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 25, 2015

Overview

1 Context
- Some definitions
- Order and orient
- Challenging problem

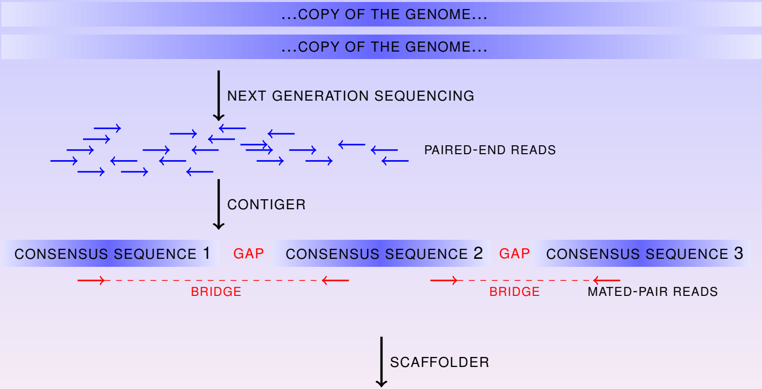
2 GST features
- Scripting for the GST
- GST modeled graph
- Expected solution
- Scaffolding solutions

3 Benchmarking workflow for the GST

4 Results
- Large repeats
- Short repeats

5 Perspectives

Context



...copy of the genome...

...copy of the genome...

next generation sequencing

paired-end reads

contiger

consensus sequence 1 gap consensus sequence 2 gap consensus sequence 3

bridge

bridge

mated-pair reads

scaffolder

In the context of high thoughput sequencing, the scaffolding process is a difficult task for which no widely accepted methodology exists.

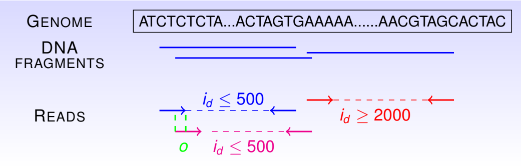
SOME DEFINITIONS

"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 (↦ reads) . . .



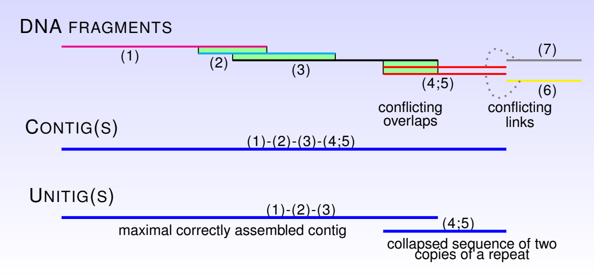
GENOME ATCTCTCTA...ACTAGTGAAAAA......AACGTAGCACTAC

DNA FRAGMENTS

READS $i_d \leq 500$ $i_d \geq 2000$ o $i_d \leq 500$

. . . reads are assembled into consensus sequences.

UNITIGS ARE HIGH CONFIDENCE CONTIGS.



DNA FRAGMENTS

(1) (2) (3) (4;5) (7) (6)

conflicting overlaps

conflicting links

CONTIG(S)

(1)-(2)-(3)-(4;5)

UNITIG(S)

(1)-(2)-(3)

maximal correctly assembled contig

(4;5)

collapsed sequence of two copies of a repeat

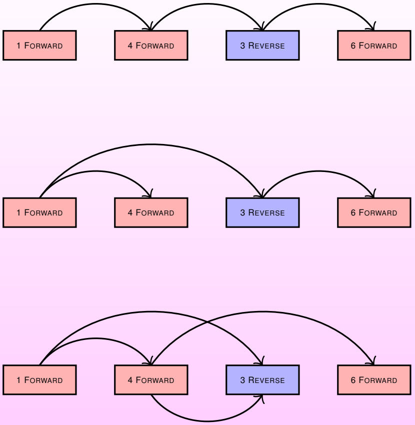
Order and orient

Mated-pair read ↦ bridge between contigs
Several correctly mapped reads ↦ link between contigs
High-confidence overlap ↦ link between contigs
All linkage information ↦ order and orient contigs

Conflicting links can and will exist.

CHALLENGING PROBLEM

What would you do in these situations?



EASY PATH

LINK COVERAGE?

HEURISTICS...

GENSCALE SCAFFOLDING TOOLS FEATURES

SEVERAL TOOLS MODEL THE PROBLEM DIFFERENTLY.

→ common features:

- modeling of the scaffolding problem as a graph
- use of unitigs instead of contigs to better compute coverage
- use of unitig coverages to duplicate nodes representing unitigs
- unitig orientations represented by separate nodes

→ differences:

- **weighted path model** focuses solely on order and orientation
- **distance based model** incorporates link length information
- **flow model** accepts intervals for unitig coverage and link length

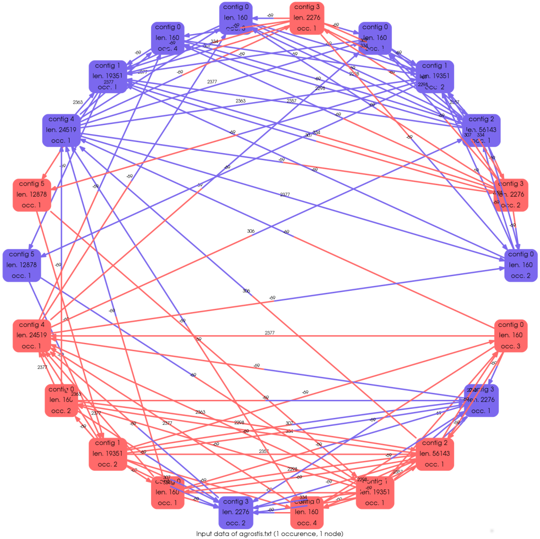
HANDLING THE MODELED GRAPHS

Automated ways to control the input data and validate the scaffolding solution:

- 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`

GST MODELED GRAPH

Agrostis stolonifera chpl. genome input data graph



contig 0 len. 160 occ. 3
contig 3 len. 2276 occ. 1
contig 0 len. 160 occ. 4
contig 0 len. 160 occ. 1
contig 1 len. 19351 occ. 1
contig 1 len. 19351 occ. 2
contig 4 len. 24519 occ. 1
contig 2 len. 56143 occ. 1
contig 5 len. 12878 occ. 1
contig 3 len. 2276 occ. 2
contig 5 len. 12878 occ. 1
contig 0 len. 160 occ. 2
contig 4 len. 24519 occ. 1
contig 0 len. 160 occ. 3
contig 0 len. 160 occ. 2
contig 3 len. 2276 occ. 1
contig 1 len. 19351 occ. 2
contig 2 len. 56143 occ. 1
contig 0 len. 160 occ. 1
contig 1 len. 19351 occ. 1
contig 3 len. 2276 occ. 2
contig 0 len. 160 occ. 4

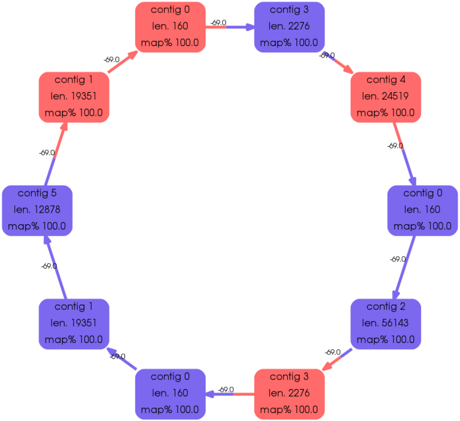
Input data of agrostis.txt (1 occurence, 1 node)

$$\sum nodes = \sum cov. \times 2$$

unitig	len	cov
1	19351	[2, 2]
0	160	[2, 4]
3	2276	[2, 2]
2	56143	[1, 1]
5	12878	[1, 1]
4	24519	[1, 1]

EXPECTED SOLUTION

Agrostis stolonifera chpl. genome expected solution



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

→USE UNITIGS THE RIGHT NUMBER OF TIMES IN THE CORRECT ORIENTATION

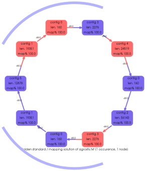
→ORDER UNITIGS TO OBTAIN AN UNINTERRUPTED CIRCULAR PATH

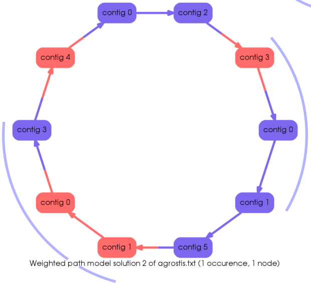
unitig	orient.	occ.
1	reverse	1
1	forward	1
3	reverse	1
3	forward	1
0	reverse	1
0	forward	2

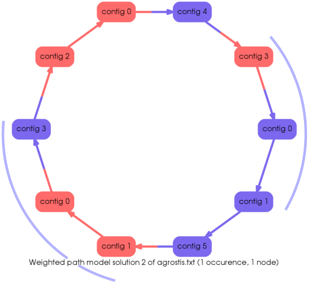
SCAFFOLDING SOLUTIONS

Agrostis stolonifera chpl. genome scaffolding solutions

UNITIGS FORMING THE INVERTED REPEATED SEQUENCE OF THE CHLOROPLASTIC GENOME ARE DUPLICATED.

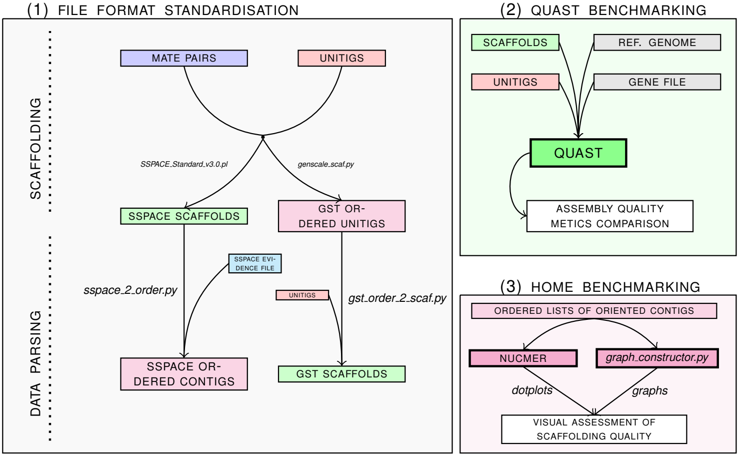

contig 0 len: 160 map% 100.0
contig 3 len: 2276 map% 100.0
contig 4 len: 24519 map% 100.0
contig 0 len: 160 map% 100.0
contig 2 len: 56143 map% 100.0
contig 3 len: 2276 map% 100.0
contig 0 len: 160 map% 100.0
contig 1 len: 19351 map% 100.0
contig 5 len: 12878 map% 100.0
contig 1 len: 19351 map% 100.0
olden standard / mapping solution of agrostis.txt (1 occurence, 1 node)





Benchmarking workflow for the GST

Solutions found with the Genscale tools are benchmarked against the SSPACE published scaffolder.



(1) File format standardisation

Scaffolding

Mate pairs → SSPACE_Standard_v3.0.pl → SSPACE scaffolds
Unitigs → genscale_scaf.py → GST ordered unitigs

Data parsing

SSPACE scaffolds + SSPACE evidence file → sspace_2_order.py → SSPACE ordered contigs
GST ordered unitigs + Unitigs → gst_order_2_scaf.py → GST scaffolds

(2) QUAST benchmarking

Scaffolds, Unitigs, Ref. genome, Gene file → QUAST → Assembly quality metics comparison

(3) Home benchmarking

Ordered lists of oriented contigs → NUCMER → dotplots → Visual assessment of scaffolding quality
Ordered lists of oriented contigs → graph_constructor.py → graphs → Visual assessment of scaffolding quality

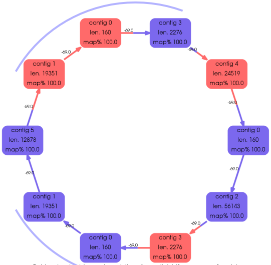
RESULTS

Using the benchmark workflow the following conclusions were drawn:

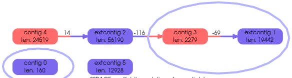
- Genomes with big repeated regions are solved a lot better with GSTs than with SSPACE
- Small repeats are very challenging to scaffold because too many conflicting links exist and GST can not take a decision or is too slow
- The GST models processing the link sequence length information perform worse than those focusing only on ordering and orientating

Large repeats - chloroplastic genomes

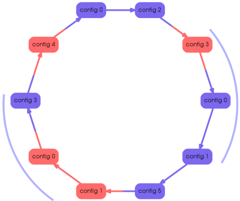
Excellent results are obtained for data sets with large repeats and a small number of unitigs.



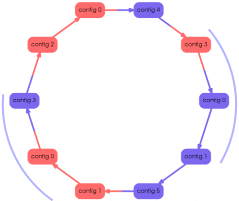
Golden standard / mapping solution of agrostis.txt (1 occurence, 1 node)



SSPACE scaffolding solution of agrostis.txt



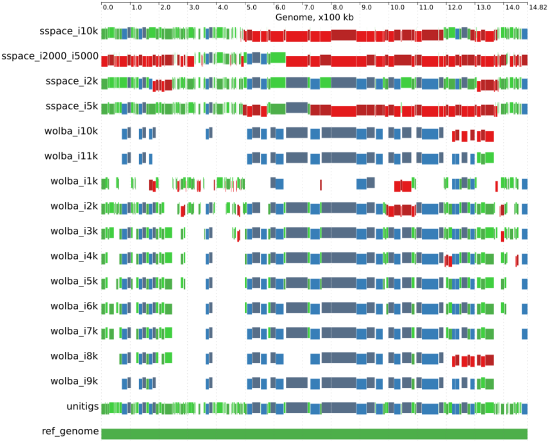
Weighted path model solution 2 of agrostis.txt (1 occurence, 1 node)



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

SHORT REPEATS - BACTERIAL GENOMES

SMALL REPEATS ARE PROBLEMATIC STARTING FROM THE UNITIG BUILDING STEP. THE WOLBACHIA ENDOSYMBIONT ORGANISM POSSESSES 444 UNITIGS AND ONLY 138 ARE LONGER THAN 1000 BASE PAIRS.



PERSPECTIVES

DEVELOP - TEST - BENCHMARK

- Find strategies which solve more challenging data
  → flow model in development
- Benchmark against other tools trying to solve repeated regions
- Test the GST with real data
- Test the GST with other genome sequencing data types

THANK YOU FOR YOU ATTENTION

SCAFFOLDING: SAFETY COMES FIRST