Projets courts : Calcul de la surface accessible au solvant (ASA)

Hocine Meraouna

Hager Suzanne Elharty

Sleheddine Kastalli

15 June 2021

Introduction

La surface des acides aminés des protéines accessible au solvant est un paramètre déterminant pour l'étude du repliement des chaînes polypeptidiques et le calcul de leur stabilité. On cherche à implémenter l'algorithme de Skrake & Rupley [4] permettant de calculer la surface accessible au solvant (ASA) un descripteur important dans l'étude des protéines (par exemple : la prédiction de la structure d'une protéine)[1][3]. L'obtention de ce descripteur est aujourd'hui obtenu par des programmes informatiques tels que Naccess [2], qui sera utilisé plus tard comme comparatif.

Principe

On parle de surface accessible au solvant lorsque qu'une sonde (dans notre étude représentée par une sphère) est accueillie à la surface de la molécule. Les atomes de la molécule de soluté (plus précisément de la protéine) seront représentés par des sphères, et on cherchera à savoir si ces derniers peuvent accueillir ou non cette sonde. Pour modéliser ces sphères, chaque atome de la protéine sera représenté par une sphère de N points donnés par l'utilisateur, et positionnées de façon uniforme à une distance égale au rayon de Van der Waals de l'atome. En effet, les protéines sont décrite comme un ensemble de sphères de rayon de Van der Waals solvatées [4]

Pour cela, nous simulerons le roulement d'une molécule d'eau (i.e la sonde), autour des sphères de Van der Waals des atomes de la protéine. Le centre de la molécule d'eau est à une distance du centre de chaque atome égale au rayon de Van Der Waals de la molécule d'eau plus le rayon Van Der Waals de l'atome de la protéine [4]. Le trajet décrit par le centre de la molécule d'eau délimite la surface accessible des atomes de la protéine (cf. Figure 1)

Objectif

Calcul de la surface accessible au solvant qui est définie comme le lieu des centres d'une sphère modélisant une molécule d'eau (de rayon 1, 4 Å) lorsque celle-ci parcourt l'ensemble de la surface de la protéine.

Matériels et Méthodes

Matériels

Python:

Ce programme a été codé en Python3 et compilé sous environnement Linux.

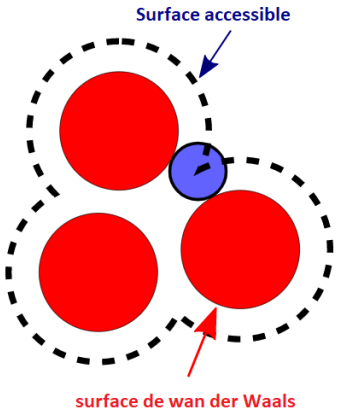

Figure 1: *La surface de Van der Waals tel que donné par le rayon atomique est représenté en rouge. La surface accessible est tracée en traits interrompus et est créé en suivant le centre de la sphère de la sonde (en bleu) qui roule le long de la surface de Van Der Waals. Le rayon de la sonde est de 1,4 Å.*

Librairies Python :

Pour exécuter ce programme nous avons utilisé des librairies Python :
- Pandas
- Numpy
- Argparse
- Matplotlib
- Scipy
- Progressbar

Données brutes :

Les données brutes utilisées pour ce programme ont été extraites de la banque de données sur les protéines ou PDB (Protein Data Bank).
Il s'agit ici de deux fichiers PDB :

- 3i40.pdb : structure de l'insuline humaine
- 1b0q.pdb : Analogue peptidique de l'alpha-mélanotropine

Trello :

Pour le suivi des taches a effectuer, nous avons utilisé l'application Trello, qui offre une interface "user friendly" pour permettre à chacun de bien attribuer sa tâche et l'état d'avancement. Pour cela nous nous sommes basés sur les méthodes AGILE, où le travail est décomposé en plusieurs tâches, et 3 catégories à savoir : "à faire", "en cours" et "terminé". Chaque jour nous mettions à jour notre avancement. Dans la figure en Annexe on peut voir les tâches terminés,

en cours ou encore à faire à la date du 14 juin 2021 (cf. Annexe Figure 3)

Naccess :

Naccess est un logciel gratuit fonctionnant sur Linux et Windows. Il permet d'obtenir l'ASA atomiques et résiduelles d'une protéine mais également des acides nucléiques à partir d'un fichier PDB.

Méthodes

Le programme va commencer par lire un fichier PDB et extraire les coordonnées de chaque atome dans un Data Frame grâce à la fonction coord().

La seconde étape consiste à construire une matrice de distance, qui contient les distances de chaque atome avec tous les autres atomes du fichier PDB. Cette matrice a été construite grâce à la fonction atom_dist_matrix() prenant en argument le Data Frame de coordonnées de chaque atome du fichier PDB.

Ensuite, nous avons déterminé les atomes de résidus spatialement voisins, à savoir ceux qui sont à une distance inférieur ou égal au diamètre d'une molécule d'eau plus celui du résidu le plus grand possible, c'est à dire le tryptophane dans notre étude. Pour ce faire nous avons utilisé la fonction threshold_dict().

A partir du dictionnaire contenant les atomes de résidus voisins, nous avons créé une sphère autour de chaque atome grâce à la fonction sphere() qui prend en argument un entier définissant le nombre de points qui constitue la sphère. Ensuite grâce à la fonction spheres_dist() nous avons calculé la distance de chaque point d'une sphère avec le centre de la sphère voisine afin de déterminer les point de la sphère exposés au solvant. Si la distance entre un point et le centre est supérieure à la somme du rayon de Van Der Waals de l'atome et du diamètre de la molécule d'eau, alors le point sera considéré comme exposé au solvant.

La dernière étape consiste à calculer la surface de chaque atome exposé au solvant. Pour cela, nous calculerons un ratio (ratio = nombre de points exposés d'une sphère d'atome /le nombre total de point constituant cette même sphère). Ce ratio sera multiplié par la surface de la sphère de chaque atome en utilisant cette formule :

$$S = ratio \times 4 \times \pi \times (rayon\ de\ VDW)^2 \tag{1}$$

Pour finir, pour calculer la surface accessible au solvant de la protéine, nous ferons la somme des surfaces accessible de chaque atome d'un même résidu.

Puis la somme de la surface accessible de chaque résidu pour obtenir la surface accessible au solvant de toute la protéine.

Résultats

Table 1: Valeurs de la surface accessible au solvant pour chaque atome de la protéine 1b0q

résidu	Notre implémentation	Naccess
CYS	85.27549	111.99
GLU	120.727667	151.35
HIS	144624399	230.92
ARG	184.331165	191.44
TRP	200.036144	205.68
CYS	80.824848	98.43
LYS	183.349323	127.48
PRO	119.65465	111.69
VAL	100.306202	213.90
Total	1231.58	1442.9

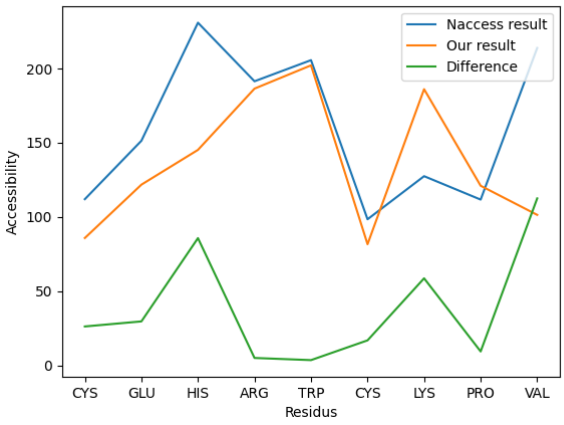

Figure 2: *Courbes représentants les valeurs d'accessibilités des résidus de 1b0q obtenues avec notre code (orange) et ceux données par Naccess (bleu) et de la différences entre les deux (vert)*

On peut voir sur la **Figure 2** et la **Table 1** que les résultats pour la majorité semblent proches avec les valeurs données par Naccess notamment pour les résidus ARG4, TRP5, CYS6 et PRO8. En effet on peut voir sur la courbe de différence que les valeurs sont pour la plupart très faibles à part pour les résidus HIS3, LYS7 et VAL9.

Conclusion & Discussion

En conclusion, nous pouvons dire que l'algorithme de Shrake & Rupley,donne des résultats proches de ceux obtenus avec Naccess qui est implémenté à partir de l'algorithme de Lee & Richard (1971).

Optimisation du code :

Pour l'exécution de ce code nous avons utilisé deux protéines de tailles différentes la 3i40 qui est constitué de 53 résidus et de 448 atomes et la 1b0q qui est constituée de 9 résidus et de 145 atomes. Le temps de calcul pour la protéine 3i40 et pour une sphère de 96 points est de 14 secondes, et avec 1000 points 2 minutes et 21 secondes. A savoir que notre premier code prenait 33 minutes et 56 secondes pour donner les resultats pour une sphère de 100 points.
Nous avons donc optimisé le temps de calcul du programme, il est aussi a noter qu'il faut privilégier un nombre de points d'environ 96 pour un meilleur résultats, lorsque le nombre de points est trop élevé le résultat est moins précis. De plus le code a aussi été amélioré suivant les règles de PEP8 (pycodestyle).

Perspectives:

Comme perspectives, nous pourrions allez plus loin, en calculant à l'instar de NACCESS la surface accessible au solvant des atomes de la chaine principale et celle des atomes de la chaine latérale.

Bibliography

[1] html?fbclid=IwAR0aERMICUZQ5NPDK6hPrEXsMwM2aURycSqhHBlIIm7DVppvcXUkwkLa1qk.

[2] http://www.bioinf.manchester.ac.uk/naccess/.

[3] Definiton de l'asa. https://boowiki.info/art/chimie-computationnelle/aire-de-surface\-accessible.html?fbclid=IwAR1WMGDKX6VmkrysOGmB1dlDOQhZoNzC6j-_46aah545y4QGB6ogaGftorQ.

[4] Shrake & Rupley. Environment and exposure to solvent of protein atoms, lysozyme and insulin. *Journal of Molecular Biology*, 1973.

Annexe

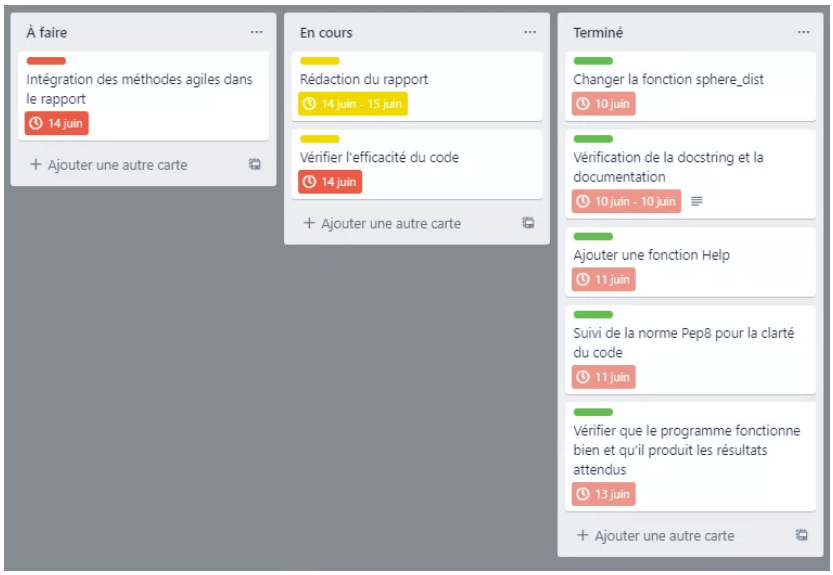

Figure 3: Schéma de notre organisation suivant les méthodes Agiles