Linear integer programming approach for chloroplast genome scaffolding

Victor Epain, Rumen Andonov

Univ. Rennes, Inria, IRISA, F-35000 Rennes, France

{victor.epain, rumen.andonov}@irisa.fr

Keywords : *de novo genome assembly, contigs graph, inverted repeats, nested pairs*

1 Introduction

Genome assembly aims to assemble the genome from a large amount of small DNA sequences named reads. This process can be roughly separated in two main steps. First, reads are merged into sequences named contigs. They correspond to paths in a graph that models relations between reads. For any contig, its orientation defines which DNA strand it belongs to. There are two orientations as the two DNA strands are read in reverse orientations.

This study exclusively focuses on the second stage of the genome assembly, scaffolding, in which contigs are put in correct order and orientation towards the completion of the final assembly. Usually this step uses additional data *e.g.* mate pairs' distances, or homology references from near-species. In contrast to that, we demonstrate here that chloroplasts' genomes can be assembled without any raw additional data. It is well known that chloroplasts' genomes possess highly conserved circular and quadripartite structures [2] (with a pair of dispersed inverted repeat regions, separated by two unique regions, see Figure 1) is sufficient.

2 Method

The first step outputs an assembly graph where each vertex corresponds to a contig and is provided with an estimated multiplicity number. This number corresponds to an upper bound of repeats' contig in the genome. In the sequel we use another graph where each vertex is duplicated according to its multiplicity number and to the two contig orientations. Edges modelising contigs' successions between in provided assembly graph are duplicated respectively. We model the genome assembly as an elementary circuit in this graph, see Figure (2). We formulate the dispersed repeats with linear constraints and we search for such a circuit using Integer Linear Programming similarly to [1].

Inverted repeats correspond to occurrences of contigs paired with other occurrences of them but in reverse orientation, see Figure (1). Therefore paired contigs positions on the assembled sequence must satisfy nested-pairs pattern. We formulate the above constraints in terms of linear program where the objective is to maximise the nested-pairs number. This results in finding a couple of longest contiguous inverted repeats. Thus, we generalise a similar approach applied for RNA folding [4]. Indeed, in contrast to the later approach where the vertices correspond to bases with known sequence indices, in our case the positions of the contigs are variables. Our tool is implemented with Python 3 and uses the open-source PuLP package which integrates a free solver CBC [5] to solve the above optimisation problem.

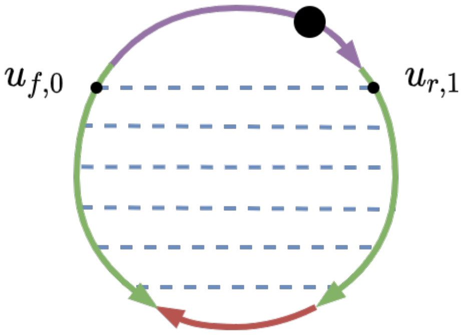

FIG. 1: An illustration of chloroplast genome structure. It is a quadripartite circuit with a pair of dispersed inverted repeat regions (in green), separated by two unique regions (in purple and red). $u_{f,0}$ is an occurence of contig u, paired with its reversed complement $u_{r,1}$ — another occurence. Blue dash lines visualise other nested pairs. The big black dot corresponds to the begining and the end of the circuit. This data is provided thanks to biological knowledge.

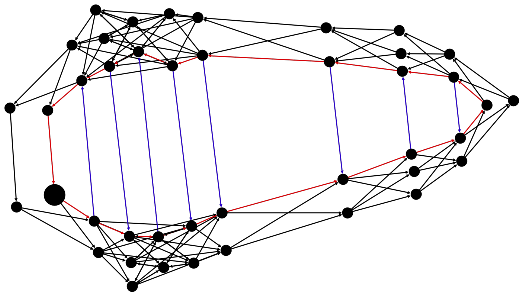

FIG. 2: The input contigs were multiplied by their multiplicity number, then doubled according to two DNA strands. The obtained graph illustrated here possesses 42 nodes and 130 edges. Contigs candidate to participate in reverse repeats have their two oriented versions linked by a blue edge. The solution path (the assembled genome) is represented in red. It begins from the biggest node (a given starter) and finishes to the same node as chloroplast genomes are circular.

3 Results

We verified our method with the well-known assembly evaluation tool QUAST [3]. We run 8 instances on a laptop (16GB RAM, 8 cores), and we obtained very encouraging preliminary results, with high genome coverage (mostly $> 99\%$), and very low mismatches and indels rates. In the case where all links between contigs are provided, our approach permits to finish pre-assembled genomes in just a few seconds, for graphs that not exceed 160 nodes and 284 edges and a dozen of candidate nested pairs.

References

[1] Rumen Andonov, Hristo Djidjev, Sebastien François, and Dominique Lavenier. Complete assembly of circular and chloroplast genomes based on global optimization. *Journal of Bioinformatics and Computational Biology*, 17(3):1950014, June 2019.

[2] Ralph Bock and Volker Knoop, editors. *Genomics of Chloroplasts and Mitochondria*, volume 35 of *Advances in Photosynthesis and Respiration*. Springer Netherlands, Dordrecht, 2012.

[3] Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, April 2013.

[4] Dan Gusfield. The RNA-Folding Problem. In Dan Gusfield, editor, *Integer Linear Programming in Computational and Systems Biology: An Entry-Level Text and Course*, pages 105–121. Cambridge University Press, Cambridge, 2019.

[5] johnjforrest, Stefan Vigerske, Haroldo Gambini Santos, Ted Ralphs, Lou Hafer, Bjarni Kristjansson, jpfasano, EdwinStraver, Miles Lubin, rlougee, jpgoncal1, h-i gassmann, and Matthew Saltzman. coin-or/Cbc: Version 2.10.5, March 2020.