CS4710 - Programming Assignment 3

This assignment focuses on the problem of protein secondary structure prediction. You will using an existing web tool to predict secondary structures, i.e. helices and strands. The goal of the assignment is to evaluate the performance of such tool. In the following I will be referring to ***stride***, but you may use any other prediction tool. The accuracy of the prediction of *stride* will be evaluated based on the fraction of correctly predicted amino acids on real protein data, for which the actual secondary structures are available.

Evaluating the accuracy of Stride

The classification accuracy of the Stride program will be computed based on the fraction at which the three different secondary structure elements were correctly predicted.

Given the PDB file of a protein P, we consider the secondary structures assignment in PDB file as gold standard. Let N_H *and* N_E be the number of amino acids that are part of helices and beta-strands as reported in the PDB file, respectively. We evaluate the results of a prediction tool based on this gold standard. Let P_H *and* P_E be the number of amino acids that the web tool has **correctly predicted** to be part of helices and beta-strands over protein P, respectively.

NOTE. P_H *(* P_E*)* *is not the number of all amino acids predicted to be on a secondary structure helix by stride, but of the subset consisting of those amino acids that are also assigned to a secondary structure helix (strand) in the PDB file.*

Compute and report the following quantities:

- $R_H = P_H/N_H$
- $R_E = P_E/N_E$
- $Q_3 = (P_H + P_E)/(N_H + N_E)$

The first two quantities correspond to per-secondary-structure-element classification performance measures, whereas the last quantity gives an overall measure of the performance.

Your program

Input: A set of protein names contained in a file. Each protein is written at a different line of the file.

Your program does the following:

I. Reads pdb files in a designated folder (will be provided).
II. For each protein it
    1. Calls the external program stride
    2. Parses the corresponding stride output and computes the quantities
        a. $R_H = P_H/N_H$
        b. $R_E = P_E/N_E$
        c. $Q_3 = (P_H + P_E)/(N_H + N_E)$
III. Computes the average value of each of the above quantities over all proteins.

YOUR OUPUT

For each input protein print the following on separate lines

- The name of the protein
- A simplified output of stride, with only the SEQ and STR lines
- The quantities R_H , R_E *and* Q_3

Print the average values of $\mathrm{R_H}$ $\mathrm{R_E}$ and Q_3 over all proteins

An example output

1atp

```
SEQ  1    VKEFLAKAKEDFLKKWETPSQNTAQLDQFDRIKTLGTGSFGRVMLVKHKE   50    1ATP
STR       HHHHHHHHHHHHHHHHH        GGGEEEEEEEEETTTTEEEEEEETT         1ATP
SEQ  51   SGNHYAMKILDKQKVVKLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNS  100    1ATP
STR       TTEEEEEEEEEHHHHHHH HHHHHHHHHHHHHH  TTTB  EEEEEEETT         1ATP
```

R_H = …

..

avgR_H

Implementation details:

See the textbook "Beginning Perl for Bioinformatics - Chapter 11 Section: The Stride Secondary Structure Predictor " for the use of an external program to calculate the secondary structure from the 3D coordinates of a PDB file.

IMPORTANT:

1. You are required to use "hashes" in your perl program. For a given protein, the hash entries will contain the 3 quantities above, $\mathrm{R_H}$ $\mathrm{R_E}$ and Q_3. Points will be taken off if you do not use hashes.
2. Treat "H", "G", "I" as helix (see later)

Output format of STRIDE

Excerpts from the STRIDE DOCUMENTATION

```
STRIDE produces output that is easily readable both visually and with
computer programs. The side effect of this conveniency is larger file size
of individual STRIDE entries. Every record is 79 symbols long and has the
following general format:


Position          Description

       1-3        Record code
       4-5        Not used
       6-73       Data
       74-75      Not used
       75-79      Four letter PDB code (if available)

 Below follows the description of each record type.
```

Code Description and format of data
REM Remarks and blank lines Format: free
HDR Header. Protein name, date of file creation and PDB code
SEQ Amino acid sequence

Format: 6-9 First residue PDB number
11-60 Sequence
62-65 Last residue PDB number

STR Secondary structure summary

Format: 11-60 Secondary structure assignment
One-letter secondary structure code
H Alpha helix
G 3-10 helix
I PI-helix
E Extended conformation (NOTE: in the PDB file this is a beta SHEET)
B or b Isolated bridge
T Turn
C Coil (none of the above)

For each record (data line) except those with codes REM and STR the number of fields is consistent and is readily suitable for processing with external tools, such as awk, perl, etc.