
Projets courts : Sujet n° 2 : Calcul de la surface accessible au solvant



Trinôme	Hocine Merouana Hager Elharty Sleheddine Kastalli
Année universitaire	2020/2021 Université Sorbonne Paris 7

Introduction

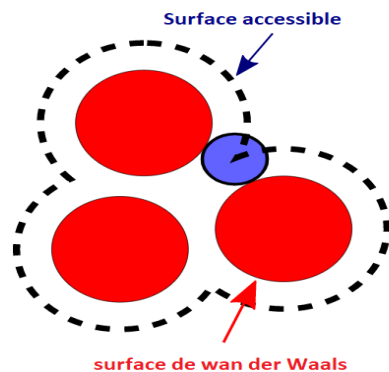
La minimisation de la surface accessible aux solvants, appliquée à l'interaction entre les surfaces des protéines constituent une contrainte puissante pour la prédiction de la conformation de ces assemblages macromoléculaires (Chothia & Janin, 1975 ; Wodak & Janin, 1978 ; Richmond & Richards, 1978 ; Cohen et al., 1979,1982). Le site de la surface d'une molécule de soluté accessible aux solvants est la surface la plus proche des centres des molécules de solvant où le soluté et le solvant sont représentés par des sphères dures. Calculée, cette surface est définie comme la surface de van der Waals enveloppe de la molécule de soluté dilatée par le rayon de la sphère de solvant sur chaque centre d'atome soluté (Lee & Richards, 1971 ; Richards, 1977).

Le modèle des sphères de van der Waals est largement utilisé pour l'interprétation des structures des protéines et pour la corrélation avec les paramètres physico-chimique. L'algorithme le plus simple est probablement la méthode des incréments : Un incrément de volume est associé à chacun des n atomes, et le volume total est la somme des n contributions.

Principe

Rouler une molécule d'eau, représentée par une sphère, autour des atomes, également représentée par une sphère dont le rayon est celui du rayon de Van Der Waals de chaque atome, de la protéine.

Le trajet décrit par le centre de la molécule d'eau délimite la surface accessible des atomes de la protéine.



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.

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.

Matériel et Méthodes

Matériel :

Python :

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

Librairies Python :

Pour exécuter ce programme nous avons utilisé des librairies non-standards pour 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 : HUMAN INSULIN
- 1b0q : DITHIOL ALPHA MELANOTROPIN PEPTIDE CYCLIZED VIA RHENIUM METAL COORDINATION

Conda

Conda est un puissant gestionnaire de packages et d'environnement utilisé avec les commandes de ligne de commande à l'invite Anaconda pour Windows, ou dans un terminal pour macOS ou Linux.

Git et Github

Git est un logiciel de gestion de versions décentralisé.

GitHub est un service web d'hébergement et de gestion de développement de logiciels.

Git et Github ont été utilisés dans ce projet pour faciliter le travail en groupe.

Tous les codes, ainsi que la documentation de ce sujet ont été déposés dans le dépôt Github ci-dessous :

https://github.com/hocinebib/Projet_Court_H-H/tree/master

NACCESS

Naccess est un programme autonome qui calcule la surface accessible d'une molécule à partir d'un fichier au format PDB (Protein Data Bank).

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 atomes du fichier pdb. Cette matrice a été construite grâce à la fonction

atom_dist_matrix() prenant en argument le *dataframe* de coordonnées.

Ensuite, nous avons déterminé les atomes spatialement voisins, afin de vérifier si le diamètre d'une molécule d'eau passe entre ces deux atomes. Pour ce faire nous avons utilisé les fonctions : **threshold_dict()** et **spheres_dist()**.

A partir du dictionnaire contenant les atomes voisins, nous avons créé une sphère autour de chaque atome grâce à la fonction **sphere()**. En effet cette fonction permet de calculer la distance de chaque point d'une sphère avec chaque point des sphères voisines afin de déterminer les points de ces dernières exposés au solvant. Si la distance entre deux points 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

$$Ratio = \frac{N \text{ points exposés}}{N \text{ points total d'une sphère}}$$

Ce ratio sera multiplié par la surface de la sphère de chaque atome en utilisant cette formule : $S = Ratio \times 4 \times \pi \times (r_{VDW})^2$

Finalement, pour calculer la surface accessible au solvant, 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 :

Comparaison avec NACCESS

```
naccess: using default vdwr.radii
naccess: using default STD FILE
(base) meraouna@meraouna-Satellite-C50-A545:~/Téléchargements$ cat 1b0q.
1b0q.asa 1b0q.log 1b0q.pdb 1b0q.rsa
(base) meraouna@meraouna-Satellite-C50-A545:~/Téléchargements$ cat 1b0q.rsa
REM Relative accessibilities read from external file "/home/meraouna/Téléchargements/naccess2.1.1/standard.data"
REM File of summed (Sum) and % (per.) accessibilities for
REM RES _NUM All-atoms Total-Side Main-Chain Non-polar All polar
REM ABS REL ABS REL ABS REL ABS REL ABS REL
RES CYS A 1 111.99 83.4 42.12 43.5 69.88 186.2 44.80 45.7 67.19 184.8
RES GLU A 2 151.35 87.9 36.85 101.6 14.50 38.7 56.35 93.5 95.01 84.9
RES HIS A 3 230.92 126.3 67.52 113.9 63.40 177.1 142.47 146.7 88.45 103.2
RES ARG A 5 191.44 80.2 46.26 72.7 45.17 120.4 87.36 112.3 104.08 64.7
RES TRP A 6 205.68 82.5 100.55 94.9 5.12 13.4 171.19 90.3 34.49 57.8
RES CYS A 7 98.43 73.3 71.98 74.4 26.45 70.5 75.85 77.5 22.58 62.1
RES LYS A 8 127.48 63.5 09.84 67.3 17.64 47.0 65.59 56.3 61.89 73.5
RES PRO A 9 111.69 82.0 09.64 91.4 2.05 12.6 109.64 90.6 2.05 13.5
RES VAL A 10 213.90 141.2 63.80 143.3 50.11 134.8 185.56 160.7 28.34 78.8
END Absolute sums over single chains surface
CHAIN 1 A 1442.9 1148.5 294.3 938.8 504.1
END Absolute sums over all chains
TOTAL 1442.9 1148.5 294.3 938.8 504.1
(base) meraouna@meraouna-Satellite-C50-A545:~/Téléchargements$
```

Figure 1 : Résultats NACCESS avec la protéine 1b0q

```
hage@GLAPTOP-F1030B: ~/src/chem/hage/OneDrive/projet_cstas/Projet_Cstas_H_H
Python 3.8.3 (default, May 19 2020, 18:47:26)
[CC 7.3.0] :: Anaconda, Inc. on linux
Type "help", "copyright", "credits" or "license()" for more information.
>>> import surface2 as wc
>>> atcwc.comb(1b0q.pdb)
>>> atcwc.atom_dist_mtx(atc)
>>> mntwc.threshold(dict(mtx))
[0.0] [=====] spherewc.link_fct(atc, mntwc, 100)
[=====] time elapsed :12 min 9 sec
>>> ratiwc.ratio(spherewc, 100)
>>> surfwc.acf_surf(rati, atc)
>>> asa = wc.res_surf(surf, atc)
>>> asa
['CYS': 117.73196149237984, 'GLU': 174.3376699676473, 'HIS': 200.3454305398364, 'ARG': 278.8520998749846, 'TRP': 200.17094113914544, 'LYS': 273.4042308575278, 'PRO': 171.00491593308368]
>>>
>>> for keys in asa:
...     print(keys, asa[keys])
...
CYS 117.73196149237984
GLU 174.3376699676473
HIS 200.3454305398364
ARG 278.8520998749846
TRP 200.17094113914544
LYS 273.4042308575278
PRO 171.00491593308368
```

Figure 2 : Valeurs de la surface accessible au solvant pour chaque atome de la protéine 1b0q

La surface accessible totale pour la protéine 1b0q est de **1505.047 Å²** comparée à celle de NACCESS qui est de **1442.9 Å²**.

Discussion & conclusion :

1. Problèmes d'optimisation du code

Pour l'exécution de ce code nous avons utilisé deux protéines de taille très différentes la 3i40 qui est constitué de 53 résidus et de 448 atomes et la 1b0q qui est

constituée de 10 résidus (peptide) et de 145 atomes.

Le temps de calcul pour la protéine 3i40 et pour une sphère de 30 points est d'environ 6 min. Pour 100 points le temps de calcul est de 33 min et 56 secondes. Il est à noter que plus on augmente le nombre de points, plus les résultats sont précis, mais le temps de calcul devient trop long.

Un autre critère qui influe le temps de calcul est le choix de la protéine, en effet pour la protéine 1b0q (10 résidus), le temps de calcul pour une sphère de 30 points est de 5 min, et pour la sphère de 100 points de 12 min et 9sec.

Pour conclure, les résultats semblent cohérents, cependant, le code aurait pu être optimisé au mieux pour minimiser le temps de calcul.

2. Perspectives

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

Références

1. F. M. Richards, Annu. Rev. Biophys. Bioeng. 6,151 (1977).
2. Chothia. C. & Janin, J. (1975). Xutiuw (London). 256. 705 708
3. Cohen. F. E.. Stern berg. M. J. E. & Taylor, W. R. J.Mol. Bio. 156. 821-862.
4. <https://boowiki.info/art/chimie-computationnelle/aire-de-surface-accessible.html?fbclid=IwAR1WMDGKX6Vmkr50GmB1dDOQhZoNzC6j-46aah545y4QGB6ogaGftorQ>
5. <http://biochimej.univ-angers.fr/Page2/TexteTD/9BioinfoSTRUCTURALE/2BioinformatiqueLicence/1BioinfoLicenceS6.html?fbclid=IwAR0aERMICUZQ5NPDK6hPrEXsMwM2aURycSqHBIllm7DVppvcXUkwkLa1qk>