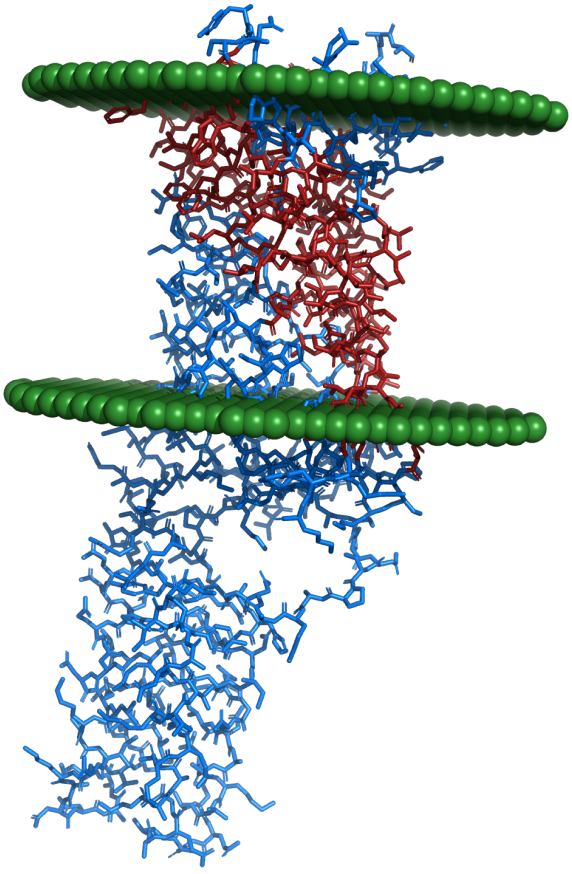
Université :
Université de Paris,
UFR Sciences du Vivant
Master de Bioinformatique
UE Programmation et gestion de projet

Assignation et détection des parties transmembranaires d'une protéine



Source : personnelle.
Légende : détection de la membrane de la protéine 4iar au cours du projet.

Enseignant :
Jean-Christophe Gelly

Étudiants.es :
Thibaul DUGAUQUIER
Lara HERRMANN

Année :
2020 - 2021

Tables des matières

1. Introduction

Les protéines sont scindées en trois principales classes : fibreuses, globulaires et membranaires. Les protéines membranaires et globulaires ont longtemps été difficilement identifiables. C'est dans ce contexte que G.E. Tusnady et al. ont développé, en 2004, l'algorithme TMDET dans l'objectif d'identifier et de classifier les protéines transmembranaires de la Protein Data Bank (PDB). L'objectif du projet est d'implémenter un outil capable de déterminer les zones transmembranaires d'une protéine à la façon de l'algorithme TMDET. Pour identifier la position la plus pertinente de la membrane, le nombre de résidus exposés au solvant et le nombre de résidus hydrophobes sont maximisés dans un espace entre deux plans parallèles.

2. Matériels et méthodes

2.1. Lecture du fichier et stockage des données PDB

La première étape du projet consiste en l'extraction des informations du fichier PDB. Pour simplifier l'exécution du programme, seuls les carbones alphas sont stockés puis traités dans la suite de l'analyse. Une valeur leur est associée, en fonction de leur hydrophobicité. Cette valeur a été fixée à 1 pour les résidus hydrophobes (F, G, I, L, M,V, W, Y) et à -1 pour les résidus hydrophiles (A, C, D, E, H, K, N, P, Q, R, S, T).

2.2. Calcul des zones de surface accessible

La détermination de la position de la membrane au sein de la protéine nécessite l'identification des résidus accessibles au solvant. C'est la fonction DSSP du module de Biopython qui a été utilisée pour déterminer la surface accessible relative ou Accessible Surface Area (ASA) de chaque acide aminé. Par défaut, seuls les résidus dont l'ASA est supérieure à 0.5 sont conservés dans la suite de l'analyse.

2.3. Création de vecteurs

Pour identifier le plan de l'espace le plus à même d'accueillir la membrane au sein de la structure tridimensionnelle de la protéine, la première étape consiste à déterminer les coordonnées du centre de masse. Il s'agit du point de l'espace par rapport auquel la masse de la protéine est uniformément répartie. Dans le cas de cette étude, seuls les carbones alphas sont pris en compte, donc les coordonnées du centre de masse correspondent à la moyenne des coordonnées de tous les résidus de la protéine. Le but est de parcourir un maximum de directions à partir de ce point, et de successivement déterminer l'hydrophobicité des plans tridimensionnels perpendiculaires à ces directions. Pour calculer les directions, une sphère de rayon 1 est utilisée, sur laquelle sont réparties de manière uniforme un nombre de point fini. Cette répartition homogène est obtenue grâce à l'algorithme de Fibonacci (Fig. 1). Le nombre de points à la surface de la sphère est déterminé par l'utilisateur. La valeur par défaut est de 20.

Les vecteurs directeurs permettant de déterminer les directions mentionnées précédemment sont calculés à partir de deux points, l'un étant le centre de masse, et l'autre

étant un des points de la surface de la sphère. Les plans permettant de déterminer la position de la membrane sont calculés à partir des directions pointées par chacun de ces vecteurs. Aussi, plus le nombre de points à la surface de la sphère est important, plus on explore de possibilités au sein de la protéine.

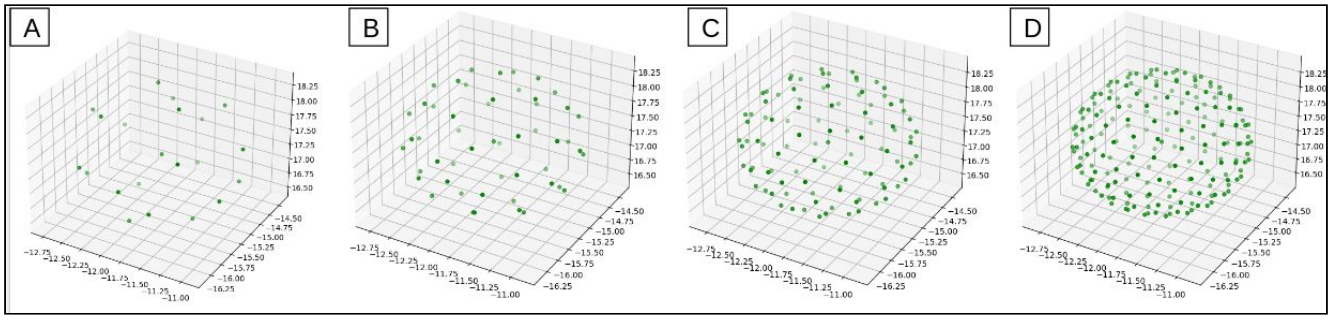


Figure 1 - Répartition de points à égale distance les uns des autres à l'aide de l'algorithme de Fibonacci et affichage dans un espace à 3 dimensions. (A) Répartition sur la sphère de 20 points. (B) Répartition sur la sphère de 50 points. (C) Répartition sur la sphère de 100 points. (D) Répartition sur la sphère de 200 points. Source : personnelle.

2.4. Détermination des membranes

La membrane est définie par deux plans parallèles entre eux et normaux à un vecteur (Fig 2). Ils sont séparés par une distance de 15Å. Les deux plans vont être décalés le long de l'axe du vecteur normal, par pas de 1Å. Le parcours s'interrompt lorsque la distance maximale des résidus au centre de masse est dépassée. Puis ils vont parcourir les différents vecteurs créés.

À partir des résidus accessibles aux solvants et présents dans l'intervalle des deux plans, l'hydrophobicité de la membrane est calculée. Son emplacement idéal est déterminé par le score le plus élevé.

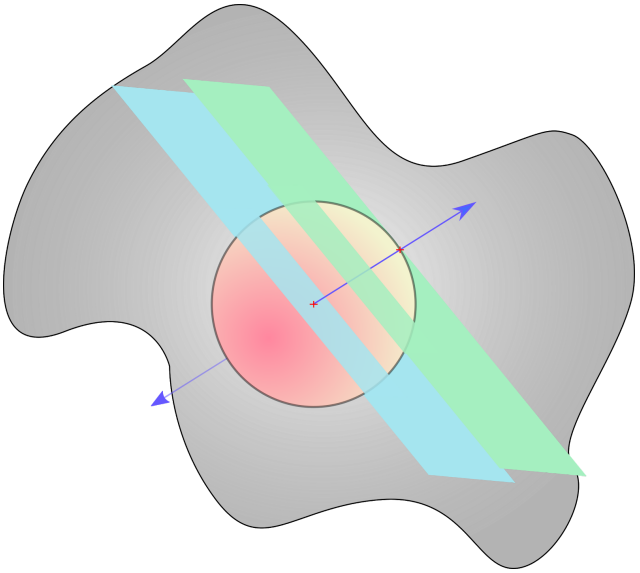


Figure 2 - Schéma de la détermination du plan membranaire, à partir d'un vecteur normal. Le vecteur normal (représenté par une flèche bleue) est défini à l'aide du centre de masse de la protéine (croix rouge) et d'un point (croix rouge) à la surface d'une sphère de rayon 1 (sphère orange). La membrane est représentée par deux plans (en cyan et vert) séparés par une distance de 15Å. Les résidus de la protéine (en noire) présents entre ces deux plans constituent la membrane. Source : personnelle.

2.5. Affichage des résultats

Deux types de résultats sont fournis par le programme. Le premier correspond à un fichier au format txt répertoriant les tranches consécutives de résidus qui sont transmembranaires. Le second résultat utilise le module PyMOL pour générer une image au format png de la protéine et de ses résidus en sticks. La région transmembranaire de la protéine est colorée en rouge, et le reste de la protéine en cyan.

2.6. Jeu de données

Le programme a été testé sur les 4 protéines suivantes : 5b2n, 4pxk, 4iar, et 2lly. Respectivement constituées de 279, 258, 401 et 137 résidus. Ces protéines sont considérées comme des protéines membranaires de référence.

3. Résultats

3.1. Analyse des résultats via PyMOL

La première analyse des résultats consiste en la visualisation des régions transmembranaires à l'aide de PyMOL. Les résidus sont contenus dans un intervalle de 15Å, entre deux plans parallèles.

3.2. Comparaison avec les fichiers pdb d'OPM

Pour vérifier si l'orientation et la position des membranes sont correctes, les résultats obtenus ont été comparés avec les fichiers pdb issus de la base de données Orientations of Proteins in Membranes (OPM). Pour chaque protéine analysée, la membrane détectée est correctement positionnée. Seule l'orientation du plan n'est pas optimisée. En effet, dans de nombreux cas, le plan membranaire défini n'est pas parallèle aux plans issus des fichiers de l'OPM (Fig. 3).

3.3. Discussion

Pour augmenter la précision de la position de la membrane, deux possibilités sont à explorer. La première consiste en la détermination de l'hydrophobicité à partir de l'échelle d'hydrophobicité des résidus et non à partir d'une valeur binaire. Néanmoins, selon G.E. Tusnady et al. 2004, les résultats obtenus sont similaires. Le seconde vise à ajuster la taille de la membrane de façon à ce qu'elle soit optimale et non fixée à 15Å. Finalement, pour pouvoir déterminer une orientation plus précise du plan membranaire dans l'espace, plusieurs points d'origine des vecteurs peuvent être explorés.

Lors de cette étude, un seuil d'accessibilité de 0.5 a été utilisé. Il a été montré qu'un seuil supérieur ou inférieur menait à des résultats moins probants car ils étaient respectivement trop restrictifs et trop permissifs. C'est pour cette raison que le seuil de 0.5 a été conservé pour la présentation des résultats. Le nombre de points choisi est aussi un facteur déterminant. En effet, en dessous de 20 points, le nombre de directions parcourues par le programme n'est pas assez intéressant, en particulier dans les protéines de moyenne et grande

taille. Plus on explore de direction au sein de la protéine, plus le résultat sera précis. Le temps de résolution de l'algorithme sera quant à lui plus conséquent.

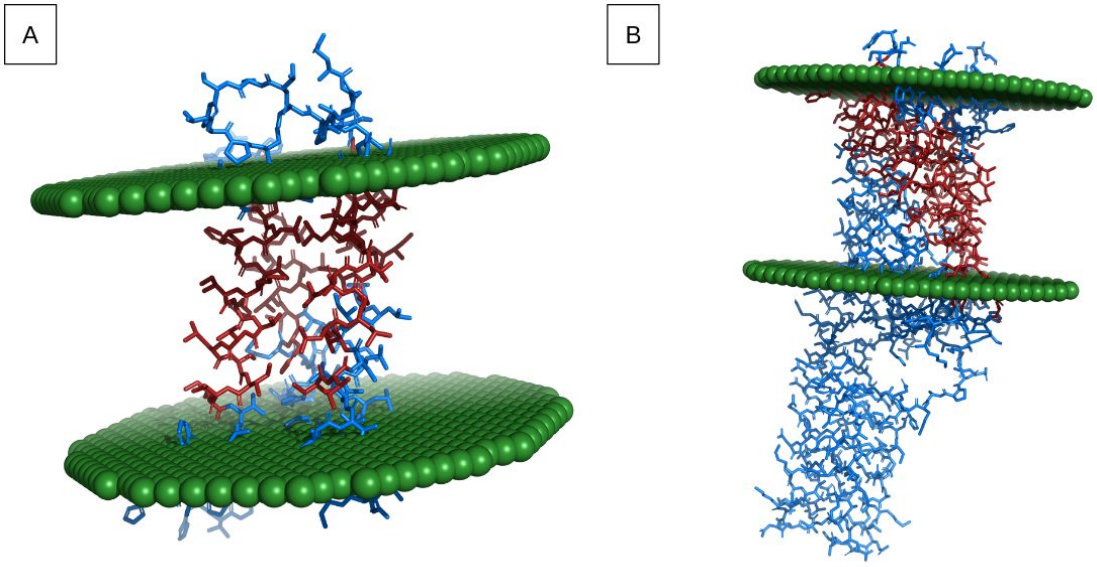


Figure 3 **-** Représentation des protéines 2lly (A), et 4iar (B). à l'aide du logiciel PyMOL. La membrane issue de la base de données OPM est représentée en sphères vertes, la région transmembranaire de la protéine déterminée lors de cette étude en sticks rouges et la région non membranaire en sticks bleus. Source : personnelle.

3.4. Temps d'exécution

Le temps d'exécution de l'algorithme est relativement faible (de quelques secondes pour les petites protéines à quelques minutes pour les plus imposantes), ce qui est comparable aux résultats présentés dans l'article de G.E. Tusnady et al, 2004. Quant à la génération de l'image png de la protéine par PyMOL, son temps d'exécution est inférieur à 10 secondes. Dans les deux cas, ce temps de traitement est largement influencé par la taille de la protéine traitée.

4. Conclusion

L'algorithme créé permet la détermination des régions transmembranaires d'une protéine avec une résolution relativement bonne. La comparaison entre les résultats obtenus et ceux recensés dans les bases de données de protéines transmembranaires, telles qu'OPM, met en avant la fiabilité de l'algorithme notamment pour l'identification de l'orientation du plan membranaire. Néanmoins, la largeur fixe de la membrane à 15Å ne permet pas la détermination exacte des coordonnées des acides aminés transmembranaires, et ce, plus particulièrement chez les petites protéines.

Annexes

Annexe 1 - Difficultés rencontrées

- Positionnement des points sur la sphère. Il nous a été nécessaire de trouver un algorithme existant et de l'adapter (algorithme de *Fibonacci*)
- Utilisation de la géométrie dans l'espace. Nécessité de se replonger dans les formules du lycée (détermination d'un vecteur à partir de 2 points, détermination de l'équation cartésienne d'une droite, calcul de la distance entre deux plans)
- Affichage des résultats et utilisation de PyMOL.
- Problèmes d'identifiants des dictionnaires et notamment faire correspondre les lignes du fichier avec un identifiant PDB (Utilisation de DSSP)

Annexe 2 - Structure du programmes réalisés

1. Gestion des paramètres en entrée
2. Extraction des informations du fichier : carbones alphas et coordonnées
3. Calcul du centre de masse de la protéine
4. Calcul de la surface accessible au solvant pour chaque résidus (fonction DSSP de BioPython)
5. Répartition des points sur une sphère de rayon 1 (algorithme de Fibonacci)
6. Détermination du plan membranaire le plus pertinent
   - 6.1. Parcours des vecteurs et détermination de leurs coordonnées
   - 6.2. Parcours des plans le long du vecteur jusqu'à la distance maximale
   - 6.3. Calcul de l'hydrophobicité de chaque plan
   - 6.4. Détermination du plan avec le meilleur score
   - 6.5. Stockage de l'hydrophobicité max, de la liste des carbones composant le plan
7. Génération d'une image PyMOL
8. Segmentation des résultats
9. Écriture d'un fichier résultats

Annexe 3 - Exemple d'utilisation du programme

Affichage de l'aide :

```
python3 projet_SSDC.py -h
```

Utilisation des paramètres par défaut :

```
python3 projet_SSDC.py -f 2lly.pdb
```

Ligne de commande complète :

```
python3 projet_SSDC.py -f 2lly.pdb -o resultats.txt -p 200 -s 0.5 -m 15
```