BITS :: Call for Abstracts 2021 - Oral communication

Type	Oral communication
Session	Algorithms for Bioinformatics
Title	A methodology for capturing pangenomic content among incomplete genomes
All Authors	Bonnici V(1), Motterle G(1), Franco G(1), Giugno R(1)

Affiliation

(1) Department of Computer Science, University of Verona, Italy.

Motivation

In the last two decades the field of pangenomics has acquired a notably increased interest by the scientific community. It was firstly defined as the problem of understanding the genetic composition of a set of bacterial isolates belonging to the same species [1]. A significant advancement is reached by computing genetic homology. It is shown that the identification of widely shared genes or, in contrast, strain specific genetic material can bring advances in several biological problems. For example, such an identification can provide a methodology for discovery of bacterial biomarkers or for developing targeted therapeutics and vaccines [2].

Several computational approaches have been proposed, and recent studies have compared their performance [3]. Among the benchmarked tools, PanDelos [4] and GET_HOMOLOGUES [5] have shown a better ability in computing genetic homology and thus in recognizing gene presence among the analyzed genomes. Unfortunately, most of the available approaches can only work under the condition that a complete genetic information is extracted from analysed genomes. Completeness means that the entire nucleotide sequence of each examined gene is known. On the other hand, there are a plethora of incomplete genomes that are available on public resources. They are in the form of draft sequences, due to the complexity in assembling whole genomes, or because of the lack of economic resources. The opportunity to manage incomplete information may turn out to be useful in tempestive and chip responses to bacterial epidemics. At the state of the art, there are only two methodologies able to deal with incomplete genomes, GenAPI [6] and Pan4Draft [7]. GenAPI is mainly based on directly comparing genetic sequences in their incomplete form. As a result, it can only be applied to analyse isolates belonging to the same species, and it may produce unexpected results in less related genomes. Pan4Draft performs an online query for identifying the known gene whose sequence is the most similar to the incomplete one. The main issue in such an approach is that genes belonging to the same incomplete genome are mapped to known genes of different genomes.

Methods

PanDelos has been originally developed for analysing genomes with highly heterogeneous similarity among them. However, its good performance has been shown also for closely related genomes. Here, we propose an adaptation of the PanDelos approach in order to deal with incomplete genomes. The adaptation is aimed at performing intra- and inter-species pangenomic analyses. A preliminary step for partially reconstructing the missing information has been developed. The reconstruction is based on the recognition of the complete genome that is the most similar to the incomplete one. The complete genome can be provided by the user, or it can be automatically selected among those that are publicly available. A measure for comparing the similarity among the set of incomplete fragments of a given isolate and a complete genome is proposed. Fragments are split into pieces at most 10Kb long. Each piece is blasted against a reference database of genomes, and top hits are recorded. The genome which receives the higher number of best hits is accounted as the reference genome. Then, the missing information is obtained after aligning the fragments of the incomplete genome to the complete one. Moreover, the original homology computation is refined in order to exploit the surrogate information obtained by the preliminary step. Similarity between genes is weighted by the percentage of the sequence that has been reconstructed. In this way, genes with a lower reconstructed percentage produce more confident similarities.

Results

The proposed methodology has been compared with GenAPI and Pan4Draft on intra- and inter-species benchmarks. The benchmarks were built by artificially removing part of the genetic information. Tool performances were evaluated by measuring their capability in reconstructing the original pangenomic content. In both situations, the proposed approach has outperformed GenAPI and Pan4Draft in reconstructing the correct pangenomic information.

Info

References
[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] Bonnici, Vincenzo, et al. Briefings in Bioinformatics (2020).

[4] Bonnici, Vincenzo, et al. BMC bioinformatics 19.15 (2018): 47-59.

[5] Contreras-Moreira, Bruno, et al. Applied and environmental microbiology 79.24 (2013): 7696-7701.

[6] Gabrielaite, Migle, et al. BMC bioinformatics 21.1 (2020): 1-8.

[7] Veras, Allan, et al. Scientific reports 8.1 (2018): 1-8.

Figure

-

Availability	-

Corresponding Author

Name, Surname	Vincenzo, Bonnici
Email	vincenzo.bonnici@univr.it
Submitted on	29.04.2021

Società Italiana di Bioinformatica
C.F. / P.IVA 97319460586
E-mail bits@bioinformatics.it

Sede legale Viale G. Mazzini, 114/B - 00195 Roma
Website bioinformatics.it