

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

SCAFFOLDING
BENCHMARK-
ING

ALEXANDRINA
BODRUG

Context

Context

Some definitions
Order and orient

Genscale
scaffolding
tools features

The raw input data
GST modeled graph
Expected solution

Challenging
problem

Scripting for
the GST

Benchmarking
workflow for
the GST

Results

Scaffolding solutions
example: Agrostis

1 Context

2 Context

- Some definitions
- Order and orient

3 Genscale scaffolding tools features

- The raw input data
- GST modeled graph
- Expected solution

4 Challenging problem

5 Scripting for the GST

6 Benchmarking workflow for the GST

7 Results

- Scaffolding solutions example: *Agrostis stolonifera*

8 Perspectives

??

SCAFFOLDING BENCHMARK- ING

ALEXANDRINA
BODRUG

Context

Context

Some definitions
Order and orient

Genscale scaffolding tools features

The raw input data
GST modeled graph
Expected solution

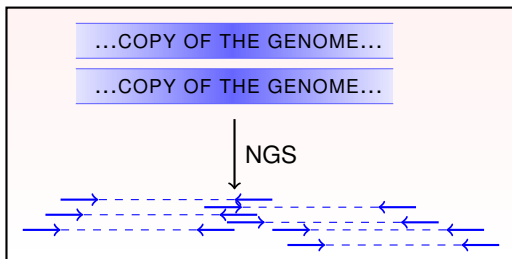
Challenging problem

Scripting for the GST

Benchmarking workflow for the GST

Results

Scaffolding solutions
example: Agrostis



??

SCAFFOLDING BENCHMARK- ING

ALEXANDRINA
BODRUG

Context

Context

Some definitions
Order and orient

Genscale scaffolding tools features

The raw input data
GST modeled graph
Expected solution

Challenging problem

Scripting for the GST

Benchmarking workflow for the GST

Results

Scaffolding solutions
example: *Agrostis*

"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

??

SCAFFOLDING BENCHMARK- ING

ALEXANDRINA
BODRUG

Context

Context

Some definitions
Order and orient

Genscale
scaffolding
tools features

The raw input data
GST modeled graph
Expected solution

Challenging
problem

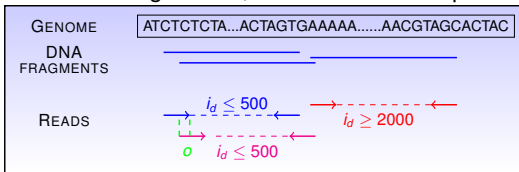
Scripting for
the GST

Benchmarking
workflow for
the GST

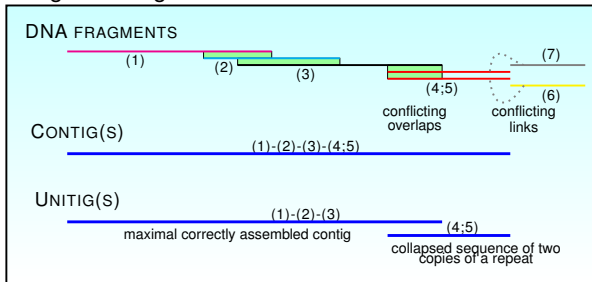
Results

Scaffolding solutions
example: Agrostis

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



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



??

SCAFFOLDING BENCHMARK- ING

ALEXANDRINA
BODRUG

Context

Context

Some definitions
Order and orient

Genscale
scaffolding
tools features

The raw input data
GST modeled graph
Expected solution

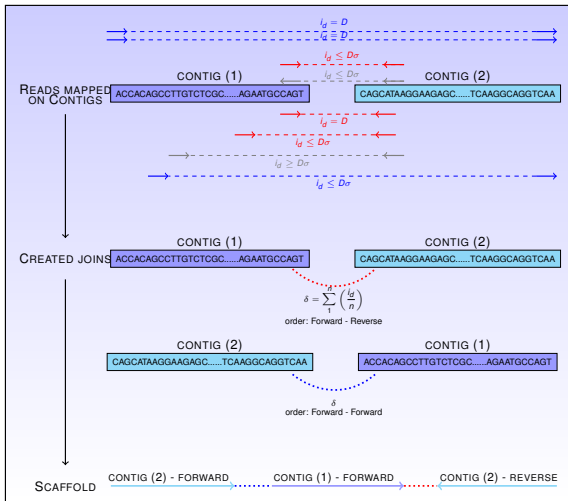
Challenging
problem

Scripting for
the GST

Benchmarking
workflow for
the GST

Results

Scaffolding solutions
example: Agrostis



??

SCAFFOLDING BENCHMARK- ING

ALEXANDRINA
BODRUG

Context

Context

Some definitions

Order and orient

Genscale scaffolding tools features

The raw input data

GST modeled graph

Expected solution

Challenging problem

Scripting for the GST

Benchmarking workflow for the GST

Results

Scaffolding solutions

example: Agrostis

- 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 occurrence they create a node
- ... and for each unitig orientation, a different node is yet again created

??

SCAFFOLDING BENCHMARK- ING

ALEXANDRINA
BODRUG

Context

Context

Some definitions
Order and orient

Genscale
scaffolding
tools features

The raw input data
GST modeled graph
Expected solution

Challenging
problem

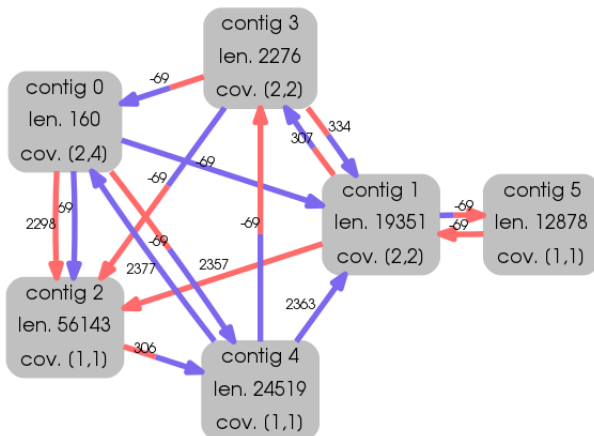
Scripting for
the GST

Benchmarking
workflow for
the GST

Results

Scaffolding solutions
example: Agrostis

RAW INPUT DATA OF AGROSTIS STOLONIFERA



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

??

SCAFFOLDING BENCHMARK- ING

ALEXANDRINA BODRUG

Context

Context

Some definitions
Order and orient

Genscale scaffolding tools features

The raw input data
GST modeled graph
Expected solution

Challenging problem

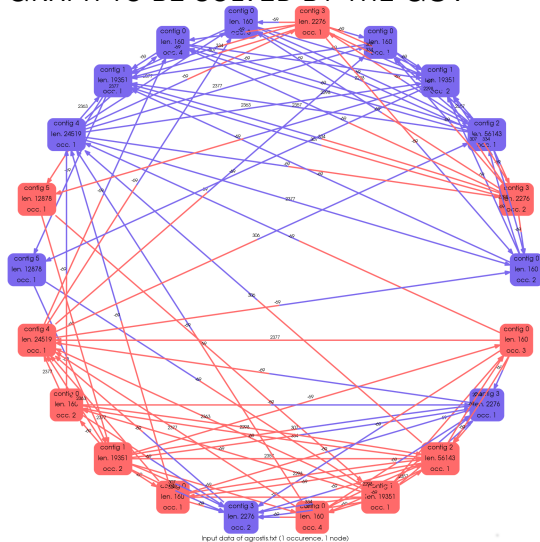
Scripting for the GST

Benchmarking workflow for the GST

Results

Scaffolding solutions
example: Agrostis

GRAPH TO BE SOLVED BY THE GST



??

SCAFFOLDING BENCHMARK- ING

ALEXANDRINA
BODRUG

Context

Context

Some definitions
Order and orient

Genscale scaffolding tools features

The raw input data
GST modeled graph
Expected solution

Challenging problem

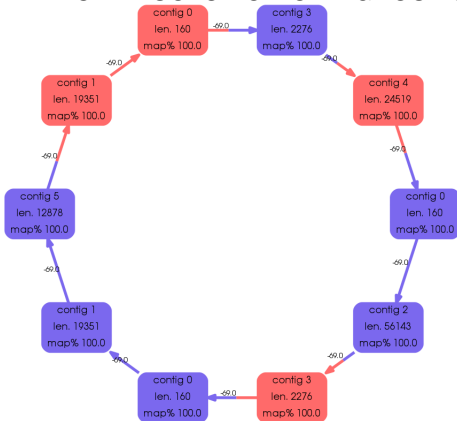
Scripting for the GST

Benchmarking workflow for the GST

Results

Scaffolding solutions
example: Agrostis

EXPECTED SOLUTION OF AGROSTIS STOLONIFERA



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

??

SCAFFOLDING BENCHMARK- ING

ALEXANDRINA
BODRUG

Context

Context

Some definitions
Order and orient

Genscale scaffolding tools features

The raw input data
GST modeled graph
Expected solution

Challenging problem

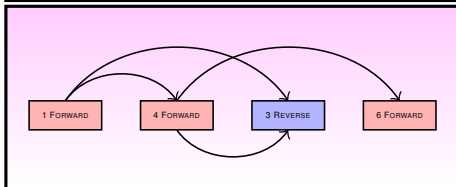
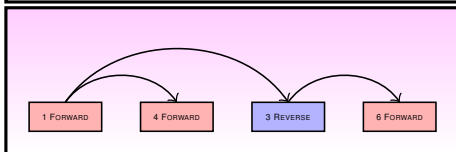
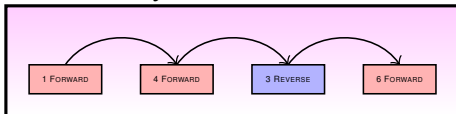
Scripting for the GST

Benchmarking workflow for the GST

Results

Scaffolding solutions
example: Agrostis

What would you do in these situations?



??

SCAFFOLDING BENCHMARK- ING

ALEXANDRINA
BODRUG

Context

Context

Some definitions
Order and orient

Genscale scaffolding tools features

The raw input data
GST modeled graph
Expected solution

Challenging problem

Scripting for the GST

Benchmarking workflow for the GST

Results

Scaffolding solutions
example: Agrostis

- 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`

??

SCAFFOLDING BENCHMARKING

ALEXANDRINA
BODRUG

Context

Context

Some definitions
Order and orient

Genscale scaffolding
tools features

The raw input data
GST modeled graph
Expected solution

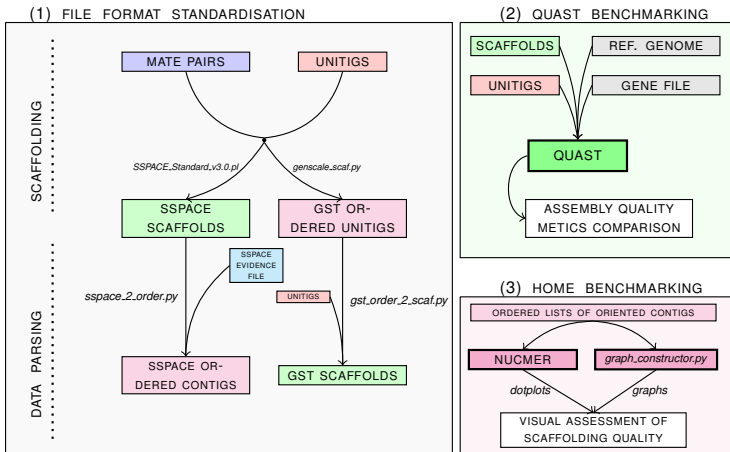
Challenging
problem

Scripting for
the GST

Benchmarking
workflow for
the GST

Results

Scaffolding solutions
example: Agrostis



??

SCAFFOLDING BENCHMARK- ING

ALEXANDRINA
BODRUG

Context

Context

Some definitions
Order and orient

Genscale scaffolding tools features

The raw input data
GST modeled graph
Expected solution

Challenging problem

Scripting for the GST

Benchmarking workflow for the GST

Results

Scaffolding solutions
example: Agrostis

- 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

??

SCAFFOLDING BENCHMARK- ING

ALEXANDRINA
BODRUG

Context

Context

Some definitions
Order and orient

Genscale scaffolding tools features

The raw input data
GST modeled graph
Expected solution

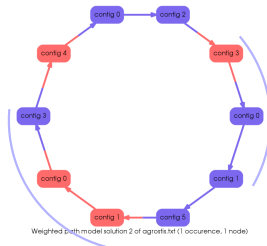
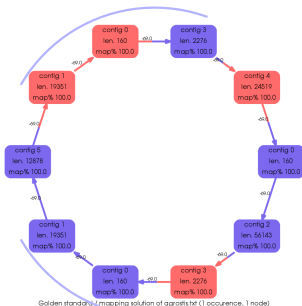
Challenging problem

Scripting for the GST

Benchmarking workflow for the GST

Results

Scaffolding solutions
example: Agrostis



??

SCAFFOLDING BENCHMARK- ING

ALEXANDRINA
BODRUG

Context

Context

Some definitions
Order and orient

Genscale scaffolding tools features

The raw input data
GST modeled graph
Expected solution

Challenging problem

Scripting for the GST

Benchmarking workflow for the GST

Results

Scaffolding solutions
example: Agrostis

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

SCAFFOLDING BENCHMARK- ING

ALEXANDRINA
BODRUG

Context

Context

Some definitions

Order and orient

Genscale
scaffolding
tools features

The raw input data

GST modeled graph

Expected solution

Challenging
problem

Scripting for
the GST

Benchmarking
workflow for
the GST

Results

Scaffolding solutions

example: Agrostis

Thanks!

The End