Project 4

GOR III

Step 1 - parser

- parse file CATH_info.txt
  - first four characters correspond to the file name
  - the fifth character to the chain that should be used

- parse files in dssp directory
  - only take those chains indicated in the fifth character in CATH_info.txt
  - check columns 3 to 5 to recuperate sequence, AA and corresponding class
  - regroup class information
    - H, G, I -> H
    - E, B -> E
    - T -> T
    - C, S, “ “ -> C

Step 2 - GOR III

- general formula

$$I(\Delta S; R) = I(S; R) - I(n\text{-}S; R) = \log(f_{S,R}/f_{n\text{-}S,R}) + \log(f_{n\text{-}S}/f_S) \quad (3)$$

- GOR III adaptation

$$I(\Delta S_j; R_1, \ldots, R_n) \approx I(\Delta S_j; R_j) + \Sigma_{m,m\neq 0}\, I(\Delta S_j; R_{j+m}|R_j) \quad (8)$$

$$I(\Delta S_j; R_{j+m}|R_j) = \log(f_{S_j,R_{j+m},R_j}/f_{n\text{-}S_j,R_{j+m},R_j}) + \log(f_{n\text{-}S_j,R_j}/f_{S_j,R_j}) \quad (9)$$

- count the following
  - frequency of the structure (f_S)
  - frequency of the pair (structure, AA: f_{S, R})
  - frequency of the triplet (structure, AA neighbor, AA: f_{S,Rm,R)
    - neighborhood of 8 AA to the left and 8 aa to the right
- use the frequencies to compute predictions: that conformation S with the highest value in equation (8) will be the predicted conformation

Step 3 - Quality of predictions

- compute Q3 and MCC
  - **Q3**: number of correctly predicted residues/total number of residues
  - **MCC**

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

- visualize your prediction quality using ROC curve(s)