Module `search_proteome.py`

Given a UniProt proteome or FASTA file and a queried peptide motif, the Python script performs the following:

1. counts the frequency of said motif
2. computes the expected frequency of the motif based on the amino-acid fractions in the input proteome
3. compares the frequency of the motif in structured and disordered regions (if a list of disordered regions is supplied)
4. identifies proteins that contain the queried motif
5. and performs statistical tests to assess the significance of the observed frequency of the motif.

Index

```
class Proteome
```

Search for motifs within proteomes and compare the observed vs. expected motif frequencies.

Methods

```
expected_motifs(self, query_proteome, query_motif)
```

Calculate the expected number of a queried motif based on amino acid frequencies in the proteome.

Parameters

query_proteome : pandas.DataFrame, the proteome of interest.

query_motif : str or list, the motif(s) of interest.

Returns

final_dataframe : pandas.DataFrame, expected fraction and count of the motif.

```
find_motifs(self, data_frame, query)
```

Search a proteome for a specific amino acid or motif.

Parameters

data_frame : pandas.DataFrame, contains the proteome of interest

query : str, the amino acid or motif to be searched

Returns

sum : int, total number of hits for the queried residue/motif

fraction : float, hits (sum) divided by the total residues or motifs of the same size

```
find_flex_motifs(self, data_frame, list_of_motifs)
```

Search a proteome for a list of motifs.

Parameters

data_frame : pandas.DataFrame, the proteome of interest

list_of_motifs : list of str, the amino acids or motifs to be searched.

Returns

counted_motifs : pandas.DataFrame, the queried motifs and number of instances of each motif.

find_proteins(self, data_frame, query, proteome_type)

Search for proteins in a proteome that contain the queried motif.

Parameters

data_frame : pandas.DataFrame, the proteome of interest

query : str or list of strings, the amino acid or motifs to be searched
(e.g. 'IPV' or ['IPI', 'IPV', 'VPI', 'VPV'])

proteome_type : str, either 'FASTA' or 'UniProt' depending on the proteome

Returns

found_proteins : list, array of UniProt/FASTA IDs that contain the query

iterate_motif(self, query)

Generate a list of all possible motifs that are consistent with an input motif

Note: the "_" symbol separates positions in the motif, "X" designates any residue, and multiple residues within a given motif position indicate "or".

Example: MILV_X_DE means Met, Ile, Leu, or Val followed by any residue followed by Asp or Glu.

Parameters

query : str, the amino acid motif to be queried
use "_" to separate positions in the motif (e.g. I/V-X-I/V would be "IV_X_IV")

Returns

all_possibilities : list of str, all variations of the inputted motif
(e.g. IV_X_IV yields ['IAI', 'ICI', 'IDI', … , 'VVV', 'VWV', 'VYV']

load_fasta(self, fasta)

Read a FASTA file with identifiers and sequences.

Parameters

fasta : file, standard FASTA format with identifiers (e.g. >) and sequences

Returns

fasta_proteome : pandas.DataFrame, the FASTA IDs, sequences, and length of sequences

load_uniprot(self, uniprot)

Read a UniProt proteome file.

Note that the input proteome file should contain the following column order:
(1) Entry (2) Entry name (3) Status (4) Protein names (5) Gene names (6) Organism (7) Length (8) Sequence

Parameters

uniprot : file downloaded from UniProt containing a proteome of interest

Returns

uniprot_proteome : pandas.DataFrame, FASTA IDs, sequences, and sequence length

proteome_AA_fractions(self, query_proteome)

Calculate the fractional amino acid composition of a proteome.

Parameters

query_proteome : pandas.DataFrame, the proteome of interest

Returns

AA_fractions : pandas.DataFrame, the amino acid frequencies in the proteome

proteome_length(self, query_proteome)

Calculate the total number of amino acids in a proteome.

Parameters

query_proteome : pandas.DataFrame, the proteome of interest

Returns

total_length : int, the total number of residues in the queried proteome

---