There are two primary ways for the user to interact with the program. First, the user is able to generate an artificial data set containing all the neccesary information for the algorithm to work with it, and second, perform the estimation using either the generated data, or the user's data set, provided it is formatted similarly.

Data generation

To run data generation, type in the command line a script of the form

```
python motif_generate.py --params [params] --output [output]
```

params

A JSON file with parameters of the model to be used when generating the data. The file has to contain at least the following objects:

- 'w': a positive integer - length of sequences in the sample.
- 'k': a positive integer - size of the sample to be generated.
- 'alpha': a floating-point number from $(0, 1)$ - the probability of a single observation following the motif distribution as opposed to the background distribution.
- 'theta': a list of lists (of length w), representing a $4 \times w$ matrix describing the motif distribution, where the element in i-th row and j-th column is the probability that on the j-th position in the sequence, i-th base is drawn.
- 'theta_b': a list of length 4, representing the background distribution - i-th term is the probability that the i-th base is drawn.

There is an example file 'params_set1.json' included with the program, it is used as a default parameters file if the argument is not specified.

output

A file location where a JSON file with the generated data set will be put. By default, if the location is left unspecified, the output will be stored as 'generated_data.json'.

Parameters estimation

To run the actual estimation algorithm, type in a script in the following format

```
python motif_estimate.py --input [input] --output [output]
```

input

A JSON file with a data set of DNA sequences. It has to contain at least the following objects:

- 'alpha': a floating-point number from $(0, 1)$ - the probability of a single observation following the motif distribution as opposed to the background distribution.
- 'X': a list of lists, representing a $k \times w$ matrix, where the rows are DNA sequences, with integers or floats from $\{1, 2, 3, 4\}$ as terms.

There is an example file 'generated_data.json' included in the files, with data generated with the same parameters as the ones specfied in the 'params_set1.json' file. It is used by default if the argument is not specified.

output

A file location where a JSON file with the estimated parameters will be placed. By default, if the argument is left unspecified, the estimated will be put in 'estimated_params.json' file.

For the theoretical details of the model and estimation method, check the 'report' PDF.