BONSAI

Bioinformatics and Sequence Analysis

M. Salson

L'équipe BONSAI

Au sein de CRIStAL (Univ Lille, CNRS, Centrale)

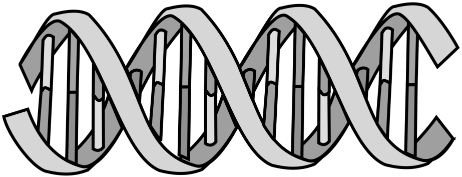
9 permanent·e·s

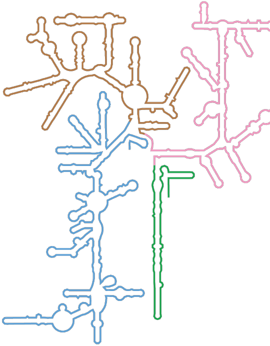
2 chercheuses, 1 chercheur, 5 enseignants-chercheurs, 1 ingénieur de recherche

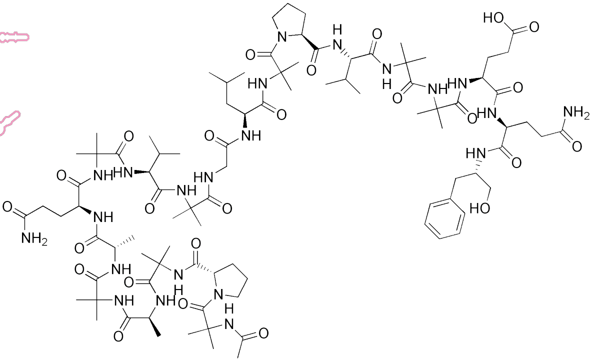
9 non permanent·e·s

7 doctorant·e·s, 1 ingénieur, 1 post-doctorant

Analyzing biological sequence data

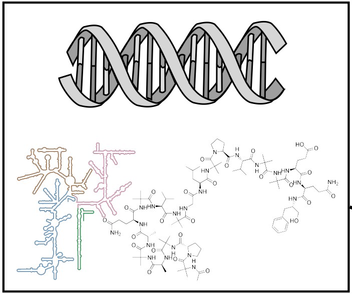






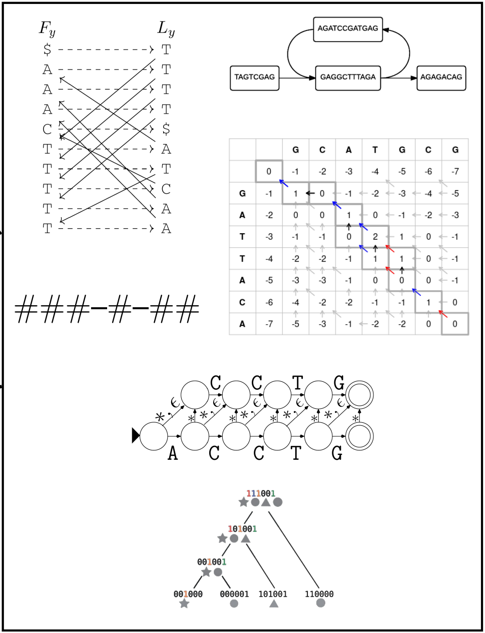
Images: Forluvoft et Leyo, CC BY SA Petrov et al, Edgar 181

BiOiNformatics Sequence AnalysIs

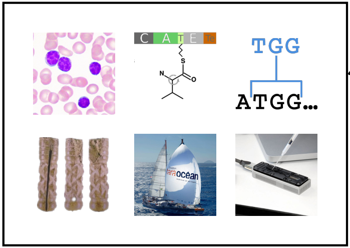


In the cell

In our brains



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



Among our applications

Efficient algorithms matter

Let's compare two sequences

A T A C T G A

T A C G A C

Efficient algorithms matter

Let's compare two sequences

A T A C T G A

T A C G A C

The optimal[*] solution is

A	T	A	C	T	G	A	
	T	A	C	–	G	A	C

[*] Optimal in terms of minimising the number of differences

Efficient algorithms matter

Let's compare two sequences

A T A C T G A

T A C G A C

The optimal* solution is

```
A T A C T G A
  T A C – G A C
```

To find it we need to compute all possibilities

* Optimal in terms of minimising the number of differences

Efficient algorithms matter

Let's compare two sequences

A T A C T G A

T A C G A C

The optimal* solution is

```
A T A C T G A
  T A C – G A C
```

To find it we need to compute all possibilities

This takes $\geq$**30 microseconds** for two 300nt sequences

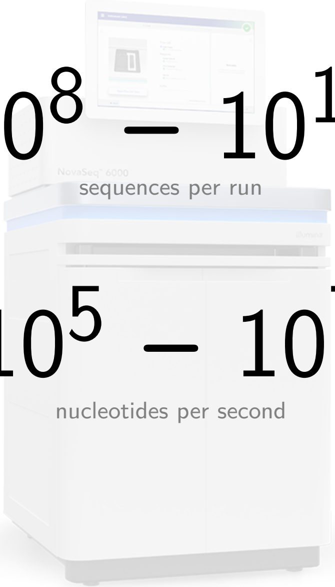
* Optimal in terms of minimising the number of differences

Sequencing data is produced at a (very) high throughput

$$10^{8} - 10^{10}$$

sequences per run

Sequencing data is produced at a (very) high throughput



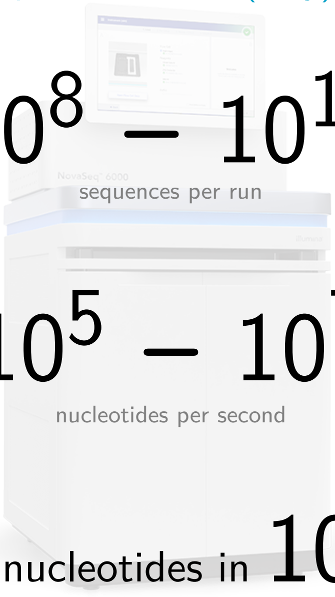
$$10^8 - 10^{10}$$

sequences per run

$$10^5 - 10^7$$

nucleotides per second

Sequencing data is produced at a (very) high throughput



$$10^8 - 10^{10}$$

sequences per run

$$10^5 - 10^7$$

nucleotides per second

up to 300 nucleotides in 10 microseconds

Efficient algorithms cope with the rising throughput

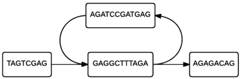
Processing the data could be **10 times longer** than sequencing

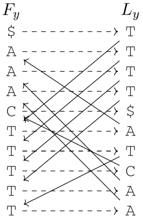
Efficient algorithms cope with the rising throughput

Processing the data could be **10 times longer** than sequencing

It's not

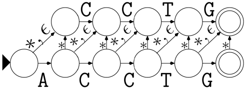
Efficient algorithms cope with the rising throughput

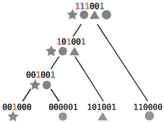




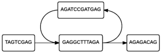
Processing the data could be **10 times longer** than sequencing

It's not





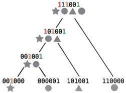
A wide variety of applications



Graphs

###-#-##

Seeds



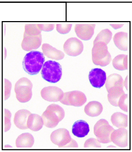
Indexes

High-throughput sequencing

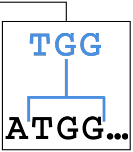
Peptides

3rd gen sequencing

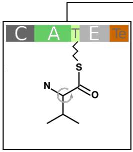
Metagenomics



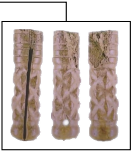
Oncohematology



Large-scale search



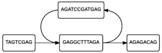
Non-ribosomal peptides



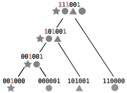
Paleoproteomics

Images: © Univ of Oxford, © Sacha Bollet, CC BY SA Ytngargar, © Maude Pupin, © Olivier Sarrazin

A wide variety of applications



###-#-##



Graphs

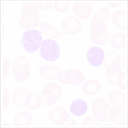
Seeds

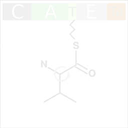
Indexes

High-throughput sequencing

Peptides

3rd gen sequencing

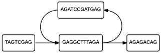




Splicing variants
Multiple alignment assessment
Virus mapping

Images: © Univ of Oxford, © Sacha Bollet, CC BY SA Ytngargar, © Maude Pupin, © Olivier Sarrazin

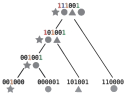
A wide variety of applications



Graphs

###-#-##

Seeds

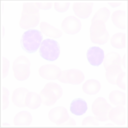


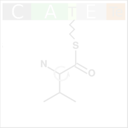
Indexes

High-throughput sequencing

Peptides

Metagenomics



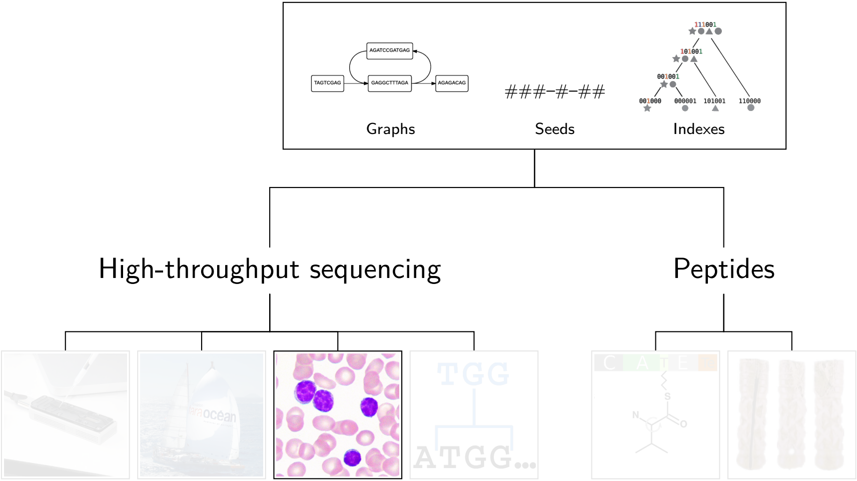


rRNA filtering

Marker gene reconstruction

Images: © Univ of Oxford, © Sacha Bollet, CC BY SA Ytngargar, © Maude Pupin, © Olivier Sarrazin

A wide variety of applications



Graphs

Seeds

Indexes

High-throughput sequencing

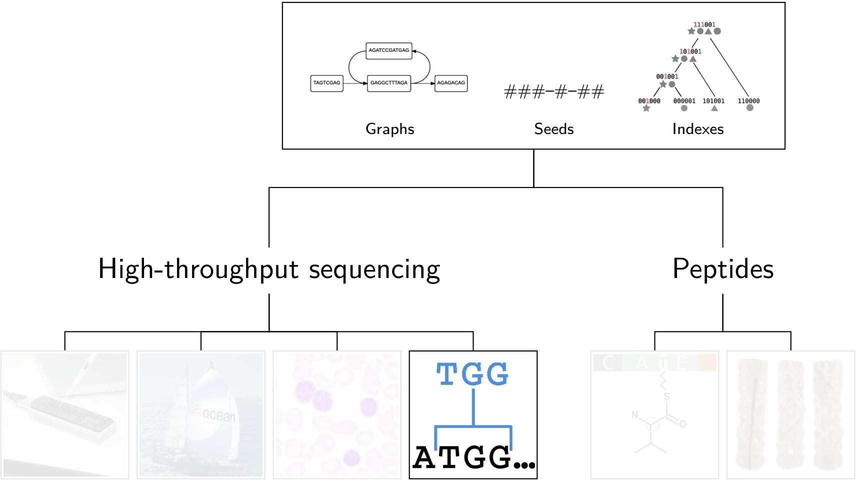
Peptides

Oncohematology

Marker identification

Images: © Univ of Oxford, © Sacha Bollet, CC BY SA Ytngargar, © Maude Pupin, © Olivier Sarrazin

A wide variety of applications



Graphs

Seeds

Indexes

High-throughput sequencing

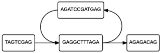
Peptides

Large-scale search

Searching into collections of runs/genomes

Images: © Univ of Oxford, © Sacha Bollet, CC BY SA Ytngargar, © Maude Pupin, © Olivier Sarrazin

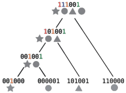
A wide variety of applications



Graphs

###-#-##

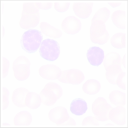
Seeds

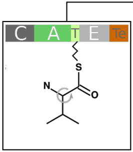


Indexes

High-throughput sequencing

Peptides



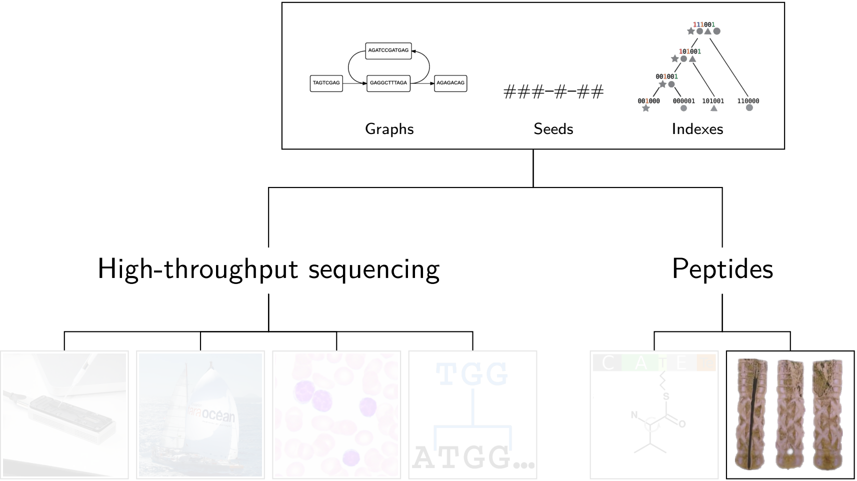


Non-ribosomal peptides

Reference resource
Search of relevant peptides
Expert sourcing

Images: © Univ of Oxford, © Sacha Bollet, CC BY SA Ytngargar, © Maude Pupin, © Olivier Sarrazin

A wide variety of applications



Graphs

Seeds

Indexes

High-throughput sequencing

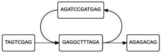
Peptides

Paleoproteomics

Identification of ancestral species

Images: © Univ of Oxford, © Sacha Bollet, CC BY SA Ytngargar, © Maude Pupin, © Olivier Sarrazin

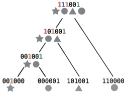
A wide variety of applications



Graphs

###-#-##

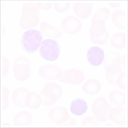
Seeds

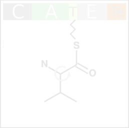


Indexes

High-throughput sequencing

Peptides





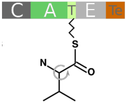
Images: © Univ of Oxford, © Sacha Bollet, CC BY SA Ytngargar, © Maude Pupin, © Olivier Sarrazin

A large software production

Our software production is available online
`https://bioinfo.univ-lille.fr`

Mainly released under open-source licenses

Some success stories of our software

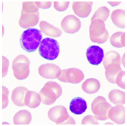


Norine

- 2K queries/month
- 50+ scientists registered to the expert sourcing application
- *service delivery plan* of the european ELIXIR network

SortMeRNA

- Cited more than 2,000 times.
- Distributed in the Qiime2 pipeline.



Vidjil

- Non-profit VidjilNet consortium in InriaSoft: 8 subscribing hospitals fund two engineers
- More than 50,000 samples analyzed

Bonsai – Bioinformatics sequence analysis

cristal.univ-lille.fr/bonsai – @Bonsai_Bioinfo

Efficient methods devised in collaboration with biology/health labs for analyzing biological sequence data

M2 Internship available: *Bioinformatics analysis of SARS-CoV-2: evaluating and improving sgRNA detection methods*