k-Seq Analysis

This tool calculates *A* and *k*t*, according to the equation *A(1-Exp(-k[S]t))*, for every sequence in a relevant *k*-Seq data set. It requires several input files: a file with sequence data/counts for a *k*-Seq "start" round, a file for each tested *k*-Seq round after selection has occurred, files describing the reaction conditions, and known or approximate normalization constants for each round based on the expected amount of DNA/RNA/protein present before and after the reaction in each sample.

How to use the script to calculate *k*-Seq results:

To reproduce the numerical results reported in this publication, the python script `kseq_tools_v01.py` can be run as follows:

```
python kseq_tools_v01.py start_round kseq_rounds output_file normalization_list
substrate_concs rounds_to_average rounds_to_error -v -p
```

where `start_round` is R5c-counts.txt (in Count reads files for original selection), `kseq_rounds` is example-rounds.txt, `normalization_list` is example-normalization.txt, `substrate_concs` is example-subst-concs.txt, `rounds_to_average` is example-rnds-to-avg.txt and `rounds_to_error` is example-rnds-to-err.txt (in k-seq Analysis - example input files).

The code requires "counts" files as input, consisting of three lines of metadata followed by one line per unique sequence in the pool in the following format: sequences in the first column and counts (an integer number) in the second column. Such files are produced by our Galaxy tools, currently available at the Chen Lab website. The files corresponding to the start round (R5c-counts.txt), each tested *k*-Seq round after selection (called in `kseq_rounds`), as well as every other input file can be found in this repository.

For more information on usage, read below or run `python kseq_tools_v01.py -h` in the terminal.

Advance usage information

Goal:

The kseq_tools script calculates catalytic kinetics for a population of sequences, using the *k*-Seq methodology. The script takes as input a *k*-Seq 'start round' and a list of additional rounds (each corresponding to selection under known conditions). The output consist of predicted constants *A* and *k*t* for catalysis following *[surviving fraction] = A(1-Exp(-k*[S]*t))*.

Input:

The script can be generally run as follows:

```
python kseq_tools_v01.py start_round kseq_rounds output_file normalization_list
substrate_concs rounds_to_average rounds_to_error
```

Required arguments (positionally dependent):

`start_round` - File containing sequence counts for the pre-*k*-Seq population (e.g. `R5c-counts.txt`). The code requires a "counts" file consisting of three lines of metadata followed by one line per unique sequence in the pool, in the following format: sequences in the first column and counts (an integer number) in the second column. Such files are produced by our Galaxy tools, currently available at the Chen Lab website.

`kseq_rounds` - File containing a list of filenames, each of which contains sequence counts for a post-*k*-Seq population. Each of these files must share the same format as `start_round`.

`output` - Name of the output file (e.g. `kseq-data.csv`)

`normalization_list` - List of normalization factors for each round, starting with the start round. One value per line. For the start round, should typically be set equal to *1/[total amount of DNA/RNA/protein]* present at start of *k*-Seq selection rounds, and for all other rounds should be *1/[amount of DNA/RNA/protein]* remaining after selection step. The units here will correspond to the units used in the script's output.

`substrate_concs` - File containing list of substrate concentrations for each set of kseq rounds. One row corresponds to each unique substrate concentration, not to each round/experimental sample (if duplicates are used). Must be the same number of rows as `rounds_to_averge`.

`rounds_to_average` - File containing comma-separated lists of sets of rounds to average together for fits (e.g. row 1 as "1,2,3" and row 2 as "4,5,6" will average the abundances of rounds 1, 2, and 3, and then also average 4, 5, and 6). If only a single replicate was carried out, then the file should only contain one number per round (e.g. row 1 as "1," row 2 as "2", etc.) For abundance without kseq calculations, this file should be left blank. Must be the same number of rows as `substrate_concs`.

`rounds_to_error` - File containing comma-separated lists of sets of rounds used for each replicate. If no replicates, this file should be left blank (standard deviation will be based on goodness of fit instead).

Optional arguments:

The following options can be utilized to adjust the behavior of the script. Most options require an additional argument. A comma indicates that an option can be called multiple ways.

`-h`, `--help` - Show help message and exit.

`-s SEARCH_SET [SEARCH_SET ...]`, `--search_set SEARCH_SET [SEARCH_SET ...]` - If used, this option will cause the script to only run *k*-Seq over a subset of all sequences. By default, search set is set to `all`, generating *k*-Seq data for all sequences in the start round. If set to `center CENTER_SEQUENCE DISTANCE` (requires three arguments, the second being a sequence and the third being an integer), will only generate kseq data over all sequences within a fixed distance of the center. If set to `list SEQUENCE_LIST_LOCATION`, will only generate data over the sequences listed in a file at the given location (this file should contain one sequence per row).

`-v`, `--verbose` - If this flag is used, output file will include data on sequence concentration at every kseq round.

`-o OUT_TYPE`, `--out_type OUT_TYPE` - Set output file type to `csv` (comma-separated values) or `tsv` (tab-separated values). Default output file type is csv.

`--min_count MIN_COUNT` - Minimum count of sequences searched (must be an integer). Default value is 1 (search through all sequences present in the "*k*-Seq start" round).

`-p`, `--track_progress` - If this flag is used, progress will be printed in the terminal.

Output:

The output file from this script contains all sequences for which *k*-Seq data is calculated. The first row corresponds to column headers, the second row provides the number of unique sequences in each round, and the third row provides the total numbers of sequences.

For each sequence's *k*-Seq output (i.e., each row), the values in each column provide the following information:

`Seq. Name` - The sequence identity. Same as sequence provided in input counts file.

`Sequence amount` - Displayed only if the flag `-v` is enabled. The total amount of this sequence present in the

test tube at the start of *k*-Seq. Units correspond to normalization constants (e.g. if the normalization units are *1/ng*, the units of 'sequence amount' will be in *ng*).

`Surv. fraction ROUND_NAME` - Displayed only if `-v` is enabled. The fraction of this sequence that survives selection to be sequenced in sample `ROUND_NAME`. Should be < 1 for all sequences in all rounds.

`A by avg` - The value of *A* calculated from fitting averaged data sets.

`k*t by avg` - The value of *k*t* calculated from fitting averaged data sets.

`A st. dev` - The standard deviation of *A* calculated from the values fit to multiple replicates.

`k*t st. dev` - The standard deviation of *k*t* calculated from values fit to multiple replicates.

`distance from ct` - The distance of this sequence from the peak center (only if `SEARCH_SET` is `center`).