MASTER'S THESIS
UNIVERSITY OF RENNES 1
BIOINFORMATICS AND GENOMICS MASTER'S DEGREE
(2014 - 2015)

TEST AND BENCHMARKING OF A NEW SCAFFOLDING METHODOLOGY

INSTITUTE FOR RESEARCH IN IT AND RANDOM SYSTEMS, GENSCALE
263 AVENUE GENERAL LECLERC, 35000 RENNES, FRANCE

Author:
Alexandrina BODRUG

Supervisors:
Pr. R. ANDONOV
Pr. D. LAVENIER

22 JUNE, 2015

Thanks

Abbreviations & github link

Contents

Introduction

Backgroud

De novo whole genome shotgun assembly is the process which pieces together overlapping small fragmented DNA sequences (reads) produced by Next Generation Sequencing methods into larger sequences (contigs). Contigs highly vary in number and length depending on the sequencing method, sequencing depth and inherent genome characteristics. Contigs have to be ordered and relatively oriented, a step which is metaphorically called scaffolding. The aim is to obtain complete genomes (or chromosomes) containing gaps of known lengths because the less fragmented the genome is, the easier the downstream analysis are[1]. Because this uninterrupted genome sequence is a precious information in many research fields, there has been an important effort made to improve the performance of assembly algorithms and the quality of NGS data. The 2011 and 2013 Assemblathon projects[2][3] aimed at benchmarking existing tools against high coverage diploid genomes with a focus on the contig building step. What is commonly called an assembly tool is often only the contig building step. It can include a built-in scaffolding step although it is not always clear how to access and configure it. However building contigs and scaffolding are two very different problems that should be performed separately. Since 2004 a dozen of stand alone scaffolding tools have been developed with the first being Bambus[4], originally part of the MetAMOS[5] assembly and analysis pipeline. Despite its simply formulated goal - order and orient contigs - scaffolding is a challenging computational problem. Most tools model the problem as a graph where vertices represent contigs and links represent bundles of pairs of reads linking two contigs. Erroneous data (fake links due to poorly assembled contigs, low quality libraries), missing data (low quality libraries, unfit insert size, low genome coverage) and inherent genome characteristics (repeated regions, heterozygosity) stand in the way of a perfect and easy scaffolding process.

Project goal

In this report the results and prospects of the scaffolding tools developed by the Genscale tool at IRISA are discussed.

Datasets and input data

Features of the chloroplast genomes

Chloroplasts are small organelles in plant photosynthetic tissues which possess their own DNA. The chloroplast genomes are small ($\approx 150kpb$), circular and have a large inverted repeated sequence of around $25kpb$.

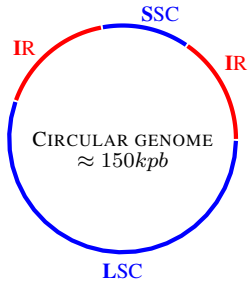

Figure 1: Chloroplast genome structure

Inverted Repeat (IR $\approx 23kpb$) ; Long Single Copy (LSC); Small Simple Copy (SSC $\approx 85kpb$)

The instances used for this study are the chloroplast genomes of the following organisms: *Eucalyptus globulus, Acorus calamus, Atropa belladonna, Agrostis stolonifera, Cucumis sativus, Lecomlella madagascariensis, Oenothera elata, Pinus koraiensis* and *Euglena gracilis*. Despite having the same general structure, the amount of repeated sequences is different from an instance to another, making them more or less easy to assemble.

Input data

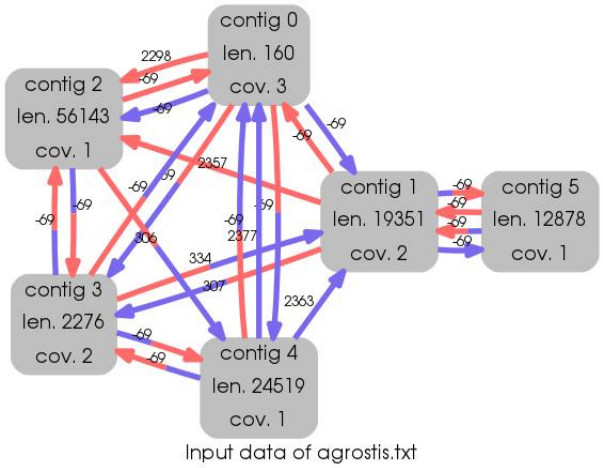

Figure 3: Input data of *Agrostis stolonifera* viewed in form of a graph

The image was generated with the *graph_generator.py* script. The color on the links indicate if the links points to the contigs in forward (blue) or reverse (red) orientation.

Material and methods

Inspecting the input data

Benchmarking

Chosen scaffolding tools to benchmark against Benchmarking workflow Comparisons QUAST Comparison function Visualization MUMMER Visualization tool

Results

Data sets compared

QUAST and comparison function

Visualization

Conclusion

References

[1] Hunt, M., Newbold, C., Berriman, M. & Otto, T. D. A comprehensive evaluation of assembly scaffolding tools **15**, R42 (2014).

[2] Earl, D. *et al.* Assemblathon 1: A competitive assessment of de novo short read assembly methods **21**, 2224–2241 (2011-12).

[3] Bradnam, K. R. *et al.* Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species **2**, 10 (2013-07-22).

[4] Pop, M., Kosack, D. S. & Salzberg, S. L. Hierarchical scaffolding with bambus **14**, 149–159 (2004-01-01).

[5] Treangen, T. J. *et al.* MetAMOS: a modular and open source metagenomic assembly and analysis pipeline **14**, R2 (2013-01-15).