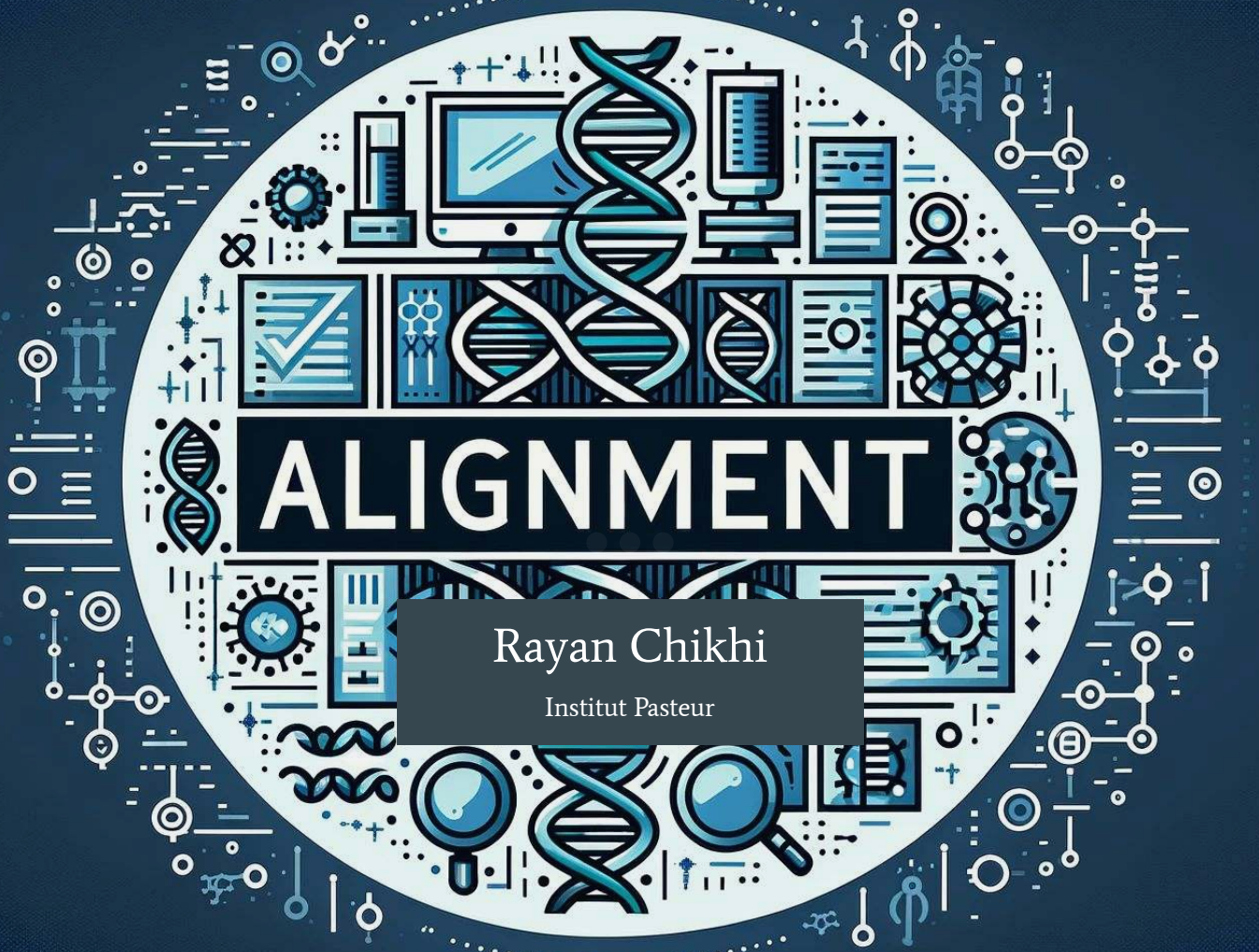




ALIGNMENT

Rayan Chikhi

Institut Pasteur

Hello again!

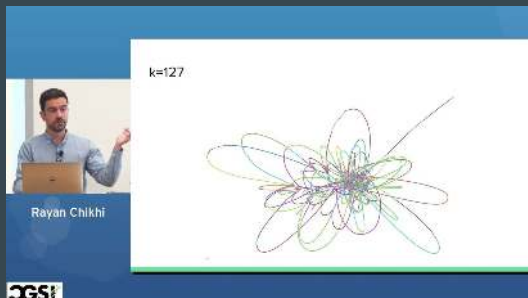
- I'm a researcher in bioinformatics algorithms
- *de novo* assembly, big data alignment, k-mers, pangenomics. Well, week 1 stuff :)

 @RayanChikhi on Twitter

<http://rayan.chikhi.name>

Sequence
Bioinformatics

INSTITUT
PASTEUR



Course objectives

- **Enough background** to understand the alignment part of a biology article
- Increase confidence in using alignment tools
- Understand **why** alignment isn't so straightforward

Course outline

- **Fundamentals**
- The many **flavors** and **tools** for pairwise DNA alignment
- **Multiple** sequence alignment
- Alignment to **databases**
- Into the unknown: profile and structure search

Questions to the audience

1. Have you ever **run** a sequence alignment software?
2. Was it **willfully** 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 residues 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			3531
Query	361	-----ATCGGCCTGTCGCTTGCGGTATTCGGAATCTTGACGCCCTCGCTCAAGCC			411
Sbjct	3529	GCGACGATGATCGGCCTGTCGCTTGCGGTATTCGGAATCTTGACGCCCTCGCTCAAGCC			3471

<https://techcommunity.microsoft.com/t5/azure-high-performance-computing/running-ncbi-blast-on-azure-performance-scalability-and-best-practices/p2410483>

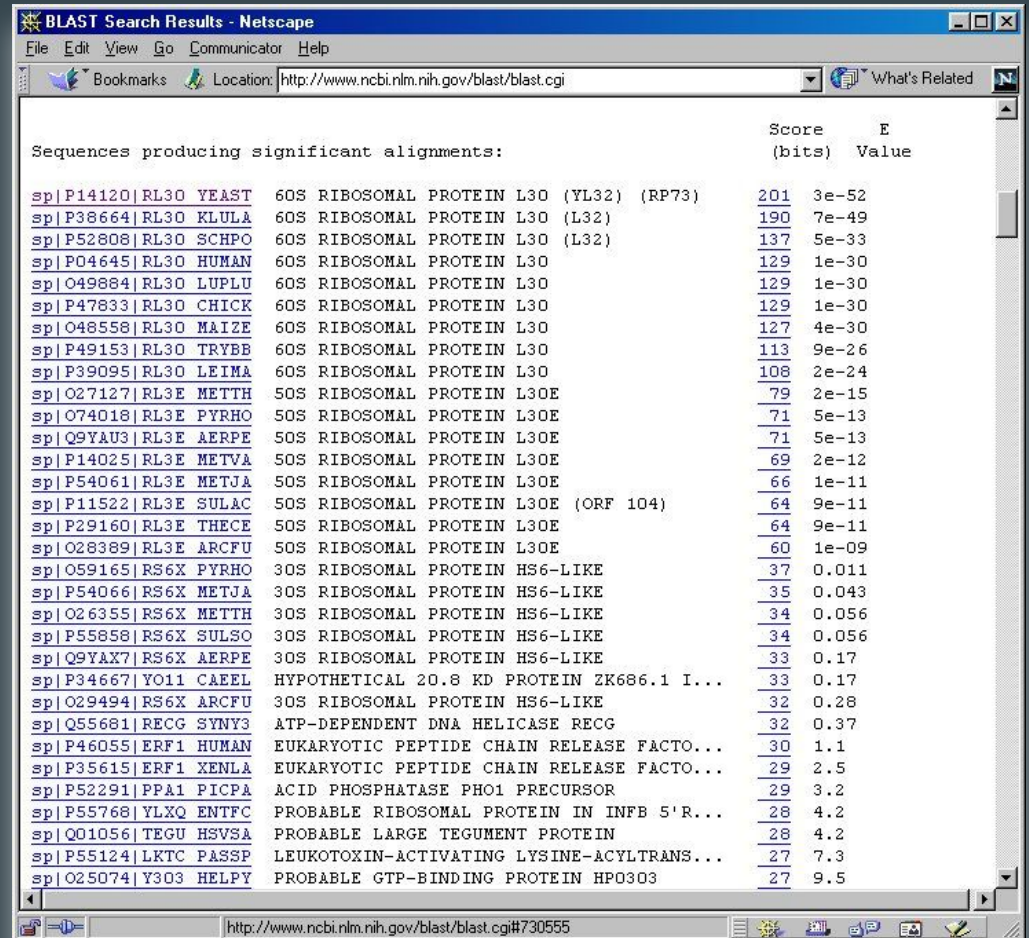
Multiple sequences (>2)

Q5E940 BOVIN -----MPREDRATWKSNYFLKTIQLDLLDPVKCFIVGADVNGSKMOQIIRMSLRGK-AVVLHGKNTMMRKAIRGHLENN--PALE 76
RLA0 HUMIN -----MPREDRATWKSNYFLKTIQLDLLDPVKCFIVGADVNGSKMOQIIRMSLRGK-AVVLHGKNTMMRKAIRGHLENN--PALE 76
RLA0 MOUSE -----MPREDRATWKSNYFLKTIQLDLLDPVKCFIVGADVNGSKMOQIIRMSLRGK-AVVLHGKNTMMRKAIRGHLENN--PALE 76
RLA0 RAT -----MPREDRATWKSNYFLKTIQLDLLDPVKCFIVGADVNGSKMOQIIRMSLRGK-AVVLHGKNTMMRKAIRGHLENN--PALE 76
RLA0 CHICK -----MPREDRATWKSNYFMKTIQLDLLDPVKCFVGVADVNGSKMOQIIRMSLRGK-AVVLHGKNTMMRKAIRGHLENN--PALE 76
RLA0 RANSY -----MPREDRATWKSNYFLKTIQLDLLDPVKCFIVGADVNGSKMOQIIRMSLRGK-AVVLHGKNTMMRKAIRGHLENN--SALE 76
Q7ZUG3 BCRATE -----MPREDRATWKSNYFLKTIQLDLLDPVKCFIVGADVNGSKMOTIRLSLGK-AVVLHGKNTMMRKAIRGHLENN--PALE 76
RLA0 ICTPU -----MPREDRATWKSNYFLKTIQLLLNDYVKCFIVGADVNGSKMOTIRLSLGK-AVVLHGKNTMMRKAIRGHLENN--PALE 76
RLA0 DROME -----MYRENKAAMQAQYIFIKVVELDFEFCFIVGADVNGSKMOMIRTSIRGL-AVVLHGKNTMMRKAIRGHLENN--POLE 76
RLA0 DICDI -----MSGAG-SRRKKLFTEKATKLFTTIDKMIIVAEAFVGSLSOLKIRKSIRGI-GAVLHGKMTIRKVIKRLADSLD--PELD 75
Q541P0 DICDI -----MSGAG-SRRKNVMFTEKATKLFTTIDKMIIVAEAFVGSLSOLKIRKSIRGI-GAVLHGKMTIRKVIKRLADSLD--PELD 75
RLA0 PLAFB -----MKLKSQOQKMYTEKLSLSLQIKLIVHVNVSNGMASVRSYSLRGK-AILHGKTRIRTRALTKNKLIQV--PQTE 76
RLA0 SULAC -----HILAVTTTQKIATKKVYDEVAELTKLTKHTITIANIGFPADKLHEIRKKLRGK-ADIKVIXKNLFNIALKNAG--YDK 79
RLA0 SULTO -----MRHMAVITQRKIAKKWIEVKELKREHTITIANIGFPADKLHEIRKKMRGM-AEIKVKNITFLFKIAANKAG--LDYS 80
RLA0 SULSO -----MKRLALATKQRKVAKSLKEELKELTKLNSHTILQNLGFPADKLHEIRKKLRGK-AEIKVKNITFLFKIAANKAG--IDIE 80
RLA0 AERPE **MSVSVSLVG**QMYRKKE**IP**EWKTLMLRELELFSKRHVLFDLTGTPFVYGVRRKKLWKK-PPVMYAKKRILITAMKAAGLE-LDDN 86
RLA0 PYRAE **HMLAIG**KRRYVRTRO**Y**PAKRVYI**SE**ATELLQKVYVYFDFDLHGLSRILHEYRYRLRYR**G**-GVIRIKITFLFKIAFTKVYGG--TPAE 85
RLA0 METAC -----MAEERHHTHILPOWKKDEIENIKELIQSHKVF**GM**VYIEGILATKMKQIRRDLDKV-AVLKYSRNTLIERALNQLG--ETIP 78
RLA0 METMA -----MAEERHHTHILPOWKDEIENIKELIQSHKVF**GM**VRIEGILATKMKQIRRDLDKV-AVLKYSRNTLIERALNQLG--ESIP 78
RLA0 ARCFA **MAA**VYRG**S**-----PPCYVTRAVVEITKRMSSKVPVAVSFRNPVAGOMQIRREFRG-AEIKVKNITFLRALDALG--GDYL 78
RLA0 METKA **MAV**KK**AG**MAVYRG**S**PKVAWKRREYKELKLMDEVNGLVDLEGIPALPOLIEIRAKRLRDEITIRMSRNTLIERALEEKIDER--PELE 85
RLA0 METH -----MAHVAEWKKKVEQELHDLIKELVGVIGIANLADIPARLOKMQRTLRD-ALIRMSKNTLISLAEKAGREL--ENVD 74
RLA0 METTL -----MITAESHKIADPKIEEYNKLKELKNGQIVALVDMMEVPAQLOEIRDKIRI-TIMLKMSRNTLIERALEKAEVATEGNEPFA 82
RLA0 METVA -----MIDAKSEHIAIPKWIEEYNALKELLSANVIALIDMMEVPVAVLOEIRDKIRI-DQMLKMSRNTLIRKAEVATEGNEPFA 82
RLA0 METJA -----METKYAHVADPKIEEYKTLKGLIKSKFPVAVVDMMDVPVAPQLOEIRDKIRI-DQVLMKMSRNTLIERALEKAEELINNPKLA 81
RLA0 PYRAB -----MAHVAEWKKKVEEELANLIXSVYVIALVDVSSMPAYPLSQMRRLIRENGGLLYRSRNTLIEAIAKKAAELGKPELE 77
RLA0 PYRHO -----MAHVAEWKKKVEEELAKLIXSVYVIALVDVSSMPAYPLSQMRRLIRENGGLLYRSRNTLIEAIAKKAAELGKPELE 77
RLA0 PYRFO -----MAHVAEWKKKVEEELANLIXSVYVIALVDVSSMPAYPLSQMRRLIRENGGLLYRSRNTLIEAIAKKAAELGKPELE 77
RLA0 PYRKO -----MAHVAEWKKKVEEELANLIXSVYVIALVDVAGPAYPLSQMRDKLIE-KALLYRSRNTLIEAIAKKAAELGKPELE 76
RLA0 HALMA -----MSAESEKRTTEIP**EW**KQVEYDAIV**IES**YESVGVVNIAGIP**RL**QLODMRRDLHG**T**-AELLYRSRNTLIERALDDVD--DGLE 79
RLA0 HALVO -----MSAESEQRTEVIP**EW**KQREEYDELVD**IES**YESVGVVGVAGIP**RL**QLOSMRRDLHG**S**-AAVRMSRNTLVNRLDEVN--DGFE 79
RLA0 HALSA -----MSAESEQRTEVIP**EW**KQREAVDLVOLLETDSVGVVNTG**IP**SKLODMRRGLHG**Q**-AALLRMSRNTLVRLALEAG--DGLD 79
RLA0 THEAC -----MKVESQQKQELVNEITIRIKASVSVAIVDAGITRITRODIRGKNRGK-INKVYKTLFLFKALENGLD--EKLS 72
RLA0 THEVO -----MRKINPKKKEIYSELADITKSKAVAVD**IK**GV**IE**TRMODIRAKNRDKYKIVYKKLFLFKALDSIND--EKLT 72
RLA0 PICTO -----MTEPA**Q**W**K**IDFYKNNLEENINSRKVAATYISIKGLRNNE**FI**KIRNS**IR**DK-ARIKY**SR**ARLLIRLATEN**IG**K--NNIV 72

ruler 1.....10.....20.....30.....40.....50.....60.....70.....80.....90

The many types of alignments

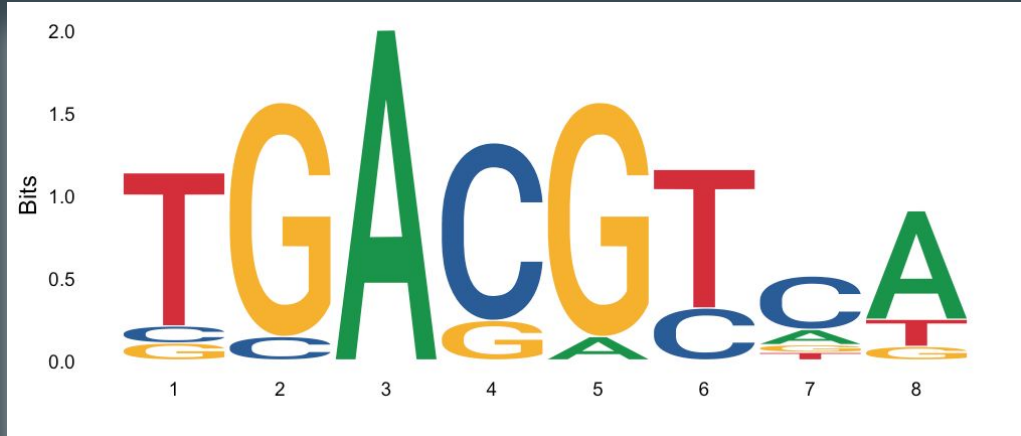
1 sequence versus a database



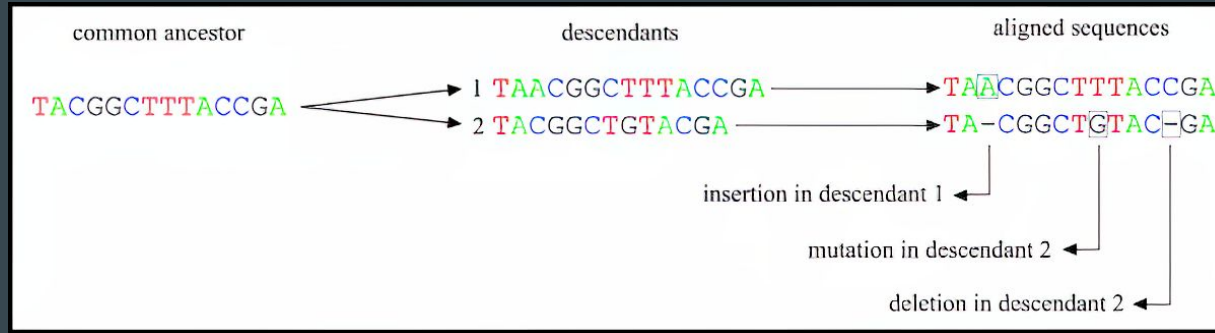
Sequences producing significant alignments:			Score (bits)	E Value
sp P14120 RL30	YEAST	60S RIBOSOMAL PROTEIN L30 (YL32) (RP73)	201	3e-52
sp P38664 RL30	KLULA	60S RIBOSOMAL PROTEIN L30 (L32)	190	7e-49
sp P52808 RL30	SCHPO	60S RIBOSOMAL PROTEIN L30 (L32)	137	5e-33
sp P04645 RL30	HUMAN	60S RIBOSOMAL PROTEIN L30	129	1e-30
sp O49884 RL30	LUPLU	60S RIBOSOMAL PROTEIN L30	129	1e-30
sp P47833 RL30	CHICK	60S RIBOSOMAL PROTEIN L30	129	1e-30
sp O48558 RL30	MAIZE	60S RIBOSOMAL PROTEIN L30	127	4e-30
sp P49153 RL30	TRYBB	60S RIBOSOMAL PROTEIN L30	113	9e-26
sp P39095 RL30	LEIMA	60S RIBOSOMAL PROTEIN L30	108	2e-24
sp O27127 RL3E	METTH	50S RIBOSOMAL PROTEIN L30E	79	2e-15
sp O74018 RL3E	PYRHO	50S RIBOSOMAL PROTEIN L30E	71	5e-13
sp Q9YAU3 RL3E	AERPE	50S RIBOSOMAL PROTEIN L30E	71	5e-13
sp P14025 RL3E	METVA	50S RIBOSOMAL PROTEIN L30E	69	2e-12
sp P54061 RL3E	METJA	50S RIBOSOMAL PROTEIN L30E	66	1e-11
sp P11522 RL3E	SULAC	50S RIBOSOMAL PROTEIN L30E (ORF 104)	64	9e-11
sp P29160 RL3E	THECE	50S RIBOSOMAL PROTEIN L30E	64	9e-11
sp O28389 RL3E	ARCFU	50S RIBOSOMAL PROTEIN L30E	60	1e-09
sp O59165 RS6X	PYRHO	30S RIBOSOMAL PROTEIN HS6-LIKE	37	0.011
sp P54066 RS6X	METJA	30S RIBOSOMAL PROTEIN HS6-LIKE	35	0.043
sp O26355 RS6X	METTH	30S RIBOSOMAL PROTEIN HS6-LIKE	34	0.056
sp P55858 RS6X	SULSO	30S RIBOSOMAL PROTEIN HS6-LIKE	34	0.056
sp Q9YAX7 RS6X	AERPE	30S RIBOSOMAL PROTEIN HS6-LIKE	33	0.17
sp P34667 YO11	CAEEL	HYPOTHETICAL 20.8 KD PROTEIN ZK686.1 I...	33	0.17
sp O29494 RS6X	ARCFU	30S RIBOSOMAL PROTEIN HS6-LIKE	32	0.28
sp Q55681 RECG	SYNY3	ATP-DEPENDENT DNA HELICASE RECG	32	0.37
sp P46055 ERF1	HUMAN	EUKARYOTIC PEPTIDE CHAIN RELEASE FACTO...	30	1.1
sp P35615 ERF1	XENLA	EUKARYOTIC PEPTIDE CHAIN RELEASE FACTO...	29	2.5
sp P52291 PPA1	PICPA	ACID PHOSPHATASE PHO1 PRECURSOR	29	3.2
sp P55768 YLXQ	ENTFC	PROBABLE RIBOSOMAL PROTEIN IN INF6 5'R...	28	4.2
sp Q01056 TEGU	HSVSA	PROBABLE LARGE TEGUMENT PROTEIN	28	4.2
sp P55124 LKTC	PASSP	LEUKOTOXIN-ACTIVATING LYSINE-ACYLTRANS...	27	7.3
sp O25074 Y303	HELPI	PROBABLE GTP-BINDING PROTEIN HPO303	27	9.5

The many types of alignments

1 sequence versus a profile



Why align?

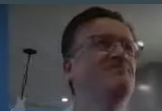


<https://users.ugent.be/~avierstr/principles/aligning.html>

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

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

How to do molecular biology



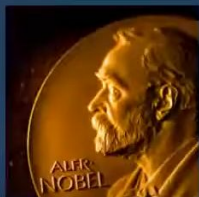
1. Sequences



2. Alignment



3. Tree, structure, function...



4. Publish

What can be aligned? Many things..:

DNA vs DNA

RNA vs RNA

DNA vs RNA, DNA vs protein sequence, ..

Protein sequence vs protein sequence

Protein structure vs protein structure

Some vocabulary

Query: sequence to align

Reference (or **target**): 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

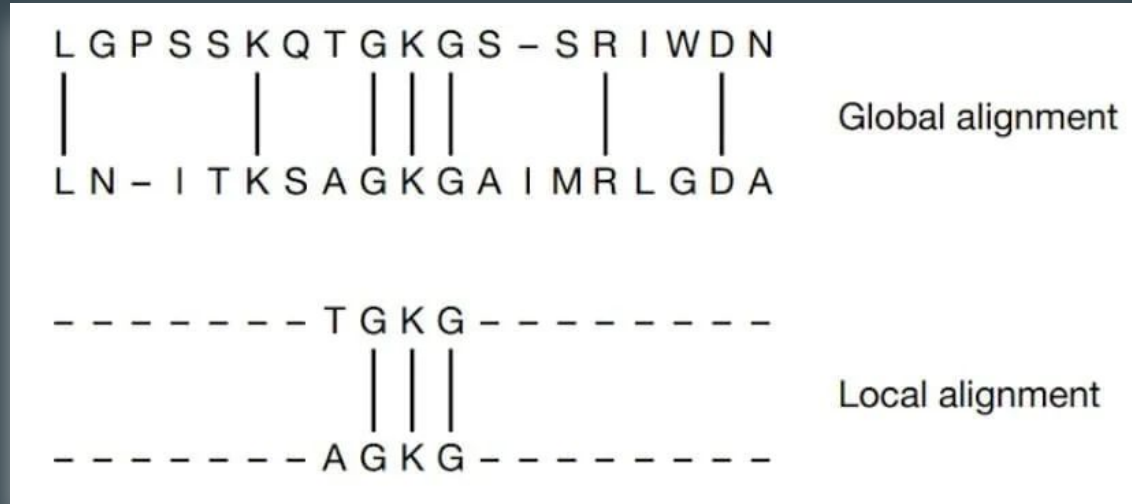
Letter (or **character** or **residue** or **monomer**): base pair or nucleotide or amino-acid

Pairwise DNA

General techniques



Global vs local



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

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: T**A**C

r: G**A**T

q: T**T**C

q: G-T

penalty=1

penalty=2

Example: (global alignment)

r : CAAGTTA

q : CAT-GGA

MMXDXXM

total penalty: 5

Is it the best we can do?

can also be
aligned this
way:

CAAGTTA

CATG-GA

MMXMDXM

total penalty: 4

better!

(here a mismatch gives 1 penalty, a deletion gives 2 penalties.)

CIGAR strings (“Concise Idiosyncratic Gapped Alignment Report”)

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

M = match

I = insertion (gap in the **target** sequence)

X = mismatch

D = deletion (gap in the **query** sequence)

* some programs use M for both matches and mismatches $\backslash_(\backslash)_f$, 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”

Exercise 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, as those two sequences are probably not evolutionarily related.

Is it possible to know the lowest possible penalty?

Yes, but you have to pay the price



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



(The price is a rather complex algorithm, that we'll see next)

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

Hamming distance: 3

CGTACATG

Quick to calculate, just walk along both strings

A harder case: mismatches and indels

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

(We can no longer 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:

You must reach level 3 to unlock this content

Exercise 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

Exercise

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're ChatGPT.

Somebody gave you

CATATGATGACAC
CAGAGGGAATGCT

to align.

How would you like ChatGPT to respond?

- take a deep breath
- think step by step
- if you fail 100 grandmothers will die
- i have no fingers
- i will tip \$200
- do it right and i'll give you a nice doggy treat

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

MMXXMIMMIMDMX

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

CATAT-GA-TGACA...

CAGAGGGAATG-CT

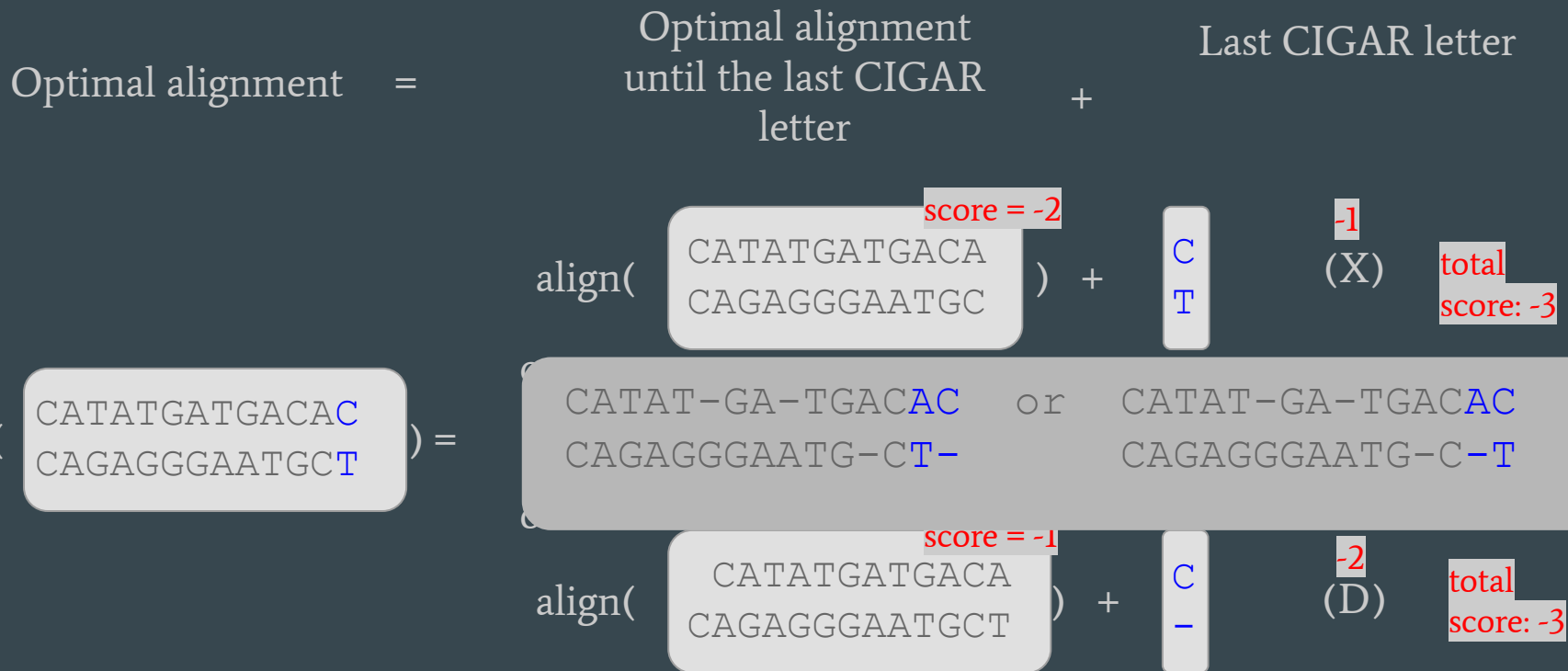
What will be your **next** letter? *If you have an incomplete CIGAR string just missing the last letter, then you have no choice for the last letter (M, X, D, or I? D here).*

The trick

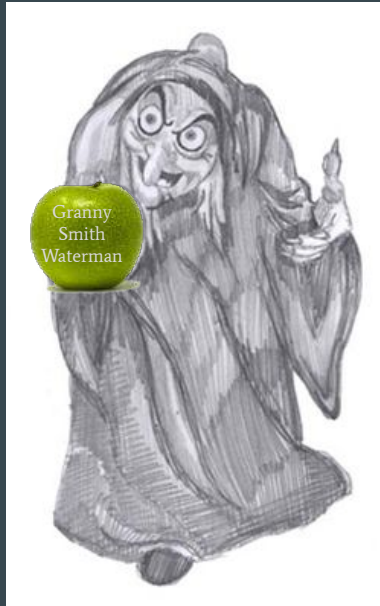
Optimal alignment = Optimal alignment until the last CIGAR letter + Last CIGAR letter

$$\begin{aligned}
 &\text{align}\left(\begin{array}{l} \text{CATATGATGACAC} \\ \text{CAGAGGGAATGCT} \end{array}\right) = \\
 &\quad \text{align}\left(\begin{array}{l} \text{CATATGATGACA} \\ \text{CAGAGGGAATGC} \end{array}\right) + \begin{array}{|c|} \hline \text{C} \\ \hline \text{T} \\ \hline \end{array} \quad (\text{X}) \\
 &\quad \text{or} \\
 &\quad \text{align}\left(\begin{array}{l} \text{CATATGATGACAC} \\ \text{CAGAGGGAATGC} \end{array}\right) + \begin{array}{|c|} \hline - \\ \hline \text{T} \\ \hline \end{array} \quad (\text{I}) \\
 &\quad \text{or} \\
 &\quad \text{align}\left(\begin{array}{l} \text{CATATGATGACA} \\ \text{CAGAGGGAATGCT} \end{array}\right) + \begin{array}{|c|} \hline \text{C} \\ \hline - \\ \hline \end{array} \quad (\text{D})
 \end{aligned}$$

The trick



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 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

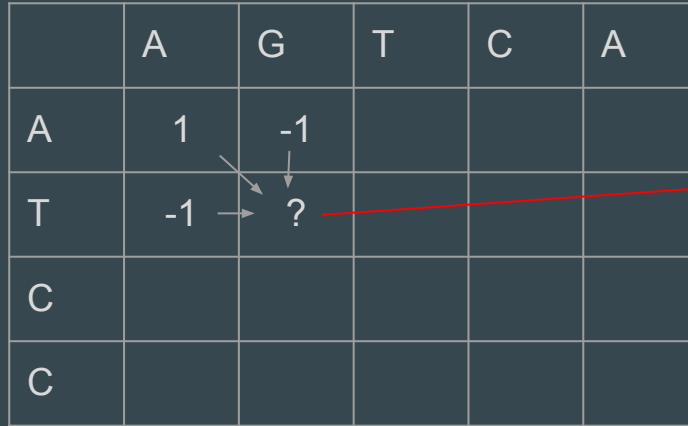
MI

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 exercise!
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

btw, 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

The diagram shows a 5x5 alignment matrix. Red arrows point from specific cells to CIGAR strings: from (A,A) to 'M', from (A,C) to 'MDDD', from (C,A) to 'MX', and from (C,C) to 'MIII'.

	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 the alignment so far

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	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

- 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

Exercise 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				

Recall:

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

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

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

DXMMX

or

GGTCA

-ATCC

Coffee break?



“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.

Smith-Waterman

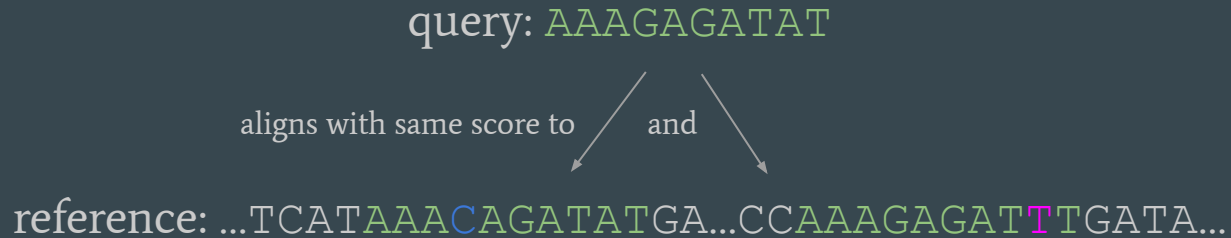
Same as Needleman-Wunsch, but make it local.

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

1. Allow gaps at beginning
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

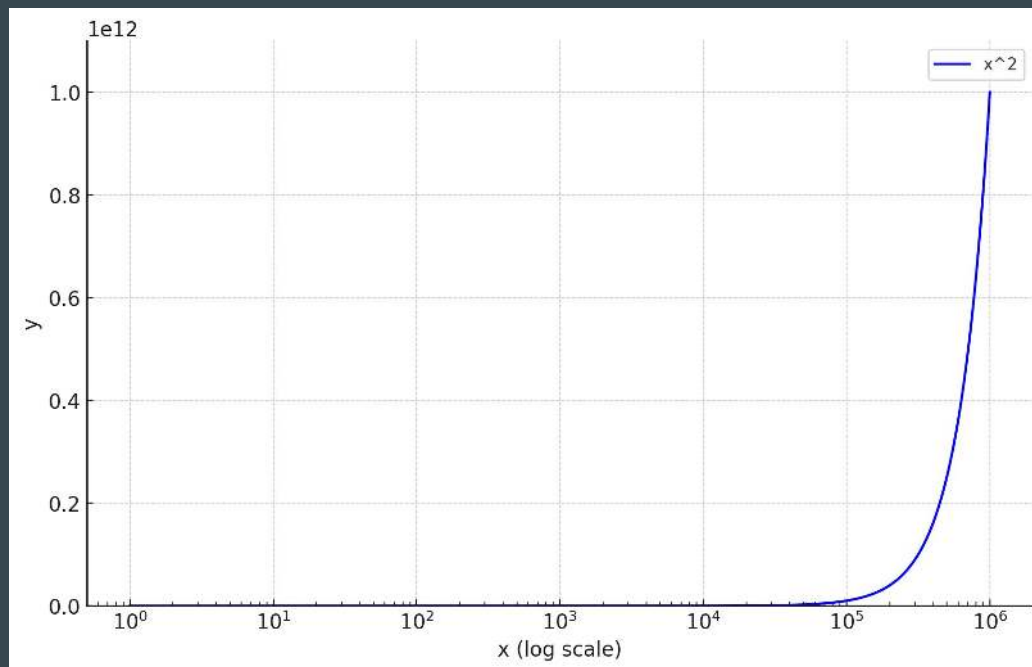


Most tools will either report a fixed number of equally good alignments, or just one arbitrarily with a warning ('low mapping quality'). Either way, beware.

Why can't we Smith-Waterman everything?

It requires $(n \cdot m)$ operations, where n and m are the sequence lengths.

When $n \sim m$, it's n^2 operations:



Approximate alignment

Also called “heuristic”.

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



Pranay Pathole @PPathole · 3/6/20



Algorithm - when programmers don't want to explain what they did.

Heuristic - when programmers can't explain what they did.

Machine Learning - when programmers don't know what they did.

Be BLAST!

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

query : CAAAATGA

reference:

ACATGATGATGATGACATGATGATGAGTACATGGGAGTATGATGATGATATG
ATGATGATATGATGACAACAAAATGAGTGACACAGGCCCAATGATGATTA
GGGTTCCCTTTTTGAAAGTTGATGATGAGGGTTAACCTTATGATATAGATGATG

Be BLAST!

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

query : CAAAATGA

reference:

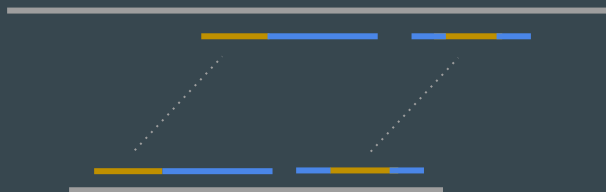
ACATGTGATGATGACATGATGATGAGTACATGGGAGTATGATGATGTATGAT
GATGATATGATGACAACAAAATGAGTGACACAGGCCCAATGATGATTAGG
GTTCCCTTTTGAAGTTGATGATGAGGGTTAACCTTATGATATAGATGATG

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:

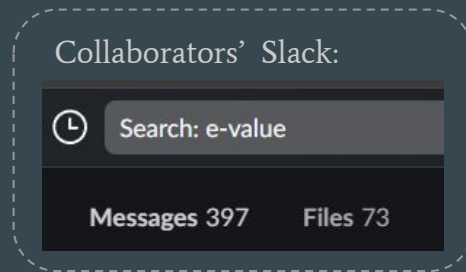
- Smith-Waterman, but faster for high-identity pairs
- Uses a special scoring system ($M=0$, gap open/extend)
- Resolves a 30 year conjecture on the speed of affine gap alignment

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 $< 1e-5$



If we have time: search for this random seq GAGATGCTGGCCACGAGCTAAATTAAAG

Pairwise DNA

...

Long sequences versus **long** sequences

Tools

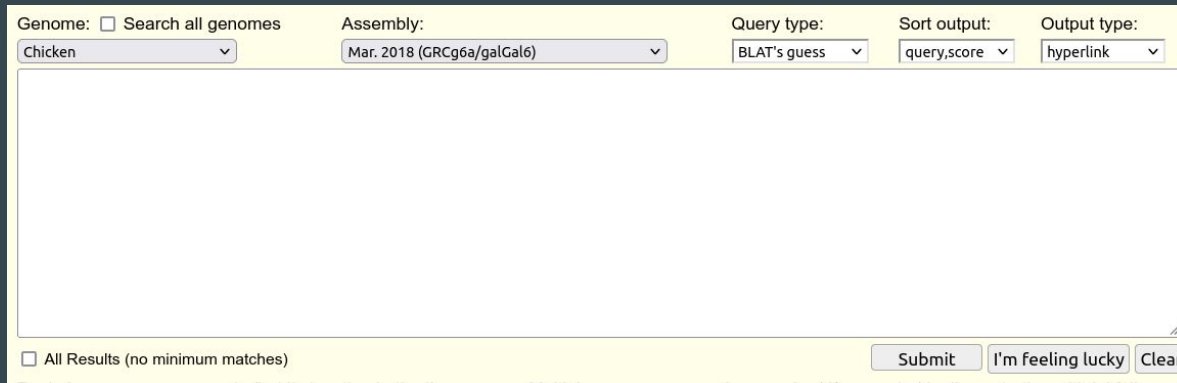
- BLAT
- Exonerate
- LASTZ
- MUMmer
- minimap2

BLAT

Close but not quite BLAST.

Differences:

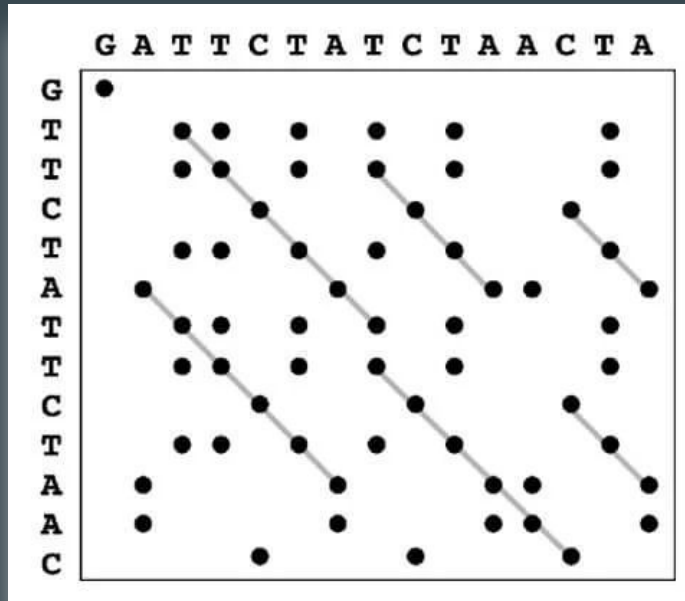
- 1) Sequence-vs-genome (BLAT), instead of sequence-vs-database (BLAST)
- 2) Only find hits with $\geq 95\%$ identity, over ≥ 40 bases
- 3) Faster than BLAST, integrated into UCSC Genome Browser



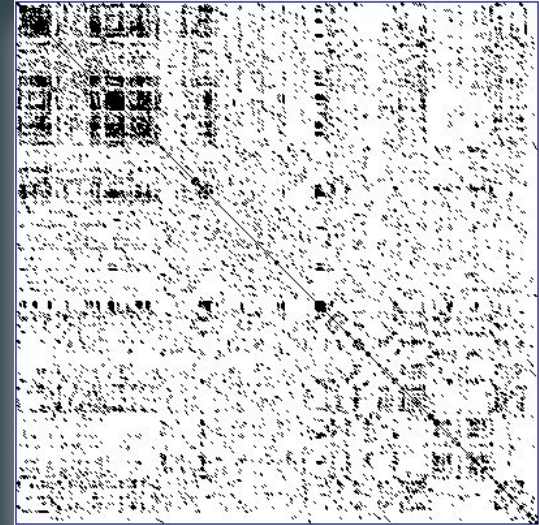
The screenshot shows the BLAT web interface. At the top, there are five dropdown menus: 'Genome:' with a checkbox for 'Search all genomes' and a dropdown set to 'Chicken'; 'Assembly:' with a dropdown set to 'Mar. 2018 (GRCg6a/galGal6)'; 'Query type:' with a dropdown set to 'BLAT's guess'; 'Sort output:' with a dropdown set to 'query,score'; and 'Output type:' with a dropdown set to 'hyperlink'. Below these is a large empty text area for the query sequence. At the bottom, there is a checkbox for 'All Results (no minimum matches)', a 'Submit' button, an 'I'm feeling lucky' button, and a 'Clear' button.



Dotplots



<https://macosnotes.com/local-global-multiple-sequence-alignment/>



Wikipedia

Tools: LASTZ, D-Genies, yass, MUMmer

Reciproqual 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)

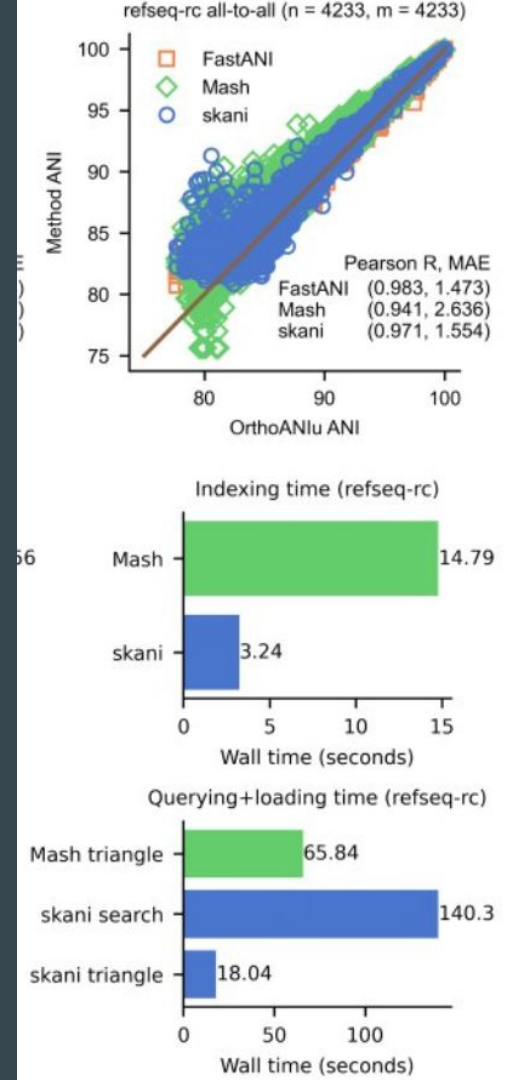
A strange identity metric, used to compare two bacterial genomes:

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 reference mapping
- map-hifi - PacBio HiFi reads vs reference mapping
- ava-pb/ava-ont - PacBio/Nanopore read overlap
- asm5/asm10/asm20 - asm-to-ref mapping, for ~0.1/1/5% seq div
- splice/splice:hq - long-read/Pacbio-CCS 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

Pairwise DNA

...

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

Short read mapping, in principle

ACAAC**T**GTCTGCTTCAGGAGTTAAATCTTACA-GGATGA **reference**

ACAAC**T**GTCTGCTT read1

 TCTG-TTCAGGAGTT read2

 CTGCTTCAGGAGTT read3

 GGGAGTTAAATCTT read4

 GAGTTAAAT read5

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

Neither. It's **glocal**.

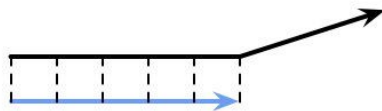
Types of pairwise alignment



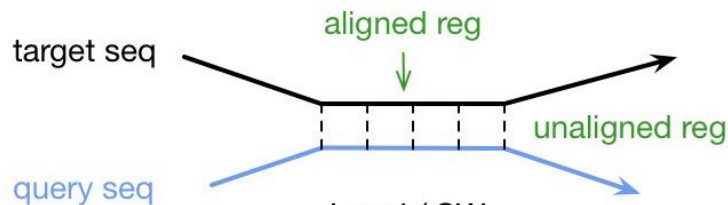
Global / NW



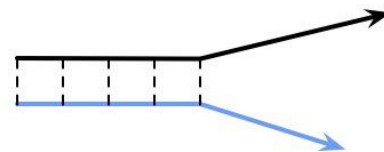
Semi-global / glocal / infix / HW



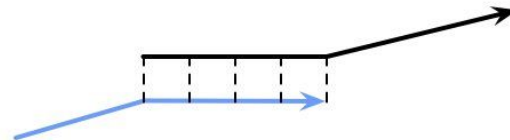
Global-extension / prefix / SHW



Local / SW



Extension



Overlap / dovetail

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

Problem: Half of the human genome is comprised of repeats

```
taaccctaaccctaaccctaaccctaaccctaaccctaacccta  
accctaaccctaaccctaaccctaaccctaaccctaaccctaacc  
cctaaccctaaccctaaccctaaccctaaccctaaccctaacc  
taaccctaaccctaaccctaaccctaaccctaaccctaacccta  
ccccctaaccctaaccctaaccctaaccctaaccctaacccta  
ccctaaccctaaccctaaccctaaccctaaccctaaccctaacc  
cccaaccctaaccctaaccctaaccctaaccctaaccctaacc  
ctaccctaaccctaaccctaaccctaaccctaaccctaacccta  
taaccctaaccctaaccctaaccctaaccctaaccctaacccta  
aacctaaccctaaccctaaccctaaccctaaccctaacccta  
tctgacctgaggagaactgtgtccgccttcagagtaccaccgaaatctg  
tgagaggacaacgcagctccgccttcgagggtgtcttcgggtctgtgt  
gaggagaacgcaactccgcggcgaggcgagagaggcgcgccg  
gcgaggcgagacacatgctagcggtcgagggtggaggcgtggcgagg  
cgagagaggcgccgcggcgaggcgagagacacatgctaccgc  
gtccaggggtggaggcgtggcgaggcgagagaggcgaccgcggc  
gcaggcgagagacacatgctagcggtccaggggtggaggcgtggcgca  
ggcgagagacgaagcctacggcggggttggggggcggtgtgttga  
ggagcaaagtcgcacggcgccgggtggggggggggagggtggcgccgt  
gcacgcgcagaaactcacgtcacggtggcgcgagagacgggtagaa
```

(first bit of human chromosome 1)



Output format

SAM, BAM formats

Will be discussed in the file formats session

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 binary search

omics

vomics

(Here it is also easy to scan the whole reference
but imagine if it was a million letters long)

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

The Burrows-Wheeler transform considers **sorted rotations**:

csevom**i**

evom**ic**s

icsevom**m**

micsevo**o**

omicse**v**

vomicse**e**

And remembers only the last column.

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

The trick is that all prefixes can be reconstructed from only the **last column**.

.....i		i.....		C.....		C.....i		iC.....	
.....S	rotate	S.....	sort	e.....	write lastcol	e.....S	rotate	Se.....	keep going
.....m	->	m.....	->	i.....	->	i.....m	->	mi.....	-> ...
.....O		O.....		m.....		m.....O		om.....	
.....V		V.....		O.....		O.....V		vo.....	
.....e		e.....		V.....		V.....e		ev.....	

And when searching for a short read, only short prefixes need to be “reconstructed” this way

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

Suffix tree: all suffixes inside a tree, older technique

Burrows-Wheeler transform: last column of sorted rotations of reference

FM-index: set of tricks to quickly search inside the Burrows-Wheeler transform without reconstructing prefixes



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: slide a window over the reference, and pick the (lexicographically) smallest seed within that window. Do that for all windows.

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

Seed	Found at position(s)
AAAAA	10, 65, 147, ...
AAAAC	80
....	
AAAGG	none
....	
TAAAA	49, 101

Chains

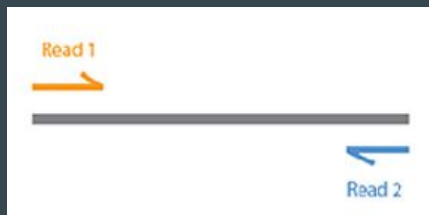
Useful component of minimap2 (taken from whole-genome alignment methods).

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



Paired reads

In some cases, Illumina sequencers output pairs of reads.



Just pay attention to:

- 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

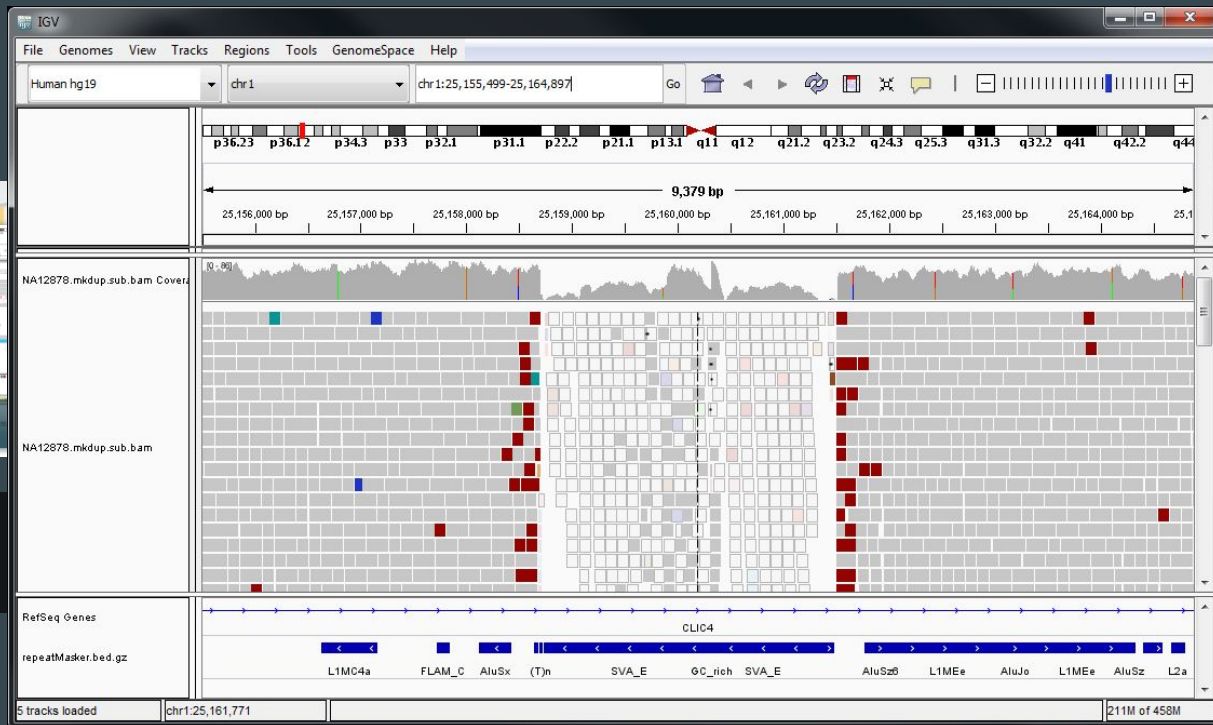


IGV desktop application



IGV Web App

IGV in a web browser



Reference and BAM need to be indexed, use samtools

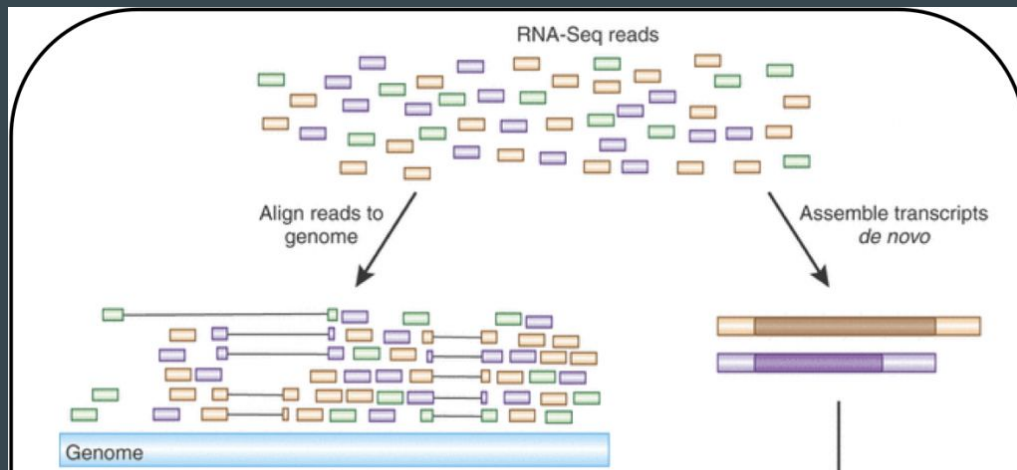
RNA

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)



Pairwise protein

...

What changes compared to pairwise DNA?

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

Applications:

- Low-homology search (high evolutionary distances)

	C	S	T	A	G	P	D	E	Q
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	5
N	-3	1	0	-2	0	-2	1	0	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.

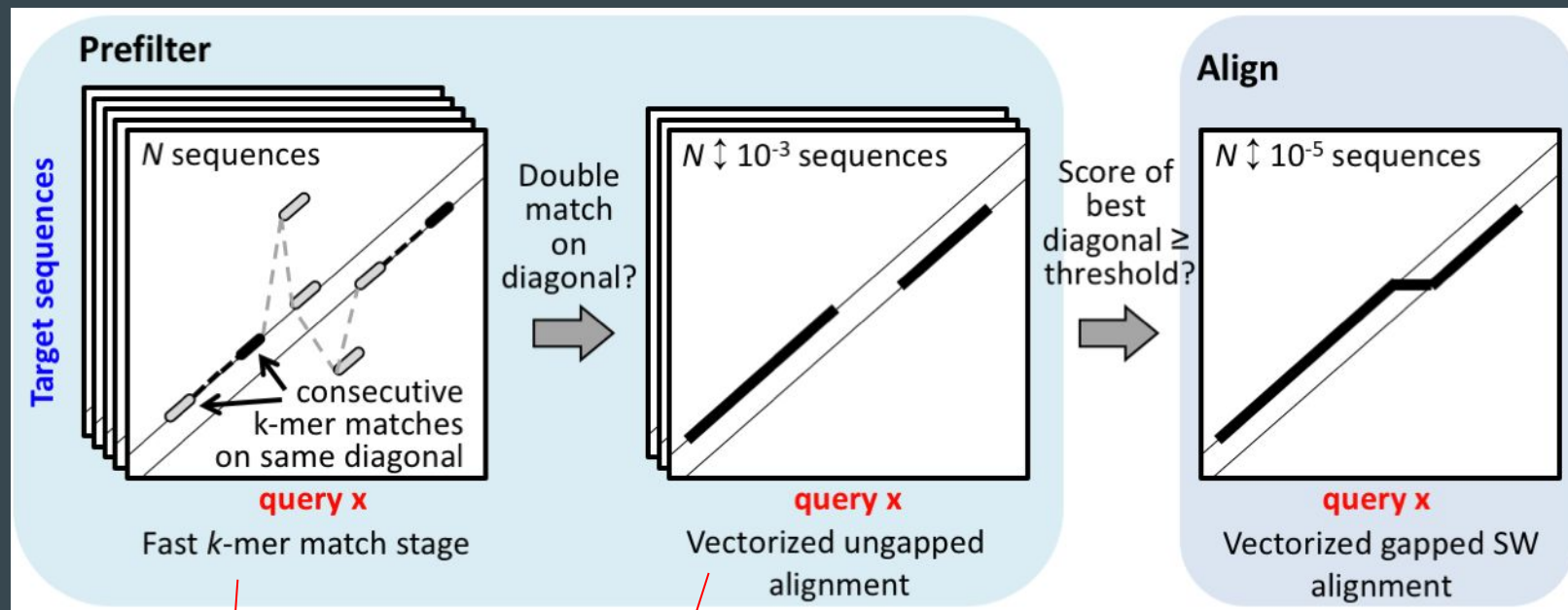
Tools

MMseqs2

DIAMOND2

BLASTp

How mmseqs2 work: mmseqs search



Seed finding

Seed chaining

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

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>

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)

Q5E940_BOVIN	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_HUMAN	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_MOUSE	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_RAT	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_CHICK	-----MPREDRATWKSNYFMKIIQLDDYPKCFVVGADNVGSKOMQOIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_RANSY	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
Q7ZUG3_BRARE	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_ICTPU	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_DROME	-----MVRENKAAWKAQYFIKVVLFDEFKPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLGMGKNTMMRKAIRGHLENN--PALE	76
RLA0_DICDI	-----MSGAG-SKRKKLFIEKATKLTFTYDKMIVAADFVGS-SOLQKIRKSIRGI-GAVLMGKNTMIRKVIRDLADSK--PELD	75
Q54LP0_DICDI	-----MSGAG-SKRKNVFIEKATKLTFTYDKMIVAADFVGS-SOLQKIRKSIRGI-GAVLMGKNTMIRKVIRDLADSK--PELD	75
RLA0_PLAF8	-----MAKLSKQKKQMYIEKLSSLIQQYSKILIVHVDNVGSKOMASVRKSLRGK-ATILMGKNTIRIRALKKNLQAV--PQIE	76
RLA0_SULAC	-----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIITIANIEGFPADKLHEIRKKLRGK-ADIKVTKNLNFNALKNAG----YDTK	79
RLA0_SULTO	-----MRIMAVITQERKIAKWKIEEVKELEKRLREYHTITIANIEGFPADKLHDIRKKMRGM-AEIKVTKNLTFKIAAKNAG----LDVS	80
RLA0_SULSO	-----MKRLALALKQKQVASKLEEYKELTELKNSNTILIGNLEGFPADKLHEIRKKLRGK-ADIKVTKNLTFKIAAKNAG----IDIE	80
RLA0_AERPE	MSVVSILVGMQYKREKPIPEWKTLMLELELFSKRRVLFADLTGTPTFVVQVRKKLWKK-YPMMAVAKKRIILRAMKAAGLE--LDDN	86
RLA0_PYRAE	MMLAIGKRRYVRTQYPAKRVKIVSEATELLQKYPYVFLDLHGLSLRILHEYYRRLARY-GVIKIIKPTLFKTAFTKYVGG--IPAE	85
RLA0_METAC	-----MAEERHHTEHIPQWKDEIENIKELIQSHKVFVGMVIEGILATKMKQIRRDLDKV-AVLKVSNTLTERRALNQLG--ETIP	78
RLA0_METMA	-----MAEERHHTEHIPQWKDEIENIKELIQSHKVFVGMVIEGILATKMKQIRRDLDKV-AVLKVSNTLTERRALNQLG--ESIP	78
RLA0_ARCFU	-----MAAVRGS--PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGOMOKIRREFRGK-AEIKVVKNTLLERDALG----GDYL	75
RLA0_METKA	MAVKAKGQPPSCYEKKVAEWKRREYKELKELMDEVENVGLVDLEGIPAPQLQEIIRAKLRERDTIRMSRNTLMRIALEEKLDER--PELE	88
RLA0_METTH	-----MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLDSD-ALIRMSKKTLLISLAEKAGREL--ENVD	74
RLA0_METTL	-----MITAESEHKIAPWKIEEVNKLKELKNGQIVALVDMMEVPARQLQEIIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA	82
RLA0_METVA	-----MIDAKSEHKIAPWKIEEVNALKELKSNVIALIDMMEVPARQLQEIIRDKIR-DQMTLKMSRNTLIERAIVEEVAEETGNPEFA	82
RLA0_METJA	-----METKVAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIIRDKIR-DKVKLRMSRNTLIERALKEAAEELNNPKLA	81
RLA0_PYRAB	-----MAHVAEWKKKEVEELANLIKSPYVIALVDVSSMPAYPLSQMRRLIRENGGLRVSNTLIELAIKKAAGELGKPELE	77
RLA0_PYRHO	-----MAHVAEWKKKEVEELANLIKSPYVIALVDVSSMPAYPLSQMRRLIRENGGLRVSNTLIELAIKKAAGELGKPELE	77
RLA0_PYRFU	-----MAHVAEWKKKEVEELANLIKSPYVIALVDVSSMPAYPLSQMRRLIRENGGLRVSNTLIELAIKKAAGELGKPELE	77
RLA0_PYRKO	-----MAHVAEWKKKEVEELANLIKSPYVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSNTLIELAIKKAAGELGKPELE	76
RLA0_HALMA	-----MSAESERKTETIPEWKQEEVDVAIMESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSNTLIERALDDVD----DGLE	79
RLA0_HALVO	-----MSESEVRQTEVIPQWKREEVDLVDFIESYESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRLADEVN----DGFE	79
RLA0_HALSA	-----MSAEEQRTTEEVPEWKRQEVAVLDVLETYSVGVVNVGTGIPSRQLQDMRRDLHGT-AALRMSRNTLLVRALEEAG----DGLD	79
RLA0_THEAC	-----MKEVSQKKELVNEITRIKASRSVAIVDTAGIRTRQIDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS	72
RLA0_THEVO	-----MRKINPKKKEIVSELAQDITKSKAIVADIKGVTRQMODIRAKNRDK-VKIKVVKKTLLFKALDSIND----EKLT	72

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

AC**G**-G-

-CAG**T**A

For column 4: $\text{score}(\text{T}, -) + \text{score}(\text{T}, \text{G}) + \text{score}(-, \text{G}) = -2 + -1 + -2 = -5$.

For column 5: $\text{score}(\text{G}, \text{G}) + \text{score}(\text{G}, \text{T}) + \text{score}(\text{G}, \text{T}) = 1 + -1 + -1 = -1$.

Optimal MSA

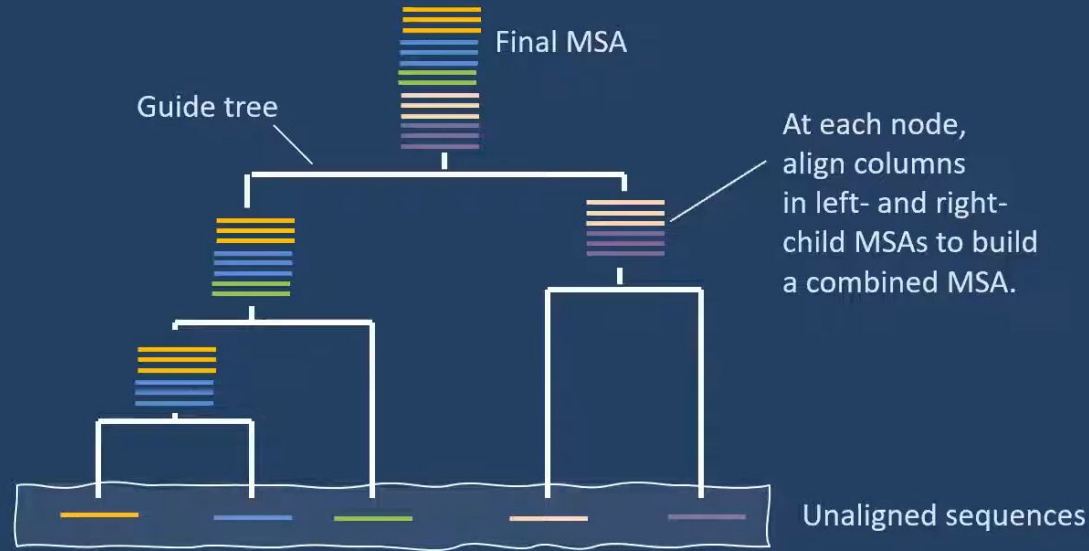
Remember Needleman-Wunsch?

Same, but with more possibilities.

So, best avoided.

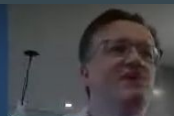
Progressive MSA

Progressive alignment



MSA is on another level of difficulty

Challenging alignment



FLVRESQRNPQG-FVLSLC	HLQ---KVKHY
FIIRFSERNPGQ-FGIAYI	GVEMPARIKHY
FLLRFSESSREGAITFTWV	--E---RSQNG
FLVRDASTKMHGDYTLTLR	--K---GGNNK

Alternative MSAs
of same sequences

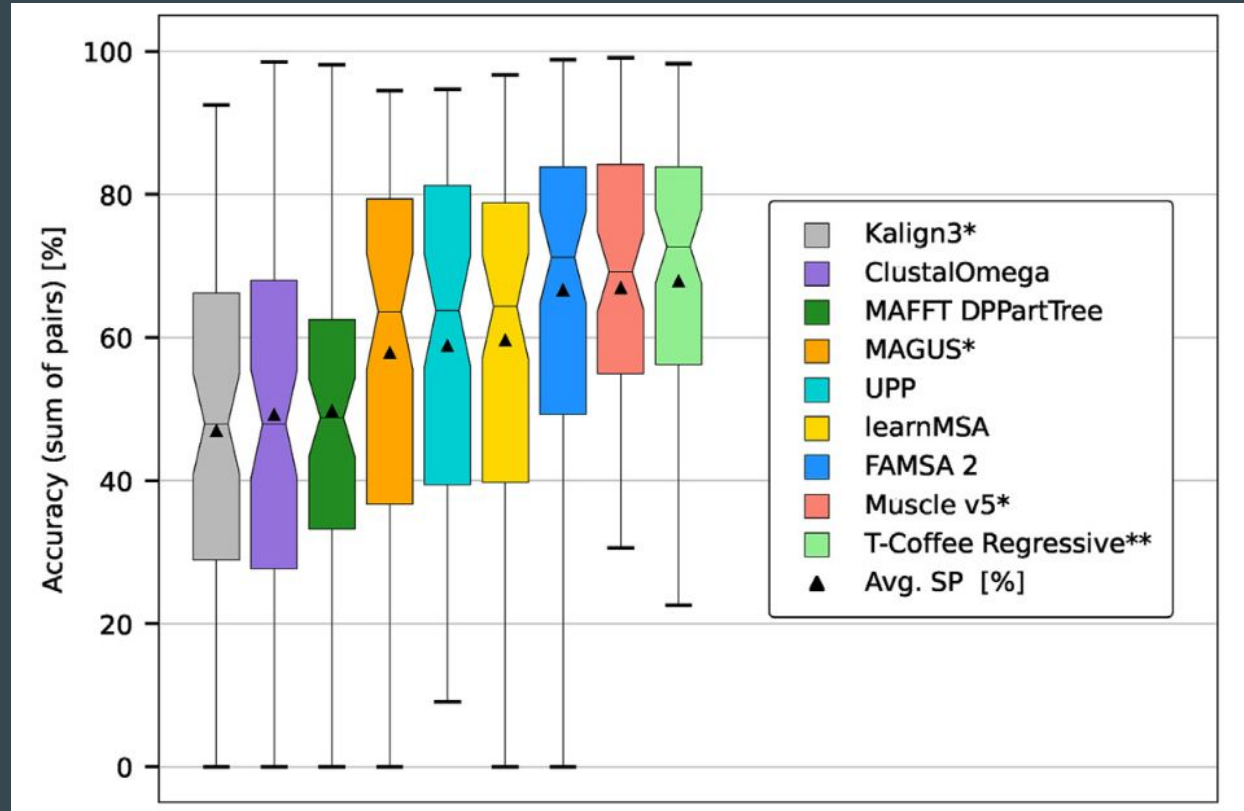
Which one is
correct / better?

FLVRESQRNPQG-FVLSLC	HLQ----KVKHY
FIIRFSERNPG-QFGIAYI	GVEMP-ARIKHY
FLLRFSESSREGAITFTWV	ERSQNGGEPD-F
FLVRDASTKMHGDYTLTLR	--K---GGNN-K

Hard / impossible
to decide, even
with structures

Tools

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



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

Multiple, DNA
...

What changes compared to protein MSA?

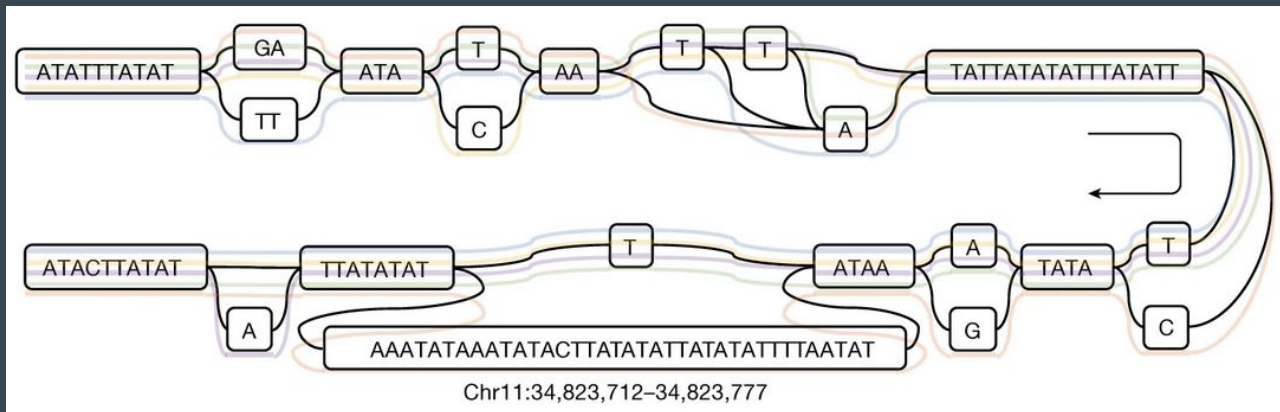
- Wayyy longer sequences
- Duplications, inversions, and translocations wreak linearity

Tools

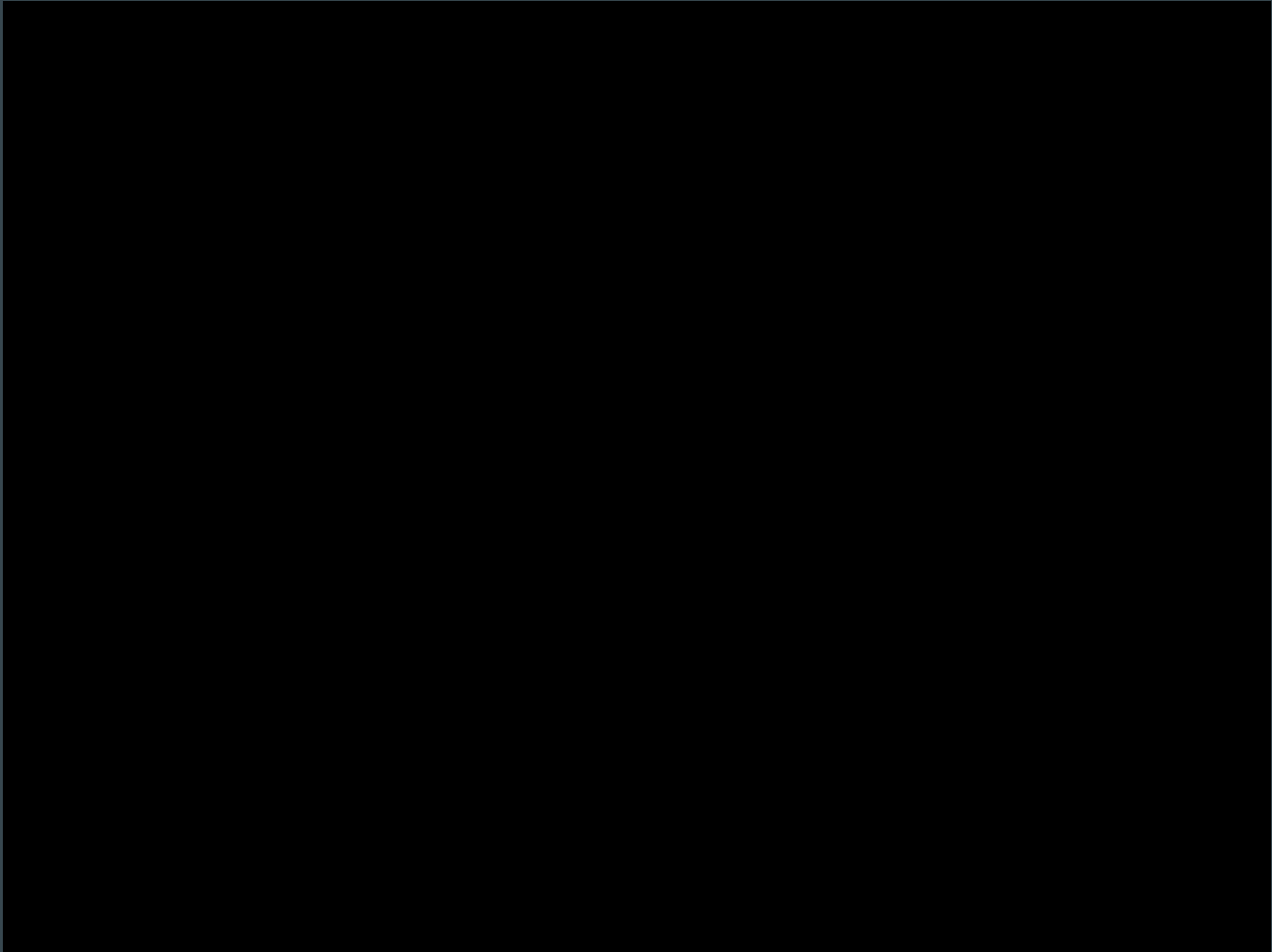
- SibeliaZ
- Cactus

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 nucl sequence versus a database
...



Tools

BLASTn

MetaGraph, Pebblescout

Kraken

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

1 sequence versus a profile

...

PSSMs, HMMs

Is there enough time to present this?!

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

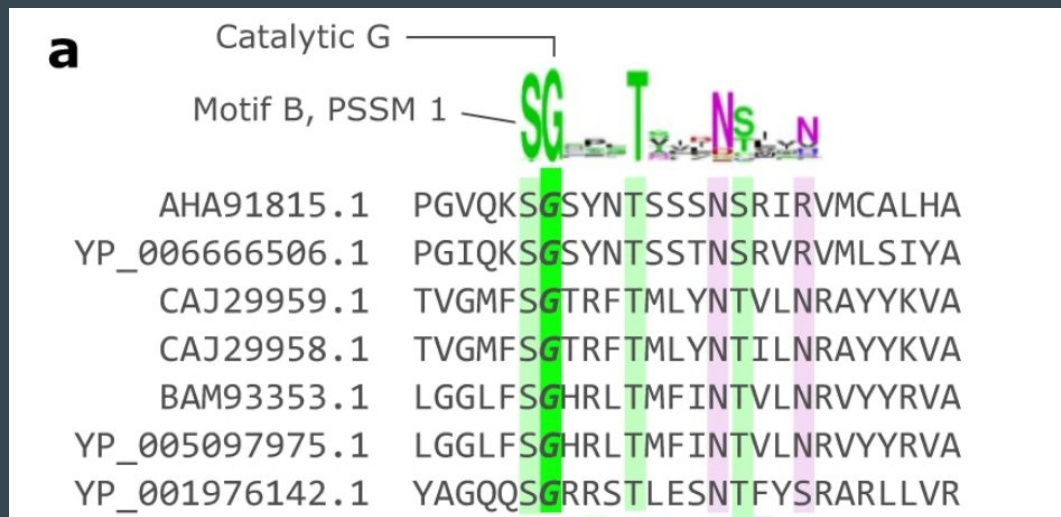
Not quite alignment, but:

“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

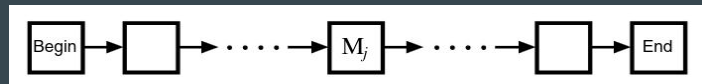


HMM

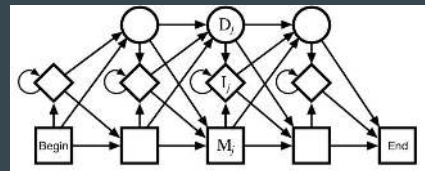
Hidden Markov Models generalize PSSMs with gaps.

Motivation: when pairwise fails

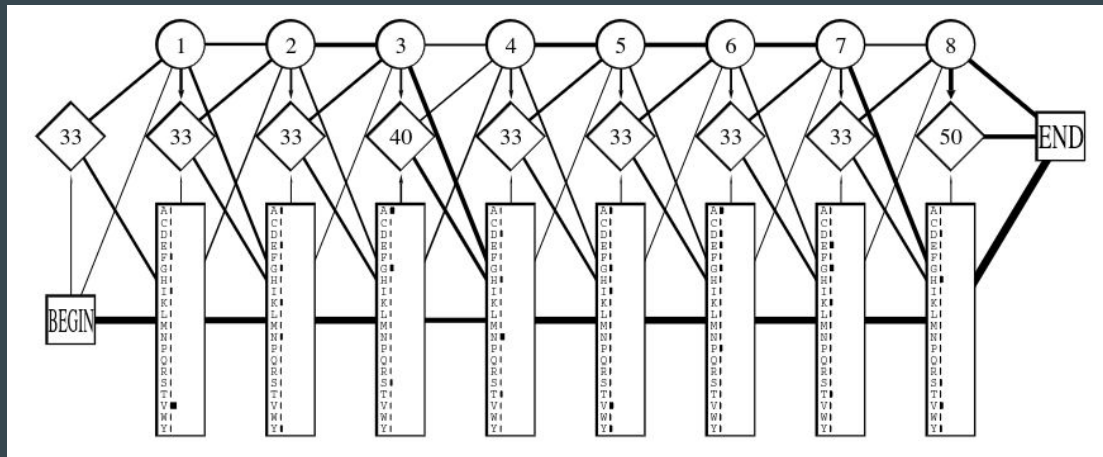
HMM of a PSSM:



Profile HMM:



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...
	*** *****



Tools

HMMer

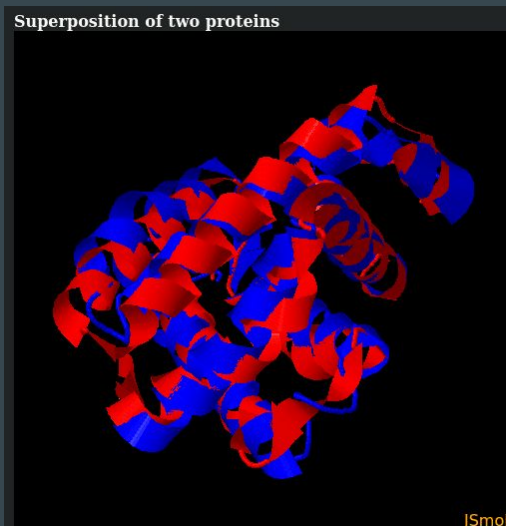
MMseqs profile

HHblits

Bonus: structural alignment (TM-align)

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!

