PROJET COURT M2 BIB juin 2020

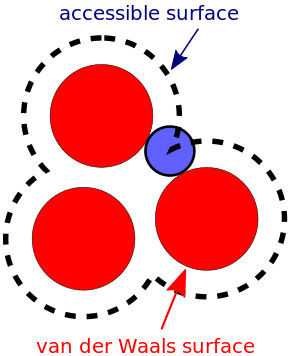
Sujet 2 : Calcul de la surface accessible au solvant d’une protéine

Julie GORSE BOGOIN

Lien GitHub vers le projet :

https://github.com/juugeebee/Projet\_court\_S2\_M2

1. **Introduction :**
   1. **Présentation du sujet**

L'algorithme Shrake-Rupley (2) est un procédé numérique qui dessine un maillage de points équidistants entre chaque atome d’une molécule et utilise le nombre de points qui sont accessibles au solvant afin de déterminer l'aire de surface accessible au solvant. Cette donnée est utilisée pour déterminer la fonction d’une protéine dont la séquence en acides aminés est connue. Un des paramètres utilisé dans cette détermination est la capacité de deux ou plus monomères à se lier entre eux. Il est donc nécessaire d’identifier les atomes participant à l’interface avec les solvants, afin de prédire la formation de complexes protéiques. Le modèle de Van der Waals est une méthode de représentations des énergies entre chaque atome (figure 1). Chacun d’eux est représenté par une boule dont le rayon dépend de ses propriétés. La surface d’une molécule représentée par des boules de Van der Waals est constituée par la frontière de l’ensemble de ces boules. La surface accessible par le solvant est l’ensemble des positions possibles d’une boule représentant le solvant, qui roulerait sur leur surface. Le centre de cette boule dessine une nouvelle surface, dénommée surface accessible au solvant.

L’objectif de ce projet est de réaliser un programme permettant de calculer la surface accessible au solvant (absolue et relative) à partir des coordonnées d’une protéine issue d’un fichier PDB (site répertoire de toutes les structures connues - Protein Data Bank).

Figure 1 : Représentation graphique de la surface disponible. La zone accessible est représentée par les lignes en pointillés, la molécule de solvant en bleu et les sphères centrées sur les atomes en rouge. (6)

* 1. **La protéine CD59**

La glycoprotéine CD59 est une protéine codée par le gène CD591. Elle contient 128 acides aminés. Elle appartient à la famille de protéines LY6/uPAR/alpha-neurotoxine2. Chez l’humain, elle est présence à la surface des hématies et des monocytes qui inhibent la constitution du complexe d'attaque membranaire du complément. Un déficit CD 59 provoque donc un défaut de protection de la membrane des hématies contre la lyse par le complément induisant une pathologie nommée hémoglobinurie paroxystique nocturne (HPN).

Les virus quant à eux, tels que le VIH, cytomégalovirus humain et la vaccine intègrent CD59 dans leur propre enveloppe virale pour éviter la lyse par le système du complément.

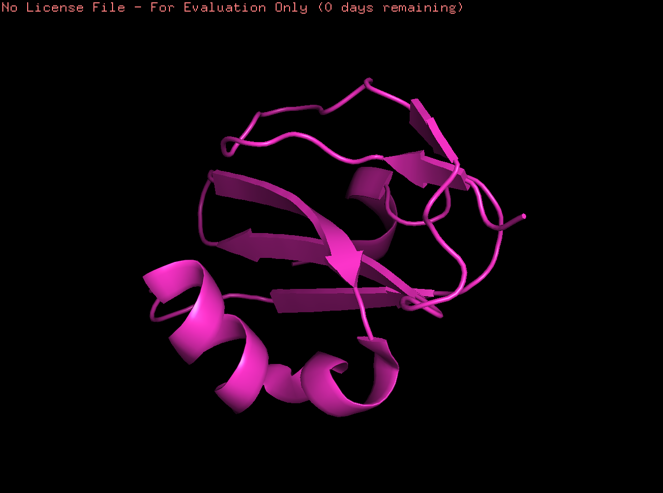
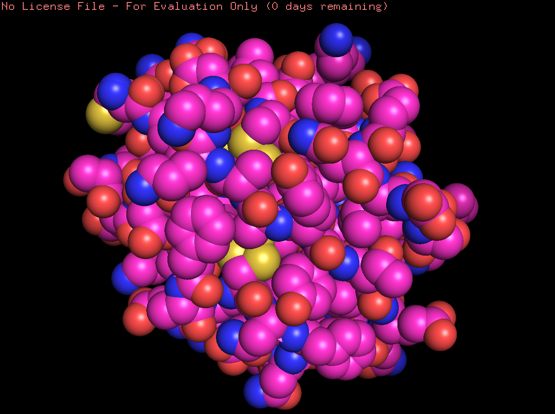


Figure 2 et 3 : Représentation de la protéine CD59 avec PyMOL

1. **Matériels et méthodes**
   1. **Description de l’algorithme**

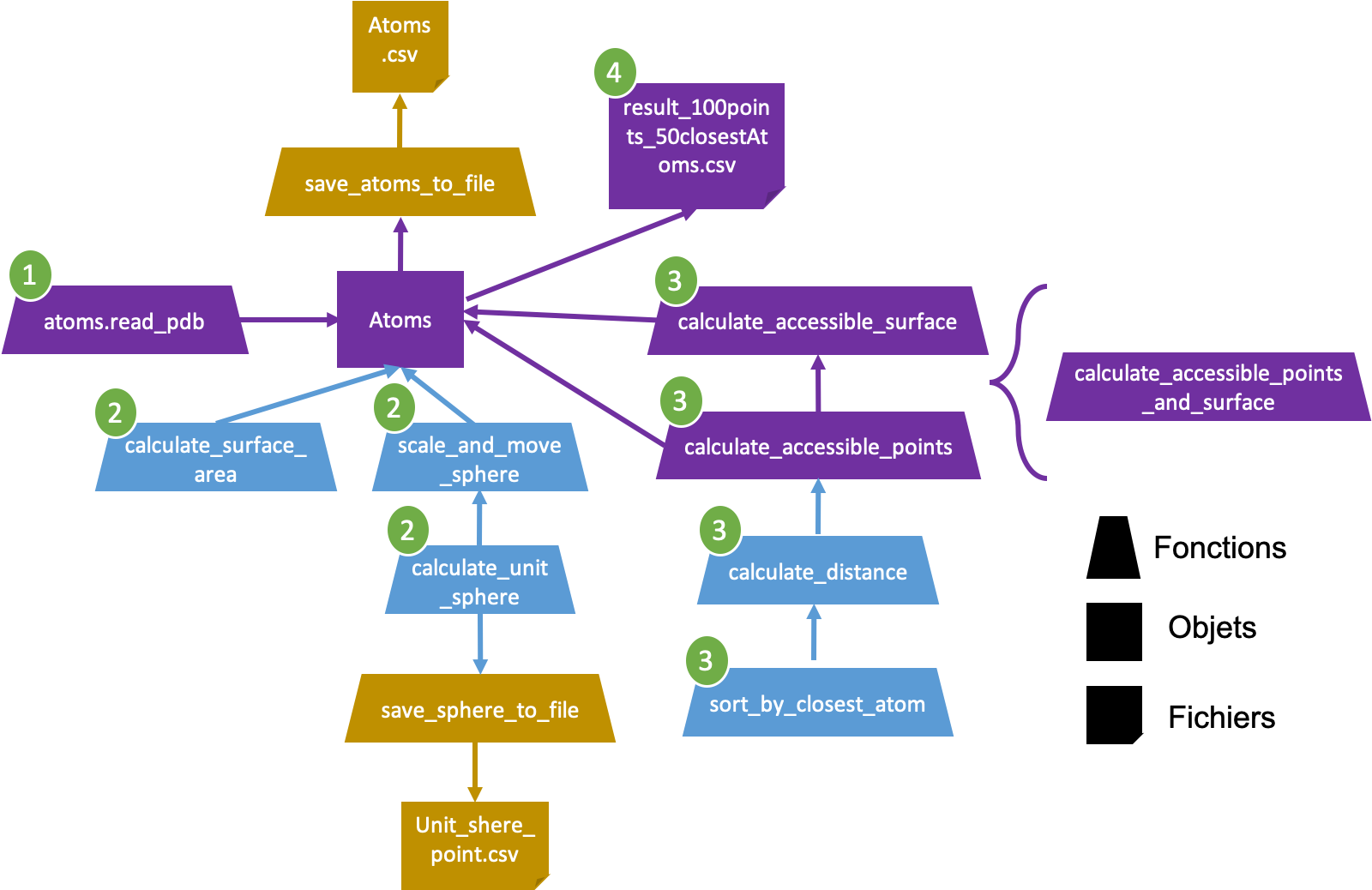


Figure 6 : Workflow.

En violet, le relatif à l’objet Atom, en bleu les fonctions générales et en moutarde le relatif aux visualisations 3D.

* Génération des objets Atoms issu de la lecture du fichier pdb avec les attributs suivants :

- index

Numéro de l’atome dans le pdb. S’incrémente à chaque ligne.

- name

Nom de l’atome

Résidu auquel appartient l’atome.

- location

Localisation dans la chaine.

- radius\_w

Rayon de Van der Walls de l’atome.

- surface\_area

Surface de la surface totale de la sphere autour de l’atome.

- atom\_sphere

Création d’une sphere de rayon 1 autour de l’atome.

- atom\_plus\_solvant\_sphere

Sphère de rayon atome + solvant.

- accessible\_points

Table contenant tous les points de la sphere.

- accessible\_points\_count

Nombre de points dans la sphère accessibles au solvant.

- accessible\_surface\_area

Surface accessible au solvant.

- accessible\_surface\_area\_pct

Pourcentage d’accessibilité.

* Lecture du fichier pdb
* Écriture dans le fichier atoms.csv. Il contient le symbole de l’atome ainsi que ces 3 coordonnées.
* Comptage du nombre d’atomes total.
* Création d’un nuage de 100 points en forme de sphère de rayon 1 autour d’un point central (algorithme de Saff et Kuijlaars).
* Écriture dans le fichier unit\_sphere\_point.csv. Il contient les coordonnées de chaque point de la sphère de rayon 1 autour de l’atome.

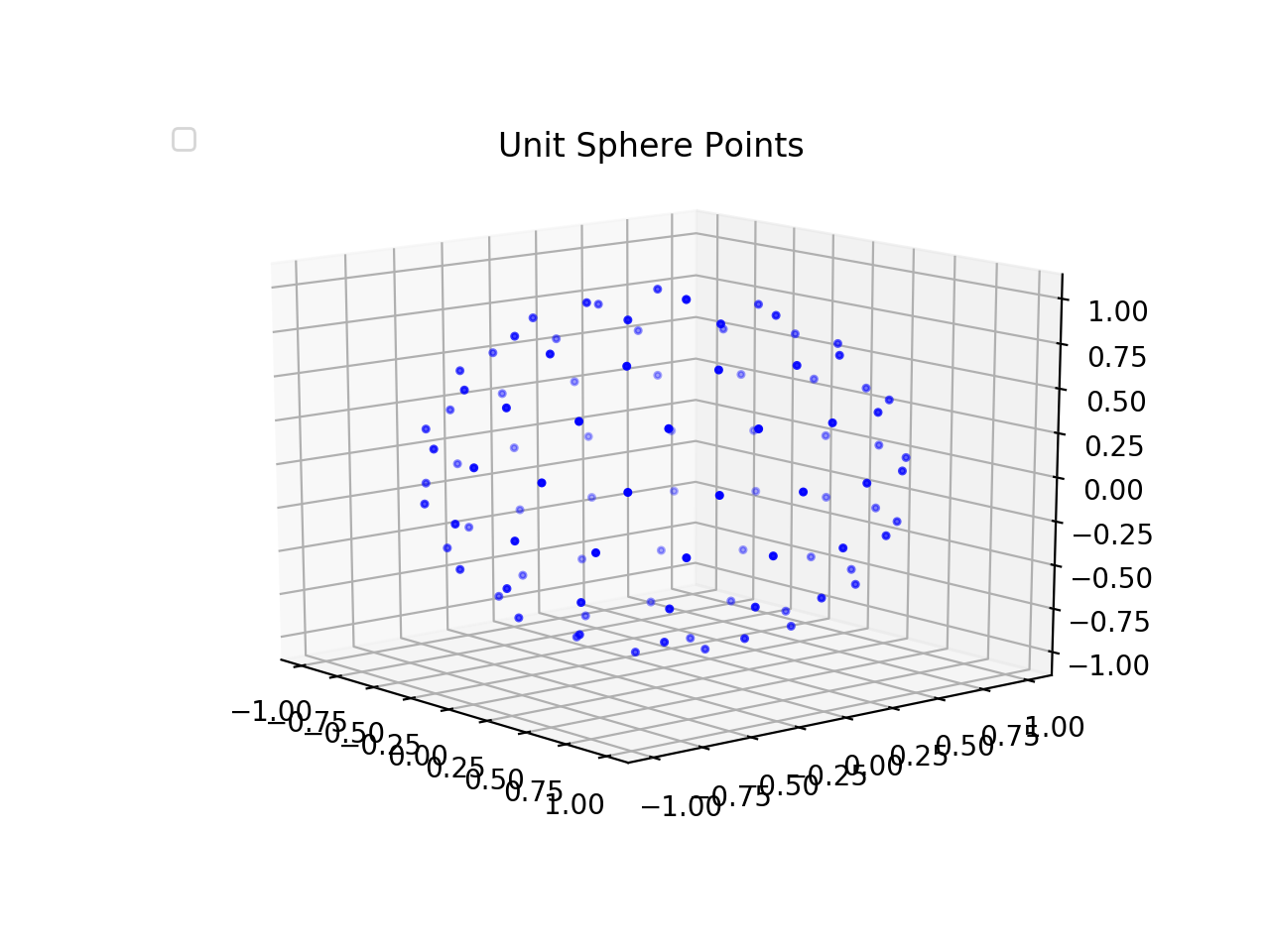


Figure 4 : Représentation 3D d’une sphère de 100 points et de rayon 1.

Image générée par le script unit\_sphere\_points\_3dplot.py à partir du fichier unit\_sphere\_point.csv.

* Déplacement de la sphère ainsi créée sur chaque atome. Le rayon de chacune d’elle prend la valeur du rayon de Van der Waals de chaque atome.

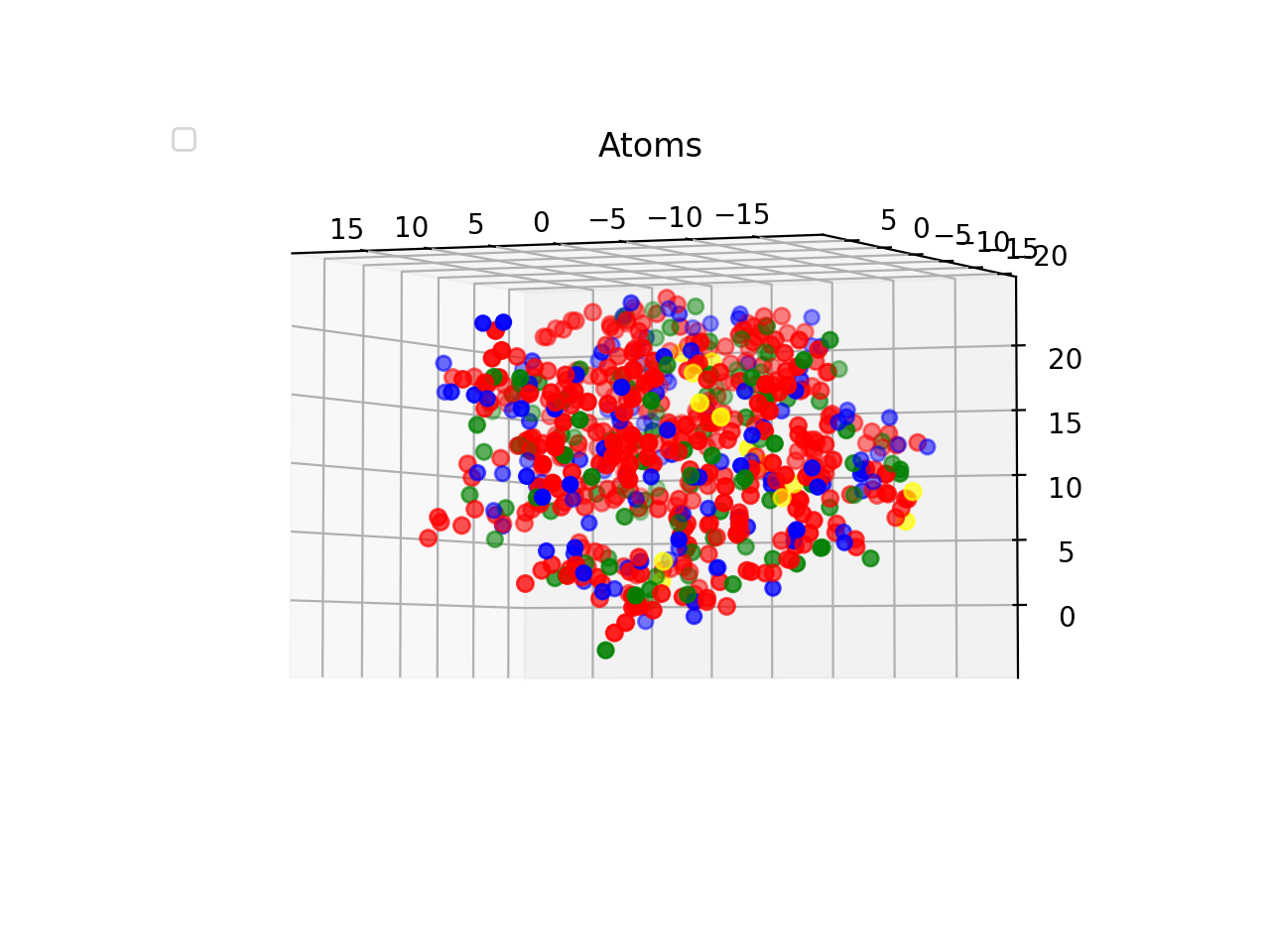


Figure 5 : Représentation des sphères centrées sur chaque atome.

En bleu les oxygènes, en rouge les carbones, en vert les azotes et en jaune les souffres.

Image générée par le script atoms\_3dplot.py à partir du fichier atoms.csv.

* Calcul des distances entre chaque point de la sphère **atome+solvant** et les 50 plus proches autres points des sphères **atome+solvant** de la protéine. Si la distance entre eux est inférieure au rayon d’une molécule d’eau, on considère que le point est sans contact.
* Tri par distance croissante entre atomes.
* Pour chaque point de chaque sphère **atome+solvant**, l’attribut accessible\_points de chaque objet Atom est par défault défini sur True.
* Les distances entre chaque point de la sphère et celles des 50 sphères les plus proches sont calculées. Si cette distance est supérieure au rayon d’une molécule d’eau, la valeur de accessible\_points est changée en False et le point est éliminé.
* A partir des points sans contact, calcul de la surface accessible totale.
* Calcul de la surface accessible relative : surface accessible / surface totale.
* Calcul du pourcentage d’accessibilité́ au solvant : (surface accessible \*100) / surface totale.
* Écriture du fichier de résultats : result\_100points\_50closestAtoms.csv synthétisant pour chaque atome le résidu auquel il appartient, ses coordonnées, X, Y, Z, sa surface totale, sa surface accessible et son accessibilité.

1. **Résultats**
   1. **Résultats du script**

Surface totale : 22395,3 Å ²

Surface totale accessible par le solvant : 1442,5 Å ²

Pourcentage d'accessibilité au solvant : 6,4%

Ces résultats ne sont pas aberrants mais on pourrait penser qu’un pourcentage de 6,4% d’accessibilité au solvant est plutôt faible.

* 1. **Comparaison avec Naccess**

Naccess est un programme autonome qui calcule la zone accessible d'une molécule à partir d'un fichier au format PDB (Protein Data Bank). Il peut calculer les accessibilités atomiques et résiduelles pour les protéines et les acides nucléiques. Il est disponible gratuitement à partir de son site internet pour les chercheurs d'établissements universitaires et à but non lucratif.

ACCALL - Accessibility calculations

MAX RESIDUES 8000

MAX ATOMS/RES 100

PDB FILE INPUT /Users/julieb/M2-BIB/Projet\_court/CD59\_2J8B.pdb

PROBE SIZE 1.40

Z-SLICE WIDTH 0.050

VDW RADII FILE vdw.radii

EXCL HETATOMS

EXCL HYDROGENS

EXCL WATERS

READVDW 32 residues input

ADDED VDW RADII

CHAINS 1

RESIDUES 78

ATOMS 629

SOLVA: PROGRAM ENDS CORRECTLY

CALCULATED ATOMIC ACCESSIBILITES

RELATIVE (STANDARD) ACCESSIBILITIES READFOR 20 AMINO ACIDS

SUMMED ACCESSIBILITIES OVER RESIDUES

Figure 7 : Fichier log de sortie de Naccess

Selon les calculs de Naccess, la surface absolue accessible au solvant est de : 4814,6 Å2 et un pourcentage d'accessibilité au solvant de 21,5%. Ces valeurs sont très envisageables et certainement plus cohérentes que celles de mon script quand on tient compte de la taille et de la composition en acides aminés de la protéine.

1. **Conclusion**

La surface accessible est utilisée dans le calcul de l'énergie associée à l'effet hydrophobe et pour améliorer la prévision de structures secondaires d’une protéine. La méthode de calcul décrite dans l’article de *Shake et al.* semble très efficace et très robuste. D’ailleurs dans l’exemple décrit dans la publication, on voit que les résultats de la simulation sont assez proches des valeurs empiriques. Il était donc très raisonnable de s’attendre à ce que les deux méthodes (*Shrake-Rupley* et Naccess) donnent des résultats similaires. Malheureusement, dans les faits, il semblerait que mon script sous-estime assez largement la surface accessible aux solvant. La méthode de discrimnation entre résidus exposés ou enfouis doit être trop stringente.

1. **Bibliographie**

1.

Saff EB, Kuijlaars ABJ. Distributing many points on a sphere. *The Mathematical Intelligencer*. 1997;19(1):5-11. doi:[10.1007/BF03024331](https://doi.org/10.1007/BF03024331)

2.

Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of Molecular Biology*. 1973;79(2):351-371. doi:[10.1016/0022-2836(73)90011-9](https://doi.org/10.1016/0022-2836(73)90011-9)

3.

Leath KJ, Johnson S, Roversi P, et al. High-resolution structures of bacterially expressed soluble human CD59. *Acta Crystallogr F Struct Biol Cryst Commun*. 2007;63(8):648-652. doi:[10.1107/S1744309107033477](https://doi.org/10.1107/S1744309107033477)

4.

Lee B, Richards FM. The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*. 1971;55(3):379-IN4. doi:[10.1016/0022-2836(71)90324-X](https://doi.org/10.1016/0022-2836(71)90324-X)

5.

Roth CM, Neal BL, Lenhoff AM. Van der Waals interactions involving proteins. *Biophysical Journal*. 1996;70(2):977-987. doi:[10.1016/S0006-3495(96)79641-8](https://doi.org/10.1016/S0006-3495(96)79641-8)

6.

http://boowiki.info/art/chimie-computationnelle/aire-de-surface-accessible.html

7.

https://www.uniprot.org/uniprot/P13987