PROJET TUTORE UNIVERSITAIRE

Exploration de nouvelles approches pour définir et évaluer une famille de régions répétées par alignement en utilisant l'annotation UniProt et les prédiction AlphaFold

Figure n°0 : beta propeller A2RUS52

WITTENMEYER Guillaume
MOSER Mathilda
KESHAVARZ-NAJAFI Moshen

I) Introduction

Les protéines dites WD repeats constituent une large famille de protéines, présente dans tous les organismes eucaryotes. Ces protéines possèdent de nombreuses fonctions cellulaires comme la transduction du signal ou la régulation de la transcription. Ces protéines agissent comme une plateforme de fixation réversible pour la catalyse de réaction protéine-protéine[1]. Par exemple, les protéines à répétitions WD jouent un rôle clé dans des complexes multi-protéiques associés au cycle cellulaire et à la signalisation intracellulaire. L'appartenance à de nombreuses voies cellulaires pourrait montrer une longue évolution de cette famille de protéines. Cette plateforme de fixation est fournie par une structure tridimensionnelle en anneau appelée bêta-propeller. Un bêta-propeller est formé par l'agencement de quatre à huit structures composées majoritairement de quatre brins β antiparallèles. Ces structures sont appelées des blades ou repeats. La formation de la structure du bêta-propellers n'est actuellement pas bien comprise. Quatre à huit blades se rassemblent autour d'un axe central pour former l'anneau du bêta-propeller. Chaque repeat fait environ une quarantaine de résidus. Les différents articles composant la documentation nous proposent des motifs caractéristiques de ces séquences qui permettent de définir ces repeats. Ainsi, dans chaque repeat, nous sommes censés observer un dipeptide GH en début de séquence et un dipeptide WD (ou équivalent) en fin de séquence.[2] Il est également possible que certains repeats possèdent plusieurs fois ces motifs dans leur séquence. Aucun de ces dipeptides ne semblent cependant parfaitement conservés.

Pour ce projet, nous nous intéresserons dans un premier temps à l'étude des alignements des différentes séquences de repeats pour tenter d'en extraire des positions hautement conservées. Nous chercherons ensuite à extraire ces positions hautement conservées au sein des différentes classes de repeats, ce qui nous permettra d'éventuellement de mettre en avant des différences significatives entre ces différentes classes. Enfin, nous entamerons une étude des structures générées par Alphafold via différentes analyses structurales et statistiques si le temps nous le permet. Le but est

[1] Andrew M Hudson and Lynn Cooley, 'Phylogenetic, Structural and Functional Relationships between WD- and Kelch-Repeat Proteins' in Christoph S Clemen, Ludwig Eichinger and Vasily Rybakin (eds), *The Coronin Family of Proteins: Subcellular Biochemistry* (Springer 2008) <https://doi.org/10.1007/978-0-387-09595-0_2> accessed 2 December 2024; Temple F Smith, 'Diversity of WD-Repeat Proteins' in Christoph S Clemen, Ludwig Eichinger and Vasily Rybakin (eds), *The Coronin Family of Proteins: Subcellular Biochemistry* (Springer 2008) <https://doi.org/10.1007/978-0-387-09595-0_3> accessed 2 December 2024.

[2] Temple F Smith and others, 'The WD Repeat: A Common Architecture for Diverse Functions' (1999) 24 Trends in Biochemical Sciences 181.

d'essayer de proposer des classifications de ces différentes classes de repeats pour tenter d'éclaircir la définition des repeats

II) Matériel et méthodes

Nous avons récupéré les séquences avec les annotations d'Uniprot et avons constitué 3 jeux de données. Le premier correspond aux séquences des repeats sans aucunes extensions. Le second set correspond aux séquences des repeats avec 10 acides aminés d'extension de leur séquence. Le troisième set possède 20 acides aminés d'extension de chaque côté de leur séquence. Les trois sets de données ont été réalisés dans le but de compenser les éventuelles erreurs d'annotations présentes dans la banque Uniprot.

Nous cherchons à déterminer si tous les repeats d'un bêta-propeller sont équivalents ou si des différences significatives peuvent être mises au jour. Ainsi, nous commençons tout d'abord par aligner toutes nos séquences issues de ces trois sets.

L'ensemble de nos scripts Python ont été exécutés dans un environnement conda spécifique au projet dans une version de Python 3.9.20. Tous nos scripts et toutes nos données sont disponibles sur le github du projet à cette adresse : https://github.com/EnceladusII/PTU_Project_5

a. Alignements de séquences

i. Alignements initiaux

Pour nos alignements de séquences, nous avons essayé plusieurs outils d'alignements multiples. Les premiers alignements ont été obtenus grâce à ClustalΩ. Ces alignements n'ont pas permis de tirer de conclusions sur les éléments ou les motifs communs entre chaque séquence. Ces alignements ne permettent pas de faire ressortir d'éléments particuliers du fait de la mauvaise prédiction d'alignement. Nous possédons beaucoup de séquences à longueurs et séquences variables. Dans ce cas, ClustalΩ a tendance à insérer beaucoup de gaps qui ne sont pas toujours cohérents. En parallèle, nous avons aussi testé les outils d'alignements MAFFT, MUSCLE et KAlign. Ces trois outils n'ont pas pu donner de résultats, car nos sets de données comportaient un trop grand nombre de séquences. Cela est dû au fait que ces alignements ont été réalisés sur le site web de l'EBI qui est limité dans le nombre de séquences à aligner. Vous pouvez retrouver les quatre outils d'alignements cités ci-dessus à l'adresse suivante : https://www.ebi.ac.uk/jdispatcher/ avec la version Version: 2.1.1+c6bfd1f0. Nous avons ensuite réalisé des alignements avec l'outil MAFFT disponible sur Galaxy Version 7.526+galaxy0.

ii. Alignements filtrés par classes de repeats

Les alignements globaux ne donnant que peu d'informations sur la conservation des résidus de ces repeats, nous avons décidé de rassembler et séparer chaque classe de repeats afin de les étudier indépendamment. Pour garder une uniformité dans nos données, nous avons choisi de nous intéresser uniquement aux domaines à 7 repeats. Ces domaines représentent la majorité des séquences que nous avons récolté dans nos données initiales (environ 143 protéines) et permettent de réduire le bruit induit par les protéines possédant un nombre de repeat[3] variable et rare.

Pour ce faire, nous avons créé un script python (disponible sur le serveur LBGI ou sur le GitHub dont l'adresse est mentionnée plus haut) permettant d'abord de trier nos fichiers fasta initiaux afin de ne garder que les protéines possédant un seul domaine comportant sept blades. Ce script va ensuite rassembler chaque classe de repeat ensemble. Ainsi, nous obtenons un fichier fasta par classe de repeat que nous utiliserons par la suite pour effectuer de nouveaux alignements MAFFT sur Galaxy

De même que pour les alignements sur les séquences non filtrés, les alignements pour les séquences filtrées ont été obtenus grâce à l'outil MAFFT disponible sur Galaxy version 7.526+galaxy0.

b. Création de weblogos

Tous les weblogos ont été générés en ligne via l'interface de seq2logo 2.0. Nous avons réalisé des weblogos pour chaque alignement de repeats (filtrés ou non). Cependant, nous avons décidé de nous intéresser uniquement aux alignements filtrés car ces derniers nous permettront de détecter les différences éventuelles entre chaque classe de repeats et nous permettrons de garder une certaine cohérence dans nos analyses.

Nous avons ensuite voulu épurer les weblogos en ne gardant que les positions les plus conservées et fréquentes. Pour cela, nous avons appliqué deux méthodes pour comparer les résultats obtenus.

Premièrement, nous n'avons gardé uniquement les positions les plus conservées. Ce travail de nettoyage a été fait à la main sur Jalview v2.11.4.0. Nous avons visualisé les alignements avec la coloration Clustal permettant de visualiser les positions qui ont plus de

[3] Hudego, 'Hudego/PTU_Project_3' <https://github.com/Hudego/PTU_Project_3> accessed 7 December 2024.

60 % d'acides aminés avec les mêmes caractères physico-chimiques. Nous avons ensuite sélectionné les positions colorées et qui possédaient également une forte occupancy . L'occupancy représente la fréquence de résidus à une position donnée . Le nettoyage a été réalisé par suppression des positions ne présentant pas de coloration et/ou ne possédant pas une forte occupancy. Le travail étant visuel, nous ne nous sommes pas fixés de valeur seuil pour l'occupancy.

Secondement, nous avons réalisé un script python qui permet de garder uniquement les positions qui ont une occupancy définie. Le script a été lancé avec des seuils d'occupancy de 60 %, 70 % et 80 %. Les alignements résultant du script ont été analysés sous Jalview mais aussi sous forme de Weblogos.

En parallèle, nous avons récupéré via Jalview, les positions conservées avec les pourcentages de présence des acides aminés. Cette analyse a été réalisée exclusivement sur les alignements sans extensions et avec 10 acides aminés d'extension. Nous nous sommes contentés des acides aminés ayant une conservation de plus de 10 % dans les alignements. Nous avons également extrait les pourcentages de chacun des acides aminés pour ces positions conservée. À partir de ces pourcentages, nous avons cherché à comparer les taux de présence de chaque acide aminés à une position donnée et entre chaque classe de repeats. Cette analyse nous permet de voir si nous retrouvons bien les mêmes motifs entre chaque classe de repeats. Elle nous permet ainsi de confirmer et de compléter les weblogos avec des valeurs chiffrées. Par la suite nous nous sommes posé la question d'étudier les différentes variances et moyennes entre les classes de repeats à partir des positions conservées à 10% pour analyser si ces différentes classes divergent ou non. Nous avons généré une version weblogo pour tous les alignements cités ci-dessus. Cependant, nos analyses porteront uniquement sur les alignements et weblogos des séquences ne possédant aucune extension. Ce choix a été fait pour une simplicité de lecture des alignements et weblogos. Les extensions de 10 ou 20 acides aminés, initialement créés pour rattraper une éventuelle mauvaise annotation des repeats dans Uniprot et ne sont pas utiles car occasionne des problèmes dans les alignements (mésalignements, création de gap, disparition de position normalement conservés). De plus, comme les repeats se suivent dans la structure, nous retrouvons souvent les débuts des repeats suivant dans les extensions amino-terminales. Pour ces dernières raisons, il est préférable d'étudier les séquences ne possédant aucune extension.

c. Récupération des structures des protéines d'intérêts

Pour effectuer une analyse structurale, nous avons décidé de récupérer les structures des protéines possédant un seul domaine de 7 blades pour faciliter les études. Pour cela nous avons décidé d'utiliser l'API d'AlphaFold en utilisant Python dans notre environnement conda. Les WD repeats étant des séquences très présentes, ces structures sont généralement bien prédites par cet outil, sauf dans le cas des structures très dynamiques comme peuvent l'être les boucles de jonctions entre les sous-unités de ces domaines protéiques. Il faudra donc faire attention à la qualité des structures prédites.

d. Superpositions de structures

Les structures que nous avons obtenues par AlphaFold représentent les protéines complètes et seule la superposition de chaque classe de repeats nous intéresse. Il nous a donc fallu couper ces domaines dans les fichiers pdbs en utilisant Python et la librairie BioPython 1.84 dans notre environnement conda.

Nous avons ensuite effectué des superpositions de ces structures en utilisant la librairie Python pymol 3.0.0 en utilisant la fonction super() de cette librairie. Le choix de cette fonction de superposition était évident du fait de la grande différence entre chaque séquence de repeats tant en longueur qu'en séquence.

En utilisant les données de RMSD et du nombre d'atomes pris en compte durant la superimposition de ces structures, nous avons décidé de visualiser la répartition de ces superposition en nous basant sur le nombre d'atoms pris en compte durant la superimposition en fonction du log10 du RMSD de cette superposition grâce à la librairie Python Matplotlib 3.9.2.

III) Résultats

Dans cette partie, nous allons aborder les différents résultats obtenus tout au long de ce projet.

a. Alignements de séquences

Les alignements initiaux n'étant visuellement pas concluants, nous avons choisi de simplifier la visualisation des positions conservées en produisant des weblogos basés sur un seuil d'occupancy. La figure 1 est une représentation visuelle sous forme de weblogo de l'alignement des séquences initiales, brutes et ré arrangées pour une occupancy des résidus de 60 à 80%. On peut donc observer dans ces weblogos certains motifs très conservés attendus comme le motif tryptophane-aspartate (WD) de fin de séquence des

repeats pour une occupancy de 80%. Cependant, on peut également observer que le second motif glycine-histidine (GH) de début de séquence n'est pas conservé pour une occupancy de 80% ni de 70% et n'apparaît clairement que pour une occupancy de 60%. On remarque également que la majorité des résidus hautement conservés sont des acides aminés ayant des propriétés hydrophobes qui pourraient avoir un lien avec le site catalytique de ce domaine. Il serait donc intéressant de comparer ces observations avec les weblogos relatifs à chaque classe de repeats pour essayer de visualiser les différences de conservations pour un certain seuil d'occupancy entre ces différentes classes.

Si l'on se penche sur l'observation des weblogos générés pour chaque classe de repeats, nous observons certaines différences. Prenons la figure 2 qui représente les différents weblogos générés pour la classe 6 des repeats. On observe ainsi que pour cette classe de repeats que le dipeptide glycine-histidine n'est pas strictement conservé au-delà de 70% d'occupancy. Cela montre encore une fois que les motifs caractéristiques des WD repeats ne sont pas strictement conservés dans chacune des classes. On observe également que pour une occupancy de 80%, la classe 6 ne semble pas conserver le domaine entier tryptophane-aspartate. Pour pouvoir trouver un motif général et consensus de ces conservations pour ces classes, nous allons donc nous pencher sur l'étude des weblogos ayant une occupancy de 70% pour chaque classe de repeats.

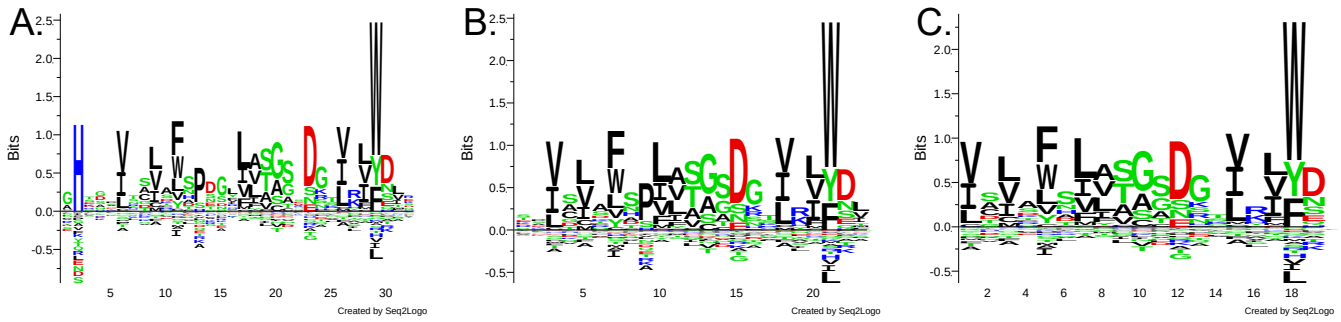

Figure n°1 : Weblogos de l'alignement globaux de séquences sans extensions

A. weblogo de l'alignement de séquences globales pour une occupancy de 60% B. weblogo de l'alignement de séquences globales pour une occupancy de 70% C. weblogo de l'alignement de séquences globales pour une occupancy de 80%

Représentation graphique des alignements de séquences sans extensions et filtré pour des occupacy de 60 % (A.), 70 % (B.) et 80 % (C.). Dans les trois niveaux de filtre, nous constatons la présence d'un motif WD très conservé. Nous constatons aussi la présence de plusieurs positions qui semble être hydrophobe.

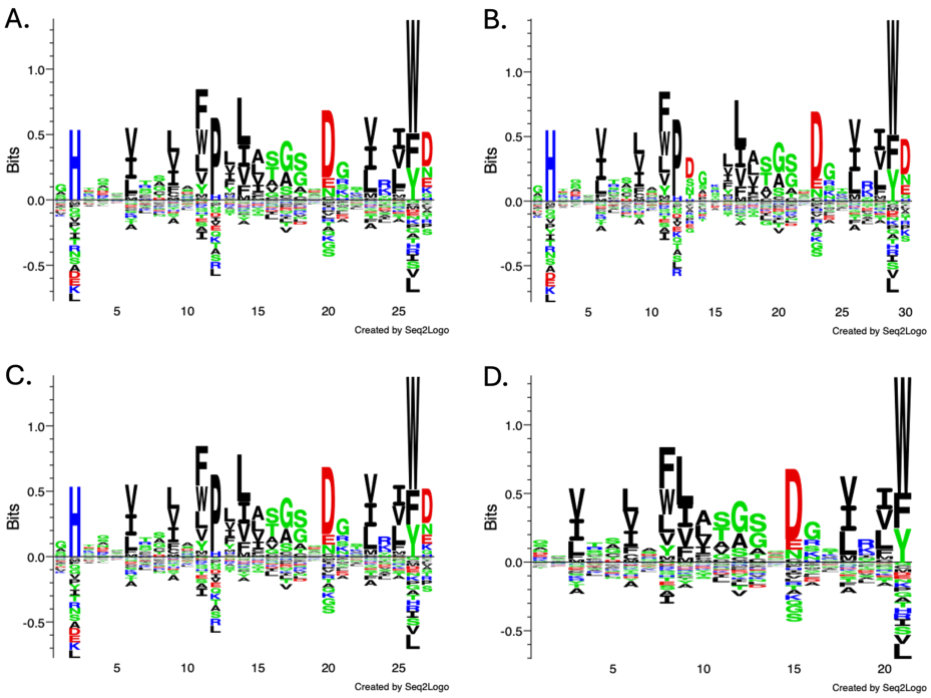

Figure n°2 : Weblogos de l'alignement de séquences pour la classe N°6

A. weblogo de l'alignements de séquences pour la classes 6 nétoyés à la main ;

B. weblogo de l'alignements de séquences pour la classes 6 ne conservant que les positions ayant au minimum 60% d'occupancy ;

C. weblogo de l'alignements de séquences pour la classes 6 ne conservant que les positions ayant au minimum 70% d'occupancy

D. weblogo de l'alignements de séquences pour la classes 6 ne conservant que les positions ayant au minimum 80% d'occupancy

Nous constatons ici que les dipeptides WD et GH apparaissent bien sauf pour le weblogo d'occupancy à 80 %.

La figure numéro 3 représente les weblogos pour une occupancy de 70% générés pour les alignements de chaque classe de repeats. On y observe que la classe numéro 7 possède un aspartate du motif WD très timide, et cela même pour une occupancy de 70%. Le fait que la classe numéro 6 et 7 ne conservent pas tout le temps ce motif WD à de très hautes occupancy peut expliquer les résultats obtenus plus haut sur les alignements initiaux montrant un non conservation systématique des motifs caractéristiques des repeats. Nous pouvons ainsi observer au sein de tous les repeats un motif récurrent :

G-H-[V,I,L]-x-[L,V,I]-[F,W]-polaire-[L,I,V]-hydrophobe-[S,T]-G-S-D-G-[V,I,L]-[R,K]-[L,V,I]-W-D

En complément de cela nous avons également manuellement annoté les conservations supérieures à 10% dans les alignements des séquences de chaque classe de repeats pour les protéines de nos alignements pour pouvoir en calculer les variances et moyennes différences de conservations entre ces différentes classes. La figure 4 représente la variance entre les classes pour la conservation d'un résidu à une position. On y remarque qu'il y a deux résidus pour lesquels la variance change de façon significative par rapport aux autres positions:

- Une glycine en milieu de séquence qui ne semble pas appartenir à un motif particulier, ce qui pourrait expliquer cette variance
- L'aspartate du motif tryptophane-aspartate de fin de séquence.

La variance plus élevée de cet aspartate pourrait signifier que certaines classes de repeats conservent moins ce residue que d'autres classes. Cela rejoint les quelques observations précédentes pour lesquelles nous avons mis en évidence qu'à certains seuils d'occupancy, ce motif n'était pas strictement conservé ou alors présente une faible présence en occupancy de 70%.

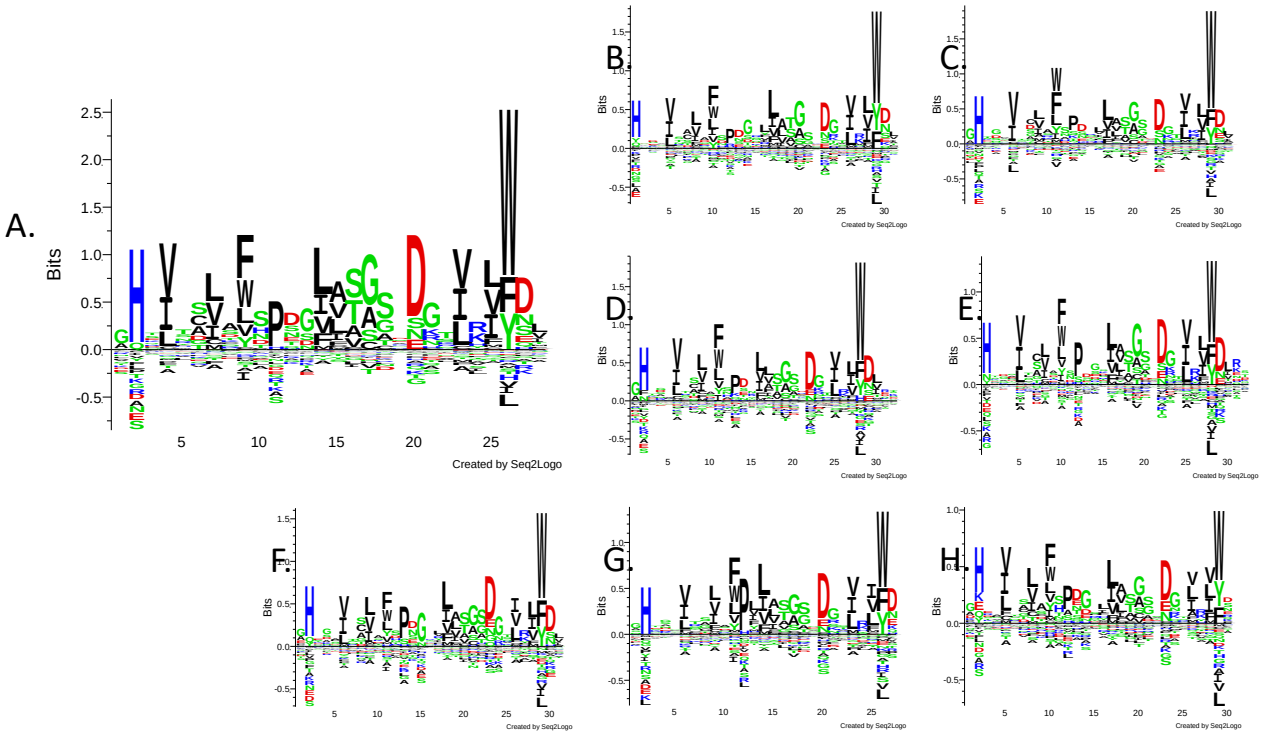

Figure n°3 : Weblogos des alignements de séquences de chaque classes de repeat pour les domaines à sept blades sans extensions pour une occupancy de 70%

A. Weblogo pour l'alignement global rassemblant tous les repeats des domaines à sept blades ; B. Weblogo pour l'alignement pour les repeat de classe 1 ; C. Weblogo pour l'alignement pour les repeat de classe 2 ; D. Weblogo pour l'alignement pour les repeat de classe 3 ; E. Weblogo pour l'alignement pour les repeat de classe 4 ; F. Weblogo pour l'alignement pour les repeat de classe 5 ; G. Weblogo pour l'alignement pour les repeat de classe 6 ; H. Weblogo pour l'alignement pour les repeat de classe 7

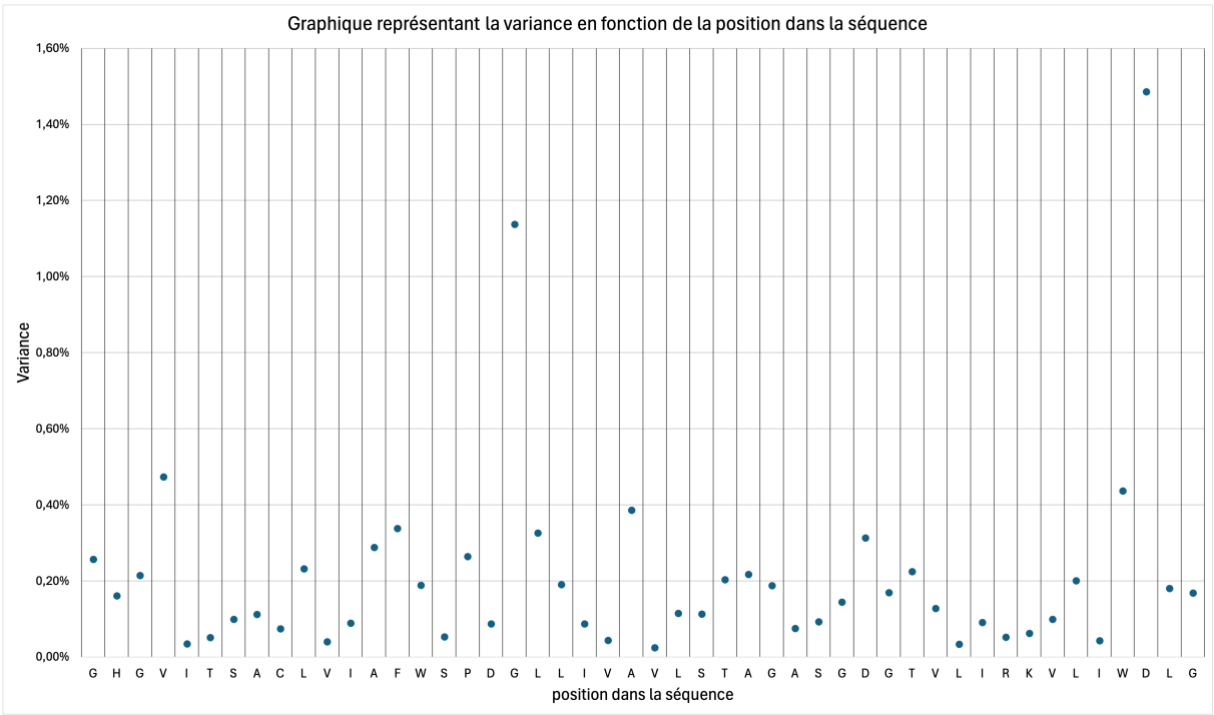

Figure n°4 ; Graphique représentant la variance en fonction de la position dans la séquence

b. Analyse structurale des repreat entre eux

Nous avons ensuite entreprit de superimposer les structures dans un logiciel de visualisation moléculaire comme PyMol. Pour cela nous avons pour chaque classe déterminé quelle structure de référence il était préférable d'utiliser grâce à un script Python. Pour cela, nous avons déterminé comme structure de référence la structure pour laquelle les moyennes des RMSD calculées durant la superimposition est la plus faible. On obtient donc la figure 5 dans laquelle on peut observer que les structures de références pour chaque classe de repeats sont différentes exceptées pour les classes 5 et 6. On remarque que les moyennes des RMSD sont comprises entre 0.94 et 2.06 Å. On observe aussi que le nombre moyen d'atomes pris en compte durant ces superimpositions est compris entre 36.9 et 157.03 ce qui représente des différences de maxima grands. Un nombre d'atomes pris en compte élevé associé à une valeur de moyenne de RMSD faible est synonyme d'une superimposition réussie sur une grande partie de la structure pour un grand nombre de repeats. Mais on remarque surtout que certaines classes de repeats ne semblent pas assez bien se superimposer. En effet, on remarque pour la classe 4 un RMSD d'environ 2 Å, ce qui est plus que les autres classes. Cela pourrait s'expliquer par le fait que le nombre d'atomes moyen pris en compte durant la superimposition est environ 3 fois inférieur par rapport aux autres classes (= 57 atomes en moyenne). La classe numéro 7 est aussi étrange car, bien qu'elle possède une moyenne des RMSD proche des autres classes (environ 1.25 Å) elle possède un nombre d'atomes pris en compte dans la superimposition très faible (environ 37) ce qui pourrait être dû à l'utilisation d'une mauvaise structure de référence.

Pour nous donner un aperçu de la qualité des superimpositions par rapport à la référence de chaque classe de repeats, nous avons généré des graphiques représentant pour chaque classe le nombre d'atomes pris en compte dans la superimposition en fonction du log10 des RMSD (figure 6). On remarque grâce à ces graphiques que les superimposition de chaque classe de repeats possèdent des superimpositions très hétérogènes avec pour les classes 4 et 7 des répartitions très faibles dûes au faible nombre d'atomes pris en compte dans chaque superimposition, ce qui se traduit par de mauvaises superimposition globales (figure 7). On pourrait aussi parler de la classe 1 qui semble avoir une répartition des superimpositions très dispersée. Cela pourrait se traduire par des structures qui sont plus éloignées dans cette classe de repeats et donc qui s'aligneraient très différemment les unes des autres. Cela peut aussi être dû à un nombre variable de feuillets bêta dans les structures téléchargées sur AlphaFold .

Class of WD repeats	Reference PDB ID	RMSD Avg	Atoms number pick during super() Avg
1	Q6ZMY6	1.5221855251023368	153.5182481751825
2	Q9UKT8	1.160593188374582	130.86861313868613
3	Q9NRL3	0.9477170253322073	157.02919708029196
4	Q9BRP4	2.0674683901950393	56.802919708029194
5	Q9Y4P8	1.2597052256266277	148.5
6	Q9Y4P8	1.421951580306758	141.7173913043478
7	P57081	1.2092456939446665	36.89051094890511

Figure n°5 : tableau présentant pour chaque classe le modèle PDB choisit comme référence, le moyenne des RMSD et la moyenne du nombre de point utilisé pour la super-imposition des structures

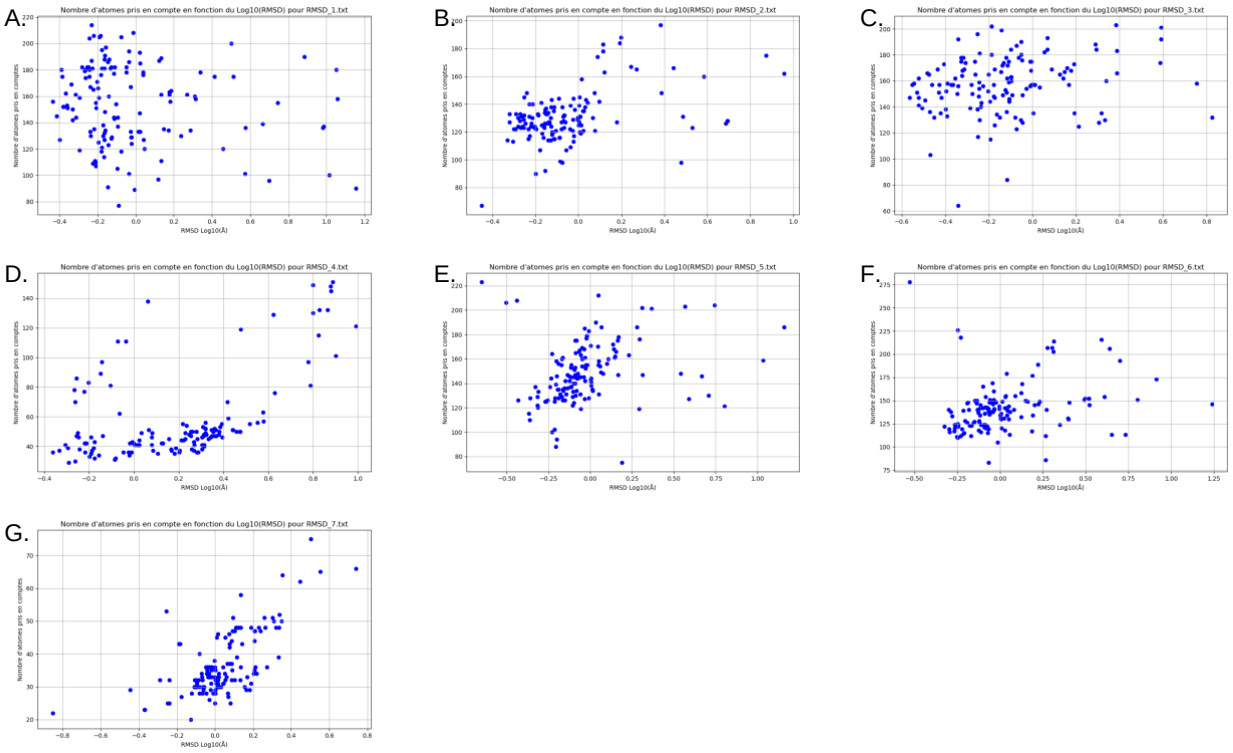

Figure n°6 : Graphiques représnetant le nombre d'atome pris en compte pour le superpose par rapport au Log10 du RMSD en Å

A. Représentation pour la classe 1

B. Représentation pour la classe 2

C. Représentation pour la classe 3

D. Représentation pour la classe 4

E. Représentation pour la classe 5

F. Représentation pour la classe 6

G. Représentation pour la classe 7

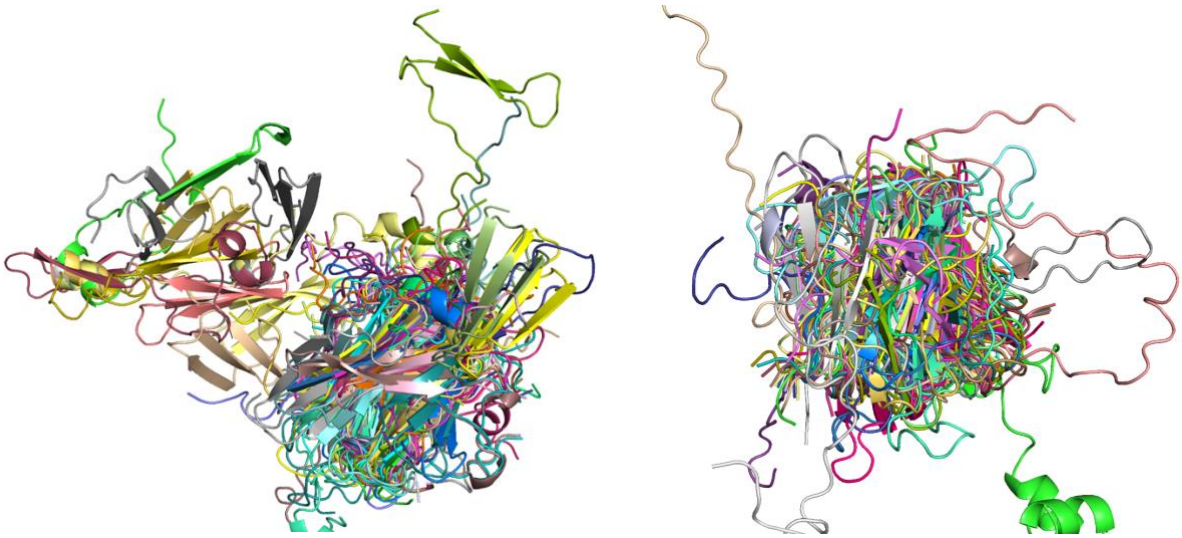

Figure n°7 : Super-imposition des structures des classes 7 et 4

A. Super-impose des structures de repeats de la classe 7

B. Super-impose des structures de repeats de la classe 4

IV) Discussion

Dans cette étude nous avons donc mis en évidence que les différentes classes de WD repeats qui composent les bêta-propellers sont difficiles à étudier dans des alignements du fait de leur grande variabilité de longueurs et de séquences. Cependant, nous avons mis en évidence que la majorité des résidus conservés étaient de nature hydrophobe. De part nos alignement nous avons tout de même réussi à extraire un motif qui semble conservé à une occupancy de 70% pour l'ensemble des classes de repeats:

G-H-[V,I,L]-x-[L,V,I]-[F,W]-polaire-[L,I,V]-hydrophobe-[S,T]-G-S-D-G-[V,I,L]-[R,K]-[L,V,I]-W-D

Nos études ont également mis en évidence la non conservation systématique des motifs caractéristiques G-H de début des séquences de ces repeats et du motif W-D de par les différents weblogos des alignements ainsi que l'analyse des variances entre les classes de repeats. Ces observations peuvent potentiellement être améliorées en peaufinant les paramètres utilisés lors de la génération des alignements par MAFFT, ce qui améliorerait les alignements, faisant ressortir à de plus grandes occupancy ces motifs caractéristiques. Lors de nos observations, nous n'avons également pas mis en évidence de répétition des motifs W-D.

Les analyses des superimpositions de ces différentes classes nous ont aussi appris que certaines structures de repeats de même classe semblent moins bien se superimposer les unes aux autres et donc être plus différentes les unes des autres comme les classes 4 et 7 qui ont un faible nombre d'atomes utilisés dans la superimposition. Ces résultats peuvent potentiellement s'expliquer par le choix de la référence utilisée pour la superimposition. En effet, même si elle représente la structure dont la moyenne des RMSD est la plus faible, ce choix ne prend pas en compte le nombre d'atomes pris en compte durant ces superimposition. Il suffit donc que cette structure de référence soit très différente structurellement mais que les scores de RMSD associés soient faibles pour induire en erreur le reste des superimpositions. Il serait donc intéressant de mettre au point un protocole permettant de déterminer quelle structure parmi la liste des repeats d'une même classe représente la meilleure référence de superimposition en prenant en considérations les scores de RMSD et le nombre d'atomes pris en compte lors de cette superimposition.

Par soucis de temps dans le cadre de ce projet, nous n'avons pas pu effectuer d'analyses structurales plus approfondies permettant de réellement améliorer la séparation des WD

repeats en plusieurs classes. Néanmoins, il serait intéressant pour continuer ce projet de créer de nouveaux protocoles d'analyses structurelles qui se basent sur les observations de conservation de la figure 4 et qui mettent en évidence ces résidus dans un logiciel de visualisation moléculaire comme PyMol ou Discovery Studio qui permettront d'observer les différentes interactions effectuées par ces résidus et émettre des hypothèses sur les raisons de leurs conservations ou non.

V) Bibliographie

Guillaume W, 'EnceladusII/PTU_Project_5' <https://github.com/EnceladusII/PTU_Project_5> accessed 7 December 2024

Hudego, 'Hudego/PTU_Project_3' <https://github.com/Hudego/PTU_Project_3> accessed 7 December 2024

Jumper J and others, 'Highly Accurate Protein Structure Prediction with AlphaFold' (2021) 596 Nature 583

Madeira F and others, 'The EMBL-EBI Job Dispatcher Sequence Analysis Tools Framework in 2024' (2024) 52 Nucleic Acids Research W521

Smith TF, 'Diversity of WD-Repeat Proteins' in Christoph S Clemen, Ludwig Eichinger and Vasily Rybakin (eds), *The Coronin Family of Proteins: Subcellular Biochemistry* (Springer 2008) <https://doi.org/10.1007/978-0-387-09595-0_3> accessed 2 December 2024

——, 'The WD Repeat: A Common Architecture for Diverse Functions' (1999) 24 Trends in Biochemical Sciences 181

The Galaxy Community, 'The Galaxy Platform for Accessible, Reproducible, and Collaborative Data Analyses: 2024 Update' (2024) 52 Nucleic Acids Research W83

Thomsen MCF and Nielsen M, 'Seq2Logo: A Method for Construction and Visualization of Amino Acid Binding Motifs and Sequence Profiles Including Sequence Weighting, Pseudo Counts and Two-Sided Representation of Amino Acid Enrichment and Depletion' (2012) 40 Nucleic Acids Research W281

Zhang Z and others, 'Function and Regulation of F-Box/WD Repeat-Containing Protein 7' (2020) 20 Oncology Letters 1526