

# Compte rendu Projet Barstar

---

Team : **Les Mugiwaras**

Auteurs : ARBES Hugo et HENRIQUES Adrien  
M1-BIBS

## Introduction

La Barstar est une protéine synthétisée par une bactérie (*Bacillus amyloliquefaciens*) afin d'inhiber l'activité de la ribonucléase de la Barnase en formant un complexe avec celle-ci. En formant ce complexe, la Barstar couvre le site actif de la Barnase empêchant ainsi la Barnase de dégrader l'ARN dans le cytoplasme de la cellule.

Dans cette étude, nous allons nous intéresser à la protéine Barstar et notamment à ses changements conformationnels lors de l'interaction Barstar-Barnase. En effet, la majorité des algorithmes de docking considèrent que les deux protéines testées sont rigides et ne subissent donc pas de changement conformationnels. Le but est donc de créer un algorithme qui va nous permettre d'étudier la stabilité de la structure de la protéine Barstar au cours du temps ainsi que ses changements conformationnels. Pour cela, nous allons donc dans un premier temps s'intéresser aux changements conformationnels globaux, puis nous nous intéresserons aux changements conformationnels locaux.

## Manuel d'utilisation

Dans cette partie, nous allons aborder l'utilisation du programme et principalement des fichiers d'entrée ainsi que des pré-requis nécessaires pour pouvoir exécuter le code.

### Pré-requis

Le programme est composé de plusieurs fonctions pour calculer, afficher, récupérer des données. Ce programme est créé en langage Python (Python 2.7), pour que celui-ci puisse être compris par l'ordinateur, il est nécessaire d'avoir installé Python. De plus, certaines de ses fonctions utilisent des fonctions graphiques telles que Matplotlib. Pour pouvoir les utiliser il est donc nécessaire d'avoir installé Matplotlib. Pour cela, ouvrez un terminal et tapez :

- Unix : `$ sudo apt-get install python-matplotlib`
- Fedora : `$ sudo yum install python-matplotlib`
- Windows : `python -m pip install -U steuptools`  
`python -m pip install matplotlib`

Ainsi pour exécuter le code, il nous suffira de taper la commande :

```
$ Python BarStar.py
```

qui exécutera le code en appelant les différentes fonctions.

## Fichiers d'entrée

Les fichiers qui vous seront demandés lors de l'exécution du programme doivent absolument suivre une structure spéciale. En effet, ceux-ci doivent être au format PDB suivant :

Colonne	Data Type	Field	Definition
1 - 6	Strings	"ATOM" ou "MODEL"	Type de données de la ligne
7 - 11	Integers	Atome	Identifiant de l'atome
13-16	Strings	Atome	Type d'atome
17 - 20	Strings	Type résidu	Type du résidu
21 - 26	Integers	Résidu	N° du résidu
32 - 38	Integers	Coordonnées	Position X
40 - 46	Integers	Coordonnées	Position Y
48 - 54	Integers	Coordonnées	Position Z
5 - 14	Integers	Model	N° du modèle

Si les fichiers d'entrée ne suivent pas ce format, le processus de parsing ne se fera pas correctement et le programme risque d'avoir des erreurs ou de ne pas donner les bonnes sorties.

De plus il est nécessaire de suivre les indications qui s'affichent dans le terminal et notamment lorsque l'on écrit les fichiers, il est nécessaire de respecter l'ordre d'entrée des fichiers. Tout d'abord on entre le fichier qui contiendra la Barstar ayant la conformation référent, puis dans un deuxième temps, le fichier contenant les différentes conformations de la Barstar dans le temps.

## Les tests

Le programme a été testé sous plusieurs angles. En effet, nous l'avons testé en entrant mal un des 2 fichiers pour tester si la reconnaissance d'erreur est détectée. De plus, nous avons testé les parties plus calculatoires (RMSD, Rayon de giration et calcul de distance) sur des jeux de données restreint pour pouvoir vérifier les résultats et ainsi la robustesse des calculs. Et enfin nous avons testé ce programme avec les données fournies afin de voir si celui-ci fonctionne.

Pour avoir un code d'une robustesse accrue, nous avons fait le choix de garder la possibilité d'avoir des chaînes différentes au sein de la protéine observée, dans le fichier pdb d'entrée. Cependant, ceci n'est pas nécessaire dans le cadre de l'analyse du fichier que nous avons utilisé pour notre projet.

## Calcul du Rayon de Giration

Le rayon de giration correspond à la distance la plus élevée entre le centre de masse de la protéine et celui des acides aminés qui la composent. Ainsi, il rend compte du dépliement et de la compaction de la protéine car plus celle-ci sera dépliée, plus la distance entre son centre de masse et le résidu le plus éloigné sera importante.

Nous nous y intéressons donc pour tenter d'observer une corrélation entre la RMSD (abordé plus loin) et le rayon de giration des différentes conformations étudiées pour la protéine d'intérêt.

Pour pouvoir effectuer le calcul d'obtention du rayon de giration, il faut tout d'abord calculer le centre de masse de la protéine d'intérêt. Celui-ci est obtenu par la somme des centres de masse de chaque résidu composant la protéine divisée par le nombre de résidus.

Pour obtenir ces centres de masse de chaque acide aminé, on fait la somme des coordonnées de chacun de ses atomes multipliée par sa masse atomique, et l'on divise le tout par la masse cumulée des atomes.

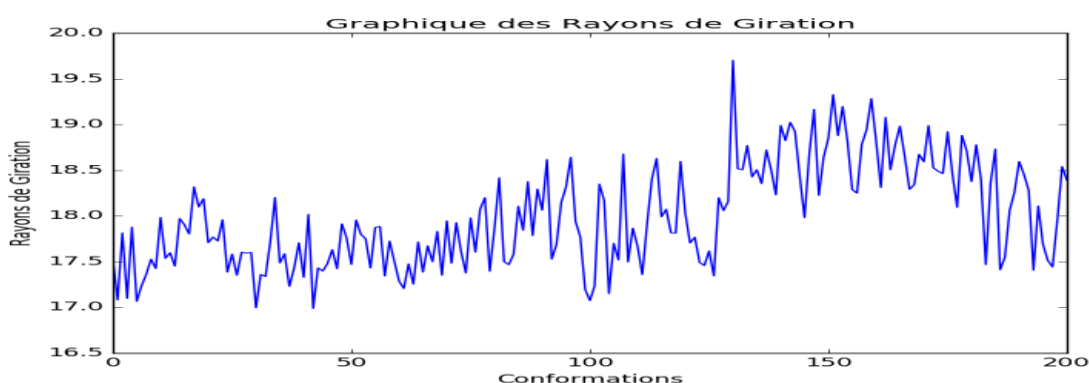
$$\text{Rayon de Giration} = \frac{\sum_{i=1}^N \text{Centre de Masse}[i]}{N}$$

Pour N le nombre d'acides aminés dans la protéine d'intérêt et i un résidu.

$$\text{Centre de masse résidu} = \frac{\sum_{i=1}^n \text{Coordonnées}[i] * \text{Masse atomique}[i]}{\sum_{i=1}^n \text{Masse atomique}[i]}$$

Pour n le nombre d'atomes dans le résidu observé et i un atome.

Pour notre étude, ces calculs nous permettent ainsi d'obtenir les résultats suivants pour les différentes conformations observées de la protéine Barstar :



*Figure1 : Angles de Giration des différentes conformations de la Barstar*

Cette analyse globale nous permet ainsi d'observer que les premières conformations [De 0 à 126] semblent avoir des rayons de giration plus petits et seraient donc des conformations plus repliées que les suivantes [De 127 à 200].

Une analyse des données du fichier de sortie nous permettrait de faire une analyse plus précise des données dans un second temps.

### **Calcul des distances moyennes de chaque résidu par rapport au centre de masse de la protéine**

Une analyse locale de la protéine consiste à calculer les distances moyennes entre le centre de chaque acide aminé et le centre de masse de la protéine (dans sa conformation correspondante).

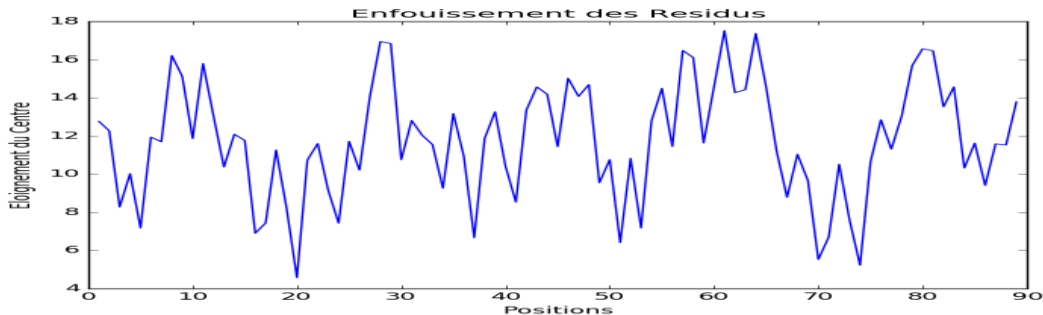
Ceci témoigne de l'enfouissement de chaque résidu dans la protéine car plus la distance moyenne est élevée, plus le résidu observé a tendance à se retrouver vers la surface de la protéine. A contrario, si la distance moyenne est faible, cela montre que le résidu se trouve préférentiellement vers le centre de la protéine.

Nous effectuons donc la moyenne des distances entre les différentes conformations pour chaque résidu de la protéine :

$$DistanceMoyenne = \frac{1}{N} \sqrt{\sum_{i=1}^N (CentredeMasseGlobal - CentredeMasse[i])^2}$$

Pour N le nombre de résidus dans la protéine d'intérêt et i un résidu.

On obtient donc pour chaque acide aminé :



*Figure 2 : Distance de chaque résidu de la protéine par rapport à son centre de masse*

Cela permet de voir les régions possédant des acides aminés en surface ou vers le centre de la protéine.

Ces résultats peuvent ensuite être comparés aux différentes RMSD moyennes obtenues pour chaque résidu (voir plus loin) pour tenter de trouver une corrélation entre l'enfouissement de certains résidus et les différences de positions entre les différentes conformations de la protéine.

## Calcul de RMSD

Dans cette partie nous allons détailler le calcul de la RMSD afin de voir comment celui-ci est obtenu et à quoi il peut servir.

La RMSD (Root Mean Square Deviation) correspond à la racine de la moyenne des distances élevées au carré. En effet, nous disposons de la Barstar de référence ainsi que d'un autre fichier avec les différentes conformations de la Barstar, dans lesquels les séquences ont été alignées. Le principe consiste donc à comparer la Barstar de référence à ses conformations en les superposant et à calculer la distance entre toutes les paires d'atomes.

## Calcul de RMSD globale

La RMSD globale d'une protéine correspond à la distance spatiale globale entre les 2 protéines. Pour la calculer nous avons donc calculé, pour chaque paire d'atomes, la RMSD pour ensuite additionner cette RMSD de chaque paire d'atomes et ainsi obtenir avoir la RMSD de chaque résidu. Ensuite ces RMSD de chaque résidu sont additionnées pour avoir la RMSD de chaque conformation.

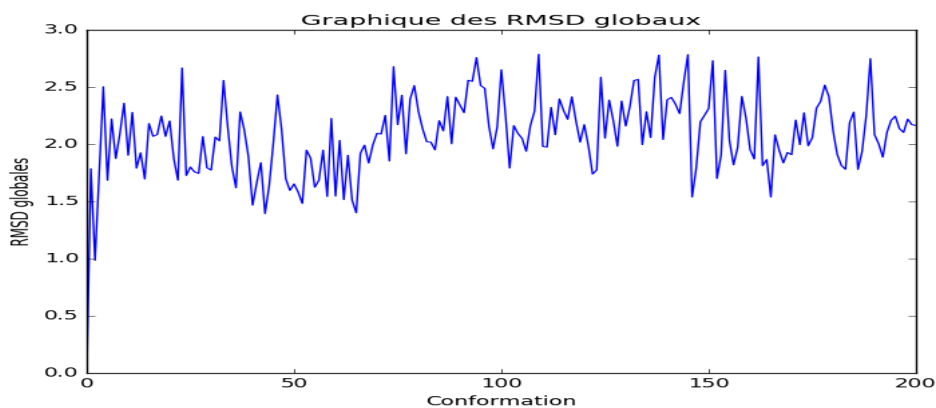
$$RMSDglob = \sum RMSDresidu = \sum \sum RMSDpairatome$$

$$Avec RMSDpairatome = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta^2 i}$$

Avec N le nombre de paires d'atomes à chaque résidu.

La RMSD ainsi obtenue nous permet de voir quelle conformation s'écarte beaucoup de la conformation de référence et de déterminer si cette protéine est stable car si la RMSD change beaucoup, cela signifierait que la distance entre la Barstar de référence et la conformation change et donc que la protéine est instable.

Pour faciliter l'analyse, nous avons intégré une fonction graphique pour pouvoir observer la RMSD globale en fonction des conformations et donc en fonction du temps (car les conformations possèdent des identifiants et sont rangées dans l'ordre croissant) (voir figure 1).



*Figure3 : Exemple de graphique RMSD globaux*

Nous pouvons ainsi avoir une vision globale des RMSD obtenue et pouvons interpréter plus facilement. Nous avons aussi fait une fonction pour récupérer les données générées après l'analyse globale, dans un fichier texte pour pouvoir avoir une analyse plus approfondie.

### Calcul de RMSD locales.

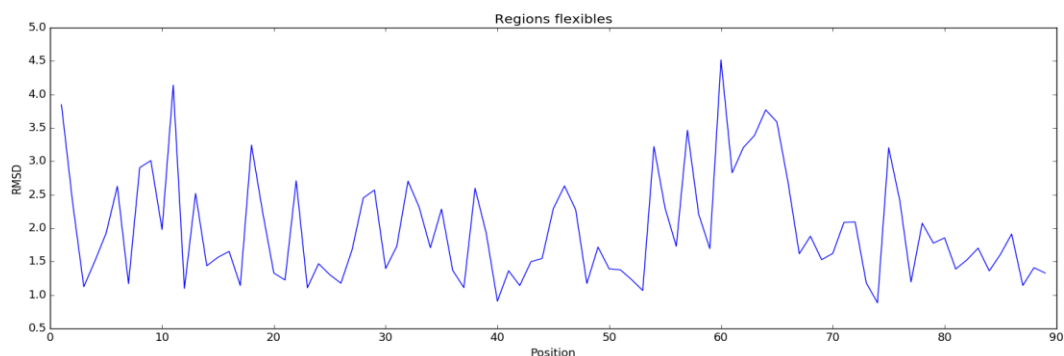
La RMSD locale d'une protéine correspond quant à elle aux RMSD moyennes à chaque résidu. Pour la calculer, il nous faut donc calculer la RMSD pour chaque paire d'atomes puis sommer la RMSD à chaque paire d'atomes pour avoir la RMSD à chaque résidu. Ainsi nous obtenons la RMSD pour chaque conformation de chaque résidu et ensuite nous avons additionné ces RMSD à chaque conformation pour obtenir la somme de toutes les RMSD de chaque position. Enfin nous avons divisé chaque RMSD par le nombre de conformations pour obtenir une RMSD moyenne à chaque positions.

$$RMSD_{locale} = \frac{RMSD_{residu}}{nbdeconformation}$$

$$\text{Avec } RMSD_{residu}(j, i) = \sum_{j=0}^{Nbres} RMSD_{pairatome}(i)$$

Avec Nbres correspondant au nombre de résidus dans les conformations.

La RMSD ainsi obtenue nous permet de voir la variation moyenne de RMSD dans la protéine. Ainsi nous pourrions déterminer quelles parties de la protéine sont les mieux conservées et quelles parties ne sont au contraire pas conservées. Pour faciliter l'analyse de cette partie, nous avons introduit une fonction graphique nous permettant d'observer la variation des RMSD moyennes en fonction des positions et pouvoir ainsi voir visuellement cette variation (voir figure 2). Nous avons aussi créé une fonction nous permettant de récupérer les données générées par l'analyse locale dans un fichier texte pour avoir la possibilité 'effectuer une analyse plus approfondie par la suite.



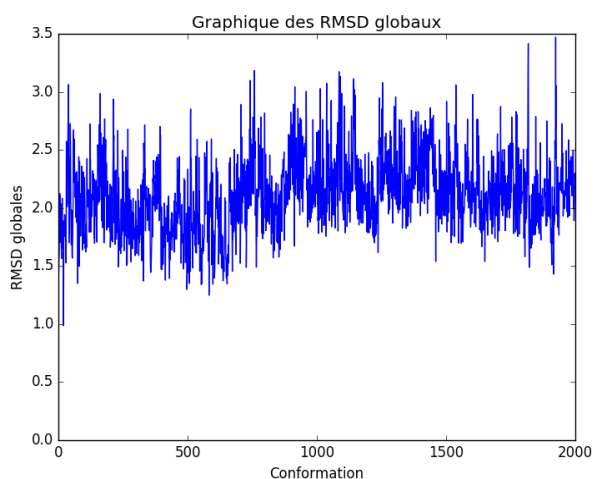
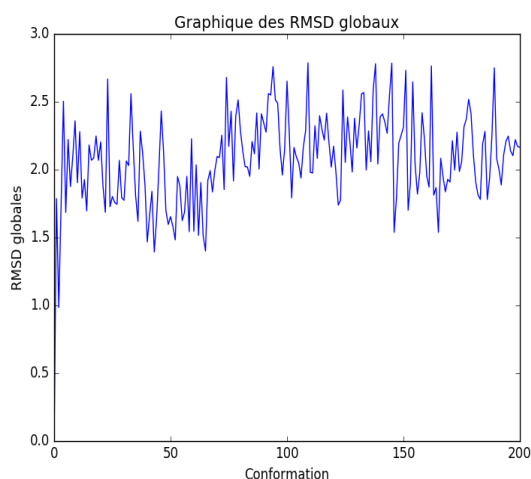
**Figure 4 :** Exemple de graphique RMSD local

## Résultats biologiques et interprétation

Grâce à l'exécution de ce programme, nous avons obtenu plusieurs résultats que nous allons analyser au cours de cette partie.

### Résultats globaux

Tout d'abord analysons la sortie de RMSD globales. Afin d'avoir un aperçu général de la RMSD, nous pouvons regarder le graphique des RMSD globales par rapport aux conformations. Nous pouvons voir que celles-ci sont comprises majoritairement entre 1.4 à 3.5 Å selon le fichier, avec énormément de fluctuation entre les conformations. Nous pouvons aussi voir que seule la première conformation a une RMSD très basse et donc cette conformation est très proche de la protéine de référence. Cela montre donc que la protéine subit bien des changements conformationnels au cours du temps.



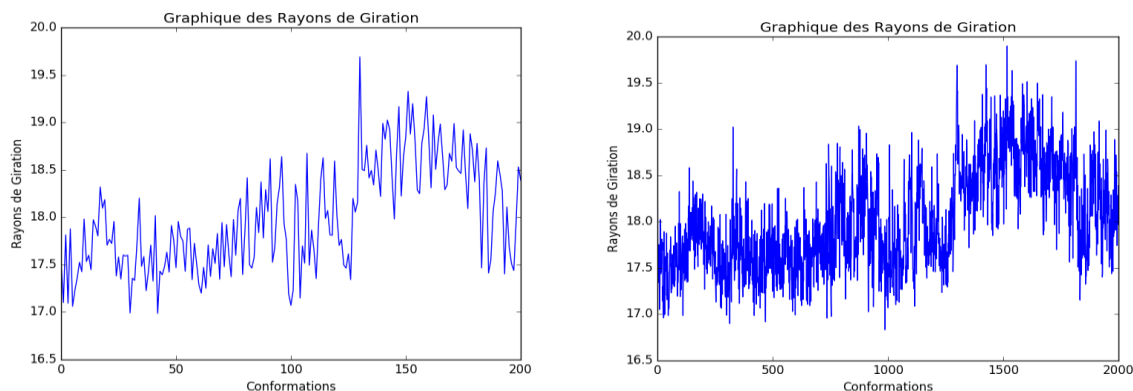
**Figure3 :** Graphiques des RMSD Globales en fonction des conformations (1) avec le fichier `md_prot_only_skip100` (2) avec le fichier `md_prot_only_skip10`

Ces changements peuvent en général être dus à des mouvements de boucles d'assez grandes amplitudes pour provoquer des différences de distances spatiales entre les protéines et donc une augmentation de la RMSD. Cela est visible quelque soit le nombre de conformations observées et en augmentant le nombre de conformations, nous n'arrivons donc pas à une RMSD globale stable qui pourrait montrer une conformation préférentielle après un certain temps. En effet, nous ne distinguons pas de tendance ni de valeurs stable dans ces valeurs de RMSD même avec un très

grand nombre d'observations mais nous pouvons tout de même remarquer que les fluctuations semblent diminuer entre les conformations 1900 et 2000 vers une RMSD de 2.1 Å (variance peut-être également plus faible car pics et creux plus petits).

Pour vérifier à quoi sont dus ces mouvements, nous allons donc regarder les résultats des rayons de giration des différentes conformations.

En observant ces résultats, on remarque que le rayon de giration varie énormément entre 17 et 20 Å. Comme nous observons beaucoup de variations, cela signifie que la distance du résidu le plus éloigné du centre de la protéine diffère énormément : soit un même résidu se rapproche et s'éloigne du centre, soit un autre résidu devient le plus éloigné (cela dépend de la conformation). Cela décrit à la fois un mouvement de compaction mais aussi un mouvement de dépliement de la protéine donc lorsque le rayon de giration augmente, le mouvement est un dépliement, alors que lorsque le rayon de giration diminue, le mouvement est une compaction. De plus, en observant les résultats sur les différents graphiques, il n'y a pas de valeur de rayon de giration stable même avec 2000 conformations. Cela montre donc que la protéine est en perpétuelle mouvement de dépliement et de compaction.



**Figure4 :** *Graphiques des Rayons de giration en fonction des conformations (1) avec le fichier md\_prot\_only\_skip100 (2) avec le fichier md\_prot\_only\_skip10*

Cela est cohérent avec les observations précédentes car du fait que la protéine soit en perpétuel mouvement de compaction/dépliement, ceci explique que la RMSD globale fluctue énormément aussi avec la compaction et le dépliement des boucles qui font alors un mouvement de grande amplitude.

Nous allons désormais nous intéresser aux résultats locaux pour observer ces modifications au niveau de chaque résidu afin de voir quels sont les résidus impliqués dans ces changements conformationnels mais aussi où sont placés ces résidus dans la protéine (leur distance par rapport au centre global de la protéine).

## Résultats locaux

Dans cette partie, les résultats ont été obtenus grâce à des moyennes, en comparant les résultats avec les fichiers md\_prot\_only\_skip100 et le fichier md\_prot\_only\_skip10, nous nous apercevons que nous obtenons des résultats quasiment égaux car les 200 premières conformations suffisent pour faire une moyenne et les 1800 autres conformations contenues dans le fichier md\_prot\_only\_skip10 ne changent presque pas ces moyennes étant donné que les valeurs sont globalement similaires.

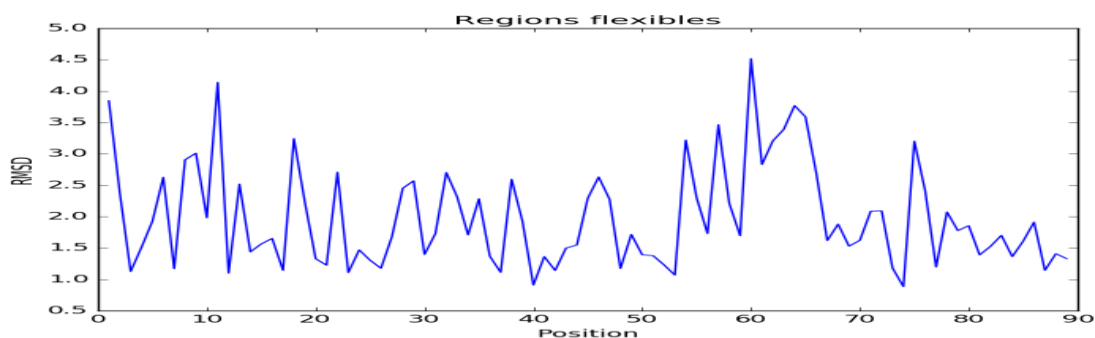


Dans cette partie, nous allons donc prendre uniquement les données de md\_prot\_only\_skip100 mais ces résultats et interprétations sont tout aussi bien applicables au fichier md\_prot\_only\_skip10.

Dans un premier temps, observons les résultats de RMSD locales obtenus en fonction du résidu. Nous pouvons ainsi remarquer que la RMSD varie entre 0.98 et 4.5 Å. Cependant nous pouvons observer des régions où les variations sont moins grandes (par exemple les résidus 48 à 53) alors qu'à d'autres endroits, la RMSD est plus importante (par exemple les résidus 61 à 68). C'est ainsi que nous avons pu déterminer que les régions suivantes subissent le plus de variation car leurs RMSD est grande :

- 0 à 3 (ce qui est normal car c'est l'extrémité C-ter)
- 6 à 11 : succession de pics
- 27 à 39 : succession de pics
- 43 à 46 : large pic
- 54 à 72 : large succession de pics

Les autres régions que l'on peut trouver sur ce graphique sont majoritairement une succession de creux ou des régions comprenant moins de variations et pour des RMSD plus faibles.



**Figure5 :** Graphiques des RMSD locales en fonction des résidus avec le fichier md\_prot\_only\_skip100

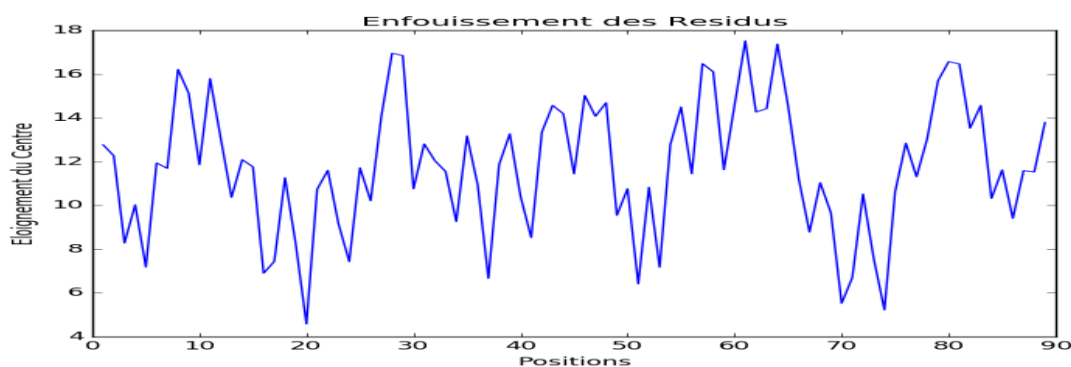
Nous avons donc déterminé les régions qui semblent plus flexibles que d'autres et désormais nous allons observer l'enfouissement de ces régions afin de déterminer si celles-ci sont enfouies ou non dans la protéine.

Pour cela, nous disposons des résultats de calcul de l'éloignement du centre en fonction des résidus. Plus la valeur de l'éloignement est grande et moins le résidu est enfouie dans la protéine, et donc plus il est en surface.

Dans ce graphique nous pouvons observer des fluctuations, ce qui est normal car la protéine est repliée sur elle-même et donc certains résidus sont enfouies et d'autres en surface. Ce qui nous intéresse donc est de savoir si les régions déterminées ci-dessus sont enfouies. Nous regardons donc les résidus :

- 6 à 11 : la valeur de l'éloignement varie entre 11 et 16 ce qui est plutôt élevé, donc la région semble plutôt en surface
- 27 à 39 : les valeurs varient entre 11 et 17 donc la région est aussi en surface
- 43 à 46 : les valeurs varient entre 12 et 15 donc la région est en surface
- 54 à 72 : les valeurs varient entre 6 et 17 il y a donc une partie en surface et une partie plus enfouie



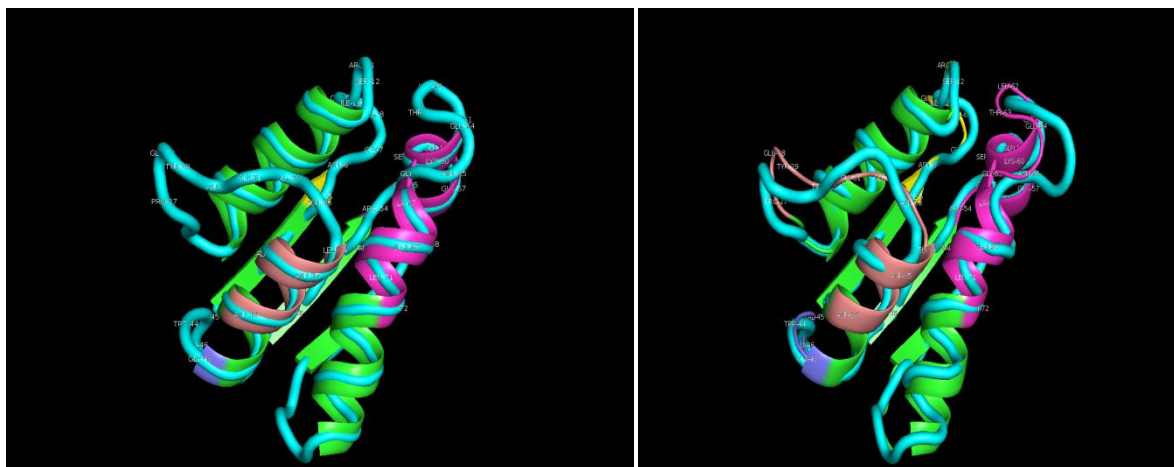


**Figure5 :** Graphiques de l'enfouissement du centre en fonction des résidus avec le fichier *md\_prot\_only\_skip100*

Nous avons donc vu que la plus part des résidus qui subissent le plus de changement sont situés en surface mais qu'il y a aussi des parties plus internes qui changent bien que cela reste une minorité.

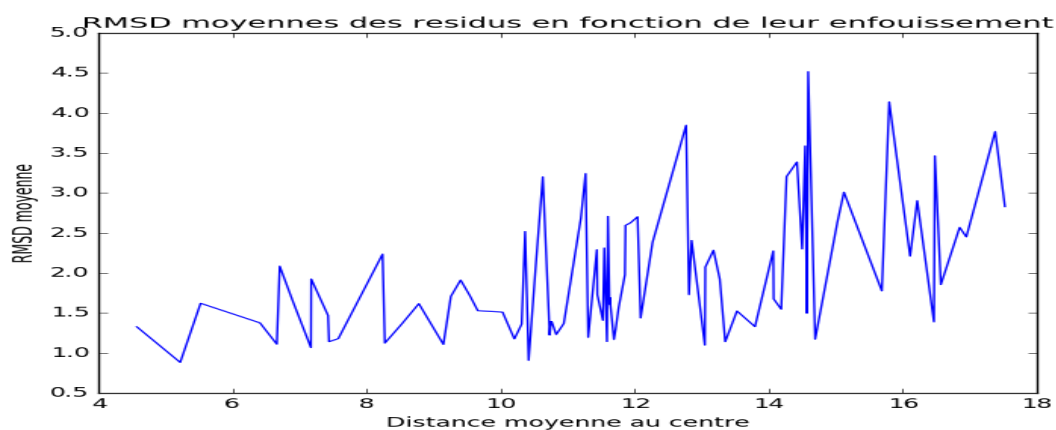
Nous avons donc vu que la protéine Barstar subit des changements conformationnels au cours du temps et ces changements ont lieu pour la majeure partie sur des résidus de surface mais il arrive que certains résidus enfouis subissent des changements.

Pour appuyer ces résultats, nous avons modélisé cette protéine via le logiciel PyMol, ainsi que marqué les régions précédemment retenues comme étant les régions à fortes flexibilités, et nous avons superposé des conformations de la Barstar au-dessus de la protéine Barstar de référence.



**Figure6 :** Photographie PyMol de la Barstar de référence (structure 3D schématisée) et de l'une des ses conformation (structure bleue continue).

Nous pouvons donc, grâce à ces photographies, voir que sur la photographie de gauche, la conformation observée de la Barstar concorde parfaitement avec la structure de la Barstar de référence. Par contre, pour une autre conformation, cela a changé et la nouvelle observation ne concorde plus à certains endroits et principalement aux endroits colorés correspondant aux régions supposées plus flexibles.



*Figure 7 : RMSD moyenne de chaque résidu en fonction de son enfouissement moyen dans la protéine*

D'après ces résultats, on peut voir que la RMSD semble plus élevée et varie bien plus pour les résidus semblant plutôt situés vers la surface.

Le site permettant l'association avec la Barnase se trouve donc probablement en surface car ce correspond à la région dont les RMSD varient le plus, ce que nous cherchons. En effet, nous possédons l'information (grâce à l'énoncé) que 2 acides aminés subissent un fort changement conformationnel pour pouvoir fixer la Barnase. Les pics les plus importants [Position 60 : RMSD = 4.517] [Position 11 : RMSD = 4.1] sont donc probablement ces acides aminés clés et qu'ils permettent donc cette fixation en vue d'une inhibition par le site actif.

## Conclusion

L'analyse des changements conformationnels de la protéine Barstar au cours du temps nous a permis d'observer qu'elle pouvait effectuer diverses actions de compaction/dépliage attestant des mouvements de boucles de grandes amplitudes observés lors de l'analyse des RMSD globaux. Suite à ces observations nous avons réalisé des calculs correspondant à une analyse plus approfondie, à l'échelle des résidus et non de la protéine globale.

Ceci nous a permis de voir quelles régions étaient flexibles (RMSD importante) et ainsi rapprocher cela à l'enfouissement des acides aminés dans ces régions d'intérêt au sein de la protéine. Cette observation nous a donc permis d'identifier 2 résidus subissant des changements conformationnels plus importants, sur lesquels nous pourrions effectuer une analyse plus approfondie pour confirmer leur rôle dans la fixation de la Barnase dans le complexe Barstar-Barnase.

Une fois la Barnase fixée sur la Barstar, l'activité d'inhibition de cette dernière sur la première doit se faire dans une région fortement conservée. Le centre des protéines étant constitué, en général, d'un cœur hydrophobe correspondant au site actif, on peut soumettre l'hypothèse que ce centre de la Barstar serait son site actif permettant ainsi l'inhibition du substrat qu'est la Barnase.