Protein Interaction Calculator

Projet court M2-BI : Debbah Nagi, Vander Meersche Yann
Git-Projet-Court

I. Introduction

Les interactions protéiques intra- et inter-chaînes sont indispensables à la stabilisation des protéines, et par conséquent à leur fonction. Ces informations pouvant être déterminées à partir des coordonnées atomiques d'un fichier PDB, un outil calculant ces différentes interactions serait particulièrement intéressant. Le serveur web *PIC*[1] (Protein Interactions Calculator Implementation - PIC) a donc été mis en place afin de répondre à ce besoin. Celui-ci permet à l'utilisateur de facilement détecter les liaisons disulfures, les interactions hydrophobes, les interactions ioniques, les liaisons hydrogènes, les interactions aromatiques-aromatiques, les interactions aromatiques-soufre et les interactions cation-π. Dans ce projet, nous avons ré-implémenté ces calculs afin d'obtenir une sortie similaire à partir d'une structure 3D correcte et utilisable localement.

II. Matériel et Méthodes

Caractéristiques du script

Notre version de *PIC* est codée en Python 3, et utilise différents modules : *Pandas* pour le stockage des données, *NumPy/SciPy* pour les calculs, *tabulate* afin de formater facilement le style des tableaux de sorties et *Selenium/geckodriver* (nécessite une connexion internet), nous permettant d'automatiser la récupération du fichier de sortie de *HBOND*[2], l'outil externe utilisé par Tina et. al.[1] dans *PIC*. Tous ces modules sont présents dans l'environnement *Conda* présent sur *GitHub* (Git-Projet-Court).

Notre script requiert un fichier de coordonnées au format PDB. Il accepte des fichiers de protéines simple (*simple-chaîne*) ou complexe (*multi-chaînes*), et peut ouvrir des fichiers PDB issus de RMN, en ne considérant que le premier modèle. Notre script présente deux paramètres optionnels, permettant de rechercher les interactions entre les différentes chaînes (*-interchain*), interne à une chaîne (*-intrachain*), ou l'ensemble des interactions possibles lorsque ces options sont omises.

Le projet a été majoritairement développé en suivant la méthode *Kanban* (voir Fig. S1 et Tab. S2). *PIC* permettant de calculer 7 types d'interactions différentes, nous avons pu coder ces différentes parties du code indépendamment avant de les fusionner.

Définition des différentes interactions

Les interactions étudiées peuvent être définies de différentes façons. Ici nous avons choisi de nous baser sur les critères de PIC[1], eux même basés sur les critères standards de la littérature.

Tout comme PIC, nous avons utilisé HBOND[2] (*fichier .hbd*) pour calculer les **liaisons hydrogènes**. Trois types sont calculées : chaîne principale-chaîne principale, chaîne

principale-chaîne secondaire et chaîne secondaire-chaîne secondaire. Parmis ces trois types, il y aura une liaison hydrogène lorsque la distance donneur-accepteur est inférieure à 3.5 Å.
Une **liaison disulfure**[6] est présente lorsque deux soufres de cystéines sont à moins de 2.2 Å.
L'**interaction hydrophobe** est définie lorsque qu'au moins deux carbones de chaînes latérales des acides aminés hydrophobes (ALA, VAL, LEU, ILE, MET, PHE, TRP, PRO, TYR) sont à une distance inférieure à 5 Å.
Une **interaction ionique** entre deux résidus est quant à elle présente lorsque l'azote de la chaîne latérale d'un résidu cationique (ARG, LYS, HIS) se trouve proche d'un oxygène d'un résidu doté d'anion (ASP, GLU) dans un seuil inférieur à 6 Å.
Pour calculer les **interactions aromatiques-aromatiques**[3] (PHE, TYR, TRP), nous prendrons en compte la distance entre les centres de gravité des cycles aromatiques (moyenne des coordonnées des six carbones). Cette distance devra alors être comprise entre 4.5 et 7 Å.
Suivant ce principe, nous pouvons identifier les interactions **aromatique-soufre**[4] lorsque la distance entre le centroïde du cycle aromatique et l'atome de soufre d'un résidu soufré (CYS, MET) est à inférieure à 5.3 Å, et les **interactions cation-π**[5] lorsque la distance entre le centre de gravité du cycle aromatique et l'atome d'azote de la lysine, ou de la moyenne des coordonnées des atomes d'azote de l'arginine est inférieurs à 6 Å. Ces définitions d'interactions peuvent donc être légèrement différente de celle de PIC, car manquant de documentation précise, nous avons dû choisir.

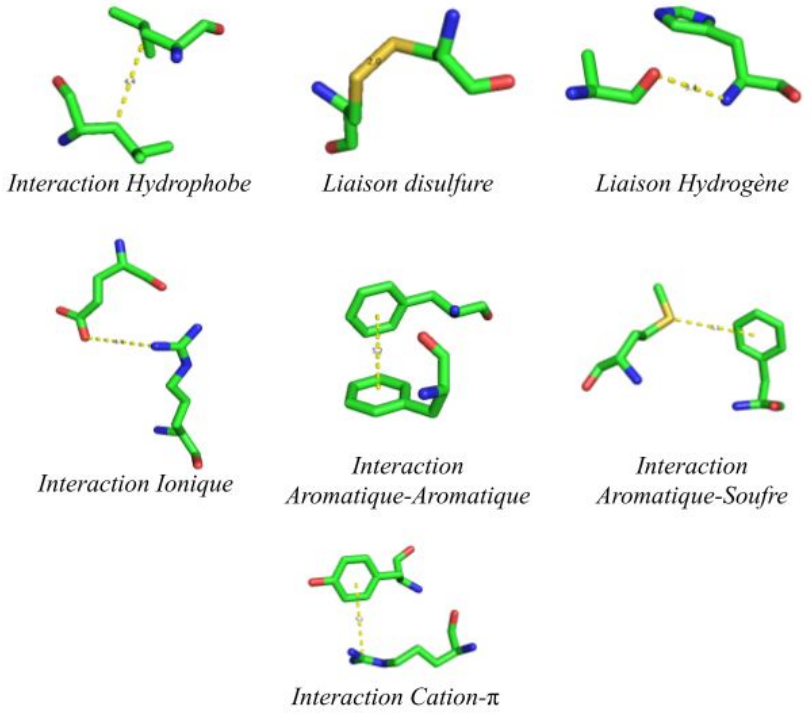

Fig 1 : Représentation Pymol des différentes interactions calculées (2dsq.pdb). *Interaction hydrophobe entre deux leucines; Liaison disulfure entre deux cystéines; Liaison hydrogène entre une alanine et une histidine (chaîne principale avec chaîne principale); Interaction ionique entre une arginine et un acide glutamique; Interaction aromatique-aromatique entre deux phénylalanines; Interaction aromatique-soufre entre une phénylalanine et une méthionine; Interaction cation-π entre une tyrosine et une arginine.*

III. Résultats et discussion

Résultats et discussion

Pour tester les performances de notre script, nous avons calculé les interactions de dix protéines, avec *PIC* et avec notre script puis nous avons compilé ces résultats dans la table 1.

		Hydrophobe		Ionique		Arom-Arom		Arom-sulfure		Cation-π		Pont disulphure	
ID PDB		Script	PIC	Script	PIC	Script	PIC	Script	PIC	Script	PIC	Script	PIC
1A3N	Compte	463	465	43	41	20	20	6	6	3	3	0	0
	% erreur	0,4		4,9		0		0		0		0	
1BTA	Compte	88	88	7	7	3	3	2	2	0	0	0	0
	% erreur	0		0		0		0		0		0	
1EJG	Compte	17	17	1	1	0	0	4	5	1	4	3	3
	% erreur	0		0		0		20		75		0	
1GZ2	Compte	95	95	11	11	11	11	4	4	3	3	3	3
	% erreur	0		0		0		0		0		0	
1IGY	Compte	794	794	105	103	46	46	29	29	22	24	12	14
	% erreur	0		1,9		0		0		8,3		14,3	
1KEW	Compte	608	609	99	93	41	41	0	0	30	30	0	0
	% erreur	0,2		6,5		0		0		0		0	
2DSQ	Compte	175	175	18	18	6	6	5	5	4	5	23	23
	% erreur	0		0		0		0		20		0	
3GP2	Compte	90	91	12	12	8	8	0	0	1	1	0	0
	% erreur	1,1		0		0		0		0		0	
3I40	Compte	22	22	0	0	0	0	0	0	0	0	1	2
	% erreur	0		100		100		0		0		50	
3Q06	Compte	584	584	126	124	13	13	0	0	18	19	0	0
	% erreur	0		1,6		0		0		5,3		0	
MOYENNE	Compte	2936	2940	422	410	148	148	50	51	82	89	42	45
	% erreur	0,1		2,9		0		2		7,9		6,7	

Table 1: Table comparative des résultats de notre programme et de PIC (intra-chaîne) sur 10 protéines (voir Tab. S1 pour les résultats inter-chaine). *Pour chaque protéine, nous calculons le nombre d'interaction identifiées par les deux outils, ainsi que le pourcentage d'erreur (\|nbScript - nbPIC\| / nbPIC). Nous n'affichons pas les liaisons hydrogènes puisque nous parsons exactement le même fichier HBOND, et que nous avons donc aucune erreur.*

Nous pouvons voir que nos résultats sont très proches de ceux de PIC, mais bien que notre code soit basé sur les critères définis par PIC, nous observons de légères différences. Ces différences peuvent être causée par les définitions légèrement différentes de certaines liaisons que nous avons implémentés par manque de documentation. En effet, pour les liaisons cation-π, nous avons choisi d'utiliser la moyenne des azotes de l'arginine par soucis de cohérence, puisque nous utilisons le centroïde du cycle aromatique pour les calculs, alors que PIC calcule les distance depuis le carbone lié aux deux azotes. Ces différences peuvent aussi être causées par un manque de précision dans le calcul des distances de PIC.

La taille du fichier proposé ainsi que la configuration de l'ordinateur employé font varier le temps d'exécution de quelques secondes à quelques minutes.

IV. Conclusion

Ce script remplit parfaitement son rôle, mais reste tout de même perfectible. D'une part, en plus d'être peu documentée, la définition des différentes interactions manque cruellement de précision. En effet, une interaction n'est pas définie uniquement par une distance. De nombreux paramètres entrent en compte, comme l'angle et l'énergie de liaison par exemple.

D'autre part, le code de notre programme est loin d'être optimal. Il aurait été intéressant de réaliser ce projet en programmation orienté objet, ce qui aurait été plus efficace et plus rapide, mais beaucoup plus compliqué à coder, ce que nous avons préféré éviter vu le temps très limité dont nous disposions.

Enfin, nous aimerions, dans une prochaine version du script implémenter le calcul des angles dièdre et des angles classiques, ce que nous avons préféré éviter ici car il nous a été impossible de comprendre la méthodologie utilisés par PIC pour ces calculs d'angles et obtenions des résultats incohérent avec ceux de PIC.

Bibliographie

1. Tina, K. G., Bhadra, R. & Srinivasan, N. PIC: Protein Interactions Calculator. *Nucleic Acids Res.* **35**, W473–W476 (2007).
2. JP, O., Johnson, M., Sali, A. & Blundell, T. Tertiary Structural Constraints on Protein Evolutionary Diversity: Templates, Key Residues and Structure Prediction. *Proc. Biol. Sci.* **241**, 132–45 (1990).
3. Burley, S. & Petsko, G. Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science* **229**, 23–28 (1985).
4. Reid, C., Lindley, P. F. & Thornton, J. M. Sulphur-aromatic interactions in proteins. *FEBS Lett.* **190**, 5 (1985).
5. Sathyapriya, R. Interaction of DNA with clusters of amino acids in proteins. *Nucleic Acids Res.* **32**, 4109–4118 (2004).
6. Sowdhamini, R. *et al.* Stereochemical modeling of disulfide bridges. Criteria for introduction into proteins by site-directed mutagenesis. *Protein Eng. Des. Sel.* **3**, 95–103 (1989).

Annexe 1

ID PDB		Hydrophobe		Ionique		Arom-Arom		Arom-sulfure		Cation-π		Pont disulphure	
		Script	PIC	Script	PIC	Script	PIC	Script	PIC	Script	PIC	Script	PIC
1A3N	Compte	32	32	4	4	2	2	0	0	8	8	0	0
	% erreur	0		0		0		0		0		0	
1BTA	Compte	88	88	7	7	3	3	2	2	0	0	0	0
	% erreur	0		0		0		0		0		0	
1EJG	Compte	0	0	0	0	0	0	0	0	0	0	0	0
	% erreur	0		0		0		0		0		0	
1GZ2	Compte	0	0	0	0	0	0	0	0	0	0	0	0
	% erreur	0		0		0		0		0		0	
1IGY	Compte	54	56	14	14	15	15	0	0	2	2	5	5
	% erreur	3,6		0		0		0		0		0	
1KEW	Compte	26	31	4	4	0	0	0	0	0	0	0	0
	% erreur	16,1		0		0		0		0		0	
2DSQ	Compte	61	61	14	14	4	4	1	1	2	2	0	0
	% erreur	0		0		0		0		0		0	
3GP2	Compte	33	33	5	5	0	0	0	0	0	0	0	0
	% erreur	0		0		0		0		0		0	
3I40	Compte	14	14	1	1	1	1	0	0	0	0	2	2
	% erreur	0		0		0		0		0		0	
3Q06	Compte	29	34	21	21	6	6	0	0	6	7	0	0
	% erreur	14,7		0		0		0		14,3		0	
MOYENNE	Compte	337	349	70	70	31	31	3	3	18	19	7	7
	% erreur	3,4		0		0		0		5,3		0	

Table S1: Table comparative des résultats de notre programme et de PIC (inter-chaîne) sur 10 protéines. *Pour chaque protéine, nous calculons le nombre d'interaction identifiées par les deux outils, ainsi que le pourcentage d'erreur (\|nbScript - nbPIC\| / nbPIC).*

Annexe 2

Nom de la tache	DATE DE DÉBUT	DATE DE FIN	DÉBUT AU JOUR	JOURS DE TRAVAIL	MEMBRES	POURCENTAGE COMPLÉTÉ
Mise en place du Projet						
Initialisation du GIT	9/8	9/8	0	1	Nagi/Yann	100 %
Recherche Bibliographique	9/9	11/9	1	3	Nagi/Yann	100 %
Design du programme	9/10	9/11	2	2	Nagi/Yann	100 %
Programmation	9/9	9/15	1	7	Nagi/Yann	100 %
Optimisation – Documentation	9/13	9/16	5	4	Nagi/Yann	100 %
Rapport						
Choix de structure	9/13	9/15	5	3	Nagi/Yann	100 %
Rédaction	9/14	9/16	6	3	Nagi/Yann	100 %
Choix des annexes	9/15	9/15	7	1	Nagi/Yann	100 %
Relecture	9/15	9/16	7	2	Nagi/Yann	100 %
Soutenance						
Préparation des slides	9/16	9/17	8	2	Nagi/Yann	100 %
Répétition	9/17	9/17	9	1	Nagi/Yann	100 %
Soutenance	9/18	9/18	10	1	Nagi/Yann	100 %

Nom de la tache	DATE DE DÉBUT	DATE DE FIN	DÉBUT AU JOUR	JOURS DE TRAVAIL	MEMBRES	POURCENTAGE COMPLÉTÉ
Mise en place du Projet						
Initialisation du GIT	9/7	9/7	0	1	Nagi/Yann	100 %
Recherche Bibliographique	9/7	9/9	0	3	Nagi/Yann	100 %
Design du programme	9/7	9/8	0	2	Nagi/Yann	100 %
Programmation	9/9	9/14	2	6	Nagi/Yann	100 %
Optimisation – Documentation	9/12	9/14	5	3	Nagi/Yann	100 %
Rapport						
Choix de structure	9/12	9/12	5	1	Nagi/Yann	100 %
Rédaction	9/13	9/15	6	3	Nagi/Yann	100 %
Choix des annexes	9/14	9/14	7	1	Nagi/Yann	100 %
Relecture	9/14	9/15	7	2	Nagi/Yann	100 %
Soutenance						
Préparation des slides	9/14	9/17	7	4	Nagi/Yann	100 %
Répétition	9/15	9/17	8	3	Nagi/Yann	100 %
Soutenance	9/18	9/18	11	1	Nagi/Yann	100 %

Table S2 : *Tâches accomplies et leur période de temps (accompli à gauche, prévu à droite).*

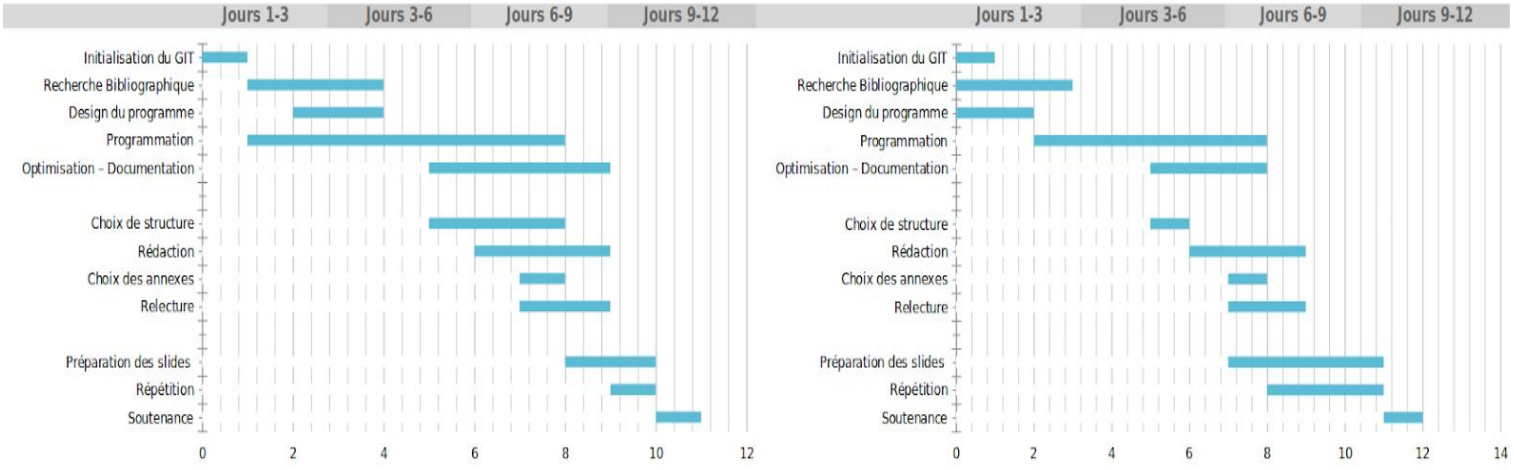

Figure S1 : *Diagramme de Gantt du travail accompli (à gauche) et prévu (à droite).*

Annexe 3

Le programme se structure de la façon suivante :

- ➢ **Traitement des arguments apportés** par la commande à l'aide d'Argparse dans la fonction *args(),* sous forme de variables et vérification de l'existence du fichier. **Sortie** du programme si l'aide est choisie ou que le fichier n'existe pas.

- ➢ **Appel de la fonction *parse_pdb*** lisant le fichier et stockant les informations d'intérêts dans une *Array NumPy* et une liste.

- ➢ **Appel de la fonction *distance_matrix*** *(SciPy)* sur les coordonnées d'intérêt créant une matrice de distance de chaque paire d'atome possible du *PDB*.

- ➢ **Création d'un DataFrame** associé aux résidus protéiques et la matrice de distance, respectant un *threshold* défini par le plus grand *cut-off* ajouté de 3*, afin de stocker toutes les interactions possibles, tout en profitant du calcul rapide de la matrice de distance.*

- ➢ **Parsing du DataFrame** répondant aux critères des *cut-off* pour les interactions Hydrophobes et les liaisons Disulfures sous la forme de deux **DataFrames**.

- ➢ **Appel de *launching_HBOND()*** permettant l'exécution de *HBOND* directement sur internet par le biais du module *Selenium*, à l'aide du navigateur *Mozilla Firefox*, avec une interface graphique désactivée.

- ➢ **Appel de *body_to_list()*** permettant l'extraction et le parsing des résultats de *HBOND* sous forme de **DataFrames correspondants**.

- ➢ **Parsing du DataFrame** répondant aux critères de *cut-off* pour les interactions ioniques stocké en DataFrame.

- ➢ **Calcul et parsing** des centroïdes d'acides aminés aromatiques, recherche des paires correspondantes aux interactions aromatiques-aromatiques sur le DataFrame puis **création** de **DataFrame**.

- ➢ **Création des DataFrames** correspondant aux interactions aromatiques-soufre et cation-π après analyse et calcul selon les *cut-off*.