SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

3I005

PROJET 2

Statistique en bioinformatique :

Analyse statistique d'une famille de protéines

Encadré par Ari Ugarte

Réalisé par :

AMROUCHE Sara (3523540)

RIABI Arij (3702151)

Sommaire

- Introduction
- Modélisation par PSWM
  - A. Estimer une PSWM
  - B. Conservation
  - C. Evaluation d'une nouvelle séquence
- Coévolution de résidus en contact
- Conclusion

Introduction :

Les protéines sont un type de macromolécule propre au vivant et présent dans chaque organisme.Elles sont constituées d'une suite de résidus d'acides aminés Elle est caractérisée par Sa structure primaire qui n'est autre que la séquence d'acides aminés de la protéine .Dans le cadre de notre projet, nous nous intéresserons à l'homologie entre les protéines : deux protéines sont dites homologues lorsque les gènes qui les codent ont une origine commune du point de vue de l'évolution

La recherche et la classification des protéines homologues sont des enjeux importants de la bio-informatique. Plusieurs méthodes ont été exploitées pour résoudre ce problème. Le développement de l'informatique permet le stockage et l'organisation (structuration) des données biologiques. L'identification des protéines aujourd'hui possible de manière automatique

L'objectif de notre projet alors est la détermination de la structure de protéines par des techniques intermédiaires entre les données expérimentales utilisant des données spécifiques à la protéine étudiée et les méthodes de prédiction utilisant le plus possible de données génériques communes l'ensemble des protéines puisque la plupart d'entre elles reposent sur la conversion de données statistiques, issues d'expériences, en potentiels de contrainte ou de tri. L'approche proposée pour résoudre le problème de la prédiction de structure des protéines est d'utiliser la structure d'une ou plusieurs protéines dont la séquence présente une grande similarité. Cette solution est appelée modélisation comparative, aussi dénommée modélisation par homologie ;La modélisation par homologie utilise les outils d'alignement de séquences pour trouver des protéines dont la séquence est proche de la protéine que l'on cherche { modéliser la a définition des domaines d'homologie date des premières études sur les protéines à activité tyrosine kinase (PTK). Les domaines les plus répandus et les plus étudiés sont les domaines SH2 et SH3 (Src homology 2 et 3), souvent trouvés de concert, voire en plusieurs exemplaires dans une même protéine. Dans notre projet on va s'intéresser au domaine SH3.

Première partie : Modélisation par PSWM

A. Estimer une PSWM

Si l'on se place dans le cadre d'un modèle de Bernoulli, où les différentes positions sont indépendantes entres elles, une matrice poids-position permet de représenter l'information : en chaque position, on calcule la fréquence d'apparition de chacun des caractères. Ainsi la matrice PWM est un bon modèle à utiliser.

Description de la fonction nombre_occurence :

Cette fonction sert à calculer pour chaque position i allant de 0 à L (48) et chaque acide aminé appartenant à l'ensemble alpha_acide le nombre d'occurrence de chaque acide dans la position i, elle renvoie un dictionnaire de clé-valeur tel que chaque clé est sous forme d'un couple (i, a) avec comme valeur le nombre d'occurrence associé à l'acide a dans la position i.

Description de la fonction poids :

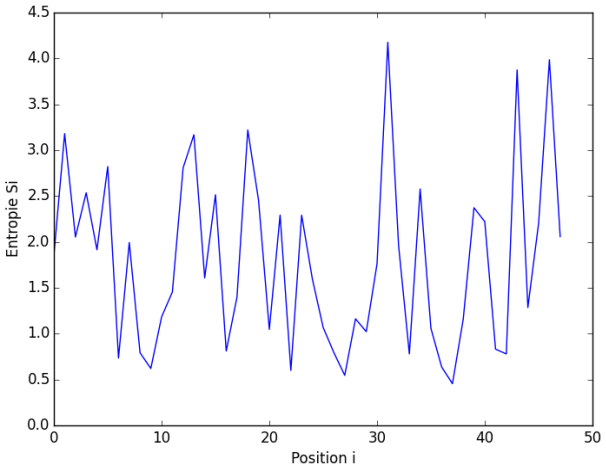
Cette fonction calcule le poids de chaque acide a dans une position i selon la formule ωi(a) = (ni(a) + 1) / (M + q), elle prend comme entrée le dictionnaire d'occurrence calculé par nombre_occurence, et elle construit une matrice M de poids telle que le poids d'un acide a pour une position i, situé a la case M[i] [position de a dans alpha_acide].

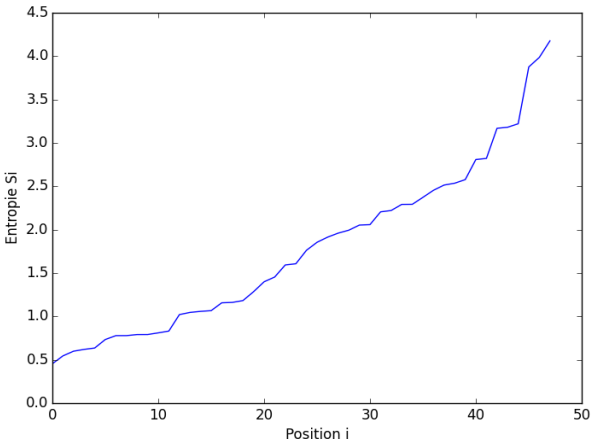
B. Conservation

La conservation d'une position dans un alignement multiple est donc une information importante dans l'étude d'une famille de séquences puisque la modélisation par homologie s'appuie sur l'hypothèse que des protéines homologues ont en général conservé la même fonction et une topologie et une structure proches.

Description de la fonction entropie_relative :

Entropie_relative, comme l'indique son nom, permet de trouver les positions qui sont conservées, autrement dit, avec le poids le plus élevés. Cette fonction va prendre en entrée la matrice de poids, et va calculer suivant la formule
Si = log2 (q) $+\sum_{a\in A} w_i(a) \cdot log[w_{i(A)}]$ une liste qui va contenir toutes les entropies pour chaque position i allant de 0 à 48.
On a donc tracer la courbe représentant les variations des entropies relatives en fonctions de la position i, et on a obtenu ces résultats .





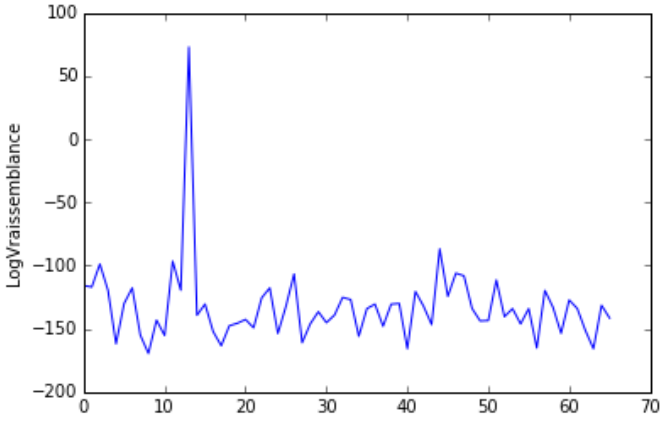
On a à gauche la courbe représentant les entropies relatives sans les avoir triés, et à droite la courbes des entropies après avoir effectué le tri. On constate pleins de variation des entropies des entropies, avec des positions ayant une Entropie allant jusqu'à 46 ! et on peut dire que plus l'entropie relative pour une colonne est élevée, plus cette position est conservée c'est-à-dire la protéine rencontre moins de mutations sur celle-ci. Et après avoir implémenter la fonction qui calcule les trois plus grandes entropies, on a constaté que P, G et W avaient les plus grandes entropies et donc étaient les plus conservées

C. Evaluer une nouvelle séquence :

Le moyen le plus simple de caractériser une famille de séquences est d'identifier les conservations de séquence au sein de la famille. Pour cela on construit un alignement dit alignement multiple de séquences maximisant la vraisemblance.

Description de la fonction Log Vraisemblance :

Cette fonction va calculer à partir d'un fichier de séquences et des poids déjà calculés des données train, une liste qui va contenir toutes les « log vraisemblance » des séquences afin de déterminer si une séquence donnée dans ce fichier est plus probable dans le modèle spécifié. Pour cela elle fera appel à la fonction **fonctionFmodeleNul** (calcule la liste de fréquence de chaque de toutes les lettres dans un alignement du train sans voir la position i, et ensuite grâce à la formule elle renvoie les « logvraisemblance » cherchées. On a tracé la courbe de log vraisemblance en fonction de la première position i (de 0 à N-L) et on a obtenu le résultat suivant :



On sait bien que plus le log vraisemblance de la séquence b est grand, plus celle-ci a de chance d'appartenir à la famille, et on constate plusieurs variations dans cette courbe, et on peut voir un pic au niveau de la position 13, ce qui veut dire que la séquence associée a plus de chance d'appartenir à la famille.

Deuxième partie : Co-evolution de RÉSIDUES EN CONTACT

L'appartenance à une même famille laisse entendre que ces protéines dérivent d'un ancêtre commun. Certains éléments de séquence, comme par exemple la position conservée des résidus cystéine (qui jouent un rôle important dans le repliement des protéines) peuvent être pris en compte pour compenser une faible identité des séquences. On peut sur le même principe regrouper certaines familles en superfamilles comme par exemple celle des récepteurs nucléaires.

Le repliement d'une chaîne polypeptidique étant un processus spontané sous le contrôle de la séquence, il était admis comme corollaire que la structure native d'une protéine est la structure la plus stable dans un environnement donné.

Description de la fonction nombre_occurence2 :

Cette fonction va retourner un dictionnaire de clé-valeur telle que chaque clé est sous la forme d'un quadruplé (i, j, a, b) où i et j sont les positions respectives des acide aminés a et b, quant à la valeur, elle représente nombre de séquences avec l'acide aminée a en cette position i et avec acide aminée b en position j.

Description de la fonction poids2 :

Cette fonction va prendre en argument le dictionnaire retourné par la fonction nombre_occurence2 et qui va renvoyer un dictionnaire avec le quadruplet (i, j, a, b) précédent et auquel elle associe le poids correspondant qui sera calculé à partir de la formule suivante ωij (a, b) = (nij (a, b) + 1)/(q M + q).

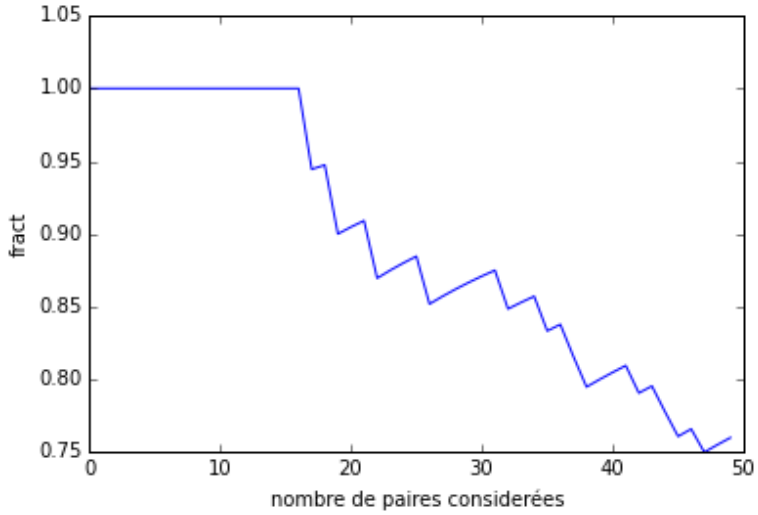
Description de la fonction info mutuelle :

Cette fonction prend en argument la matrice de poids renvoyé par la fonction de poids de la première partie ainsi que le dictionnaire de poids calculé à la deuxième partie **et** elle construit une matrice ou elle associe à chaque paire de position (i,j) la valeur dite information mutuelle qui est calculé avec la formule qui lui est associé dans l enoncé.

Description de la fonction Distance :

Le but de cette fonction est de calculer, la fraction des 50 paires de positions ayant l'information mutuelle M la plus grande, avec une distance inférieure à 8, et donc au début, on va trier les informations mutuelles trouvé, prendre les 50 plus grandes valeurs, puis trouver la distance correspondante a ces paires, on calcule ensuite la fraction des paires ayant une distance plus petite que 8, et on renvoie une liste contenant ces fractions.

On trace maintenant la courbe de fraction en fonction du nombre de paires considérées et on obtient le résultat suivant :



On remarque à partir de cette courbe que jusqu'à un nombre de pair 16, on a une fraction qui est constance à la valeur de 1, puis elle commence a diminué jusqu'à arriver à 0.75, et donc on peut remarquer que les paires les plus corrélées ont une probabilité élevée d'être en contact.

On peut conclure que Plus la valeur associée au nombre de paires de positions sélectionnées est grande, plus ces paires auront une probabilité élevée d'être en contact. Donc plus on a de paires, moins on a de proba qu'ils soient en contact, comme dans ce cas , on a une grande valeur jusqu'à 16 paire, puis elle commence à diminuer.

CONCLUSION :

La principale motivation à la base de ce Projet était de mettre à profit les alignements multiples de séquences protéiques pour analyser la conservation des interactions au sein des structures protéiques, mais aussi pour apporter des outils automatiques qui peuvent aider à analyser la pertinence de ces alignements afin de détecter les séquences qui appartiennent à la même famille et de corrélations entre colonnes différentes de l'alignement, et de leur relation avec les distances entres acides aminés dans la structure 3D d'une protéine représentative de la famille.