Regroupement par les k-means de structures à l'aide d'un alphabet structural

Nicolas Zimmermann - M2BI

Vendredi 20 Septembre 2019

Contents

1 Introduction **1**

2 Matériel et méthodes **1**

2.1 Description des algorithmes 1

2.1.1 PBxplore 1

2.1.2 k-medoids 1

2.2 Description des données 2

3 Résultats **2**

4 Conclusion **3**

1 Introduction

Le but de ce projet a été de développer un module permettant de classifier les conformations d'une protéine. Nous nous sommes basés sur des une dynamique moléculaire du domaine calf-1 pour tester le module. Le module a été développé en python et reprend plusieurs autres module, principalement PBxplore[**?**]. Grâce au module PBxplore[**?**], les structures 3D des conformations de protéines sont encodées en séquences 1D de protein blocks. Des matrices de distances entre séquences sont ensuite calculées soit par identité soit par différence en utilisant une matrice de substitution. Les matrices ainsi obtenus sont ensuite utilisé par un algorithme de k-medoids (il était impossible d'utilisé les séquences 1D tel quel en entrée de k-means) pour réaliser des clusters de séquences.

2 Matériel et méthodes

2.1 Description des algorithmes

2.1.1 PBxplore

PBxplore[**?**] se base sur la conformation spatiale de la chaîne alpha de la protéine pour la simplifier en blocs. Chaque cinq résidus en 3D sont ainsi simplifié en un bloc 1D ce qui réduit grandement les dimensionnalités des données. La similitude entre les différents bloques est disponible sous la forme d'une matrice de substitution dans le module, cette matrice fut utilisé pour généré une matrice de distance par dissimilarité.

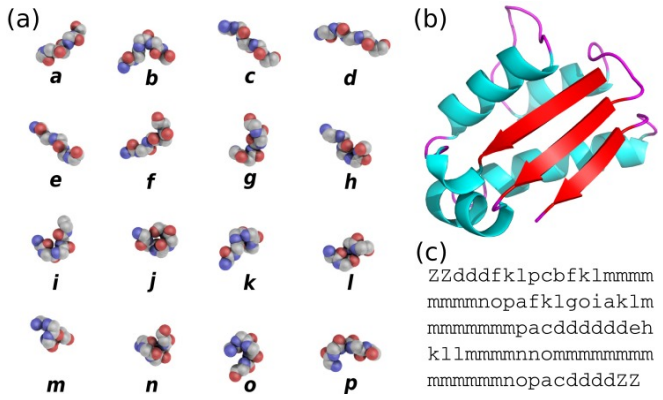

Figure 1: (a) Les 16 blocs de l'alphabet PB (b) Structure 3D d'une protéine (c) Séquence 1D de la protéine '(b)' encodées en blocs[**?**]

2.1.2 k-medoids

D'après le sujet initial, l'algorithme des k-means devait être utilisé pour obtenir des clusters des différentes conformations de la protéine. Cependant l'algorithme des k-means prend en entrée des objets ayant des variables quantitatives. Les séquences 1D étant une suite de variables qualitatives, nous avons

du nous tourné vers un autre algorithme afin de réalisé la classification. Nous avons ainsi choisi d'utilisé l'algorithme des k-medoids provenant du module pyclustering[**?**] pour réalisé la classification. Cet algorithme propose l'utilisation d'une matrice de distances entre les objets à classifier pour réaliser les clusters. Les matrices de distances sont calculés soit pas identité/différences des blocs, soit en utilisant une matrice de substitution. Les deux matrices peuvent êtres calculées et donnent des résultats.

De la même manière qu'un k-means, les groupes minimise la distance intra-groupe des éléments et maximise la distance inter-groupe. Les médoides sont les éléments des groupes les plus proches de tout les éléments du groupes. Dans notre cas, ces éléments sont intéressant puisqu'ils sont les conformations de la protéines les plus représentatives des groupes. On peut ainsi observer ces conformations "d'intérêt".

2.2 Description des données

Les données utilisées proviennent d'une dynamique moléculaire[**?**] du domaine Calf-1. Les structures des conformations ont été enregistré au format pdb en utilisant le logiciel gromacs, la procédure étant décrite dans le code source du projet.

Il résulte de cette dynamique 501 conformations du domaine Calf-1 qui seront classifié par notre module

3 Résultats

Les données ont pu être classifié par k-medoids mais d'avantage de temps est nécessaire pour obtenir des résultats probants.

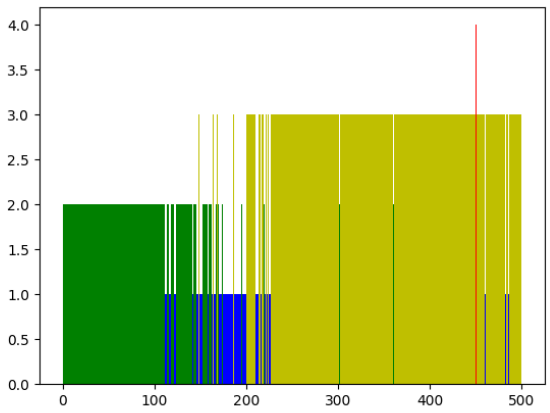

Figure 2: Clusters obtenus à partir de la matrice de distance la plus simple. Sur l'axe des abscisses chaque position correspond à une conformations, les conformations sont présentés dans l'ordre du déroulement de la dynamique moléculaire. L'axe des abscisses peut ainsi être vu comme une dimension temporelle. La valeur(hauteur) de l'axe des ordonnées ainsi que la couleurs ne servent qu'à identifié les clusters

On observe sur cette figure la division des conformations en 4 clusters. Les conformations proches temporellement ont tendance à se retrouver dans les mêmes clusters, cela pourrait être expliqué par le temps nécessaire à la protéine pour se retrouver dans une conformation significativement différente.

4 Conclusion

L'utilisation de l'algorithme des k-medoids[**?**] permet de classifier les différentes conformations des protéines. Les résultats sont intéressant mais plus de temps est nécessaire afin de mieux exploité les potentiels de la méthode.