Framework for functional profile prediction
User Guide

Requirements:
To run this tool you have to have Python 3 installed. If you want to use the bash script intended for trying out the tool, you also have to have bash command line.

Input:
A table with OTU identifiers and abundances in the given sample, closed-reference picked against Greengenes, format .tab. *(Provided in the data/test directory as sample*.tab)*

Output:
KO profile table. *(Expected output to compare accuracy provided in the data/test directory as expected*.tab)*

Running the tool:
The easiest way to run the tool is from the command line from folder framework:

```
$ ./test.sh python-alias [-m method-name] [-i input-filename] [-o output-filename]
```

Running the script with `-h` parameter will show help, including description of parameters and available methods for functional prediction.

Available settings:
If you want to change other parameters than input, output and method, you need to change them in file settings.py. *(For the purpose of trying out this tool, you do not have to change anything, default settings will be used)*

Following is the description of all parameters that can be changed in settings.py.
Global settings (same for each method):

- **KO_PROFILE_FILENAME** - file with reference ko profiles
- **INPUT_FILENAME** - input sample
- **UNKNOWN_METHOD** - method to use for prediction
- **RESULT_FILENAME** - result file
- **COUNT_16S_FILENAME** - file with 16s counts

Methods for prediction and specific settings:

- alignment_simple
    - **SIMILAR** - how many similar sequences should be averaged
    - **SCORE_FILENAME** - distance matrix file
- random

    - **SIMILAR** - how many similar sequences should be averaged
- tree
    - **TREE_FILE** - file with phylogenetic tree
- treshold
    - **TRESHOLD** - similarity treshold in percent
- weighted
    - **MAX_DIFF** - maximal difference to consider sequences as similar
    - **SCORE_FILENAME** - distance matrix file
    - **SIMILAR** - how many similar sequences should be averaged
- regression_by_sequence
    - **RESTS** - how much of the reference table is known