Protein Secondary Structure Prediction

This code is used to predict the secondary structure of a protein based on its amino acid sequence. The secondary structure refers to the local conformation of the protein chain, specifically identifying regions that form alpha-helices or beta-strands.

Input

The input to the code is the amino acid sequence of the protein. The sequence is represented by a string variable named `seq`. It should be provided as input before running the code.

Amino Acid Representation

Each amino acid in the sequence is represented by a single-letter symbol, following the standard convention. Here are the symbols used for some amino acids:

- Alanine (Ala) - A
- Cysteine (Cys) - C
- Aspartic Acid (Asp) - D
- Glutamic Acid (Glu) - E
- Phenylalanine (Phe) - F
- Glycine (Gly) - G
- Histidine (His) - H
- Isoleucine (Ile) - I
- Lysine (Lys) - K
- Leucine (Leu) - L
- Methionine (Met) - M
- Asparagine (Asn) - N
- Proline (Pro) - P
- Glutamine (Gln) - Q
- Arginine (Arg) - R
- Serine (Ser) - S
- Threonine (Thr) - T
- Valine (Val) - V
- Tryptophan (Trp) - W

- Tyrosine (Tyr) - Y

Propensity Values

Propensity values are pre-defined scores indicating the likelihood of an amino acid being part of an alpha-helix or beta-strand. The code provides two dictionaries: p_helix and p_strand.

- p_helix: This dictionary stores the propensity values for each amino acid residue for alpha-helix formation.
- p_strand: This dictionary stores the propensity values for each amino acid residue for beta-strand formation.

Function Definitions

The code includes several function definitions to perform the secondary structure prediction:

1. extend(start, end, structure, seq, p_helix, p_strand, helix, strands)

This function is responsible for extending helices and strands to the left and right. It takes the start and end indices of a structure (helix or strand), the structure type, the sequence, the propensity dictionaries, and the helix and strand lists as inputs. It extends the structure in both directions based on specific criteria and updates the corresponding lists.

2. print_secondary_structure(seq, secondary_structure)

This function prints the secondary structure of the protein based on the input sequence and the predicted secondary structure. It takes the sequence and the secondary structure as inputs and prints them in a readable format.

3. finding_overlaps(seq, helix, strands, p_helix, p_strand)

This function determines the overlaps between helices and strands and assigns the corresponding secondary structure to each residue. It takes the sequence, helix and strand lists, and the propensity dictionaries as inputs. It returns the final secondary structure as a list.

Main Code Flow

The code initializes the helix and strands lists with the letter 'Z', representing the absence of a helix or strand at each position. The code iterates through the sequence using a sliding window and identifies helix nucleation sites. If a window contains four or more helix-forming amino acids, the entire window is marked as a helix. The extend function is then called to extend the helix to the left and right. Similarly, the code identifies strand nucleation sites by using a sliding window. If a window contains three or more strand-forming amino acids, the window is marked as a strand. The extend function is called to extend the strand. The finding_overlaps function is called to determine the overlaps between helices and strands and assign the secondary structure (H for helix, S for strand) to each residue. The print_secondary_structure function is called to print the final secondary structure prediction based on the input sequence and the assigned secondary structure.