User manual: MotifQuest (Updated 4/29/2024)

About MotifQuest:
MotifQuest is an open-source Python-based algorithm designed for identification of motifs from endogenous peptide databases.

License:
MotifQuest is freely available for download from GitHub (https://github.com/lingjunli-research/MotifQuest).

Computational requirements:

MotifQuest is only supported for use with Windows OS, though limited functionality may be available with MacOS and Linux/UNIX. MotifQuest is capable of running on a standard personal computer. For design purposes, the program was validated on a Windows Desktop computer with the following specifications:

Processor: Intel(R) Xeon(R) CPU E5-1607 v3 @ 3.10GHz

Installed RAM: 128 GB

System Type: 64-bit OS

System: Windows 10

Setup:

The MotifQuest Python code can be downloaded from the associated GitHub repository (https://github.com/lingjunli-research/MotifQuest). To run this, command line knowledge or familiarity with a python IDE is required. While we recommend running in a virtual environment in Anaconda, any command line should be operational. For successful operation, the following packages are required, some of which come standard with Python:

- Bio
- Pandas
- CSV
- OS
- Subprocess
- Numpy
- Re
- Math
- Scipy

MotifQuest also relies on the Clustal Omega package. This package, along with all installation instructions, can be installed here: http://www.clustal.org/omega/#Download. Upon running, if a `[WinError 2]` populates, this is likely a result of incomplete or incorrect installation of Clustal Omega.

Usage:

To run MotifQuest, first input all variables. These variables, along with their lines within the code, are detailed below.

The primary script for MotifQuest is `MotifQuest_code.py`. Following this, the `partial_motif_corrector.py` code can be run, which will make all partial motifs compatible with EndoGenius. This script is optional if not running with EndoGenius. This script also combines full and partial motif files into a single, complete, motif database.

Phase I: `MotifQuest_code.py`
Lines to modify
Line 21: `output_folder_path`
- Replace with desired output path. This path must also contain a copy of your FASTA database.

Line 22: `t_val`
- Replace with desired T-value. This is a list variable, and multiple values can be input. *Recommended value is* `[0,0.4]` *to account for optimized values for full and partial motifs.*

Line 23: `min_motif_len`
- Replace with desired minimum length for a motif (integer). *Recommended value is 2.*

Line 24: `motif_instances`
- Replace with desired minimum number of instances a motif must appear within the FASTA database (integer). *Recommended value is 2.*

Line 25: `minimum_flank_size`
- Replace with desired minimum number amino acids that must flank a variable amino acid in a partial motif (integer). *Recommended value is 1.*

Line 26: `in_file`
- Replace with path to FASTA database path.

Phase II: partial_motif_corrector.py
Lines to modify
Line 12: `partial_motif_path`
- Replace with path to the partial motif database generated in **Phase I**

Line 13: `full_motif_path`
- Replace with path to the full motif database generated in **Phase I**

Line 14: `fasta_path`
- Replace with path to FASTA database used in **Phase I**

Line 15: `output_path`
- Replace with desired output path.

Line 16: `min_len`
- Replace with desired minimum length for a motif (integer). *Recommended value is 2.*

Output files:

For a standard run, exports will include:
- Distance_matrix.txt: contains the ClustalW distance matrix generated from the aligned FASTA files
- Full_motif.csv: export of all full motif outputs from FASTA database
- Partial_motif.csv: export of all partial motif outputs from FASTA database
- Full_motif_report.csv: export of all full motifs, score, T-value, and sequences that the full motif was found in from FASTA database
- Partial_motif_report.csv: export of all full motifs, score, T-value, and sequences that the partial motif was found in from FASTA database

Other issues:

This document will be routinely updated as user issues arise. If your issue is not displayed in this document, feel free to either contact the current maintainer (tcdang@wisc.edu), or open an issue on the GitHub page for this program (https://github.com/lingjunli-research/MotifQuest/issues).