Computational Pan-Genomics

Eric Rivals - LIRMM, CNRS, Montpellier

July 1st 2019

Abstract

The notion of pan-genome or genome graph, understood as the representation of all variant genomes of one or more closely related species, is gaining attractiveness to replace the classical, sequence-based, linear reference genome. Indeed, one of main reasons why most analyses are based on the linear reference genome is simplicity: simplicity of processing – for most bioinformatic algorithms target a single linear sequence as reference – and simplicity of understanding – our minds prefer linear structures. Due evolution, related sequences may share common or highly similar regions, which should not be considered as independent, but are currently considered so, for instance when one blasts a query sequence against a collections of genomes.

A sequence graph representation summarises in a single reference the similarities and differences of the multiple collected variants. For instance, read mapping on a sequence graph directly tells in which variants a read maps and in which it does not, thereby revealing variant specificity (if any). Even inside the class of graph based representations, several options are available for the data structure to use. With increasing sequencing throughput, the number of sequences to represent grows and efficiency becomes an issue. Practical uses of sequence graphs and pan-genomes depends heavily of the underlying indexing data structures, especially of their memory footprint. I will present an overview of computational issues related to such sequence graph based approaches to pan-genome representation, and illustrates their complexity and advantages (which reach far beyond efficiency aspects).

Pointers and references

1. Computational pan-genomics: status, promises and challenges
   The Computational Pan-Genomics Consortium
   Briefings in Bioinformatics, Vol. 19(1), p. 118–135, 2018.
   https://doi.org/10.1093/bib/bbw089

2. Rebooting the Human Genome
   MIT tech review 2015
   https://www.technologyreview.com/s/537916/rebooting-the-human-genome/