Manual Repeat profiles

This manual provides the necessary information to run the scripts employed for Noel *et al.* "Repetitive DNA profiles of the amphibian mitogenome". The scripts, outputs examples, and datasets can be downloaded from: https://github.com/SalmonellaIIB/Repeat_profiles.git or you can contact Noel Cabañas for technical assistance at ncabanas@iibiomedicas.unam.mx.

Requirements

You need to install Python 3.4 or superior and some python modules.

- Datetime
- Os
- Re
- Random
- Biopython

All these modules are installed by default when installing Python. You can install python using Anaconda Distribution too. Biopython is a module that should be installed after installing python since it did not is installed with the standard python installation. To install biopython easily you can use the Python package management tool pip, which is preinstalled with python. If you do not have pip tool installed you can download it from https://pip.pypa.io/en/stable/installing/

To Installing biopython run this line in the command line:

```
$ pip install biopython
```

Running an example

The scripts should run correctly after the installation of these requirements. To verify you can run the example of the analyses to repeat sequences. In the folder name example, you will find an automatized example of how to identify direct and inverted repeat sequences. Only run the script from the command line or Python interpreter and the work is done.

Step 1. Download the folder example.

The folder contains the following files:

- ds.txt: This file contains the NCBI access numbers to download the mitogenomes. It needs to be in the same folder with get_mitogenomes.py, mito_repeats.py and repeats.py.
- get_mitogenomes.py: This scrip is the automatized version. It runs download the mitogenomes and runs the repeat sequences analysis.
- mito_repeats.py: This is the main Script that identifies the DRS and IRS and generates .cvs files
- repeats.py: This Script is used by the mito_repeats.py.py script, both need to be in the same directory, otherwise you won't be able to run the scripts correctly.
- result_by_size.py: create tables where you can find the number of repeat and direct repeat sequences of different sizes for every mitogenome.

Step 2. Enter into the folder. From /example/ directory execute in command line the following line>

```
$ ./python get_mitogenomes.py
```

The script will create a folder called 'repeats', with two subfolders. The folder named 'fasta' contains the mitogenomes in fasta format. The folder called 'results' contains the direct and inverted repeat sequences identified.

** another way to run the example is opening the script get_mitogenomes.py with a python interpreter and run, which will create the same action as the line $ ./python get_mitogenomes.py.

Running your data

To run your own data, I recommend to use download the folder example and replace "ids.txt" with a file with the same name. This file must contain the NCBI access numbers of the sequences that you want to analyze. Once you already made this change run again the script.

```
$ ./python get_mitogenomes.py
```

This process will run as exactly as the example, but analyzing your data. Every time that you run the script with different data I recommend downloading again the folder example, otherwise you might generate redundancies in your files and errors can show up.

These scripts have been automatized to make your life easier. I encourage you to modify the scripts to obtain the results according to your necessities.

Finally, if you use the scripts please cite the article of Noel *et al.* "Repetitive DNA profiles of the amphibian mitogenome".