

Projet informatique : Étude du complexe sNRP H/ACA

UE : Langage de script (Python)
M1 BIBS 2017
Université Paris-Sud

Maxime Chaput
Jonathan Torel Salat

Introduction

La pseudouridine ψ est un nucléotide dérivé de l'uridine, celle-ci n'est pas incorporée lors de la transcription mais résulte d'une modification post-transcriptionnelle de certains résidus d'uridine. Elle est abondante dans les ribosomes et est indispensable au bon fonctionnement de la traduction. Cette modification est une isomérisation de l'uridine ce qui permet à la pseudo-uridine de comporter un second groupe NH, utile pour stabiliser leur structure.

Ce sont les complexes sRNP H/ACA qui catalysent cette transformation, ils sont composés de plusieurs protéines dont notamment la protéine Cbf5 (aCBF5 chez les archaeae) qui porte l'activité enzymatique composée de plusieurs domaines.

L'objectif de ce projet est de comprendre grâce à l'analyse des changements conformationnels globaux et locaux de ce complexe et de ses différents domaines, l'impact et l'utilité de ceux-ci ainsi que de leurs résidus dans l'isomérisation de l'uridine.

Matériel et Méthode

Données

Dans ce but nous disposons :

- d'un fichier au format PDB comportant les structures des 500 conformations de la protéine en solution, obtenu par dynamique moléculaire de 10ns.
- d'un fichier au format PDB contenant la structure de référence du complexe sRNP H/ACA

Puisqu'il n'existe pas de structure 3D de ce complexe chez les Eucaryotes, celui étudié correspond aux archaeae.

Analyse des changements conformationnels globaux

La première étape consiste à écrire un programme permettant d'étudier si la protéine subit des changements conformationnels majeurs en solution. Pour cela la comparaison via la superposition de structure entre la structure de référence et les structures des 500 conformations sera la méthode utilisée. En effet le RMSD global de la protéines et pour chacun des cinq domaines tout au long de la dynamique pourra ainsi être calculé.

Le RMSD (Root-Mean-Square Deviation) où « déviation de la racine de la moyenne des carrés » consiste à comparer des valeurs théoriques (référence) avec des valeurs observées (conformation étudiée).

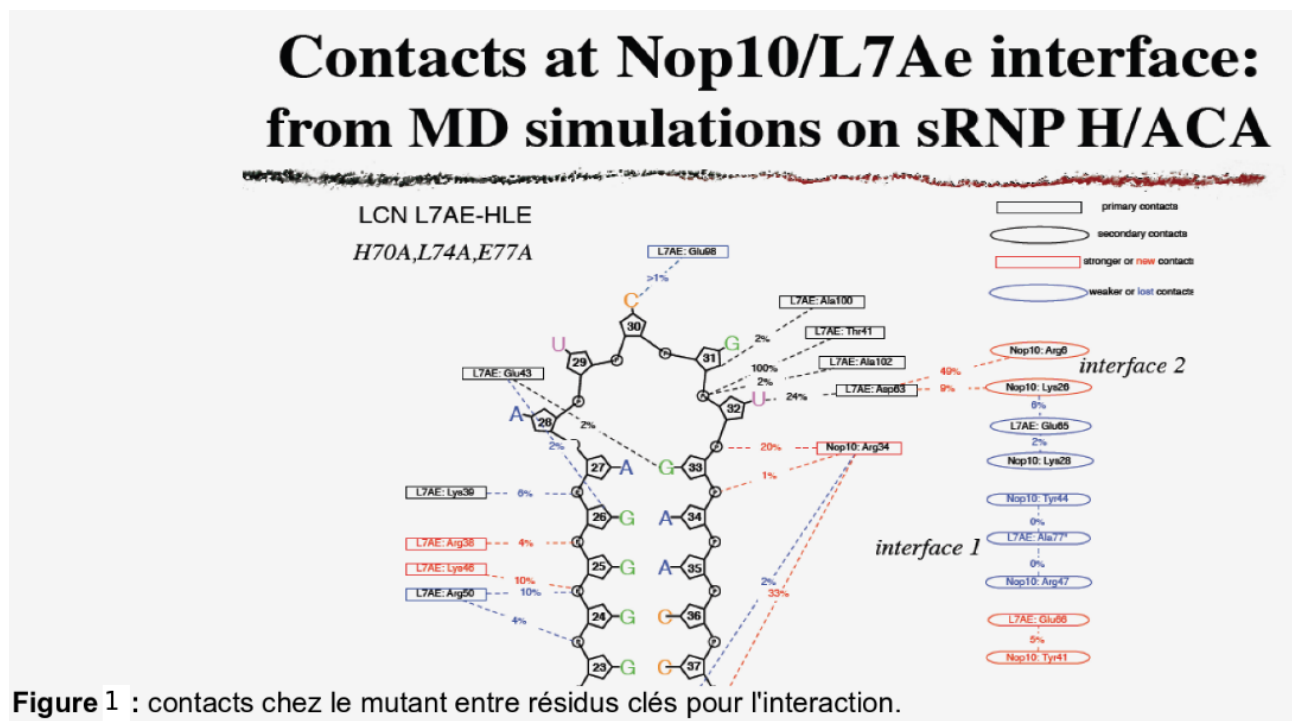
$$RMSD = \sqrt{\frac{\sum_{c=1}^n ((x_r - x_c)^2 + (y_r - y_c)^2 + (z_r - z_c)^2)}{n}}$$

Dans cette formule x_r , y_r et z_r (resp. x_c , y_c et z_c) sont les coordonnées spatiales des atomes de la conformation de référence (resp. la conformation étudiée), n étant le nombre d'atomes comparés. Le but étant de donner une idée de la distance qui les sépare, le RMSD est calculé en Angström (Å).

Afin de diminuer le temps de calcul notre programme calculera au préalable les centre de masse des résidus concernés et s'en servira dans le calcul du RMSD.

Analyse des changements conformationnels locaux

Les temps de contact entre plusieurs résidus de la protéines et certains résidus de la protéines avec l'ARN ont été mesurés au cours de la dynamique moléculaire avec le mutant et ont permis d'identifier des résidus jouant un rôle clé dans l'interaction avec l'ARN (**Figure1**).



Dans un premier temps en identifiant pour chaque conformation les résidus des domaines protéiques appartenant à l'interface (distance inférieure à 10 Å entre ces

résidus et le brin d'ARN) et en calculant leurs fréquence pour l'ensemble des conformations il sera possible d'obtenir de façon précise les changements conformationnels locaux de ces résidus clés pour lors de l'interaction avec l'ARN.

Dans un second temps le calcul du temps de contact entre différentes paires de résidus clés permettra une analyse plus fine du processus. Pour cela nous avons calculé pour chaque paire de résidus clés étudiés le nombre de conformations dans lesquelles ces résidus sont en contact. Comme ces résultats ont été obtenus à partir d'une dynamique de 10ns il suffit de diviser ce nombre de conformation par le nombre total de conformations et de multiplier le résultat par la durée totale de la dynamique.

$T = (N_c/N_t) * 10$ avec T en ns et Nc le nombre de conformation

Le programme

Notre programme est divisé en plusieurs parties :

- un fichier qui regroupe les différentes fonctions utilisées dans le programme
- un fichier main.py qui fait appel à ces fonctions et gère les sorties

Tout d'abord à l'aide d'un parseur qui prend en entrée les fichiers .pdb nous récupérons les informations de la conformation de référence et des 500 configurations issues de la dynamique moléculaire. Ce parseur renvoi ses informations sous la forme d'un dictionnaire.

Ensuite pour chaque résidu de chacune des conformations le Centre de Masse est calculé à partir des informations des deux dictionnaires et est utilisé pour calculer le RMSD. Nous avons une fonction spécifique pour le RMSD global et une fonction spécifique pour le RMSD concernant un domain spécifique. Les résultats seront dans des fichiers RMSD_<Domaine>.png

Afin de calculer les fréquences de contact entre les domaines protéiques et l'ARN tout en optimisant le nombre d'opération à effectuer par l'algorithme nous présélectionnons des résidus dans la configuration de référence, nous ne gardons que ceux qui sont à moins de 14 Å de l'ARN. En effet en règle générale les résidus peuvent marier de maximum 3/4 Å et comme le suil d'interaction est fixé à 10 Å cela évite de calculer la distance pour tout les résidus des 500 configurations. Le résultat de fréquence est stocké sous la forme d'une bibliothèque.

Résultats & Interprétation

Analyse des changements conformationnels globaux

Voici le graphe du RMSD global calculé à partir des centres de masse des résidus, au cours du temps(**Figure2**).

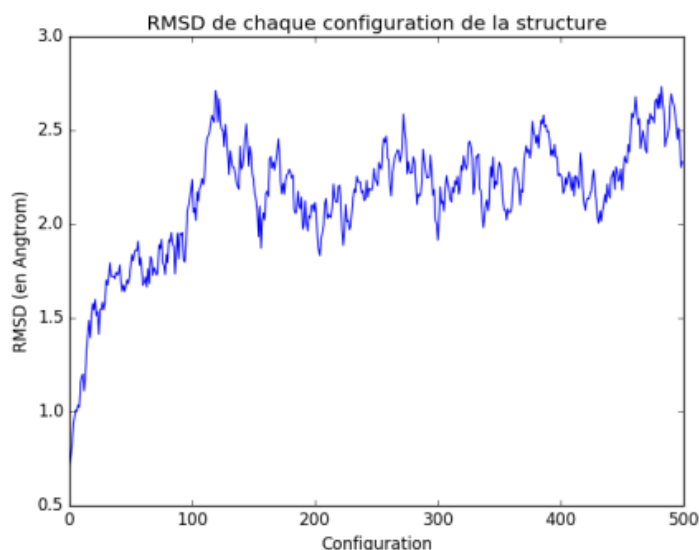


Figure 2

RMSD: Distance entre les atomes de la configuration étudiée avec celle de référence

Configuration: Les 500 conformations étudiées dans l'ordre chronologique

Nous pouvons observer que le RMSD augmente très rapidement de 0 à 1,6 Å lors des cinquante premières conformations avant de rester relativement stable jusqu'à la 110ème configurations avec la présence d'un pic à 2,7 Å à la configuration n°130 puis retourne à un état stationnaire entre 2 et 2,7 Å durant les configurations suivantes.

Nous pouvons supposer qu'il y a une phase de fixation entre les conformations 1 et 50 puis une activité catalytique à partir de la 110ème configuration. Après la 180ème configuration le complexe semble avoir plus ou moins atteint son état attendu lors de la liaison avec le sRNA.

Le RMSD pour chacun des domaines protéiques A1,A2,A3 et A4, correspondant respectivement à aCBF5, PUA, aNOP10 et L7Ae (Muller, 2007) a également été calculé (**Figure3**) :

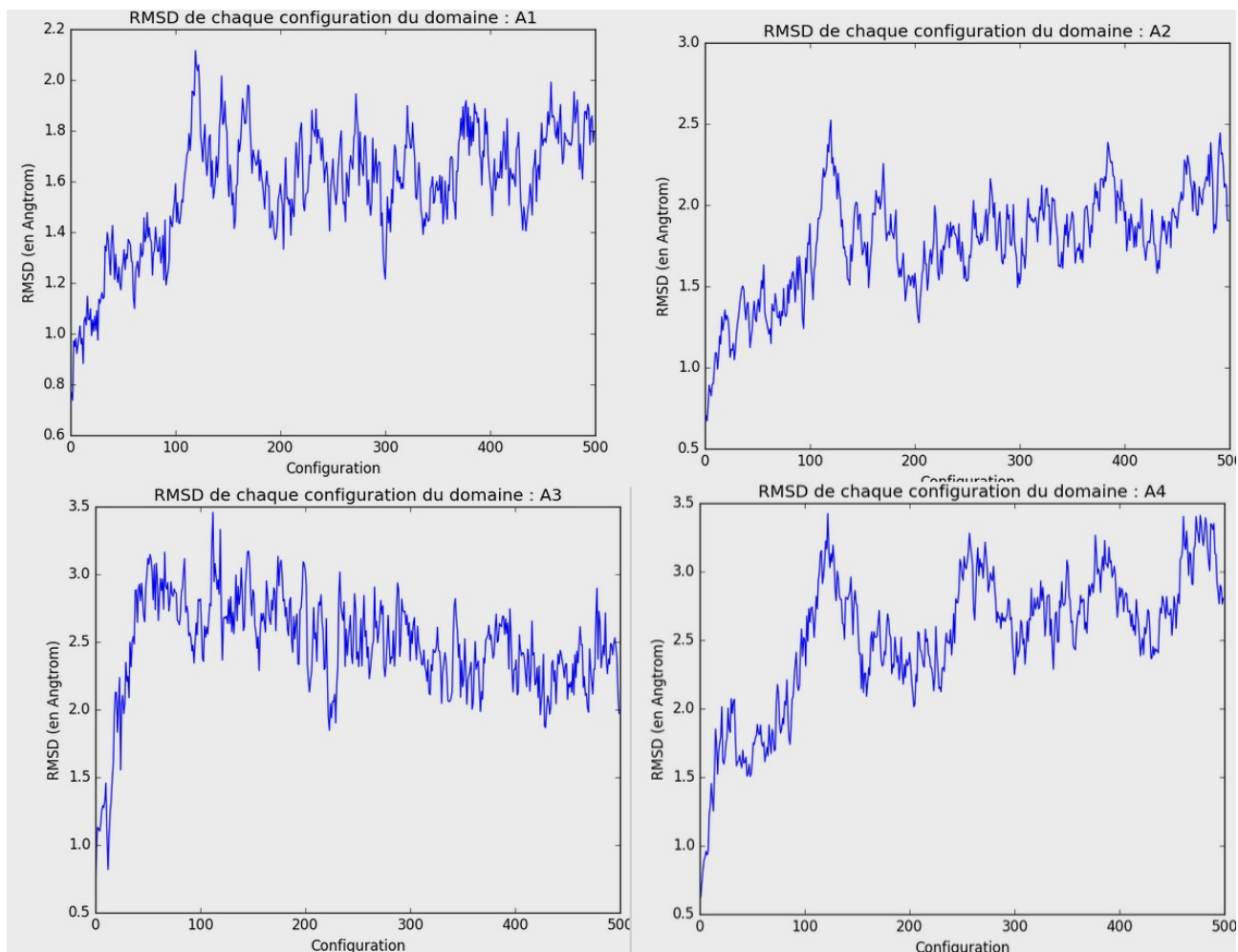


Figure3 : RMSD des différents domaines

Trois tendances se détachent nettement :

- Le RMSD des domaines A1 (aCBF5) et A2 (PUA) augmentent nettement moins vite que le RMSD global. Celui du A1 atteint un état stationnaire après environ 130 configuration (présence d'un pic) et reste entre 1,5 et 2 Å. Celui du A2 augmente relativement rapidement jusqu'à sa 130ème configuration (avec un pic) puis reste relativement stable avec néanmoins une légère augmentation de 1,5 à 2,1 Å entre les configurations 200 et 500.

- Le RMSD du domaine A3 (aNOP10) augmente très rapidement jusqu'à 2,9 Å en à peine 80 configurations, puis reste relativement stable jusqu'à la présence d'un pic à la configuration n°130. Ensuite le RMSD va diminuer de façon constante jusqu'à 2,5 Å à la configuration n°500.

- Le RMSD du domaine A4 (L7Ae) augmente rapidement (1,8 Å) jusqu'à la configuration 50 avant de rester relativement stable jusqu'à la configuration 110 puis augmente très rapidement jusqu'à 3,3 Å à la configuration n°130 avant de redescendre à 2,2 Å jusqu'à la conformation 170. Celui-ci augmente ensuite progressivement jusqu'à 2,7 Å pour la conformation n°500 avec des

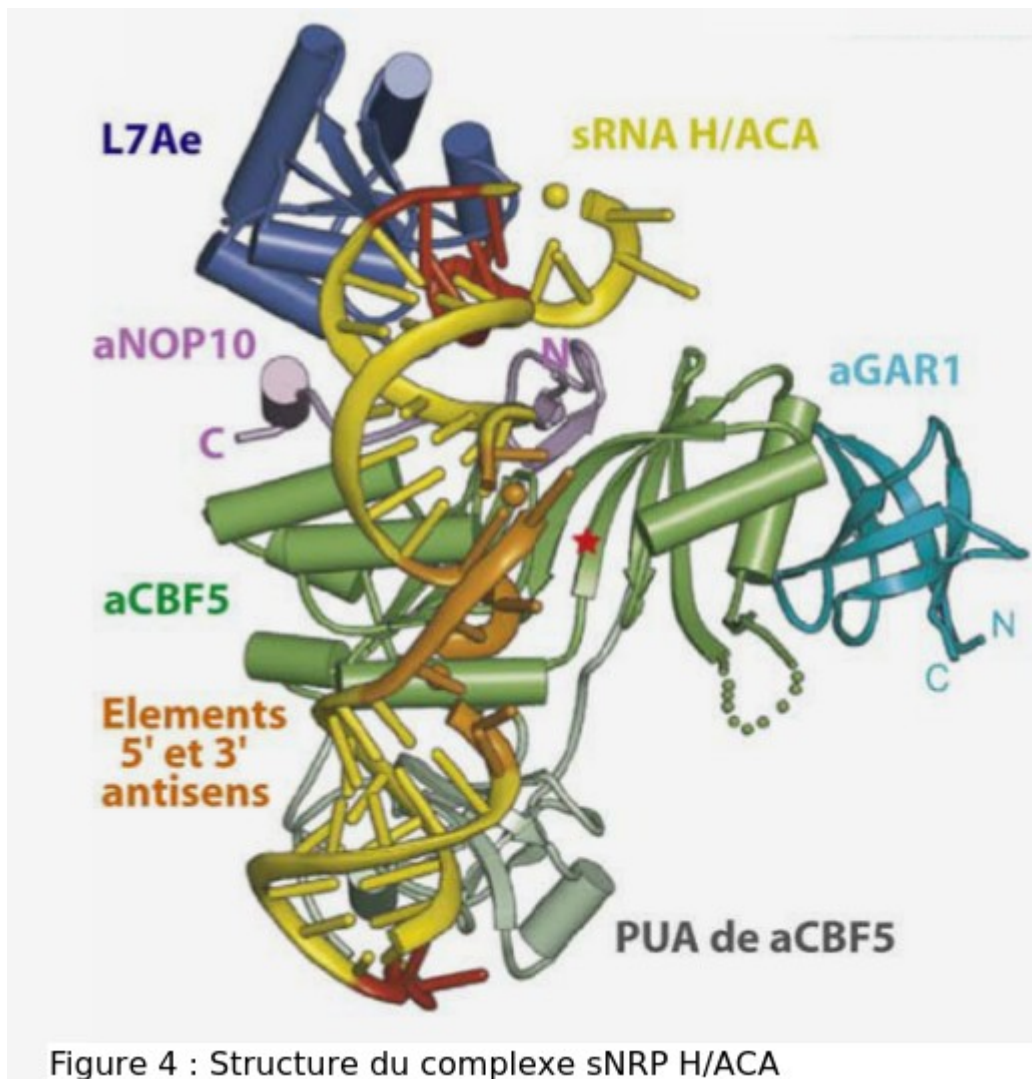
pics secondaires vers les configurations n°260, n°380 et n°450.

Nous pouvons donc supposer que puisque les domaines A1 et A2 varient le moins ils sont les moins flexibles et subissent moins de changements conformationnels que les domaines A3 et A4.

Si nous prenons comme hypothèse que les domaines les plus flexibles sont ceux qui interagissent le plus avec le sRNA alors cela signifie que le domaine A3 rentre rapidement en contact avec celui-ci et le reste tout le long des 500 conformations. Nous pouvons donc prédire que celui-ci est utile à la fixation. Le domaine A4 attend la configuration numéro 110 avant de vraiment atteindre un changement conformationnel important. Nous pouvons donc supposer que le domaine A3 interagit tôt avec le sRNA puis le domaine A4 intervient à son tour dans la fixation.

Quels que soient les domaines on observe un pic du RMSD vers la configuration n°130, ce qui peut signifier le début de l'activité catalytique suite à la stabilité du complexe.

Des études ont montrés que aNOP10 (A3) est pris en « sandwich » entre L7Ae (A4) et aCBE5(A1), Muller, 2007 (**Figure4**) il est donc normal que suite au changement conformationnel de celui-ci A1 et A4 subissent également un changement conformationnel. De plus comme A2 (PUA) est un domaine de A1 il est compréhensible que leur RMSD varient de façon semblable. De plus comme il a été prouvé que L7Ae (Watkins *et al*, 2000) ne se lie qu'au motif k-turn du sRNA, il est normal qu'il intervienne très rapidement après que aNOP10 (A3) ai approché celui-ci. Une fois les fixations faites il semble logique qu'un pic du RMSD ai lieu et indique le début de la catalyse cela doit correspondre à la poche de pseudouridification qui se positionne au site catalytique de aCBF5 (Muller, 2007).



Analyse des changements conformationnels locaux

Afin d'étudier les résidus ayant une grande fréquence de contact avec l'interface du fragment d'ARN nous avons répertorié dans un dictionnaire tout les résidus ayant eu un contact avec leur interface associée.

L'hypothèse qui en découle est que les résidus ayant une forte fréquence sont ceux participant à la liaison de l'ARN et ils appartiennent à aCBF5, aNOP10 et L7Ae.

Cette hypothèse est appuyé par la thèse de Muller, 2007, qui explique que la tige-boucle de l'ARN est coincé entre L7Ae, aCBF5 et aNOP10 et également par le domaine PUA de aCBF5.

Ensuite concernant les temps de contact des résidus, il aurait été pertinent de prendre les paires résidus-ARN parmi les résidus s'étant démarqué de par leur forte

interaction avec l'ARN (**Figure1**).

En étudiant les différences de temps de contact entre les paires de résidus chez le mutant et le sauvage ainsi que les résultats associés nous aurions pu en déduire l'utilité de certaines de ces liaisons lors de l'interaction. Par exemple une liaison forte chez un mutant et nulle chez le sauvage entraînant une activité catalytique faible serait significatif d'une liaison d'importance.

Conclusion

Au vu des résultats il apparaît que l'analyse des changements conformationnels globaux permet de montrer l'importance de chaque domaine du complexe sNRP H/ACA. En effet, nous pouvons supposer que le domaine aCBF5 (A1), correspondant à CBF5 chez les eucaryotes porte l'activité catalytique de l'isomérisation de l'uridine en pseudouridine, le domaine PUA (A2) de aCBF5 partage ce même rôle. Le domaine aNOP10, NOP10 chez les eucaryotes, serait le premier à subir un changement conformationnel du fait de sa position centrale en aCBF5 et L7Ae participant ainsi la fixation du complexe protéique à l'ARN. Par la suite le domaine L7Ae fixe complètement le complexe en se liant au motif k-turn. On assiste alors à une courbure de 50° de la tige-boucle de l'ARN (Muller, 2007) qui se retrouve coincé entre ces trois domaines protéiques. La poche de pseudouridification se positionne alors au site catalytique de aCBF5 et la catalyse à lieu.

L'image de Pymol a montré que globalement certains sites spécifiques avaient une grande fréquence de liaison à l'ARN et il est probable que ce soient des sites de liaison et à activité catalytique.

Il pourrait être judicieux de réaliser un temps de contact des paires de résidus clés afin de pouvoir statuer sur l'utilité de ces paires. De plus il pourrait être pertinent d'effectuer des mutations spécifiques de ces sites de liaison et de catalyse afin d'obtenir des résultats plus poussés sur le rôle de chacun.

Bibliographie

Watkins NJ, et *al.* (2000) A common core RNP structure shared between the small nucleolar box C/D RNPs and the spliceosomal U4 snRNP.

Muller S. (2007) Identification of determinants in the protein partners aCBF5 and aNOP10 necessary for the tRNA:Psi55-synthase and RNA-guided RNA:Psi-synthase activities. *Nucleic Acids Res* 35(16):5610-24