

# Mini projet 2 : Analyse statistique d'une famille de protéines

Fabien Tang  
Valentin Colliard

2018

## 1.1 Introduction

Lors de ce projet, nous avons cherché à analyser statistiquement une famille de protéines donnée par un alignement de séquence en détectant :

- les positions conservées,
- les séquences appartenant à la même famille,
- les corrélations entre les différentes colonnes de l'alignement et leur relation avec les distances entre acides aminés.

## 1.2 Données

- Dtrain.txt contient M=5643 séquences de protéines d'une même famille. Chaque séquence a pour longueur L= 48 positions et chaque acide aminé appartient à  $A = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, -\}$ .
- testseq.txt contient une séquence de longueur N = 114.
- distances.txt contient les distances entre paires d'acides aminés sous forme position 1, position 2, aa.

## 1.3 Modélisation par PSWM

### 1.3.1 Matrice de poids spécifiques des positions

Dans un premier temps, nous avons chargé nos données dans python grâce aux fonctions de lecture de fichier puis nous avons défini les fonctions  $n(i,a,liste)$  et  $w(i,a,liste)$  permettant respectivement de compter le nombre d'occurrences et de calculer le poids d'un acide aminé  $a$  à une position  $i$  donnée.

Les fonctions  $n\_global(liste)$  et  $w\_global(liste)$  utiliseront les fonctions précédentes pour la création des différentes matrices pour chaque position  $i = 0, \dots, L-1$  et chaque acide aminé  $a$  appartenant à  $A$ .

### 1.3.2 Conservation

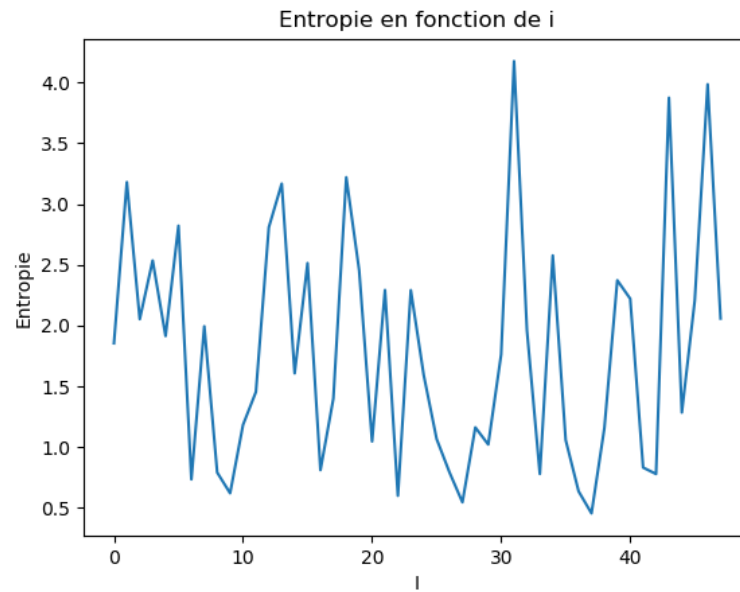
Une fois les différentes matrices obtenues, nous avons codé la fonction  $s(i,liste)$  calculant l'entropie relative en fonction d'une position  $i$ .

Cette fonction est appelée par la fonction  $s\_global\_trie(liste)$  afin de déterminer les différentes positions qui ont un poids très élevé pour un acide aminé.

$ai(liste)$  se charge de déterminer les 3 acides aminés les plus conservés.

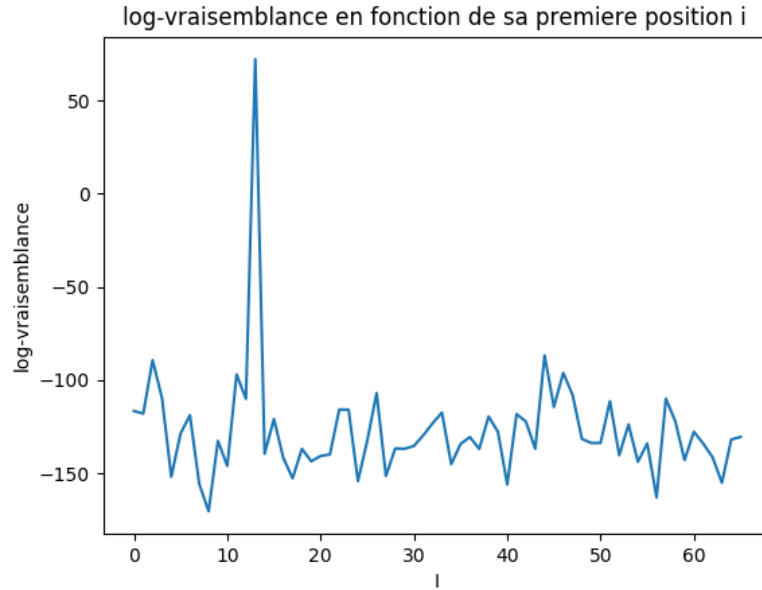
Les trois positions et acides aminés les plus conservés :

(31, 4.176827384837058), (46, 3.9862728301572674), (43, 3.875513964134912)  
'W', 'P', 'G'



### 1.3.3 Evaluer une nouvelle séquence

Afin de décider si une nouvelle séquence  $b$  fait partie de la même famille d'une protéine, nous avons calculé la probabilité...



## 1.4 Coévolution de résidues en contact

Dans cette seconde partie, nous avons amélioré notre modèle en permettant la détection de corrélations entre les occurrences des acides aminés pour 2 positions données. La fonction eq10 et eq11 calculent ainsi le nombre de séquences et le poids pour 2 acides aminés aux positions  $i$  et  $j$ .

Ces deux fonctions nous permettent donc de calculer l'information mutuelle  $M_{ij}$  (eq12) pour chaque paire de positions.  $0 \leq i < j \leq L - 1$

## 1.5 Conclusion