MBMG
Metabarcoding & Metagenomics

Metabarcoding and Metagenomics 4: 19–21
DOI 10.3897/mbmg.4.53535

Software Description

BOLDigger – a Python package to identify and organise sequences with the Barcode of Life Data systems

Dominik Buchner[1], **Florian Leese**[1]

1 *University of Duisburg-Essen, Aquatic Ecosystem Research, Universitaetsstr. 5, 45141 Essen, Germany*

Corresponding author: Dominik Buchner (dominik.buchner@uni-due.de)

Academic editor: Dirk Steinke | **Received** 22 April 2020 | **Accepted** 29 May 2020 | **Published** 3 June 2020

Abstract

DNA metabarcoding workflows produce hundreds to ten-thousands of Operational Taxonomic Units (OTUs) or Exact Sequence Variants (ESVs) per analysis. In most workflows, a taxonomic assignment to these generated sequences is needed. This is typically done using publicly available databases. Especially, yet not exclusively, for Eumetazoan metabarcoding, the Barcode of Life Data system (BOLD) is the most comprehensive and curated reference barcode database and, therefore, typically the first choice for taxonomic assignment. While an application programme interface (API) exists to query data in large batches, no information on the many and important unpublished data are obtained through the API. The alternative approach using the BOLD identification engine on the website provides full access, yet it is restricted to 100 sequences at once. We developed a small platform-independent and graphical user interface (GUI) software package, BOLDigger, which aims to solve this problem by automating the process of sending successive requests of up to 100 sequences without surpassing the capacities of BOLD. BOLDigger can be used to download the results of the identification engine, as well as metadata for the obtained hits. For the selection of the best fitting hit, three different methods are implemented. A new approach, combining a threshold-based approach with the metadata information, was implemented to make use of the metadata.

Key Words

metabarcoding, species identification, BOLD, OTUs, taxonomic assignment, database

Introduction

DNA metabarcoding is a cost- and time-effective method to assess species diversity of bulk or environmental samples (Taberlet et al. 2012; Yu et al. 2012; Elbrecht and Steinke 2019). DNA metabarcoding datasets often consist of hundreds or even thousands of Operational Taxonomic Units (OTUs) or Exact Sequence Variants (ESVs), which need to be queried against databases to assign taxonomy. The Barcode of Life Data System (BOLD) offers such a database with more than 7 million reference sequences (Barcodes) for the primary barcode sequence in the animal kingdom, the mitochondrial cytochrome c oxidase I gene fragment (COI) (Ratnasingham and Hebert 2007). The database also supports plant reference barcodes, with about 500,000 sequences of the ribulose bisphosphate carboxylase and maturase K genes (rbcL & matK) and fungi, with about 150,000 reference sequences of the Internal Transcribed Spacer region (ITS).

The BOLD Identification System (IDS) can be used to identify an unknown query sequence via the website or the provided (fast) API by tracing and returning the nearest neighbours to the query sequence from a global alignment of all reference sequences (Ratnasingham and Hebert 2007). While the identification engine of the website is limited to 100 sequences at once, one downside of using the faster API for sequence identification is that it only provides access to published COI records while the website also provides private and early release data that represent about 50% of all records on BOLD (Weigand et al. 2019). Even though these records are less trustworthy, since the underlying data are not accessible, they still hold valuable information that can be used for sequences that lack publicly available reference data. While it is assumed that, with growing data, the IDS will deliver a definite species-level hit for a given sequence (Ratnasingham and Hebert 2007), this is still very often not the case. For example, Weigand et al. (2019) showed that, of the 4504 freshwater macroinvertebrates used for routine monitoring in Europe, about 65% of the species are represented by at least one barcode, showing that there are still large gaps to fill, even for important groups like freshwater macroinvertebrates. Therefore, BOLD applies conservative rules to return a so-called top-hit that solely rely on sequence similarity and gives access to all available information about the chosen reference sequence (Ratnasingham and Hebert 2007). Most often the chosen top-hit simply is the first hit of the first 99 nearest neighbours, even if there are other records with a similarity above 99%. To avoid this, a threshold-based approach, including thresholds for different taxonomic ranks that also consider the metadata, was implemented in BOLDigger.

Sequence similarity thresholds are used for taxonomic assignment across all domains of life (Hebert et al. 2003; Venter et al. 2004; Fazekas et al. 2008). Despite being criticised to not be applicable for all taxonomic groups and amplicon lengths (Mahé et al. 2015) or being different between taxonomic groups (Kvist 2016; Meyer and Paulay 2005), they have strong empirical support, especially for species level, for large groups, such as birds, fish and several insect orders (Hebert et al. 2003; Virgilio et al. 2010; Ward et al. 2005). For genus, family, order and higher ranks, the sequence similarities differ between taxonomic groups, due to the different evolutionary histories and mutational speeds. However, by using conservative threshold values for the different taxonomic levels, false positives can effectively be prevented while losing taxonomic resolution (Ratnasingham and Hebert 2007). More comprehensive reference databases can solve this challenge.

The presented Python package BOLDigger aims to act as an interface for species identification, to download additional data and organisation of these. As a platform-independent, open-source tool, it can be used to collect IDS results from BOLD, including private and early release data. It also provides the user with additional data for all public references in the dataset, as well as implementing a safer way to determine the top-hit by combining a threshold-based approach with the additional information provided by BOLD. To improve user-friendliness, a BOLDigger comes with a GUI (Fig. 1).

Package description

The Python package BOLDigger (version 1.1.5) is available from the Python Package Index (PyPI) at https://pypi.org/project/boldigger/. It can be installed using the Python package installer (pip) with the command **pip install boldigger**. In case both python version 2 and 3 are installed on the operating system, the correct version of pip has to be used (**pip3 install boldigger**). All operating systems (Windows, Linux and MacOS) are supported, as long as Python 3 is installed. It can be started with the command **boldigger** from the command line after installation. Updates can be automatically downloaded and installed with the command **pip install -- upgrade boldigger**. Further information about installation, the current version and troubleshooting are provided via the GitHub repository page (https://github.com/DominikBuchner/BOLDigger).

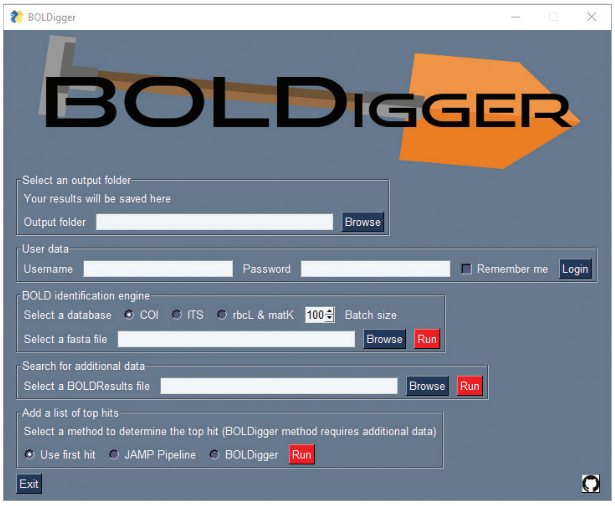

Figure 1. Graphical User Interface of BOLDigger.

BOLDigger comes with a GUI for easy operation (Fig. 1). All output will be saved to the output folder. Since a login is required to use the IDS for more than one sequence, an account at BOLD and its user data is required by BOLDigger. BOLDigger can query all three databases of BOLD by using the "BOLD identification engine" command. The batch size controls the number of sequences to be queried at once (e.g. a fasta file containing 1000 sequences will send 10 successive requests). All results are saved to an excel file. This file can be used to download additional data with the "Search for additional data" command. Additional data will simply be added to this file. The "Add a list of top hits" command adds a list of top-hits to a new worksheet of the result file with different methods. For a detailed description of the different ways, please consult the GitHub repository (https://github.com/DominikBuchner/BOLDigger).

Conclusions

BOLDigger is a platform-independent GUI software package that allows users to query metabarcoding data against the BOLD sequence database in a simple fashion. It facilitates data analysis and provides alternative approaches for the assignment of the best hit.

Project description

Title: BOLDigger – a Python package to identify and organise sequences with the Barcode of Life Data systems

Study area description: Metabarcoding, eDNA, Barcoding, Biomonitoring
Download page: https://pypi.org/project/boldigger/
Programming language: Python 3
Licence: MIT Licence

Author contributions

Conceived and designed the study: DB; Wrote the Python package: DB; Wrote the paper: DB, FL

Acknowledgements

We thank the leeselab members, especially Arne J. Beermann and Till-Hendrik Macher for comments and feedback on the programme. We thank the BOLD support team for support with respect to the data mining via the IDS. Till-Hendrick Macher kindly designed the BOLDigger logo. FL is member of and supported by COST Action DNAqua-Net (CA15219).

References

Elbrecht V, Steinke D (2019) Scaling up DNA metabarcoding for freshwater macrozoobenthos monitoring. Freshwater Biology 64: 380–387. https://doi.org/10.1111/fwb.13220

Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM, Hajibabaei M, Barrett SCH (2008) Multiple Multilocus DNA Barcodes from the Plastid Genome Discriminate Plant Species Equally Well. PLOS ONE 3: e2802. https://doi.org/10.1371/journal.pone.0002802

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London. Series B: Biological Sciences 270: 313–321. https://doi.org/10.1098/rspb.2002.2218

Kvist S (2016) Does a global DNA barcoding gap exist in Annelida? Mitochondrial DNA Part A 27: 2241–2252.

Mahé F, Rognes T, Quince C, Vargas C de, Dunthorn M (2015) Swarm v2: highly-scalable and high-resolution amplicon clustering. PeerJ 3: e1420. https://doi.org/10.7717/peerj.1420

Meyer CP, Paulay G (2005) DNA Barcoding: error rates based on comprehensive sampling. PLoS Biol 3: e422. https://doi.org/10.1371/journal.pbio.0030422

Ratnasingham S, Hebert PDN (2007) bold: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular Ecology Notes 7: 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology 21: 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y-H, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66–74. https://doi.org/10.1126/science.1093857

Virgilio M, Backeljau T, Nevado B, De Meyer M (2010) Comparative performances of DNA barcoding across insect orders. BMC Bioinformatics 11: 206. https://doi.org/10.1186/1471-2105-11-206

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PDN (2005) DNA barcoding Australia's fish species. Philosophical Transactions of the Royal Society B: Biological Sciences 360: 1847–1857. https://doi.org/10.1098/rstb.2005.1716

Weigand H, Beermann AJ, Čiampor F, Costa FO, Csabai Z, Duarte S, Geiger MF, Grabowski M, Rimet F, Rulik B, Strand M, Szucsich N, Weigand AM, Willassen E, Wyler SA, Bouchez A, Borja A, Čiamporová-Zaťovičová Z, Ferreira S, Dijkstra K-DB, Eisendle U, Freyhof J, Gadawski P, Graf W, Haegerbaeumer A, van der Hoorn BB, Japoshvili B, Keresztes L, Keskin E, Leese F, Macher JN, Mamos T, Paz G, Pešić V, Pfannkuchen DM, Pfannkuchen MA, Price BW, Rinkevich B, Teixeira MAL, Várbíró G, Ekrem T (2019) DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. Science of the Total Environment 678: 499–524. https://doi.org/10.1016/j.scito tenv.2019.04.247

Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods in Ecology and Evolution 3(4): 613–623. https://doi.org/10.1111/j.2041-210X.2012.00198.x