PROJET LONG

MÉTHODE DE GÉNÉRATION DE SCORE DE COMPATIBILITÉ 1D-3D PAR UN ALGORITHME DÁPPRENTISSAGE AUTOMATIQUE

January 8, 2018

Safia SAFA-TAHAR-HENNI
Master II Bio-informatique
Université Paris Diderot

Contents

INTRODUCTION

0.0.1 Etat de l'art et problématique

La compréhension des mécanismes de repliement des protéines est souvent considérée comme un objectif important qui permettra aux biologistes structuraux de découvrir la relation mystérieuse entre la séquence, la structure et la fonction des protéines. Les connaissances sur le repliement des protéines pourraient faciliter davantage la conception de médicaments. Prédire la structure 3D (repliement) d'une protéine est un problème clé en biologie moléculaire. La détermination du reploiement d'une protéine repose principalement sur des méthodes expérimentales moléculaires. Avec le développement de techniques de séquençage de nouvelle génération, la découverte de nouvelles séquences protéiques a rapidement augmenté. Avec un tel nombre de protéines, l'utilisation de techniques expérimentales pour déterminer le repliement des protéines est extrêmement difficile parce que ces techniques sont longues et coûteuses. La capacité de prédire les taux de repliement des protéines sans avoir besoin de travaux expérimentaux aiderait le travail de recherche des biologistes structuraux. Ainsi, il est urgent de développer des méthodes de prédiction par ordinateur capables de classer automatiquement, rapidement et avec précision des séquences de protéines inconnues dans des catégories de plis spécifiques. Les méthodes de calcul récentes, en particulier des méthodes basées sur l'apprentissage automatique, pour la reconnaissance des replis protéiques ont été développés.(Recent Progress in Machine Learning-Based Methods for Protein Fold Recognition) Les méthodes basées sur l'apprentissage automatique peuvent être classées en deux classes selon les algorithmes d'apprentissage utilisés dans les modèles de prédiction : (1) méthodes basées sur un seul classificateur ; et (2) des méthodes basées sur un classificateur d'ensemble. Les méthodes basées sur un seul classificateur utilisent un seul algorithme d'apprentissage spécifique pour construire des modèles de prédiction, tandis que les méthodes basées sur un classificateur d'ensemble utilisent des algorithmes d'apprentissage multiples, similaires ou différents, pour construire des modèles de prédiction. La plupart des méthodes actuelles de classification unique utilisées dans la reconnaissance des repliements protéiques sont basées sur un classificateur SVM. Chen et al. ont récemment proposé une méthode de reconnaissance des replis protéiques à base de RF appelée PFP-RFSM. Le cadre de PFP-RFSM implique un algorithme complet de représentation des caractéristiques qui peut capturer des informations séquentielles et structurelles distinctes à partir des séquences primaires et des structures prédites, respectivement. Lampros et al. proposent une nouvelle méthode d'optimisation pour la classification des repliements protéiques; le modèle de prédiction de cette méthode est construit sur la base d'une chaîne

de Markov formée avec la structure primaire des protéines et sur un HMM à espace réduit.

De nombreux outils bio-informatiques ont vu le jour au cours de la dernière décennie et chacun a présenté des caractéristiques différentes. L'article de (Towards more accurate prediction of protein folding rates: a review of the existing Web-based bioinformatics approaches.), présente une revue et une comparaison d'outils (SFoldRate FOLD-RATE Pred-PFR FoldRate K-Fold) de prédiction qui sont actuellement disponibles sur le Web et prédisent principalement le taux de repliement des protéines.

Ce projet porte sur l'implémentation d'une méthode de génération de score de compatibilité 1D-3D grâce un algorithme dápprentissage automatique. Pour cela, la base de données utilisées est composée d'une structure de chaque famille de la base de donnée Homstrad. À chaque acide aminé (de chaque séquence de chaque structure) est associé un vecteur d'environnement. Ce vecteur est composé d'information sur la structure secondaire, l'accessibilité au solvant, la déformabilité de la protéines (grâce aux alphabets structuraux), et le nombre et types d'interaction créés. Ces vecteurs vont être utilisés afin de faire un regroupement par la méthode de Kmeans ou Kmedoids afin de caractériser les différents environnements (Enfoui? Accessible ?). Un score pourra être déterminé en calculant la fréquence d'un acide aminé dans chaque cluster par rapport à la fréquence attendu. Grâce à une prédiction par arbre de décision, on pourra déterminer un acide aminé en fonction de son environnement, la prédiction d'une séquence à partir de vecteurs d'environnement seulement pourrais permettre de mettre en avant des acides aminés qui pourraient stabiliser la structure.

MATÉRIEL & MÉTHODES

Le programme est disponible à l'adresse suivante : `https://github.com/SafiaSafa/projet_long`

0.0.2 Données

0.0.2.1 Homstrad

Les données utilisées sont issues de la base de données Homstrad. HOMSTRAD (HOMologous STRucture Alignment Database) est une base de données organisée d'alignements basés sur la structure pour les familles de protéines homologues. Toutes les structures protéiques connues sont regroupées en familles homologues (c'est-à-dire, une ascendance commune) et les séquences des membres représentatifs de chaque famille sont alignées sur la base de

leurs structures 3D en utilisant les programmes MNYFIT, STAMP et COMPARER. Ces alignements basés sur la structure sont annotés avec JOY et examinés individuellement.(Reference : Mizuguchi K, Deane CM, Blundell TL, Overington JP. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. Protein Science 7:2469-2471.) Dans le cadre de ce projet, la base de données est composée d'un représentant(une structure) par famille (qui possède sur Homstrad au moins 2 PDB).

0.0.2.2 PBxplore

L'utilisation d'un alphabet structural, les Blocs Protéiques (PBs), est un très bon outil pour donner une idée de la déformabilité de la protéine. Les alphabets structuraux sont des librairies de fragment 3D qui permettent d'approximer une structure protéique. Les PBs sont des petits prototypes structurels qui vont permettre d'approximer localement la structure protéique. L'API PBxplore (Disponible sur https://github.com/pierrepo/PBxplore) est composé d'un ensemble d'outils qui permette d'analyser les Dynamique Moléculiare(MD) et la déformabilité structurale des protéines en utilisant les PBs. Il permet de retranscrire chaque snapshots (structures obtenues au cours de la MD) en séquences de Blocs Protéiques, et propose une analyse statistique des variations conformationnelles. (PBxplore: A Tool To Analyze Local Protein Structure And Deformability With Protein Blocks)

0.0.2.3 DSSP

Afin d'avoir accès aux informations concernant la structure secondaire et l'accessibilité au solvant, j'ai utilisé le programme DSSP, conçu par Wolfgang Kabsch et Chris Sander. DSSP est une base de données d'affectations de structures secondaires (et beaucoup plus) pour toutes les entrées de protéines dans la PDB.

0.0.2.4 Arpeggio

Arpeggio, un serveur Web qui permet de calculer les interactions entre protéines et protéines, ADN ou de petites molécules ligands, comprenant les interactions de van der Waals, ioniques, carbonyles, métaux, hydrophobes, liaisons halogènes, les liaisons hydrogène et les interactions spécifiques aux atomes des cycles aromatiques et anneau aromatique-anneau aromatique. Je me suis plus particulièrement intéressait aux interactions polaires, ioniques, hydrophobes, de van der Waals et aux liaisons hydrogènes.

0.0.3 Algorithme

0.0.3.1 Clustering

1. Kmeans

   Le regroupement par K-means est l'algorithme d'apprentissage automatique non supervisé le plus couramment utilisé pour partitionner un ensemble de données donné en un ensemble de k groupes (c'est-à-dire k clusters), où k représente le nombre de groupes pré-spécifiés par l'analyste. Il classe les objets dans plusieurs groupes (clusters), de sorte que les objets d'un même cluster soient aussi similaires que possible (ie, haute similarité intra-classe), alors que les objets des différents clusters sont aussi différents que possible (ie faible similarité inter-classe). En regroupement par K-means, chaque cluster est représenté par son centre (ie, centroïde) qui correspond à la moyenne des points assignés au cluster. L'idée de base de la classification k-means consiste à définir des clusters afin de minimiser la variation totale intra-cluster. Il existe plusieurs algorithmes k-means disponibles, l'algorithme standard est l'algorithme de Hartigan-Wong (1979).

2. Kmedoids

   Le regroupement par k-Medoids se différencie des k-means par le fait qu'un cluster est représenté avec son centre dans l'algorithme k-means, mais avec l'objet le plus proche au centre du cluster dans l'algorithme k-medoids. Il est plus robuste que k-means en présence de valeurs aberrantes. PAM (Partitioning Around Medoids) est un algorithme classique pour k-medoids clustering. Dans ce programme, le package R pam() a été utilisé.

0.0.3.2 Arbre de décision

Il y a beaucoup de paquets dans R pour modéliser les arbres de décision: rpart, parti, RWeka, ipred, randomForest, gbm, C50. Dans ce programme, le package R rpart qui implémente le partitionnement récursif a été utilisé. Les algorithmes d'apprentissage basés sur l'arbre sont considérés comme l'une des méthodes d'apprentissage supervisé les plus utilisées. Contrairement à d'autres algorithmes basés sur des techniques statistiques, l'arbre de décision est un modèle non-paramétrique, n'ayant aucune hypothèse sous-jacente pour le modèle. Les arbres de décision sont de puissants classificateurs non linéaires, qui utilisent une structure arborescente pour modéliser les relations entre les entités et les résultats potentiels. . Cette

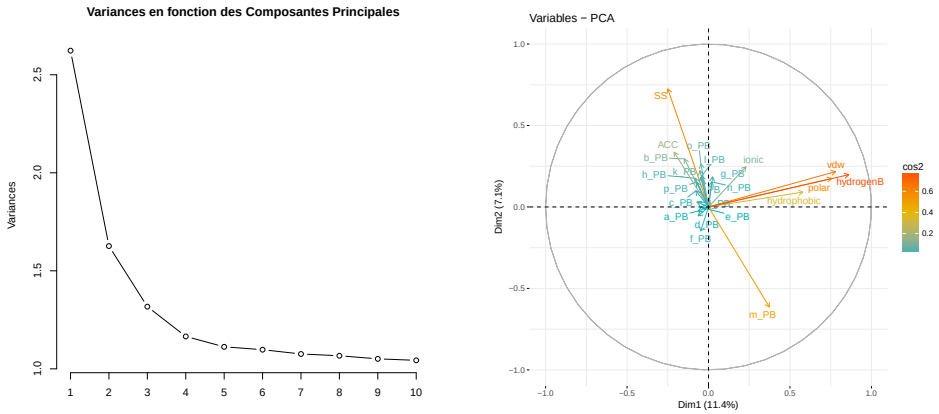
approche d'apprentissage machine est utilisée pour classer les données dans des classes et pour représenter les résultats dans un organigramme, tel qu'une structure arborescente.

RESULTATS

0.0.4 Clustering

Avant d'effectuer la clusterisation, une analyse en composante principale a été réalisé(Fig 1). On peut voir que les composante 1 et 2 suffisent pour représenter la majorité de la variance (Fig 1.A). De plus le graphique de corrélation des variables(Fig 1.B) montre les relations entre toutes les variables. Il peut être interprété comme suit:

Figure 1: Analyse en Composantes Principales: A:Graphique de la variance porté par Composantes Principale. B: Graphique de corrélation des variables. La qualité de représentation des variables sur la carte de l'ACP s'appelle cos2 (cosinus carré).

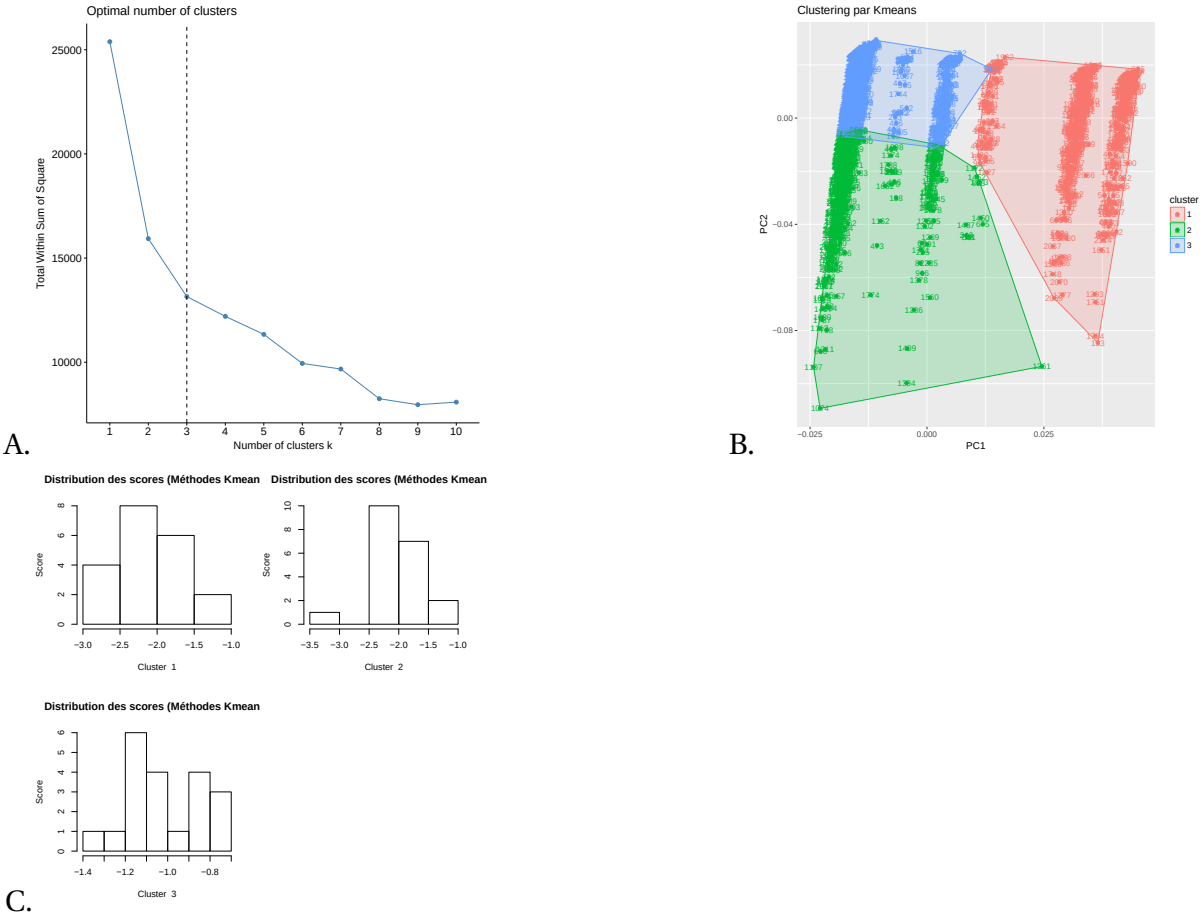


A. B.

- Les variables positivement corrélées sont regroupées.
- Les variables négativement corrélées sont positionnées sur les côtés opposés de l'origine du graphique (quadrants opposés).
- La distance entre les variables et l'origine mesure la qualité de représentation des variables. Les variables qui sont loin de l'origine sont bien représentées par l'ACP.

De plus un cos2 élevé indique une bonne représentation de la variable sur les axes principaux en considération. Dans ce cas, la variable est positionnée à proximité de la circonférence du cercle de corrélation. C'est le cas ici des variables représentants les liasions hydrogènes, vand der Waals ou bien la structure secondaire. Un faible cos2 indique que la variable n'est pas parfaitement représentée par les axes principaux. Dans ce cas, la variable est proche du centre du cercle.C'est le cas ici des variables représenant certain Proteins Blocks

Figure 2: Regroupement par Kmeans. A:Détermination du nombre de cluster optimal par la méthode du "genou". B:Résultats de la Regroupement par les Kmeans. C: scores de chaque acides aminées dans chaque cluster



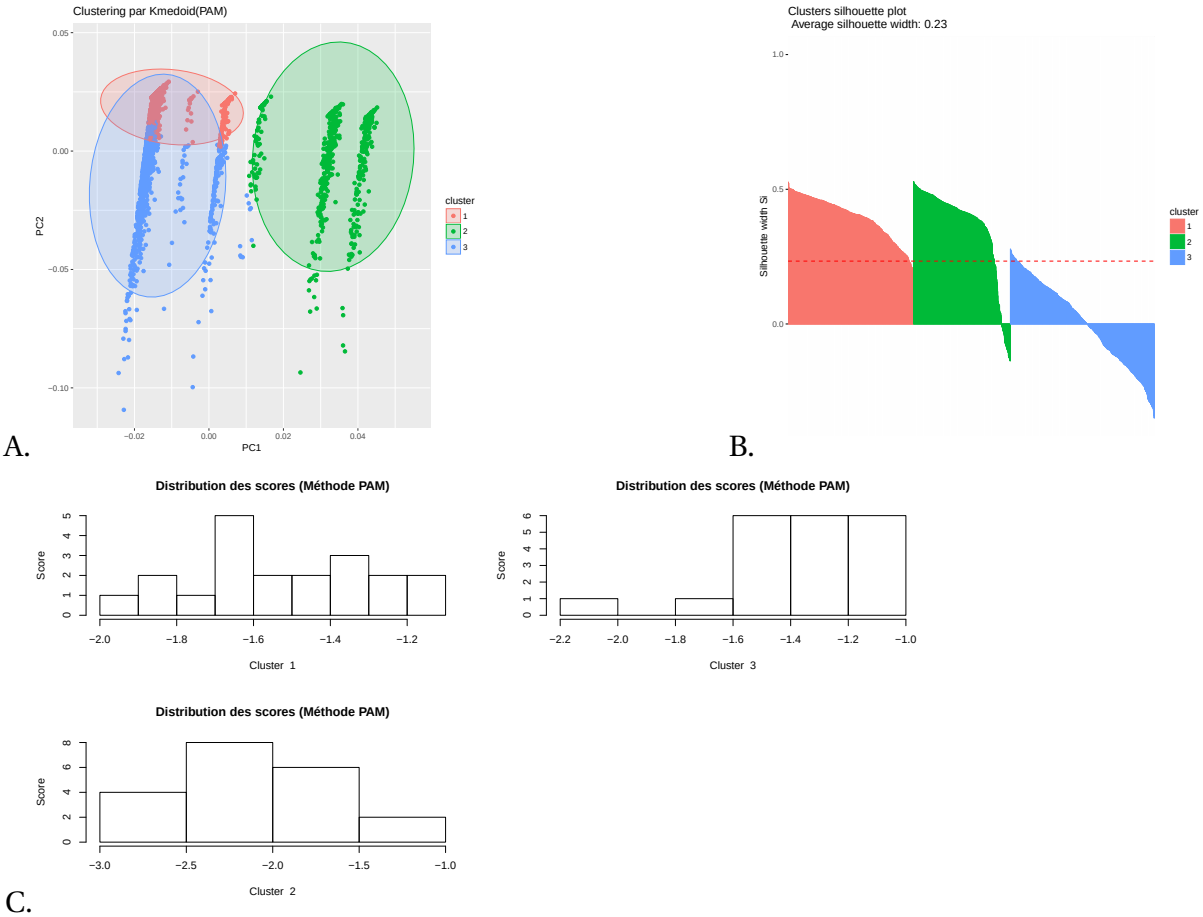
Comme dit précedement, il est nécessaire de spécifier le nombre de clusters à utiliser a priori dans la méthode de regroupement par Kmeans. La figure 2.A présente la méthode du "genou" utilisé pour déterminer le nombre de groupes optimales. L'emplacement d'une courbe (genou) dans le graphique est généralement considéré comme un indicateur du nombre approprié de groupes. La figure 2.B présente le graphique de regroupement par les Kmeans, on peut voir que les trois groupe ne sont pas totalement distinct. ON pourrais proposer un modèle avec 6 groupes. La figure 2.C présente la distribution des score dans chque cluster. Le score est calculer de la manière suivante:

$$log2(frequence_observé/frequence_attendu)$$

. La fréquence observé correspond à la fréquence de l'acide aminé dans le cluster. La

fréquences attendu correspond à la fréquence de l'acide aminée dans la base de données.

Figure 3: Regroupement par K-medoids. A: Graphique de Regroupement par les K-medoids. B:Graphique des silhouettes des cluster C: scores de chaque acides aminées dans chaque cluster

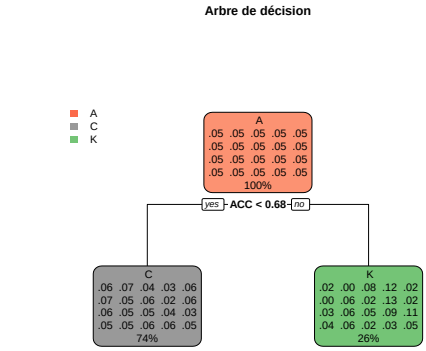


Le graphique **B** de la Fig.3 montre les silhouettes associés à chaque cluster. Une grand silhouettes *si* (proche de 1) suggère que les observations correspondantes sont très bien groupées, un petit *si* (autour de 0) signifie que l'observation se situe entre deux groupes, et les observations avec un *si* négatif sont probablement placées dans le mauvais groupe. Puisque la moyenne des *Si* est respectivement de 0.23 , les groupes ne sont pas identifiés.

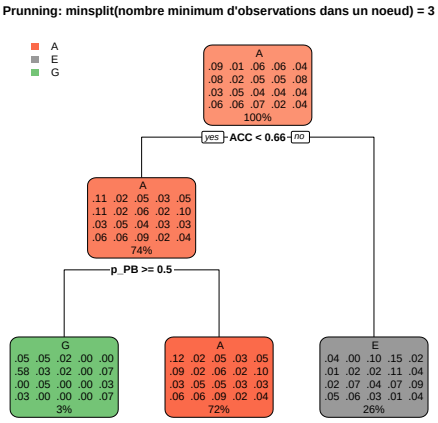
0.0.5 Prédiction

Une prédiction de la séquence dune globine (1a4fb.pdb) à l'aide d'un arbre de décision à été réaliser en partant de sa structure (fig4).

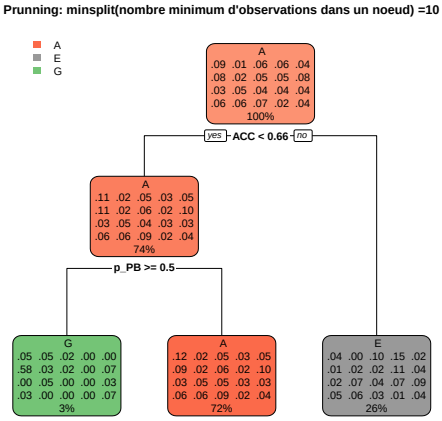
Figure 4: titre



A.



B.



C.

CONCLUSION

Ce qu'il reste à faire: Décrire les environnements (enfoui/accessible) en regardant la possition de cette environemment dans la structure, afin de caractériser les différents cluster. Comparer les mutations stabilisant une structures au résultats prédit par l'arbre de décision. Utiliser un réseau neuronal pour générer des scores de compatibilité 1D-3D. Réaliser l'apprentissage en partant de la séquences (1D) pour obtenir l'environnment associer (3D).

REFERENCES