SCAFFOLDING BENCHMARK-ING

ALEXANDRINA BODRUG

Some definitions

Genscale

scaffolding

The raw input data

GST modeled graph

Challenging

Scripting fo

Benchmarking workflow for

Results

Example of Agros

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 Some definition Order and orien

Genscale scaffolding

The raw input data
GST modeled graph

Challengin

Scripting for

Benchmarking workflow for tge GST

Results
Example of Agrosti

- 1 Context
 - Some definitions
 - Order and orient
- 2 Genscale scaffolding tools features
 - The raw input data
 - GST modeled graph
 - Expected solution
- 3 Challenging problem
- 4 Scripting for the GST
- 5 Benchmarking workflow for tge GST
- 6 Results
 - Example of Agrostis stolonifera
- 7 Perspectives

CONTEXT

SCAFFOLDING. **RENCHMARK-**ING

AI EXANDRINA BODRUG

Context

The raw input data GST modeled graph

workflow for

"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

SCAFFOLDING BENCHMARK-ING ALEXANDRINA

Bodrug

Some definitions

Order and orien

Genscale scaffolding

The raw input data
GST modeled graph

Challenging

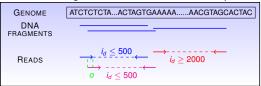
problem

Scripting fo the GST

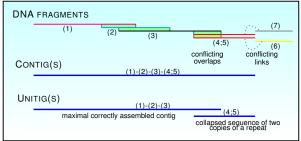
Benchmarking workflow for tge GST

Example of Agrosti stolonifera

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



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



ORDER AND ORIENT

SCAFFOLDING BENCHMARK-ING

ALEXANDRINA BODRUG

Some definition

Order and orient

Genscale scaffolding

tools features
The raw input data

GST modeled graph

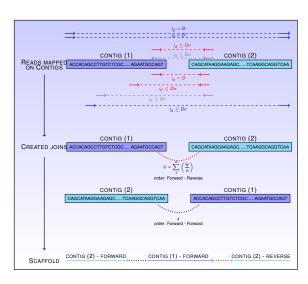
Challenging

problem

Scripting for the GST

Benchmarking workflow for

Results
Example of Agrost stolonifera



SCAFFOLDING BENCHMARK-ING

ALEXANDRINA BODRUG

Context Some definitions Order and orient

Genscale scaffolding tools features

The raw input data

GST modeled graph Expected solution

Challenging problem

Scripting fo

Benchmarking workflow for

Results

Example of Agros stolonifera

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 uniting orientation, a different node is yet again created

THE RAW INPUT DATA

SCAFFOLDING BENCHMARK-ING

ALEXANDRINA BODRUG

Context Some definition

Genscale

scaffolding tools features

The raw input data
GST modeled graph
Expected solution

Challenging problem

Scripting fo the GST

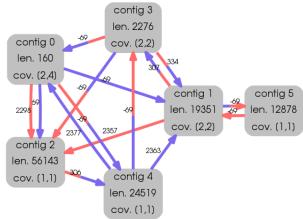
Benchmarking workflow for tge GST

Results

Example of A

Davanastivas

RAW INPUT DATA OF AGROSTIS STOLONIFERA



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

THE RAW INPUT DATA

SCAFFOLDING BENCHMARK-ING

ALEXANDRINA BODRUG

Context

Some definition Order and orien

Genscale scaffolding

The raw input data

GST modeled graph Expected solution

problem

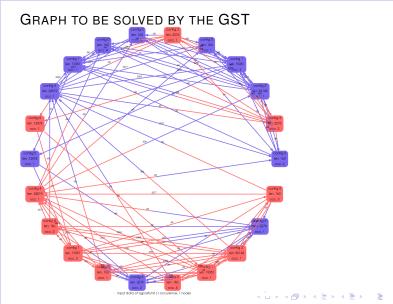
Scripting for the GST

Benchmarking workflow for tge GST

Results

Example of Agr

Doronostivos



THE RAW INPUT DATA

SCAFFOLDING BENCHMARK-ING

ALEXANDRINA BODRUG

Some definition

Genscale

scaffolding

The raw input data

Expected solution

Challenging

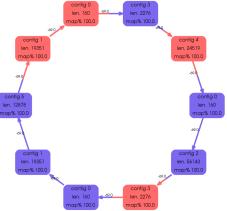
Scripting fo

Benchmarking workflow for

Results
Example of A

stolonitera

EXPECTED SOLUTION OF AGROSTIS STOLONIFERA



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

CHALLENGING PROBLEM

SCAFFOLDING BENCHMARK-ING

ALEXANDRINA BODRUG

Context Some definition

Genscale

scaffolding tools feature

The raw input data
GST modeled graph
Expected solution

Challenging

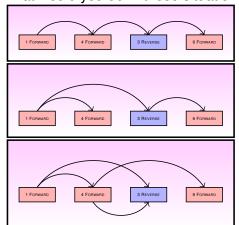
Scripting for

Benchmarking workflow for tge GST

Example of A

Stolorillera

What would you do in these situations?



SCRIPTING FOR THE GST

SCAFFOLDING BENCHMARK-ING

ALEXANDRINA BODRUG

Context
Some definitions
Order and orient

scaffolding tools features The raw input data GST modeled graph

Challenging problem

Scripting for the GST

Benchmarking workflow for

Results

Example of Agros

- 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

SCAFFOLDING BENCHMARK-ING

ALEXANDRINA BODRUG

Context Some definition

Genscale scaffolding

The raw input data
GST modeled grap

GST modeled graph Expected solution

problem

Scripting fo the GST

Benchmarking workflow for tge GST

Example of Agrost stolonifera

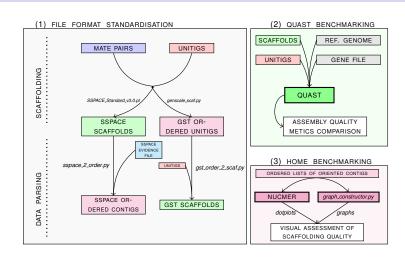


Figure: Benchmarking workflow

RESULTS

SCAFFOLDING BENCHMARK-ING

ALEXANDRINA BODRUG

Context
Some definitions
Order and orien

scaffolding tools features The raw input data GST modeled graph Expected solution

Challenging problem

Scripting for

Benchmarking workflow for tge GST

Results

Example of Agros

 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

SCAFFOLDING BENCHMARK-ING

ALEXANDRINA BODRUG

Conte

Some definition

Genscale scaffolding

scaffolding tools featur

The raw input data
GST modeled graph
Expected solution

Challengin problem

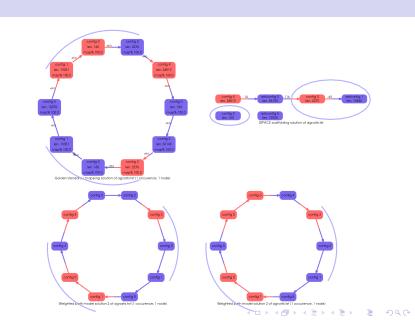
Scripting for the GST

Benchmarking workflow for tge GST

Results

Example of Agrostis stolonifera

arenactiva



PERSPECTIVES

SCAFFOLDING BENCHMARK-ING

ALEXANDRINA BODRUG

Context
Some definitions
Order and orient

Genscale scaffolding

The raw input data
GST modeled graph

GST modeled graph Expected solution

problem

Scripting for

Benchmarking workflow for

Results

Example of Agros

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

SCAFFOLDING BENCHMARK-ING

AI EXANDRINA BODRUG

The raw input data GST modeled graph

Expected solution

workflow for

Thanks!

The End