**Manual of FuzzyID2**

FuzzyID2 is a software package for species identification via DNA barcoding. We here devise a two-step searching strategy, which firstly use Hidden Markov Models (HMM) algorithm to narrow the searching scope, and then find out its corresponding species of queries using minimum K2P (Kimura-two-parameter) distance.

The software is written by C++ and uses Bio++ library to estimate genetic distance. We have compiled the software on 32-bit and 64-bit linux distributions. FuzzyID2 as a free software is released under CNU General Public License and is available at **https://github.com/zhangab2008/FuzzyID2.git**. The FuzzyID2 distribution includes a user manual and a sample dataset.

FuzzyID2 runs under command-line interface. Interaction with the software involves command-line arguments.

**FuzzyID2 installation**

1. Upload the FuzzyID2 software package “FuzzyID2.tar.gz” to the linux server.
2. Open the linux terminal console in the FuzzyID2 software direction.
3. Type “tar zxvf FuzzyID2.tar.gz” to decompress the software package, and then type “cd FuzzyID2” to change direction.
4. Type “./FuzzyID2\_i686\_setup.sh” to configure and install the software for 32 bit computing environment or “./FuzzyID2\_X86-64\_setup.sh” for 64 bit environment.

**FuzzyID2 options:**

-c Specify an operational mode to run.

There are three operational modes to run in FuzzyID2.

-c Theta1 Calculate maximum intraspecific genetic distance of each species which contains in the reference database.

Example: ./FuzzyID2 -c Theta1 -m K2P -d Lepidoptera

-c Theta2 Calculate minimum interspecific genetic distance between each species and its nearest-neighbor in the reference database.

Example: ./FuzzyID2 -c Theta2 -m K2P -d Lepidoptera

-c MD Calculate the pairwise genetic distance between query sequence and reference barcodes. The nearest-neighbor barcode from reference dataset which has the smallest K2P distance with its query sequence will become the finally hit.

Example: ./FuzzyID2 -c MD -in query.fas -m K2P -out outPut.csv -d Lepidoptera

-d Specify the reference database.

-in Specify the query sequences file name.

The sequence file must be *fasta* format.

-m Specify the nucleotide substitution models.

There are three nucleotide substitution models available in FuzzyID2: K2P (default), JC69 and GTR.

-out Specify the species identification result file name.

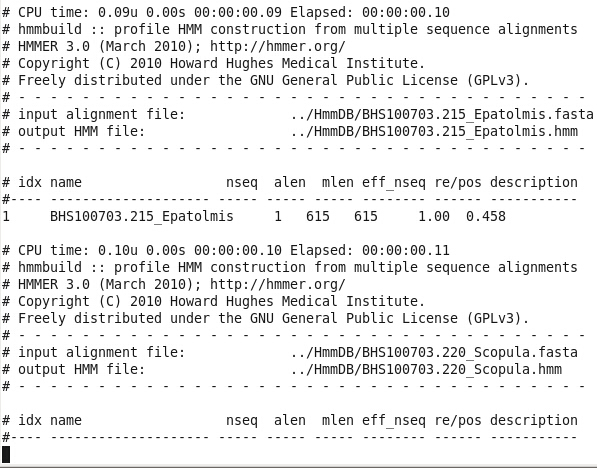
The result file is provided in CSV format.

**Construct the local reference databases**

The reference barcode datasets have to be provided in *fasta* format or BOLD’s *tsv* format. It will be firstly conserved in sqlite3 database by a python script, “makeDB.py”, to build local reference database and Hmm database automatically.

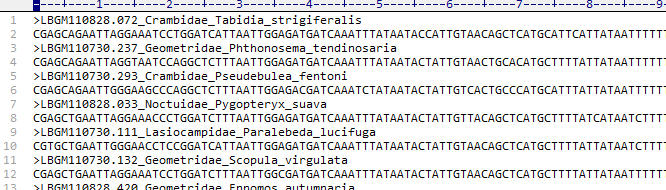
The sequence name of reference barcode in *fasta* format file is required as “>specimenID\_familyName\_genusName\_speciesName”.

1. Upload reference barcode file to the linux server and conserved under “DB” folder in the FuzzyID2 directory.
2. Open the linux terminal console in the FuzzyID2 directory and type “cd DB” command to change direction to “DB” folder.
3. Type “python3 makeDB.py” command in the consoleand press Enter.
4. Type “1”(for BOLD’s *tsv* format reference barcodes file) or “2”(for *fasta* format reference barcodes file) and press Enter.
5. Type the reference barcodes file name which is conserved in “DB” folder, and press Enter.
6. Enter the maximum number of samples in each species to build rough HMM database. This parameter can be decided according to the number of reference size and the default value is 5. For example, set 10 samples when the number of barcodes is less than 2,000; set 5 samples when the number of barcodes is between 2,000 to 5,000; set 2 samples when the size is larger than 10,000. And then press “Enter” to start the construction of reference database.

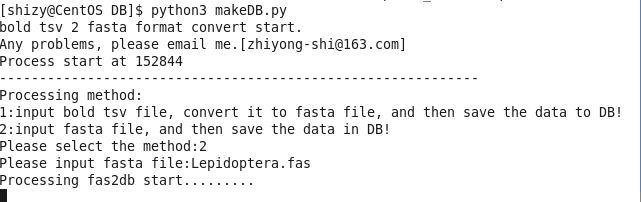


**Example analysis:**

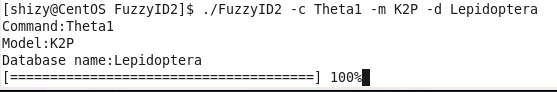
1. The reference barcodes file “Lepidoptera.fas” is uploaded to the linux server and conserved under folder “DB” in FuzzyID2 directory.

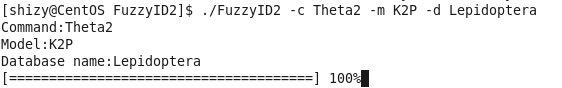


1. In “DB” directory, Python script “makeDB.py” is used to build local reference database.



1. Then, return to FuzzyID2 directory and calculate *theta1* and *theta2* by typing the following commands.

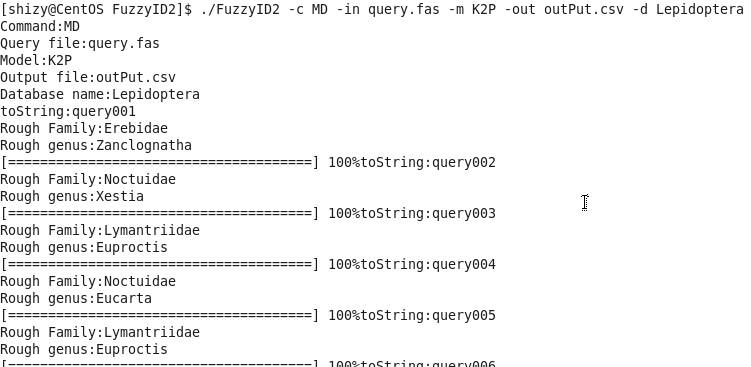




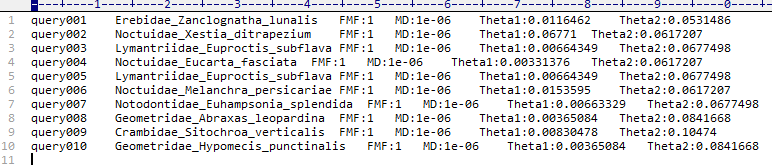
1. The query sequences file “query.fas” is uploaded to the linux server and conserved in FuzzyID2 directory.



1. Identified the query sequence by typing the following command.



1. The content of the result file “outPut.csv” is shown as following picture.



The name of these columns are: query sequence ID, assigned taxonomy of the query sequence, fuzzy membership value, minimum genetic distance between query sequence and assigned species, maximum intraspecific genetic distance of the assigned species and minimum interspecific genetic distance of the assigned species.