Arnaud LY
M2BI

# Project M2BI

Sujet : Conception d'un programme d'évaluation de la qualité d'un modèle 3D de protéine

## Introduction

On peut aujourd'hui obtenir des informations sur une structure protéique à partir des banques de donnée et faire des analyses sur les interactions entre les acides aminés, de la modélisation de structure etc... En ce qui concerne cette modélisation protéique, plusieurs méthodes ont été mises au point, en se basant sur différentes théories parmi lesquelles, la théorie de *Discrete Optimized Protein Energy (DOPE)* est celle qui nous intéresse. DOPE se base sur le potentiel statistique entre 2 acides aminés et est dépendant de la distance entre ces molécules, de l'interaction entre les carbones alpha ainsi que des chaines latérales.

Pour cette étude, on ne prend en compte que les interactions entre les carbones alpha de chaque acide aminé. Ainsi, en utilisant le potentiel statistique DOPE et la méthode d'analyse décrite dans (Shen et al, 2006)[1], on réalise un programme permettant d'évaluer un modèle protéique 3D.

## Matériels et méthodes

### Support logistique

Pour réaliser ce programme, le langage Python 3 sera utilisé et l'écriture du programme sera fait dans l'éditeur Visual Studio Code. On utilisera Miniconda comme gestionnaire de packages et pour la création de l'environnement de travail. Les bibliothèques utilisées sont la bibliothèque standard, Pandas pour la gestion de DataFrame et NumPy pour la gestion des données en tableau. L'utilisation du programme sera sous environnement Linux.

### Recueil de donnée

Les données protéiques sont obtenues grâce à la banque de données sur les protéines RCSB, sous format .pdb. Les protéines utilisées pour l'étude sont 1PDC et 1GCN. Un document texte contenant les valeurs de DOPE suivant la méthode décrite sur l'article précédemment cité a été récupéré à l'adresse : http://www.dsimb.inserm.fr/~gelly/data/dope.par. Ce document texte contient toutes les valeurs de DOPE de tous les couples possibles d'acide aminé, associées à une distance entre 0 et 15 Å avec une incrémentation de 0,5 Å, donnant ainsi 30 valeurs de DOPE par interaction (*voir Annexe*)

[1] Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006 Nov;15(11):2507-24. doi: 10.1110/ps.062416606. PMID: 17075131; PMCID: PMC2242414.

Conception du programme

Lors de la conception de ce programme, une modélisation des étapes du programme a été réalisée (*Cf Figure 1*). En se basant sur cette modélisation, la première étape a été de créer une/des classe(s) afin de catégoriser nos variables et méthodes. Suivi de cela, le document texte contenant les données DOPE a été ouvert puis arrangé afin de contenir, dans les deux premières colonnes d'un DataFrame, les résidus de carbone alpha (CA) des acides aminés, suivis des valeurs DOPE avec pour index de colonne, la valeur en Angstrom correspondant. Ensuite, pour obtenir des valeurs dites "expérimentales", venant de la séquence d'origine, on récupère les coordonnées x, y et z de chaque CA et on calcule la distance euclidienne entre tous les CA, entre i et i + n avec n>2. Cette distance étant associée à un couple de CA, cette correspondance permet d'obtenir une valeur DOPE. La somme de ces valeurs donne un score qui sera comparé aux scores dits de "référence".

Ces scores de "référence" sont obtenus à partir de la même séquence source mais avec un ordre de CA obtenu aléatoirement. Les distances et valeurs DOPE leurs sont ensuite associés puis un z-score est calculé pour ces valeurs DOPE. La moyenne et écart-type obtenus permettent ensuite d'évaluer le score "expérimental", avec la formule ((moyenne – valeur) / écart-type). La valeur obtenue après utilisation de cette formule indique la position relative du score "expérimentale" dans une distribution gaussienne, donc d'une distribution de hasard. Ainsi, plus la valeur est élevée, plus le modèle 3D obtenu par les coordonnées étudiées est viable.

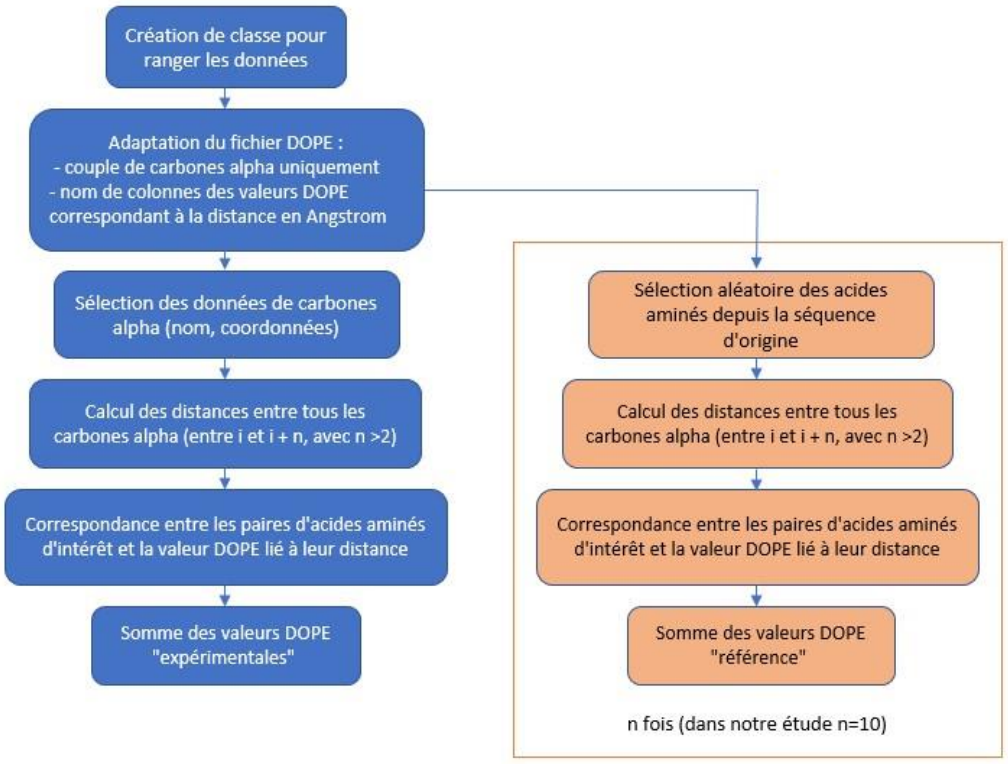


Figure 1 : Schéma représentatif de la conception du programme de cette étude

## Résultats

À l'issue de l'utilisation de ce programme, on obtient un score "expérimental" correspondant au score de la séquence d'origine, une liste de 10 valeurs de sommes de valeurs DOPE provenant de séquences aléatoires, la moyenne et l'écart-type de cette liste ainsi qu'un score de comparaison du score "expérimental" avec la distribution de référence. Ces valeurs sont transcrites dans *compil_results* se trouvant dans le dossier *results.*

Pour le fichier exemple "1gcn.pdb", on obtient un score DOPE "expérimental" de –11,4. La liste de valeurs de "référence" est différente à chaque utilisation du programme. Cependant, la moyenne et l'écart-type restent assez stable, avec respectivement pour valeur –60 et 4. Cela donne un score final d'environ -17. Quant au fichier '1pdc.pdb', on obtient une somme de –44,8, une moyenne d'environ -105, un écart-type d'environ 2 et un score final de –24.

## Discussion et Conclusion

Les valeurs obtenues par ce programme permettent d'évaluer la qualité de la modélisation d'une protéine à partir de ses coordonnées 3D. Les exemples utilisés dans cette étude possèdent un nombre faible d'acide aminés : 1GCN et 1PDC ayant respectivement 29 et 45 acides aminés. De plus, leur structure tertiaire est relativement simple. Avec pour comparaison 2 exemples de petites protéines, il est difficile de confirmer la qualité du score final, malgré le fait que ce dernier soit relativement élevé. Dans l'étude de (Kumari *et al*, 2020)[2], une mesure de score DOPE a été réalisée pour évaluer leur modèle de protéine. Ces derniers ont des scores plus élevés (environ -14 500), probablement dus à des structures plus complexes et un nombre de résidus plus élevé (138 pour Seq.B99990002).

Cependant, il est à préciser que dans notre étude, le calcul de ce score s'est basé uniquement sur les interactions entre CA et ne prend pas en compte les autres données comme les chaines latérales, manquant ainsi des valeurs DOPE d'interaction. De plus, l'association de la valeur DOPE avec la distance mesurée entre les CA n'est pas précise. En effet, dans le programme de notre étude, la distance est arrondie au supérieur pour attribuer la valeur DOPE. Pour augmenter la précision de cette valeur, une interpolation linéaire serait nécessaire.

Pour conclure, bien que des scores DOPE ont été générés pour les fichiers PDB, ces valeurs ne sont pas représentatives de la précision de la technique de qualification de modèle. L'utilisation de protéine plus longue, de la totalité des données fournies avec le fichier PDB ainsi que l'utilisation d'interpolation linéaire ne sont qu'un début dans l'amélioration de ce processus de calcul.

---

[2] Kumari R, Chaudhary A, Mani A. Casuarictin: A new herbal drug molecule for Alzheimer's disease as inhibitor of presenilin stabilization factor like protein. Heliyon. 2020 Nov 21;6(11):e05546. doi: 10.1016/j.heliyon.2020.e05546. PMID: 33294689; PMCID: PMC7689514.

# Annexe

| | res1 | res2 | 0.5 | 1.0 | 1.5 | 2.0 | ... | 12.5 | 13.0 | 13.5 | 14.0 | 14.5 | 15.0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | ALA | ALA | 10.0 | 10.0 | 10.0 | 10.0 | ... | -0.08 | 0.01 | -0.02 | -0.08 | -0.12 | -0.02 |
| 1 | ALA | ARG | 10.0 | 10.0 | 10.0 | 10.0 | ... | -0.01 | 0.02 | -0.00 | -0.00 | -0.10 | -0.02 |
| 2 | ALA | ASN | 10.0 | 10.0 | 10.0 | 10.0 | ... | -0.01 | 0.04 | 0.03 | -0.02 | -0.05 | -0.02 |
| 3 | ALA | ASP | 10.0 | 10.0 | 10.0 | 10.0 | ... | -0.03 | -0.00 | -0.02 | -0.02 | -0.11 | -0.02 |
| 4 | ALA | CYS | 10.0 | 10.0 | 10.0 | 10.0 | ... | -0.06 | -0.03 | -0.07 | -0.05 | -0.11 | -0.02 |
| 5 | ALA | GLN | 10.0 | 10.0 | 10.0 | 10.0 | ... | -0.04 | 0.00 | 0.01 | -0.00 | -0.11 | -0.02 |
| 6 | ALA | GLU | 10.0 | 10.0 | 10.0 | 10.0 | ... | -0.04 | -0.01 | -0.03 | -0.06 | -0.10 | -0.02 |
| 7 | ALA | GLY | 10.0 | 10.0 | 10.0 | 10.0 | ... | 0.02 | -0.02 | -0.01 | -0.03 | -0.07 | -0.02 |
| 8 | ALA | HIS | 10.0 | 10.0 | 10.0 | 10.0 | ... | 0.01 | 0.03 | 0.03 | -0.01 | -0.11 | -0.02 |
| 9 | ALA | ILE | 10.0 | 10.0 | 10.0 | 10.0 | ... | -0.09 | -0.06 | -0.04 | -0.02 | -0.08 | -0.02 |

Annexe 1 : Partie du DataFrame des valeurs DOPE

### Difficultés rencontrées

- Compréhension des différentes étapes à la réalisation du programme
- Il s'agit de mon 1er projet et script en bioinformatique (réel apprentissage du codage python depuis aout)
- Utilisation des OOP
- Mise en place des docstrings, documentation, fonction help
- Rendre le code plus concis, plus clair, moins brouillon
- Difficulté à programmer pour une utilisation sous Linux