

# 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 définis 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'une acide aminée a à une position i donné.

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

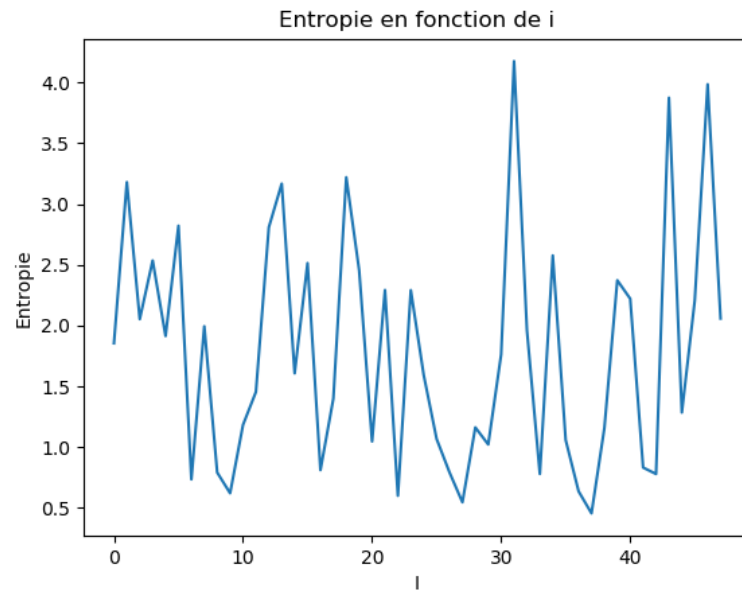
### 1.3.2 Conservation

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

Cette fonction est appelé par la fonction  $s\_global\_trie(liste)$  afin de déterminer les différentes positions qui ont un poids très élevé pour une acide aminée.  $ai(liste)$  se charge de déterminer les 3 acides aminés les plus conservés.

Les trois positions plus conservées et les acides aminés conservés :

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



### 1.3.3 Evaluer une nouvelle séquence

## 1.4 Conclusion