JAN 09, 2019

OPEN ACCESS

DOI:

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

External link:

<https://www.biorxiv.org/content/early/2018/10/05/435644>

Protocol Citation: Michael MG Gruenstaeudl 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>

License: This is an open access protocol distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Usage of EMBL2checklists

PLOS One

Michael MG Gruenstaeudl¹

¹Freie Universität Berlin

Michael MG Gruenstaeudl

Freie Universität Berlin

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

[Gruenstaeudl.and.Hartmaring.2018_EMBL2checklists_PREPRINT.pdf](#)

Created: Dec 05, 2018

Last Modified: Jan 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

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:

[h](#)

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
```

```
File Edit View Terminal Tabs Help
michael_science /home/michael_science/git/michaelgruenstaeudl_EMBL2checklis
ts $
```

Video tutorial of the correct execution of the EMBL flat file validator on an input file to EMBL2checklists.

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
```

Video tutorial of the correct installation of EMBL2checklists, including the testing of software initialization upon installation.

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
```

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

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](#).