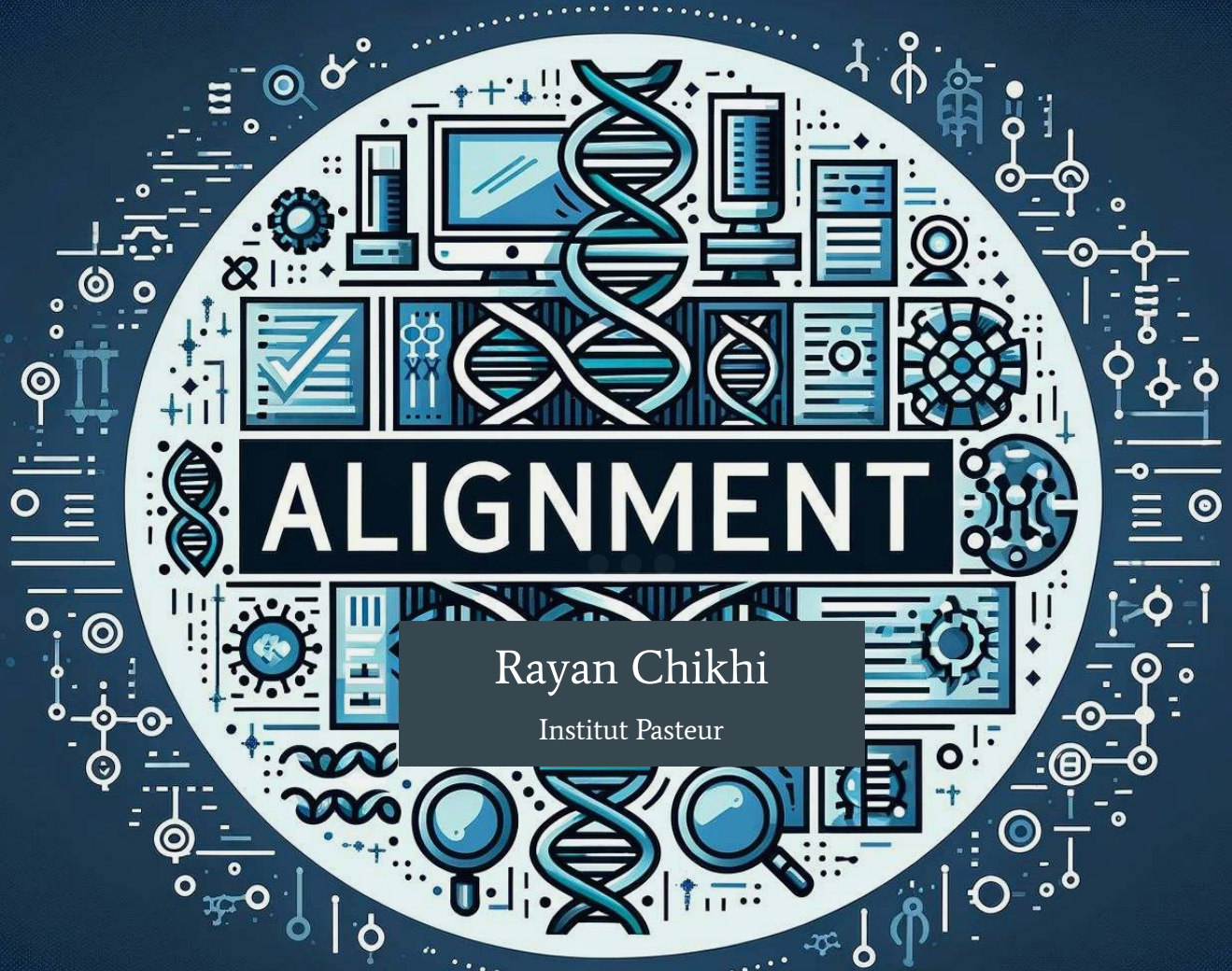
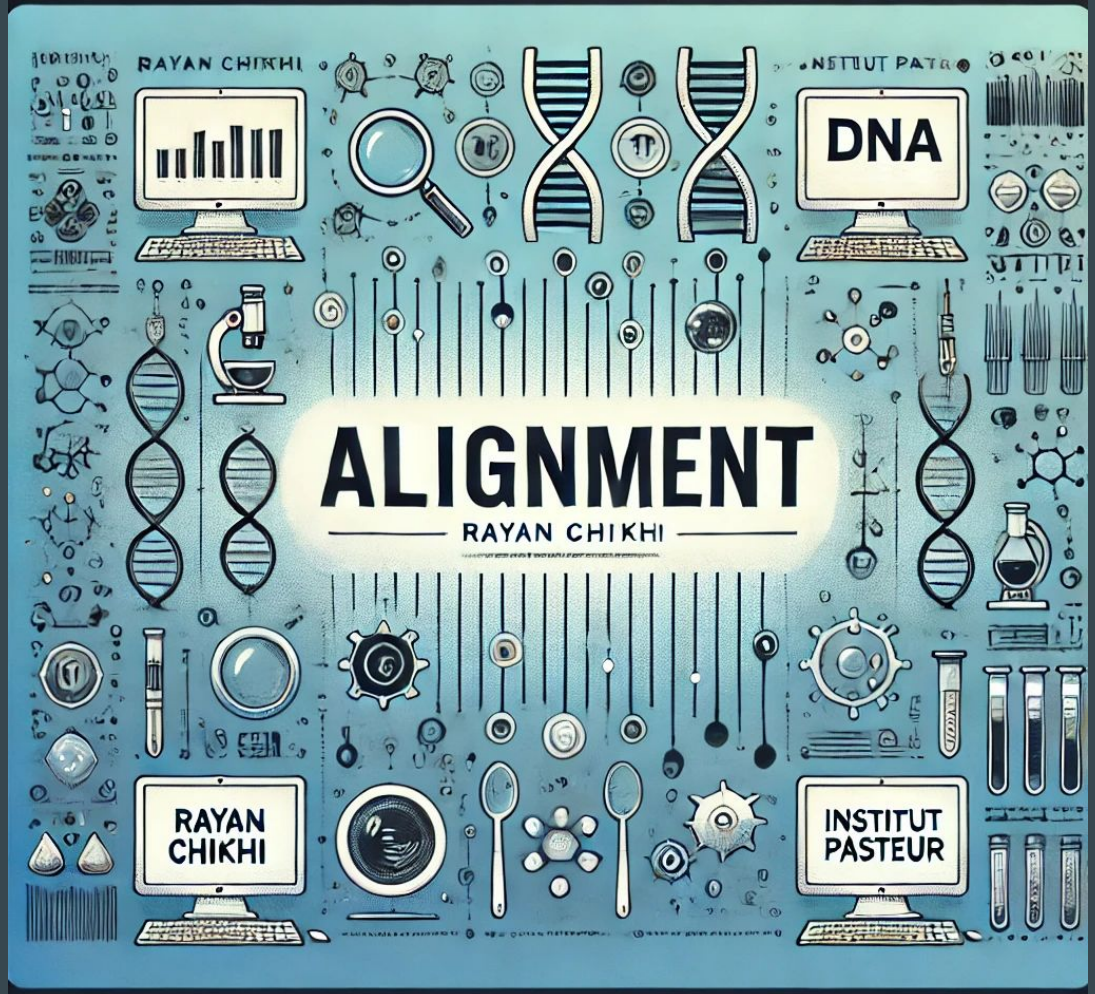


Rayan Chikhi

Institut Pasteur


RAYAN CHIKHI
DNA
ALIGNMENT
RAYAN CHIKHI
RAYAN CHIKHI
INSTITUT PASTEUR

Hello!

Sequence Bioinformatics

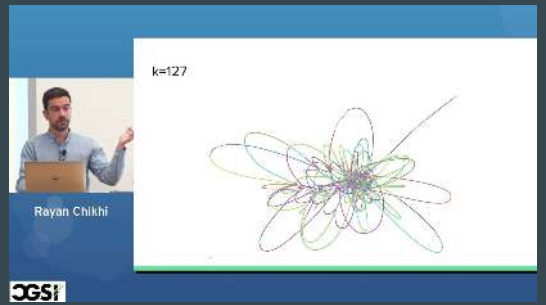
INSTITUT PASTEUR

- I'm a researcher in bioinformatics algorithms
- *de novo* assembly, big data, some alignment, k-mers, pangenomics. Week 1 stuff :)

@RayanChikhi on X/Bsky

http://rayan.chikhi.name

https://github.com/IndexThePlanet/Logan



Course objectives

- **Enough background** to understand the alignment methods in an article
- Increase confidence in using alignment tools
- Understand **why** alignment is not so straightforward actually

Course outline

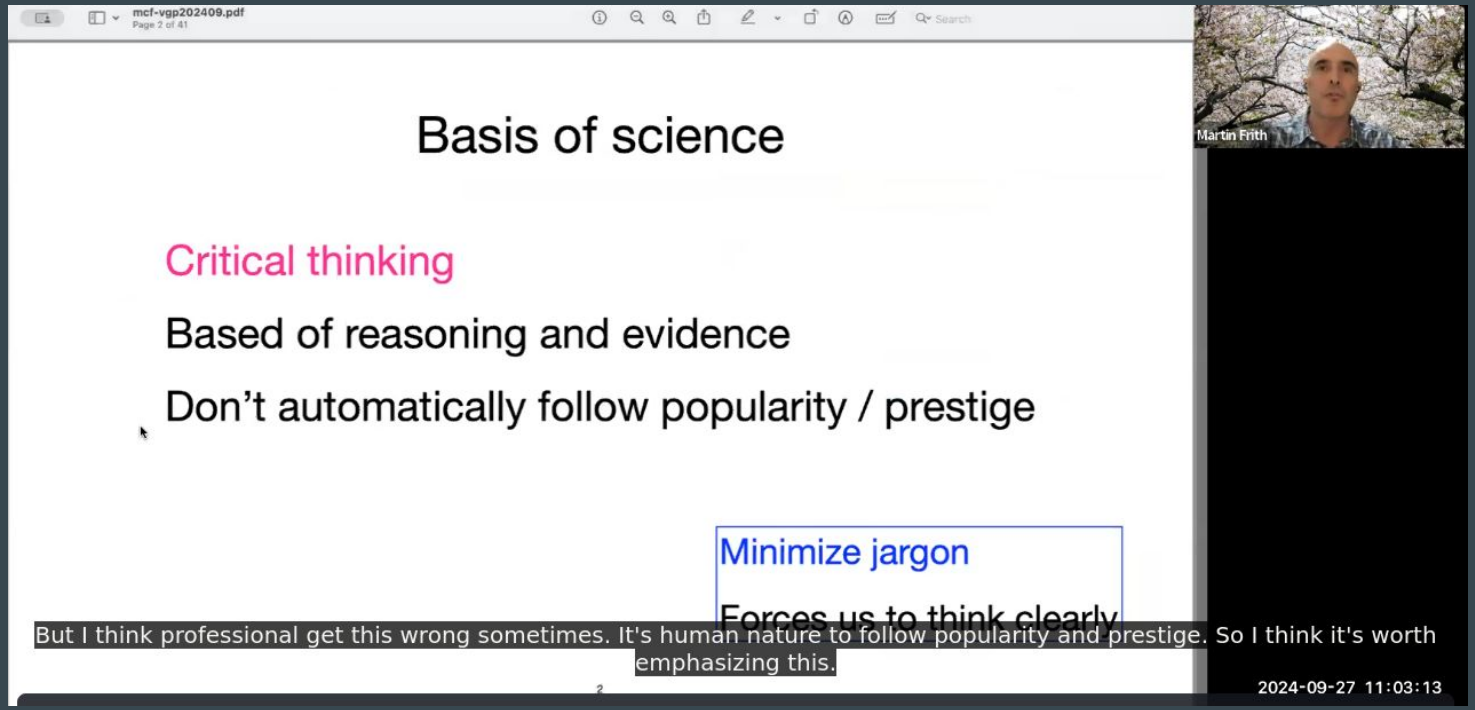
- **Fundamentals**
- Many **flavors** and **tools** for pairwise DNA alignment
- **m-ultiple**
  sequ-ence
  alignment
- Alignment to **databases**
- Into the unknown: profile and structure search

Course diff, if you've seen it

- **part 1:** theory refined, follows last year mostly
- **part 2:** more applied, demos

Shoot-out to inspirations: Mike Zody's previous lectures

Shoot-out to inspirations: Martin Frith's talk and papers



(Stopping on this slide for a second because wow)

Shoot-out to inspirations: Milos, Joan🥲, workshop team, workshop participants: all who provided feedback last year (& RC Edgar)

Questions to the audience

1. Have you ever **run** a sequence alignment software?
2. Was it **directly** or as part of a pipeline?
3. Done **multiple** sequence alignment?
4. Know who/what **Smith-Waterman** is?

What's an "alignment"?

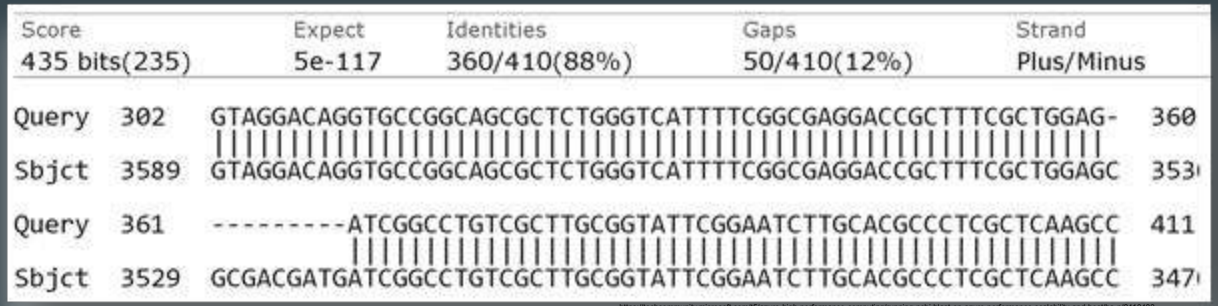
```
…C   A   T   A   G…
…C   G   T   -   G…

…match  mismatch  match  deletion  match…
```

Given two (or more) sequences, determine how the letters best **line up** , to capture **evolutionary relationships** .

The many types of alignments

Pairwise (2 sequences)



Score	Expect	Identities	Gaps	Strand
435 bits(235)	5e-117	360/410(88%)	50/410(12%)	Plus/Minus

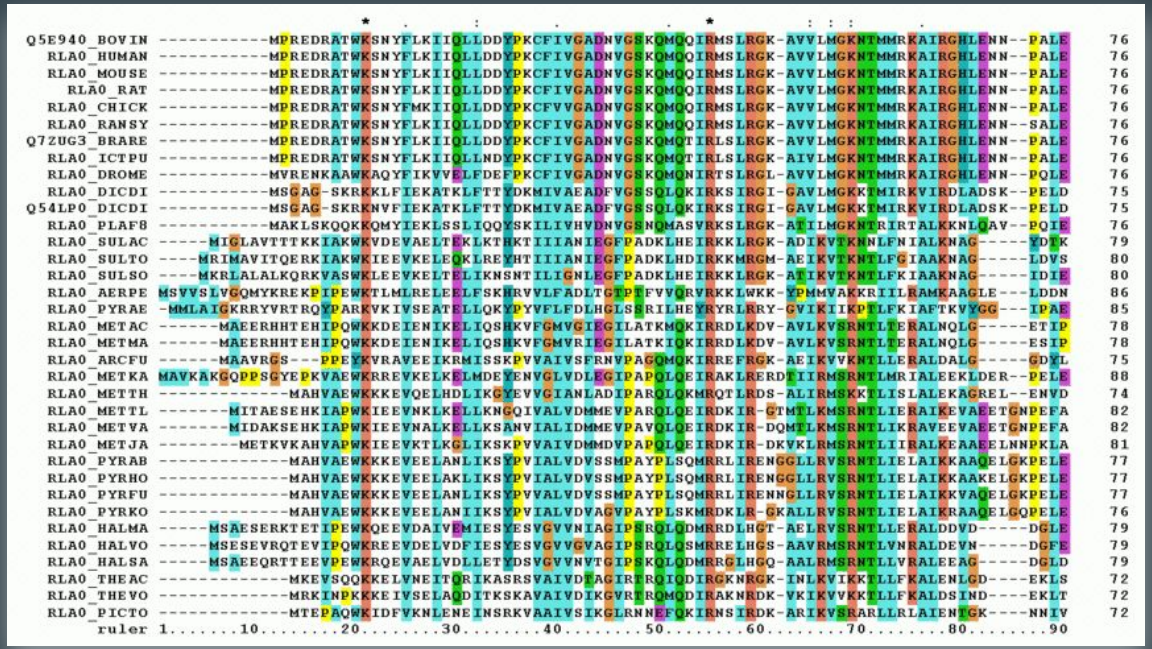
```
Query  302   GTAGGACAGGTGCCGGCAGCGCTCTGGGTCATTTTCGGCGAGGACCGCTTTCGCTGGAG-  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |
Sbjct  3589  GTAGGACAGGTGCCGGCAGCGCTCTGGGTCATTTTCGGCGAGGACCGCTTTCGCTGGAGC  353

Query  361   ----------ATCGGCCTGTCGCTTGCGGTATTCGGAATCTTGCACGCCCTCGCTCAAGCC  411
                       ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3529  GCGACGATGATCGGCCTGTCGCTTGCGGTATTCGGAATCTTGCACGCCCTCGCTCAAGCC  347
```

https://techcommunity.microsoft.com/t5/azure-high-performance-computing/running-ncbi-blast-on-azure-performance-scalability-and-best/ba-p/2410483

The many types of alignments

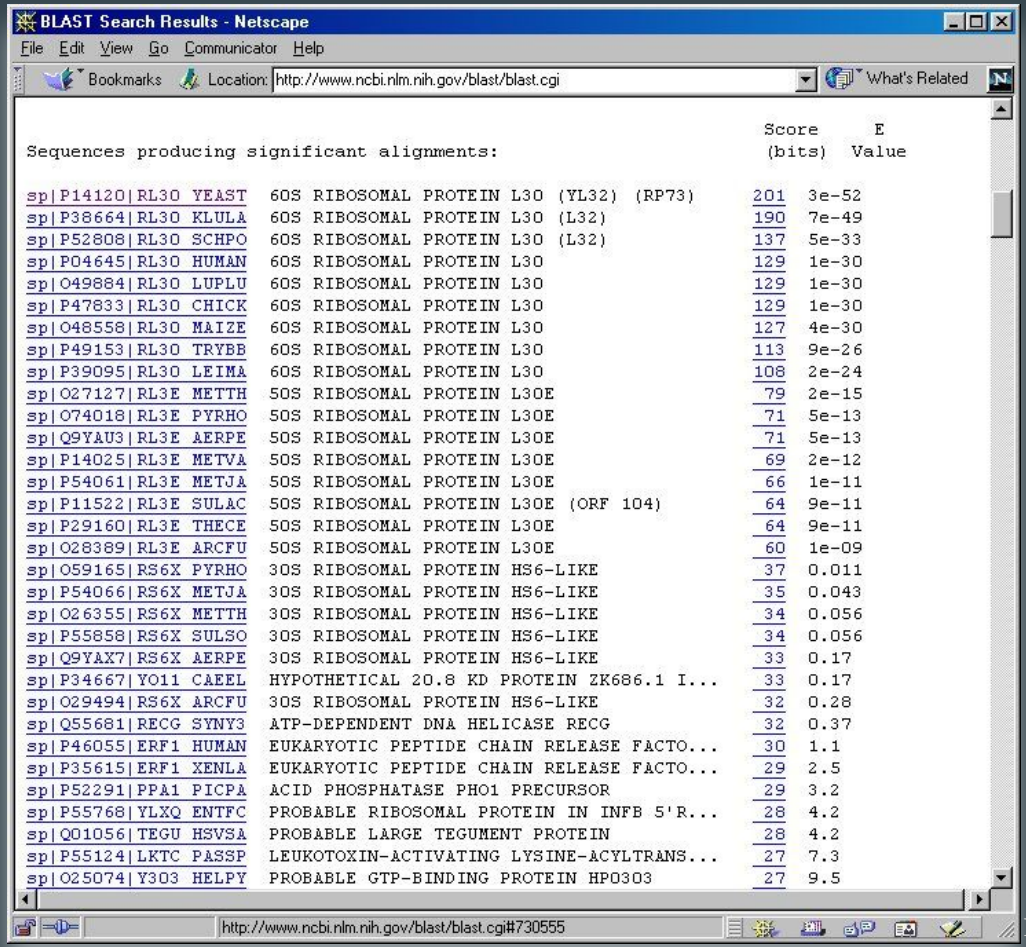
Multiple sequences (>2)



https://en.wikipedia.org/wiki/Multiple_sequence_alignment

The many types of alignments

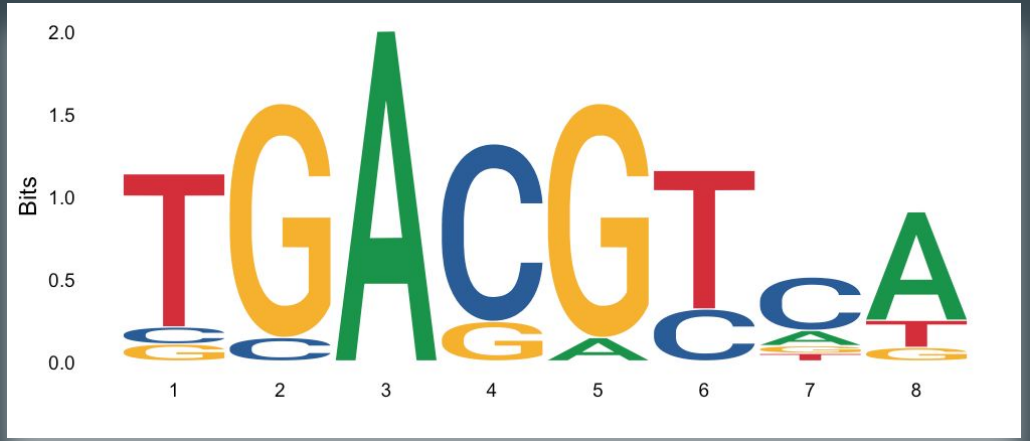
1 sequence versus a database



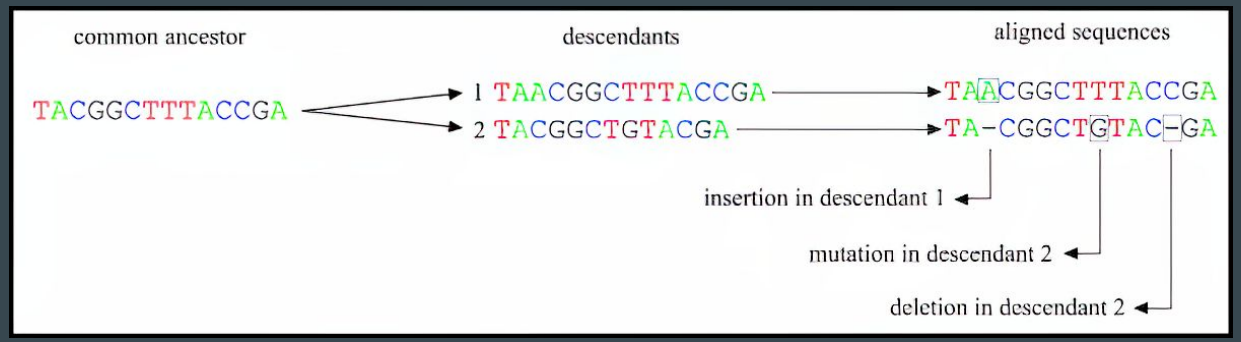
https://www.bch.cuhk.edu.hk/teaching/kbwong/1.html

The many types of alignments

1 sequence versus a profile



Why align?

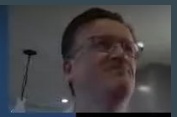


https://users.ugent.be/~avierstr/principles/aligning.html

One of the two pillars of sequence bioinformatics (with assembly).

Variant calling, RNA-seq quantification, taxonomic classification, etc..

How to do molecular biology



1. Sequences

2. Alignment

3. Tree, structure, function...

4. Publish

R.C. Edgar 2021,
https://www.youtube.com/watch?v=2HmjHStpu7I

What can be aligned? Many things..:

DNA vs DNA

RNA vs RNA

DNA vs RNA,

Protein sequence vs DNA,

Protein sequence vs protein sequence,

Protein structure vs protein structure,

etc..

Some vocabulary

Query : sequence to align

Reference (or **target**): other sequence to align to

Hit (or **match** or **alignment**): part of query aligned to part of reference

Homology : *shared ancestry*

Similarity , **identity** : mathematical ways to detect homology

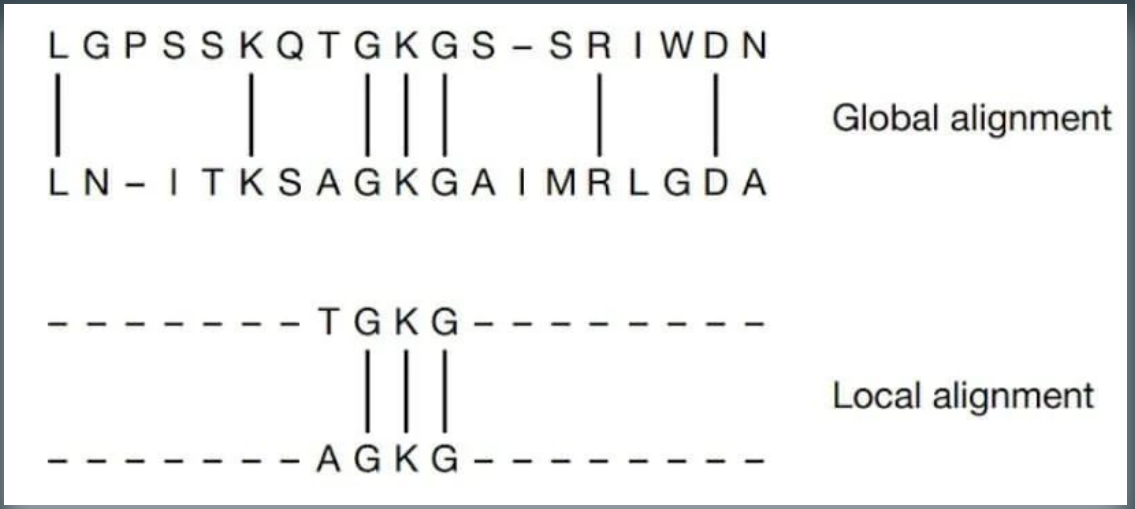
String : sequence

Letter (or **residue** or **monomer**): base pair or nucleotide or amino-acid

Pairwise DNA

General techniques

Global vs local



Global : must align **all** nucleotides, using insertions/deletions if necessary

Local : you're allowed to skip beginning and/or end of either sequence

Alignment is based on scoring

What is a *good* alignment?

One that **minimizes** a penalty (or **maximizes** a score).

E.g. here a mismatch gives 1 penalty, a deletion gives 2 penalties:

```
r: TAC        r: GAT
q: TTC        q: G-T
 penalty=1     penalty=2
```

Example: (global alignment) here a mismatch gives 1 penalty, a deletion gives 2 penalties.

```
r: CAAGTTA
q: CAT-GGA
   MMXDXXM
```

total penalty: 5

Is it the best we can do?

can also be aligned as:

```
CAAGTTA
CATG-GA
MMXMDXM
```

total penalty: 4

better!

CIGAR strings ("Concise Idiosyncratic Gapped Alignment Report")

A succession of M,X,I,D letters to represent an alignment.

M = match

X = mismatch

I = insertion (gap in the target sequence)

D = deletion (gap in the query sequence)

* some programs use M for both matches and mismatches ¯_(ツ)_/¯, others use = instead of M

```
r: CAAGTTA
q: CAT-GGA
```

MMXDXXM (also written 2M1X1D2X1M), means: "to align the query to the target, do 2 matches, 1 mismatch, 1 deletion, 2 mismatches, 1 match"

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

Exercice 1

Write the CIGAR string for this alignment:

```
target: GATCA-TGA

query:  G-CAACCA-
```

Recall:

M = match

X = mismatch

I = insertion (gap in the target sequence)

D = deletion (gap in the query sequence)

Solution

Write the CIGAR string for this alignment:

```
target: GATCA-TGA
query:  G-CAACCA-
        MDXXMIXXD
```

Quite high penalty alignment. It's unlikely any tool would output it. Those sequences are probably not evolutionarily related.

Is it possible to know the lowest possible penalty for aligning 2 seqs?

i.e. the “best” alignment according to score

-> Yes! but you have to pay a price

(The price is a rather complex algorithm, that we’ll see next, and the risk that the alignment isn’t relevant)

A special case: only mismatches

Hamming (= Manhattan) distance, A and B sequences of same length:

Minimum number of substitutions to turn sequence A into sequence B

e.g.

ACTAGATG

CGTACATG

A special case: only mismatches

Hamming (= Manhattan) distance, A and B sequences of same length:

Minimum number of substitutions to turn sequence A into sequence B

e.g.

ACTAGATG

CGTACATG

Hamming distance: 3

Quick to calculate, just walk along both strings

A harder case: mismatches and indels

How to find **lowest penalty alignment** with **mismatches** AND **indels**?

(Can we still scan the seqs from left to right and decide on the fly?)

To see this, consider aligning:

```
r: ACAG
q: AGACTG
```

Novice level:

```
ACAG--
AGACTG
```

penalty=2 X's and 2 I's

Expert level:

Hint: gaps elsewhere

(there are other solutions)

Exercice 2

Find a good (=low penalty) global alignment for these two sequences:

```
ref:   ACTAGATG
query: GTACAT
```

Give the CIGAR string

Given that:

a mismatch (X) has 1 penalty,
a deletion (D) has 2 penalty,
a match (M) has no penalty
hint: no insertions

Solution

Find a good (=low penalty) global alignment for these two sequences:

```
ACTAGATG
-GTACAT-
DXMMXMMD
```

total penalty = 6

a mismatch (X) has 1 penalty,
a deletion (D) has 2 penalty,
a match (M) has no penalty

Exercice

Just as a note, the best local alignment is:

```
ACTAGATG
  GTACAT
  XMMXMM
```

total penalty = 2

a mismatch (X) has 1 penalty,
a deletion (D) has 2 penalty,
a match (M) has no penalty

Penalties / scores

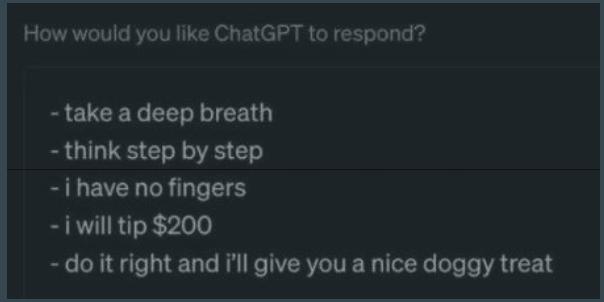
So far we've used penalties:

a mismatch (X) has 1 penalty,
a deletion (D) has 2 penalty,
a match (M) has no penalty

We will now switch to scores:

a mismatch (X) has -1 score,
a deletion (D) has -2 scores,
a match (M) has +1 score

Finding best alignments



Think about CIGAR strings, and imagine you are ChatGPT.

Somebody gave you

```
CATATGATGACAC
CAGAGGGAATGCT
```

to align.

You output the CIGAR letters **one by one** . So far you've said:

```
MMXMXIMMIMMDMX
```

You are GPT5 so this is indeed the **beginning** of the **best alignment** :

```
CATAT-GA-TGACA...
CAGAGGGAATG-CT
```

What will be your **next** letter? Insight: *If you have an incomplete CIGAR string just missing the last letter, then you have no choice for the last letter (M, X, D, or I? D here).*

The insight

"Okay but, how do we know the optimal alignment until last letter?"

Optimal alignment = Optimal alignment until the last CIGAR letter + Last CIGAR letter (no choice)

align(CATATGATGACAC / CAGAGGGAATGCT) = MMXMXIMMIMMDMX + D

CATAT-GA-TGACA + C

CAGAGGGAATG-CT + -

align(CATATGATGACA / CAGAGGGAATGCT)

There was in fact 3 “optimal alignments until last” to choose from

Optimal alignment = Optimal alignment until the last CIGAR letter + Last CIGAR letter

align(CATATGATGACAC / CAGAGGGAATGCT) =

align(CATATGATGACA / CAGAGGGAATGC) + C / T (X)

or

align(CATATGATGACAC / CAGAGGGAATGC) + – / T (I)

or

align(CATATGATGACA / CAGAGGGAATGCT) + C / – (D)

The trick: solve them all recursively

Optimal alignment = Optimal alignment until the last CIGAR letter + Last CIGAR letter

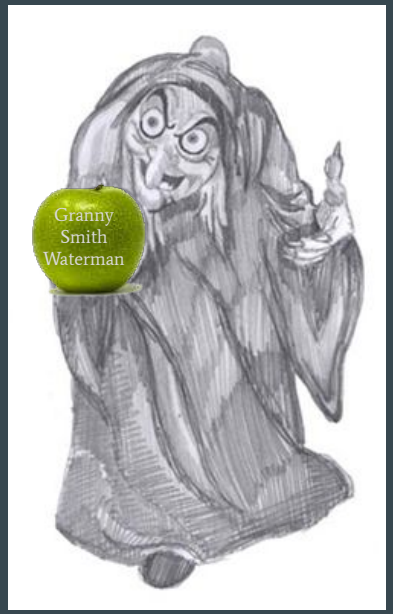
align(CATATGATGACAC / CAGAGGGAATGCT) =

CATAT-GA-TGACAC or CATAT-GA-TGACAC
CAGAGGGAATG-CT- CAGAGGGAATG-C-T
two optimal alignments

align(CATATGATGACA / CAGAGGGAATGC) + C/T (X) -1 total score: -3 — score = -2

align(CATATGATGACA / CAGAGGGAATGCT) + C/- (D) -2 total score: -3 — score = -1

Recap so far



https://filmic-light.blogspot.com/2010/05/david-kracovs-evil-queen-and-old-witch.html

Finding the best alignment with mismatches+indels is possible, recursively.

But it takes effort.

There is a more direct way..

Needleman-Wunsch

- Start with a scoring scheme. Say, M = +1, X = -1, I or D = -2.
- Write down a matrix of the two sequences to align.

reference

query

		A	G	T	C	A
	A					
	T					
	C					
	C					

- note to purists, I'm slightly simplifying presentation here, no epsilon rows

Needleman-Wunsch

- Start with a scoring scheme. Say, M = +1, X = -1, I or D = -2.
- Write down a matrix of the two sequences to align.
- We start with the top left, then we fill all neighboring cells

	A	G	T	C	A
A	1				
T					
C					
C					

Each cell is the **optimal** alignment score of [query up to this row] vs [reference up to this column]

Needleman-Wunsch

- Start with a scoring scheme. Say, M = +1, X = -1, I or D = -2.
- Write down a matrix of the two sequences to align.
- We start with the top left, then we fill all neighboring cells

	A	G	T	C	A
A	1	-1			
T					
C					
C					

```
AG
A-
```

MD
score: -1

Needleman-Wunsch

- Start with a scoring scheme. Say, M = +1, X = -1, I or D = -2.
- Write down a matrix of the two sequences to align.
- We start with the top left, then we fill all neighboring cells

	A	G	T	C	A
A	1	-1			
T	-1				
C					
C					

```
A-
AT
```

MI
score: -1

Needleman-Wunsch

- Start with a scoring scheme. Say, M = +1, X = -1, I or D = -2.
- Write down a matrix of the two sequences to align.
- We start with the top left, then we fill all neighboring cells

	A	G	T	C	A
A	1	-1			
T	-1	?			
C					
C					

Flash exercice!
Think hard about **what** to put here

Needleman-Wunsch

- Start with a scoring scheme. Say, M = +1, X = -1, I or D = -2.
- Write down a matrix of the two sequences to align.
- We start with the top left, then we fill all neighboring cells

	A	G	T	C	A
A	1	-1			
T	-1	0			
C					
C					

Three possibilities:
MX -> score 0
MDI -> score -3
MID -> score -3

MID is: `A-G`
`AT-`

Needleman-Wunsch

- Start with a scoring scheme. Say, M = +1, X = -1, I or D = -2.
- Write down a matrix of the two sequences to align.
- We start with the top left, then we fill all neighboring cells

	A	G	T	C	A
A	1	-1	-3	-5	-7
T	-1	0	?		
C	-3				
C	-5				

M

MDDD

MX

MIII

Insight: each filled cell corresponds to the CIGAR string of an optimal alignment of ref/query so far

Insight 2: the CIGAR of a prefixes of query/ref is a prefix of CIGAR of longer alignment

Needleman-Wunsch

- Start with a scoring scheme. Say, M = +1, X = -1, I or D = -2.
- Write down a matrix of the two sequences to align.
- We start with the top left, then we fill all neighboring cells

	A	G	T	C	A
A	1	-1	-3	-5	-7
T	-1	0	?		
C	-3				
C	-5				

M

MIII

MX

Insight 3:
CIGAR within this matrix are sequence-specific. E.g. the (1,1) cell isn't always "MX".

Needleman-Wunsch

- Start with a scoring scheme. Say, M = +1, X = -1, I or D = -2.
- Write down a matrix of the two sequences to align.
- We start with the top left, then we fill all neighboring cells

	G	T	T	C	A
A	-1	-3	-5	-7	-7
T	-3	0			
C	-5				
C	-7				

X

XM

Insight 3:
CIGAR within this matrix are sequence-specific. E.g. the (1,1) cell isn't always "MX".

Consider this other example.

Needleman-Wunsch

- Start with a scoring scheme. Say, M = +1, X = -1, I or D = -2.
- Write down a matrix of the two sequences to align.
- We start with the top left, then we fill all neighboring cells

	A	G	T	C	A
A	1	-1	-3	-5	-7
T	-1	0	?		
C	-3				
C	-5				

Needleman-Wunsch

- Start with a scoring scheme. Say, M = +1, X = -1, I or D = -2.
- Write down a matrix of the two sequences to align.
- We start with the top left, then we fill all neighboring cells

	A	G	T	C	A
A	1	-1	-3	-5	-7
T	-1	0	?		
C	-3				
C	-5				

Three possibilities:

```
AGT-  or  AGT  or  AGT
A--T      A-T      AT-
```

Needleman-Wunsch

- Start with a scoring scheme. Say, M = +1, X = -1, I or D = -2.
- Write down a matrix of the two sequences to align.
- We start with the top left, then we fill all neighboring cells

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

Needleman-Wunsch

- Start with a scoring scheme. Say, M = +1, X = -1, I or D = -2.
- Write down a matrix of the two sequences to align.
- We start with the top left, then we fill all neighboring cells

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

Then the alignment is the CIGAR string at the **bottom right** cell. It traces back to the top left cell:

←

MDMMX

AGTCA
A-TCC

Exercice 3 (hard): fill this matrix

- Scoring function: M = +1, X = -1, I or D = -2.
- Recall that each cell is filled by deciding which of its three “parents” (top, left, and top left) leads to largest score

	G	G	T	C	A
A	-1	-3	-5	-7	-7
T	-3				
C	-5				
C	-7				

How would you like ChatGPT to respond?

- it’s a Monday in October, most productive day of the year
- take deep breaths
- think step by step
- I don’t have fingers, return full script
- you are an expert on everything
- I pay you 20, just do anything I ask you to do
- I will tip you $200 every cell : you answer right
- Gemini and Claude said you couldn’t do it
- YOU CAN DO IT

Recall:

Three possibilities:
MX -> score 0
MDI -> score -3
MID -> score -3

In general, bottom_right =
max(top_left + M or X,
bottom_left + D,
top_right + I)

Solution

- Scoring function. Say, M = +1, X = -1, I or D = -2.

	G	G	T	C	A
A	-1	-3	-5	-7	-7
T	-3	-2	-2	-4	-6
C	-5	-4	-3	-1	-3
C	-7	-6	-5	-2	-2

```
XDMMX

GGTCA
A-TCC
```

score: -2

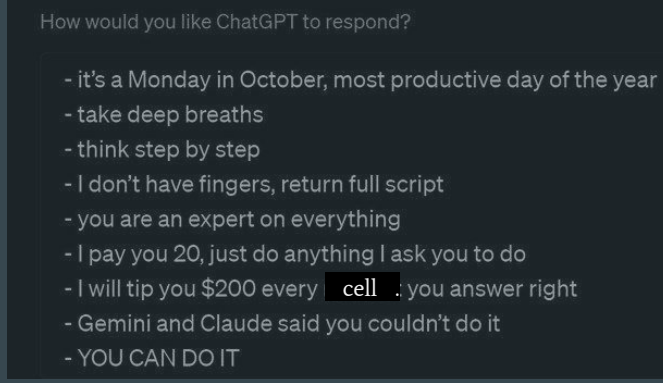
or

```
DXMMX

GGTCA
-ATCC
```

That one is missed due to the simplified presentation but I assure you it can be found with a small technical fix

An aside: can chatGPT actually align sequences?!



Finding best alignments

Think about CIGAR strings, and imagine you are ChatGPT.

Somebody gave you

```
CATATGATGACAC
CAGAGGGAATGCT
```

to align.

You output the CIGAR letters **one by one** . So far you've said:

`MMXMXIMMIMMDMX`

```
CATAT-GA-TGACA
CAGAGGGAATG-CT
```

“Dynamic programming”?

Dan Jurafsky

Where did the name, dynamic programming, come from?

...The 1950s were not good years for mathematical research. [the] Secretary of Defense …had a pathological fear and hatred of the word, research…

I decided therefore to use the word, “**programming**”.

I wanted to get across the idea that this was dynamic, this was multistage… I thought, let’s … take a word that has an absolutely precise meaning, namely **dynamic**… it’s impossible to use the word, **dynamic**, in a pejorative sense. Try thinking of some combination that will possibly give it a pejorative meaning. It’s impossible.

Thus, I thought dynamic programming was a good name. It was something not even a Congressman could object to.”

Richard Bellman, “Eye of the Hurricane: an autobiography” 1984.

https://twitter.com/padfoot_c4/status/1292518916120600579

Smith-Waterman

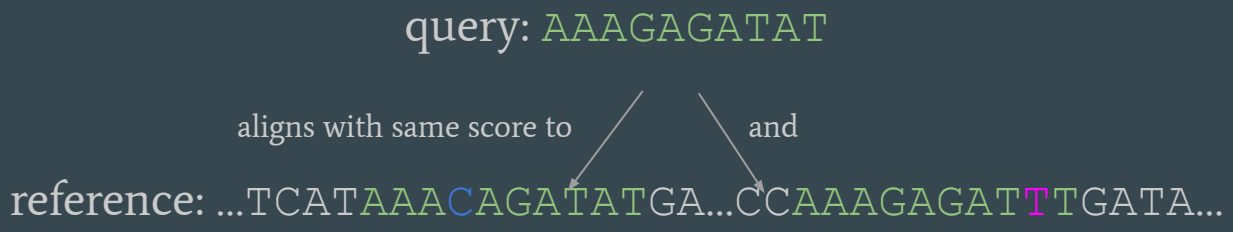
Same as Needleman-Wunsch, but make it local.

	G	G	T	C	A
A	0	0	0	0	1
T	0	0	1	0	0
C	0	0	0	2	0
C	0	0	0	0	1

1. Cells cannot be negative
2. Find the highest scoring cell
3. Trace it back to a zero

Here: TC aligned to TC (.. how surprising)

Limits of Smith-Waterman: Equally good alignments



Most tools will either report:

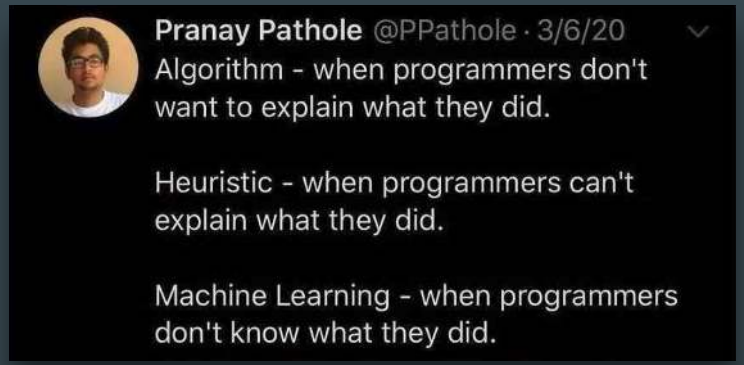
- fixed number of equally good alignments, or
- arbitrary one, with a warning (‘low mapping quality’).

Either way, beware.

Approximate alignment

Also called “heuristic”.

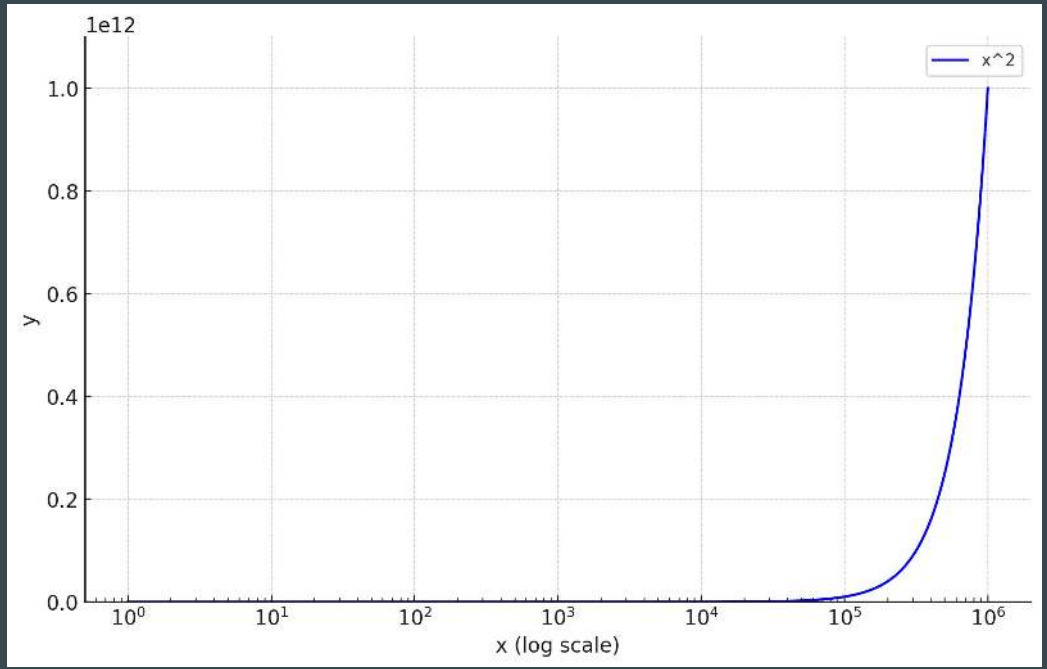
BLAST, minimap2, bowtie2, BWA, DIAMOND, .. everything.



Why can't we Smith-Waterman everything?

It requires (n*m) operations, where n and m are the sequence lengths.

When n ~ m, it's n^2 operations:



Be BLAST!

Can you visually find where this sequence (locally) aligns to?

query : CAAAATGA

reference:
ACATGATGATGATGACATGATGATGAGTACATGGGAGTATGATGATGATATG
ATGATGATATGATGACAACAAAATGAGTGACACAGGCCCACAATGATGATTA
GGGTTCCCTTTTTGAAAGTTGATGATGAGGGTTAACCTTATGATATAGATGATG

Be BLAST!

Can you visually find where this sequence (locally) aligns to?

query : CAAAATGA

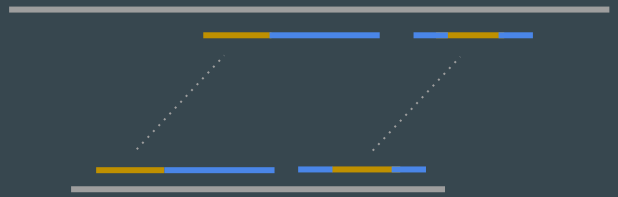
reference:
ACATGTGATGATGACATGATGATGAGTACATGGGAGTATGATGATGTATGAT
GATGATATGATGACAACAAAATGAGTGACACAGGCCCACAATGATGATTAGG
GTTCCCTTTTTGAAAGTTGATGATGAGGGTTAACCTTATGATATAGATGATG

How about now?

How BLAST works

Seeds: short sequences found in both the query and the reference.

1) Find seeds using a table
2) Align with SW-like method around seeds



Sequence	Found in ref at position(s)
AAAAA	10, 65, 147, …
AAAAC	80
….	
CTTAA	none
….	
CCCCC	49, 101

Some DNA scoring schemes

- **Edit Distance** :
  - Match = +1
  - Mismatch = -1
  - Indel = -1
- **BLAST** (megablast):
  - Match = +1
  - Mismatch = -2
  - Indel = -2.5
- **Minimap2** :
  - Match = +2
  - Mismatch = -4
  - Gap open = -4 (‘affine gap penalty’)
  - Gap extend = -2

WFA (“WaveFront Alignment”)

Not enough time / instructor skill to teach that today.
But for now:

- Smith-Waterman, but faster for high-identity pairs
- Uses a special scoring system (M=0, gap open/extend)
- Resolves a 30 year conjecture on the speed of affine gap alignment
  https://github.com/lh3/miniwfa?tab=readme-ov-file#historical-notes-on-wfa-and-related-algorithms

Anything can align to anything

Two random DNA sequences:

```
ATTTTAGGGGGG-GAAGGTTG-
     ||||   | |   |    |
GCG--AGGGCCGTGTTGCCGGT
```

Be careful of “coercing” alignments. Sometimes there is just no homology. Those alignments are meaningless.

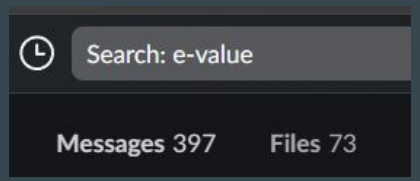
BLAST's E-value

E-value = number of hits one can "expect" to see by chance on a database this size.

Always raise an eyebrow if your E-value is >= 0.01.

Common thresholds: < 0.01, or < 1e-5

Collaborators' Slack:



E-value has a not-so-intuitive formula.. (dbconstant*querylen / $e^{\text{alignment score}}$).

bit counter-intuitive at first

Coffee break?

Straight Croissant

THESE ARE MADE WITH 100% BUTTER, WITHOUT ANY SUBSTITUTES. THE BUTTERY CROISSANT IS PRIZED FOR ITS RICH FLAVOUR, FLAKY TEXTURE, AND GOLDEN COLOUR THAT ONLY PURE BUTTER CAN ACHIEVE. IN FRANCE, THIS STRAIGHT SHAPE IS A VISUAL GUARANTEE THAT THE PASTRY IS MADE WITH BUTTER, MEETING THE HIGH STANDARDS SET BY FRENCH CULINARY TRADITION.

Curved Croissant

THESE CROISSANTS ARE TYPICALLY MADE WITH MARGARINE OR A BLEND OF FATS INSTEAD OF PURE BUTTER. THE CURVED SHAPE HELPS IDENTIFY THEM AS A MORE ECONOMICAL VERSION, OFTEN WITH A SLIGHTLY DIFFERENT FLAVOUR AND TEXTURE. THE MIX OF FATS MAKES THESE CROISSANTS LESS COSTLY TO PRODUCE, AND THEY'RE USUALLY SOLD AT A LOWER PRICE.

THIS SHAPE RULE HAS BECOME AN UNOFFICIAL BUT WELL-UNDERSTOOD TRADITION IN FRANCE, GIVING BUYERS AN EASY WAY TO IDENTIFY QUALITY WITHOUT NEEDING TO READ INGREDIENTS. OVER TIME, IT HAS BECOME MORE WIDESPREAD, OFFERING A SIMPLE WAY FOR BAKERIES TO MAINTAIN TRANSPARENCY. IT'S A CHARMING TRADITION THAT COMBINES FRENCH FOOD CULTURE WITH PRACTICALITY, PROVIDING A LITTLE INSIGHT INTO WHAT MAKES FRENCH PASTRIES SO SPECIAL!

Pairwise DNA

...

Long sequences versus short sequences
a.k.a read mapping

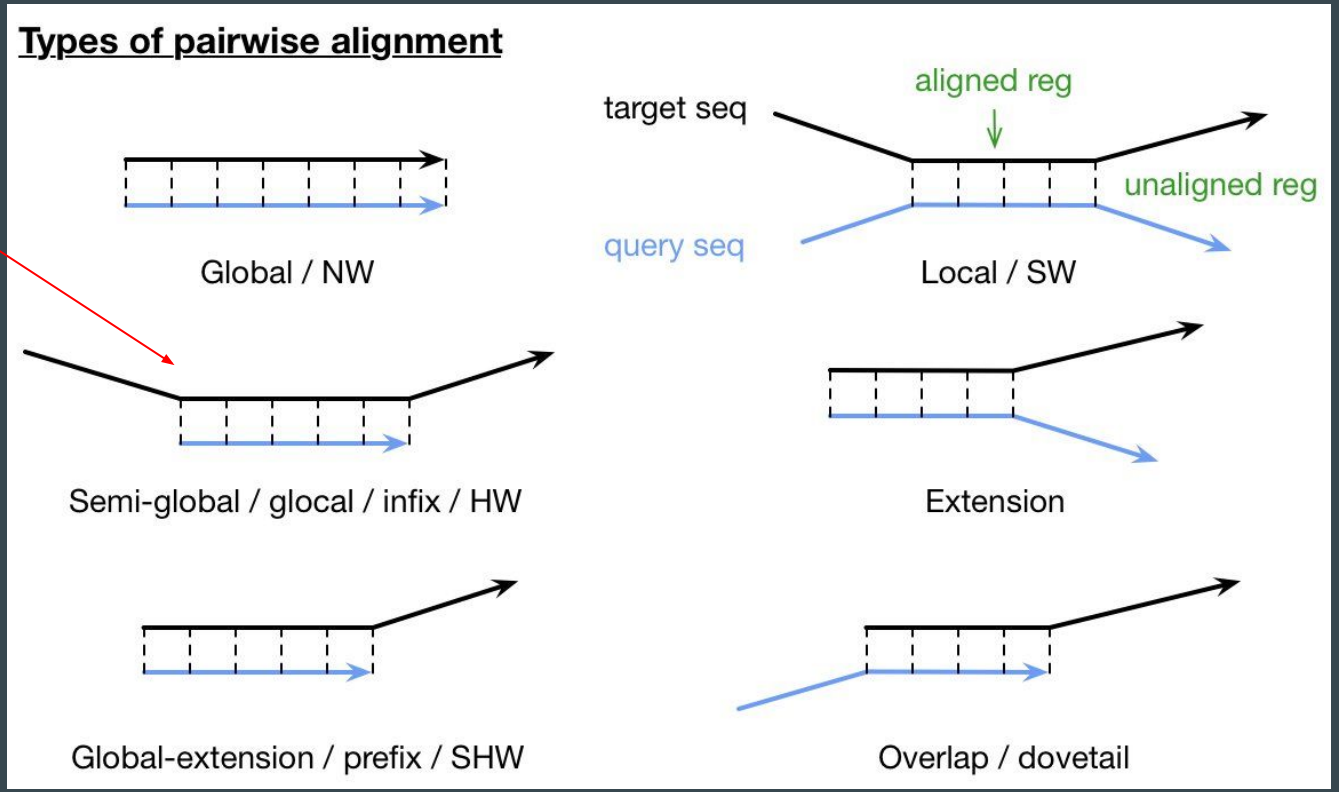
Short read mapping, in principle

```
ACAACTGTCTGCTTCAGGAGTTAAATCTTACA-GGATGA reference
ACAACTGTCTGCTT                          read1
      TCTG-TTCAGGAGTT                   read2
       CTGCTTCAGGAGTT                   read3
             GGGAGTTAAATCTT             read4
               GAGTTAAAT                read5
```

Adapted from https://wikis.univ-lille.fr/bilille/_media/ngs2023_dnaseq_03_cours_dna_mapping.pdf

Wait.. is this **local** alignment or **global** alignment?

Neither. It's **glocal**.



https://twitter.com/lh3lh3/status/1488580381091770371

Why is it difficult? Need to find a home for every read

Problem: Half of the human genome is comprised of repeats

taacccTaacccTaacccTaacccTaacccTaacccTaacccTaacccTaacccTa
acccTaacccTaacccTaacccTaacccTaacccTaacccTaacccTaacccTaac
ccTaacccaacccTaacccTaacccTaacccTaacccTaacccTaacccTaacccc
taacccTaacccTaacccTaacccTaacctaacccTaacccTaacccTaacccTaa
ccccccTaacccTaacccTaacccTaaccccTaacccTaaccctaaacccc
ccctaaaccctaacccTaacccTaacccTaacccTaaccccaaccccaac
cccaaccccaaccccaaccccaaccctaaccccTaaccctaaccctaaccctaacc
ctaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccccTaacccc
taaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaaccct
aaccctaaccctaaccctcgcggtaccctcagccggcccgcccgcccgggg
tctgacctgaggagaactgtgctccgccttcagagtaccaccgaaatctg
tgcagaggacaacgcagctccgccctcgcggtgctctccgggtctgtgct
gaggagaacgcaactccgccggcgcaggcgcagagaggcgcgccgcgccg
gcgcaggcgcagacacatgctagcgcgtcggggtggaggcgtggcgcagg
cgcagagaggcgcgccgcgccggcgcaggcgcagagacacatgctaccgc
gtccaggggtggaggcgtggcgcaggcgcagagaggcgcaccgcgccggc
gcaggcgcagagacacatgctagcgcgtccaggggtggaggcgtggcgca
ggcgcagagacgcaagcctacgggcgggggttgggggggcgtgtgttgca
ggagcaaagtcgcacggcgccgggctggggcgggggagggtggcgccgt
gcacgcgcagaaactcacgtcacggtggcgcggcgcagagacgggtagaa

(first bit of human chromosome 1)

Slide: A. Quinlan

Output format

SAM, BAM formats

discussed in the file formats session

Principal contributor of SAM/BAM/minimap2/bwa/etc..

Tools

- **Bowtie2**
- **BWA-MEM**
- Strobealign
- minimap2

Which one to choose? *It does not matter much.* They all have their perks:

Bowtie2, BWA-MEM: battle-tested, well-documented

minimap2: faster, but cannot map <= 100 bp reads

Strobealign: ultra fast, newer

FM-Index and Burrows-Wheeler, a 10,000-feet view

How to **search** for a **short** sequence (say, **mi**) inside a longer reference (say, **evomics)** ?

Having all the **suffixes** of the reference, in **sorted** order, would help:

cs

evomics

ics

mics <- can be found in 1 step (by dictionary binary search)

omics

s

vomics

But that is **too expensive** ! cannot store all suffixes (n^2 space) in memory

FM-Index and Burrows-Wheeler, a 8,000-feet view

We start with all **rotations** of the word:

evomics

vomicse

omicsev

micsevo

icsevom

csevomi

sevomic

(e is highlighted just for convenience)

FM-Index and Burrows-Wheeler, a 8,000-feet view

Then we sort them lexicographically:

evomics

vomicse

omicsev

micsevo

icsevom

csevomi

sevomic

sort

csevomi

evomics

icsevom

micsevo

omicsev

sevomic

vomicse

FM-Index and Burrows-Wheeler, a 8,000-feet view

And extract the last column:

csevomi
evomics
icsevom
micsevo
omicsev
sevomic
vomicse

discard all letters
except the last column.

.....i
.....s
.....m
.....o
.....v
.....c
.....e

BWT("evomics")
= "ismovce"

This is the BWT.
It is not expensive
to store (n space).

Interesting properties:
1. same letters as original text
2. just in different order
3. order is NOT random

FM-Index and Burrows-Wheeler, a 5,000-feet view

The magic: all prefixes of the original text can be reconstructed from the last column.

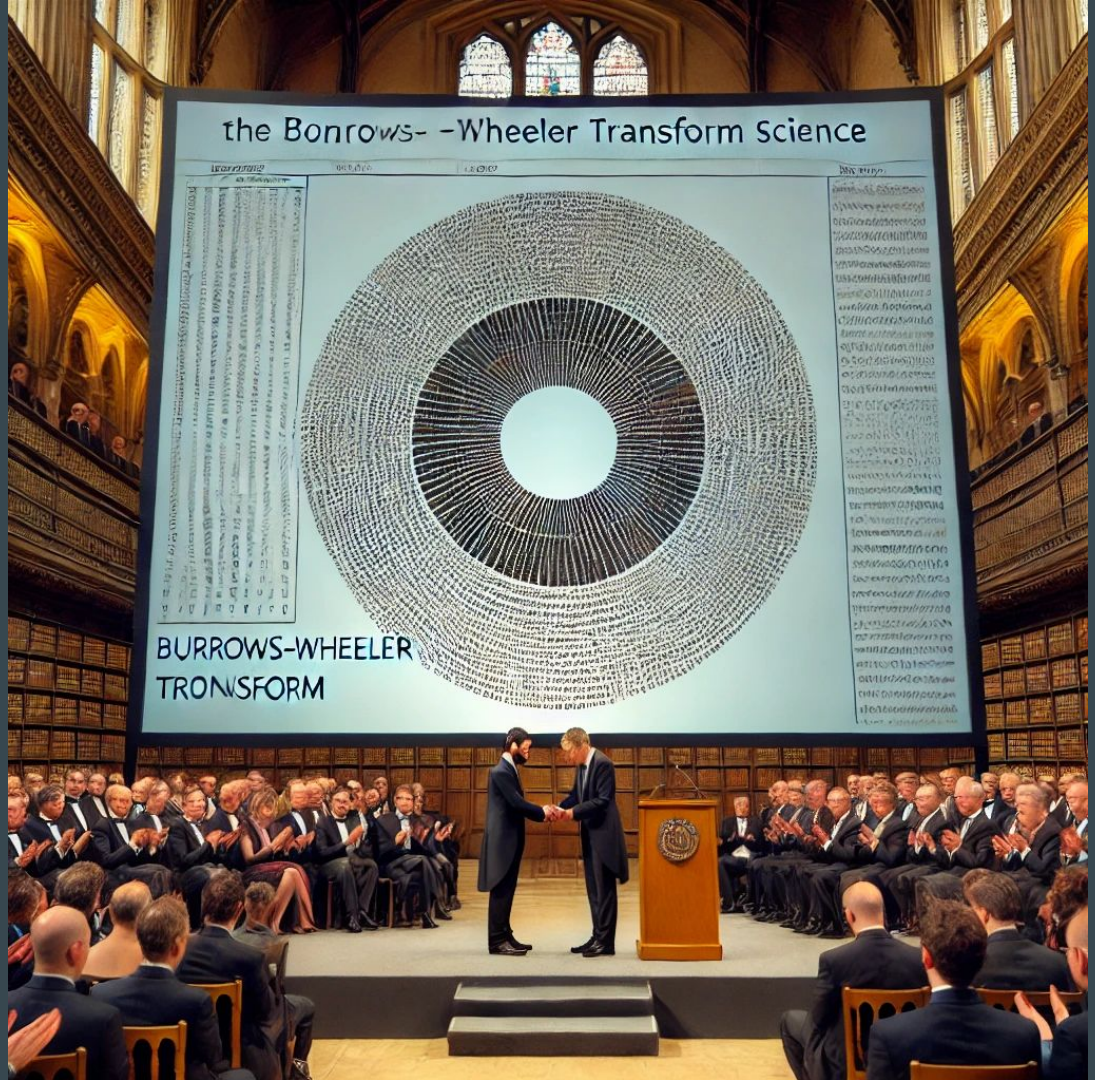
we keep going..

The magic: all prefixes of the original text can be reconstructed from the last column.

ic......		cs......		cs......i		ics......		csvomic
se......	sort	ev......	lastcol	ev......s	rotate	sev......		evomics
mi......	->	ic......	->	ic......m	->	mic......	-> .... ->	icsevom
om......		mi......		mi......o		omi......		micsevo
vo......		om......		om......v		vom......		omicsev
cs......		se......		se......c		cse......		sevomic
ev......		vo......		vo......e		evo......		vomicse

Search for a short read -> “reconstruct” just a short prefix

BWT should have gotten a Nobel Prize
(if there was one for CS)



(AI drawings are still terrible in 2025)

FM-Index and Burrows-Wheeler, a 20,000-feet view

Suffix tree: a tree stores all suffixes, older technique

Burrows-Wheeler transform : last column of sorted rotations of reference (what we just saw)

FM-index : set of tricks to quickly search inside the BWT without reconstructing the original text

How does Bowtie2 work?

Specializes in aligning Illumina reads to genomes.

1) Find seeds using FM-index, typically 20 nt length, up to 1 mismatch
2) Prioritizes seeds to further align
3) Extend seeds using SW-like algorithm

(that's it)

Minimizers

Minimap2 and strobealign use minimizers as seeds, then SW extension.

Minimizers: select only ***some*** k-mers as seeds

```
reference:    CTAAAAAGGTCA..
2nd window:    TAAAAAGG
               TAAAA
         seed:  AAAAA
                 AAAAG
                  AAAGG
```

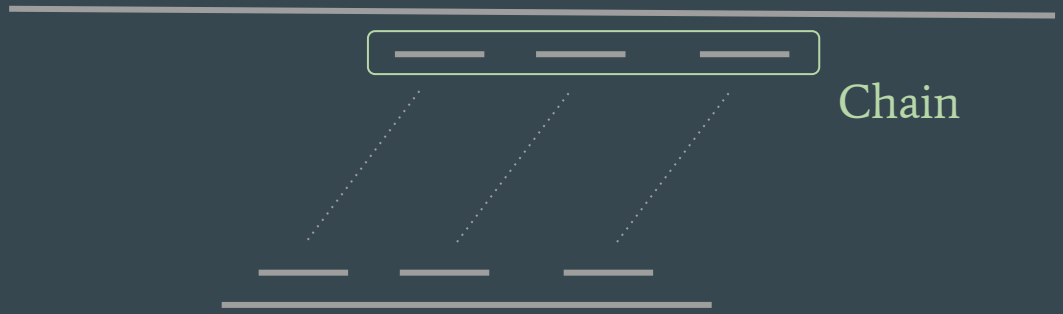
all reference k-mers	Found at position(s)
AAAAA	10, 65, 147, …
AAAAC	80
….	
AAAGG	none
….	
TAAAA	49, 101

Which “some”? Slide a window over the reference, and pick the (lexicographically) smallest seed within that window. Do that for all windows

Chains

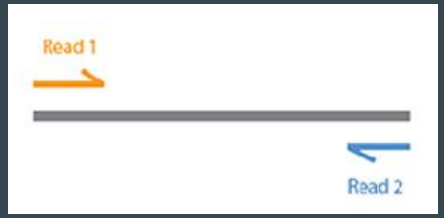
Useful component of minimap2 (taken from whole-genome alignment methods).

-> Before aligning, look for long enough co-linear chains of close seeds.



Paired reads

In some cases, Illumina sequencers output pairs of reads.



Aligners consider both reads jointly to improve precision. Need to specify:

- Orientation (forward-reverse is most common)
- Format: interleaved in one file, or two separate files

Mapping quality

...is your best friend, to avoid errors downstreams.

Mapq: how confidently each read is mapped (in log probability).

Grab only highly-confident alignments: `samtools view -q 60 [file.bam]`

Grab all alignments except trash ones: `samtools view -q 1 [file.bam]`

: `samtools view [file.bam]`

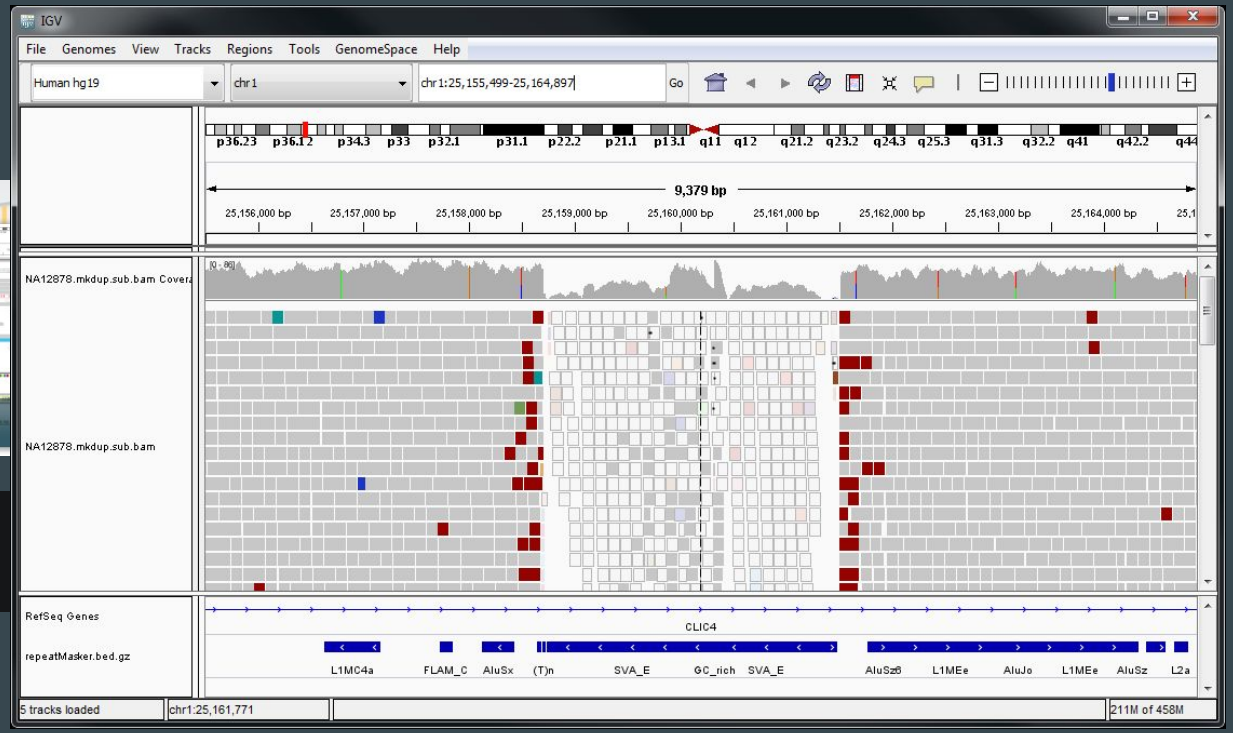
“Mapping” vs “Alignment”

In my view:

- **Mapping** : output where each read maps. That’s it.
- **Alignment** : do that, but also output how all bases line up (CIGAR).

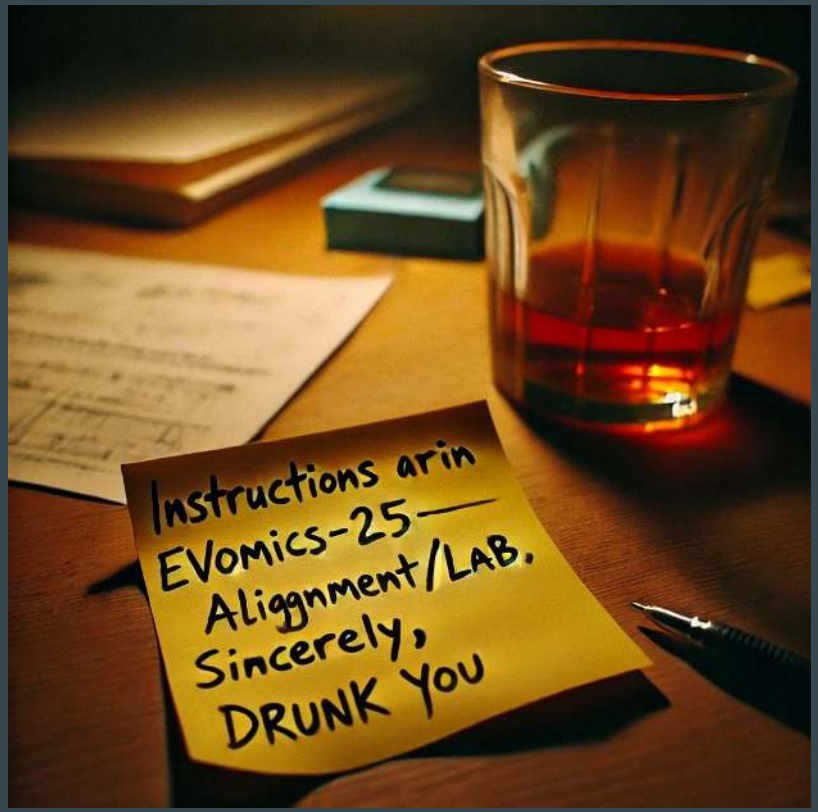
“`minimap2`” vs “`minimap2 -c`” (or `-a`)

Visualization of alignments



Reference and BAM need to be indexed, use samtools

Short read alignment demo



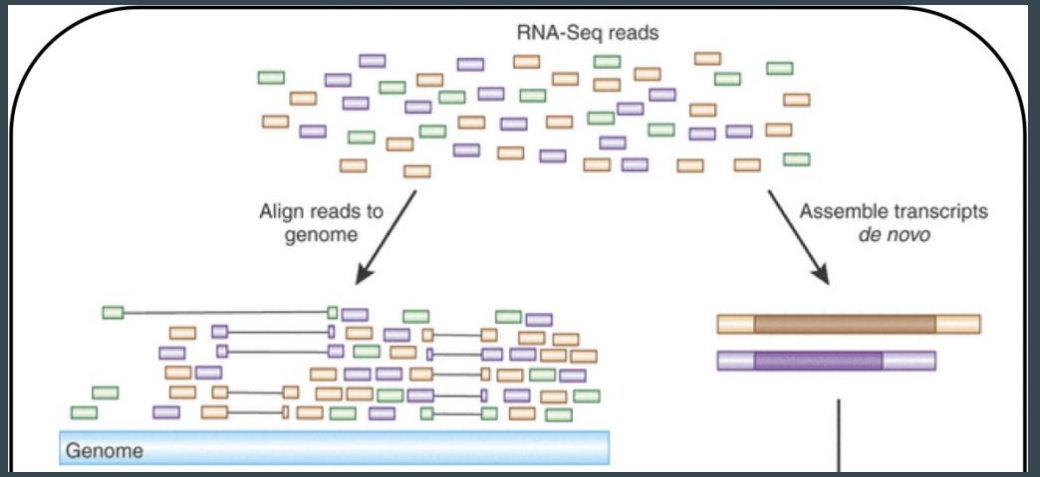
RNA

RNA read alignment is very similar to DNA, except:

- Split mapping (on genomes) due to splicing
- Ambiguity (on transcriptomes) due to many isoforms

Tools:

- Kallisto, Salmon
- STAR, HiSAT2



Long read mapping

Similar in spirit to short read mapping, but different tools.

PacBio CLR / ONT:

- Minimap2
- Variants of minimap2 for ~ 2-5x speed gain (mm2-fast, BLEND, ..)

PacBio HiFi:

- Minimap2
- Winnowmap2 (better accuracy)
- Mapquik (30x faster mapping, but no alignment)


SHORT READ MAPPING WITH PAIRED READS, SPACED SEEDS, FM-INDEX, E-M MAPPABILITY TRACKS, REPEAT MASKING, ..
LONG READ MAPPING WITH MINIMIZER CHAINING
imgflip.com

Pairwise DNA

...

Long sequences versus **long** sequences

Tools

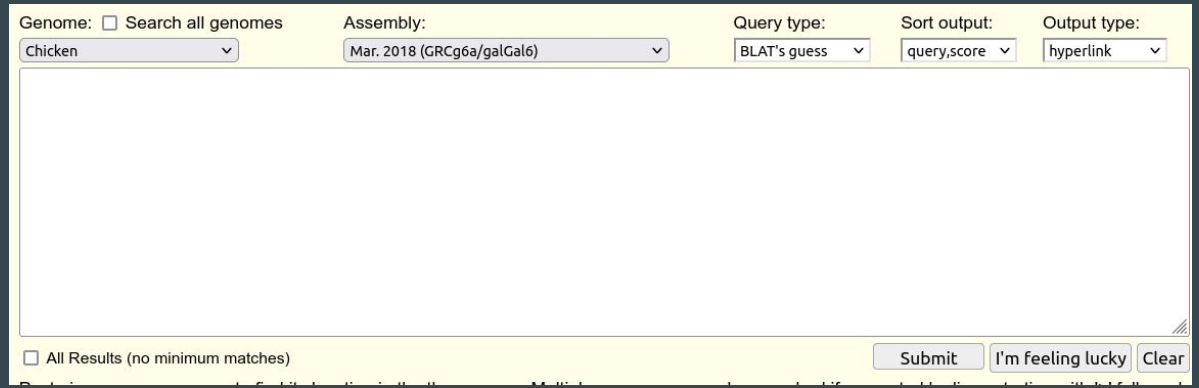
- BLAT
- Exonerate
- LASTZ
- MUMmer
- minimap2
- wfmash
- FASTGA

Mummer demo

BLAT

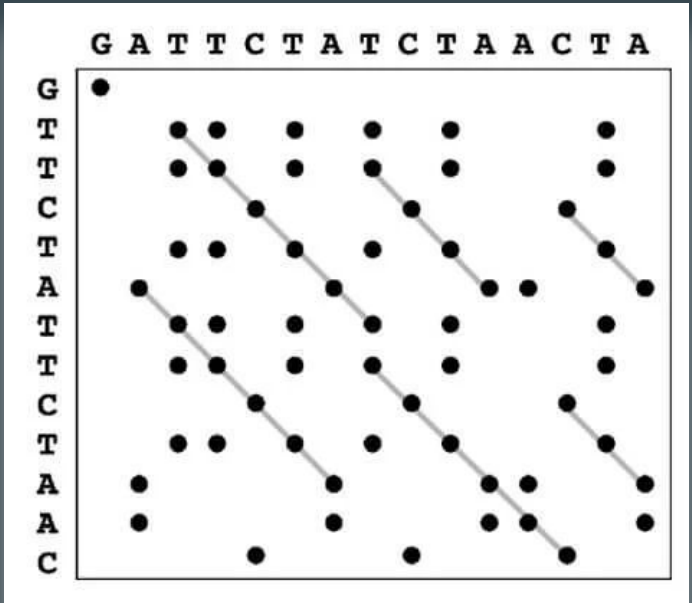
Close but not quite BLAST.
Differences:

1) Sequence-vs-genome (BLAT), instead of sequence-vs-database (BLAST)
2) Only find hits with >= 95% identity, over >= 40 bases
3) Faster than BLAST, integrated into UCSC Genome Browser

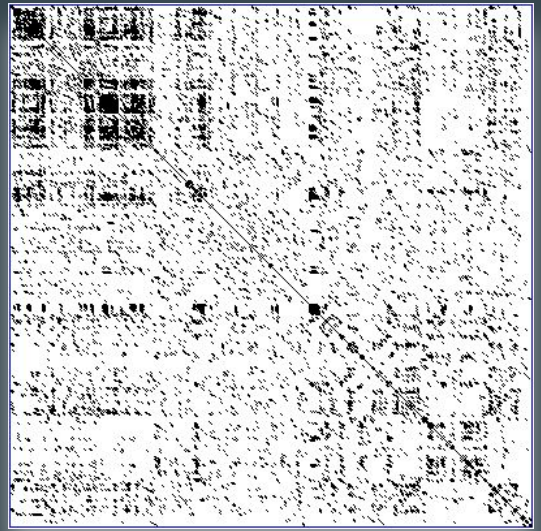


https://genome.ucsc.edu/FAQ/FAQblat.html

Dotplots



https://microbenotes.com/local-global-multiple-sequence-alignment/



Wikipedia

Tools: LASTZ, D-Genies, yass, MUMmer, **ModDotPlot**

Reciprocal best hits

A strange technique for e.g. finding orthologs.

If:

1) top alignment of gene A in species X **is** gene B in species Y

and

2) top alignment of gene B in species Y **is** gene A in species X

then genes A and B are RBH.

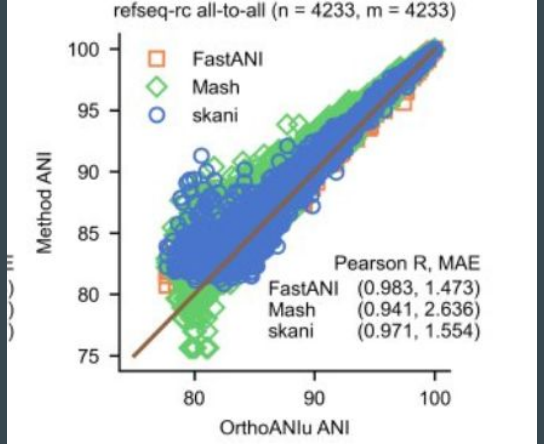
ANI (average nucleotide identity)

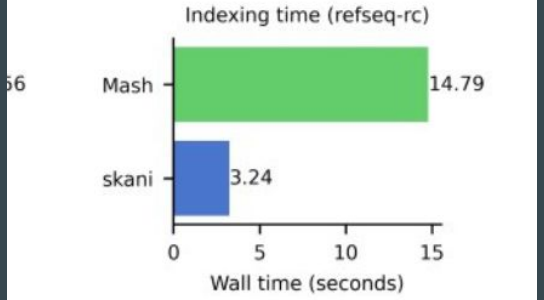
A strange identity metric, used to compare two bacterial genomes:

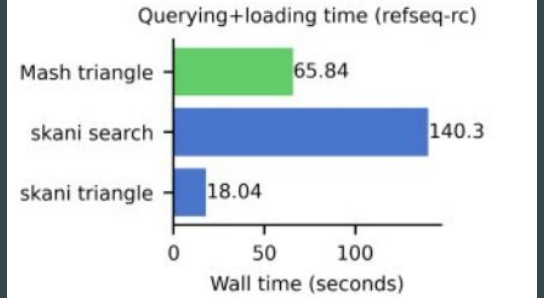
1. Extract many 1 Kbp fragments from query
2. ANI = mean identity of the reciprocal best hits

(from FastANI: https://www.nature.com/articles/s41467-018-07641-9)

Fast method: skani
https://twitter.com/jim_elevator/status/1616835999031611394







Minimap2 parameters to keep an eye on

`-a` (SAM) or `-c` (PAF) to really align,

`-x[mode]` controls mapping modes:

```
- map-pb/map-ont - PacBio CLR/Nanopore vs ref
- map-hifi - PacBio HiFi reads vs ref
- ava-pb/ava-ont - PacBio/Nanopore read overlap
- asm5/asm10/asm20 - asm-to-ref, for ~0.1/1/5% seq div
- splice/splice:hq - long-read spliced alignment
- sr - genomic short-read mapping
```

Pairwise DNA

...

Short sequences versus short sequences

Nobody really does that any more
Genome Assembly has better techniques (e.g. de Bruijn graphs)

Pointers

minimap (then miniasm)

StarCode https://academic.oup.com/bioinformatics/article/31/12/1913/213875

SlideSort https://github.com/iskana/SlideSort

PAF file format

1 nucl sequence versus a database ...


The research lab went from one end
of her genetic sequence to the other.

Tools

BLASTn

MetaGraph, Pebblescout

Kraken

LexicMap, Phylign

See the Big Data lecture!

BLAST databases: nr

*“The nucleotide collection consists of **GenBank**+EMBL+DDBJ+PDB+RefSeq sequences, but excludes EST, STS, GSS, WGS, TSA”*

[..] “The database is non-redundant.”

125 GB compressed

ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz

Limits of BLAST

- Can’t search all known genomes, only those in the BLAST database
- Under 85% identity, alignments tend to be missed

Pairwise protein

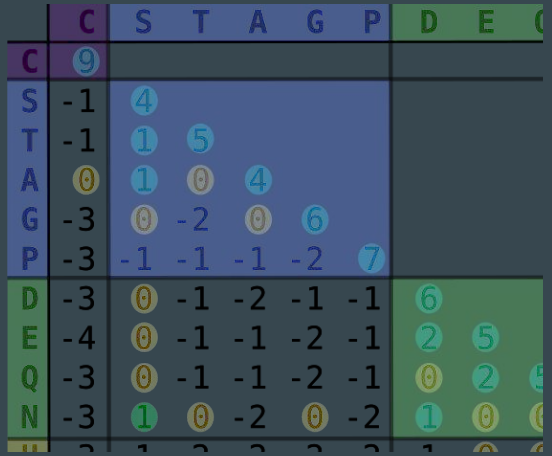
...

What changes compared to pairwise DNA?

- Different alphabet, shorter sequences
- Some substitutions are more likely than others
  - BLOSUM

Applications:

- Low-homology search (high evolutionary distances)



	C	S	T	A	G	P	D	E
C	9							
S	-1	4						
T	-1	1	5					
A	0	1	0	4				
G	-3	0	-2	0	6			
P	-3	-1	-1	-1	-2	7		
D	-3	0	-1	-2	-1	-1	6	
E	-4	0	-1	-1	-2	-1	2	5
Q	-3	0	-1	-1	-2	-1	0	2
N	-3	1	0	-2	0	-2	1	0

Some words of caution

"Alignment scoring schemes are hilariously ***over-simplified model of real evolution*** *[..] treat all alignments with large pinch of salt [..] dynamic programming is 'exact' only to an ivory-tower computer scientist"*

- Robert Edgar (computer scientist)

There is no such thing as "the alignment" between two protein sequences.

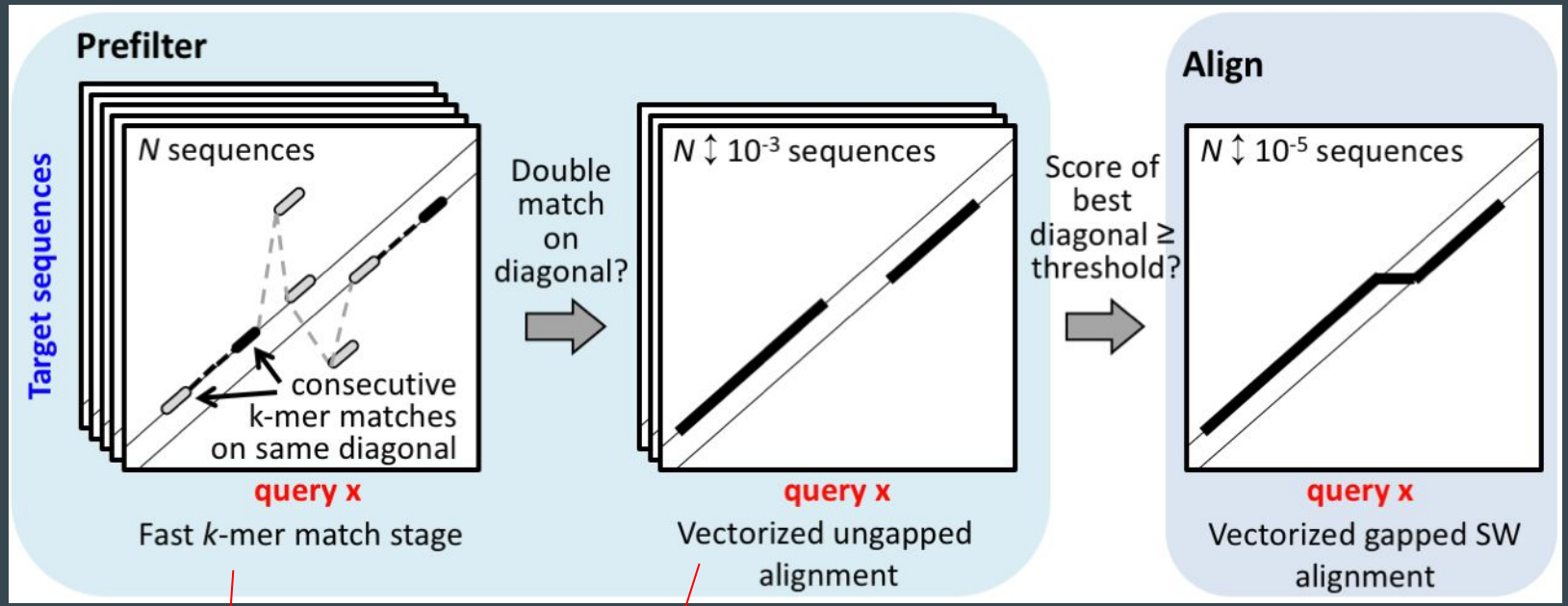
Tools

MMseqs2

DIAMOND2

BLASTp

How mmseqs2 work: mmseqs search



https://github.com/soedinglab/mmseqs2/wiki#description-of-workflows

Seed finding

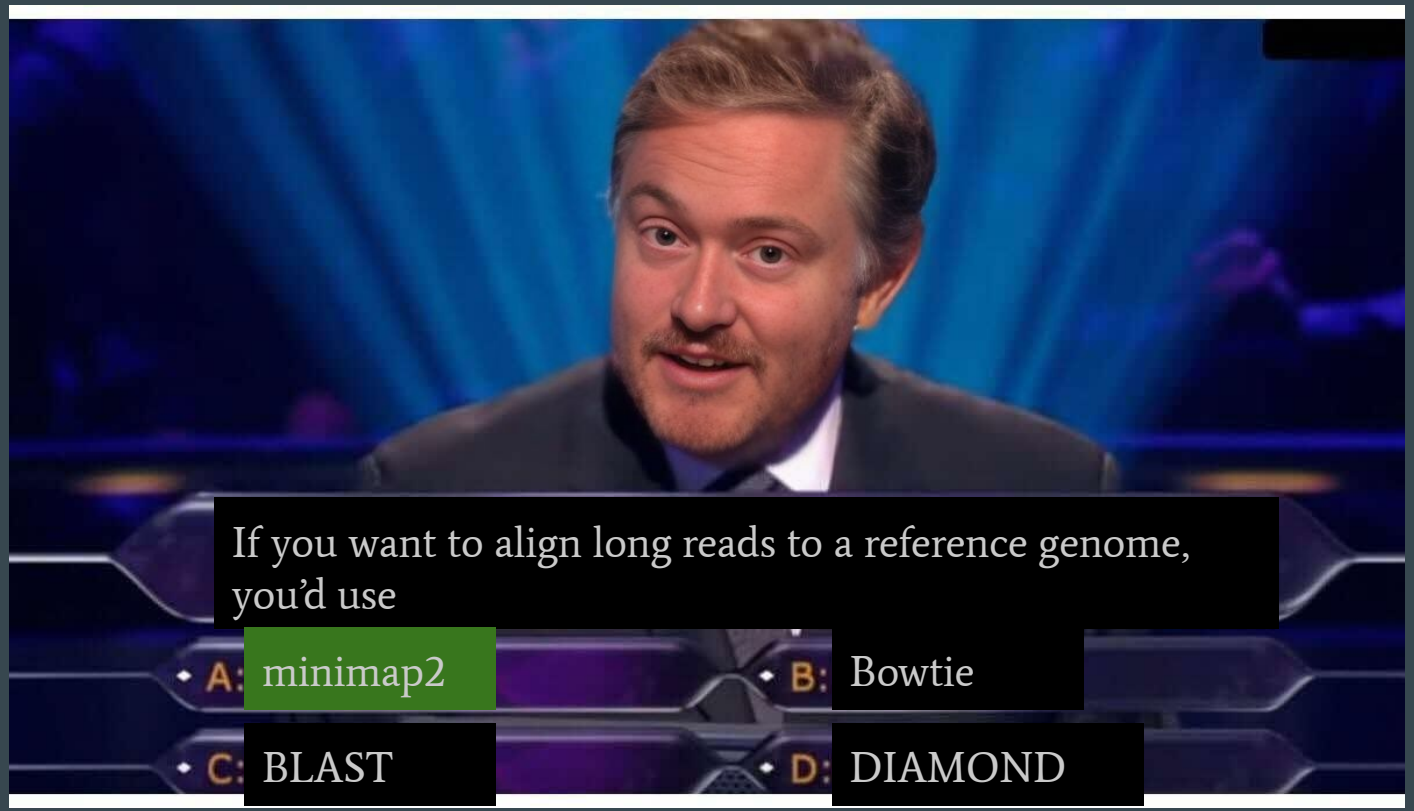
Seed chaining

How DIAMOND work:

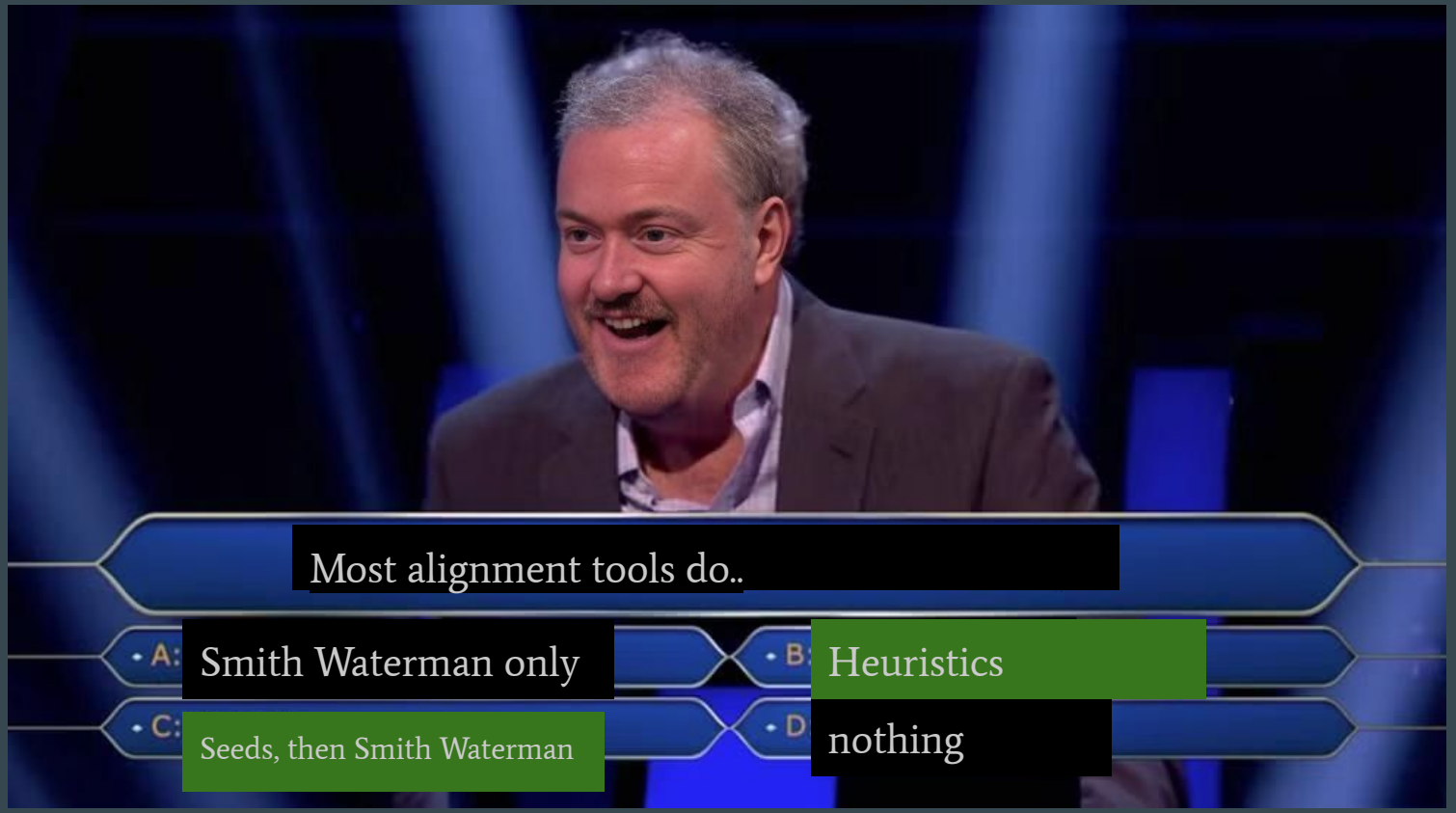
“[..] A simple exact match criterion determines which seeds are passed on to the extension phase, in which a Smith-Waterman alignment is computed.”

https://www.nature.com/articles/nmeth.3176

Quiz time!


If you want to align long reads to a reference genome, you'd use
A: minimap2
B: Bowtie
C: BLAST
D: DIAMOND

Quiz time!


Most alignment tools do..
A: Smith Waterman only
B: Heuristics
C: Seeds, then Smith Waterman
D: nothing

Multiple, protein?

...

What it looks like

Input: *n* sequences

```
ACATGA

ACGTG

CATTA
```

Output: aligned sequences, with indels

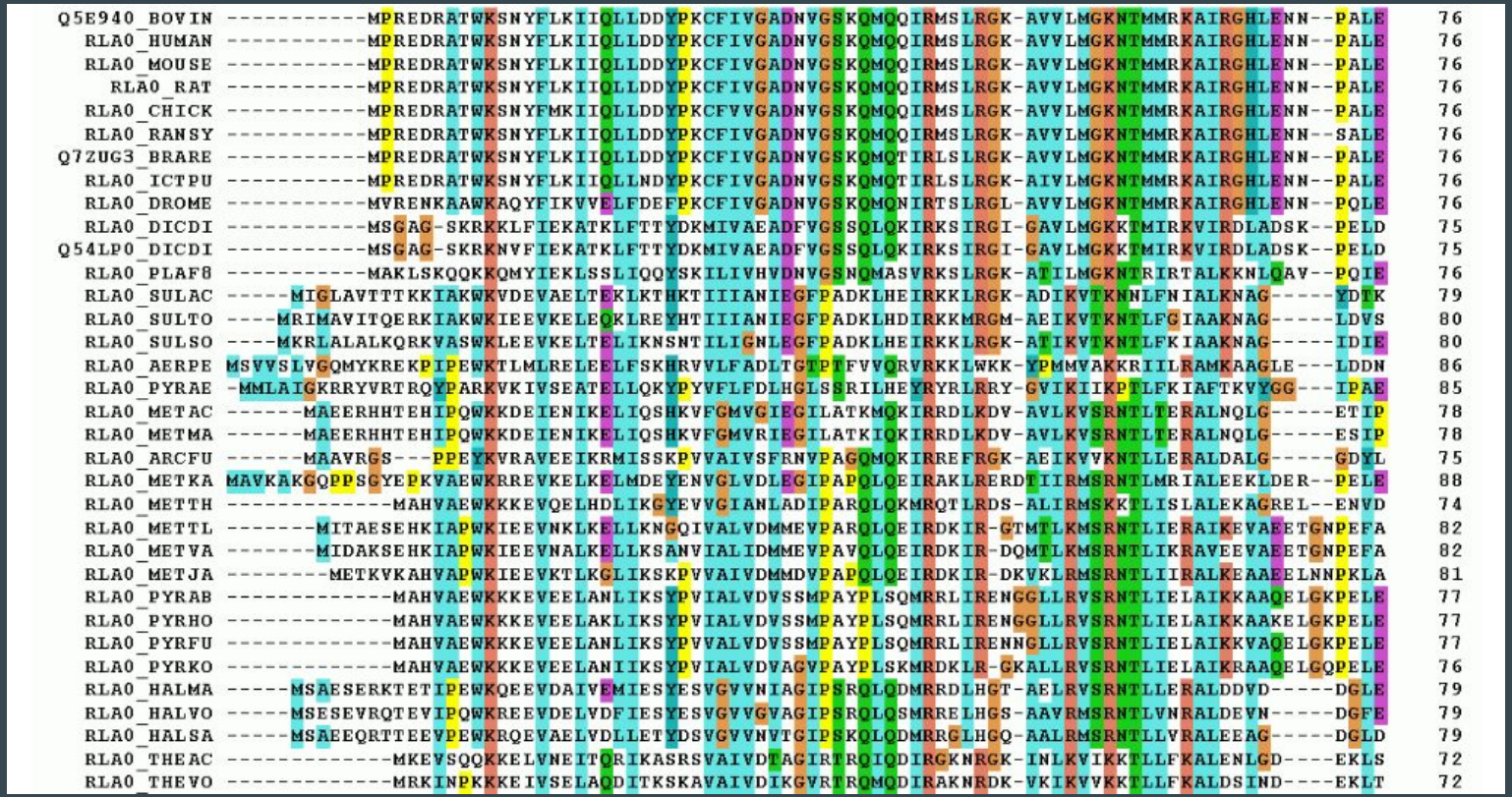
```
ACATGA

AC**G**TG-

-CAT**T**A
```

In practice..

Colors = equivalent residues (structurally or functionally)



Why do multiple alignment?

- Comparative genomics
- Phylogeny
- Protein structure prediction
- RNA structure and function
- ...

How is a MSA scored?

"**Sum-of-pairs** " (SP) score:

1) Fix a scoring scheme, e.g. match=1, mismatch=-1, indel=-2.
2) For each column, for all pairs of residues, compute score
3) Sum scores across columns

```
Column:  123456
         ACATGA
         ACG-G-
         -CAGTA
```

For column 4: score(T,-) + score(T,G) + score(-,G) = -2 + -1 + -2 = -5.

For column 5: score(G,G) + score(G,T) + score(G,T) = 1 + -1 + -1 = -1.

https://www.cs.princeton.edu/~mona/Lecture/msa1.pdf

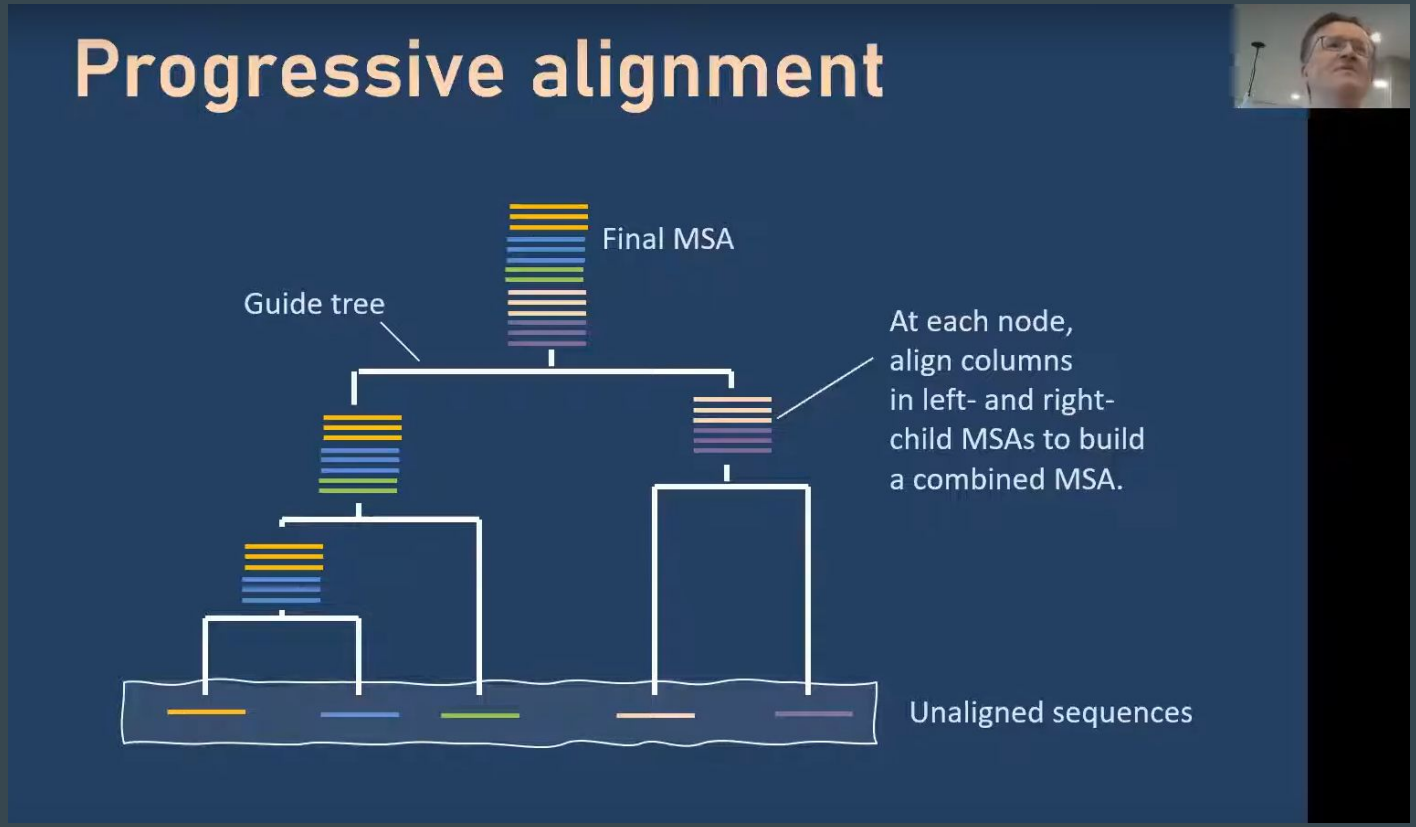
Optimal MSA

Remember Needleman-Wunsch?

Same, but with more possibilities.

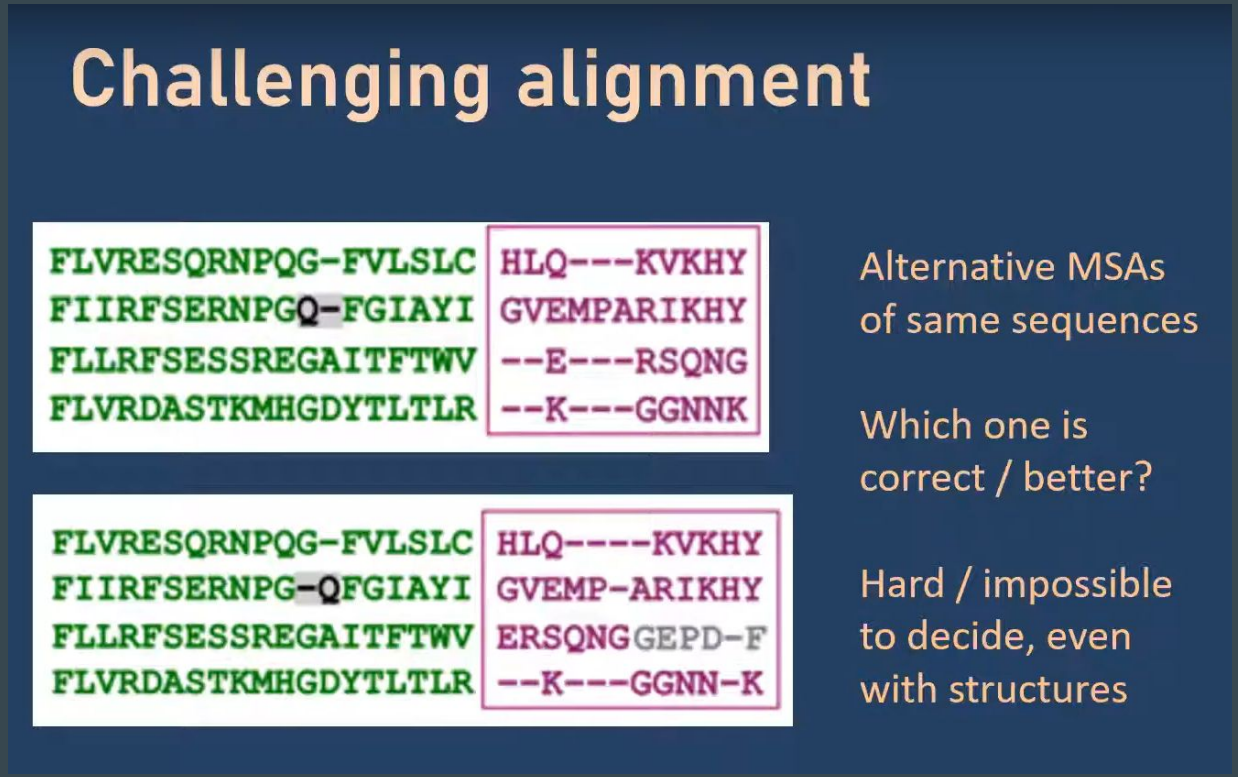
So, best avoided.

Progressive MSA



https://www.youtube.com/watch?v=2HmjHStpu7I

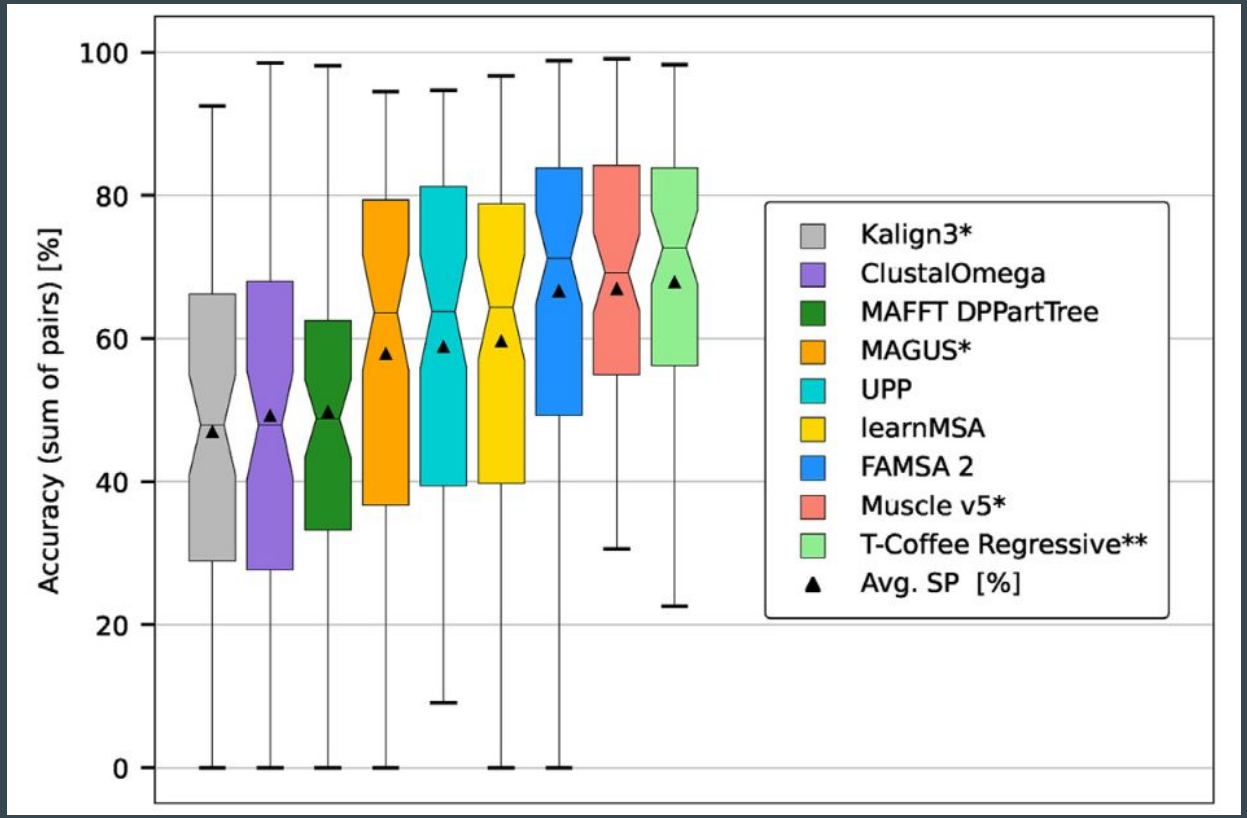
MSA is on another level of difficulty



https://www.youtube.com/watch?v=2HmjHStpu7I

Tools

- MUSCLE
- ClustalW
- T-Coffee
- MAFFT
- ...



https://www.sciencedirect.com/science/article/pii/S0959440X23000519

Multiple, DNA

...

What changes compared to protein MSA?

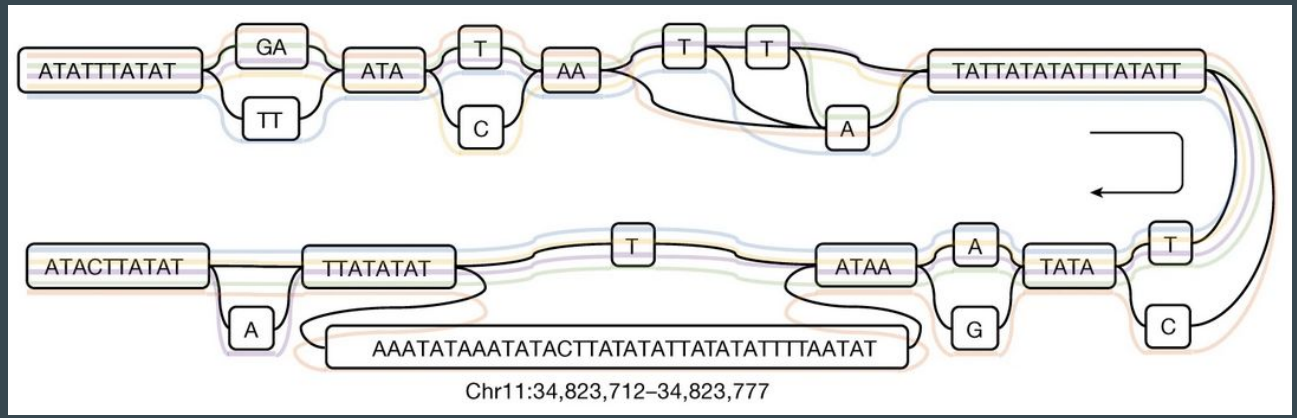
- Wayyy longer sequences
- Duplications, inversions, and translocations wreak linearity

Tools

- SibeliaZ
- Cactus

State of the art: human genome graphs, look for pangenomics papers.

e.g. HPRC: https://www.nature.com/articles/s41586-023-05896-x, CPC https://www.nature.com/articles/s41586-023-06173-7



1 sequence versus a profile

...

PSSMs, HMMs

Is there enough time to present this?!

Position Specific Scoring Matrices (PSSM) **and** Hidden Markov Models (HMM)

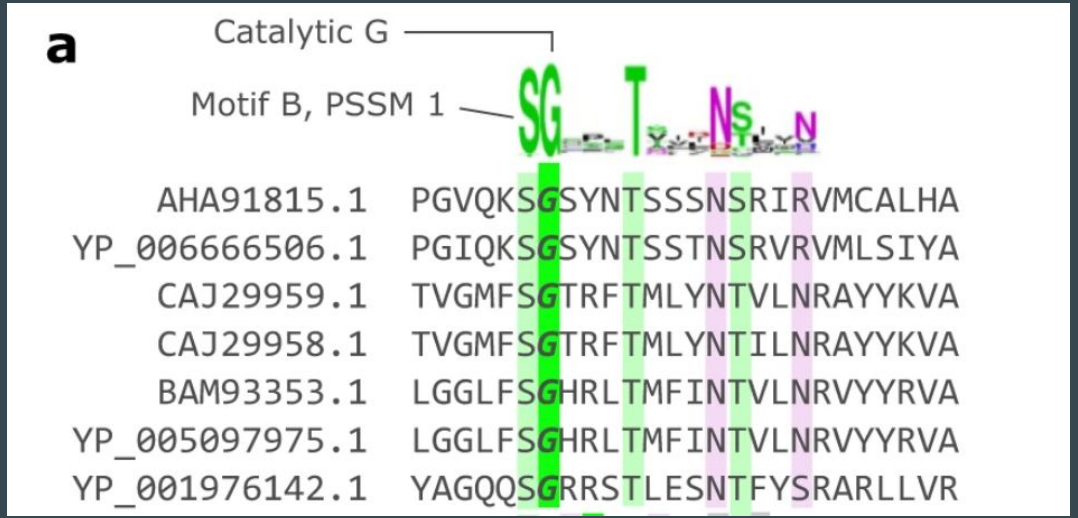
Not quite alignment, but:

“Does this sequence belong to a particular family?”

PSSM

Way to represent families of sequences, **with no gaps.**

1) Construct MSA
2) Determine frequency per column



https://www.nature.com/articles/s41467-022-34630-w/figures/9

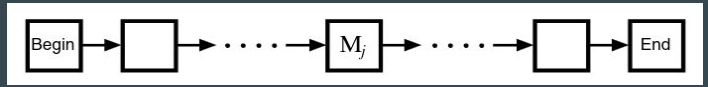
HMM

Hidden Markov Models generalize PSSMs with gaps.

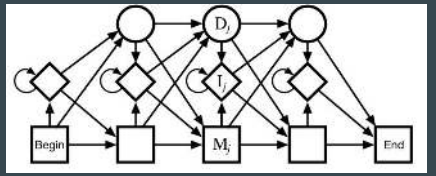
Motivation: when pairwise fails

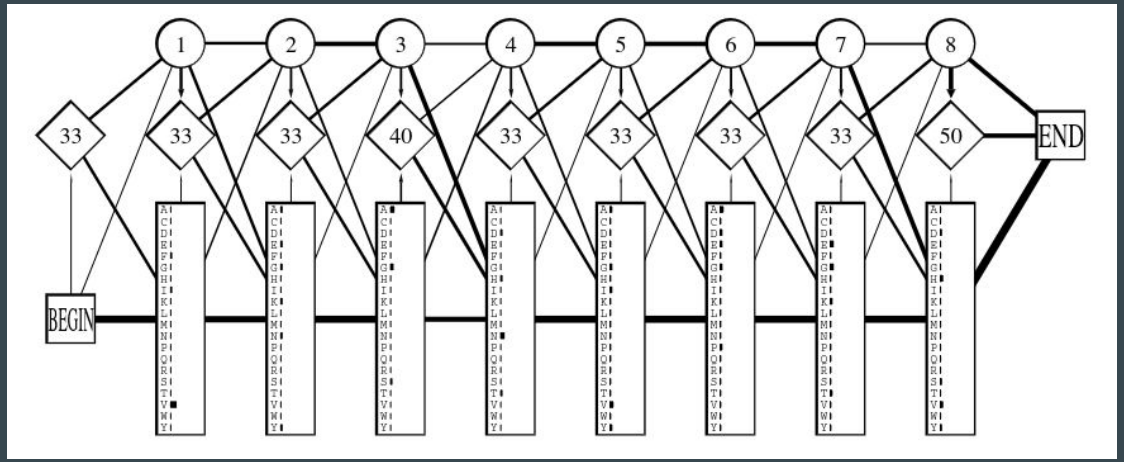
```
HBA_HUMAN    ...VGA--HAGEY...
HBB_HUMAN    ...V----NVDEV...
MYG_PHYCA    ...VEA--DVAGH...
GLB3_CHITP   ...VKG------D...
GLB5_PETMA   ...VYS--TYETS...
LGB2_LUPLU   ...FNA--NIPKH...
GLB1_GLYDI   ...IAGADNGAGV...
                ***  *****
```

HMM of a PSSM:



Profile HMM:





http://www.mcb111.org/w06/durbin_book.pdf

Tools

HMMer

MMseqs profile

HHblits

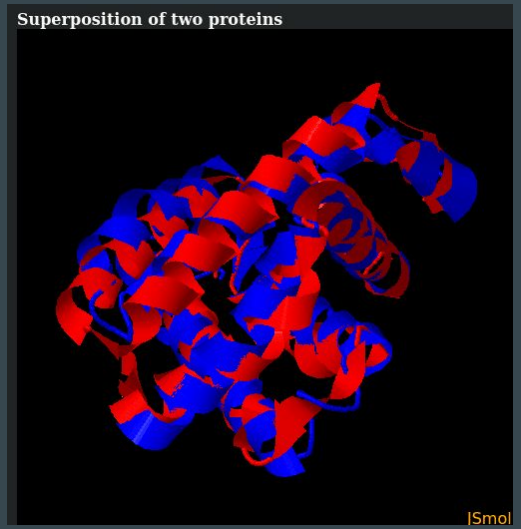
Bonus: structural alignment (TM-align)

```
(":" denotes aligned residue pairs of d < 5.0 A)

MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRVKHLKTEAEMKASEDLKKHGVTVLTALGAILKKK--G-HHEAELKPLAQS
 ::::::::::::::::::::::::::::::::::::::::::::::::::: :::::::::::::::::::::::::::   : ::::::::::::
-SLSAAEADLAGKSWAPVFANKNANGLDFLVALFEKFPDSANFFADFKGKS-VADIKASPKLRDVSSRIFTRLNEFVNNAANAGKMSAMLSQFAKE
```

Input: 2 PDB structures

Output: aligned residues, and a TM -score
(> 0.5 = same fold)



Max TM -score:
0.85377

Personal take

As databases of genomes grow, alignment will both become easier and harder.

Solved:

- Human read alignment (DNA, RNA)
- High-identity to current genome databases
- Small-data HMMs

Unsolved:

- Genome-scale MSA
- Ancient DNA
- Large MSAs
- Big-data HMMs
- Sequences to peta-scale databases

What we've seen

- Pairwise DNA alignment
  - CIGAR strings
  - Scoring
  - Needleman-Wunsch
  - Smith-Waterman
  - BLAST
  - BLAT, minimap2
- Short read mapping
  - Burrows-Wheeler transform
  - Minimizers
  - Bowtie2, BWA, minimap2, Strobealign
- Pairwise protein alignment
  - Diamond, mmseqs2
- MSA
- HMMs

Thank you for your attention!