: Genome is first transformed from plain text into some different form that is more suitable for fast alignment.

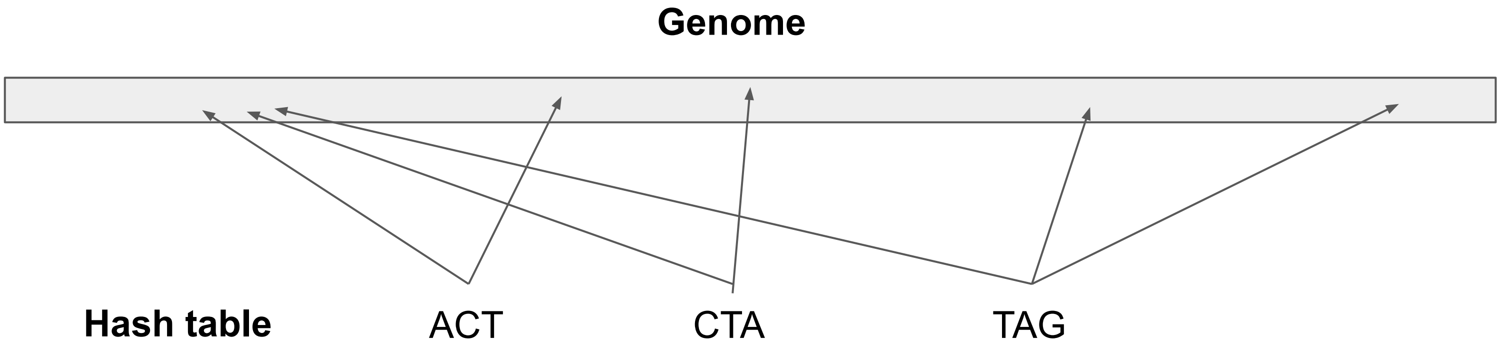


Indexing - Making local alignment faster

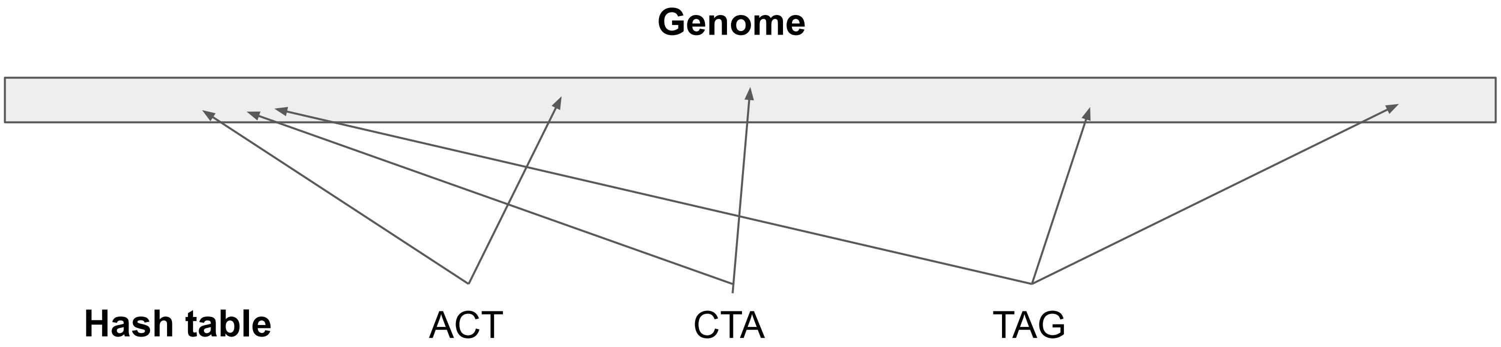
Idea: Genome is first transformed from plain text into some different form that is more suitable for fast alignment.



Indexing - hash table



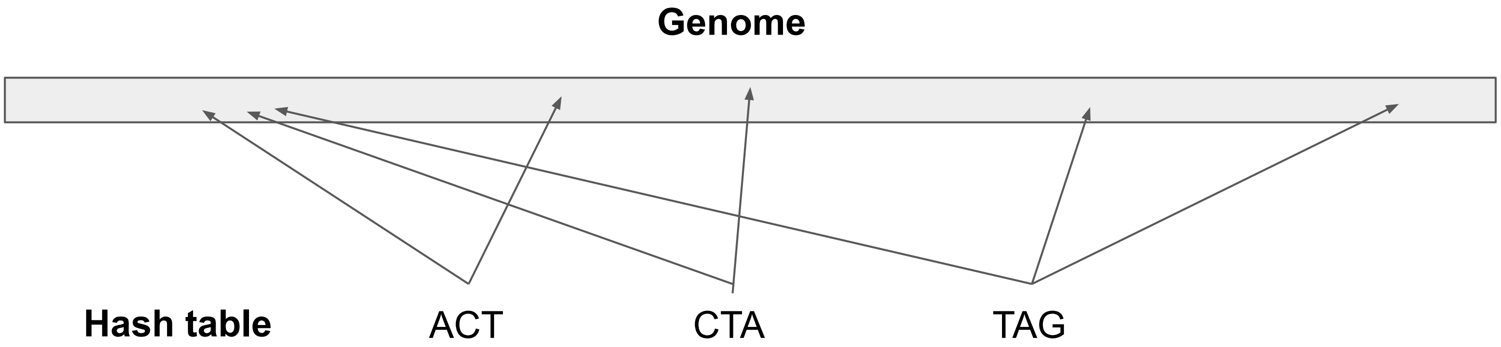
Indexing - hash table



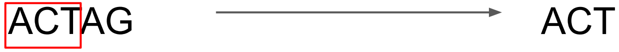
Input sequence

ACTAG

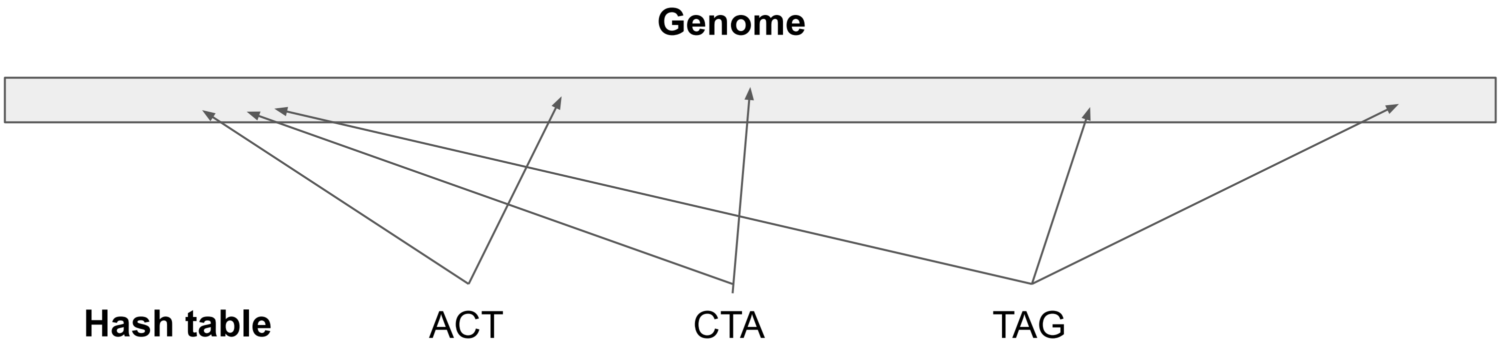
Indexing - hash table



Input sequence



Indexing - hash table



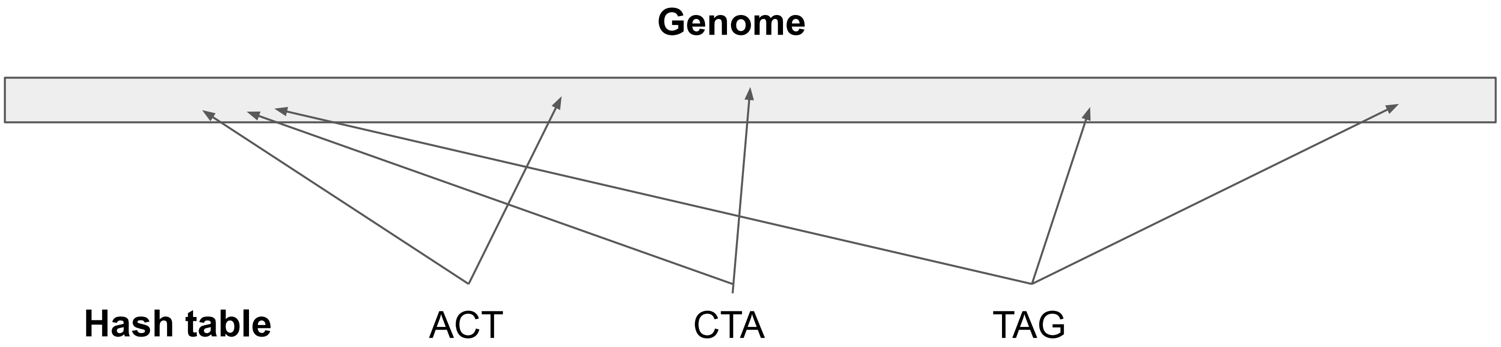
Genome

Hash table ACT CTA TAG

Input sequence

ACTAG → ACT CTA

Indexing - hash table



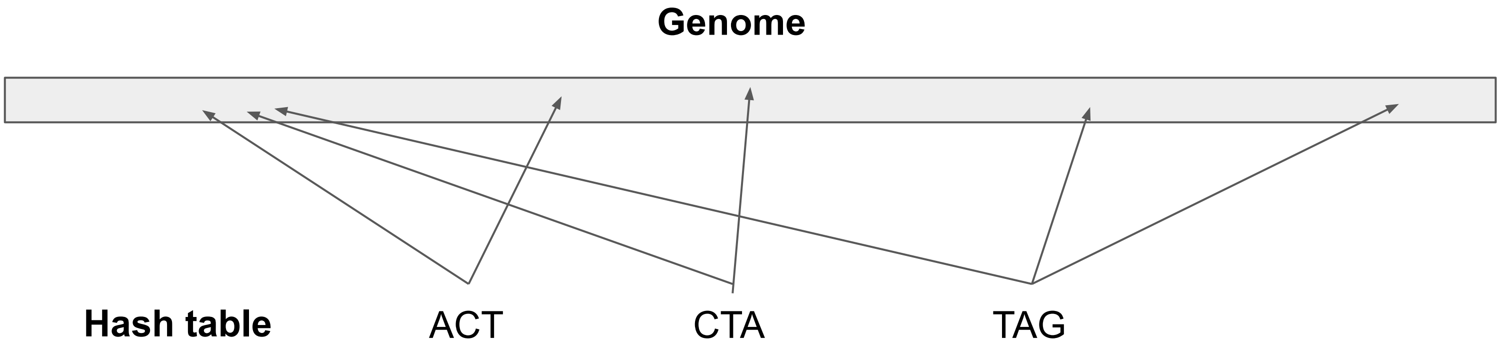
Genome

Hash table ACT CTA TAG

Input sequence

ACTAG → ACT CTA TAG

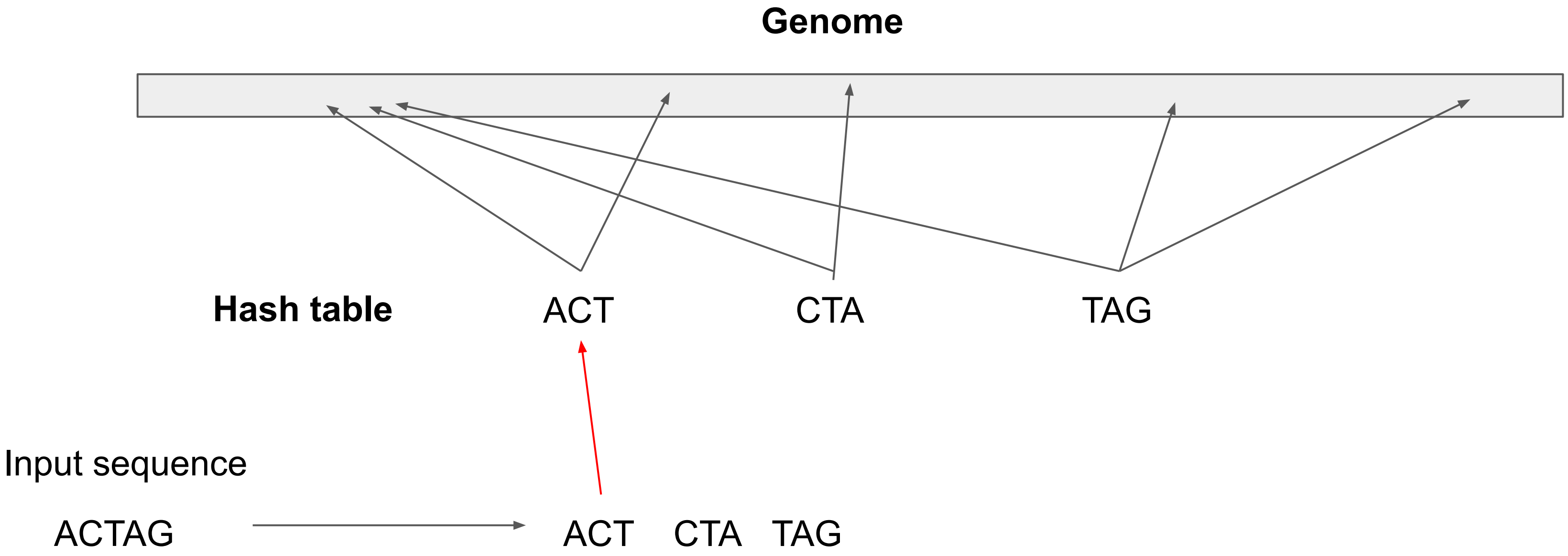
Indexing - hash table



Input sequence

ACTAG → ACT CTA TAG

Indexing - hash table



Genome

Hash table

ACT

CTA

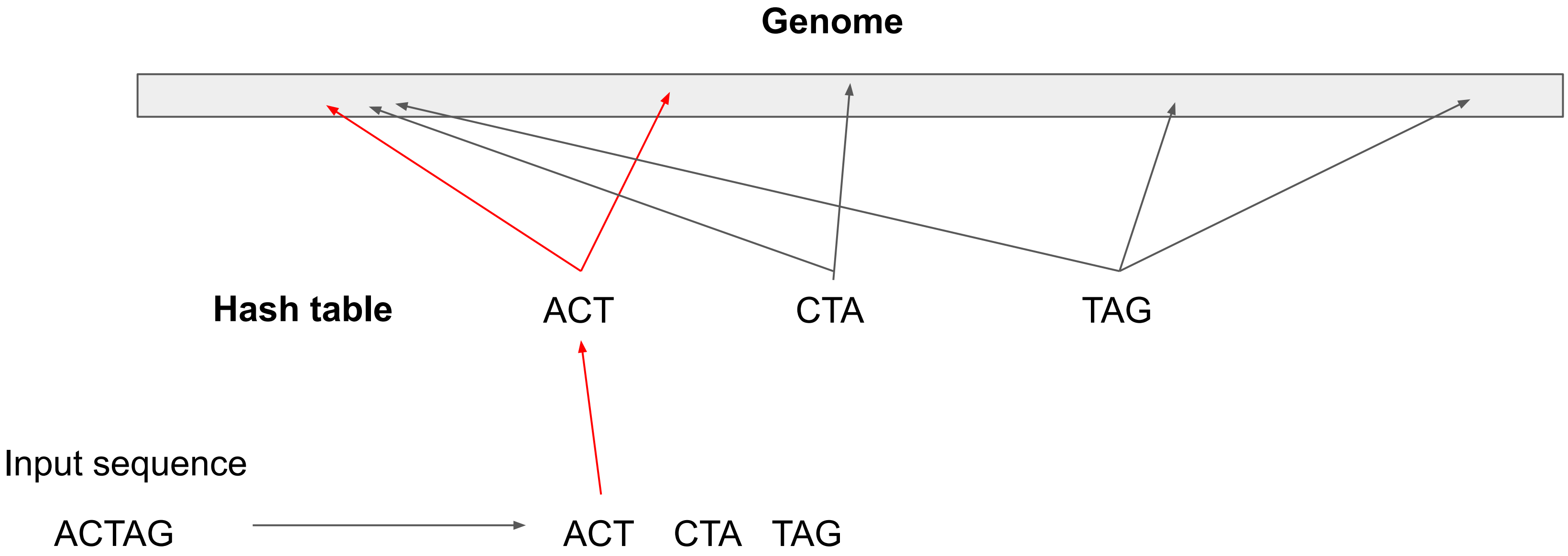
TAG

Input sequence

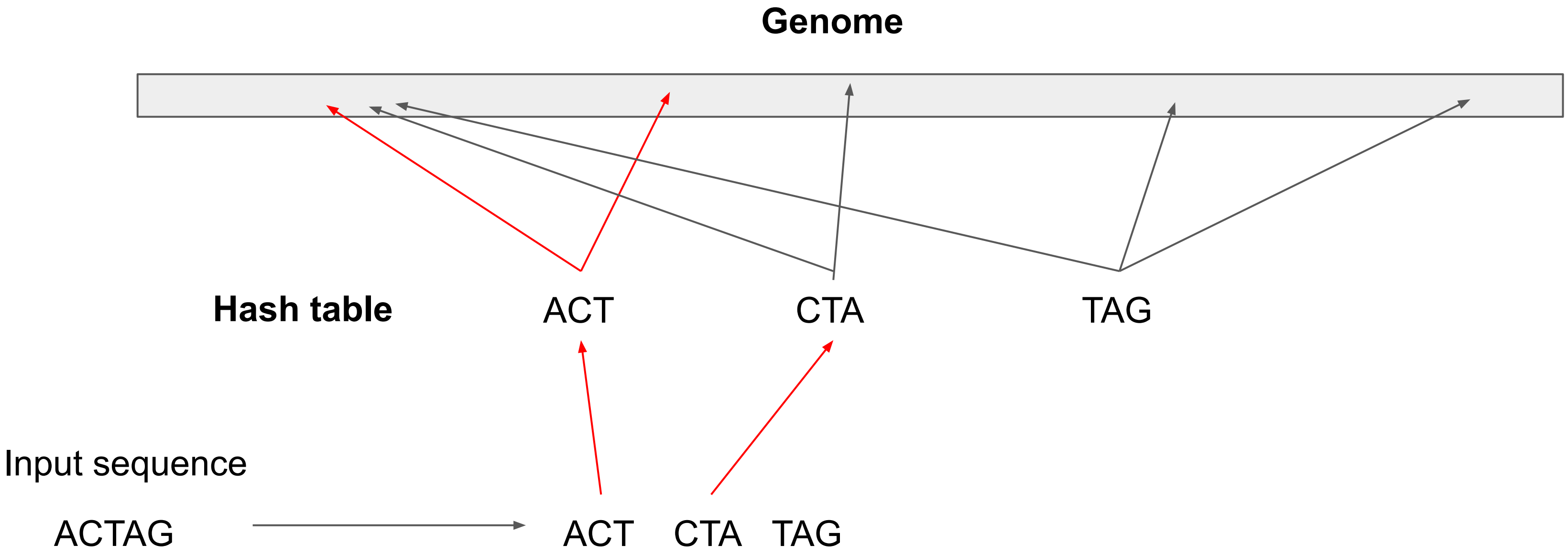
ACTAG

ACT CTA TAG

Indexing - hash table



Indexing - hash table



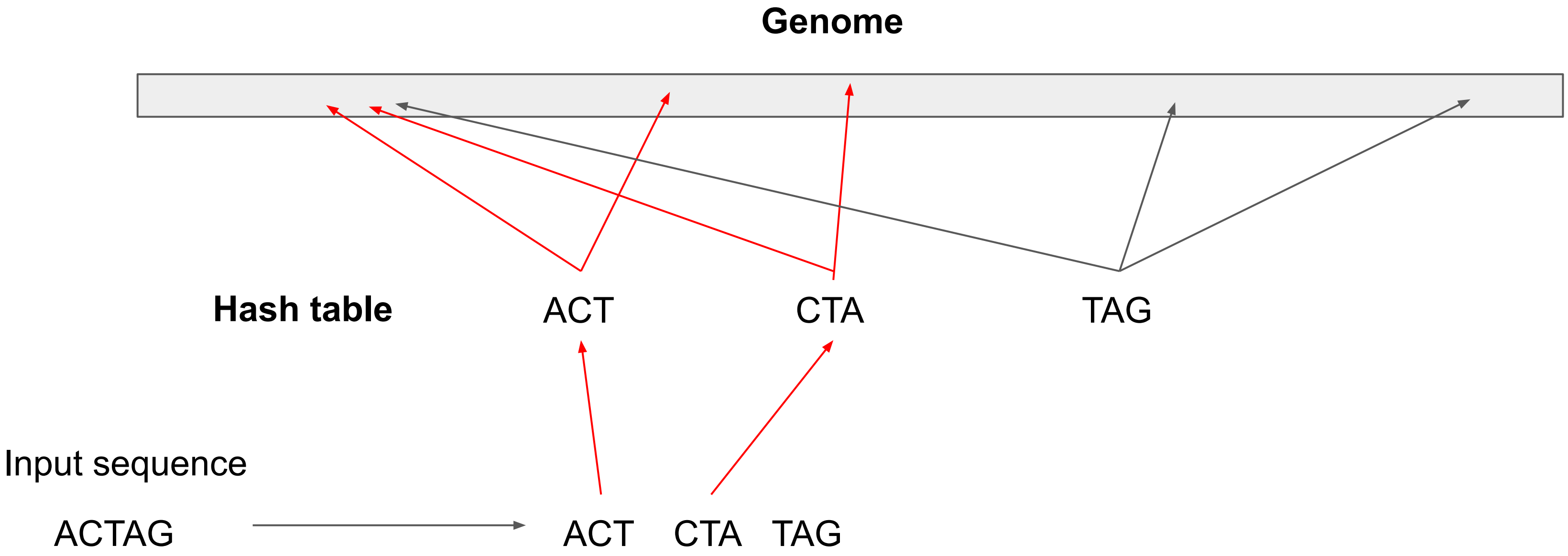
Genome

Hash table ACT CTA TAG

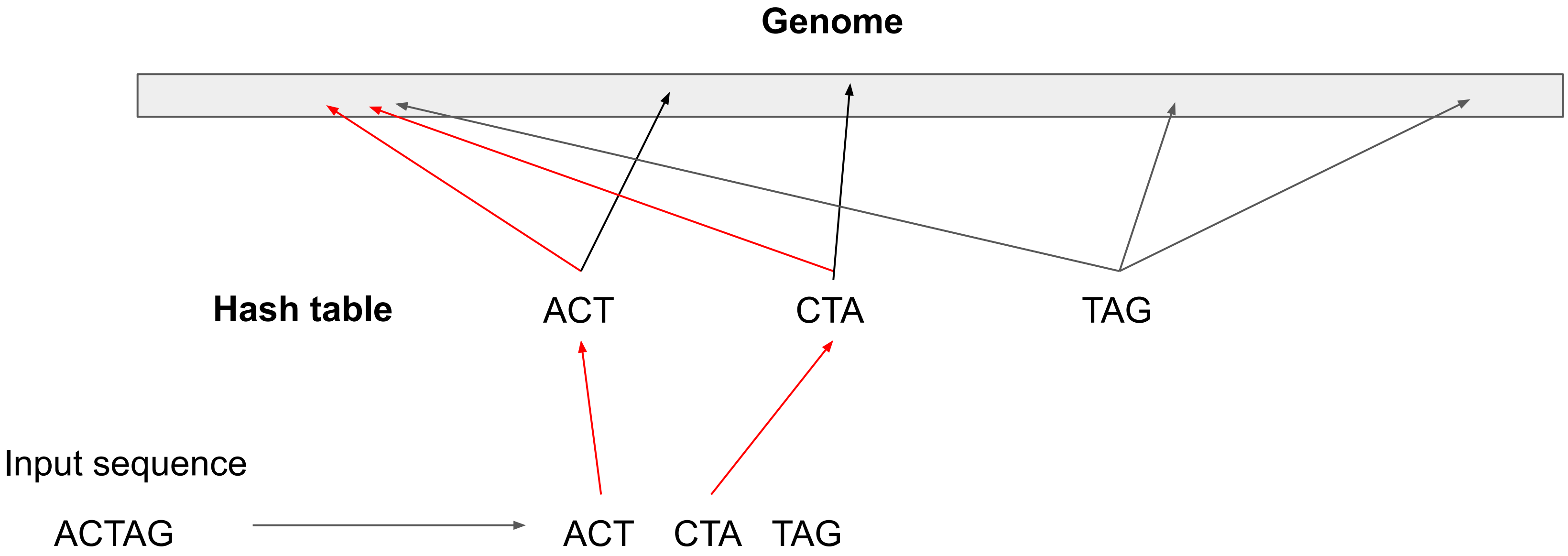
Input sequence

ACTAG → ACT CTA TAG

Indexing - hash table



Indexing - hash table



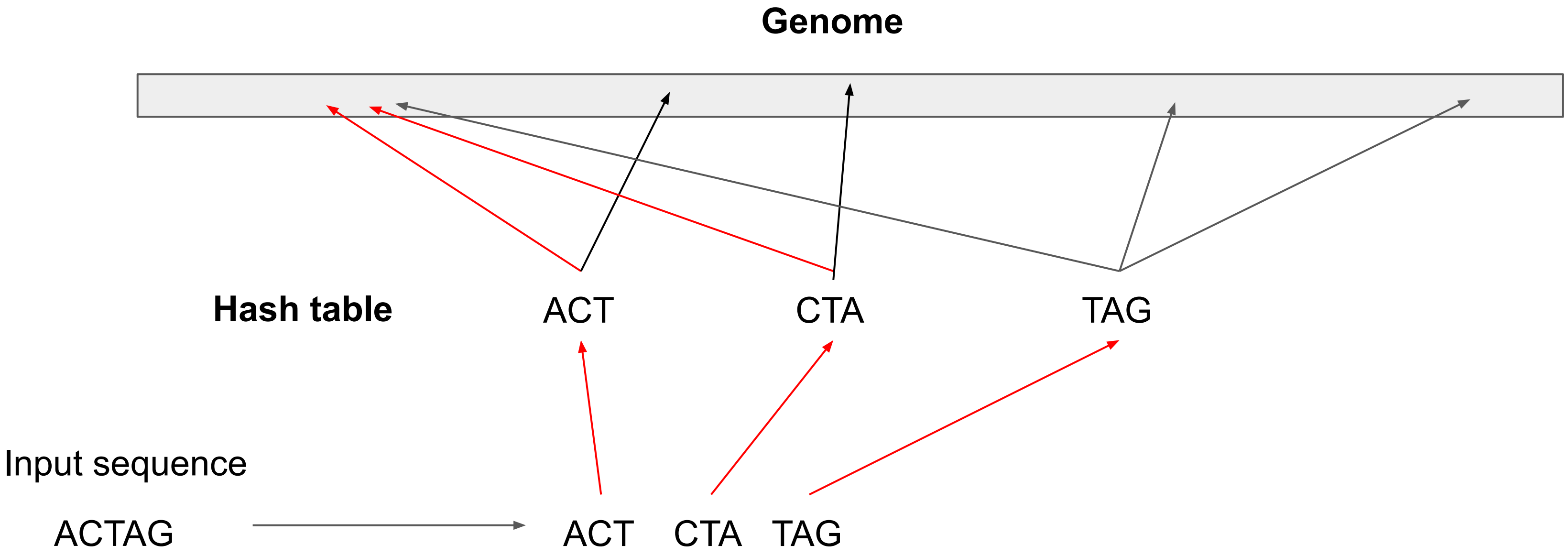
Genome

Hash table ACT CTA TAG

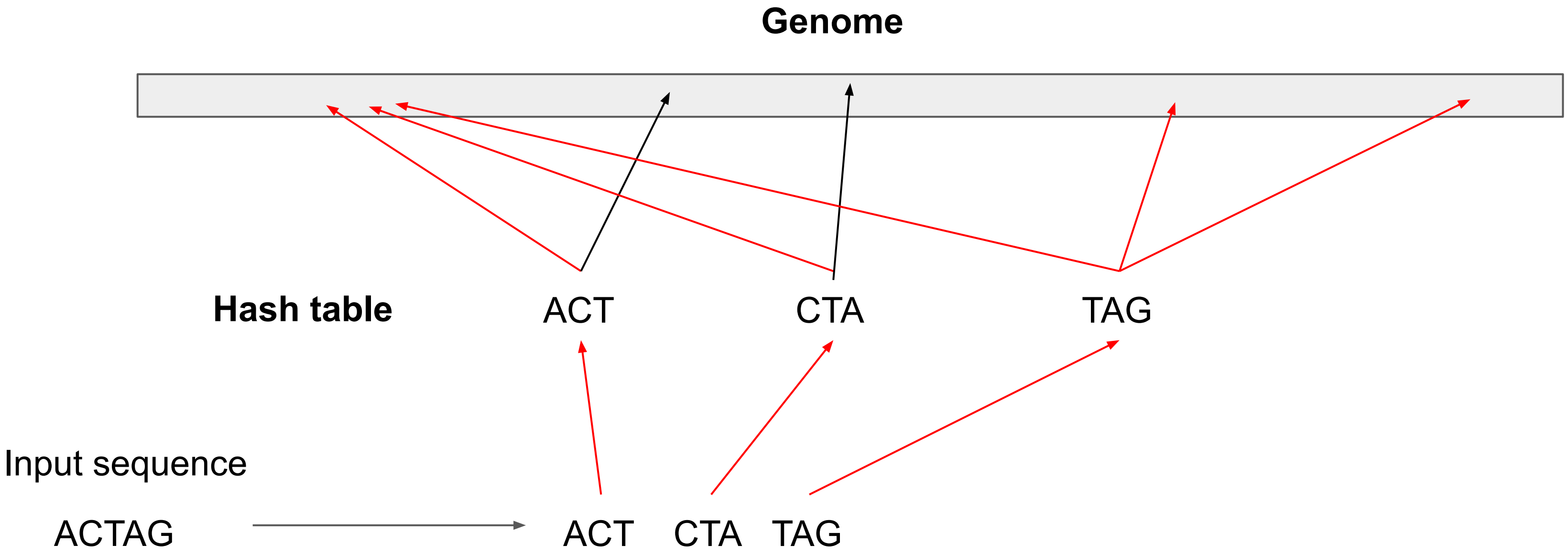
Input sequence

ACTAG → ACT CTA TAG

Indexing - hash table



Indexing - hash table



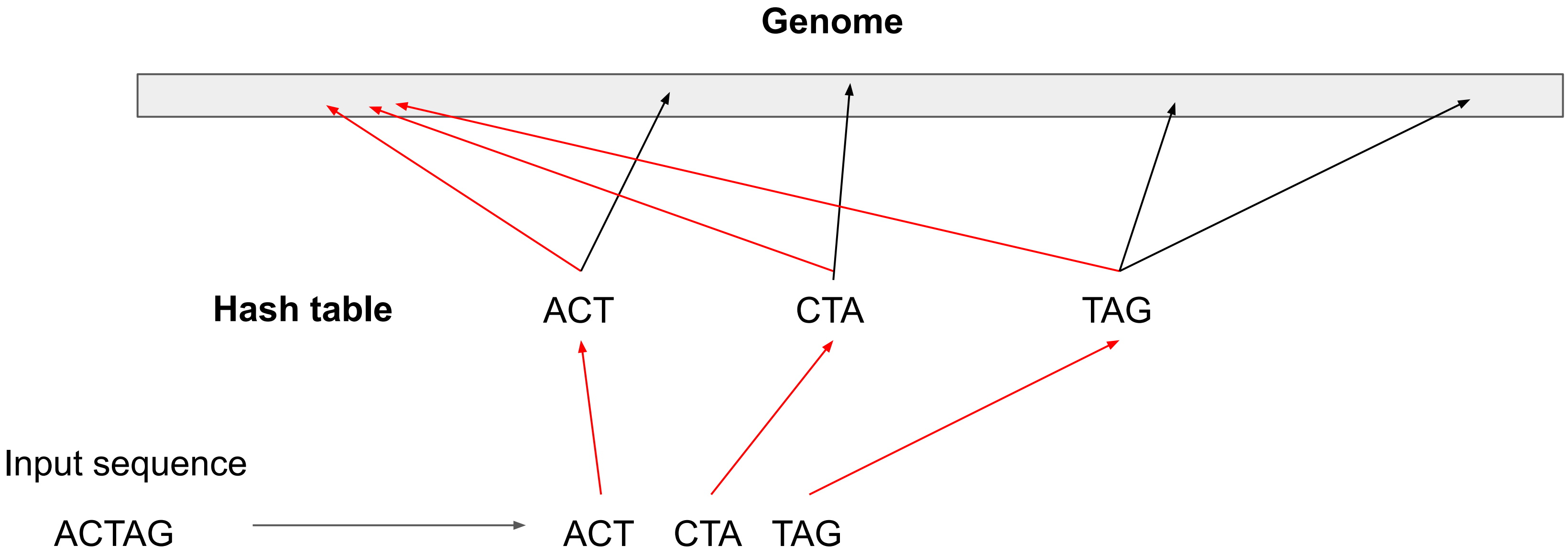
Genome

Hash table ACT CTA TAG

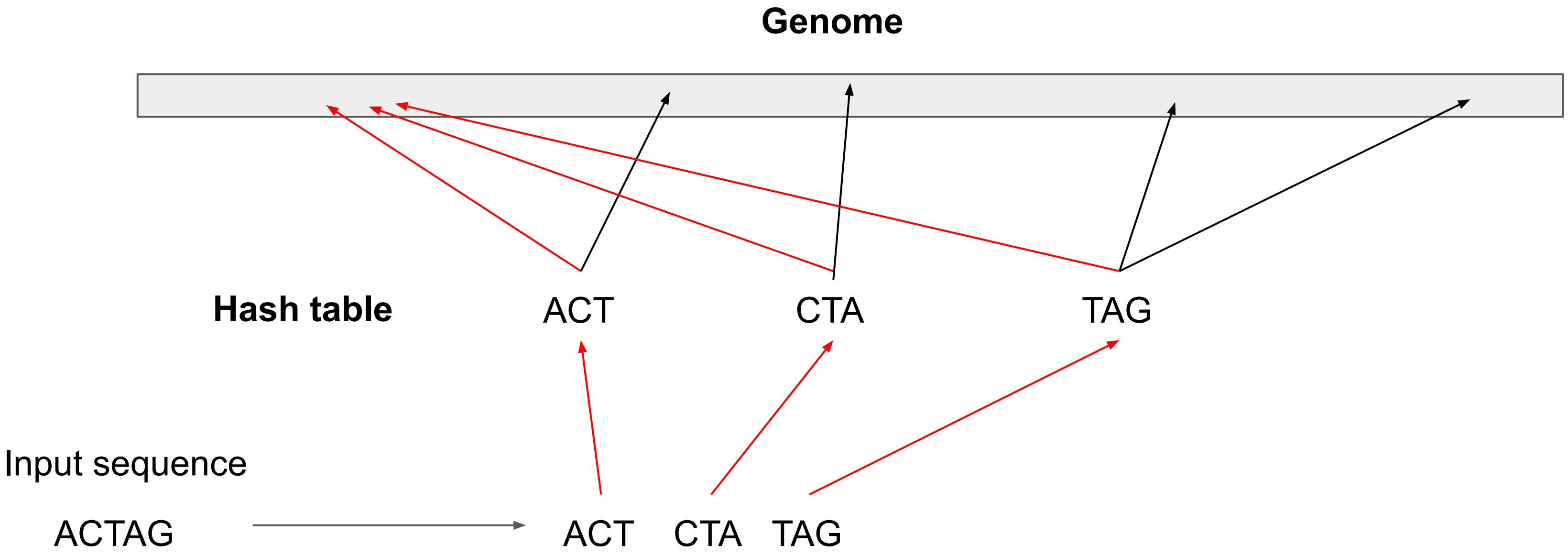
Input sequence

ACTAG → ACT CTA TAG

Indexing - hash table



Indexing - hash table



Genome

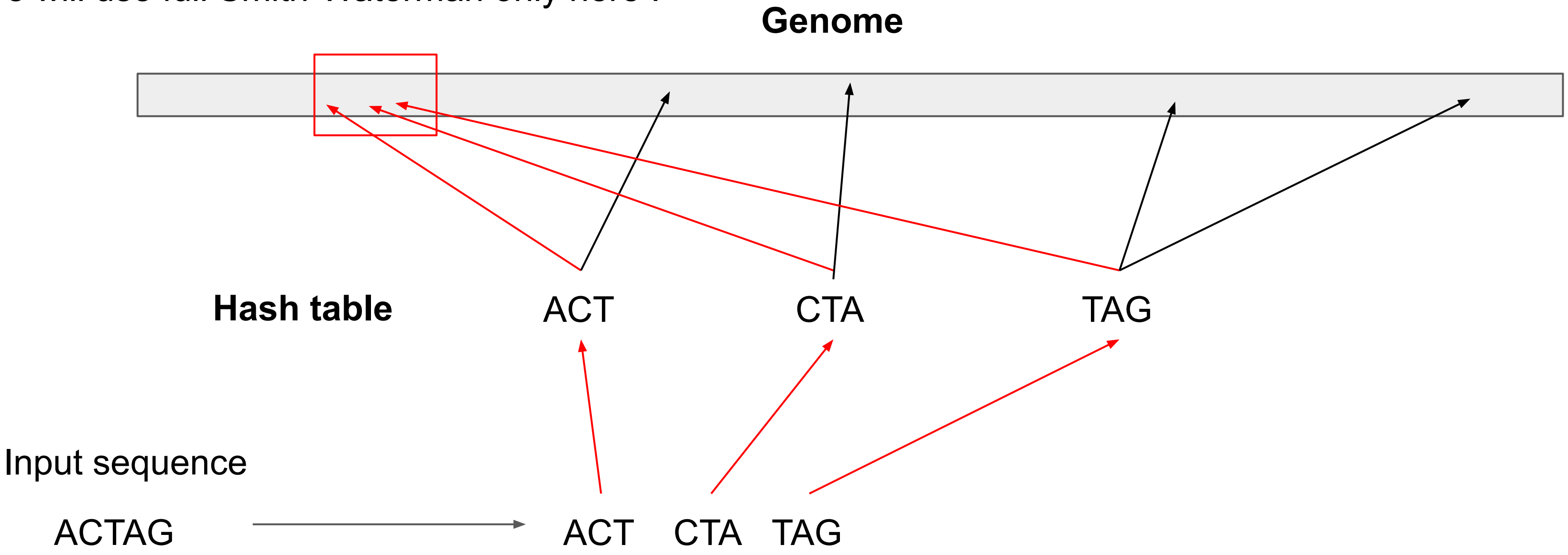
Hash table ACT CTA TAG

Input sequence

ACTAG → ACT CTA TAG

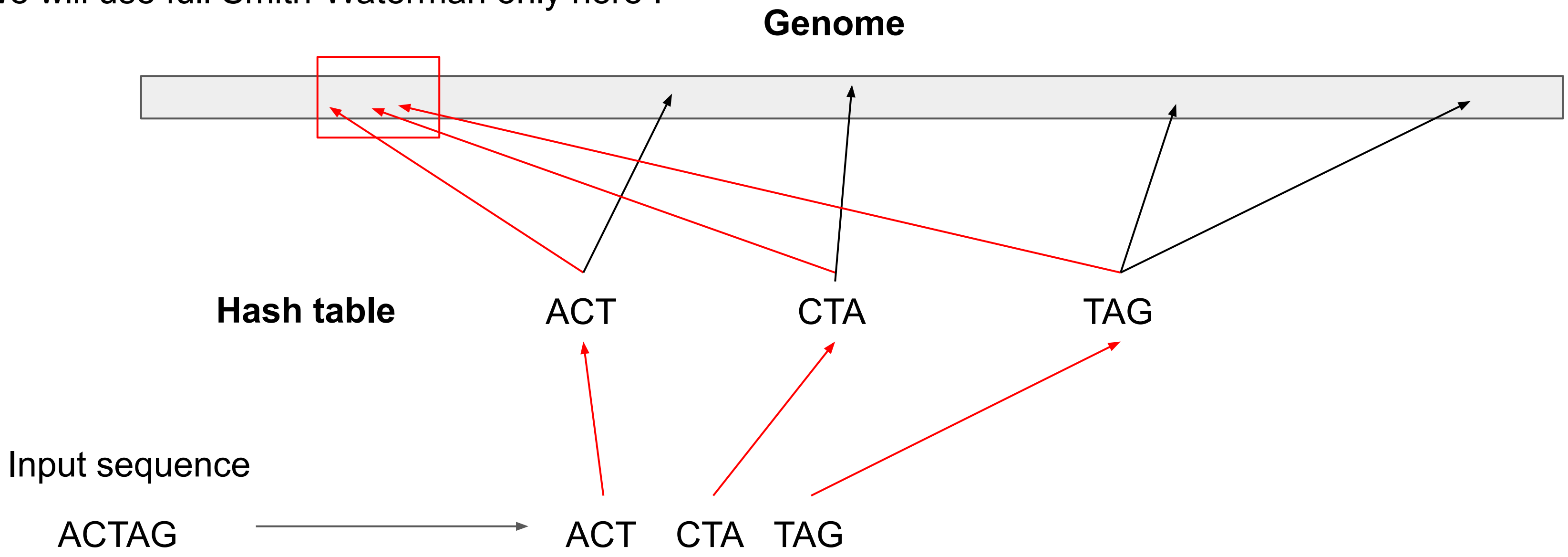
Indexing - hash table

Our sequence will be somewhere in this region,
we will use full Smith-Waterman only here !



Indexing - hash table

Our sequence will be somewhere in this region,
we will use full Smith-Waterman only here !



Genome

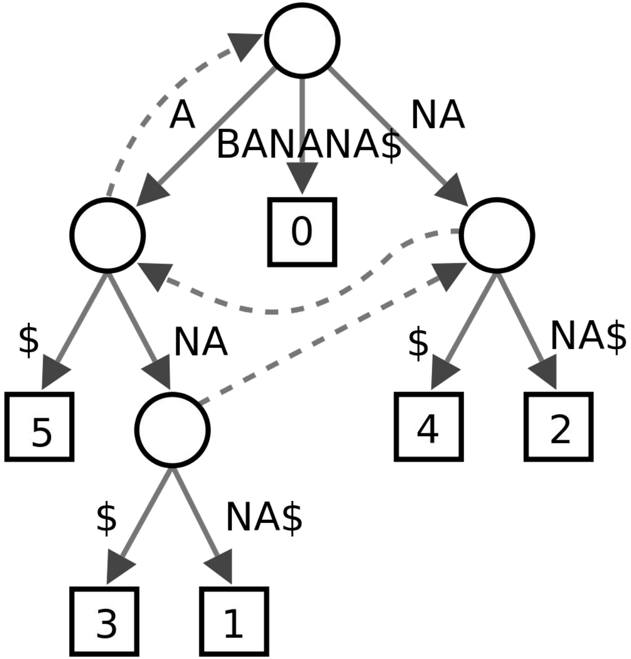
Hash table ACT CTA TAG

Input sequence

ACTAG → ACT CTA TAG

Fast, but what price do we pay for this ?

Indexing - suffix tree



Substring of a given string is a **prefix of one of its suffixes**.

String: BANANA

Substring: NAN

Suffixes:

BANANA$
ANANA$
NANA$
ANA$
NA$
A$

Similar strategies: suffix array, Burrows-Wheeler transform

Indexing - how does it look in practice ?

Example using bowtie2 genome mapper:

1. Build index for reference genome

```
bowtie2-build my_reference.fasta my_index_name
```

2. Align reads using the index

```
bowtie2 -U my_reference.fasta -x my_index_name
```

Note: Some aligners do the index creation implicitly.

General Aligners vs Genome Aligners

General aligners

BLAST, HMMER, MMSeqs2, ...

- Typically used for search in large databases (e.g. NCBI nt)
- Do not make use of paired reads
- Do not make use of sequence quality information
- Intended for general search of sequences, not only short reads

NGS Aligners (mappers)

bowtie2, STAR, bwa, ...

- Used to align large number of short reads to genome
- Can take advantage of sequence quality information
- Can make use of paired reads
- Optimized for the task of short read mapping
- Produce output in standardized format (SAM/BAM)

General Aligners vs Genome Aligners

General aligners

BLAST, HMMER, MMSeqs2, ...

- Typically used for search in large databases (e.g. NCBI nt)
- Do not make use of paired reads
- Do not make use of sequence quality information
- Intended for general search of sequences, not only short reads

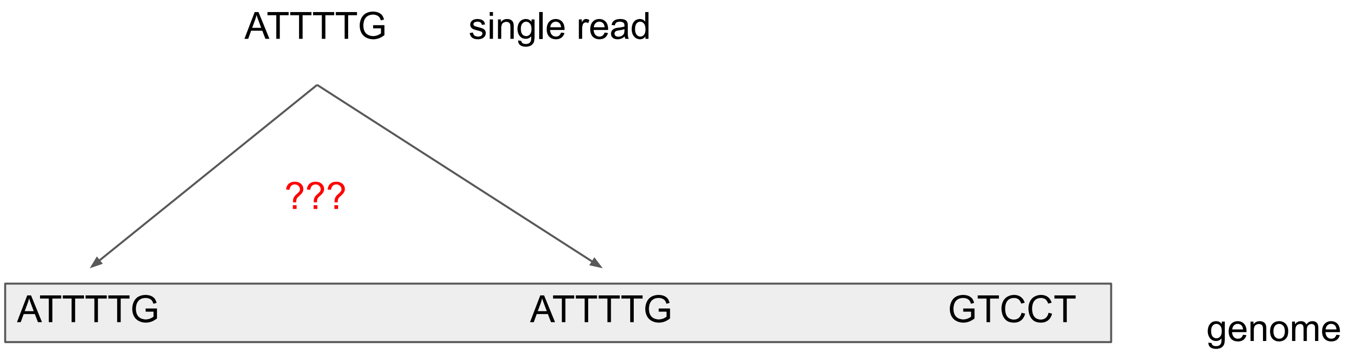
NGS Aligners (mappers)

bowtie2, STAR, bwa, ...

- Used to align large number of short reads to genome
- Can take advantage of sequence quality information
- **Can make use of paired reads**
- Optimized for the task of short read mapping
- Produce output in standardized format (SAM/BAM)

NGS Aligners - how can paired reads help

One of the biggest challenges for alignments to genome are repetitions.



ATTTTG single read

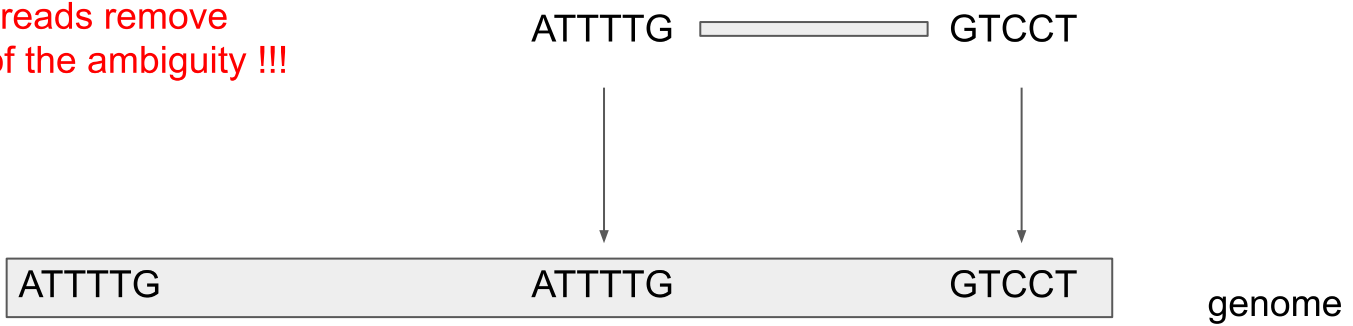
???

ATTTTG ATTTTG GTCCT genome

NGS Aligners - how can paired reads help

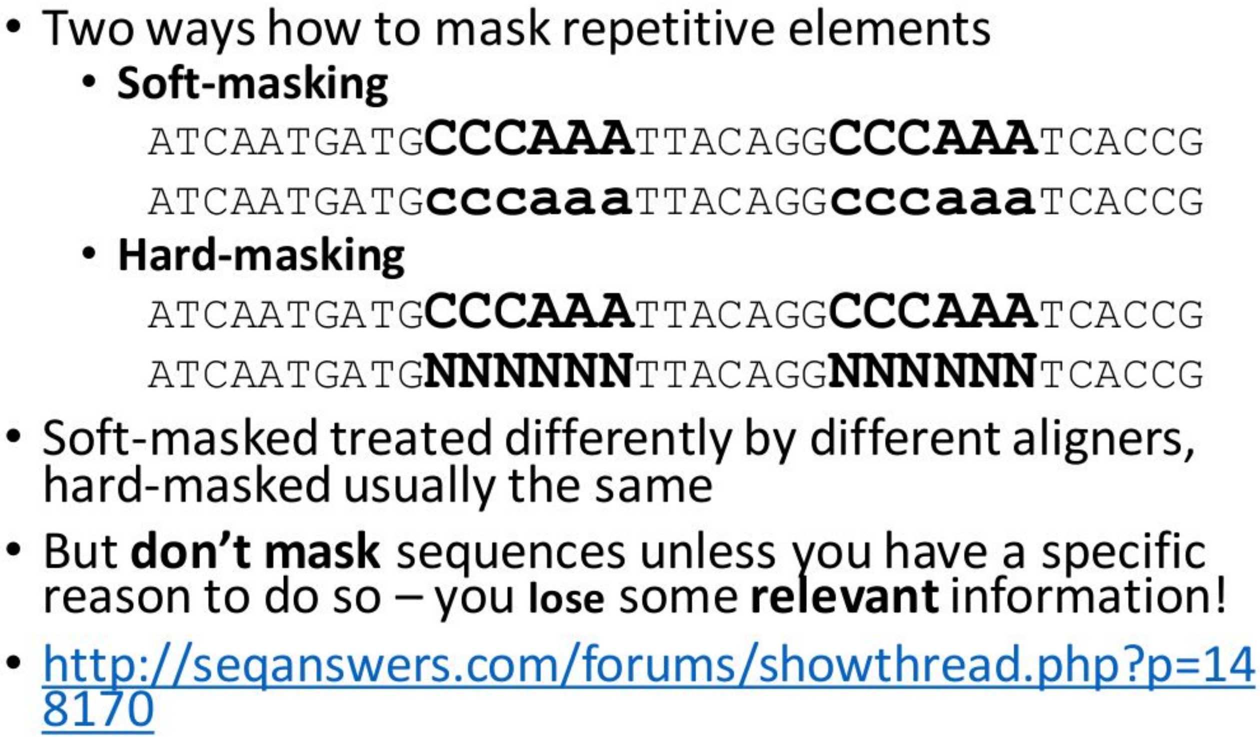
One of the biggest challenges for alignments to genome are repetitions.

Paired reads remove some of the ambiguity !!!



NGS aligners - repeat masking

Another way to deal with repeats is to mask them.



- Two ways how to mask repetitive elements
  - **Soft-masking**

    ATCAATGATG**CCCAAA**TTACAGG**CCCAAA**TCACCG

    ATCAATGATG**cccaaa**TTACAGG**cccaaa**TCACCG
  - **Hard-masking**

    ATCAATGATG**CCCAAA**TTACAGG**CCCAAA**TCACCG

    ATCAATGATG**NNNNNN**TTACAGG**NNNNNN**TCACCG
- Soft-masked treated differently by different aligners, hard-masked usually the same
- But **don't mask** sequences unless you have a specific reason to do so – you **lose** some **relevant** information!
- http://seqanswers.com/forums/showthread.php?p=148170

General Aligners vs Genome Aligners

General aligners

BLAST, HMMER, MMSeqs2, ...

- Typically used for search in large databases (e.g. NCBI nt)
- Do not make use of paired reads
- Do not make use of sequence quality information
- Intended for general search of sequences, not only short reads

NGS Aligners (mappers)

bowtie2, STAR, bwa, ...

- Used to align large number of short reads to genome
- Can take advantage of sequence quality information
- Can make use of paired reads
- Optimized for the task of short read mapping
- Produce output in standardized format (SAM/BAM)

General Aligners vs Genome Aligners

General aligners

BLAST, HMMER, MMSeqs2, ...

- Typically used for search in large databases (e.g. NCBI nt)
- Do not make use of paired reads
- Do not make use of sequence quality information
- Intended for general search of sequences, not only short reads

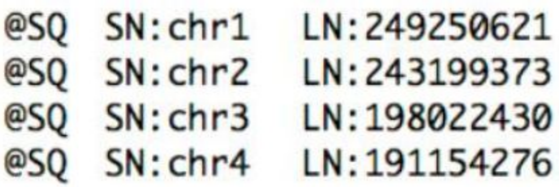
NGS Aligners (mappers)

bowtie2, STAR, bwa, ...

- Used to align large number of short reads to genome
- Can take advantage of sequence quality information
- Can make use of paired reads
- Optimized for the task of short read mapping
- **Produce output in standardized format (SAM/BAM)**

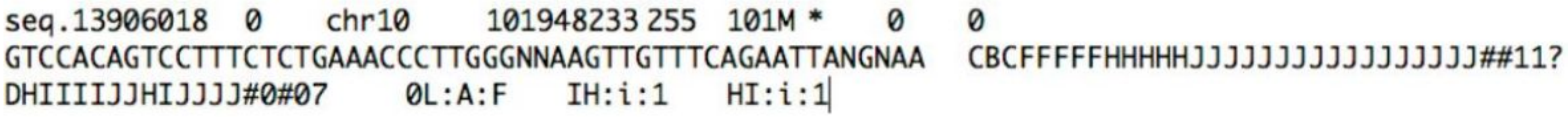
NGS Aligners - standardized output (SAM/BAM)

- Header



```
@SQ  SN:chr1   LN:249250621
@SQ  SN:chr2   LN:243199373
@SQ  SN:chr3   LN:198022430
@SQ  SN:chr4   LN:191154276
```

- Body



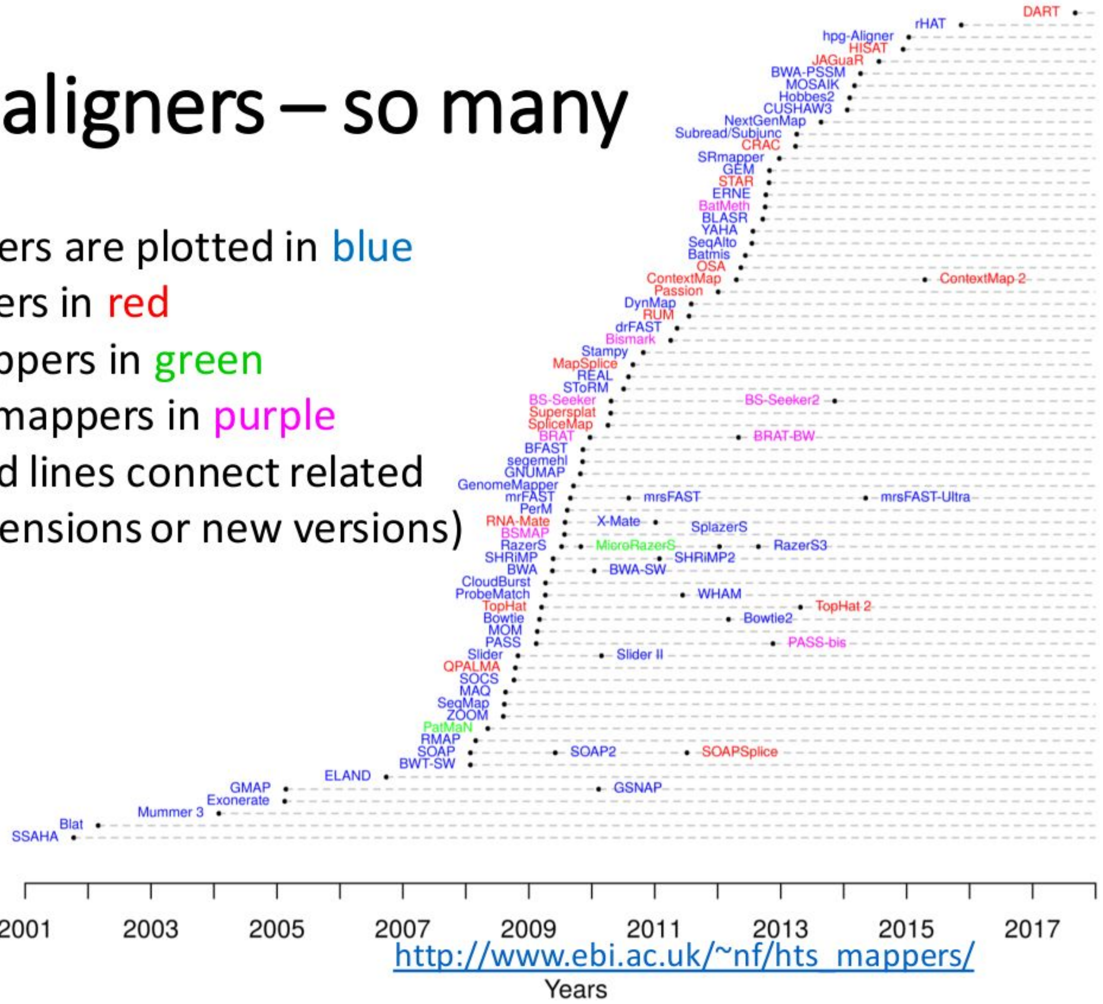
```
seq.13906018   0    chr10     101948233 255   101M *    0    0
GTCCACAGTCCTTTCTCTGAAACCCTTGGGNNAAGTTGTTTCAGAATTANGNAA   CBCFFFFFHHHHHJJJJJJJJJJJJJJJ##11?
DHIIIIJJHIJJJJ#0#07      0L:A:F    IH:i:1    HI:i:1
```

- One line per mapped read
- BAM = binary version of SAM (compression)

https://samtools.github.io/hts-specs/SAMv1.pdf

NGS aligners – so many

- DNA mappers are plotted in blue
- RNA mappers in red
- miRNA mappers in green
- Bisulphite mappers in purple
- Grey dotted lines connect related mappers (extensions or new versions)



http://www.ebi.ac.uk/~nf/hts_mappers/

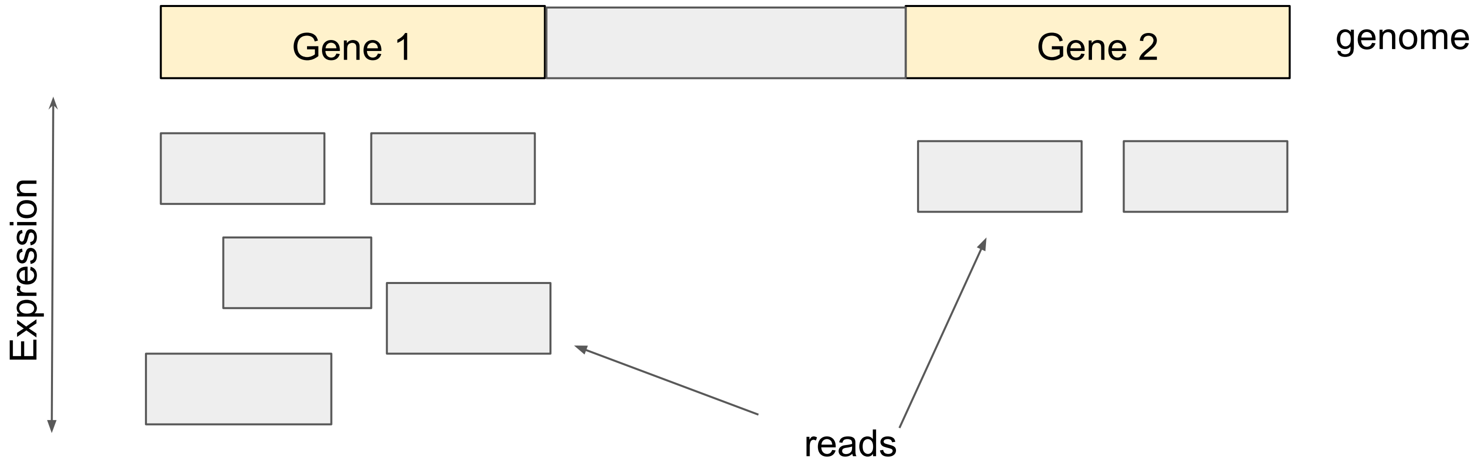
NGS Alignment - what can we use alignment for ?

- **Whole genome sequencing** - we map reads onto reference to find variation
- **Exome sequencing** - same as before, but only **exomic** DNA is captured. Saves a lot of money if you are only interested in genes.
- **ChIP-Seq/CLIP** - sequencing of DNA regions where binding of proteins happens.
- **Transcriptome sequencing -** sequencing of transcribed RNA in order to get expression profile of genes

NGS Alignment - what can we use alignment for ?

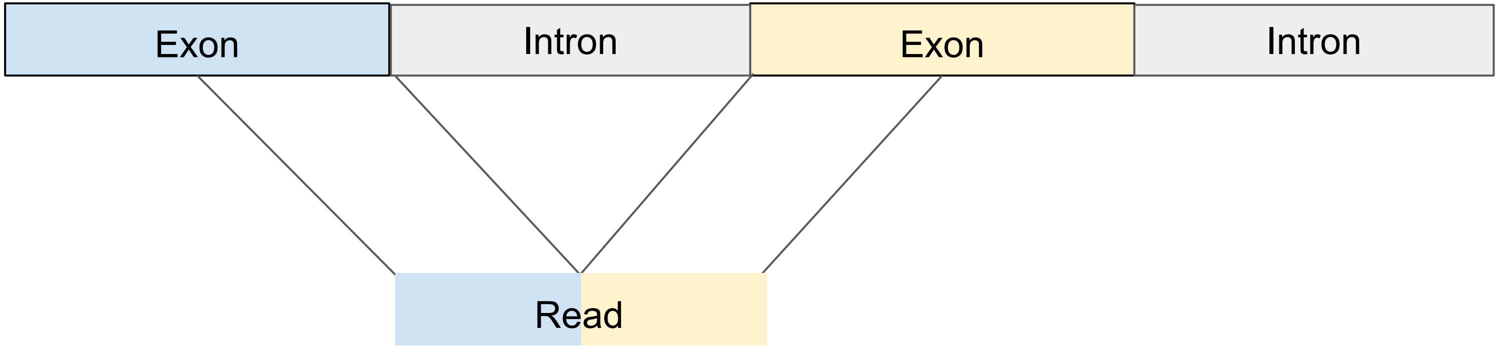
- **Whole genome sequencing** - we map reads onto reference to find variation
- **Exome sequencing** - same as before, but only **exomic** DNA is captured. Saves a lot of money if you are only interested in genes.
- **ChIP-Seq/CLIP** - sequencing of DNA regions where binding of proteins happens.
- **Transcriptome sequencing -** sequencing of **transcribed RNA** in order to get expression profile of genes

Transcriptome sequencing



Transcriptome sequencing - splicing

But how about Eukaryotes and their splicing ?



What to do about this ?

Transcriptome sequencing - splicing

1. Do not map reads onto a reference genome, but **reference transcriptome.**
   Reference transcriptome contains whole continuous transcribed sequences after splicing. No worries about introns.

2. Use **splice-aware** aligner
   Use aligner which is designed for transcriptome alignment and takes splicing into account. Examples: bowtie2, STAR, BWA, HISAT2 ...

Never use **non splice-aware aligner** to map RNA-seq reads onto a reference genome ! (e.g. bowtie, BFAST, …)

NGS aligners - summary

- Choose right aligner for the task at hand !
  - splice-aware vs non splice-aware
  - gapped vs non-gapped alignment
  - exact alignment vs fast approximate location (e.g. Kallisto)
- Aligners have often optimized default parameters for **specific reference.**
- Always read the manual.
- Never use settings without knowing what they are !
- Read the reviews !

http://bioinformatics.oxfordjournals.org/content/27/20/2790
http://www.ncbi.nlm.nih.gov/pubmed/24185836
http://www.biomedcentral.com/1471-2105/14/184
http://bib.oxfordjournals.org/content/11/5/473.full
http://omictools.com/read-alignment-c83-p1.html

Reference sequences

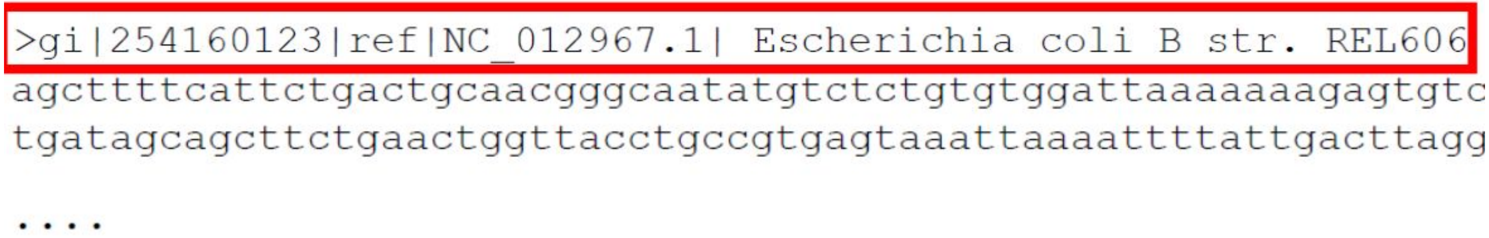
- It can be
  - reference genome
  - reference transcriptome
  - just some collection of sequences
- Usually a FASTA file
- Usually one long sequence per chromosome
- Unassembled parts of the genome at the end
- Naming of records is important !

Reference sequences

- It can be
    - reference genome
    - reference transcriptome
    - just some collection of sequences
- Usually a FASTA file
- Usually one long sequence per chromosome
- Unassembled parts of the genome at the end
- **Naming of records is important !**

Reference sequences - naming

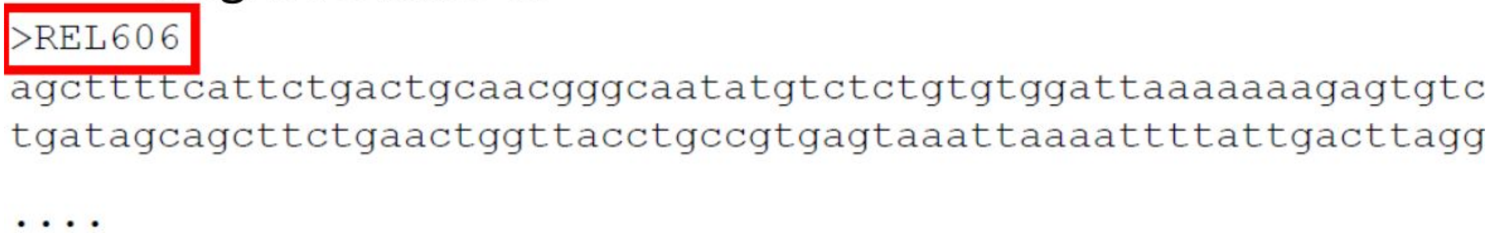
- FASTA Format



```
>gi|254160123|ref|NC_012967.1| Escherichia coli B str. REL606
agcttttcattctgactgcaacgggcaatatgtctctgtgtggattaaaaaaagagtgtc
tgatagcagcttctgaactggttacctgccgtgagtaaattaaaattttattgacttagg

....
```

- Using complex reference sequence names is a common problem during analysis
- Might rename to

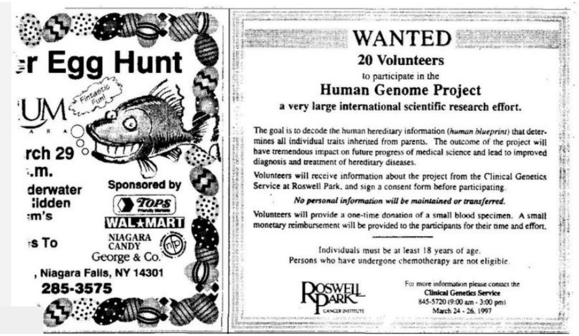


```
>REL606
agcttttcattctgactgcaacgggcaatatgtctctgtgtggattaaaaaaagagtgtc
tgatagcagcttctgaactggttacctgccgtgagtaaattaaaattttattgacttagg

....
```

Reference sequences - human genome

- One representative human genome reference sequence
    - Derived from DNA of 13 volunteers from Buffalo, NY
- Maintained by the **Genome Reference Consortium (GRC)**
    - New versions are released periodically
    - Results from different versions are not compatible !
    - Releases are provided by UCSC and NCBI
    - **Different sources use different chromosome identifiers (chr1 vs 1) !**



r Egg Hunt
Fintastic Fun!
UM
A R A
rch 29
.m.
derwater
idden
m's
s To
Sponsored by
TOPS
WAL★MART
NIAGARA CANDY
George & Co.
, Niagara Falls, NY 14301
285-3575

WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

The goal is to decode the human hereditary information (*human blueprint*) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

No personal information will be maintained or transferred.

Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

ROSWELL PARK CANCER INSTITUTE

For more information please contact the
Clinical Genetics Service
845-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

Reference sequences - annotations

- **Additional description of reference** - e.g. annotations of different regions on the reference
- **GTF** and **GFF** file format
- Names of chromosomes/sequences **have to match** the names in reference
- Different types of features
  - Manually verified genes
  - Predicted genes
  - Introns
  - ...

Considerations

- How many mismatches to allow ?
    - Vary depending on biology or genome completeness
- How many matches to report ?
    - Are you interested in multiple matches ?
- Require best match, first/any match ?
    - First match only is usually much faster.
- Quality of the reference sequence
    - How much can I trust my results ?
      If the reference is bad no aligner can save me !

You have to think about these questions !