Debugging long-read genome and metagenome assemblies using string graph analysis

Pierre Marijon, Jean-Stéphane Varré, Rayan Chikhi

Debugging long-read genome and metagenome assemblies using string graph analysis

Pierre MARIJON[1], Jean Stéphane VARRÉ[2] and Rayan CHIKHI[2]

[1] Inria, Université de Lille, CNRS, Centrale Lille, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France

[2] Univ. Lille, CNRS, Centrale Lille, Inria, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, F-59000 Lille, France

Corresponding author: pierre.marijon@inria.fr

Third-generation long-read sequencing technologies tackle the repeat problem in genome assembly by producing reads that are long enough to span most repeat instances. In principle one expects that with such reads most bacterial genomes will be assembled into a single contig [1]. However in practice, some datasets fail to be perfectly assembled even with leading assemblers, and are fragmented into a handful of contigs. As a mean to investigate those cases, we consider the string graphs that are generated by assemblers during intermediate stages of the assembly process. We seek to establish a coherent framework for analyzing these graphs, in the hope that they will help us determine the biological causes that led the assembler to output shorter contigs. This poster presents some preliminary results of such an analysis.

We visualized, analyzed and compared assembly graphs generated by *Canu* [2] and *Miniasm* assemblers [3] on biological (MBRAC-26 [4]) and synthetic datasets (created with LongISLND [5]). We introduce the concept of *graph projection* of an assembly graph onto another, taking advantage of the recent GFA format. We are thus able to observe how reads that are neighbors of contigs extremities overlap, in terms of error rate and overlap length. We implemented an automatic and user-friendly *snakemake* pipeline that generates a HTML report for each assembly. We identified cases of contigs that were not joined by the assembler despite indications in the string graph that such joins could have been made. These cases highlight potential directions on how to improve the assembly process. In future work we will take advantage of this investigation to propose alternative assembly hypotheses based on string graph analysis.

References

[1] Sergey Koren and Adam M Phillippy. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology*, 23:110–120, 2015.

[2] Sergey Koren, Brian P. Walenz, Konstantin Berlin, Jason R. Miller, Nicholas H. Bergman, and Adam M. Phillippy. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, page gr.215087.116, 2017.

[3] Heng Li. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, 32(14):2103–2110, 2016.

[4] Esther Singer, Bill Andreopoulos, Robert M. Bowers, Janey Lee, Shweta Deshpande, Jennifer Chiniquy, Doina Ciobanu, Hans-Peter Klenk, Matthew Zane, Christopher Daum, Alicia Clum, Jan-Fang Cheng, Alex Copeland, and Tanja Woyke. Next generation sequencing data of a defined microbial mock community. *Scientific Data*, 3:160081, 2016.

[5] Bayo Lau, Marghoob Mohiyuddin, John C. Mu, Li Tai Fang, Narges Bani Asadi, Carolina Dallett, and Hugo Y. K. Lam. LongISLND: in silico sequencing of lengthy and noisy datatypes. *Bioinformatics*, 32(24):3829–3832, 2016.