COPLA, a taxonomic classifier of plasmids

Santiago Redondo-Salvo[1], Roger Bartomeus[1], Luis Vielva[2], Kaitlin A. Tagg[3,4], Hattie E. Webb[3,4], Raúl Fernández-López[1], and Fernando de la Cruz[1,✉]

[1]Instituto de Biomedicina y Biotecnología de Cantabria (IBBTEC), Universidad de Cantabria-CSIC, C/Albert Einstein 22, 39011 Santander, Spain
[2]Departmento de Ingeniería de las Comunicaciones, Universidad de Cantabria, Santander, Spain
[3]Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta, USA
[4]WDS, Inc., 1600 Clifton Road, Atlanta, USA

The Plasmid Taxonomic Unit (PTU) concept is an initial step for a natural classification of plasmids. Here we present COPLA, a software for plasmid assignation to existing PTUs. To assess its performance, we used a sample of 1,000 plasmids missing from its current database. Overall, 41% of samples could be assigned an existing PTU (63% within the most abundant order, *Enterobacterales*), while 4% of samples could help to define new PTUs once COPLA database was updated.

Correspondence: ***delacruz@unican.es***

Introduction

Plasmids are important elements in the dissemination of genes such as antibiotic resistance determinants. The short-range evolution of bacterial genomes (as in epidemiological outbreaks) occurs more often by acquisition of mobile genetic elements carrying, for example, resistance determinants, than by point mutations creating new alleles with a selective advantage (1). Recently we reported a procedure for the taxonomic classification of plasmids (2), and we defined PTUs (plasmid taxonomic units) as the equivalent to plasmid species. A simple and rapid method for automatic assignation of plasmids to PTUs will help in the definition of the plasmid species responsible for outbreaks of antibiotic resistant strains (3).

Methods

Plasmid clustering algorithm (sHSBM). The PTU definition algorithm was based on a DNA homology network constructed from the pairwise comparison of all RefSeq84 plasmids (NCBI, dataset from September 2017), using Average Nucleotide Identity (ANI), as detailed in (2). An edge was drawn between two plasmid nodes if homology was detected over 50% of the shorter genome length. The graph-tool library was used to apply Hierarchical Stochastic Block Modeling (HSBM) to this network, with plasmids clustering based on the statistical significance of the graph topological information (4). HSBM clusters were further divided into components as, conceptually, PTU members should present DNA homology. Only clusters with at least four members were considered representative enough to be included in the next step. A home-made algorithm was added in (2) to join the previously divided clusters into a single PTU if two biological specifications were met:

(1) Size compatibility: the difference on median size of both clusters is less than half of the larger median size.

(2) Intercluster density: two HSBM clusters are joined if the number of edges between them is >50% of the maximum number of edges between both clusters, adjusted for their relative density.

As an example, 76% of the *Enterobacterales* plasmids in RefSeq84 could be assigned a PTU. More details are given in Suppl. Table 1.

PTU prediction algorithm (COPLA). The algorithm for predicting the PTU of new plasmids is based on the ANI homology network, processed by HSBM, and tuned as explained above. First, ANI between the query and database plasmids is calculated. Draft plasmid genomes are supported by concatenating their contigs. Next, a search for the plasmid relaxases is performed using MOBscan (5) to provide the plasmid MOB class. If not provided by the user, amino acid sequences of plasmid CDSs are predicted with Prodigal (6). Mating pair formation (MPF) and plasmid replication typing is performed using CONJscan (7) and PlasmidFinder (8). Antimicrobial resistance (AMR) genes are identified with a blastn search (>80% identity, <1e-20 e-value) against the CARD database (9).

After data input, the query plasmid is inserted into the sHSBM plasmid network described above. Next, 1,000 iterations of a multilevel Monte Carlo algorithm are performed to get a new graph partitioning that improves the Minimum Description Length (MDL) of the graph (4). The Description Length is the amount of information required to describe a graph, based on its partitioning into different clusters. It is calculated by turning the probability distribution of the different graph's partitions into entropies. This search is constrained by impeding the creation or deletion of new partitions to ensure that the query is assigned to the already described PTUs. However, as MDL optimization is not restricted to the query but is a global process and, in addition, query introduction will change the underlying network's topology, partition changes may be larger than expected. Finally, the HSBM output is transformed into PTUs using the same requirements described above. The query is assigned the most frequent PTU label of the members of the updated cluster. The PTU assignation is scored based on the partition overlap (10) between all database plasmids with an annotated PTU appearing in the updated query's cluster, which indicates how much the clustering has changed due to inclusion of the query.

The COPLA output provides several files, the result of ANI comparisons, a list of plasmids related to the query and, if

Table 1. Benchmark for 1,000 new plasmids of RefSeq200 dataset for the most abundant bacterial orders

Outcome	*Enterobact.*	*Lactobac.*	*Bacillales*	All samples
PTU assigned [Score]	259 (63%) [0.98 ± 0.06]	40 (46%) [0.94 ± 0.11]	25 (30%) [0.96 ± 0.1]	408 (41%) [0.97 ± 0.08]
New PTU [Score]	19 (5%) [0.84 ± 0.22]	6 (7%) [0.89 ± 0.17]	2 (2%) [0.75 ± 0.35]	41 (4%) [0.83 ± 0.24]
Not assigned [Score]	131 (32%) [0.99 ± 0.05]	41 (47%) [1 ± 0.01]	55 (67%) [1 ± 0.0]	551 (55%) [1 ± 0.05]

Number of cases (and percentage) for each prediction outcome. Mean and standard deviation of the prediction scores for each outcome class are additionally provided (in square brackets). More detailed results in Suppl. Table 3.

not provided by the user, a file with the plasmid ORFome computed by Prodigal. In addition, it provides the user with information about the plasmid relaxase, MPF type, replication formula, AMR genes and predicted PTU host-range.

Results

We tested COPLA performance with 1,000 randomly chosen plasmid sequences uploaded to NCBI since RefSeq84 was released (see Table 1). For this we downloaded RefSeq200 release (23,309 sequences) and removed those from RefSeq84 release. Sequences with NG accession numbers were further removed as these are genomic regions, resulting in a dataset of 12,561 plasmids (Suppl. Table 2). Three possible outcomes may occur (more details in the GitHub README.md file):

(1) A plasmid sequence can be assigned to an existing PTU. This happened globally in 408/1,000 cases (41%), or 259/409 (63%) for the most abundant host order (*Enterobacterales*). Output is expressed as the PTU to which the plasmid belongs, plus the PTU prediction score.

(2) A plasmid sequence clusters into a group of plasmids with no previously assigned PTU. This is an interesting situation as this plasmid could be part of a potentially new PTU. Output gives the plasmids related to the query. This happened in 41 cases (4%) – or 19 cases (5%) for the *Enterobacterales*.

(3) A plasmid sequence clusters with less than 4 plasmids and, so, it is below the algorithm's minimum threshold for PTU definition.

Availability and Implementation

Source code is freely available at https://github.com/santirdnd/COPLA under the GPL v3.0 license. An online service is available at https://castillo.dicom.unican.es/copla.

Supplementary information

Suppl_Table_1.xlsx: Metadata of RefSeq84 plasmids used for sHSBM clustering and PTUs definition.
Suppl_Table_2.xlsx: Metadata of RefSeq200 plasmids used for COPLA performance evaluation.
Suppl_Table_3.xlsx: Detailed results for the PTU assignation of the samples.

Acknowledgements

Funding

This work was supported by the Spanish Ministerio de Ciencia e Innovación [BFU2017-86378-P to FdlC, DI-17-09164 to SR-S]; and USA Centers for Disease Control and Prevention [200-2019-06679 to FdlC].

Conflict of Interest: none declared.

Bibliography

1. Marie Touchon, Amandine Perrin, Jorge André Moura de Sousa, Belinda Vangchhia, Samantha Burn, Claire L. O'Brien, Erick Denamur, David Gordon, and Eduardo PC Rocha. Phylogenetic background and habitat drive the genetic diversification of Escherichia coli. *PLOS Genetics*, 16(6):e1008866, June 2020. ISSN 1553-7404. doi: 10.1371/journal.pgen.1008866. Publisher: Public Library of Science.
2. Santiago Redondo-Salvo, Raúl Fernández-López, Raúl Ruiz, Luis Vielva, María de Toro, Eduardo P. C. Rocha, M. Pilar Garcillán-Barcia, and Fernando de la Cruz. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids. *Nature Communications*, 11(1):3602, July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17278-2. Number: 1 Publisher: Nature Publishing Group.
3. Elizabeth A. Miller, Ehud Elnekave, Cristian Flores-Figueroa, Abigail Johnson, Ashley Kearney, Jeannette Munoz-Aguayo, Kaitlin A. Tagg, Lorelee Tschetter, Bonnie P. Weber, Celine A. Nadon, Dave Boxrud, Randall S. Singer, Jason P. Folster, and Timothy J. Johnson. Emergence of a Novel Salmonella enterica Serotype Reading Clonal Group Is Linked to Its Expansion in Commercial Turkey Production, Resulting in Unanticipated Human Illness in North America. *mSphere*, 5(2), April 2020. ISSN 2379-5042. doi: 10.1128/mSphere.00056-20. Publisher: American Society for Microbiology Journals Section: Research Article.
4. Tiago P. Peixoto. Bayesian Stochastic Blockmodeling. In *Advances in Network Clustering and Blockmodeling*, pages 289–332. John Wiley & Sons, Ltd, 2019. ISBN 978-1-119-48329-8. doi: 10.1002/9781119483298.ch11. Section: 11 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119483298.ch11.
5. M. Pilar Garcillán-Barcia, Santiago Redondo-Salvo, Luis Vielva, and Fernando de la Cruz. MOBscan: Automated Annotation of MOB Relaxases. In Fernando de la Cruz, editor, *Horizontal Gene Transfer: Methods and Protocols*, Methods in Molecular Biology, pages 295–308. Springer US, New York, NY, 2020. ISBN 978-1-4939-9877-7. doi: 10.1007/978-1-4939-9877-7_21.
6. Doug Hyatt, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119, March 2010. ISSN 1471-2105. doi: 10.1186/1471-2105-11-119.
7. Sophie S. Abby, Jean Cury, Julien Guglielmini, Bertrand Néron, Marie Touchon, and Eduardo P. C. Rocha. Identification of protein secretion systems in bacterial genomes. *Scientific Reports*, 6:srep23080, March 2016. ISSN 2045-2322. doi: 10.1038/srep23080.
8. Alessandra Carattoli, Ea Zankari, Aurora García-Fernández, Mette Voldby Larsen, Ole Lund, Laura Villa, Frank Møller Aarestrup, and Henrik Hasman. In Silico Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. *Antimicrobial Agents and Chemotherapy*, 58(7):3895–3903, January 2014. ISSN 0066-4804, 1098-6596. doi: 10.1128/AAC.02412-14.
9. Brian P. Alcock, Amogelang R. Raphenya, Tammy T. Y. Lau, Kara K. Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, Anna-Lisa V. Nguyen, Annie A. Cheng, Sihan Liu, Sally Y. Min, Anatoly Miroshnichenko, Hiu-Ki Tran, Rafik E. Werfalli, Jalees A. Nasir, Martins Oloni, David J. Speicher, Alexandra Florescu, Bhavya Singh, Mateusz Faltyn, Anastasia Hernandez-Koutoucheva, Arjun N. Sharma, Emily Bordeleau, Andrew C. Pawlowski, Haley L. Zubyk, Damion Dooley, Emma Griffiths, Finlay Maguire, Geoff L. Winsor, Robert G. Beiko, Fiona S. L. Brinkman, William W. L. Hsiao, Gary V. Domselaar, and Andrew G. McArthur. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 48(D1):D517–D525, January 2020. ISSN 0305-1048. doi: 10.1093/nar/gkz935. Publisher: Oxford Academic.
10. Tiago P. Peixoto. Revealing consensus and dissensus between network partitions. *arXiv:2005.13977 [physics, stat]*, June 2020. arXiv: 2005.13977.