Master 2 de Biologie Informatique 2020/2021

Parcours UFR - Sciences du Vivant (SDV)

Université de Paris

Projet Cours

Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features

Laëtitia HOLLIER & Lynda MESSAD DIT MAHTAL

Responsables: Catherine Etchebest & Jean-Christophe Gelly

I. INTRODUCTION

Cette étude se consacre sur l'analyse de la reconnaissance de modèles des bonds d'hydrogène et des caractéristiques géométriques des protéines. Autrement dit, nous nous intéressons à l'assignation des structures secondaires de protéines, à comprendre de quels motifs ces dernières sont composées. Ce rapport se base sur un article scientifique nommé ***Dictionary of Protein Secondary Structure:Pattern Recognition of Hydrogen-Bonded and Geometrical Features***,Kabsch W, Sander C, 1983. Cinq protéines issues de cet article ont été choisies. Grâce à la méthode DSSP (Hydrogen Bond Estimation Algorithm), la structure de protéines sélectionnées pourra alors être déterminée et les résultats obtenus seront comparés avec ceux de DSSP.

État de l'art:

Toutes les protéines fonctionnelles sont structurées en structures secondaires. De ce fait, déterminer la structure secondaire d'une protéine à partir de sa séquence en acides aminés est un des sujets les plus importants de la biologie structurale et qui permet d'étudier le repliement de la protéine et sa fonctionnalité. De nombreuses méthodes peuvent être utilisées telles que les propriétés des Carbones alpha. Une autre méthode est celle utilisée dans la DSSP qui est basée sur les liaisons hydrogènes et l'identification de paternes.

II. MATERIELS & METHODES

Pour ce projet, cinq protéines ont été choisies et leurs fichiers PDB ont été téléchargés (**Protein Data Bank**): ***1EST*** (hydroxylase), ***2SGA*** (anciennement 1SGA, protéase), ***2MHB*** (hémoglobine), ***6LDH*** (4LDH, oxydoréductase), ***7FAB*** (1FAB, immunoglobuline). Ces dernières ont été sélectionnées car elles possèdent plusieurs types de structures: elles peuvent être constituée que d'hélice, de feuillet bêta ou alors avoir des structures mixtes. Un script ***dssp.py*** a été élaboré afin de déterminer ces structures secondaires et de connaître la position des différents atomes les composant.

Pour exécuter ce code, une simple commande shell qui exécute un script python suffit. Elle prend comme arguments le nom du script et le fichier pdb :

python3 dssp.py file.pdb.

Il est impératif d'activer l'environnement conda "DSSPenv" avant l'exécution car celui-ci contient les packages essentiels au fonctionnement du programme (voir github).

De là, une première étape a dû vérifier la longueur des liaisons peptidiques afin d'éviter de travailler sur des protéines présentant des ruptures de liaisons. Par la suite, le programme ***Reduce*** a été utilisé afin de rajouter les atomes d'hydrogène au fichier pdb "file.pdb" car il est très rare que les fichiers pdb contiennent les atomes d'hydrogène et ils doivent t'être ajouter afin de pouvoir détecter les liaison hydrogènes. Cette étape se fait avec le script pdbHydrogene.py qui nous permet d'utiliser Reduce autant que module python et de vérifier son installation.

Ensuite, les atomes C, O, N et H impliqués dans les structures secondaires ainsi que les atomes CA caractérisant eux la chiralité et les courbures ont été récupérés des fichiers PDB.Ceci va nous permettre de déterminer les Hbond (liaisons d'hydrogène) pouvant se former entre plusieurs résidus. Une fois les distances entre les atomes calculées, il est primordial de vérifier l'énergie de liaison grâce à la formule de Coulomb ci-dessous avec q1=0.42e et q2=0.20e. Afin de détecter un Hbond entre deux résidus, cette Énergie doit être inférieure à 0.5kcal/mol.

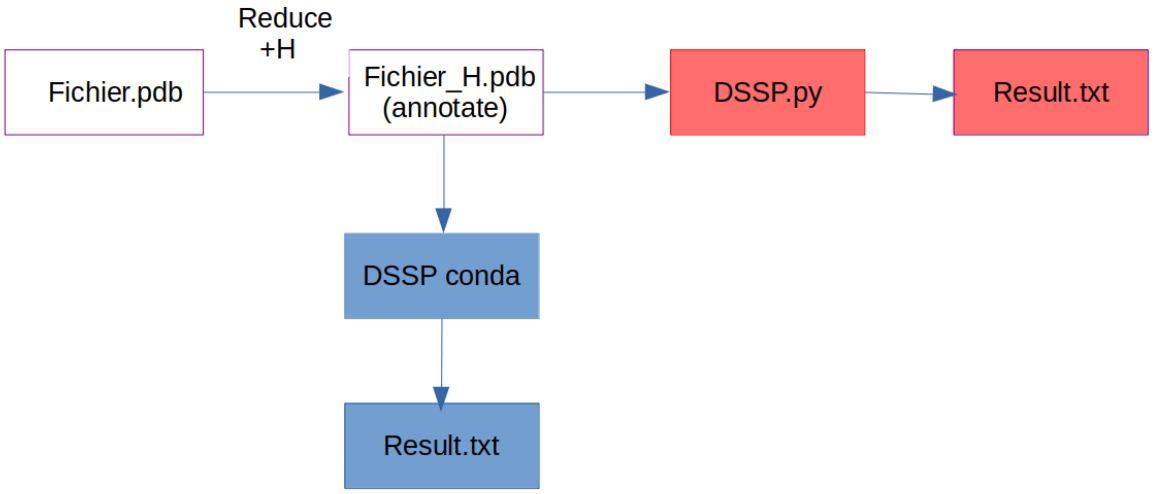
$$E = q_1 q_2 (1/r(\mathrm{ON}) + 1/r(\mathrm{CH}) - 1/r(\mathrm{OH}) - 1/r(\mathrm{CN}))*f$$

La distance entre deux atomes est obtenue à partir des coordonnées x,y et z obtenues par cristallographie aux rayons X et qu'on a extrait du fichier PDB de chaque protéine.

Dans ce projet, les pattern comme les n-turn: 3-turn, 4-turn, les 5-turn et les brins, ont pu être trouvés et de là les hélices présentes dans les peptides caractérisées. Les n-turn sont caractérisés par le nombre n qui correspond au nombre de résidus impliqués dans la liaison Hydrogène entre le résidu i et i +n. En utilisant les caractéristiques des n-turns on sait qu' une hélice est détecter des n-turns de la même classe sont successifs, c-à-d par exemple une 4-hélice (hélice alpha) est présente si on détecte un 4-turn à la position du résidu i et celui qui le précède. Pour avoir plus de certitude, les angles définissant la chiralité, les angles de courbures (bend) et les angles dièdres ont été déterminés en s'appuyant sur les distances et les caractéristiques de chaque structure secondaire. Les résultats sont présents dans les fichiers de sorties sur Github sous le nom de "*peptide_parsing_results.txt*" avec peptide le nom des protéines citées plus haut. Une comparaison des résultats obtenus dans ce travail et de DSSP a été réalisée (l'assignation complète des feuillets beta n'a pas été établie mais approximation des brins est possible avec ce même script, les fonctions y sont déjà implémenter).

Pour la gestion de ce projet, on s'est appuyé sur les principes et outils de la méthode Agile.

Stratégie d'implémentation de la méthode DSSP



III. RÉSULTATS

La longueur de ces liaisons peptidiques se situe entre 1 et 3 Angstrom pour les peptides ***1EST***, ***2SGA*** et ***6LDH***. Les protéines 2MHB et 7FAB n'ont pas pu être étudiées car d'après notre code, elles présentent des ruptures de liaisons peptidiques. Trois fichiers de sorties correspondant à nos 3 protéines précédentes ont été récupérés grâce à ce programme.

Les résultats obtenus pour les différentes structures sont les 3-turns,4-turns,5-turns, 3-hélice, 4-hélice, 5-hélice ainsi que les atomes nom structurés et qui n'interviennent donc pas dans les structures secondaires (nommé None). Les différentes étapes décrites auparavant dans Matériels et méthodes nous ont permis d'avoir un acheminement des

Notre première observation est que pour définir les structures secondaires d'une protéine plusieurs paramètres doivent être pris en compte. En effet, durant ce travail, il a été difficile de séparer avec certitude une hélice alpha d'une 5-hélice par exemple. Un résidu en particulier peut être pris entre deux turns ou deux "hélices" théoriquement par l'algorithme. Vient alors une étape délicate d'attribution des structures.

Une comparaison des structures secondaires trouvés par notre algorithme et celles de DSSP_conda (installer sur conda) a été établie pour la protéine 1EST.

structure	4-hélice (H)	3-hélice (G)	5-hélice (I)	4-tunr (>)	3-tunr (<)
DSSP	22	31	5	25	32
DSSP_conda	13	12	9	29	35

Tableau récapitulatif des nombres de structures trouvés pour la protéine 1EST par notre algorithme DSSP et le programme DSSP sous conda.

Pour cette protéine, notre algorithme a eu une surestimation des hélices. pour les brin on en a détecté que quelques uns (environ 3). Ces structures de feuillets trouvés sont dans un intervalle très large et on les retrouve aussi dans les feuillets trouvés par DSSP_conda. Mais la majorité n'ont pas été détectés. On a aussi trouvé des résiduels qui n'ont pas été structurés.

Le calcul de chiralité, des angles bend et des angles dièdres sont quasiment identiques aux résultats de DSSP_conda.

Le fichier de sortie nous permet de récupérer les atomes et leurs coordonnées atomiques, ainsi que toutes les structures secondaires trouvées dans la protéine.

Pour les hélices alpha, DSSP conda présente deux hélices alpha que on a réussi à détecter avec notre programme aux mêmes positions : DSSP_conda: 165-168 et 235-243 , notre algorithme : 165-168,235-242.

Pour les hélice 3-hélice, DSSP conda en situe quelque une à : 57-59 puis 173-175 et nous à 56-58 puis 172-176.

IV. CONCLUSION/DISCUSSION

L'assignation des structures secondaires en s'appuyant sur les liaisons hydrogènes est une procédure assez minutieuse car elle se fait sur des calculs de distances et d'énergie. Une des principale contrainte peut être la qualité des fichiers pdb, il n'est pas rare de trouver des fichiers pdb contenant des erreurs ou des imprécisions. Le passage par reduce est primordial et le résultat peut donc varier selon la version du reduce.

Ensuite, nous n'avons réussi à détecter les n-turns essentiels à la détection des hélices et des feuillets beta. Les hélices détectées sont comparables à celles détectées par DSSP avec quelques chevauchements entre les différents types d'hélices.

La détection des brins et des feuillets beta n'a pas été optimale sur notre algorithme et présente beaucoup de différences avec le programme de référence.

Néanmoins, nous avons rencontré des incohérences comme la présence d'un meme résidu dans différentes structures. Améliorer cet algorithme pour être plus précis et prendre en compte les autres variables telles que la chiralité est une suite logique de ce projet.