SORBONNE UNIVERSITÉ

Rapport

Mini_Projet2 :
Analyse Statistique D'une Famille De Protéines

Yuchen ZHONG

2018

SOMMAIRE

1. INTRODUCTION

Dans ce mini-projet, on va chercher à analyser statistiquement une famille de protéines donnée par un alignement de séquences en ayant 3 objectifs:

- Détection des positions conservées
- Détection de séquences qui appartiennent à la même famille,
- Détection de corrélations entre colonnes différentes de l'alignement, et de leur relation avec distances entres acides aminées dans la structure 3D d'une protéine représentative de la famille.

Le principe est d'essayer d'extraire information statistique sur structure et fonction.

2. DONNÉES

Avant de commencement, on va d'abord expliquer simplement les données étudiées.

Dans le fichier « Dtrain.txt », il contient M = 5643 séquences de protéines d'une seule famille en format FASTA, une ligne de commentaire commençant par la caractère « > » et une ligne de séquence alignée de même longueur L = 48 dont chaque position contient un élément dans l'ensemble alphabet q = 21, où il existe 20 acides aminés et un trou.

Lors de fichier « distances.txt », il y a des informations sur les distances entre paires d'acides aminées afin d'obtenir les informations importantes en utilisant notre modélisation PSWM.

3.MODÈLISATION PAR PSWM

A. Estimer une PSWM

« position-specific weight matrix » soit PSWM, une matrice de poids spécifiques des positions. Du coup, on a lu des acides aminées de la fichier « Dtrain.txt » et les mis à une matrice LxM. Et il nous faut

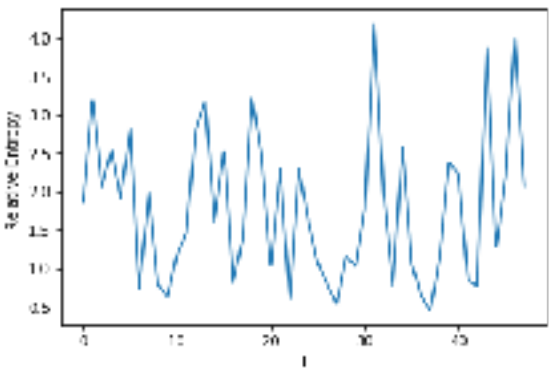
aussi de savoir le nombre d'occurrences d'acide aminée « a » en colonne i, afin de calculer son poids avec une supposition que toutes les positions sont indépendantes. Et on a testé que P(a0,...,aL-1|w) = 1 et alors $w_0(' -') = 0.3132$ correspond au résultat. Les trois importantes acides aminées sont ['W', 'P', 'G'].

B. Conservation

Dans cette partie, on essaye de trouver les positions qui ont un poids très élevé pour une acide aminée, soit une position conservée et importante dans la biologie. Par conséquent, on utilise une notion entropie relative S et elle se servit à déterminer laquelle acide aminée plus probable dans une spécifique position. Et on a testé que l'entropie relative S à la colonne 0 S_0 est égale à 1.8547 correspondant au résultat.

C. Evaluer une nouvelle séquence

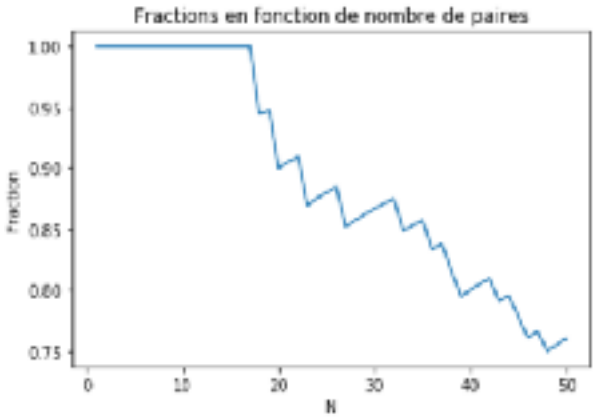
On utilise une nouvelle séquence à tester b dans la fichier « test_seq.txt » et notre modèle PSWM en vue de calculer la probabilité $P(b_0, \ldots, b_{L-1} \mid w)$ en comparant avec un modèle nul $f^{(0)}(b)$ qui n'est pas spécifique dans les positions. En raison de la facilité de calcul, on introduit la log-vraisemblance $l(b_0, \ldots, b_{L-1})$. Et on a testé que sa valeur de 48 premières éléments de la séquence de fichier « test_seq.txt » est égale à -115.7964. Et on a bien trouvé une séquence appartenant à la même famille dont l = 72.843. Et voici les deux dessins ci-dessous qu'il s'agit de l'entropie et lot-vraisemblance en fonction de sa position l.



4. CO-ÉVOLUTION DE RÉSIDUES EN CONTACT

Après fini de notre partie de simple modèle factorisé des PSWM, on commence à chercher la co-évolution de deux positions qui sont en contact. Afin de détecter les corrélations, on calcul les nombres de co-occurences avec acide aminée a en position i et b en position j. En théorique, $\sum_{b} w_{ij}(a,b) = w_i(a)$ que l'on a bien testé.

Ensuite, on continue à calculer l'information mutuelle M_{ij} qui quantifie les corrélations des paires de positions. Plus que elles sont corrélées, plus que M_{ij} est grande. Et à la comparaison avec les distances correspondantes, on pourra déterminer la probabilité élevée d'être en contact. Et on a testé que $M_{0,1}$ = 0.40404 correspondant au résultat. Et voici le dessin ci-dessous qu'il s'agit de la fraction en fonction du nombre de paires considérées :



5. CONCLUSTION

Alors cette analyse statistique nous permet de trouver certaines informations sur structure et sur différentes de acides aminés à l'aide de notre modélisation PSWM et et de la corrélation. Cela a vraiment du sens dans la biologie afin de découvrir les acides aminés importantes dans de différentes séquences et dans le futur nous pouvons peut-être de remplacer certaine acide aminée par une autre différente.

6. Guide D’Utilisation

Il est simple de faire fonctionner notre code en utilisant une commande « jupyter notebook » dans le terminal et trouvant notre fichier dans le répertoire de notebook sur le navigateur.