A systematic approach for dissecting the molecular mechanisms of transcriptional regulation in bacteria.

doi: 10.1101/239335

The *jupyter_notebook/* folder contains a variety of Jupyter Notebooks related to the work in the appendices. In addition, the following two Jupyter notebooks can be used to visualize the Sort-Seq data.
- Sort-Seq_promoter_data_visualization.ipynb can be used to view plots of expression shifts, information footprints, and mutation rate for each promoter.
- Sort-Seq_energy_matrix_visualization.ipynb can be used to visualize the inferred energy matrices.

All processed data is available in tidy formatted .csv files that are most conveniently viewed using pandas in Python, but can also be loaded with a spreadsheet editor such as Microsoft Excel.

All code used for analysis and plotting was written in Python. The following dependencies may be required to run the files. Note that energy matrix processing and some of the figure plotting *.py* code requires PyMC 2.3.X and requires a Python 2.X installation.
- matplotlib
- numpy
- pandas
- scipy
- seaborn
- ipython
- biopython
- pymc - corner

Sort-Seq experiments

Processed results

Processed Sort-Seq data can be found in *code/sortseq/*. Each folder contains processed Sort-Seq data (from a single sequencing run, and may contain multiple experiments for different promoters and experimental conditions). Expression shift, information footprints, mutation rates are found in files that end in summary.csv. Energy matrices are found in seperate .csv files.

Processing new data

Pipeline to calculate expression shifts, information footprints, etc:

For each Sort-Seq experiment, a configuration file is made to list the experimental details associated with it. These .cfg files are placed in the *code/sortseq/(sortseq_experiment_name)/cfg_files/* folder.

To process new Sort-Seq data (multiple quality filtered ...bin*.fasta or ...bin*.fastq files; must end with bin number as shown):
- create new folder for Sort-Seq data in the *code/sortseq/* directory.
- create .cfg file in a new folder called *cfg_files/*. Add in appropriate details and location of sequencing files (see others for example)
- In new folder, run python script in command prompt (i.e. Terminal in Mac):

```
python ../../processing/processing_seq.py cfg_files/(config_filename).cfg
```

- Once that has completed (several hours with current scripts), run the following to generate plots:

```
python ../../processing/analysis.py cfg_file/(config_filename).cfg
```

Processing .sql files from MPAthic:

The energy matrices were inferred by MCMC using the MPAthic software (doi: https://doi.org/10.1101/054676), which provided .sql files (20 MCMC runs per inference). Note that it expects a certain file naming format and may need modification for new files. In *code/sortseq/(sortseq_experiment_name)/cfg_files/*, the associated config file is edited to include several lines related to matrix identify, position, and length; also included is location of .sql files. For example:

```
emat_dir_sql = ../../../data/sortseq_MPAthic_MCMC/
# position information for each energy weight matrix model
[CRP]
TF = crp
TF_type = 1
mut_region_start = 26
mut_region_length = 26
```

To process the .sql files from each MCMC, in *code/sortseq/(Sort-Seq folder name)/* run the following command in the command prompt:

```
python ../processing_emat.py cfg_files/(name of cfg file).cfg CRP
```

DNA affinity Chromatography and Mass Spectrometry experiments

Processed results

Processed data can be found in *code/mass_spec/*. Each folder contains a summary .csv file with protein enrichment and details related to the experiment such as the DNA traget sequence.

data analysis:

Protein enrichment values are obtained from the 'ProteinGroups.txt' file that is generated by the software MaxQuant (http://www.maxquant.org), used to analyzed Thermo '.raw' LC/MS/MS data files. Within each experimental folder (contained in *code/mass_spec/*), there is a .py Python file used to extract the relevant data from the 'ProteinGroups.txt' that is found in the MaxQuant analysis txt folder. This compiles a summarized '.csv' file that also will contain addition experimental details.

Miscellaneous

code/processing/
- Scripts used to process Sort-Seq .fastq or .fasta files, and .sql files associated with energy matrix MCMC

code/analysis/
- Jupyter Notebooks and other analysis used in work.

code/flow/
- Histogram data from flow cytometry experiments to measure expression of promoter plasmids. Data is used in several figures.

code/figures/
- Contains all Python scripts used to generate figures for main text and SI material. - Note that in many cases, formatting or arrangement of figures were modified in Adobe Illustrator. - Note also that several of the SI figures require the full sequence data files. These are available upon request.

misc/plasmid_sequences/
- Promoter plasmid sequences in .gb GenBank format

misc/primers_oligo_sequences/
- Primer and oligo sequences used in work

data/mass_spec

- Contains the ProteinGroup.txt output files from MaxQuant analysis of raw Thermo LC/MS/MS data files.

data/sortseq_raw
- Contains raw sequencing files from the Sort-Seq experiments (one sorted bin per file)

data/sortseq_pymc_dump
- Contains *mar* promoter file of sequences across all sorted bins, used for library analysis in Supplemental FigS1E,F.

data/sortseq_MPAthic_MCMC
- Contains the *.sql* files obtained from running MCMC for energy matrix inference with the MPAthic software.