UNITE Species Hypotheses Matching Analysis

Kessy Abarenkov[‡], Urmas Kõljalg[§], R. Henrik Nilsson[|]

‡ University of Tartu Natural History Museum, Tartu, Estonia
§ University of Tartu, Tartu, Estonia
| University of Gothenburg, Göteborg, Sweden

Corresponding author: Kessy Abarenkov (kessy.abarenkov@ut.ee)

Abstract

UNITE (**U**ser-friendly **N**ordic **ITS** **E**ctomycorrhizal Database) (Nilsson et al. 2018) is a theoretical and practical platform for calculating, identifying and communicating DNA-based species hypotheses that may not have been described as formal species yet. UNITE species hypotheses (SH) matching analysis is a digital service for the global species discovery from environmental DNA (eDNA). SH matching service is based on UNITE datasets hosted in PlutoF. The tool places a user's unknown DNA sequences in FASTA format into existing UNITE species hypotheses and forms SHs not yet present in the system. Its output in a series of CSV and HTML files (Fig. 1) includes information about what species are present in eDNA samples, whether they are potentially undescribed new species, where they are found in other studies, whether they are alien or threatened species, etc. The service is very useful for understanding species distribution patterns, host range, etc.

Analysis output will provide Digital Object Identifier (DOI)-based stable identifiers for communicating species hypotheses found in eDNA. DOIs are connected to the taxonomic backbone of PlutoF and GBIF (Global Biodiversity Information Facility). In this way every DOI is accompanied by a taxon name, which is still widely used for the communication of species. In the case of undescribed species, DOIs will be issued by the PlutoF system. All UNITE services are focused on fungi but cover all Eukarya by using publicly available rDNA ITS (ribosomal DNA Internal Transcribed Spacer) marker sequences accompanied by sample metadata. SH matching analysis results can be published in GBIF as DNA-derived occurrence data where occurrences are linked to SH identifiers present in GBIF backbone taxonomy.

Source code of the SH matching analysis tool is available in GitHub with an implementation accessible online on the PlutoF platform for registered users. PlutoF (Abarenkov et al. 2010) is an open data management platform for biology and related disciplines. Users can manage their bodiversity datasets through a full data life cycle—from uploading to publishing and archiving in FAIR (Findable, Accessible, Interoperable, Reusable) way. PlutoF also features an analysis module by providing analytical services for molecular sequence identification and species discovery from eDNA samples. There

are more than 8 000 registrered users in PlutoF with more than 500 users who have used its analysis module (as of August, 2022).

Keywords

eDNA, sequence classification, species hypothesis

Presenting author

Kessy Abarenkov

Presented at

TDWG 2022

Conflicts of interest

References

- Abarenkov K, Tedersoo L, Nilsson RH, Vellak K, Saar I, Veldre V, Parmasto E, Prous M, Aan A, Ots M, Kurina O, Ostonen I, Jõgeva J, Halapuu S, Põldmaa K, Toots M, Truu J, Larsson K, Kõljalg U (2010) PlutoF—a Web Based Workbench for Ecological and Taxonomic Research, with an Online Implementation for Fungal ITS Sequences. Evolutionary Bioinformatics 6: 189-196. https://doi.org/10.4137/ebo.s6271
- Nilsson RH, Larsson K, Taylor AF, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tedersoo L, Saar I, Kõljalg U, Abarenkov K (2018) The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Research 47 (D1): D259-D264. https://doi.org/10.1093/nar/gky1022

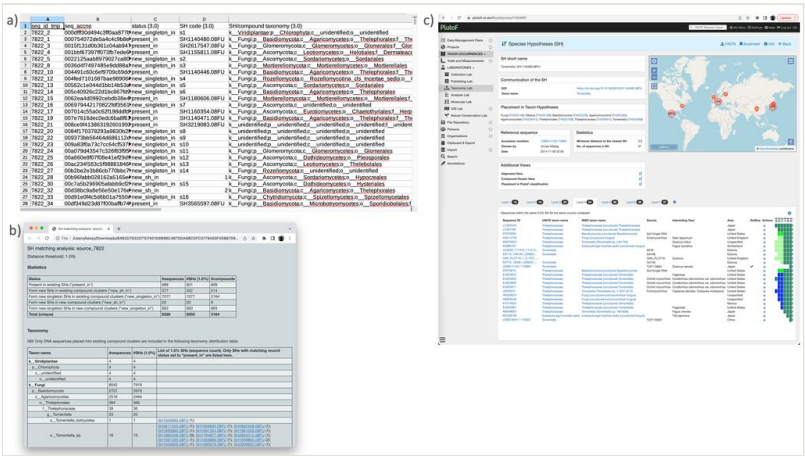

Figure 1.

UNITE species hypotheses (SH) matching service output - a) CSV-formatted data matrix describing user's query sequences placed into existing SHs or forming new SHs, b) screenshot of the HTML output summarizing the analysis results present in the accompanying CSV files, c) screenshot of the SH1140480.08FU - one of the species hypotheses where query sequences were placed.