BITS :: Call for Abstracts 2022 - Oral communication

Type	Oral communication
Session	Pangenomics
Title	A systematic evaluation of computational tools for gene-oriented pangenome detection in fragmented genomes.
All Authors	Bonnici V(1), Mangoni M(2), Franco G(2) and Giugno R(2)

Affiliation

(1)Department of Mathematical, Physical and Computer Sciences, University of Parma, Italy
(2)Department of Computer Science, University of Verona, Italy.

Motivation

Pangenomics was originally defined as the problem of comparing the composition of gene sets within a set of bacterial isolates belonging to the same species [1]. The problem requires the calculation of genetic sequence homology among such genes. Pangenomics has gained an increasing interest by the scientific community, because the identification of strain-specific genes, made possible via pangenomic analyses, provides a methodology to discover bacterial biomarkers and develop targeted therapeutics vaccines [2]. When combined with metagenomics, namely for human microbiome composition analysis, gene-oriented pangenome detection becomes a promising method to decipher ecosystem functions and population-level evolution [3].

Current computational tools are able to investigate the genetic content of isolates for which a complete genomic sequence is available [4]. A comparison of computational tools on fragmented genomes involves the definition of statistical measurements to assess the performance [5]. To this aim, synthetic benchmarks are built by means of computational methodologies that, starting from one single root genome, simulate evolution to create a synthetic progeny of the root genome. PanProva [6] is a recent tool that, differently from previous approaches, embeds horizontal gene transfer (HGT) of bacterial populations into the evolutionary process. It produces synthetic bacterial populations that better reflect pangenomics distributions.

Beside analysis of complete genomes, there is a plethora of incomplete genomes that are available on public resources. Incomplete means that the process for reconstructing their genomic sequence is not complete, and only fragments of the partial DNA sequence are currently available. However, the information contained into these fragments may play an essential role in current and future pangenomic analyses. Namely, the possibility of managing incomplete information may turn out to be useful in tempestive and chip responses to bacterial epidemics. GenAPI [7] is a tool for extracting pangeomic content from incomplete genomes, mainly based on the comparison between the gene sequences in their incomplete form. The tool can only be applied to analyze isolates that are highly related to each other phylogenetically, such as isolates belonging to the same bacterial strain. Pan4Draft [8] is a more sophisticated tool that, given an incomplete gene, runs a query over an online database in order to recognize the known gene that is most similar to the incomplete gene. However, it is not checked that all the extracted genes belong to the input fragments and are not artifacts introduced by the reference.

Methods

A recently developed methodology called PanDelos-frags deals with incomplete genomes, is based on one of the better performing tools for complete sequences (PanDelos [9]), and overcomes the issues of GenAPI and Pan4Draft. In detail, given an incomplete genomes, in terms of a set of fragments, the tool recognizes the complete genome that is mostly similar to it within a given database. The default database is the entire collection of bacterial reference genomes available via NCBI. Then, the tool aligns the fragment to the retrieved reference genome and runs a gene detection algorithm to extract genetic sequences. In this way, the portion of sequence that is reconstructed from the reference genome is ensured to be the most probable according to the current knowledge. Moreover, differently from GenAPI and Pan4Draft, the tool is able to recognize genes for which essential parts were missing in the fragments, such as those needed for a gene detection algorithm to recognize the presence of a gene.

Results

In this work we have developed a pipeline to systematically compare pangeomic tools able to manage fragmented genomes, which requires an ad hoc solution for measuring their performance. In particular, we used PANPROVA to generate synthetic complete genomes, from which complete gene sequences were extracted. Then, we simulated fragmentation by discarding a given percentage of genomic sequence. Fragments were used as input of computational tools, and the output genes retrieved by the tools were mapped back to the original genetic sequences. Statistical measures were calculated for homology relations. Namely, we evaluated if two phylogenetically related genes were captured by the tool as homologues. We applied the developed pipeline to compare currently available tools by statistical measures. Preliminary results confirm that, beside the recognition of a higher number of genes, PanDelos-frags outperforms the competitors in computing the homology among all the genetic sequences, by reaching an average f1score of 0.8.

Info

[1] Medini, Duccio, et al. Current opinion in genetics & development 15.6 (2005): 589-594.
[2] Anani, Hussein, et al. Microbial pathogenesis 149 (2020): 104275.
[3] Zhong, Chaofang, et al. Computational and Structural Biotechnology Journal 19 (2021): 1458-1466.
[4] Contreras-Moreira, Bruno, et al. Applied and environmental microbiology 79.24 (2013): 7696-7701.
[5] Bonnici, Vincenzo, et al. Briefings in Bioinformatics (2020).
[6] Bonnici, Vincenzo, and Giugno, Rosalba. Bioinformatics 38.9 (2022): 2631–2632,
[7] Gabrielaite, Migle, et al. BMC bioinformatics 21.1 (2020): 1-8.
[8] Veras, Allan, et al. Scientific reports 8.1 (2018): 1-8.
[9] Bonnici, Vincenzo, et al. BMC bioinformatics 19.15 (2018): 47-59.

filename	-
Figure	
-	
Availability	-

Corresponding Author

Name, Surname	Vincenzo, Bonnici
Email	vincenzo.bonnici@unipr.it
Submitted on	11.05.2022