Etude du complexe sRNP H/ACA

Marine Figarol & Christelle Viguier

UE Langage de Script Python M1 BIBS Université Paris-Saclay

24 mai 2018



Table des matières

Table des matières	1
Introduction	2
Résultats	
Analyses globales	
Changement conformationnel	
Flexibilité intra-domaine	
Analyses locales	4
Distance de contact protéine-ARN	4
Fréquence de contact protéine-ARN	5
Fréquence de contact des résidus clé entre eux et avec l'ARN	6
Discussion et conclusion	8
Annexes	9

Introduction

Le complexe sNRP H/ACA est connu pour son rôle dans l'isomérisation de l'uridine (U) en pseudo-uridine (ψ) sur les brins d'ARN. Il s'agit d'une modification post-transcriptionnelle appelée pseudo-uridylation. La pseudo-uridine possède un groupe NH supplémentaire capable de former une liaison hydrogène, et permet ainsi la stabilisation de l'hélice d'ARN. Le complexe sNRP H/ACA est un complexe ribonucléique qui permet donc l'interaction entre un sRNA et cet ARN cible. La protéine est constituée de 4 domaines : aCBF5 (A1), PUA (A2), aNOP10 (A3) et L7Ae (A4) et aGAR1 (absent des données) liés à l'ARN (annexe 1).

L'objectif de ce projet est d'analyser la flexibilité locale et globale du complexe sNRP H/ACA lors de l'interaction avec l'ARN, et de mesurer le temps de contact entre des résidus clés déterminés à partir d'une étude sur un mutant. Ces études sont menées dans le but de déterminer le rôle des différents domaines du complexe dans l'interaction avec l'ARN. Afin de répondre à ces questions, nous avons développé des scripts Python permettant de manipuler les fichiers sources et d'effectuer les calculs nécessaires.

Objectifs et principe

Nous disposons de la structure de référence du complexe sRNP H /ACA, ainsi que de 500 conformations du complexe issues d'une dynamique moléculaire de 10ns en milieu aqueux dans deux fichiers au format PDB. Deux approches sont explorées : une approche globale analysant les changements de conformation des domaines, et une approche locale permettant de focaliser l'analyse au niveau des résidus totaux et spécifiques.

Dans un premier temps, la création d'un parseur de fichier PDB a été réalisé. Les informations pertinentes des atomes sont stockées et ordonnées dans des dictionnaires. Pour la suite du projet, nous travaillons à partir de ces dictionnaires pour répondre aux différentes problématiques.

L'étape suivante a pour but de déterminer les changements conformationnels de la protéine en solution aqueuse. Cette étape permet de mettre en évidence les résidus ayant une importance au sein du complexe, traduisant une forte stabilité de celui-ci. Pour ce faire, nous effectuons un calcul de RMSD (Root Mean Square Deviation) entre les mêmes atomes de la structure de référence et ceux des 500 conformations de la protéine. Le calcul du RMSD se fait sur les carbones alphas de chaque atome. Ainsi nous obtenons la distance moyenne entre les atomes de la dynamique protéique.

Une mesure de RMSD faible indique qu'il y a peu de déviation structurale entre la protéine de référence et une conformation de cette protéine à un temps donné dans un milieu aqueux. Le calcul est le suivant :

$$ext{RMSD} = \sqrt{rac{1}{N}\sum_{i=1}^N \delta_i^2}$$

Avec : - N le nombre de pair d'atomes

- δi la distance entre un atome i et la structure de référence Le résultat est en Ångström (Å).

Deux seuils de RMSD sont définis :

RMSD < 3Å: les deux structures sont à peu près identiques
 RMSD > 5Å: l'écart entre les deux structures est significatif

Lors de la troisième étape, nous effectuons des calculs de RMSD croisé afin de visualiser la flexibilité intra-domaine.

Par la suite, une recherche de résidus constituant l'interface protéine/ARN est réalisée. Dans un premier temps, la distance de contact entre la protéine et l'ARN est calculée. Puis, la fréquence de contact protéine-ARN est déterminée afin de visualiser tous les résidus appartenant à l'interface.

Dernièrement, nous calculons la fréquence de contact entre différentes paires de résidus clés. Ces résidus clé sont ceux jouant un rôle clé dans l'interaction avec l'ARN constituant l'interface protéine-ARN. Ces derniers sont déterminés à partir de l'analyse sur un mutant.

Résultats Analyses globales

Changement conformationnel

Le calcul du RMSD a été réalisé pour chaque domaine protéique des 500 conformations. On observe ainsi l'évolution des différents états de conformation de la calibration de la protéine dans son système.

Concernant les domaines A1 et A2 (annexe 2), on observe une augmentation globale du RMSD jusqu'à atteindre un plateau. L'augmentation est due à la mise en solution de la protéine à partir de la forme cristallisée. Ce plateau signifie que les résidus du domaine sont stables, donc le domaine est peu flexible. Les résultats indiquent que la distance entre les résidus par rapport à la conformation initiale a différée d'environ 1,7Å. On reste donc en dessous du seuil de 3Å signifiant que la protéine en milieu aqueux a une structure très proche de la protéine de référence.

Les domaines A3 et A4 possèdent un comportement similaire aux domaines A1 et A2, mais avec un RMSD plus important au niveau du plateau (environ 2,5Å). Ces domaines sont donc plus flexibles mais restent globalement inférieurs au seuil.

Flexibilité intra-domaine

Des calculs de RMSD croisés ont également été réalisés : le RMSD est ainsi calculé sur chaque domaine, entre une conformation donnée et toutes les autres. Les résultats sont présentés dans l'annexe 3. Les résultats sont complémentaires à ceux de la partie *Changement conformationnel*. En effet, on remarque que les domaines ont un RMSD global similaire : A1 et A2 ont un RMSD maximal de 2,5Å et A3 et A4 possèdent un RMSD maximal de 3,5Å.

Mis en milieu aqueux, le domaine A3 est très stable au cours du temps. Les 3 autres domaines apparaissent relativement plus flexibles, ce qui est corrélé avec les résultats de l'annexe 2.

Analyses locales

Distance de contact protéine-ARN

Le calcul de la matrice de distance entre chaque domaine protéique et l'ARN est représenté en HeatMap dans l'annexe 4. Les calculs ont été effectués sur le barycentre de chaque molécule. Il est ainsi possible de visualiser globalement les zones protéiques en interaction avec des résidus précis d'ARN. On remarque notamment que le domaine A4 est en forte interaction avec une extrémité du brin d'ARN (distance faible entre les résidus et les acides ribonucléiques 20 à 40). En effet, la visualisation PyMol (annexe 1) montre que le domaine est localisé à une extrémité de l'ARN. Les résultats sont donc cohérents avec la visualisation PyMol.

Cette matrice permet donc de localiser les résidus en interaction proche avec l'ARN. Ce calcul est nécessaire pour déterminer la fréquence de contact protéine-ARN étudié dans les parties suivantes

Fréquence de contact protéine-ARN

En calculant la distance de contact protéine-ARN pour chacune des 500 conformations à partir du barycentre, nous obtenons la fréquence de contact de chaque résidu. Différents seuil de distance en Ångström sont testés afin de déterminer la distance limite pour considérer un résidu comme appartenant à l'interface. Les résultats sont enregistrés dans le(s) fichier(s) *interface.txt*.

Dans un premier temps, nous avons pris comme distance maximale 4Å car elle correspond à une distance inter-atome moyenne⁽¹⁾. Comme nous n'avons obtenu aucun résidu appartenant à l'interface avec ce seuil, plusieurs tests ont été effectués avec des seuils plus élevés, reportés dans le tableau 1.

Distance maximale (en Å)	Nombre de résidus appartenant à l'interface
4	0
6	22
8	62
9	81
10	93
14	167

Tableau 1 : Résultats des différents tests de seuil maximal de distance

Au vu des résultats obtenus, le seuil de 9Å semble être un bon compromis. En effet, il y a très peu de différence entre un seuil à 9Å et un seuil à 8 ou 10Å. Afin de contrôler le choix du seuil, une visualisation sur PyMol est réalisée en mettant en évidence les résidus estimés comme appartenant à l'interface (figure 1).



Figure 1 : Représentation des résidus appartenant à l'interface avec un seuil de distance de 9Å (en rouge) et résidus non considérés comme appartenant à l'interface (en vert).

Visuellement, les résidus à l'interface pour un seuil de 9Å semblent cohérents. On considère donc ce seuil comme étant satisfaisant.

Fréquence de contact des résidus clé entre eux et avec l'ARN

La fréquence de contact des résidus clés a été calculée sur les 500 conformations de la protéine. Le calcul s'est fait par rapport aux autres résidus de cette protéine ainsi qu'avec les acides ribonucléiques de l'ARN. La fréquence de contact entre deux résidus est obtenue en calculant le nombre de conformations où il y a contact (distance <= 9Å) par rapport au nombre total de conformations (500). Les résultats sont ainsi stockés dans le fichier *residus_cles.txt*.

Les résidus clés ont été choisis arbitrairement parmi ceux présentés dans la l'annexe 5, issu de la thèse scientifique sur laquelle se base notre étude ⁽¹⁾. Les résultats sont reportés dans le tableau 2.

⁽¹⁾ Thèse: Sébastien MULLER, Développement et application de méthodes bioinformatiques pour la recherche de gènes et sRNA à boîtes H/ACA dans les génomes d'archaea et étude fonctionnelle des ARN et sRNP H/ACA correspondants, 2007.

(2)

Résultats obtenus par algorithme			Résultats décrits dans la littérature
Résidu (domaine)	Résidu (domaine)	Fréquence de contact mesurée	Etat de la liaison après mutation
34 (A3)	33 (ARN)	0.0	Création d'une nouvelle liaison
100 (A4)	31 (ARN)	1.0	Aucun changement
98 (A4)	30 (ARN)	0.97	Affaiblissement ou perte de la liaison
6 (A3)	63 (A4)	0.972	Renforcement de la liaison
41 (A4)	32 (ARN)	1.0	Aucun changement
26 (A3)	65 (A4)	0.828	Affaiblissement ou perte de la liaison
66 (A4)	41 (A3)	0.68	Renforcement de la liaison
46 (A4)	25 (ARN)	0.988	Renforcement de la liaison

Tableau 2 : Résultats de la recherche des résidus clé en comparaison avec ceux de la littérature

Ces résultats sont corrélés avec une analyse qualitative par visualisation sur PyMol. Par exemple, nous n'obtenons aucun contact entre le résidu 34 du domaine A3 et le résidu 33 de l'ARN sur toutes les 500 configurations analysées. Ce résultat est concordent avec l'observation 3D que nous avons pu faire (figure 2).

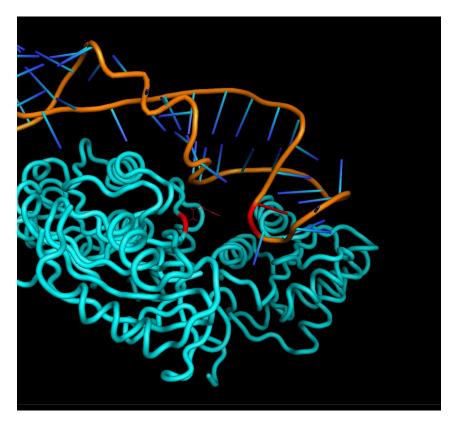


Figure 2 : Analyse 3D sur PyMol. En rouge : les résidus analysées. Résidu 34 du domaine A3 (gauche) et résidu 33 de l'ARN (droite).

D'après la figure 2, nous observons en effet une distance importante entre ces deux résidus, rendant toute interaction impossible.

Discussion et conclusion

Lors de cette étude, de nombreux aspects de l'interaction de la protéine sRNP avec l'ARN ont pu être mis en évidence.

Tout d'abord, il s'agit d'une protéine globalement stable au cours du temps lorsqu'elle est liée à l'ARN. En effet, tous les domaines et résidus ont une faible variation du RMSD, n'atteignant jamais le seuil de 5Å sur les 500 conformations étudiées.

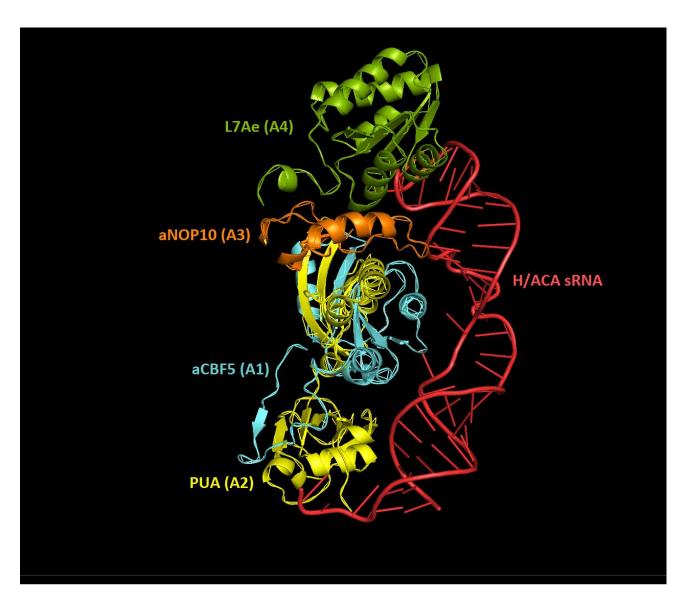
Cette stabilité sur l'ensemble de la protéine est explicable par le fait que tous les domaines sont en contact avec l'ARN, bien que le domaine A3 ait moins de résidus formant l'interface de contact (justifié par le fait que le domaine est relativement petit).

Afin d'obtenir ce résultat, nous avons choisi un seuil minimal de distance à 9Å, ce qui correspond à une distance élevée. En effet, nous avons calculé les distances inter-résidus à partir de leur barycentre (centre de masse de la molécule) et non pas pour chacun des atomes de chaque résidu. De ce fait, les plus proches atomes ne sont pas pris en comptes seuls dans le calcul. Cependant, en augmentant le seuil de distance, nous obtenons un résultat cohérent avec l'étude sur laquelle le projet est basé.

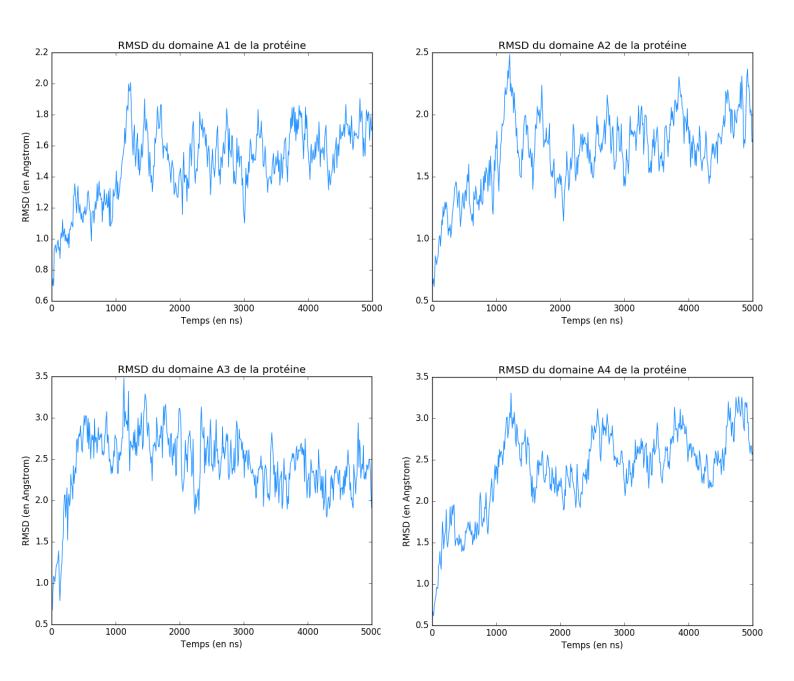
Il est intéressant de constater qu'il y a des différences d'interactions protéine/ARN entre la protéine sauvage et sa forme mutée. Parmi les 8 résidus clé analysés lors de ce projet, seuls 2 n'ont pas de modification dans leur interaction avec un résidu donné. Il y a donc un changement de conformation net chez le mutant, ce qui expliquerait l'altération de l'activité catalytique de la protéine.

Annexes

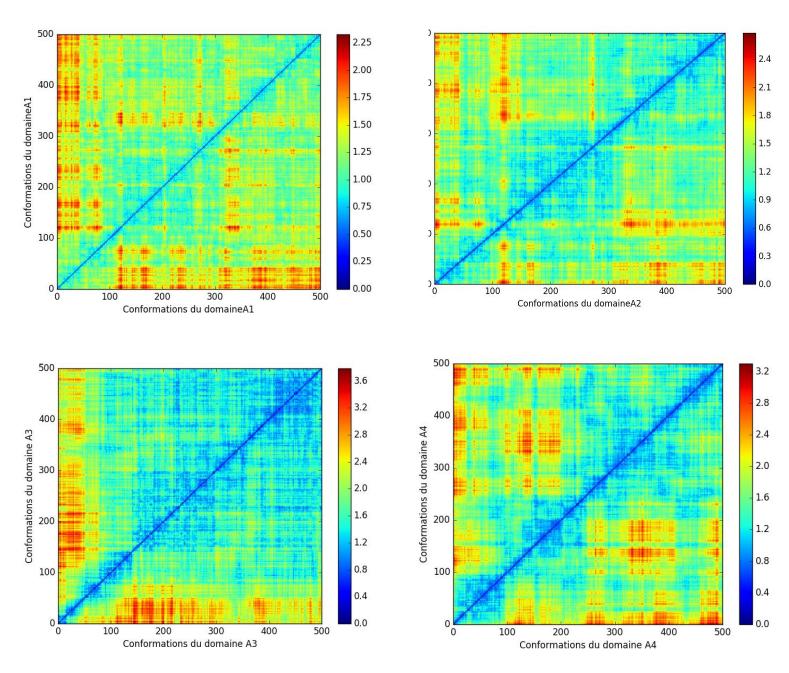
- Annexe 1 : Visualisation de la structure de référence du complexe sNRP H/ACA sur PyMol.
- Annexe 2 : Dynamique des domaines quatre domaines de la protéine en milieu aqueux au cours du temps.
- Annexe 3 : HeatMap de distance intra-domaine au cours du temps. Rouge : distance élevé et bleu : distance faible (en Å).
- Annexe 4 : HeatMap de distance entre l'ARN et les résidus des différents domaines protéiques (en Å).
- Annexe 5 : Contacts et résidus clés chez le mutant.



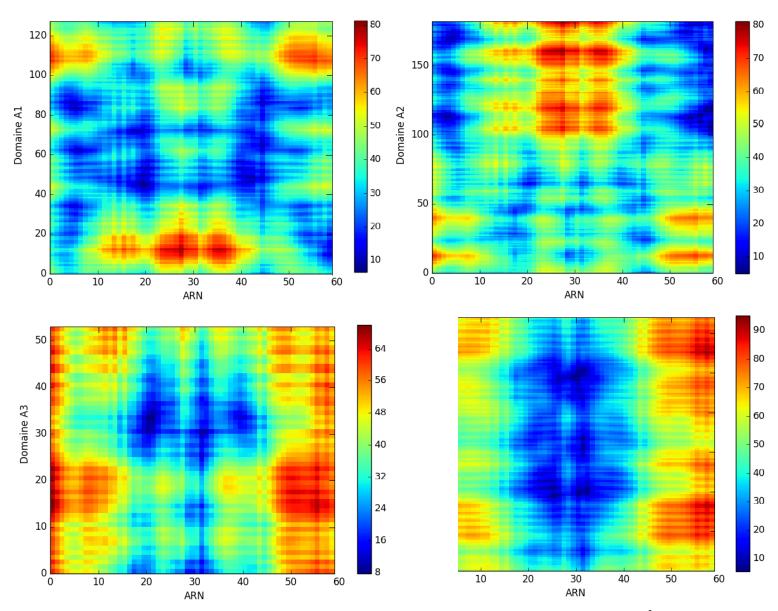
Annexe 1 : Visualisation de la structure de référence du complexe sNRP H/ACA sur PyMol.



Annexe 2 : Dynamique des domaines quatre domaines de la protéine en milieu aqueux au cours du temps.

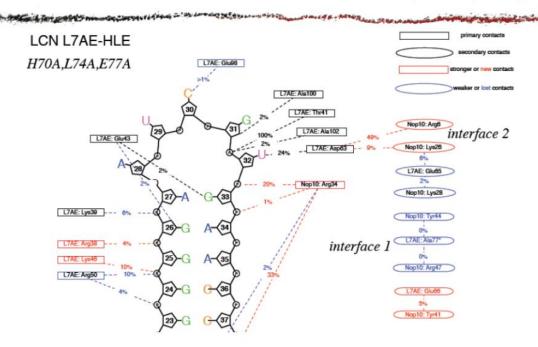


Annexe 3 : HeatMap de distance intradomaine au cours du temps. Rouge : distance élevé et bleu : distance faible (en Å).



Annexe 4 : HeatMap de distance entre l'ARN et les résidus des différents domaines protéiques (en Å).

Contacts at Nop10/L7Ae interface: from MD simulations on sRNP H/ACA



Annexe 5 : Contacts et résidus clés chez le mutant.