

Projet 3

GOR III

01/12/17

Parser les données

- Parse CATH_info.txt pour obtenir le nom du fichier et la chaine à utiliser
- Parse les fichiers de dssp
 - Uniquement la chaine indiquée dans CATH_info.txt
 - Colonnes 3 à 5 pour la chaine, le résidu et la structure secondaire correspondante
 - H, G et I ➔ Hélice (H)
 - E et B ➔ Feuillêt/brin bêta (E)
 - T, C, S, “ “ ➔ Coude/non structuré (C)
- Format :
 - > identifier|protein name|organism
 - MTAEPSIVARSNFNVCRLPGTPEAICATYTGSIIPGATSPGDYAN
 - CCEECCCCHHHHHHHHHCCCCCHHHHHHHHCCEECCCCCCHHHCC
 - > ...

GOR III

- Structures secondaires dépendent des acides aminés, mais aussi du voisinage de ces acides aminés
- Probabilité pour qu'un acide aminé R adopte une structure S

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

Avec f_S la fréquence de la structure, $f_{S,R}$ la fréquence de l'acide aminé R dans une structure S et les correspondantes fréquences pour les autres structures que S pour f_{n-S} et $f_{n-S,R}$

GOR III

- Probabilité tenant en compte le voisinage

$$I(\Delta S_j; R_1, \dots, R_n) \approx I(\Delta S_j; R_j) + \sum_{m, m \neq 0} I(\Delta S_j; R_{j+m} | R_j)$$

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

Avec f_{S_j, R_{j+m}, R_j} la fréquence d'observer R_{j+m} et R_j dans une structure S et les fréquences correspondantes pour les structures non S avec f_{n-S_j, R_{j+m}, R_j} et avec $-8 \leq m \leq 8$ and $m \neq 0$

- La première équation vous donne les probabilités pour calculer les prédictions → la conformation avec la plus grande valeur est la structure prédite

Quality of predictions

- Q3: nombre de résidus avec la structure prédite correcte / nombre total de résidus
- MCC :

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

**“Think twice,
code once.”**

- ANONYMOUS

Date des séances d'aide :

- 08/12/17
- 15/12/17

Remise le 18/12/17 à 12h !