CSI: Cleavage Site Investigator

Features

- Run straight from command line
- Compatible with FASTA file format (.fa and .fasta)
- Determine top and bottom strand cleavage events
- Export results to .csv files
- Create visual event distributions as heatmaps and strand linkage plots

Contents

Installation

1. Install Python (tested with Python 3.9.1)
2. Install required libraries (BioPython, Seaborn, SVGWrite and TQDM)
   - Either using Pip

```
pip install biopython==1.79
pip install tqdm==4.55.1
pip install seaborn==0.11.1
pip install svgwrite==1.4
```

   - Or using the provided Anaconda environment file ("csi.yml" in "resources" folder)

```
conda env create -f csi.yml
```

Usage

Notes

- Example files for testing CSI can be downloaded from TODO. These files are:
    - "TODO" - Template sequence (must contain one sequence)
    - "TODO" - Cassette sequence (must contain one sequence)
    - "TODO" - Consensus sequence(s) (can contain multiple sequences)
- The above files are used throughout the following code demos
- Each program (csi.py, heatmap.py and strandlinkageplot.py) can be run entirely from command line. Full argument documentation is accessible using the `-h` (or `--help`) flag (e.g. `python csi.py -h`).

Running CSI (basic)

- The main CSI program is run using csi.py. This will analyse the specified consensus sequences and optionally output event distributions, summary statistics and plots (advanced plotting options available by running heatmap.py and strandlinkageplot.py directly).
- CSI requires a minimum of three arguments, specifying paths to the cassette (`-c` or `--cass_path`), reference (`-r` or `--ref_path`) and test (`-t` or `--test_path`) files.

```
python csi.py -c TODO -r TODO -t TODO
```

- With default parameters (no optional arguments specified) a basic summary will be displayed with the following sections
    - "TS position" - Position of the top-strand cleavage event
    - "BS position" - Position of the bottom-strand cleavage event
    - "Split seq" - `True` if the cleavage event spanned the start/end of the reference sequence, `False` otherwise
    - "Count" - Number of identified events matching this cleavage event (% of total identified events shown in parenthesis)
    - "Type" - Type of cleavage event (either "Blunt end", "3′ overhang" or "5′ overhang")
- An example output is shown below

```
RESULTS:
    Full sequence frequency:

        TS position: 2503
        BS position: 2503
        Split seq:   False
        Count:       753/756 (99.6% of events)
        Type:        Blunt end
```

```
TS position: 2503
BS position: 1932
Split seq:   False
Count:       1/756 (0.1% of events)
Type:        3' overhang

TS position: 2503
BS position: 2495
Split seq:   False
Count:       1/756 (0.1% of events)
Type:        3' overhang

...
```

Running CSI (advanced)

- CSI offers optional command line parameters to specify execution settings (e.g. the number of bases to fit) as well as additional outputs (e.g. summary CSV files or rendered heatmap plots).
- Optional arguments are listed below:

Argument|Description|Default

h or help|Show help message (lists all required and optional arguments)|NA

Generating strand linkage plots directly

Generating heatmap plots directly