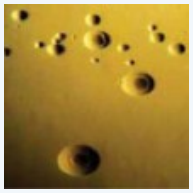
cgfb

BIOINFORMATIQUE

Bacterial genome finishing with Mix and the MolliGen database

« Microbial Bioinformatics day»
Institut Pasteur - 25/09/14

BARRE Aurélien – SEVIN Emeric



Mycoplasma study

- Bacteria lacking a cell wall and parasites for humans, animals, plants or insects
- Small genome → 1 Gb
- Unaffected by many common antibiotics
- Cell culture contamination

- Minimum cell and synthetic biology project based on M. genitalium (C. Venter)
- Local collaboration with INRA 1332 - GDPP

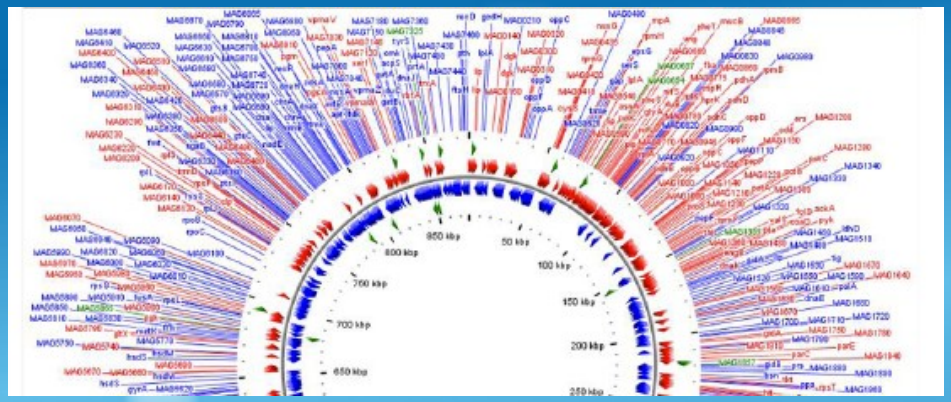
cgfb
BIOINFORMATIQUE

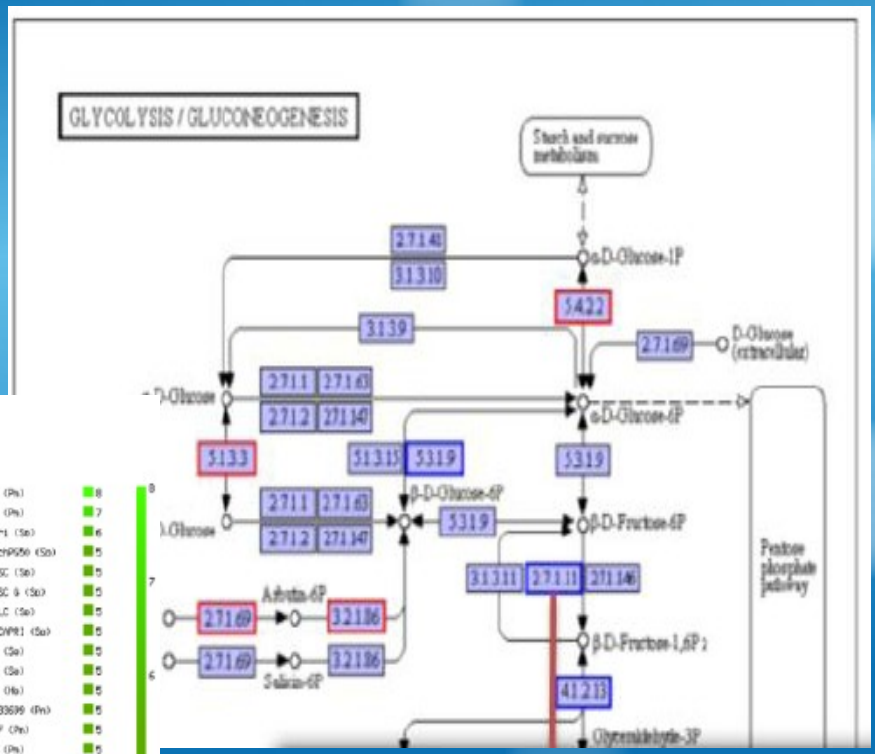
MolliGen

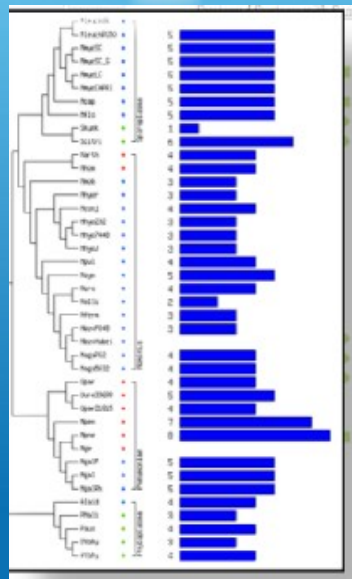
- Web platform dedicated to the comparative genomics of Mollicutes www.molligen.org
- Evolution of the Pasteur « xxxList » web sites

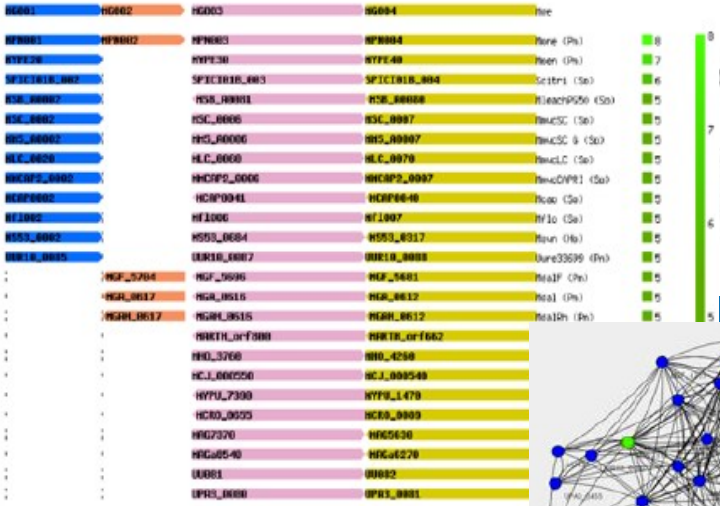
- Over 60 genomes stored, with multiple strains
- Biological experts for curation and annotation
- Large toolbox to analyse and compare genomes
- Private section

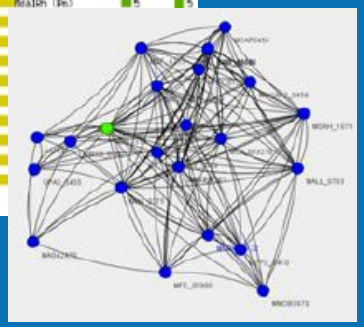
cgfb BIOINFORMATIQUE

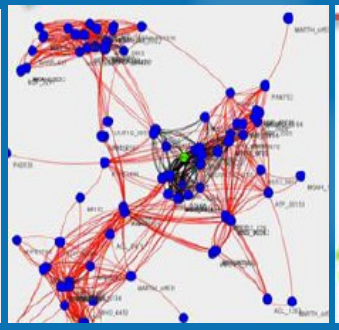


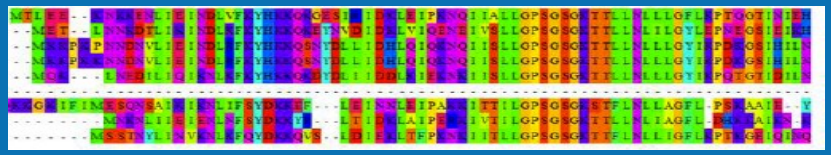

GLYCOLYSIS / GLUCONEOGENESIS
Starch and sucrose metabolism
D-Glucose-1P
D-Glucose-6P
β-D-Fructose-6P
β-D-Fructose-1,6P2
Pentose phosphate pathway









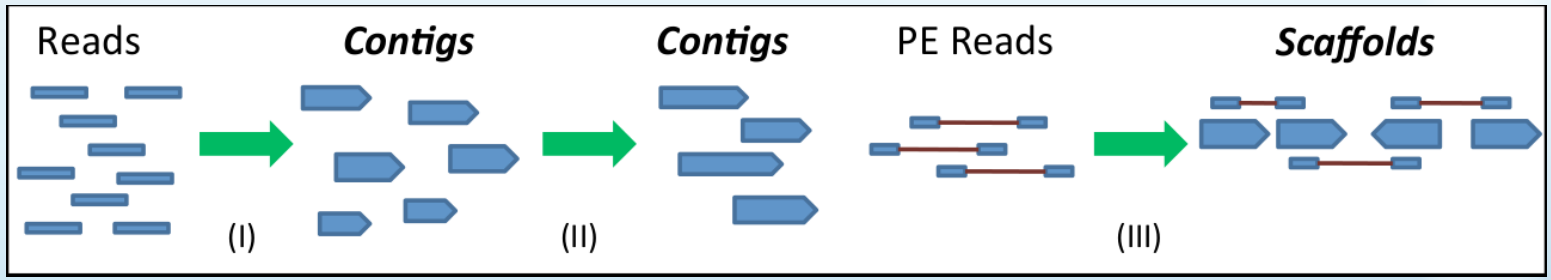


Evolmyco

- ANR evolmyco 2008–2011 : NGS sequencing of 10 new species or strains
- Unfinished genomes produced, with a large number of contigs
- Evident problem in assemblies in regard of existing genomes in MolliGen
- Results not usable as such → finishing strategy

Finishing of genome assemblies

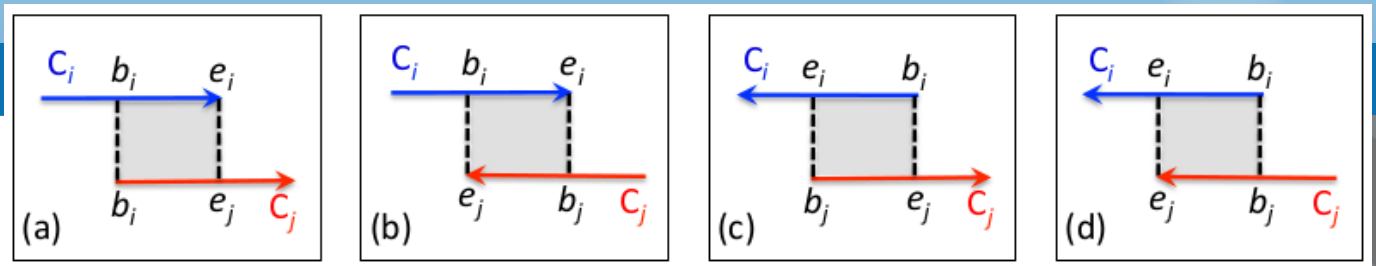
Genome assembly overview



- Goals :
  - Fill the gaps and extend contigs
  - Targeted sequencing to resolve misassemblies and sequence gaps
- In silico finishing
  - Achieve a better quality assembly (w.r.t. fragmentation)
  - Typically by assembly combination

How we *Mix* assemblies

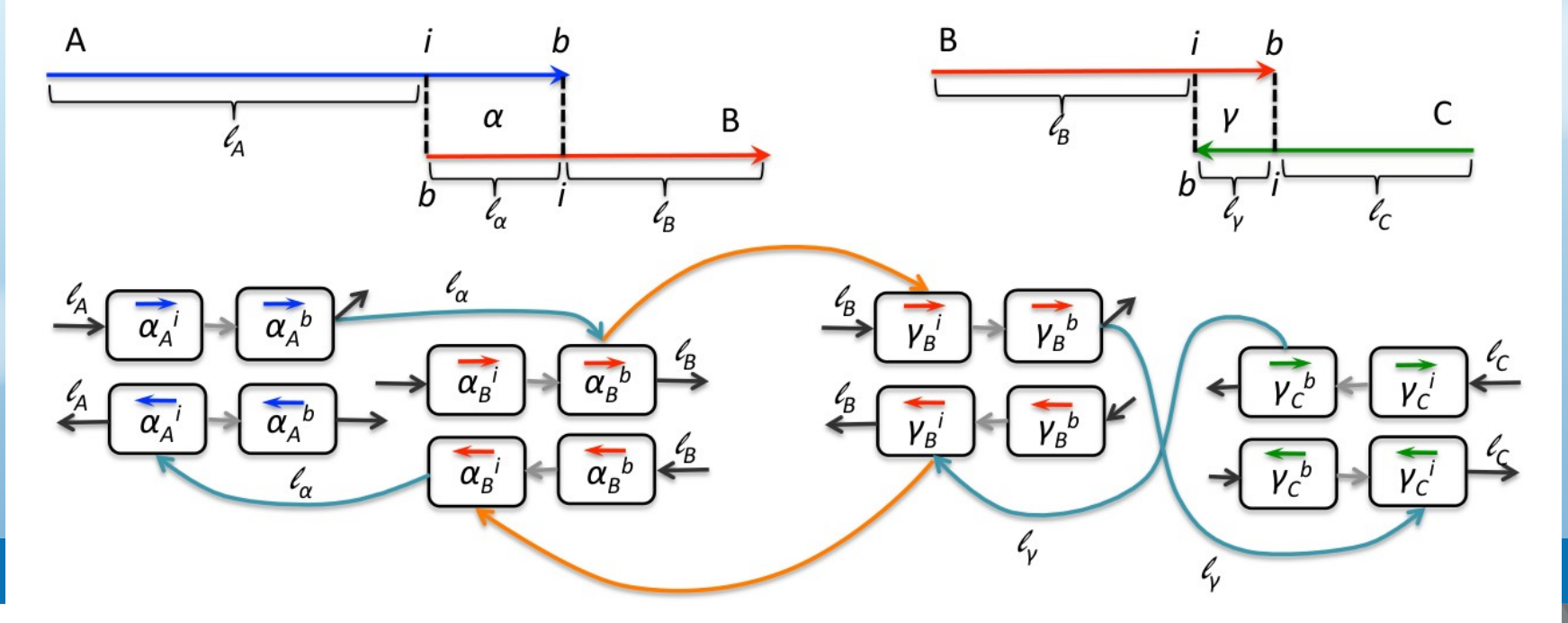
- 2 or more input assemblies - No reference genome, no additional raw reads, no « privileged » assembly
- *Mix* : Input data
  - Assemblies $\mathbb{A}_i = \{C_i\}$ pooled together $\mathbb{A} = \cup \mathbb{A}_i$
  - Alignments set $\mathcal{A}$: each $\mathbb{A}_i \in \mathbb{A}$ aligned against the others
- 4 nodes to represent an alignment zone on a contig
  - Limits of the alignment zone : *b* (boundary) and *i* (internal)
  - How the alignment zone is read : → (forward) and ← (reverse)



cgfb BIOINFORMATIQUE

Extension graph

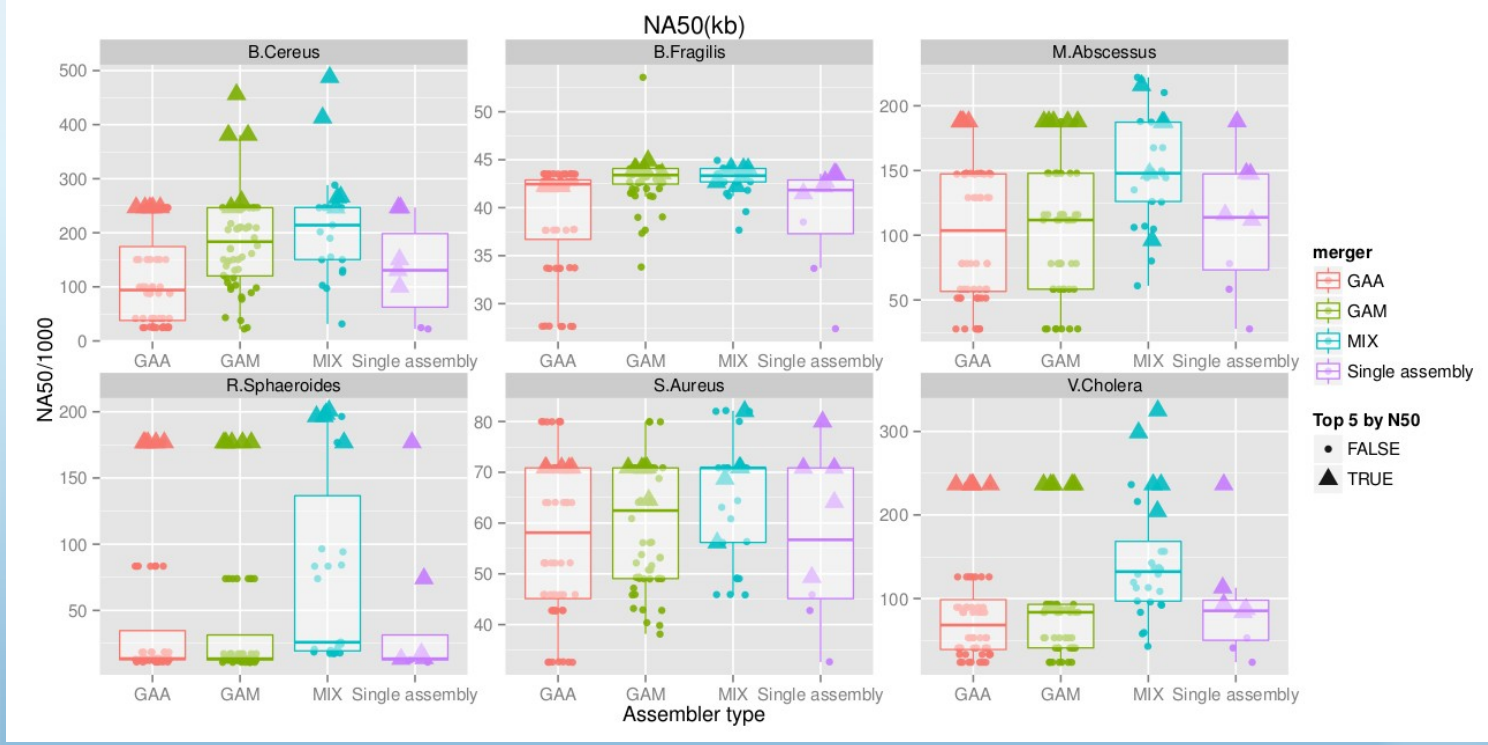
- What is *Mix*ed?
  - Create an extension graph
  - Look for paths that maximize the path length



cgfb BIOINFORMATIQUE

Performance evaluation : the GAGE-B benchmark (Magoc *et al.*, 2013)

- Compares 8 assemblers and 2 assembly mergers



→ Fully correlated with N50

cgfb BIOINFORMATIQUE

Application to the *Mycoplasma*

- 10 *Mycoplasma* genomes
- Three assemblers (CLC, ABySS, MIRA)
- 454 (130~170 nt) and Illumina (36 nt)

- Different assemblers correctly assemble different regions
- *Mix* systematically reduces genome fragmentation
- Validation with core genome
- Annotation and submission to the MolliGen database

https://github.com/cbib/MIX

cgfb BIOINFORMATIQUE

… next steps

- Resulting assemblies are integrated in MolliGen
- Comparison with existing genomes : core gene set & evolution scenario
- MIX will be used as a service in MolliGen to integrate new genomes directly from reads files

cgfb
BIOINFORMATIQUE

Thanks to

From CBiB

- Florence MAURIER
- Hayssam SOUEIDAN
- David BENABEN
- Macha NIKOLSKI

cgfb BIOINFORMATIQUE