User Guide for the analysis of amino acids surrounding unmodified and UbKEKSylated lysines in human proteome.

This package contains 2 short codes which can be run in any software or application supporting Python language:

- **Analysis_sequences_from_diGly_IP.py** : Measure the number of appearance for each amino acids depending on the position for known sequences measuring exactly 31 residues.
- **Analysis_lysines_from_protein_database.py** : Measure the number of appearance for each amino acids depending on the position for full protein sequences gathered in a FASTA format.

System requirements:

Operating systems: Works on Windows 10 (or higher) and on macOS Catalina 10.15.6 (or higher), when using any software capable of running short Python scripts (Anaconda, Jupyter, IDLE, etc…).

This script version has been tested on: Two data sets with different sizes. The small data set is composed of 655 sequences (each one being exactly 31 residues long) and contains 655 lysine environments . The large data set features 20 386 proteins sequences under a FASTA format and contains 651 306 lysine environments.

Additional software (optional): To obtain a list of sequences containing exactly 31 residues and study the UbKEKSylation sites in the human proteome, we identified modified lysines via the MaxQuant software (version 1.6.17.0). By doing so, we obtained an Excel file containing the sequence from position -15 to +15 surrounding each modified lysine.

Installation guide:

If a software or application capable of running short Python script is already available on your computer, no installation is required. Just run the script.

Demo:

Instruction to run on demo data: To run our data, be sure to place furnished demo files in the same directory as the Python script. Next, run the script !

Expected output: A matrix in a format 21 x 31 is expected as a final output. Each line represents a position in the sequence. Line 16 is expected to be only lysine residues (all amino acids except lysine should present a value equal to 0. Each column indicates which amino acid is considered (in the following order A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, X ; where X corresponds to a gap).

Exemples of expected output with our 2 data sets is available in this package. Please note that small intermediary files may appear, to allow you to access every step of the analysis.

Expected run time for demo files: For our small demo data set (655 lysine environments), output was obtained instantly. For our large demo data set (651 306 lysine environments), output was obtained within 15 seconds of run time.

Instructions for use:

How to run the script on your data: Please follow those simple steps:

1. Register your data under a .txt file and be sure to place this file in the same directory as our Python script. Partial sequence of 31 residues should be organized as our demo file (1 line = 1 sequence) for the script Analysis_sequences_from_diGly_IP.py. Full proteins sequences should be under a FASTA format for the script Analysis_lysines_from_protein_database.py.
2. Depending on the chosen script, replace the name of the text file to open by the name of your data in line 59 (for Analysis_sequences_from_diGly_IP.py) or line 11 (for Analysis_lysines_from_protein_database.py).
3. Run the script. Run time may vary depending on the size of your data set.
4. The script will generate a matrix with your results. Each line represents a position in the sequence. Line 16 is expected to be only lysine residues (all amino acids excepte lysine should presented a value equal to 0. Each column indicates which amino acid is found (in the following order A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, X ; where X corresponds to a gap).

Reproduction instructions: Files used for our published analysis were provided as demo files. Please refer to the demo paragraph if you wish to reproduce our results.

CREATORS: LAB BOISVERT, UNIVERSITY OF SHERBROOKE (QC, CANADA)