Optimal de novo assemblies for chloroplast genomes based on inverted repeats patterns

Rumen Andonov, Victor Epain, Dominique Lavenier

Optimal de novo assemblies for chloroplast genomes based on inverted repeats patterns

Rumen Andonov[1*], Victor Epain[1*], Dominique Lavenier[1]

[1]Univ Rennes, Inria, CNRS, IRISA, F-35000 Rennes, France

*rumen.andonov@irisa.fr ; victor.epain@laposte.net

Background
Chloroplast genome assembly remains challenging because sequencing step outputs short reads both from plant and plastid genomes. Some recent dedicated assemblers [1,2] use the information of a highly conserved circular and quadripartite structure with a pair of dispersed inverted repeat regions in chloroplast genomes.
Materials and methods
We designed a dedicated pattern-driven *de novo* assembler which requires short unpaired reads uniquely (distances provided by paired-reads are not needed), sequenced from both the plant and its chloroplasts. A first step consists in separating the chloroplasts reads from the reads specific to plant. To this end we use the observation that the chloroplast genomes are over-represented compared to the plant genome. Then we compute an estimated coverage of the pre-assembled contigs and we keep the ones with higher coverage. The first step outputs an assembly graph where each vertex corresponds to a contig and is provided with an estimated multiplicity number. In the sequel we use another graph where each vertex is duplicated according to its multiplicity number and to the two possible contig orientations. The edges are duplicated respectively. In our approach the genome assembly is modelled as finding an elementary path in this graph. We formulate the dispersed repeats as linear constraints and we search for an elementary path using Integer Linear Programming similarly to [3]. In our approach inverted repeats correspond to occurrences of contigs paired with other occurrences of them but in reverse orientation. Their 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 maximize the nested-pairs number. Thus, we generalize 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 to solve the above optimization problem.
Results
We tested our program with QUAST [5] and we obtained very encouraging preliminary results, with high genome coverage (mostly >99%), and very low mismatches and indels rates.
Conclusions
We designed a chloroplast genome dedicated pattern-driven *de novo* assembler using only short unpaired reads. We formulate the conserved circular and quadripartite structure as linear constraints and implemented this model in an open-source program. Finally, QUAST evaluation returned some encouraging preliminary results.

Acknowledgements
The two first authors have equally participated to the presented work. The first step of the method was designed and implemented by Dominique Lavenier, the second was designed by Rumen Andonov and Victor Epain who implemented it.

References
1. Jin J-J, Yu W-B, Yang J-B, Song Y, dePamphilis CW, Yi T-S, et al. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biology. 2020 Sep 10;21(1):241.
2. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res. 2017 Feb 28;45(4):e18.
3. Andonov R, Djidjev H, François S, Lavenier D. Complete assembly of circular and chloroplast genomes based on global optimization. J Bioinform Comput Biol. 2019 Jun;17(3):1950014.
4. Gusfield D. The RNA-Folding Problem. In: Integer Linear Programming in Computational and Systems Biology: An Entry-Level Text and Course. Cambridge: Cambridge University Press; 2019. p. 105–121.
5. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013 Apr 15;29(8):1072–5.