Projet 2 : Statistique en BioInformatique

Gauvain Jodie et Bielle Benjamin

13 novembre 2013

Sommaire

Chapitre 1

Analyse statistique d'une famille de protéines

1.1 Introduction

Ce projet a pour but de faire tester la pertinence d'un modèle de données représentant un domaine de protéines.
Le modèle que nous avons choisi est une matrice de poids pondérés (PWM). Elle exprime, pour chaque position, la proportion de chaque acide amineé (AA). Nous montrerons la pertinence ou non de ce modèle dans la suite du rapport.

1.2 Expérimentations

Pour étudier notre modèle, nous faisons la chose suivante :
Une instance de PWM est créée, celle-ci exprime (voir Introduction)la proportion de chaque AA pour chaque position.
La structure de donnée utilisée pour modèliser la PWM est une arraylist de arraylist (ainsi notre matrice peut être très modulable). Elle utilise pour les calculs de n_{ij} une arraylist de caractères pour effectuer le dénombrement. Notre programme comporte un menu permettant à l'utilisateur de faciliter son utilisation du programme (seul les tests d'informations mutuelles et de fraction ne s'affichent pas mais sont redirigés vers des fichiers).
Notre programme est constitué de différents fichiers ayant tous un rôle très précis.
Le fichier ***LectureDTrain.java*** permet de lire le fichier ***Dtrain.txt*** et ainsi d'initialiser la matrice associée.
Le fichier ***LectureTSeq.java*** permet de lire le fichier ***test seq.txt*** et ainsi d'initialiser la matrice associée.
Le fichier ***Test.java*** comporte simplement les différents tests du programme.
Et enfin le fichier ***Statistiques.java*** contient les différentes fonctions (poids, entropie, vraisemblance, ...) de statistiques et de calcul de notre programme.

1.3 Résultats

Le calcul de l'entropie relative pour chaque position fait apparaître que les positions les plus conservées sont :

- la position 31.
- la position 43.

– la position 47.

pour les acides aminés W , G et P (respectivement).
Voir la courbe ***Entropie.png*** dans le dossier ***graph***.

Ensuite des 50 couples les plus corrélés, on extrait la fraction d'entre eux qui sont effectivement en contact (c'est à dire ceux dont la distance est inférieure a huit) dans le fichier ***distances.txt***.
Voir la courbe ***Fraction.png*** dans le dossier ***graph***.

Enfin, depuis la chaine de test contenue dans le fichier ***test seq.txt***, on extrait les sous-chaîne qui ont une vraisemblance à zéro. Sur le graphe ***Vraisemblance.png*** (dans le dossier ***graph***), on observe un pic, ce pic montre que cette sous-chaîne à plus de chance d'appartenir à un domaine considéré.
Sur notre graphe, on remarque que la sous-chaîne 13 appartient probablement au domaine considéré.

1.4 Conclusion

On peut en conclure, d'après nos résultats, que le modèle choisi est pertinent.
Ce modèle est capable de déterminer l'appartenance d'une chaîne au domaine, de repérer les positions les plus conservées dans le domaine et de détecter (avec un taux de succès correcte :) les possibles contacts entre différentes positions.