Assembly strategies for the recovery of Metagenome-Assembled Genomes

Benjamin Churcheward[1], Guillaume Fertin[1], and Samuel Chaffron[1,2]

[1]Nantes Université, CNRS UMR 6004, LS2N, F-44000, France
[2]Research Federation (FR2022) Tara Oceans GO-SEE, Paris, France

Abstract

Since the late 2000s, development of next-generation sequencing allowed to facilitate the reconstruction of genomes for an increasing number of organisms. However, due to the limitation of cultivation techniques to isolate and sequence microbial genomes, only a little part of microbial diversity has been accessible so far. The rise of metagenomics is helping to access a wider part of bacterial diversity, to better determine and understand microbial communities structure, and also enables the reconstruction of genomes from metagenomics data, or Metagenome-Assembled Genomes (MAGs). The reconstruction of MAGs implies two steps, first the assembly of the metagenomic samples, and second the binning, which consists in grouping contigs more likely to belong to a common genome, based on their compositional features and differential coverage between several metagenomics samples. While an increasing number of samples is thought to enhance quality of reconstructed MAGs, resources consumption also increases with the number of samples. Thus, a relevant grouping of samples into co-samples sets has to be performed. Using TARA oceans datasets, recent studies have combined samples using geographic distances [1, 2] nevertheless no consensus exists on how to co-assemble samples [1, 2, 3].

In this work, we explored an approach relying on ecological distances between marine metagenomic samples sequence similarities to group our samples in co-samples sets. These distances were computed using sequence similarities between the different samples, using Simka[4]. Optimal clustering approaches found several solutions that satisfy both needs to group ecologically related samples, and to limit number of samples per co-samples sets. We then tested three different assembly strategies, on three toy co-samples sets we gathered using distances previously computed. Assembly strategies has an influence on the number and the quality of the MAGs. While a coassembly approach using several samples allows to reconstruct more MAGs than single-sample assembly, quality of MAGs can decrease with the increasing distances between coassembled samples. An approach combining both single-assembly and coassembly approaches can offset their disadvantages.

References

[1] Tom O. Delmont, Christopher Quince, Alon Shaiber, Özcan C. Esen, Sonny TM Lee, Michael S. Rappé, Sandra L. McLellan, Sebastian Lücker, and A. Murat Eren. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology*, 3(7):804–813, July 2018.

[2] Benjamin J. Tully, Elaina D. Graham, and John F. Heidelberg. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific data*, 5:170203, 2018.

[3] Javier Tamames and Fernando Puente-Sanchez. Squeezem, a highly portable, fully automatic metagenomic analysis pipeline. *Frontiers in microbiology*, 9:3349, 2018.

[4] Gaëtan Benoit, Pierre Peterlongo, Mahendra Mariadassou, Erwan Drezen, Sophie Schbath, Dominique Lavenier, and Claire Lemaitre. Multiple comparative metagenomics using multiset k -mer counting. *PeerJ Computer Science*, 2:e94, November 2016.