

Hello again!

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



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"?



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

Pairwise (2 sequences)

Score 435 bits(235)		Expect 5e-117	Identities 360/410(88%)	Gaps 50/410(12%)	Strand Plus/Minu	s
Query	302	GTAGGACAGGTGC	CGGCAGCGCTCTGGGTCA	TTTTCGGCGAGGACCGCTT	TCGCTGGAG-	360
Sbjct	3589	GTAGGACAGGTGC	CGGCAGCGCTCTGGGTCA	TTTTCGGCGAGGACCGCTT	TCGCTGGAGC	353
Query	361	ATCG	GCCTGTCGCTTGCGGTAT	TCGGAATCTTGCACGCCCT	CGCTCAAGCC	411
Sbjct	3529	GCGACGATGATCG	GCCTGTCGCTTGCGGTAT	tcggAATcTTgcAcgccct	CGCTCAAGCC	347

Multiple sequences (>2)

Q5E940 BOVIN	MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENNPALE	76
RLAO HUMAN	MPREDRAT#KSNYFLKIIQLLDDYPKCFIVGADNVGSKOMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENNPALE	76
RLAO MOUSE	MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMOQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENNPALE	76
RLAO RAT	MPREDRATWKSNYFLKIIQLLDDYPKCFIYGADNYGSKOMOQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENNPALE	76
RLA0 CHICK	MPREDRATWKSNYFMKIIGLLDDYPKCFVVGADNVGSKOMOQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENNPALE	76
RLAO RANSY	MPREDRATWKSNYFLKIIQLLDDYPKCFIYGADNYGSKOMOQIRMSLRGK-AYYLMGKNTMMRKAIRGHLENNSALE	76
Q7ZUG3 BRARE	MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKOMOTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENNPALE	76
RLA0 ICTPU	MPREDRATWKSNYFLKIQLLNDYPKCFIYGADNYGSKOMOTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENNPALE	76
RLA0 DROME	MURENKAAWKAQYFIKUUELFDEFPKCFIUGADNUGSKOMONIRTSLRGL-AUVLMGKNTMMRKAIRGHLENNPOLE	76
RLA0 DICDI	MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSOLOKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSKPELD	75
Q54LP0 DICDI	MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSOLOKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSKPELD	75
RLAO PLAF8	MAKLSKQQK <mark>K</mark> QMYIEKLSSLIQQYSKILIVHVDNVG <mark>S</mark> NQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAVPQIE	76
RLA0 SULAC	MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKHNLFNIALKNAGYDTK	79
RLA0 SULTO	MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAGLDVS	80
RLA0 SULSO	MKRLALALKORKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAGIDIE	80
RLA0 AERPE	MSVVSLVGQMYKREK <mark>PIPEWK</mark> TLMLRELE <mark>ELF</mark> SKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLELDDN	86
RLAO PYRAE	-MMLAIGKRRYVRTRQYPARKYKIVSEATELLQKYPYVFLFDLHGLSBRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKYYGGIPAE	85
RLAO METAC	MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMOKIRRDLKDV-AVLKVSRNTLTERALNQLGETIP	78
RLAO METMA	MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDIKDV-AVLKVSRNTLTERALNQLGESIP	78
RLA0 ARCFU	MAAVRGSPPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGOMOKIRREFRGK-AEIKVVKNTLLERALDALGGDYL	75
RLAO METKA	MAVKAK <mark>GOPP</mark> SGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPOLOEIRAKLRERDTIIRMSRNTLMRIALEEKLDERPELE	88
RLA0 METTH	MAHVAEWKKEVQELHDLIKGYEVVGIANLADIPAROLOKMROTLRDS-ALIRMSKKTLISLALEKAGRELENVD	74
RLA0 METTL	MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPAROLOEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA	82
RLAO METVA	MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVOLOEIRDKIR-DOMTLKMSRNTLIKRAVEEVAEETGNPEFA	82
RLAO METJA	METKYKAHYAPWKIEEVKTLKGLIKSKPYYAIYDMMDYPAPOLOEIRDKIR-DKYKLRMSRNTLIIRALKEAAEELNNPKLA	81
RLAO PYRAB	MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAQELGKPELE	77
RLA0 PYRHO	MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAKELGKPELE	77
RLA0 PYRFU	MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSOMRRLIRENNGLLRVSRNTLIELAIKKVAGELGKPELE	77
RLA0 PYRKO	MAHVAEWKKKEVEELANIIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKRAAQELGOPELE	76
RLAO HALMA	MSAESERKTETIPEWKQEEVDAIVEMIESYESYGVVNIAGIPERQLQDMRRDLHGT-AELRVSRNTLLERALDDVDDGLE	79
RLAO HALVO	MSESEVRQTEVIPQWKREEVDELVDFIESVESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRALDEVNDGFE	79
RLA0 HALSA	MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEEAGDGLD	79
RLAO THE AC	MKEVSQQKKELVNEITORIKASRSVAIVOTAGIRTROIODIRGKNRGK-INLKVIKKTLLFKALENLGDEKLS	72
RLA0 THE VO	MRKINPKKKEIVSELAODITKSKAVAIVDIKGVRTROMODIRAKNRDK-VKIKVVKKTLLFKALDSINDEKLT	72
RLAO PICTO	MTEPAQWKIDFVKNLENE INSRKVAAIVSIKGLRNNEFOKIRNSIRDK-ARIKVSRARLLRLAIENIGKNNIV	72
ruler	110	

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

1 sequence versus a database

💥 BLAST Search Results - Netscape					
File Edit Yiew Go Communicator Help					
👔 🏒 Bookmarks 🦼 Location: http:///	www.ncbi.nlm.nih.gov/blast/blast.cgi	•	What's Related	N	
5. O P P P00		Score	e E		
Sequences producing signifi	cant alignments:	(bits	s) Value		
		- 172-51 Z	n 298		
sp P14120 RL30 YEAST 60S P	(IBOSOMAL PROTEIN L30 (YL32) (RP73)	201 3	3e-52		
sp P38664 RL30 KLULA 605 F	(IBOSOMAL PROTEIN L30 (L32)	190 7	/e-49		
SPIPS2808 RL30 SCHPO 605 F	(IBOSOMAL PROTEIN L30 (L32)	120 1	be-33		
api04098410130 10010 605 F	TBOSOMAL PROTEIN 130	129 1	LE-30		
an P478331 PL30 CHICK 605 P	TBOSOMAL PROTEIN LSO	129 1	LE-30 Le-30		
sn104855818130 MAIZE 60S F	TBOSOMAL PROTEIN L30	127 4	le-30		
sp P49153 RL30 TRYBB 60S F	BOSOMAL PROTEIN L30	113 9	9e-26		
sp P39095 RL30 LEIMA 60S F	BOSOMAL PROTEIN L30	108 2	2e-24		
sp 027127 RL3E METTH 50S F	BOSOMAL PROTEIN L30E	79 2	2e-15		
sp 074018 RL3E PYRHO 50S F	BOSOMAL PROTEIN L30E	71 5	5e-13		
sp Q9YAU3 RL3E AERPE 50S F	BOSOMAL PROTEIN L30E	71 5	5e-13		
sp P14025 RL3E METVA 50S F	BOSOMAL PROTEIN L30E	69 2	2e-12		
sp P54061 RL3E METJA 50S F	BOSOMAL PROTEIN L30E	66 1	le-11		
sp P11522 RL3E SULAC 50S F	BOSOMAL PROTEIN L3DE (ORF 104)	64 9	9e-11		
sp P29160 RL3E THECE 50S F	BOSOMAL PROTEIN L30E	64 9	9e-11		
sp 028389 RL3E ARCFU 50S F	IBOSOMAL PROTEIN L30E	60 1	Le-09		
sp 059165 RS6X PYRHO 30S F	IBOSOMAL PROTEIN HS6-LIKE	37 0	0.011		
sp P54066 RS6X METJA 30S F	IBOSOMAL PROTEIN HS6-LIKE	<u>35</u> C	0.043		
sp 026355 RS6X METTH 30S F	RIBOSOMAL PROTEIN HS6-LIKE	<u>_34</u> C	0.056		
sp P55858 RS6X_SULSO_30S_F	IBOSOMAL PROTEIN HS6-LIKE	<u>34</u> C	0.056		
sp Q9YAX7 RS6X AERPE 30S F	IBOSOMAL PROTEIN HS6-LIKE	33 0	0.17		
sp P34667 Y011 CAEEL HYPO1	HETICAL 20.8 KD PROTEIN 2K686.1 I	33 0	J.17		
sploz9494 RS6X ARCFU 30S F	TEDSOMAL PROTEIN HS6-LIKE	<u></u>	J.28		
SPIQSS601 RECG SINIS AIP-D	VEPENDENI DNA HELICADE RECO	<u>34</u> U	1.37		
SPIP46035 ERFI HUMAN LUKAR	VOTIC PEPTIDE CHAIN RELEASE FACTO	20 2			
SPIPSSOIS ERFI AENLA LORAF	DUOGDUNTIGE DUO1 DEFCUESOD	20 2			
an P557681 VI YO FNTEC PROBA	BIE DIBOSOMAL DEOTETN IN INFE 5'D	28 4	1 2		
SNIO01056 TEGIL HSVS1 PROBA	BLE LARGE TEGIMENT PROTEIN	28 4	1.2		
sn P55124 LKTC PASSP LEUKC	TOXIN-ACTIVATING LYSINE-ACYLTRANS	27 7	7.3		
Sp 025074 Y303 HELPY PROBA	BLE GTP-BINDING PROTEIN HP0303	27 9	9.5	_ 1	
1			1	•	
http://www.ncl	bi.nlm.nih.gov/blast/blast.cgi#730555				
j	bttps://www.	wheeh outple and			

9

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
- \wedge
- 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, 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



15

Global vs local



Global: must align **all** nucleotides, using insertions/deletions if necessary **Local**: you're allowed to skip beginning and/or end of either sequence

Alignment is based on scoring

What is a *good* alignment? One that **minimizes** a penalty (or **maximizes** a score).

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

r: T A C	r:	G A T
-----------------	----	--------------

q: T T C	q: G-T
-----------------	--------

penalty=1 penalty=2

Example: (global alignment)

	Is it the best we can de	o?	better!
	total penalty: 5		total penalty: 4
	MMXDXXM		MMXMDXM
d:	CAT-GGA	way:	CATG-GA
r:	CAAGTTA	can also be	CAAGTTA

(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.

I = insertion (gap in the target sequence) M = matchD = deletion (gap in the query sequence) X = mismatch

* some programs use M for both matches and mismatches $1_(\mathcal{V})_f$, 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"

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, 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



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



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

(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 <u>same length</u>:

Minimum number of substitutions to turn sequence A into sequence B

e.g.





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**? (*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



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

26



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



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 to align. CAGAGGGAATGCT 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: 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? *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 Last CIGAR letter until the last CIGAR Optimal alignment +letter CATATGATGACA С (X) align(+CAGAGGGAATGC Т or CATATGATGACAC CATATGATGACAC — (I) align(align() = +CAGAGGGAATGC CAGAGGGAATGCT Т or CATATGATGACA С (D) align(CAGAGGGAATGCT



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

 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



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


- 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



- 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



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

- 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



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

- 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



- 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	Т	С	A
A	1	-1	-3	-5	-7
Т	-1	0 -	* ?		
С	-3				
С	-5				

- 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

	А	G	Т	С	А
A	1	-1	-3	-5	-7
Т	-1	0	0	-4	-6
С	-3	-2	-1	1	-1
С	-5	-4	-3	0	0

- 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



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	Т	С	A
A	-1	-3	-5	-7	-7
т	-3				
С	-5				
С	-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

Red	call:					
	A	G	Т	С	A	Three possibilities:
A	1	-1				$MX \rightarrow score 0$
т		? -				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	Т	С	A	ſ		1
A	-1	-3	-5	-7	-7		XDMMX	
Т	-3	-2	-2	-4	-6		GGTCA	or
С	-5	-4	-3	-1	-3		A-TCC	
С	-7	-6	-5	-2	-2 _		score: -2	

Coffee break?



"Dynamic programming"?



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	т	С	А
A	0	0	0	0	1
т	0	0	1	-1	-1
С	0	-1	-1	2	0
С	0	-1	-2	0	1

Allow gaps at beginning
Find the highest scoring cell
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: ...TCATAAACAGATATGA...CCAAAGAGATTTGATA...

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*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

Be BLAST!

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

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
СТТАА	none
CCCCC	49, 101

Some DNA scoring schemes

• Edit Distance:

- Match = +1
- Mismatch = -1
- \circ Indel = -1

• **BLAST** (megablast):

- \circ Match = +1
- Mismatch = -2
- \circ Indel = -2.5

• Minimap2:

- $\circ \quad \text{Match} = +2$
- 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

$\bullet \bullet \bullet$

Long sequences versus long sequences

Tools

- BLAT
- Exonerate
- LASTZ
- MUMmer
- minimap2



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

Genome: 🔲 Search all genomes		Assembly: Query type:			Sort output:	Output type:	
Chicken	~	Mar. 2018 (GRCg6a/galGal6) 🗸 🗸		BLAT's guess 🗸 🗸	query,score 🗸	hyperlink	~
							11.
All Results (no r	minimum matches)				Submit I'm	feeling lucky	Clear



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

Dotplots





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

$\bullet \bullet \bullet$

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 <u>https://github.com/iskana/SlideSort</u>

PAF file format

Pairwise DNA

$\bullet \bullet \bullet$

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

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



Neither. It's **glocal**.

<u>69</u>

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

Problem: Half of the human genome is comprised of repeats

taaccctaaccctaaccctaaccctaaccctaaccctaacccta accctaaccctaaccctaaccctaaccctaaccctaaccctaac cctaacccaaccctaaccctaaccctaaccctaaccctaacccc taaccctaaccctaaccctaaccctaaccctaaccctaaccctaa ccccctaaccctaaccctaaccctaaccctaaccctaaccctaaccc ccctaaaccctaaccctaaccctaaccctaaccccaaccccaac cccaaccccaaccccaacccctaacccctaaccctaaccctaacc ctaccctaaccctaaccctaaccctaaccctaacccctaacccc taaccctaaccctaaccctaaccctaaccctaaccctaacccctaaccct aaccctaaccctcgcggtaccctcagccggcccgcccggc tctgacctgaggagaactgtgctccgccttcagagtaccaccgaaatctg tgcagaggacaacgcagctccgccctcgcggtgctctccgggtctgtgct gaggagaacgcaactccgccggcgcaggcgcagagaggcgcgccgccgcg gcgcaggcgcagacacatgctagcgcgtcggggtggaggcgtggcgcagg cgcagagaggcgcgccgcgccggcgcagggcgcagagacacatgctaccgc gtccaggggtggaggcgtggcgcaggcgcagaggggcgcaccgcggc gcaggcgcagagacacatgctagcgcgtccaggggtggaggcgtggcgca gcacgcgcagaaactcacgtcacggtggcgcggcgcagagacgggtagaa

(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

<u>minimap2</u>: 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		
e vomics		
ics		
mics	<- can be found in 1 step by binary search	
o mics		
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**:

csevomi

evomics

icsevom

micsevo

omicsev

vomicse

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.



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



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: 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
ΤΑΑΑΑ	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



Reference and BAM need to be indexed, use samtools



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)



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.



MMseqs2

DIAMOND2

BLASTp

How mmseqs2 work: mmseqs search



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..

Q5E940 BOVIN	MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNYG <mark>8</mark> KQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENNPALE	76
RLA0 HUMAN	MPREDRATWKSNYFLKIIQLLDDYPKCFIYGADNYG <mark>8</mark> KQMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENNPALE	76
RLA0 MOUSE	MPREDRATWKSNYFLKIIQLLDDYPKCFIYGADNYG <mark>SKQMQQIR</mark> MSL <mark>RGK-AVVLMGKNTMMRKAIRGHLE</mark> NNPALE	76
RLAO RAT	MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNYG <mark>S</mark> KQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENNPALE	76
RLA0 CHICK	MPREDRATWKSNYFMKIIQLLDDYPKCFYYGADNYG <mark>3</mark> KQMQQIRMSLRGK-AVYLMGKNTMMRKAIRGHLENNPALE	76
RLA0 RANSY	MPREDRATWKSNYFLKIIOLLDDYPKCFIVGADNYGSKOMOQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENNSALE	76
Q7ZUG3 BRARE	MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNYG <mark>SKQMQ</mark> TIRLSLRGK-AVVLMGKNTMMRKAIRGHLENNPALE	76
RLA0 ICTPU	MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNYG <mark>SKQMQTIR</mark> LSLRGK-AIVLMGKNTMMRKAIRGHLENNPALE	76
RLA0 DROME	MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVG <mark>S</mark> KOMONIRTSLRGL-AVVLMGKNTMMRKAIRGHLENNPOLE	76
RLA0 DICDI	MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSKPELD	75
Q54LP0 DICDI	MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSDLOKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSKPELD	75
RLA0 PLAF8	MAK LSKQQKKQMY IEKLSSL IQQYSK ILI VHV DNYGSNQMASVRKSLRGK-AT ILMGKNTR IRTALKKNLQAVPQ IE	76
RLA0 SULAC	MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAGYDTK	79
RLA0 SULTO	MRIMAVITQERKIAKWKIEEVKELEOKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAGLDVS	80
RLA0 SULSO	MKRLALALKORKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAGIDIE	80
RLA0 AERPE	MSVVSLVGQMYKREK <mark>PIPEWK</mark> TLMLRELE <mark>ELFSKHRVVLFADLTGTPI</mark> FVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLELDDN	86
RLA0 PYRAE	-MMLAIGKRRYVRTRQ <mark>YP</mark> ARKVKIVSEATELLQKYPYVFLFDLHGLS <mark>BRILHEYR</mark> YRLRRY-GVIKIIKPTLFKIAFTKVYGGIPAE	85
RLA0 METAC	MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMOKIRRDLKDV-AVLKVSRNTLTERALNQLGETIP	78
RLAO METMA	MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLGESIP	78
RLA0 ARCFU	MAAYRGSPPEYKYRAVEEIKRMISSKPYVAIYSFRNYPAGOMOKIRREFRGK-AEIKYYKNTLLERALDALGGDYL	75
RLA0 METKA	MAVKAKGOPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPOLOEIRAKLRERDTIIRMSRNTLMRIALEEKLDERPELE	88
RLA0 METTH	MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGRELENVD	74
RLA0 METTL	MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA	82
RLA0 METVA	MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVOLOEIRDKIR-DOMTLKMSRNTLIKRAVEEVAEETGNPEFA	82
RLA0 METJA	METKYKAHYA <mark>PWK</mark> IE EVKTLK <mark>GLIKSKPYYAIYDMMDYPAPQLQ</mark> EI <mark>R</mark> DKI <mark>R</mark> -DKYKL <mark>RMSRNTLIIRALKEAAE</mark> ELNN <mark>P</mark> KLA	81
RLA0 PYRAB	MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAQELGKPELE	77
RLA0 PYRHO	MAHVAEWKKKEVEELAKLIKS YPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAKELGKPELE	77
RLA0 PYRFU	MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRLIRENNGLLRVSRNTLIELAIKKVAQELGKPELE	77
RLA0 PYRKO	MAHVAEWKKKEVEELANIIKS <mark>YP</mark> VIALVDVAGVPAYPLSKM <mark>R</mark> DKL <mark>R-GKALLRVSRNT</mark> LIELAIKRAAQELGQPELE	76
RLAO HALMA	MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSROLODMRRDLHGT-AELRVSRNTLLERALDDVDDGLE	79
RLA0 HALVO	MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRALDEVNDGFE	79
RLA0 HALSA	MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEEAGDGLD	79
RLA0 THEAC	MKEVSQQ <mark>K</mark> KELVNEITORIKASRSVAIVOTAGIRIROIODIRGKN <mark>RG</mark> K-INLKVIKKILLF <mark>K</mark> ALENLGDEKLS	72
RLA0 THE VO	MRKINPKKKEIVSELAODITKSKAVAIVDIKGVRIROMODIRAKNKKKILLFKALDSINDEKLT	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: 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.

96



Remember Needleman-Wunsch?

Same, but with more possibilities.

So, best avoided.

Progressive MSA



MSA is on another level of difficulty

Challenging alignment



FLVRESQRNPQG-FVLSLC	HLQKVKHY
FIIRFSERNPGQ-FGIAYI	GVEMPARIKHY
FLLRFSESSREGAITFTWV	ERSQNG
FLVRDASTKMHGDYTLTLR	KGGNNK

Alternative MSAs of same sequences

Which one is correct / better?

FLVRESQRNPQG-FVLSLC	HLQKVKHY
FIIRFSERNPG-QFGIAYI	GVEMP-ARIKHY
FLLRFSESSREGAITFTWV	ERSQNGGEPD-F
FLVRDASTKMHGDYTLTLR	KGGNN-K

Hard / impossible to decide, even with structures

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

Tools

- MUSCLE
- ClustalW
- T-Coffee
- MAFFT

•••

igodol



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

 $\bullet \bullet \bullet$



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

 $\bullet \bullet \bullet$

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?"



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

- 1) Construct MSA
- 2) Determine frequency per column

a Catalytic G		1				
Motif B, PSSM 1						
AHA91815.1	PGVQKS	SYN	TSSSN	I <mark>S</mark> RI	RVMCALHA	
YP_006666506.1	PGIQKS	<mark>s</mark> syn	TSSTN	ISRV	RVMLSIYA	
CAJ29959.1	TVGMFS	T RF	TMLYN	ITVL	NRAYYKVA	
CAJ29958.1	TVGMFS	T RF	TMLYN	ITIL	NRAYYKVA	
BAM93353.1	LGGLFS	<mark>9</mark> HRL	TMFIN	ITVL	NRVYYRVA	
YP_005097975.1	LGGLFS	<mark>7</mark> HRL	TMFIN	ITVL	NRVYYRVA	
YP_001976142.1	YAGQQS	RRS	TLESN	ITFY	SRARLLVR	

HMM

Hidden Markov Models generalize PSSMs with gaps.

Motivation: when pairwise fails

HBA_HUMAN	VGAHAGEY
HBB_HUMAN	VNVDEV
MYG_PHYCA	VEADVAGH
GLB3_CHITP	VKGD
GLB5_PETMA	VYSTYETS
LGB2_LUPLU	FNANIPKH
GLB1_GLYDI	IAGADNGAGV
	*** *****

HMM of a PSSM:



Profile HMM:







HMMer

MMseqs profile

HHblits

Bonus: structural alignment (TM-align)

(":" denotes aligned residue pairs of d < 5.0 A, "." denotes other aligned residues)

MVLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRVKHLKTEAEMKASEDLKKHGVTVLTALGAILKKK--G-HHEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGNFGADAQGAMNKALELFRKDIAAKYKELGYQG

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!

