CpG Island Discovery

Specification

Due to the general simplicity of the task for finding CpG islands, the best approach seems to be a simple python command line tool controlled with command line arguments coming with well written README and well documented functionality.

Application Functionality

1. Read the DNA record from either of the following sources:
    - Online from the NCBI database based on the gene name
    - Local file
    - Command line argument
2. Use the algorithm to find CpG islands.
    - The algorithm to find CpG islands has not yet been standardized or universally accepted.
    - Some things about the algorithm are common across implementations. In short:
        - bp length >= 200
        - CG content >= 50%
        - Observed to Expected CpG >= 60%
    - The algorithm proposed in the original assignment is very simple, but it does not produce satisfying results that need to be further processed. It only finds islands of size 200 and a bigger island gets interpreted as multiple 200 bp islands offset by 1bp. It is also rather slow performance-wise.
    - Instead, I propose to use a more complex algorithm denoted as Takai and Jones' algorithm, which can find bigger islands. There exists at least one python (cython) implementation which seems very well-optimized and will likely be very fast, but it lacks the user-friendly nature of the tool and the additional features that this implementation may provide. If the application that is to be made suffers performance-wise given long records, this implementation should be examined further as it uses many neat tricks to run fast including parallel computing. It is posted under MIT licence so using its code directly or turning it into a library is also an option.
3. Output the information about found islands.
    - Print the results to the terminal.
    - Save the results to a file (csv).

Usage

The general strategy for this application is to make it simple by default only introducing the options in the README or -help option.

Example usage cases will be provided once the specific functionality is decided and implemented. For now, a placeholder example will be provided and explained.

```
python islands.py -io DQ011153.1
```

Explanation: Get the **i**nput record with the name `DQ011153.1` from the **o**nline database. The output option is not specified, defaults to printing to the screen

output:

```
 Begin  end    gcper   obs_exp win_length
189862  190062  0.505   1.176 200
189863  190063  0.505   1.176 200
189864  190164  0.51    1.153   300
...
```

Additional Features

The following features may improve the usability of the application but are not required for the operational prototype / proof-of-concept. They increase the complexity of the original simple command line tool and thus should be turned off by default.

Multi Input

Allow the user to specify multiple input records to evaluate all of them in a batch. This will be different for any input method. For NCBI online import, the user should give multiple names as arguments. For local file, the user will instead give a list of files or a directory where all records store in it will be examined. For command line input, the user should simply provide multiple records as multiple arguments.

Parallel Evaluation

Use multithreading to parallelize the task of finding CpG islands. This may happen in two layers. First, the multi input feature can be supported with this, where the parallelization of work on different genes is always possible, so multiple records may be inspected simultaneously. The second may be the parallel evaluation within each single evaluated gene, which may or may not be possible given the specific algorithm.

Other than CpG Islands

Perhaps finding "islands" formed with other components, like GpC, ApT, TpA, CpA, and so on, may be of interest. The user could specify what kind of the island they want to search for.

Visualize Output Information

The resulting list of islands in the gene may be visualized when nature of this visualisation should be discussed further.

Parametrize Algorithm Execution

Allow the user to change the parameters of the selected algorithm. For example to change the minimal island length from 200 base pairs, raise the required CG percentage and so on. As the definition of CpG islands varies greatly across the sources, this may prove to be useful to "change" the algorithm and match its results to the ones produced by a different implementation.

Installation

The installation will differ based on the application complexity. For a simple application, *python 3* will by required and a list of additional dependencies will be contained in the repository, along with a manual on how to install them. Should the number of dependencies grow large, it may be better to run the application in a virtual environment where all there dependencies will be preinstalled and the application will run in it.