Welcome to Your First Step of the Challenge!

Total Points: 100

In this challenge, we will focus on pre-processing the data for protein classification. The dataset provided is `metadata_org.csv`.

This challenge is divided into two parts that should be addressed in the Notebook:

1. Coding: Focused on addressing the challenge of data analysis and preprocessing. (60 points)
2. Questions: Designed to explore and address the subcategories within the challenge. (40 points)

Data Overview:

The dataset contains the following four columns:

1. Entry
2. Sequence
3. Selected_PDB
4. proteinClass

Key expectations for the data:

- Entry, Sequence, and Selected_PDB should contain unique entries.
- The proteinClass column is expected to have 10 unique classes.

Questions to Explore

1. Unknown Amino Acids

- Why is it good practice to replace unknown amino acids? (2 points)
- Why does this notebook use alanine to replace unknown amino acids? (1 point)
- What other amino acids could substitute unknown amino acids? (2 points)

2. Add Sequence Length

- Why calculate sequence length? (1 point)
- How to verify if `df_seq` contains the necessary data? (1 point)
- How does sequence length influence protein stability and function? (2 points)
- Coding section (5 points)

3. All Possible Dipeptide Frequencies

- What is a dipeptide? (2 points)
- What type of bond is formed between two amino acids in a dipeptide? (1 point)
- Why focus on dipeptides instead of tripeptides or polypeptides? (2 points)
- How might dipeptide frequency contribute to machine learning-based protein classification? (2 points)
- Coding section (11 points)

4. Reduced Amino Acid Alphabet Frequencies

- Why is reducing the amino acid alphabet beneficial in bioinformatics? (2 points)
- What criteria are used to group amino acids? (e.g., hydrophobicity, charge, size) (2 points)
- Impact of reduced alphabet on sequence alignment and phylogenetic analysis? (2 points)
- Coding section (11 points)

5. N-Gram Profiles of Reduced Amino Acid Alphabet

- What are N-grams in the context of protein sequences? (2 points)
- Why are N-grams helpful in studying protein sequences? (2 points)
- How does increasing the N-value affect N-gram complexity? (2 points)

- Why use RAAA-based N-grams instead of the full 20 amino acid set? (2 points)
- Applications of N-gram profiles in protein classification into functional families? (2 points)
- Coding section (11 points)

6. Add Protein Properties Using BioPython

- What is the `ProteinAnalysis` class in BioPython used for? (1 point)
- How does secondary structure composition influence protein stability and interactions? (2 points)
- How can physicochemical properties aid in training classification models? (1 point)
- Coding section (11 points)

7. Save the Cleaned Data

- How to ensure no data loss during saving and reloading? (1 point)
- Differences between train-test split and cross-validation? (1 point)
- When to apply standardization or min-max scaling to properties? (2 points)
- Coding section (11 points)