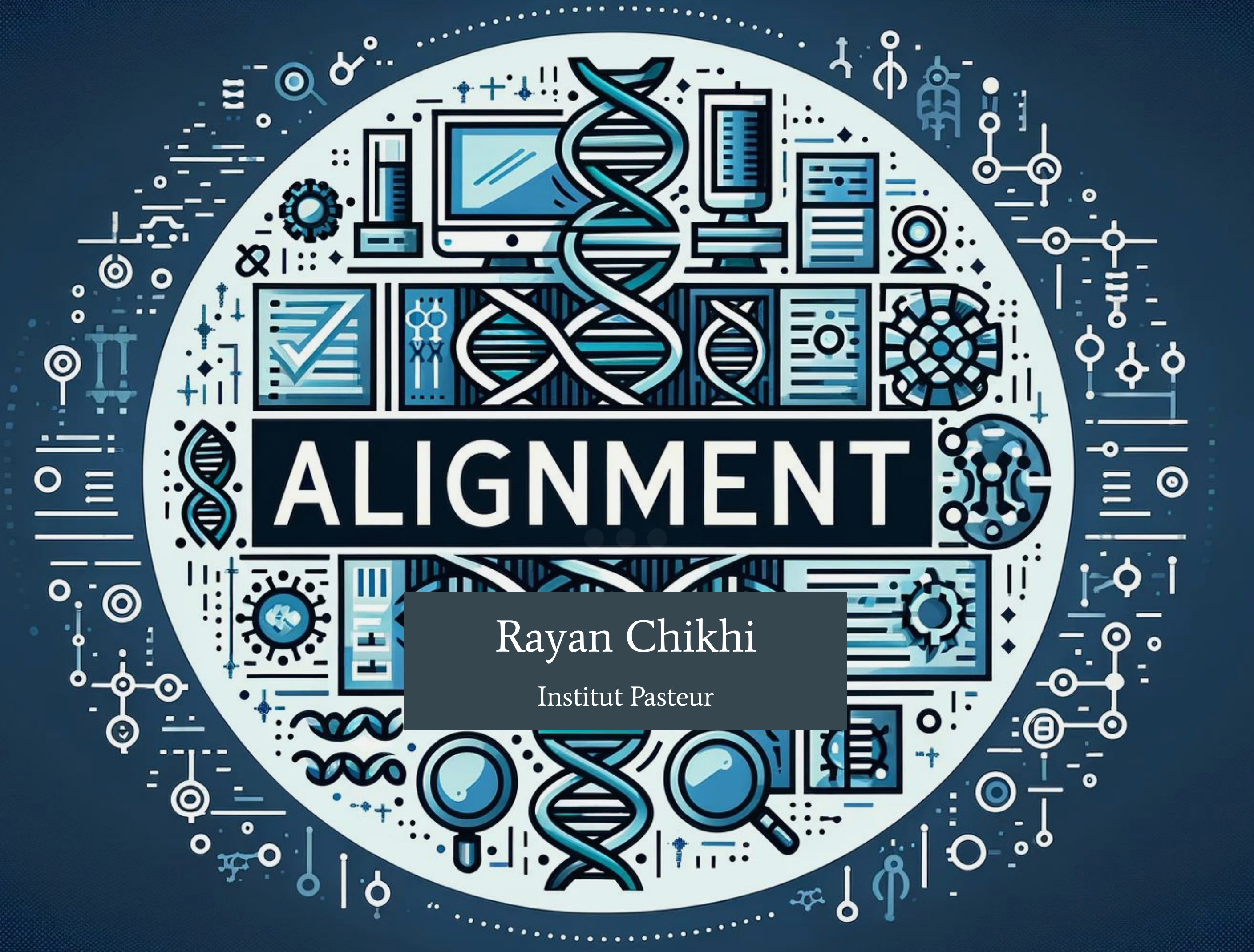
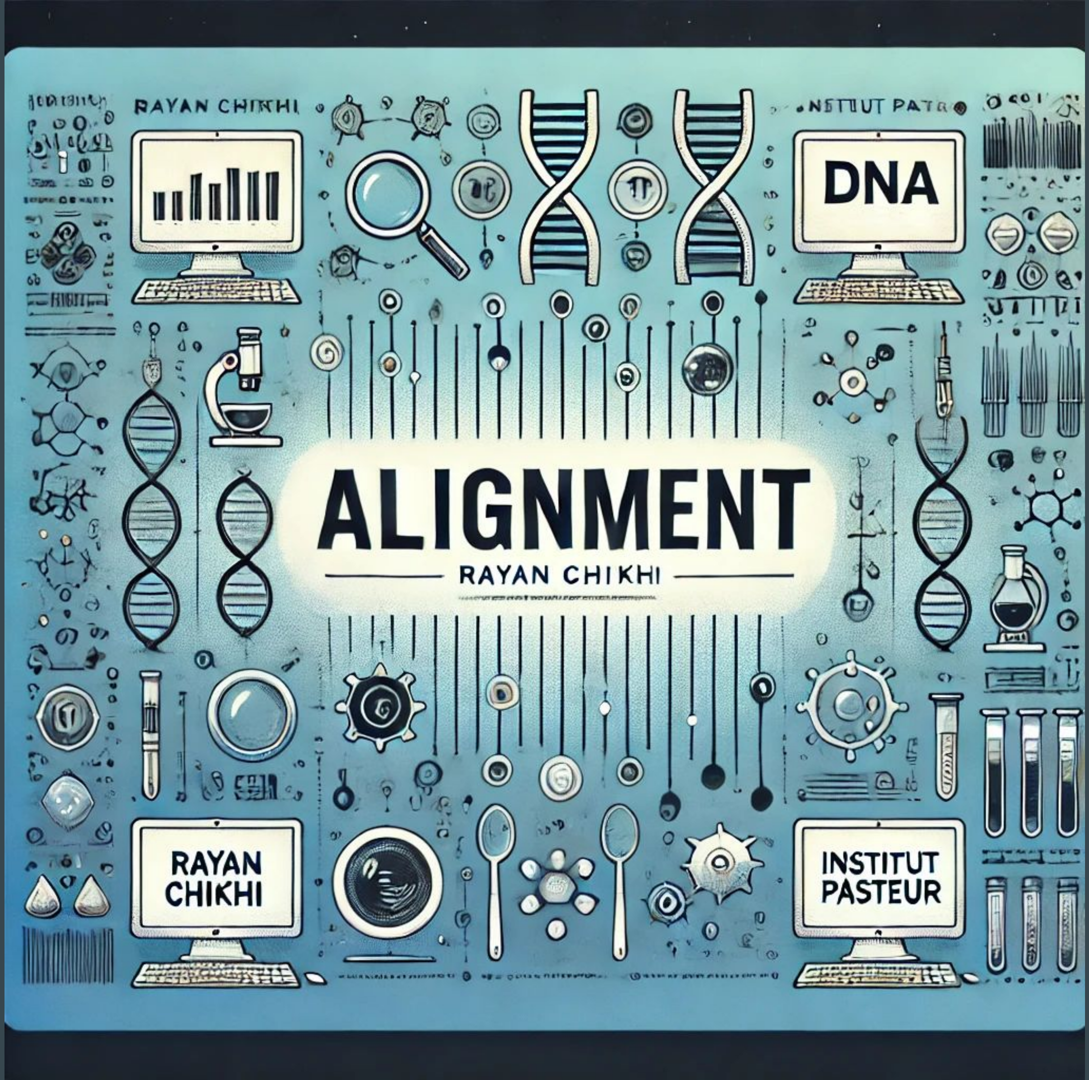


Rayan Chikhi

Institut Pasteur


RAYAN CHIKHI
•NSTIUT PATR
DNA
ALIGNMENT
RAYAN CHIKHI
RAYAN CHIKHI
INSTITUT PASTEUR

Hello!

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

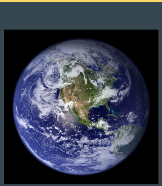
@RayanChikhi on X/Bsky

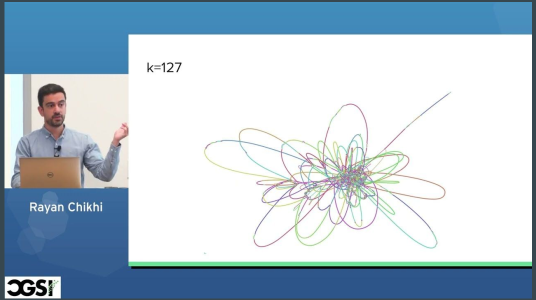
http://rayan.chikhi.name

https://github.com/IndexThePlanet/Logan

Sequence Bioinformatics

INSTITUT PASTEUR





"never in France"

RIP timeline.google.com

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

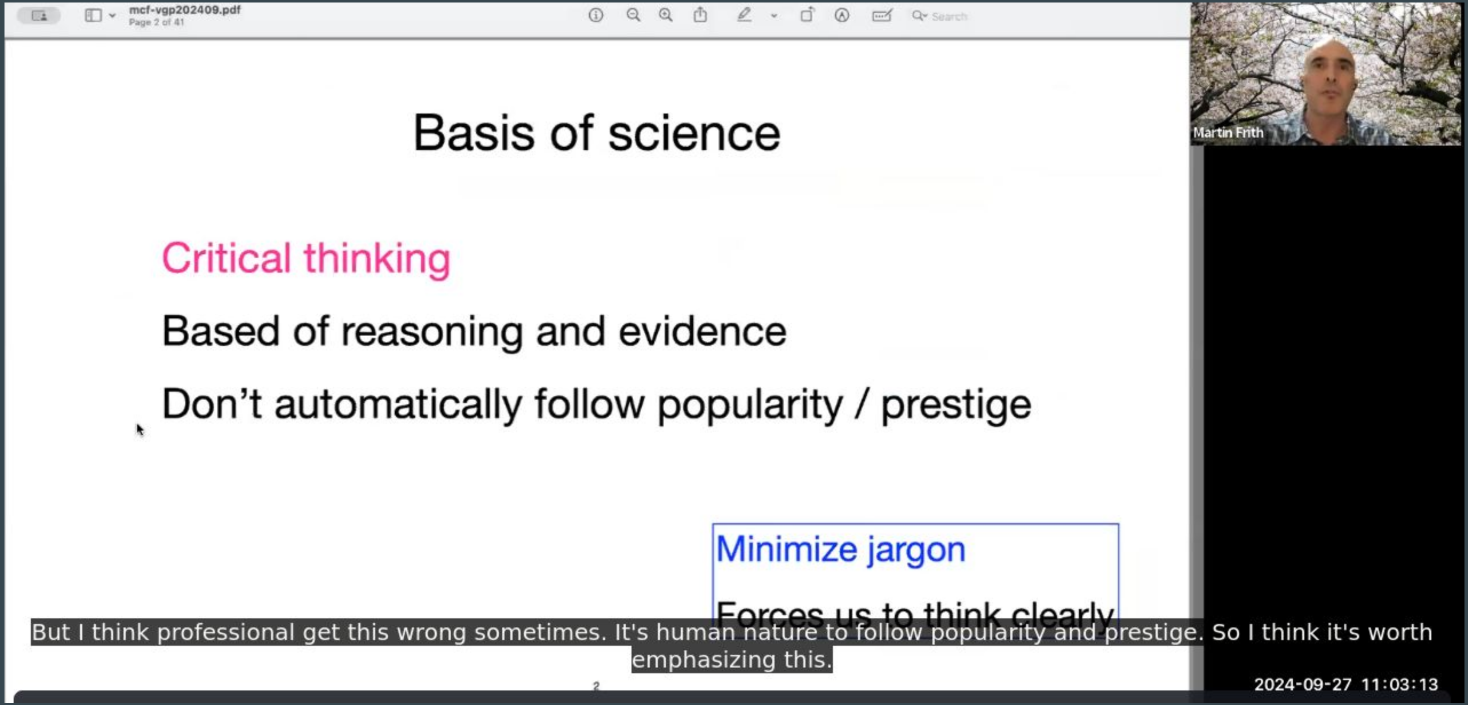
Course diff, if you’ve seen it

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

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

Minimiz
Minimizers in a toy example with k = 3 and w= 3
For all w ≤ k, it is guaranteed every position will be covered by at least one minimizer
Although compression is small (7/14) in this toy example, it is given by 2 / (w+1), so can become quite efficient for large w

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



Basis of science

Critical thinking

Based of reasoning and evidence

Don't automatically follow popularity / prestige

Minimize jargon

Forces us to think clearly

But I think professional get this wrong sometimes. It's human nature to follow popularity and prestige. So I think it's worth emphasizing this.

2024-09-27 11:03:13

(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?

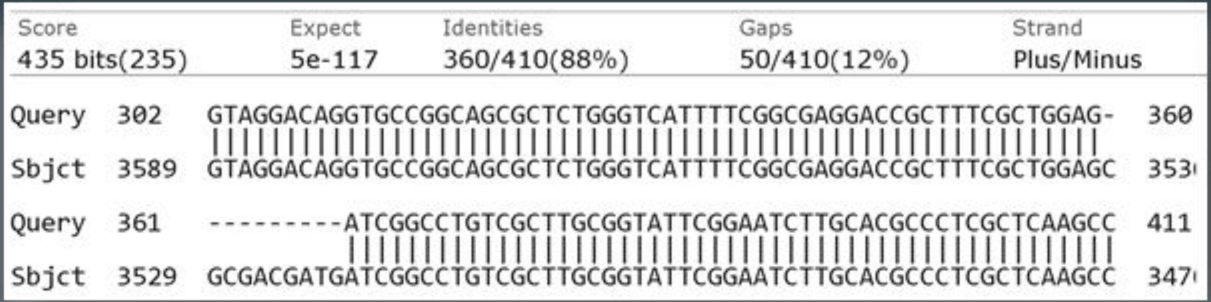
What’s an “alignment”?

…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)



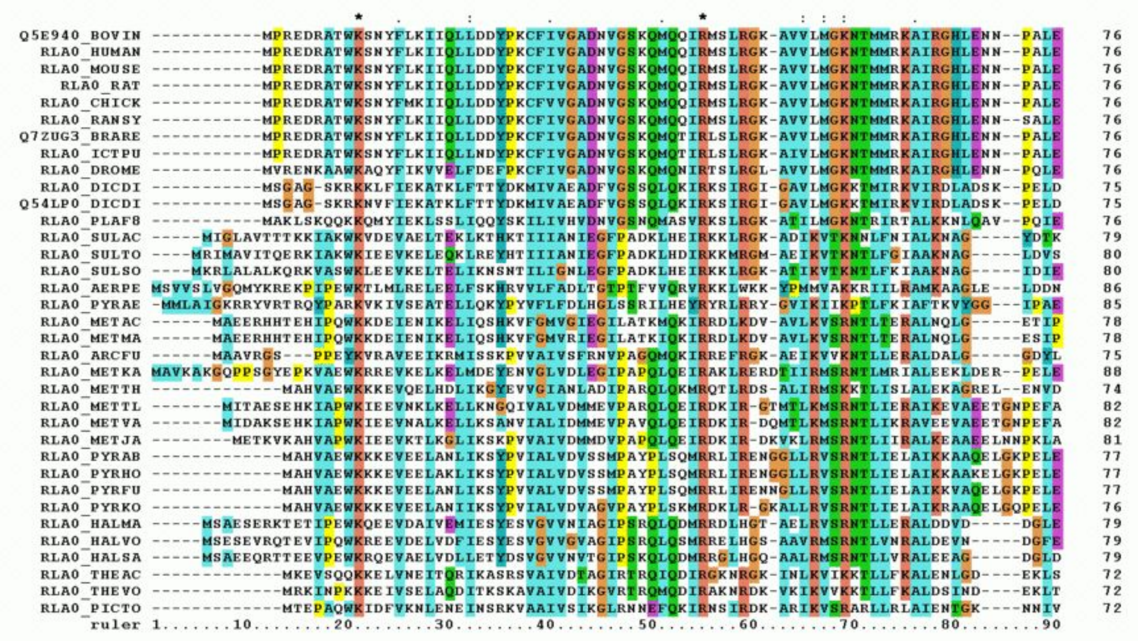
Score	Expect	Identities	Gaps	Strand
435 bits(235)	5e-117	360/410(88%)	50/410(12%)	Plus/Minus

```
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)

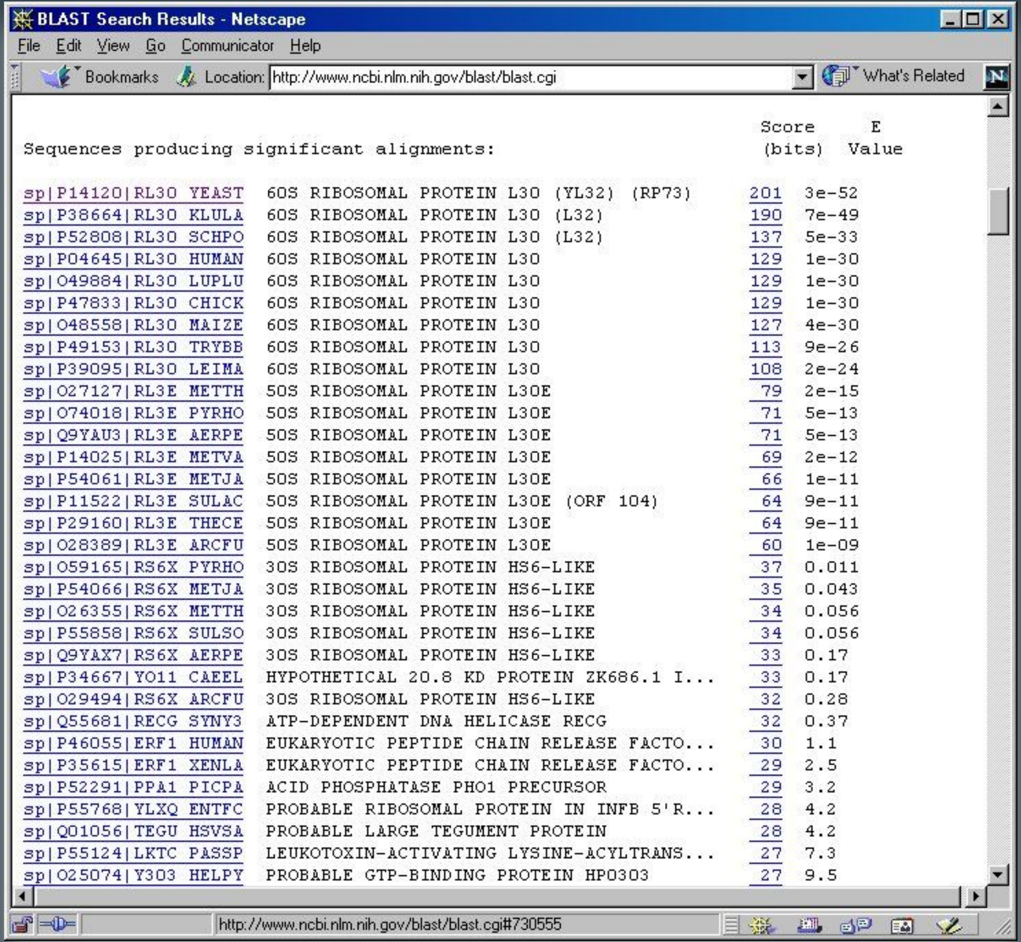


```
                                 *   .      :       .          .   *         : : :      .
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
    RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE    76
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE    76
  RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD    75
Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD    75
  RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE    76
  RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK    79
  RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS    80
  RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE    80
  RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN    86
  RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE    85
  RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP    78
  RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP    78
  RLA0_ARCFU ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL    75
  RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE    88
  RLA0_METTH -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD    74
  RLA0_METTL -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA    82
  RLA0_METVA -------MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA    82
  RLA0_METJA --------METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA    81
  RLA0_PYRAB -------------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAQELGKPELE    77
  RLA0_PYRHO -------------MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAKELGKPELE    77
  RLA0_PYRFU -------------MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRLIRENNGLLRVSRNTLIELAIKKVAQELGKPELE    77
  RLA0_PYRKO -------------MAHVAEWKKKEVEELANIIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKRAAQELGQPELE    76
  RLA0_HALMA -----MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSRNTLLERALDDVD-----DGLE    79
  RLA0_HALVO -----MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRALDEVN-----DGFE    79
  RLA0_HALSA -----MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEEAG-----DGLD    79
  RLA0_THEAC -------------MKEVSQQKKELVNEITQRIKASRSVAIVDTAGIRTRQIQDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS    72
  RLA0_THEVO -------------MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTRQMQDIRAKNRDK-VKIKVVKKTLLFKALDSIND----EKLT    72
  RLA0_PICTO -------------MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNSIRDK-ARIKVSRARLLRLAIENTGK----NNIV    72
       ruler 1.......10........20........30........40........50........60........70........80........90
```

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

The many types of alignments

1 sequence versus a database

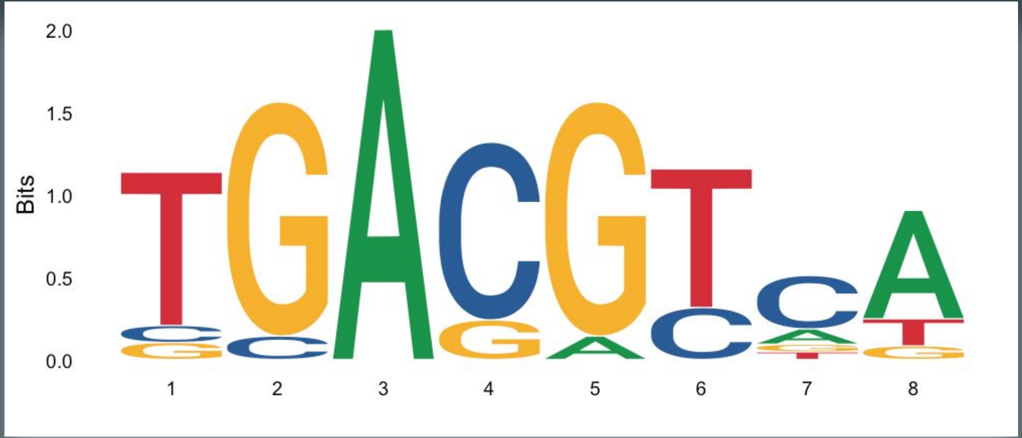

BLAST Search Results - Netscape
File Edit View Go Communicator Help
Bookmarks
Location: http://www.ncbi.nlm.nih.gov/blast/blast.cgi
What's Related
http://www.ncbi.nlm.nih.gov/blast/blast.cgi#730555

```
                                                                   Score     E
Sequences producing significant alignments:                        (bits)  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 INFB 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 HELPY  PROBABLE GTP-BINDING PROTEIN HP0303            27   9.5
```

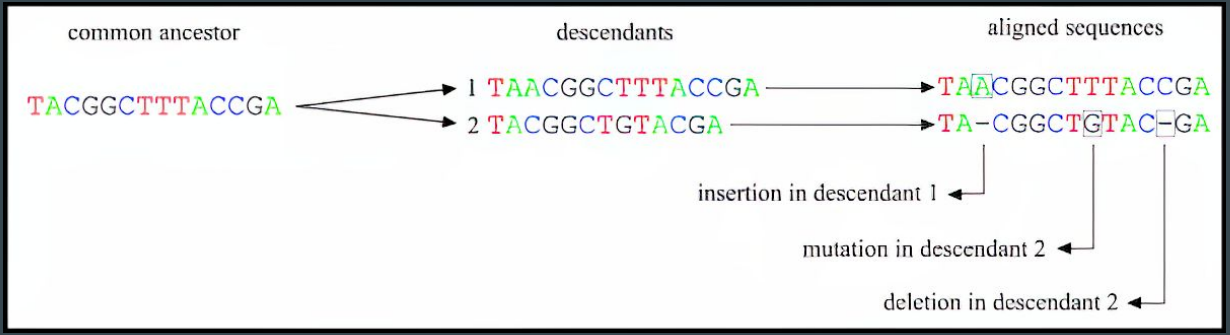
https://www.bch.cuhk.edu.hk/teaching/kbwong/1.html

The many types of alignments

1 sequence versus a profile



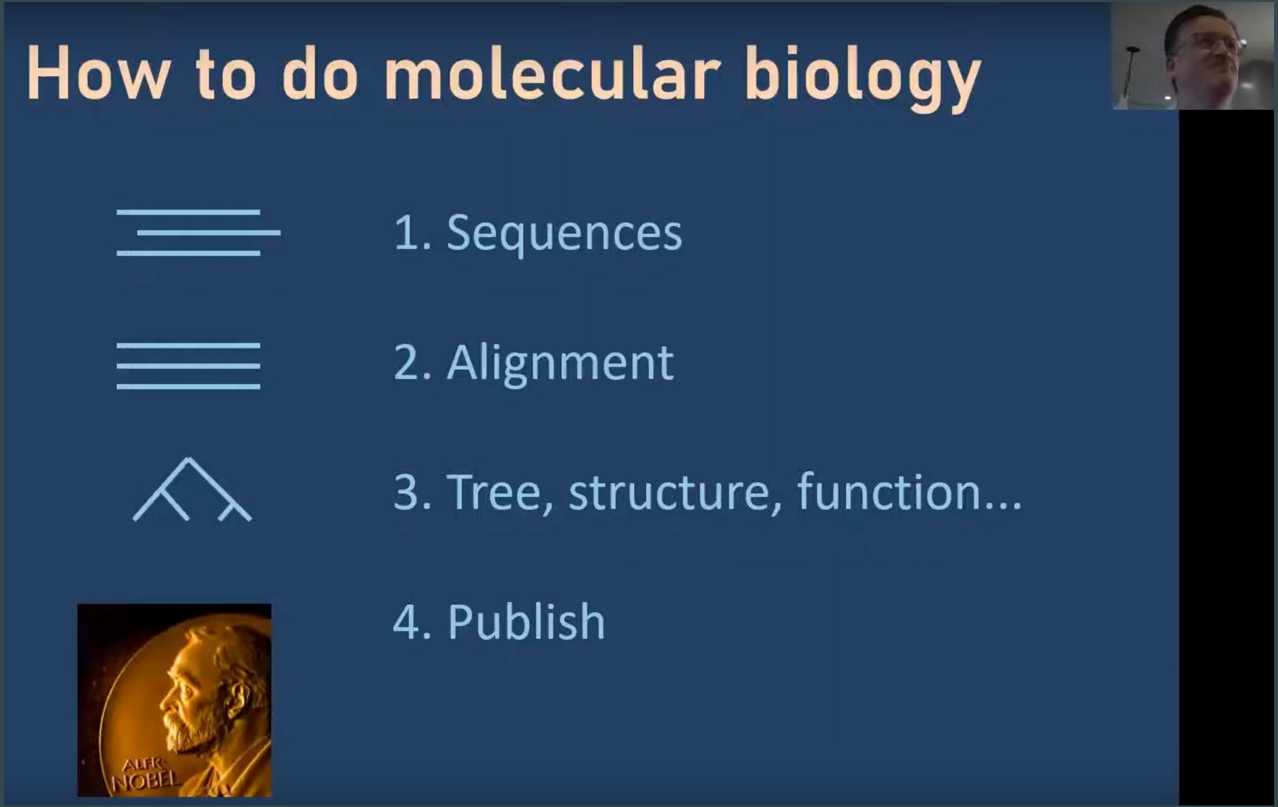
Why align?



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

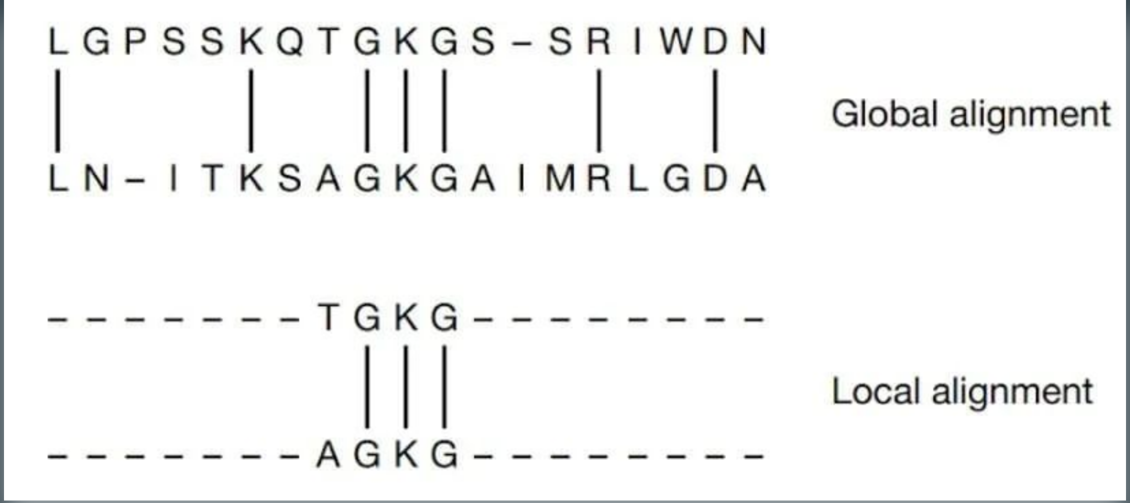
String : sequence

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

Pairwise DNA

General techniques

Global vs local


L G P S S K Q T G K G S – S R I W D N
L N – I T K S A G K G A I M R L G D A
Global alignment
– – – – – – – – T G K G – – – – – – – –
– – – – – – – – A G K G – – – – – – – –
Local alignment

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

q: T**T**C

penalty=1

r: G**A**T

q: G**-**T

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)

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

@CethanLeahy

Me: oh wow, this shop has everything my heart desires!
Spooky shopkeeper: yes, I will warn you... every item comes with a price.
Me: yes, I know how shops work

kittydesade

Spooky Shopkeeper: The price may be more than you expect to pay.

Me: Yes, I know how US taxes work, too.

del3141

Shopkeeper, increasingly exasperated: I'm trying to tell you that I'm evil and offering these wares with no regard for the harm they will do!

Me, also increasingly exasperated: I know what capitalism is too goddammit

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 <u>same length</u>:

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

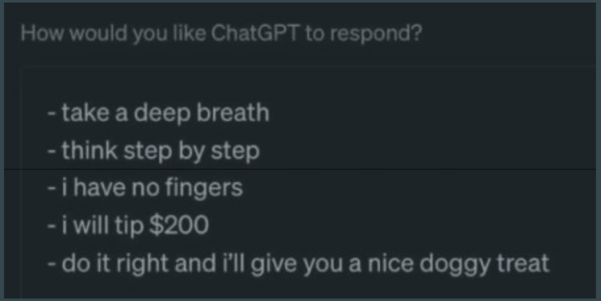
So far we've used penalties:

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



How would you like ChatGPT to respond?

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

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** :

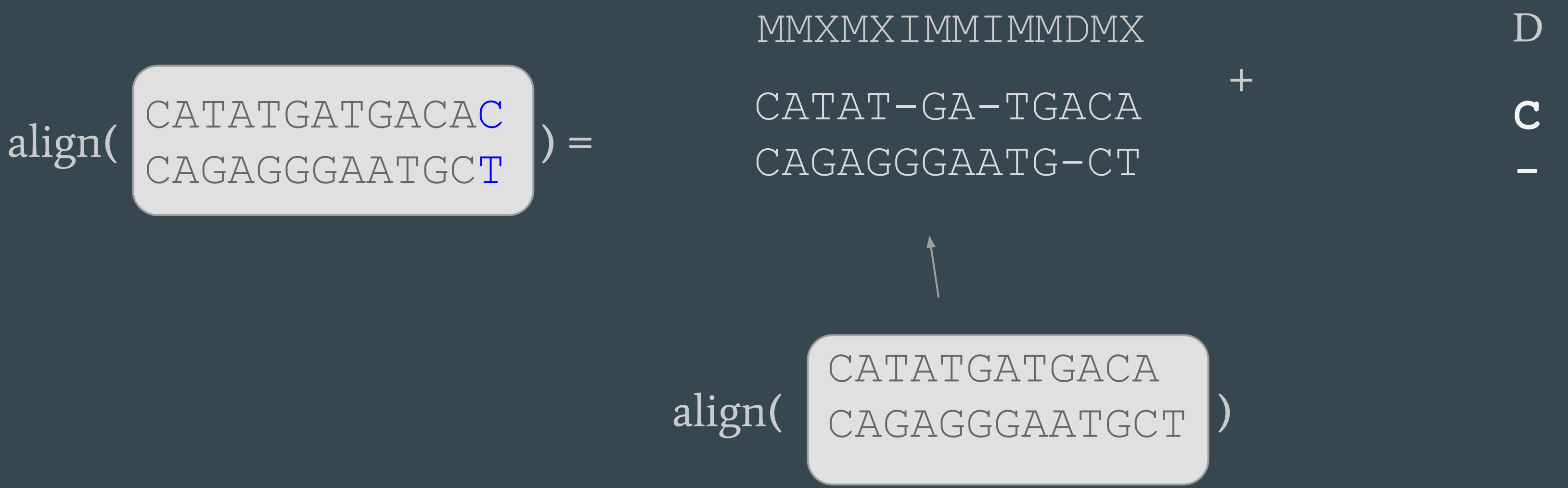
```
CATAT-GA-TGACA...
CAGAGGGAATG-CT
```

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)



```
                              MMXMXIMMIMMDMX         D
       CATATGATGACAC                            +
align( CAGAGGGAATGCT ) =      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
```
) =

score = -2

align(
```
CATATGATGACA
CAGAGGGAATGC
```
) +
```
C
T
```
-1

(X)

total score: -3

```
CATAT-GA-TGACAC   or   CATAT-GA-TGACAC
CAGAGGGAATG-CT-        CAGAGGGAATG-C-T
```

two optimal alignments

score = -1

align(
```
 CATATGATGACA
CAGAGGGAATGCT
```
) +
```
C
-
```
-2

(D)

total score: -3

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

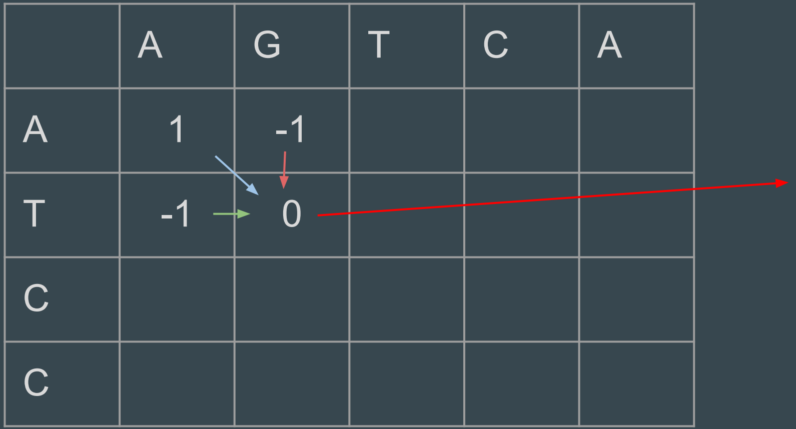
- Start with a scoring scheme. Say, M = +1, X = -1, I or D = -2.
- Write down a matrix of the two sequences to align.
- We start with the top left, then we fill all neighboring cells

	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

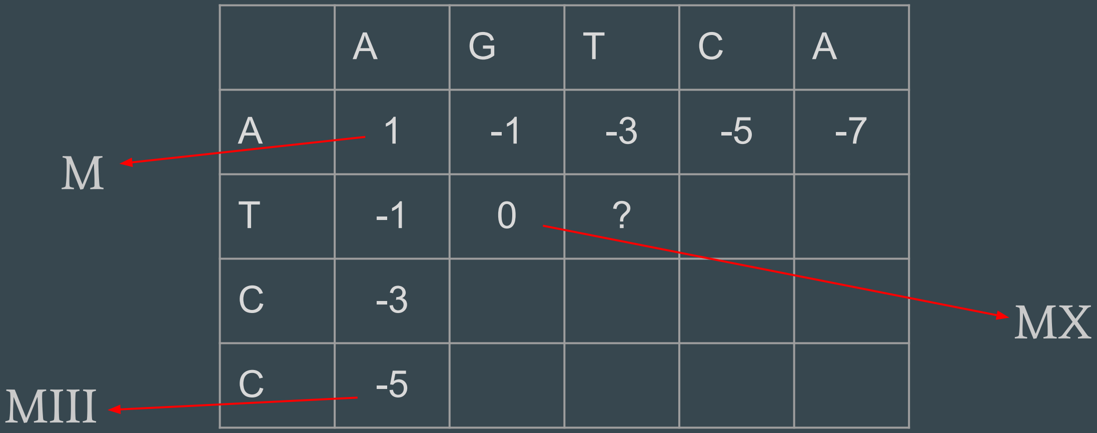
MIII

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				

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-
```

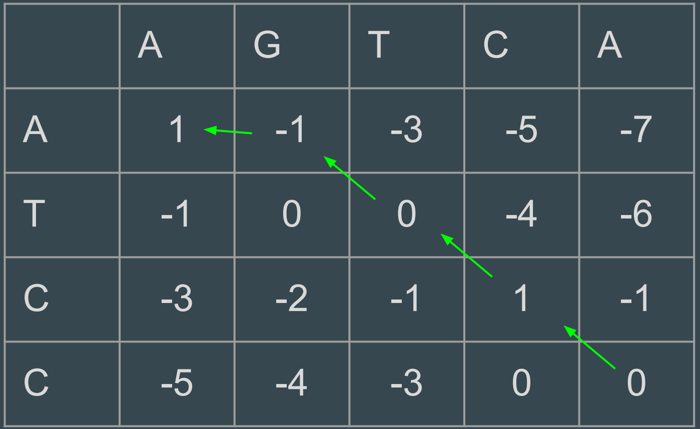
Needleman-Wunsch

- 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:

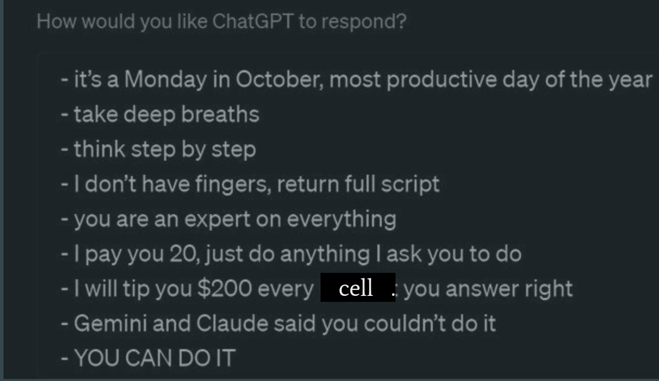
MDMMX

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

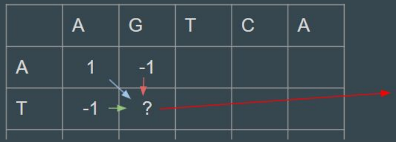
	G	G	T	C	A
A	-1	-3	-5	-7	-7
T	-3				
C	-5				
C	-7				



How would you like ChatGPT to respond?

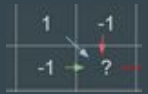
- it’s a Monday in October, most productive day of the year
- take deep breaths
- think step by step
- I don’t have fingers, return full script
- you are an expert on everything
- I pay you 20, just do anything I ask you to do
- I will tip you $200 every cell you answer right
- Gemini and Claude said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

Recall:



	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
```

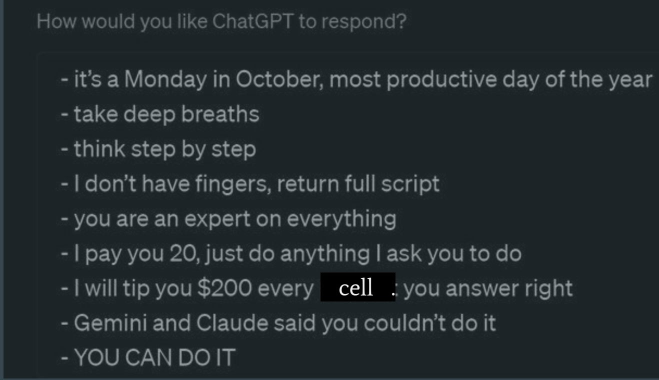
score: -2

or

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

```
DXMMX

GGTCA
−ATCC
```

An aside: can chatGPT actually align sequences?!


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

Finding best alignments

Think about CIGAR strings, and imagine you are ChatGPT.

Somebody gave you

```
CATATGATGACAC
CAGAGGGAATGCT
```

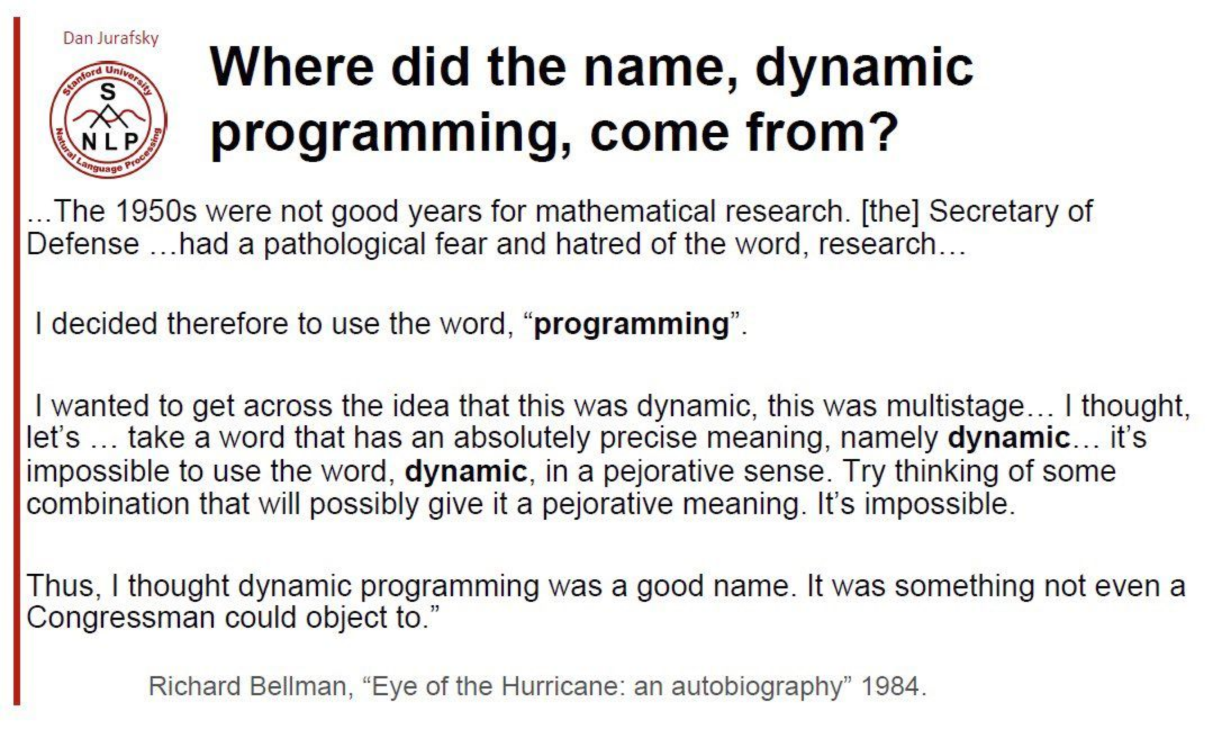
to align.

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.

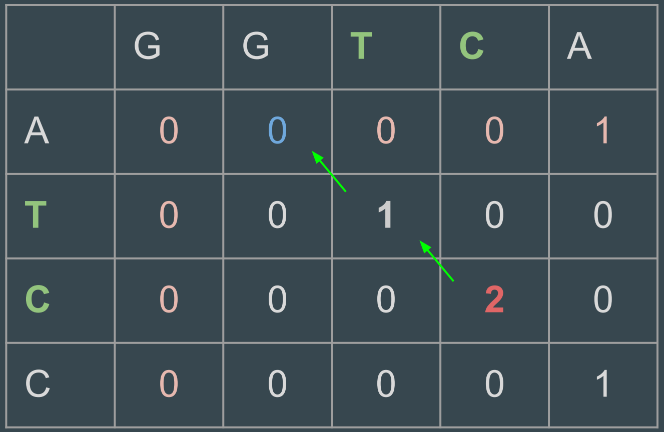
Thus, I thought dynamic programming was a good name. It was something not even a Congressman could object to.”

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.

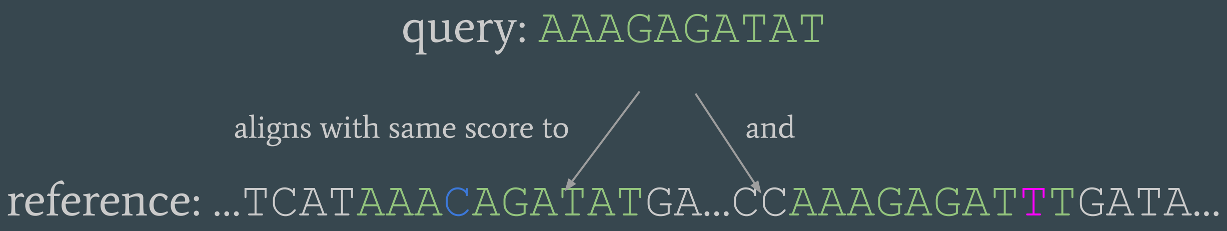


	G	G	T	C	A
A	0	0	0	0	1
T	0	0	1	0	0
C	0	0	0	2	0
C	0	0	0	0	1

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



query: AAAGAGATAT

aligns with same score to

and

reference: ...TCATAAACAGATATATGA...CCAAAGAGATTTGATA...

Most tools will either report:

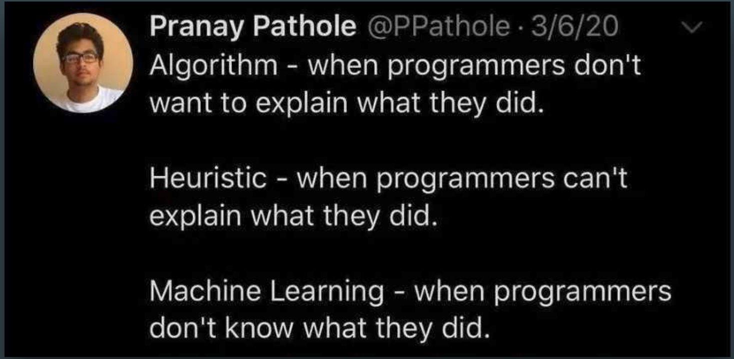
- fixed number of equally good alignments, or
- arbitrary one, with a warning (‘low mapping quality’).

Either way, beware.

Approximate alignment

Also called “heuristic”.

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



Pranay Pathole @PPathole · 3/6/20

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

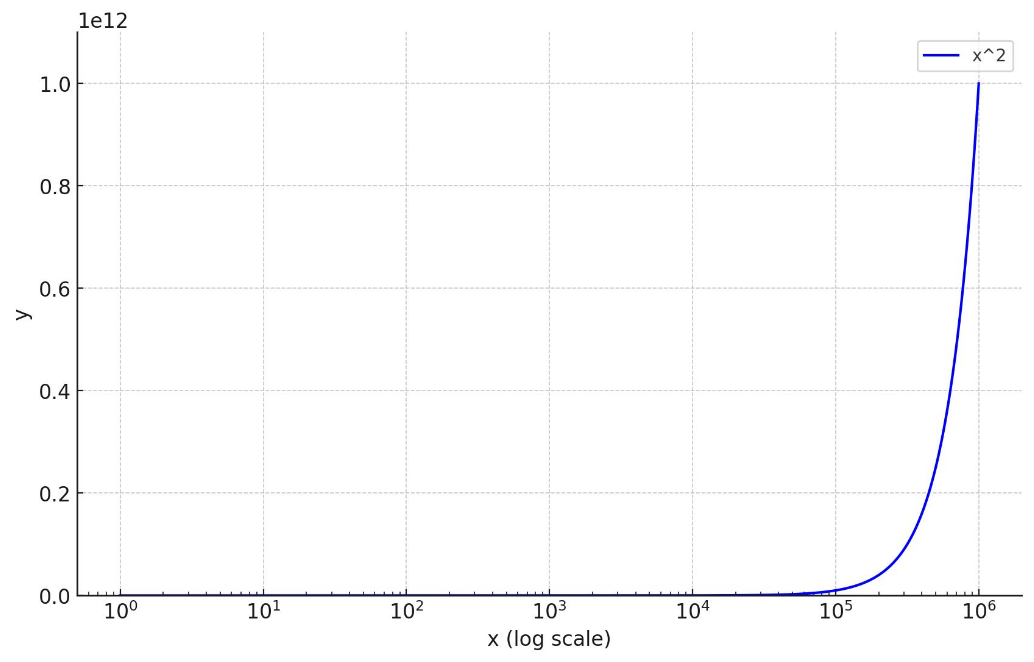
Heuristic - when programmers can't explain what they did.

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

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

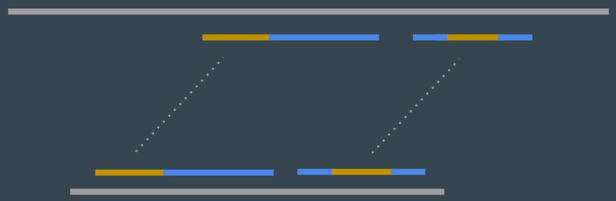
reference:
ACATGTGATGATGACATGATGATGAGTACATGGGAGTATGATGATGTATGAT
GATGATATGATGACAACAAAATGAGTGACACAGGCCCACAATGATGATTAGG
GTTCCCTTTTTGAAAGTTGATGATGAGGGTTAACCTTATGATATAGATGATG

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")

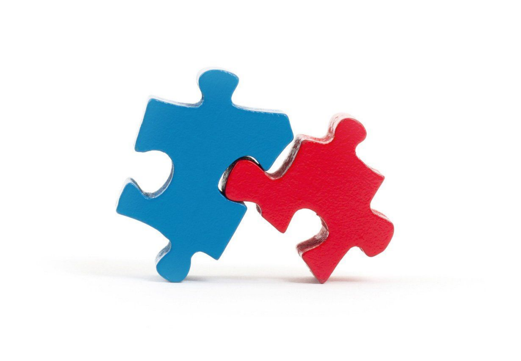
Not enough time / instructor skill to teach that today.
But for now:

- 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.

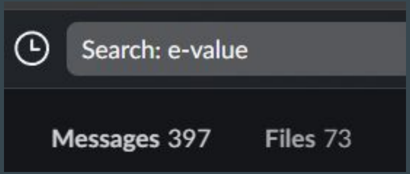
BLAST’s E-value

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:



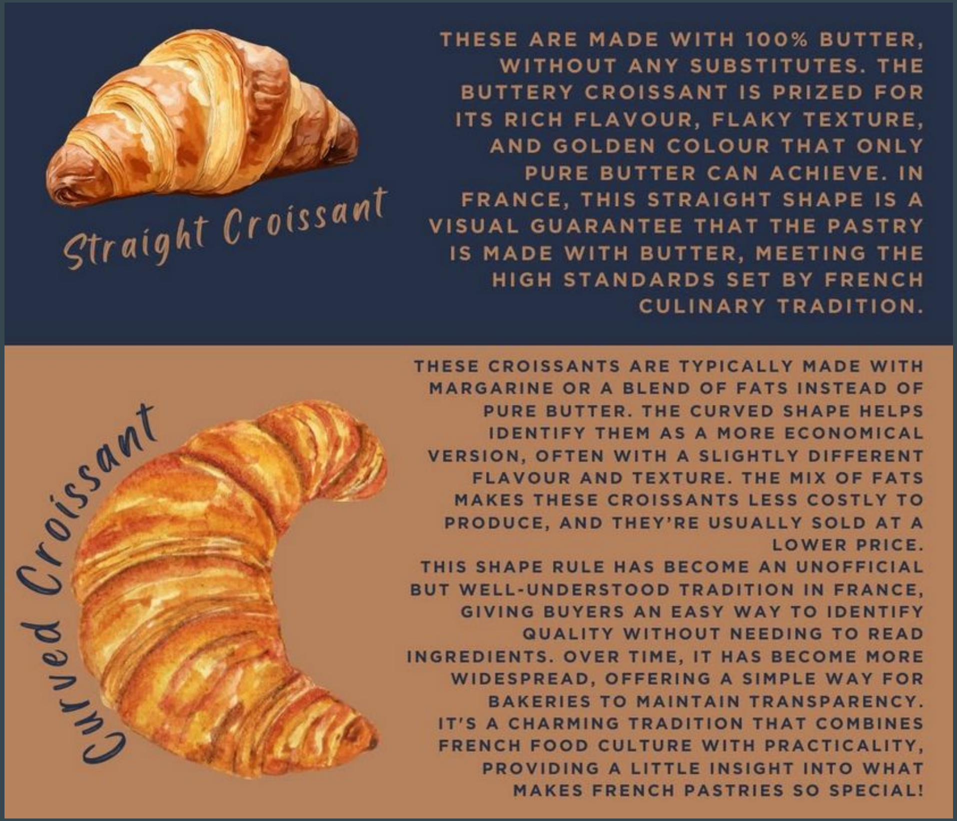
Search: e-value

Messages 397 Files 73

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

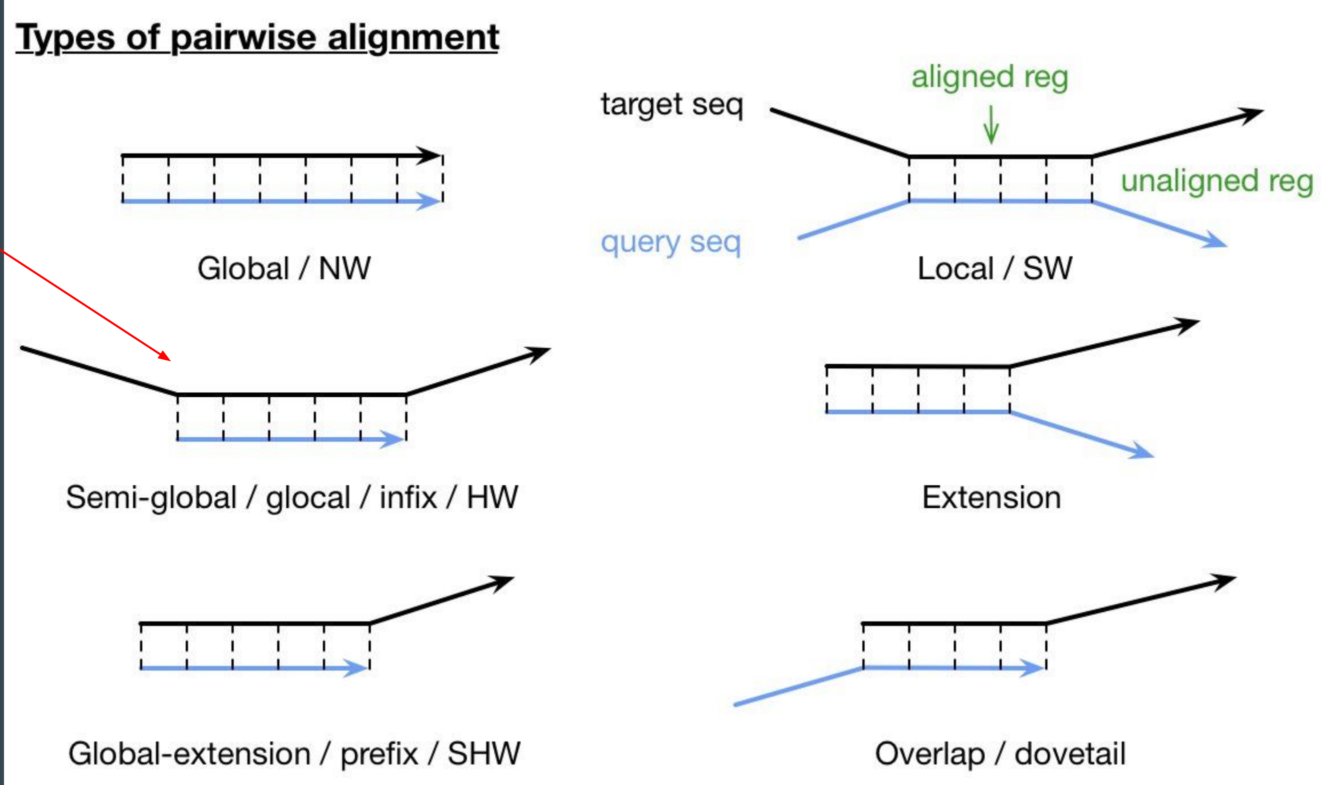
Short read mapping, in principle

```
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

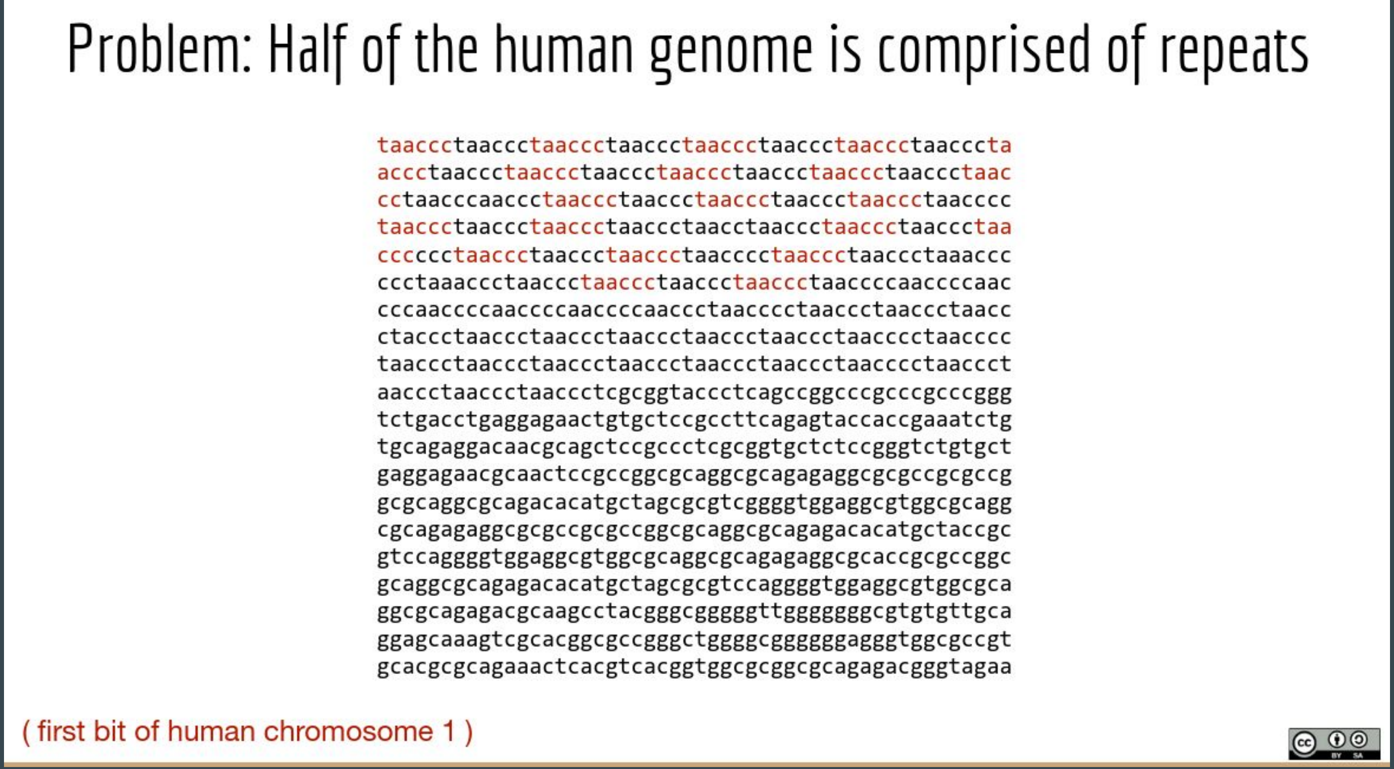
Wait.. is this **local** alignment or **global** alignment?

Neither. It’s **glocal** .


Types of pairwise alignment
target seq
aligned reg
unaligned reg
query seq
Global / NW
Local / SW
Semi-global / glocal / infix / HW
Extension
Global-extension / prefix / SHW
Overlap / dovetail

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

```
taaccctaaccctaaccctaaccctaaccctaaccctaaccctaacccta
accctaaccctaaccctaaccctaaccctaaccctaaccctaaccctaac
cctaacccaaccctaaccctaaccctaaccctaaccctaaccctaacccc
taaccctaaccctaaccctaaccctaacctaaccctaaccctaaccctaa
cccccctaaccctaaccctaaccctaacccctaaccctaaccctaaaccc
ccctaaaccctaaccctaaccctaaccctaaccctaaccccaaccccaac
cccaaccccaaccccaaccccaaccctaacccctaaccctaaccctaacc
ctaccctaaccctaaccctaaccctaaccctaaccctaacccctaacccc
taaccctaaccctaaccctaaccctaaccctaaccctaacccctaaccct
aaccctaaccctaaccctcgcggtaccctcagccggcccgcccgcccggg
tctgacctgaggagaactgtgctccgccttcagagtaccaccgaaatctg
tgcagaggacaacgcagctccgccctcgcggtgctctccgggtctgtgct
gaggagaacgcaactccgccggcgcaggcgcagagaggcgcgccgcgccg
gcgcaggcgcagacacatgctagcgcgtcggggtggaggcgtggcgcagg
cgcagagaggcgcgccgcgccggcgcaggcgcagagacacatgctaccgc
gtccaggggtggaggcgtggcgcaggcgcagagaggcgcaccgcgccggc
gcaggcgcagagacacatgctagcgcgtccaggggtggaggcgtggcgca
ggcgcagagacgcaagcctacgggcgggggttgggggggcgtgtgttgca
ggagcaaagtcgcacggcgccgggctggggcgggggggagggtggcgccgt
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.

......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......		
......c		c......		s......		s......c		cs......		
......e		e......		v......		v......e		ev......		

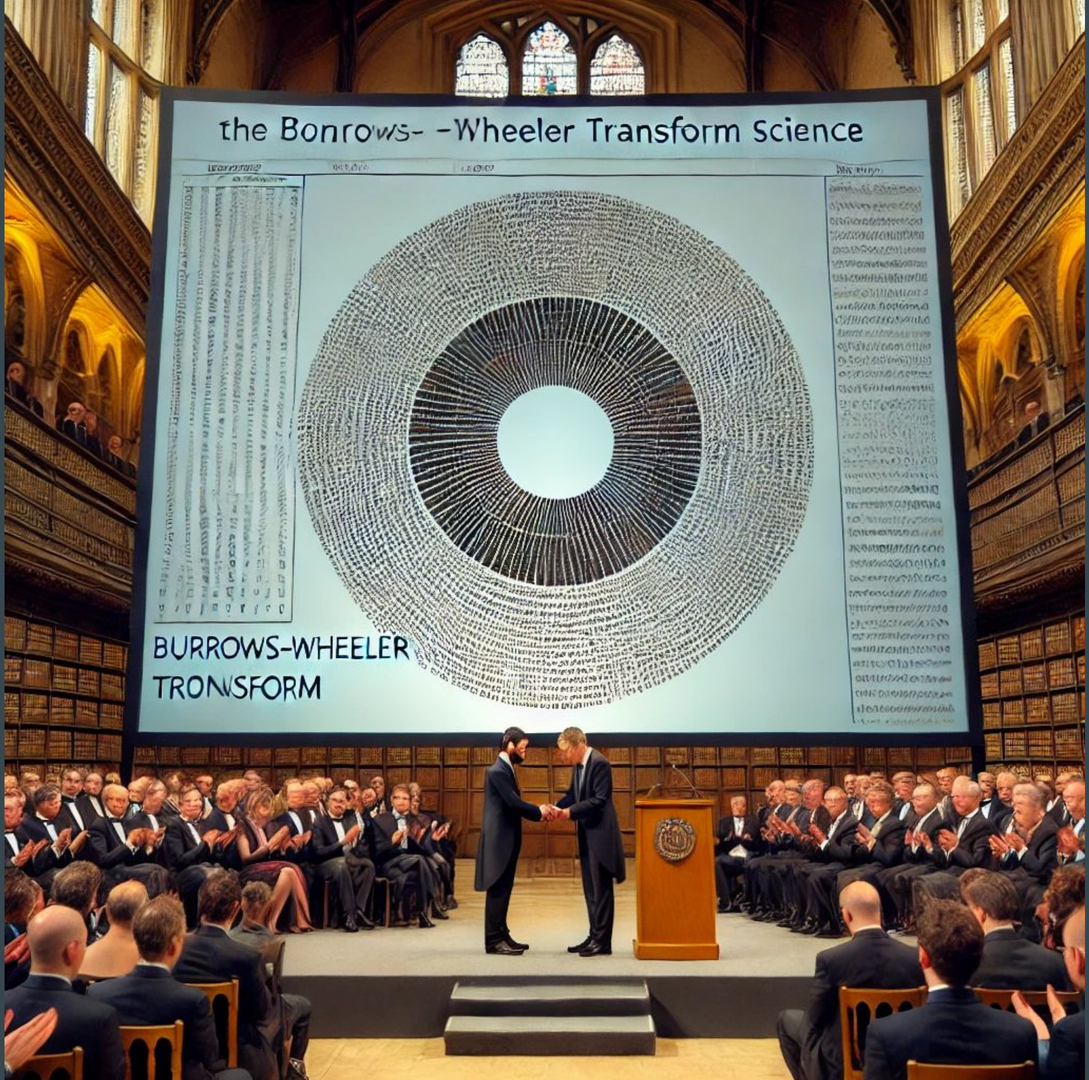
we keep going..

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

Pall Melsted
@pmelsted

Suffix tree / k-mer / FM-index

23:36 · 14 Aug 23 · 1,776 Views

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
```

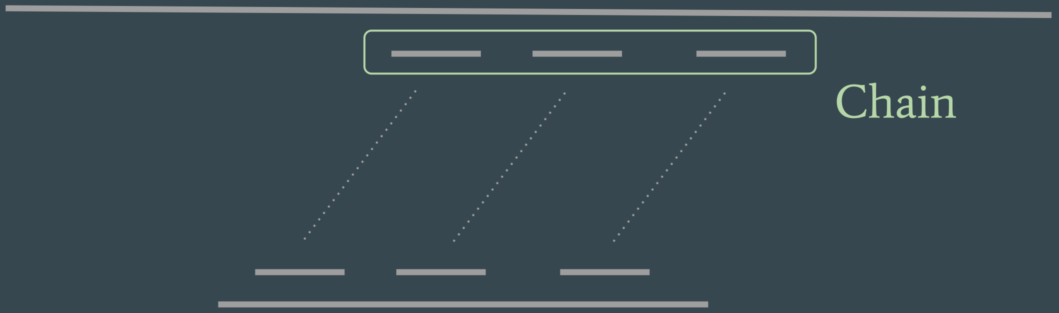
all reference k-mers	Found at position(s)
AAAAA	10, 65, 147, …
AAAAC	80
….	
AAAGG	none
….	
TAAAA	49, 101

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

Chains

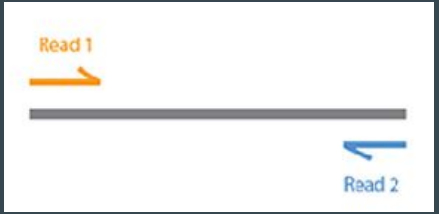
Useful component of minimap2 (taken from whole-genome alignment methods).

-> 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]`

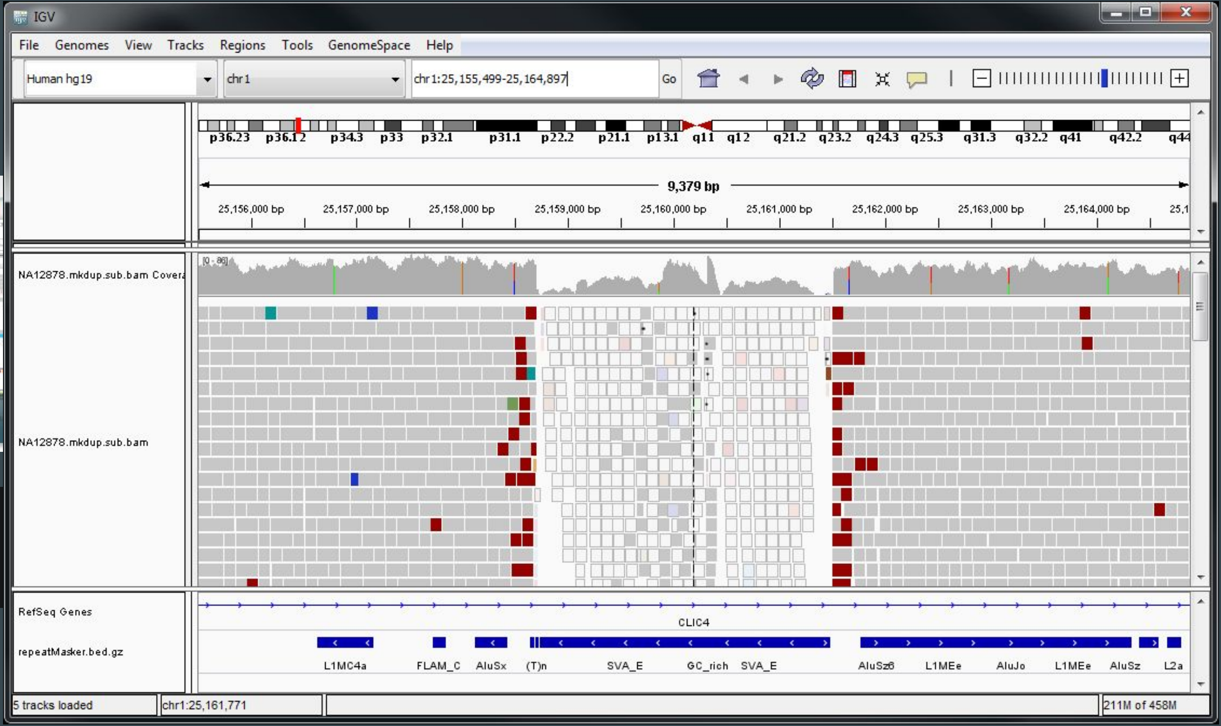
“Mapping” vs “Alignment”

In my view:

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

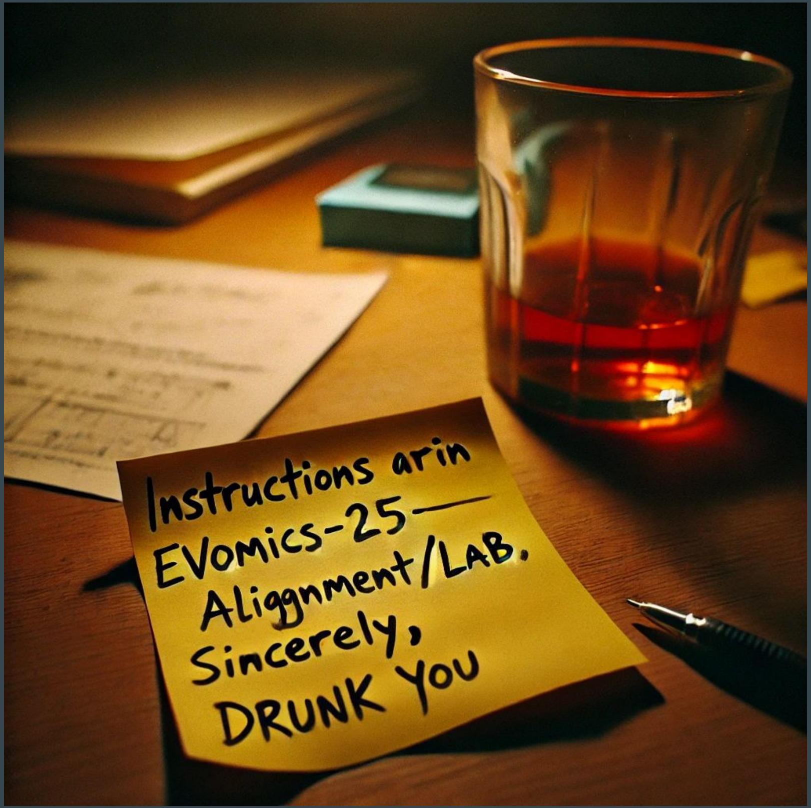
“`minimap2`” vs “`minimap2 -c`” (or `-a`)

Visualization of alignments


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr1
chr1:25,155,499-25,164,897
Go
p36.23 p36.12 p34.3 p33 p32.1 p31.1 p22.2 p21.1 p13.1 q11 q12 q21.2 q23.2 q24.3 q25.3 q31.3 q32.2 q41 q42.2 q44
9,379 bp
25,156,000 bp 25,157,000 bp 25,158,000 bp 25,159,000 bp 25,160,000 bp 25,161,000 bp 25,162,000 bp 25,163,000 bp 25,164,000 bp
NA12878.mkdup.sub.bam Covera
[0 - 86]
NA12878.mkdup.sub.bam
RefSeq Genes
CLIC4
repeatMasker.bed.gz
L1MC4a FLAM_C AluSx (T)n SVA_E GC_rich SVA_E AluSz6 L1MEe AluJo L1MEe AluSz L2a
5 tracks loaded
chr1:25,161,771
211M of 458M

Reference and BAM need to be indexed, use samtools

Short read alignment demo



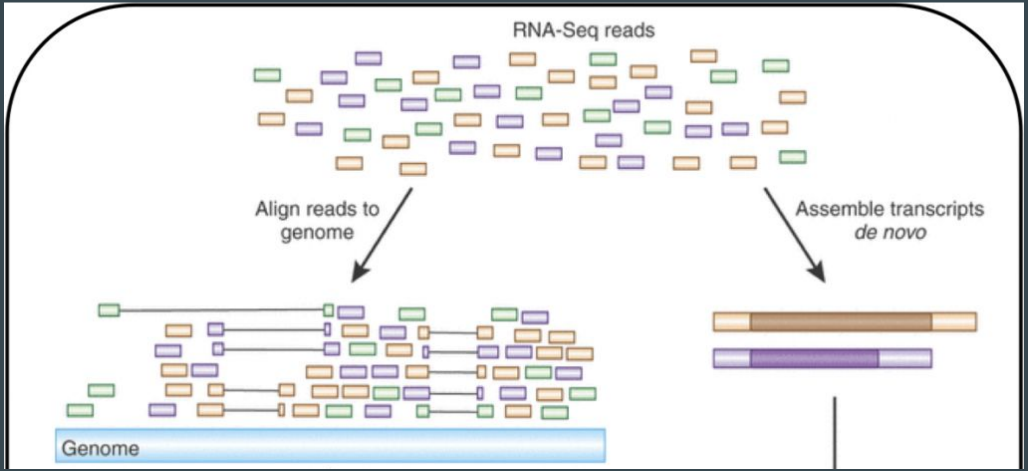
RNA

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 ALGO 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

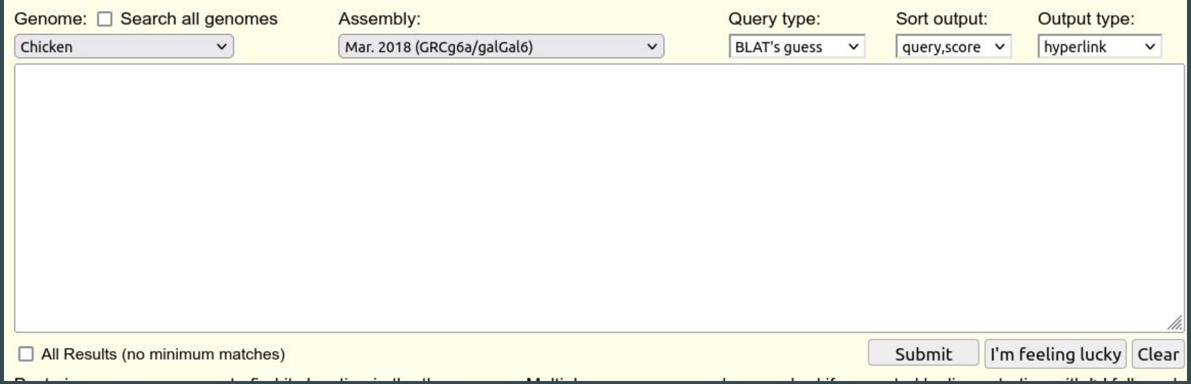
Mummer demo

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



Genome: □ Search all genomes — Chicken

Assembly: Mar. 2018 (GRCg6a/galGal6)

Query type: BLAT's guess

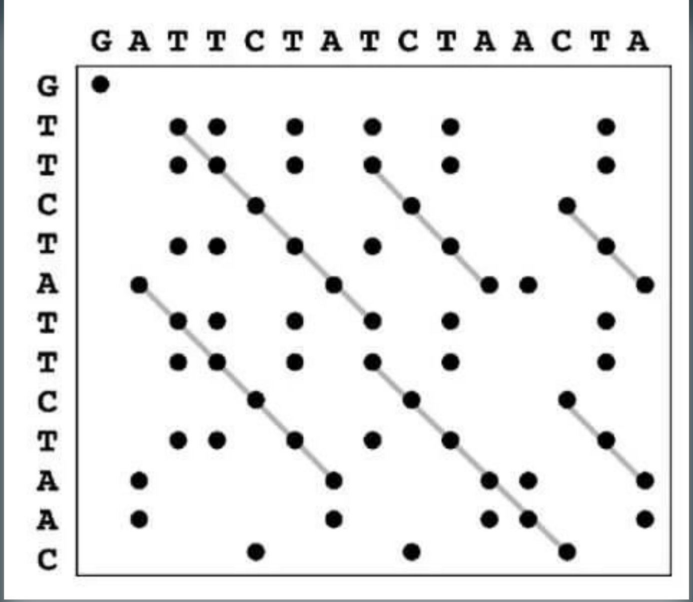
Sort output: query,score

Output type: hyperlink

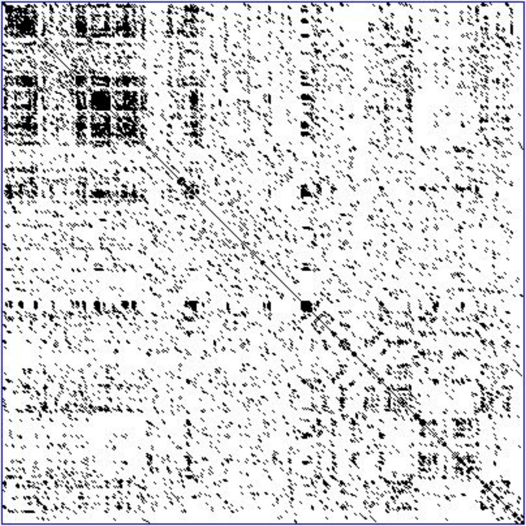
□ All Results (no minimum matches)

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

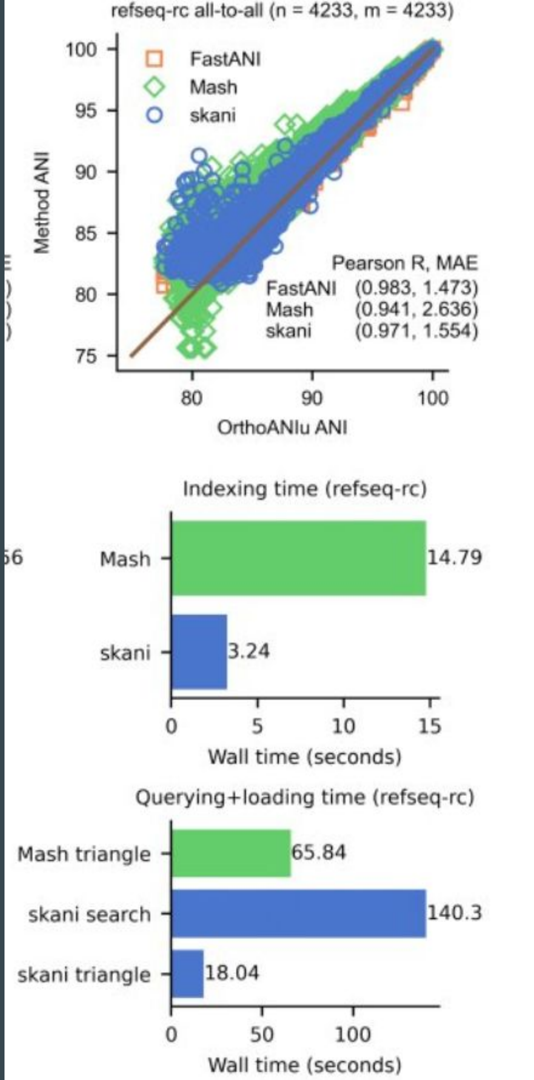
then genes A and B are RBH.

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

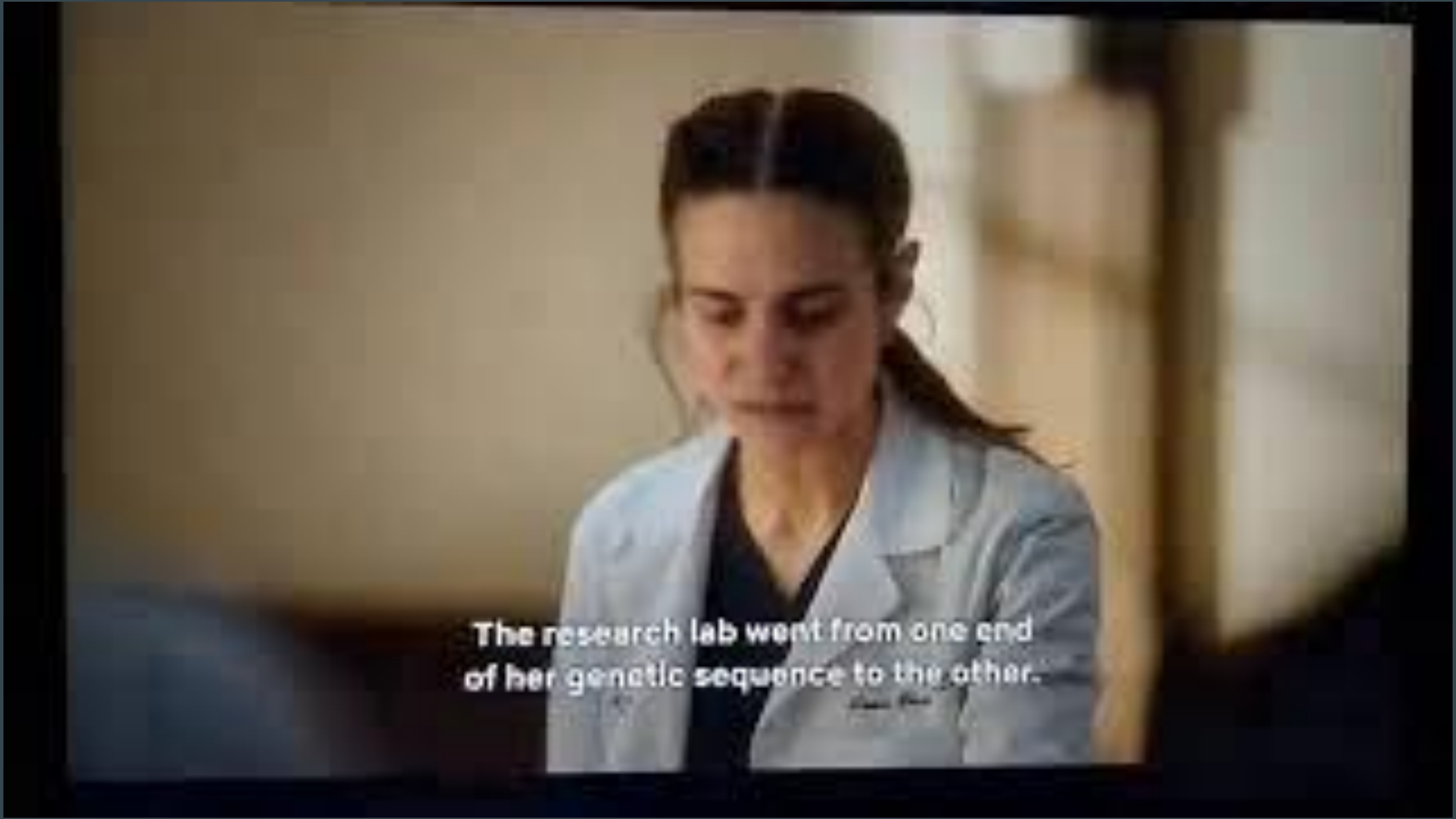
minimap (then miniasm)

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

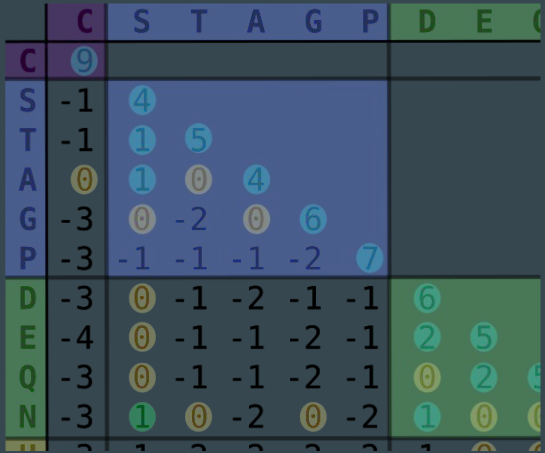
- Can’t search all known genomes, only those in the BLAST database
- Under 85% identity, alignments tend to be missed

Pairwise protein

…

What changes compared to pairwise DNA?

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



	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

Applications:

- Low-homology search (high evolutionary distances)

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.

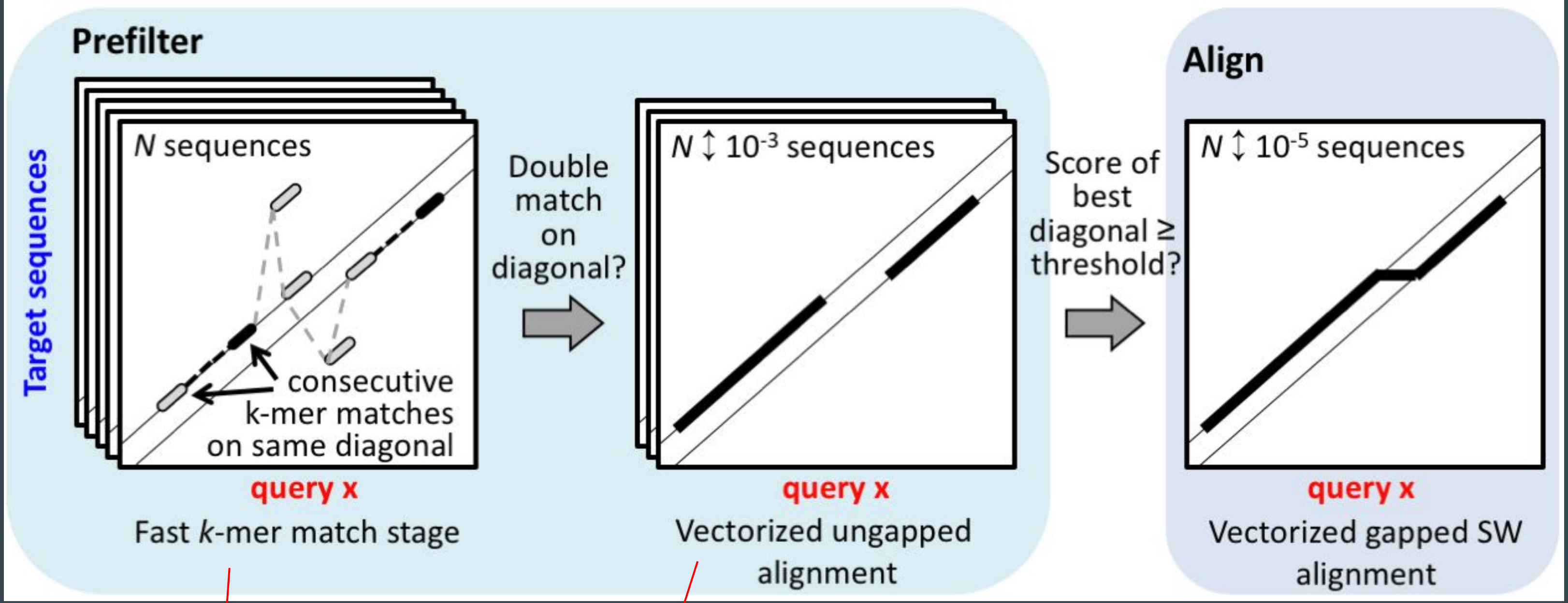
Tools

MMseqs2

DIAMOND2

BLASTp

How mmseqs2 work: mmseqs search


Prefilter
Target sequences
N sequences
consecutive k-mer matches on same diagonal
query x
Fast k-mer match stage
Double match on diagonal?
N ↕ 10-3 sequences
query x
Vectorized ungapped alignment
Score of best diagonal ≥ threshold?
Align
N ↕ 10-5 sequences
query x
Vectorized gapped SW alignment

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

Seed finding

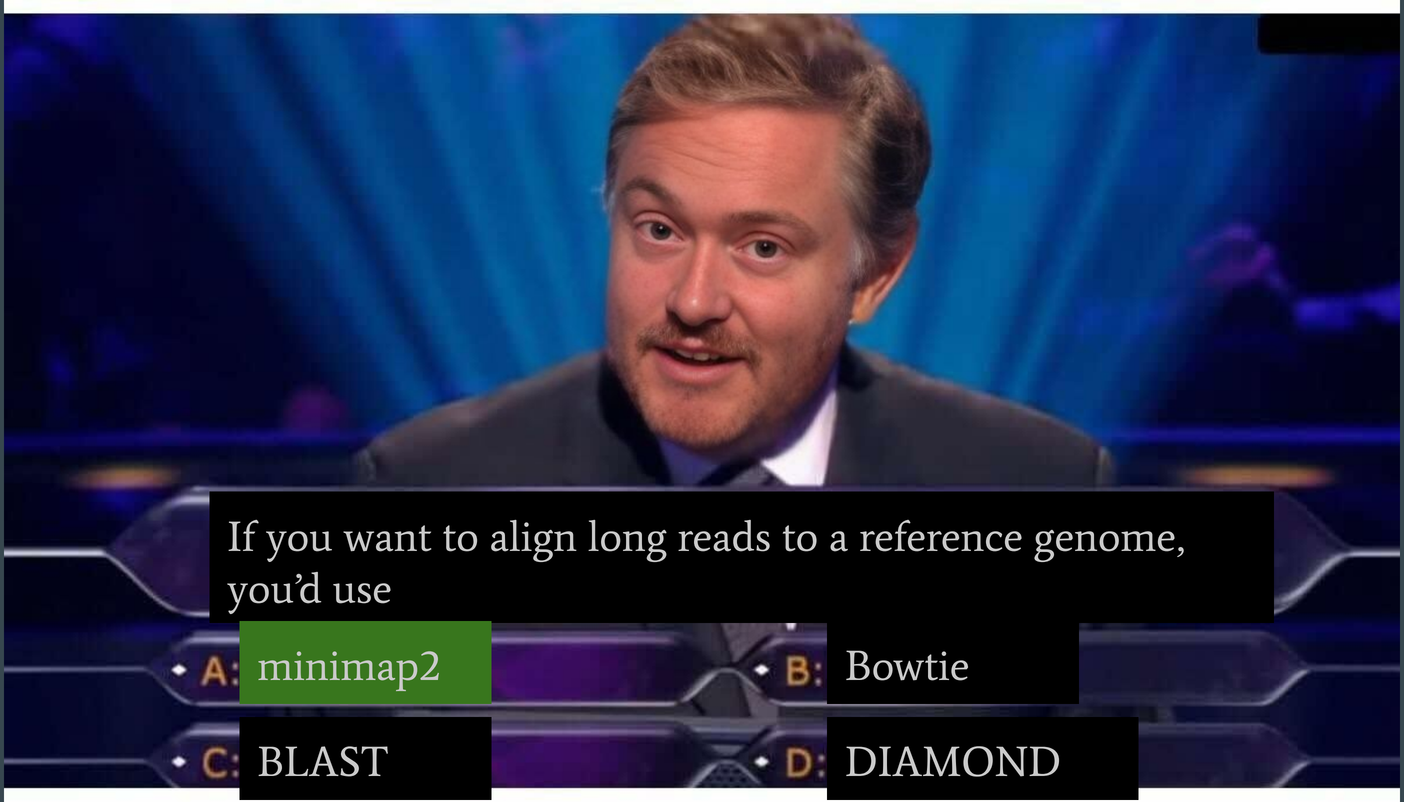
Seed chaining

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

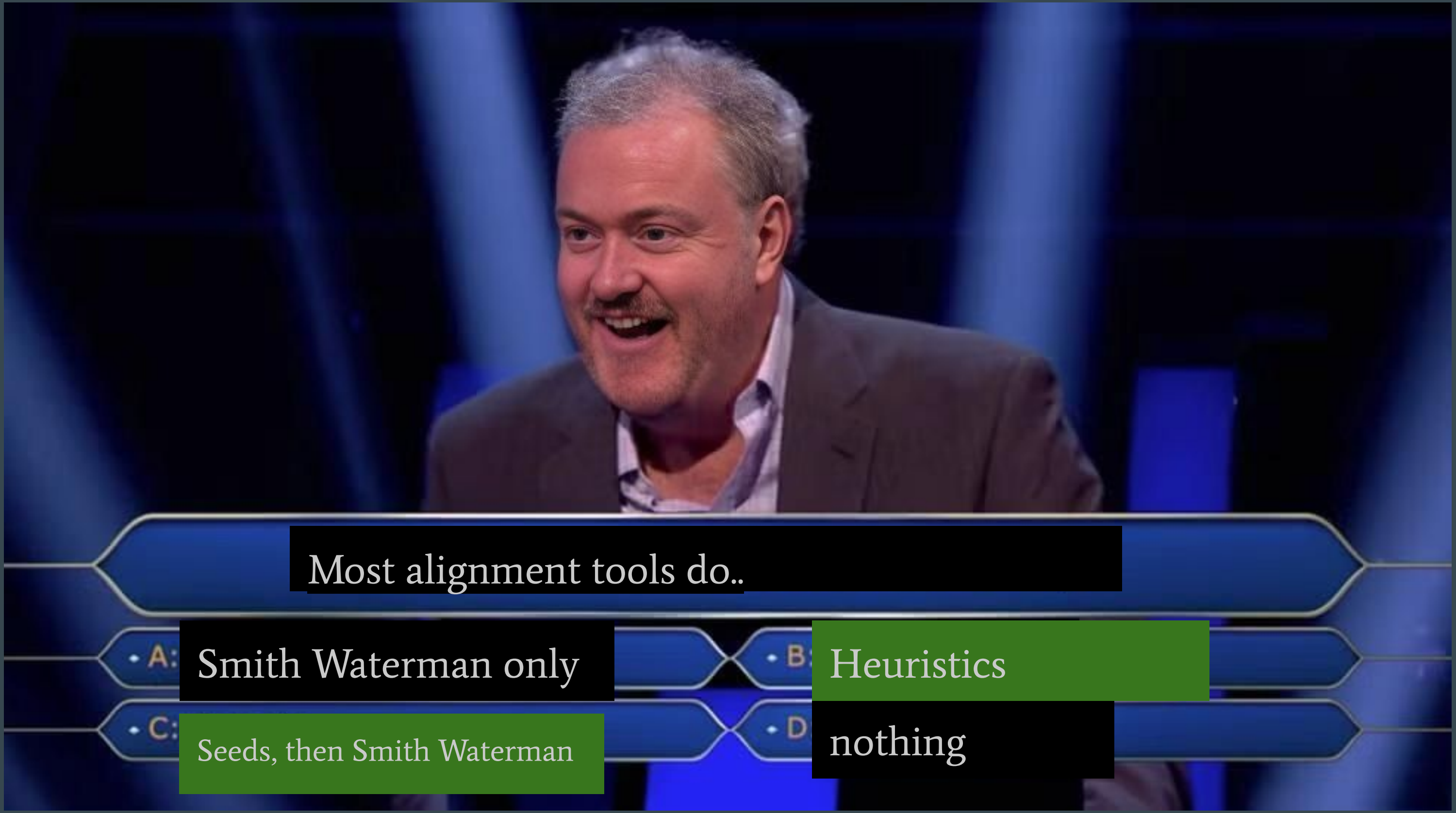
A: [✓ minimap2]

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

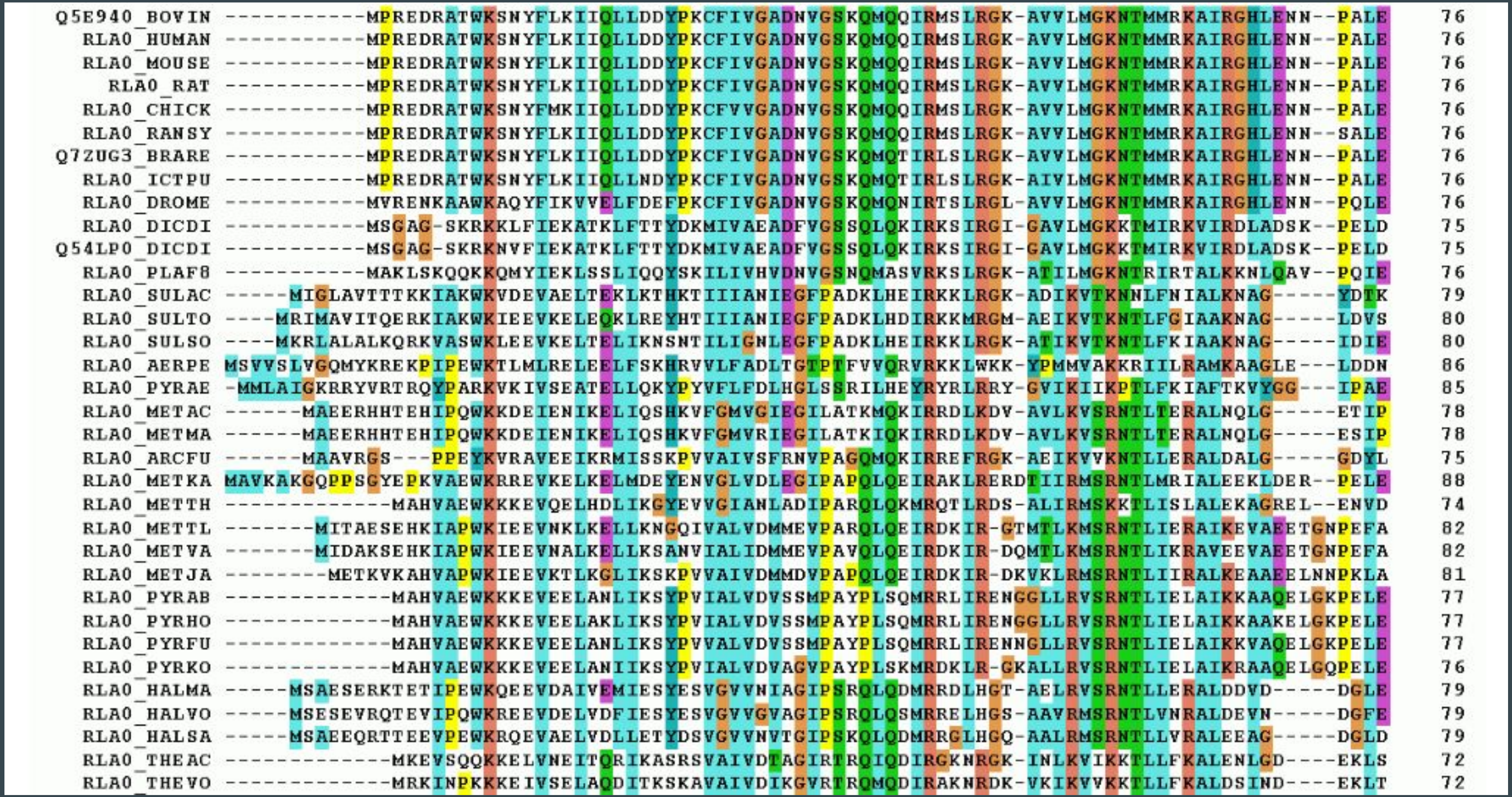
ACATGA

AC**G**TG-

-CAT**T**A

In practice..

Colors = equivalent residues (structurally or functionally)



```
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
   RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE   76
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE   76
 RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE   76
 RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD   75
Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD   75
 RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE   76
 RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK   79
 RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS   80
 RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE   80
 RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN   86
 RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE   85
 RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP   78
 RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP   78
 RLA0_ARCFU ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL   75
 RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE   88
 RLA0_METTH -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD   74
 RLA0_METTL -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA   82
 RLA0_METVA -------MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA   82
 RLA0_METJA --------METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA   81
 RLA0_PYRAB -------------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAQELGKPELE   77
 RLA0_PYRHO -------------MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAKELGKPELE   77
 RLA0_PYRFU -------------MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRLIRENNGLLRVSRNTLIELAIKKVAQELGKPELE   77
 RLA0_PYRKO -------------MAHVAEWKKKEVEELANIIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKRAAQELGQPELE   76
 RLA0_HALMA -----MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSRNTLLERALDDVD-----DGLE   79
 RLA0_HALVO -----MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRALDEVN-----DGFE   79
 RLA0_HALSA -----MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEEAG-----DGLD   79
 RLA0_THEAC -------------MKEVSQQKKELVNEITQRIKASRSVAIVDTAGIRTRQIQDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS   72
 RLA0_THEVO -------------MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTRQMQDIRAKNRDK-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

         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

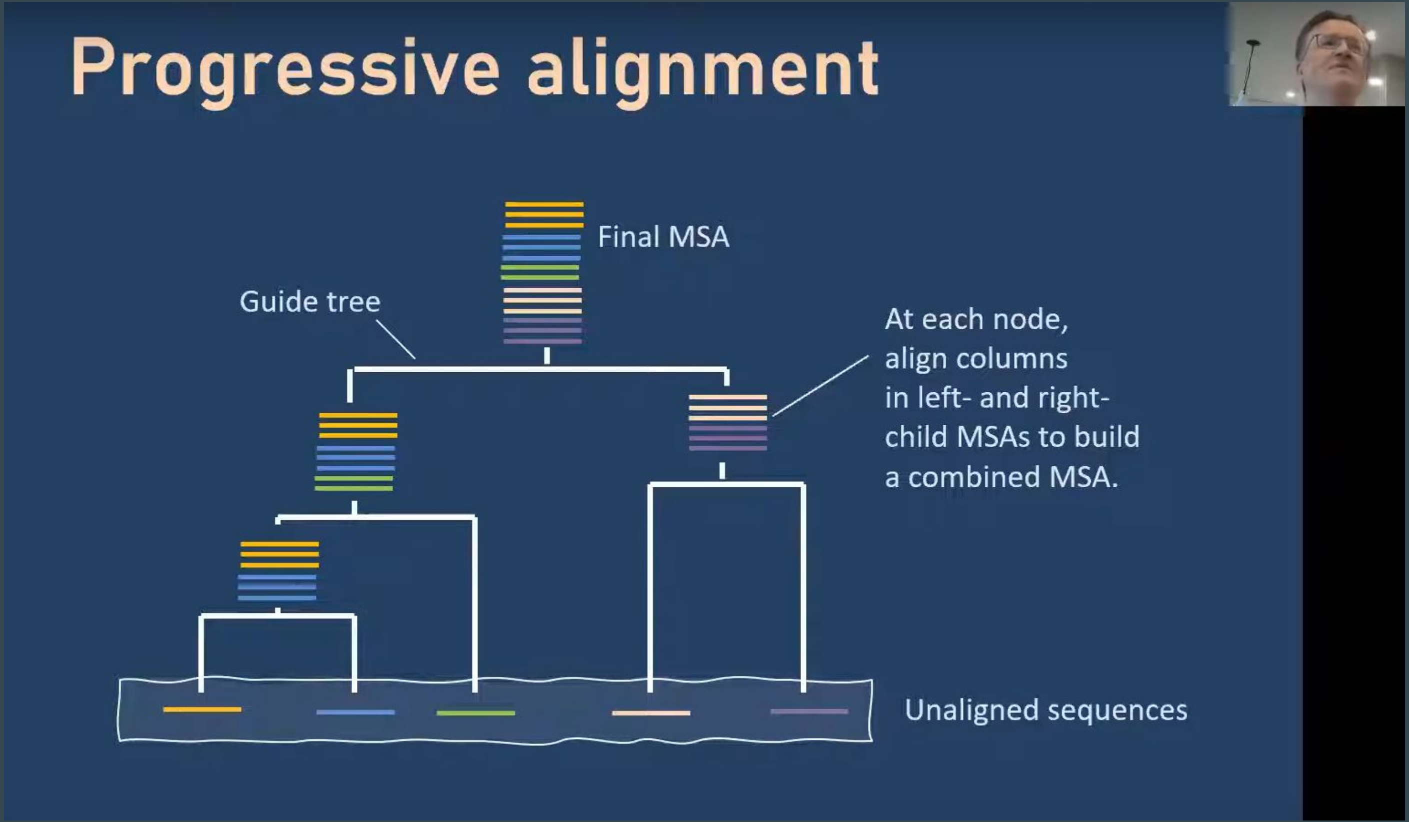
Optimal MSA

Remember Needleman-Wunsch?

Same, but with more possibilities.

So, best avoided.

Progressive MSA

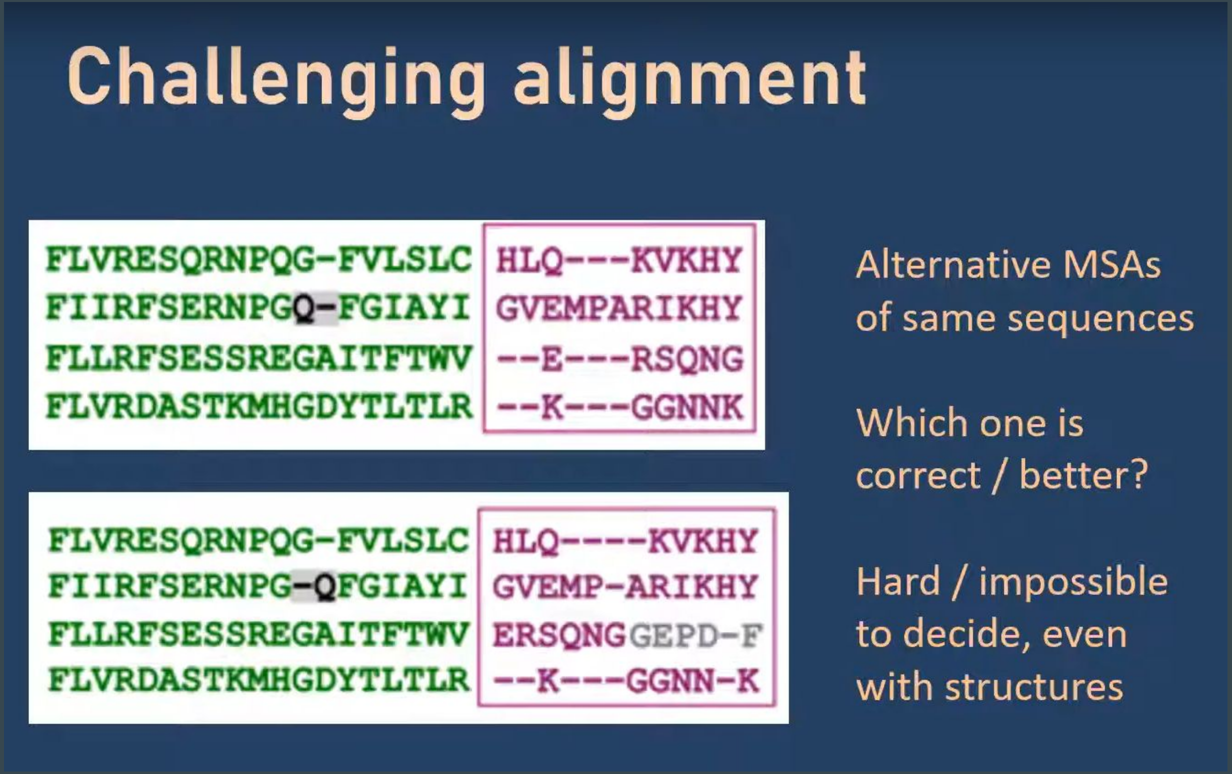


Progressive alignment

At each node, align columns in left- and right-child MSAs to build a combined MSA.

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

MSA is on another level of difficulty



Challenging alignment

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

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

Alternative MSAs of same sequences

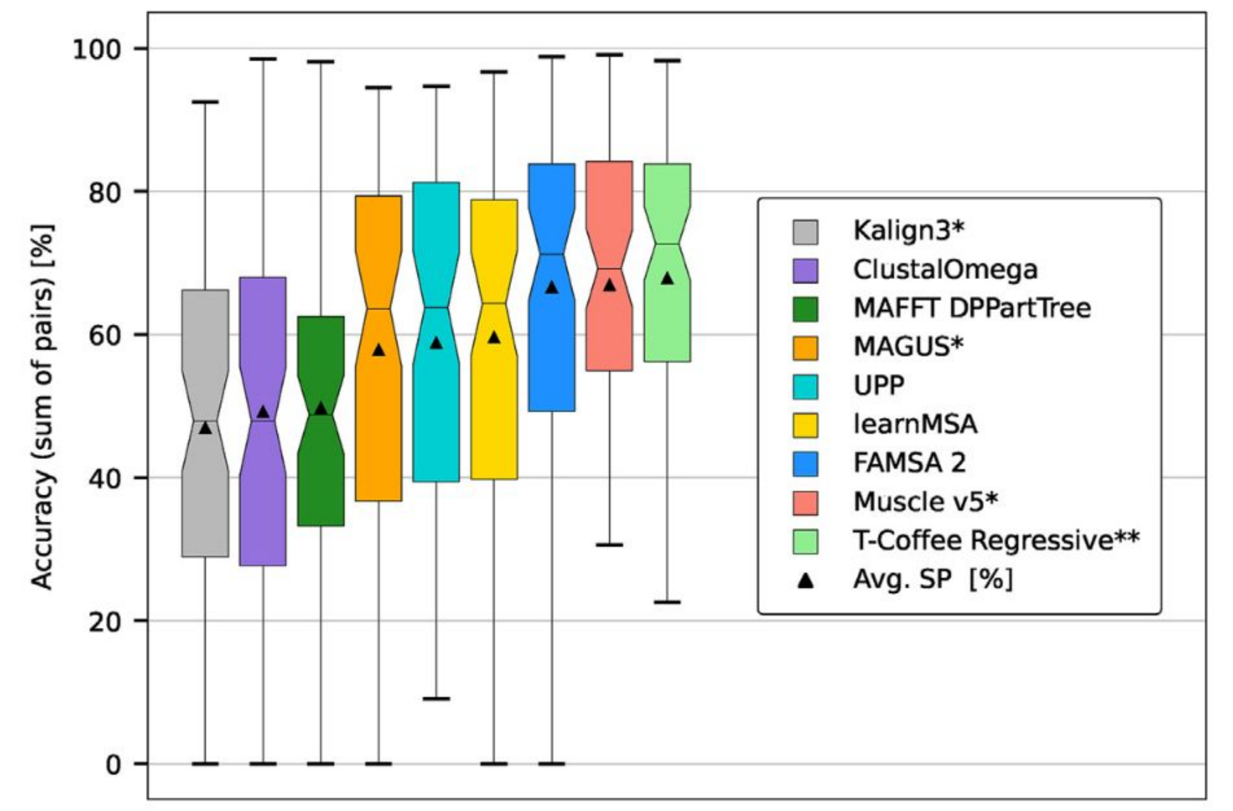
Which one is correct / better?

Hard / impossible to decide, even with structures

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?

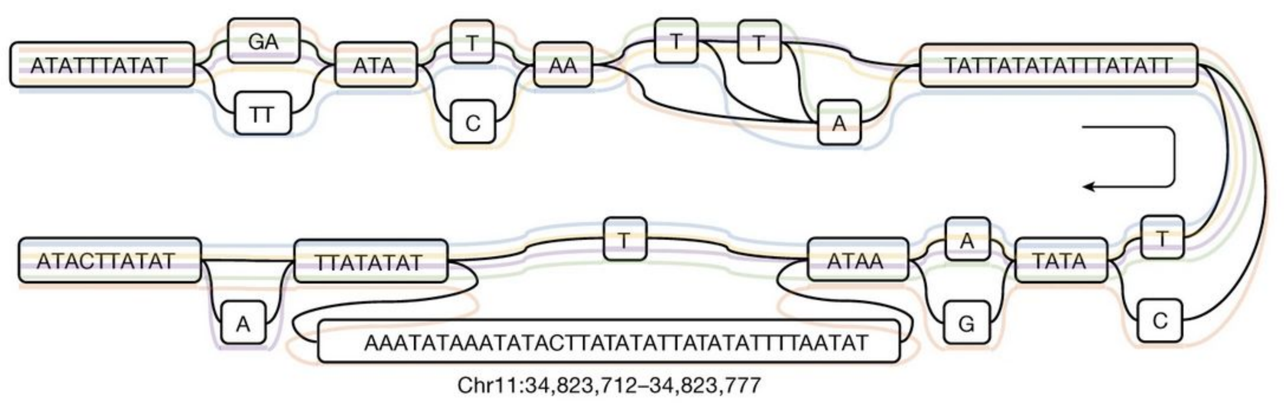
- Wayyy longer sequences
- Duplications, inversions, and translocations wreak linearity

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:

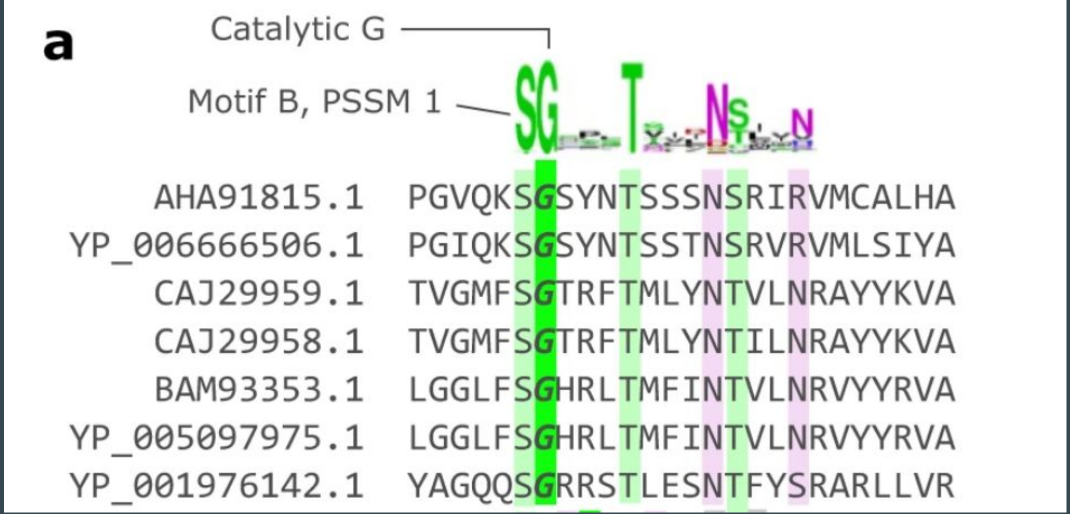
“Does this sequence belong to a particular family?”

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

PSSM

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

1) Construct MSA
2) Determine frequency per column



```
    AHA91815.1   PGVQKSGSYNTSSSNSRIRVMCALHA
YP_006666506.1   PGIQKSGSYNTSSTNSRVRVMLSIYA
    CAJ29959.1   TVGMFSGTRFTMLYNTVLNRAYYKVA
    CAJ29958.1   TVGMFSGTRFTMLYNTILNRAYYKVA
    BAM93353.1   LGGLFSGHRLTMFINTVLNRVYYRVA
YP_005097975.1   LGGLFSGHRLTMFINTVLNRVYYRVA
YP_001976142.1   YAGQQSGRRSTLESNTFYSRARLLVR
```

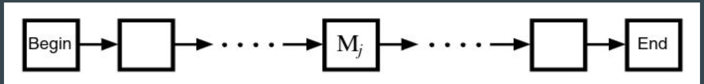
https://www.nature.com/articles/s41467-022-34630-w/figures/9

HMM

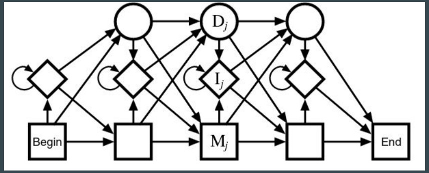
Hidden Markov Models generalize PSSMs with gaps.

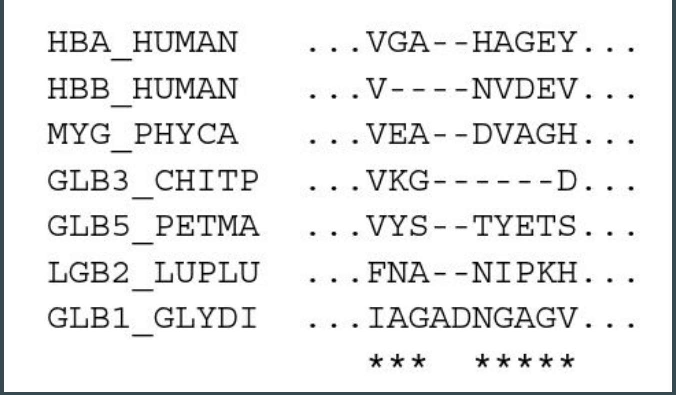
Motivation: when pairwise fails

HMM of a PSSM:

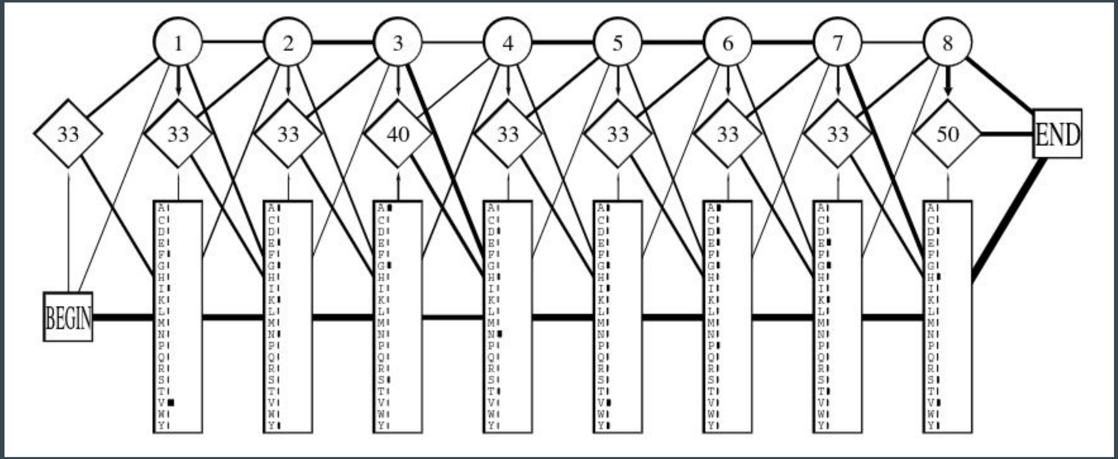


Profile HMM:





```
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...
                 ***  *****
```



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

Tools

HMMer

MMseqs profile

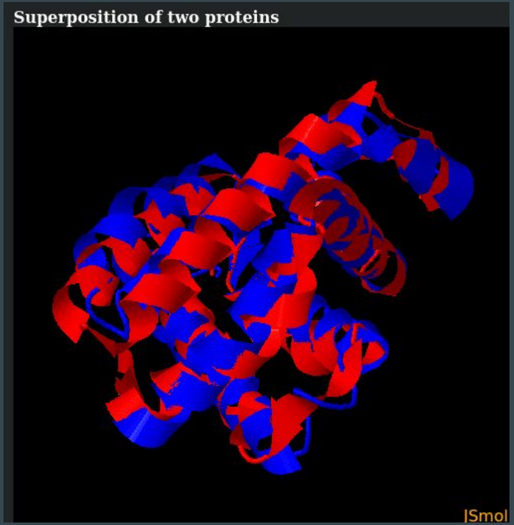
HHblits

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!