***Rapport projet : Calcul de la surface accessible au solvant d’une protéine***

*GOULANCOURT Rebecca*

*Lien Github : https://github.com/Rebbekkah/Projet.git*

1. Introduction

Dans ce projet nous chercherons par l’intermédiaire d’un script python à déterminer la surface accessible, absolue et relative, d’une protéine par le solvant à partir de ses coordonnées qui seront extraites à partir d’un fichier PDB. Le but de ce projet est d’avoir une idée de la structure de la protéine, de l’exposition ou l’enfouissement de ses résidus mais aussi pouvoir déterminer avec quels résidus le solvant pourrait rentrer en contact. Cela pourrait permettre de comprendre les perturbations structurales liées au solvant.

Ainsi le but final serait d’obtenir une surface exposée au solvant, ou surface accessible. À l’issue du calcul de la surface accessible au solvant, nous pourrons comparer nos résultats avec ceux obtenus grâce à NACCESS.

On cherhce à savoir si un mol d’eau passe….. Nous prendrons comme référence de solvant une molécule d’eau (H20) dont la taille est de 1.4 Ä2.

1. Matériel et méthodes
2. Matériel

* La Proteine Data Bank (PDB)

La PDB est une banque de données protéique permettant de regrouper toutes les informations obtenues sur ces protéines, comme par exemple leurs structures 3D. Toutes ces informations peuvent être téléchargées en format PDB (*identifiant\_de\_la\_protéine.pdb*). Dans notre cas c’est en téléchargeant ce fichier PDB que l’on peut récupérer les coordonnées atomiques d’une protéine d’intérêt.

* Python 3.9 et les bibliothèques

La version de python que nous avons utilisée est la version 3.9. Les bibliothèques (ou encore modules ou librairies) sont des fonctions open sources téléchargeables à l’aide de la commande « import ». Elles permettent de palier des problèmes liés à un code trop lourd à écrire en important un code déjà écrit. Dans notre cas nous avons importé les modules pandas, numpy, sys, scipy, matplotlib, mpl\_toolkits et collections.

La bibliothèque pandas permet la manipulation et l’analyse de données et numpy aide à la manipulation des vecteurs, matrices et tableaux. Ces deux bibliothèques sont complémentaires et seront clés dans notre script car les données et output seront lus sous forme de DataFrame.

Le module sys permet d’accéder à des variables de l’interpréteur, dans notre cas le nom du fichier en entrée tandis que scipy (extension de numpy) est un ensemble de bibliothèques servant à résoudre des formules et problèmes mathématiques.

La librairie matplotlib est une librairie complète donnant accès à des représentations graphiques, alors que mpl\_toolkits permet la création de graphes complexes en 3 dimensions comme par exemple des sphères. Enfin le module collection permet rapidement de créer un « dictionnaire de dictionnaire ».

* Github

Pour ce qui est de la gestion du script celui-ci a été régulièrement déposé sur Github. C’est une plateforme essentiellement pour la collaboration où l’on peut s’inscrire pour déposer ses fichiers.

* Environnement utilisé

Pour ce projet j’ai utilisé l’environnement de base auquel j’ai installé les librairies citées plus haut lorsqu’elles n’étaient pas déjà présentes.

* FreeSASA

FreeSASA est une librairie écrite en langage C et accessible par des commandes en C++ dans le terminal. Elle permet d’obtenir la surface accessible d’un solvant de molécules biologiques par l’intermédiaire de la méthode de Lee & Richard’s. Une démo est disponible sur le site officiel de FreeSASA (<https://freesasa.github.io/>) où l’on peut rentrer l’identifiant PDB d’une protéine et obtenir en output sa surface accessible au solvant. Nous pouvons ainsi comparer des résultats avec ceux obtenus via FreeSASA.

1. Méthodes

En premier lieu nous allons choisir dans la banque PDB une molécule. Ensuite il faudra récupérer le fichier correspondant en format .pdb et le stocker dans le même dossier que le script. Pour le faire tourner il faudra l’appeler et ensuite passer en argument l’identifiant de la molécule étudiée (en format .pdb).

Le script devra être construit de sorte à respecter les étapes clés présentées ci-dessous.

i) Parsing des fichiers PDB afin de récupérer les coordonnées atomiques de chaque atome présent dans la protéine.

ii) Construire une matrice de distance des atomes deux à deux de la protéine à partir de leurs coordonnées (méthode euclidienne). On obtiens alors une matrice de la distance séparant chaque atome. Ensuite on définit un seuil à partir duquel les atomes ne seront plus considérés comme voisins l’un de l’autre. Ce seuil est définit essentiellement par le plus grand Rayon de Wan der Waals atomique (ici le phosphore P) tel que : seuil = rayonWdW\_max \*2 +1.4. Si la distance séparant les couples d’atomes est inférieure au seuil alors on pourra les considérer voisins.

iii) On récupère ensuite pour chaque atome une liste de ses voisins et on en fait un « dictionnaire de listes ».

iv) On étudie ensuite un atome à la fois et on place sur chacun de ses voisins une sphère dont le diamètre correspond au Rayon de Wan der Waals de l’atome voisin. Pour créer la sphère il s’agit de créer une nouvelle matrice qui contient de nouvelles coordonnées tel que :

|  |  |  |
| --- | --- | --- |
| Atome voisin | Coordonnées initiale | Coordonnées sur la sphère |
| 0 | {x0, y0, z0} | {x0 + rWdW\_atome, y0 + rWdW\_atome, z0 + rWdW\_atome} |
| 1 | {x1, y1, z1} | {x1 + rWdW\_atome, y1 + rWdW\_atome, z1 + rWdW\_atome} |
| … | {xn, yn, zn} | {xn + rWdW\_atome, yn + rWdW\_atome, zn + rWdW\_atome} |

On dispose alors de trois points sur lesquels on peut tracer une sphère. On pourra y placer par la suite un nombre de points désiré. Plus il y aura de points sur la sphère plus les résultats seront précis car un nombre de points élevé permet de couvrir et calculer sur une plus grande surface et éviter de biaiser les résultats lors des calculs de surface accessibles.

v)