Université Paris Diderot

5 rue Thomas Mann

75003 Paris

Calcul automatisé d'angles entre deux sous domaines protéiques

Auteur:
Benjamin Millot
benjamin.millot@outlook.fr

Encadrants:
Thibault Tubiana
thibault.tubiana@gmail.com
Catherine Etchebest
catherine.etchebest@univ-paris-diderot.fr
Jean-Christophe Gelly
jean-christophe.gelly@univ-paris-diderot.fr

Vendredi 08 Janvier 2016

Contents

1 Introduction

Le Norovirus, plus connu sous le nom du virus de la gastro-entérite, est responsable d'infections inflammatoires du système digestif et entraine entre autre diarrhées, vomissements et céphalées[1]. Le Norovirus se structure de la manière suivante: une capside, d'environ 30-40 nm de diamètre, entoure et protège l'ARN viral[2, 3]. La capside du Norovirus est constituée de 180 protéines dont la séquence, nommée VP1, est identique. L'étude de cette capside est donc un élément clé dans la compréhension et la lutte contre le Norovirus.

De récentes études [4, 5] ont montrés que la protéine VP1 pouvait être décrite en deux domaines bien distincts, S (pour *Shell*) et P (pour *Protruding*). Ces deux domaines témoignent d'une certaine flexibilité l'une vis-à-vis de l'autre. Dans cette étude, nous nous proposons donc de développer une méthode originale visant à calculer l'angle entre ces deux domaines au cours d'une dynamique moléculaire.

1.1 État de l'art sur le Norovirus

La gastro-entérite peut être d'origine bactérienne (comme par exemple la contamination par les *colibacilles*) ou virale. Parmi les différents virus responsables, le Norovirus est connu pour être la cause la plus courante de diarrhée dans les pays développés[1]. L'être humain est actuellement le seul réservoir connu.

Le Norovirus est virus non enveloppé à ARN: une capside entoure l'ARN viral et va permettre la fixation puis l'infection de l'hôte. La capside VP1 du Norovirus est constituée de deux domaines. Le domaine S, pour *Shell*, est une structure de 191 acides aminés (29-220) qui constitue la coque interne de la capside. L'autre domaine, le domaine P (230-520), pour *Protruding*, s'élève du plan du domaine S à la manière d'un pic érigé. Ces deux domaines sont reliés par une dizaine d'acides aminés[4, 5].

Des études menées par *Hansmann et al.* et *Katpally et al.* [4, 5]. ont montrées qu'il existe une flexibilité entre ces deux domaines. Cette flexibilité entraine une variation d'angle entre S et P. C'est cette variation que nous chercherons à mesurer au cours de l'étude d'une dynamique moléculaire de VP1. Notre modèle sera la la protéine VP1 du Norovirus (code PDB: 1IHM) cristallographiée par *Prasad et al.* Trois monomères A, B, et C, de séquence identiques, présentent des différences dans l'orientation des plans de S et de P, entrainant de fait de légères variations structurales, qui permettent par symétrie de reconstituer la capside du Norovirus. Nous chercherons donc à vérifier si la molécule A peut tendre vers une conformation C au cours d'une dynamique moléculaire par l'étude de l'angle et de la distance entre les deux domaines (*cf. Figure 1*).

1.2 Calcul des angles et intérêt de l'approche proposée

Le calcul d'un angle entre deux plans d'une même protéine n'est en effet pas un problème trivial. Certains outils de calcul d'angles existent actuellement, mais ne répondent pas tout à fait

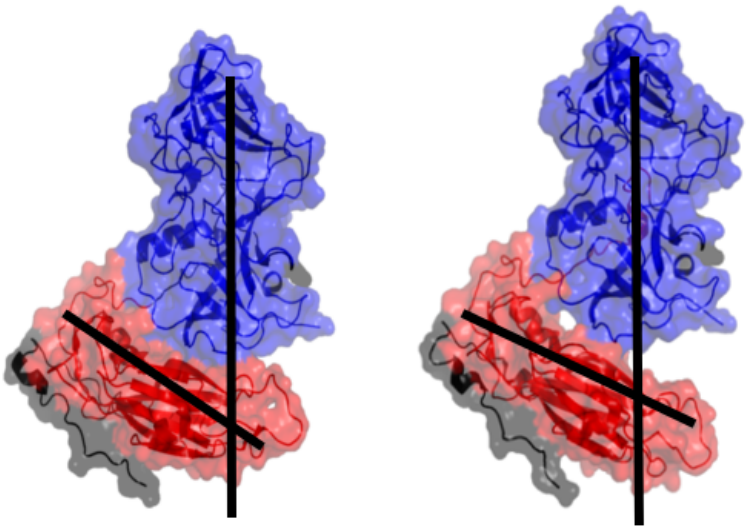

Figure 1: **Variations des plans de S et de P chez la protéine A (à gauche) et la protéine C (à droite).**

à notre besoin ou nécessitent des connaissances sur préalables sur les plans de la protéine. L'outil gmx gangle de Gromacs calcule différents types d'angles sur entre deux vecteurs au cours d'une trajectoire, mais nécessite un point de référence ainsi qu'un vecteur normal au point. Le logiciel de visualisation Chimera permet au contraire de calculer un angle entre deux plans d'une protéine, mais pas de réaliser le calcul au cours d'une dynamique moléculaire. Certaines approches peuvent être envisageables par le logiciel de visualisation VMD, mais nécessitent un scripting important au préalable du fait du manque de fonctions relatives aux angles du logiciel (actuellement, un seul script personnalisé a été trouvé, calculant l'angle entre trois points attribués empiriquement).

L'approche que nous développerons ici permet le calcul de l'angle entre les deux domaines sans connaissance préalable des plans de la protéine: l'utilisateur doit simplement spécifier les limites respectives de chacun des domaines. Il n'est pas nécessaire de connaitre un point de référence ou l'équation du plan. L'algorithme utilisé emploie en effet les vecteurs propres et non les plans à proprement parler pour générer ses résultats, ce qui simplifie à la fois les calculs et réduit les approximations. Le programme a aussi été pensé pour être adaptable à d'autre protéines.

2 Matériel et méthodes

2.1 Génération des données

Les données de références proviennent des résultats d'une dynamique moléculaire à réalisée sur un monomère A de la protéine 1IHM, pendant 5 nanosecondes avec Gromacs et le champ de force Amber 99SB-ILDN. La dynamique a été traitée ensuite traitée afin d'être nettoyée: les ions et l'eau ont été supprimés et la trajectoire à été *smoothée*. Nous disposons ainsi de deux fichiers d'entrée: le fichier de topologie et le fichier binaire de la trajectoire.

2.2 Langage et dépendances

Python 2.7 a été choisi comme langage de programmation, à la fois pour sa simplicité syntaxique et ses capacités de parsing des fichiers d'entrée. Le langage est aussi compatible avec le package de Konrad Hinsen *GromacsTrajectory*, capable de lire les fichiers binaires de trajectoire, qui nous a été fourni. Nous avons cherché à nous adapter au mieux aux données brutes: ainsi *GromacsTrajectory* sera le seul paquet externe utilisé. Les autres dépendances (numpy, re, matplotlib, argparse) sont inclus de base dans la version 2.7 de Python.

Le programme se concentrant sur le calcul d'angle et de distances au cours d'une trajectoire, il n'a pas été optimisé à proprement dit pour la vitesse de production des résultats. A l'inverse, il a été pensé comme un module de lecture des trajectoire à part entière, à la manière des logiciels *mdtraj* ou *MDAnalysis*. Il offre ainsi à l'utilisateur (au pré-requis de connaissance minimales en programmation orientée objet) deux avantages majeurs: pouvoir consulter et vérifier les résultats des calculs à n'importe quel moment de la trajectoire et pouvoir développer avec facilité ses propres fonctions.

2.3 Organisation du programme

Le programme a été pensé avec le paradigme objet afin de faciliter sa compréhension et son utilisation (*cf. Figure 2*). Une brève description des classes est présentée ici dans un souci de clarté.

- Un objet **Md**, parent, qui va contenir toutes les informations de la dynamique. C'est à la fois le point de départ du programme (remplissage des données) ainsi que le point d'arrivée (strockage des valeurs requises pour le tracé des graphes).
- Un objet **Topology** qui va contenir les informations relatives à la topologie (residus, positions, atomes).
- Un objet **Trajectory**, relatif à la trajectoire, qui va contenir une succession de **Frame**.
- Des objets **Frame**, contenant l'état de la protéine à l'instant t (résidus, atomes, coordonnées, domaines). Les résidus et atomes sont extraits de la **Topology**, tandis que les coordonées

sont lues à partir du paquet *GromacsTrajectory*. Chaque **Frame** contient deux objets **Domain**.

- Des objets **Domain** contenant les informations relatifs aux domaines étudiés (position de début et de fin, résidus inclus, atomes inclus, coordonnées relatives). Dans le but de diminuer le bruit et de limiter les opérations de calculs complexes, ce sont sur ces objets que la sélection des carbones alpha (ainsi que leurs coordonnées respectives) et l'affinage par MSF (*Mean Square Fluctuation*) sera réalisée. Chaque objet **Domain** contient un unique objet **Eig**.
- Des objets **Eig**, qui contiennent les valeurs propres et vecteurs propres des coordonnées référencées dans le **Domain** parent. Ces vecteurs seront utilisés pour calculer les angles et les distances qui seront remontés en amont dans l'objet **Md** parent.

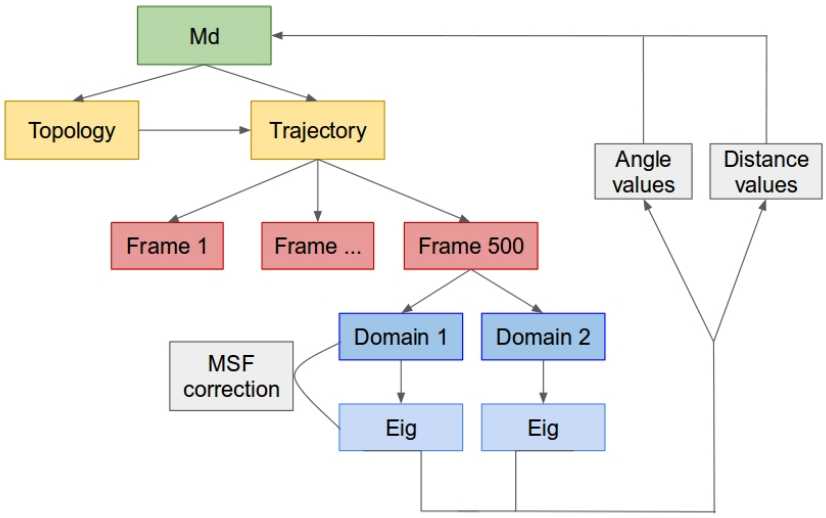

Figure 2: **Organisation et structure du programme**

2.4 Choix algorithmiques et calculs vectoriels

Le module opère de la manière suivante: a chaque frame, les deux domaines S et P sont délimités. Dans l'optique de réduire le bruit, seul les carbones alpha des acides aminés sont considérés. La matrice de coordonnées des atomes d'un domaine est alors centrée et transformée en matrice d'inertie en multipliant la matrice elle-même par sa transposée. A partir de cette matrice trois valeurs propres liées à trois vecteurs propres sont calculés. Les vecteurs propres, formant un axe orthonormé, donnent les trois directions de l'information pour les points considérés. Ainsi, partant du principe que les domaines S et P forment de relativement "bons" plans, les deux premiers vecteurs propres décrivent les deux directions du plan sous-jacent. Par définition, le troisième vecteur propre (associé à la troisième valeur propre) est orthogonal aux deux premiers. Il décrit donc le vecteur normal au plan.

C'est ce troisième vecteur qui nous sera utile ici. Par projection, l'angle entre les troisièmes vecteurs propres des deux domaines est égal à l'angle entre les plans de S et P. Néanmoins, les vecteurs propres ne donnent d'informations sur le sens de l'information, ce qui influe sur l'angle mesuré et qui parfois conduit a la mesure de deux angles différents (le petit et le grand) au cours de la même trajectoire. Nous employons le signe du produit scalaire entre les deux vecteurs pour déterminer s'ils sont orientés dans le même sens ou non. Cet effet est ainsi pris en compte dans le programme, qui s'assure de toujours vérifier le même angle au cours de la dynamique.

Pour affiner les résultats, nous avons aussi inclus des fonctions permettant d'éliminer les résidus ayant un fort mouvement au cours de la dynamique. L'algorithme utilisé est celui du MSF, pour *Mean Square Fluctuation*, qui attribue une valeur de mouvement a chacun des résidu en comparant sa position à la frame i par rapport à une position de référence moy (position moyenne du résidu au cours de la dynamique). Le MSF est ainsi décrit par la fonction suivante:

$$MSF_{resi} = \frac{\sum_{i=1}^{N_{frames}} pos_i - pos_{moy}}{N_{frames}}$$

Une fois les valeurs de MSF pour chaque résidus calculée, un nettoyage des résidus présentant les plus grands MSF est effectué. La définition d'une "trop grande valeur" est déterminée par l'algorithme de Box-Whisker: une fois les données classés, les quartiles Q1, Q2 et Q3 sont calculés. Les résidus à supprimer sont ceux dépassant le seuil suivant:

$$Seuil = Q_3 + 1.5 \times (Q_3 - Q_1)$$

3 Résultats

3.1 Angles entre les deux domaines au cours de la dynamique

L'angle entre les domaines S et P à ainsi été calculé au cours de la trajectoire (*cf. Figure 3*). Nous avons délimité les résidus du domaine S entre 29-201 et ceux du domaine P entre 230-520. Le choix de ces domaines sera discuté dans la suite de l'étude. Seuls les carbones alpha ont été considérés afin de diminuer le bruit. La totalité des résidus de ces deux domaines n'ont pas été pris en compte: nous avons réalisé un affinage en supprimer les résidus ayant un MSF trop important, c'est-à-dire ceux démontrant le plus de mouvement durant la dynamique.

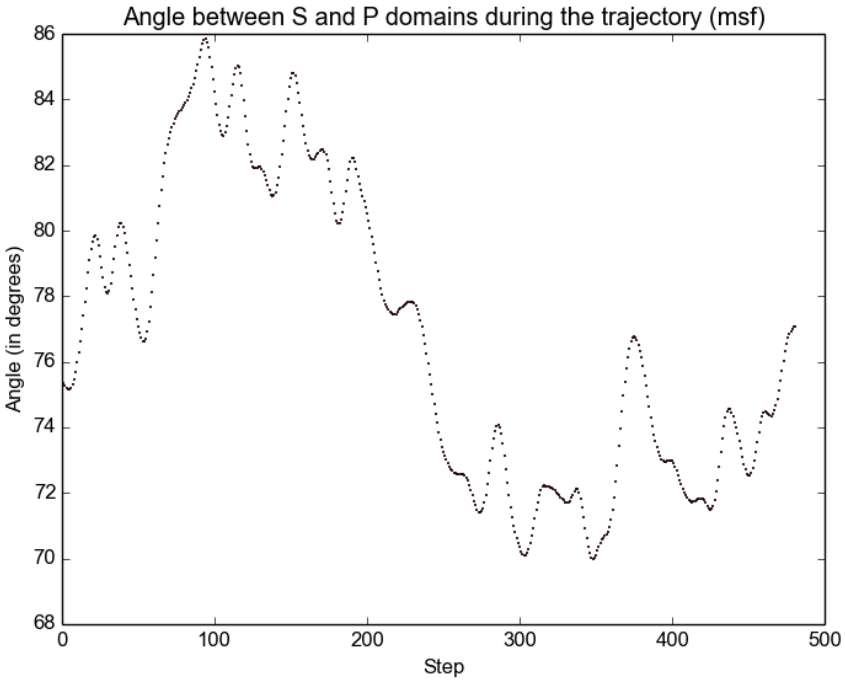

Figure 3: **Angle entre les domaines S et P mesurés au cours de la trajectoire.** Les résidus présentant de grandes variations de mouvements ont été supprimés.

Il existe une variation d'angle de 16 degrés durant la trajectoire de 5 nanosecondes. Durant la première ns, l'angle entre les deux plans de S et P augmente (jusqu'à 86°), puis diminue jusqu'à la troisième ns (70°). L'angle entre les deux plans évolue alors de manière constante en dents de scie, mais ayant une légère tendance à augmenter, jusqu'à la fin de la dynamique. Au regard de la question initiale, il ne nous est pas possible d'affirmer que le monomère A tend vers une conformation C. Nous avons néanmoins prouvé qu'il existe bien **in silico** une flexibilité de la protéine VP1 entre les deux domaines.

3.2 Distance entre les deux domaines au cours de la dynamique

Afin de disposer d'une autre référence de comparaison de S et P, nous avons calculé, à chaque frame de la dynamique, la distance entre les barycentres des deux domaines respectifs (*cf. Figure 4*). La distance entre les barycentres semble évoluer de manière inverse à l'angle mesuré: diminution de la distance (de 3.70 à 3.60 nm) pendant la première ns, puis augmentation (3.80 nm) jusqu'à 3ns, avant d'évoluer de manière constante en dents de scies jusqu'à la fin de la trajectoire.

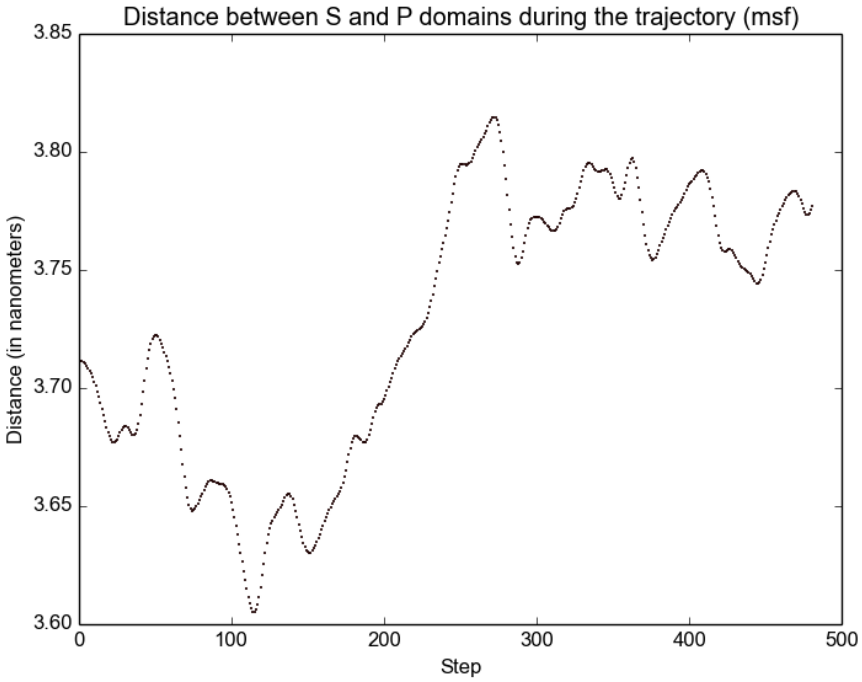

Figure 4: **Distance entre les barycentres des domaines S et P mesurés au cours de la trajectoire.** Les résidus présentant de grandes variations de mouvements ont été supprimés. Les barycentres ont été ré-estimés à chaque frame.

Il est difficile de tirer des conclusions significatives de ces résultats seuls, car la variation totale de 2 Angstrom est suffisamment petite pour être expliquée aussi bien par une évolution vers une conformation C, conformationnellement très proche, qu'un mouvement faible de la protéine au cours de la dynamique. L'étude du RSMD de la protéine présentée ci-dessous ne nous permet pas de conclure quant à l'une ou l'autre des hypothèses, mais nous pouvons néanmoins affirmer que les résultats sont cohérents avec la variation angulaire entre les deux domaines dans la mesure où tous deux montrent la flexibilité de VP1.

3.3 Mesures du RMSD

Pour enquêter plus en détail sur la question, nous avons comparé le RMSD (Root-Mean Square Deviation) des frames de la trajectoire avec la molécule A de référence (*cf. Figure 5*), puis la molécule C extraite de 1IHM (*cf. Figure 6*). Le RMSD est une mesure de l'alignement de deux structures entre elles: plus la valeur est petite, plus les molécules sont structurellement proches. Notons que nous considérons les valeurs suivantes significatives dans la mesure où, si les valeurs en elle-même sont faibles, la molécule est petite (520 résidus), les changements faibles ont donc un impact digne d'intérêt.

Nous avons tout d'abord réalisé la mesure du RMSD vis-à-vis de la molécule A de référence (*cf. Figure 5*). La variation du RMSD oscille entre 0.12 et 0.25 nm au cours de la trajectoire, et augmente globalement au cours de la trajectoire. Cette observation mène a deux conclusion. Premièrement, les valeurs d'angles et plus particulièrement les faibles valeurs de distances sont bien cohérents avec la dynamique: le faible mouvement observé au cours de la trajectoire produit de faibles valeurs de distances. En second lieu, le RMSD mesuré est croissant sur le temps et n'atteins pas un plateau d'équilibre. Il est possible que cette dynamique sur 5 ns soit trop courte (10 ns aurait peut-être été préférable) pour être stable. En conséquence, même si nous considérons la variation du RMSD comme particulièrement intéressante pour confirmer nos précédents résultats, nous prendrons soin de n'affirmer aucune corrélation directe entre la variation du RMSD et le fait que la protéine A tende vers une conformation C.

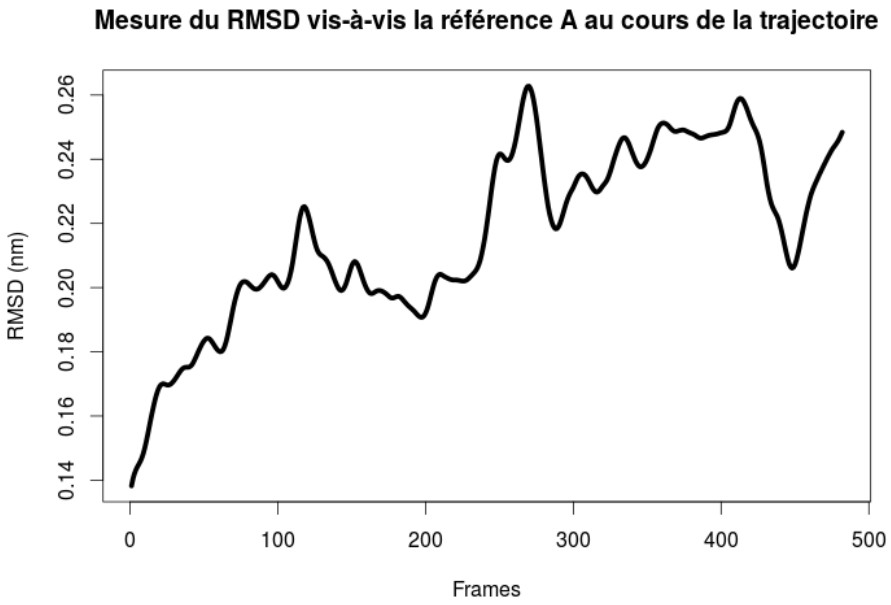

Figure 5: **RMSD de la trajectoire vis-à-vis de la molécule A de référence**. Le RMSD est mesuré sur un alignement en tout-atomes.

Gardant cela à l'esprit, nous avons mesuré le RMSD de la dynamique par rapport à la molécule C extraite du fichier PDB 1IHM (*cf. Figure 6*). Il semble qu'aux alentours de la 180^{eme} frame, la molécule A est la plus proche de la C (RMSD = 3.045 nm). Les valeurs grandes valeurs d'angles entre S et P semble s'accompagner d'un meilleur alignement avec la molécule C. Cependant la variation du RMSD est très faible (de l'ordre de 0.2 Angström) et nous empêche d'affirmer explicitement qu'un passage à l'état C de la molécule A est possible, mais laisse néanmoins entrevoir la possibilité d'une flexibilité capable d'"orienter" les domaines de la protéine.

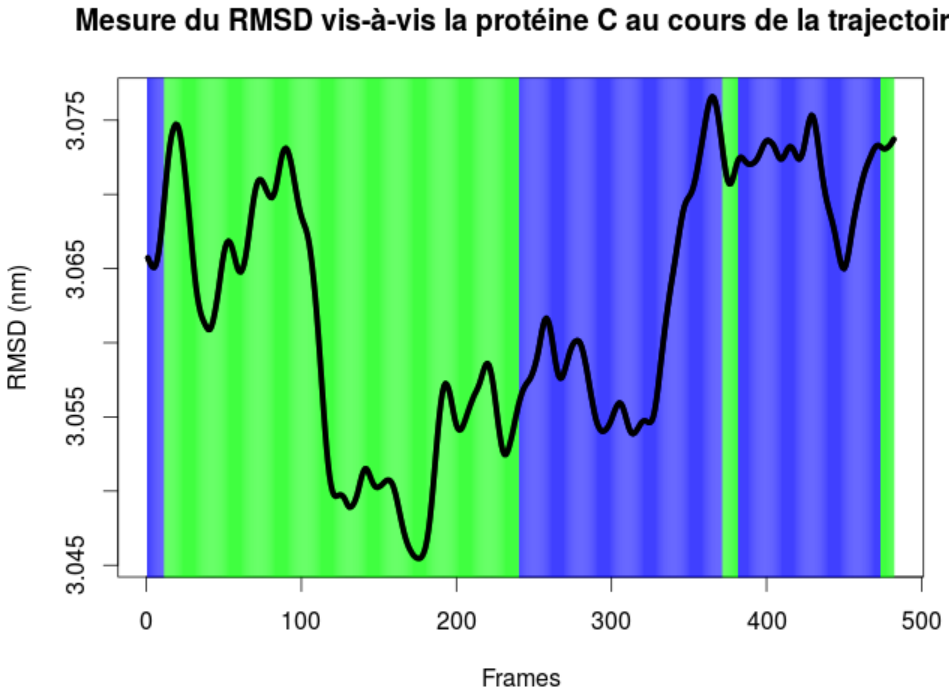

Figure 6: **RMSD de la trajectoire vis-à-vis de la molécule C de 1IHM**. Les couleurs de fond renseignent sur la valeur de l'angle mesurée. Quand l'angle est plus grand que celui de référence (supérieur à 76°) un coloration verte est définie; quand l'angle est plus petit (inférieur à 76°), une coloration bleue est définie. Les protéines sont alignées en tout-atomes.

4 Discussion

4.1 Qualité des résultats

De manière globale, nous sommes satisfaits des résultats de obtenus au cours de cette étude. Notre superviseur M. Thibault Tubiana a eu la gentillesse de nous transmettre ses résultats personnels (basés sur un algorithme créant un plan passant au mieux par les résidus du domaines, *cf. Figure 7*), dans la mesure où cette étude s'inscrit dans un projet qui n'a pas encore été publié à ce jour. Nos valeurs d'angles mesurées, même si elles ne sont pas exactement identiques, sont relativement proches. De plus, les profils et les variations des courbes sont sensiblement les mêmes, ce qui nous conforte dans l'interprétation de nos résultats. En effet, cela prouve à la fois que nos calculs sont corrects et que l'approche par vecteurs propres est équivalente à celle passant par les plans.

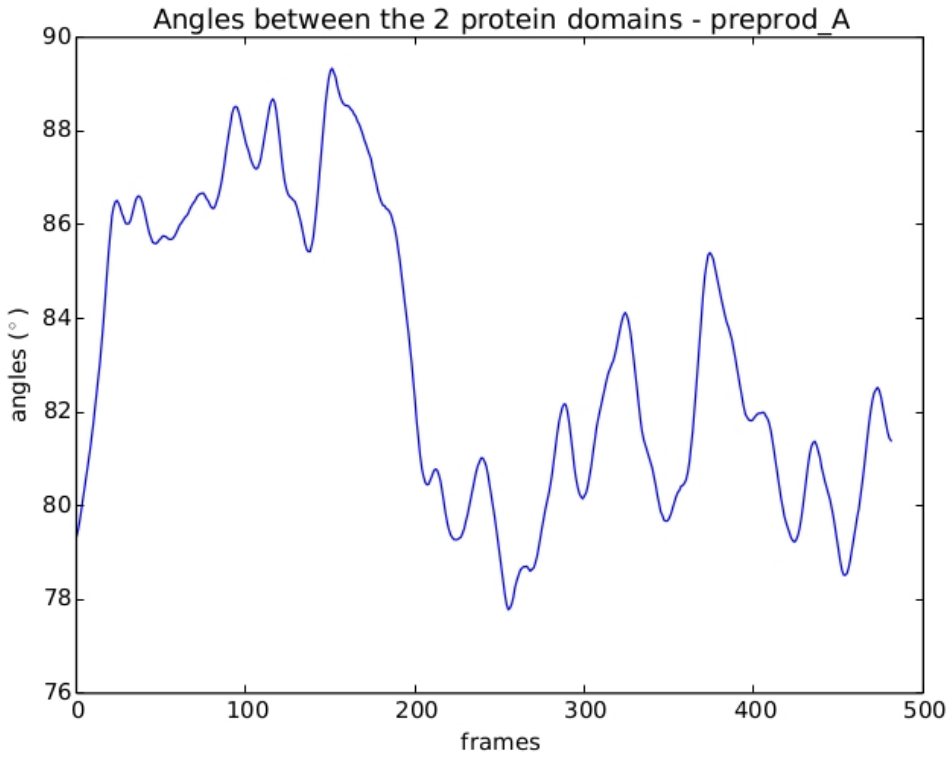

Figure 7: **Variation de l'angle entre S et P au cours de la trajectoire**. Ces résultats nous ont été fourni par notre superviseur et sont basés sur une méthode du "fit" de plans sur les domaines.

4.2 Importance du choix de S/P et limites de l'algorithme

L'approche développée ici n'est néanmoins pas exempte de défauts. En effet, nous avons du réduire le domaine S de 19 acides aminés (29-201 au lieu de 29-220) afin d'obtenir de tels résultats. En effet, l'approche par vecteurs propres assimile les directions de l'information à un plan. Ainsi, plus le domaine ressemble de manière inhérente à un plan, plus le résultat sera précis. Ici, les derniers résidus forment une "queue" reliée à la jointure avec le domaine P et sortant clairement du "plan" formé par le reste du domaine S.

En conséquence, le plan formé par S est différent, et impacte donc l'angle formé avec P. Si l'on considère les résidus 29-220, sans correction par MSF, les profils et écarts sont sensiblement similaires, mais les valeurs d'angles sont plus élevées (oscillant entre 106° et 120°, *cf. Figure 8*).

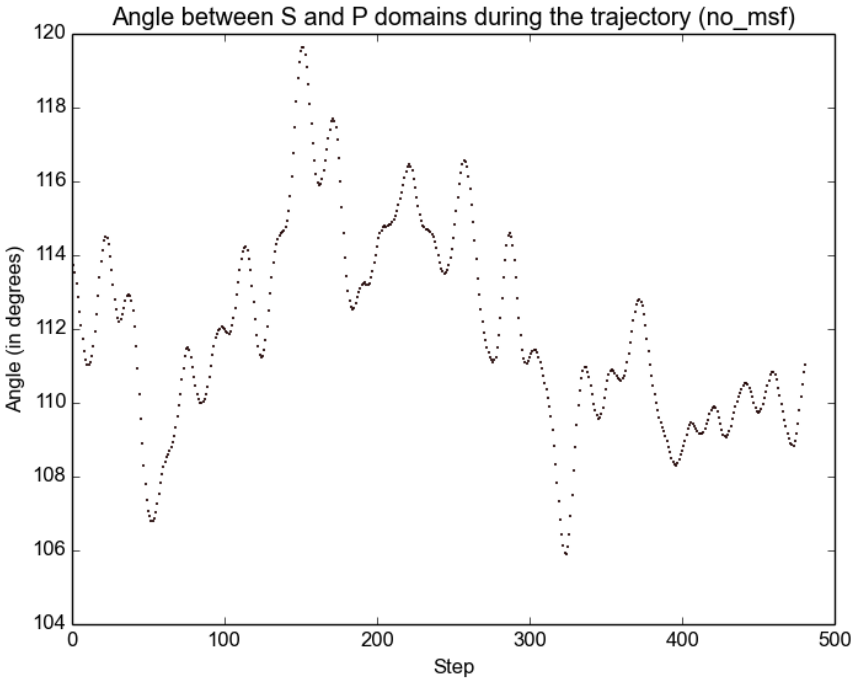

Figure 8: **Mesures plus élevées d'angles entre les domaines S et P au cours de la trajectoire.** Les valeurs d'angle calculées ici dont plus élevées car i) les résidus présentant de grandes variations de mouvements n'ont été supprimés et ii) les residus formant une queue sortant du plan du domaine S ont été conservés.

Un autre écueil de la méthode se situe dans la définition même des vecteurs propres, et peut être vu en analysant le graphe des angles quand l'on considère le domaine S entre 29-220 après correction par le MSF (*cf. Figure 9*). En effet, au cours de la dynamique, il est possible que le vecteur propre mesuré change au cours des frames. La protéine bougeant peu, les valeurs propres doivent rester proches au cours de la dynamique. Hors, si le domaine étudié n'a pas une forme

plane, alors l'algorithme a du mal à différencier le deuxième et troisième sens de l'information. En conséquence, les valeurs propres 2 et 3 peuvent s'intervertir selon les frames, décrivant alors deux repères formés par des vecteurs propres différents et entrainant la mesure de deux angles différents au cours d'un même protocole (*cf. Table 1*). Il aurait été possible de développer une validation pour vérifier la bonne succession de ces valeurs propres (et donc, en aval, des vecteurs propres). Cette observation mathématique montre un aspect très spécifique du programme: pour qu'il fonctionne, le domaine doit être relativement plan. Il est donc impossible d'utiliser notre algorithme pour analyser une protéine globulaire, qui ne présente pas à proprement parler de sens prédominant de son information.

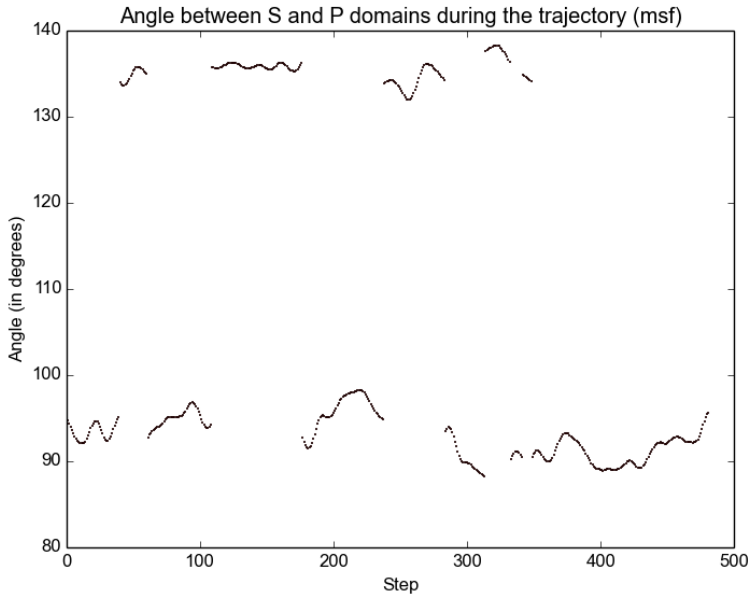

Figure 9: **Mesures plus élevées d'angles entre les domaines S et P au cours de la trajectoire.** Les valeurs d'angle calculées ici dont plus élevées car i) les résidus présentant de grandes variations de mouvements n'ont été supprimés et ii) les residus formant une queue sortant du plan du domaine S ont été conservés.

Frame	VP 1	VP 2	VP 3	Angle (°)
39	83.36	344.43	254.40	94.86
40	83.30	345.21	253.98	95.13
41	83.20	253.59	345.97	133.97
42	83.09	253.27	346.71	133.76

Table 1: **Calcul des valeurs propres au cours de la dynamique.** Les frames ont été choisies pour montrer les deuxièmes et troisièmes valeurs propres qui s'intervertissent et biaisent le calcul de l'angle

5 Conclusion

La capside du Norovirus (protéines VP1) est un élément essentiel nécessaire à l'infection de l'hôte. De récentes études montrent la flexibilité de deux domaines présents dans la protéine: le domaine S (*Shell*) et le domaine P (*Protruding*). La capside est constituée de protéines similaires en terme de séquences mais disposant de légères variations structurales dues à de très subtiles différences dans l'orientation des domaines S et P. L'idée de projet est de vérifier si une molécule A de 1IHM impliquée dans la capside peut tendre vers la conformation d'une molécule C au cours d'une dynamique moléculaire.

Au cours de cette étude, nous avons mis en place un programme facilement déployable visant à mesurer l'angle entre S et P au cours de la trajectoire. Les résultats obtenus sont cohérents et ont été validés par notre superviseur. Nous avons aussi mesuré la distance entre les deux domaines et étudié les variations du RMSD afin de s'assurer des mouvements effectués par la protéine. Il ne nous est pas possible d'affirmer avec certitude que le monomère A tend vers la conformation C, mais les résultats obtenus renforcent fortement l'idée que la protéine est flexible et qu'un changement d'état, même très léger, peut être envisageable.

6 Remerciements

Je remercie particulièrement M. Thibault Tubiana qui m'a accompagné dans le déroulement du projet et dans l'analyse des résultats obtenus.

References

[1] L. Lindesmith, C. Moe, S. Marionneau, N. Ruvoen, X. Jiang, L. Lindblad, P. Stewart, J. LePendu, and R. Baric, "Human susceptibility and resistance to Norwalk virus infection," *Nature Medicine*, vol. 9, pp. 548–553, May 2003.

[2] "ViralZone: Norovirus."

[3] S. G. Morillo and M. d. C. S. T. Timenetsky, "Norovirus: an overview," *Revista da Associação Médica Brasileira*, vol. 57, pp. 462–467, Aug. 2011.

[4] U. Katpally, N. R. Voss, T. Cavazza, S. Taube, J. R. Rubin, V. L. Young, J. Stuckey, V. K. Ward, H. W. Virgin, C. E. Wobus, and T. J. Smith, "High-Resolution Cryo-Electron Microscopy Structures of Murine Norovirus 1 and Rabbit Hemorrhagic Disease Virus Reveal Marked Flexibility in the Receptor Binding Domains," *Journal of Virology*, vol. 84, pp. 5836–5841, June 2010.

[5] G. S. Hansman, D. W. Taylor, J. S. McLellan, T. J. Smith, I. Georgiev, J. R. H. Tame, S.-Y. Park, M. Yamazaki, F. Gondaira, M. Miki, K. Katayama, K. Murata, and P. D. Kwong, "Structural Basis for Broad Detection of Genogroup II Noroviruses by a Monoclonal Antibody That Binds to a Site Occluded in the Viral Particle," *Journal of Virology*, vol. 86, pp. 3635–3646, Apr. 2012.