Rapport 3I005
Statistique en Bio-informatique :
Analyse statique d'une famille de protéines

MENG Fanshuo :3403051
XIANG Yangfan :3300401

I. Préliminaire

Une famille de protéine est stockée comme une matrice de taille $M \times L$ avec comme valeur dans chaque case une lettre d'alphabet qui appartient à
$\mathcal{A} = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, -\}$.
Les séquences de protéines sont stockées comme les chaines de caractères, regrouper dans un tableau de pointeur qui les pointes.
La distance entre les paires d'acides aminées sont stockées dans une matrice de taille $L \times L$ mais seul la partie supérieure au diagonal sont utilisés, dans chaque case on a la valeur de la distance entre la première position donnée par l'abscisse et la deuxième position donné par l'ordonné.
On a des fonctions pour lire ces données dans les fichiers correspondant, fessant des prétraitements nécessaires et finalement les stockées comme décrit ci-dessus.
Les fonctions sont respectivement : -lireDtrain
-lireTestSeq
-lireDistanceFile

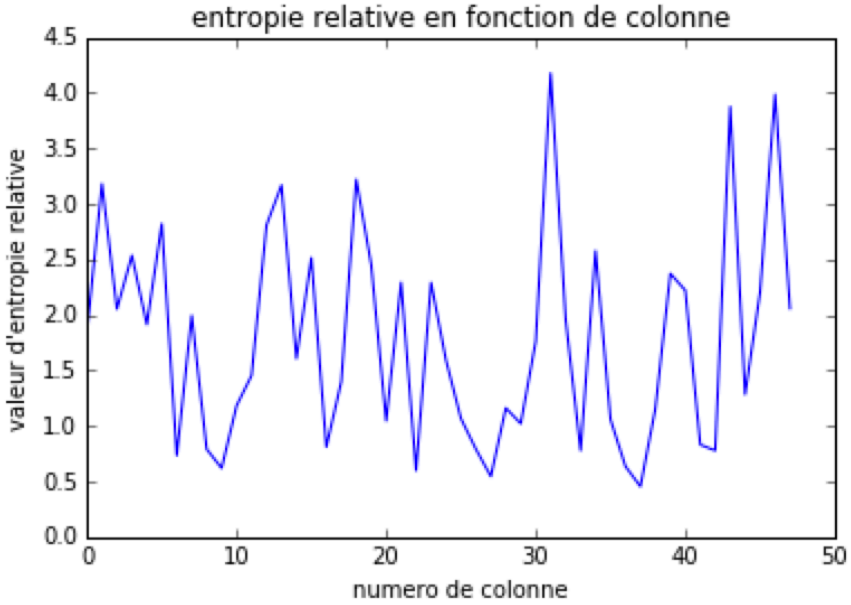
II. Modélisation par PSWM

1) Modélisation

PSWM appelé "position-specific weight matrix" est une matrice qui pour chaque position en colonne, calcule la probabilité d'apparition de chaque alphabet dont le poid. Dans notre cas la matrice considérer est notre matrice qui représente une famille de protéine.
Les poids sont calculés en utilisant les nombres d'occurrence et un pseudo count pour assuré que tous les poids ont une valeur supérieure à 0. Le but est de facilité les calcules avenir tout en gardant la validité des poids.
La structure de donné adopter est un tableau de dictionnaire, chaque case du tableau correspond à une colonne du PSWM et de choix de dictionnaire est que les clés sont des lettres d'alphabet donc un dictionnaire est parfaitement adapter la représentation de ce genre de donné.
La fonction correspondant est –poidDtrain.

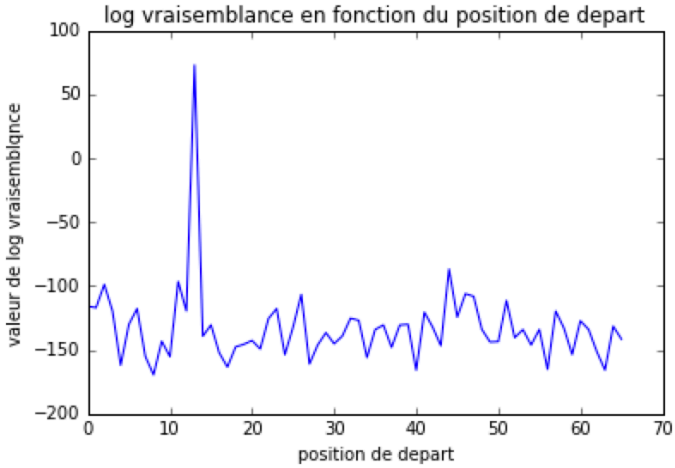
2) Entropie des positions

A partir de PSWM on peut faire une première analyse, l'entropie relative pour chaque colonne. C'est à dire l'information que la colonne nous apporte, plus la valeur de l'entropie relative est grand plus la colonne nous apporte de l'information. Il est caractérisé par un présence en majorité ou non d'un alphabet, si les alphabets sont présents de manière équiprobable alors on ne peut déduire de l'information donc la valeur de l'entropie relative est petit. Si au contraire, la forte présence d'un alphabet nous renseigne l'importance cette colonne, dans ce cas la valeur de l'entropie relative est plus élevé.
L'entropie relative est stockée sous forme d'un tableau, l'index représente les colonnes.
La fonction correspondant est : -listEntropie
On peut ainsi tracer la courbe des valeurs d'entropie relative en fonction de la colonne.



3) Log-vraisemblance

Maintenant on peut déterminer si une séquence de protéine appartient ou non à une famille de protéine mais il nous faut un seuil de probabilité pour en décider de l'appartenant, le seuil est donc calculé par un *modèle nul*. Le calcule est le suivant, on calcule ensuite le produit de poids pour la position que chaque alphabet de la séquence occupe. On considère que tous alphabet sont indépendant et on calcule la probabilité d'apparition de la séquence dans la famille entière donc c'est le produit des poids de chaque alphabet de la séquence dans la famille. La valeur de log vraisemblance est le log de leur fraction. Si la valeur est positive et grande, on peut dire que la séquence appartient à la famille.
Les valeurs sont dans un tableau avec l'index la position de départ de la séquence.
La fonction correspondant est : - tabLogVarSeq
On peut ainsi tracer la courbe des valeurs de log vraisemblance en fonction de la position de départ.



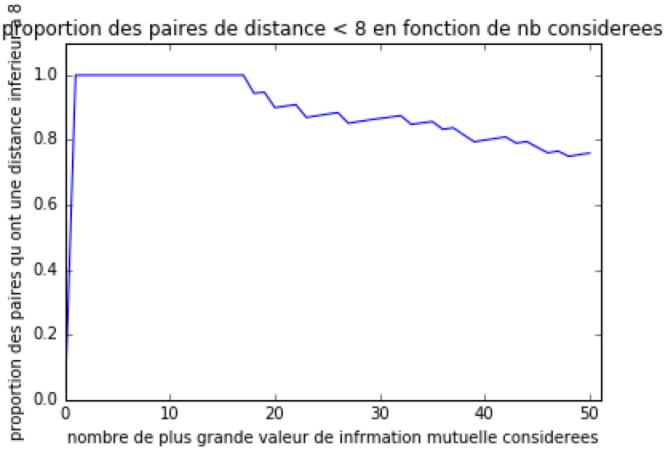
Une pic pour la position 14 donc la séquence de position de départ 14 appartient à la famille.

III. Coévolution de résiduel en contact

On calcule les cooccurrences et les poids respectifs de même manière que PSWM avec un pseudo count.
La fonction correspondant est : -mat_co_occu
Stocké sous forme d'une matrice de taille $L \times L$ ou seul la partie supérieure au diagonal est utilisé, dans dans chaque case l'abscisse et l'ordonné correspond aux positions et comme valeur un pointeur vers une collections Counter qui est une extension de la classe dictionnaire qui nous permet d'éviter de représenter les combinaisons de lettre d'alphabet qui ont une occurrence nul et aussi de conter avec une meilleure efficacité.

Calculons à présent la matrice *d'information mutuelle*, il vaut nul si et seulement si les deux positions sont statiquement indépendantes et prends une valeur positive sinon.
La fonction correspondant est : - inf_Mutuelle
Stocké sous forme d'une matrice de taille $L \times L$ ou seul la partie supérieure au diagonal est utilisé, dans dans chaque case l'abscisse et l'ordonné correspond aux positions et la valeur d'information mutuelle.

Faisons le lien entre la distance (en Angstrom) entre les positions lu dans le fichier distance.



On en conclue que les paires les plus corrélées ont une probabilité élevée d'être en contact.
Plus la valeur d'information mutuelle diminue, moins il y a des paires en contact.