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 colonnes différentes de l'alignement et leur relation avec les distances entres acides aminées dans la structure 3D d'une protéine representative de la famille.

1.2 Données

- Dtrain.txt contenant 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 contenant une séquence de longueur N = 114.
- distances.txt contenant les distances entre paires d'acides aminées sous la forme (position 1, position 2, aa)

1.3 Modélisation par PSWM

Dans un premier temps, nous avons chargé nos données dans python grâce aux fonctionss de lecture de fichier puis nous avons définis une fonction n(i,a,liste) permettant de compter le nombre d'occurences d'acide aminée a à une position i donné. Suite à cela,

1.4 Conclusion