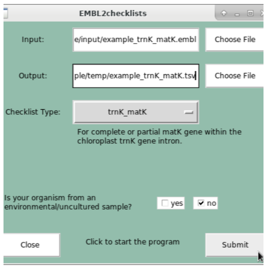
Jan 09, 2019

Usage of EMBL2checklists

PLOS One

DOI
dx.doi.org/10.17504/protocols.io.v6me9c6



Input: e/input/example_trnK_matK.embl

Output: ple/temp/example_trnK_matK.tsv

Checklist Type: trnK_matK

For complete or partial matK gene within the chloroplast trnK gene intron.

Is your organism from an environmental/uncultured sample? □ yes ☑ no

Click to start the program

Michael Grünstäudl[1]

[1]Freie Universität Berlin

Michael Grünstäudl
Fort Hays State University

DOI: https://dx.doi.org/10.17504/protocols.io.v6me9c6

External link: https://www.biorxiv.org/content/early/2018/10/05/435644

Protocol Citation: Michael Grünstäudl 2019. Usage of EMBL2checklists. **protocols.io** https://dx.doi.org/10.17504/protocols.io.v6me9c6

Manuscript citation:
Gruenstaeudl M., Hartmaring Y. (2018). EMBL2checklists: A Python package to facilitate the user-friendly submission of plant DNA barcoding sequences to ENA. bioRxiv 435644; doi: https://doi.org/10.1101/435644

Protocol status: Working
We use this protocol and it's working

Created: December 05, 2018

Last Modified: January 09, 2019

Protocol Integer ID: 18349

Keywords: bioinformatics, DNA sequence submission, ENA, European Nucleotide Archive, plant DNA barcoding, plant systematics, Python, sequence database, software tool, Webin, ready checklist files from raw dna sequence data, disseminating novel dna sequence data, raw dna sequence data, novel dna sequence data, converts dna sequences from the common embl, sequence section of the european nucleotide archive, european nucleotide archive, metadata via embl2checklist, genbank flat file formats to submission, public sequence database, submission of dna sequence, converts dna sequence, dna sequence, embl2checklist, file preparation for database submission, fungal dna barcoding, usage of embl2checklist, annotated sequence section, specific ena checklist, file preparation, metadata into the idiosyncratic format, dna, ready checklist file, interactive webin submission system of ena, associated metadata, step protocol of the bioinformatic step, flat file format, bioinformatic step, delimited spreadsheet, friendly software tool, au

Abstract

The submission of DNA sequences to public sequence databases is an essential, but insufficiently automated step in the process of generating and disseminating novel DNA sequence data. A user-friendly software tool is needed that streamlines the file preparation for database submissions of DNA sequences that are commonly generated in plant and fungal DNA barcoding. A Python package was developed that converts DNA sequences from the common EMBL and GenBank flat file formats to submission-ready, tab-delimited spreadsheets (so-called "checklists") for a subsequent upload to the annotated sequence section of the European Nucleotide Archive (ENA). The package, titled "EMBL2checklists", automatically converts DNA sequences, their annotation features, and associated metadata into the idiosyncratic format of marker-specific ENA checklists and, thus, generates output that can be uploaded via the interactive Webin submission system of ENA. Here, we present a step-by-step protocol of the bioinformatic steps necessary to generate submission-ready checklist files from raw DNA sequence data and associated metadata via EMBL2checklists.

Attachments

PDF

Gruenstaeudl.and.Har...

1.2MB

Troubleshooting

1 **Annotation of DNA sequences**

Add sequence features and feature qualifiers to each DNA sequence using **INSDC-compatible feature table** keywords.
Use *any* of the following software tools:

Software	
Geneious	NAME
Kearse et al. (2012) Geneious Basic: An integrated and extendable desktop software platform for ... Bioinformatics 28: 1647-1649	DEVELOPER

Geneious is among the most efficient and best-documented tools to **adding sequence features and feature qualifiers** to DNA sequences. For more information regarding the automatic annotation of sequences with Geneious, see the following video tutorial on Youtube:

https://www.youtube.com/embed/CY1e2RkULas

Software	
Artemis	NAME
Rutherford et al. (2000) Artemis: Sequence visualization and annotation. Bioinformatics 16: 944--945	DEVELOPER

Software	
DnaSP	NAME
Rozas et al. (2017) DnaSP v6: DNA Sequence Polymorphism Analysis of Large Datasets. Mol. Biol. Evol. 34: 3299-3302	DEVELOPER

For exporting annotated DNA sequences as GenBank- or EMBL-formatted flat files, see pages 44 to 46 of the **user manual of DnaSP v.6.12**.

2 **Saving sequences as flat file**

Save multiple DNA sequences of the same barcoding marker as an single, multi-sequence flat file in EMBL or GenBank format. Use *any* of the following software tools:

Software	
Geneious	NAME
Kearse et al. (2012) Geneious Basic: An integrated and extendable desktop software platform for ... Bioinformatics 28: 1647-1649	DEVELOPER

Geneious is among the most efficient and best-documented tools to **export annotated DNA sequences** as a GenBank flat file format.

Software	
Artemis	NAME
Rutherford et al. (2000) Artemis: Sequence visualization and annotation. Bioinformatics 16: 944--945	DEVELOPER

Software	
DnaSP	NAME
Rozas et al. (2017) DnaSP v6: DNA Sequence Polymorphism Analysis of Large Datasets. Mol. Biol. Evol. 34: 3299-3302	DEVELOPER

For exporting annotated DNA sequences as GenBank- or EMBL-formatted flat files, see page 93 of the **user manual of DnaSP v.6.12**.

3 **Validation of flat file**

Test the validity of the file format, the feature table syntax or the taxonomic status of organism names, among other aspects, using one of two software tools.

For validation of *EMBL-formatted flat files*, use:

Software

EMBL flat file validator NAME

Command

Command to validate the format and content of an EMBL-formatted flat file (which is used as input to EMBL2checklists in the subsequent protocol step) via the EMBL flat file validator.

```
java -jar embl-api-validator-1.1.155.jar example_ITS.embl
```

For validation of *GenBank-formatted flat files*, use:

Software

SBOL Validator NAME

4 **Installation of EMBL2checklists**

It is recommended that you install EMBL2checklists via **pip** (i.e., the recommended installer of the **Python Package Index**). On Linux and MacOS, this can be achieved via the following command:

Command

Command to install EMBL2checklists via the default Python installer.

```
(sudo) pip2 install EMBL2checklists
```

Conversion from flat file to checklist

Convert the EMBL- or GenBank-formatted flat file into the Webin checklist file format using the software:

Software

EMBL2checklists

NAME

Command

Command to run EMBL2checklists on an example dataset provided as part of the Python package.

```
EMBL2checklists_CLI \
    -i example/input/example_trnK_matK.embl \
    -o example/temp/example_trnK_matK.tsv \
    -c trnK_matK \
    -e no
```

```
michael_science /home/michael_science/git/michaelgruenstaeudl_EMBL2checklis
ts $
```

Video tutorial of the correct execution of EMBL2checklists via the command-line on an example dataset provided as part of the Python package.

Command

Command to run EMBL2checklists on an example dataset provided as part of the Python package.

```
EMBL2checklists_GUI
```

```
michael_science /home/michael_science/git/michaelgruenstaeudl_EMBL2checklis
ts $ 
```

Video tutorial of the correct execution of EMBL2checklists via the GUI on an example dataset provided as part of the Python package.

5 **Upload checklist to ENA**

Upload the resulting checklist to the sequence section of **ENA** using the **interactive route of the Webin submission system**.