Graph analysis of fragmented long-read bacterial genome assemblies

Pierre Marijon, Rayan Chikhi, Jean-Stéphane Varré

Inria, University of Lille

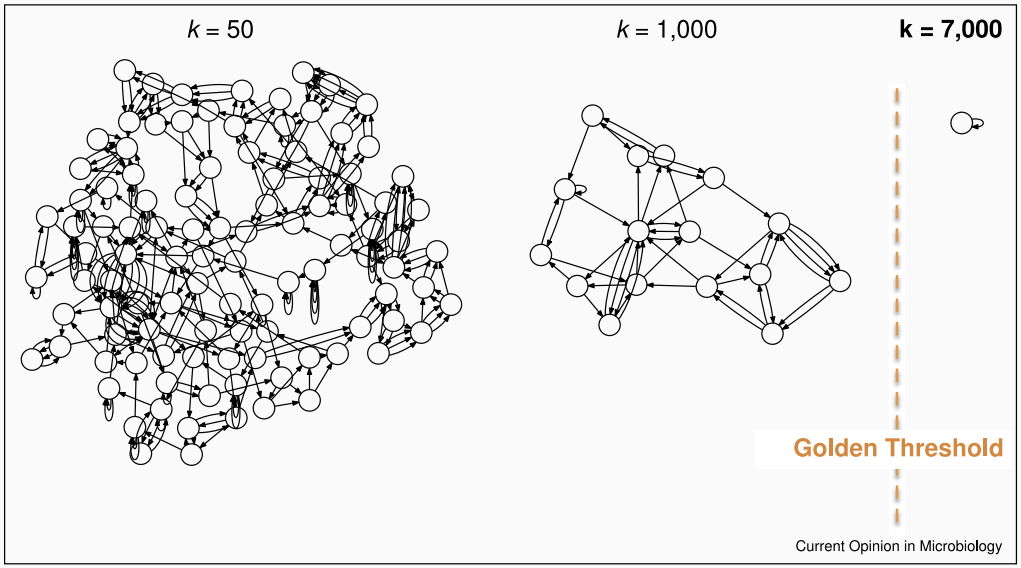
Introduction

Introduction: *de novo* assembly problem, solved ?

Assembly of 3rd generation sequencing data

- requires correction (not my problem today)
- solves almost all genomic repetitions

Assembly graph of the *E. coli* genome[1]:



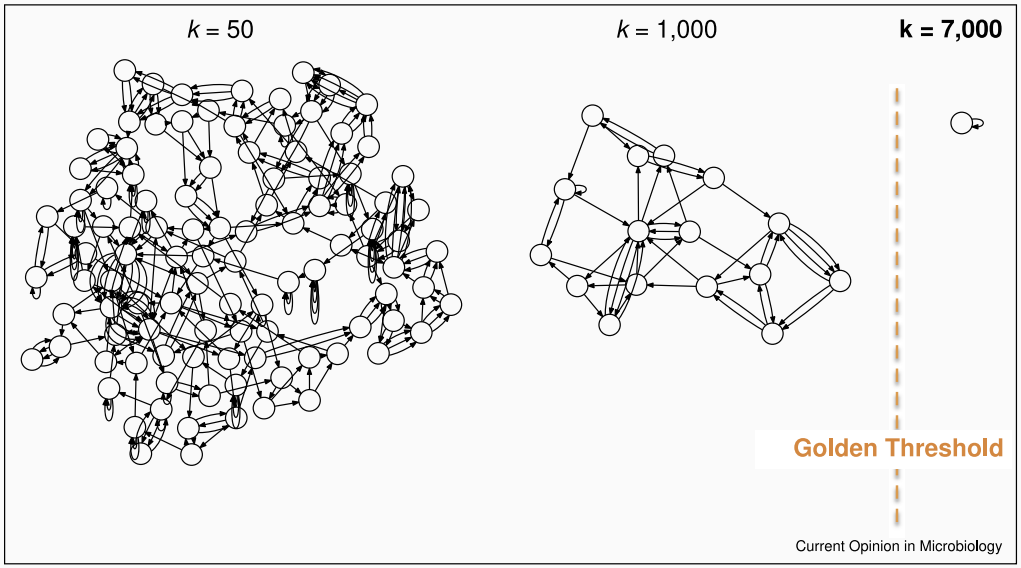
[1] One chromosome, one contig [Koren and Phillippy, 2015]

Introduction: *de novo* assembly problem, solved ?

Assembly of 3rd generation sequencing data

- requires correction (not my problem today)
- solves almost all genomic repetitions

Assembly graph of the *E. coli* genome[1]:



But in reality ...

[1] One chromosome, one contig [Koren and Phillippy, 2015]

Introduction: *de novo* assembly problem, solved ?

NCTC : 3000 bacteria cultures sequenced with PacBio, and assembled with HGAP[2]

599 / 1136 (34 %) assemblies are not single-contig (as of Feb 2019)

Species	Strain	Sample	Runs	Automated Assembly	Manual Assembly	Manual Assembly Chromosome Contig Number	Manual Assembly Plasmid Contig Number	Manual Assembly Unidentified Contig Number
Achromobacter xylosoxidans	NCTC10807	ERS451415	ERR550491 ERR550506 ERR550507	Pending	EMBL	1	0	0
Budvicia aquatica	NCTC12282	ERS462988	ERR581162	Pending	EMBL	2	0	0
Campylobacter jejuni	NCTC11351	ERS445056	ERR550473 ERR550476	Pending	EMBL	1	0	0
Cedecea neteri	NCTC12120	ERS462978	ERR581152 ERR581168 ERR597265	Pending	EMBL	7	1	0
Citrobacter amalonaticus	NCTC10805	ERS485850	ERR601566 ERR601575	Pending	EMBL	1	2	0
Citrobacter freundii	NCTC9750	ERS485849	ERR601559 ERR601565	Pending	EMBL	1	0	0
Citrobacter koseri	NCTC10849	ERS473430	ERR581173	Pending	EMBL	1	1	0
Corynebacterium diphtheriae	NCTC11397	ERS451417	ERR550510	Pending	EMBL	1	0	0
Cronobacter sakazakii	NCTC11467	ERS462977	ERR581151 ERR581167	Pending	EMBL	4	3	0

[2][Chin et al., 2013]

Introduction: *de novo* assembly problem, solved ?

NCTC : 3000 bacteria cultures sequenced with PacBio, and assembled with HGAP[2]

599 / 1136 (34 %) assemblies are not single-contig (as of Feb 2019)

Species	Strain	Sample	Runs	Automated Assembly	Manual Assembly	Manual Assembly Chromosome Contig Number	Manual Assembly Plasmid Contig Number	Manual Assembly Unidentified Contig Number
Achromobacter xylosoxidans	NCTC10807	ERS451415	ERR550491 ERR550506 ERR550507	Pending	EMBL	1	0	0
Budvicia aquatica	NCTC12282	ERS462988	ERR581162	Pending	EMBL	2	0	0
Campylobacter jejuni	NCTC11351	ERS445056	ERR550473 ERR550476	Pending	EMBL	1	0	0
Cedecea neteri	NCTC12120	ERS462978	ERR581152 ERR581168 ERR597265	Pending	EMBL	7	1	0
Citrobacter amalonaticus	NCTC10805	ERS485850	ERR601566 ERR601575	Pending	EMBL	1	2	0
Citrobacter freundii	NCTC9750	ERS485849	ERR601559 ERR601565	Pending	EMBL	1	0	0
Citrobacter koseri	NCTC10849	ERS473430	ERR581173	Pending	EMBL	1	1	0
Corynebacterium diphtheriae	NCTC11397	ERS451417	ERR550510	Pending	EMBL	1	0	0
Cronobacter sakazakii	NCTC11467	ERS462977	ERR581151 ERR581167	Pending	EMBL	4	3	0

Assembly problem is solved for many bacteria but not for all.

[2][Chin et al., 2013]

KNOT: Knowledge Network Overlap exTraction

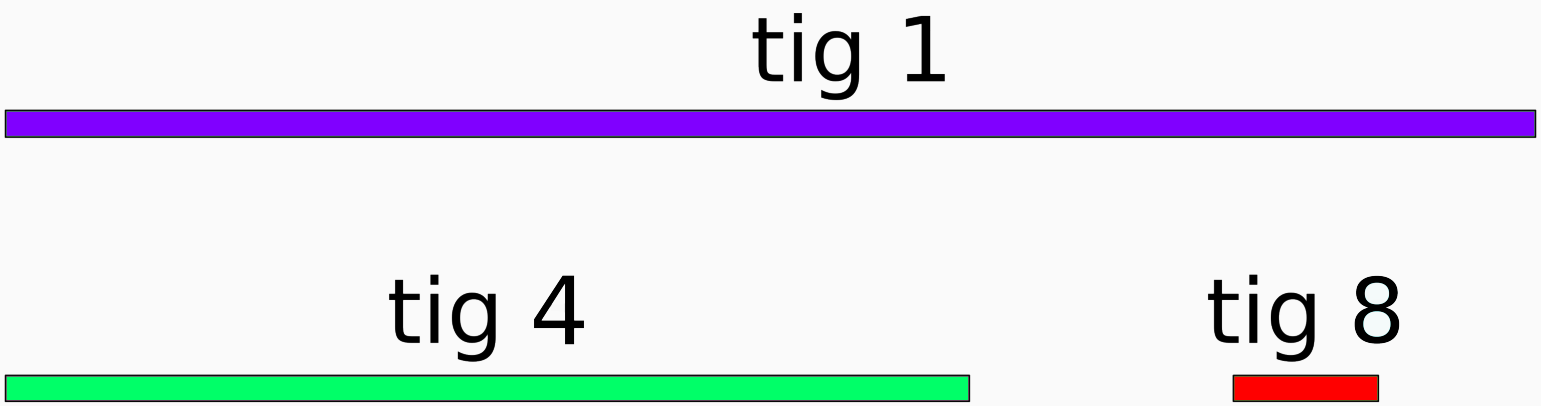
KNOT: A synthetic example

- **Dataset**: *Terriglobus roseus* synthetic pacbio, 20x coverage (LongISLND[3])
- **Assembly tools**: Canu [4]

[3][Lau et al., 2016]
[4][Koren et al., 2017]

KNOT: A synthetic example

- **Dataset**: *Terriglobus roseus* synthetic pacbio, 20x coverage (LongISLND[3])
- **Assembly tools**: Canu [4]

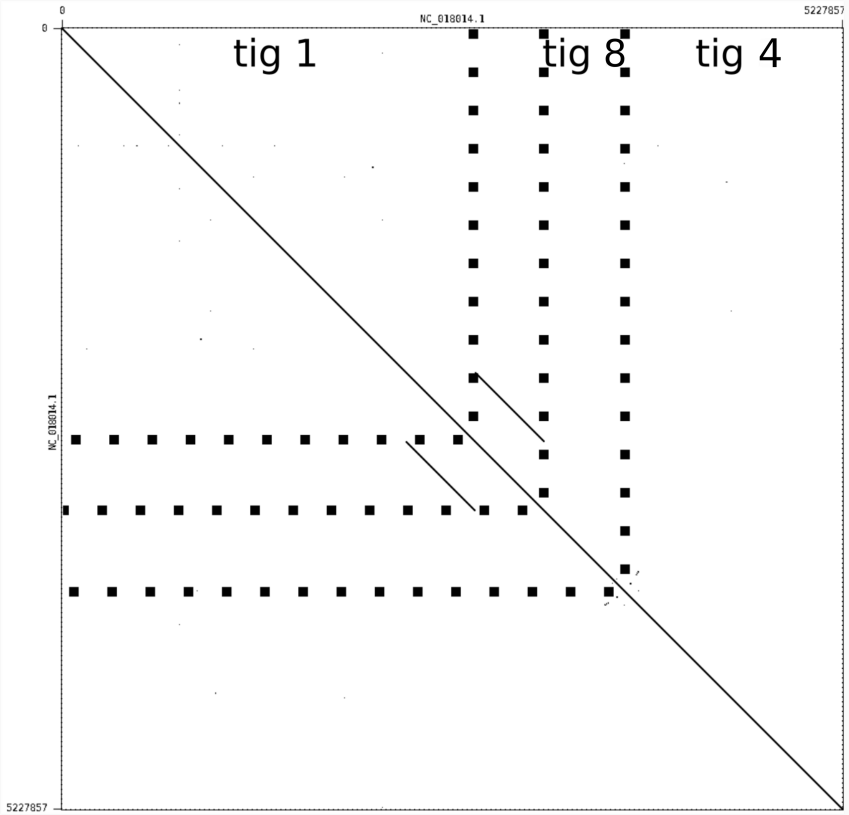


Can we recover missing edges between contigs?

[3][Lau et al., 2016]
[4][Koren et al., 2017]

Not even a repetition problem..

Dotplot of *T. roseus* genome against itself.



Length of the tandem repeat is 460 kbp. The repetition explains only one of the two contig breaks.

KNOT: A synthetic example

An assembly graph can be defined as :

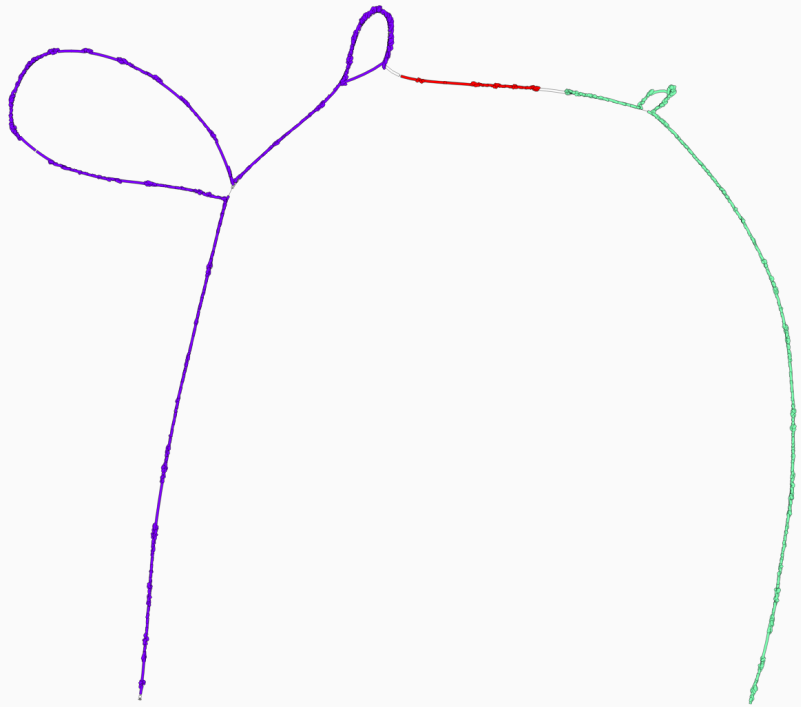
- nodes → reads
- edges → overlaps

[5][Li, 2018]

KNOT: A synthetic example

An assembly graph can be defined as :

- nodes → reads
- edges → overlaps



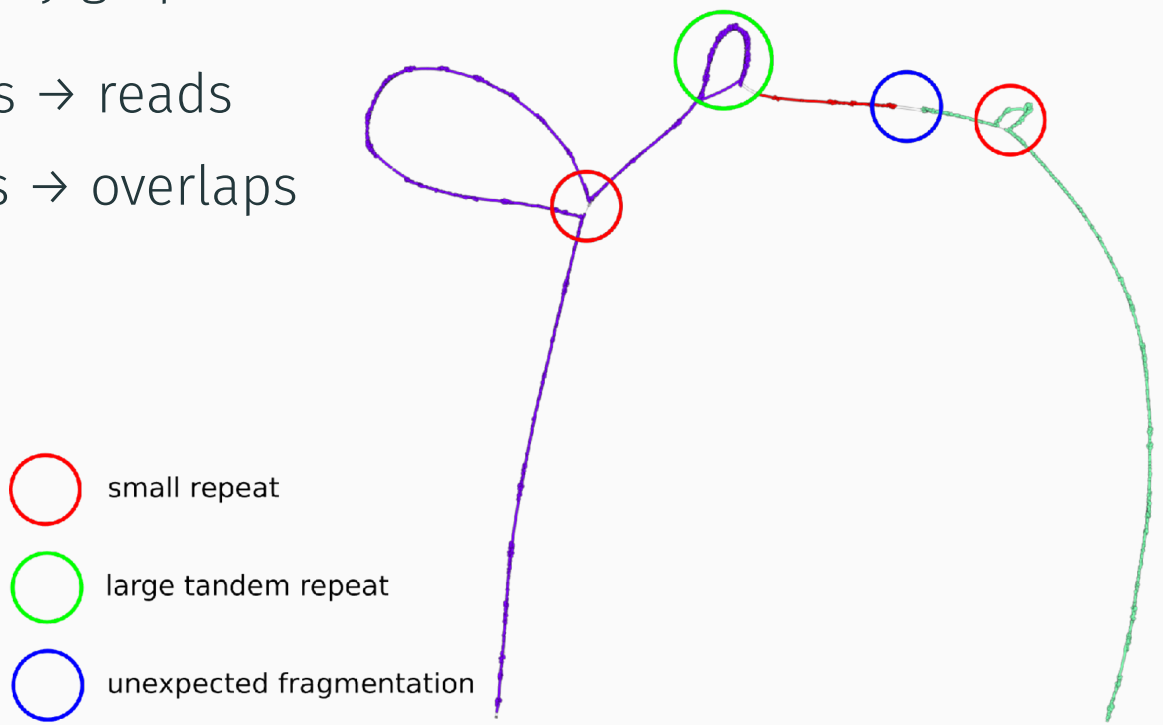
Overlap graph (constructed by `Minimap2` [5]), reads are colored by `Canu` contig.

[5] [Li, 2018]

KNOT: A synthetic example

An assembly graph can be defined as :

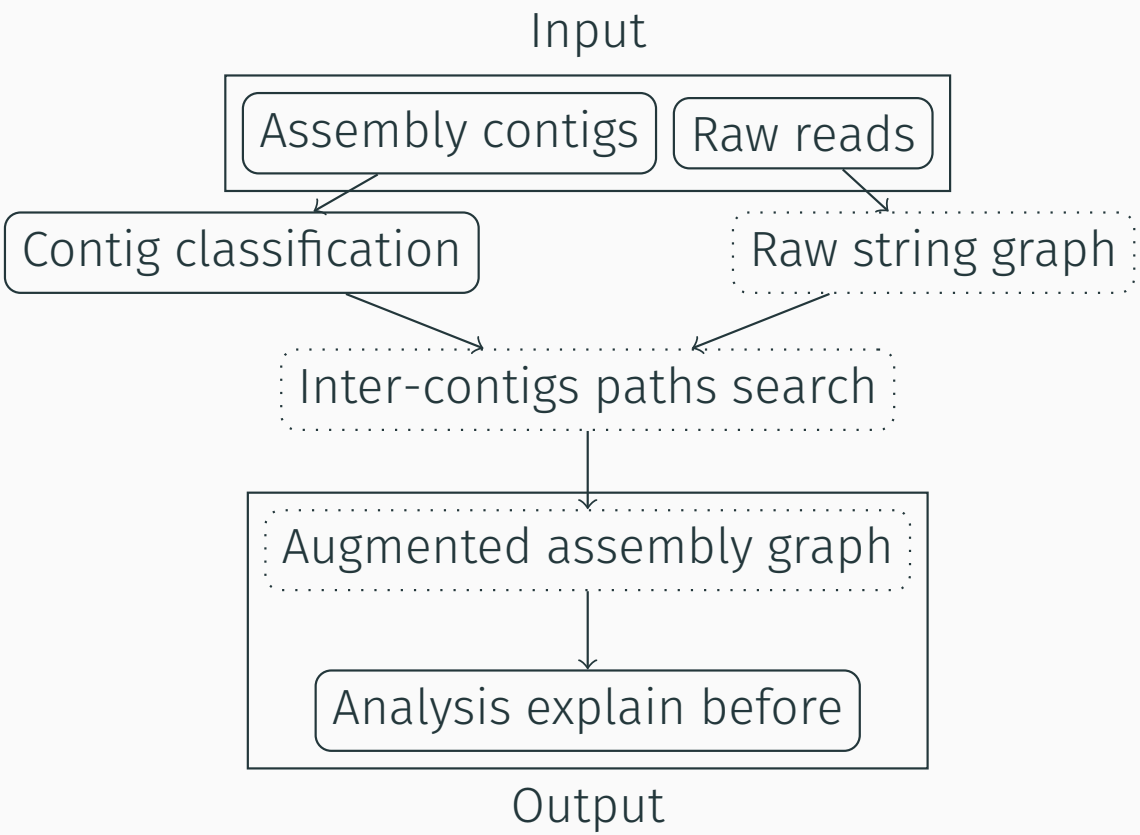
- nodes → reads
- edges → overlaps



Overlap graph (constructed by `Minimap2` [5]), reads are colored by `Canu` contig.

[5] [Li, 2018]

KNOT: Pipeline



KNOT: definition of an Augmented Assembly Graph

The AAG is an undirected, weighted graph:

nodes: contigs extremities

edges:

- between extremities of a contig (weight = 0),
- paths found between contigs (weight = path length in bases)

KNOT: definition of an Augmented Assembly Graph

The AAG is an undirected, weighted graph:

nodes: contigs extremities

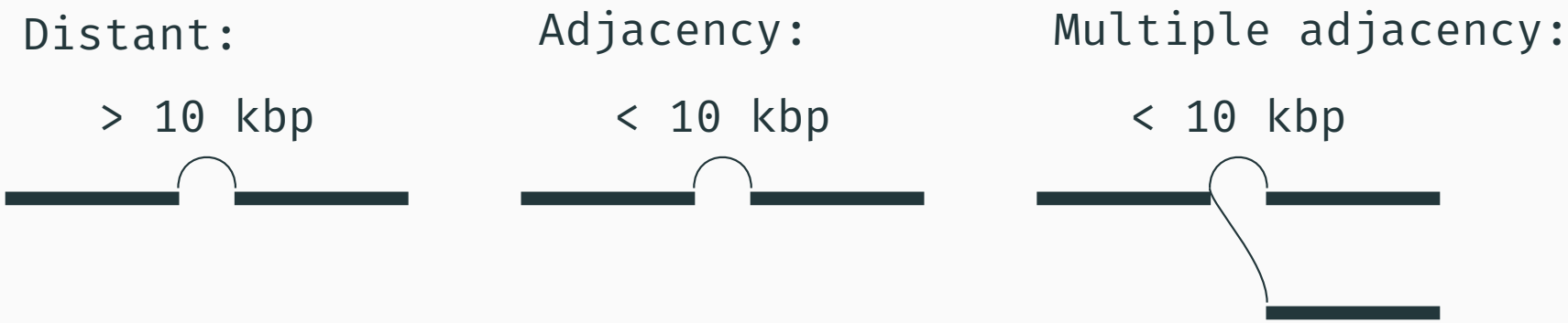
edges:

- between extremities of a contig (weight = 0),
- paths found between contigs (weight = path length in bases)

Plain links are paths compatible with true order of contigs, dotted links are other paths.

KNOT: Path classification

We classify paths based on their length (in base pairs):



In prokaryotes, most repetitions are < 10 kbp [6]

[6] [Treangen et al., 2009]

KNOT: Hamilton walk

AAG's are generally complete graphs. We can enumerate all their Hamilton walks.

The weight of a walk is the of sum of all edge weights.

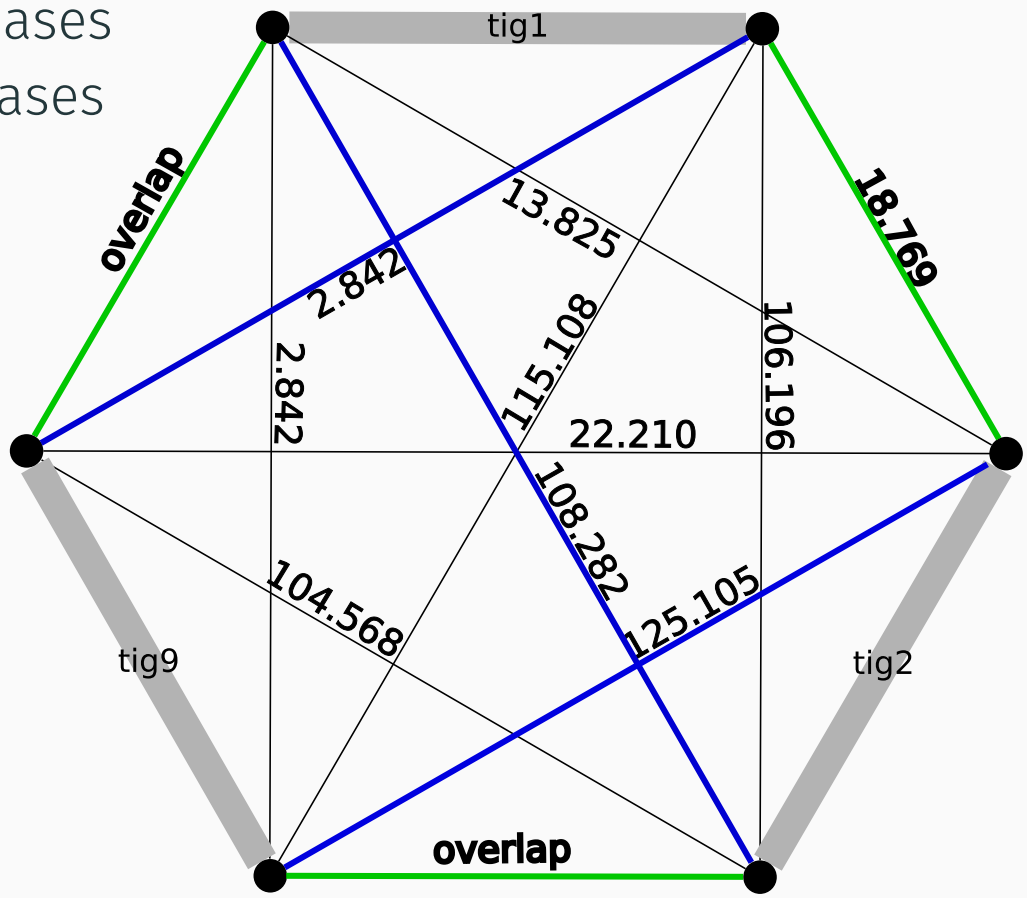
KNOT: Hamilton walk

AAG's are generally complete graphs. We can enumerate all their Hamilton walks.

The weight of a walk is the of sum of all edge weights.

Supposedly: We assume that **lowest-weight walk** is the true genome.

- Green walk weight: 18,769 bases
- Blue walk weight: 136,229 bases



Results on 38 datasets from NCTC3000

We selected 38 datasets from NCTC3000, where `Canu`, `Miniasm` and `Hinge` didn't produce the expected number of chromosomes (*i.e. unsolved assemblies*).

- 19 datasets were *manually solved* by NCTC
- 17 remained fragmented
- 2 with no assembly attempt by NCTC

KNOT: Path classification

Across 38 datasets:

Mean number of	
Canu contigs	4.32
Edges in AAG	32.67
Theoretical max. edges in AAG	41.83
Distant edges	28.64
Adjacency edges	4.02
Dead-ends in **Canu** contigs	4.94
Dead-ends in AAG, adjacency edges	2.70

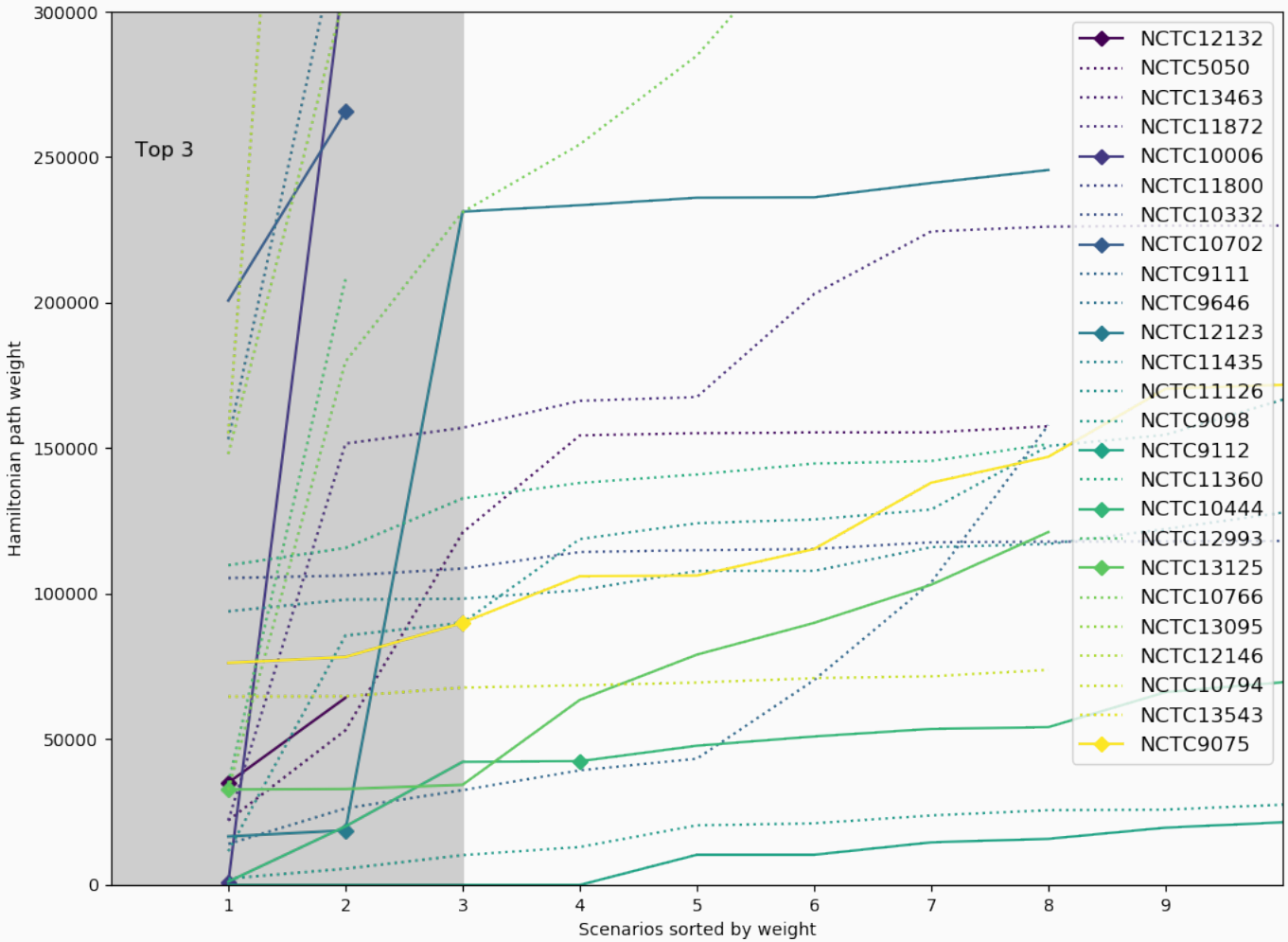
KNOT: Path classification

Across 38 datasets:

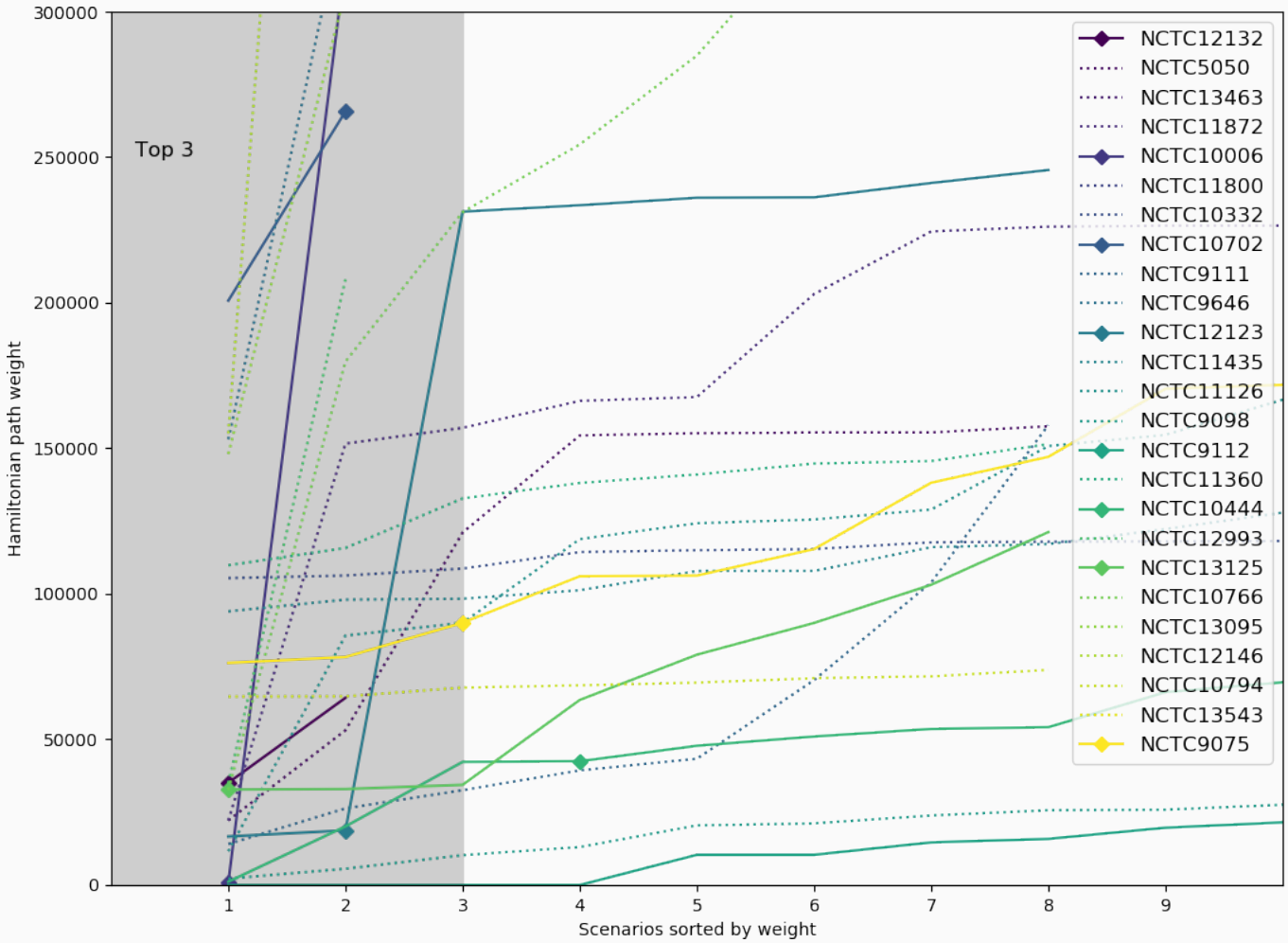
Mean number of	
Canu contigs	4.32
Edges in AAG	32.67
Theoretical max. edges in AAG	41.83
Distant edges	28.64
Adjacency edges	4.02
Dead-ends in **Canu** contigs	4.94
Dead-ends in AAG, adjacency edges	2.70

Almost half of the missing paths in contigs graph are recovered.

KNOT: Hamilton walk



KNOT: Hamilton walk



Generally, the true contig ordering is a low-weight Hamiltonian walk

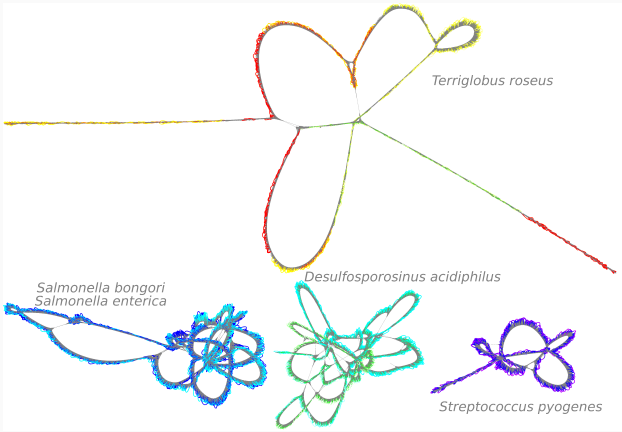
Graph analysis of fragmented long-read bacterial genome assemblies

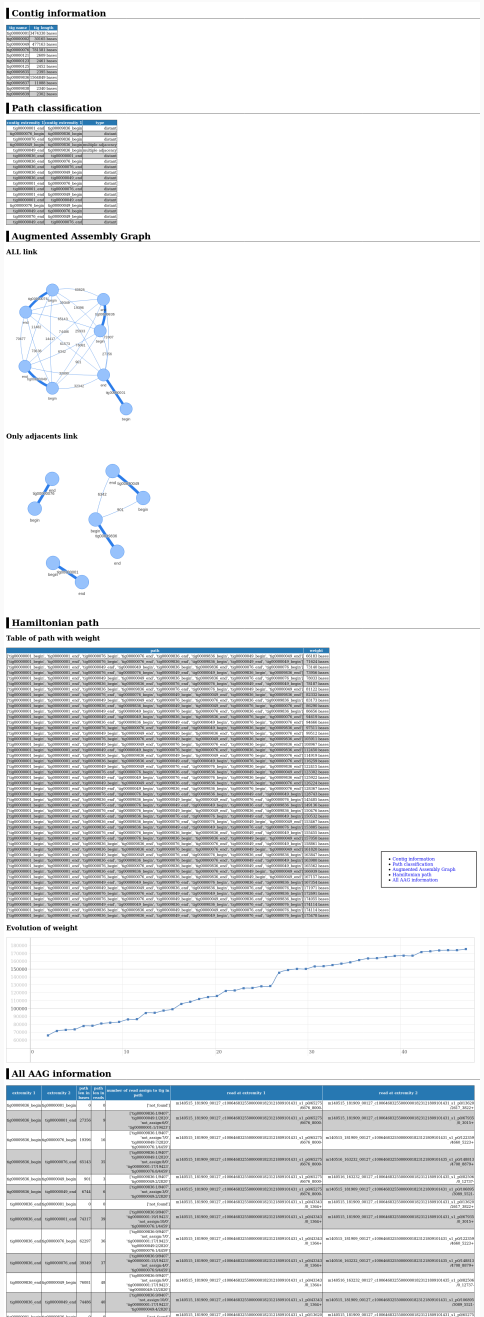
Summary:

- Bacterial assembly is not solved for all datasets
- Build and analyse **Augmented Assembly Graph** can help

Future:

- Assembly graph between contig
- Biological validation (we search collaboration)
- Application to larger genome/metagenome
- Performance improvement (path search step)





https://gitlab.inria.fr/pmarijon/knot

@pierre_marijon

References i

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E. E., Turner, S. W., and Korlach, J. (2013).
Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data.
Nature Methods, 10(6):563–569.

Koren, S. and Phillippy, A. M. (2015).
One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly.
Current Opinion in Microbiology, 23:110–120.

References ii


Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017).
Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.
Genome Research, 27(5):722–736.


Lau, B., Mohiyuddin, M., Mu, J. C., Fang, L. T., Asadi, N. B., Dallett, C., and Lam, H. Y. K. (2016).
LongISLND:in silicosequencing of lengthy and noisy datatypes.
Bioinformatics, 32(24):3829–3832.


Li, H. (2018).
Minimap2: pairwise alignment for nucleotide sequences.
Bioinformatics, 34(18):3094–3100.

References iii


Treangen, T. J., Abraham, A.-L., Touchon, M., and Rocha, E. P. (2009).
Genesis, effects and fates of repeats in prokaryotic genomes.
FEMS Microbiology Reviews, 33(3):539–571.