EVALUATION AND BENCHMARKING OF A NEW SCAFFOLDING METHODOLOGY

ALEXANDRINA BODRUG

SUPERVISORS: PR. RUMEN ANDONOV & DR. DOMINIQUE LAVENIER

UNIVERSITY RENNES 1

BIOINFORMATICS AND GENOMICS MASTER

JUNE 23, 2015

Overview

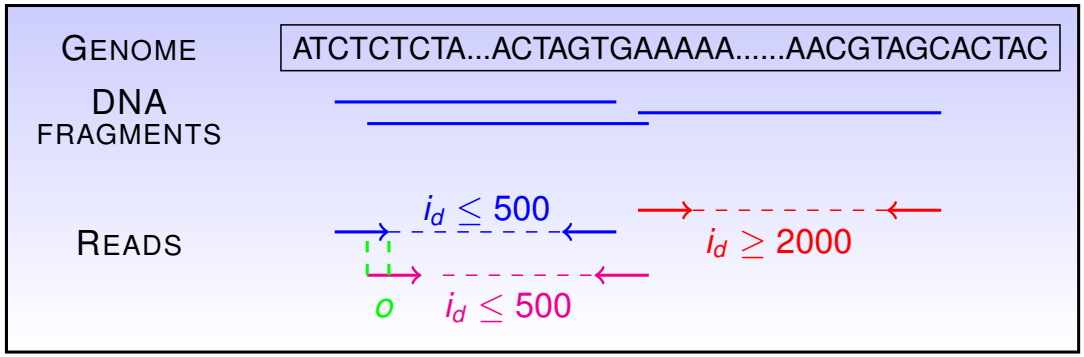
CONTEXT

"The *Contig Scaffolding Problem* is to order and orientate the given contigs in a manner that is consistent with as many mate-pairs as possible".

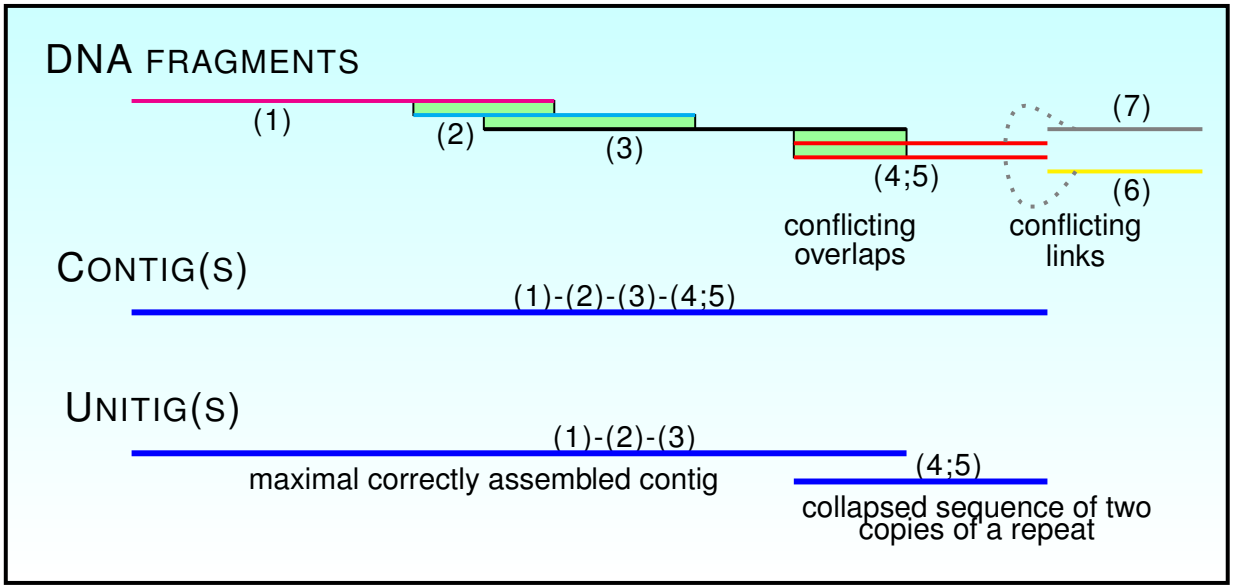
Hudson *et al.* 2002

SOME DEFINITIONS

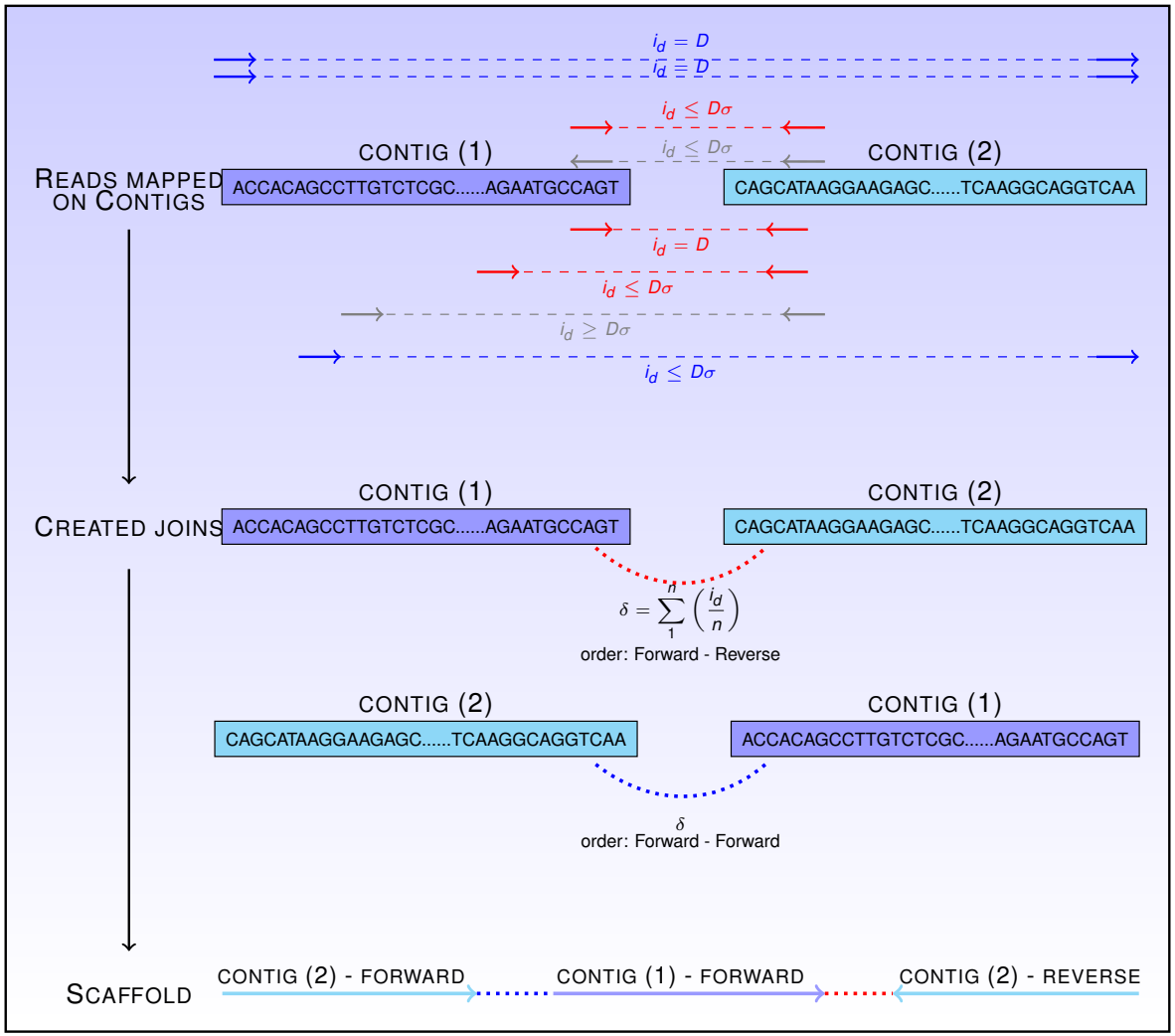
Genome is fragmented, extremities are sequenced ($\mapsto$ reads) . . .



. . . reads are assembled though high-confidence overlappings into contigs or unitigs.



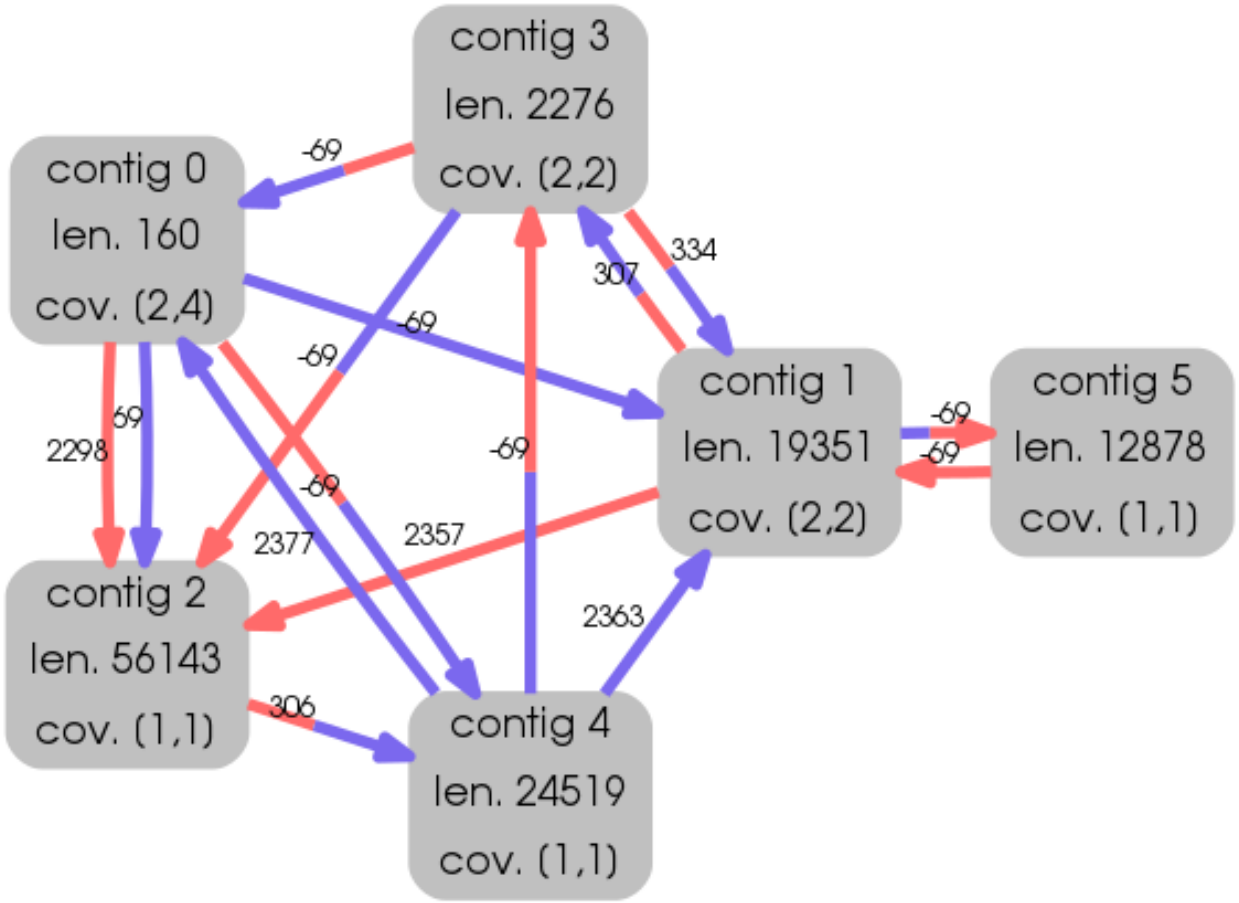
ORDER AND ORIENT



- uses unitigs instead of contigs to better compute unitig coverage
- uses unitig coverages to duplicated regions
- several models exist, their common point is that for each unitig occurence they create a node
- ... and for each unitig orientation, a different node is yet again created

THE RAW INPUT DATA

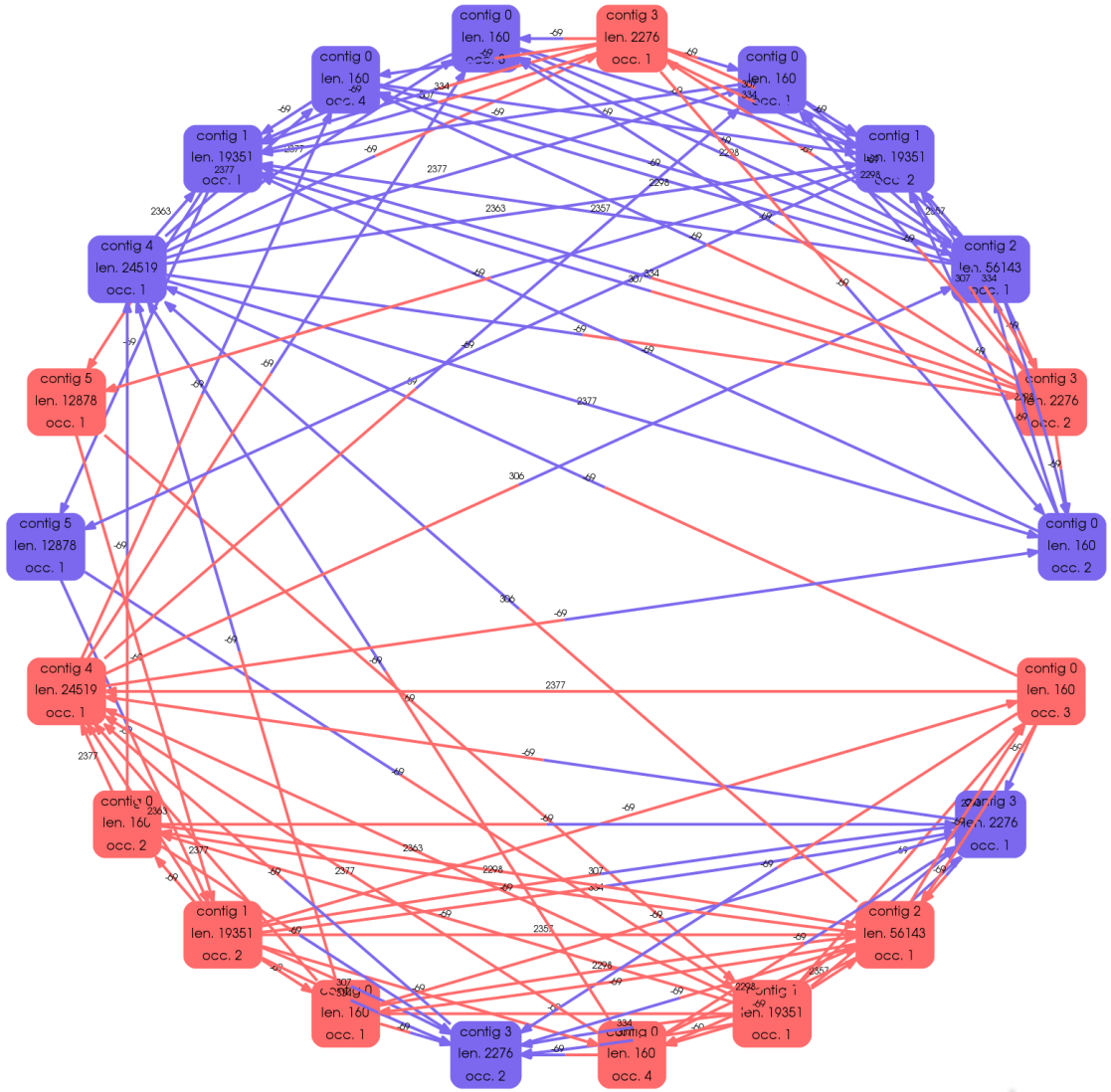
RAW INPUT DATA OF AGROSTIS STOLONIFERA



Input data of agrostis.txt (1 contig, 1 node)

THE RAW INPUT DATA

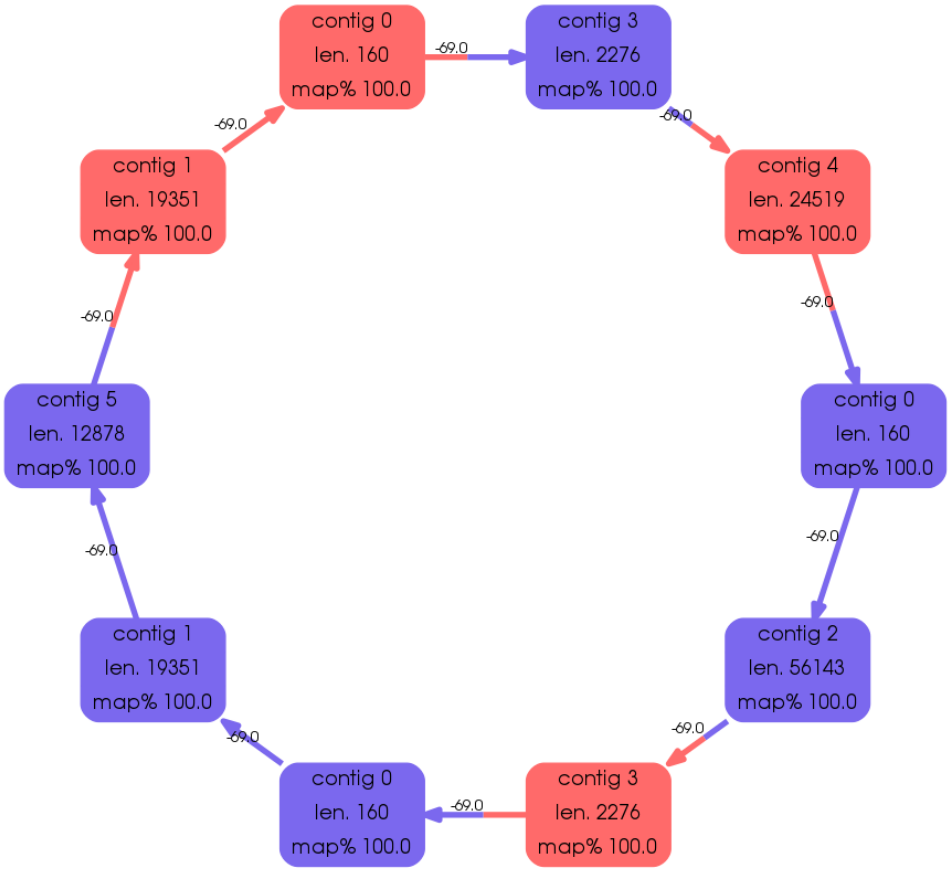
GRAPH TO BE SOLVED BY THE GST



Input data of agrostis.txt (1 occurence, 1 node)

The raw input data

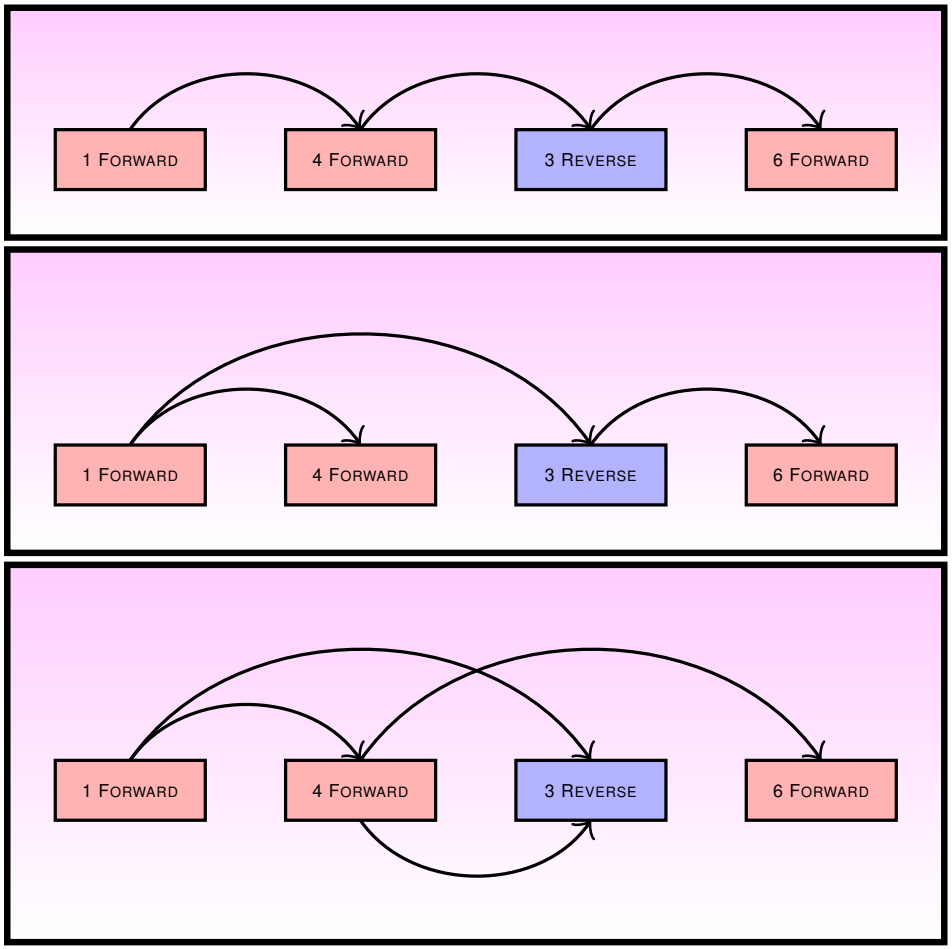
Expected solution of Agrostis stolonifera



Golden standard / mapping solution of agrostis.txt (1 occurence, 1 node)

CHALLENGING PROBLEM

What would you do in these situations?



SCRIPTING FOR THE GST

- a script to visualize input data and GST solutions: `graph_generator.py`
- a script to inspect the features of the modeled input graph: `graph_inspector.py`
- a script to automatically detect correctly solved instances: `graph_comparator.py`

BENCHMARKING WORKFLOW FOR TGE GST

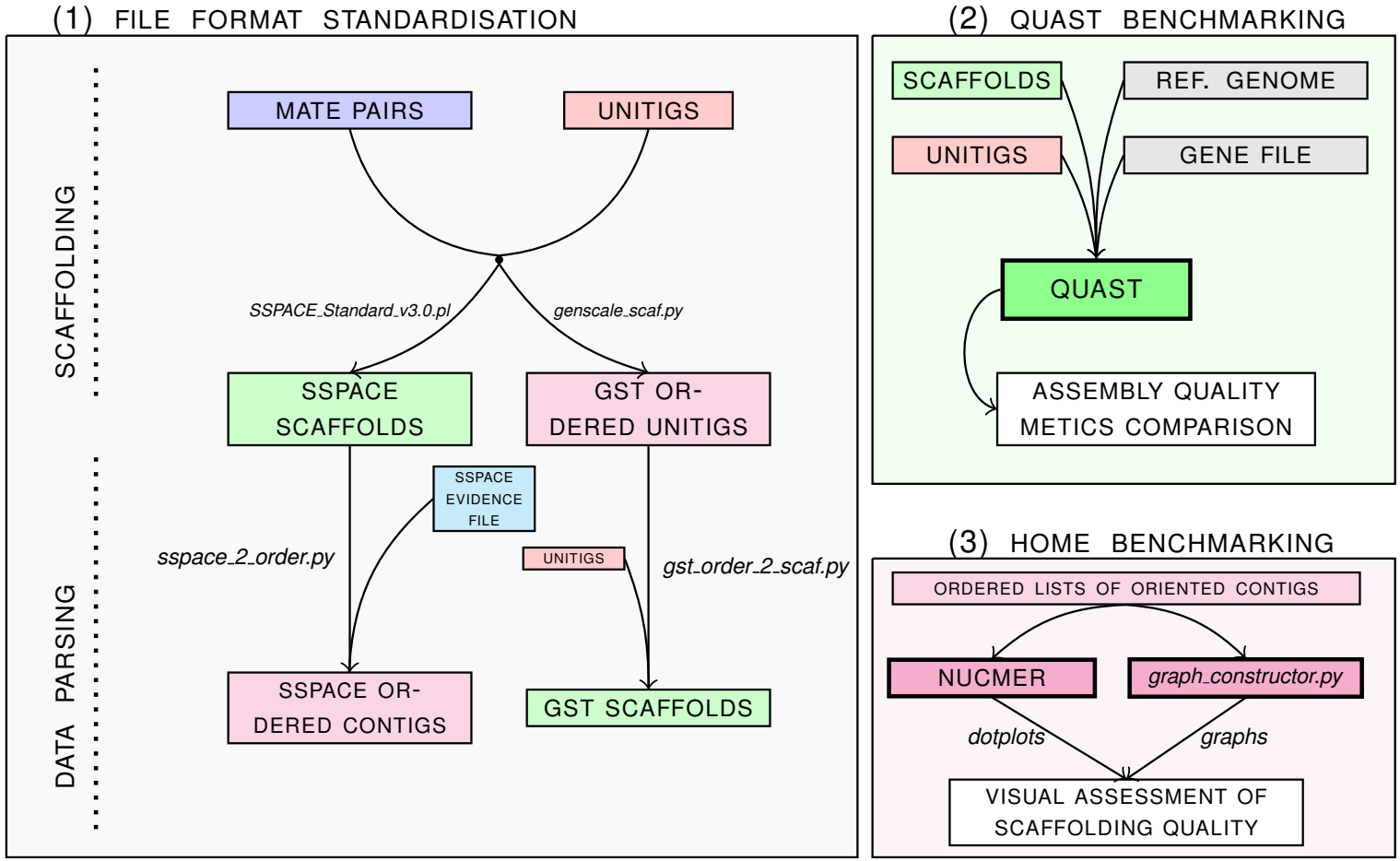
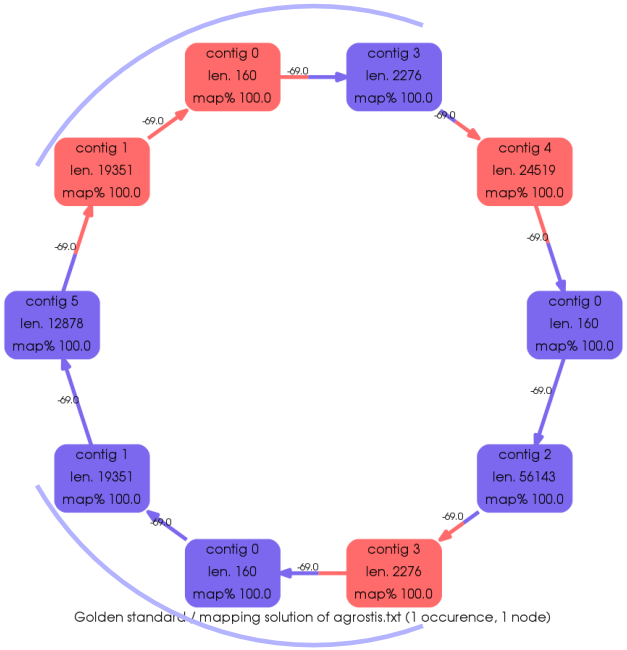


Figure : Benchmarking workflow

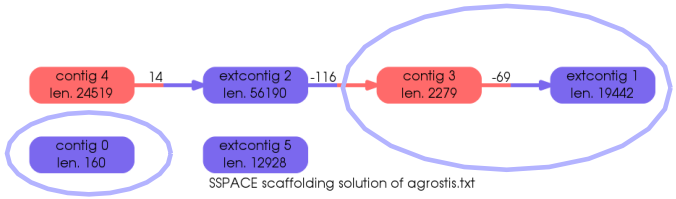
Results

- Genomes with big repeated regions were solved a lot better than SSPACE
- Small repeats are very challenging to assemble because too many conflicting links exists and GST can not take a decision or is too slow

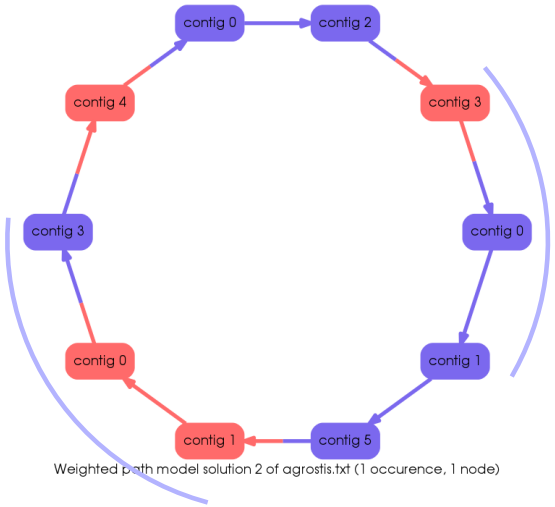
Example of Agrostis stolonifera



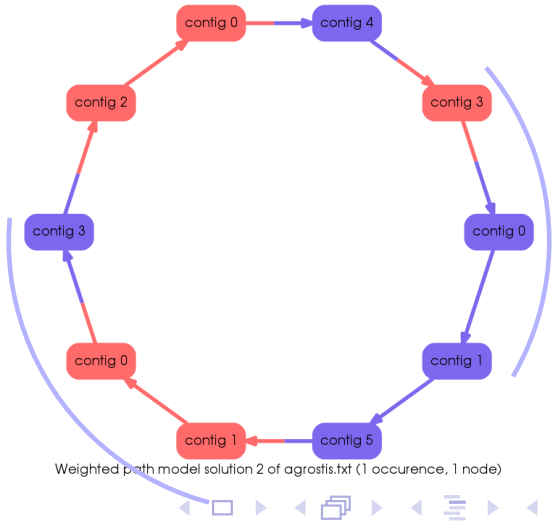
Golden standard / mapping solution of agrostis.txt (1 occurence, 1 node)



SSPACE scaffolding solution of agrostis.txt



Weighted path model solution 2 of agrostis.txt (1 occurence, 1 node)



Weighted path model solution 2 of agrostis.txt (1 occurence, 1 node)

PERSPECTIVES

- Find strategies which solve more challenging data (flow model)
- Scaffold bacterial data
- Test the GST with real data

Thanks!

The End