ARPHA Conference Abstracts 4: e64936
doi: 10.3897/aca.4.e64936

Conference Abstract

From DNA sequences to operational reference databases: an opinionated approach using R

François Keck[‡], Florian Altermatt[‡,§]

‡ Eawag: Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
§ Department of Evolutionary Biology and Environmental Studies, University of Zürich, Zürich, Switzerland

Corresponding author: François Keck (francois.keck@gmail.com)

Received: 24 Feb 2021 | Published: 04 Mar 2021

Citation: Keck F, Altermatt F (2021) From DNA sequences to operational reference databases: an opinionated approach using R. ARPHA Conference Abstracts 4: e64936. https://doi.org/10.3897/aca.4.e64936

Abstract

Reference databases of sequences that have been taxonomically assigned are a key element for DNA-based identification of organisms. Accurate and complete reference databases are necessary to associate a correct taxonomic name to the sequences obtained in studies using metabarcoding. Today many research projects using DNA metabarcoding include the development of a custom reference database, often derived from large repositories like GenBank. At the same time, many projects are focussing on the development of ready-to-use databases validated by experts and targeting specific markers and taxonomic groups.

While mainstream tools such as spreadsheet softwares may be suitable to manage small databases, they quickly become insufficient when the amount of data increases and validation operations become more complex. There is a clear need for providing user-friendly and powerful tools to manipulate biological sequences and manage reference databases. The R language which is a free software and has already been adopted by many researchers to perform their analyses is highly suitable to develop such tools.

In this talk, we will outline the approach we recommend to handle small- to middle-sized reference databases, currently still making the majority of projects. We will advocate that a simple tabular approach where each sequence constitutes an observation may be the most adequate. While such a single table may be less flexible and less optimized than relational databases or more complex data structures, it is easy to maintain and allows the direct use

of modern dataframe centric tools. We will specifically present and discuss two R packages that can be used jointly to make reference database development more accessible and more reproducible. First, we will briefly introduce bioseq (Keck 2020) which is dedicated to biological sequence manipulation and analysis. The package implements classes and functions to make analyses of complex datasets including DNA, RNA or protein sequences as simple as possible. The strength of bioseq is to provide standard and more advanced functions to perform low level operations through a simple and consistent programming interface. Then we will present refdb, which has been developed as an environment for semi-automatic and assisted construction of reference databases. The refdb package is a reference database manager offering a set of powerful functions to import, organize, clean, filter, audit and export the data. We will outline how these two packages together can speed up reference database generation and handling, and contribute to standardization and repeatability in metabarcoding studies.

Keywords

DNA metabarcoding; reference database; biological sequences, taxonomy; R package

Presenting author

François Keck

Presented at

1st DNAQUA International Conference (March 9-11, 2021)

References

- Keck F (2020) Handling biological sequences in R with the bioseq package. Methods in Ecology and Evolution 11 (12): 1728-1732. https://doi.org/10.1111/2041-210X.13490