**Using the same rules as applied for previous homeworks. You will have over two weeks to complete this homework, but do not wait to start. This homework has been lightened with the intent of assigning Homework 5 *before* Homework 4 is due.**

You could use a Hidden Markov Model (HMM) to predict secondary structure (the hidden state) from the amino acid sequence (the observed state). The HMM model assumes a Markov chain model from the unobserved secondary structure sequence. In this homework, we want to see how appropriate the Markov assumption is for secondary structure. To do so, we will study protein sequences with known secondary structure.

There are seven secondary structural states (from the DSSP website):

**H** = $\alpha$-helix

**B** = residue in isolated $\beta$-bridge

**E** = extended strand, participates in $\beta$ ladder

**G** = 3-helix (310 helix)

**I** = 5 helix ($\pi$-helix)

**T** = hydrogen bonded turn

**S** = bend

Further, there is one unstructured/loop/missing data state, which I assign the character 'C' in the data. This state is described on the DSSP website as such:

> A blank in the DSSP secondary structure determination stands for loop or irregular. Loops and irregular elements are often, very incorrectly, called "random coil" or "coil". Many programs, including the PDBFINDER, replace this blank by a C (doing undue justice to the historical artefactual naming of loops and irregular stretches) because one never knows if a blank means loop or no-output, or something-went-wrong.

You will be working with a secondary structure dataset that was pulled from PDB (`https://www.rcsb.org/pdb/static.do?p=download/http/index.html#ss`) and cleaned into the file `pdb.fa` with the following form.

```
>name1
HHHHHHHHHHHHHHHGGGHHHHHHHHHHHHHHHHCGGGGGGCTTTTTCCSHHHHHHCHHHHHHHHHHH
>name2
HHHHHHHHHCCEEEEEECTTSCEEEETTEEEESSSCTTTHH
```

The "`>name`" line holds the PDB name of the protein. The following line contains the sequence of known secondary structure states for this protein.

Here are R libraries, including a custom library from Zeb, that can help you with this homework. Note, you will need to have administrative privileges to install Zeb's library, or appropriately change the local library path as in the following code.

```
local.libpath <- "/home/kdorman/local/R/library"
withr::with_libpaths(new = local.libpath,
        devtools::install_github('arendsee/zwc', quiet = T))
require(magrittr, quietly = T, warn.conflicts = F)
```

```
require(dplyr, quietly = T, warn.conflicts = F)
require(reshape2, quietly = T, warn.conflicts = F)
require(readr, quietly = T, warn.conflicts = F)
require(stringr, quietly = T, warn.conflicts = F)
require(zwc, lib.loc = local.libpath, quietly = T, warn.conflicts = F)
```

# Part I: Markov chains and secondary structure prediction

1. **A first order Markov model.** If you are using R, you can get word counts very quickly with the `zwc` program. For example:

```
zwc::fasta_wc(k=1, 'test.fa')

##   word count
## 1    I    33
## 2    E 48389
## 3    G  8545
## 4    S 20770
## 5    C 43370
## 6    B  2794
## 7    T 26393
## 8    H 76765
```

   A small file, `test.fa`, is provided for testing your code. These data consist of 1000 sequences randomly drawn from the total dataset of 385,460 sequences.

   (a) Model secondary structure with a first order Markov chain. Estimate the transition matrix and vector of initial state probabilities using the `pdb.fa` data.

   (b) Given your estimated first-order model, what is the probability of observing the sequence "CHC"?

2. **Extend your model to an $m$-order Markov model.**

   (a) Model secondary structure with a 2nd-order Markov chain. Show the transition matrix and vector of initial probabilities.

   (b) What is the probability of CHC with the estimated second order model?

3. **Estimating Markov chain order.**

   (a) Simulate 1000 structures of the same length of the sequences in `test.fa` from the fitted model assuming order 2.

   (b) Use the likelihood ratio test to find the optimal order for the simulated data of Part (a).

   (c) Scale up. Use the likelihood ratio test to find the optimal order for the data in `pdb.fa`.

4. **Analysis.** Is there evidence of dependence between secondary structure states along a protein sequence? Over how many positions does the dependence extend? Is the Markov chain an appropriate model for secondary structure along a protein sequence? How else might you verify the appropriateness of the Markov model for this application?