Assigning Latin scientific names to OTUs based on sequence clusters (Topic 4)

Aim/problem/goal

Curated sequence databases are important tools in modern taxonomy. They are used to identify sequences at the Operational Taxonomic Unit (OTU) level. OTUs are usually represented by some stable identifier, such as the Species Hypothesis (SH) in UNITE (Kõljalg et al. 2020) or the Barcode Index Number (BIN) in the Barcode of Life Data System (BOLD). In principle, these identifiers represent Species/Taxon concepts. In order to answer the question "What species does this sequence represent?" a linkage from an OTU identifier to a latin scientific name is needed, if existing. The taxon name for an OTU in the reference database has to somehow be derived from the taxonomic annotation of the sequences constituting the OTU. Currently, BOLD does not provide a single consensus taxon name for each BIN. In order to apply a taxon name, users therefore have to inspect all taxonomic annotations within the BIN and pick one. When blasting a single or a few sequences, this approach may suffice. However, in a taxonomic classification pipeline for many sequences (e.g. metabarcoding) this approach is impossible. Similarly, in order to place BINs or SHs into classic taxonomies such as the GBIF backbone taxonomy or the Catalogue of Life all bins must be unambiguously linked to a (parent) taxon. Therefore, the aim is to explore the algorithms for taxonomic assignment currently used by UNITE/PlutoF and the International Barcode of Life project (iBOL) Barcode Index Numbers (BINs) dataset in GBIF (The International Barcode of Life Consortium 2016) and discuss shortcomings and advantages. This project also aims to explore improvements based on the underlying data.

Method

A set of NCBI accessions with taxon labels assigned was used as input data. This could be imagined to either be the members of an OTU or top ‘X’ matches of a blast result set.

1. Clean/normalise names. i.e. informal names like ‘Bactrocera sp.27’ should be discarded at species level and be snapped to a higher taxon, in this example Bactrocera. This was best done using the GBIF species match API (https://www.gbif.org/developer/species#searching).
2. For all species level names, find the year of description and synonymy. Here, we used the Catalogue of Life (COL) nameusage search API (http://api.catalogueoflife.org/#/default/searchDataset).
3. For each accession with a species level taxon assigned, find out if the sequence was derived from type material. We accomplished this using a combination of the NCBI Entrez API ESearch and EFetch methods (Entrez Programming Utilities Help, https://www.ncbi.nlm.nih.gov/books/NBK25501/).

Results

During the hackathon we did "proof of concept" implementations of each of the three method steps in the R and Nodejs programming languages. Apart from further testing and improving error handling, the outstanding work would be chaining the steps into a pipeline that would fulfill the objective of the topic. As an outcome of the work on step 3 i.e. retrieving type information from NCBI, we found that the NCBI Targeted Loci RefSeq projects are high quality

data sources for Type specimens of Fungi and Prokaryotes. Hence a spin-off project in the form of an API adapter was written to make these projects available through GBIF (see Robbertse 2022 and McVeigh 2022) where they now contribute DNA sequences as well as bibliographic references to the clustered specimen view in GBIF.

Conclusion

APIs are already available to fulfill the goals of this topic (Fig. 4). However, these are spread across three different infrastructures and some subtasks require quite detailed knowledge of the underlying data structures. A full pipeline implementation of the proposed algorithm in for example R would therefore be a useful tool for taxonomic annotation of OTUs/sequence clusters.

References

- Kõljalg, Urmas, Henrik R. Nilsson, Dmitry Schigel, Leho Tedersoo, Karl-Henrik Larsson, Tom W. May, Andy F. S. Taylor, et al. 2020. 'The Taxon Hypothesis Paradigm—On the Unambiguous Detection and Communication of Taxa'. *Microorganisms* 8 (12): 1910. https://doi.org/10.3390/microorganisms8121910.
- McVeigh, Richard. 2022. 'Bacterial 16S Ribosomal RNA RefSeq Targeted Loci Project'. National Center for Biotechnology Information (NCBI). https://doi.org/10.15468/K2C8EN.
- Robbertse, Barbara. 2022. 'Fungal Internal Transcribed Spacer RNA (ITS) RefSeq Targeted Loci Project'. National Center for Biotechnology Information (NCBI). https://doi.org/10.15468/CM2GBP.
- The International Barcode of Life Consortium (2016). International Barcode of Life project (iBOL) Barcode Index Numbers (BINs). Checklist dataset https://doi.org/10.15468/wvfqoi accessed via GBIF.org on 2022-02-15.