doi 10.18699/bgrs2024-06-36

A unified data preprocessing framework to address inconsistencies in comparative plastid genome studies

Emirsaliev A.[1, 2, 3]*, Salina E.[1, 2], Mitrofanova I.[3], Afonnikov D. [1, 2]

[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*

[2] *Kurchatov Genomic Center of the Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia;*

[3] *N.V. Tsitsin Main Botanical Garden, RAS, Moscow, Russia*

* *emirsaleh@bionet.nsc.ru*

Key words: plastome; chloroplast genome; data standardization; data harmonization; normalization; comparative genomics; phylogenetics; phylogenomics

Motivation and Aim: Plastid genomes (plastome) are the most underused data sources for phylogenetic analysis of plants. There are also a lot of widely used universal barcodes for plant genotyping which are derived from plastid genome data. With the rise of high throughput sequencing technologies ever increasing number of the complete plastid genome sequences greatly expanded the amount of available data to be analyzed. After the first complete plastome sequence was published [1] thousands of complete organellar genomes of plants were obtained by a global research community. In the long historical way of new plastid genomes assembling, describing and analyzing, diverse approaches, tools and standards were made use of at different time periods. As a result, the corpus of data obtained may be inconsistent due to different methods used in data processing. Such a differences affect the feature names, sequences and annotations of the records deposited in databases. The analysis tools used in comparative genome studies expect the data to be comparable, which means for the sequences to be of proper linear representation, and for features to be annotated and named according to standards. The sequence splits, as well as the orientation and order of large substrings also referenced as genome regions might affect the alignment scores. Inconsistent naming greatly impacts the records to be retrieved from databases and annotations to be involved in analysis. Some analyses fail when features are not properly named and ordered. We argue for benefits of *a priori* plastome data standardization to eliminate shortcomings related to the origin of data being analyzed. The aim of this work is to present a unified data preprocessing framework that standardizes plastid genome data, enabling more accurate and meaningful comparative analyses, and advancing our understanding of plant genomics and phylogeny.

Methods and Algorithms: The proposed framework handles common sources of discrepancies in plastid genome data through the following stages:

1. Data querying and retrieval from Nucleotide Database of NCBI/EMBL-EBI/DDBJ via Entrez API.
2. Records deduplication.
3. Assessing the coverage of annotation.
4. Validating the linearized representation of plastome sequences.
5. Defining the records to be kept.
6. Properly ordering and orienting genome regions.
7. Filling missing but required fields in the annotation.
8. Extracting data for analysis based on the purpose.

The framework is implemented in an interactive Jupyter Notebook environment, orchestrating Biopython, Pandas, and external tools like cpDNA_standardization [2], GeSeq [3], and NOVOWrap [4] for validation and annotation. One can easily configure each step if needed by modifying particular cell in the notebook.

Results: Here we describe a framework addressing some data quality issues, enabling more accurate and meaningful comparative analyses of plastid genomes, potentially yielding novel insights and advancing our understanding of plant genomics and phylogeny. The steps formalized in the framework allows to prepare plastome data for comparative genome studies, addressing some common issues and providing control at stages that might affect the analysis quality, like deduplication, fixing naming issues and sequences regions shifting, reordering and reorienting. The framework handles issues occuring systematically in the plastome data that were empirically identified while analyzing sample plastome data of the Cichorieae tribe (Asteraceae family, Eudicots). The framework documentation and code, as well as sample data, are available at github.com/asan-emirsaleh/plastome-preprocessing.

Conclusion: The proposed unified data preprocessing framework aims to eliminate shortcomings related to the use of diverse tools for data preparation and annotation before depositing in databases, by standardizing and ensuring data comparability. With addressing inconsistencies in data processing, the framework facilitates reliable and reproducible comparative genomic studies. Furthermore, this approach can potentially be extended to other small genomes of complex structure which fit the circular model, like mitochondrion genomes. The framework could also be integrated in plastome data analysis pipelines.

Funding: The study is supported by the budget project No. FWNR-2022-0017.

References

1. Shinozaki K., Ohme M., Tanaka M., Wakasugi T., Hayashida N., Matsubayashi T., Zaita N., Chunwongse J., Obokata J., Yamaguchi-Shinozaki K., Ohto C., Torazawa K., Meng B.Y., Sugita M., Deno H., Kamogashira T., Yamada K., Kusuda J., Takaiwa F., Kato A., Tohdoh N., Shimada H., Sugiura M. The complete nucleotide sequence of the tobacco chloroplast genome: Its gene organization and expression. *EMBO J*. 1986;5(9):2043-2049. doi 10.1002/j.1460-2075.1986.tb04464.x
2. Turudić A., Liber Z., Grdiša M., Jakše J., Varga F., Šatović Z. Towards the Well-Tempered Chloroplast DNA Sequences. *Plants*. 2021;10(7):1360. doi 10.3390/plants10071360
3. Tillich M., Lehwark P., Pellizzer T., Ulbricht-Jones E.S., Fischer A., Bock R., Greiner S. GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res*. 2017;45(W1):W6-W11. doi 10.1093/nar/gkx391
4. Wu P., Xu C., Chen H., Yang J., Zhang X., Zhou S. NOVOWrap: An automated solution for plastid genome assembly and structure standardization. *Mol. Ecol. Resour*. 2021;21(6):2177-2186. doi 10.1111/1755-0998.13410