Bioinformatics, 34(21), 2018, 3753–3754
doi: 10.1093/bioinformatics/bty454
Advance Access Publication Date: 5 June 2018
Applications Note

Sequence analysis

MIDORI server: a webserver for taxonomic assignment of unknown metazoan mitochondrial-encoded sequences using a curated database

Matthieu Leray[1], Shian-Lei Ho[2], I-Jeng Lin[2] and Ryuji J. Machida[2,*]

[1]Smithsonian Tropical Research Institute, Smithsonian Institution, Panama City, Republic of Panama and [2]Biodiversity Research Centre, Academia Sinica, Taipei 115-29, Taiwan

*To whom correspondence should be addressed.
Associate Editor: John Hancock

Received on January 20, 2018; revised on May 28, 2018; editorial decision on May 30, 2018; accepted on June 1, 2018

Abstract

Summary: We present MIDORI server, a user-friendly web platform that uses a curated reference dataset, MIDORI, for high throughput taxonomic classification of unknown metazoan mitochondrial-encoded gene sequences. Currently three methods of taxonomic assignments: RDP Classifier, SPINGO and SINTAX, are implemented.
Availability and implementation: The web server is freely available at {http://reference-midori.info/server.php}.
Contact: ryujimachida@gmail.com

1 Introduction

The era of massive sequencing has transformed our ability to study Earth's bio-diversity (Taberlet *et al.*, 2012). DNA extracted from various environments (i.e. soil, water, air, food products) can be analyzed and compared to public databases of annotated reference sequences to determine the presence of microbial, plant and animal taxa. The possible applications are extremely diverse. PCR-based (i.e. metagenetics or metabarcoding) and PCR-free (i.e. metatranscriptomics or metagenomics) DNA sequencing approaches can be used to study diversity patterns of overlooked microscopic taxa (Al-Rshaidat *et al.*, 2016), study the response of biological communities to environmental changes (Ji *et al.*, 2013), investigate illegal trade of endangered wildlife (Arulandhu *et al.*, 2017) and detect species mislabelling in food products (Raclariu *et al.*, 2017). The robustness of these technics, however, largely depends on our ability to rapidly and reliably assign taxonomy to sequences recovered from the environment.

The realization by the scientific community that public repositories of genetic data (e.g. GenBank) contained a significant number of taxonomically mislabelled sequences has promoted the creation of curated databases with higher quality standards. For example, reference datasets were built for nuclear-encoded ribosomal RNA genes [e.g. PR^2 (Guillou *et al.*, 2012), Silva (Quast *et al.*, 2013)]. Recently, we assembled the first curated database of mitochondrial-encoded genes, MIDORI, for taxonomic assignments of metazoan sequences (Machida *et al.*, 2017). Mitochondrial genes provide higher taxonomic resolution for most metazoan groups than nuclear-encoded genes. As a result, they have been increasingly targeted in metagenetics and metagenomics studies (Leray and Knowlton, 2016). MIDORI was built by retrieving all nucleotide sequences from GenBank BLAST NT and, after quality filtration, includes metazoan mitochondrial sequences for 13 protein-coding (ATP synthase sub-unit 6 and 8; Cytochrome oxidase sub-unit I, II and III; Cytochrome b apoenzyme; NADH dehydrogenase sub-units 1–4, 4L, 5 and 6) and two ribosomal RNA genes (Large and Small ribosomal sub-unit RNA) with species-level level taxonomic information (see details in Machida *et al.*, 2017).

2 Server description

Here, we present MIDORI server, a user-friendly platform to facilitate taxonomic classification of mitochondrial-encoded gene

sequences with MIDORI. The server currently performs taxonomic assignments with three algorithms that predict taxonomy using *k-mer* similarity: SPINGO (Allard *et al.*, 2015), RDP classifier (Wang *et al.*, 2007) and SINTAX (Edgar, 2016). A maximum of 10 000 sequences in a FASTA format can be uploaded at once, and all of them must be shorter than 4000 base pairs. Each algorithm can be run using two versions of each of the 15 mitochondrial-encoded gene reference datasets: MIDORI-Unique and MIDORI-Longest. MIDORI-Unique contains all haplotypes of every species while MIDORI-Longest contains a single haplotype per species, the longest one. For example, MIDORI-Longest for the COI gene contains the longest sequence for every species represented in the COI dataset. Using 1336 zooplankton sequences (Machida *et al.*, 2009, 500 bp), we estimated the time required for assignments using the three algorithms with default settings (reference: COI-Longest). As a result, relatively longer calculation time was required for RDP classifier (630 s), compared to SPINGO (90 s) and SINTAX (100 s). Assigned phyla were compared between the results obtained from RDP classifier and SINTAX. The result indicated that about 10% of assignments were inconsistent between the results (most likely the groups with fewer reference sequences). Furthermore, we have also deposited the results of Leave One Out Test in MIDORI web site (http://www.reference-midori.info/download.php, Wang *et al.*, 2007). These results indicated that possibility of mis-assignment increases with the supporting bootstrap values decrease, demonstrating the importance of careful interpretation of results obtained for the analyses.

The server is designed to give full flexibility to the user and functions with recent major browsers. A range of options is available for each algorithm such as assignment confidence cut-off (RDP), *k-mer* size (SPINGO) and bootstrap cut-off (SINTAX). A question mark button located next to each option provides hint details to the user when hovered by the cursor. The user can provide an e-mail address to receive the text-formatted result of the analysis.

The server was extensively tested using mock sample and real environmental data. It is easy to use and does neither require any registration nor specific software to be installed locally. The RDP classifier is pre-trained with each of the reference datasets.

3 Conclusion

As bio-monitoring and bio-surveillance increasingly rely on mitochondrial-encoded sequence data, the ability to rapidly and reliably assign metazoan sequences to taxonomic groups has become indispensable. MIDORI server enables the classification of large number of unknown metazoan reads to taxa represented in the curated reference database. MIDORI will be regularly updated. We also intend to implement several additional taxonomic assignment algorithms on MIDORI server in the near future [e.g. SAP (Munch *et al.*, 2008), RTAX (Soergel *et al.*, 2012), METAXA2 (Bengtsson-Palme *et al.*, 2015)].

Acknowledgements

The authors would like to thank Chao-Yu Pan and Peter Hsiao for technical assistances. They would also like to thank three anonymous referees for thoughtful and insightful comments.

Funding

This work was supported by Ministry of Science and Technology, Taiwan [grant number 105-2621-B-001-003]; and Academia Sinica.

Conflict of Interest: none declared.

References

Al-Rshaidat,M.M.D. *et al.* (2016) Deep COI sequencing of standardized benthic samples unveils overlooked diversity of Jordanian coral reefs in the northern Red Sea. *Genome*, **59**, 724–737.

Allard,G. *et al.* (2015) SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics*, **16**, 324.

Arulandhu,A.J. *et al.* (2017) Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. *GigaScience*, **6**, 1–18.

Bengtsson-Palme,J. *et al.* (2015) METAXA2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol. Ecol. Res.*, **15**, 1403–1414.

Edgar,R. (2016) SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*, doi: 10.1101/074161.

Guillou,L. *et al.* (2012) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.*, **41**, D597–D604.

Ji,Y. *et al.* (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecol. Lett.*, **16**, 1245–1257.

Leray,M. and Knowlton,N. (2016) Censusing marine eukaryotic diversity in the twenty-first century. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **371**, 20150331.

Machida,R.J. *et al.* (2017) Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Sci. Data*, **4**, 170027.

Machida,R.J. *et al.* (2009) Zooplankton diversity analysis through single-gene sequencing of a community sample. *BMC Genomics*, **10**, 438.

Munch,K. *et al.* (2008) Statistical assignment of DNA sequences using Bayesian phylogenetics. *Syst. Biol.*, **57**, 750–757.

Quast,C. *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596.

Raclariu,A.C. *et al.* (2017) Comparative authentication of Hypericum perforatum herbal products using DNA metabarcoding, TLC and HPLC-MS. *Sci. Rep.*, **7**, 1291.

Soergel,D.A.W. *et al.* (2012) Selection of primers for optimal taxonomic classification of environmental 16S rRNA sequences. *Ismej.*, **6**, 1440–1444.

Taberlet,P. *et al.* (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.*, **21**, 2045–2050.

Wang,Q. *et al.* (2007) Naïve bayesian classifier for rapid assignment of rRNA sequences into the new bacteral taxonomy. *Appl. Environ. Microb.*, **73**, 5261–5267.