UE - Langage de Script (Python)
M1 BIBS 2017

Etude du Complexe sRNP H/ACA

Auteurs :

ROBIEUX Arthur
ROZIERE Julien

- Université Paris Sud -

Introduction

L'étude menée dans le cadre du projet porte sur le complexe sNRP H/ACA. Ce complexe est connu pour son rôle dans l'isomérisation de l'uridine en pseudo-uridine des brins ARN. Ces pseudo-uridine sont elles impliquées dans la stabilisation des hélices alpha des ARN, autrement dit dans la structure secondaire. Une étude de dynamique a été préalablement réalisée sur la protéine d'un individu sauvage et des individus mutants, altérant l'activité du complexe, afin de déterminer les potentiels sites de flexibilité de ce dernier. Des sites d'interaction chez le mutant avec l'ARN ont été révélé et seront utiles pour la deuxième partie de notre travail. Nous allons dans un premier temps nous attarder sur les changements conformationnels globaux de la protéine chez le sauvage en analysant la RMSD globale ainsi que celle obtenue pour chaque domaine afin de peut être mettre en évidence des domaines plus flexibles. Pour cela, nous nous sommes basés sur les conformations du complexe obtenues grâce à l'étude de la dynamique menée avant notre projet. Dans un second temps, nous avons étudié les changements conformationnels locaux, toujours chez le sauvage, au niveau des sites montrés comme étant fortement impliqués dans l'interaction avec l'ARN chez le mutant. Nous avons pour cela analysé le temps de contact entre chacun des résidus que nous avons choisi arbitrairement, à savoir l'Arg38, la Lys46, l'Arg34 et l'Ala100. Ainsi le temps de contact de chacun de ces résidus avec l'ARN pour les 500 conformations a été calculé.

Notre rapport se divise donc en 2 grandes parties, la première consacrée à l'aspect informatique du projet et donc à l'explication de la conception du programme nécessaire pour obtenir les résultats pour l'interprétation biologique et enfin la deuxième s'attarde sur l'aspect biologique de notre travail et exposera donc les conclusions sur les différentes problématiques posées par le complexe sNRP H/ACA, exposées dans le paragraphe précédent.

I / Création du programme

1) Parser

Pour débuter cette étude, il a été nécessaire de développer des outils permettant de manipuler informatiquement les protéines appartenant au complexe. Nous disposions de deux fichiers à parser : le complexe RNA H/ACA de référence, puis 500 conformations de ce complexe lorsqu'il se trouve en solution.

Nous avons mis en place plusieurs parseurs afin de s'adapter aux différentes situations que nous avons pu rencontrer pendant le projet.

Premièrement, le parseur du fichier de référence. Celui ci prend en entrée un fichier ".pdb" et retourne un dictionnaire contenant les informations triées de façon hiérarchique afin d'être lisible par Python. La hiérarchie du dictionnaire retourné par ce premier parseur a été un élément qui nous a conduit à le modifier par la suite. La sortie de ce parser est représentée schématiquement dans la Figure 1.a.

Ensuite, nous avons trouvé judicieux de créer une fonction reconnaissant une unique conformation dans le fichier pab21_500frames.pdb afin de pouvoir les parser une à une. Nous l'appellerons le parseur multiple. Pour cela, à chaque "model", nous récupérons

le contenu du pdb associé à cette conformation dans une chaine de caractère. Puis nous envoyons cette chaîne à un second parseur, identique au premier, mais prenant en entrée une chaîne de caractères (“string”).

Enfin, lors de la dernière étape du projet, nous avons rencontré un problème lié au parseur, le domaine de chaque résidu se trouvait trop bas dans la hiérarchie et était trop difficilement accessible. Pour palier à cela, nous avons modifié le parseur multiple en remontant les domaines dans la hiérarchie. Ceux ci étaient accessibles au niveau de l’atome. Pour remédier à ce problème d’accessibilité, nous les avons placé avant les résidus dans la hiérarchie, ce qui semble plus logique car un domaine contient des résidus. Cette hiérarchie (Figure 1.b) est bien plus optimale pour les calculs liés à l’interface (Partie 3).

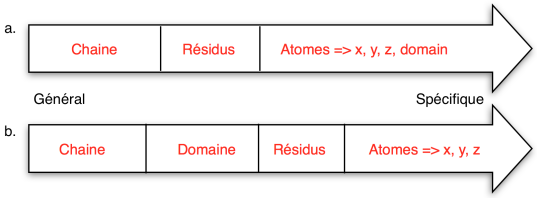

Figure 1 : Hiérarchie des sorties des parseurs
a : dictionnaire retourné par les 2 premiers parseurs
b : dictionnaire retourné par le 3ême parseur

2) RMSD

Le RMSD est une mesure rendant compte de la déviation structurale entre deux structures alignées. Cette mesure est obtenue à l’aide de la formule suivante :

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2}$$

avec : - N le nombre de paires d’atomes
- δi la distance spatiale séparant les deux atomes i de chacune des structures

a)Global

Pour calculer le RMSD globale sur l’ensemble des conformations, nous avons calculé le RMSD entre la structure de référence du complexe fourni dans le fichier “pab21_ref.pdb” et chacune des 500 conformations disponibles. Nous avons ainsi obtenu 500 résultats de RMSD, un pour chaque conformation, ce qui nous a permit d’obtenir le graphique en Figure 2.

Ces calculs de RMSD se font dans une fonction prenant en attribut deux protéines parsées. Celle ci va accéder à chaque résidu commun à ces deux protéines, puis pour chacun des atomes de ces résidus, on calcul la distance spatiale les séparant en utilisant leurs coordonnées. Enfin, on applique la formule donnée ci-dessus pour obtenir la déviation structurale entre la structure de référence et cette même structure dans une autre conformation.

Afin de s'y retrouver et de contenir tous ces résultats de RMSD de façon organisée, nous avons créé un dictionnaire général, contenant pour chaque forme du complexe, son RMSD globale et spécifique à chaque domaine par rapport à la structure de référence.

b) Local à chaque domaine

Pour le calcul du RMSD local, nous avons procédé de la même manière que pour le calcul global, en comparant les 500 conformations de la dynamique avec la conformation de référence, mais en séparant le calcul pour chacun des domaines afin de mettre, ou non, en évidence une flexibilité accrue pour certains d'entre eux. Les graphiques obtenus sont également analysés dans la seconde partie.

Ce calcul se fait globalement de la même façon que le RMSD global. Cependant, il a été nécessaire de créer une seconde fonction afin de pouvoir différencier chaque domaine. Le principe est exactement le même, à l'exception du calcul de la distance spatiale entre les atomes. En effet, ces atomes ne sont comptabilisés et utilisés dans les calculs uniquement s'ils appartiennent au domaine choisi lors de l'appel de la fonction. Cette vérification de l'appartenance de l'atome au domaine suffit au calcul de RMSD spécifique.

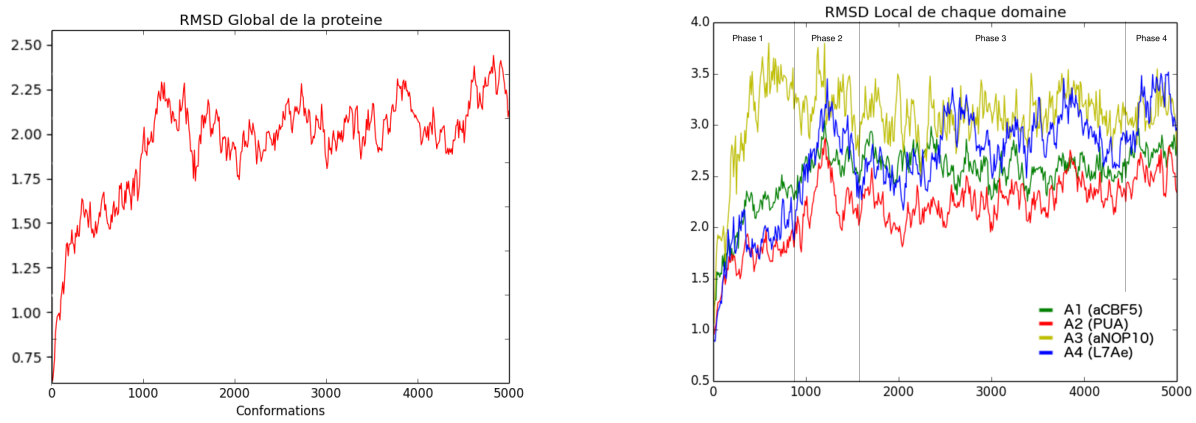

Figure 2 : RMSD Globale et spécifique à chaque domaine

3) Interface

a)Appartenance des résidus à l'interface

Dans cette deuxième partie d'étude, nous avons tenté de mettre en évidence la fréquence d'appartenance de chacun des résidus des domaines protéiques à l'interface avec le brin d'ARN. Pour réaliser ce calcul de fréquence, il a été nécessaire de développer une fonction calculant la distance entre chaque résidus des domaines protéiques A1, A2, A3 et A4 avec chaque résidus du brin d'ARN. Si la distance obtenue est inférieure à un seuil établi (ici 9A), on considère que le résidu en question fait partie de l'interface à la conformation considéré. Ces calculs ont été réalisés pour les 500 conformations et le nombre d'appartenance à l'interface obtenu pour chaque résidu a été divisé par 500 afin de donner la fréquence d'appartenance à l'interface de chacun des résidus recherchée.

Dans un premier temps, nous avons créé un dictionnaire, prenant en clés les noms des résidus appartenant à l'interface et en valeur le nombre d'occurrences de ces derniers dans l'interface. Ensuite nous avons créé une fonction "distanceResidus" prenant en argument deux résidus (dictionnaires comportant la liste des atomes appartenant au résidu considéré ainsi que les coordonnées pour chacun de ces atomes) et renvoyant la distance minimale entre chaque paire d'atomes possible entre les deux résidus. Le résidu un envoyé à cette fonction correspond à un résidu d'un des complexe protéique et le second à un résidu de l'ARN car les comparaisons doivent être réalisées entre les résidus de la protéine et ceux de la ARN pour déterminer la fréquence d'appartenance à l'interface.

Enfin si cette distance, caractérisant la distance entre les résidus, est inférieure au seuil, le nom du résidu appartenant à un domaine protéique est mis comme clé du dictionnaire si celui ci ne s'y trouvait pas déjà et la valeur associée devient 1, si il s'y trouvait déjà, on ajoute ajoute 1 à sa valeur.

Ces calculs sont réalisés 500 fois, dans le but de vérifier si les résidus des domaines protéiques appartiennent à l'interface dans chacune des conformations. Une fois cela terminé, on divise par 500 le nombre de fois où les résidus ont été dans l'interface, afin d'obtenir leurs fréquences d'appartenance à l'interface.

b) Temps de contact

Enfin, nous nous sommes penché plus précisément sur les sites révélés comme étant important dans l'interaction avec l'ARN chez la protéine du mutant. Pour les sites choisis nous avons calculé le temps de contact avec le site de l'ARN associé pour les 500 conformations. Ceci dans le but de révéler si ces sites, à l'instar du complexe chez le mutant, sont importants pour le complexe du sauvage et ainsi apporté de nouvelles conclusions biologiques (Partie II). Pour cela, il a été nécessaire de calculer la distance de chacune des paires et de vérifier si la valeur est inférieure au seuil défini précédemment. Ceci pour les 500 conformations. Le temps de contact correspond donc au nombre de conformations, sur les 500, pour lesquelles les sites sont à une distance inférieure à notre seuil.

Pour cette partie, nous nous avons réutilisé la fonction "distanceResidus". Cette dernière prenait comme arguments les paires de résidus d'intérêt expliqués dans l'introduction de ce rapport à savoir :

Arg38 / Résidu 26 (ARN)
Lys46 / Résidu 25 (ARN)
Arg34 / Résidu 33 (ARN)
Ala100 / Résidu 31 (ARN)

Pour ces 4 paires, la distance a été calculé grâce à notre fonction pour les 500 conformations. Si la distance à une conformation donnée est inférieure à notre seuil, la paire de résidus est considérée en contact et dans un dictionnaire créé à cet effet, on stocke en clés les noms des résidus constituants la paire et en valeur le nombre de fois où ces derniers sont en contact, ceci nous donne alors le temps de contact pour nos 4 paires de résidus.

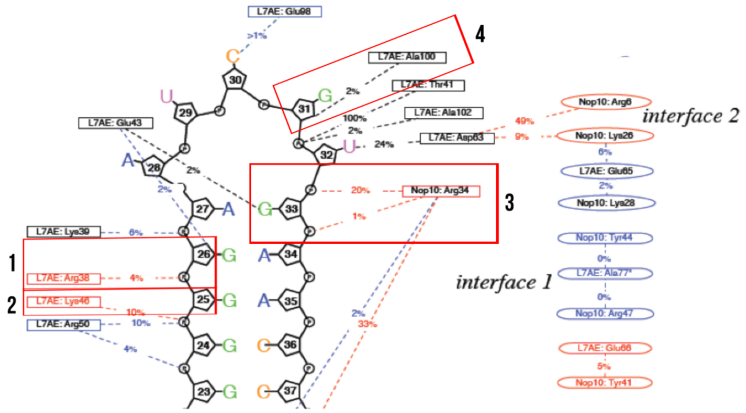

Figure 3 : Contacts chez le mutant entre résidus clés pour l'interaction, les rectangles correspondent aux paires de résidus traités dans notre étude.

II / Analyse des résultats

1) Changements conformationnels globaux

Pour l'étude des changements conformationnels globaux, il a été réalisé un calcul de RMSD sur la protéine globale entre la structure de base et les 500 conformations. (Figure 2). Le graphique montre clairement une forte augmentation de la RMSD et ce dès les premières conformations. En effet entre 0 et 50 conformations environ, l'augmentation de la RMSD globale est très forte (valeur multipliée par 2), puis une augmentation de la RMSD plus lente jusqu'à environ la 100ème conformation. Ceci nous amène à penser qu'il s'agit de la phase de fixation des différents domaines sur l'ARN, ceci explique donc l'augmentation forte du RMSD. Puis l'augmentation reste « constante », nous faisant penser à la phase d'activité du complexe dans laquelle les modifications conformationelles sont moindres. De manière générale, la RMSD globale décrit bien une activité catalytique avec dans un premier temps une étape de fixation et enfin une étape de réaction spécifique à ce complexe.

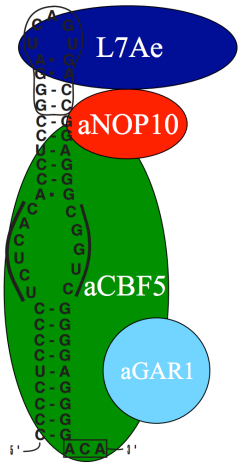

Figure 4 :
Représentation schématique du complexe

Intéressons nous à la RMSD de chaque domaine. On constate dans la première phase de la dynamique (100 premières conformations) que la RMSD du domaine A3 (aNOP10) est bien supérieure à celles des autres domaines. Ceci laisse supposer que celui ci serait potentiellement le domaine servant à la liaison entre les autres domaines composant le complexe comme le suggère la figure 4 montrant effectivement que aNOP10 relie A4 (L7Ae) et A1 (aCBF5). Ensuite il est possible de remarquer que la RMSD du domaine L7Ae augmente fortement à partir de la 100ème conformation supposant qu'il s'agit de la phase de reconnaissance de l'ARN grâce à ce domaine. Ceci est également appuyé par la figure 3 nous montrant que ce dernier est en contact avec la partie K-turn, région pouvant permettre une reconnaissance du brin d'ARN. Enfin, une phase de stabilisation du RMSD est observable appuyant l'hypothèse selon laquelle les conformations associées décrivent l'activité d'isomérisation de l'uridine.

De manière générale, on peut remarquer 4 phases. La première, de la 1ere à la 100ème conformation, indiquerait une formation du complexe avec la création d'une liaison entre les domaines grâce à aNOP10 rattachant L7Ae et aCBF5. Ceci se confirme également avec le RMSD du domaine aCBF5, potentiellement responsable de l'activité catalytique, qui n'augmente fortement qu'au début de la dynamique, là où le domaine se fixe avant qu'il soit actif. La deuxième phase correspondrait à la phase de reconnaissance de l'ARN par L7Ae. La troisième phase, quant à elle, désignerait la phase d'activité du complexe d'où une pseudo-stabilité du RMSD de chacun des domaines,particulièrement aCBF5, suggérant d'autant plus que ce domaine est responsable de la réaction. Enfin la dernière tendance correspondrait eu détachement du domaine de reconnaissance L7Ae indiquant la potentielle fin de l'activité. En effet on constate que la tendance du RMSD pour ce domaine dépasse dorénavant celle de aNOP10. Quant au domaine A2 (PUA), ce dernier est associé à aCBF5 expliquant l'évolution de son RMSD similaire.

2) Changements conformationnels locaux

a) Résidus dans l'interface

Afin d'étudier les changements conformationnels locaux, nous avons commencé par étudier l'interface du fragment d'ARN. Le but étant de déterminer les résidus ayant tendance à être fréquemment présent dans celle ci.

La fonction mise en place pour calculer ces fréquences est fonctionnelle sur de petit échantillon (si l'on prend que quelque conformations). Cependant, en utilisant les 500 conformations comme souhaité, le programme devient extrêmement long, nous n'avons donc pas pu obtenir de résultats pour la dynamique complète du complexe.

Notre théorie quant aux résultats seraient que les résidus à forte fréquence d'appartenance à l'interface avec l'ARN sont ceux des domaines L7Ae, aNOP10 et aCBF5 pour les domaines protéiques et pour les résidus de l'ARN, seuls ceux d'un même brin ont une fréquence haute car il semblerait que la fixation des domaines ne se fasse que sur l'un d'entre eux (voir Figure 5).

Figure 5 : Représentation du complexe sous PyMol

b) Analyse des résidus clés

Lors de l'étude dynamique menée sur le mutant, plusieurs sites ont été montré comme étant important dans l'interaction avec l'ARN (cf Figure 3). Il a donc été intéressant de tester ces derniers dans l'interface chez le sauvage afin de mettre en évidence, ou non, des différences. Notre choix d'analyse s'est porté sur 4 résidus dont 2 ont une interaction avec l'ARN forte chez le mutant (l'Arg38 et la Lys46 du domaine L7AE) afin de voir si cette force d'interaction est respectée chez le sauvage. Nous avons également choisi un résidu du domaine Nop10 (Arg34) afin de pouvoir tester un résidu d'un autre domaine avec des interactions primaires avec l'ARN. Enfin le résidu Ala100 du domaine L7AE, comme les deux premiers, afin de voir si ce résidu en interaction plus faible chez le mutant que les deux autres résidus du même domaine, se révélera différent chez notre sauvage. Sur chacun de ces résidus, nous avons calculé le temps de contact avec le résidu de l'ARN associé sur la Figure 6.

On observe que la liaison forte qu'il existait entre la Lys46 et l'ARN chez le mutant est respectée chez le sauvage. Son temps de contact est au maximum, c'est à dire que ce résidu a été en contact avec l'ARN durant l'intégralité de la dynamique. Cependant, la seconde liaison forte que nous avons sélectionnée (Arg38 - ARN), n'est plus du tout présente. En effet, le temps de contact calculé est de 0. Nous pouvons donc penser que celui ci n'était pas présent chez le sauvage ou encore que son interaction avec l'ARN pourrait causer l'inactivation de la réaction.

De plus, on remarque que l'Ala100 du domaine L7AE a désormais un temps de contact de 500. Ceci révèle que cette liaison a très probablement son importance dans le complexe et l'activité de ce dernier et ce car avec la protéine du mutant, l'interaction de ce résidu avec l'ARN était considéré comme faible (voir Figure 3).

Enfin, l'Arg34 du domaine Nop10 est en liaison quasi constamment avec l'interface de l'ARN (temps de contact = 498). Ce qui était aussi le cas chez notre mutant, ce résidu n'a donc pas subi de modification au niveau de son interaction avec l'ARN.

```
31-100   500
26-38    0
25-46    500
33-34    498
```

Figure 6 - Temps de contacts des 4 paires

Ceci nous amène à penser que les deux liaisons ayant des propriétés différentes entre le complexe du sauvage et celui du mutant sont impliquées dans l'inactivation du complexe. Ceci n'est néanmoins pas suffisant pour le conclure et d'autres études seraient intéressantes pour valider ou infirmer cette hypothèse. Par exemple, il serait intéressant de réaliser une mutagenèse dirigée sur ces sites et ainsi regarder l'activité des protéines obtenues. Enfin, il serait d'autant plus intéressant de regarder les autres liaisons avec l'ARN montrées dans la Figure 3 afin de révéler d'autres anomalies comme les deux déjà observées.

Conclusion

Au travers de cette étude, il a été possible d'inférer les différentes fonctions des domaines étudiés. En ce qui concerne le domaine aCBF5, il serait chargé de la réaction d'isomérisation de l'uridine en pseudo-uridine. Le domaine L7AE, quant à lui, permet la reconnaissance de la région K-turn de l'ARN. Enfin, aNOP10 serait le domaine joignant les deux domaines évoqués précédemment. Ces déductions ont été possibles grâce aux calculs de RMSD nous permettant ainsi de comparer les déviations structurales. L'imagerie PyMol a également compensée le manque de résultats pour les calculs de fréquences d'appartenance à l'interface et permettent, même difficilement, de voir que l'interaction ne se fait que sur un brin de l'ARN. Nous avons également mis en évidence des différences au sein des liaisons domaine-ARN entre le complexe mutant et sauvage. Ces deux différences sont intéressantes et suggèrent que les deux résidus du domaine sont inversement impliquées dans l'interface avec l'ARN et expliqueraient en partie l'inactivité du complexe mutant.

Afin d'aller plus loin, il serait judicieux de réaliser des mutagenèses dirigées sur sites révélés comme étant fortement impliqués dans l'interface avec l'ARN. Ceci pourrait confirmer leurs importances ou non dans l'activité du complexe. De plus, nous pourrions réaliser le même protocole pour calculer les temps de contact chez le sauvage, sur les autres résidus très fortement impliqués dans l'interaction avec l'ARN, présents dans la Figure 3 et non traités dans notre analyse.