Rapport de projet – Sujet 3

Assignation des structures secondaire de protéines

Introduction

Ce sujet consiste à proposer une implémentation de la méthode DSSP (*Define Secondary Structure of Proteins*) en utilisant les méthodes de Bio-informatiques apprises dans le cadre du Master, et en appliquant les méthodes de gestion de projet usuelles en Bio-informatiques.

La méthode DSSP permet d'assigner des structures secondaires à une protéine, à partir des coordonnées atomiques de cette dernière. Cette méthode est publiée pour la première fois en 1983 par Kabsch et Sander (Institut de recherche médicale Max Planck d'Heidelberg)[1]. Notre implémentation devra se baser sur l'algorithme décrit dans cet article, avec pour objectif d'identifier les Hélices et les feuillets.

Etat de l'art

Déterminer la structure secondaire d'une protéine à partir de sa séquence d'acides aminés est une étape cruciale dans la compréhension de ses repliements et donc de sa fonctionnalité.

D'autres travaux[2], antérieurs à la méthode DSSP, permettaient déjà d'identifier les Hélices et les feuillets de manière automatisée. Mais seules les coordonnées des Carbones α étaient utilisés, ce qui impactaient considérablement la précision de la méthode.

Dans leur méthode, Kabsch et Sander développe un algorithme capable de déduire les structures secondaires d'une protéine donnée en se basant sur les patterns de liaisons hydrogènes existantes entre les groupements CO et NH des liaisons peptidiques.

Méthode

L'algorithme

Afin de statuer sur l'existence d'une liaison hydrogène entre deux résidus, l'algorithme de Kabsch et Sander calcule l'énergie électrostatique E (kcal/mol) entre le CO du donneur et le NH de l'accepteur, selon la formule suivante :

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

Avec : q1 = 0.42e ; q2 = 0.20e ; f=332 ; r(AB) = distance en Angstroms entre l'atome A et l'atome B.

La distance entre deux atomes est obtenue à partir des coordonnées x,y et z obtenues par cristallographie aux rayons X et disponibles dans le fichier PDB de la protéine. Seules les liaisons avec une énergie inférieure à -0.5 kcal/mol sont des liaisons hydrogènes.

Pour l'identification des structures secondaires en hélice ou en feuillet, l'algorithme se base sur des patterns de profils de liaison H. Ces patterns qui doivent se répéter au moins une fois de façon successive sont basés sur la distance entre les résidus impliqués dans les liaisons hydrogènes.

Pour identifier une hélice, il faut identifier la répétition successive du même n-turn. Un n-turn se caractérise par le nombre n qui sépare les deux résidus impliqués dans la liaison H : résidu i et i+n. Il existe 3 types de n-turn en fonction du sous-type d'hélice (n= 3, n=4, et n=5)

Pour identifier les feuillets, la gymnastique est plus complexe. Pour simplifier, nous considérons que la répétition successive de liaisons hydrogènes entre 2 résidus significativement éloignés correspond à un feuillet (n>10).

Le code (cf. annexe1)

L'implémentation que nous proposons se base sur un script en python, avec un prétraitement préalable du fichier PDB par l'outils reduce. La construction du code vous est détaillée en annexe 1.

La gestion de projet (cf. annexe2)

Nous avons choisi de gérer ce projet en utilisant les principes et outils de la méthode Agile. Veuillez trouver plus de détails en annexe 2.

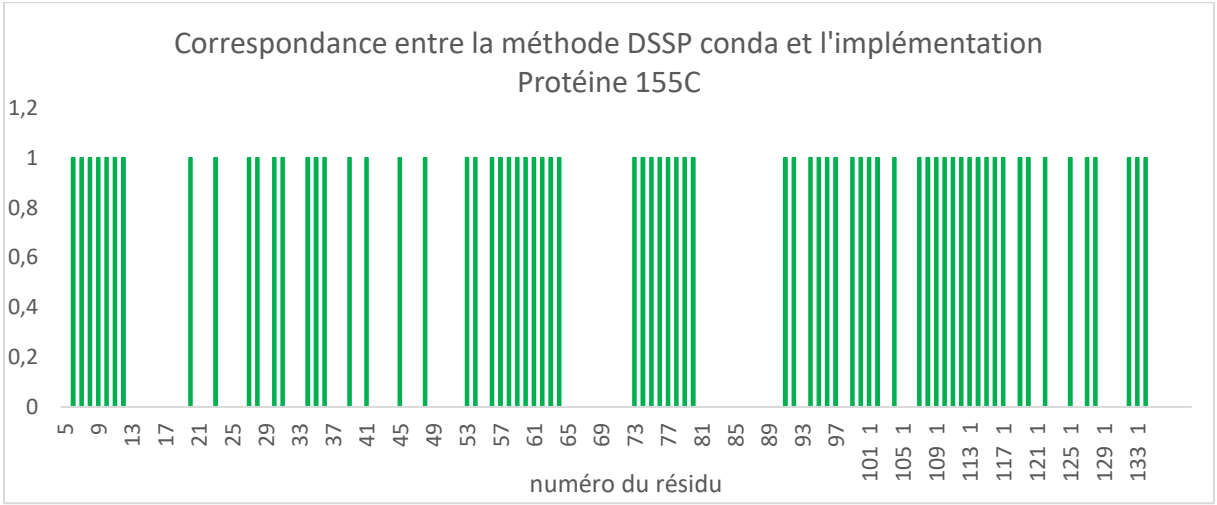
Résultats

Choix des protéines (fichiers PDB disponibles sous https://www.rcsb.org/)

Nous choisissons d'analyser la Pancreatic Trypsin Inhibitor (3PTI), Le Cytochrome C550 (155C) et l'Adenylate Kinase(2ADK).

Corrélation avec la méthode DSSP de référence

Seuls les résultats obtenus pour la protéine 155C vous seront présentés ici. Les autres seront présentés lors de l'oral. La comparaison entre la méthode DSSP de référence et notre proposition d'implémentation montre un taux de correspondance de **54%** pour cette protéine. Vous pouvez observer ci-dessous une représentation des correspondances en structure secondaires par résidu.



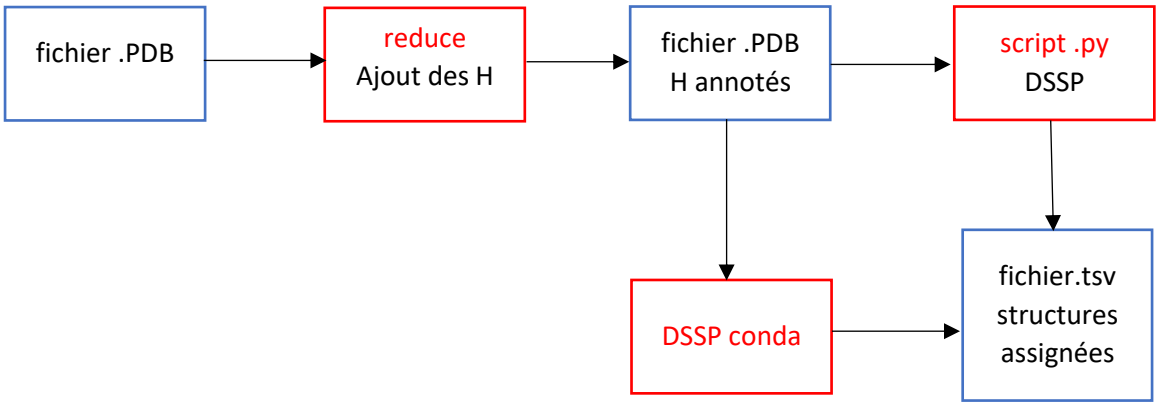
Conclusion

La méthode d'implémentation que nous proposons permet d'identifier les liaisons hydrogènes d'intérêt et d'assigner des structures secondaires. Néanmoins nous relevons certaines incohérences telles que l'implication du même résidu à la fois dans une structure en Hélice et dans une structure en feuillet. Cela peut expliquer que le taux de correspondance obtenu pour la protéine 155C ne soit pas optimale.

Nous avons comme objectif d'améliorer notre méthode afin d'augmenter le taux de correspondance pour la protéine 155C, et analyser toutes les protéines du panel.

Annexe 1 : Stratégie d'implémentation de la méthode DSSP

Approche globale



Script python

fichier .PDB
H annotés

```
# Modules utiles
import sys
import os
import math
```

```
def extraction_atome(fichier_pdb):
    """ Lit un fichier pdb et renvoie les coordonées de tous les atomes
        impliqués dans la liaison peptidique (C, O, H, N).
```

```
def distance(res1, res2):
    """ Permet de calculer la distance entre deux points
        ayant leurs cooronnées x, y et z
```

```
def liaison_hydrogene(dico):
    """ Permet de determiner l'existence d'une liaison
       hydrogene (LH) entre le CO du residu i et le NH du i+n,
       ou entre le NH de i et le CO de i-n, "n" allant de 1 au
       nombre total de residu de la chaine,en se basant sur
       l'energie electrostatistique (E) et en appliquant la
       condition: presence de LH si E < 0.5kcal/mol
```

```
def pattern(liaison_hydroge):
    """ Permet d'attribuer les differents motifs, selon
        la position des atomes impliqués dans la LH
        3<=|i-i+n|<=5: il s'agit de turns
        |i-i+n|<3: pas de stucture, d'apres notre comprehension
        de l'article ("None")
        |i-i+n|>5: liaison eloignée, suspicion de brin beta
```

```
def structure(pat):
    """ Permet d'assigner le type de stucture secondaire
    helice (H) ou feuillet beta (B), suivant la localisation
    de la lh et son nombre de repetition (succession)
```

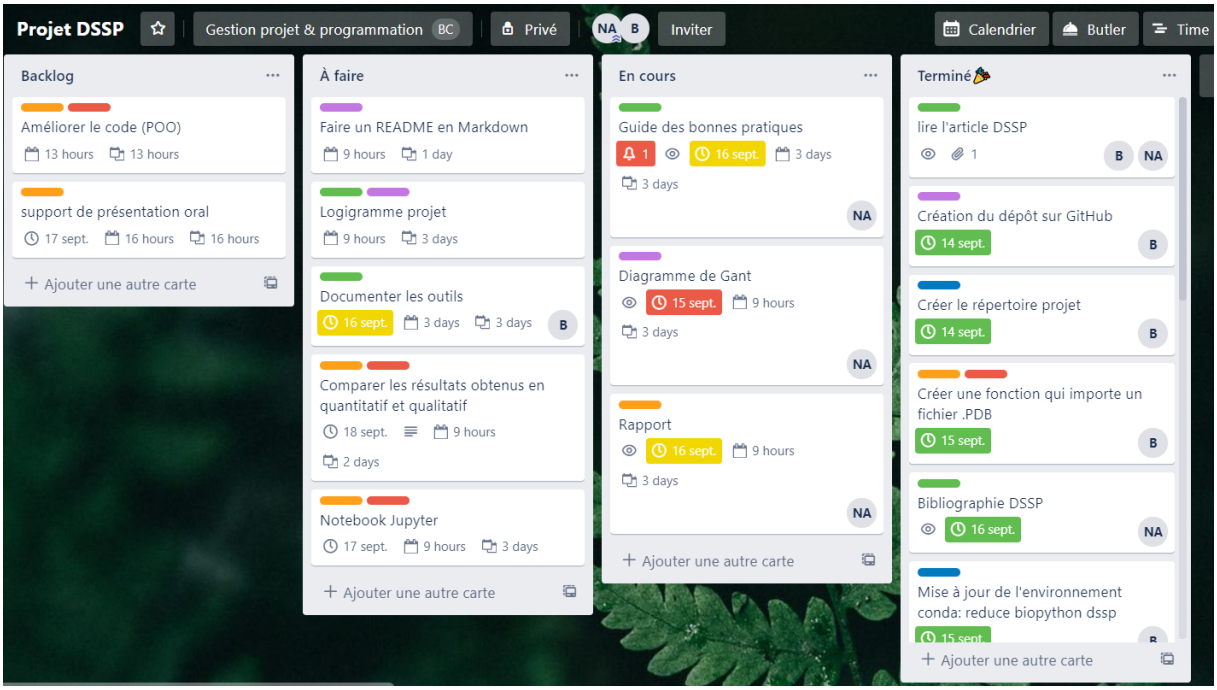
```
def fichier_dssp(recap, struc):
    """ Permet de generer un fichier de sortie avec 2 colonnes:
        le couple de residu, et le type de structure secondaire
        dans lequel il est impliqué (ici en l'occurence H et B et nturns)
```

Environnement conda
Dssp.yml

Annexe 2 : Gestion de projet

Outils Agiles utilisés dans le cadre de ce projet :

- Kanban via l'outils en ligne Trello : https://trello.com/b/9AYbYOnS



- Réunions périodiques et fréquentes (2 par jours) pour faire le point sur les taches en cours et les points bloquants.
- Définition d'une durée de sprint pour la livraison du script (1 jour). A chaque livraison le script doit être fonctionnel.

Références

(1) W. Kabsch & C. Sander, 1983, Biopolymers, Vol. 22,2577-2637.

(2) Levitt, M. & Greer, J. (1977) J. Mol. Biol. 114,181-239.

https://anaconda.org/salilab/dssp

https://anaconda.org/conda-forge/reducer

https://swift.cmbi.umcn.nl/gv/dssp/

https://www.rcsb.org/