Scalability in bioinformatics

Antoine Limasset

Bonsai Team,CRIStAL, Lille University, CNRS, France

antoine.limasset@gmail.com
@BQPMalfoy

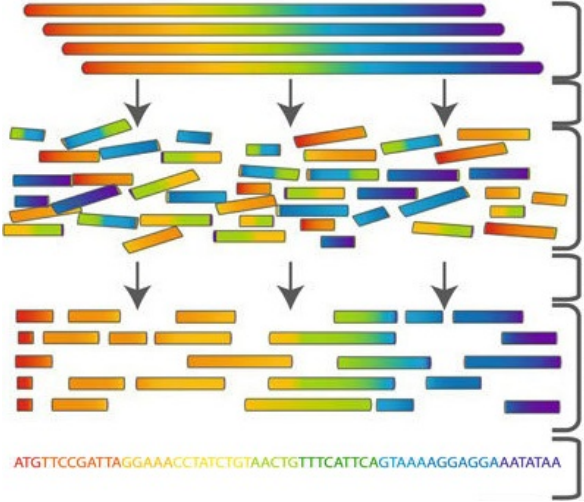
Take home messages

Bioinformatics is a exiting research field for computer scientists

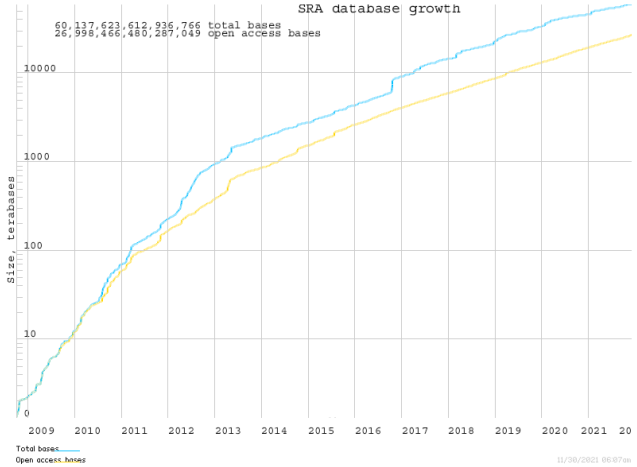
How?

- **Drowned** by data deluge
- Algorithms and data structures **matter**
- Very rewarding to help biologists

Sequencing


ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA

Big Data


SRA database growth
60,137,623,612,936,766 total bases
26,998,466,480,287,049 open access bases
Size, terabases
10000
1000
100
10
0
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
20
Total bases
Open access bases
11/30/2021 06:07am

Presentation Leitmotiv

We sequenced a weird bacteria!
Let's see if it looks like something known!

Bacterial database

One million available genomes (Genbank)
$\approx$ 10 megabases each
Database size estimation: 10 Terabases (10^{13})
Using 2bit per bases $\approx$ 3 TeraBytes

Idea 1: Use alignment

Use Smith–Waterman algorithm to compare our query to each genome

		A	C	A	C	A	C	T	A
	0	0	0	0	0	0	0	0	0
A	0	2	1	2	1	2	1	0	2
G	0	1	1	1	1	1	1	0	1
C	0	0	3	2	3	2	3	2	1
A	0	2	2	5	4	5	4	3	4
C	0	1	4	4	7	6	7	6	5
A	0	2	3	6	6	9	8	7	8
C	0	1	4	5	8	8	11	10	9
A	0	2	3	6	7	10	10	10	12

A - C A C A C T A

A G C A C A C - A

Idea 1: Use alignment

Use Smith-Waterman algorithm to compare our query to each genome

Complexity

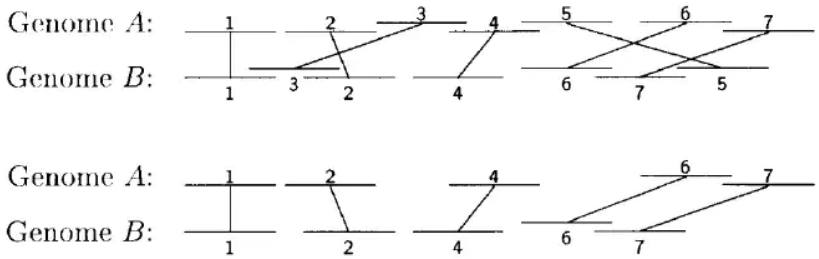
$\mathcal{O}(G^2.N)$
(G the genomes size and N the amount of genomes)

Cost of our search

$\approx 10^{20}$ operations

Idea 2: Use Longest increasing subsequence

Use LIS algorithm to compare our query to each genome



Idea 2: Use Longest increasing subsequence

Use LIS algorithm to compare our query to each genome

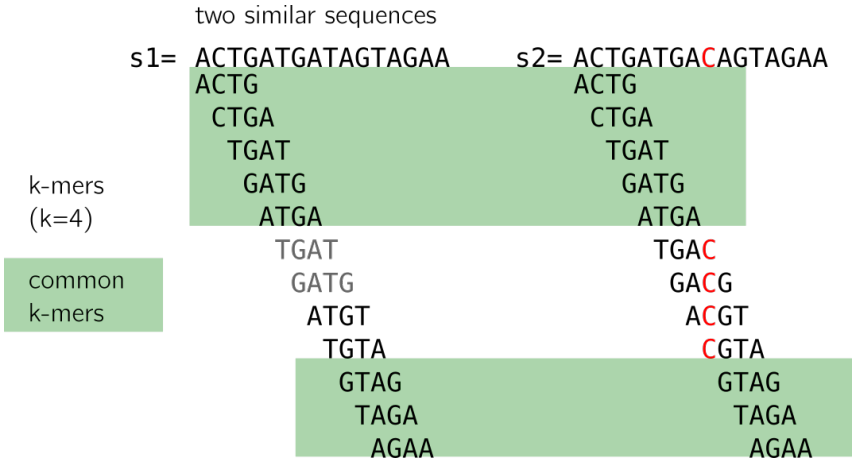
Complexity

$\mathcal{O}(G.ln(G).N)$
(G the genomes size and N the amount of genomes)

Cost of our search

$\approx 10^{14}$ operations

Idea 3: Count shared words



Idea 3: Count shared words

Index fixed size word (k-mer) from each genome with a hash table and count

Complexity

$\mathcal{O}(G.N)$
(G the genomes size and N the amount of genomes)

Cost of our search

$\approx 10^{13}$ operations

Idea 4: Index k-mers

We can build an index associating to each k-mer its originating datasets

Color matrix

k-mer	Color set
ACTG	0110010101
ACTT	1000011111
CTTG	0011110000
TTTC	0110010101
GCGT	0111110101
AGCC	0110010101

Idea 4: Index k-mers

We consider only the query time because we got a pre-built index

Complexity

$\mathcal{O}(G)$
(G the genomes size)

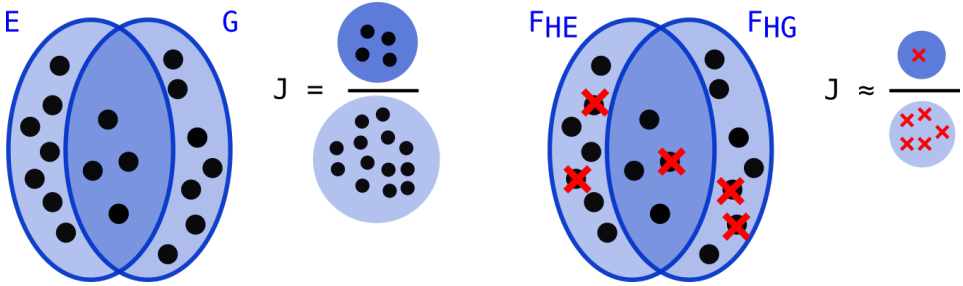
Cost of our search

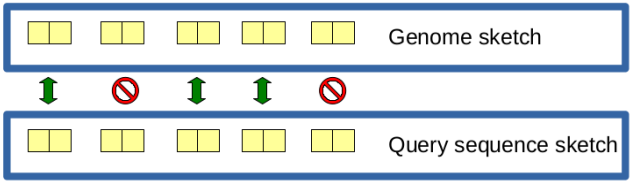
$\approx 10^7$ operations

Memory cost

Up to$\mathcal{O}(G.N.K)$ nucleotide in theory
In practice, counting 100 billions distinct k-mers, using 64bits per kmer $>$ 1 TeraByte

Idea 5: Use minhash sketches





Idea 5: Use minhash sketches

Represent each genome with a Minhash sketch

Complexity

$\mathcal{O}(S.N)$
(S the sketch size, N the number of genomes)

Cost of our search

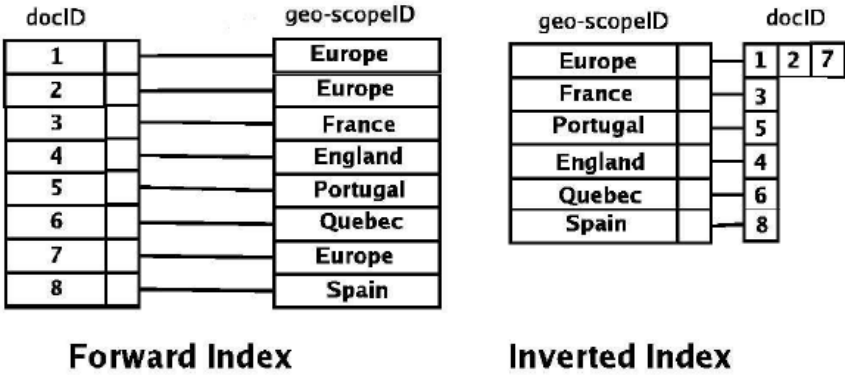
Using 1,000 fingerprints
$\approx 10^9$ operations
$\approx 10^6$ Random access

Memory cost

$\mathcal{O}(H.G)$ integers
In practice,using sketches of 1,000 32 bits integers, $\approx$ 4 GigaBytes

Idea 6: Index fingerprints

Index fingerprint using inversed index



Idea 6: Index fingerprints

Index integer using inversed index

Complexity

$\mathcal{O}(H)$
(H the sketch size)

Cost of our search

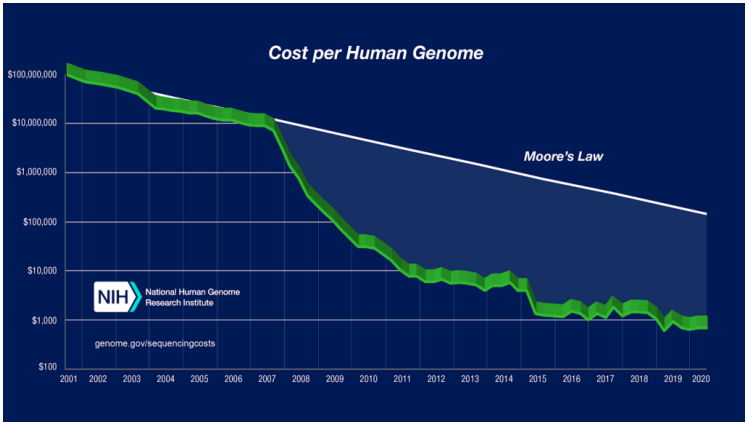
$\approx 10^3$ operations

Memory cost

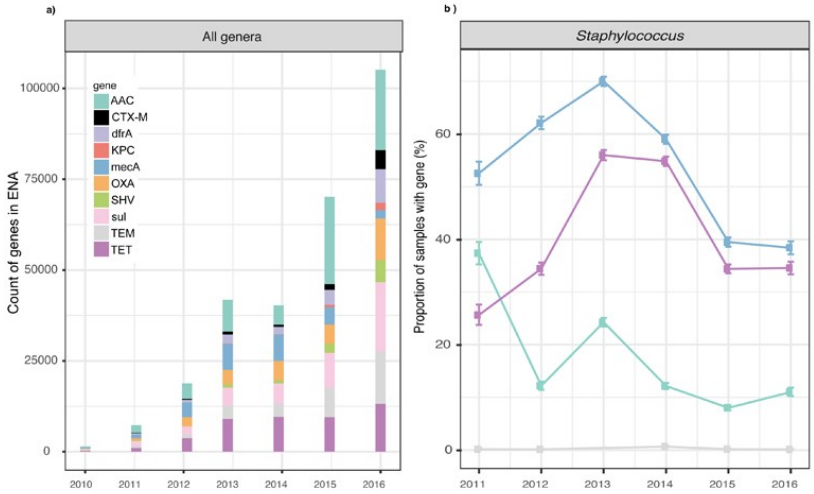
$\mathcal{O}(H.G)$ integers
In practice,using sketches of 1,000 32 bits integers, $\approx$ 4 GigaBytes

Take home message

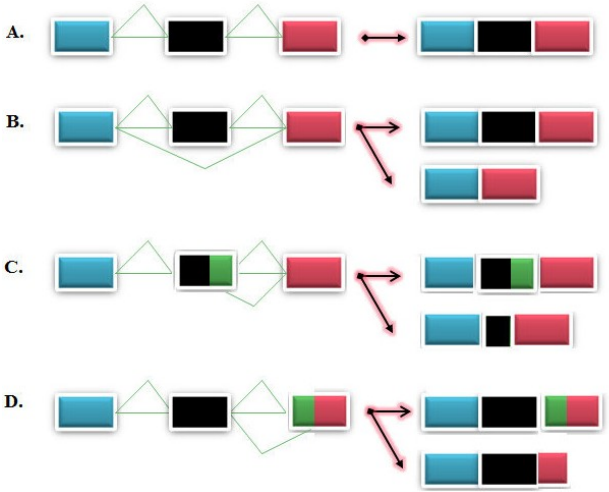
Algorithmic and data structure improvement are the cornerstones of being able to scale with the current databases sizes



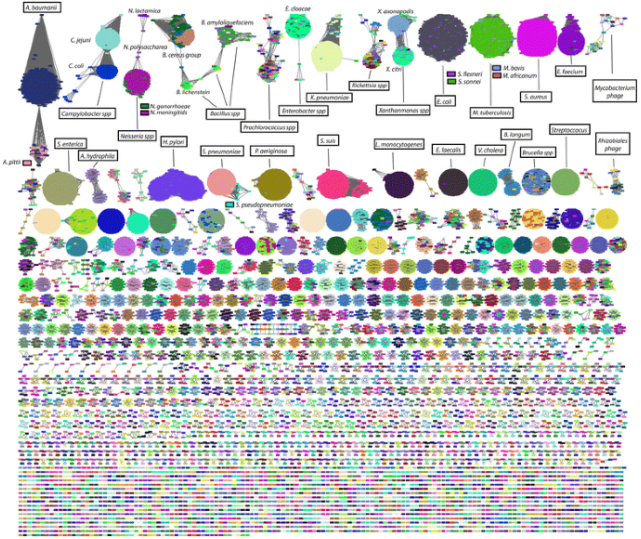
Application to antibiotic resistance gene surveillance



Application to biomarker detection



Comparison and clustering of all know genomes



Take home messages

Bioinformatics is a exiting research field for computer scientists

How?

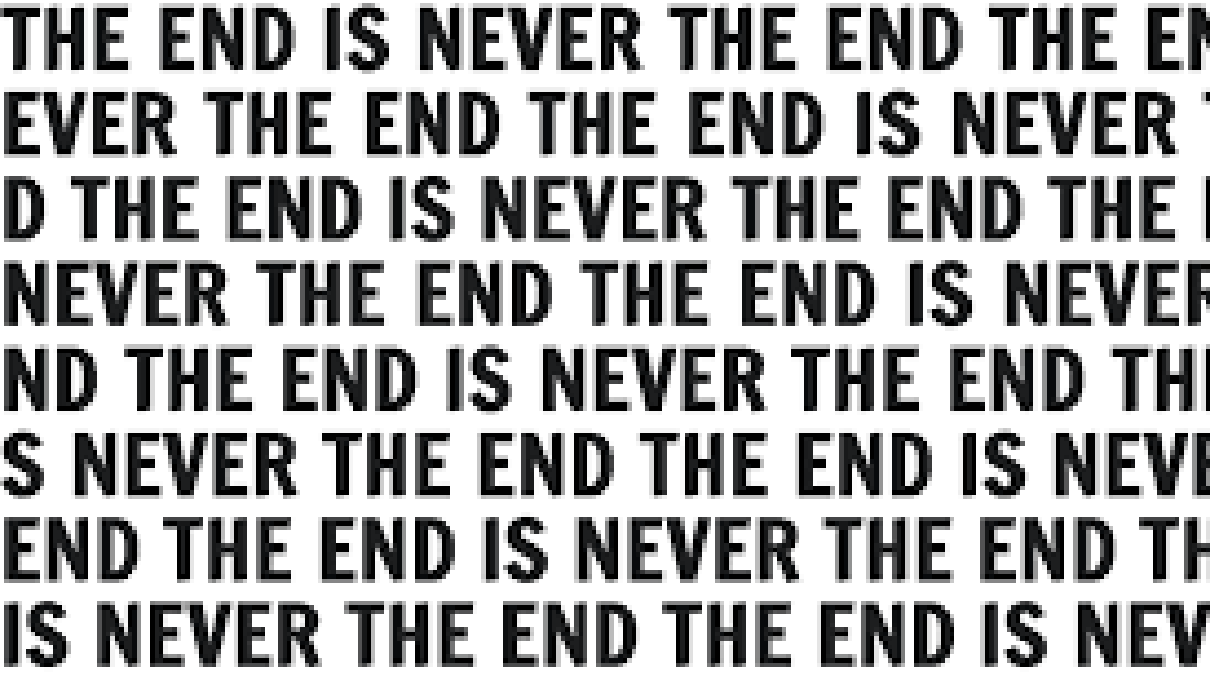
- **Drowned** by data deluge
- Algorithms and data structures **matter**
- Very rewarding to help biologists

Join Us!

Open subjects:
(Internship/PhD Thesis)

- Index data structure
- Specialized architecture
- Sequence analysis

Contact:
Antoine.Limasset@univ-lille.fr
@BQPmalfoy on Twitter


THE END IS NEVER THE END THE EN
EVER THE END THE END IS NEVER
D THE END IS NEVER THE END THE
NEVER THE END THE END IS NEVER
ND THE END IS NEVER THE END TH
S NEVER THE END THE END IS NEVE
END THE END IS NEVER THE END TH
IS NEVER THE END THE END IS NEV