



HAL
open science

De l'unité d'assemblage à la capside : application in silico au norovirus et au virus de l'hépatite B

Jean-Charles Carvaillo

► **To cite this version:**

Jean-Charles Carvaillo. De l'unité d'assemblage à la capside : application in silico au norovirus et au virus de l'hépatite B. Modélisation et simulation. Université Paris-Saclay, 2021. Français. NNT : 2021UPASQ029 . tel-03455620

HAL Id: tel-03455620

<https://theses.hal.science/tel-03455620>

Submitted on 29 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

De l'unité d'assemblage à la capside : application
in silico au norovirus et au virus de l'hépatite B
*From assembly unit to capsid: in silico application to
norovirus and hepatitis B virus*

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 569, Innovation Thérapeutique : du Fondamental à
l'Appliqué (ITFA)

Spécialité de doctorat: Biochimie et Biologie Structurale

Unité de recherche : Université Paris-Saclay, CEA, CNRS, Institute for Integrative
Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France.

Référent : Faculté de pharmacie

**Thèse présentée et soutenue à Paris-Saclay,
le 25, juin, 2021, par**

Jean-Charles CARVAILLO

Composition du Jury

Sylvie NESSLER

Professeure,
Université Paris-Saclay

Présidente

Anja BÖCKMANN

Directrice de Recherche,
Université de Lyon

Rapporteuse & Examinatrice

Chantal PRÉVOST

Chargée de Recherche, HDR,
Université de Paris

Rapporteuse & Examinatrice

Jean COGNET

Professeur
Sorbonne Université

Examineur

Direction de la thèse

Stéphane BRESSANELLI

Directeur de Recherche,
Université Paris-Saclay

Directeur de thèse

Yves BOULARD

Chargé de Recherche,
Université Paris-Saclay

Co-Directeur de thèse

De l'unité d'assemblage à la capside : application *in silico* au norovirus et au virus de l'hépatite B

Jean-Charles Carvaillo

Rapporteur Dr. Anja Böckmann

*Molecular Microbiology and Structural Biochemistry (MMSB -
Université Claude Bernard Lyon 1, CNRS)*

Rapporteur Dr. Chantal Prevost

Laboratoire de Biochimie Théorique (LBT - Université de Paris, CNRS)

Membre du Jury Pr. Sylvie Nessler

*Institut de Biologie Intégrative de la Cellule (I2BC -
Université de Paris-Saclay, CNRS)*

Membre du Jury Pr. Jean Cognet

Laboratoire Jean Perrin (LJP - Sorbonne Université, CNRS)

Directeurs de Thèse Stéphane Bressanelli
CNRS

Yves Boulard
CEA

De l'unité d'assemblage à la capside : application in-silico au norovirus et au virus de l'hépatite B

Jean-Charles Carvaillo © 30 avril 2021

Thèse de doctorat en Sciences de la Vie

Rapporteurs : Dr. Anja Böckmann et Dr. Chantal Prévost

Directeur de thèse : Dr. Stéphane Bressanelli

Co-Directeur de thèse : Dr. Yves Boulard

Université Paris Sud

École doctorale Innovation Thérapeutique : du fondamental à l'appliqué

Équipe IMAPP : Interactions et Mécanismes d'Assemblage des Protéines et des Peptides

I2BC - CNRS UMR9198

1 avenue de la terrasse

91190 Gif-Sur-Yvette

REMERCIEMENTS

Je remercie tout d'abord mes directeurs de thèse, Stéphane Bressanelli et Yves Boulard, pour leur mentorat au cours de ces 4 dernières années. Je remercie également Thibault Tubiana pour m'avoir formé, guidé et conseillé tout au long de mon parcours. Merci à Sella Detchanamourtty et Fernando Luìs Barroso da Silva, pour leur contribution à l'étude d'assemblage de la capsid du Norovirus.

Je tiens aussi à remercier Pierre Chervy et Maelenn Chevreuil pour leur aide lors de la purification des protéines de capsid du VHB. Je remercie Valérie Geertsen pour les expériences d'ICP-MS et Benoit d'Autreaux, pour m'avoir aidé à titrer le zinc.

J'aimerais remercier tous les membres de l'équipe IMAPP pour leur bonne humeur et le partage de leurs connaissances. Merci à Kalouna Kra, Adrien Royet, Rémi Ruedas, Samira Agouda, Ordy Gnewou, Marion Schvartz, Sonia Fieulaine, Maïté Paternostre, Laura Pieri et Virginie Gervais. Merci également aux anciens membres de l'équipe : Cristina Petcut, Johan Habersetzer, Kaouther Ben Ouirane, Keinny François et Huy Nguyen Khac Minh.

Je remercie tout particulièrement Sarah Gibilaro pour m'avoir accompagné tout au long de cette aventure. Elle a été d'un soutien sans faille et m'a donc permis d'aborder sereinement cette étape de ma vie. Ses connaissances en français m'ont aussi été très utiles ! Je tiens également à remercier Wolfy Carvaillo, mon compagnon à 4 pattes et mon meilleur ami. Il m'accompagne depuis 11 ans déjà et continue de me supporter malgré lui... Pour bien longtemps encore, j'espère.

Merci également à Catherine Gallwa qui a effectué un énorme travail de mise en forme sur ce manuscrit et qui connaît désormais toutes mes figures par cœur. Elle a aussi su me motiver lorsque j'en ai eu besoin.

Je tiens à remercier ma famille, mon père Patrick Carvaillo et ma sœur Marie Bocquet, qui m'ont accompagné et m'ont toujours témoigné leur confiance. Je remercie mes grands-parents, Michelle et Jean Pastissier, pour m'avoir aidé de bien des manières.

Je remercie aussi Charlotte Périn pour les longues conversations scientifiques en terrasse ainsi que, Rodrigue Carvaillo et la famille Vautier.

Je tiens à remercier David Prager, enseignant de Mathématiques au Lycée André Paillot (LEGTA) de Saint-Genis-Laval, qui m'a posé mon tout premier problème algorithmique. Je remercie également Patrick Fuchs pour avoir rendu aussi intéressant la modélisation moléculaire. Ils m'ont conduit à en arriver là où je suis aujourd'hui.

Je remercie, enfin, l'Agence Nationale de Recherche sur le Sida et les hépatites virales pour le financement de mes recherches.



JEAN-CHARLES CARVAILLO

MODÉLISATEUR MOLÉCULAIRE

GITHUB

<https://github.com/jecarvaillo>

COMPÉTENCES

Langages Informatiques

- Maîtrise : Bash, C, Python3
- Bonnes connaissances : HTML, JAVA, MySQL, Perl, R
- Notions : JavaScript, PHP

Logiciels

- Maîtrise : Microsoft Office, Open Office

Systèmes d'exploitation

- Maîtrise : Linux, Mac, Windows

Anglais

Niveau professionnel et scientifique

PUBLICATIONS

- Premier auteur de l'article : "[Linking Bisphenol S to Adverse Outcome Pathways Using a Combined Text Mining and Systems Biology Approach](#)"
- Co-auteur de l'article : "[TTClust: A Versatile Molecular Simulation Trajectory Clustering Program with Graphical Summaries](#)"
- Co-auteur de l'article : "[Deciphering adverse outcome pathway network linked to Bisphenol F using text mining and systems toxicology approaches](#)"

CONTACT

Adresse : 106 boulevard de Charonne,
75020 Paris
Téléphone : +33 6 30 82 93 26
Mail : carvaillojeancharles@gmail.com

EXPÉRIENCE PROFESSIONNELLE

Doctorat en Bioinformatique

CNRS - Institut de Biologie Intégrative de la cellule | Mars 2018 - Juin 2021

De l'unité d'assemblage à la capsid : applications *in-silico* au norovirus et au virus de l'hépatite B.

Tuteur et Enseignant vacataire

Université Paris Diderot | Septembre 2015 - Juin 2021

Outils pour la bureautique et internet, Python, statistiques.

CDD Ingénieur Biologiste en minage de texte

INSERM - Toxicologie, Pharmacologie et Signalisation Cellulaire | Septembre 2017 - Février 2018

Projet HBM4EU Biosurveillance de l'Homme en Europe – WP 13.1.

Stage de fin de Master

CNRS - Institut de Biologie Intégrative de la cellule | Janvier - Juin 2017

Simulations gros grains (CG) d'intermédiaires d'assemblage de la capsid de Norovirus.

Stage de Master 1

INSERM - UMR-S 973 - Molécule Thérapeutiques *in-silico* | Mars - Juin 2016

Compréhension des effets pathogènes induits par des mutations faux sens de la O-linked N-acetylglucosamine (O-GlcNAc) transférase humaine.

EDUCATION

Doctorat Bioinformatique, Modélisation Moléculaire - Spécialité Biochimie et Biologie Structurale

Université Paris-Sud - École doctorale Innovation thérapeutique du fondamental à l'appliqué
Janvier 2018 - Juin 2021

Master Sciences, Technologies, Santé - Mention Biologie informatique / Bioinformatique

Université Paris Diderot
Septembre 2015 - Juin 2017

Licence Sciences, Technologies, Santé - Mention Sciences du Vivant

Université Paris Diderot
Septembre 2013 - Juin 2015

DUT Génie Biologique - Option Bioinformatique

Université d'Auvergne IUT Biologie - Antenne d'Aurillac
Septembre 2011 - Juin 2013

COMMUNICATIONS SCIENTIFIQUES

Articles

- Tubiana T, Carvaillo J-C, Boulard Y, Bressanelli S. TTClust: A Versatile Molecular Simulation Trajectory Clustering Program with Graphical Summaries. *J Chem Inf Model*. 2018;58: 2178–2182. doi:10.1021/acs.jcim.8b00512
- Carvaillo Jean-Charles, Barouki Robert, Coumoul Xavier, Audouze Karine. Linking Bisphenol S to Adverse Outcome Pathways Using a Combined Text Mining and Systems Biology Approach. *Environmental Health Perspectives*. 2019;127: 047005. doi:10.1289/EHP4200
- Rugard M, Coumoul X, Carvaillo J-C, Barouki R, Audouze K. Deciphering adverse outcome pathway network linked to Bisphenol F using text mining and systems toxicology approaches. *Toxicol Sci*. 2019. doi:10.1093/toxsci/kfz214

Présentations orales

- **Atelier de Modélisation des Molécules d'Intérêt Biologique de Paris-Saclay (AMMIB)** - mai 2018 à Evry-Courcouronnes
- **Réunion annuelle AC42 - Réseau national hépatites (France. REcherche. Nord & sud. Sida-hiv. Hépatites FRENH)** - février 2019 à Paris
- **21^e congrès du Groupement de Graphisme et de Modélisation Moléculaire (GGMM)** - avril 2019 à Nice
- **Atelier de Modélisation des Molécules d'Intérêt Biologique de Paris-Saclay (AMMIB)** - avril 2019 à Gif-Sur-Yvettes
- **Journées Des Doctorants Joliot (CEA)** - juin 2019 à Saclay
- **1^{ère} réunion du GdR MéDynA** - octobre 2019 à Sainte-Montaine
- **Réunion annuelle AC42 - Réseau national hépatites (France. REcherche. Nord & sud. Sida-hiv. Hépatites FRENH)** - février 2020 à Paris
- **Séminaire au Laboratoire de Biochimie Théorique** - février 2021 à Paris

Posters

- **20^e congrès du Groupement de Graphisme et de Modélisation Moléculaire (GGMM)** - mai 2017 à Reims
- **Journée de l'Ecole Doctorale 569 (Innovation thérapeutiques du Fondamental à l'appliqué)** - juin 2019 à Châtenay-Malabry

LISTE DES ABRÉVIATIONS

ARD :	Régions ou domaines riches en arginine
CAM :	Modulateurs allostériques de capsid
CEA :	Commissariat à l'énergie atomique et aux énergies alternatives
CG :	Gros grains
Cp149 :	Protéine de capsid du VHB recombinante tronquée comportant le résidu 1 à 149
Cp183 A80K :	Protéine de capsid du VHB complète mutée par une lysine sur l'alanine 80.
Cp183 E77K :	Protéine de capsid du VHB complète mutée par une lysine sur l'aspartate 77.
CTD :	Domaine C-terminal de la protéine Core
HBGA :	Antigènes du groupe histo-sanguin de l'épithélium muqueux du tractus gastro-intestinal
HOD :	Hexamère de dimères de la capsid du Norovirus
ICP-MS :	Spectrométrie de masse couplée à un plasma inductif
IDR :	Région intrinsèquement désordonnée
MD :	Simulation de dynamique moléculaire
NTD :	Domaine N-terminal de la protéine Core
NSD :	Divergence spatiale normalisée, mesure quantitative de similarité entre deux ensembles de points tridimensionnels
PAR :	4-(2-pyridylazo) resorcinol, chromophore permettant de titrer le zinc
PDB ID :	Identifiant d'une structure issu d'une base de données protéique
POD :	Pentamère de dimères de la capsid du Norovirus
REMD :	Simulations de dynamique moléculaire avec échanges de répliques
RMSD :	Écart quadratique moyen, mesure quantitative de similarité entre deux ensembles de points tridimensionnels à la même résolution
RMSF :	Fluctuation quadratique moyenne
SASA :	Surface accessible au solvant
TER :	Derniers résidus du bras C-terminal
TMD :	Simulations de dynamique moléculaire dirigée
TR-SAXS :	Diffusion des rayons X aux petits angles en temps résolu
VHB et HBV :	Virus de l'hépatite B
XTAL :	Structure cristallographique

TABLE DES FIGURES

Figure 1. Constituants essentiels d'un virus représenté sous forme de schéma.	16	POD.	89
Figure 2. Schéma simplifié du cycle viral.	17	Figure 31. Comparaison du docking d'un dimère A-B ou C-C sur le POD-D.	90
Figure 3. Icosaèdre et symétrie icosaédrique.	19	Figure 32. Comparaison du docking d'un dimère A-B ou C-C sur le POD-D ₂ .	91
Figure 4. Représentation à plat d'un capsid e icosaédrique.	20	Figure 33. Comparaison du docking d'un dimère A-B ou C-C sur le POD-D ₃ .	92
Figure 5. Détermination du nombre de triangulation.	21	Figure 34. RMSD des dimères A-B et C-C.	94
Figure 6. Exemples d'antiviraux.	24	Figure 35. Exemple d'analyse de clustering réalisée avec TTClust, sur les trajectoires des dimères A-B et C-C concaténées.	95
Figure 7. Représentation schématique de la cinétique d'assemblage d'une capsid e.	26	Figure 36. Évolution du RMSD entre les dimères dans une conformation vers une autre conformation.	97
Figure 8. Représentation schématique de l'assemblage d'une capsid e.	27	Figure 37. Dynamique des structures secondaires du dimère VP1.	98
Figure 9. Simulation basée sur un système à particules, modèles de sous-unités en interaction, nommées nanoparticules.	28	Figure 38. Représentation de l'orientation du domaine S par rapport au domaine P pour les trois conformations de VP1.	99
Figure 10. Simulation basée sur un système à particules, modèle d'une sous-unité en triangle.	28	Figure 39. Évolution de l'Incurvation entre les domaines P et S au cours de la TMD (k = 0,01).	100
Figure 11. Simulation basée sur un système à particules 1.	29	Figure 40. Évolution de l'Incurvation entre les domaines P et S au cours de la TMD (k = 0,05).	101
Figure 12. Simulation basée sur un système à particules 2.	30	Figure 41. Évolution de l'Incurvation entre les domaines P et S au cours de la TMD (k = 0,10).	102
Figure 13. Simulation basée sur un système à particules 3.	31	Figure 42. Schéma simplifié et nouvelle hypothèse d'assemblage de la capsid e du Norovirus	105
Figure 14. Description structurale de la capsid e du HIV-1. Clichés au cours de la simulation gros grains (CG) de l'assemblage de la capsid e du HIV-1	33	Figure 43. Nanoindentation in-silico de la capsid e du Norovirus.	106
Figure 15. Structure cristallographique de la protéine VP1 du virus de Norwalk (norovirus GI.1).	38	Figure 44. Assemblage de la capsid e du norovirus en conditions d'équilibres.	107
Figure 16. Conformations de la protéine VP1.	39	Figure 45. Assemblage de la capsid e en conditions d'assemblage.	108
Figure 17. Organisation de la capsid e du Norovirus.	40	Figure 46. Rupture de symétrie du POD au cours des 20 µs.	109
Figure 18. Modèle d'assemblage de la capsid e du virus de Norwalk (GI) de Prasad et al.	41	Figure 47. Solutions de docking des dimères A-B et C-C sur le POD D ou le POD-D ₂ 2 3.	110
Figure 19. Modèle de nucléation-croissance de la capsid e du Norovirus.	41	Figure 48. Procédure générale de docking flexible.	111
Figure 20. Hypothèse d'assemblage de la capsid e du Norovirus de Prasad et al. et intermédiaires observés durant les expériences de TR-SAXS de Tresset et al.	42	Figure 49. Structures des dimères dissociés du GI.1, générées par modélisation de corps rigides,- à partir des données SAXS.	112
Figure 21 - Protocole de simulation de dynamique moléculaire gros grains.	44	Figure 50. MD d'assemblage d'une capsid e virale dodécaédrique autour d'un polymère chargé.	113
Figure 22. Stabilité de la structure globale, alignement sur les domaines S centraux.	45	Figure 51. Prévalence de HBsAg (%).	117
Figure 23. Stabilité des dimères dans le contexte des capsomères.	46	Figure 52. Intervention de la protéine Core tout au long du cycle viral.	118
Figure 24. Évolution des interfaces d'amarrage du POD.	48	Figure 53. Structure de la protéine Core du génotype D.	119
Figure 25. POD-D ₃ issus de notre stratégie.	83	Figure 54. Organisation de la capsid e de HBV.	120
Figure 26. Docking d'un dimère sur le POD-D ₃ 2-3-4 cristallographique ou après 10 µs de simulation.	85	Figure 55. Croissance théorique de la capsid e du VHB.	121
Figure 27. Docking d'un dimère sur le POD-D ₃ 2-3-8.	86	Figure 56. Assemblage du NTD suivi par TR-SAXS.	122
Figure 28. Intermédiaires d'assemblage compatibles avec notre stratégie.	87	Figure 57. Résumé de l'attribution des signaux de localisations du CTD de la protéine Core.	123
Figure 29. Résumé des chemins d'assemblage obtenus à partir de notre stratégie, du POD vers le POD-D ₃ hypothétique.	88	Figure 58. Hypothèses d'expositions des CTDs en dehors des pores.	124
Figure 30. Comparaison du docking d'un dimère A-B ou C-C sur le			

Figure 59. Consensus de la prédiction de structure secondaire et composition en acides aminés de Core.	127	la conformation cible du pore 3.	160
Figure 60. Composition en acides aminés des domaines de Core.	128	Figure 87 . Exposition du CTD replié par le pore 3.	161
Figure 61. Déviation (RMSD) et fluctuation quadratique moyenne (des résidus : RMSF) du NTD tronqué (Cp140).	131	Figure 88. Évolution du RMSD entre la conformation au temps t (ps) et la conformation cible du pore 3.	162
Figure 62. Analyse des modes normaux selon la méthode des blocs rigides non-linéaires (NOLB).	132	Figure 89. Exposition du CTD par le pore q3.	163
Figure 63. Modélisation du CTD et dynamique de la protéine Core.	133	Figure 90. Évolution du RMSD entre la conformation au temps t (ps) et la conformation cible du pore q3.	163
Figure 64. Modélisation par homologie avec MODELLER du CTD du modèle de Core.	134	Figure 91. Exposition du CTD allongé et replié par le pore 3.	164
Figure 65. Modélisation par homologie avec Protein Model Portal du CTD du modèle de Core.	134	Figure 92. Évolution du RMSD entre la conformation au temps t (ps) et la conformation cible du pore q3.	165
Figure 66. Première modélisation par enfilage avec I-TASSER du CTD du modèle de Core.	134	Figure 93. Localisation des résidus potentiellement impliqués dans la chélation du Zn ²⁺ .	167
Figure 67. Deuxième modélisation par enfilage avec I-TASSER du CTD du modèle de Core.	135	Figure 94. Représentation en weblogo de l'alignement multiple des 12 294 séquences de Core.	168
Figure 68. Première modélisation combinant l'homologie et l'enfilage avec ROSETTA du CTD du modèle de Core.	135	Figure 95. Résultats de TEM et DLS.	170
Figure 69. Deuxième modélisation combinant l'homologie et l'enfilage avec ROSETTA du CTD du modèle de Core.	135	Figure 96. Gammes d'étalonnages de la BSA et mesures d'absorbance des échantillons.	177
Figure 70. MD avec échange de réplique (REMD) du modèle MODELLER sur UNRES.	136	Figure 97. Spectres de l'absorbance du PAR complexé avec le Zn ²⁺ contenu dans l'échantillon Cp149.	181
Figure 71. MD avec échange de réplique (REMD) du modèle MODELLER sur UNRES.	136	Figure 98. Spectres et gammes d'étalonnages du PAR complexé ou non au Zn ²⁺ provenant du ZnSO ₄ .	183
Figure 72. Exemples de repliements du CTD après MD reliés aux propriétés électrostatiques de Core.	137	Figure 99. Spectres de l'absorbance du PAR complexé ou non au Zn ²⁺ contenu dans les échantillons.	184
Figure 73. Distribution des contacts intra- et inter-domaines.	138	Figure 100. Modèle de chélation du zinc par Core selon le champ de force ZAFF.	190
Figure 74. Résumé de la structuration du CTD. contacts intra-domaine.	141	Figure 101. Localisation des points de contacts des ions zincs.	193
Figure 75. Résumé de la structuration du CTD. contacts inter-domaines.	142	Figure 102. Surface accessible au solvant (SASA) des résidus impliqués dans la chélation des ions zincs.	194
Figure 76. Modèles de Core utilisés.	144	Figure 103. Localisation des résidus chargés négativement sur le NTD de Core.	199
Figure 77. Comparaison des RMSD de 2 systèmes simulés selon les champs de force Amber 99SB-ILDN et DISP.	145	Figure 104. Composition en acides aminés de l'intérieur du pore 3.	200
Figure 78. Comparaison de l'évolution des structures secondaires de 2 systèmes simulés selon les champs de force Amber 99SB-ILDN et -DISP.	146	Figure 105. Composition en acides aminés de l'intérieur du pore q3.	200
Figure 79. Test de l'exposition des CTD par les pores à partir des dimères libres simulés.	153	Figure 106. Dimensions des pores 3 et q3.	201
Figure 80. Passage de la capsid du VHB tout-atomes à gros grains MARTINI.	155	Figure 107. Exposition du bras C-terminal allongé et non rattaché aux pores 3 et q3.	201
Figure 81. Dynamique de la capsid du VHB.	156	Figure 108. Exposition des domaines riches en arginines (ARD).	202
Figure 82. Conformations initiales des pores.	157	Figure 109. Hypothèse du mécanisme d'exposition du CTD.	204
Figure 83. Exposition du CTD par le pore 3.	158	Figure 110. Localisation des acides aminés potentiellement impliqués dans la chélation de Zn ²⁺ .	205
Figure 84. Évolution du RMSD entre la conformation au temps t (ps) et la conformation cible du pore 3	159	Figure 111. Sites de chélation putatifs du zinc.	206
Figure 85. Exposition du CTD allongé par le pore 3.	160	Figure 112. Autres sites de chélation putatifs de Core.	207
Figure 86. Évolution du RMSD entre la conformation au temps t (ps) et		Figure 113. Paysages énergétiques d'une protéine repliée et d'une protéine intrinsèquement désordonnée.	210
		Figure 114. Interaction de la nucléoporine avec l'importine β.	211
		Figure 115. Démonstration schématique des mécanismes de sélection conformationnelle et de liaison par ajustement induit.	212

TABLE DES TABLEAUX

Tableau 1. Fréquence des structures secondaires de la protéine VP1.	98	Tableau 18. Concentrations de la protéine Cp149 obtenues selon la méthode de Bradford à la 1ère campagne	177
Tableau 2 - Données structurales expérimentales de la protéine Core du VHB. .	130	Tableau 19. Concentrations de la protéine Cp149 et cp183 E77K obtenues selon la méthode de Bradford à la 2 ^{ème} campagne.	178
Tableau 3. Contacts inter-domaines sur les 25 dernières ns des 12 dynamiques.	139	Tableau 20. Concentrations de la protéine Cp149 obtenues selon la méthode BCA à la 1ère campagne.	178
Tableau 4 –Contacts intra-domaine du CTD sur les 25 dernières ns des 12 dynamiques.	140	Tableau 21. Concentrations de la protéine Cp149 et cp183 E77K obtenues selon la méthode de BCA à la 2 ^{ème} campagne.	179
Tableau 5 –Contacts natifs inter-domaines sur les 25 dernières ns des 4 dynamiques.	147	Tableau 22. Concentrations massiques et molaires de Cp149 et Cp183 déduites selon différentes méthodes.	180
Tableau 6. Contacts natifs inter-domaines sur les 25 dernières ns des 4 dynamiques.	148	Tableau 23. Mesures d'absorbance du PAR complexé et concentrations de zinc déduites.	181
Tableau 7. Contacts natifs inter-domaines sur les 25 dernières ns des 4 dynamiques.	149	Tableau 24. Quantification du Zn ²⁺ par ICP-MS.	182
Tableau 8. Contacts natifs inter-domaines sur les 25 dernières ns des 4 dynamiques.	150	Tableau 25. Mesures d'absorbance du PAR complexé et concentrations calculées du zinc à la 1ère campagne.	185
Tableau 9. Ponts-salins entre les résidus du pore 3 et le CTD rattaché, lors de son exposition.	159	Tableau 26. Mesures d'absorbance du PAR complexé et concentrations calculées du zinc à la 2 ^{ème} campagne.	185
Tableau 10. Ponts-salins entre les résidus du pore 3 et le CTD allongé, lors de son exposition.	161	Tableau 27. Quantification du zinc à la 1ère et 2ème campagne de Cp149 et Cp183 E77K par ICP-MS.	186
Tableau 11. Ponts-salins entre les résidus du pore 3 et le CTD replié, lors de son exposition.	162	Tableau 28. Ratio du zinc en fonction de la concentration de Core tronquée (Cp149).	187
Tableau 12. Ponts-salins entre les résidus du pore q3 et le CTD rattaché, lors de son exposition.	164	Tableau 29. Ratio du zinc en fonction de la concentration de Core mutée (Cp183).	188
Tableau 13. Ponts-salins entre les résidus du pore q3 et le CTD déplié ou replié, lors de son exposition.	165	Tableau 30. Interaction du dimère tronqué (Cp149) avec le Zn ²⁺ .	191
Tableau 14. Mesures d'absorbances et concentrations déduites pour la protéine Cp149 lors de la 1 ^{ère} campagne.	173	Tableau 31. Interaction du dimère complet (Cp183) avec le Zn ²⁺ .	192
Tableau 15. Mesures d'absorbances et concentrations déduites pour la protéine Cp149 lors de la 2 ^{ème} campagne.	174	Tableau 32. Ponts-salins entre le pore 3 et les domaines riches en arginines du CTD	203
Tableau 16. Mesures d'absorbances et concentrations déduites pour la protéine Cp183 E77K lors de la 2 ^{ème} campagne.	175	Tableau 33. Ponts-salins entre le pore q3 et les domaines riches en arginines du CTD.	203
Tableau 17. Principes, avantages et inconvénients des méthodes biochimiques utilisées pour doser les protéines de capsides.	176		

TABLE DES MATIÈRES

Introduction générale	15
1 Organisation minimale d'un virus	15
2 Organisation des capsides icosaédriques	19
2.1 <i>Le nombre de triangulation T et la théorie de quasi-équivalence</i>	20
3 Les protéines de capsides – cible de choix contre la lutte antivirale	23
4 Étudier la dynamique et l'auto-assemblage des capsides <i>in-silico</i>	25
4.1 <i>Modèles mathématiques</i>	25
4.2 <i>Simulations basées sur des systèmes à particules</i>	27
4.3 <i>Simulations de dynamique moléculaire gros grains</i>	32
4.4 <i>Simulations de dynamique moléculaire tout-atomes</i>	33
Résultats sur le norovirus	35
5 Dynamique d'assemblage de la capsid du norovirus	37
5.1 <i>Contexte biologique</i>	37
5.2 <i>Contexte structural</i>	37
5.3 <i>État de l'art sur l'assemblage de la capsid du norovirus <i>vide</i></i>	40
5.4 <i>Croissance de l'intermédiaire d'assemblage du norovirus de Norwalk</i>	43
5.4.1 <i>Modélisation de l'unité asymétrique</i>	43
5.4.2 <i>Modélisation des capsomères</i>	43
5.4.3 <i>Préparation des systèmes et simulations</i>	44
5.4.4 <i>Étude de la dynamique des capsomères</i>	45
5.4.5 <i>Rupture de symétrie chez le POD</i>	47
5.4.6 <i>Étude de la formation de l'intermédiaire d'assemblage</i>	50
5.4.7 <i>Complément à l'article</i>	85
5.4.8 <i>Influence du dimère docké A-B vs C-C sur l'assemblage</i>	89
5.4.9 <i>Espace conformationnel du dimère VP1</i>	92
5.5 <i>Conclusion</i>	103
6 Discussion	105
Résultats sur le virus de l'hépatite B	115
7 Dynamique de l'unité d'assemblage de la capsid du VHB et interaction avec le zinc	117
7.1 <i>Contexte biologique</i>	117

7.2	<i>Le cycle de réplication virale, Core une protéine essentielle</i>	117
7.3	<i>L'unité d'assemblage : Core</i>	119
7.4	<i>L'assemblage de la capsid du VHB</i>	119
7.5	<i>Signaux de localisation nucléaire et exposition du CTD</i>	123
7.6	<i>Analyse structurale de Core</i>	127
7.6.1	Structures primaires et secondaires de Core	127
7.6.2	Structures tertiaires de Core	129
7.7	<i>Le domaine C-terminal (CTD) de la protéine Core – un véritable couteau suisse</i>	131
7.7.1	Dynamique du domaine d'assemblage	131
7.7.2	Modélisation du CTD	133
7.7.3	Dynamique du CTD dans le contexte du dimère seul	135
7.7.4	Influence des champs de force sur le repliement du CTD	143
7.7.5	Dynamique du CTD dans le contexte de la capsid	153
8	Motif d'interaction de Core avec le matériel génétique	167
8.1	<i>Hypothèse du motif de liaison à l'ADN faisant intervenir du zinc</i>	167
8.2	<i>Chélation du Zn²⁺ par Core in-vitro</i>	169
8.2.1	Production de la protéine Core Cp149 et Cp183	169
8.2.2	Vérification de l'intégrité des capsides	169
8.2.3	Quantification de la protéine Core – un processus pas si évident	171
8.2.4	Quantification du zinc en interaction avec Core	180
8.3	<i>Conclusion sur la quantification du zinc en contact avec les protéines de capsid</i>	187
8.4	<i>Chélation du Zn²⁺ par Core in-silico</i>	189
8.4.1	Modélisation des sites de chélation de Core	189
8.4.2	Interaction de Core avec le zinc libre	190
8.4.3	Accessibilité des sites putatifs	194
8.5	<i>Conclusion</i>	197
9	Discussion	199
10	Perspectives	209
	Conclusion générale	213
	Bibliographie	215
	Annexes	223

INTRODUCTION GÉNÉRALE

Cette introduction définit les notions importantes sur les virus et les éléments communs qui les caractérisent. Le complexe macromoléculaire qui protège l'information virale, coque ou capside et les techniques *in-silico* pour étudier son assemblage, seront abordés. Ces sujets sont présentés car les protéines sur lesquelles les travaux de thèse ont été menés sont des protéines de capside. La protéine de capside du norovirus sera introduite dans un chapitre qui lui est dédié. Une problématique d'assemblage liée aux protéines de capside du norovirus sera abordée en parallèle. La protéine de capside du virus de l'hépatite B (VHB) sera également introduite et son étude suivra dans un autre chapitre.

1 ORGANISATION MINIMALE D'UN VIRUS

Les virus sont définis comme des parasites intracellulaires obligatoires, c'est-à-dire qu'ils ont besoin des cellules d'un hôte pour se répliquer. Ils peuvent être considérés comme des entités dont le but est de multiplier l'information qui est contenue dans leur matériel génétique. Ils le produisent à partir de leur propre machinerie ou en détournant celle de l'hôte à son insu et qui est nécessaire à leur survie.

L'origine des virus reste une grande question non résolue. En effet, les virus ne possédant pas de ribosomes, ils ne peuvent être classés dans la trichotomie, archées, bactéries et eucaryotes, établie selon les travaux de Carl Woese. La classification de Baltimore, faite pour les virus, se base sur le type d'acide nucléique et sur leur mode d'expression. Cette classification ne distingue pas clairement certains virus, c'est le cas du virus de l'hépatite B dont le génome viral est partiellement rétro-transcrit lors de son cycle viral. On peut donc préférer la classification établie par le Comité International de Taxonomie des Virus (ICTV [1]). La classification des virus de l'ICTV se fait selon différents niveaux de hiérarchisation : ordres, familles, genres, espèces, sous-espèces, sérotypes, souches et isolats. Un certain nombre de critères, comme le type de génome, la présence ou non d'enveloppe, les propriétés de symétrie de la capside, la nature de l'hôte et la présence d'enzymes particulières, sont utilisés pour construire la taxonomie des virus.

Les virus sont omniprésents dans le vivant et peuvent donc infectés autant l'Homme que d'autres animaux, plantes ou bactéries. Ils se composent (1) d'une chaîne d'acide nucléique simple ou double brins (ADN ou ARN) constituant le support de l'information virale (Figure 1); (2) d'une capside ou coque protéique protégeant le matériel génétique viral face à l'environnement extérieur et à l'organisme de l'hôte. Certains virus peuvent comprendre (3) une enveloppe lipidique supplémentaire qui est héritée des cellules de l'hôte et dont certaines molécules, la composant, facilitent la propagation vers d'autres cellules (Figure 1).

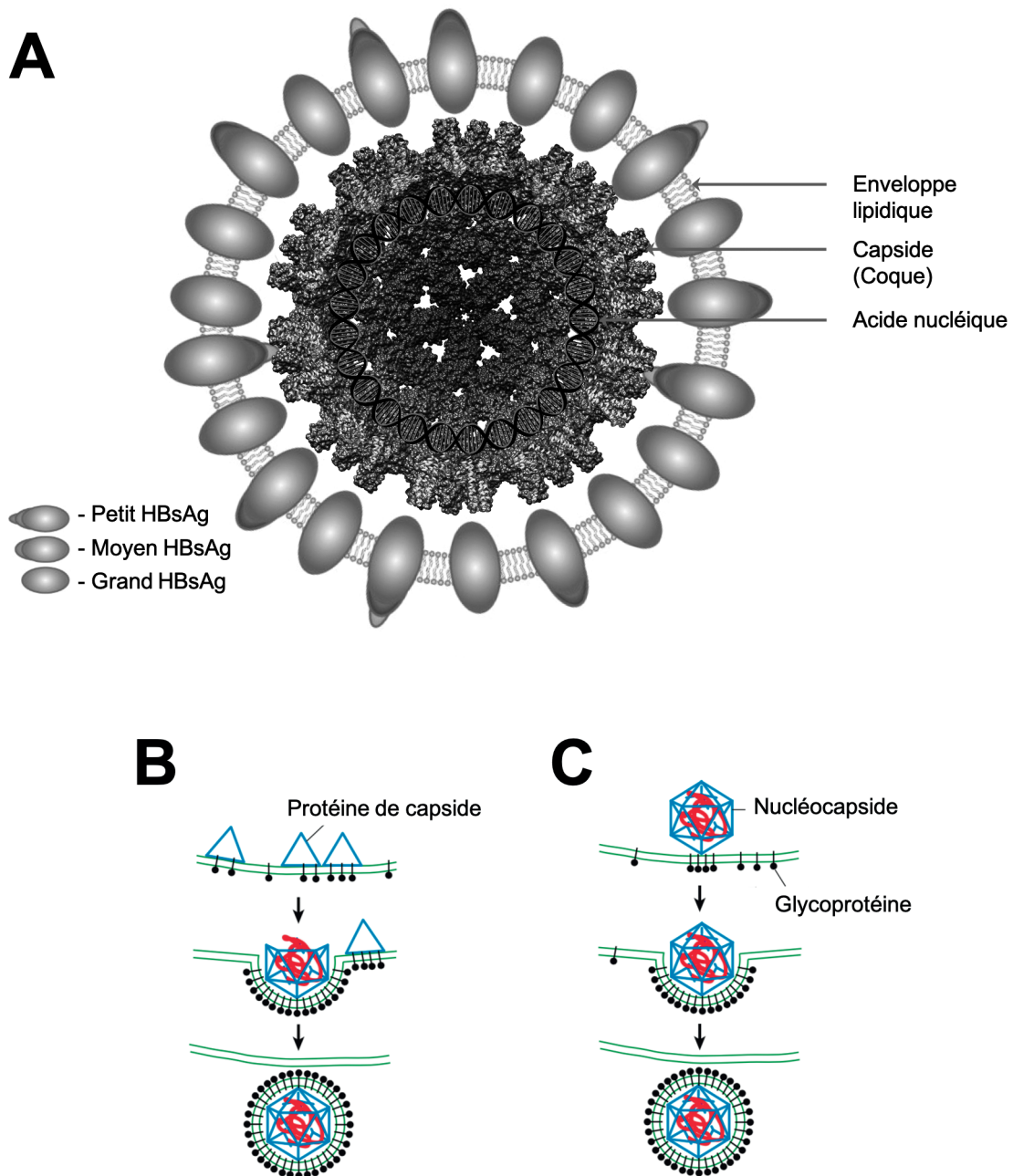


Figure 1 – Constituants essentiels d'un virus représenté sous forme de schéma. (A) Tous les virus sont au moins composés d'un acide nucléique encapsulé dans une coque. L'enveloppe lipidique ne recouvre pas toutes les familles de virus. Mécanismes d'exportation des virus (B, C). (B) La capsid se forme en contact avec la membrane plasmique de la cellule hépatique infectée. (C) La capsid est déjà formée et vient en contact avec la membrane plasmique pour être enveloppée.

Les virus infectieux requièrent l'empaquetage du génome viral dans les capsides qui peuvent être enveloppées de façon simultanée [2,3]. Le coronavirus 2 du syndrome respiratoire aigu sévère (SARS-CoV-2), lui-même, est caractérisé par un acide ribonucléique simple brin positif (ARN +), contenu dans une capsid enveloppée. En résumé, le cycle viral de la majorité des virus consiste à pirater la machinerie cellulaire de la cellule hôte. Une fois entré dans la cellule, le virus prolifère et est libéré pour parasiter d'autres cellules ou organismes cellulaires (Figure 2).

La protéine de nucléocapside de ce virus joue, comme pour les autres virus, un rôle clé dans la réplication virale, son rôle principal étant l'empaquetage du génome viral [4].

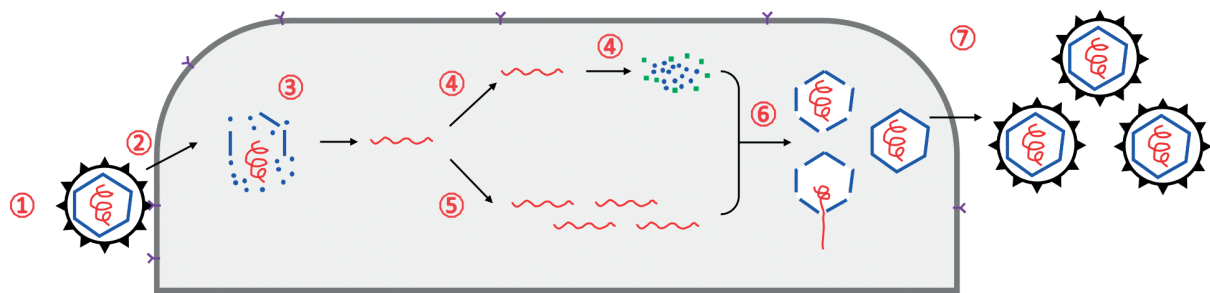


Figure 2 – Schéma simplifié du cycle viral : 1. Attachement de la capsidite ou de l'enveloppe ; 2. Pénétration 3. Décapsidation ; 4. Synthèses des protéines structurales et non structurales à partir du matériel génétique viral ; 5. Réplication du matériel génétique viral ; 6. Encapsidation ; 7. Libération des nucléocapsides.

Il existe deux grands types de structures de capsidite : (i) les capsidites de forme allongées qui sont caractérisées par une géométrie hélicoïdale, c'est le cas du virus de la mosaïque du tabac (TMV) ; et (ii) les capsidites de forme sphérique qui sont définies comme des solides de Platon. Un solide de Platon est un polyèdre régulier et convexe. Ces polyèdres peuvent être divisés en trois groupes de symétrie : (T) le groupe tétraédrique, (O) le groupe (cubique) / octaédrique et (I) le groupe (dodécaédrique) / icosaédrique [5]. Ces derniers sont contraints par les propriétés géométriques que confère un polyèdre régulier convexe. Ils composent environ la moitié des capsidites des virus sur Terre. Le groupe icosaédrique comprend le VHB et le virus de la gastroentérite virale (Norovirus). L'architecture icosaédrique nous intéresse tout particulièrement car les études menées dans ce travail sont en lien avec les protéines de capsidite du VHB et du Norovirus. Il est donc important de définir avec précision les caractéristiques d'une géométrie icosaédrique.

2 ORGANISATION DES CAPSIDES ICOSAÉDRIQUES

À la différence des capsides de géométrie hélicoïdale, qui peuvent s'accommoder à n'importe quelle longueur de chaîne polynucléotidique, les capsides icosaédriques sont limitées par leur géométrie [2,6]. Elles se composent nécessairement de 60 sous-unités identiques. Elles s'organisent en 20 faces (triangles équilatéraux), 30 arêtes, 12 sommets et possèdent 3 axes de symétrie. Un axe de symétrie 5 qui se situe au sommet, un axe de symétrie 2 sur les arêtes et un axe de symétrie 3, sur le centre de chaque face (Figure 3).

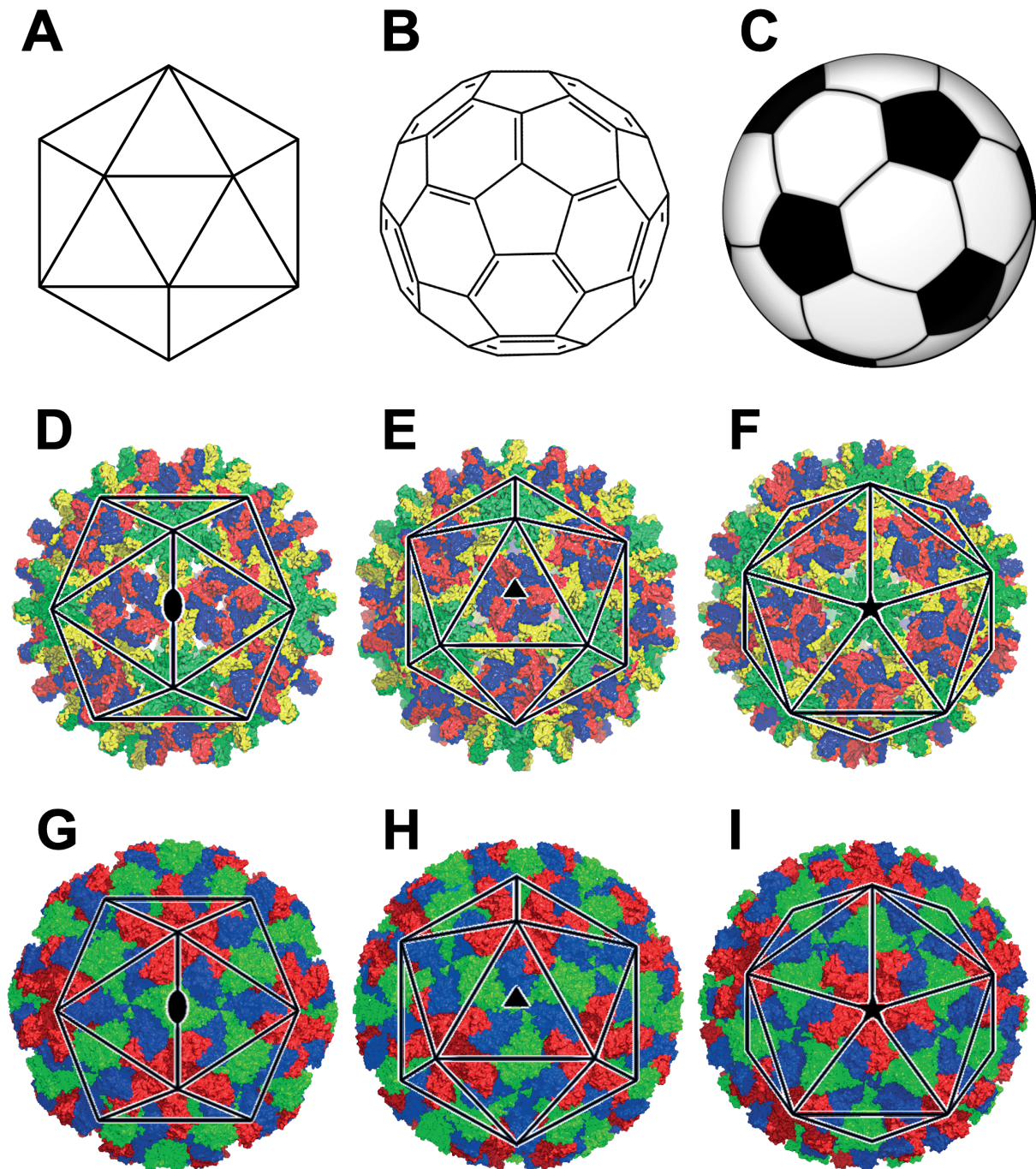


Figure 3 – Icosaèdre et symétrie icosaédrique. (A, B, C) Représentation d'icosaèdres. (A) Solide de Platon. (B) solide d'Archimède (icosaèdre tronquée). (C) Ballon de foot, icosaèdre tronquée. (D, E, F) Représentation de la capside du VHB. (G, H, I) Représentation de la capside du Norovirus. (D, G) Symétrie d'axe 2. (E, H) Symétrie d'axe 3. (F, I) Symétrie d'axe 5.

Contrairement au solide de Platon (Figure 3A) formé uniquement de polygones réguliers convexes isométriques (ex. triangles équilatéraux) le solide d'Archimède (Figure 3B) est un polyèdre convexe semi-régulier. Ce dernier est composé par 2 sortes de polygones réguliers (ex. pentamères et hexamères). Les capsides icosaédrique de virus se rapprochent davantage des icosaèdres tronqués (solide d'Archimède). En effet, une capside peut être décrite par un ensemble d'hexagones, organisés en un réseau. Mais les sommets de chaque face d'une capside doivent obligatoirement comporter un pentagone. Les pentagones sont nécessaires pour induire la courbure de cet ensemble afin d'obtenir une géométrie sphérique.

2.1 Le nombre de triangulation T et la théorie de quasi-équivalence

“The basic assumption is that shell is held together by the same type of bonds throughout, but that these bonds may be deformed in slightly different ways in the different, non-symmetry related environments” [6].

Une capside comporte nécessairement 60 sous-unités identiques. Si la taille de la capside et le nombre de sous-unités requises pour former la capside deviennent plus importantes, l'assemblage totalement symétrique de la capside devient impossible à réaliser. Caspar et Klug proposent alors une théorie de quasi-équivalence dans laquelle les sous-unités interagissent de façon quasi-équivalente. À noter qu'un icosaèdre, peu importe sa taille, est projetable sur une surface qui comporte 20 faces (Figure 4).

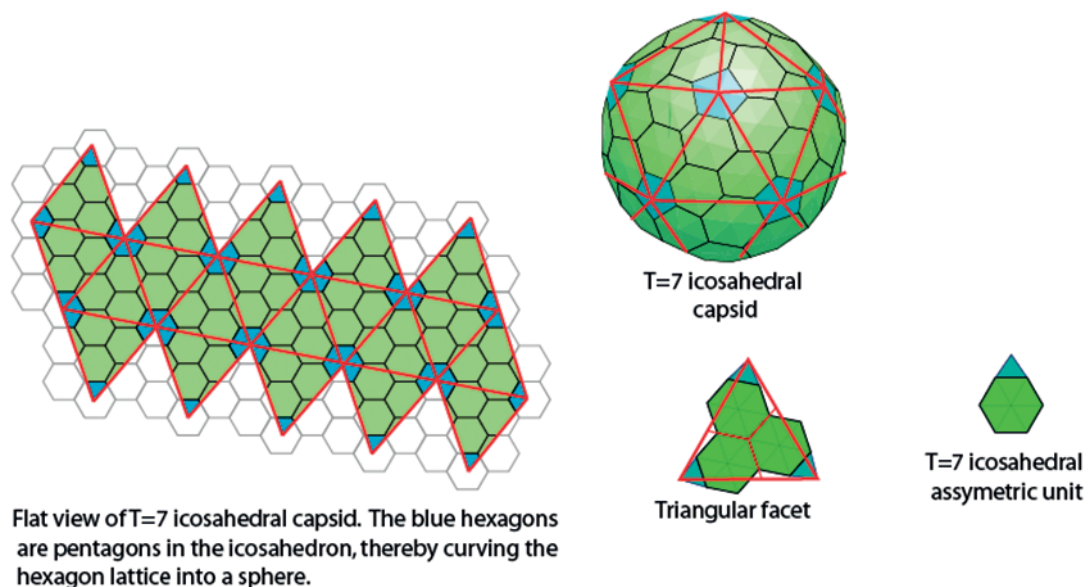


Figure 4 – Représentation à plat d’une capsid icosaédrique, issue de <https://viralzone.expasy.org/8577>. Représentation en haut d’une capsid T=7 sous forme d’une quasi-sphère. Représentation à gauche de cette même capsid en plan. Chacune des 20 faces comporte des hexamères et des pentamères sur ses sommets. Représentation à droite d’une unité asymétrique comportant tous les conformères de la même protéine de capsid. Ici 7 conformations différentes.

Dans une capsidre tronquée (solide d'Archimède), il faut donc nécessairement que les unités (monomères ou conformations) des sous-unités d'assemblage soient légèrement différentes pour instaurer la quasi-équivalence. Elle est à l'origine de la formation de ce type de capsidre. Si l'on assemble la capsidre seulement à partir d'hexagones, il est impossible d'obtenir la courbure de la capsidre. Pour obtenir une sphère, il faut donc transformer exactement 12 hexagones en pentagones. L'organisation d'une capsidre correspond, lorsqu'il y a plus de 60 sous-unités, à un ensemble d'hexamères et de pentamères. Les 12 pentamères doivent être situés aux sommets de chacune des faces. Une capsidre icosaédrique comporte $60T$ sous-unités, où T correspond au nombre de triangulation. Il correspond aussi au nombre de conformations distinctes dans une même sous-unité. Le nombre de triangulation T est défini mathématiquement comme le carré de la longueur de chaque arête des faces. Pour retrouver T , il suffit d'appliquer la formule suivante : $T=h^2+hk+k^2$ (Figure 5). h est le nombre d'hexamères qu'il faut traverser dans une première direction pour atteindre le prochain pentamère, k représente le nombre d'hexamères qu'il faut parcourir dans une autre direction pour atteindre ce même pentamère (<https://viralzone.expasy.org/8577>) [7].

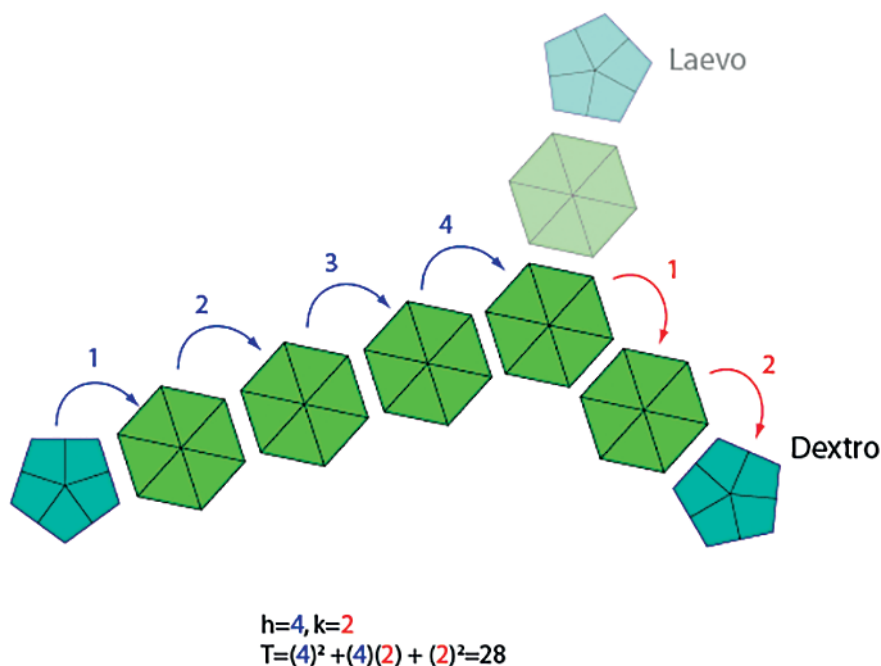


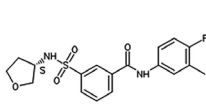
Figure 5 – Détermination du nombre de triangulation, issue de <https://viralzone.expasy.org/8577>.
 Détermination du nombre de triangulation T avec les paramètres h et k . h correspond au nombre de pas, à partir d'un pentamère, dans une même direction pour atteindre un nouveau pentamère. k correspond au nombre de pas dans une autre direction pour atteindre ce même nouveau pentamère.

3 LES PROTÉINES DE CAPSIDES – CIBLE DE CHOIX CONTRE LA LUTTE ANTIVIRALE

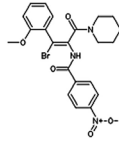
Lorsque les virus persistent, ou si un vaccin n'a pas été développé, il est nécessaire d'agir sur les mécanismes clés de prolifération des virus (Figure 2). Pour cela, il est possible de cibler plusieurs de ces mécanismes pour freiner ou inhiber totalement une ou des étapes du cycle de réplication virale. Les antiviraux sont donc souvent utilisés simultanément pour lutter contre les virus difficiles, voire impossible à totalement éliminer. On parle alors de "cocktail antiviral". Par exemple, l'information génétique virale du HIV-1 ou des hépatites virales B et C, contenue dans les cellules d'un hôte infecté, est très persistante. Les antiviraux, agissant en amont ou en aval de la production de matériel génétique, sont donc utilisés. C'est le cas des antirétroviraux, analogues de nucléosides et nucléotides ou d'antiviraux ciblant la coque des virus. Les antiviraux, tels que les interférons α ou les nucléotides, ont un effet mineur ou nul sur la forme persistante du matériel génétique du VHB (ADNccc). Il existe des molécules qui interagissent plutôt avec les protéines de capsidite du VHB, ce sont les modulateurs allostériques de protéines de capsidite (Core Allosteric Modulators, abrégé CAM). Les CAM (Figure 6A) sont de petites molécules qui vont venir perturber l'auto-assemblage des nucléocapsides. Ils provoquent la formation de protéines qui s'agrègent et conduit à la production de capsides aberrantes (CAM-A) ou provoquent la formation de capsides vides (CAM-B). Ils préviennent l'encapsulation du génome viral. Il est donc possible, chez le VHB, de potentialiser davantage une guérison fonctionnelle en combinant plusieurs antiviraux dont les CAM [8]. Des études antivirales ont été menées sur le génogroupe II du norovirus et ont montré que l'acide citrique [2.1.3] perturbe l'assemblage des capsides (Figure 6B) [9]. Les oligosaccharides du lait humain interagissent directement avec les protéines de capsidite [10]. Ces deux molécules préviennent l'interaction des nucléocapsides avec les antigènes du groupe histo-sanguin (HBGA) qui tapissent l'épithélium muqueux du tractus gastro-intestinal [11].

A**Capsid Assembly Modulators (CAMs)**

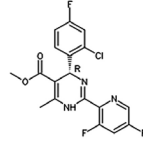
JNJ-632



AT130



BAY41-4109

**Nucleoside Analogue (NA)**

Entecavir (ETV)

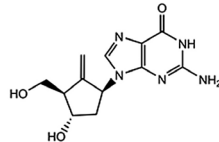
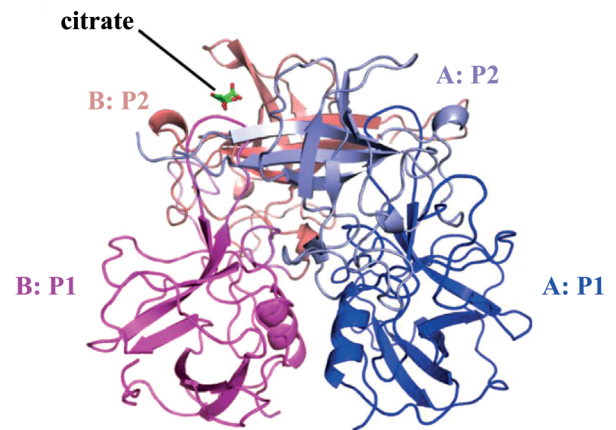
**B**

Figure 6 – Exemples d’antiviraux tirés des articles [8,9]. **A.** Exemples de modulateurs allostériques de protéines de capsides et d’analogues de nucléoside. Le JNJ-632 est un sulfamoylbenzamide (CAM-B), l’AT130 est un dérivé de phenylpropenamides (CAM-B) et enfin le BAY41-4109 est un heteroarylpyrimidine (CAM-A). **B.** Structure cristallographique du complexe entre le domaine P du norovirus du génogroupe II et le citrate. Le monomère A est coloré en bleu et le monomère B en magenta. Les sous-domaines P2 sont colorés dans des teintes plus claires.

Les protéines de capside protègent et transportent le génome viral. Elles constituent des cibles idéales et communes de la lutte antivirale. Les mécanismes d’action qui caractérisent les CAM-A et CAM-B pourraient être utilisés à plus grande échelle. Il y aurait donc une application directe sur la formation de l’ensemble des capsides virales. D’où l’intérêt de comprendre comment l’assemblage des capsides s’effectue.

4 ÉTUDIER LA DYNAMIQUE ET L'AUTO-ASSEMBLAGE DES CAPSIDES *IN-SILICO*

L'auto-assemblage des capsides est un processus au cours duquel des protéines de capsidie adoptent une organisation quaternaire indépendamment de l'intervention d'une source extérieure. Ce processus est régi par des interactions non-covalentes : liaisons hydrogènes, interactions hydrophobes, forces de van der Waals, π -stacking et électrostatiques. Dans le cas des nucléocapsides, les sous-unités et le matériel génétique vont spontanément former des particules virales. Cet assemblage correspond à leur état de plus faible énergie [6]. Les méthodes *in-silico* et les mathématiques sont des outils utiles pour construire des modèles plausibles de structure de capsidie. À titre d'exemple, un modèle de nucléocapsidie du coronavirus suivant une géométrie d'octaèdre tronqué, a été proposé à partir de données de cryo-microscopie électronique (CryoEM) et de contraintes mathématiques propres aux particules quasi-sphériques [5]. Ces outils sont également adaptés pour étudier et définir des modèles d'auto-assemblage de capsidie vide et de nucléocapsidie.

Des modèles théoriques et des méthodes biophysiques sont utilisés pour étudier les mécanismes d'assemblage des capsides virales, comme : la résonance magnétique nucléaire du solide [12], la diffusion de rayons X aux petits angles résolue en temps [13], la chromatographie d'exclusion stérique [14] ou encore, la diffusion dynamique de la lumière [14]. Pourtant, aucune de ces méthodes expérimentales détecte l'ensemble des intermédiaires d'assemblage, de la brique d'assemblage à la capsidie complète. Analyser l'ensemble des chemins empruntés et des produits au cours de l'assemblage est difficile à mettre en œuvre. Ces processus se produisent sur des gammes de longueurs et des échelles de temps très larges (de l'ångström au micromètre et de la picoseconde à la minute). De plus, la plupart d'entre eux sont transitoires [15]. Pour surmonter ce problème, des modèles théoriques indépendants, tant à l'échelle nanométrique qu'à l'échelle temporelle, ont été élaborés.

4.1 Modèles mathématiques

L'assemblage de la capsidie de virus peut être établi selon des modèles mathématiques qui décrivent des cinétiques de polymérisation. De manière générale, les modèles mathématiques qui décrivent l'assemblage de polymères suivent les cinétiques de deux étapes cruciales. (1) Une première étape de nucléation au cours de laquelle s'initie l'assemblage de n unités libres. C'est une étape dite lente. (2) Une deuxième étape d'élongation qui se décrit comme un empilement d'unités libres, les unes à la suite des autres, sur une extrémité ou sur une autre de n unités complexées. C'est un phénomène, dit, rapide [16]. Pendant l'élongation, un équilibre s'instaure entre les unités d'assemblage libres et les unités d'assemblage polymérisées. Quand l'équilibre est atteint, les vitesses d'association et de dissociation deviennent égales. Il existe une multitude de mécanismes de nucléations lentes (exemples de mécanismes chez l'actine ; Matsudaira et al., 1987).

Le modèle de polymérisation de la capsidite suit les mêmes cinétiques mais l'amorçage de la nucléation est différent. Pour la polymérisation d'un filament, cela correspond au temps nécessaire pour former la première association entre deux unités libres. Pour une capsidite, il correspond au temps pour former un intermédiaire d'assemblage considéré comme le noyau critique d'assemblage. Une fois ces intermédiaires formés, les capsidites sont capables de croître (Figures 7 et 8).

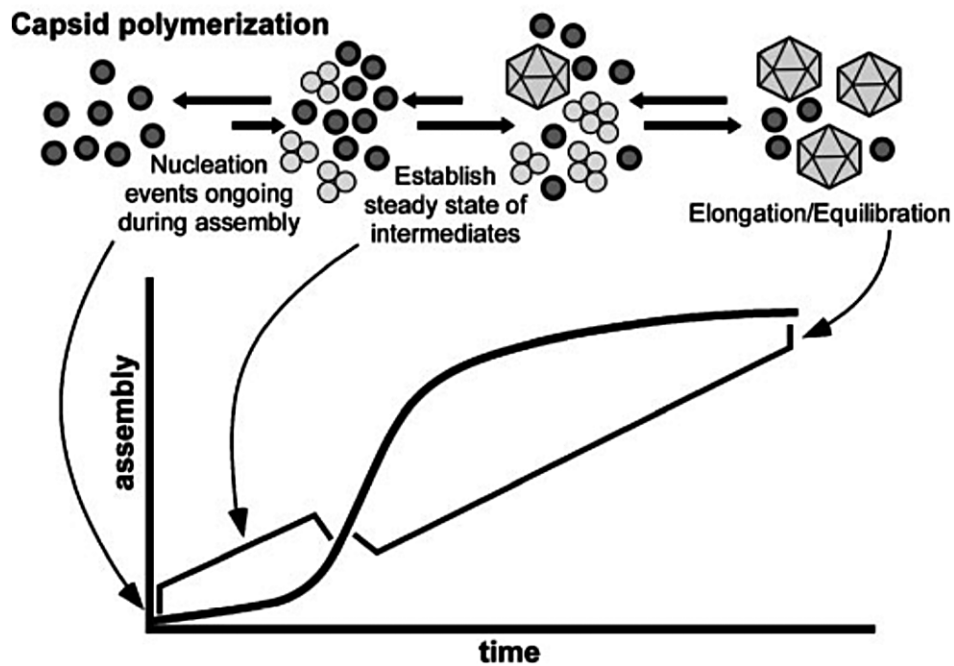


Figure 7 – Représentation schématique de la cinétique d'assemblage d'une capsidite, issue de l'article [16]. Un noyau critique d'assemblage est requis pour chaque capsidite. Les noyaux se forment de façon continue au cours de cette réaction. L'assemblage d'une capsidite suit une étape de nucléation des unités d'assemblage, une étape au cours de laquelle des intermédiaires se forment et enfin, une étape d'élongation/équilibration. À l'issue de la dernière étape, coexiste des capsidites et des unités d'assemblage.

Les modèles mathématiques nous donnent une idée de la vitesse et des réactions d'assemblage mais en ne se limitant qu'à quelques acteurs : l'unité d'assemblage, les intermédiaires d'assemblage, la capsidite complète.

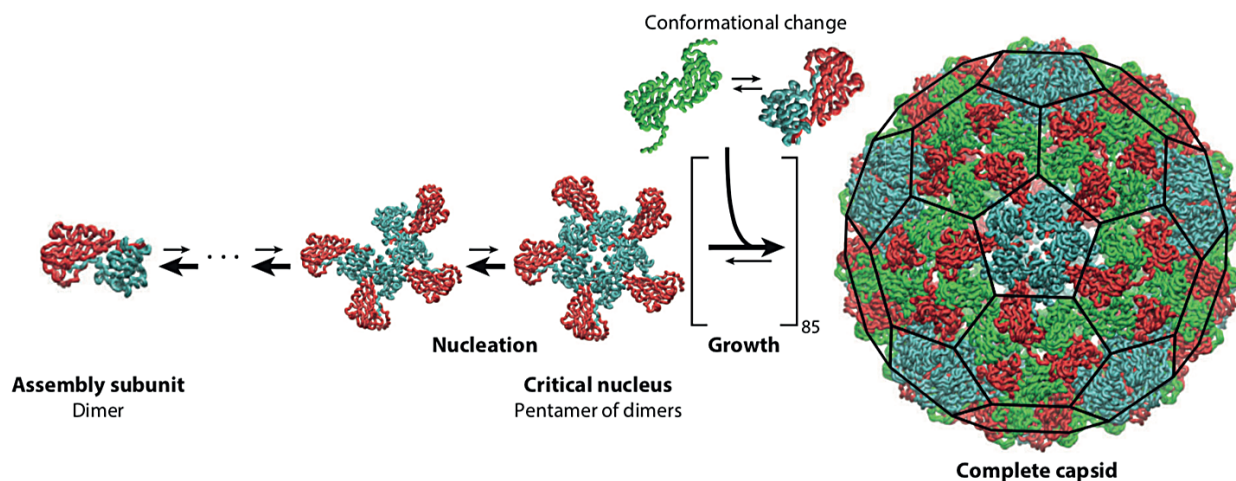


Figure 8 – Représentation schématique de l'assemblage d'une capsid, issue de l'article [2]. De l'unité d'assemblage jusqu'à la capsid complète. Les unités d'assemblage forment le noyau critique, suivi d'une étape de croissance conduisant à la formation d'une capsid complète.

Ils constituent un bon point de départ pour comprendre les réactions d'assemblage in-vitro et in-vivo. Ils ne tiennent cependant pas compte des descripteurs physico-chimiques importants qui les caractérisent. De plus, les capsides et leurs intermédiaires sont traités de façon rigide à l'aide de tels modèles.

La simulation de dynamique moléculaire (MD) modélise l'évolution d'un système à particule au cours du temps. Elle est utilisée pour analyser les mouvements physiques des atomes et des molécules. Elle donne une partie de la dynamique du système étudié [18] (p. 353).

4.2 Simulations fondées sur des systèmes à particules

Les méthodes expérimentales actuelles, pour étudier l'auto-assemblage des capsides, ne sont pas encore à même de déterminer l'ensemble de leur dynamique. La vitesse d'acquisition des données structurales ou autres (acquisition d'enveloppes, facteurs de formes...) est dans une échelle de temps supérieure aux phénomènes d'assemblage, mis en œuvre lors de la croissance de certains intermédiaires. Les simulations basées sur des systèmes à particules et les méthodes calculatoires, en général, sont un bon moyen pour palier à la problématique d'échelle de temps. Les modèles d'états Markovien (MSMs) sont l'une des méthodes computationnelles adaptées à l'étude de processus intramoléculaires. [19]. C'est une approche qui repose sur des modèles mathématiques mais qui intègre la notion de mouvement d'objets, définis comme des particules. Le modèle Markovien modélise les changements d'états d'un système, pas à pas, selon une fonction de transition probabiliste. Ces transitions conduisent potentiellement à un changement d'état et à l'évolution du système. Dans les travaux de Perkett et Hagan, deux modèles sont utilisés : un modèle sphérique modélisant une sous-unité entière (Figure 9) et un modèle plus précis de sous-unité de géométrie triangulaire (Figure 10).

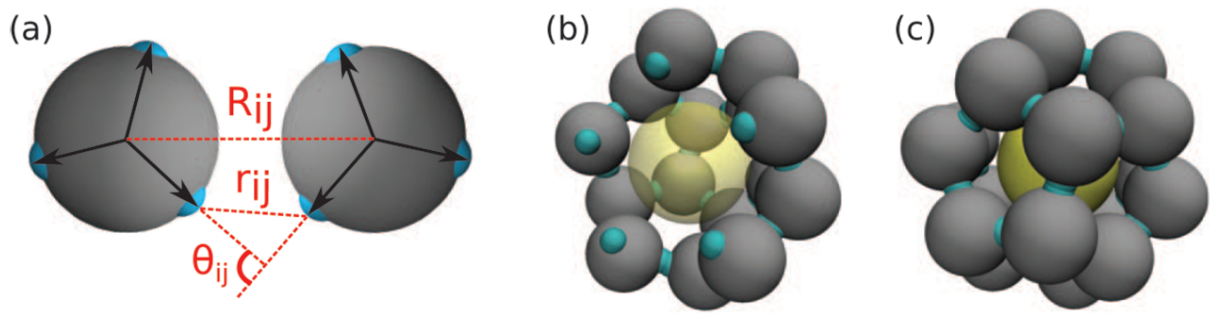


Figure 9 – Simulation basée sur un système à particules, issue de l'article [19]. Modèles de sous-unités en interaction, nommées nanoparticules. (a) Modèle de deux sous-unités en interaction, dont les liaisons sont modélisées par des vecteurs (flèches) et représentées par des protubérances en cyan. L'angle entre deux vecteurs correspond à un angle de 180° . (b) Vue coupée d'une capsid complète composée de nanoparticules. (c) Représentation d'une capsid complète contenant 20 sous-unités respectant la symétrie icosaédrique.

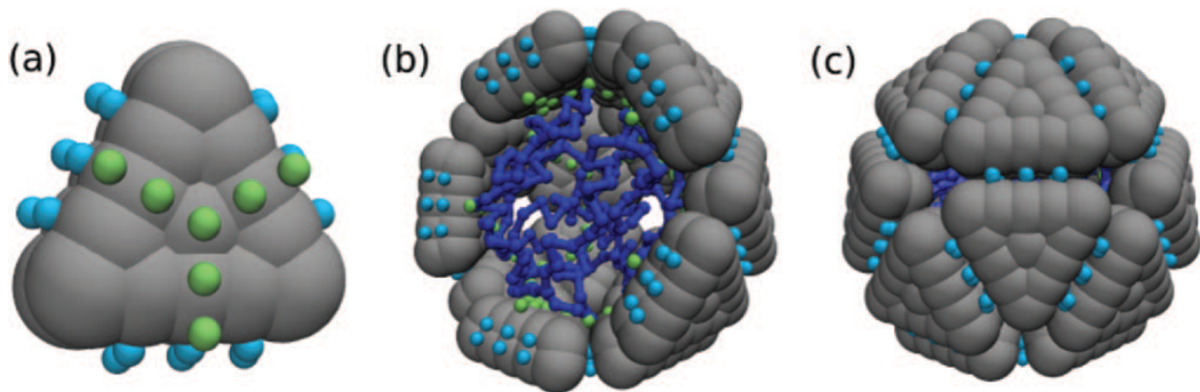


Figure 10 – Simulation basée sur un système à particules, issue de l'article [19]. Modèle d'une sous-unité en triangle. (a) Représentation d'une sous-unité, en gris, comportant des attracteurs en vert et cyan. Les billes vertes sont des attracteurs d'acides nucléiques et les billes cyan des attracteurs de sous-unités. (b) Vue coupée d'une capsid complète comportant un acide nucléique en son centre, coloré en bleu marine. (c) Représentation d'une capsid complète formée de 20 sous-unités.

Ces modèles sont simulés à l'aide de simulation de dynamiques browniennes selon une méthode probabiliste (Figure 11). Il s'agit de la méthode de Monte-Carlo.

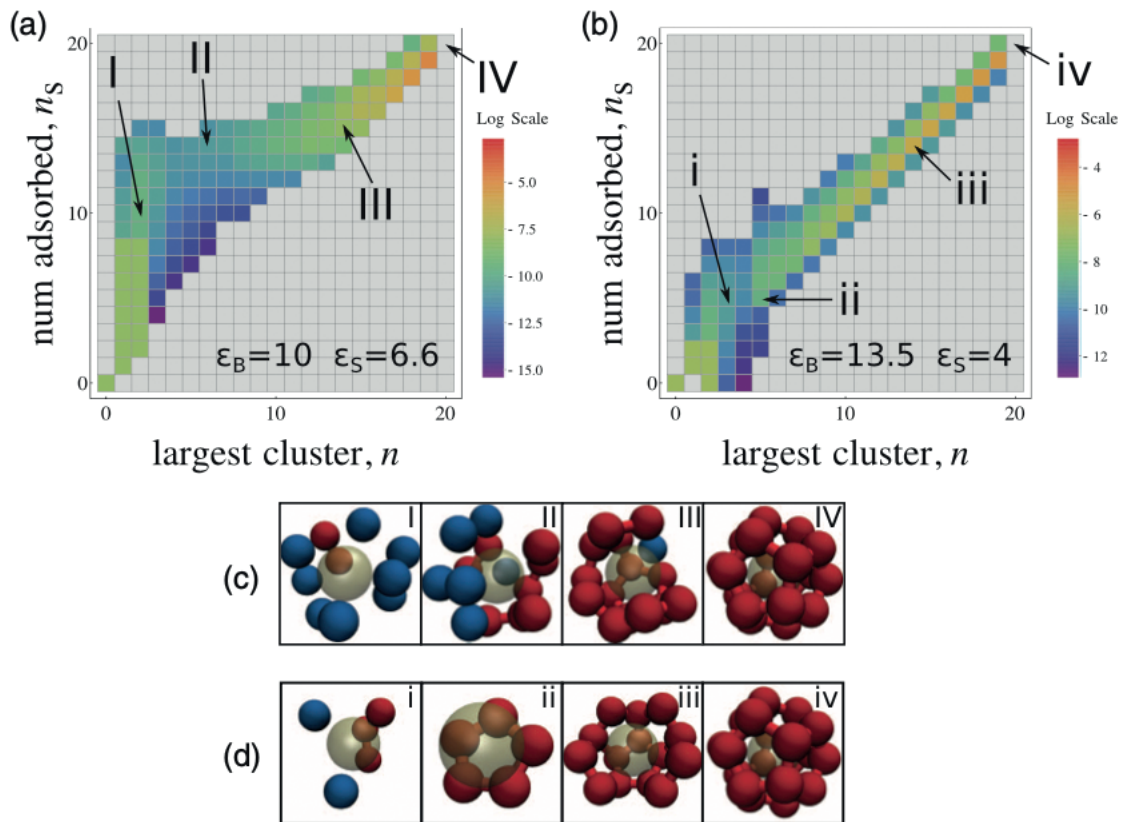


Figure 11 – Simulation basée sur un système à particules, issue de l'article [19]. (a) et (b) Suivi des changements d'états (flux) suivant les modèles d'états Markovien mis en place par Perkett et Hagan et selon les paramètres qu'ils ont définis ϵ_B et ϵ_S . Le flux est coloré selon une échelle logarithmique. (c) et (d) Représentation des chemins d'assemblage selon les paramètres en (a) et (b). Les sous-unités en interaction autour d'une nanoparticule sont en rouge et les sous-unités libres, attirées par la nanoparticule en bleu. La nanoparticule, en question est représentée de façon translucide au centre.

Dans d'autres études, le couplage d'un algorithme de diffusion-réaction et de la méthode de Monte-Carlo simule la diffusion et l'interaction des particules correspondant aux protéines de capsid, entre-elles, sans avoir à modéliser de sous-unités [19]. Les contacts quasi-équivalents au sein de la capsid T=3 sont modélisés par 4 interfaces de contact sur les particules (Figure 12). L'effet de la quasi-équivalence des capsides est aussi étudié en réalisant des simulations de dynamiques moléculaires (MD) sur des particules [20]. 3 particules, dont leur géométrie trapézoïdale est quasi-identique, ont été utilisées pour modéliser les sous-unités et mimer l'effet de la quasi-équivalence nécessaire pour former une capsid de plus de 60 sous-unités.

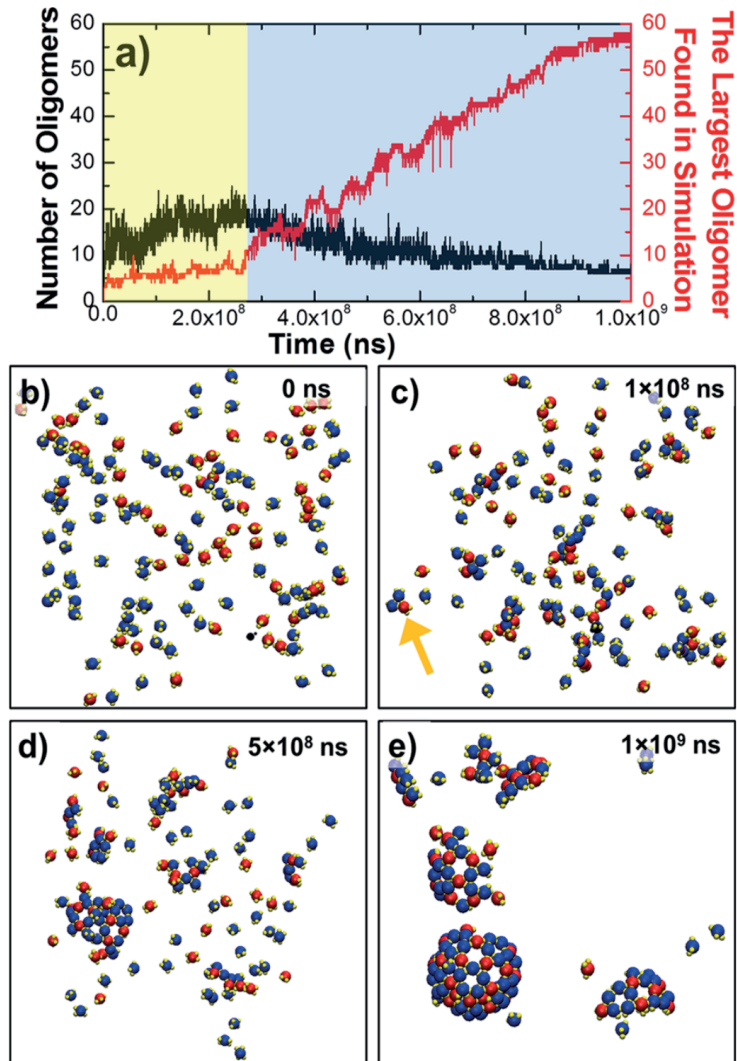


Figure 12 – Simulation basée sur un système à particules, issue de l'article [20]. Simulation de l'assemblage de la capside du bactériophage MS2 selon un modèle rigide. (a) Cinétique d'assemblage de la capside. Le nombre total d'oligomères formés au cours du temps est représenté par la courbe noire et l'oligomère le plus grand, par la courbe rouge. Deux étapes sont suggérées dans l'assemblage de la capside. L'étape de nucléation correspond à la partie de gauche, en jaune. L'étape de croissance correspond à la partie de droite, en bleu clair. (b) La simulation commence à partir d'une configuration initiale du système. (c et d) Clichés de l'assemblage de la capside au cours de la simulation. (e) Configuration du système en fin de simulation. Une capside, en bas à droite, est formée par plus de 50 dimères (a).

L'étude de Rapaport correspond à la modélisation de la dynamique d'assemblage d'une capsid T=3. Une force d'attraction couplée à un paramètre ajustable est utilisée pour simuler l'interaction des sous-unités et, dans la finalité, l'assemblage de la capsid complète (Figure 13).

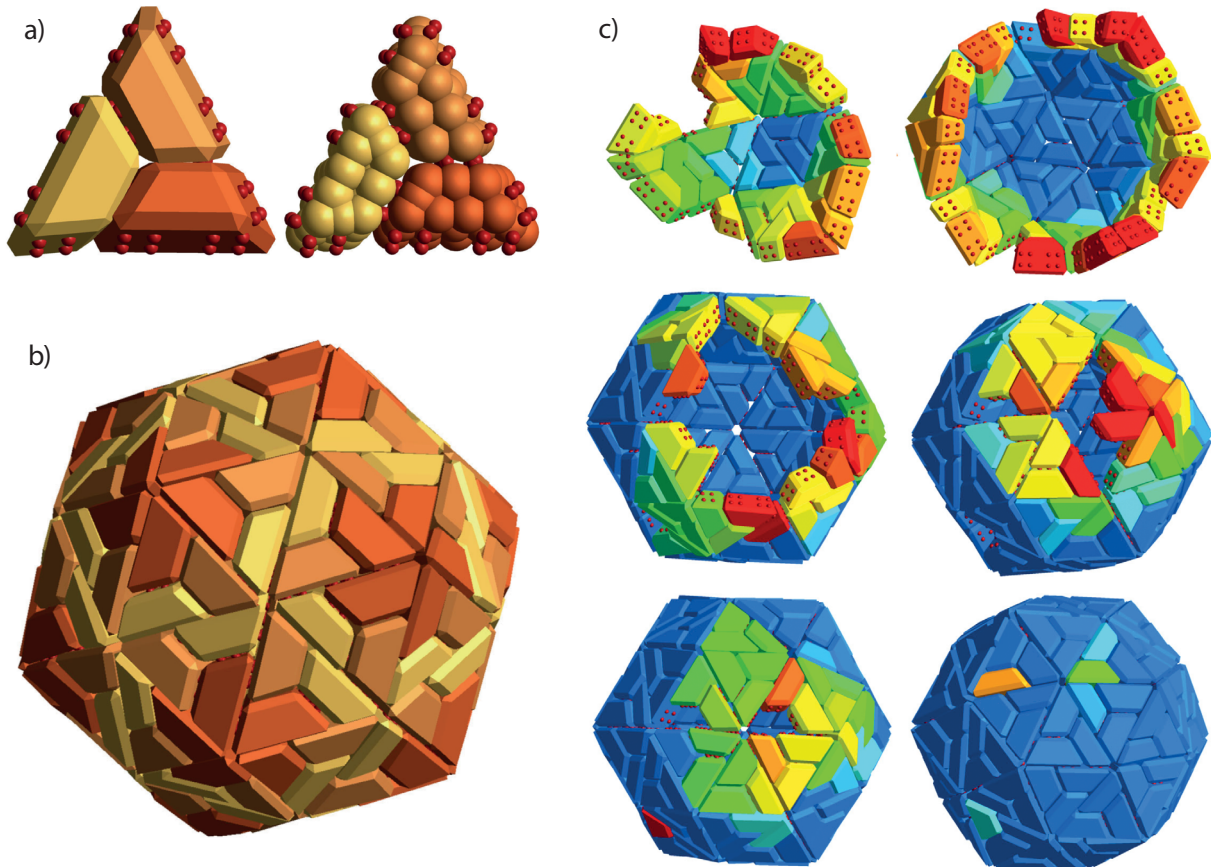


Figure 13 – Simulation basée sur un système à particules, issue de l'article [21]. (a) Représentation d'un trimère composé des 3 types de particules, colorées différemment. Les sites de liaison, des trimères et des particules sont représentés en rouge. (b) Représentation d'une capsid complète formée par 180 particules et comportant les 3 types. (c) Représentation de l'assemblage de la capsid au cours de la simulation. Un gradient du bleu vers le rouge représente l'ordre chronologique dans lequel les particules viennent se fixer sur la capsid en cours d'assemblage.

Ces méthodes simplifient l'échelle granulométrique de façon à simuler un très grand nombre de particules et d'interactions simultanées sur une échelle de temps, de l'ordre de la microseconde jusqu'à la dizaine de microsecondes. Elles nous donnent donc une très bonne idée de la cinétique et des interactions mis en œuvre pour construire la capsid. Ces méthodes sont toutefois très dépendantes de la concentration et de la stœchiométrie des particules du système étudié. Les particules que comportent de tels systèmes sont définies de façon rigide. Il n'est donc pas possible de simuler un changement de conformation au cours de ce type de simulation.

4.3 Simulations de dynamique moléculaire gros grains

Les MD gros grains sont utiles pour simuler l'auto-assemblage des capsides à une échelle de précision encore plus élevée. Cette méthode a l'avantage d'admettre un plus grand nombre de degrés de liberté et un certain degré de flexibilité pour les protéines de capsides étudiées. En effet, les atomes composant les résidus de tels modèles sont regroupés en une ou plusieurs "bille(s)", appelées "gros grains" (Coarse-Grained abrégé CG). Les grains sont maintenues entre-elles par un réseau de ressorts duquel il est possible de soustraire des liens pour instaurer une flexibilité entre domaines. Les modèles gros grains réduisent la complexité du système résolu à l'échelle atomique que l'on veut étudier. Un système de 44 000 atomes correspond à un système CG de 11 000 billes. Une étude menée sur l'assemblage de la capsidite du virus de l'immunodéficience humaine 1 (HIV-1), utilisant des simulations gros grains, a révélé des mécanismes clés. [22]. En jouant sur la concentration de la protéine de capsidite du HIV-1 et sur l'encombrement moléculaire, les auteurs parviennent à identifier le noyau critique d'assemblage. Les changements de conformation de la protéine de capsidite HIV-1, simulant l'assemblage de la capsidite entière sont contrôlés par l'ajout d'un paramètre (Figure 14).

Les MD gros grains reproduisent avec succès une grande diversité d'auto-assemblages, comme l'assemblage de lipoprotéines à haute-densité, de polymères de synthèses et de bicouches lipidiques [23–25]. Cette méthode constitue un bon compromis entre précision et coût de temps de calcul. Elle est adaptée pour prendre en compte des changements de conformation, pour étudier la flexibilité et les interactions d'un gros système ou comportant un très grand nombre de copies du même objet. Il faut retenir que lors du passage du tout-atomes au gros grains le nombre de degrés de liberté des objets simulés est inférieur à celui de la réalité et que les termes d'énergies calculés sont approximatifs.

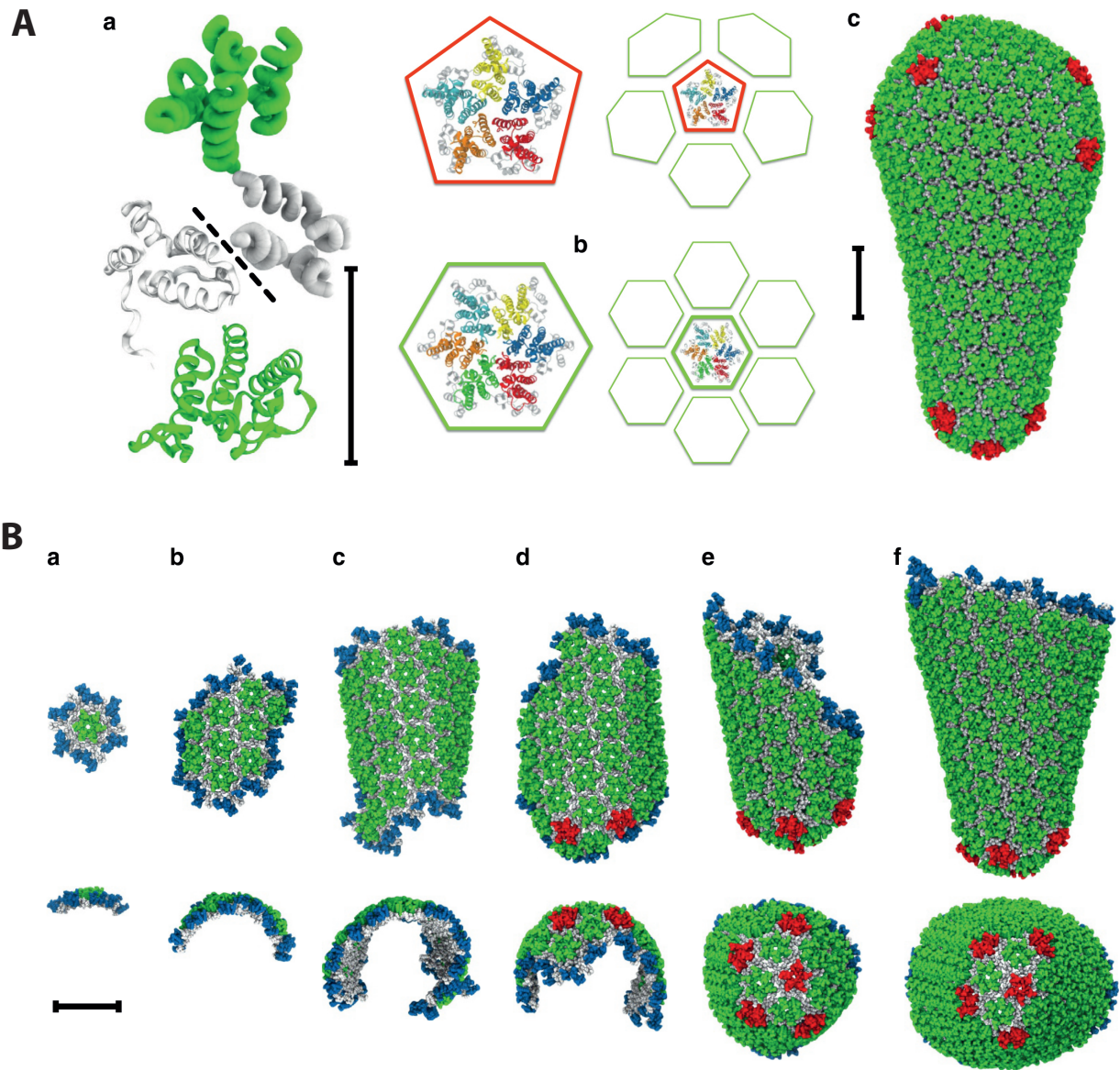


Figure 14 – Figure issue de l'article [22]. A. Description structurale de la capside du HIV-1. (a) Dimères à l'échelle gros grains et à l'échelle atomique. Le domaine N-terminal (NTD) est représenté en vert et le domaine C-terminal (CTD), en gris clair. Les deux monomères en interaction sont séparés par une ligne en pointillée au niveau de l'interface dimérique. (b) Représentation de la quasi-équivalence du pentamère et de l'hexamère. Les CTD sont représentés en gris clair. (c) Capside complète du HIV-1. Les NTD des pentamères sont en rouges et les NTD des hexamères, en vert. **B. Clichés au cours de la simulation gros grains (CG) de l'assemblage de la capside du HIV-1.** (a) 120 x 106, (b) 240 x 106, (c) 440 x 106, (d) 460 x 106, (e) 600 x 106 et (f) 1700 x 106 étapes de simulation de dynamiques moléculaires. Barre d'échelle, 20 nm.

4.4 Simulations de dynamique moléculaire tout-atomes

Étant donné le coût du temps de calcul nécessaire pour effectuer des MD de complexes biologiques macromoléculaires à l'échelle atomique, ce type de méthode n'a pas encore été utilisé pour étudier l'auto-assemblage de capside. Cependant, si on suit la loi de Moore, il sera possible de simuler à l'échelle de la microseconde des complexes de la taille d'un ribosome en 2030 [26]. On peut donc considérer qu'il sera possible de réaliser la simulation

d'auto-assemblage dans cette même décennie. Les MD à l'échelle atomique peuvent néanmoins être utilisées pour étudier la stabilité d'un complexe ou des intermédiaires d'assemblage. Les MD de la capsidite du VHB tronquée et du VIH sont de bons exemples de capsidites déjà simulées [27, 28]. Il est possible d'amarrer des unités d'assemblage sur un capsomère, à l'échelle atomique. Dans notre étude, ce procédé est utilisé pour déterminer la dynamique d'assemblage de l'intermédiaire de la capsidite du norovirus.

À partir de méthodes de simulation à différentes résolutions, j'étudie dans un premier cas l'assemblage de la capsidite du Norovirus en me basant sur des données issues de la biophysique (TR SAXS). La combinaison de méthodes tout-atomes et gros-grains constitue un bon compromis entre précision et temps de dynamique simulé pour nos systèmes de grandes tailles (plus de 40 000 atomes lourds). Dans un deuxième temps, un grand panel de méthodes computationnelles (modélisation, MD classique, TMD...) sera utilisé pour étudier la dynamique de la protéine de capsidite du VHB libre et au sein de la capsidite.

RÉSULTATS SUR LE NOROVIRUS

5 DYNAMIQUE D'ASSEMBLAGE DE LA CAPSIDE DU NOROVIRUS

5.1 Contexte biologique

Le norovirus est un virus non enveloppé de la famille Caliciviridae. Il est la cause principale de gastroentérites virales chez l'Homme et les animaux. Cette « grippe intestinale » est une des causes majeures d'infections d'origine alimentaire. Les norovirus humains sont à l'origine d'un cinquième des cas de gastroentérites dans le monde et d'environ 699 millions d'infections par an [29,30]. Ces virus, très infectieux, sont un facteur de comorbidité dans les pays en voie de développement, provoquant la mort de 200 000 personnes par an [11]. La plus grande proportion de décès survient chez les enfants. Le coût annuel mondial induit par ce virus s'élève à 64 milliards de dollars [31].

Plus de 40 souches virales sont dénombrées et divisées en 7 génogroupes (GI - GVII) [32]. Parmi ces 7 génogroupes, les génogroupes I, II et IV peuvent infecter l'Homme. Le GII est responsable de 80 à 90% des infections mondiale par le norovirus [33].

La forme infectieuse du virus encapsule un ARN simple brin de polarité positive de 7,5 kb qui contient trois cadres de lecture correspondant à des protéines structurales ou non-structurales.

5.2 Contexte structural

La protéine structurale VP1 est codée par l'un des cadres de lecture (Figure 15). Il s'agit de la protéine majeure de la capsid. Elle est composée de ~530 résidus et se caractérise par un poids moléculaire de ~57 kDa. Elle comporte deux domaines : le domaine S ("Shell") et le domaine P ("Protruding"), reliés par un segment interstitiel (résidus 221 à 229). Le domaine S (résidus 30 à 220) est le module d'assemblage. Il compose la partie interne de la capsid virale. Le domaine P (résidus 230 à 520) est exposé au milieu biologique de son hôte [34]. Il peut être décomposé en deux sous-domaines : le sous domaine P1 (résidus 230 à 278 et résidus 406 à 520) et le sous domaine P2 (résidus 279 à 405). P2 est le sous-domaine le plus exposé, il interagit avec les antigènes du groupe histo-sanguin du tractus gastro-intestinal [9]. Les parties terminales comprennent les résidus 1 à 29, qui forment le bras N-terminal, partiellement ordonné. Elles comprennent également une partie C-terminale désordonnée, qui comprend les résidus 521 à 530 pour la VP1 du virus de Norwalk, membre du génogroupe I, génotype 1 (GI.1).

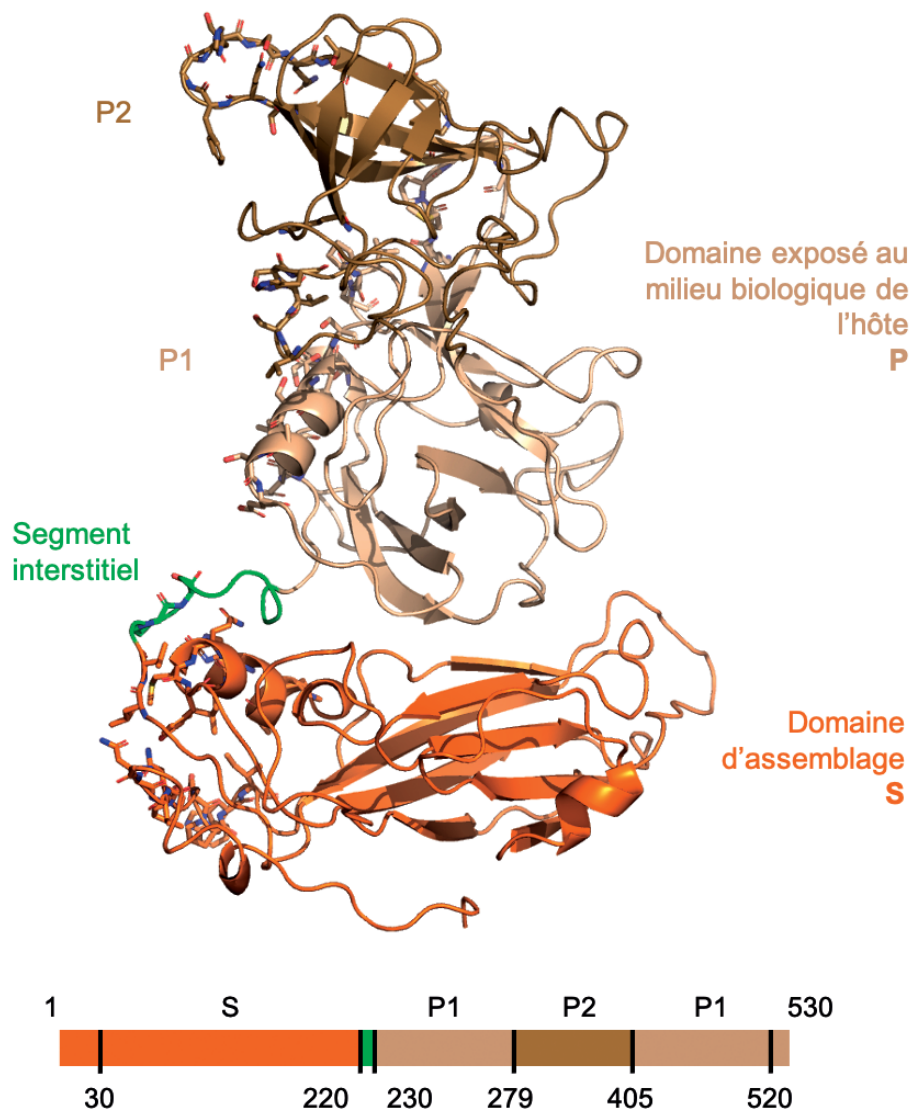


Figure 15 – Structure cristallographique de la protéine VP1 du virus de Norwalk (norovirus GI.1).
 Représentation des structures secondaires des résidus 30 et 520. Le domaine S est représenté en orange. Le domaine P est coloré en beige. Le beige clair correspond au sous-domaine P1 et le beige foncé, au sous-domaine P2. Le segment interstitiel qui relie les deux domaines est coloré en vert.

Le domaine S est la région la plus conservée. Plus de 60% de la séquence en acides aminés est identique entre les différents génogroupes [35].

La capside du Norovirus possède une géométrie icosaédrique de type T=3. Dans le contexte de la capside, la protéine VP1 adopte 3 conformations (A, B et C - Figure 16A) [34]. L'orientation entre le domaine S et P pour les conformations A, B et C est respectivement de 87,4°, 88,1° et 94,8° (Figure 16B). Cette orientation est calculée entre deux plans (domaine S et domaine P). L'angle entre les domaines S et P de la conformation C est beaucoup plus important. Les conformations A, B et C sont quasi-équivalentes et les infimes différences sont nécessaires à la constitution de la capside entière (géométrie icosaédrique tronquée).

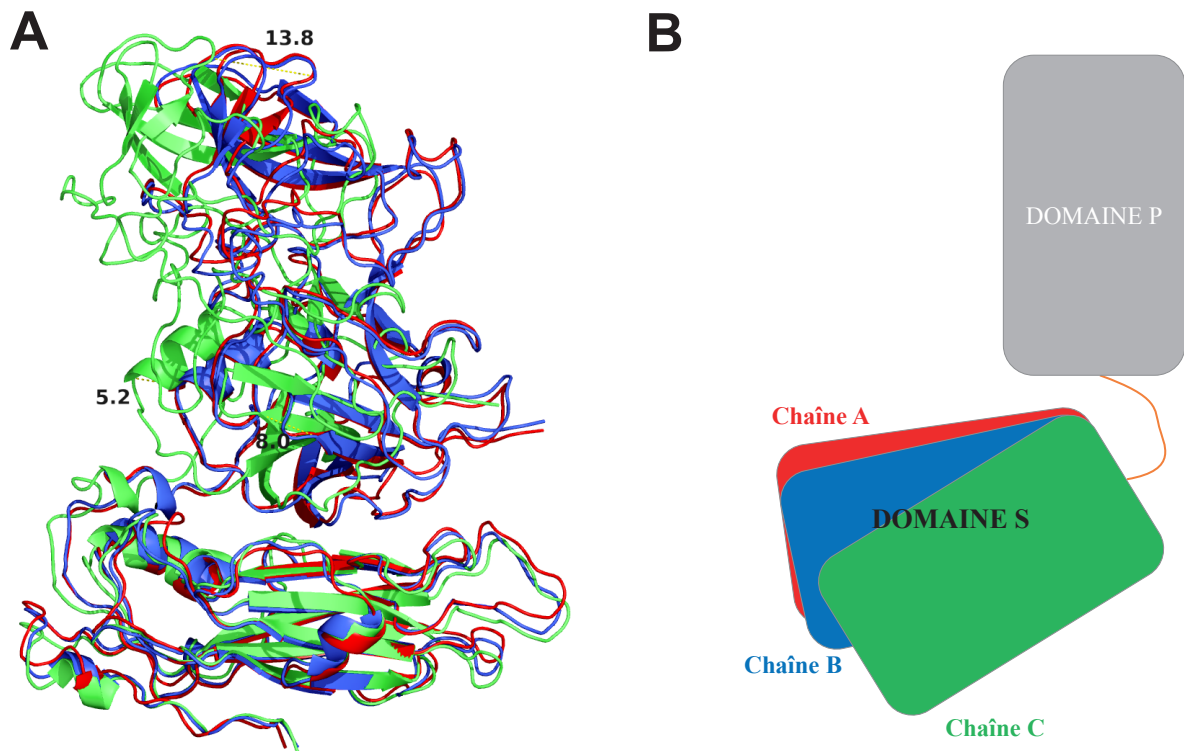


Figure 16 – Conformations de la protéine VP1. (A) Conformation A en rouge, conformation B en bleu et conformation C en vert. Les domaines S (résidus 30 à 220) sont superposés les uns sur les autres. Les conformations A, B et C ne sont pas équivalentes, les domaines P n'ont pas la même orientation. Une distance de 13,8 Å sépare la partie supérieure du domaine P de la conformation A et C. (B) Représentation de l'orientation du domaine S par rapport au domaine P pour les trois conformations de VP1. Vu du dessus. Schéma issu de la thèse de Thibault Tubiana (<https://tel.archives-ouvertes.fr/tel-01773889>).

En effet, la capside est constituée de 180 protéines VP1, 60 conformations A, 60 conformations B et 60 conformations C (Figure 17A). On retrouve ces 3 conformations parmi les unités d'assemblage (dimère A-B ou C-C).

Elles sont retrouvées, dans le contexte capsidique, sous ces deux conformations de dimères (Figure 17B). Cette capsidique comporte donc un total de 90 dimères. Dans la capsidique, les dimères sont disposés alternativement avec des contacts quasi-équivalents, selon une symétrie 5 formant un pentamère de dimères A-B (POD - Figure 17C). Les chaînes A sont localisées au centre et les chaînes B, en périphérie. Elles peuvent aussi être présentes dans une symétrie quasi-6, formant un hexamère de dimères avec 3 dimères A-B et 3 dimères C-C (HOD - Figure 17C). Les chaînes B et la moitié des chaînes C sont situées au centre. Les autres chaînes sont en périphérie.

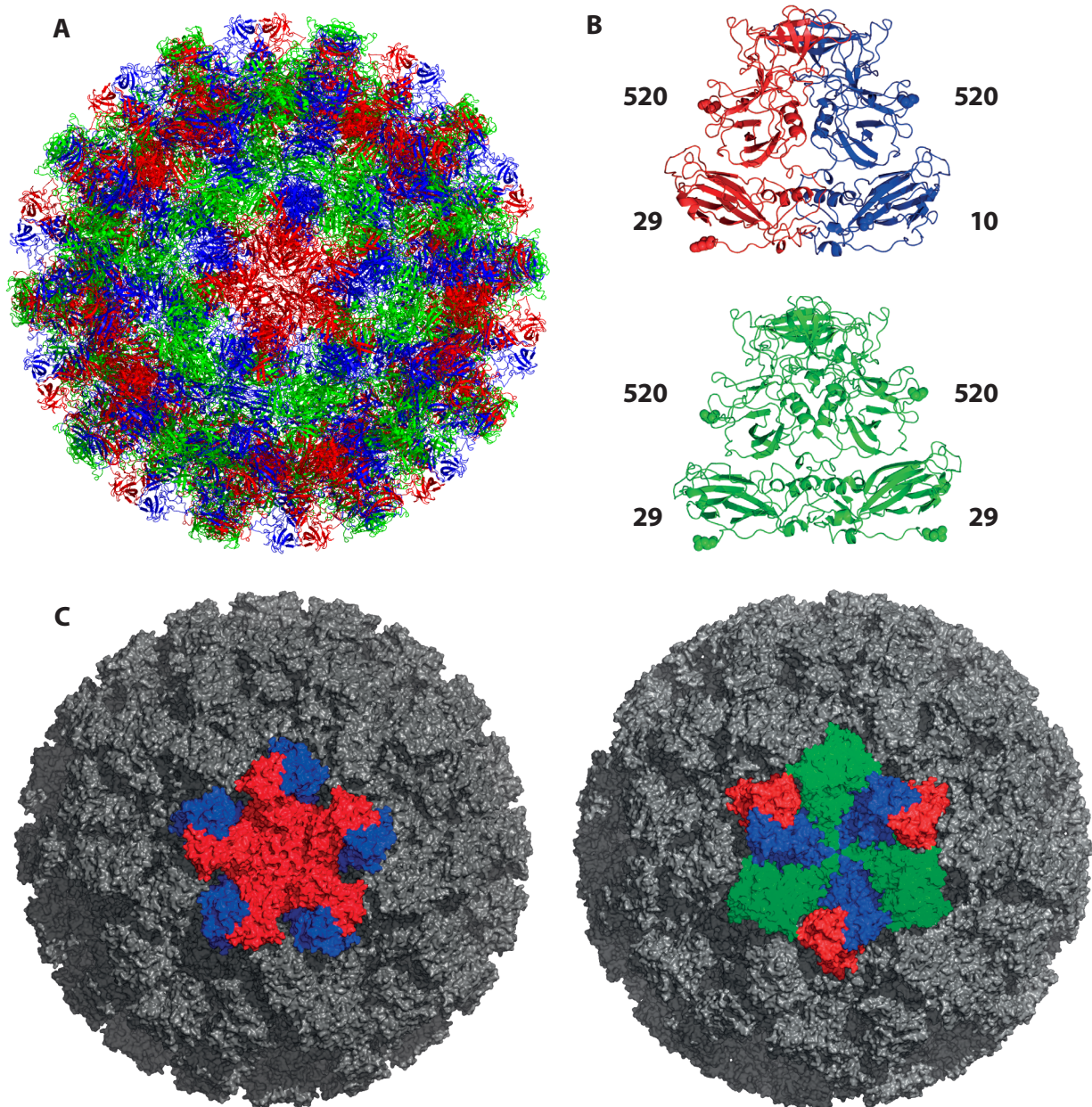


Figure 17 – Organisation de la capside du Norovirus. A. 90 dimères s'organisent en icosaèdre $T=3$. Les dimères A-B sont en rouge et bleu et les dimères C-C, de conformation légèrement différente, sont en vert. B. Les dimères A-B et C-C isolés de la capside. Les deux premiers et deux derniers résidus ordonnés sont représentés par des sphères. C. Représentation des capsomères dans la capside du Norovirus : pentamère de dimères (POD) à gauche et hexamère de dimères (HOD), à droite.

5.3 État de l'art sur l'assemblage de la capside du norovirus vide

La capside du virus de Norwalk (GI) a été résolue par Prasad et al., en 1999, par cristallographie aux rayons X (PDB ID : 1IHM) [34]. L'analyse des interactions entre domaines P a conduit les auteurs à conclure que l'unité d'assemblage est un dimère de la protéine VP1. Cette analyse est en accord avec l'observation de la dissociation de capsides vides recombinantes en une solution de dimères [35–37]. L'analyse des interactions entre les domaines S (axe 5 : environ 1800 \AA^2 de surface de contact) montre que le POD serait le capsomère le plus stable. Sur la base de ces contacts, le POD se formerait donc dans un premier temps

[34]. Les auteurs ont d'ailleurs proposé un chemin d'assemblage menant à la capside, par croissance isotrope, à partir d'un POD (Figure 18). Selon cette hypothèse, l'assemblage de la capside est essentiellement dû aux contacts entre domaines S. Il faut noter que, dans ce modèle, on considère qu'il existe un équilibre en solution entre les deux formes dimériques.

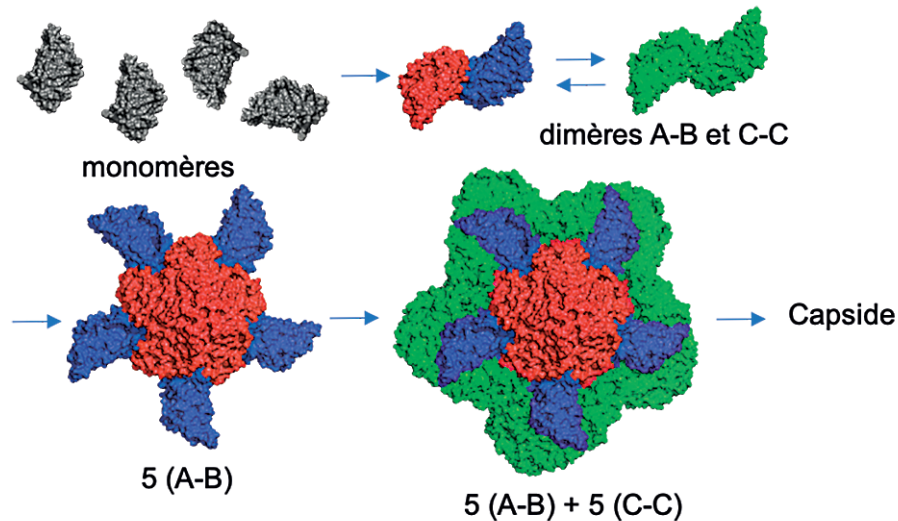


Figure 18 – Modèle d'assemblage de la capside du virus de Norwalk (GI) de Prasad et al., adapté de [34]. Seuls les domaines S sont représentés. Les monomères libres sont représentés en gris. Les chaînes des dimères A et B sont en rouge et bleu et les chaînes C, en vert.

En partant de l'hypothèse de Prasad, l'assemblage de la capside du norovirus devrait suivre le modèle de nucléation croissance (Figure 19). Les unités d'assemblage formeraient donc, par nucléation, le noyau critique : le POD. Dans le modèle de nucléation-croissance, la vitesse de dissociation est du premier ordre (constante) et la vitesse d'association, du deuxième ordre (proportionnelle à la concentration en dimère).

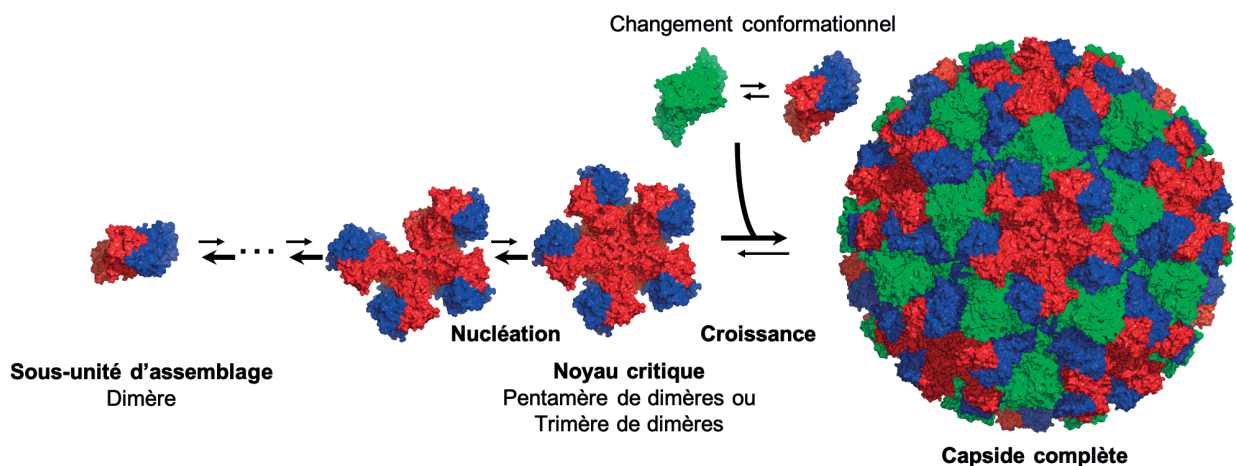


Figure 19 – Modèle de nucléation-croissance de la capside du Norovirus. L'épaisseur des flèches schématise les vitesses de réaction. Les chaînes A, B et C sont respectivement représentées en rouge, bleu et vert.

Il faut que la concentration en unités d'assemblage soit suffisante ($\sim 12 \mu\text{M}$) pour que la vitesse d'association dépasse la vitesse de dissociation et qu'ainsi, le POD se forme. La formation du noyau critique est cinétiquement limitante : il se forme lentement et contribue

directement à l'assemblage. À l'issue de la formation du noyau critique, s'ensuit une étape de croissance qui conduit à l'assemblage de la capside complète.

Une expérience de TR-SAXS, sur le norovirus bovin (GIII), a été réalisée en 2013 [38]. Cette expérience a mis en évidence 3 formes observables : un facteur de forme qui correspond au dimère VP1 de norovirus bovin, un facteur de forme correspondant à l'intermédiaire de forme allongé composé de 10 à 11 dimères et un facteur de forme caractéristique de la capside du norovirus bovin (Figure 20). Les premiers intermédiaires apparaissent entre 4 et 120 ms. Cette population d'intermédiaires est visible tout au long de l'expérience et croît jusqu'à ~2 minutes. Une fois ce temps dépassé, elle diminue avec l'apparition des premières capsides du norovirus bovin (dernier facteur de forme) qui continuent de se former pendant plusieurs heures.

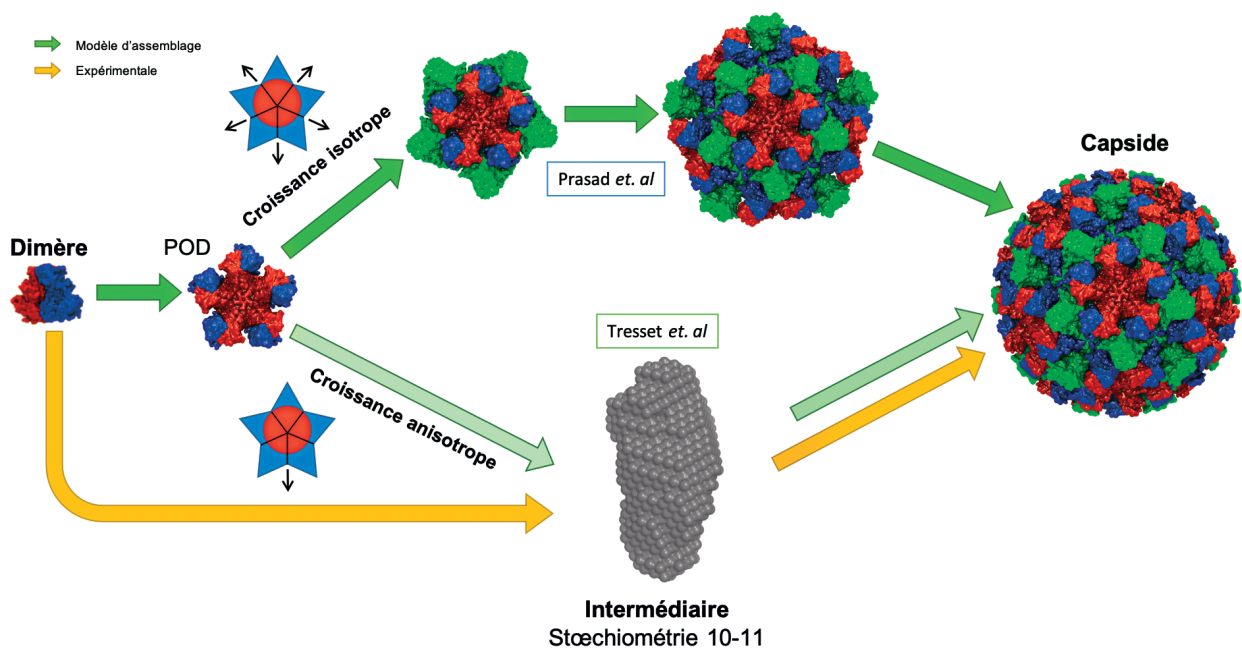


Figure 20 – Hypothèse d'assemblage de la capside du Norovirus de Prasad et al. et intermédiaires observés durant les expériences de TR-SAXS de Tresset et al. Les flèches vertes correspondent aux modèles d'assemblage. Le modèle d'assemblage de Prasad est en vert foncé. Le modèle d'assemblage que nous étudions est en vert clair. Les flèches jaunes correspondent aux facteurs de formes obtenus expérimentalement.

Le modèle d'assemblage est incompatible avec une croissance isotrope à partir du POD, comme proposé par Prasad et al. car nous obtenons un intermédiaire allongé [34]. La nouvelle hypothèse d'assemblage correspondrait donc à la formation d'un intermédiaire par croissance anisotrope, à partir du POD.

L'objectif de cette thèse est de comprendre comment passer du POD à l'intermédiaire d'assemblage observé par TR-SAXS [38]. Il est notamment question de comprendre quelles sont les interactions mises en jeu pour former l'intermédiaire.

5.4 Croissance de l'intermédiaire d'assemblage du norovirus de Norwalk

Au cours de mon stage de Master 2 Bioinformatique à l'Université de Paris, sous la co-supervision de Thibault Tubiana, j'ai étudié la dynamique des capsomères du norovirus de Norwalk (G1.1). Des analyses sur des structures statiques ont été effectuées sur la capsid du virus de Norwalk par les auteurs Prasad et al.. Leur recherche sur les interactions au sein de ces structures leur ont permis de définir que le POD est la structure la plus stable. Cependant, on peut supposer que l'HOD soit le noyau critique. Les deux capsomères (POD et HOD) sont simulés afin de voir s'il existe une différence entre eux et comprendre, comment, leur dynamique influe sur l'assemblage de l'intermédiaire.

Au cours de ma thèse, j'ai développé une stratégie qui combine différentes méthodes *in silico* de manière à comprendre comment un intermédiaire allongé se forme, à partir du POD. Cette stratégie est une procédure itérative au cours de laquelle on évalue l'amarrage d'un nouveau dimère sur un POD simulé, puis sur le POD-D résultant et ainsi de suite. La quatrième itération a été réalisée, sous ma co-supervision, par Sella Detchanamourtty lors de son stage de Licence de Physique, à l'université d'Evry.

Dans cette section, je commencerai par étudier la dynamique des capsomères du norovirus de Norwalk (G1.1), puis, je décrirai notre stratégie et les résultats obtenus à partir de celle-ci. Ces résultats mèneront à un intermédiaire allongé théorique, à partir du POD.

5.4.1 Modélisation de l'unité asymétrique

Une étape de modélisation des résidus manquants est nécessaire pour étudier la dynamique des capsomères et notamment, des régions désordonnées. Les bras terminaux pourraient intervenir dans l'assemblage de l'intermédiaire et sont donc modélisés. L'unité asymétrique d'une capsid icosaédrique a la particularité de comporter toutes les conformations d'un même monomère. L'unité asymétrique de la capsid du norovirus de Norwalk (PDB ID : 1IHM) [34] va donc nous servir pour modéliser les capsomères. Les bras terminaux de l'unité asymétrique sont partiellement ordonnés ou désordonnés (Figure 17B) :

- Les résidus 1 à 28 des monomères A et C et les résidus 1 à 9 du monomère B sont manquants pour les bras N-terminaux.
- Les acides aminés 521 à 530 des trois monomères sont manquants sur les parties C-terminales.

Ces résidus manquants sont modélisés avec Modeller 9.16 [39].

5.4.2 Modélisation des capsomères

Une fois le modèle complet de l'unité asymétrique construit, il nous sert de référence pour construire le POD et l'HOD. En dupliquant, puis, en superposant l'unité asymétrique complète sur un POD ou un HOD extrait de la structure cristallographique, nous reconstruisons les capsomères complets. Le POD comporte environ 44 000 atomes.

5.4.3 Préparation des systèmes et simulations

Après modélisation des capsomères, un passage tout-atomes à gros grains (CG) MARTINI 2.2 [40,41], couplé au réseau élastique ElNeDyn [42], est effectué. Une étape de suppression du réseau élastique entre les domaines S et P est réalisée avec domELNEDIN [43]. Les domaines protéiques P et S restent ainsi indépendants, autorisant des changements conformationnels entre domaines. Les capsomères gros grains (CG) sont enfin préparés et initient les simulations avec la suite GROMACS 5.1.4 [44] (Figure 21). Le protocole illustré est utilisé pour l'ensemble des systèmes simulés à l'échelle gros grains.

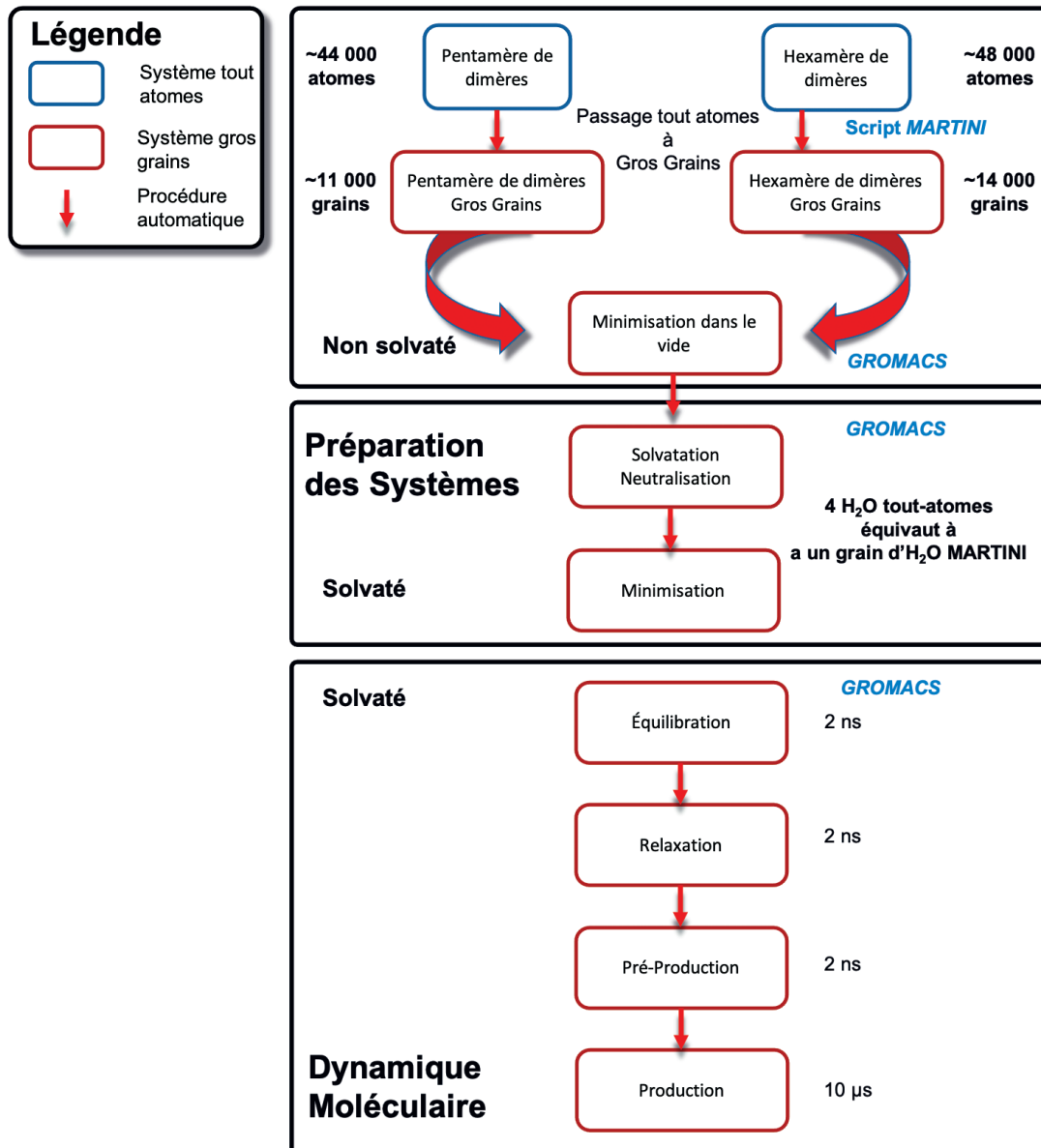


Figure 21 - Protocole de simulation de dynamique moléculaire gros grains.

Cette échelle a l'avantage de simuler de la microseconde à la 10^{aine} de microsecondes des systèmes de plus de 40 000 atomes lourds. On produit ~3,5 μs par jour avec 168 processeurs Intel(R) Xeon(R) CPU E5-2690 v4 de 2,60GHz pour un système solvate et

neutralisé de 120 000 graines. L'ensemble des paramètres utilisés dans nos simulations sont décrits dans la section 5.4.6. Le passage d'un système d'une plus haute à une plus basse résolution est facile à mettre en place. L'inverse est plus difficile car des informations sont manquantes ou sont perdues (ex. structure secondaire) lors du passage de l'un à l'autre.

5.4.4 Étude de la dynamique des capsomères

Pour étudier la dynamique des capsomères, des simulations de dynamiques moléculaires CG de 10 à 20 μ s ont été réalisées. Pour chaque simulation, une réplique a également été produite pour confirmer les observations et augmenter la statistique. La capsid du norovirus peut s'assembler avec les domaines S seuls. Les dimères interagissent essentiellement par des contacts entre les domaines S centraux. Les trajectoires sont donc superposées sur les domaines S centraux (POD : résidus 30 à 220 des chaînes A ; HOD : résidus 30 à 220 des chaînes B et C centrales ; en orange sur la Figure 22).

Le calcul du RMSD sur la structure globale (résidus 1 à 530), ainsi que sur les domaines S centraux, est un bon indicateur de stabilité (Figure 22). L'analyse du RMSD révèle que les domaines S centraux du POD et de l'HOD sont extrêmement stables. Un plateau est atteint pour les deux capsomères durant la première microseconde et les RMSD sont stabilisés à environ 2,5 Å de leur structure initiale (0 μ s – structure après pré-production). Lorsque l'on considère l'ensemble de la structure, un plateau est atteint par le RMSD du POD, au bout \sim 7 μ s. Le RMSD de l'HOD atteint également un plateau au bout de 5 μ s. La structure de l'HOD diverge de moins de 5 Å et la structure du POD diverge, elle, de plus de 5 Å. Pour les domaines S centraux (RMSD plus élevé pour l'HOD que le POD), le comportement s'inverse. Certains des bras N-terminaux du POD et de l'HOD interagissent au niveau des interfaces d'amarrage.

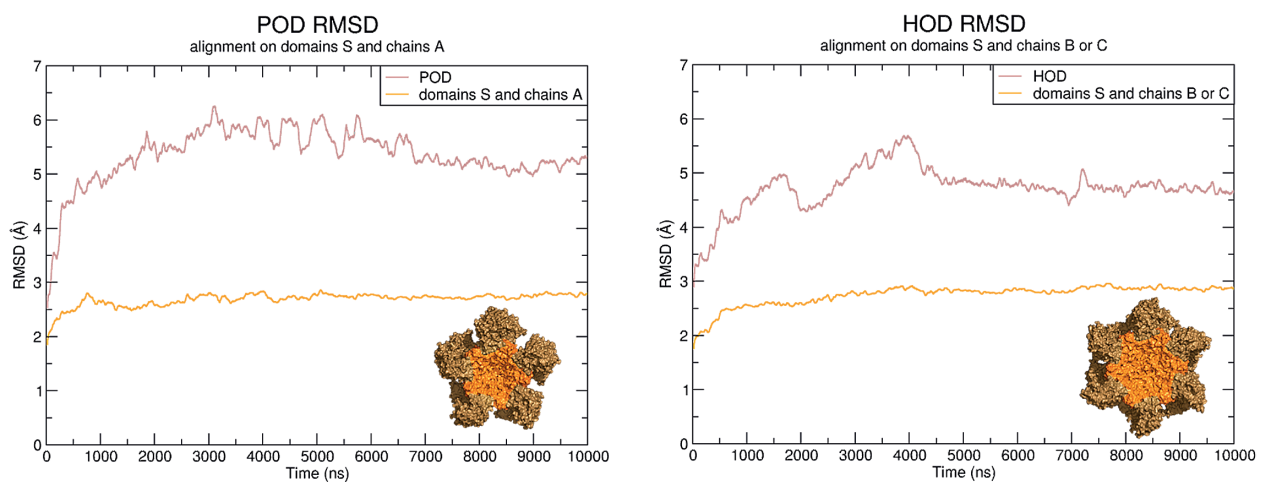


Figure 22 – Stabilité de la structure globale, superposition sur les domaines S centraux.

À gauche, le pentamère de dimères. À droite, l'hexamère de dimères. Le RMSD, calculé sur les domaines S centraux, est représenté en orange. Le RMSD calculé sur la structure globale est représenté par une courbe marron.

Ces analyses révèlent que les domaines centraux sont très stables et que les domaines en périphérie divergent davantage que les domaines S centraux. En conclusion, on peut dire :

- Qu'il n'y a pas d'instabilité du POD et l'HOD, ils ont le même niveau de stabilité ($\sim 5 \text{ \AA}$).
- Que la périphérie du POD est cependant plus mobile que celle de l'HOD.

La prochaine étape consiste à se focaliser sur le comportement des dimères au sein des capsomères (Figure 23A). En parallèle, des analyses de l'évolution des interfaces ont été effectuées. Elles servent à identifier les changements locaux au niveau des interfaces d'amarrage (Figure 23B).

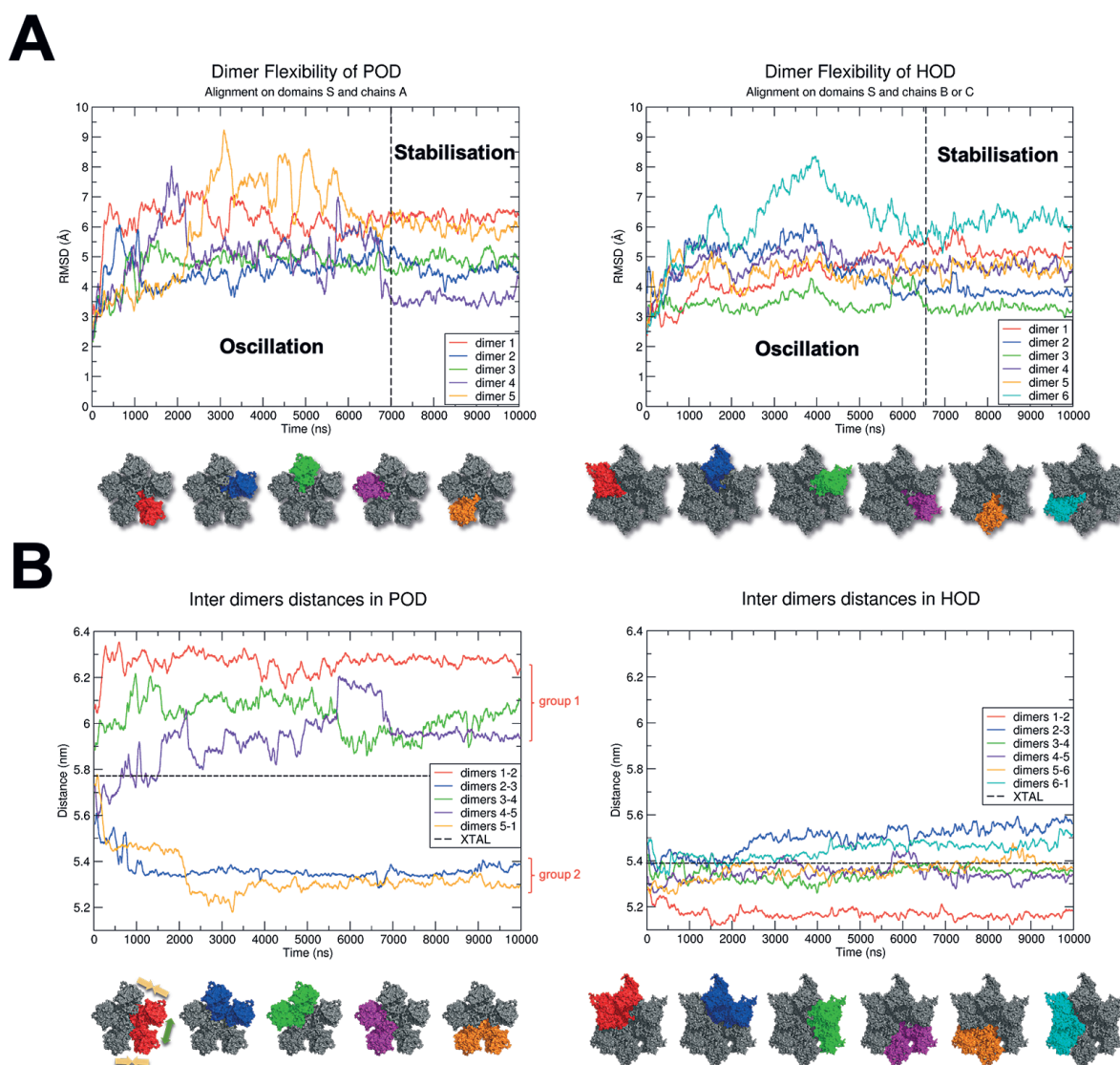


Figure 23 – Stabilité des dimères dans le contexte des capsomères. A. Flexibilité des dimères dans les capsomères. B. Distances inter-dimères pour toutes les paires de dimères adjacents. Pour les différencier, les paires de dimères sont représentées par 5 à 6 couleurs. Les couleurs des courbes correspondent aux dimères ou interfaces considérés.

La superposition des trajectoires est effectuée sur les domaines S centraux et le RMSD est calculé sur les dimères entiers. Le RMSD des dimères comporte 2 phases (Figure 23A) : une première phase d'oscillation et une seconde phase de stabilisation. Les oscillations du RMSD des dimères, dans le POD, sont plus marquées que celles de l'HOD. L'évolution du RMSD se stabilise après 7 μ s pour le POD et 6,5 μ s, pour l'HOD. Le RMSD semble indiquer que les dimères convergent vers une conformation stable.

Pour quantifier l'accessibilité des interfaces d'amarrage, on analyse l'évolution de la distance entre 2 bras (entre les centres de masses de 2 dimères) au cours de la simulation (Figure 23B). Les distances inter-dimères calculées sont comparées aux distances de références issues de la structure cristallographique (PDB ID : 1IHM) [34]. La distance de référence de chaque capsomère est indiquée par une ligne horizontale en pointillé (POD : 5,77 nm ; HOD : 5,39 nm). Les accessibilités calculées pour l'HOD varient, en moyenne, de 1 Å (0,1 nm) à 3 Å (0,3 nm), au maximum. Les distances inter-dimères calculées pour le POD varient d'environ 4 Å en moyenne à 5 Å, au maximum. On peut définir deux groupes pour le POD :

- Un premier groupe, dans lequel la distance entre dimères a augmenté.
- Un second groupe, pour lequel la distance entre dimères a diminué.

Ces deux groupes sont séparés d'environ 5 Å à 10 μ s.

En conclusion, ces analyses montrent que :

- Le POD et l'HOD sont stables, les domaines S centraux sont très stables.
- Le POD est beaucoup plus dynamique que l'HOD, en considérant la dynamique de la périphérie et des interfaces.

5.4.5 Rupture de symétrie chez le POD

L'accessibilité des interfaces, pour les dimères complets (domaines S et P), indique qu'une rupture de symétrie s'est produite chez le POD. Compte tenu de l'importance des domaines S, dans l'assemblage de la capside du norovirus, une analyse de distances inter-dimères des domaines S périphériques (chaînes périphériques), est effectuée (Figure 24A). Elle est complétée par un calcul de surface accessible au solvant (SASA) des interfaces du POD (Figure 24B). De cette manière, on peut vérifier si la rupture de symétrie ne se produit pas uniquement au niveau des chaînes périphériques.

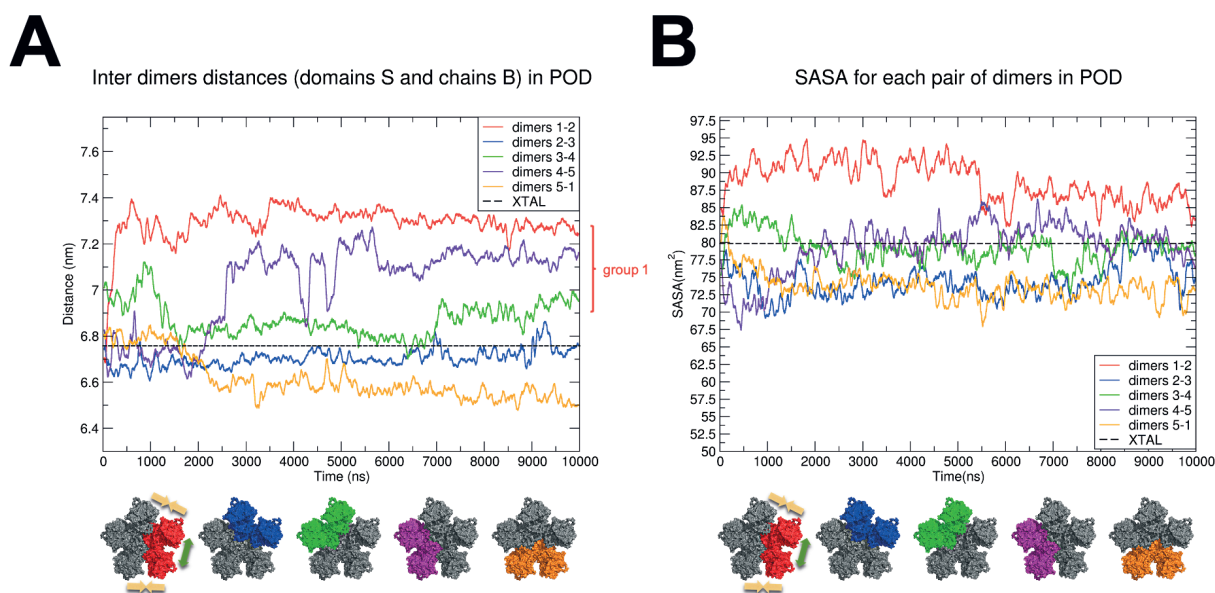


Figure 24 – Évolution des interfaces d’amarrage du POD. A. Distances inter-dimères des domaines S périphériques entre toutes les paires de dimères adjacents. B. Surface accessible au solvant de toutes les interfaces du POD. Les couleurs des courbes correspondent à celles des interfaces.

L’analyse des distances inter-dimères des domaines S périphériques sur la trajectoire montre que, pour 3 des 5 paires de dimères voisins, la distance augmente (Figure 24A - groupe 1). Les deux autres interfaces ont des distances inter-dimères égales ou inférieures à la distance cristallographique de référence : 6,76 nm. La valeur de la SASA, pour les interfaces (Figure 24B), est corrélée avec l’évolution des distances entre domaines S périphériques. En considérant la SASA de référence : 79,85 nm² (ligne horizontale en pointillé), la surface peut augmenter de 4 % (dimères 1-2) ou diminuer de 2 à 8 % (dimères 3-4, 4-5, 2-3 et 5-1 par ordre croissant de perte de surface accessible). Les interfaces directement voisines de celles qui s’écartent le plus (dimères 1-2), se rapprochent le plus (dimères 2-3 et 5-1).

5.4.5.1 Conclusion sur la rupture de symétrie du POD

Le POD est donc bien sujet à une rupture de symétrie. En effectuant les mêmes analyses sur la réplique du POD, ces résultats sont confirmés. La réplique est aussi caractérisée par une rupture de symétrie. Entre la simulation et sa réplique, les interfaces qui s’écartent ou se rapprochent, sont différentes. Si l’on effectue une rotation de la réplique et qu’on la superpose sur la dynamique, les ruptures de symétrie de ces deux systèmes sont équivalentes. La rupture de symétrie est-elle à l’origine de la croissance de l’intermédiaire ? A-t-elle une influence ?

Une série de tests d’amarrage a été effectuée avant la soumission de l’article qui est joint. Elle a montré que l’amarrage d’un dimère donne plus de résultats de docking si les parties flexibles sont retirées. Les bras N-terminaux, qui se sont placés au niveau des interfaces, gênent l’amarrage de nouvelles unités d’assemblage. Les capsomères sont alors constitués des résidus 30 à 220 et 230 à 520.

La méthode d'amarrage que nous utilisons est une méthode qui calcule la meilleure orientation d'une molécule (protéine) vers une autre pour former un complexe stable [18] (p. 661). Le principe de l'amarrage repose sur 3 étapes :

- La représentation des systèmes moléculaires
- L'implémentation d'un algorithme de recherche des interactions
- Le développement d'une fonction pour évaluer l'interaction entre les deux partenaires

Toutes les étapes d'amarrage sont corps rigides, c'est-à-dire que les deux partenaires sont considérés comme des corps rigides lors de leur interaction.

5.4.6 Étude de la formation de l'intermédiaire d'assemblage

Notre article “Combining computational methods to study biological macromolecular complexes assembly: Application to the norovirus capsid initial growth”, qui porte sur l'étude de la formation de l'intermédiaire d'assemblage observée par TR-SAXS [38], a fait l'objet d'une soumission en mars 2021. Dans le cadre de cette étude, une stratégie qui repose sur la combinaison de méthodes *in silico* a été développée. Elle combine des simulations de dynamiques moléculaires gros grains, des étapes de regroupement de structures (clustering) et des étapes d'amarrage. Cette stratégie est une répétition d'amarrage successifs (n), au cours desquels une nouvelle unité d'assemblage (D) est amarrée sur le POD de départ. Dans cet article, seront présentés les résultats d'amarrage d'un dimère C-C sur le POD, puis sur le POD-D résultant, jusqu'à 2 POD-D₂ possibles (correspond à 2 répétitions de notre stratégie sur le POD). De manière générale, les structures résultantes de notre stratégie sont nommées POD-D_n où n correspond au nombre de dimères amarrés au POD de départ. Chacune des interfaces d'amarrage composant les POD-D_n sont numérotées. Par exemple, si le POD est amarré à l'interface n°2, on le nomme POD-D (2). Si ce même POD-D (2) est amarré à l'interface n°3, on le nomme POD-D₂ 2-3. Dans le cas où le POD-D (2) est amarré à l'interface n°6, son nom est POD-D₂ 2-6 etc. Cette nomenclature indique précisément combien de dimères sont amarrés et sur quelles interfaces.

Article soumis.

Combining computational methods to study biological macromolecular complexes assembly: Application to the norovirus capsid initial growth

Studying self-assembly of norovirus capsid with in silico methods

Jean-Charles Carvaillo¹, Thibault Tubiana^{1,2,3}, Stéphane Bressanelli^{1*}, Fernando Luís Barroso da Silva^{4,5}, Yves Boulard^{1*}

1. *Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198 Gif sur Yvette cedex, France*
2. *Computational Biology Unit, Department of Informatics, University of Bergen, Norway*
3. *Department of Biological Sciences, University of Bergen, Norway*
4. *Universidade de São Paulo, Departamento de Ciências Biomoleculares, Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Av. café, s/no – campus da USP, BR-14040-903 – Ribeirão Preto – SP, Brazil*
5. *Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, NC 27695, United States*

*: authors to whom correspondence may be addressed

Abstract

Understanding how macromolecular assembly occurs is a fundamental challenge because they are at the center of most biological processes. It is an elaborate process that requires non-covalent stable interactions between partners to stabilize a desired architecture for a specific purpose. One of the advantages of virus models is that under adequate conditions capsid proteins can be efficiently assembled in vitro in the absence of any other component. The present study aims at describing the initial steps of molecular self-assembly of norovirus capsid by combining in silico computational approaches.

Introduction

Large macromolecular complexes may consist of hundreds of individual components, including mainly proteins, nucleic acids and lipids, to perform many vital tasks for the cell. Among the best known, we can cite the ribosome to synthesize proteins or spliceosome to catalyze pre-RNA splicing [1]. If these macromolecular complexes' final structures could be solved with X-ray crystallography or cryo-electron microscopy, their formation is difficult to study. It is an elaborate process that requires to define the order of assembly and to investigate the interactions between each partner. Moreover, it is a dynamic

process and the interaction between binding molecules can serve to stabilize a desired preferential conformation to favor binding of a supplementary partner. To study those assembly pathways, virus capsids can be perfect study objects because they are large complexes made of multiple copies of a few proteins packaging the viral nucleic acid. In the simplest case, the capsid is made of a single protein that can self-assemble even without the nucleic acid component.

Viruses are obligate intracellular parasites, meaning that viruses rely on intracellular resources and hijacked cellular machinery of a host. They can invade and propagate into all type of life forms: animals, plants and bacteria [2]. Their replication is both a complex and straightforward mechanism. It is based on the implementation of their information into host cells. To do so, viruses are organized by three essential components. (i) The first corresponds to the information; i.e. the genetic material. It can be found as a single or double-stranded of polynucleotide chain (DNA or RNA). (ii) A protein shell that plays a major role in protection and other biological functions such as transport, regulation of this information. A capsid containing the genetic material is formed from numerous copies of a single or a few protein sequences. (iii) The last one is not composing every virus; it corresponds to a membrane enveloping the capsid and acquired from host cell membrane [3]. This study focuses on the understanding of the capsid assembly process and especially on the formation of assembly intermediates. Most of the capsids can be classified into two classes of geometry: rodlike or spherical [3]. Rodlike capsid adopts a helical symmetry that can accommodate in principle any length of nucleic acid. Spherical capsids, as for them, usually adopt icosahedral symmetry restrained by regular polyhedron properties. There are exactly 60 identical pieces in an icosahedron, but regular icosahedral capsids can actually include a multiple T of 60 protein subunits, thus allowing larger shells to be made with the same polypeptide length. For particular values of T (e.g. 1, 3, 4 ...), the $60 \times T$ protein subunits can form the icosahedral shell by assuming T very similar (though not identical) conformations [4]. Both theoretical models and experimental techniques such as magic angle spinning nuclear magnetic resonance [5], time-resolved small-angle-X-ray scattering (TR-SAXS) [6,7], size exclusion chromatography [8] or light scattering experiments [9] have been used to study assembly mechanisms of the viral capsid. It should be noted that none of these experimental techniques allows to detect all

assembly intermediates. It remains difficult to parse all capsid assembly pathways and products that occur over large ranges of length and time scales (angstrom to micrometer and picosecond to minute), and most of which are transient [10]. To overcome this problem, independent theoretical models of both nano spatial scales and time scales have been developed. For example, mathematical models to study capsid assembly kinetics were established to determine evolution of intermediates concentrations in function of time [11]. Two major models, one to predict the assembly of empty capsids and the second to describe the assembly around their nucleic acid were developed. But both are based on nucleation-growth schemes and particularly on a critical nucleus formation [3]. A critical nucleus corresponds to the smallest intermediate which have more than 50/50 chance of growing to the complete capsid before disassembling [3,12,13]. Capsid assembly can be also studied with particle-based dynamics simulations. Monte-Carlo and molecular dynamics simulations of capsid subunits can be computed to monitor their dynamics and interactions between them [14–16]. Defining intermediates as particles or coarse-grained (CG) models enable faster sampling due to reduced degrees of freedom. Most of the time, the defined assembly subunit corresponds to the critical nucleus. With these *in-silico* techniques, plausible assembly models have been defined [17,18], especially on T=3 capsid self-assembly [14]. Another examples of *in-silico* coarse-grained capsid self-assembly studies on human immunodeficiency virus 1 (HIV-1) capsid can be cited [16]. Numerical methods are high performance computing techniques contributing to better understanding assembly or dynamic of multimer structures. We can cite the microsecond simulation of hepatitis B virus (HBV) capsid at an atomistic level as example of huge system performed on a supercomputer. The solvated and neutralized HBV capsid is composed of almost 6 million atoms [19]. This all-atom simulation indicates that the capsid is capable of asymmetric distortion and highlight the role of ionic strength to promote capsid assembly.

In 1999 Prasad and coworkers solved the Norwalk virus capsid structure at atomistic resolution [20]. After analyzing the 3D structure, they proposed that a pentamer of dimers (POD) would be a major intermediate. From this one, an assembly model was suggested based on an isotropic growth by adding the assembly unit of norovirus in all directions. The Norovirus assembly unit corresponds to a dimer of a capsid polypeptide chain call VP1. The capsid is composed of 90 dimers. Polypeptide chain

differentiates into two quasi-equivalent dimers (A-B or C-C) in capsid. It is important to note that VP1 exists in a dimer form in solution. In 2013, the TR-SAXS bovine norovirus self-assembly study performed by Tresset *et al.* revealed 3 species: (1) a form factor corresponding to bovine norovirus VP1 dimer, (2) a form factor corresponding to an elongated intermediate, and (3) a form factor characteristic of complete bovine norovirus capsid [21]. Clearly the TR-SAXS results do not correspond to the isotropic growth implied by the self-assembly model proposed by Prasad *et al.*, but to an isotropic growth of the capsid. To improve our comprehension of the formation of long elongated intermediate from the POD capsomer, we developed an *in-silico* computing strategy, combining both all-atom and coarse-grained molecular dynamics and docking calculations. During this process, the analyses of the interesting intermediate structures are based on the knowledge of the final solution extracted from the capsid structure. From a biological point of view, capsid assembly and maturation correspond to the late stages of the viral life cycle which are essentials for the formation of infectious viral particles. Consequently, the development of drug that target the viral capsid assembly is attractive. Indeed, such inhibitor was identified for HBV capsid and is used as is used as therapeutic treatment [22].

Materials and methods

Starting models. Complete asymmetric unit model of Norwalk Virus VP1 capsid protein was modeled based on crystallographic capsid structure (PDB: 1IHM) [20]. Xray structure of chain A or C has been determined from residues 29-520 and chain B comprising residues 10-520 of the 530 residues. The missing residues were completed with Modeller 9.16 program [23]. At the end, we obtained the all-atoms complete models of both VP1 A-B and C-C dimers and the pentamer of dimers (POD).

Assembling Study Method. The strategy used to study the formation of a molecular edifice with computational methods is summarized in Figure 1. It has been applied to study the assembly of Norwalk virus capsid. We can sub-divide the cyclic process in four steps. (1) The all-atom starting model (POD) was converted to Martini coarse-grained structure. Charges were assigned assuming protonation states at pH 7 because i) this is presumably the physiological pH for norovirus capsid assembly and ii) it was shown that in vitro the recombinant norovirus capsid disassembles into dimers at pH 9 and low salinity

and efficiently reassembles into capsids at pH 6 to 7 and salinity 50 to 150 mM [24,25]. This pH7 coarse-grained structure was used to initiate at least 10 microseconds production time of MD simulations with Gromacs. (2) After clustering of the trajectory, the representative structure of the last cluster was selected. (3) This structure was then converted back to an atomistic model to allow docking of a new molecular brick (VP1 dimer), and (4) the best docking solution was selected based on both the energy criterion and the root mean square difference (RMSD) in coordinates between a docking pose and the “exact” solution (crystallographic structure obtained at pH 4.8). At this step, two strategies were used. Either the new structure issued from the calculation is directly used to initiate a new cycle or to keep the curvature of the capsid, the structure extracted from capsid is used to launch the Martini CG molecular dynamic simulation.

Molecular dynamics calculations. We computed molecular dynamics (MD) simulations with the GROMACS simulation package version 5.1.4 or version 2016.4 [26] and the MARTINI 2.2 force-field [27,28] coupled to ElNeDyn elastic network [29]. Successive structures of this assembly study are coarse-grained (CG); an elastic network is effective on backbone grained to maintain secondary and tertiary structures properties. To conserve independent flexibility and movement of S and P domains, we used the domELNEDIN algorithm deleting elastic bounds between domains [30]. A first steepest-descent energy minimization in the void on CG models is performed, Coulomb interactions within a cut-off radius of 0 to 1.2 nm and non-bonded Van der Waals (Lennard-Jones) interactions within a cut-off radius of 0.9 to 1.2 nm were used. Systems were neutralized with Na⁺ ions and solvated with MARTINI standard water combined to a dielectric constant ($\epsilon = 15$) All the simulations were performed at neutral pH and at 300 K. Solvation was performed under periodic condition with a non-cubic box with a distance between the protein and the edge of the box of at least 3 nm. The structures obtained after a second energy minimization were used to initiate molecular dynamics simulations. An integration time step of 20 fs was used during the calculation and the neighbor list was updated every 10th step. Covalent bonds in the proteins were constrained using the LINCS algorithm [31]. We performed a NPT equilibration of 2 ns. The pressure of the simulation box was kept at an average of 1 bar using the isotropic Berendsen coupling algorithm [32] and the temperature is kept at 300 K using v-rescale

coupling algorithm without a heating step. After the equilibration, the position restraints were removed. All other simulation parameters were the same as during the equilibration except for the pressure coupling algorithm for which an isotropic Parinello-Rahman barostat was used [33]. After the 2 ns relaxation and 2 ns pre-production steps, microseconds (10 to 20 μ s) CG MD simulations were finally produced. VMD 1.9.3 [34] was used for RMSD (1) calculations where δ^i is the distance between atom i and either a reference structure or the mean positions of the N equivalent atoms, and Gromacs tool “rmsf” for the root mean square fluctuation (2), where T is the duration of the simulation but in frames, $x_i(t_j)$ the coordinates of atom x_i at time t_j . \tilde{x}_i corresponds to the time-averaged position of the same coordinates i .

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

(1)

$$RMSF = \sqrt{\frac{1}{T} \sum_{t_j=1}^T (x_i(t_j) - \tilde{x}_i)^2}$$

(2)

Clustering analysis. We used our *in-house* program TTClust [35] to cluster the trajectories with a hierarchical clustering. The hierarchical distance is computed from the pairwise RMSD matrix Ward variance minimization algorithm [36]. The optimal number of cluster is estimated with the elbow method and a threshold of 0.995 with K-means clustering [37]. The representative frame is defined by the structure with the lowest RMSD against all the frames which belong to the same cluster.

Docking procedure. Coarse-grained representative structures obtained after clustering step are converted to all-atom structures with a going backward dedicated tool [38]. A protein-protein rigid body docking (FFT) is performed on ClusPro 2.0 webserver with standard parameters [39] on a receptor and his ligand. After docking, ClusPro 2.0 computed several low energy clusters which contain similar structures and extracted representative conformations. To discriminate docking solutions, we used the ClusPro 2.0 balanced scoring scheme and cluster centers PIPER energy functions [40]. In addition to the balanced score provided by ClusPro, we implemented an in-house tool based on PyMOL [41] functionalities using a crystallographic reference. The receptor used in docking process and the corresponding receptor in a crystallographic reference are 3D fitted at each docking interface. Then,

local RMSD are computed between docked ligand on targeted receptor and original ligands in the crystal reference structure. This method allowed us to test each crystallographic docking interface to identify which one is docked on targeted receptor, to sort and classify them. We plotted local RMSD on y-axis and corresponding docking interface in x-axis. The number of members in a cluster is represented by the size of dots. The resulting graph synthetize the procedure described upward.

Complementary analysis based on electrostatic features

When a charged protein (e.g. a ligand) approaches or even binds to another protein (e.g., a receptor), its charge can induce changes in the ionizable properties of the titratable amino acids on the protein especially the ones at its nearby surface. Of course, due to the long-range nature of electrostatic interactions, a strong electrostatic coupling between titratable groups can be observed between groups of residues, and even buried titratable groups can be affected by the visits of the charged macromolecule during an observation time. Such shifts in the protonation states provide valuable physical chemical information, and can offer a pragmatic approach to map all the titratable protein residues that are electrostatically affected by such visits during a simulation run [42]. By the comparison of the partial charges of a protein immersed in an aqueous electrolyte solution (apo state) with their values in the presence of its charged ligand in the same physical chemical conditions (pH, ionic strength, protein concentration and temperature), such charge shifts are directly estimated and can be used to identify the key amino acids that interact with the ligand. For a pragmatic use of this information to define the specific surface of the receptor where the ligand has more often visited, a weak electrostatic coupling between superficial and buried titratable group is the ideal scenario. As stronger is this coupling, a higher number of deeper and/or just neighbors amino acids is perturbed by the ligand coming nearer the receptor, and consequentially less effective is the theoretical prediction of the preferential binding spots. The partial charges of all macromolecular systems at pH 7 and 150 mM of 1:1 salt were calculated by the fast proton titration scheme for biomolecular systems [43]. Such constant-pH Monte Carlo simulations were performed both for the apo states (receptor in the absence of the ligand) and during the ligand-receptor complexation processes [44]. We implemented a visualization

and an analyzing PyMOL tool to exploit resulting-partial charges (in elementary units) variations (δq) that is explained further in the results section.

Results and Discussion

Starting from the structure of the Norovirus capsid resolved by X-ray crystallography (PDB: 1IHM), we extracted a POD as it was identified as the most stable and therefore most likely structural intermediary that could initiate the capsid assembly [20]. The structure was then completed and minimized. The all-atom model was converted to Martini coarse-grained structure to initiate 20 microseconds production time of MD simulations with Gromacs. After trajectory clustering, we identified a stable conformation in which the initial symmetry of the POD is broken, resulting in an asymmetric structure (Figure 2). Compared to the symmetric POD for which the distance between each arm is constant, some arms of the asymmetric POD are nearest whether others are furthest. This new CG structure was then converted to an all-atom structure to allow docking of a new dimer [38]. The docking approach was also performed on the symmetric POD to compare the results. The 30 representative structures were generated among 1 000 best docked solutions performed by balanced ClusPro model and based on the PIPER energy of the new complex. These representative structures were then visually analyzed. Some of them are satisfactory, the new dimer docks between two arms of the POD as expected but some of them are not compatible with a new intermediate to form the final capsid (some examples are shown in Figure S1). To discriminate between the acceptable solutions and the false ones, we developed an analyzing tool. It is based on the calculation of the RMSD between the structure of the final exact solution that corresponds to a POD with an additional dimer extracted from the capsid X-ray structure and the selected docking solution obtained with ClusPro program. The cut-off for the RMSD was defined firstly visually on the graph but as it is shown in Figure 2B there is a well-defined gap between good and bad RMSD values. A RMSD of 16 Å clearly permits us to validate good solutions (threshold represented by a dashed line) and it excludes bad or false positive energetic solutions. This strategy was further validated by the fact that we observed a very good correlation between the best energy structures and the lower RMSD docking results (Figure 2C). The final results are shown in Figure 2 for the POD

structure. Clearly, for the symmetric POD extracted from the crystal structure, the five favorable positions (between each arm) for docking a new dimer are equivalent (Fig. 2A) whereas for the asymmetric POD resulting from 10 μ s MD simulations, one docking position is impossible and the 4 others are not equivalent (Figure 2B). Positions 1 and 2 seem to be the best. After 20 μ s MD simulations (Figure 2D), the situation was slightly different since 2 docking positions appeared favorable, positions 2 and 4. At this stage, we retained the position 2 to add a new dimer at the POD and to continue the study. Starting from this new structure named POD-D, we used the same strategy as described above to add a new dimer. The results are shown in Figure 3.

Based on the POD-D structure, we expected six favorable positions to dock a new dimer. Four of them correspond to the free remaining positions of the initial POD structure. The other two possibilities correspond to docking on the supplementary dimer. As shown on graph Figure 3A, only five of them are identified as acceptable solutions if we consider the crystal capsid structure. After 20 μ s of MD, the number of favorable docking solutions for a new dimer is reduced to three (Figure 3B). They correspond to free positions on the POD structure, the two positions adjacent to the first docked dimer and the third to one of the two positions to its opposite. We observe that there is still a good correlation between the best energy structures and the lower RMSD docking results (Figure 3C). However, some very favorable solutions for the energy criterion are now clearly excluded if we consider the RMSD criterion (blue circle in Figure 3C). At this stage, we retained the solution for which there were the most positive results with the lowest minimum RMSD (blue arrow Fig. 3B). It corresponds to the structure in which the new dimers docks near the previous one and we name it POD-D₂-2-3 (structure POD with 2 docked dimers at position 2 and 3). The results of the addition of a new dimer to this structure are shown in Figure 4. Without surprise, if we consider the X-ray structure, the three remaining positions of the initial POD structure are very satisfactory (positions 1,4 and 5 in Fig. 4A) resulting in an isotropic growth of the capsid. In contrast, two out of the four possible positions for an anisotropic growth are favorable (positions 7 and 9 in Figure 4A) but with lesser positive solutions (47 vs 273). After the MD run, the results are significantly different (Figure 4B) since only two positions are advantageous to dock a new dimer, the first corresponding to an isotropic growth (position 4 in Figure 4B) and the second to an

anisotropic growth (position 8 in Figure 4B). If these solutions are numerically and energetically quasi equivalent, the RMSD value is better for docking on the initial POD structure. As the structure growth, obviously the best solutions do not correspond to the ones selected by ClusPro with only an energy criterion.

As explained in the Materials and Methods section, we also tested our strategy when keeping the best-docked solution to initiate a new cycle of docking an additional VP1 dimer. Compared to the approach described above, several drawbacks appear. Firstly, the curvature of the viral capsid is not maintained and secondly, defects accumulate, as the molecular assembly grows compared to the reference crystal structure. This accumulation of errors makes the selection of good docking solutions more and more difficult. However, at the start of the procedure, it is quite satisfactory. The docked solutions are interesting and close to those described above. In particular, at the POD-D step, a second pathway leading up to an anisotropic growth appeared (position 6 in Figure 3D). This solution, named POD-D₂-6, was observed with our first approach as shown in Figure 3A, but the docked dimer's position was too far from the crystal structure to be retained. In general, isotropic docking solutions are preferred as shown in Table 1. But it clearly depends on our choice for the first docking dimer for which many good solutions are possible.

	Total number of solution compatible with an isotropic growth	Total number of solution incompatible with an isotropic growth
Positions from Xtal POD-D structure (Fig. 3A)	359	24
Positions after 14 μ s of molecular dynamics of the Xtal POD-D structure (Fig. 3B)	98	0
Positions from POD-D structure (Fig. 3D) issue of docking results on POD structure obtained after 8.2 μ s of molecular dynamics (Fig 2B)	28	10

Table 1 : Comparison of POD-D docking results

For the rest of the study, we explored the two paths as shown in Figures 4 and 5 for the POD-D₂ step. The results are resumed in Figure 6. All the potential docking solutions that remained to complete the

initial POD structure are observed for the POD-D₂ crystal structure as shown in Figures 4 and 5. It corresponds to positions 1, 4 and 5 in Figure 4A and to positions 1, 3, 4 and 5 in Figure 5A. These solutions are compatible with an isotropic growth for the assembly. We also observed solutions compatible with anisotropic growth (position 7 and 9 in Figure 4A and 7 and 10 in Figure 5A). After the MD run, we observed only 4 favorable docking positions for the supplementary dimer. 3 of them correspond to a docking on the initial POD structure (position 2 in Figure 4B and 1 and 3 in Figure 5B) and only one to a docking on a new interface (position 6 in Figure 4B). These solutions are close for both RMSD and energy criteria. Interestingly, the two paths have a common POD-D₃ solution, (position 6 Figure 4B and 1 Figure 5B). We have shown that exploring the best docking solutions of an additional dimer from a POD-D step with different initial structure (one compatible with an isotropic growth, the second with an anisotropic growth), leads to a same common solution as illustrated in Figure 6.

To understand better these paths and investigate its electrostatic features, we attempted to discriminate the best docking solutions employing a different computing solution following the strategy developed by Barroso da Silva team, the PROCEEDpKa method [45]. The core idea of this approach is the identification of all titratable groups of the receptor that were perturbed by the visits and presence of the ligand. When the electrostatic coupling of these ionizable groups is not so strong, the *pKa* shifts can be used to identify the key amino acids responsible for a ligand-receptor association [45]. For the sake of comparison with the other simulation approaches and due to the larger size of the macromolecules involved in the present study, the PROCEEDpKa method was simplified to analyze only the charges shifts at solution pH 7.0 instead of the *pKa* shifts. We used this method to define if there was a correlation between the best docking results obtained using a constant-charge methodology and the electrostatic properties of these structures. We thus computed the shifts in charges of ionizable residues in the POD in the presence of a sixth dimer during a complexation study. This was compared with the charges obtained from a titration study in the absence of this sixth dimer. We first plotted these charges shifts near expected interfaces, with a cutoff chosen so as to exclude residues in nearby interfaces (supplementary Figure 2). Whatever the system (POD, POD-D or POD-D₂), we found no apparent differences between expected interfaces with this criterion (supplementary Figures 3 and 4). By one

side, this implies that the electrostatic coupling between the ionizable sites is quite strong for the POD system. It also indicates that both inner and superficial titratable groups can effectively participate in the interplay of the molecular interactions among the dimers. Indeed, the importance of electrostatics is well established experimentally for norovirus capsid dimers. For instance, the above-mentioned TR-SAXS study was performed with dimers dissociated at pH 9 and low salinity by a jump to pH 6 and salinity 55-105 mM [21]. We verified this phenomenon with the constant-pH (CpH) [43,45] simulation method, and found that dimer-dimer interactions are indeed computed as repulsive at pH 9 and $I = 10\text{mM}$ and as attractive at pH 6 and $I = 55\text{mM}$ (Figure 7). By another side, it hampers the practical application of this approach to probe and reveal the existence of any preferential binding spots for this particular system, if any.

At this stage, we searched to combine the calculated shifts in charges with the docking results differently. As shown on Figure 8A for the POD system, we mapped on the 3D structure the charged shifted amino acids and the center of mass of all representative docking positions for the new dimer calculated with ClusPro. VP1 dimer is composed of two structural domains, the N-terminal shell domain (S) making up the protein capsid and the C-terminal protruding domain (P) making up the exterior surface of the shell (supplementary Figure 5). If one draws a plane separating domains S from domains P (Figure 8A, bottom), it is apparent that the most shifted POD residues (darkest grays) are generally on the side of P domains, while all ClusPro docking solutions are on the side of S domains, be they correct (green) or incorrect (orange and red) solutions. Reasoning that electrostatic interactions are long range, and that there might be a strong coupling between titratable groups as mentioned above, we calculated, for each docking position i , a total electrostatic contribution of the POD residues Q_i based simply on Coulomb's law (i.e. proportional to the magnitude of each shift and inversely proportional to the distance squared). Interestingly, we observed on average that the docking solutions corresponding to correct addition of the next dimer (in green on Figure 8B) had higher Q_i than near misses (orange). The highest Q_i were computed for some particular docking positions in red Figure 8. Indeed, these last solutions correspond to positions where the dimer is docked in the center of the POD structure, near the 5-fold axis, which implies that the dimer is roughly equidistant from one considered amino acid whatever its

distance. Consequently, its contribution to the charge shift of one residue is equivalent for each dimer belonging to the POD structure. Definitely, the good docking positions (in green) have always a high Q_i value.

Thus, this combined analysis clearly discriminates between correct and near-correct solutions at least under the assumptions behind this work. However, although the electrostatic analysis can identify the key amino acids for the association, the studied structures' electrostatic variations are not sufficient to distinguish correct solutions from widely incorrect but geometrically central solutions for this system with such a strong electrostatic coupling among its ionizable groups. It demonstrates that the docking results we defined with our strategy (based on a constant-charge approach) could not be unambiguously reproduced with the sole mapping of the electrostatics features. This is also in accordance with the predominantly hydrophobic interfaces between domains S. Now, if we plotted the Q_i value versus the PIPER energy of ClusPro (Figure 9), we observed that the best energy docking positions have globally a high Q_i value. It is particularly interesting because it suggests that without knowledge of the final solution (crystal structure), it may be possible to discriminate docking solutions based on the energy values combined with the electrostatic properties of the molecules. This result is very encouraging and demonstrates that it must be possible to describe the assembly mechanisms of biological molecular complexes by combining different simulation methods.

Conclusion

Deciphering the assembly of macromolecular complexes remains a challenging question, but our work demonstrates that taking into account the dynamical properties of the molecules, both conformational changes and variations in physico-chemical properties induced by the interactions between two molecules, are essential to select the best solutions for docking. Indeed, based on *in silico* predictions combining docking, molecular dynamics and electrostatic properties, our results demonstrate that if an isotropic growth of the capsid, starting from the POD structure is favored, the anisotropic growth as experimentally identified by Tresset *et al.* is accessible with our strategy. In this latter case, the proposed intermediate composed of two PODs related by a dimer [21] is not compatible with our calculations.

So, we push our reasoning, based on our results, to propose an intermediate composed of 10 dimers (Figure 6) compatible with an anisotropic growth of a completed capsid. This hypothesis is finally very attractive because it induces that it is possible to complete a capsid with 9 intermediates formed of 10 dimers, without additional dimers. Furthermore, this solution remains compatible with the previous experimental observations. This approach, applied on the Norovirus capsid, can be used on other virus capsids or macromolecular complexes to study their assembly process.

Acknowledgements

This work was granted access to the high-performance computing resources of TGCC under allocation A0050710605, A0070710605 and A0090710605 made by GENCI. We also thank the CEA-CCRT infrastructure for giving us access to the supercomputer COBALT. J-CC was supported by a predoctoral fellowship from ANRS (France Recherche Nord & Sud SIDA-HIV Hépatites: FRENESH), AAP 2018-1. F.L.B.S. is also deeply thankful to the support given by “Fundação de Amparo à Pesquisa do Estado de São Paulo” (Fapesp 2020/07158-2), the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq 305393/2020-0) and the Swedish National Infrastructure for Computing (SNIC 2020/14-38).

Author Contributions

Conceptualization: SB, YB

Data curation: JCC

Formal analysis: JCC, SB, YB

Funding acquisition: SB, YB

Investigation: JCC, TT, SB, FLBS, YB

Methodology: FLBS, YB

Project administration: SB, YB

Writing – original draft: JCC, YB

Writing – review & editing: JCC, TT, SB, FLBS, YB

References

1. Staley JP, Woolford JL. Assembly of ribosomes and spliceosomes: complex ribonucleoprotein machines. *Current Opinion in Cell Biology*. 2009;21: 109–118. doi:10.1016/j.ceb.2009.01.003
2. Summers WC. Virus Infection. *Encyclopedia of Microbiology*. 2009; 546–552. doi:10.1016/B978-012373944-5.00323-0
3. Perlmutter JD, Hagan MF. Mechanisms of Virus Assembly. *Annual Review of Physical Chemistry*. 2015;66: 217–239. doi:10.1146/annurev-physchem-040214-121637
4. Caspar DL, Klug A. Physical principles in the construction of regular viruses. *Cold Spring Harb Symp Quant Biol*. 1962;27: 1–24. doi:10.1101/sqb.1962.027.001.005
5. Suiter CL, Quinn CM, Lu M, Hou G, Zhang H, Polenova T. MAS NMR of HIV-1 protein assemblies. *J Magn Reson*. 2015;253: 10–22. doi:10.1016/j.jmr.2014.12.009
6. Tubiana T, Boulard Y, Bressanelli S. Dynamics and asymmetry in the dimer of the norovirus major capsid protein. *PLOS ONE*. 2017;12: e0182056. doi:10.1371/journal.pone.0182056
7. Chevreuil M, Law-Hine D, Chen J, Bressanelli S, Combet S, Constantin D, et al. Nonequilibrium self-assembly dynamics of icosahedral viral capsids packaging genome or polyelectrolyte. *Nature Communications*. 2018;9: 3071. doi:10.1038/s41467-018-05426-8
8. Meng D, Hjelm RP, Hu J, Wu J. A Theoretical Model for the Dynamic Structure of Hepatitis B Nucleocapsid. *Biophys J*. 2011;101: 2476–2484. doi:10.1016/j.bpj.2011.10.002
9. Zlotnick A, Johnson JM, Wingfield PW, Stahl SJ, Endres D. A Theoretical Model Successfully Identifies Features of Hepatitis B Virus Capsid Assembly. *Biochemistry*. 1999;38: 14644–14652. doi:10.1021/bi991611a
10. Hagan MF. Modeling Viral Capsid Assembly. *Adv Chem Phys*. 2014;155: 1–68. doi:10.1002/9781118755815.ch01
11. Zlotnick A. Theoretical aspects of virus capsid assembly. *Journal of Molecular Recognition*. 2005;18: 479–490. doi:10.1002/jmr.754
12. Prevelige PE, Thomas D, King J. Nucleation and growth phases in the polymerization of coat and scaffolding subunits into icosahedral procapsid shells. *Biophys J*. 1993;64: 824–835.
13. Perlmutter JD, Perkett MR, Hagan MF. Pathways for virus assembly around nucleic acids. *J Mol Biol*. 2014;426: 3148–3165. doi:10.1016/j.jmb.2014.07.004
14. Rapaport DC. Molecular dynamics study of T = 3 capsid assembly. *J Biol Phys*. 2018;44: 147–162. doi:10.1007/s10867-018-9486-7
15. Wang B, Zhang J, Wu Y. A Multiscale Model for the Self-Assembly of Coat Proteins in Bacteriophage MS2. *J Chem Inf Model*. 2019;59: 3899–3909. doi:10.1021/acs.jcim.9b00514
16. Grime JMA, Dama JF, Ganser-Pornillos BK, Woodward CL, Jensen GJ, Yeager M, et al. Coarse-grained simulation reveals key features of HIV-1 capsid self-assembly. *Nature Communications*. 2016;7: 11568. doi:10.1038/ncomms11568

17. Perkett MR, Hagan MF. Using Markov state models to study self-assembly. *J Chem Phys.* 2014;140: 214101. doi:10.1063/1.4878494
18. Reguera D, Hernández-Rojas J, Llorente JMG. Kinetics of empty viral capsid assembly in a minimal model. *Soft Matter.* 2019;15: 7166–7172. doi:10.1039/C9SM01593K
19. Hadden JA, Perilla JR, Schlicksup CJ, Venkatakrishnan B, Zlotnick A, Schulten K. All-atom molecular dynamics of the HBV capsid reveals insights into biological function and cryo-EM resolution limits. *eLife.* 2018 [cited 19 Jul 2018]. doi:10.7554/eLife.32478
20. Prasad BVV, Hardy ME, Dokland T, Bella J, Rossmann MG, Estes MK. X-ray Crystallographic Structure of the Norwalk Virus Capsid. *Science.* 1999;286: 287–290.
21. Tresset G, Le Coeur C, Bryche J-F, Tatou M, Zeghal M, Charpilienne A, et al. Norovirus Capsid Proteins Self-Assemble through Biphasic Kinetics via Long-Lived Stave-like Intermediates. *J Am Chem Soc.* 2013;135: 15373–15381.
22. Berke JM, Dehertogh P, Vergauwen K, Damme EV, Mostmans W, Vandyck K, et al. Capsid Assembly Modulators Have a Dual Mechanism of Action in Primary Human Hepatocytes Infected with Hepatitis B Virus. *Antimicrobial Agents and Chemotherapy.* 2017;61. doi:10.1128/AAC.00560-17
23. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics.* 2016;54: 5.6.1-5.6.37. doi:10.1002/cpbi.3
24. Ausar SF, Foubert TR, Hudson MH, Vedvick TS, Middaugh CR. Conformational stability and disassembly of Norwalk virus-like particles. Effect of pH and temperature. *J Biol Chem.* 2006;281: 19478–19488. doi:10.1074/jbc.M603313200
25. Tresset G, Decouche V, Bryche J-F, Charpilienne A, Le Cœur C, Barbier C, et al. Unusual self-assembly properties of Norovirus Newbury2 virus-like particles. *Arch Biochem Biophys.* 2013;537: 144–152. doi:10.1016/j.abb.2013.07.003
26. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX.* 2015;1–2: 19–25.
27. Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink S-J. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J Chem Theory Comput.* 2008;4: 819–834.
28. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, de Vries AH. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J Phys Chem B.* 2007;111: 7812–7824. doi:10.1021/jp071097f
29. Periole X, Cavalli M, Marrink S-J, Ceruso MA. Combining an Elastic Network With a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition. *J Chem Theory Comput.* 2009;5: 2531–2543.
30. Siuda I, Thøgersen L. Conformational flexibility of the leucine binding protein examined by protein domain coarse-grained molecular dynamics. *J Mol Model.* 2013;19: 4931–4945.
31. Hinsen K. Analysis of domain motions by approximate normal mode calculations. *Proteins.* 1998;33: 417–429.

32. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*. 1984;81: 3684–3690.
33. Parrinello M, Rahman A. Crystal Structure and Pair Potentials: A Molecular-Dynamics Study. *Phys Rev Lett*. 1980;45: 1196–1199.
34. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph*. 1996;14: 33–38, 27–28.
35. Tubiana T, Carvaille J-C, Boulard Y, Bressanelli S. TTClust: A Versatile Molecular Simulation Trajectory Clustering Program with Graphical Summaries. *J Chem Inf Model*. 2018;58: 2178–2182. doi:10.1021/acs.jcim.8b00512
36. Jr JHW. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*. 1963;58: 236–244. doi:10.1080/01621459.1963.10500845
37. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2001;63: 411–423. doi:https://doi.org/10.1111/1467-9868.00293
38. Wassenaar TA, Pluhackova K, Böckmann RA, Marrink SJ, Tieleman DP. Going Backward: A Flexible Geometric Approach to Reverse Transformation from Coarse Grained to Atomistic Models. *J Chem Theory Comput*. 2014;10: 676–690. doi:10.1021/ct400617g
39. Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, et al. The ClusPro web server for protein-protein docking. *Nat Protocols*. 2017;12: 255–278. doi:10.1038/nprot.2016.169
40. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*. 2006;65: 392–406. doi:10.1002/prot.21117
41. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. 2015.
42. Srivastava D, Santiso E, Gubbins K, Barroso da Silva FL. Computationally Mapping pKa Shifts Due to the Presence of a Polyelectrolyte Chain around Whey Proteins. *Langmuir*. 2017;33: 11417–11428. doi:10.1021/acs.langmuir.7b02271
43. Barroso da Silva FL, MacKernan D. Benchmarking a Fast Proton Titration Scheme in Implicit Solvent for Biomolecular Simulations. *J Chem Theory Comput*. 2017;13: 2915–2929. doi:10.1021/acs.jctc.6b01114
44. Poveda-Cuevas SA, Etchebest C, Barroso da Silva FL. Insights into the ZIKV NS1 Virology from Different Strains through a Fine Analysis of Physicochemical Properties. *ACS Omega*. 2018;3: 16212–16229. doi:10.1021/acsomega.8b02081
45. Poveda-Cuevas SA, Etchebest C, Barroso da Silva FL. Identification of Electrostatic Epitopes in Flavivirus by Computer Simulations: The PROCEEDpKa Method. *J Chem Inf Model*. 2020;60: 944–963. doi:10.1021/acs.jcim.9b00895

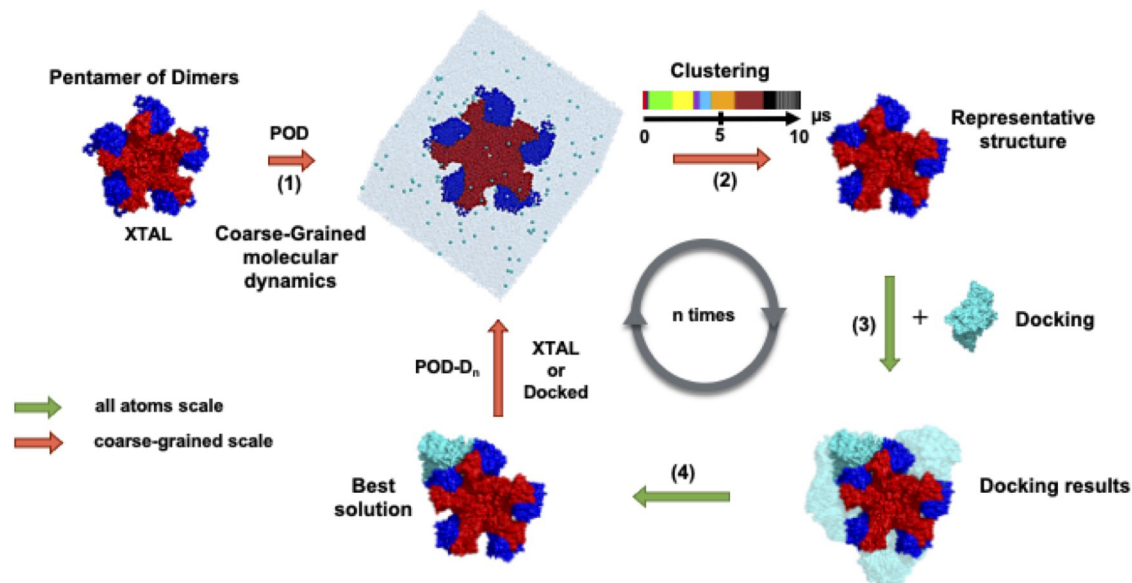


Figure 1. Computational strategy for studying the growth of the Norovirus capsid. The starting structure used is a pentamer of dimers (POD). For each dimer, chain A is colored in red and chain B in blue. Dimers added during the docking trial are colored in cyan. Our strategy has 4 steps: (1) the all-atom model is converted to a coarse-grained model, (2) a coarse-grained molecular dynamics simulation is performed, a barplot is produced during clustering step and it correspond to the distribution of structures in clusters (colored in red to grey) along the coarse-grained simulation, (3) docking of a new dimer is completed, (4) the best solution is retained to initiate a new cycle. The molecular dynamics simulation and clustering steps are at a coarse-grained scale symbolized by orange arrows. The docking step is performed at an all-atom scale (green arrows).

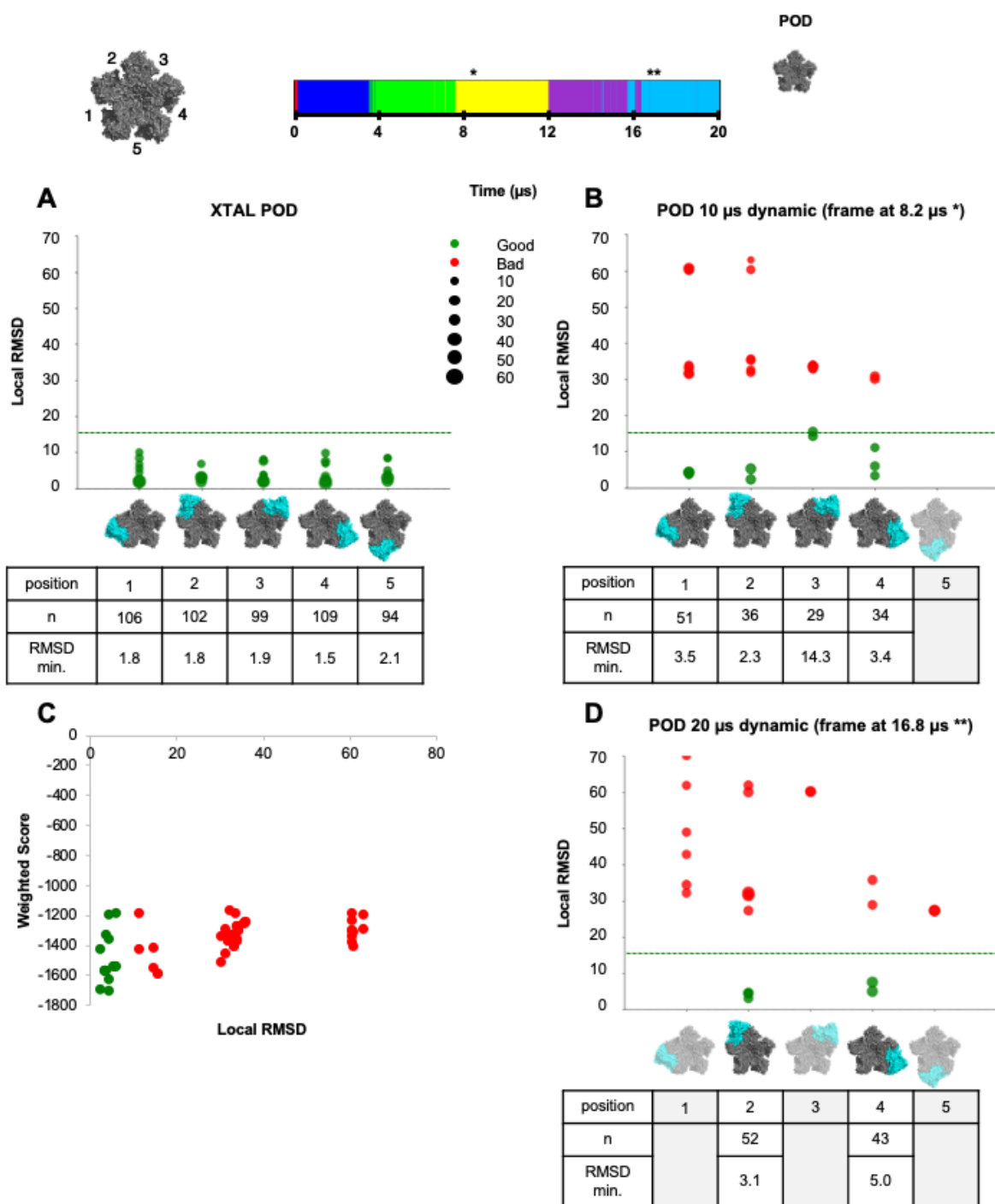


Figure 2. Docking solutions onto POD structure. Top panel, the timeline of the simulation. The asterisks denote the representative frames of two transiently stable conformations in a 10 μ s coarse-grained molecular dynamics simulation (*) or in the same simulation extended to 20 μ s (**). **(A)** Docking results onto the POD structure extracted from the X-ray crystal **(B)** Docking onto the representative structure after 8.2 μ s. **(D)** Docking onto the representative structure after 16.8 μ s. Each docking solution output by ClusPro was checked against each of the five possible positions represented below the plots, both visually and by computing a local RMSD (see materials and methods). The solution was then assigned to the position with the minimum local RMSD. The horizontal dashed line at 16 Å indicates a cutoff beyond which the docking solution does not match an acceptable solution, while it always does below this cutoff (bad solutions are in red circles and good solutions are in green circles). The sizes of circles are proportional to the number of ClusPro members for each docking solution. The tables report the total numbers of good solutions for each position 1 to 5. **(C)** PIPER weighted score computed by ClusPro as a function of local RMSD.

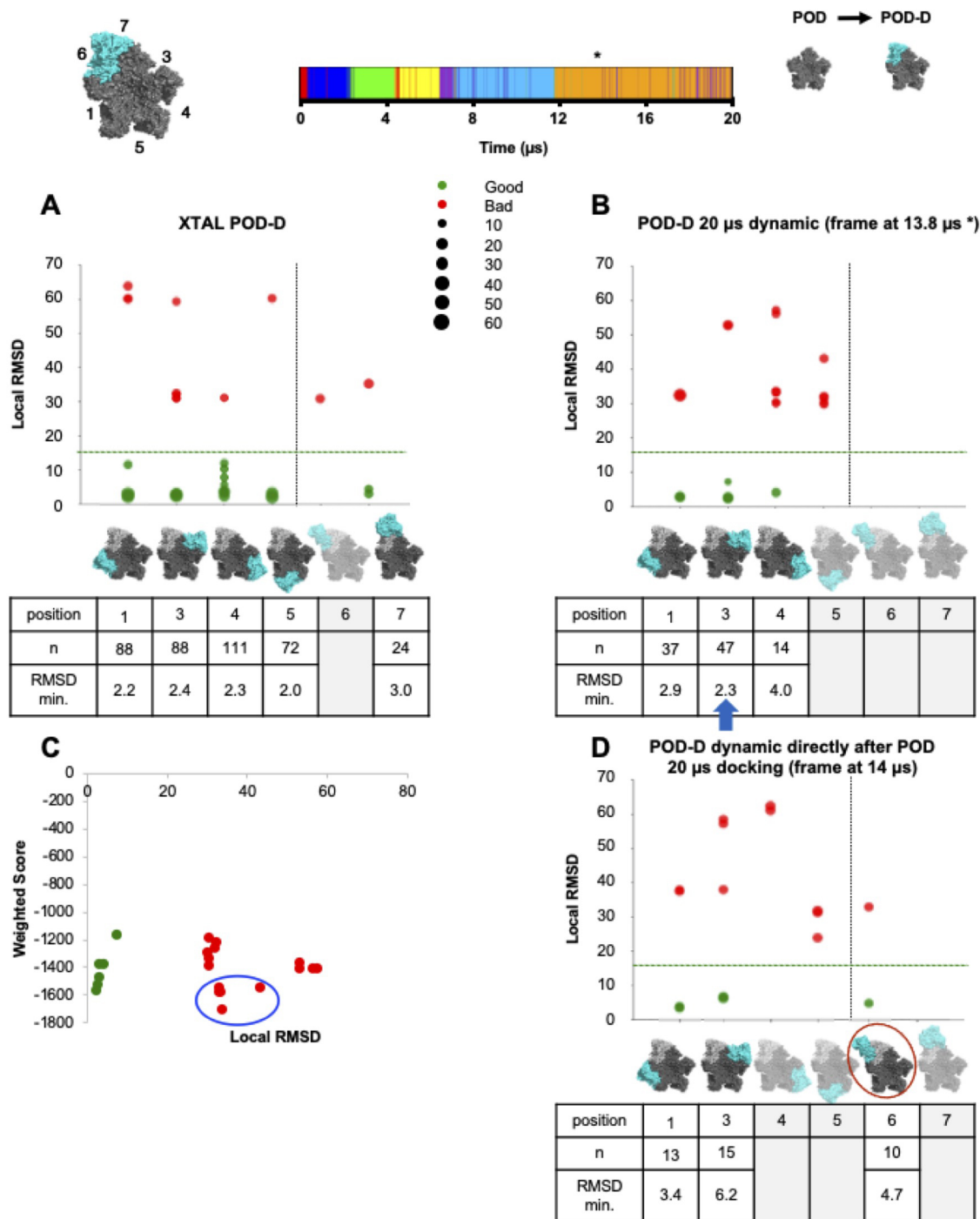


Figure 3. Docking solutions onto POD-D structure. Vertical lines separate isotropic growth from anisotropic growth. (A) Docking results onto the POD-D structure extracted from the X-ray crystal (B) Docking results onto the representative structure after 13.8 μ s coarse-grained simulation. The best solution (blue arrow) will be used to study isotropic growth. (C) PIPPER weighted score computed by ClusPro as a function of local RMSD. (D) Docking results onto the best previous docking solution selected after a 10 μ s simulation (POD docking solution at position 2 - Fig 2B) which was used to initiate a second 10 μ s molecular dynamics. The docking solution at position 6 (surrounded by a red circle) which corresponds to an anisotropic growth was selected to explore anisotropic path.

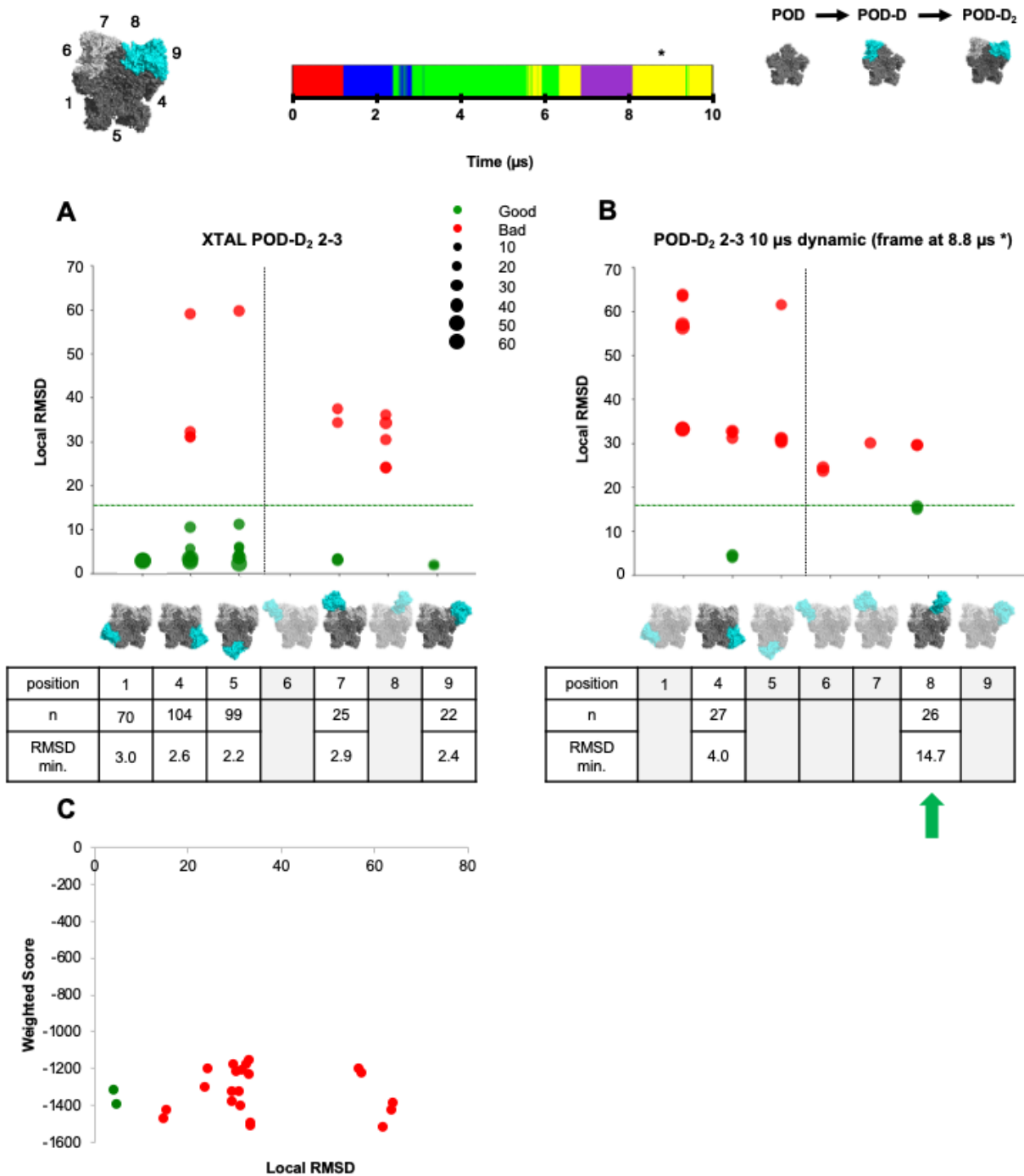
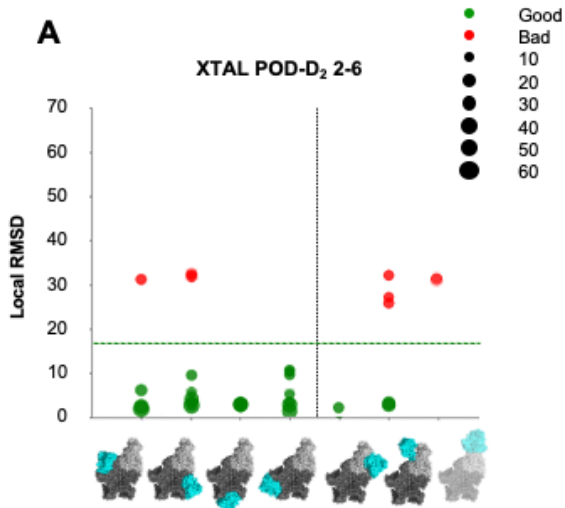
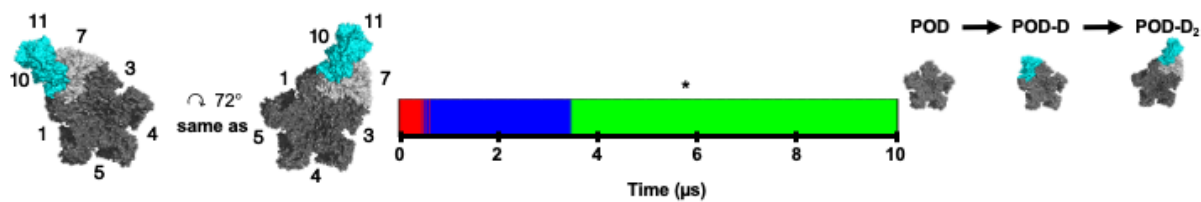
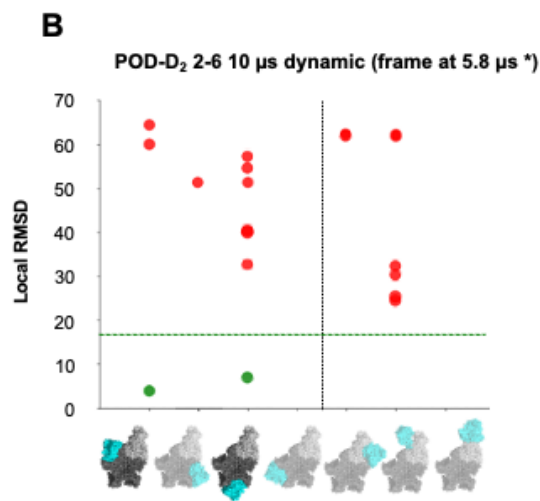


Figure 4. Docking solutions onto POD-D₂ 2-3 structure. (A) Docking results onto the POD-D₂ 2-3 structure extracted from the X-ray crystal or (B) onto the representative structure after 8.8 μs simulation. Vertical lines separate isotropic from anisotropic docking solutions. The green arrow represents a common docked structure between adjacent and continued POD-D₂. (C) PIPPER weighted score computed by ClusPro as a function of local RMSD.



position	1	3	4	5	7	10	11
n	80	109	82	108	11	44	
RMSD min.	1.9	3.0	3.0	1.7	2.5	2.6	



position	1	3	4	5	7	10	11
n	12		12				
RMSD min.	3.9		7.1				

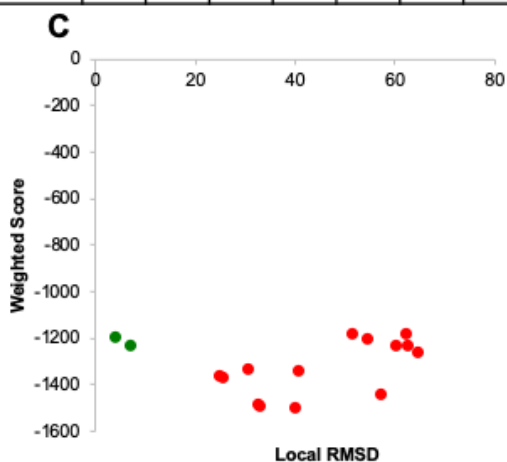
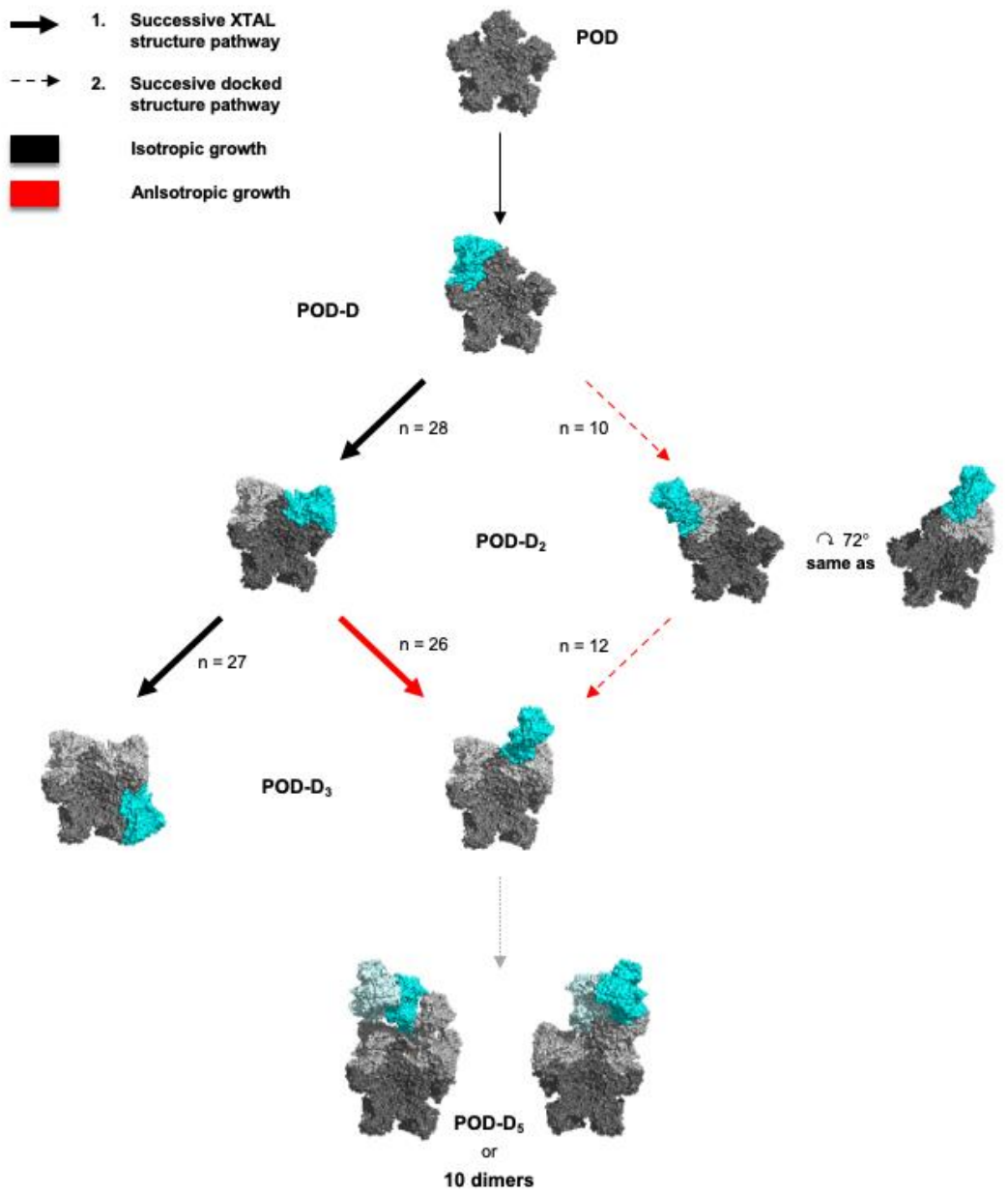


Figure 5. Docking solutions onto POD-D₂ 2-6 structure. (A) Docking results onto the POD-D₂ 2-6 structure extracted from the X-ray crystal or (B) onto the representative structure after 5.8 μ s simulation. Vertical lines separate isotropic from anisotropic docking solutions. The green arrow represents a common docked structure between adjacent and continued POD-D₂. (C) PIPPER weighted score computed by ClusPro as a function of local RMSD.



W

Figure 6. Growth pathways from the POD structure resulting from our docking strategy. The number n corresponds to the number of solutions at each step. By extrapolation, POD-D₅ corresponds to model compatible with the elongated assembly intermediate identified by Tresset *et al.* during norovirus capsid self-assembly study [21].

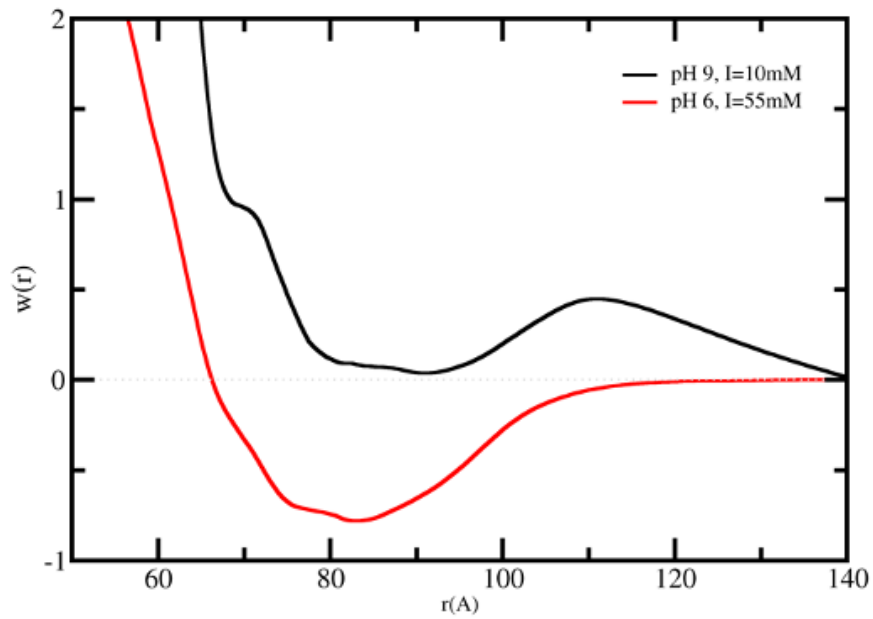


Figure 7. Free energy profiles for the interaction of the VP1 dimer homoassociation. The simulated free energy of interactions [$w(r)$] between the centers of mass of the two macromolecules at different physical chemical conditions (pH 9 and I=10mM; pH 6 and I=55mM) were obtained from CpH MC simulations [43,45]. These simulations started with the two macromolecules placed at random orientation and separation distance. Temperature was fixed at 298K. Free energies are given in $k_B T$ units (k_B is the Boltzmann constant).

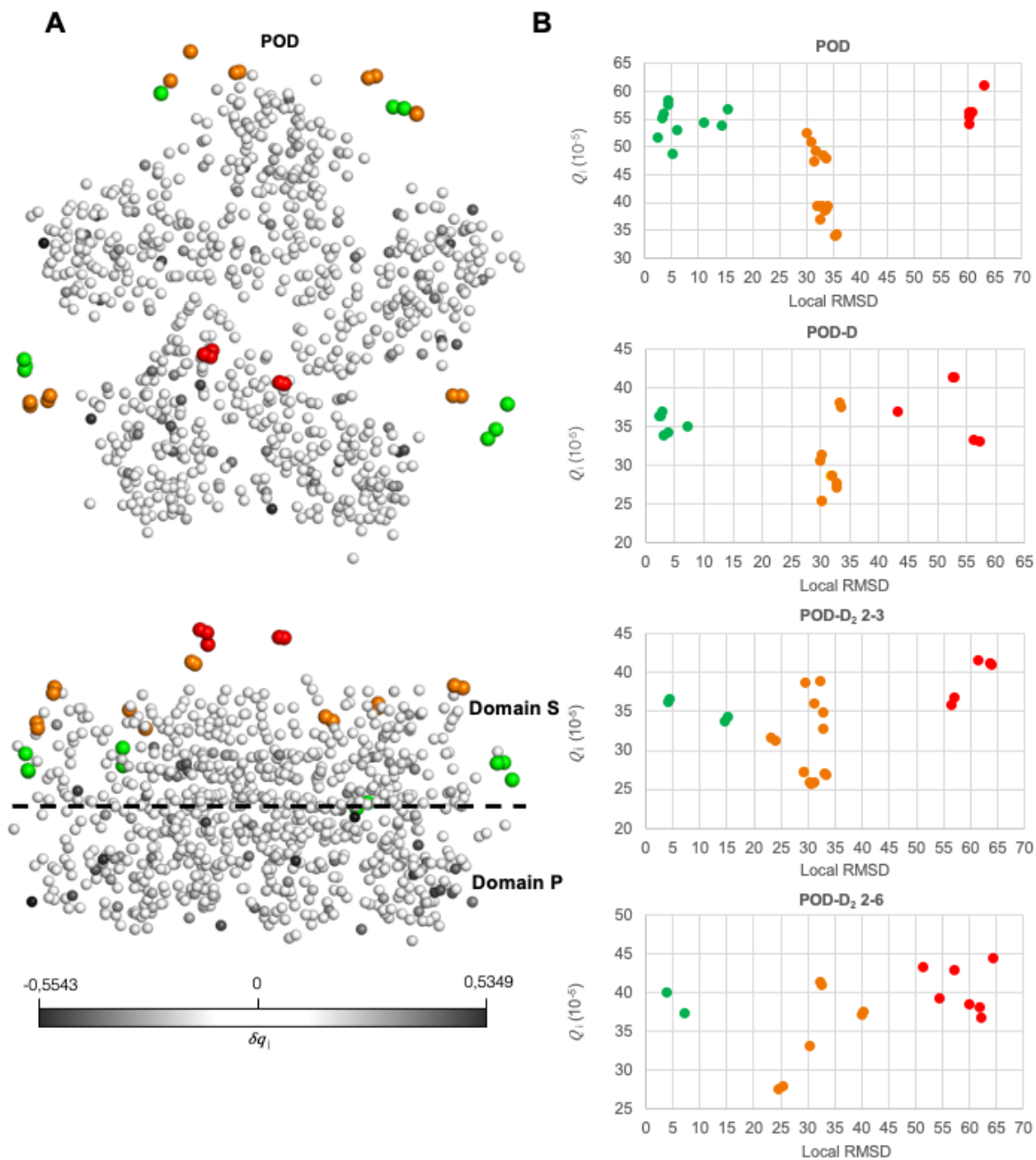


Figure 8. Relationship between docking solutions and titratable residues. (A) Localization of center of mass of docking solutions and titratable residues. (B) Total electrostatic contribution as a function of RMSD. For $i=1 \dots n$ docked dimers, and $j=1 \dots m$ titratable residues, each with a shift δq_j , we computed:

$$Q_i = \sum_{j=1}^m \frac{|\delta q_j|}{r_{ij}^2}$$

where r_{ij} is the distance between i^{th} dimer's center of mass and the j^{th} titratable residue's center of mass. The green spheres correspond to correctly docked solutions. Orange and red spheres correspond to incorrectly docked solutions.

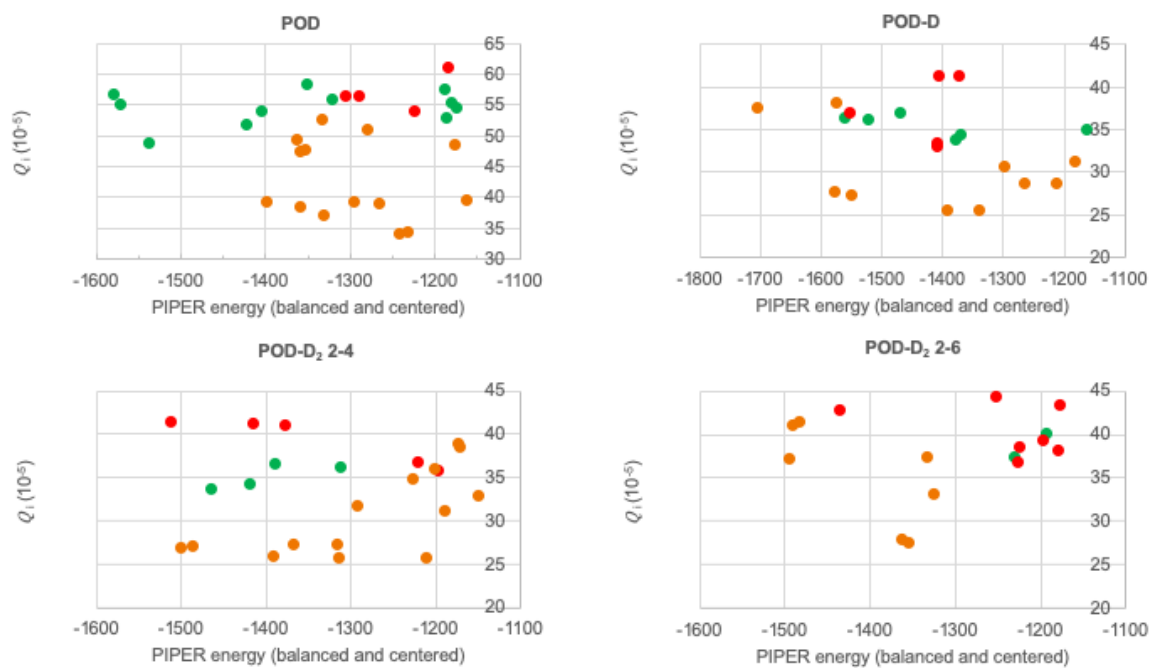


Figure 9. PIPPER weighted score computed by ClusPro as a function of total electrostatic contribution for POD, POD-D and POD-D₂ systems. The green spheres correspond to correctly docked solutions. Orange and red spheres correspond to incorrectly docked solutions.

Supporting Information

Supplementary Figure 1.

Examples of good and bad docking solutions. POD, POD-D, POD-D₂ 2-3 and POD-D₂ 2-6 are colored in grey (POD in dark grey and additional dimers in light grey). Good docking solutions are shown in green and bad docking solutions in red (domains S are in light green or light red and domains P are in dark green or dark red).

Supplementary Figure 2.

Impact of dimer approach on ionizable residues of the receptor. Charge shifts were calculated on POD, POD-D, POD-D₂ systems as illustrated in (a, b and c). Charges are firstly computed on the system without ligand (column “Reference shift (q)”) and secondly in presence of a new dimer (column “MC shift (q)”) in (d). It allows to calculate the shift of ionizable residues (column “Shifts difference (q)”) in (d).

$$q_{shift} = q_{reference} - q_{MC}$$

To identify significant shifts, we decided to use positive and negative averages as threshold (e). q_{shift} upper positive average and below negative average are considered as significant shifts. To discriminate and to quantify charge shifting at a specific interface of docking positions, we calculated both distributions of the number of ionizable residues and of significantly shifted residues as a function of the distance from the center of mass of the concerned dimers in the X-ray structure of the capsid (f). It allows us to select a radius of 60 Å around a docked dimer’s center of mass as a good limit to define the docking interface (g). Finally, violin plots (it combines boxplot and density plot) were used to visualize our data (h). cf. **Suppl. Fig 3 and 4.**

Supplementary Figure 3.

Violin plots of ionizable shifted residues at the POD or POD-D docking interface. The violin plot displays the distribution of the shifted charges of the ionizable residues. It contains a boxplot that indicates the mean in the middle and the standard deviation at extremities. The blue point corresponds to the mean of positive shifted charges and the red point corresponds to mean of negative shifted charges. Docking positions selected to define capsid growth are circled in red.

Supplementary Figure 4.

Violin plots of ionizable shifted residues at the POD-D₂ docking interface. The violin plot displays the distribution of the shifted charges of the ionizable residues. It contains a boxplot that indicates the mean in the middle and the standard deviation at extremities. The blue point corresponds to the mean of positive shifted charges and the red point corresponds to mean of negative shifted charges. Docking positions selected to define capsid growth are circled in red.

Supplementary Figure 5.

Structure of VP1 dimer. Domains S (shell, residues 20 to 220) considered as assembly domains are colored in orange, and domains P (protruding, residues 230 to 520) domain are colored in brown. The domains are connected by an interstitial loop in green.

Supplementary Figure 6.

RMSD time series. RMSD vs time of POD, POD-D, POD-D₂ 2-3 and POD-D₂ 2-6 structures are represented in black, brown, grey and indigo respectively. All dynamics are characterized by 20 μs of production time, except for POD-D₂ 2-6 (17 μs).

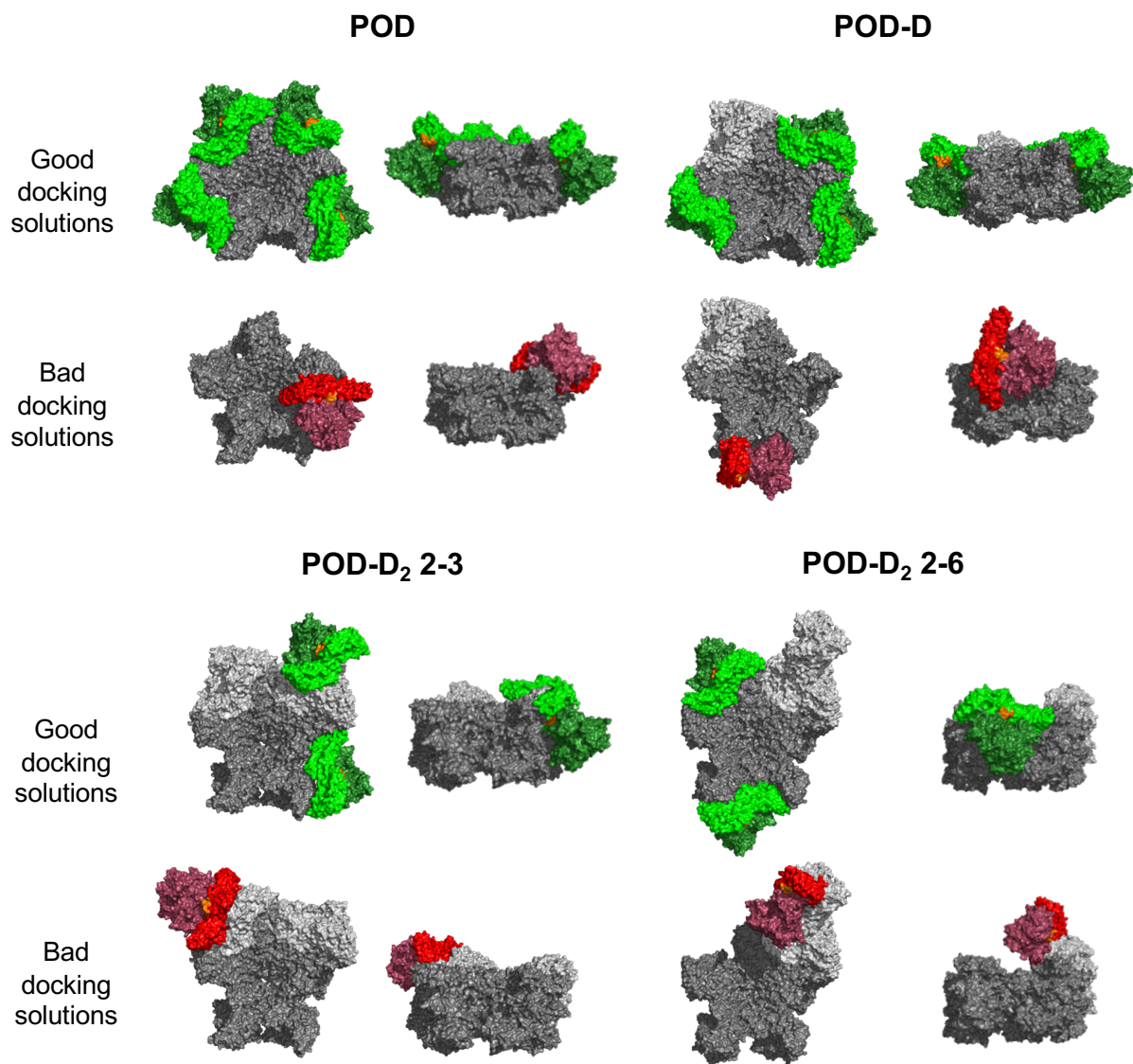


Figure S1

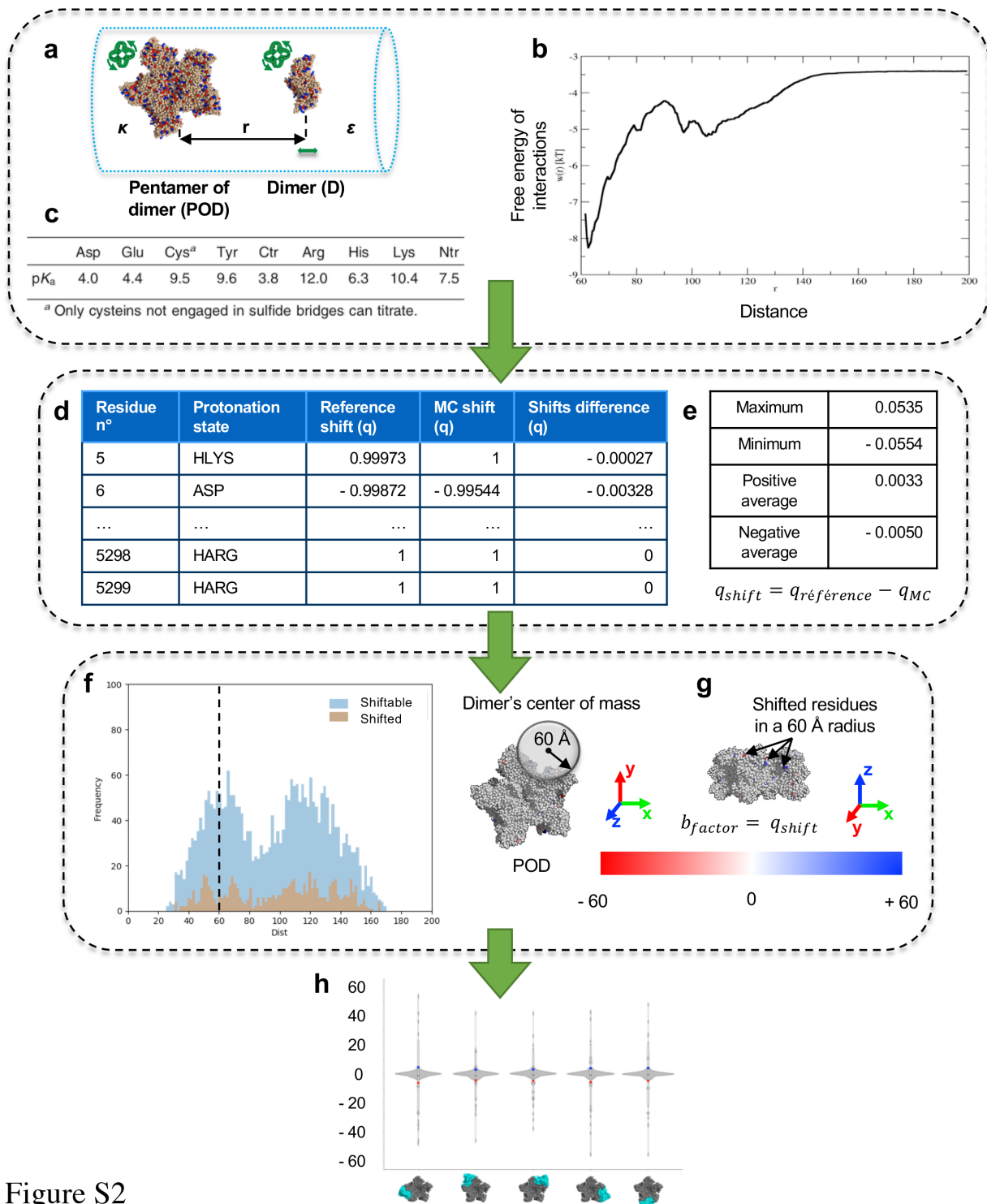


Figure S2

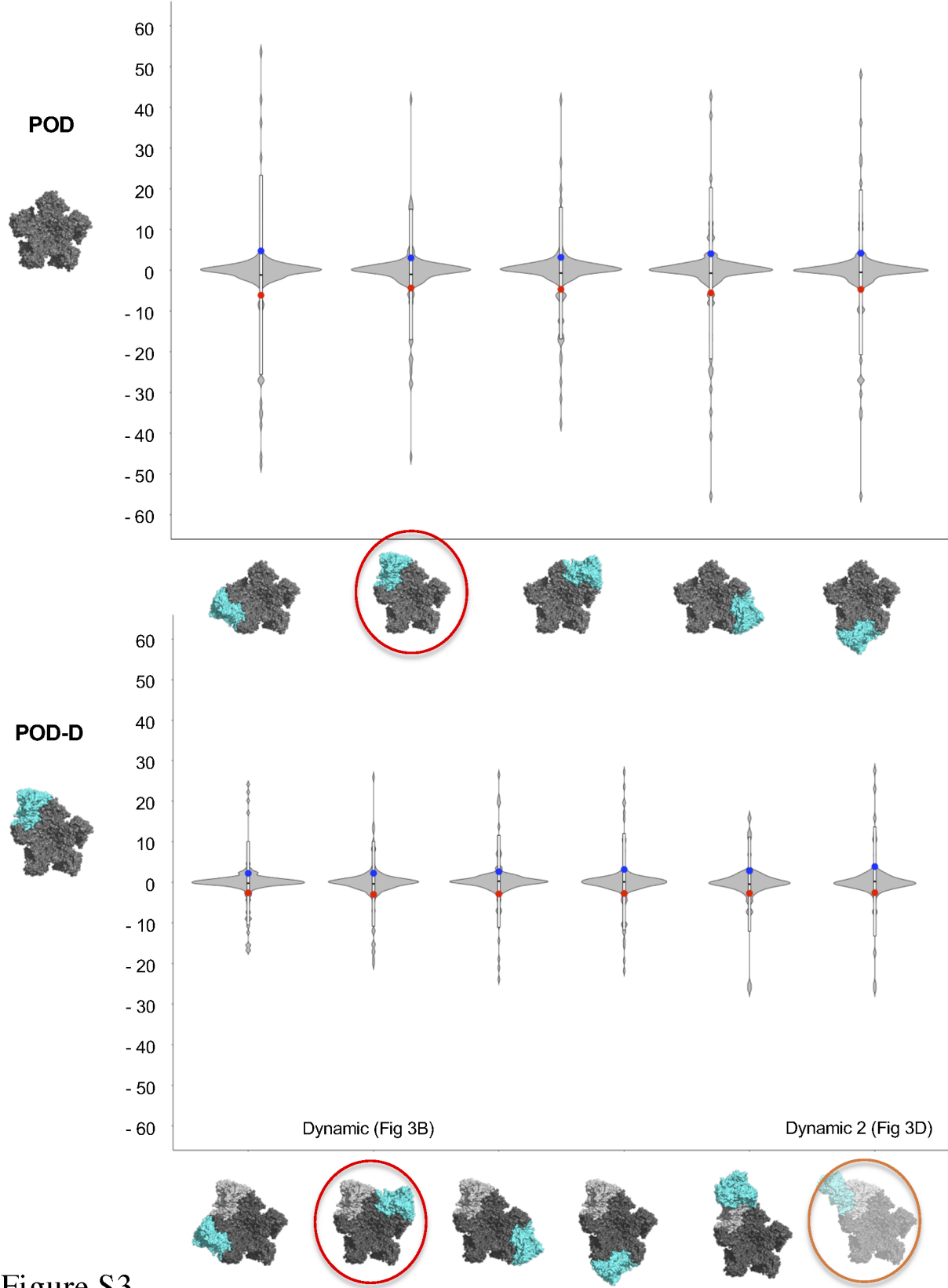
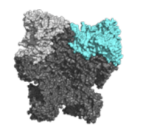
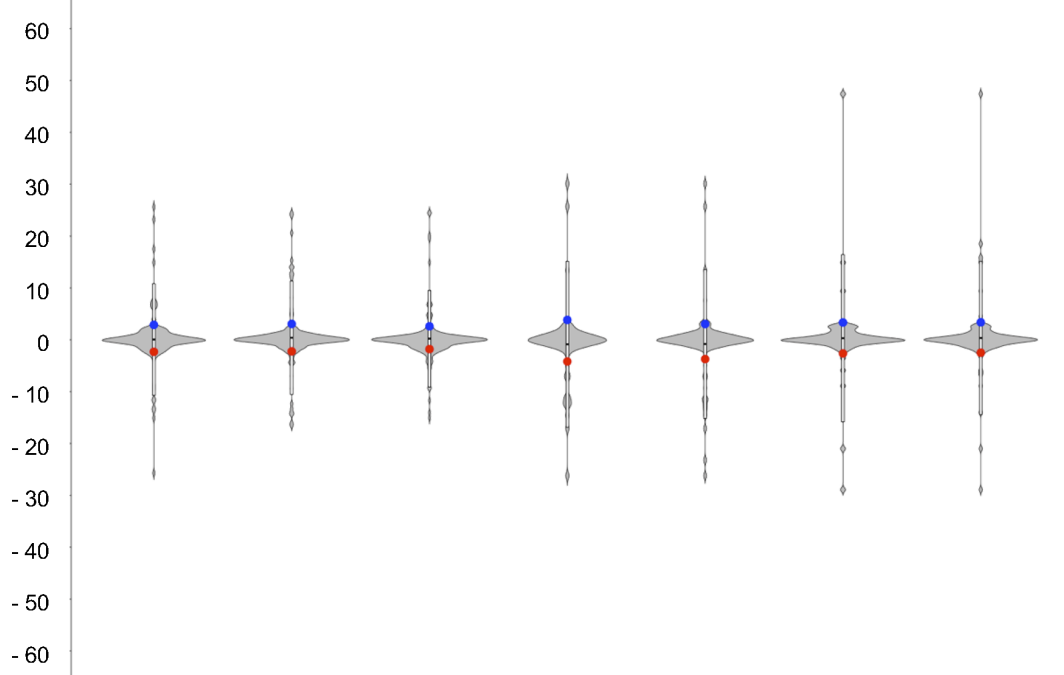


Figure S3

POD-D₂ 2-3



Dynamic (Fig 3B)



POD-D₂ 2-6



Dynamic 2 (Fig 3D)

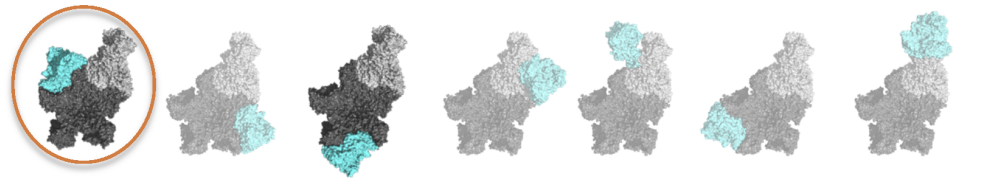
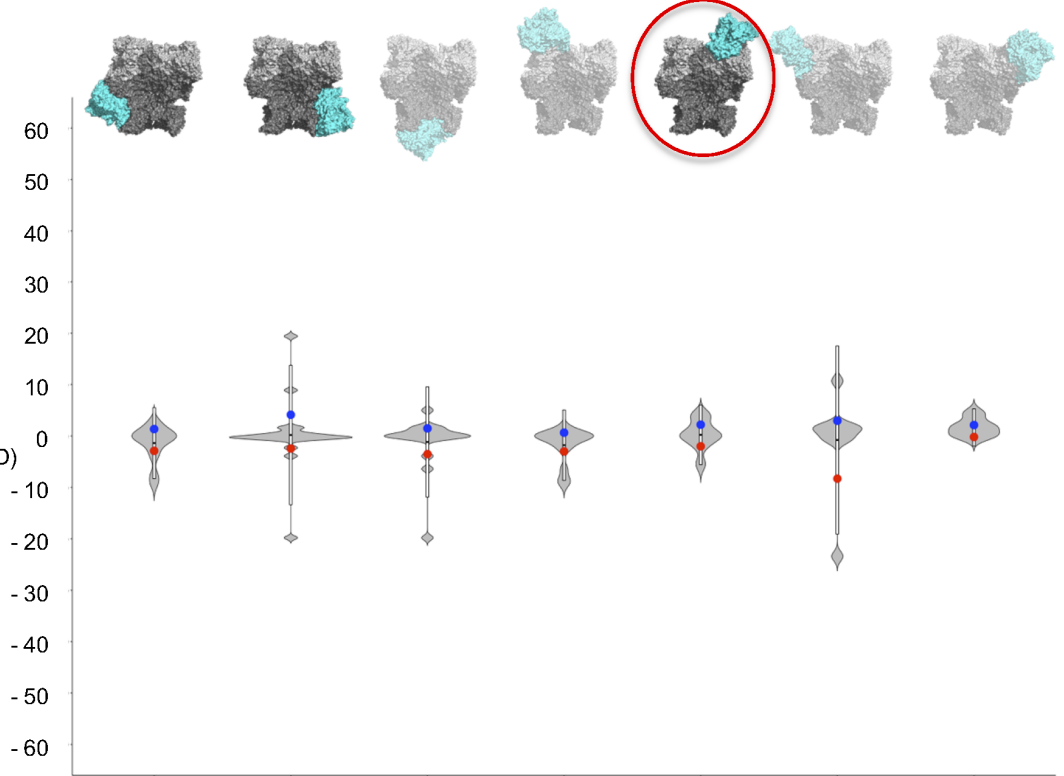


Figure S4

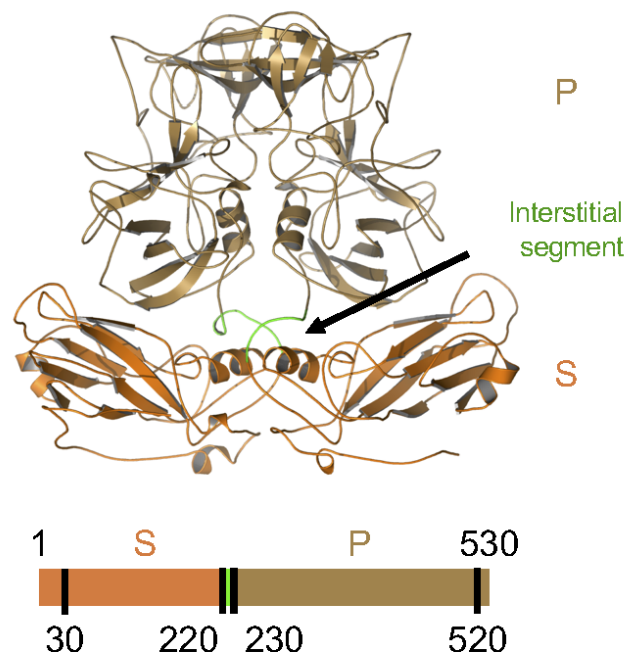


Figure S5

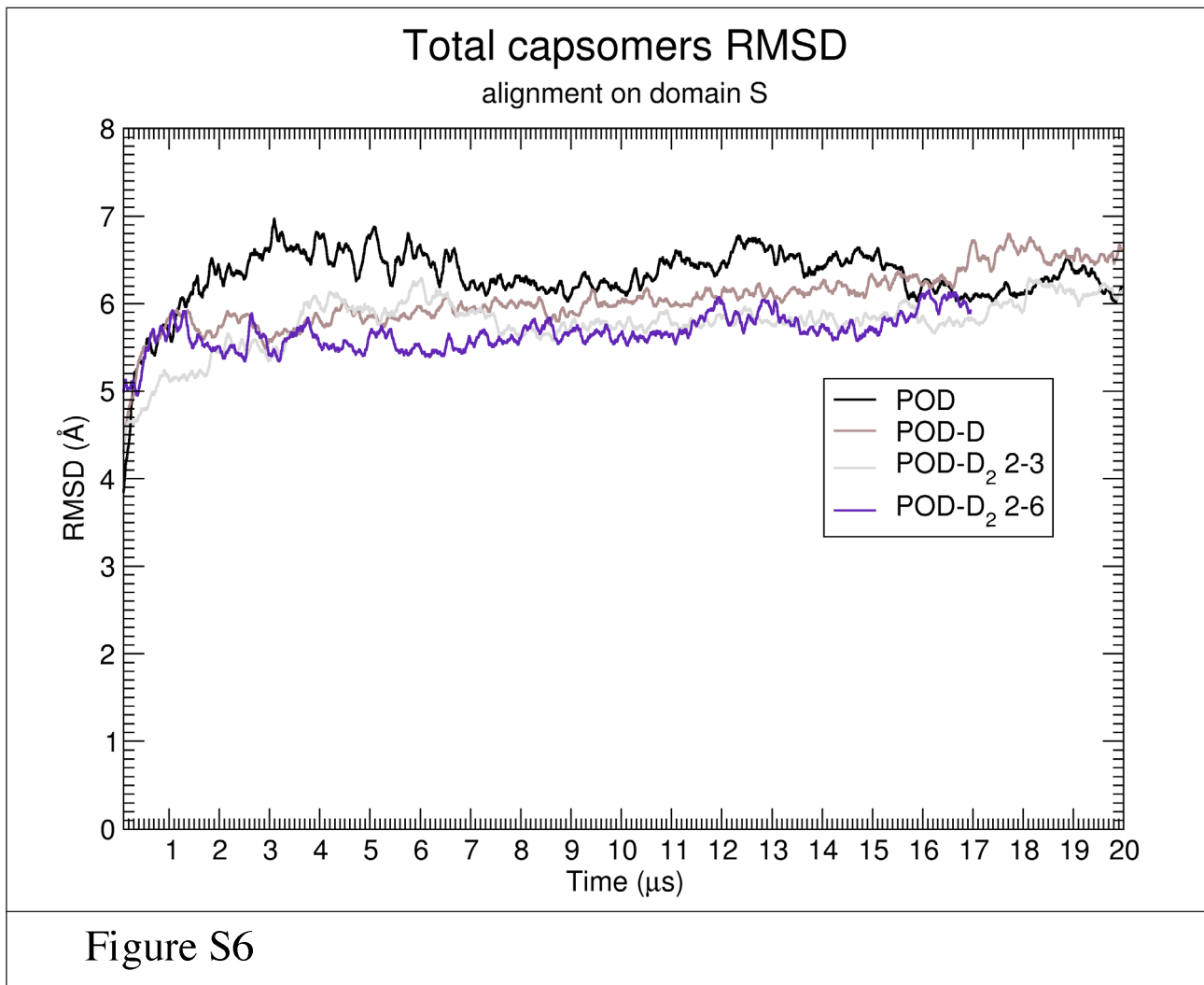


Figure S6

Les résultats obtenus dans cet article ont permis de comprendre une partie des interactions mises en jeu lors de la formation d'un intermédiaire théorique. La dynamique des POD(- D_n) influence clairement les résultats d'amarrage. Les propriétés électrostatiques des systèmes pourraient expliquer, partiellement, les résultats d'amarrage. Dans l'article, l'étude est arrêtée à l'étape $n = 2$, avec 2 POD- D_3 possibles (Figure 25) : (1) Un POD- D_3 2-3-4 qui correspond à l'amarrage de 3 dimères C-C sur le pourtour du POD, amarrés aux positions 2, 3 et 4 de façon adjacente et (2) un POD- D_3 1-2-6. Pour celui-ci, le premier amarrage est sur le pourtour du POD (position 2). Le second amarrage s'effectue sur le premier dimère amarré (position 6) et le dernier se situe sur le pourtour du POD, à côté du premier dimère amarré. Ces deux solutions correspondent à des objets qui poussent dans une direction privilégiée.

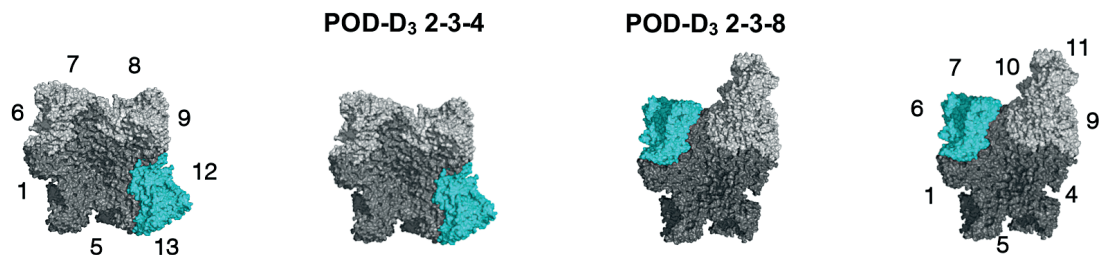


Figure 25 – POD- D_3 issus de notre stratégie. Le POD est gris foncé, les deux premiers dimères amarrés sont gris clair. Le dernier dimère amarré est cyan.

Après la soumission de cet article, des analyses complémentaires sont effectuées. Elles consistent à réaliser un nouveau cycle complet (simulation, regroupement de structures (clustering), amarrage (docking)...), puis à comparer le docking d'un dimère A-B ou C-C sur le POD, POD-D₂ ou POD-D₃ et enfin à étudier brièvement l'espace conformationnel du dimère VP1.

5.4.7 Complément à l'article

Lors de son stage, Sella Detchanamourtty a réalisé une étape supplémentaire d'amarrage d'un dimère C-C sur les 2 POD-D₃ (Figure 25). Ces 2 POD-D₃ ont été construits à partir de la même stratégie. L'itération supplémentaires à partir des 2 POD-D₃, conduit à 2 POD-D₄. En haut à gauche des figures de l'itération des POD-D₃ (Figures 25 et 26), la numérotation des interfaces est illustrée. On peut également voir la représentation des groupes majoritaires le long de l'axe du temps. L'astérisque qui figure sur l'axe des temps, correspond à une structure représentative du POD-D₃ utilisée pour l'amarrage.

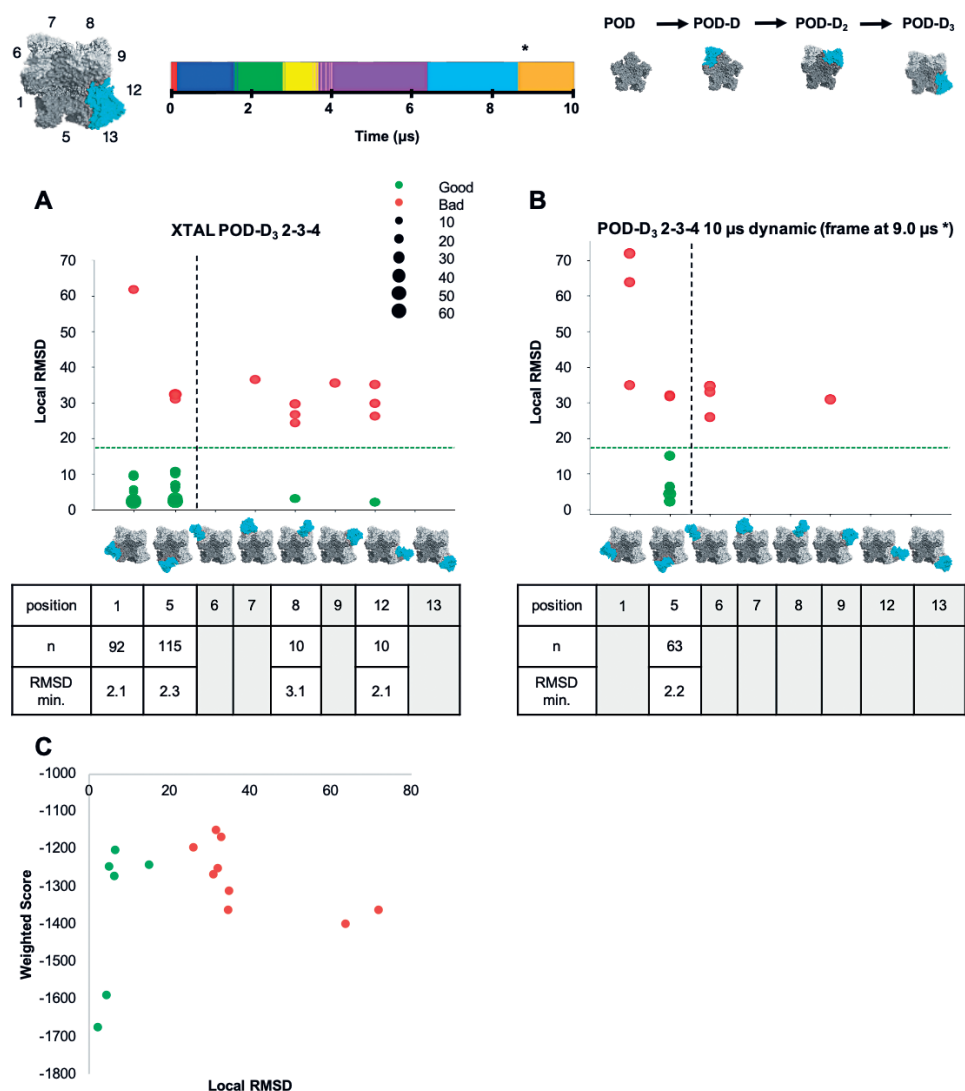


Figure 26 – Amarrage d'un dimère sur le POD-D₃ 2-3-4 cristallographique (A) ou après 10 µs de simulation (B). Haut de la figure : Représentation des groupes majoritaire le long de l'axe du temps du POD-D₃ 2-3-4. L'en-tête correspond à la timeline issue de l'étape de regroupement de structures de la trajectoire du POD-D₃ 2-3-4. Les lignes verticales séparent les solutions d'amarrage sur le pourtour du POD des autres solutions. Les lignes horizontales indiquent le seuil (16 Å) qui sépare les solutions bien classées des moins bien classées. (C) Représentation du Score ClusPro des solutions d'amarrage en fonction du RMSD local calculé.

Le POD-D₃ 2-3-4 correspond à la solution d'amarrage la mieux classée à l'itération sur le POD-D₂ 2-3.

Un dimère est amarré sur les deux dernières interfaces (positions 1 et 5 dans la Figure 26A) du pourtour du POD et sur deux autres interfaces (positions 8 et 12 dans la Figure 26A) du POD-D₃ 2-3-4 cristallographique. Ces dernières correspondent à des interfaces faisant intervenir des dimères déjà amarrés sur le POD initial. Le POD-D₃ 2-3-4, après simulation, a des solutions d'amarrage bien classées sur une seule interface (position 5 dans la Figure 26B). Cette interface est localisée sur le pourtour du POD. Les dimères sont mal amarrés aux positions 1, 6 et 9 et aucun dimère n'est amarré aux positions 7, 8, 12 et 13.

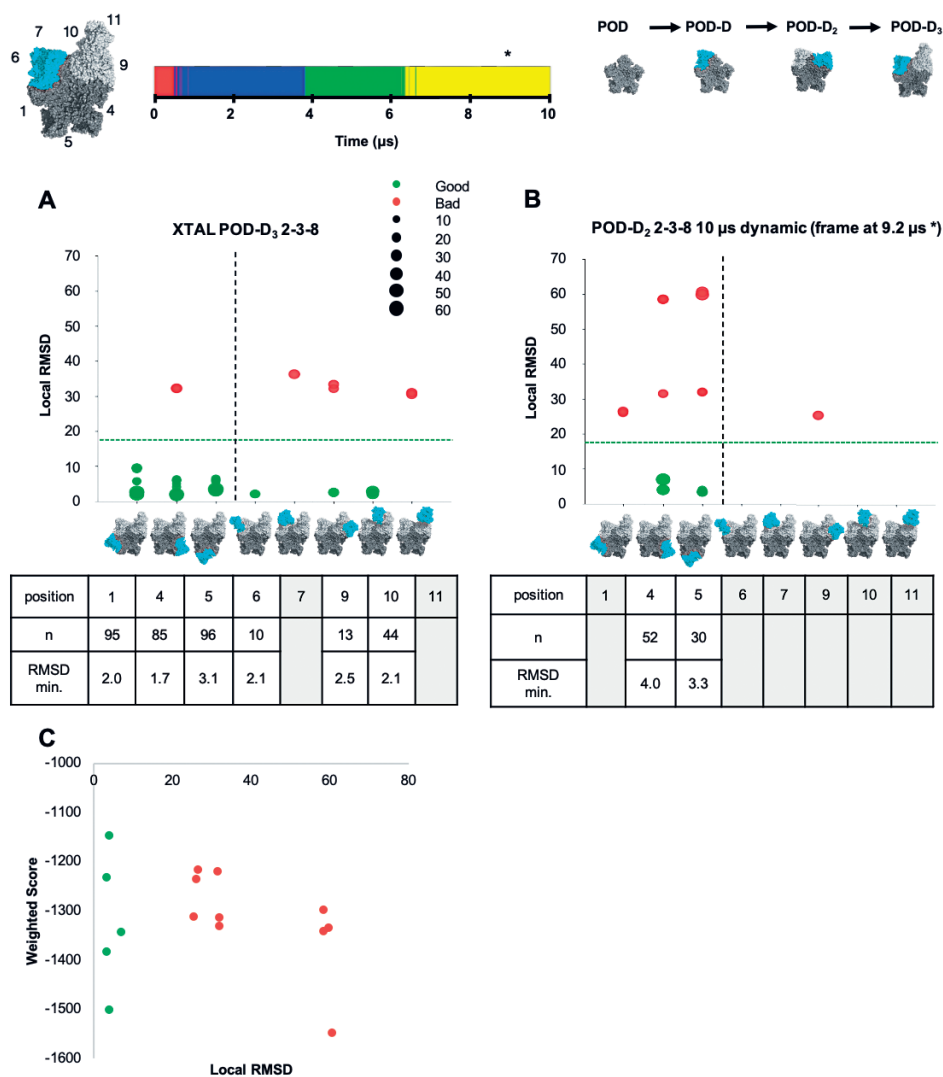


Figure 27 – Amarrage d'un dimère sur le POD-D₃ 2-3-8. (A) Cristallographique (B) Après 10 µs de simulation. (C) Représentation du Score ClusPro des solutions d'amarrage en fonction du RMSD local calculé.

Le POD-D₃ simulé a moins de solutions d'amarrage, bien classées ou moins bien classées, que le POD-D₃ cristallographique. On a pu voir, précédemment, que le score ClusPro n'est pas suffisant pour sélectionner la meilleure solution d'amarrage en terme de classement (section 5.4.6).

À cette itération en particulier, il aurait été possible de sélectionner une solution bien classée (Figure 26C).

Concernant le POD-D₃ 2-3-8 cristallographique, des dimères sont amarrés sur les dernières interfaces (positions 1, 4 et 5 dans la Figure 27A) libres du pourtour du POD, et sur les 3 autres interfaces (positions 6, 9 et 10 dans la Figure 29A), en dehors du pourtour du POD. Les 3 dernières interfaces sont localisées sur les dimères déjà amarrés sur le POD initial et induisent une croissance dans une direction privilégiée. Après MD du POD-D₃ 2-3-8, des dimères sont seulement amarrés sur 2 des 3 interfaces (positions 4 et 5 dans la Figure 29B) restantes de la bordure du POD. Ici encore, les résultats montrent que le POD-D₃ simulé a moins de solutions, bien ou moins bien classées, que le POD-D₃ cristallographique. Le meilleur score ClusPro correspond à une solution d'amarrage moins favorable en terme de classement lors de cette itération. Cela confirme, une fois de plus, que le score ClusPro n'est pas suffisant pour sélectionner les solutions d'amarrage bien classées. Les meilleures solutions de docking pour le POD-D₃ 2-3-4 et POD-D₃ 2-3-8 ne convergent pas vers le même POD-D₄ mais sont situées sur la même interface (position 5).

5.4.7.1 Résumé de l'ensemble de l'étude d'assemblage du POD à l'intermédiaire

À la suite des itérations du POD au POD-D₂ figurant dans l'article, une étape supplémentaire a été effectuée pour les deux POD D₃ résultants (Figure 29). Les résultats d'amarrage diffèrent entre un POD(-D_n) cristallographique et un POD simulé. De manière générale et de façon attendue, il y a plus de solutions de docking sur les structures cristallographiques. Cela indique que la dynamique a un effet sur l'amarrage d'une nouvelle unité d'assemblage. Les solutions d'amarrage du POD-D₃ 2-3-8, après simulation, peuvent être associées à une croissance anisotrope, tandis que celles du POD-D₃ 2-3-4, correspondent à une croissance isotrope (Figure 29).

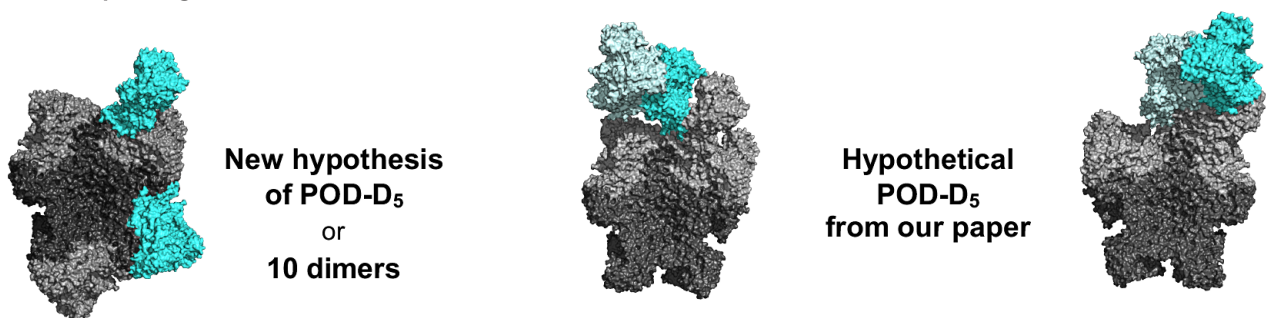


Figure 28 – Intermédiaires d'assemblage compatibles avec notre stratégie. À gauche, nouveau POD-D₅ plausible à partir des résultats de l'itération des POD-D₃. Au milieu et à droite, POD-D₅ plausibles issus de notre article.

3 décamères de dimères ont été construits, 2 issus de l'article et 1 issu de l'itération POD-D₄ (Figure 28). Ce dernier est proposé comme solution commune à partir des 2 POD-D₄ (Figure 28 - New POD-D₅). Il faut chercher à savoir si ces 3 POD-D₅ sont compatibles avec l'enveloppe obtenue par TR-SAXS [38]. Pour le déterminer, les structures tout-atomes des POD-D₅ sont superposées sur l'enveloppe TR-SAXS.

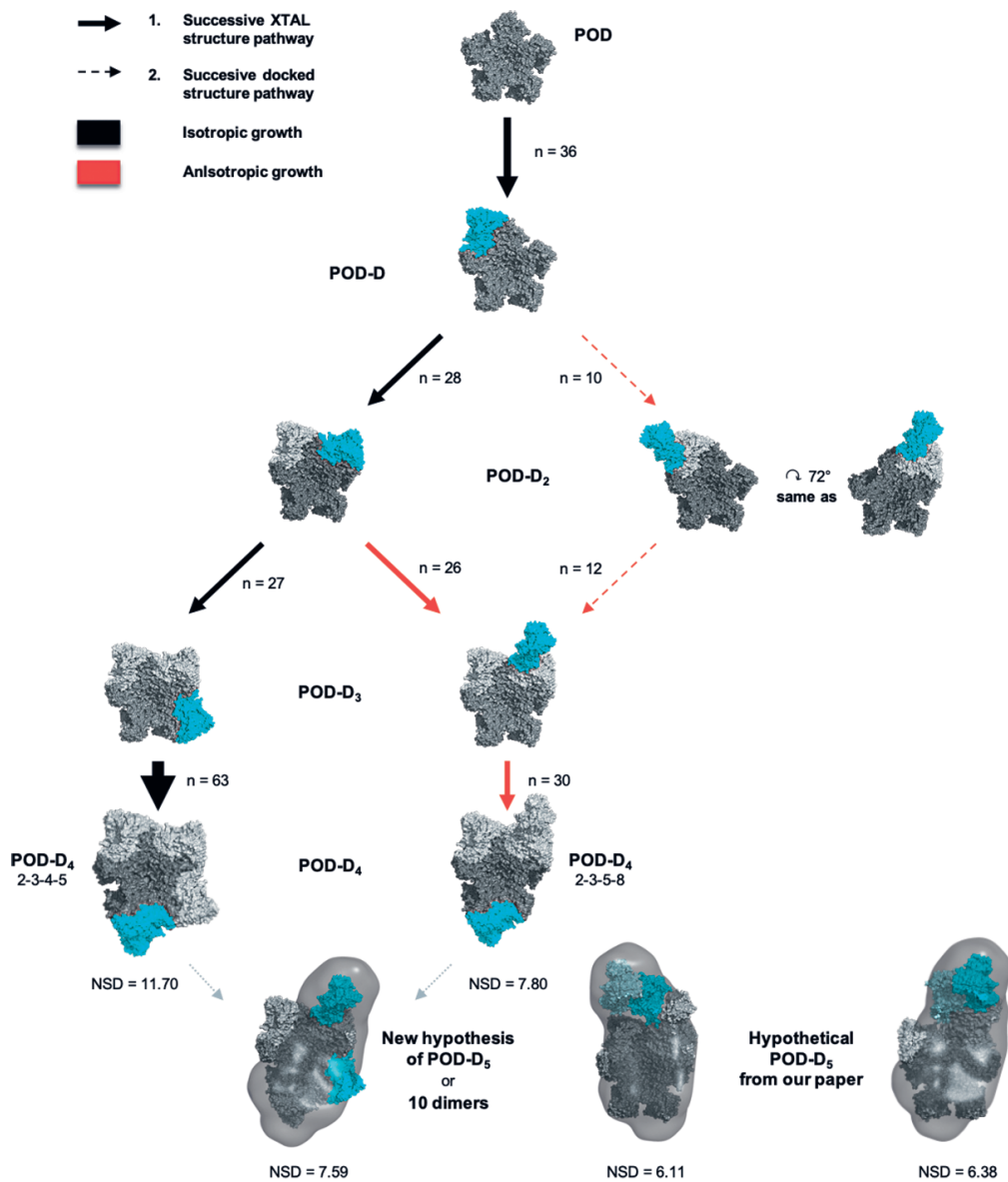


Figure 29 – Résumé des chemins d’assemblage obtenus à partir de notre stratégie, du POD vers le POD-D₅ hypothétique. L’épaisseur des flèches est proportionnelle au nombre de solutions n d’amarrage menant à l’intermédiaire suivant. Hormis le POD-D₅, toutes les solutions sont issues de la stratégie d’étude d’assemblage. NSD est la mesure de similarité entre les POD-D₄ et les POD-D₅ avec l’enveloppe TR-SAXS [38].

Ce processus est réalisé avec le programme SUPCOMB 23 [45] de la suite ATSAS 3.0.3 1. Cet outil considère chaque structure comme un ensemble de points. Il calcule la plus petite divergence spatiale normalisée (NSD) pour trouver la meilleure superposition entre 2 modèles. La NSD correspond donc à une mesure quantitative de similarité entre 2 ensembles de points tridimensionnels (3D).

Le POD-D₅ qui résulterait des deux POD-D₄, se caractérise par une NSD de 7,59 (“New hypothesis of POD-D₅” dans la Figure 29). Les POD-D₅ proposés dans l’article ont des NSD calculées, qui sont égales à 6,11 et 6,38 (“Hypothetical POD-D₅ from our paper” dans la Figure 29). Pour comparer les deux POD-D₄ et identifier celui qui est le plus compatible

avec l'enveloppe TR-SAXS, la NSD est utilisée. Le POD-D₄ 2-3-4-5 a une NSD égale à 11,70 et le POD-D₄ 2-3-5-8 a une NDS de 7,80. On peut donc considérer que le POD-D₄ 2-3-5-8 est davantage compatible avec l'enveloppe. C'est aussi le cas pour les POD-D₅ proposés dans l'article, qui ont une NSD de ~6,25. Nous pensons que l'enveloppe pourrait correspondre à plusieurs intermédiaires d'assemblage (Figure 29 – "New POD-D₅" et "Hypothetical POD-D₅ from our paper"). Le type d'intermédiaire d'assemblage ne serait donc pas unique.

5.4.8 Influence du dimère amarré A-B vs C-C sur l'assemblage

Les résultats qui sont présentés dans l'article et dans la section correspondent à l'ajout successif de dimères C C sur le POD de départ, jusqu'au POD-D₄ ainsi formé. Le dimère C-C utilisé, correspond à la conformation des dimères en interaction avec un POD cristallographique (PDB ID : 1IHM). La structure du POD de départ jusqu'au POD-D₄ a évolué lors des simulations. Certaines des interfaces des POD(-D_n) cristallographiques sont en interaction avec un dimère A-B. Notre stratégie a donc été suivie à l'identique, mais cette fois ci, en utilisant un dimère A-B lors des étapes d'amarrage (Figures 30, 31, 32 et 33).

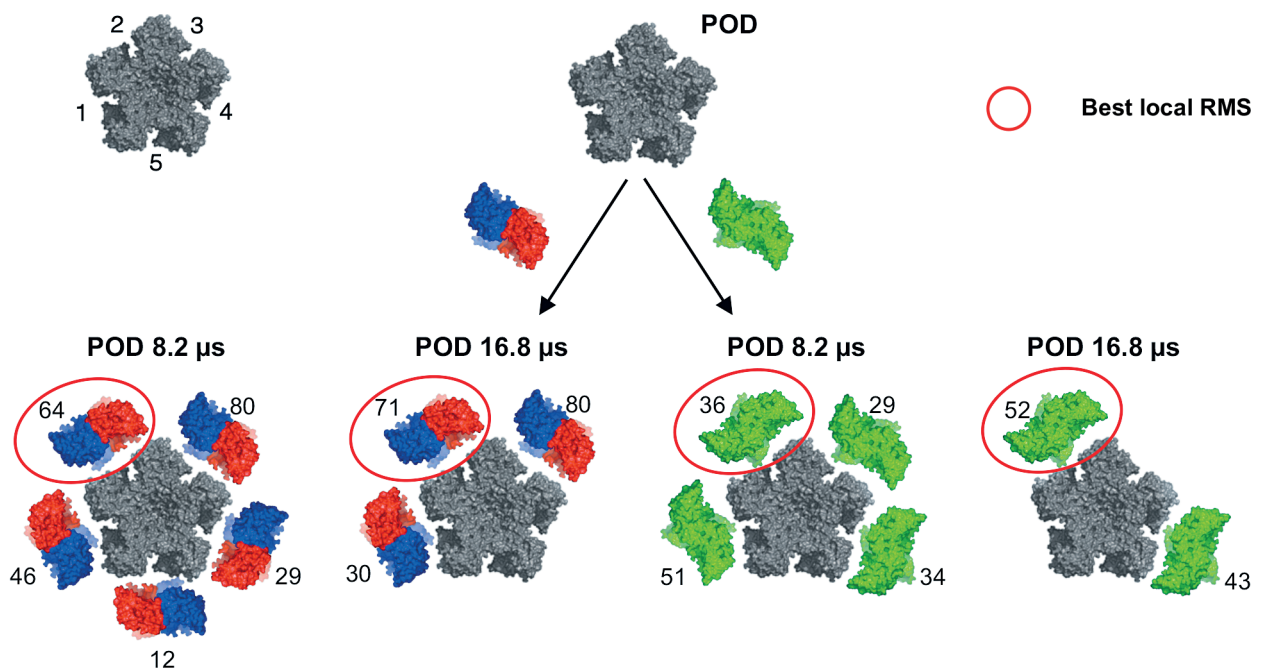


Figure 30 – Comparaison de l'amarrage d'un dimère A-B ou C-C sur le POD. Les interfaces d'amarrage sont numérotées de 1 à 5. Les temps en microsecondes figurent pour les structures représentatives utilisées. Les nombres au-dessus des solutions d'amarrage correspondent aux nombres de membres. Les solutions entourées sont les mieux classées en termes de RMSD.

Les amarrages du dimère A-B sur le POD, à 8,2 μs et 16,8 μs, admettent des solutions sur une interface supplémentaire, en comparaison avec les amarrages du dimère C-C (Figure 30). Il s'agit respectivement des positions 5 et 3. En termes de RMSD, les meilleures solutions d'amarrage pour le POD, à 8,2 et 16,8 μs, quel que soit le dimère amarré, sont situées sur la même interface (position 2). De manière générale, l'amarrage du dimère A-B sur la structure

représentative du POD à 10 μ s (POD 8,2 μ s), est possible sur toutes les interfaces. La perte de symétrie du POD pendant les 10 μ s n'influe pas sur l'amarrage du dimère A-B. Pour les POD-D, il existe aussi des solutions sur une interface supplémentaire lorsque l'on amarre un dimère A-B (Figure 31). Elles sont sur le pourtour du POD et correspondent respectivement aux positions 5 du POD-D ("XTAL") et 4, du POD-D ("successive"). En ce qui concerne les RMSD, les meilleures solutions pour le POD-D ("XTAL") se trouvent, pour le dimère A-B, à la position 1 et pour le dimère C-C, à la position 3. Ces positions d'amarrage sont équivalentes car les POD-D₂, obtenus dans les deux cas, sont identiques. Ils sont symétriques (rotation de 72°). Les meilleures solutions de docking du POD-D ("successive") sont situées sur la même interface (position 1). Pour les deux dimères (A-B ou C-C), une solution de docking en dehors du pourtour du POD apparaît (position 6).

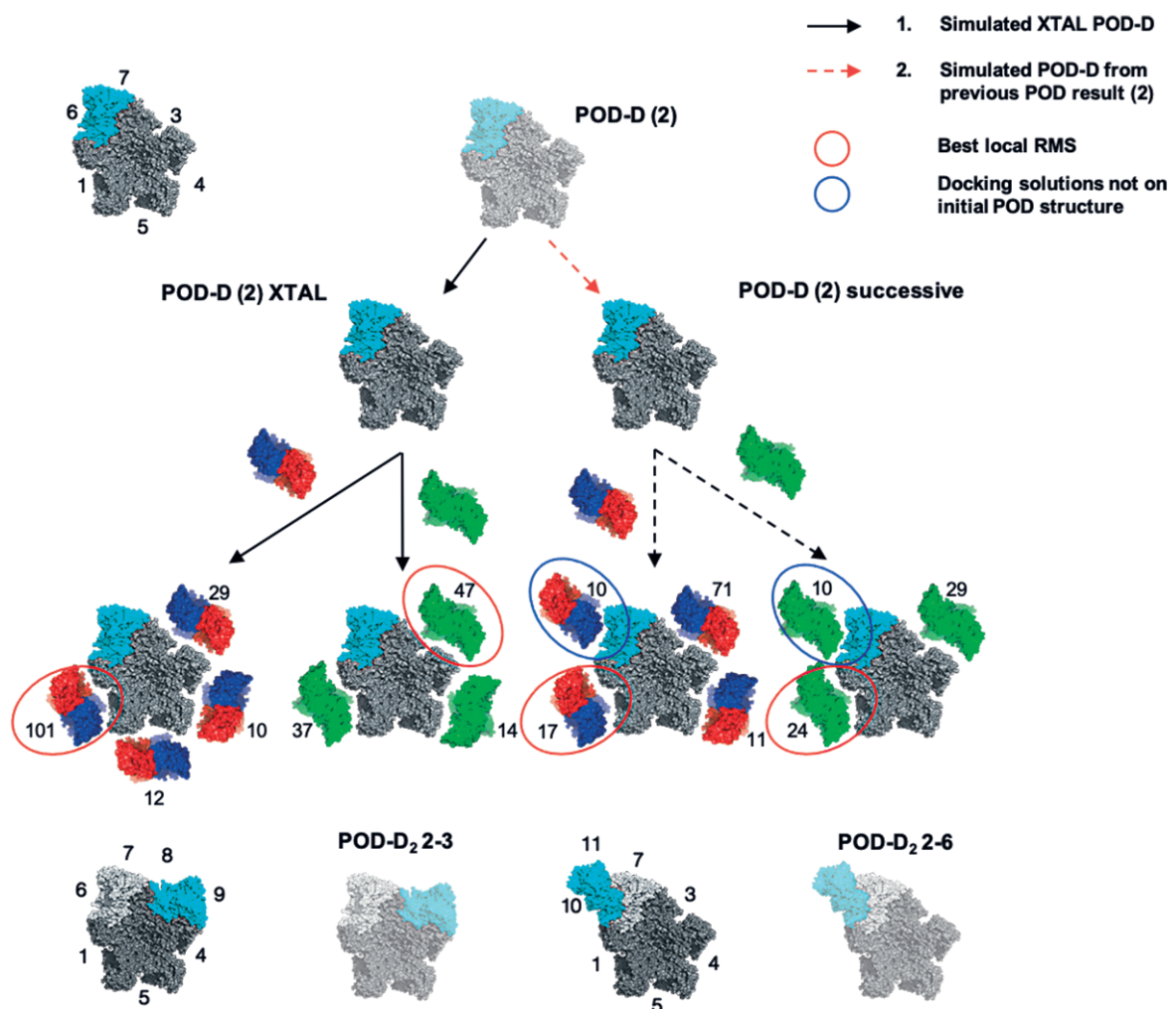


Figure 31 – Comparaison de l'amarrage d'un dimère A-B ou C-C sur le POD-D. Les interfaces d'amarrage sont numérotées de 1 à 7. Un dimère a été amarré à la position 2 du POD, le résultat est le POD-D (2). Les nombres au-dessus des solutions d'amarrage correspondent aux nombres de membres. Les solutions entourées en rouge sont les meilleures en terme de RMSD. Les solutions entourées en bleu correspondent à celles en dehors du pourtour du POD initial.

Le POD-D, issu de la structure cristallographique simulé, peut interagir avec un dimère sur toutes les interfaces restantes du pourtour du POD initial. Comme la solution d'amarrage en dehors du pourtour du POD initial apparaît sur l'autre POD-D (POD-D (2) successive), on en déduit que la durée de simulation des POD-D_n influe sur la croissance de l'intermédiaire.

Pour les POD-D₂, il y a des solutions de docking sur une interface supplémentaire de la bordure du POD initial (Figure 32). Il s'agit de la position 1 sur le POD-D₂ 2-3, en amarrant avec le dimère A-B et de la position 4 sur le POD-D₂ 2-6, en amarrant avec le dimère C-C. Les meilleures solutions de docking sur le POD-D₂ 2-3, sont aux mêmes interfaces (position 4) pour les deux dimères. Il en est de même, pour le POD-D₂ 2-6. Les meilleures solutions sont à la même interface (position 1). Pour les deux types de dimères, des solutions d'amarrage qui ne sont pas sur le pourtour du POD initial du POD-D₂ 2-3 apparaissent (position 8).

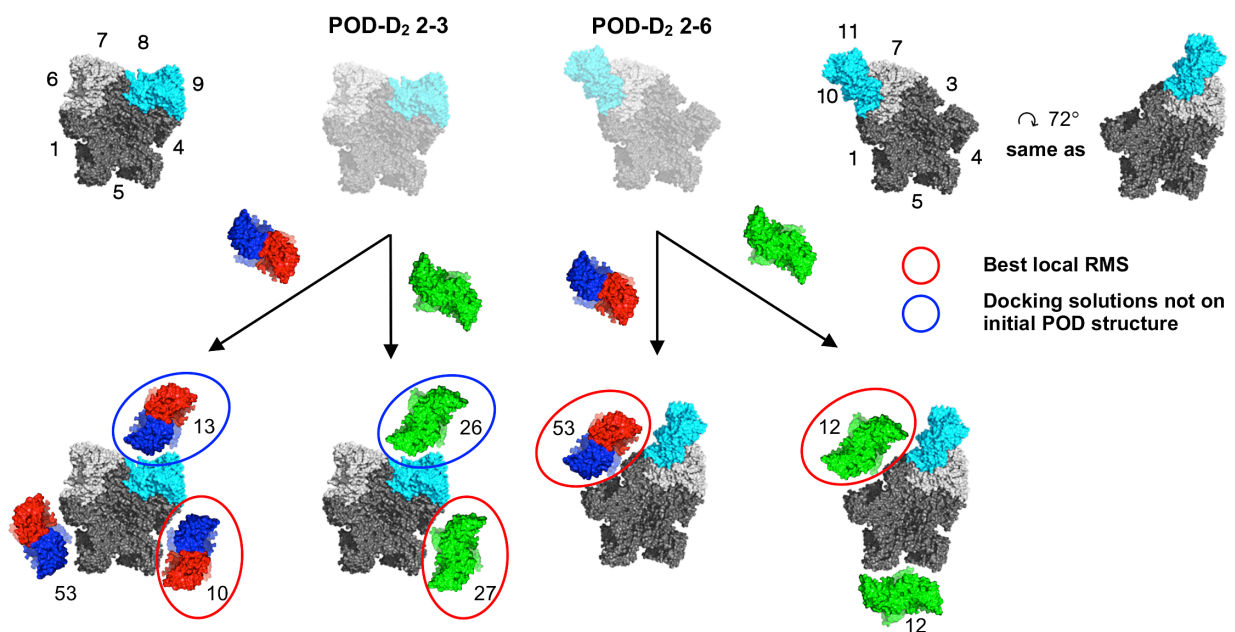


Figure 32 – Comparaison de l'amarrage d'un dimère A-B ou C-C sur le POD-D₂. Les interfaces d'amarrage sont numérotées de 1 à 9 ou de 1 à 11. Le POD-D (2) est soit amarré à la position 3 (POD-D₂ 2-3), soit à la position 6 (POD-D₂ 2-6). Les nombres au-dessus des solutions d'amarrage correspondent aux nombres de membres. Les solutions entourées en rouge sont les meilleures solutions en termes de RMSD. Les solutions entourées en bleu correspondent à celles, en dehors du pourtour du POD initial.

Les deux POD-D₃, issus des meilleures solutions d'amarrage des POD-D₂, sont quant à eux caractérisés par le même nombre d'interfaces amarrées (Figure 33). Une seule interface peut interagir avec un dimère sur le POD-D₃ 2-3-4 (position 5) et deux interfaces peuvent interagir avec un dimère, sur le POD-D₃ 2-3-8 (positions 1 et 4 avec un dimère A-B et positions 4 et 5 avec dimère C-C). Les solutions sont placées sur les interfaces du POD initial. La meilleure solution d'amarrage est commune sur le POD-D₃ 2-3-4. Pour le POD-D₃ 2-3-8, elle est à la position 1, avec le dimère A-B et à la position 5, avec le dimère C-C. C'est le seul POD (-D_n) qui ne converge pas vers la même solution en amarrant avec les deux

conformations de dimères A-B et C-C. À cette itération, il n'y a pas de solution en dehors du pourtour du POD initial. Le POD-D₃ amarré, par un dimère en dehors du pourtour du POD initial, (POD-D₃ 2-3-8) est davantage amarré, que le POD-D₃, uniquement amarré sur le pourtour du POD initial (POD-D₃ 2-3-4).

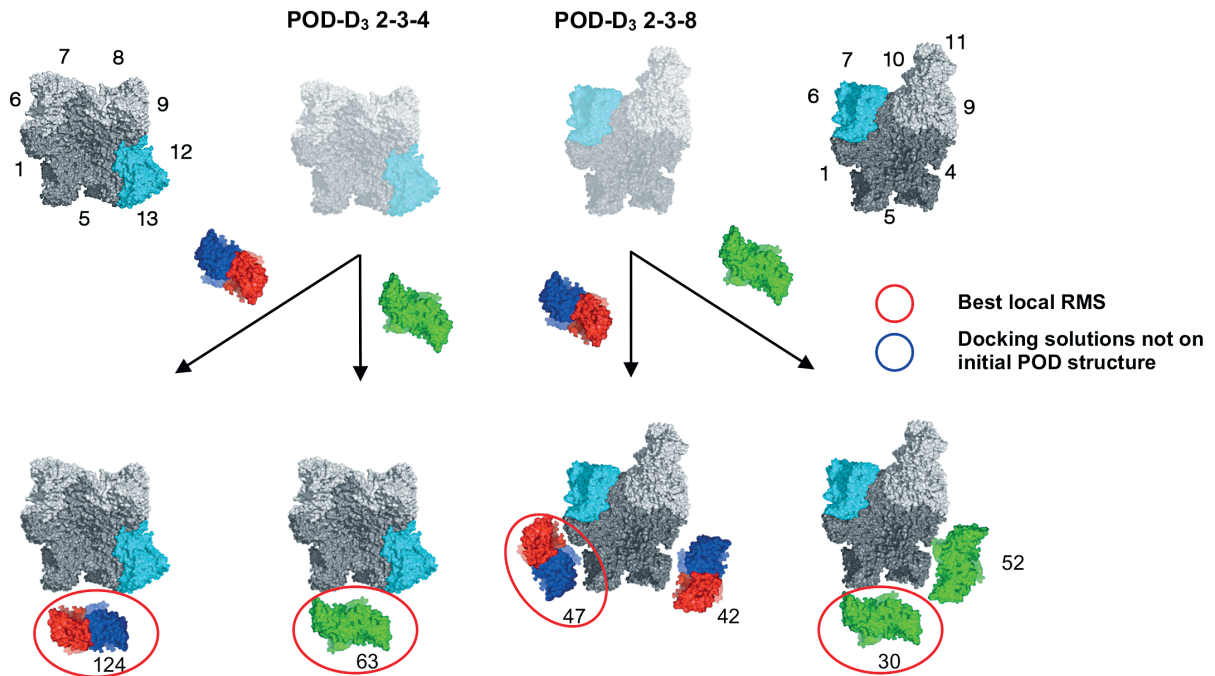


Figure 33 – Comparaison de l'amarrage d'un dimère A-B ou C-C sur le POD-D₃. Les interfaces d'amarrage sont numérotées de 1 à 13 ou de 1 à 11. Le POD-D₂ est soit amarré à la position 4 (POD-D₃ 2-3-4), soit à la position 8 (POD-D₃ 2-6-8). Les nombres au-dessus des solutions d'amarrage correspondent aux nombres de membres. Les meilleures solutions, en termes de RMSD, sont entourées en rouge. Les solutions entourées en bleu correspondent à celles en dehors du pourtour du POD initial.

5.4.8.1 Conclusion sur l'influence de l'amarrage avec un dimère A-B ou C-C

Pour la majorité des itérations, les étapes d'amarrage avec un dimère A-B comptent une interface d'amarrage supplémentaire sur les POD(-D_n). En tenant en compte de l'ensemble des analyses, on voit que la durée de simulation et la conformation du dimère utilisé pour l'amarrage, ont une influence sur la croissance de l'intermédiaire.

Les meilleures solutions d'un dimère A-B ou C-C sur un POD(-D_n) sont quasi-identiques. On peut construire le même chemin d'assemblage avec un dimère A-B qu'avec un dimère C-C.

5.4.9 Espace conformationnel du dimère VP1

Les analyses des résultats d'amarrage, à partir des dimères A-B et C-C, ont révélé que la conformation du dimère utilisé a une influence sur la croissance de l'intermédiaire. Il faut donc se demander :

- Quelle est la conformation du dimère libre en solution ?
- Quel serait alors son impact sur la croissance ?

Un dimère libre pourrait adopter des conformations différentes des dimères contraints dans le contexte de la capsid. Pour avoir une idée de l'espace conformationnel du dimère libre, Il faudrait s'intéresser à la MD du dimère A-B et du dimère C-C, ainsi qu'au passage de l'un vers l'autre.

5.4.9.1 Dynamiques des dimères cristallographiques

Les dimères A-B et C-C ont été simulés à deux échelles granulométriques, gros grains et tout-atomes. Les simulations sont calculées, soit avec GROMACS 5.1.4 [44], soit avec Amber16 [46].

Les RMSD des MD (Figure 34) des dimères A- B et C-C vont nous servir à déterminer si les dimères sont caractérisés par un espace conformationnel commun.

Les trajectoires sont superposées sur les domaines S d'un dimère cristallographique non simulé ou en fin de simulation (A-B ou C-C). Elles sont superposées sur leurs propres domaines S ou sur les domaines S d'un dimère dans une autre conformation (A-B vs C-C ou C-C vs A-B).

Le calcul du RMSD se fait sur les résidus 30 à 220 (domaines S) ou les résidus 1 à 530 (entier), tout-atomes ou gros grains.

Les figures 34A et B correspondent aux calculs des RMSD des MD tout-atomes, réalisées avec GROMACS [44]. La figure 34C montre les RMSD calculés des MD tout-atomes produites avec Amber16 [46].

À l'échelle atomique et à l'issue des 105 ns, le domaine S du dimère A-B diverge d'environ 2,3 Å et celui du dimère C-C, d'environ 3,8 Å des dimères cristallographiques (Figure 34A). Lorsque les dimères sont superposés sur les conformations cristallographiques qui ne correspondent pas à leur conformation initiale (A-B vs C-C ou C-C vs A-B), les RMSD ne diminuent pas au cours du temps. Lorsque le calcul du RMSD se fait sur la totalité de la structure des dimères (Figure 34B), la divergence est plus importante. Après 105 ns, le dimère A-B diverge de 7,5 Å et le dimère C-C, de 11 Å.

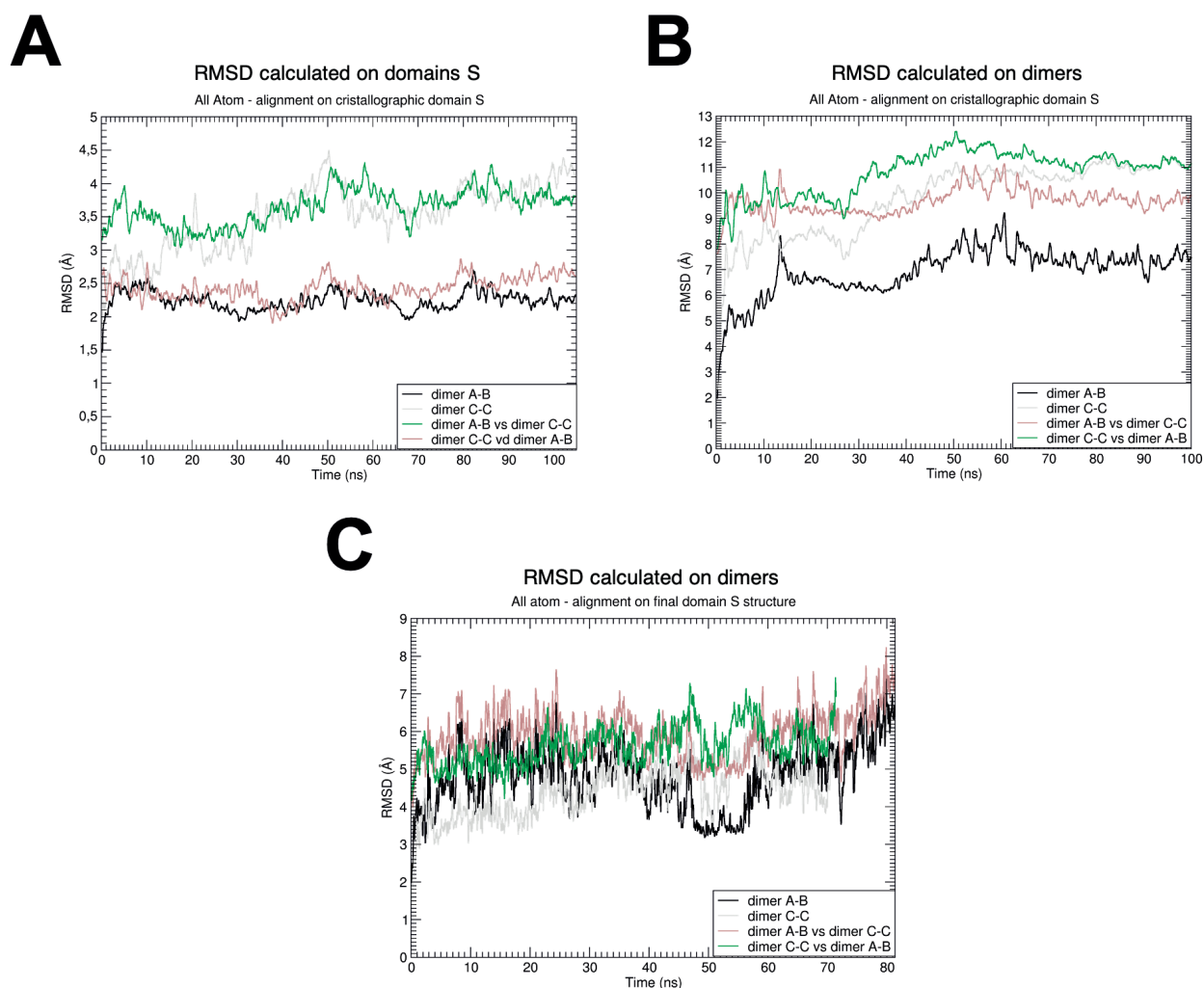


Figure 34 – RMSD des dimères A-B et C-C. A. Le calcul du RMSD est sur les domaines S. La superposition est effectuée sur les domaines S des dimères cristallographiques. B. Calcul du RMSD des dimères, superposition sur les domaines S des dimères cristallographiques. C. Calcul du RMSD des dimères, superposition sur les conformations des dimères en fin de simulation.

Lorsque la trajectoire est superposée sur le domaine S d'un autre type de dimère, en fin de simulation (Figure 34C), le RMSD ne diminue pas. De manière générale, les RMSD indiquent que les dimères A-B et C-C, en termes de conformation, ne se rapprochent pas l'un vers l'autre.

Après 80 ns de MD, les dimères A-B et C-C sont toujours différents. À cette échelle de temps, il n'y a pas de conformation commune pour les deux dimères A-B et C-C. On pouvait pourtant s'attendre à ce que les RMSD des dimères, lorsqu'ils sont superposés sur un autre type de dimère, diminuent au cours du temps.

5.4.9.2 Regroupement des structures des MD des dimères A-B et C-C

Nous cherchons à comparer les trajectoires des dimères A-B et C-C et à identifier un espace conformationnel commun, entre les deux dimères. Nous utilisons comme référence la structure de départ ($t=0$) du dimère A-B.

Nous avons analysées ces trajectoires avec l'outil TTClust, développé dans notre laboratoire

[47]. Le regroupement des structures est effectué selon la méthode de classification hiérarchique de Ward [48] (p. 258). Le nombre de groupes est estimé selon la méthode du coude (elbow) ou fixé à 10. Nous avons calculé l'écart-type pour chaque groupe (RMSD moyen entre toutes les paires de conformations dans un groupe). La structure représentative du groupe est identifiée, comme étant la plus proche du barycentre. Les résultats des analyses sont présentés dans la figure 34.

Les analyses portant sur des trajectoires A-B ou C-C révèlent qu'il n'y a aucun espace conformationnel commun.

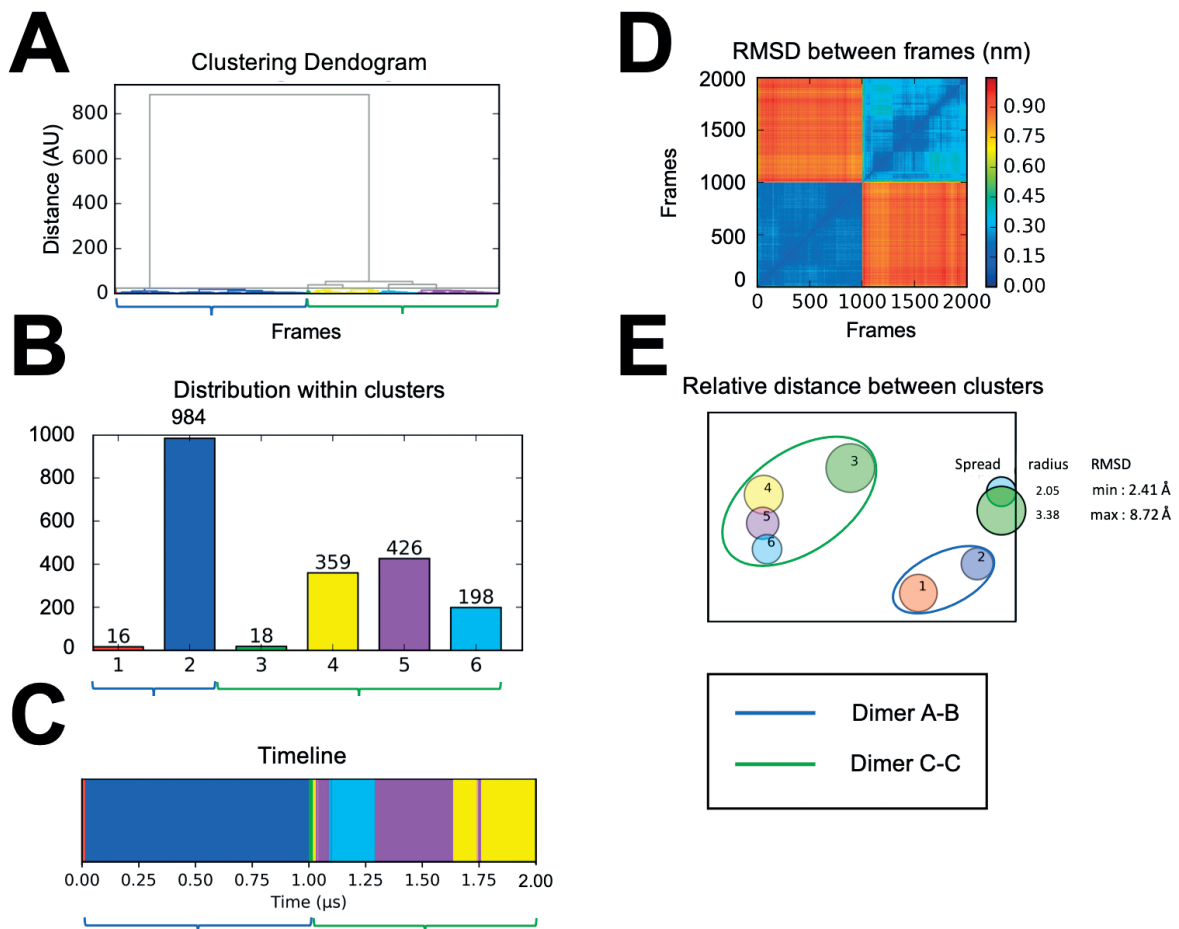


Figure 35 – Exemple d'analyse de regroupement de structures réalisée avec TTClust [47], sur les trajectoires des dimères A-B et C-C concaténées. A. Dendrogramme de la distance hiérarchique entre les groupes majoritaires. B. Histogramme des 6 groupes de conformation. C. Représentation des 6 groupes majoritaires le long de l'axe du temps. D. Carte des corrélations croisées dynamiques entre toutes les conformations (DCCM) où les 1000 premières nanosecondes correspondent à la trajectoire du dimère A-B et les 1000 suivantes à la trajectoire du dimère C-C. E. Projection 2D des distances entre les 6 groupes, où l'aire du disque est proportionnel à l'écart-type. Les accolades sur l'axe des abscisses correspondent aux conformations de la dynamique du dimère A-B en bleu et du C-C en vert (fig A, B et C); les ellipses regroupent les sous-groupes avec le même code couleur (fig E).

La superposition est effectuée sur les domaine S ou sur le squelette carboné de tout le dimère. Le RMSD est calculé sur :

- Tous les atomes du dimère ;

- Ou carbones α du domaine S ;
- Ou carbones α du dimère, hormis les bras N- et C-terminaux ;
- Ou hélices et carbone α ;
- Ou brin et carbone α .

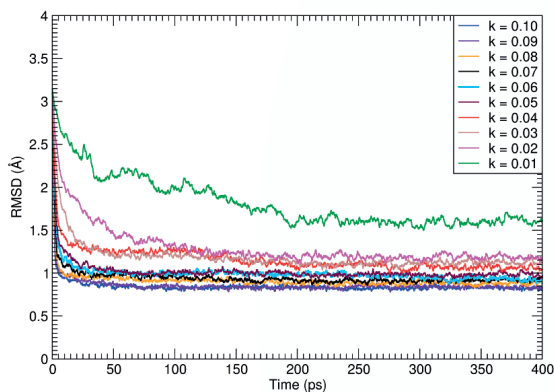
Dans l'exemple ci-dessus, les résultats d'une procédure de regroupement de structures sont montrés. Pour ce regroupement de structures, la superposition de la trajectoire des deux dimères a été effectuée sur le domaine S. Le RMSD a été calculé sur l'ensemble du dimère. On voit que les conformations des dimères A B (accolades et cercle en bleu) et C C (accolades et cercle en vert) pendant les simulations sont bien séparées (Figure 35A et C). La dynamique du dimère A B est divisée en deux groupes (1 et 2, Figure 35B et E). Les groupes 3, 4, 5 et 6 (Figure 35B et E) correspondent à la dynamique du dimère C C. On peut définir un premier sous-ensemble (groupes (clusters) 1 et 2 entourés en bleu) et un second (groupes 3, 4, 5 et 6) (Figure 35E), très éloignés l'un de l'autre. Enfin, la carte des RMSD croisés (Figure 35D) nous montre que les structures rencontrées dans la trajectoire ont des conformations proches (couleurs de température froide ; RMSD compris entre 0 et 0,45 nm). Lorsque l'on croise les conformations des simulations des deux dimères (A B vs C C ou C C vs A B), les RMSD calculés sont beaucoup plus élevés (couleurs de température chaude ; RMSD compris entre 0,7 et 1 nm). On peut également noter que les conformations du dimère A B au cours de sa MD divergent moins entre elles que celles du dimère C C. Tous les résultats de regroupement confondus montrent que les conformations des dimères A-B et C-C n'ont jamais de structure commune. Il n'y a donc pas d'espace conformationnel commun entre les deux dimères.

5.4.9.3 Passage de la conformation A-B vers C-C et inversement

Pour provoquer et observer le changement conformationnel du dimère A B vers le dimère C C et inversement, des simulations de dynamique moléculaire dirigée (TMD) ont été réalisées (Figure 36). D'une TMD à l'autre, la constante harmonique k est ajustée pour augmenter ou diminuer la force appliquée sur les atomes des dimères A-B (Figure 36A) et C-C (Figure 36B). Dans notre cas, la force est appliquée sur les atomes du squelette carboné, qui ne composent ni les bras N- et C-terminaux, ni les segments interstitiels. Cette force permet de provoquer le passage d'une conformation à l'autre.

A

TMD RMSD evolution between dimer A-B and dimer C-C

**B**

TMD RMSD evolution between dimer C-C and A-B

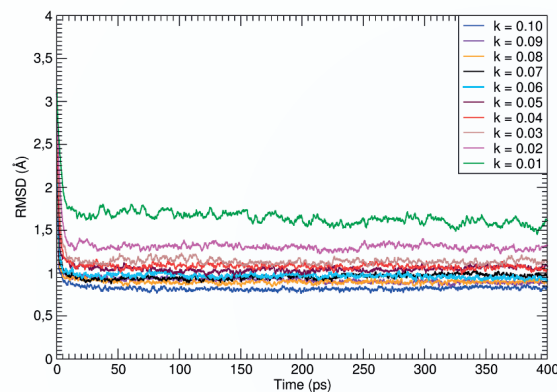


Figure 36 – Évolution du RMSD entre les dimères dans une conformation vers une autre conformation.
 A. Passage du dimère A-B vers le dimère C-C. B. Passage du dimère C-C vers le dimère A-B.

On s'attend à ce que, dans l'ensemble des TMD, les dimères dans une conformation de départ (dimère A-B ou dimère C-C) convergent vers la conformation cible (dimère C-C ou dimère A-B). En fonction de la constante harmonique utilisée, les dimères sont censés converger vers un même point, dans un temps plus ou moins long (de $\sim 3,13$ Å à 0 Å). Pourtant, tous les RMSD des TMD d'un même dimère montrent que la structure initiale ne converge pas vers la même conformation finale. Cela indique que les dimères ne parviennent pas totalement à passer de la conformation de départ vers la structure cible. Si les dimères avaient totalement convergés, les RMSD seraient de ~ 0 Å.

5.4.9.4 Composition en structure secondaire du dimère VP1

La composition en structures secondaires de la protéine VP1 aide à comprendre pourquoi le passage du dimère A-B à C-C et inversement, n'est pas total. Les données structurales de la capside du Norovirus (PDB ID : 1IHM) [34] montrent que plus de la moitié de la protéine VP1 est composée de boucles (Figure 37A et Tableau 1).

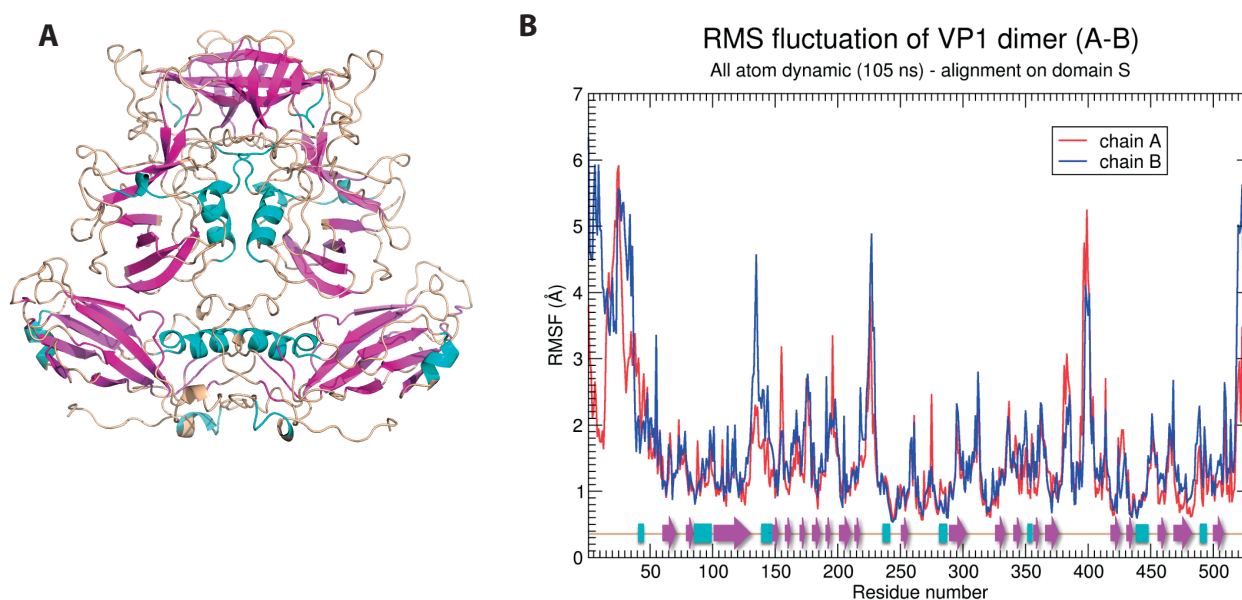


Figure 37 – Dynamique des structures secondaires du dimère VP1. A. Structures secondaires du dimère VP1. B. Fluctuation quadratique moyenne des résidus du dimère VP1 (A-B). Les flèches correspondent aux brins et les rectangles ou ressorts, aux hélices. Les hélices sont colorées en cyan, les brins sont représentés en magenta et les boucles, ou coude, sont en beige.

Structure Secondaire	Protéine VP1 totale (en %)	Protéine VP1 résolue, résidus 30 à 520 (en %)
Hélice	10,19	11,02
Brin	28,30	30,61
Boucle, coude, ...	61,51	58,37

Tableau 1 – Fréquence des structures secondaires de la protéine VP1.

Des analyses ont été effectuées sur les MD tout-atomes de 105 ns des dimères A-B et C-C. Le calcul de la fluctuation quadratique moyenne (RMSF) des résidus du dimère VP1 indique que les boucles qui relient les domaines ou qui composent les parties terminales, ont un facteur de flexibilité plus important que les domaines structurés (Figure 37B). Les brins et hélices de la protéine VP1 sont parfois connectés par de longues boucles (en moyenne : ~8 résidus ; au maximum : 39 résidus). Cette analyse confirme que les parties non structurées qui relient les domaines ou structures secondaires, sont très flexibles. Cela indique que les domaines et sous-domaines ont une grande mobilité.

En repartant des structures cristallographiques, les superpositions des domaines S ou P entre les dimères A-B et C-C montrent que la différence entre les domaines P est plus faible que celle entre les domaines S (RMSD des domaines S : 1,582 ; RMSD des domaines P :

0,489). Le segment interstitiel (résidus 221 à 229) connecte les domaines S aux domaines P et induit une flexibilité, l'un par rapport à l'autre. Cette flexibilité est à l'origine d'une différence d'orientation entre les domaines S et P pour les 3 conformations de VP1, dans le contexte de la capside (Figure 38) [35].

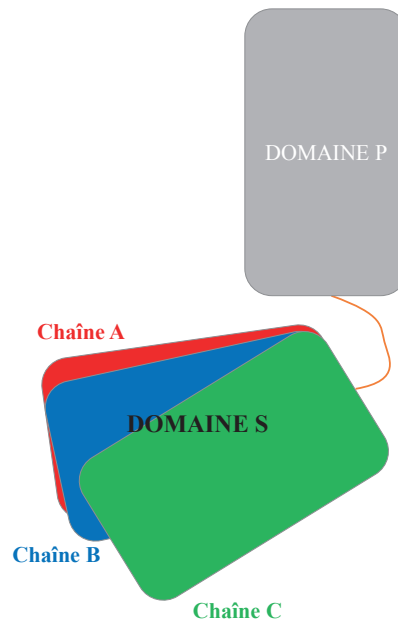


Figure 38 – Représentation de l'orientation du domaine S par rapport au domaine P pour les trois conformations de VP1. Vue du dessus. Schéma issu de la thèse de Thibault Tubiana (<https://tel.archives-ouvertes.fr/tel-01773889>).

L'orientation du domaine S par rapport au domaine P est suivi au cours des TMD du passage du dimère A-B vers C-C. Deux plans sont définis : l'un à partir des résidus ordonnés du domaine S et l'autre avec les résidus ordonnés du domaine P. L'orientation (angle) est calculée en utilisant ces deux plans et en suivant l'axe z (vue du dessus des domaines) . L'analyse s'est focalisée sur les TMD dont les constantes harmoniques sont égales à 0,01, 0,05 et 0,10 kcal mol⁻¹ Å⁻² (Figures 39 à 41). Les trajectoires sont superposées sur le squelette carboné des résidus 231 à 519. L'angle est calculé entre un premier plan (domaine S) et un deuxième plan (domaine P). Le premier plan est construit en ne considérant que les structures secondaires du domaine S. Le deuxième plan est établi à partir du squelette carboné des résidus 231 à 519.

Les RMSD sur les figures 39 à 41 sont calculés entre la conformation initial (au temps t) et la conformation cible. Sur la figure 39, l'angle de la conformation A passe progressivement de ~85° à ~91°. La conformation B passe progressivement de ~84° à ~91°.

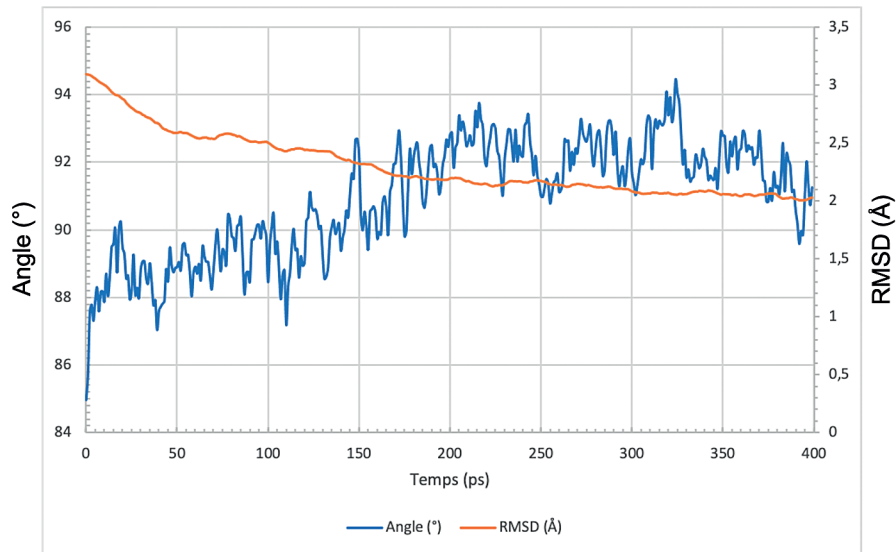
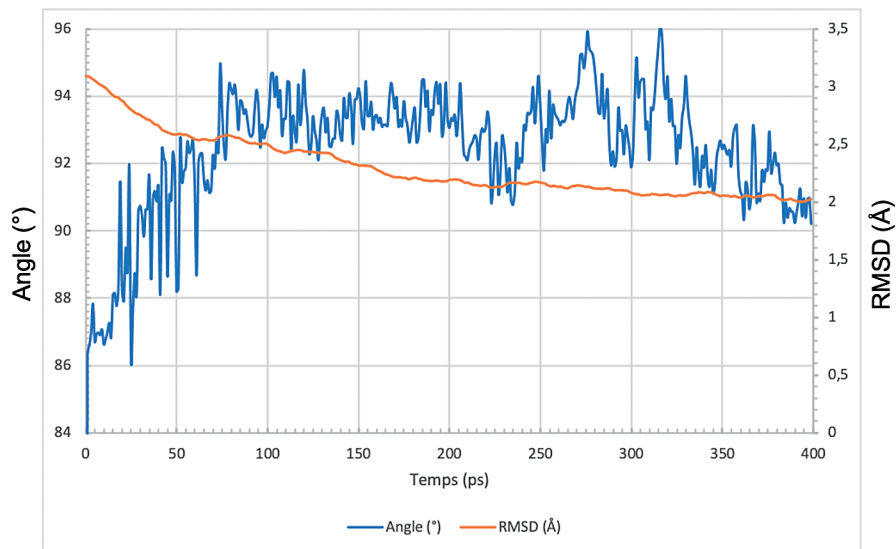
A**B**

Figure 39 – Évolution de l'Incurvation entre les domaines P et S au cours de la TMD ($k = 0,01 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$). A. Évolution du RMSD et de l'angle entre les 2 domaines de la chaîne A. B. Évolution du RMSD et de l'angle entre les 2 domaines de la chaîne B.

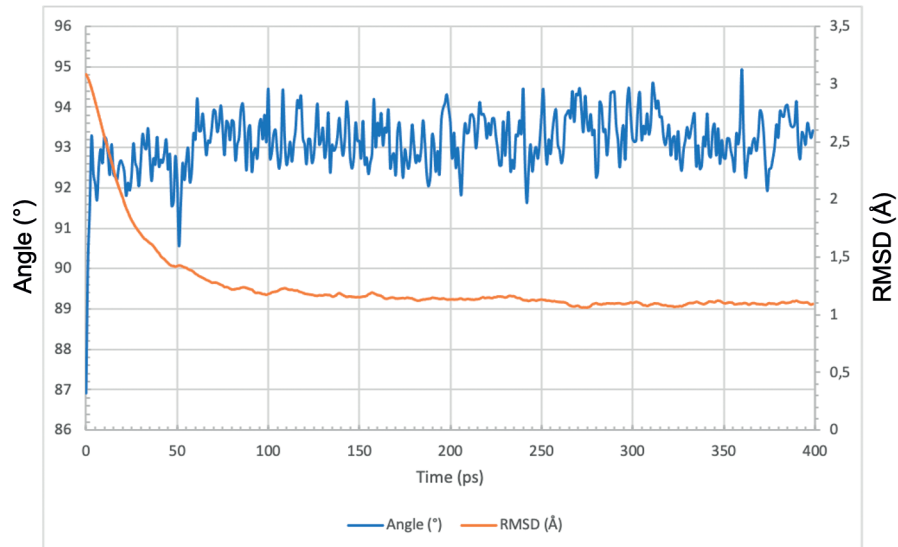
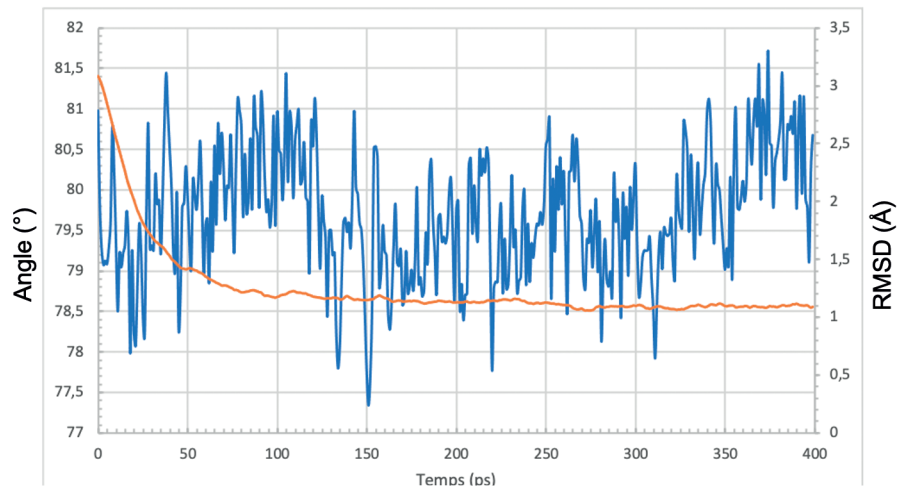
A**B**

Figure 40 – Évolution de l’Incurvation entre les domaines P et S au cours de la TMD ($k = 0,05 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$). A. Évolution du RMSD et de l’angle entre les 2 domaines de la chaîne A. B. Évolution du RMSD et de l’angle entre les 2 domaines de la chaîne B.

Sur la figure 40, l’angle de la conformation A passe de $\sim 85^\circ$ à $\sim 93^\circ$ à la 1^{ère} ps. Cet angle est très proche de la valeur cible ($\sim 94^\circ$ - conformation C). L’angle de la conformation B fluctue autour de $\sim 80^\circ$.

Pour $k = 0,10 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ (Figure 41), les angles de la conformation A et B fluctuent respectivement autour de $\sim 99^\circ$ et $\sim 98^\circ$.

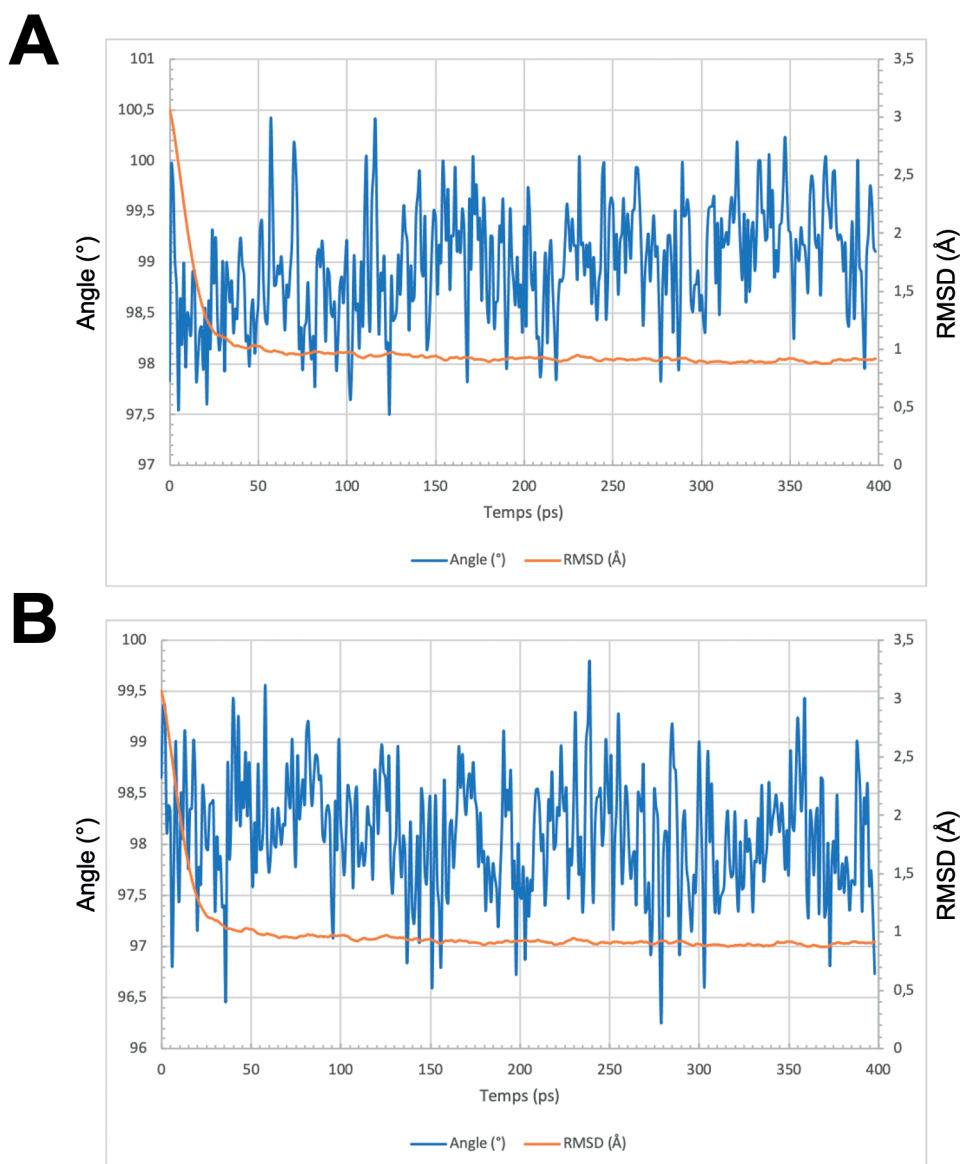


Figure 41 – Évolution de l’Incurvation entre les domaines P et S au cours de la TMD ($k = 0,10 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$). A. Évolution du RMSD et de l’angle entre les 2 domaines de la chaîne A. B. Évolution du RMSD et de l’angle entre les 2 domaines de la chaîne B.

Lors des TMD, où k est égale à $0,01$ ou à $0,05 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$, l’orientation des domaines de la conformation A se convertit en une orientation compatible avec la conformation C. De manière générale, l’orientation des domaines de la conformation B ne parvient pas à passer à l’orientation de la conformation C. Pour $k = 0,01 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ et $k = 0,05 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$, il y a une corrélation entre le changement d’orientation des domaines et l’évolution du RMSD de la conformation A et pas pour la B.

5.4.9.5 Conclusion sur l’espace conformationnel du dimère VP1

De manière générale, les dimères A-B divergent moins de leur structure initiale, que les dimères C-C. Les domaines S divergent moins que les domaines P. L’orientation de ces deux domaines évolue au cours des simulations et la superposition s’effectue sur les domaines S. Les MD indiquent aussi que la conformation d’un dimère A-B ne converge pas vers la

conformation d'un dimère C-C et inversement, au cours d'une MD classique. Les TMD nous montrent également que le passage d'une conformation à l'autre nécessite un réarrangement des boucles et coudes, en plus des régions ordonnées. La convergence d'une conformation à l'autre, même avec une TMD, n'est pas totale. Le pourcentage de parties non structurées de la protéine VP1 explique cette difficulté pour passer d'une conformation à l'autre.

5.5 CONCLUSION

Cette stratégie nous donne une idée des chemins empruntés pour aboutir à l'intermédiaire observé par TR-SAXS. Même si notre procédure montre que le chemin isotrope est privilégié, un chemin anisotrope est possible et conduit à des intermédiaires compatibles avec l'enveloppe. L'étude montre que la dynamique et les propriétés électrostatiques des POD(-D_n) participent à la formation de l'intermédiaire d'assemblage (section 5.4.6).

La croissance de l'intermédiaire est également influencée par la conformation du dimère amarré. C'est pourquoi il a fallu chercher à identifier une partie de l'espace conformationnel du dimère libre en solution. La proportion de la protéine structurale VP1, en régions intrinsèquement désordonnées (plus de 60 %), montre une grande flexibilité des domaines et sous-domaines. Elle peut expliquer la difficulté pour passer de la conformation du dimère A-B à la conformation du dimère C-C et inversement. L'interconversion des conformations, à l'aide de TMD, n'est pas totale.

6 DISCUSSION

Tresset et al. ont étudié l'assemblage de la capside vide du norovirus bovin (GI.1). Dans cet article, les auteurs mettent en évidence un intermédiaire allongé de 10-11 dimères par TR SAXS. La question de la formation d'un intermédiaire de ce type se pose, notamment à partir d'un POD du virus de Norwalk (GI.1) [34]. La stratégie développée dans le cadre de l'article repose sur les données cristallographiques du norovirus GI.1 et des méthodes *in-silico*. Étant donné que la structure des dimères libres en solution n'est pas connue, nous avons utilisé les structures cristallographiques des dimères A-B et C-C du GI.1 (PDB ID : 1IHM) sont utilisées [34]. Elles constituent les meilleures bases d'unités d'assemblage, à amarrer sur le POD.

Cette stratégie montre que l'assemblage *in-silico* d'un intermédiaire de forme allongée, comme observé en TR-SAXS [38], est possible. Il est intéressant de voir que les intermédiaires obtenus avec ces travaux réfutent l'hypothèse d'un intermédiaire d'assemblage formé de 2 POD et reliés par un dimère interstitiel [38]. En effet, les analyses montrent que plusieurs dimères sont amarrés sur le pourtour du POD initial. Tous les intermédiaires proposés correspondent davantage à la formation d'un axe hexamérique adjacent au POD initial (Figure 42). Ces résultats montrent que l'intermédiaire d'assemblage correspond à l'un des décimères de dimères (POD-D₅) proposés. La superposition des structures, issues de cette méthode, est tout à fait compatibles avec l'enveloppe TR SAXS. Les structures caractérisées par des NSD égales à 6,11 et 6,38 sont d'autant plus compatibles. Elles correspondent aux structures proposées dans l'article soumis, "Combining computational methods to study biological macromolecular complexes assembly : Application to the norovirus initial growth".

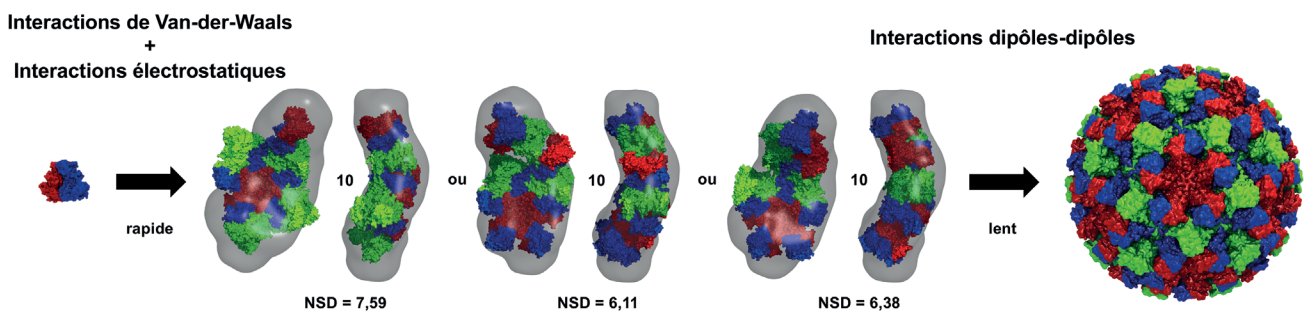


Figure 42 – Schéma simplifié et nouvelle hypothèse d'assemblage de la capside du Norovirus.

Cinétique d'assemblage simplifiée de la capside du Norovirus. Les différentes structures correspondent aux enveloppes modélisées à partir des facteurs de formes déterminés par TR-SAXS [38]. La NSD est une mesure de similarité : plus elle est proche de 0, plus la structure tout-atomes proposée est compatible avec l'enveloppe TR-SAXS.

Il n'est pas exclu que l'enveloppe TR-SAXS corresponde à plusieurs agencements de POD D₅. Si l'intermédiaire n'est pas unique, il y aurait plusieurs chemins d'assemblage possibles. Si les intermédiaires d'assemblage sont bien composés de 10 dimères [38], il faut 9 intermédiaires pour former la capside complète (90 dimères). Dans cette nouvelle hypothèse, il n'y a pas besoin de dimères supplémentaires pour former la capside. Cela nous conforte dans l'idée

que la formation de la capside du norovirus repose sur un mécanisme coopératif. C'est dans cette deuxième étape que l'incorporation du génome médiée par VP2 pourrait intervenir. Le point de départ de notre stratégie d'étude d'assemblage in-silico suit le postulat de Prasad et al. [34]. Sur la base des contacts entre dimères, l'étude part du POD pour étudier la formation de l'intermédiaire. Savoir si la formation de l'intermédiaire peut passer par l'HOD est l'une des questions qui se pose. Les analyses menées sur les 2 capsomères (POD et HOD) montrent qu'il n'y a pas de raison de penser que le POD est un noyau plus vraisemblable que l'HOD. À l'échelle de temps simulé (20 μ s), les 2 convergent rapidement et ne se dissocient pas. S'ils étaient simulés plus longtemps (milliseconde), ils pourraient se dissocier. La voie de l'HOD n'a donc pas été poursuivie. Si l'on avait écarté l'hypothèse de Prasad et al., le choix du noyau critique aurait pu se porter sur l'HOD. Boyd et al. ont mené une étude de nanoindentation in-silico sur les axes 2, 3 et 5 de la capside du virus de Norwalk [49]. Cette expérience est réalisée à une autre échelle de temps (500 ns) et dans un autre contexte. Elle met en évidence la force nécessaire pour dissocier les contacts entre dimères pour chacun des axes. Elle montre que les interfaces entre les domaines S des chaînes B et C sont plus faciles à dissocier. Le POD serait plus stable que l'HOD sur la base de la rupture des interfaces par nanoindentation (Figure 43). Cet article est manifestement en accord avec le postulat de Prasad et al. [34].

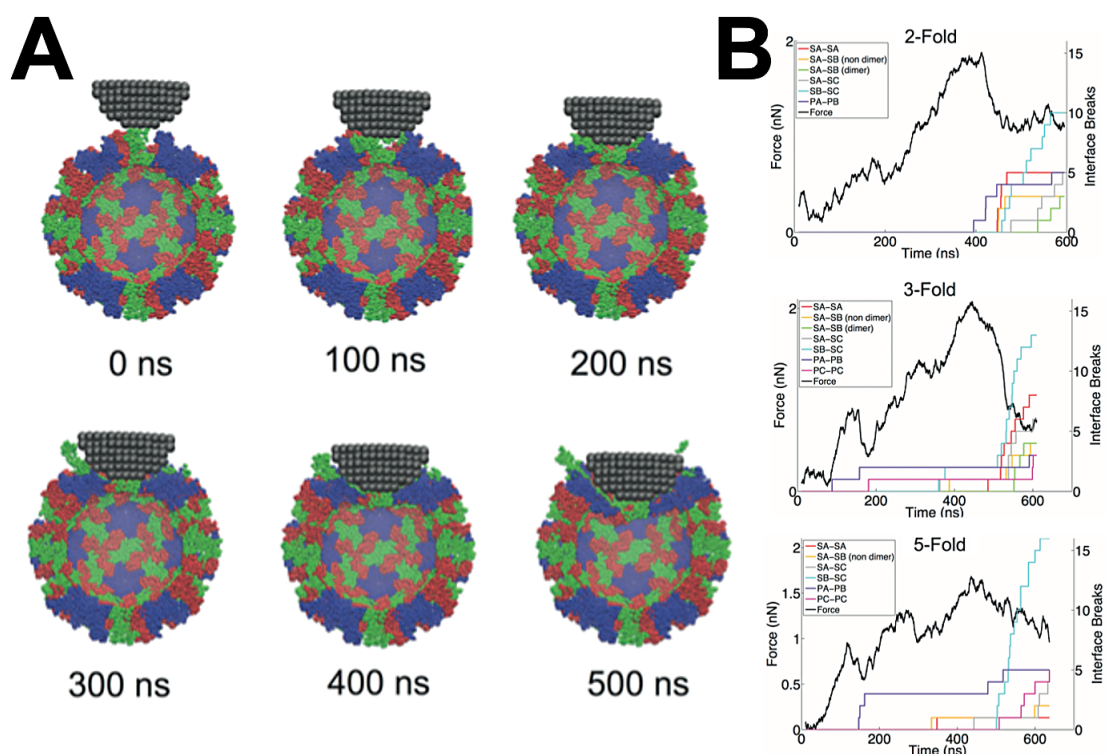


Figure 43 – Nanoindentation in-silico de la capside du Norovirus, issue de l'article [49]. A. Simulation de la nanoindentation de la capside de 0 à 500 ns. Les chaînes A sont colorées en bleu, les B en rouge et les C, en vert. Rupture des interfaces au cours des simulations de nanoindentation. B. Les interfaces considérées vont de SA-SA à PC-PC ; S pour "Shell" et P pour "Protruding" ; et A, B et C pour les chaînes. On peut voir que les interfaces SB-SC se cassent beaucoup plus facilement que les interfaces SA-SB.

La protéine de capside du norovirus est un bon candidat pour étudier l'assemblage des capsides seules car elle est capable de s'auto-assembler en absence de matériel génétique viral. Le TR-SAXS et la spectrométrie de masse à mobilité ionique sont des techniques *in-vitro* qui peuvent aider à répondre à la coordination de l'assemblage. Une étude d'assemblage de la capside du norovirus GI.1, effectuée dans des conditions d'équilibre (conditions de non-assemblage) et par spectrométrie de masse à mobilité ionique, identifie des oligomères de la taille du dimère au POD (Figure 44) [50].

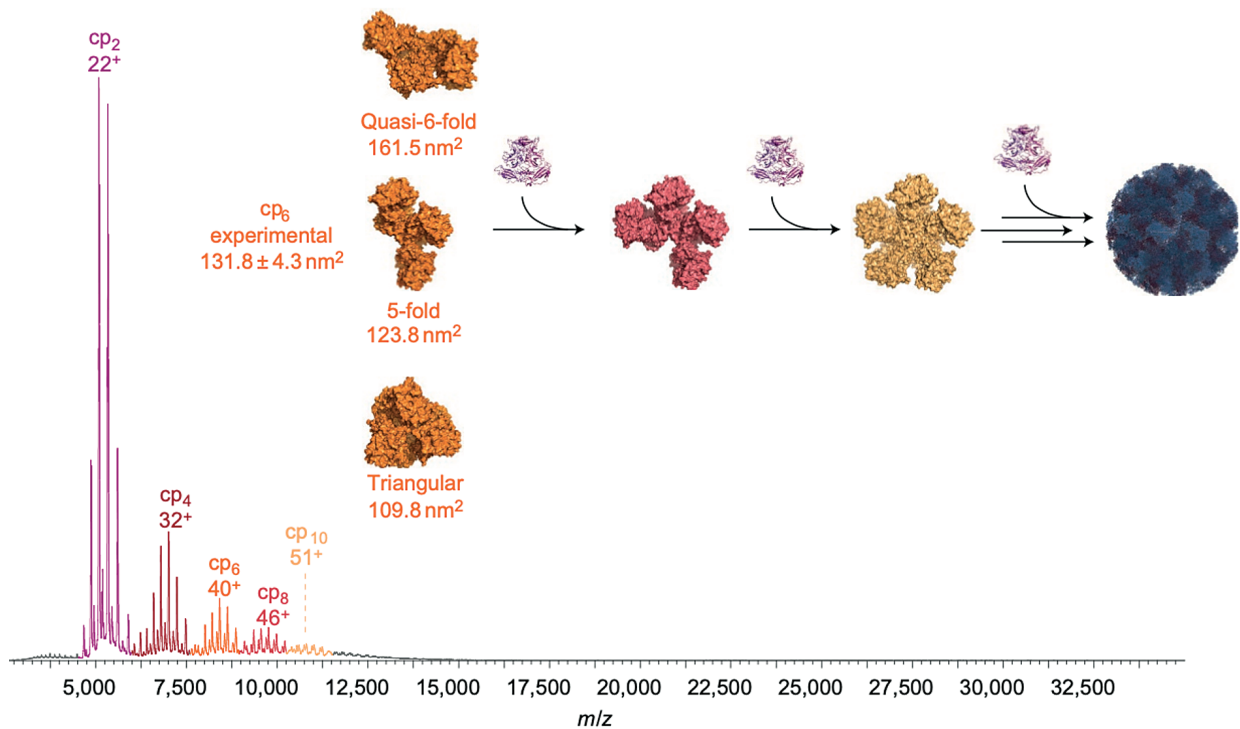


Figure 44 – Assemblage de la capside du norovirus en conditions d'équilibre. Spectrométrie de masse de la protéine de capside dans des conditions de non-assemblage, issue de l'article [50]. "cp" correspond à protéine de capside. Il est suivi du nombre de monomères. m/z correspond au rapport masse/charge. Seules les petits oligomères de dimères sont détectés (cp_2 à cp_{10}) Ces oligomères sont en accord avec le chemin proposé par Prasad et al. pour former le noyau critique.

Lors des expériences de TR-SAXS sur la protéine de capside du norovirus bovin (GI.1.2), seuls 3 facteurs de forme sont identifiés : les dimères, les intermédiaires de forme allongée et les capsides (Figures 45A.b et 45B). On peut notamment voir, sur la figure 45A.c, que les dimères sont très vite consommés et que la concentration d'intermédiaires augmente proportionnellement dans les 100 premières secondes. Ensuite, la concentration d'intermédiaires diminue et les capsides apparaissent sans que l'on détecte d'objet supplémentaire entre l'intermédiaire et la capside. Les noyaux critiques d'assemblage ne sont pas détectés car ils sont très vite consommés pour former les intermédiaires [38]. On n'a donc pas la preuve que la formation de l'intermédiaire passe par le POD, sur cette expérience.

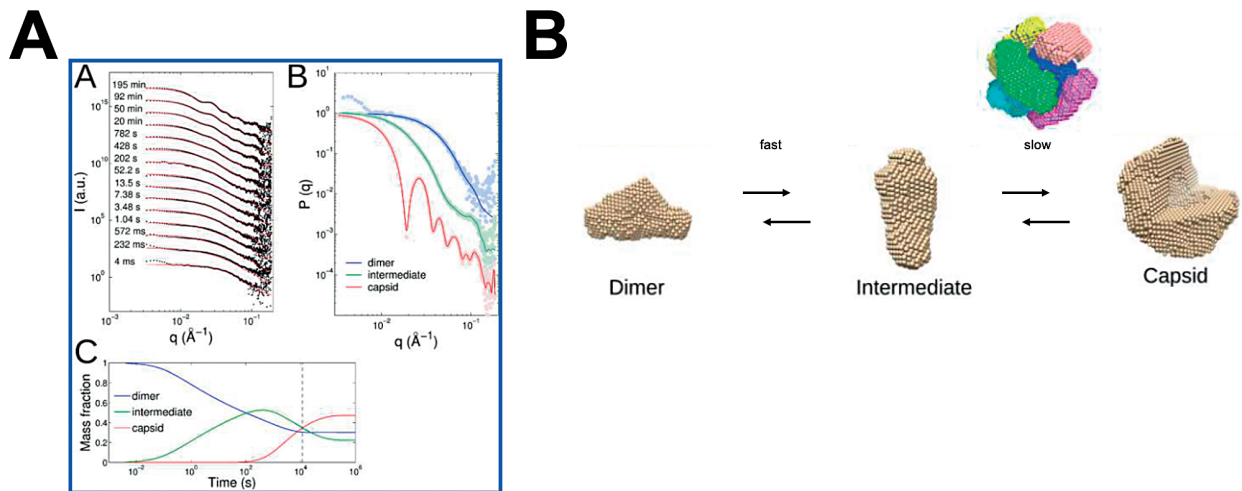


Figure 45 – Assemblage de la capside en conditions d'assemblage. Traitement des données de TR-SAXS de l'assemblage de la capside du norovirus bovin (GIII), issu de l'article [38]. (A) Comparaison des intensités de diffusion expérimentales (points noirs) et celles reconstruites à partir de l'ajustement global (ligne rouge). (B) Facteurs de forme $P(q)$, extraits de la simulation des 3 espèces en solution : dimères en bleu, intermédiaires en vert et capsid en rouge. (C) Fractions de masse pour les 3 espèces en fonction du temps, à partir du modèle cinétique. La ligne en pointillé indique les dernières données expérimentales. (D) Schéma simplifié de la cinétique d'assemblage et hypothèse de mécanisme d'assemblage [38].

Les souches virales de l'expérience de TR-SAXS (GIII.2) et de l'étude d'assemblage par spectrométrie de masse (GI.1), sont différentes. Il y a seulement 50 % d'identité de séquence entre les 2 souches. La stratégie mise en place contribue à la compréhension de l'assemblage de l'intermédiaire allongé à partir du POD du GI.1. Elle aide aussi à observer la formation ou la rupture de ponts-salins et le changement de charge des résidus du complexe à l'approche d'un dimère en solution.

La rupture de symétrie, qui se produit chez le POD, n'est pas le point de départ de la croissance (Figure 46). La stratégie montre que les interfaces d'amarrage non amarrées sont occupées par un dimère (A-B ou C-C), à l'étape suivante (POD-D). Elle met également en évidence qu'une isotropie apparaît à partir d'un noyau de ce type. De fait, à l'étape suivante, la meilleure solution de docking est amarrée de façon adjacente au premier dimère amarré.

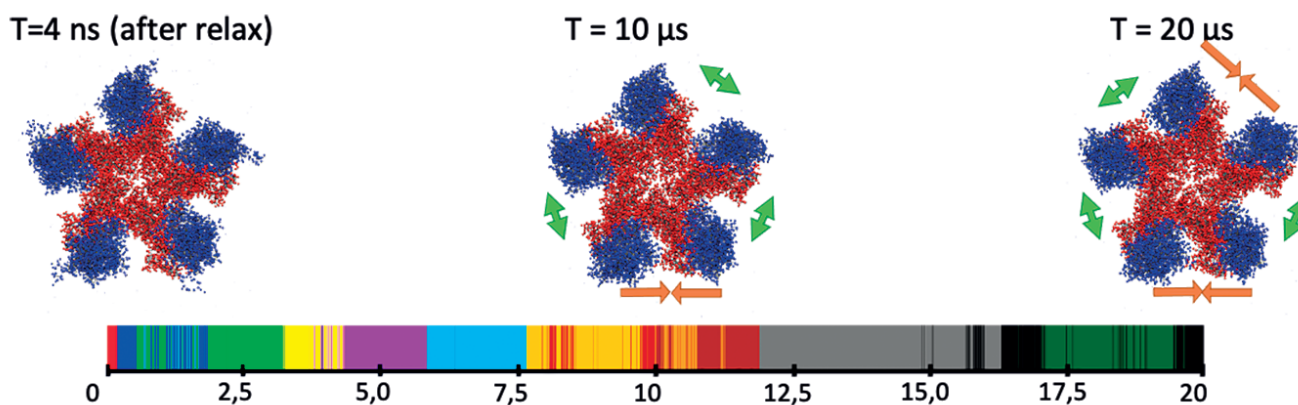


Figure 46 – Rupture de symétrie du POD au cours des 20 μ s. Les structures du POD à 4 ns, 10 μ s et 20 μ s sont représentées. Les chaînes A sont en rouge et les chaînes B, en bleu. La représentation des groupes majoritaires le long de l'axe du temps est représentée en dessous.

Le parti pris, dans l'étude d'assemblage, est d'utiliser la structure cristallographique (PDB ID : 1IHM) [34] pour filtrer les solutions d'amarrage. La structure nous sert de référence. À chaque étape, une structure cristallographique correspondant au résultat d'amarrage précédent, est utilisée. Cela induit le rétablissement de la courbure de la capsid. Comme l'étude se base sur l'enveloppe de l'intermédiaire de la capsid du GIII.2, les solutions compatibles avec celle-ci sont retenues. Pour obtenir les intermédiaires allongés, il faut donc sélectionner des solutions d'amarrage en dehors du pourtour du POD initial. Les complexes simulés sont superposés sur la structure cristallographique de référence. Les RMSD, entre les dimères amarrés et les dimères autour du complexe cristallographique de référence, sont calculés. Les meilleures solutions en termes de RMSD ne sont pas toujours retenues. Les solutions amarrées en dehors du pourtour du POD initial sont choisies comme par exemple, à l'étape du POD-D (2) et du POD-D₂ 2-3 (Figure 47). Ces solutions admettent toujours un RMSD plus élevé que les solutions sur le pourtour du POD initial. Elles sont trouvées plus rarement que les solutions sur le pourtour (voire non trouvées à certaines étapes). Le test en parallèle des solutions "pourtour" et des solutions "allongées", quand il y en a, montre que les 2 complexes se rejoignent à l'étape suivante. C'est le cas du POD-D, amarré sur le pourtour mais aussi, en dehors.

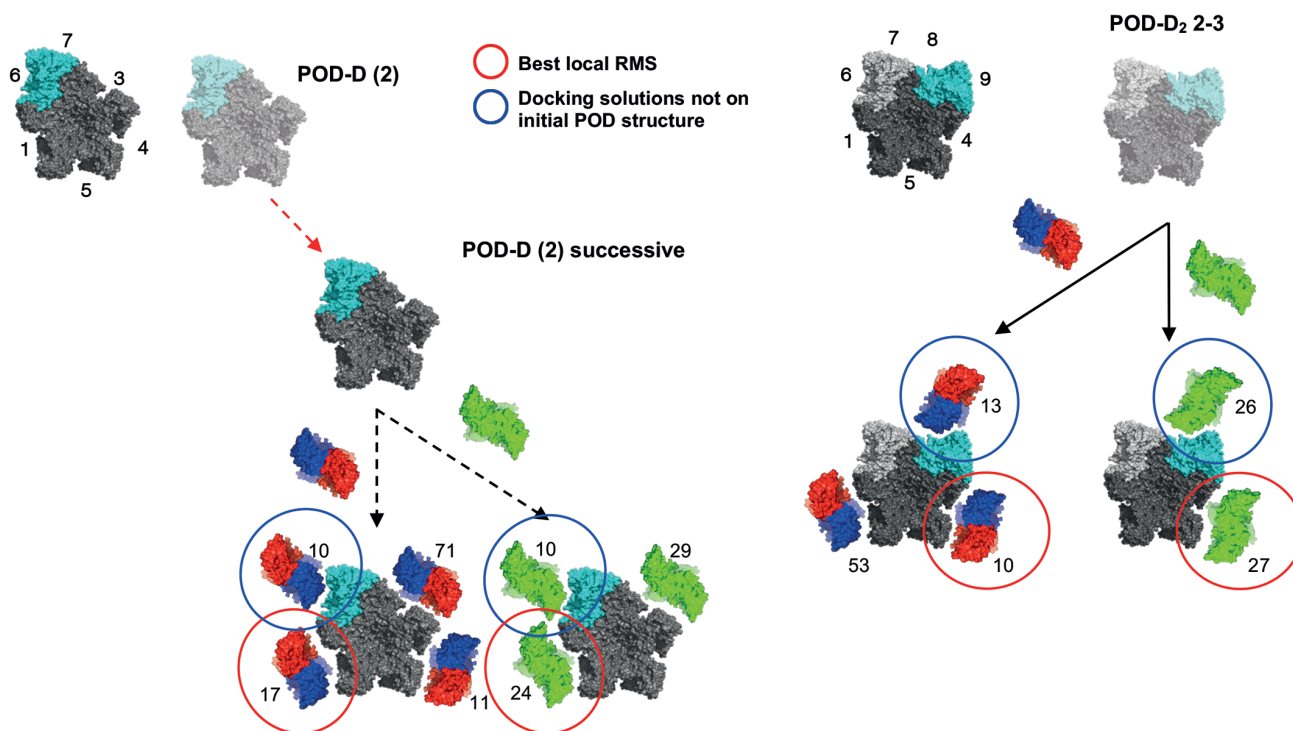


Figure 47 – Solutions d'amarrage des dimères A-B et C-C sur le POD D ou le POD-D₂, 2 3. Le premier dimère amarré est en gris clair, le dernier est en cyan. Le POD est coloré en gris foncé. Le dimère A-B est en rouge et bleu et le C-C, en vert. Les meilleures solutions de docking sont entourées en rouge et les solutions, en dehors du pourtour du POD initial, sont entourées en bleu.

Les scores ClusPro [51,52], seuls, ne sont pas suffisants pour discriminer les bonnes solutions d'amarrage que l'on retrouve dans la structure cristallographique de la capsid du norovirus (PDB ID : 1IHM) [34]. En effet, cette stratégie dépend de la structure cristallographique du complexe assemblé. La contribution électrostatique totale (Q_i) des POD(-D_n) sur les dimères a été déterminée en traitant les données, issues de la méthode PROCEEDpKa de Fernando Luis Barroso Da Silva [53]. La prise en compte simultanée du score ClusPro [51,52] et de la contribution électrostatique totale (Q_i) est une piste d'amélioration envisageable pour sélectionner les bonnes solutions d'amarrage. Cette approche donne des résultats intéressants pour l'amarrage du POD avec un dimère. Même s'il n'y a pas beaucoup de résidus titrables parmi les complexes, le Q_i est un atout pour améliorer le score ClusPro. De cette manière, la stratégie ne dépendrait plus de la structure cristallographique de l'assemblage complet (PDB ID : 1IHM) [34] pour identifier les bonnes solutions, et pourrait être généralisée à n'importe quelle capsid ou même de systèmes "inconnus".

Une méthode d'amarrage moléculaire de type "corps rigides" [51] est utilisée. ClusPro a été bien noté à la 7^{ème} édition du concours CAPRI ("Critical Assessment of Predicted Interactions") [54]. Lors de cette édition, il est arrivé premier dans la catégorie des serveurs pour toutes les cibles d'amarrage à prédire. Cependant, au fur et à mesure des itérations, le POD(-D_n) croît mais le taux d'échantillonnage reste constant : 70 000 positions d'amarrage sont toujours constamment testées même si le nombre de positions d'amarrage et la taille du système augmentent [51]. Il faudrait donc augmenter le nombre de positions testées, en fonction de la taille du récepteur POD(-D_n).

La méthode d'amarrage moléculaire de type "corps rigides" ne prend pas en compte de modification locale du récepteur $POD(D_n)$ ou du ligand (D). Cette flexibilité n'est pas permise. La principale amélioration de cette stratégie serait d'utiliser un serveur ou logiciel qui bénéficie d'une méthode d'amarrage flexible [55] (ATTRACT [56,57], RosettaDock [58], FlexDock [59]...).

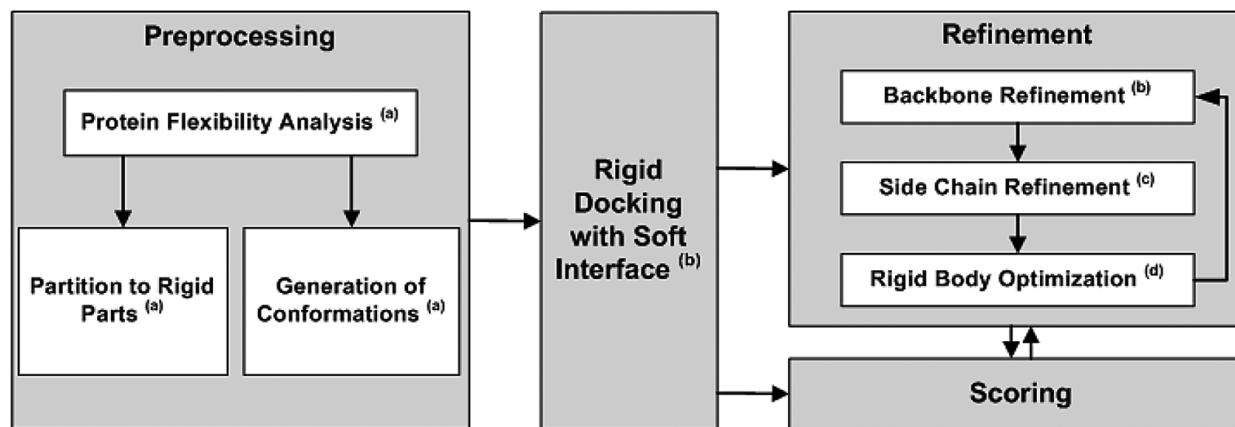


Figure 48 – Procédure générale d'amarrage flexible, tirée de l'article [55].

L'étape d'amarrage flexible est normalement précédée d'une étape d'analyse structurale de flexibilité du récepteur et du ligand (Figure 48). Le principe de cet amarrage est quasi-équivalent car il repose essentiellement sur un amarrage de type "corps rigides", avec des ajustements consécutifs. Ils consistent à affiner les complexes (affinement du squelette et des chaînes latérales des protéines, optimisation de l'amarrage). Les scores sont déterminés au cours de l'affinement et de l'amarrage. Cette amélioration pourrait entraîner la génération de meilleures et de plus nombreuses solutions (avec un affinement de l'interface des récepteurs). Malheureusement, cette solution n'est pas envisageable pour nos systèmes qui comportent plus de 30 000 atomes.

La question de la conformation de l'unité d'assemblage seule en solution est un autre point à approfondir. Il s'agit d'un réel problème et d'une question difficile à répondre. Une situation idéale serait d'amarrer un dimère libre ou indifférencié en solution qui adopterait spontanément une conformation A-B ou C-C lors de l'interaction avec le POD seul ou le $POD-D_n$. En l'absence de ces dimères libres, on pourrait utiliser des conformations du dimère A-B et du dimère C-C simulés au lieu des conformations cristallographiques. La conformation de ces unités d'assemblage influence nécessairement l'étape d'amarrage. Mais pour accéder à la conformation du dimère libre en solution, il faudrait résoudre la question de l'interconversion des dimères cristallographiques. Il serait envisageable d'utiliser une combinaison de techniques d'échantillonnage accéléré comme celle de la métadynamique et de l'"umbrella sampling" [60,61]. Même si cette méthode est extrêmement puissante, elle est utilisée principalement sur de petits systèmes (petits

ligands : ~1500 atomes) [62]. D'autres méthodes pourraient échantillonner davantage l'espace conformationnel [63], comme : la MD avec échanges de répliques (REMD) [64,65]. Dans notre étude, nous avons utilisé la MD ciblée (TMD) pour passer d'une conformation de dimère cristallographique à l'autre (dimère A-B vers dimère C-C). Elle ne parvient pas à convertir complètement le dimère A-B en C-C. La réorientation des domaines P et S ne se produit pas. La méthode in-silico utilisée (force sur les coordonnées des atomes), n'est pas suffisante pour l'interconversion. Pour résoudre cette question, on pourrait effectuer des TMD successives sur des sous-domaines à définir des unités d'assemblage pour réaliser l'interconversion globale. Il faudrait tout de même définir l'ordre de ces TMD successives.

Le chemin d'assemblage du GIII, observé par TR-SAXS, s'effectue dans des conditions bien particulières. Le tampon dans lequel se trouvent les dimères (tampon de désassemblage : pH 9.0) est rapidement mélangé, à volume égal, dans un tampon d'assemblage (pH 6.0-7.0 avec 100 mM NaCl). Le tampon de départ induit des changements de protonation sur les résidus chargés des unités d'assemblage, entraînant des changements physico-chimiques. Le premier effet du tampon est de moduler la répulsion électrostatique entre unités d'assemblage. Le pH et la salinité peuvent également induire des changements structuraux [35,37,38,66]. Tubiana et al. ont observé, par SAXS, que l'enveloppe du dimère dissocié est plus étendue que prévu [35] (Figure 49). La superposition de la structure cristallographique du dimère dans l'enveloppe le montre.

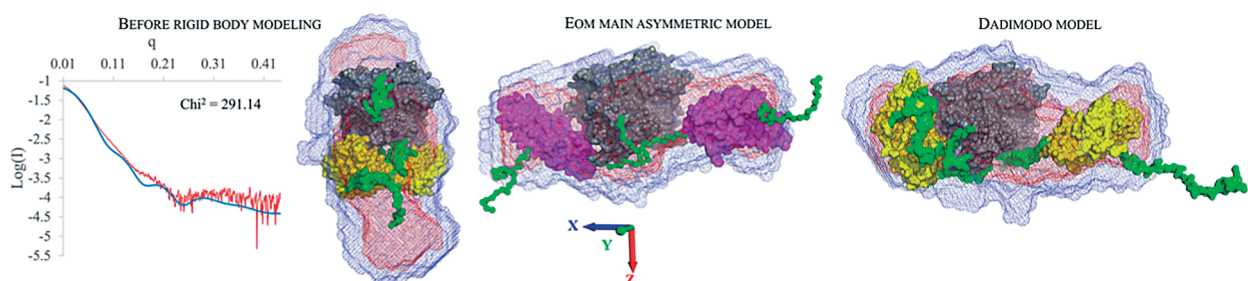


Figure 49 – Structures des dimères dissociés du GI.1, générées par modélisation de corps rigides, à partir des données SAXS, tirées de l'article [35]. Comparaison des dimères GI.1 : le premier, avant l'étape de modélisation de corps rigides, le second un modèle EOM (asymétrique) et le troisième, un modèle Dadimodo. La meilleure superposition de l'enveloppe ab-initio, calculée à partir des données SAXS, est affichée en bleu pour chaque modèle. Les domaines S sont représentés en jaune ou en violet, les domaines P, en gris et les bras terminaux, en vert.

L'état du dimère libre n'est pas connu or, il est à la base même de l'assemblage de la capsid. Pour cette raison, des MD des 2 dimères cristallographiques du GI.1 (PDB ID : 1IHM) [34] sont effectuées pour identifier des conformations communes entre les 2. À cette échelle de temps, les MD (105 ns) montrent qu'ils n'en admettent pas (aucune structure commune). L'état du dimère en solution reste une question à approfondir.

Le noyau critique d'une capsidie icosaédrique peut avoir une forme compacte et symétrique (Le virus de la marbrure chlorotique de la cornille (CCMV) : pentamère de dimères [67], HBV : trimère de dimères [50]). À partir d'une même protéine de capsidie, des capsidies de plusieurs géométries icosaédriques peuvent en résulter (CCMV [67], Norovirus GII.4 [68], HBV du génogroupe D [69]). Le CCMV passe par un POD pour former une capsidie entière [67]. Le CCMV a 2 types de géométrie icosaédrique tronquée qui peuvent se former à partir d'un POD : une pseudo T=2, par l'association coopérative de 12 POD, ou une T=3, par l'addition de dimères. La définition des deux modèles de croissance, coopérative et nucléation-croissance, est donnée dans la figure 49. Cela suggère que la formation de la capsidie du CCMV peut se faire de façon isotrope. L'assemblage de la capsidie T=3 du CCMV illustre très bien le mécanisme d'assemblage par nucléation-croissance. Chevreuil et al. vérifient expérimentalement par TR-SAXS que l'assemblage de la nucléocapsidie de CCMV relève principalement du modèle de nucléation-croissance [13]. Tout comme Zlotnick et al. l'avaient conclu sur la capsidie vide [67].

L'assemblage de la capsidie de norovirus ne relève pas des mêmes mécanismes. Il correspondrait à un mécanisme coopératif. L'auto-assemblage de la protéine de capsidie et l'encapsidation du génome sont des mécanismes distincts, qui doivent être coordonnés par des facteurs supplémentaires et spécifiques, probablement VP2. La protéine VP2 est présente dans les particules du norovirus infectieux, en faible quantité et intervient certainement à l'empaquetage de l'ARN viral, durant l'assemblage [35]. Les nucléocapsidies icosaédriques de virus peuvent s'assembler de façon isotrope ou anisotrope, selon un mécanisme de nucléation-croissance (Figure 50b) ou un mécanisme "en masse" (Figure 50a).

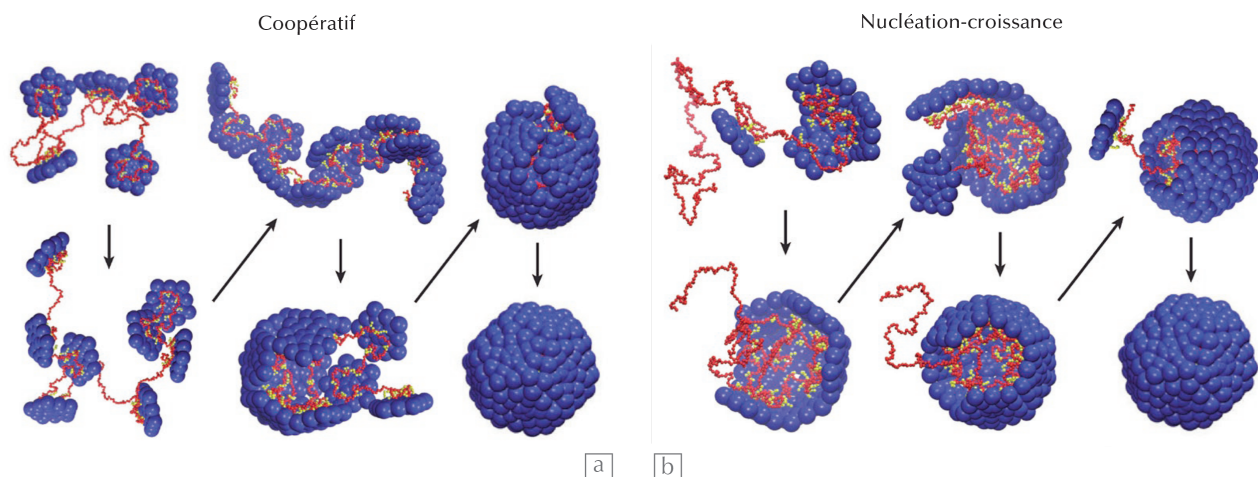


Figure 50 – MD d'assemblage d'une capsidie virale dodécaédrique autour d'un polymère chargé (d'après [2] et tirée de [3]). Les sous-unités protéiques de la capsidie sont ici modélisées par des pentagones formés de billes bleues et le polymère chargé par une chaîne de billes rouges. (a) Le modèle coopératif résulte d'abord d'une forte interaction entre les sous-unités de la capsidie et d'un polymère encapsulable (matériel génétique et/ou protéine spécifique) et ensuite d'une interaction plus faible entre les sous-unités. (b) Le modèle de nucléation-croissance au contraire résulte d'abord d'une forte interaction entre les sous-unités de la capsidie, puis d'interactions plus faibles entre les sous-unités ou les blocs de sous-unités en croissance et le polymère.

RÉSULTATS SUR LE VIRUS DE L'HÉPATITE B

7 DYNAMIQUE DE L'UNITÉ D'ASSEMBLAGE DE LA CAPSIDE DU VHB ET INTERACTION AVEC LE ZINC

7.1 Contexte biologique

Le virus de l'hépatite B (VHB) est un problème de santé majeur provoquant des infections chroniques chez l'Homme et conduisant à des maladies hépatiques, des cirrhoses ou encore au développement de carcinomes hépatocellulaires (HCC). Le VHB touche plus de 257 millions de personnes dans le monde (Figure 51). En association à d'autres facteurs de comorbidité, il conduit à plus de 700 000 morts par an [27,69,70]. Bien qu'un vaccin soit disponible, celui-ci n'est pas curatif. De nouvelles options thérapeutiques sont en développement et ciblent différentes étapes du cycle viral (réplication du génome viral, assemblage de la capside). C'est notamment le cas des modulateurs allostériques de la protéine Core du VHB (CAM) [71,72].

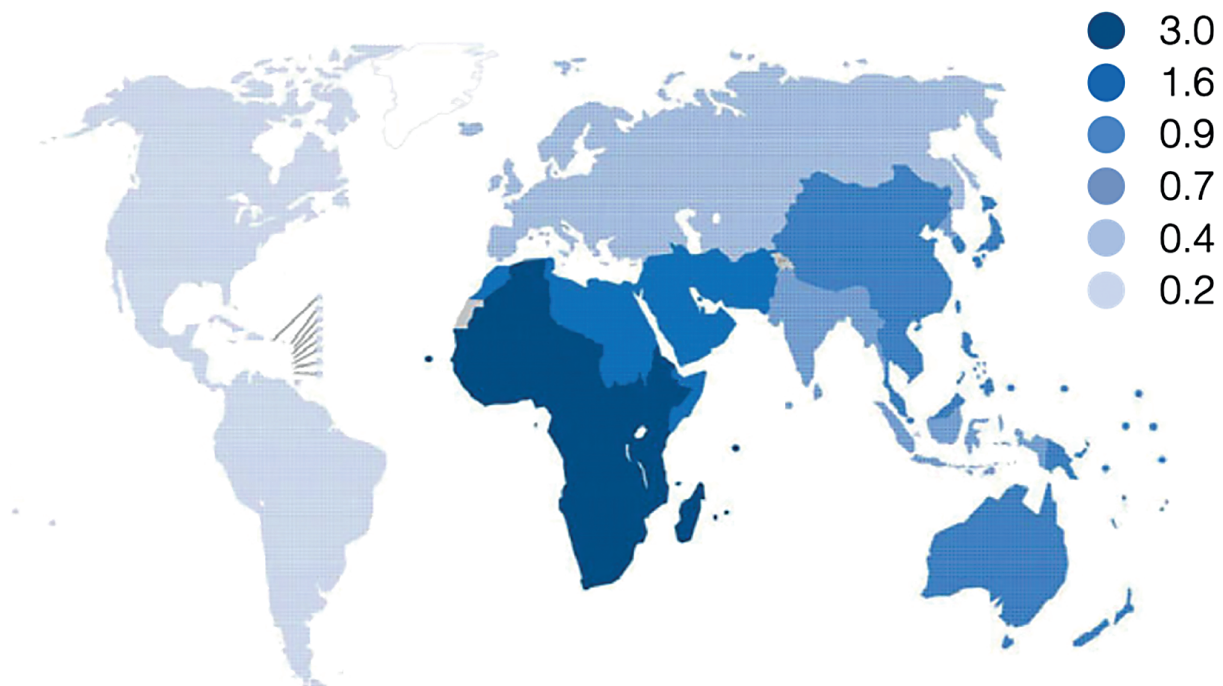


Figure 51 – Prévalence de HBsAg (%). Figure adaptée du *Global hepatitis report* [73].

7.2 Le cycle de réplication virale, Core une protéine essentielle

Le VHB est doté d'une enveloppe dont la membrane est héritée de la cellule hôte et contient les protéines d'enveloppe du virus. À l'intérieur de celle-ci se trouve une coque protéique qui encapsule le génome viral. La capside en solution s'auto-assemble à partir d'une seule protéine, ou unité d'assemblage. Cette protéine est un homodimère. Il s'agit de la protéine Core (HBc) qui est aussi l'antigène HBcAg. Elle est produite par le ribosome de l'hôte qui est détournée par le virus. Cette protéine peut aussi être exprimée artificiellement par

Escherichia Coli. La protéine Core joue un rôle essentiel et nécessaire tout au long du cycle viral (Figure 52) :

- (1) le transport et le relargage du génome viral dans le noyau,
- (2) la formation d'un mini-chromosome dans le noyau de la cellule hôte,
- (3) la régulation de l'expression des gènes de l'hôte,
- (4) La production et l'encapsulation du génome viral,
- (5) suivi d'une étape de transcription inverse et
- (6) le recyclage des nucléocapsides matures transportées vers le noyau pour amplifier le mini-chromosome [74].

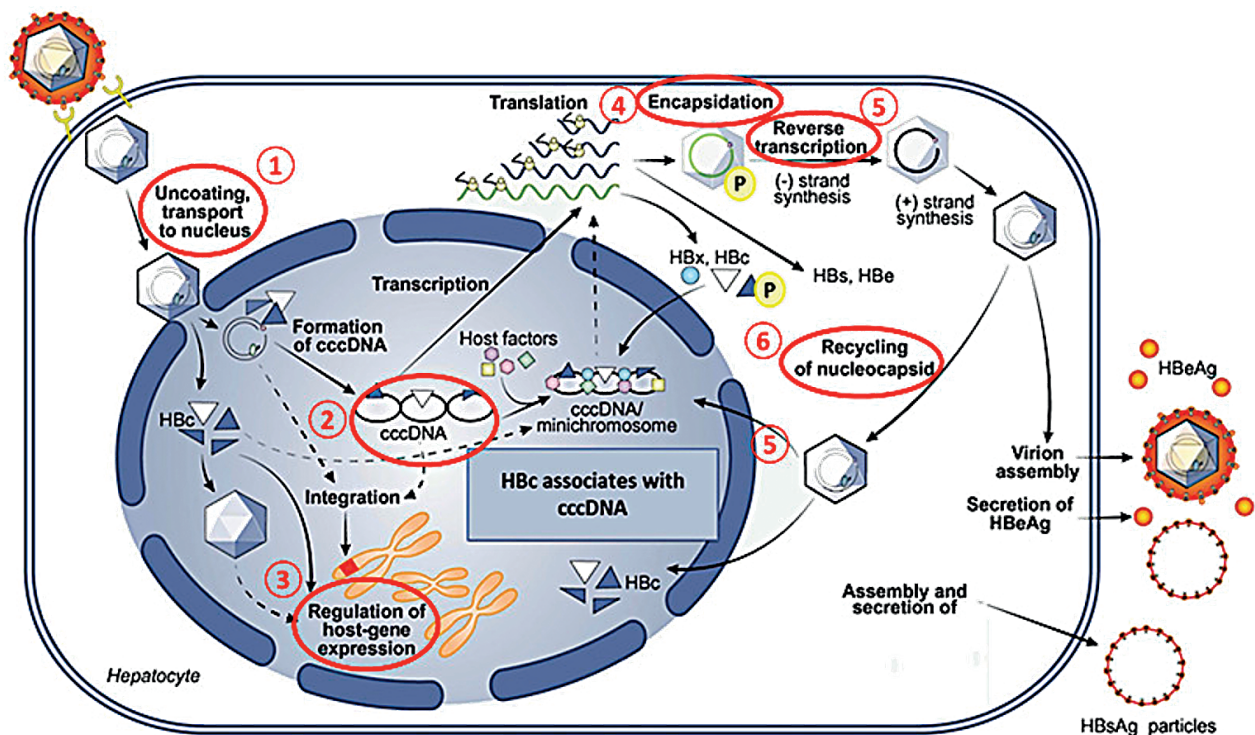


Figure 52 – Intervention de la protéine Core tout au long du cycle viral, issue de l'article [74].

Le matériel génétique existe sous 3 formes au cours de ces étapes. La nucléocapside mature contient le génome viral sous forme d'ADN relâché circulaire (ADNrc). Il est comparable à un ADN double brin circulaire mais le brin positif est incomplet. Une fois l'enveloppe du virus fusionnée avec la membrane de la cellule, la nucléocapside est transportée vers le noyau où l'ADNrc est relargué. Il est complété et réparé par la machinerie cellulaire résultant en un ADN circulaire covalentement clos (ADNccc). L'ADNccc est superenroulé et s'associe à des histones, protéines X et Core pour former un mini-chromosome. Ce mini-chromosome est la matrice de synthèse des transcrits viraux. Lors de l'étape de réplication, les transcrits les plus abondants correspondent à l'ARN pré-génomique (ARNpg). L'ARNpg est rétrotranscrit en ADNrc dans la nucléocapside [70,75].

7.3 L'unité d'assemblage : Core

La protéine Core est constituée d'un domaine d'assemblage N-terminal (résidus 1 à 149, NTD) ordonné, de structure connue et d'un domaine d'interaction avec les acides nucléiques C-terminal (résidus 150 à 183, CTD), pour lequel il n'y a pas d'information structurale [69] (Figure 53). Les structures primaires, secondaires, tertiaires et quaternaires de Core seront décrites plus amplement par la suite.

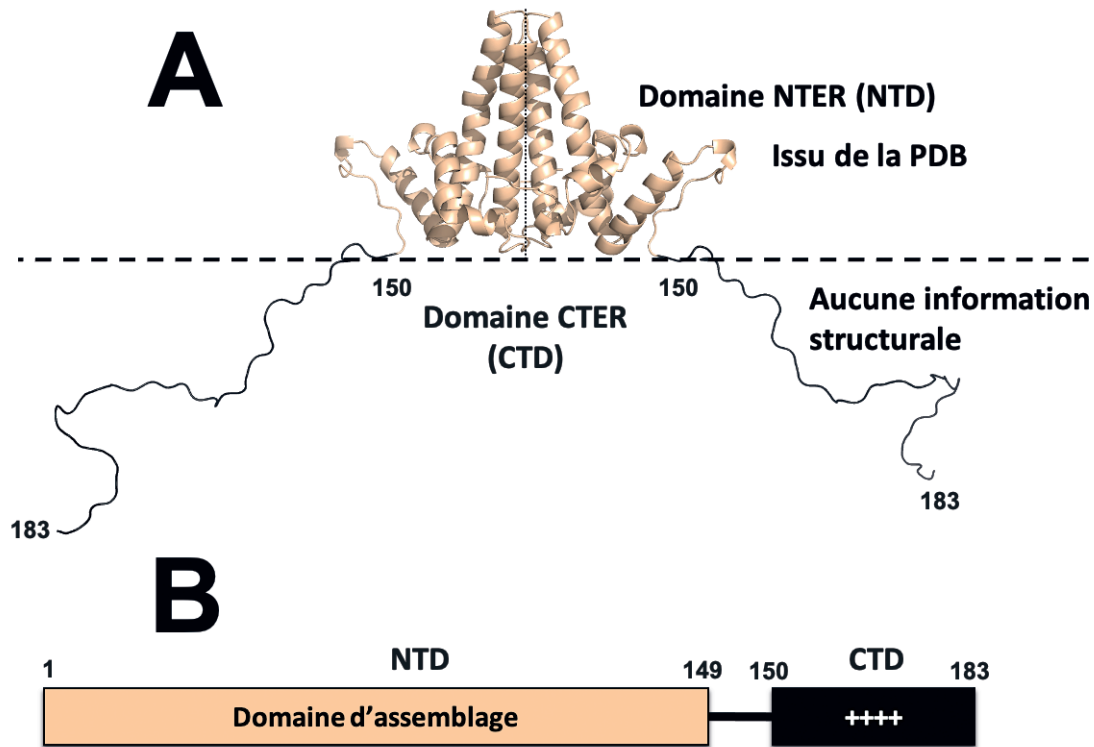


Figure 53 – Structure de la protéine Core du génotype D. (A) Structure de Core. (B) Répartition des domaines NTD et CTD. Le domaine N-terminal (NTD) est en beige. Le domaine C-terminal (CTD) est en noir. La structure du NTD est connue, contrairement à celle du CTD. On ne sait pas comment interagit le matériel génétique avec le domaine CTD chargé positivement.

7.4 L'assemblage de la capsid du VHB

La capsid, de géométrie icosaédrique, existe in-vivo (lors de l'infection) et in-vitro (quand Core est produite en *Escherichia Coli*), sous 2 formes : la T=3 ou la plus commune, la T=4 [69]. L'architecture "sphérique" T=4 est constituée de 240 sous-unités divisées en 4 monomères (A, B, C et D). Les structures primaires et secondaires des monomères sont communes mais leur structure tertiaire diverge. Dans le contexte de la capsid (T=4 - Figure 54A), 4 conformations de Core sont retrouvées sous forme de dimères A-B ou C-D (Figure 54B). Il faut 60 dimères A-B et 60 dimères CD pour former la capsid du VHB (Figure 54C).

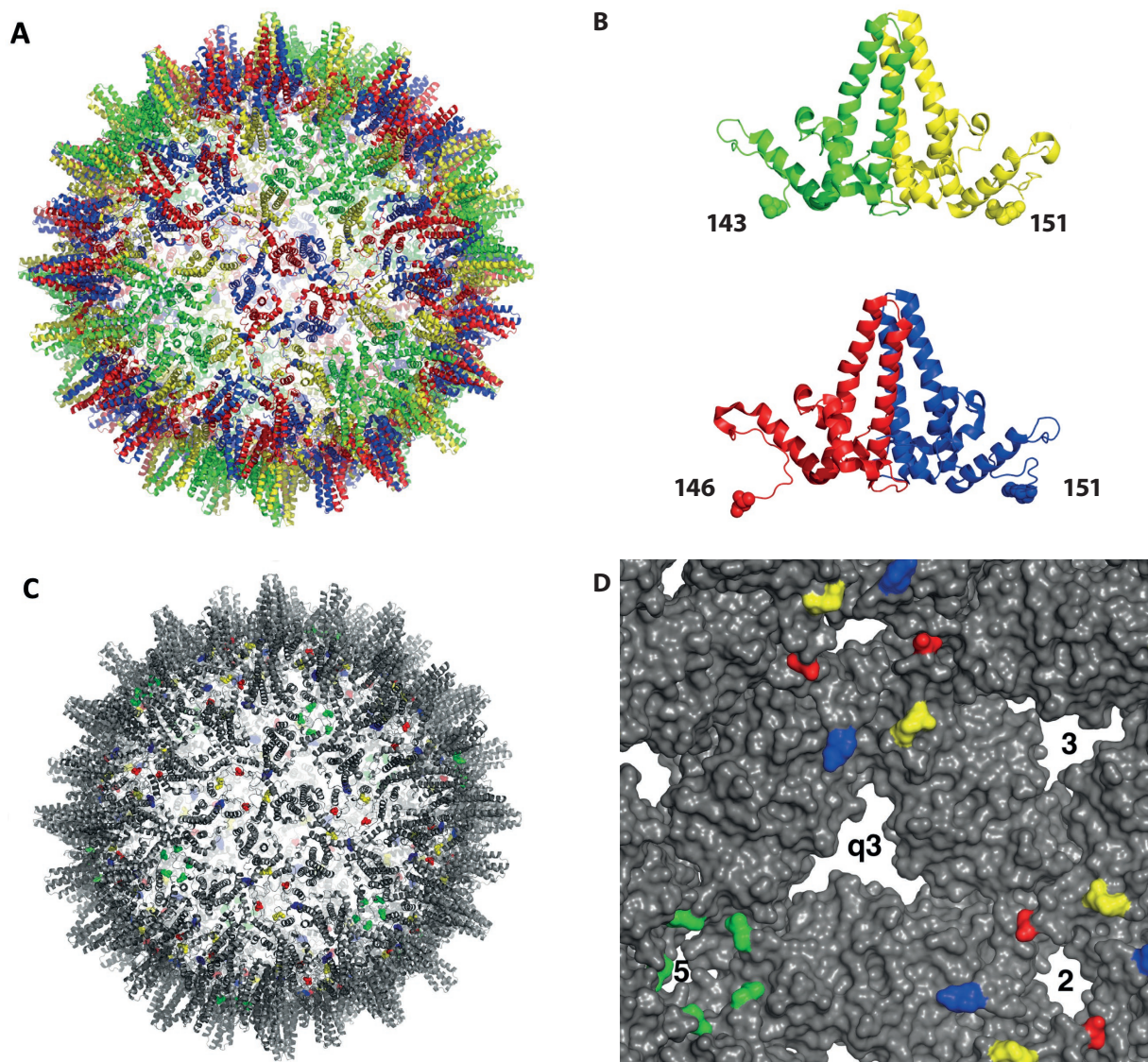


Figure 54 – Organisation de la capside de HBV. A. 120 dimères de NTD s’organisent en icosaèdre $T=4$. Les dimères A-B sont en vert et jaune. Les dimères C-D, de conformation légèrement différente, sont en rouge et bleu. B. Les dimères A-B et C-D isolés de la capside. Les derniers résidus ordonnés sont représentés par des sphères. C. Localisation des derniers 240 résidus ordonnés dans la capside, vue dans la même orientation qu’en A. D. Zoom sur l’intérieur de la capside en représentation de surface. Dans cette représentation, on voit clairement que les pores, dans la capside, sont situés aux axes quasi-6 (q6 ou 2), quasi (3) (q3 ou l3), 3 et 5, avec des ouvertures plus importantes au niveau de q3 et 3.

Dans le contexte de la nucléocapside, les CTD sont situés à l’intérieur de la capside, en contact avec le matériel génétique. La nucléocapside est close mais fenestrée avec des pores assez grands pour permettre le passage de CTD (Figure 54D).

La formation de la capside du VHB suivrait un modèle de nucléation-croissance (Figure 55) [14]. Dans ce modèle, la vitesse de dissociation est supérieure à la vitesse d’association des unités d’assemblage. Les sous-unités d’assemblage produiraient, par nucléation, les noyaux critiques. Un type de noyau est susceptible de s’assembler. Il s’agit d’un trimère de dimères (qui correspondrait soit au pore q3, soit au pore 3). Dans le modèle de nucléation-croissance, la vitesse de dissociation est du premier ordre (constante) et la vitesse d’association, du deuxième ordre (proportionnelle à la concentration en dimère). La formation du noyau

critique est cinétiquement limitante : il se forme lentement et se consomme vite. Pour que le noyau critique se forme, il faut que la concentration en unités d'assemblage soit suffisante ($\sim 50 \mu\text{M}$) pour que la vitesse d'association dépasse la vitesse de dissociation. Le noyau critique pourrait alors croître en amarrant successivement des unités d'assemblage. La conformation de ces unités pourrait changer pour s'adapter à la morphologie de la capsid en cours de construction. À noter que, seul le NTD suffit à l'assemblage de la capsid.

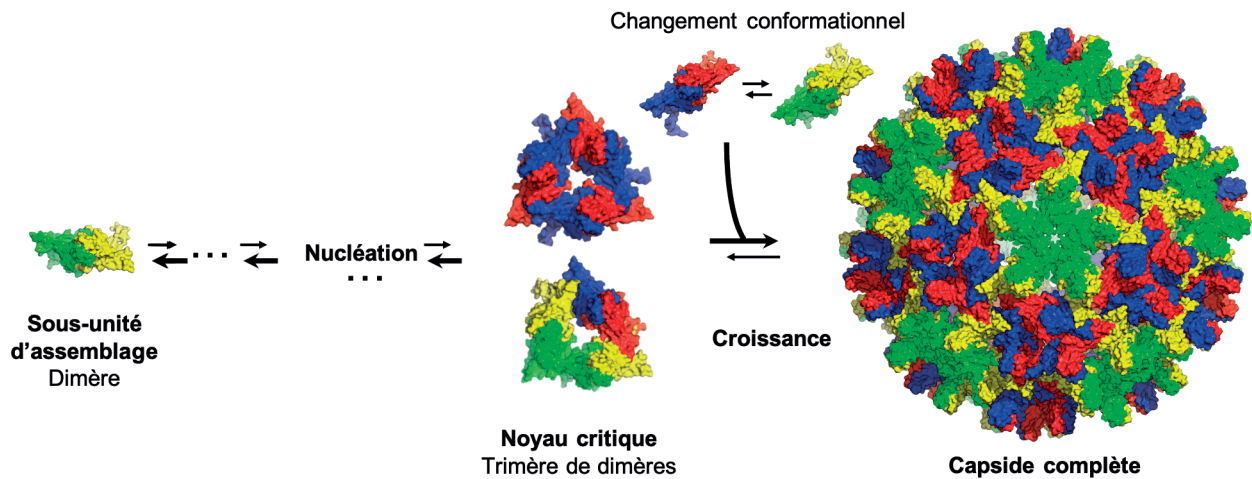


Figure 55 – Croissance théorique de la capsid du VHB.

Une étude de l'assemblage de la capsid du VHB par TR-SAXS confirme que l'auto-assemblage de la capsid suit un modèle de nucléation croissance [76] (Figure 56). Elle révèle aussi une troisième phase lente, dite, de "relaxation". Au cours de cette phase, les unités d'assemblage, incorporées dans la capsid en formation, pourraient s'auto-organiser pour conduire à une capsid de géométrie icosaédrique T=4. Cette phase pourrait aussi correspondre à la capture des dernières sous-unités. Il est plausible qu'elle corresponde à ces 2 processus, se déroulant simultanément.

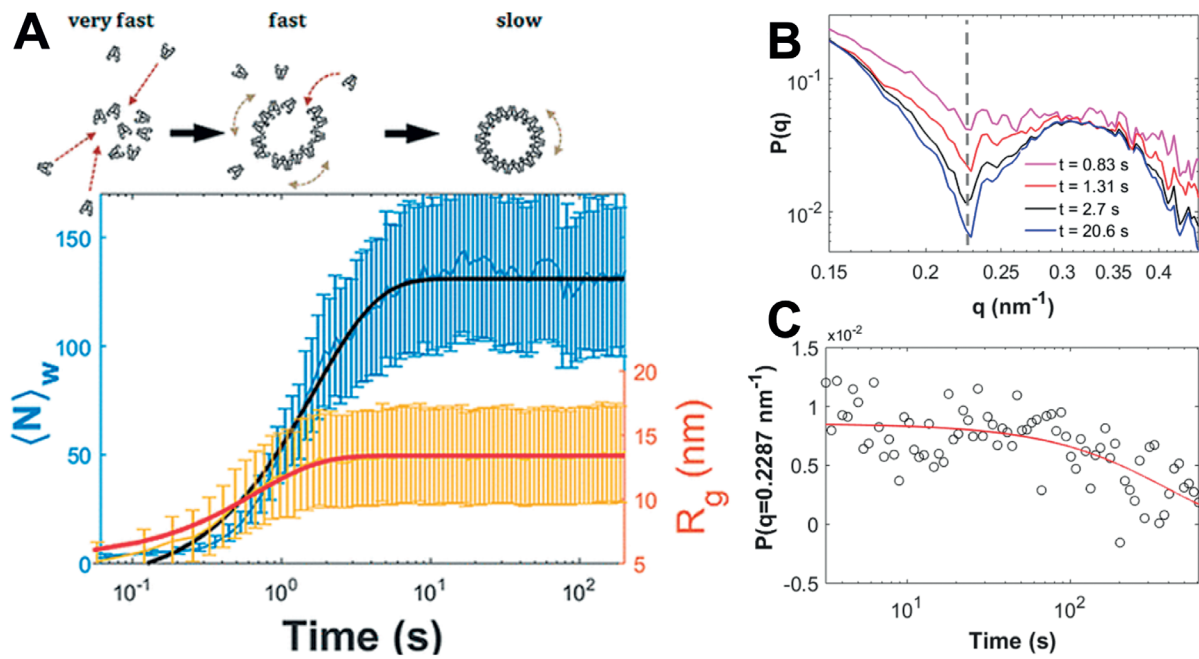


Figure 56 – Assemblage du NTD suivi par TR-SAXS, issu de l'article [76]. A. Évolution du nombre d'agrégation et du rayon de giration au cours du temps. $\langle N \rangle_w$ correspond au nombre d'agrégation moyenne et R_g au rayon de giration. B. Facteurs de formes $P(q)$ pour différents points de temps. La ligne verticale, en pointillé, indique $q=0,2287 \text{ nm}^{-1}$. C. Valeurs du facteur de forme à $q=0,2287 \text{ nm}^{-1}$.

L'assemblage *in vitro* de la protéine Core tronquée (Cp149) est couramment déclenché à partir de dimères dissociés en ajoutant du sel. Dans leur publication, Stray et al. mettent en évidence que l'assemblage de Cp149 est déclenché à partir de dimères dissociés, en ajoutant du Zn^{2+} , à plusieurs ordres de grandeur inférieures à Na^+ , K^+ , Mg^{2+} et Ca^{2+} [77].

7.5 SIGNAUX DE LOCALISATION NUCLÉAIRE ET EXPOSITION DU CTD

In vivo, c'est une Core phosphorylée sur son CTD qui encapside l'ARNpg et la polymérase. *In vitro*, une étude montre que la mutation de 3 sérines (3 sites de phosphorylation majeurs : S155, S162 et S170) en glutamate, de façon à mimer la phosphorylation constitutive des sérines, montre que Core est compétente pour encapsider l'ARNpg mais déficiente pour d'autres étapes [78,79]. Les étapes successives de la conversion de l'ARNpg en ADNrc [80] s'accompagnent de déphosphorylations successives du CTD. Ceci conduit à l'exposition du CTD à l'extérieur de la nucléocapside mature.

La protéine Core comporte des signaux de transports intracellulaires sur son CTD [27,70]. Le CTD joue un rôle extrêmement important dans le cycle viral du VHB et notamment dans le transport du matériel génétique vers le noyau. Le CTD ne comporte pas moins de quatre motifs riches en arginines (ARD ou Arginine Rich Domains - Figure 57) [81].

Les ARD constituent des signaux de localisation et d'exportation nucléaire. Lorsque les deuxième (ARD-II, résidus 157-159) et quatrième (ARD-IV, résidus 172-175) signaux sont mutés, ils semblent conjointement affecter l'exportation de Core. Cette exportation s'effectue du noyau vers le cytosol. Core s'accumule alors dans le noyau des cellules hôtes. Par élimination, les premier (ARD-I, résidus 150-152) et troisième (ARD-III, résidus 164-167) motifs riches en arginines sont associés comme des signaux de localisation nucléaire. Ces 2 motifs doivent agir de façon co-dépendante et synergique car seuls, ils ne suffisent pas à l'importation de Core vers le noyau.

Les ARD constituent des signaux de localisation et d'exportation nucléaire. Lorsque les deuxième (ARD-II, résidus 157-159) et quatrième (ARD-IV, résidus 172-175) signaux sont mutés, ils semblent conjointement affecter l'exportation de Core. Cette exportation s'effectue du noyau vers le cytosol. Core s'accumule alors dans le noyau des cellules hôtes. Par élimination, les premier (ARD-I, résidus 150-152) et troisième (ARD-III, résidus 164-167) motifs riches en arginines sont associés comme des signaux de localisation nucléaire. Ces 2 motifs doivent agir de façon co-dépendante et synergique car seuls, ils ne suffisent pas à l'importation de Core vers le noyau.

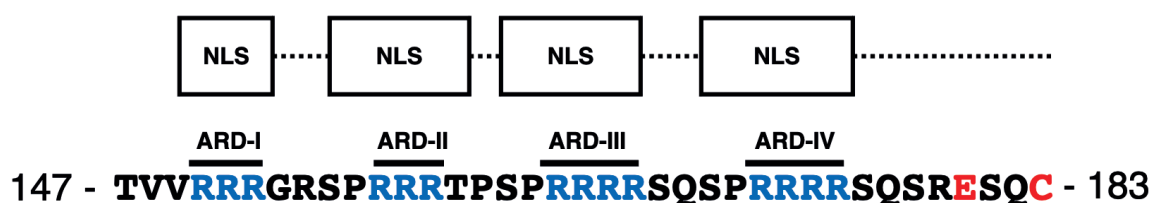


Figure 57 – Résumé de l'attribution des signaux de localisations du CTD de la protéine Core, issu de l'étude [81]. Dans la légende, NES : Signal d'exportation nucléaire, NLS : Signal de localisation nucléaire.

Des études montrent que les CTD peuvent être exposés en dehors de la capsid [82,83]. Une chromatographie, en phase liquide-spectrométrie de masse (LC-MS) [83] et une

électrophorèse sur gel (SDS-PAGE ou NAGE) [82], après une digestion à la trypsine, ont permis de le mettre en évidence. La phosphorylation du CTD influence l'exposition des bras C-terminaux par les pores. Si on mute les 3 sites de phosphorylations majeurs du CTD (sérines S155E, S162E et S170E), cela joue également sur l'exposition des bras [83]. En résumé, plus le CTD est phosphorylé, moins il est capable de sortir par les pores. Selon ces études, la partie du CTD majoritairement exposée correspond à la région immédiatement après le résidu 157, notamment quand les capsides (recombinantes) sont déphosphorylées (-SRPK1 - Figure 58) [82]. L'une des hypothèses d'exposition du CTD serait qu'une boucle, comportant le résidu 157, se forme sur le CTD et soit exposée par les pores. La partie qui suit cette boucle pourrait donc être enfouie dans la capside (Figure 58). Selon les auteurs, il est aussi possible que toute la partie du CTD, au-delà du résidu 157, soit exposée à l'extérieur de la capside.

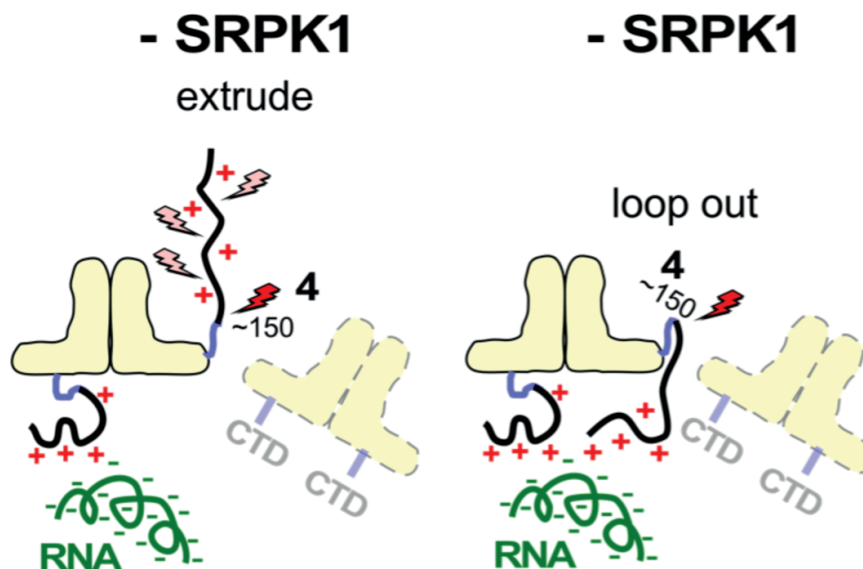


Figure 58 – Hypothèses d'expositions des CTDs en dehors des pores, issues de l'étude [82]. À gauche, exposition de tout le CTD. À droite, exposition d'une boucle sur le CTD.

Il existe un modèle théorique qui décrit la distribution radiale du CTD des dimères, au cours de la reverse transcription virale [84]. Ce modèle se base sur la théorie de la fonctionnelle de la densité (DFT) et d'un modèle gros grains. Il montre une proportion significative de CTD exposés à la surface des nucléocapsides immatures (ARNpg), tandis que les CTD sont, pour la plupart, confinés à l'intérieur des nucléocapsides matures (ADN). À noter, que dans le cas des nucléocapsides immatures, les sérines (S155, S162 et S170) sont chargées négativement. La reconstruction d'image Cryo-EM montre également que le CTD des dimères peut être exposée à la surface de la capside pour venir interagir avec des importines β seules. Les CTD comportant les signaux de localisation nucléaire, une fois exposés, contribuent au transport de la protéine Core ou des nucléocapsides vers le noyau. Cela se fait grâce à l'interaction des CTD exposés avec un complexe, qui inclut l'importine α et l'importine β [85].

L'objectif de cette étude est de comprendre : (1) quelle est la dynamique du CTD dans le contexte du dimère libre en solution ? (2) Quelle est la dynamique du CTD dans le contexte d'une capsid ? (3) Comment Core, via son CTD, interagit avec du matériel génétique ? J'ai mené une étude, majoritairement computationnelle et également expérimentale, est menée sur la protéine Core du VHB pour répondre à ces questions.

en 3 groupes : résidus chargés négativement, résidus chargés positivement et enfin, autres résidus polaires. Ce classement indique que Core est une protéine amphiphile (Figure 59C). Ce même type d'analyse sur les domaines montre que le NTD comporte une quantité quasi-équivalente de résidus hydrophobes et polaires (Figure 60B). Les résidus hydrophobes sont localisés dans un cœur et au niveau de l'interface de dimérisation des monomères de Core (Figure 60A). Les résidus polaires tapissent toute la surface du NTD. Le CTD ne comporte pas de résidu hydrophobe. Il est principalement composé de résidus polaires chargés positivement (Figure 60C et D).

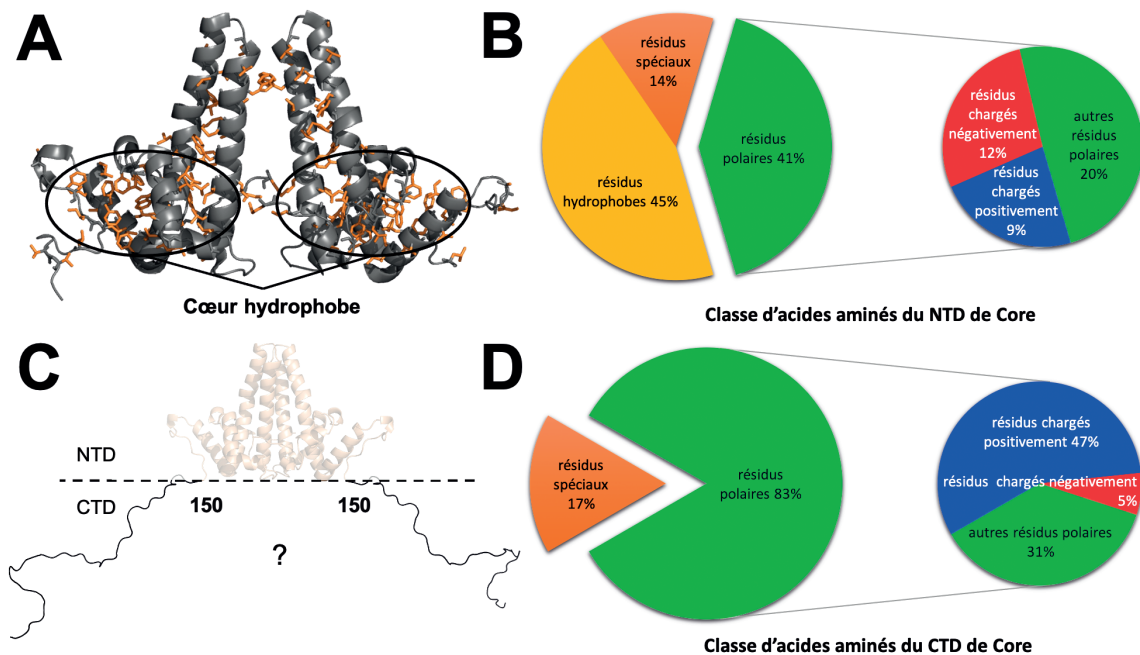


Figure 60 – Composition en acides aminés des domaines de Core. A. Domaine NTD de Core résolu [69], les monomères sont représentés en gris et les résidus hydrophobes, en orange. B. Classement des acides aminés du NTD. C. Domaine CTD de structure inconnue du résidu 150 à 183. D. Classement des acides aminés du CTD.

S'il est considéré que la charge d'une arginine est égale à 1 et que la charge d'une cystéine terminale ou d'un glutamate est égale à -1, la charge positive globale du CTD serait de +14 (pour 34 résidus ; 150 à 183). Cette charge pourrait jouer un rôle majeur dans l'interaction de Core avec le matériel génétique, qui lui, est chargé négativement.

7.6.2 Structures tertiaires de Core

Il existe une trentaine de modèles de la structure tridimensionnelle (3D) de la protéine Core du VHB, en 2021 (Tableau 2).

Ces modèles sont issus d'expériences de cristallographie aux rayons X (p. ex. [69]) ou de cryo-microscopie électronique (p. ex. [87]). Les modèles ont (1) une résolution inférieure à 5 Å et (2) une structure du NTD parfaitement définie, alors que le CTD ne l'est pas. Pour certaines études où la protéine Core tronquée (Cp149) a été produite, on ne dispose donc que des données structurales du domaine d'assemblage. Il est difficile de résoudre le CTD car il n'est pas composé d'éléments structurants. Il est très flexible et très hydrophile. Dans la plupart des structures cristallographiques et en cryo-microscopie électronique, la protéine Core se trouve dans une capsidie assemblée. Seul le mutant Y132A cristallise dans un assemblage partiel (trimère de dimères), qui permet d'avoir des données à haute résolution (<2 Å).

Le CTD de Core joue un rôle dans des étapes du cycle viral. En son absence, il n'y a pas d'interaction avec du matériel génétique. Or, Core est constamment en interaction avec du matériel génétique (ARNpg, ADNrc, ADNccc, régulation des gènes de l'hôte) au cours du cycle viral. Les fonctions multiples de Core sont assurées grâce à une capacité à ajuster sa structure. La dynamique de Core et plus particulièrement à celle du CTD, ainsi que les contacts pouvant se former sont étudiés à la suite.

PDB ID [REF]	TECHNIQUE	RÉSOLUTION (en Å)	CORE	MUTATION	GENOTYPE
1QGT [69]	XRAY	3,3	Cp149	0	D
2G33 [88]	XRAY	3,96	Cp149	C48A, C61A C107A	D
2G34 [88]	XRAY	5,05	Cp149	C48A, C61A C107A	D
2QIG [89]	XRAY	8,9	Cp149	0	D
3J2V [87]	CryoEM	3,5	Cp183	0	A2
3KXS [90]	XRAY	2,25	Cp149	Y132A	D
4BMG [91]	XRAY	3	Cp149	Y132A	D
4G93 [92]	XRAY	4,2	Cp149	C48A, C61A C107A	D
5D7Y [93]	XRAY	3,89	Cp149	C48A, C61A, V93M, C107A	D
5E0I [94]	XRAY	1,95	Cp149	Y132A	D
5GMZ [95]	XRAY	1,7	Cp149	Y132A	D
5T2P [96]	XRAY	1,69	Cp149	Y132A	D
5WRE [96]	XRAY	1,95	Cp149	Y132A	D
5WTW [96]	XRAY	2,62	Cp149	Y132A	D
6BVF [72]	CryoEM	4	Cp149	C48A, C61A C107A	D
6BVN [72]	CryoEM	4	Cp149	C48A, C61A C107A	D
6ECS (Woodchuck) [97]	XRAY	2,90	Cp149	Y132A	Isolate 1
6EDJ (Woodchuck) [97]	CryoEM	4,52	Cp149	0	Isolate 1
6HU4 [98]	CryoEM	2,64	Cp183	F97L	D
6HU7 [98]	CryoEM	2,80	Cp183	F97L	D
6HTX [98]	CryoEM	2,66	Cp183	0	D
6J10 [99]	XRAY	2,3	Cp149	Y132A	D
6TIK [100]	CryoEM	3,40	Cp149	0	D
6UI6 [101]	CryoEM	3,53	Cp149	0	A2
6UI7[101]	CryoEM	3,65	Cp149	0	A2
6VZP [102]	CryoEM	3,60	Cp149	0	D
6W0K [102]	CryoEM	4,60	Cp149	D78S	D
6WFS [103]	CryoEM	4,60	Cp149	C48A, C61A, C107A, C150	D
6YGI (Duck) [104]	CryoEM	3,00	Cp149	R124E	DHBV-16
6YGH (Duck) [104]	CryoEM	3,70	Cp149	0	DHBV-16

Tableau 2 - Données structurales expérimentales de la protéine Core du VHB. Les lignes en gras correspondent aux structures de la protéine Core utilisées pour les étapes de modélisation et de simulation. Cp149 : protéine Core tronquée. Cp183 : Protéine Core complète.

7.7 Le domaine C-terminal (CTD) de la protéine Core – un véritable couteau suisse

Cette partie est une synthèse des travaux que j'ai mené sur la caractérisation de la dynamique et de la structure du CTD. La structuration du CTD est étudiée dans le contexte du dimère isolé et dans le contexte d'une capsidie assemblée.

7.7.1 Dynamique du domaine d'assemblage

Comme NTD constitue le domaine d'assemblage, il est important de mettre en évidence sa dynamique afin d'avoir une meilleure idée de sa flexibilité.

Une simulation de dynamique moléculaire (MD) tout atomes et un calcul de modes normaux sont produits à partir de la structure cristallographique d'un dimère AB résolu en 2015 ; $< 2 \text{ \AA}$ [94]. Le domaine NTD de cette structure est résolu jusqu'au résidu 148. Après avoir rétabli la tyrosine de la mutation Y132A de la structure, le NTD tronqué (Cp140), des résidus 1 à 140 est simulé durant 105 ns.

La déviation (RMSD) et la fluctuation quadratique moyenne (des résidus : RMSF) du NTD tronqué (résidus 1 à 140) sont calculées (Figure 61). Le RMSD indique que la structure du NTD converge en un peu plus de 10 ns, puis se stabilise (Figure 61A). Le domaine ne diverge que de $\sim 2,5 \text{ \AA}$ de la structure au temps $t = 0 \text{ ns}$ (structure après pré-production). Le RMSF montre que les résidus composant les hélices et les boucles qui les relient divergent très peu de cette même structure initiale, environ $2,5 \text{ \AA}$ au maximum (Figure 61B). Les résidus, au-delà de l'hélice $\alpha 5$, comportent une région qui diverge davantage par rapport au reste du NTD (résidus 120 à 140 : $\sim 4 \text{ \AA}$ au maximum).

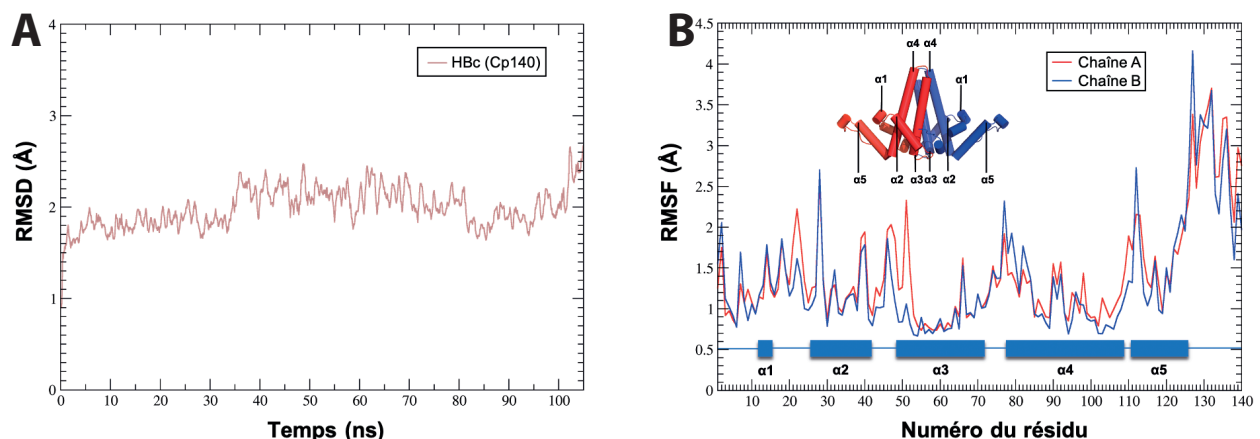


Figure 61 – Déviation (RMSD) et fluctuation quadratique moyenne (des résidus : RMSF) du NTD tronqué (Cp140).

Cette région correspond à l'interface d'interaction entre les dimères. À cette échelle de temps, le domaine d'assemblage de la capsidie est extrêmement stable.

Les simulations de dynamique moléculaire sont utilisées pour étudier les mouvements d'origine thermique d'une petite protéine (~ 100 -600 résidus), de la centaine de nanosecondes

à la microseconde. Pour aller au-delà de la microseconde, il est possible d'étudier les mouvements d'une protéine, en faisant une approximation des petits déplacements. Le mouvement de chacun des atomes d'un système est décrit comme une combinaison de modes de vibration indépendants les uns des autres (les modes normaux).

L'analyse des modes normaux par la méthode des blocs rigides non-linéaires (NOLB) [105], révèle que les hélices qui forment la spicule ($\alpha 3$ et $\alpha 4$) peuvent être soumises à des mouvements élémentaires qui les courbent ou les tordent (1, 2 et 3 de la Figure 62). Les hélices $\alpha 5$ peuvent également se courber ou se tordre (3 et 4 de la Figure 62).

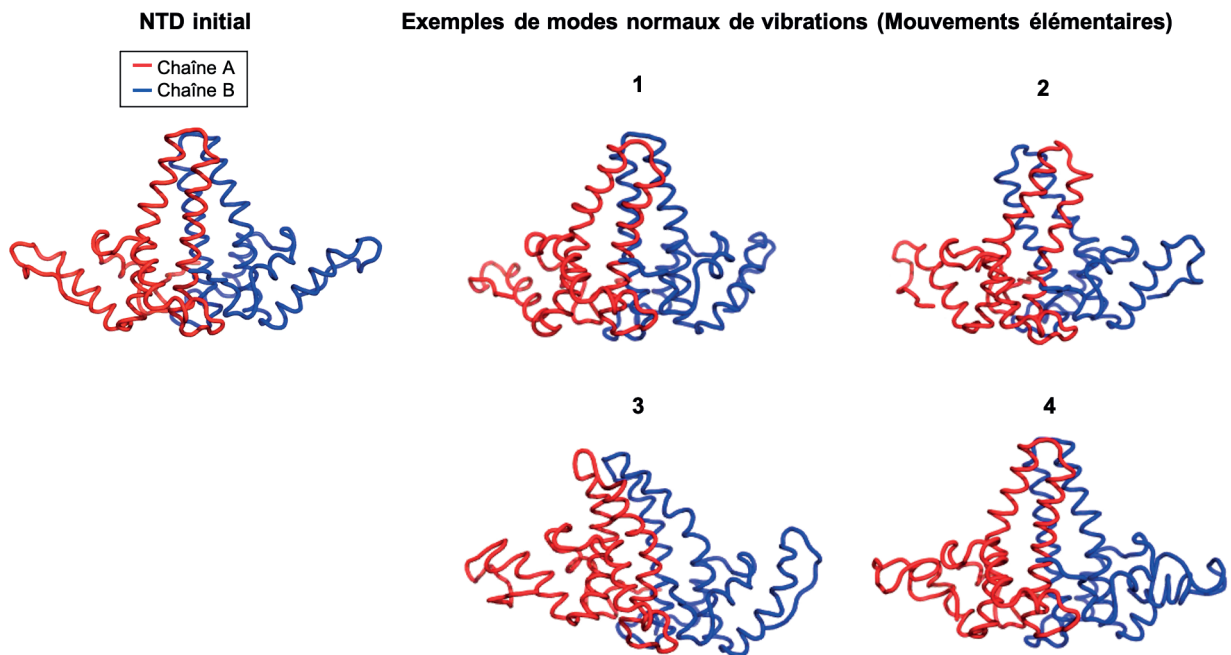


Figure 62 – Analyse des modes normaux selon la méthode des blocs rigides non-linéaires (NOLB).
La chaîne A est en rouge et la chaîne B, en bleu.

En conclusion, le NTD est à la fois stable et flexible. Les mouvements élémentaires sont localisés au niveau de la spicule et des zones de contacts entre dimères.

7.7.2 Modélisation du CTD

Pour étudier la protéine Core complète, j'ai rajouté les résidus manquants (CTD) de la structure cristallographique ayant la meilleure résolution en 2018 (PDB ID : 5E0I) [94] en suivant différentes méthodes de modélisation ; résolution à 2.0 Å. Le dimère AB de cette structure a été résolu des résidus 1 à 148 et 152 à 156. Les résidus 152 à 156 n'ont pas été inclus dans le "template" pour modéliser le CTD. La tyrosine mutée Y132A a, elle été rétablie. Pour explorer au maximum l'espace conformationnel du CTD, 4 serveurs ou programmes de modélisation sont utilisés (Figure 63). Deux d'entre eux reposent sur une modélisation par homologie : MODELLER 9.19 [39] et Protein Portal Model [106]. L'un repose sur la modélisation par enfilage : I-TASSER [107] et l'autre combine les 2 méthodes de modélisation : ROBETTA [108].

Le choix de modéliser le CTD avec plusieurs outils a permis d'obtenir un grand nombre de modèles. Cette démarche a l'avantage de nous permettre à explorer, plus largement, l'espace conformationnel du CTD de Core. Les modèles, caractérisés par les structures 3D de CTD les plus différentes, sont sélectionnés (Figures 63 à 69). Ces modèles présentent déjà des propriétés intéressantes.

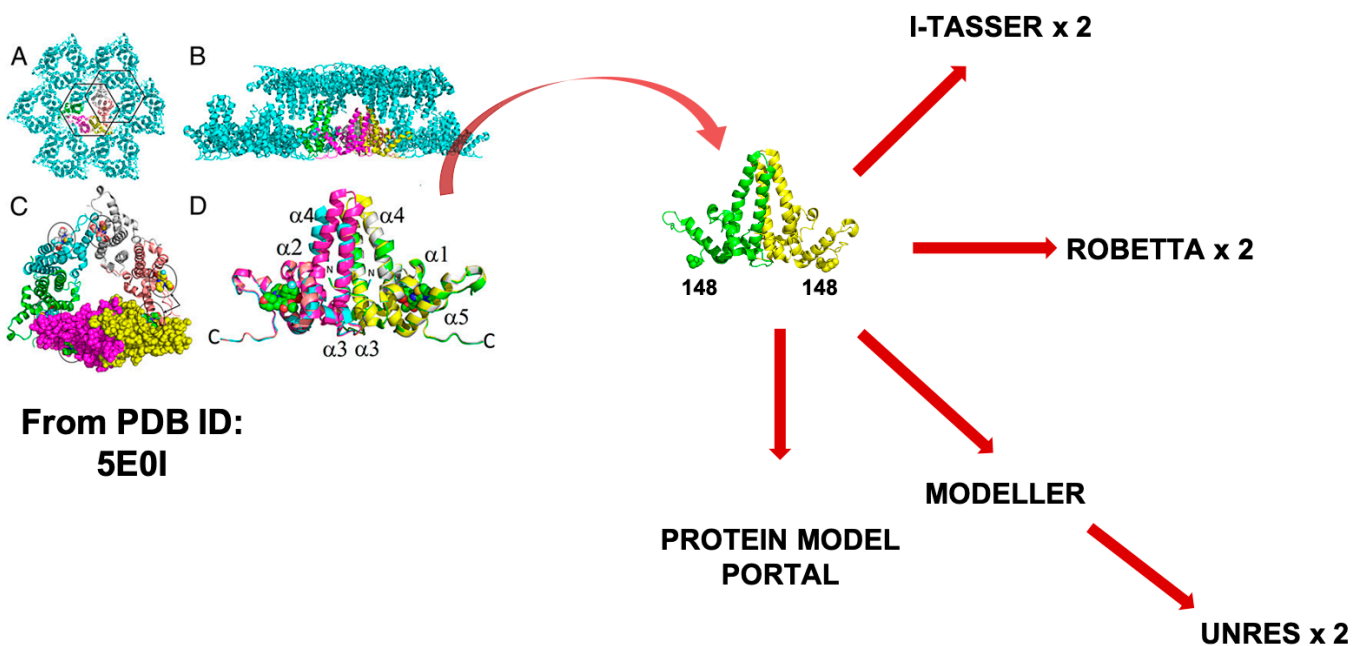


Figure 63 – Modélisation du CTD et MD de la protéine Core. Un modèle est produit par Protein Model Portal [106] et MODELLER [39]. 2 modèles sont sélectionnés parmi 5 produits par ROBETTA [108] et I-TASSER [107].

Les structures, issues de la modélisation par homologie, figurent ci-dessous : modèle MODELLER [39] (Figure 64) et modèle Protein Model Portal [106] (Figure 65).

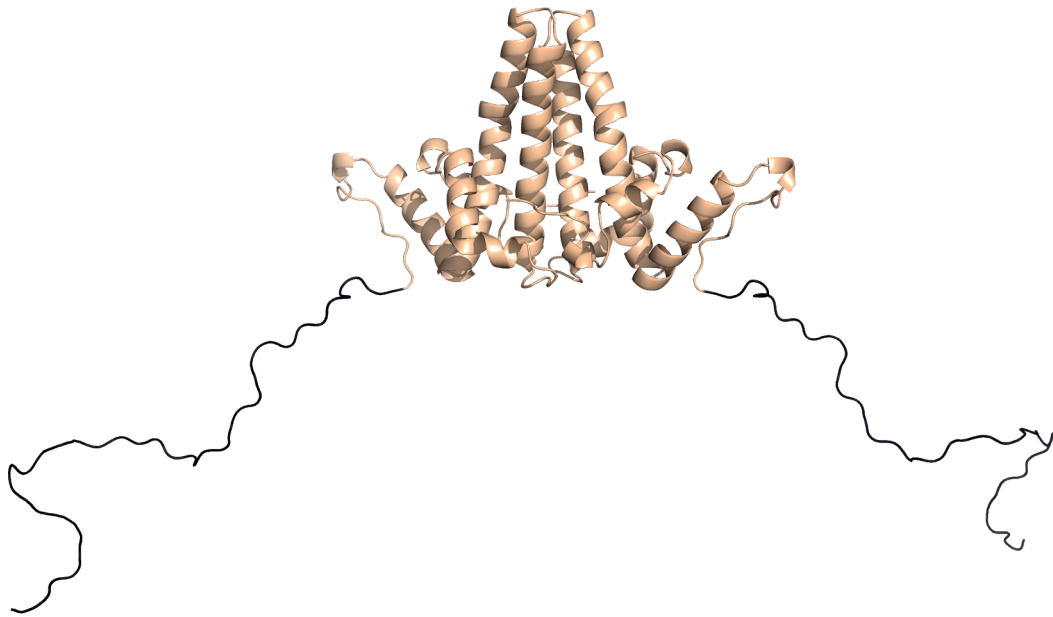


Figure 64 – Modélisation par homologie avec MODELLER [39] du CTD du modèle de Core. Le NTD est coloré en beige. Le CTD est en noir, n'est pas structuré. Les bras C-terminaux sont complètement dépliés.

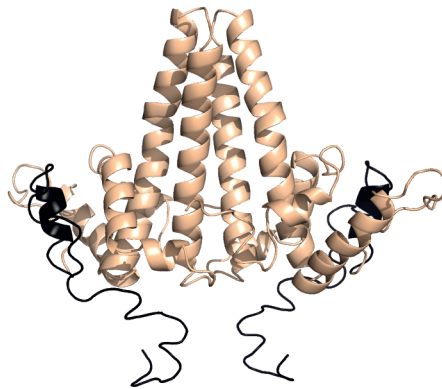


Figure 65 – Modélisation par homologie avec Protein Model Portal [106] du CTD du modèle de Core. Le NTD est coloré en beige. Le CTD est en noir. Les bras C-terminaux sont, en partie, repliés en contact avec les hélices $\alpha 5$ et en dessous du NTD. Une petite hélice se forme sur les 2 bras.

Les structures, issues de la modélisation par enfilage, figurent ci-dessous : le premier modèle produit par ITASSER [107] (Figure 66) et le deuxième modèle généré par ITASSER (Figure 67).

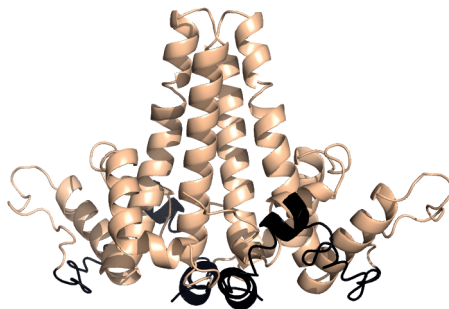


Figure 66 – Première modélisation par enfilage avec I-TASSER [107] du CTD du modèle de Core. Le NTD est coloré en beige. Le CTD est en noir. Les bras C-terminaux sont en partie repliés en dessous du NTD. Ils sont partiellement structurés en hélices.

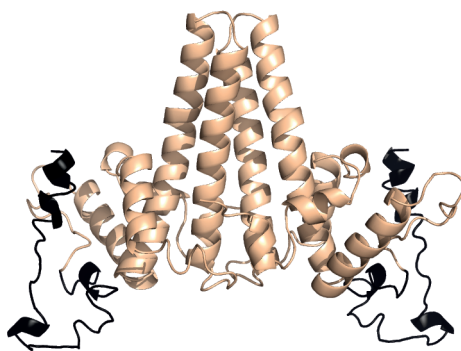


Figure 67 – Deuxième modélisation par enfilage avec I-TASSER [107] du CTD du modèle de Core. Le NTD est coloré en beige. Le CTD est en noir. Les bras C-terminaux sont, en partie, repliés en contact et en dessous des hélices $\alpha 5$. Ils se structurent partiellement en hélices.

Les structures, issues de la combinaison des 2 méthodes de modélisation, figurent ci-dessous : le premier modèle construit par ROBETTA [108] (Figure 68) et le deuxième modèle généré par ROBETTA (Figure 69).

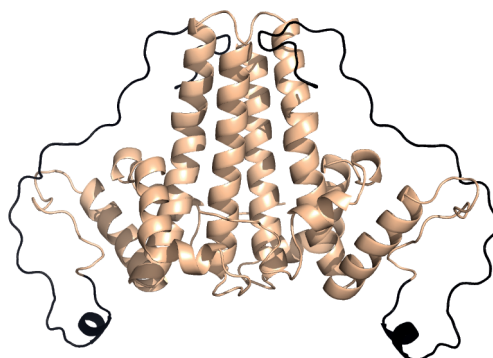


Figure 68 – Première modélisation combinant l'homologie et l'enfilage avec ROBETTA [108] du CTD du modèle de Core. Le NTD est coloré en beige. Le CTD est en noir. Les bras C-terminaux sont en contact avec la spicule. Il n'y a quasiment pas de structures secondaires.

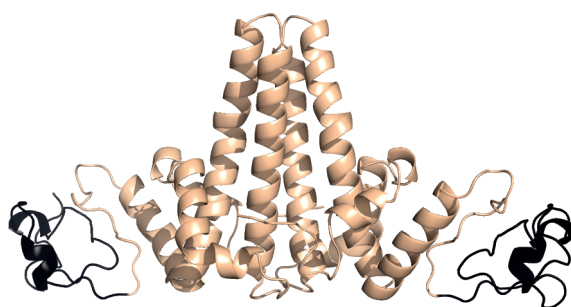


Figure 69 – Deuxième modélisation combinant l'homologie et l'enfilage avec ROBETTA [108] du CTD du modèle de Core. Le NTD est coloré en beige. Le CTD est en noir. Les bras C-terminaux sont en contact avec les hélices $\alpha 5$ (zones d'interactions entre dimères). Ils se structurent partiellement en hélices.

7.7.3 Dynamique du CTD dans le contexte du dimère seul

Des MD tout-atomes, de 105 ns, sont produites pour étudier la dynamique du CTD (~15 000h consommées pour 105 ns de MD sur des processeurs Intel® Xeon® CPU E5-

2690 v4 de 2,60GHz pour les plus gros systèmes : 481 279 atomes – MODELLER). Les MD sont réalisées à partir des modèles cités précédemment. Elles sont calculées à l'aide de GROMACS [44] et du champ de force Amber 99SB-ILDN [109]. Une simulation et une réplique sont produites pour les modèles ROSETTA (2 x 2 simulations) et MODELLER (2 simulations). Pour les autres modèles, une seule MD est réalisée (3 simulations). Une MD supplémentaire, pour le modèle MODELLER, est effectuée à une température de 330K (1 simulation). 10 MD sont calculées à partir de ces 6 modèles.

Pour aller plus loin dans l'exploration de l'espace conformationnel du CTD, une MD avec échanges de répliques (REMD) du modèle MODELLER est réalisée sur le serveur UNRES [110]. 2 modèles, issus de UNRES (Figures 70 et 71), sont simulés (2 MD).

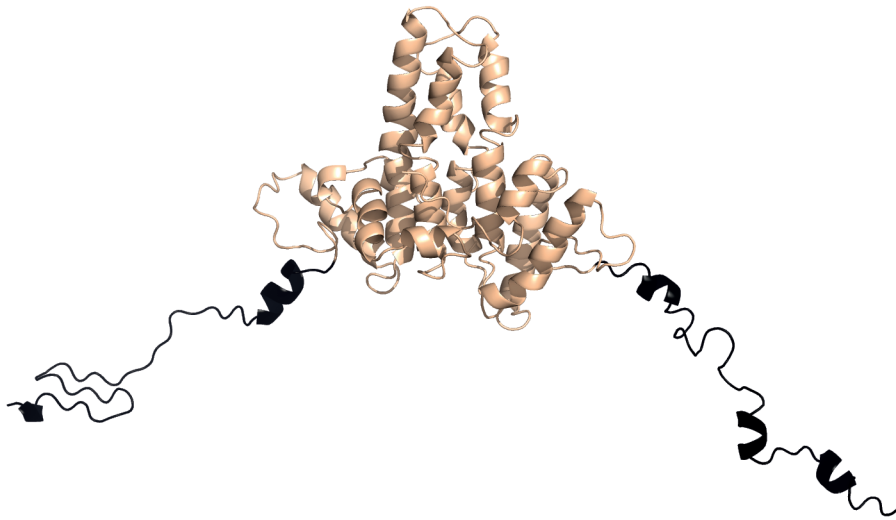


Figure 70 – MD avec échange de réplique (REMD) du modèle MODELLER sur UNRES [110]. Le NTD déformé est coloré en beige. Les bras C-terminaux, majoritairement dépliés, sont en noir. Les contacts d'un des 2 bras sont formés sur lui-même. Les bras se structurent partiellement en hélices.

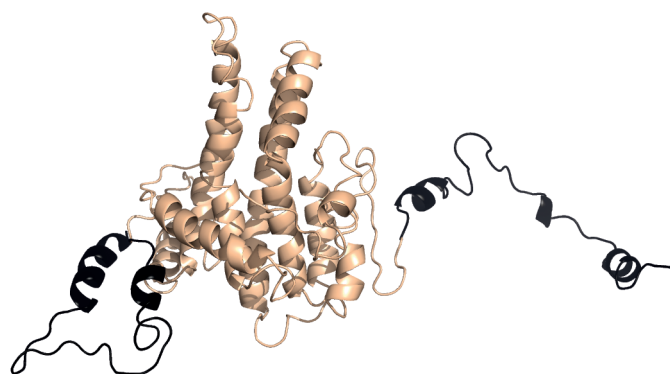


Figure 71 – MD avec échange de réplique (REMD) du modèle MODELLER sur UNRES [110]. Le NTD déformé est coloré en beige. Le CTD est en noir. L'un des deux bras C-terminaux est replié contre le NTD, l'autre est déplié. Les contacts d'un des deux bras sont formés sur lui-même. Les bras se structurent partiellement en hélices.

Au total, 12 simulations sont donc réalisées. Les simulations les plus intéressantes, en termes de repliement du CTD, sont la dynamique et la réplique du modèle MODELLER (Figure 72A). La visualisation de toutes les MD montre que les CTD se replient contre le NTD et

sur eux-mêmes. Pour mieux comprendre la dynamique des CTD, une carte électrostatique de la protéine Core est générée (Figure 72C). Les CTD, riches en arginines, sont chargés très positivement. Il y a des régions chargées très négativement sur le NTD : le sommet de sa spicule et sa partie inférieure (base des hélices α_4 et α_5). Cette carte explique les contacts qui se forment entre le CTD et le NTD (Figure 72).

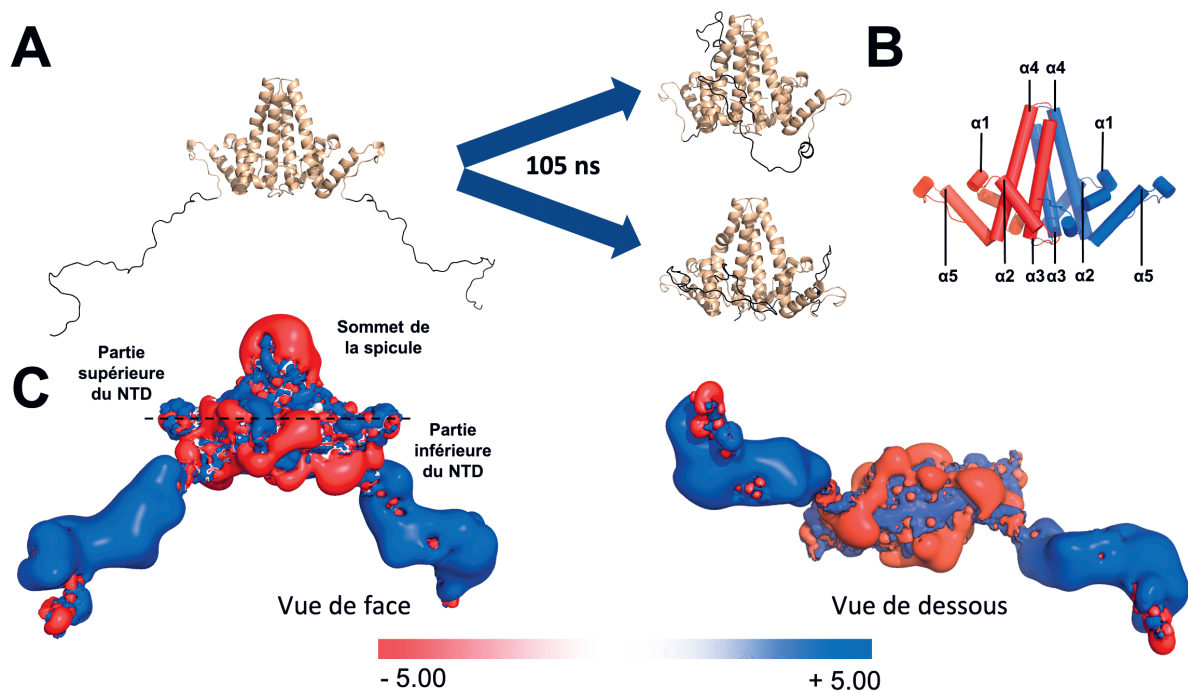


Figure 72 – Exemples de repliements du CTD après MD reliés aux propriétés électrostatiques de Core.
 A. Conformation initiale et conformations finales, après MD, du modèle MODELLER. Le NTD est en beige et le CTD est en noir. B. Localisation des hélices de α_1 à α_5 du NTD ; la chaîne A est en rouge et la chaîne B, en bleu. C. Carte électrostatique de la protéine Core ; les régions chargées négativement sont colorées en rouge et les régions chargées positivement sont colorées en bleu.

7.7.3.1 Contacts intra- et inter-domaines communs

Dans le but d'identifier les contacts communs entre toutes les MD, les interactions intra- et inter-domaines du CTD sont quantifiées, en analysant les contacts avec CPPTRAJ [111].

Cette analyse est effectuée sur les 25 dernières nanosecondes des 12 MD (Tableaux 3 et 4). On détermine les contacts de tous les résidus dans un rayon de 4,5 Å. Dans un contact, 2 résidus doivent être séparés par au moins un résidu. Les durées de tous les contacts, entre les résidus 143 à 183 et les résidus 1 à 183 (pour les 2 chaînes), sont calculées. Pour chaque paire de résidus impliqués dans un contact, on calcule la durée moyenne. Il s'agit d'une moyenne des durées d'un contact, entre une même paire de résidus, pour l'ensemble des simulations. Les contacts sont très nombreux (5547) et ne peuvent donc être tous pris en compte.

À partir de toutes les durées moyennes des contacts, un seuil égal à 3 % est établi (Figure 73). Considérer les contacts qui ont une durée moyenne supérieure à ce seuil est privilégié.

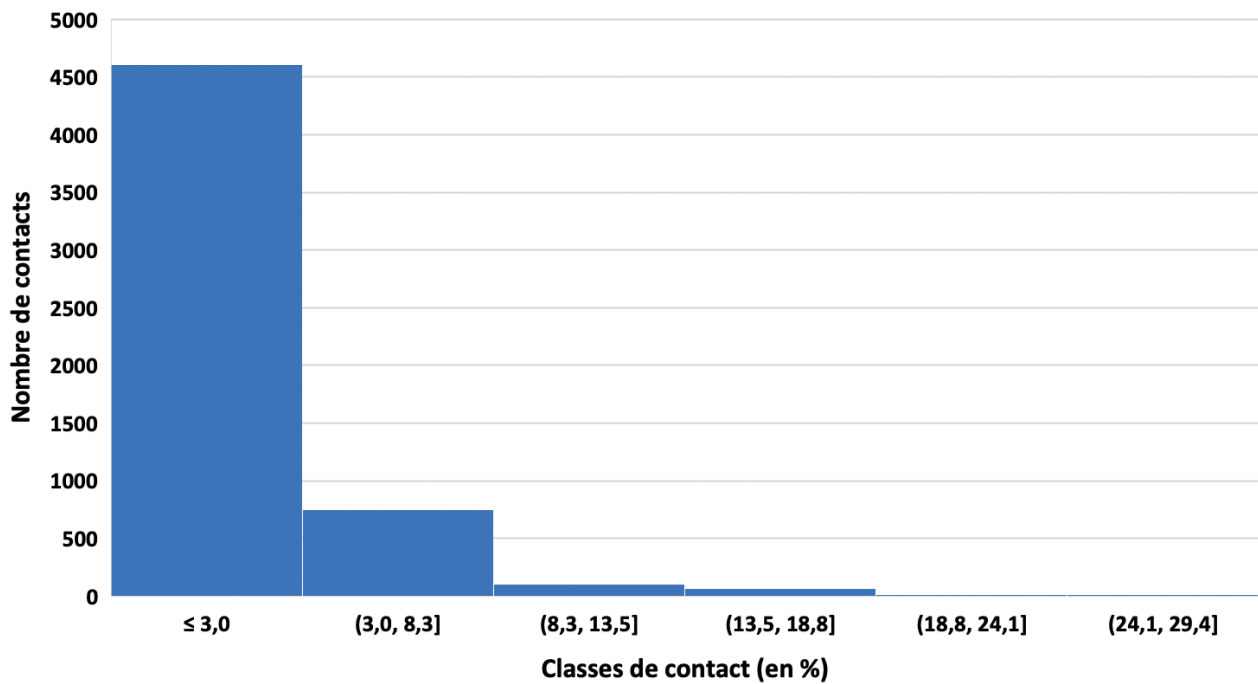


Figure 73 – Distribution des contacts intra- et inter-domaines.

Dans les tableaux 3 et 4, la somme des durées de contacts entre 2 régions (ou groupes d'acides aminés), sont calculés. Une région peut se définir comme (1) un domaine riche en arginines (ARD), (2) comme au voisinage d'un résidu ou (3), comme à un endroit précis (sommet de spicule, extrémité du CTD, hélices...). La durée de contact entre 2 groupes (Tableaux 3 et 4 : acides aminés 1 et 2) est additionnée pour toutes les MD dans lesquelles le contact s'effectue. Plus la somme est grande, plus les groupes d'acides aminés 1 et 2 se contactent.

Acides aminés 1 (CTD)	Acides aminés 2 (NTD)	Localisation	Somme des durées de contacts (en %)	Nombre de MD
164R (ARD III)	14E,17S	partie inférieure	11,33	2
158R,164R (ARD II ou III)	18F		11,73	2
165R (ARD III)	20P		4,65	3
152R (ARD I)	23F		4,84	5
173R,175R (ARD IV)	80A	sommets spicule	9,76	3
174R (ARD IV)	83D		4,50	3
171R,173R,174R (ARD IV)	84L		18,75	2
172R, 174R (ARD IV)	87S		10,74	3
174R (ARD IV)	88Y		4,99	2
182Q	91T		4,57	2
165R (ARD III)	92N		3,46	1
175R,183C (ARD IV)	112R		10,09	2
150R,154R (ARD I)	117E		10,91	4
152R (ARD I)	120V	6,16	2	

Tableau 3 – Contacts inter-domaines sur les 25 dernières ns des 12 MD. ARD : Domaine riche en arginines (cf. Figure 57). Dans la colonne “Localisation”, les régions correspondant à “partie inférieure” et “sommets spicule” sont localisées sur la Figure 72C.

L'analyse des contacts inter-domaines (Tableau 3) montre que :

- Les domaines C-terminaux interagissent sur le sommet de la spicule, pendant 19% des 25 ns au maximum, pour plus de 14% (3) des simulations.

Ces contacts correspondent exclusivement aux MD issues du modèle MODELLER et du premier modèle ROBETTA. Les CTD de ce modèle ROBETTA sont déjà en interaction avec le sommet de la spicule.

- Les domaines C-terminaux interagissent avec la partie inférieure du NTD, pendant 12% des 25 ns au maximum, pour plus de 13% (3) des simulations.

Ces contacts correspondent, en majorité, aux MD issues d'un modèle ROBETTA et des modèles I-TASSER. Les CTD de ces modèles sont initialement en interaction avec la partie inférieure du NTD.

L'analyse des contacts intra-domaine (Tableau 4) indique que le CTD se replie sur lui-même. Au maximum 56% (6) des simulations montrent que les 7 derniers acides aminés (TER) se replient sur des ARD (41% des 25 ns au plus). Pour 42% (5) des MD, au maximum, TER se replie autrement contre le CTD (plus de 11% des 25 ns).

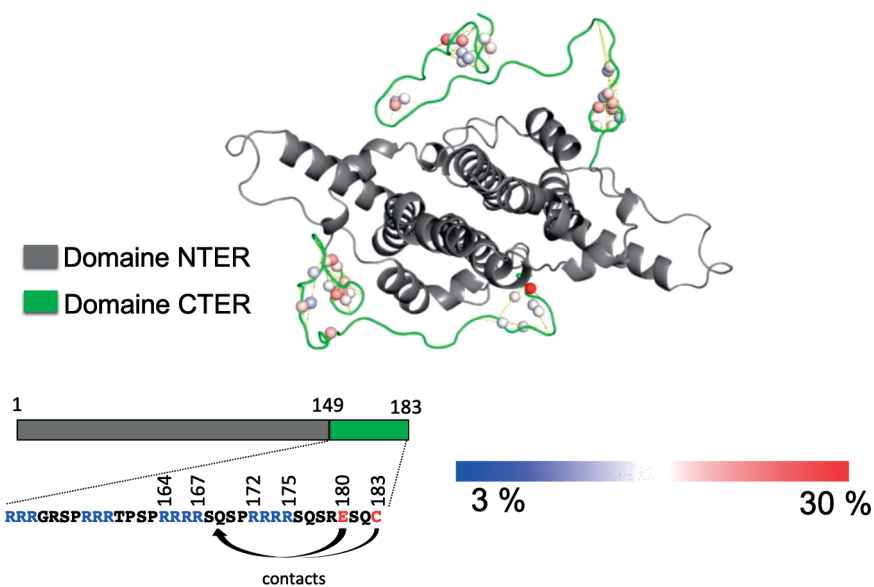
Acides aminés 1 (CTD)	Acides aminés 2 (CTD)	Localisation	Somme des durées de contacts (en %)	Nombre de MD
158R,159R (ARD II)	182Q	TER	12,79	3
165R (ARD III)	172R	ARD IV	9,00	2
165R (ARD III)	182Q	TER	10,04	3
167R (ARD III)	182Q,183C		16,62	5
169Q	179R		5,66	4
169Q	181S,182Q		6,88	5
169Q-171P	182Q		25,13	6
172R (ARD IV)	177Q,179R,180E		40,96	7
172R (ARD IV)	183C		6,10	7
174R (ARD IV)	180E		16,10	6
175R (ARD IV)	182Q		6,67	4
177Q	182Q		6,51	8
178S	181S		5,71	9
179R	183C		10,65	5

Tableau 4 –Contacts intra-domaine du CTD sur les 25 dernières ns des 12 MD. ARD : Domaine riche en arginines (cf. Figure 57).

En résumé, les figures 74 et 75 montrent que la structuration du CTD sur lui-même est due à l'interaction de C183 (cystéine terminale) et de E180, avec l'une des régions riches en arginines. Pour au moins 6 simulations, une boucle se forme entre ces résidus chargés négativement et l'une des régions riches en arginines. L'interaction du CTD avec le NTD est représentée sur cette même figure. Les interactions se font majoritairement entre les résidus chargés négativement du NTD et les résidus chargés positivement du CTD. Les contacts peuvent se faire sur la spicule ou sur la partie inférieure du NTD.

Contacts internes au domaine C-terminal

Vue de dessus



Ponts-salins essentiels et communs

Vue de face

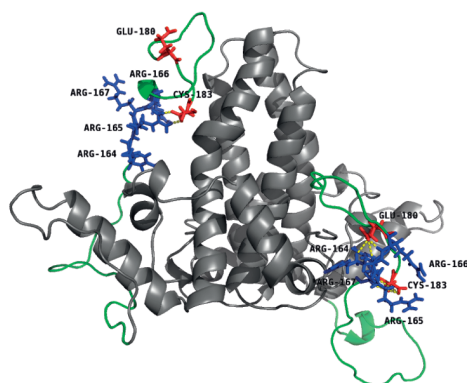
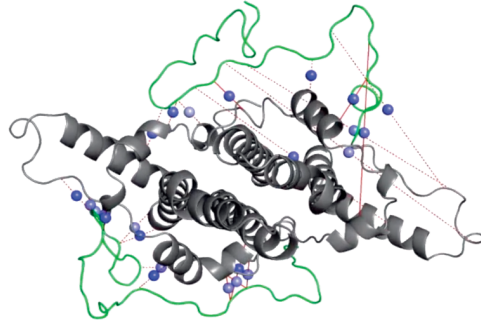


Figure 74 – Résumé de la structuration du CTD – contacts intra-domaine. Une bille correspond à un centre de masse entre deux résidus qui se contactent. La fréquence du contact est représentée par le gradient du bleu au rouge. Les structures utilisées sont issues des simulations des modèles MODELLER.

Contacts avec la partie inférieure de NTD

Vue de dessus



■ Domaine NTER

■ Domaine CTER

3 %

30 %

Contacts sur le haut de la spicule

Vue de face

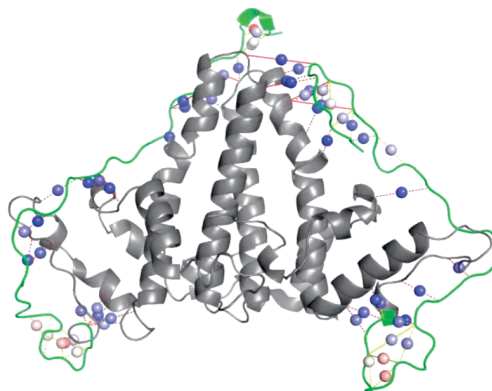


Figure 75 – Résumé de la structuration du CTD – contacts inter-domaines. Une bille correspond à un centre de masse entre deux résidus qui se contactent. La fréquence du contact est représentée par le gradient du bleu au rouge. Les structures utilisées sont issues des simulations des modèles MODELLER (en haut) et ROBETTA (en bas).

De manière générale, les contacts qui se forment entre les 2 domaines sont moins fréquents (de 3 à 19% pour ~2 des simulations) que les contacts induisant le repliement du CTD sur lui-même (de 15 à 30% de la moitié les MD).

7.7.4 Influence des champs de force sur le repliement du CTD

Un champ de force est un ensemble de forces (répulsions et attractions) qui s'exercent sur chaque atome qui compose une interaction. Il est caractérisé par un ensemble de potentiels, de variables et de paramètres empiriques qui décrivent l'énergie potentielle d'un système d'atomes ou de particules à gros-grains en mécanique moléculaire [18] (p. 165). Il existe un large panel de champs de force (MARTINI, Amber, Charmm, ...) utilisés dans le cadre de la MD. Au cours de l'étude du repliement du CTD dans le contexte du dimère libre, la question de l'influence du champ de force sur le repliement se pose. Core est constituée d'un domaine structuré (NTD) et d'un domaine non structuré (CTD). Le champ de force Amber 99SBILDN [109], qui a été utilisé jusqu'à maintenant, n'a pas été développé dans la perspective d'étudier la dynamique des régions intrinsèquement désordonnées (IDRs). On se demande alors si la simulation du CTD changerait sa dynamique avec un champ de force plus adapté. Le champ de force Amber 99SB-DISP [112], développé par Robustelli et al. en 2018, a semblé idéal pour tester le repliement du CTD. En effet, ce champ de force améliore considérablement l'état de l'art et la description de la dynamique des protéines possédant des IDRs, sans sacrifier la structuration des parties repliées. Les champs de force ont un rôle non négligeable dans le repliement des parties non structurées et dans le maintien de celles qui sont structurées [112]. Les modèles d'eaux jouent également un rôle sur le repliement des protéines [113]. Robustelli et al., adaptent le champ de force classique (Amber 99SB-ILDN) [109] combiné au modèle d'eau TIP4P-D, en optimisant les paramètres de torsion et en introduisant de petits changements dans les termes d'interaction de van der Waals de la protéine et de l'eau. Ils modifient également les charges partielles des atomes des résidus chargés. Les fichiers de paramètres que j'ai généré à partir des données du champ de force Amber 99SB-DISP et du modèle d'eau TIP4P-D sont disponibles au lien suivant : <https://github.com/jecarvaill/Amber-99SB-DISP>

Pour savoir si Amber 99SB-DISP modifie la dynamique du CTD (IDR), elle est testée sur 2 systèmes avec, comme référence, le champ de force Amber 99SB-ILDN. Les 2 modèles utilisés sont issus de MODELLER et ROBETTA (CTD en interaction avec le sommet de la spicule du NTD). Toutes les simulations sont réalisées dans les mêmes conditions (même graine aléatoire, même vitesse de départ, même boîte d'eau...). De cette manière, seul le champ de force influence le repliement du CTD. Les systèmes comportent la même boîte d'eau TIP4P-D [113]. Ce modèle donne de meilleurs résultats sur les IDR [112,113]. L'ensemble des simulations est effectué sur le supercalculateur OCCIGEN, avec le même type de processeur et le même nombre de nœuds (140 processeurs). Les systèmes, une fois neutralisés et équilibrés, sont simulés durant une période de 105 ns. Cela correspond à ~140 000 h consommées pour simuler 105 ns sur des processeurs Intel(R) Xeon(R) CPU E5-2690 v4 de 2,60GHz par système (ex. MODELLER : 464 626 atomes).

Afin de comparer l'effet du champ de force sur le repliement du CTD de 4 systèmes se basant sur deux modèles (Figure 76), les RMSD du NTD, du CTD et de Core sont calculés (Figure 77). Ces systèmes sont :

- Le modèle MODELLER, simulé selon champ de force Amber 99SB-ILDN.
- Le modèle MODELLER, simulé selon champ de force Amber 99SB-DISP.
- Le modèle ROBETTA, simulé selon champ de force Amber 99SB-ILDN.
- Le modèle ROBETTA, simulé selon champ de force Amber 99SB-DISP.

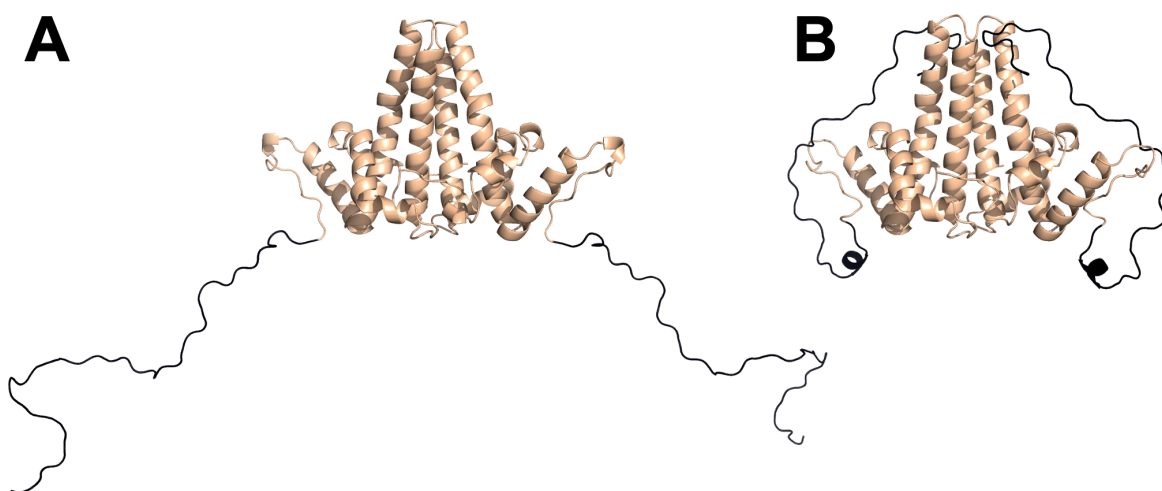


Figure 76 – Modèles de Core utilisés. A. Modèle MODELLER avec les CTD dépliés. B. Modèle ROBETTA avec les CTD repliés contre le sommet de la spicule.

L'évolution des structures secondaires de Core, sur les 35 dernières nanosecondes, est suivie (Figure 78). Enfin, une comparaison des contacts qui s'établissent au cours des 25 dernières nanosecondes, entre les champs de force Amber 99SB-ILDN et -DISP est réalisée (Tableaux 5, 6, 7 et 8).

L'analyse des RMSD montre que les champs de force ne modifient pas la dynamique du NTD. Les RMSD du CTD indiquent que leurs dynamiques changent. Pour le champ de force Amber 99SB-ILDN, le modèle MODELLER atteint un plateau à 80 ns à ~ 70 Å (Figure 77A). Le modèle ROBETTA atteint un premier plateau à 30 ns (~ 18 Å), puis, diverge de nouveau jusqu'à ~ 34 Å (Figure 77B). La conformation des CTD du modèle MODELLER, selon le champ de force Amber 99SB-DISP, atteint un plateau au bout de 35 ns (~ 50 Å) et diverge de nouveau à 75 ns (Figure 77C). Le CTD du modèle ROBETTA diverge tout au long de la simulation (~ 30 Å en fin de MD) (Figure 77D). L'évolution de Core est similaire à celle du CTD mais le RMSD est plus faible. Avec le champ de force Amber 99SB-DISP, les CTD divergent moins (MODELLER) ou autant (ROBETTA). Quoi qu'il en soit, les CTD issus du même modèle de départ n'évoluent pas de la même manière au cours des simulations,

en fonction du champ de force utilisé. Les champs de force Amber 99-SB-DISP et -ILDN semblent avoir un effet différent sur le repliement du CTD. Pour s'en assurer, il serait pertinent de réaliser des répliques.

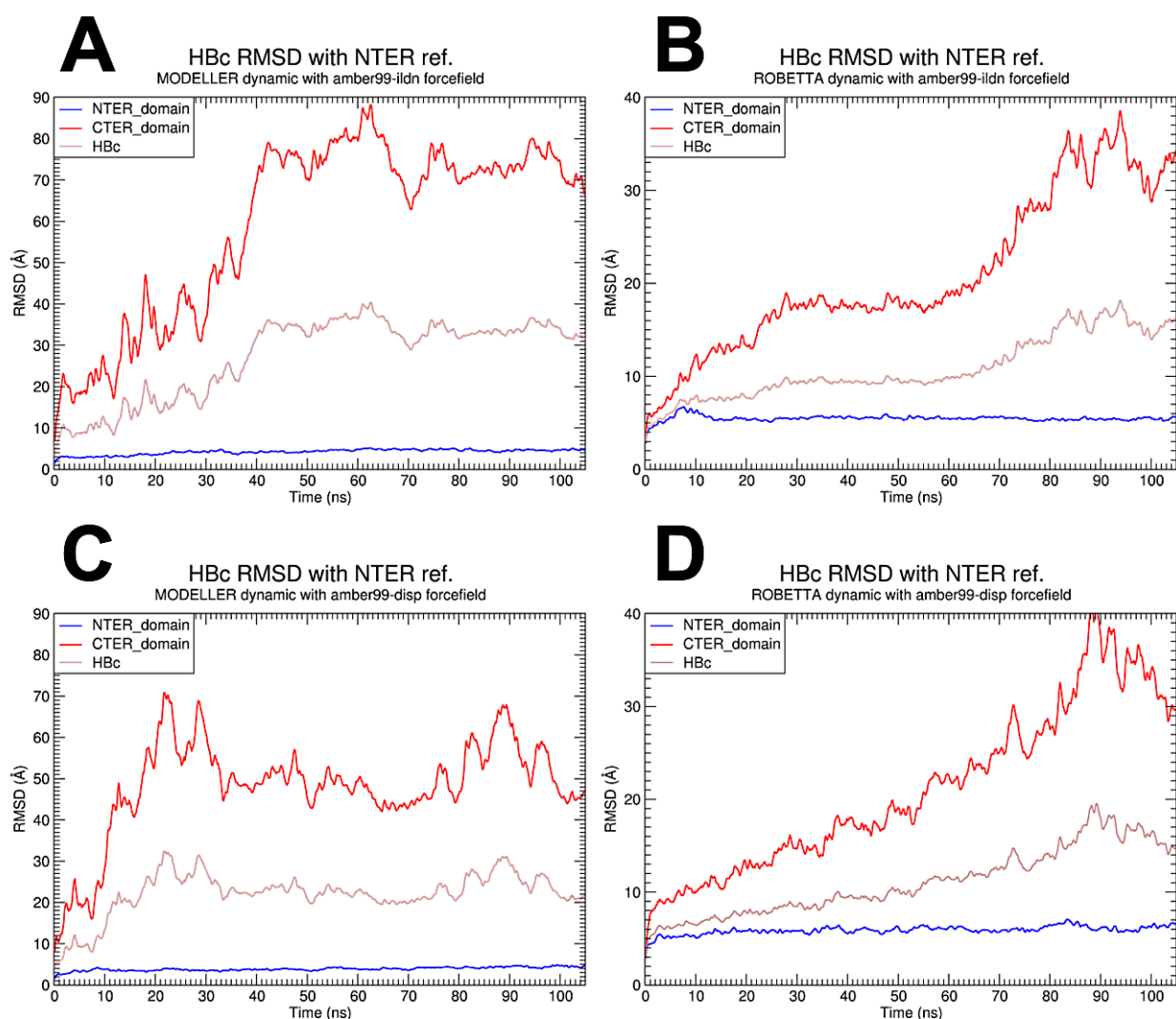


Figure 77 – Comparaison des RMSD de 2 systèmes simulés selon les champs de force Amber 99SB-ILDN [109] et DISP [112]. RMSD du NTD en bleu, RMSD du CTD en rouge et RMSD de Core, en marron. A. RMSD du modèle MODELLER simulé selon le champ de force Amber 99SB-ILDN. B. RMSD du modèle ROBETTA où les CTD sont en interaction avec le sommet de la spicule du NTD, simulé selon le champ de force Amber 99SB-ILDN. C. RMSD du modèle MODELLER, simulé selon le champ de force Amber 99SB-DISP. D. RMSD du modèle ROBETTA, simulé selon le champ de force Amber 99SB-DISP.

Les structures secondaires des 2 modèles sont tracées pendant les 35 dernières nanosecondes et en fonction du champ de force utilisé. Les 2 champs de force ont presque le même effet sur les parties structurées (NTD - résidus 1 à 128 (premier monomère) et 184 à 325 (deuxième monomère)). La structure secondaire de type hélice α est assignée aux acides aminés qui constituent le NTD. Le CTD, selon les 2 champs de force, a une structure secondaire de type : boucle, tour ou non assignée.

Il y a tout de même des différences sur la structuration secondaire du CTD entre les deux champs de force. Premièrement, les résidus terminaux de la chaîne B (résidus 360 à 366) adoptent transitoirement une structuration en hélices α ou 3-10, pour les modèles simulés selon le champ de force Amber 99SB-DISP (Figure 78C et D). Deuxièmement, d'un modèle à l'autre et d'un champ de force à l'autre, les résidus 125 à 183 (CTD - chaîne A) n'ont pas la même structuration secondaire. Le CTD de la chaîne A du modèle MODELLER a une structure secondaire de type boucle ou tour, à certains moments, avec le champ de force Amber 99SB-DISP (Figure 78C). Avec l'autre champ de force, les résidus ont très peu de structure secondaire assignées (Figure 78A). L'inverse se produit pour le modèle ROBETTA. Le CTD de la chaîne A est plus structuré avec le champ de force Amber99SB-ILDN que -DISP. Les champs de force n'induisent pas la même structuration des CTD. De manière générale, le champ de force Amber 99SB-DISP induit davantage de structuration sur la partie Cterminale. Ce champ de force fait apparaître des structures secondaires de type hélice α et 3-10 sur les CTD.

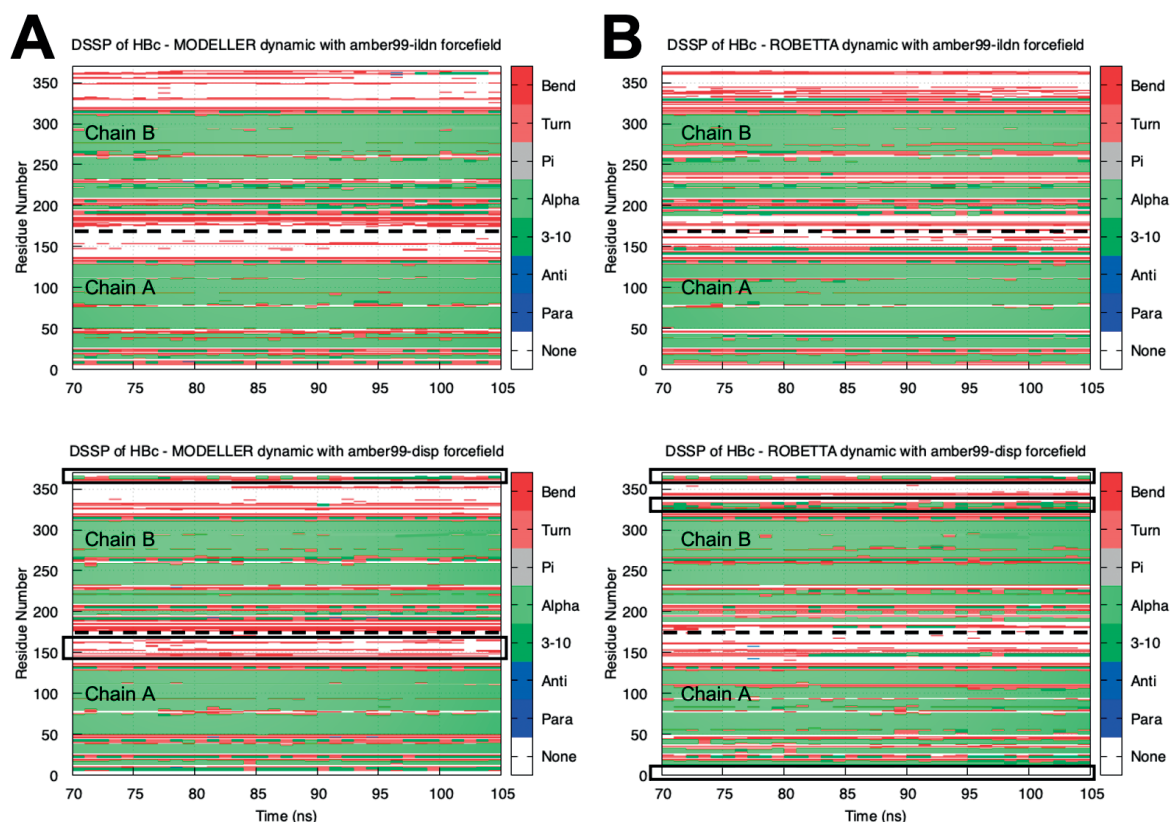


Figure 78 – Comparaison de l'évolution des structures secondaires de 2 systèmes simulés selon les champs de force Amber 99SB-ILDN [109] et -DISP [112]. A. Évolution des structures secondaires du modèle MODELLER, simulé selon le champ de force Amber 99SB-ILDN. B. Évolution des structures secondaires du modèle ROBETTA, simulé selon le champ de force Amber 99SB-ILDN. C. Évolution des structures secondaires du modèle MODELLER, simulé selon le champ de force Amber 99SB-DISP. D. Évolution des structures secondaires du modèle ROBETTA, simulé selon le champ de force Amber 99SB-DISP. Les chaînes A et B sont délimitées par une ligne horizontale en pointillé.

Les durées des contacts inter-domaines établis avant simulation sont calculées au cours des 25 dernières nanosecondes avec CPPTRAJ [111] (Tableau 5). Les résidus qui forment des contacts sont regroupés suivant leur localisation dans Core. La moyenne des durées de contacts est calculée pour chaque paire d'acides aminés 1 et 2 (Tableaux 5 à 8). Les contacts identifiés sur tous les systèmes sont comparés d'un champ de force à l'autre. Les contacts s'établissent le plus souvent entre les domaines riches en arginine (ARD) et le sommet de la spicule du NTD ou alors, au-dessous du NTD (zone qui tapisse l'intérieur de la capsid). Pour les 2 champs de force, les moyennes totales et les écarts-types sont très proches. Pour le champ de force Amber 99SB-DISP, il y a toutefois légèrement plus de contacts entre le CTD (et ARD) et le NTD.

Acides aminés 1 (CTD) et localisation	Acides aminés 2 (NTD) et localisation	Amber 99SB-DISP Moyenne des durées de contacts (en %)	Amber 99SB-ILDN Moyenne des durées de contacts (en %)
Contacts natifs			
149V (CTD)	40E,41A,46E,47H (dessous NTD)	0	24,4
152R (ARD I)	40E (dessous NTD)	0	21,26
156P,158R (ARD II)	40E (dessous NTD)	24,12	0
173R-175R (ARD IV)	80A-84L,87S (sommet spicule)	31,27	0,23
176S,177Q (CTD)	84L (sommet spicule)	15,85	0
Moyenne totale		~14,25	~9,18
Écart-type		~14,10	~12,51
Nombre de contacts nuls ou quasi-nuls		2	3

Tableau 5 – Contacts natifs inter-domaines sur les 25 dernières ns des 4 MD. ARD : Domaine riche en arginines (cf. Figure 57). Dans la colonne "Acide aminés 2 (NTD)", la case "dessous NTD" correspond à la partie sous le NTD et qui tapisse l'intérieur de la capsid et la case "sommet spicule" au sommet de Core. La moyenne des fréquences de contacts est entre parenthèses dans les colonnes "Fréquences de contacts obtenues avec Amber 99SB-DISP (%)" ou "-ILDN (%)".

Acides aminés 1 (CTD) et localisation	Acides aminés 2 (NTD) et localisation	Amber 99SBDISP Moyenne des durées de contacts (en %)	Amber 99SBILDN Moyenne des durées de contacts (en %)
Contacts non natifs			
149V (CTD)	40E (partie inférieure)	0	16,48
150R (ARD I)	46E (partie inférieure)	0	28,45
151R (ARD I)	113E,116I,117E (dessous NTD α 5)	0,13	24,94
154R,155S (CTD)	46E,48C (partie inférieure)	0	19,46
158R (ARD II)	40E,44S-46E (partie inférieure)	2,99	21,96
166R (ARD III)	33T,36A,37L (partie inférieure)	0,03	16,82
168S,169Q,171 P (CTD)	110F,113E,114T,117E (dessous NTD α 5)	19,06	0
172R-174R (ARD IV)	79P,82R-84L,87S (sommets spicule)	17,61	0,11
172R,174R (ARD IV)	109T,110F,111G (dessous NTD α 5)	29,23	0
175R (ARD IV)	80A (sommets spicule)	19,39	0
177Q (TER)	76L-78D,81S (sommets spicule)	19,46	0
179R,181S (TER)	40E (dessous NTD)	13,74	0
179R,181S,182Q (TER)	38Y,40E,41A,52H (dessous NTD)	15,84	0
182Q (TER)	108L (dessous NTD α 5)	18,73	0
182Q (TER)	126I (sommets α 5)	0	13,69
183C (TER)	41A (dessous NTD)	13,21	0
Moyenne totale		~9,76	~8,87
Écart-type		~9,87	~10,88
Nombre de contacts nuls ou quasi-nuls		6	9

Tableau 6 – Contacts natifs inter-domaines sur les 25 dernières ns des 4 MD. ARD : Domaine riche en arginines (cf. Figure 57). TER : Région C-terminale du résidu 179 à 183. Dans la colonne "Acide aminés 2", la case "dessous NTD" correspond à la partie sous le NTD ; la case " α 5" faisant partie de l'hélice α 5. La case "sommets spicule" correspond au sommet de Core. La case "partie inférieure" correspond à la partie inférieure du NTD. La case "sommets α 5" se situe au sommet de l'hélice α 5.

Les contacts non natifs inter-domaines sont quantifiés puis comparés, de façon identique, entre les 2 champs de force (Tableau 6). Les contacts s'établissent entre les ARD (ou C183) et le dessous du NTD. Ils s'effectuent également avec le sommet de la spicule. Les moyennes totales et les écarts-types sont de nouveau très proches. Le champ de force Amber 99SB-DISP comptabilise tout de même plus de contacts. Pour appuyer cette observation, il serait pertinent de réaliser des répliques.

Acides aminés 1 (CTD) et localisation	Acides aminés 2 (CTD) et localisation	Amber 99SBDISP Moyenne des durées de contacts (en %)	Amber 99SBILDN Moyenne des durées de contacts (en %)
Contacts natifs			
158R (ARD II)	161P (CTD)	17,27	0
165R (ARD III)	182Q (TER)	0	12,12
168S (CTD)	182Q (TER)	0	14,55
173R (ARD IV)	177Q,178S,181S (TER)	22,69	0
174R (ARD IV)	177Q,178S,179R, 180E, 183C (TER)	0	37,15
175R-177Q (ARD IV et CTD)	180E-183C (TER)	35,30	0,41
179R (CTD)	176S (CTD)	0	29,29
179R (CTD)	182Q (TER)	29,75	31,62
Moyenne totale		~13,13	~15,64
Écart-type		~14,96	~15,30
Nombre de contacts nuls ou quasi-nuls		4	3

Tableau 7 – Contacts natifs inter-domaines sur les 25 dernières ns des 4 MD. ARD : Domaine riche en arginines (cf. Figure 57). Dans la colonne “Acide aminés 2 (NTD)”, la case “CTD” correspond aux acides aminés du CTD, qui ne sont pas des ARD. La case “TER” correspond à la région terminale du CTD.

La quantification des contacts intra-domaine, déjà établis avant simulation, est réalisée. Elle est suivie par la comparaison de ces contacts, d’un champ de force à l’autre (Tableau 7). Ils se forment, la plupart du temps, entre les ARD et la TER. Les moyennes totales et les écarts-types sont, une fois de plus, très proches. La moyenne et l’écart-type des contacts sont légèrement plus élevés avec le champ de force Amber 99SB-ILDN.

Les types de contacts sont identiques lorsque l’on quantifie les contacts intra-domaine non natifs (Tableau 8). Les moyennes et écarts-types sont très proches entre les 2 champs de force. Elles sont, une fois de plus, plus élevées pour le champ de force Amber 99SB-DISP.

Acides aminés 1 (CTD) et localisation	Acides aminés 2 (CTD) et localisation	Amber 99SBDISP Moyenne des durées de contacts (en %)	Amber 99SBILDN Moyenne des durées de contacts (en %)
Contacts non natifs			
155S,157R (CTD et ARD II)	160T (CTD)	0,05	63,26
158R (ARD II)	161P (CTD)	24,49	0
163P (CTD)	166R (ARD III)	0,49	69,13
167R (ARD III)	170S (CTD)	16,50	0
169Q (CTD)	182Q,183C (TER)	0	15,00
169Q (CTD)	183C (TER)	15,27	0,14
170S (CTD)	174R (ARD IV)	22,84	0,14
170S (CTD)	180E (TER)	16,79	0,03
171P (CTD)	174R (ARD IV)	93,38	0,79
171P (CTD)	182Q (TER)	0	15,61
172R (ARD IV)	175R (ARD IV)	0,10	45,39
172R (ARD IV)	180E (TER)	20,93	0,26
172R (ARD IV)	182Q (TER)	0,04	12,38
173R (ARD IV)	177Q,179R-183C (TER)	10,25	23,55
174R (ARD IV)	170S,171P (TER)	58,11	0,47
174R (ARD IV)	178S-181S (TER)	0	34,03
174R,175R (ARD IV)	179R-181S,183C (TER)	0,28	39,58
175R (ARD IV)	182Q (TER)	31,12	0
175R (ARD IV)	183C (TER)	0,56	22,47
176S (CTD)	182Q (TER)	33,89	0
176S,177Q (CTD)	181S,182Q,183C (TER)	24,45	0,03
177Q (CTD)	183C (TER)	0	20,22
179R,180E (CTD)	182Q,183C (TER)	38,36	0,02
Moyenne totale		~17,73	~15,76
Écart-type		~22,86	~21,26
Nombre de contacts nuls ou quasi-nuls		10	12

Tableau 8 – Contacts natifs inter-domaines sur les 25 dernières ns des 4 MD. ARD : Domaine riche en arginines (cf. Figure 57). Dans la colonne "Acide aminés 2 (NTD)", la case "CTD" correspond aux acides aminés du CTD qui ne sont pas des ARD. La case "TER" correspond à la région terminale du CTD.

En conclusion, pour les 2 champs de force, les types de contacts les plus récurrents sont semblables. On peut les classer de la manière suivante :

- Contacts inter-domaines
 - Les régions riches en arginines, interagissent avec le dessous du NTD.
 - Les ARD interagissent avec le sommet de la spicule.
 - Les ARD interagissent avec la région C-terminale (TER).

- Contacts intra-domaine
 - Les ARD interagissent avec TER.

Les CTD ne se structurent pas davantage, cependant ils forment plus de contacts au cours des simulations calculées avec le champ de force Amber 99SB-DISP. C'est notamment le cas de la partie C-terminale de la chaîne B, qui se structure transitoirement en hélices α . Les contacts sont, en moyenne, 15% plus fréquents qu'avec le champ de force Amber 99SB-ILDN. L'étalement des valeurs des durées de contacts est également plus élevé de 3%, en moyenne.

7.7.5 Dynamique du CTD dans le contexte de la capsid

Dans cette section, l'exposition du CTD sera étudiée. En effet, le CTD comporte des signaux de localisation nucléaire d'importation (NLS) ou d'exportation (NES) (Figure 57). Les NLS contribuent au transport du matériel génétique viral vers le noyau. C'est donc une étape du cycle viral importante.

7.7.5.1 Compatibilité de conformations du CTD avec son exposition par les pores

À partir du modèle MODELLER, une simulation et une réplique sont réalisées (Chapitre 7.7.3). Comme les bras C-terminaux sont dépliés au départ, ils constituent les meilleures pistes pour voir si des conformations du CTD sont compatibles avec son exposition par les pores. 3 des 4 bras, dans les 2 simulations MODELLER, se replient sur le côté et le dernier se replie sur la spicule. Pour replacer ces conformations de Cp183 dans le contexte de la capsid, les 2 structures en fin de MD (t = 105 ns) sont superposées sur l'un des dimères A-B ou C-D de la capsid du VHB du génotype D (PDB ID : 1QGT) [69] (Figure 79).

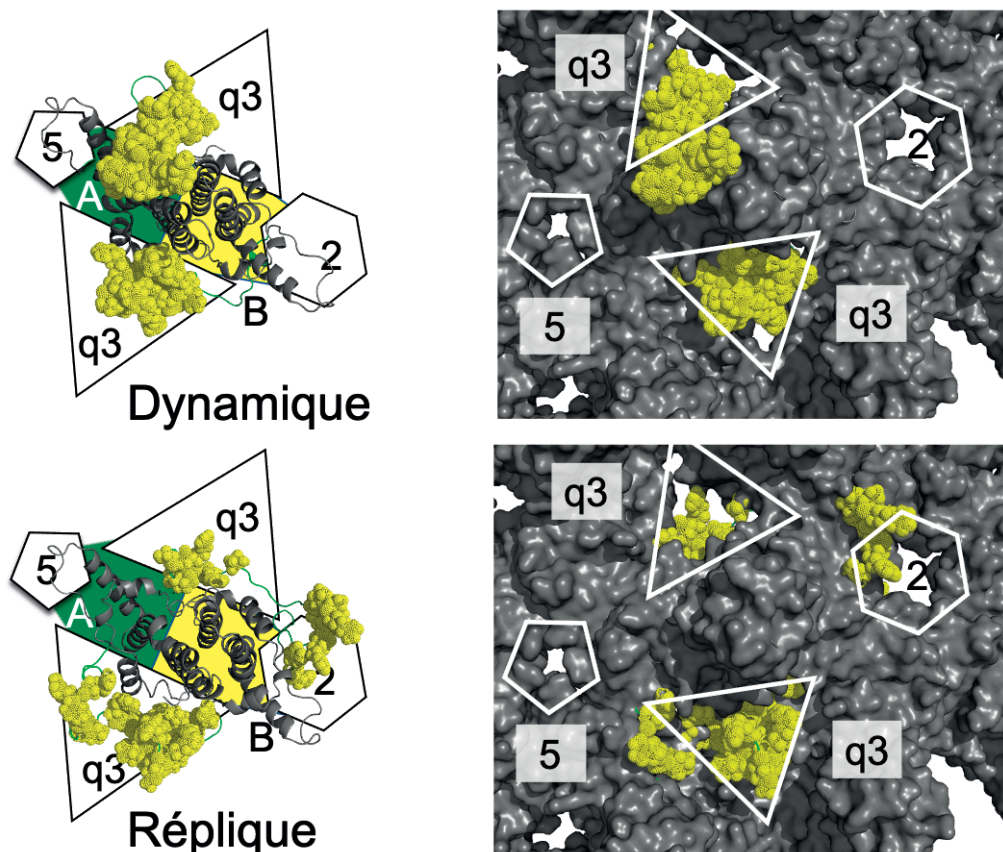


Figure 79 – Test de l'exposition des CTD par les pores à partir des dimères libres simulés. À gauche, les protéines Core en fin de simulation sont montrées (vues du dessus de la simulation et de la réplique). À droite, superposition des protéines Core simulées sur l'un des dimères A-B ou C-D de la capsid du VHB.

Les sphères de van der Waals sont modélisées (en jaune) sur les CTD pour identifier les clash stériques, entre les pores et eux (Figure 79). En se basant sur une analyse visuelle, les CTD clashent très peu avec les pores q3. Les CTD repliés contre le NTD (sur le côté ou en

contact avec le sommet de la spicule) paraissent compatibles avec son exposition (l'ARDIII et l'ARDIV sont exposés). La superposition des 2 structures, en fin de MD, sur l'un des dimères C-D de la capsid, positionne les CTD en dessous de pores 3. L'objectif est donc de chercher à caractériser l'exposition des bras C-terminaux en dehors de la capsid. Pour approfondir cette hypothèse, il serait intéressant d'étudier la dynamique de la capsid et des pores. Il serait également pertinent de réaliser des MD biaisées pour étudier l'exposition du CTD par les pores q3 et 3.

7.7.5.2 Dynamique de la capsid et des pores

La structure atomique de la capsid du VHB (PDB ID : 6HTX) [98], d'une résolution de 2,66 Å, est passée en gros grains MARTINI pour simuler la dynamique de la capsid (le protocole de simulation est similaire à celui dans la section 5.4.6). Cette structure est résolue des résidus 1 à 143 sur la chaîne A, des résidus 1 à 151 sur les chaînes B et D et des résidus 1 à 146, sur la chaîne C. S'il y a un doute sur la localisation des résidus, la localisation alternative A est sélectionnée. Dans un premier temps, la structure est passée à l'échelle gros grains, puis, simulée pendant 4 µs. Ce temps de simulation équivaut à une consommation de ~20 700 h sur des processeurs Intel(R) Xeon(R) CPU E5-2690 v4 de 2,60GHz. Le système solvaté a une concentration de 150 mM NaCl et comporte 598 142 grains. L'intérêt de cette simulation est de déterminer la flexibilité de la capsid et, plus spécifiquement, des 4 types de pores. Wynne et al. ont calculé les diamètres des pores : type 5 (le pore est presque fermé) ; type 2 (~12 Å) ; type 3 et q3 (~14 Å) [68] (Figure 80).

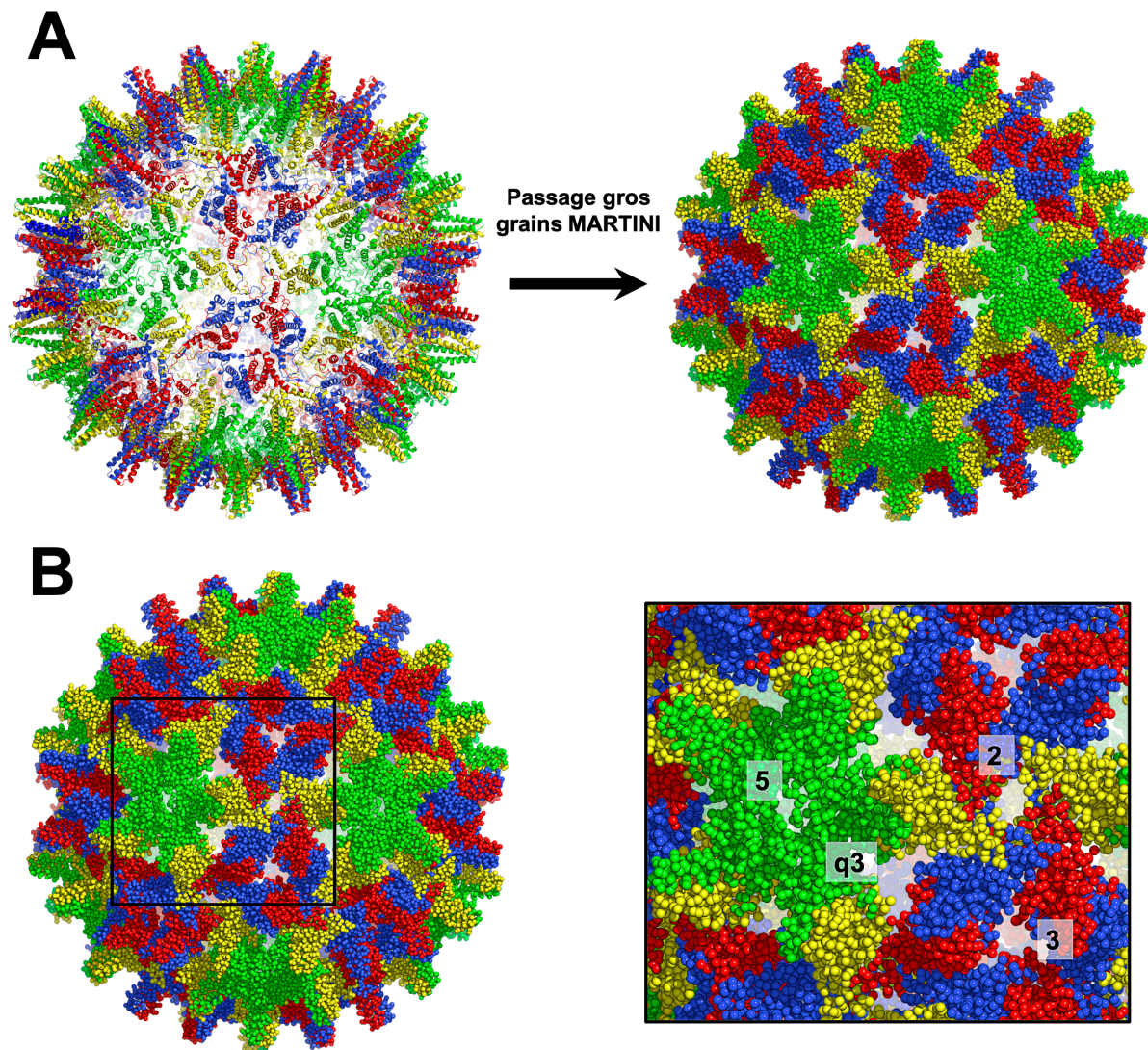


Figure 80 – Passage de la capsid du VHB tout-atomes à gros grains MARTINI. A. Passage gros grains MARTINI. B. Localisation des pores sur la capsid du VHB à l'échelle gros grains. Les chaînes A sont en vert, les chaînes B sont en jaune, les chaînes C sont en rouge et les chaînes D, en bleu.

Les rayons de giration autour des axes x, y et z, ainsi que le rayon total, sont calculés (Figure 81) avec GROMACS [44]. Les calculs du rayon de giration total et du rayon de giration autour d'un axe sont différents (<https://manual.gromacs.org/documentation/2019-rc1/reference-manual/analysis/radius-of-gyration.html>). Les rayons autour des axes de la capsid gros grains, qui est issue de la structure cristallographique (PDB ID : 6HTX) [98], sont de 11,90 nm au départ. On remarque que lors des premières 750 ns, les rayons de giration autour des axes diminuent brusquement, puis progressivement, tout au long de la simulation. L'axe z semble plus impacté que les autres. Le rayon de cet axe diminue d'environ 0,6 nm (Figure 81A). La capsid se comprime de façon asymétrique formant une sphère non régulière (ovoïde). Les rayons de giration totaux indiquent également que la capsid se comprime. Elle perd un peu plus de 3,6% (14,45 nm à 13,95 nm) de son rayon, au cours de la simulation.

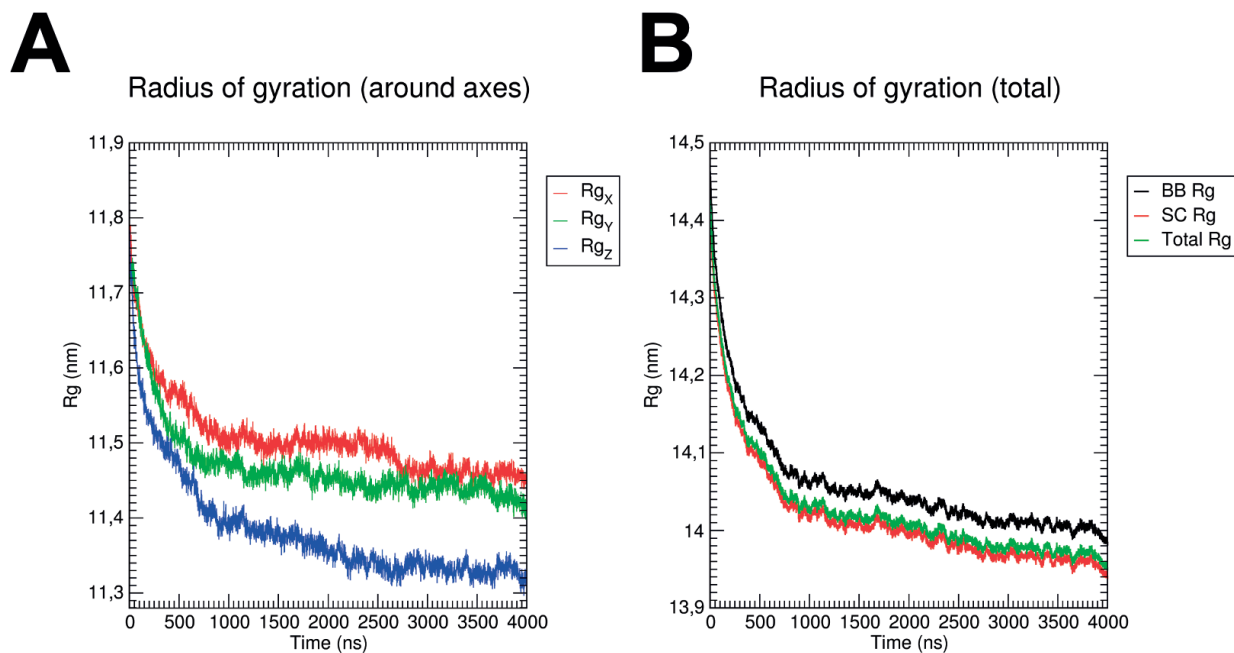


Figure 81 – Dynamique de la capside du VHB. A. Rayons de gyration sur les axes x , y et z au cours de la MD. B. Rayons de gyration totaux pour l'ensemble des grains (Total Rg), pour les grains des chaînes latérales (SC Rg) ou pour les grains du squelette carboné (BB Rg).

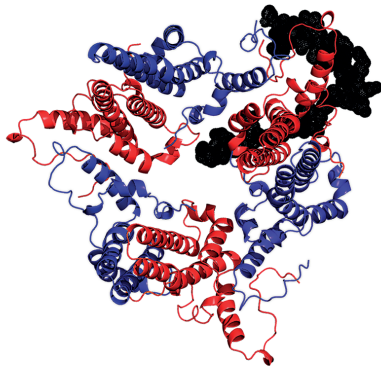
Une étude de la dynamique de la capside du VHB Cp149 (PDB ID : 2G33) [88], à l'échelle atomique, a déjà été menée [27]. Contrairement à ce que les auteurs mentionnent, les données de cette simulation ne sont pas accessibles au public. En conséquence, j'ai réalisé une simulation de 4 μ s de la capside du VHB à l'échelle gros grains est réalisée. Ces résultats ne confirment pas les observations d'Hadden et al. [27]. Pour eux, le volume de la capside augmente et la sphéricité reste constante.

Les analyses qui caractérisent la flexibilité des pores ne sont pas encore faites. Pour mettre en évidence la flexibilité des pores, une méthode reposant sur les rayons des pores est en cours de développement.

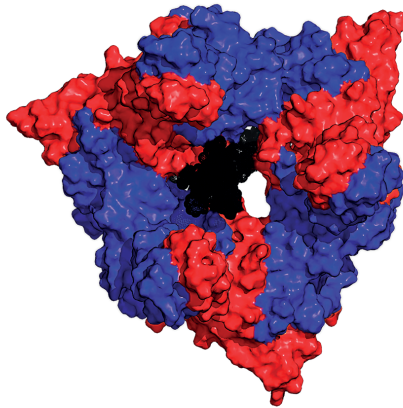
7.7.5.3 Dynamique de l'exposition du CTD par les pores de type 3 et q3

Les tests préliminaires montrent que les bras C-terminaux sont localisés au niveau de ces types de pore (Figure 79). Wynne et al. montrent, d'ailleurs, que ces pores sont les plus grands parmi les 4 [27,69]. Une expérience de cryo-microscopie électronique a révélé une densité au niveau du pore q3, qui correspondrait à un CTD [114]. Ces pores semblent être les points de passage les plus vraisemblables pour exposer les CTD [27,69,114]. Pour étudier l'exposition des CTD et des signaux d'importation nucléaire par les pores de type 3 et q3 (Figure 82), des simulations de dynamique moléculaire dirigée (TMD), de 400 ps, sont réalisées. Une de ces simulations consomme en moyenne 46h pour un système solvaté et neutralisé de 291 915 atomes (ex. pore q3).

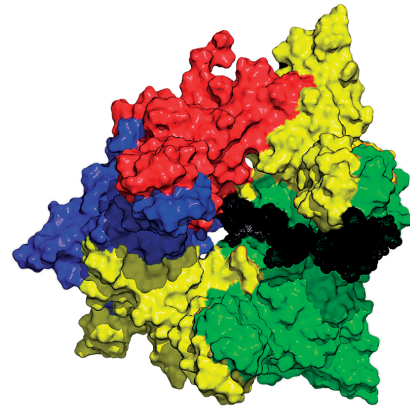
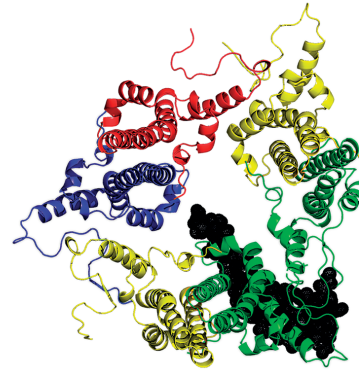
Initial - CTD sous les pores



Final - CTD exposé par les pores



Pore 3



Pore q3

Figure 82 – Conformations initiales et finales des pores et des CTD. Les chaînes A sont en jaune, les chaînes B en vert, les chaînes C en rouge et les chaînes D, en bleu. Le CTD est représenté en mode “dots” et en noir.

Des modèles du pore 3 (3 dimères C-D - Figure 82, à gauche) et q3 (2 dimères A-B et 1 dimère C-D - Figure 81, à droite), sont réalisés. Ils sont modélisés à partir d'une capsid (PDB ID : 6HTX) [98] et d'un bras C-terminal déplié, issu d'une structure des MD du modèle MODELLER (7.7.3 Dynamique du CTD dans le contexte du dimère seul). Les conformations initiales correspondent aux points de départ et les conformations finales aux points d'arrivée des TMD. Une TMD provoque le passage d'une structure de départ à une structure d'arrivée en appliquant une force sur les coordonnées des atomes. La force exercée sur les atomes est donnée par le potentiel :

$$E_{tmd_i} = \frac{1}{2} kN[RMSD_i - RMSD^*_i]^2$$

$RMSD^*_i$ correspond à la différence entre les coordonnées de l'atome i initiales et d'arrivées. N correspond au nombre d'atomes que l'on considère. k est la constante harmonique utilisée pour ajuster la force appliquée sur les atomes.

Plusieurs tests sont effectués pour savoir sur quels atomes il faut appliquer la force. Le passage du CTD, sans déformation du pore, est observé en appliquant la force sur tous les atomes du squelette carboné du pore et du CTD. D'une TMD à l'autre, la constante harmonique, k est ajustée pour augmenter ou diminuer la force appliquée sur l'ensemble du pore 3 (Figure 83), ou sur l'ensemble du pore q3 (Figure 89). Cette force provoque l'exposition du CTD au travers des pores.

7.7.5.3.1 TMD du Pore 3

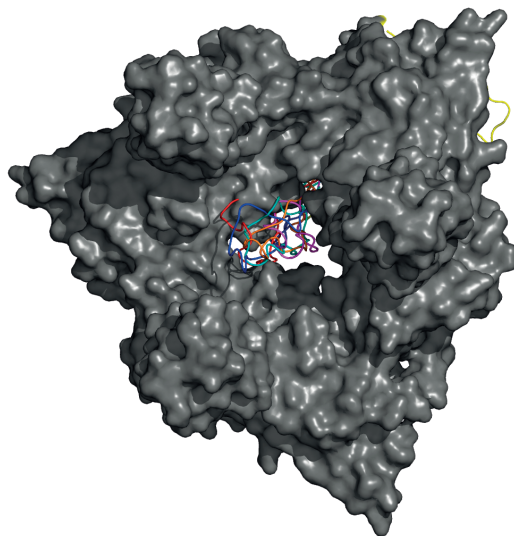


Figure 83 – Exposition du CTD par le pore 3. Conformations finales du CTD (en magenta, rouge, cyan, orange et bleu). Les couleurs sont corrélées avec les courbes sur la figure 84. Une couleur par constante harmonique est utilisée.

Le bras C-terminal d'un des dimères C-D de la structure cible du pore comporte une boucle. Cette conformation est utilisée pour tester l'hypothèse d'exposition d'Heger-Stevic et al. [82], dans laquelle une boucle (du CTD) est exposée (Figure 58, à droite).

Les TMD du pore 3 (et du CTD) révèlent que le CTD peut passer au travers de celui-ci (Figure 83). Les différentes constantes harmoniques utilisées n'ont pas le même effet sur l'exposition du CTD. Le calcul du RMSD, entre la conformation cible et une conformation au temps t (ps), montre que le CTD ne converge pas vers la conformation finale. Le bras est tout de même exposé. Les TMD, avec une constante harmonique $k = 0,06 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$; $k = 0,08 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ et $k = 0,10 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$, ne convergent pas toutes vers le même point (Figure 84).

RMSD evolution between target and initial conformation

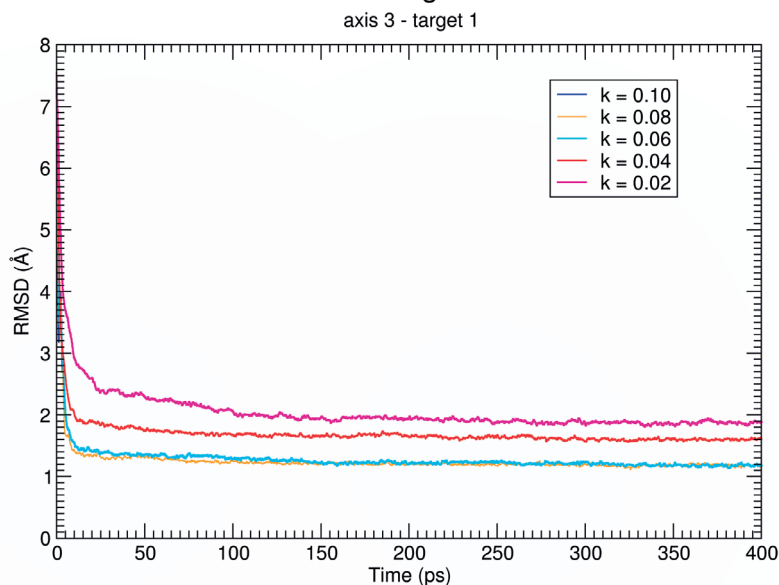


Figure 84 – Évolution du RMSD entre la conformation au temps t (ps) et la conformation cible du pore 3. k correspond à la constante harmonique utilisée dans la TMD.

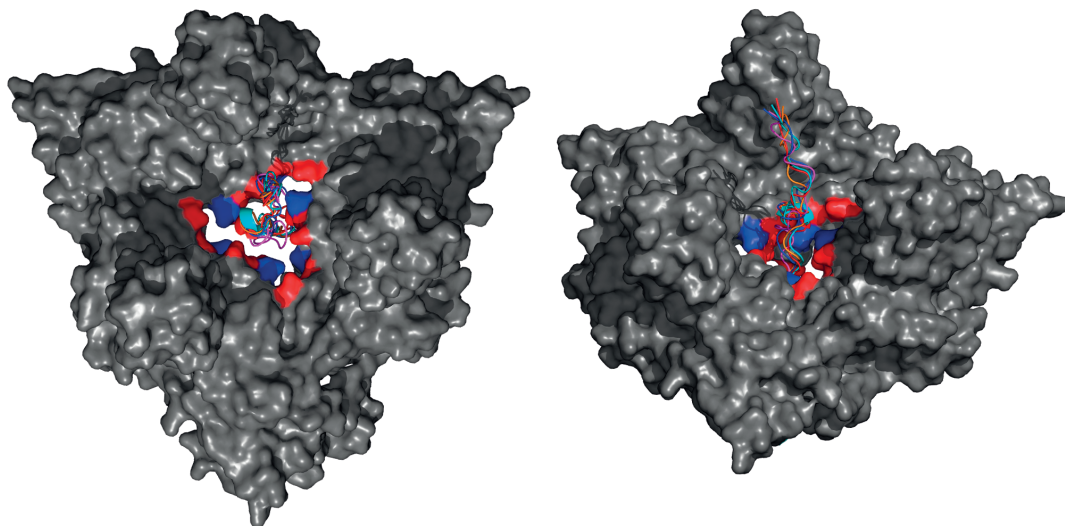
Pour mieux comprendre comment le CTD est exposé, tous les ponts-salins dans un rayon de 4 Å sont identifiés au cours de la TMD ($k = 0,01$ - Tableau 9). La conformation finale de la TMD ($k = 0,01$) est semblable à celle de la TMD ($k = 0,02$). Comme $k = 0,01 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ est la constante harmonique la plus faible, elle est idéale pour comprendre pourquoi le CTD aurait des difficultés à être exposé.

Au cours de la TMD, le CTD forme des ponts-salins avec les résidus chargés du pore 3. Les arginines (R167, R174 et R179) forment des ponts-salins avec les glutamates (E40 et E46).

Acide aminé (intérieur du pore)	Acide aminé (CTD)	Domaine riche en arginines	Durée de contact (en %)
E40	R174	ARDIV	38
E46	R167	ARDIII	11
E46	R179		87,75

Tableau 9 – Ponts-salins entre les résidus du pore 3 et le CTD, lors de son exposition. Les résultats sont issus de la TMD ($k = 0,01$).

Pour confirmer ces observations et pour voir jusqu'à quel résidu le CTD est exposé, des TMD du pore 3 et d'un fragment de CTD sont effectués. Le fragment peut être déplié (Figure 85) ou peut comporter une boucle (Figure 87). Ils seront nommés respectivement "CTD-allongé" et "CTD-replié" pour la suite. Les conformations initiales de ces 2 fragments sont situées en dessous du pore 3. Les conformations finales des TMD, pour les constantes harmoniques $k = 0,02 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$; $k = 0,04 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$; $k = 0,06 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$; $k = 0,08$ et $k = 0,10 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$, sont montrées sur les figures 85 et 87. Les RMSD sont calculés entre la conformation initiale et finale pour les mêmes constantes harmoniques. Elles sont montrées sur les figures 86 et 88.



Pore 3 – Vue de dessus

Pore 3 – Vue inclinée

Figure 85 – Exposition du CTD allongé par le pore 3. Conformations finales du CTD (en magenta, rouge, cyan, orange et bleu). Les couleurs sont corrélées avec les courbes sur la figure 86. Une couleur par constante harmonique est utilisée. Le pore est représenté en surface et les CTD, en cartoon. Les régions du pore, chargées négativement, sont colorées en rouge et les régions chargées positivement, en bleu.

Tous les “CTD-allongé” sont exposés au travers du pore 3 (Figure 85). Les RMSD indiquent que toutes les conformations initiales convergent presque totalement vers la conformation cible (~ 1 Å), sauf pour $k = 0,02$ kcal mol⁻¹ Å⁻² (Figure 86). Pour cette dernière, la conformation a un RMSD de ~ 2 Å.

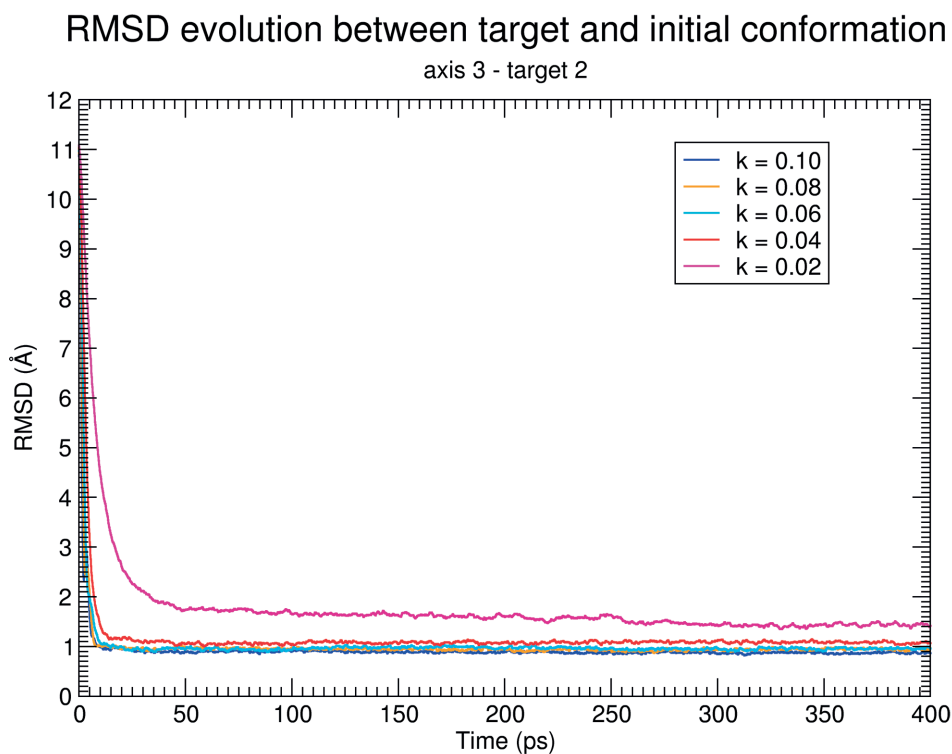


Figure 86 – Évolution du RMSD entre la conformation au temps t (ps) et la conformation cible du pore 3. Ces RMSD sont calculés sur les systèmes qui comportent un “CTD-allongé”. k correspond à la constante harmonique utilisée dans la TMD.

Des ponts-salins se forment également pour la TMD ($k = 0,01 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ - Tableau 10). Les arginines R165, R167 et R172 interagissent avec les glutamates E40 et E46.

Acide aminé (intérieur du pore)	Acide aminé (CTD)	Domaine riche en arginines	Durée de contact (en %)
E40	R167	ARDIII	12,25
E40	R172	ARDIV	4,5
E46	R165	ARDIII	9,25

Tableau 10 – Ponts-salins entre les résidus du pore 3 et le CTD allongé, lors de son exposition.

Les “CTD-replié” sont partiellement exposés au travers du pore 3 (Figure 87). Le “CTD-replié” n’est peut-être pas suffisamment exposé dans la conformation cible.

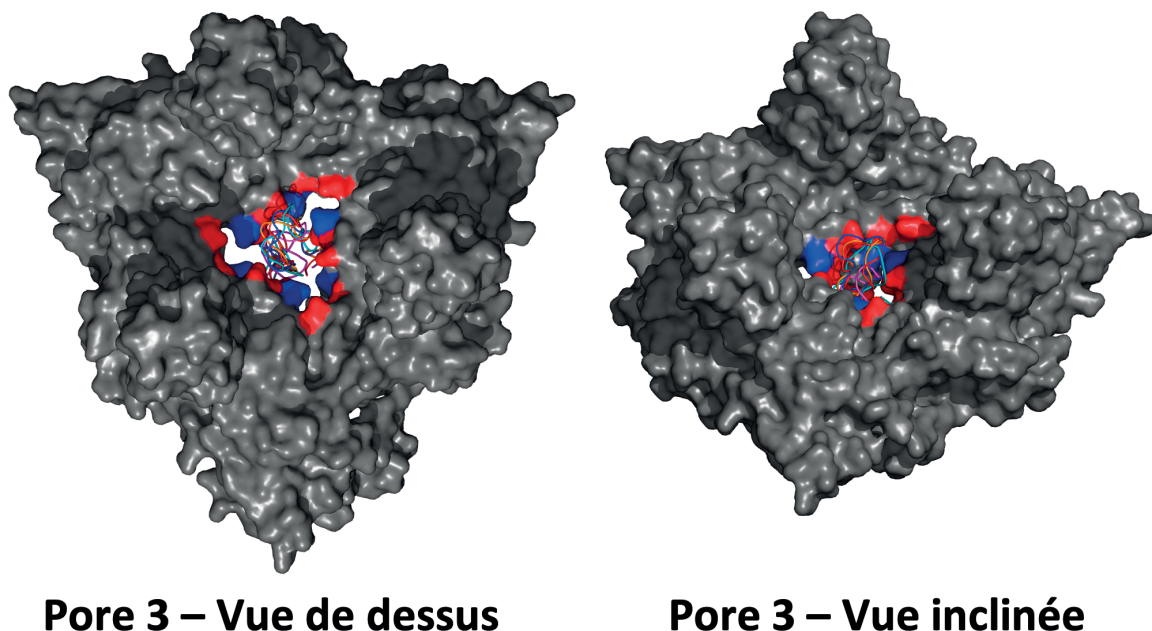


Figure 87 – Exposition du CTD replié par le pore 3. Conformations finales du CTD (en magenta, rouge, cyan, orange et bleu). Les couleurs sont corrélées avec les courbes sur la figure 88. Une couleur par constante harmonique est utilisée. Le pore est représenté en surface et les CTD, en cartoon. Les régions du pore, chargées négativement, sont colorées en rouge et les régions chargées positivement, en bleu.

Les RMSD indiquent que les conformations initiales convergent aussi presque totalement vers la conformation cible (de $\sim 1 \text{ \AA}$ à $\sim 1,4 \text{ \AA}$) (Figure 88). L’exposition partielle des boucles découle de la conformation cible utilisée.

RMSD evolution between target and initial conformation

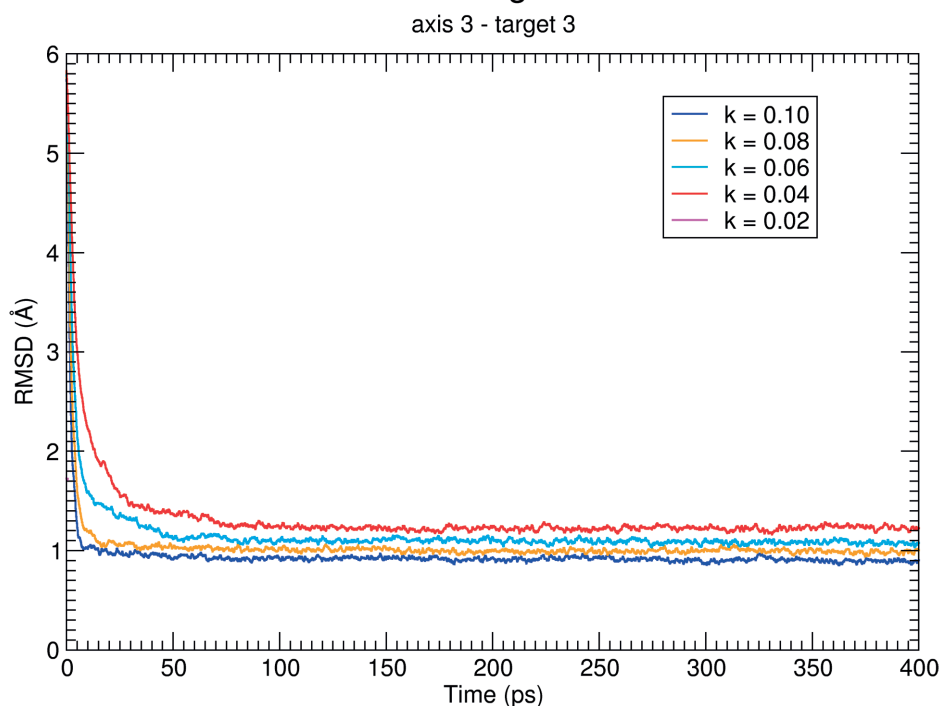


Figure 88 – Évolution du RMSD entre la conformation au temps t (ps) et la conformation cible du pore 3. Le CTD n'est pas rattaché au pore. Il est replié (présence d'une boucle). k correspond à la constante harmonique utilisée dans la TMD.

Au cours de la TMD ($k = 0,01 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$), 2 ponts-salins se forment (Tableau 11). R167 interagit avec E46 et R172 interagit avec E43.

Acide aminé (intérieur du pore)	Acide aminé (CTD)	Domaine riche en arginines	Durée de contact (en %)
E43	R172	ARDIV	20,75
E46	R167	ARDIII	20,25

Tableau 11 – Ponts-salins entre les résidus du pore 3 et le CTD replié, lors de son exposition.

7.7.5.3.2 TMD du Pore q3

En procédant de manière identique, le passage du CTD au travers du pore q3 est étudié.

Le bras C-terminal d'un des dimères A-B du pore q3 est également capable de passer au travers (Figure 89). Dans le modèle cible, le CTD est déplié et en contact avec la spicule de l'un des dimères. Cette conformation est utilisée pour tester l'hypothèse d'exposition de Heger-Stevic et al. [82], dans laquelle, le CTD exposé est déplié (Figure 58, à gauche). Les conformations finales, issues des TMD, sont repliées contre l'une des spicules (à droite sur la figure 89). Seule la constante harmonique $k = 0,02 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ n'est pas suffisante pour provoquer la sortie totale du CTD ainsi que son repliement contre la spicule (à droite sur la figure 89).

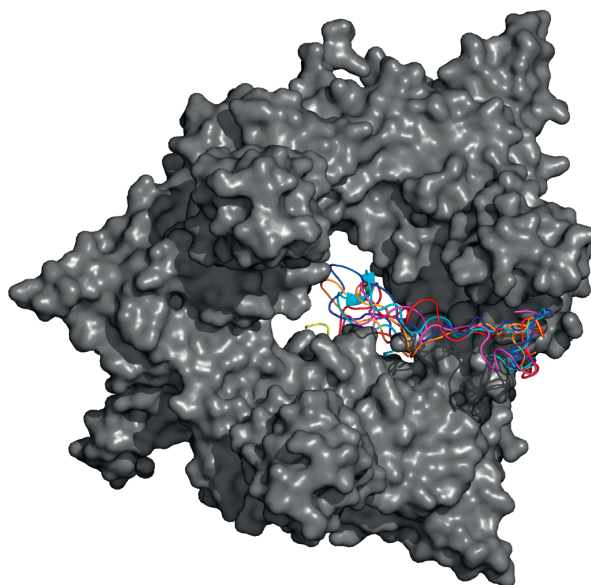


Figure 89 – Exposition du CTD par le pore q3. Conformations finales du CTD (en magenta, rouge, cyan, orange et bleu). Les couleurs sont corrélées avec les courbes sur la figure 90. Une couleur par constante harmonique est utilisée.

Les conformations initiales convergent presque toutes au même point (entre 1,3 et 1,6 Å), sauf pour $k = 0,02 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ ($\sim 2 \text{ \AA}$) (Figure 90). Les conformations du CTD finales des TMD ($k = 0,04 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ à $k = 0,10 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$) sont très exposées (interaction avec une des spicules du pore q3).

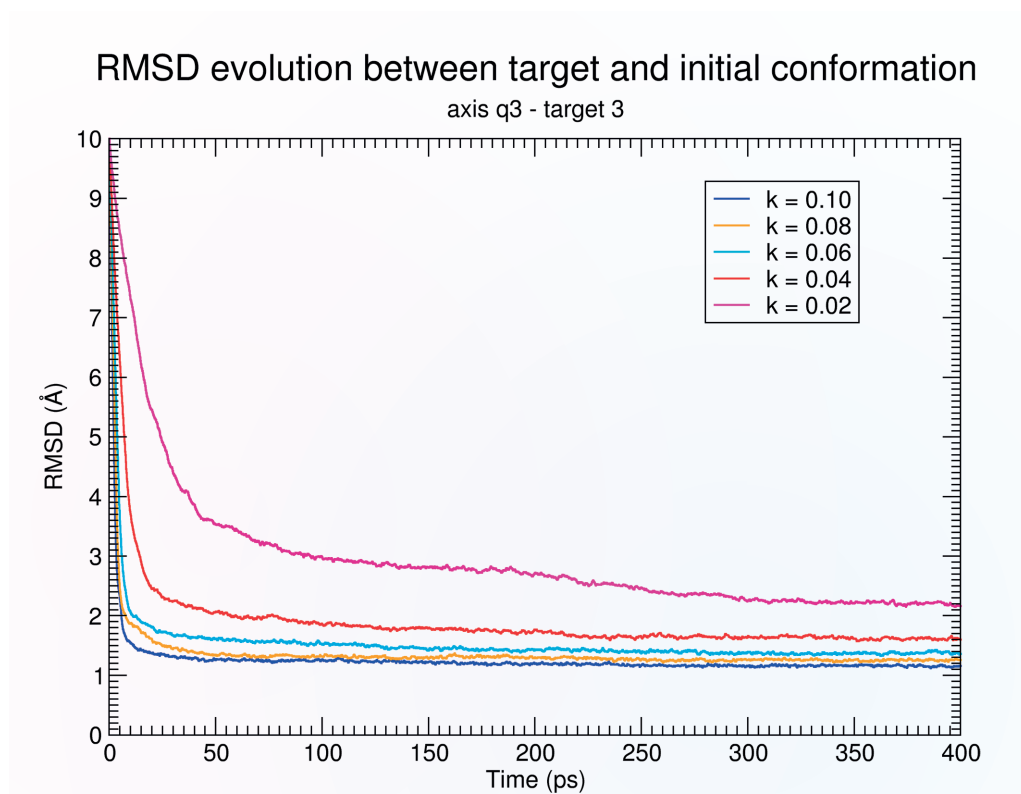


Figure 90 – Évolution du RMSD entre la conformation au temps t (ps) et la conformation cible du pore q3. k correspond à la constante harmonique utilisée dans la TMD.

3 ponts-salins se forment au cours de la TMD ($k = 0,01$ - Tableau 12). R175 interagit avec E14 pendant 20,5% de la simulation. Les autres ponts-salins (R172 - E46 et R173 - E40) sont transitoires.

Acide aminé (intérieur du pore)	Acide aminé (CTD)	Domaine riche en arginines	Durée de contact (en %)
E14	R175	ARDIV	20,5
E40	R173	ARDIV	2,25
E46	R172	ARDIV	1,75

Tableau 12 – Ponts-salins entre les résidus du pore q3 et le CTD, lors de son exposition.

Pour également confirmer ces observations et pour voir jusqu'à quel résidu le CTD est exposé, des TMD comparables à celles du pore 3 sont réalisées. Les systèmes comportent donc un fragment de CTD et le pore 3 (Figure 91). Les 2 types de fragments sont aussi respectivement nommés "CTD-allongé" et "CTD-replié" pour la suite. La procédure est équivalente à celle du pore 3.

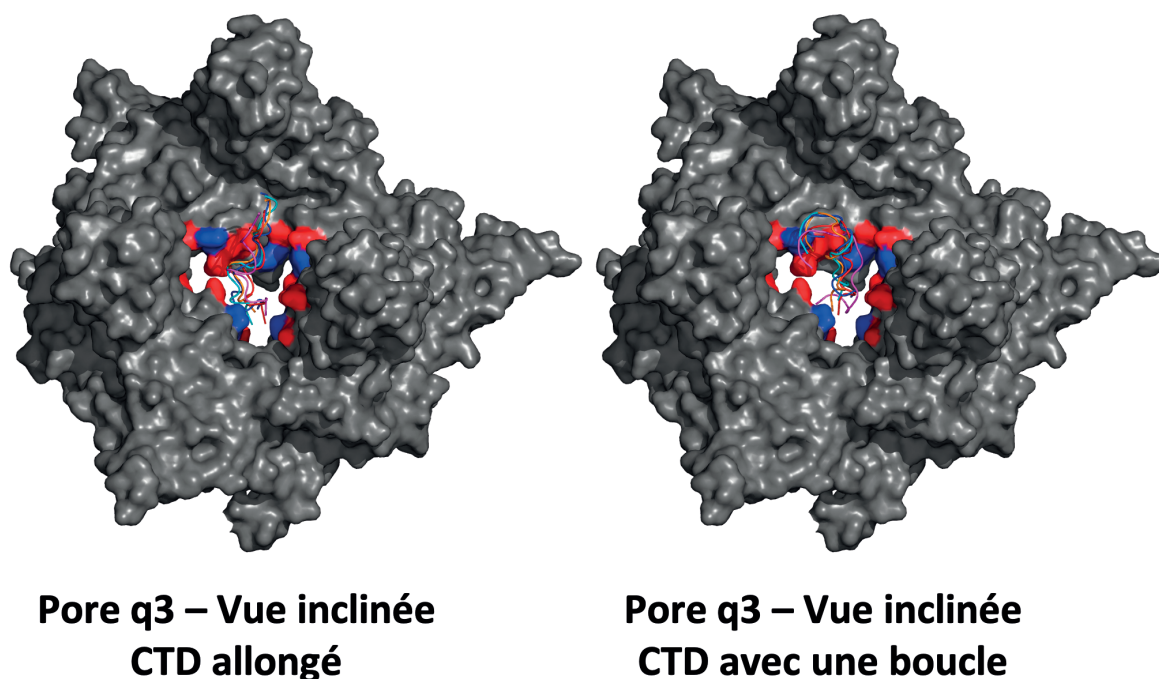


Figure 91 – Exposition du CTD allongé et replié par le pore 3. Conformations finales du CTD (en magenta, rouge, cyan, orange et bleu). Les couleurs sont corrélées avec les courbes sur la figure 92. Une couleur par constante harmonique est utilisée. Le pore est représenté en surface et les CTD, en cartoon. Les régions du pore, chargées négativement, sont colorées en rouge et les régions chargées positivement, en bleu.

Les "CTD-allongé" et les "CTD-replié" sont tous exposés à travers le pore q3, sauf dans la TMD ($k = 0,02$) où le "CTD-replié" est partiellement exposé (Figure 91). L'exposition des CTD est corrélée avec les RMSD calculés (Figure 92).

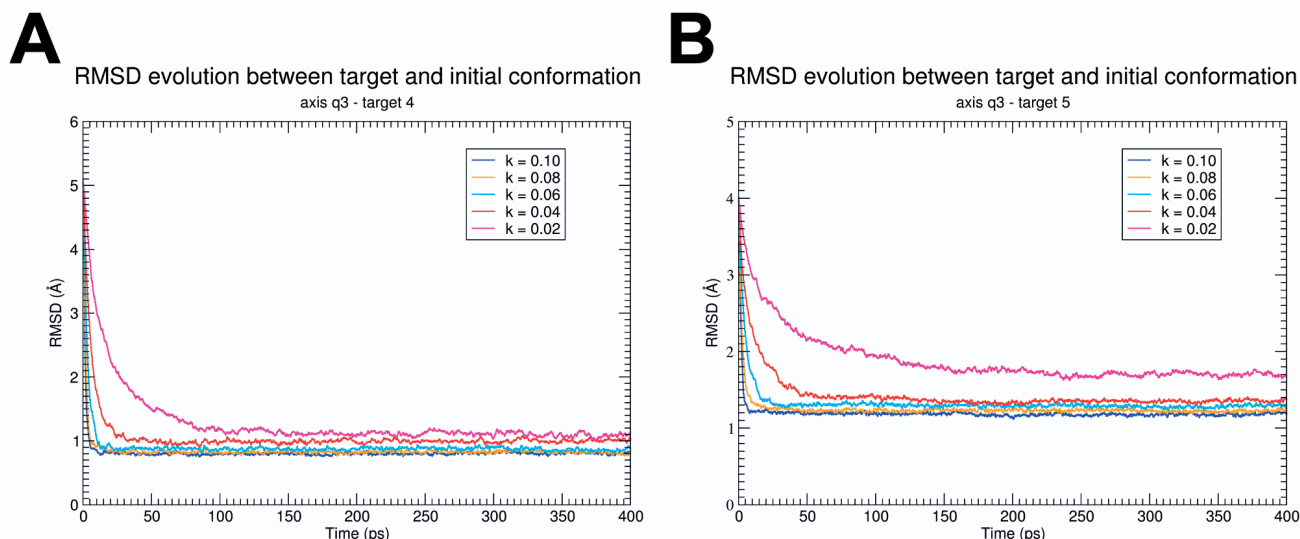


Figure 92 – Évolution du RMSD entre la conformation au temps t (ps) et la conformation cible du pore q3. A. Le CTD n'est pas rattaché au pore et n'est pas replié (allongé). B. Le CTD n'est pas rattaché et comporte une boucle. k correspond à la constante harmonique utilisée dans la TMD.

Les ponts-salins qui se forment sont presque identiques à ceux déjà identifiés pour toutes les TMD précédentes (Tableau 13). Les glutamates E40, E43 et E46 interagissent avec les R172 à R175 et R179. Ils se forment transitoirement ou durent au cours de la TMD ($k = 0,01$).

Acide aminé (intérieur du pore)	Acide aminé (CTD)	Domaine riche en arginines	Durée de contact (en %)
CTD allongé			
E40	R173	ARDIV	25,5
E40	R175	ARDIV	2
E43	R179		99
E46	R172	ARDIV	1,25
CTD comportant une boucle			
D2	R175	ARDIV	10,75
E40	R173	ARDIV	46,75
E40	R174	ARDIV	1,25
E43	R175	ARDIV	7,75
E46	R172	ARDIV	74,5

Tableau 13 – Ponts-salins entre les résidus du pore q3 et le CTD déplié ou replié, lors de son exposition.

En conclusion, les CTD sont capables de passer au travers des pores 3 et q3 en étant dépliés ou repliés. Les ARDIII et ARDIV forment des ponts-salins avec les glutamates (E40, E43 et E46) des pores. Ces résidus, chargés négativement, sont à la sortie des pores. Les ARDIII et IV sont donc majoritairement exposés. Pour améliorer la compréhension de l'exposition du CTD, il faudrait réaliser des MD avec une restriction de distance. Dans ce processus, les vitesses et conformations issues des TMD seraient utilisées. Ainsi, le CTD serait contraint d'explorer le pore et pourrait passer plus naturellement à travers (pores 3 et q3).

8 MOTIF D'INTERACTION DE CORE AVEC LE MATÉRIEL GÉNÉTIQUE

L'étude se focalise désormais sur les motifs de liaisons potentielles à l'ADN de Core et plus spécifiquement, à l'interaction de la protéine de capsid avec le zinc.

Lors des travaux menés sur Core, la question du motif d'interaction avec le matériel génétique se pose. En effet, la protéine Core est en interaction avec du matériel génétique (ARNpg, ADNrc, ADNccc...) tout au long du cycle viral. La structure primaire et les structures secondaires de Core sont, en grande majorité, des hélices α (Figure 93), Les motifs d'interactions nécessitant des feuilletts β ou constitués d'un enchaînement de leucine sont donc exclus. Par élimination, les seules possibilités sont :

- Les motifs hélice-boucle-hélice.
- Les motifs pour lesquels un dication zinc est nécessaire.

8.1 Hypothèse du motif de liaison à l'ADN faisant intervenir du zinc

Il s'avère que la base du domaine d'assemblage (NTD), partie qui tapisse l'intérieur de la capsid de Core, comporte des cystéines et des histidines (Figure 93).

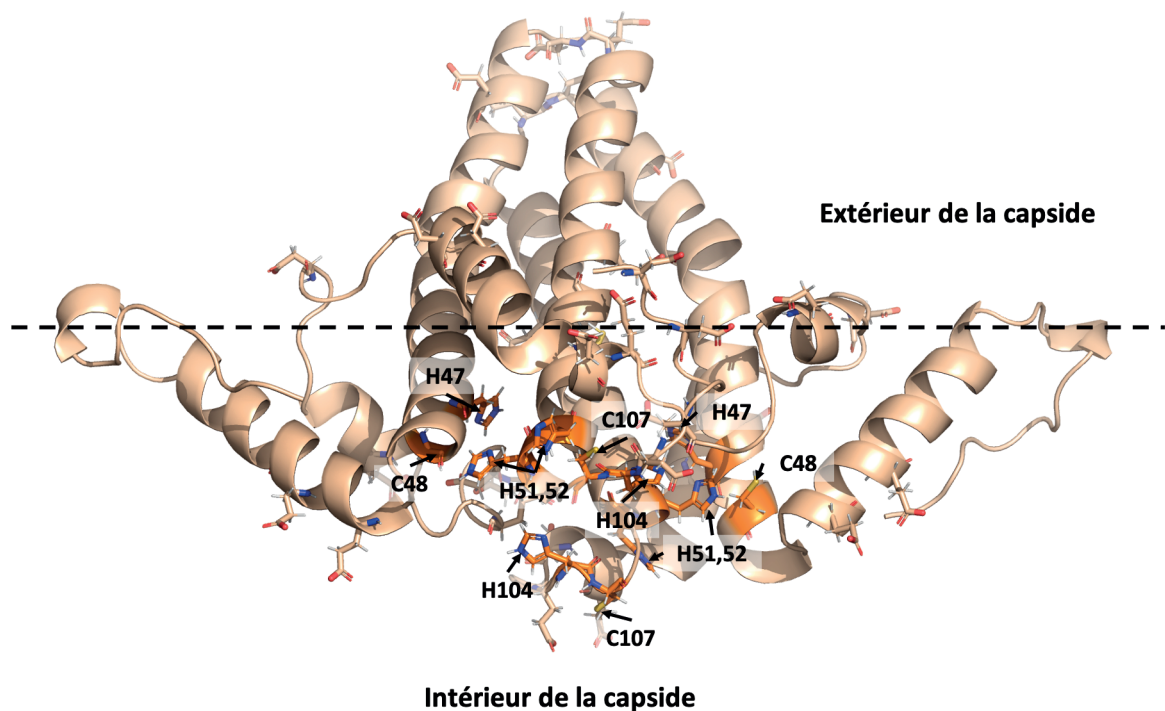


Figure 93 - Localisation des résidus potentiellement impliqués dans la chélation du Zn^{2+} . Les cystéines et histidines sont représentées en bâtonnet orange.

À proximité des histidines et cystéines se trouvent des acides aminés chargés négativement. Elles semblent accessibles pour chélater des dications métalliques et, en particulier, le zinc. Les sels ($NaCl$, $ZnCl_2$...) sont couramment utilisés pour déclencher l'assemblage de dimères dissociés. Stray et al. utilisent plusieurs sels et les comparent [77]. Leurs données de fluorescence intrinsèque sont une indication claire que Cp149 lie le zinc. Ils estiment ainsi la stoechiométrie à 3 à 4.5 Zn^{2+} par dimère, selon les conditions expérimentales [77].

Afin de déterminer si les résidus potentiellement impliqués dans la chélation d'ions Zn^{2+} sont conservés, un alignement multiple de 12 294 séquences de la protéine Core est réalisé (Figure 94). Les séquences proviennent de la base de données des VHB (HBVdb : <https://hbvdb.lyon.inserm.fr>). Pour classier les séquences, la méthode de groupe de paires non pondérées avec moyenne arithmétique est utilisée (UPGMA). L'alignement comporte 9 génotypes, du génotype A à H ainsi que le génotype RF. La plupart des séquences sont composées de 183 résidus. Chez les génotypes A et G, Core comporte respectivement 185 et 195 acides aminés. Le résultat de cet alignement indique que les cystéines et les histidines, situées en dessous du NTD, sont strictement conservées. Il montre peu de variabilité autour des cystéines et histidines. Les résidus à fort potentiel chélateur sont presque tous localisés et regroupés en dessous du NTD (Figure 93). Les cystéines et histidines sont situées au début de l'hélice α_4 et à la fin de l'hélice α_5 sauf, C183 (dernier résidu) et C61 (enfoui dans l'interface dimérique).



Figure 94 – Représentation en weblogo de l'alignement multiple des 12 294 séquences de Core.

Les résidus grisés correspondent aux régions d'insertion ou de délétion de l'alignement multiple.

Les poids des positions calculés par weblogo, révèlent qu'elles ont un poids beaucoup plus faible que les zones non grisées. Un schéma simplifié de la localisation des structures secondaires figure en dessous du weblogo.

8.2 Chélation du Zn²⁺ par Core *in-vitro*

Pour confronter l'hypothèse d'un motif d'interaction avec le matériel génétique, faisant intervenir du zinc, il est apparu nécessaire de tester *in-vitro* et *in-silico* la fixation du zinc par la protéine Core. Dans le cadre de l'identification et de la quantification du zinc, la protéine Core sauvage et deux protéines recombinantes sont exprimées, purifiées et titrées.

8.2.1 Production de la protéine Core Cp149 et Cp183

Les protéine Core recombinantes tronquées Cp149 (résidus 1 à 149) et complètes Cp183 produites en *Escherichia coli* sont purifiées. Ceci, afin d'identifier et de quantifier le zinc putativement en interaction avec la protéine Core. Pour des raisons de disponibilité 2 mutants de la boucle externe de la spicule de Cp183 A80K et E77K sont utilisés. Une quantité suffisante est purifiée (~1mg.ml⁻¹, concentration mesurée selon la méthode de Porterfield & Zlotnick – section 8.2.3) afin de réaliser des expériences de colorimétrie et de spectrométrie de masse (ICP-MS). La protéine Core Cp149 est purifiée en accord avec les protocoles établis par le Dr. Michael Nassal et adaptés par le MMSB de Lyon. En ce qui concerne les protéines complètes, les protocoles fournis par le Dr. Lauriane Lecoq du MMSB ont été adaptés par le Dr. Maëleonn Chevreuil (Numéro national de thèse : 2020UPASS043). Sous la direction et grâce à l'aide du Dr. Chevreuil, les différentes protéines Core sont purifiées au CEA de Saclay. Les protéines Core, une fois exprimées, s'assemblent spontanément sous forme de capsides.

8.2.2 Vérification de l'intégrité des capsides par TEM

Grâce à des méthodes biophysiques, il est possible de vérifier que les protéines purifiées sont bien assemblées sous forme de capsides. La réalisation de grilles de microscopies, observées avec le microscope électronique en transmission (TEM) de l'équipe, ainsi que les expériences de diffusion dynamique de la lumière (DLS), donnent des informations sur l'agrégation des protéines de capsides. Ces expériences sont réalisées avec et sous la supervision du Dr. Pierre Chervy. La réalisation des grilles de microscopie électronique se fait de la manière suivante :

- Dans un premier temps, les grilles sont ionisées (1min).
- Les échantillons dilués sont ensuite déposés et laissés en contact (30 sec).
- Une étape d'absorption de l'excès de liquide est ensuite effectuée.
- Une solution d'acétate d'uranyle est également utilisée.
- Les grilles sont ensuite séchées (10 min).
- Les grilles sont finalement observées sur "Le Redoutable" au CEA de Saclay bâtiment 532 (Figure 95A et B).

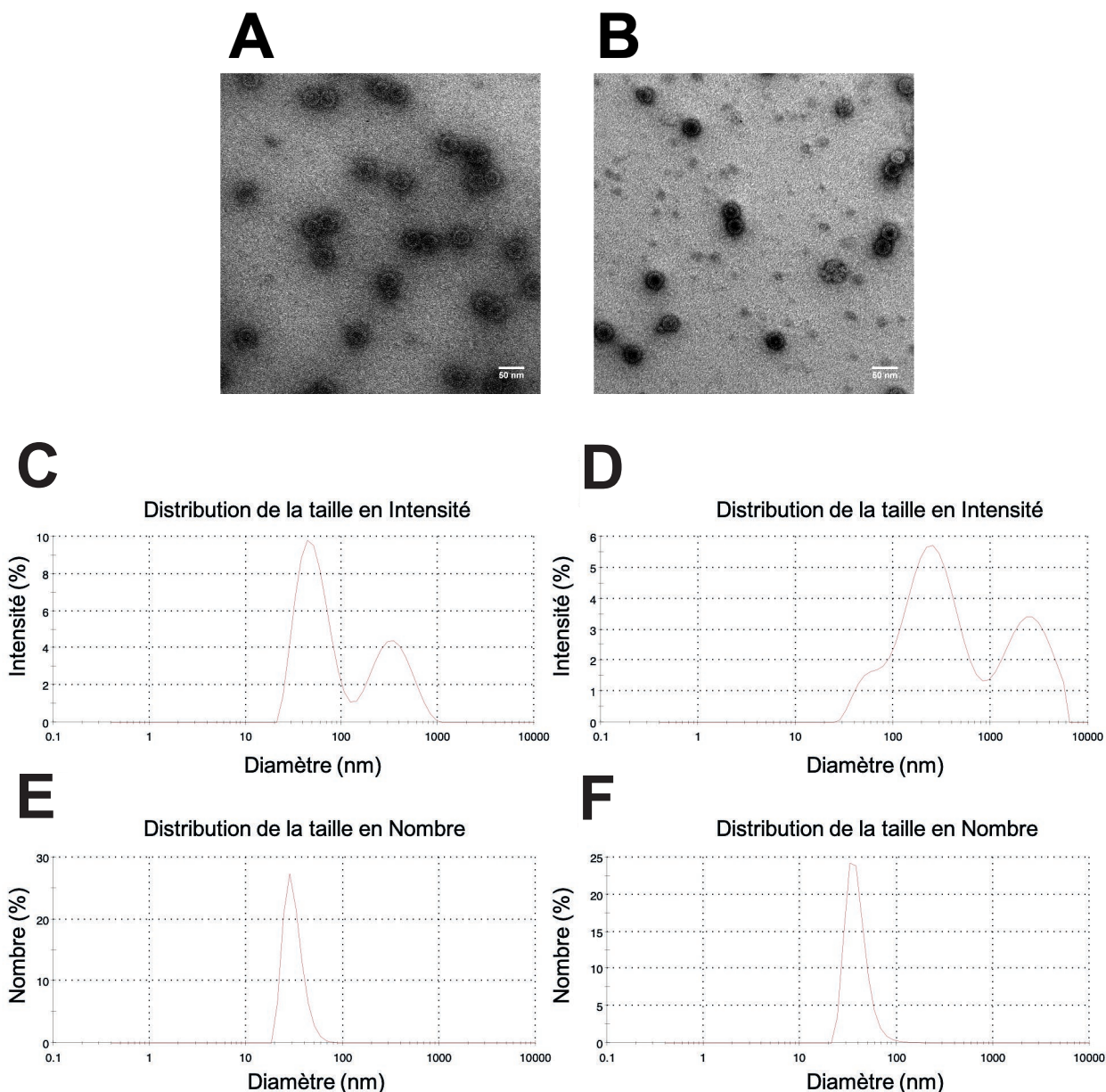


Figure 95 – Résultats de TEM et DLS. A. Observation de l'échantillon Cp149 par coloration négative. B. Observation de l'échantillon Cp183 A80K par coloration négative. C. Distribution de la taille, en Intensité, de l'échantillon Cp149 (premier pic : ~30 nm). D. Distribution de la taille, en intensité, de l'échantillon Cp183 (premier pic : ~30 nm). E. Distribution de la taille, en nombre, de l'échantillon Cp149 (~30 nm). F. Distribution de la taille, en nombre, de l'échantillon Cp183 (~30 nm).

Le diamètre moyen des capsides se situe autour de ~36 nm. Lors de DLS, les échantillons sont directement utilisés et analysés. Si l'on s'intéresse à la distribution en intensité des échantillons, plusieurs espèces semblent contribuer au signal (Figure 95A et B). Une taille de l'ordre de 35 nm des capsides Cp149 est attendue. Cela correspond au premier pic d'intensité sur la Figure 95C. L'autre pic pourrait correspondre à des agrégats de capside. La capside Cp183 a une propension plus élevée à former des agrégats. Si l'on regarde la représentation des échantillons en nombre (Figure 95C et D), les objets d'un diamètre de l'ordre de 30-40 nm dominent la distribution. Les échantillons contiennent donc majoritairement des objets de cette taille, ce qui correspond aux diamètres attendus des capsides Cp149 et Cp183.

8.2.3 Quantification de la protéine Core – un processus pas si évident

Des expériences préliminaires de titration de Core et de quantification du zinc sont réalisées. 2 campagnes sont ensuite menées pour confirmer les expériences préliminaires. Dans les travaux préliminaires, la concentration de Cp149 est estimée. Pour les 2 campagnes, la concentration des protéines de capsid est systématiquement mesurée. Elles sont quantifiées selon 2 méthodes biophysiques et 2 méthodes biochimiques. Il s'agit d'une étape importante pour rapporter la quantité de zinc à la quantité de protéines de capsid. Dans la section suivante, les méthodes biophysiques et biochimiques utilisées ainsi que les résultats de titration des protéines de capsid, seront présentés. Les résultats préliminaires seront montrés dans la section portant sur la titration du zinc.

8.2.3.1 Méthodes biophysiques

A. Méthode de Porterfield & Zlotnick

La méthode de quantification de Porterfield et Zlotnick est une technique biophysique utile pour quantifier la teneur en acides nucléiques et en protéines virales d'un échantillon [115]. Cette méthode dépend des mesures d'absorbance effectuées à 4 longueurs d'ondes : 260, 280, 340 et 360 nm.

Elle repose, en partie, sur l'absorption maximum des purines et pyrimidines à 260 nm et sur le rapport des mesures à 260 et 280 nm. Ce rapport est égal à ~2,0 pour les ARN monocaténares. L'absorbance des protéines est, elle, liée à leur composition en tryptophanes, tyrosines et ponts disulfures. Le point d'absorption maximum d'une protéine se situe autour de 280 nm. Le rapport des mesures à 260 et 280 nm est tout aussi important pour les protéines. Il est d'environ 0,6.

Cette méthode tient compte de la diffusion de la lumière pour des objets de 40 nm. Elle est estimée à partir de mesures à 340 et 360 nm, où les protéines et l'ARN n'absorbent pas. Elle est ensuite extrapolée à 260 et 280 nm pour être retirée des absorptions apparentes à ces longueurs d'onde (Équations 1a). Finalement, les concentrations sont calculées algébriquement de façon classique sur ces absorptions corrigées (Équations 1b).

Cette méthode a été spécialement développée pour déterminer la concentration des protéines de capsid virale [115]. Elle est notamment utilisée sur la protéine Core. Cette méthode est donc fiable pour quantifier les protéines de capsid.

Dans ces équations, A correspond à l'absorbance des protéines de capsides et du matériel génétique viral pour une lumière monochromatique de longueur d'onde λ . $\epsilon_{\text{protein}}$ et ϵ_{RNA} sont les coefficients d'extinctions molaires des protéines de capsides et du matériel génétique respectivement. L est la longueur du trajet parcouru par la lumière :

$$a. A_{corrected,\lambda} = A_{\lambda} - \left(\frac{A_{340} - A_{360}}{(340nm)^{-4} - (360nm)^{-4}} \right) \lambda^{-4} - A_{340} + \left(\frac{A_{340} - A_{360}}{(340nm)^{-4} - (360nm)^{-4}} \right) 340nm^{-4}$$

$$b. [protein] = \frac{(A_{corrected,260} - \left(\frac{\epsilon_{RNA260}}{\epsilon_{RNA280}} \right) A_{corrected,280})}{\left(\epsilon_{protein,260} - \left(\frac{\epsilon_{RNA260}}{\epsilon_{RNA280}} \right) \epsilon_{protein,280} \right) L}$$

$$c. [RNA] = \frac{(A_{corrected,260} - \left(\frac{\epsilon_{protein260}}{\epsilon_{protein280}} \right) A_{corrected,280})}{\left(\epsilon_{RNA,260} - \left(\frac{\epsilon_{protein260}}{\epsilon_{protein280}} \right) \epsilon_{RNA,280} \right) L}$$

Équations 1 – Détermination de la concentration en protéines de capside et d'ARN selon la méthode de Porterfield & Zlotnick. a. Correction de l'absorbance λ . b. Calcul de la concentration en protéine de capsides à partir des absorbances corrigées à 260 et 280 nm. c. Calcul de la concentration d'ARN à partir des absorbances corrigées à 260 et 280 nm.

B. Méthode optique de quantification des protéines – Absorbance à 205 nm

Il est aussi possible de mesurer la concentration d'une protéine en mesurant son absorbance à 205 nm. Cette longueur d'onde correspond au point d'absorption maximum des liaisons peptidiques [116]. La liaison peptidique absorbe entre 190 et 230 nm. La plupart des protéines ont des coefficients d'extinction massique voisins à 205 nm. Pour une concentration d'environ 1 mg.ml⁻¹ de protéine, l'absorbance à 205 nm est comprise entre 30 et 35 [117]. Par ailleurs, les acides nucléiques n'absorbent pas à cette longueur d'onde. Il aurait également fallu tenir compte de la diffusion de la lumière pour des objets de 40 nm, mais nos résultats n'en tiennent pas compte.

C. Résultats de quantification des protéines de capsides par les méthodes biophysiques

Des mesures d'absorbance sont réalisées à 205, 260, 280, 340 et 360 nm. Les Équations 1 de Porterfield & Zlotnick sont ensuite appliquées. L'absorbance à 205 nm est divisée par 31 pour obtenir la concentration des protéines de capsides.

Dans le tableau 14 figurent les mesures d'absorbance et les concentrations calculées lors de la première campagne de la protéine de capside du VHB recombinante tronquée Cp149.

Échantillon mesuré Campagne n°1	205 nm	260 nm	280 nm	340 nm	360 nm	[Cp149] mg.ml ⁻¹ (205 nm)	[Cp149] mg.ml ⁻¹ (Portfield & Zlotnick)
Blanc (1)	0	0,02	0	0,01	0,02	0	0
Blanc (2)	0	0,01	0,01	0,02	0,01	0	0
Blanc (3)	0	0,01	0,01	0	0	0	0
Éluat centrifugation (1)	0,08	0	0	0	0	0	0
Éluat centrifugation (2)	0,08	0,02	0,03	0,01	0,01	0	0
Éluat centrifugation (3)	0,1	0,02	0,02	0	0,01	0	0
Cp149 (1)	9,15	2,21	3,04	0,19	0,17	0,30	1,44
Cp149 (2)	9,35	2,23	3,06	0,24	0,19	0,30	1,40
Cp149 (3)	9,55	2,26	3,08	0,24	0,2	0,31	1,41
Cp149 dilué au 10 ^{ème} (1)	4,71	0,18	0,24	0	0	0,15	1,22
Cp149 dilué au 10 ^{ème} (2)	4,79	0,16	0,21	0	0	0,15	1,18
Cp149 dilué au 10 ^{ème} (3)	4,73	0,14	0,19	0	0	0,15	1,18

Tableau 14 – Mesures d’absorbances et concentrations déduites pour la protéine Cp149 lors de la 1^{ère} campagne.

Lors de l’étape de production, les concentrations sont déterminées selon la méthode de Porterfield & Zlotnick. Les mesures d’absorbances à 205 nm qui sont réalisées ne sont pas cohérentes. L’absorbance à 205 nm est largement sous-estimée par rapport à la concentration déterminée, par la méthode de Porterfield & Zlotnick, à l’étape de production (~1 mg.ml⁻¹). Les propriétés géométriques de la capsid et la diffusion de la lumière conduit à surestimer l’absorbance à 205 nm. La dilution induit systématiquement une erreur de détermination de la concentration de protéines de capsid.

Si on se base sur les concentrations déterminées à partir de la méthode de Porterfield & Zlotnick, en fonction de la dilution au 10^{ème} ou non, les concentrations moyennes seraient de ~1,2 mg.ml⁻¹ ou de ~1,4 mg.ml⁻¹. Il y a une différence de 19% entre les deux moyennes. Elle correspond à l’erreur de dilution.

La concentration moyenne de la Cp149 est de ~1,3 mg.ml⁻¹.

L’albumine de sérum bovin (BSA), d’une concentration connue de 2 mg.ml⁻¹, est quantifiée de la même manière. Elle a servi de contrôle. Les mesures d’absorbance, à 205 nm, ne sont pas celles attendues. Les mesures d’absorbance, à 260 et 280 nm, confirment que la protéine à une concentration de ~2,0 mg.ml⁻¹.

Échantillon mesuré Campagne n°2	205 nm	260 nm	280 nm	340 nm	360 nm	[Cp149] mg.ml ⁻¹ (205 nm)	[Cp149] mg.ml ⁻¹ (Portfield & Zlotnick)
Cp149 dilué au 10 ^{ème} (1)	4,14	0,25	0,3	0,07	0,07	0,13	1,10
Cp149 dilué au 10 ^{ème} (2)	4,26	0,25	0,31	0,09	0,07	0,14	0,94
Cp149 dilué au 10 ^{ème} (3)	4,46	0,29	0,34	0,09	0,08	0,14	1,10
Cp149 dilué au 10 ^{ème} (4)	4,00	0,09	0,15	0	0	0,13	0,90
Cp149 dilué au 10 ^{ème} (5)	4,04	0,11	0,17	0	0	0,13	0,98
Cp149 (1)	11,99	2	2,59	0,19	0,13	0,39	1,13
Cp149 (2)	12,52	2,1	2,72	0,29	0,24	0,40	1,16
Cp149 (3)	13,03	1,96	2,56	0,15	0,11	0,42	1,16
Cp149 (4)	12,14	1,89	2,5	0,1	0,05	0,39	1,15
Cp149 (5)	12,47	2,13	2,74	0,25	0,19	0,40	1,17
Cp149 (6)	11,43	2,13	2,74	0,31	0,27	0,37	1,17
Cp149 (7)	11,41	2,19	2,8	0,37	0,33	0,37	1,17
Cp149 (8)	11,61	2,12	2,73	0,32	0,26	0,37	1,14

Tableau 15 – Mesures d'absorbances et concentrations déduites pour la protéine Cp149 lors de la 2^{ème} campagne.

Les mêmes mesures sont réalisées lors de la deuxième campagne sur la protéine Cp149 (Tableau 15), issue de la même production, et sur la protéine Cp183 E77K (Tableau 16). De façon similaire, les mesures d'absorbance à 205 nm sous-estiment la concentration des protéines de capsides.

Si on prend en compte les concentrations déterminées avec la méthode de Porterfield & Zlotnick, des échantillons dilués ou non de la protéine Cp149, les concentrations moyennes seraient de ~1,0 mg.ml⁻¹ ou de ~1,2 mg.ml⁻¹ (différence de 16%).

La concentration moyenne de la Cp149 est de ~1,1 mg.ml⁻¹.

Échantillon mesuré Campagne n°2	205 nm	260 nm	280 nm	340 nm	360 nm	[Cp183] mg.ml ⁻¹ (205 nm)	[Cp183] mg.ml ⁻¹ (Portfield & Zlotnick)
Cp183 1/10 (1)	3,40	0,46	0,29	0	0	0,11	0,83
Cp183 1/10 (2)	3,50	0,55	0,38	0	0	0,11	0,73
Cp183 1/10 (3)	3,46	0,57	0,38	0,01	0	0,11	0,68
Cp183 1/10 (4)	3,49	0,54	0,37	0	0	0,11	0,83
Cp183 1/10 (5)	3,82	0,77	0,57	0,17	0,14	0,12	0,68
Cp183 1/10 (6)	3,69	0,71	0,52	0,15	0,12	0,12	0,58
Cp183 1/10 (7)	3,75	0,74	0,54	0,18	0,18	0,12	0,79
Cp183 (1)	14,48	7,62	5,71	1,57	1,33	0,47	0,85
Cp183 (2)	14,78	8	6,03	1,69	1,45	0,48	0,92
Cp183 (3)	14,62	8,02	6,05	1,63	1,39	0,47	0,96
Cp183 (4)	14,73	7,91	5,97	1,67	1,43	0,48	0,92
Cp183 (5)	14,46	7,66	5,74	1,6	1,36	0,47	0,85
Cp183 (6)	13,14	8,02	6,1	1,88	1,63	0,42	0,88
Cp183 (7)	13,28	7,4	5,61	1,72	1,48	0,43	0,79

Tableau 16 – Mesures d'absorbances et concentrations déduites pour la protéine Cp183 E77K lors de la 2^{ème} campagne.

Pour la suite la protéine Cp183 E77K sera nommée Cp183. Les concentrations moyennes des échantillons de la protéine Cp183, dilués ou non, sont aussi calculées. Elles seraient de ~0,7 mg.ml⁻¹ ou de ~0,9 mg.ml⁻¹ respectivement (erreur de dilution de 20%).

La concentration moyenne de la Cp183 est de ~0,8 mg.ml⁻¹.

8.2.3.2 Méthodes biochimiques

Le dosage colorimétrique de Bradford [118] et la méthode à l'acide bicinchoninique [119] sont les méthodes biochimiques utilisées pour titrer les protéines de capsid Cp149 et Cp183 (Tableau 17).

Dans le cadre du dosage des protéines, le tampon des protéines de capsid est changé en effectuant des concentrations/dilutions successives. Le tampon utilisé lors de cette étape est compatible avec les méthodes de dosage :

- tampon d'origine - 50 mM Tris pH 7.5, 5% sucrose, 1 mM DTT
- tampon final - 50 mM Tris pH 7.5

Des gammes étalons, avec la BSA de concentration connue à 2 mg.ml⁻¹, sont réalisées lors de la première et deuxième campagne (Figure 96).

Méthodes biochimiques et principes	Avantages	Inconvénients
<p>Méthode de Bradford</p> <p>C'est un dosage colorimétrique qui dépend de la fixation du bleu de Coomassie sur la protéine. Le bleu de Coomassie se lie aux résidus basiques (arginine, histidine, lysine) et hydrophobes. Il change de couleur en fonction de la concentration de la protéine et induit un changement d'absorbance. L'absorption optique est maximale à 595 nm.</p>	<ul style="list-style-type: none"> • Rapide, • Très simple, • Moins sensible aux interférences par divers agents. 	<ul style="list-style-type: none"> • Nécessite une gamme étalon, • Méthode linéaire sur un intervalle étroit [0,2 mg.ml⁻¹ – 0,9 mg.ml⁻¹], • Dépend intrinsèquement de la nature de la protéine, • Pas adapté en présence de matériel génétique, • Effet du tampon, bases fortes, détergents, ions ou lipides.
<p>Méthode BCA</p> <p>La méthode BCA ou méthode à l'acide bicinchoninique est une méthode mettant en œuvre la réduction des ions Cu²⁺ par les liaisons peptidiques des protéines. C'est une réaction chimique colorée de Biuret. L'absorption optique est maximale à 562 nm.</p>	<ul style="list-style-type: none"> • Relativement rapide, • Méthode sensible, • Très linéaire, • Moins de variabilité du signal selon la nature de la protéine 	<ul style="list-style-type: none"> • Nécessite une gamme étalon, • Méthode linéaire sur un intervalle étroit [0,5 µg.ml⁻¹ – 20 µg.ml⁻¹], • Effet du tampon, acides forts, agents réducteurs (DTT), chélateurs de métaux (EDTA), tampon TRIS.

Tableau 17 – Principes, avantages et inconvénients des méthodes biochimiques utilisées pour doser les protéines de capsides.

A. Concentration des protéines de capside estimées

Les gammes d'étalonnage qui sont réalisées ont un coefficient de détermination de ~0,99 (Figure 96). Elles servent à déduire les concentrations des échantillons.

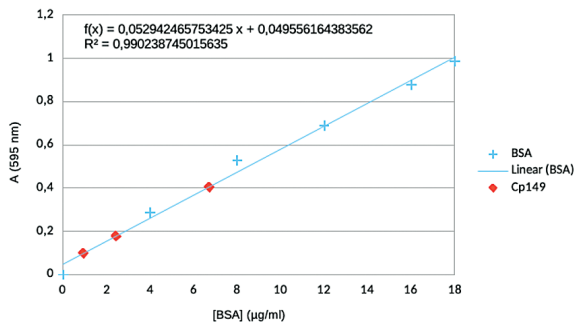
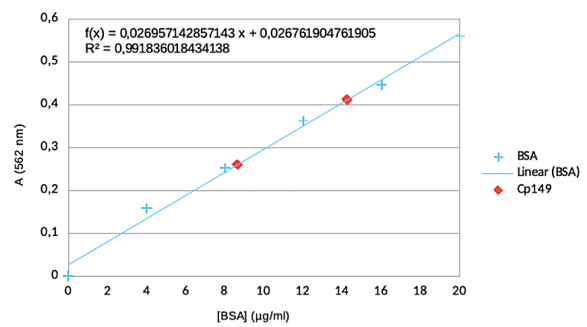
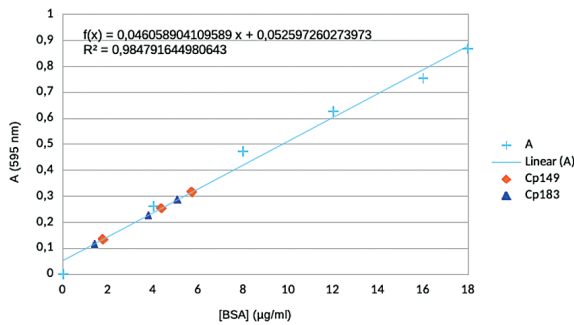
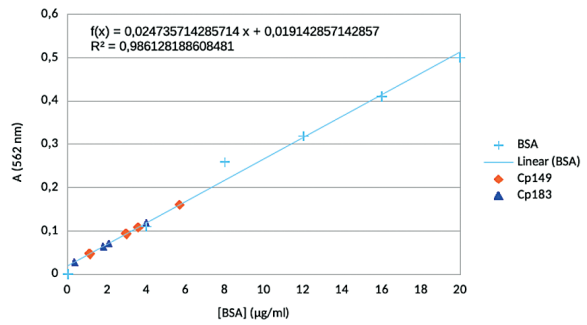
A**B****C****D**

Figure 96 - Gammes d'étalonnages de la BSA et mesures d'absorbance des échantillons. A. et B. Première campagne de quantification sur l'échantillon Cp149. C. et D. Deuxième campagne de quantification sur les échantillons Cp149 et Cp183. A. et C. Gammes d'étalonnages réalisées selon la méthode de Bradford. B. et D. Gammes d'étalonnages réalisées selon la méthode BCA.

B. Concentrations de la protéine Cp149 et Cp183 estimées par la méthode Bradford

En utilisant les gammes d'étalonnages (Figure 96A et C) de la méthode de Bradford, les concentrations des protéines de capsid sont déduites. Dans les tableaux 18 et 19 figurent les mesures d'absorbance à 595 nm des protéines Cp149 et Cp183 et les concentrations calculées.

Facteur de dilution de l'échantillon Cp149	A (595 nm)	[Cp149] lecture directe (µg.ml ⁻¹)	[Cp149] (µg.ml ⁻¹)	[Cp149] (mg.ml ⁻¹)	[Cp149] arginine* (mg.ml ⁻¹)
1000	0,099	0,93	933,92	0,93	0,74
500	0,178	2,43	1213,05	1,21	0,97
200	0,405	6,71	1342,76	1,34	1,07

Tableau 18 - Concentrations de la protéine Cp149 obtenues selon la méthode de Bradford à la 1^{ère} campagne

Composition en arginine de la BSA : 26/607 soit 4,3% ou 0,375 arginines/kDa
 Composition en arginine de la Cp149 : 8/149 soit 5,4% ou 0,476 arginines/kDa

Les mesures d'absorbance pour les dilutions ne sont pas fiables. Il y a des erreurs de dilution. La concentration moyenne de la protéine Cp149 par la méthode Bradford à la première campagne est de $\sim 1,2 \text{ mg.ml}^{-1}$. Si on pondère par rapport à la composition en arginines, la concentration est de $\sim 0,9 \text{ mg.ml}^{-1}$.

Facteur de dilution de l'échantillon Cp149	A (595 nm)	[Cp149] lecture directe ($\mu\text{g.ml}^{-1}$)	[Cp149] ($\mu\text{g.ml}^{-1}$)	[Cp149] (mg.ml^{-1})	[Cp149] arginine* (mg.ml^{-1})
500	0,134	1,77	883,70	0,88	0,70
250	0,254	4,32	1093,20	1,09	0,87
200	0,316	5,72	1143,80	1,14	0,91
Facteur de dilution de l'échantillon Cp183	A (595 nm)	[Cp183] lecture directe ($\mu\text{g.ml}^{-1}$)	[Cp183] ($\mu\text{g.ml}^{-1}$)	[Cp183] (mg.ml^{-1})	[Cp183] arginine* (mg.ml^{-1})
500	0,117	1,40	699,10	0,70	0,23
250	0,227	3,79	946,60	0,95	0,31
200	0,287	5,09	1017,80	1,02	0,33

Tableau 19 – Concentrations de la protéine Cp149 et Cp183 E77K obtenues selon la méthode de Bradford à la 2^{ème} campagne.

Composition en arginine de la BSA : 26/607 soit 4,3% ou 0,375 arginines/kDa
 Composition en arginine de la Cp149 : 8/149 soit 5,4% ou 0,476 arginines/kDa
 Composition en arginine de la Cp183 : 24/183 soit 13,1% ou 1,137 arginines/kDa

À la deuxième campagne de mesures, les concentrations de la Cp149 et de la Cp183 sont respectivement de $\sim 1,0 \text{ mg.ml}^{-1}$ et de $\sim 0,9 \text{ mg.ml}^{-1}$. Si on pondère la composition en arginines, les concentrations sont respectivement de $\sim 0,8 \text{ mg.ml}^{-1}$ et de $\sim 0,3 \text{ mg.ml}^{-1}$.

C. Concentrations de la protéine Cp149 et Cp183 estimées par la méthode BCA

Les gammes d'étalonnages de la méthode BCA (Figure 96B et D) sont aussi utilisées. Elles servent à déduire la concentration des protéines de capsid. Dans les tableaux 20 et 21 figurent les mesures d'absorbance à 562 nm et les concentrations calculées pour les protéines Cp149 et Cp183.

Facteur de dilution de l'échantillon Cp149	A (562 nm)	[Cp149] lecture directe ($\mu\text{g.ml}^{-1}$)	[Cp149] ($\mu\text{g.ml}^{-1}$)	[Cp149] (mg.ml^{-1})
100	0,26	8,652	865,22	0,87
50	0,411	14,254	712,68	0,71

Tableau 20 - Concentrations de la protéine Cp149 obtenues selon la méthode BCA à la 1^{ère} campagne.

Pour la première campagne, la concentration moyenne de la protéine Cp149 par la méthode BCA est de $\sim 0,8 \text{ mg.ml}^{-1}$.

Facteur de dilution de l'échantillon Cp149	A (562 nm)	[Cp149] lecture directe ($\mu\text{g.ml}^{-1}$)	[Cp149] ($\mu\text{g.ml}^{-1}$)	[Cp149] (mg.ml^{-1})
500	0,047	1,40	698,70	0,70
250	0,093	3,13	782,60	0,78
200	0,108	3,70	739,10	0,74
125	0,160	5,65	706,80	0,71
Facteur de dilution de l'échantillon Cp183	A (562 nm)	[Cp183] lecture directe ($\mu\text{g.ml}^{-1}$)	[Cp183] ($\mu\text{g.ml}^{-1}$)	[Cp183] (mg.ml^{-1})
500	0,028	0,68	340,80	0,34
250	0,064	2,04	509,40	0,51
200	0,071	2,30	460,30	0,46
125	0,118	4,07	509,90	0,51

Tableau 21 – Concentrations de la protéine Cp149 et Cp183 E77K obtenues selon la méthode de BCA à la 2^{ème} campagne.

À la deuxième campagne, les mesures réalisées sur la Cp183 sont très faibles et toutes hors gamme. Les concentrations des protéines Cp149 et Cp183, selon la méthode BCA, seraient de respectivement $\sim 0,7 \text{ mg.ml}^{-1}$ et de $\sim 0,5 \text{ mg.ml}^{-1}$.

8.2.3.3 Conclusion sur la quantification des protéines de capsid

Les concentrations de la protéine tronquée Cp149 et de la protéine mutée Cp183 sont montrées dans le tableau récapitulatif (Tableau 22).

Pour la méthode de Bradford, la pondération des arginines diminue beaucoup la concentration des protéines.

Si on ne tient pas compte de la pondération des arginines :

- la concentration de la protéine Cp149 est comprise entre $\sim 0,7$ et $\sim 1,3 \text{ mg.ml}^{-1}$,
- la concentration de la protéine Cp183 est comprise entre $\sim 0,3$ et $\sim 0,9 \text{ mg.ml}^{-1}$,

Sinon :

- la concentration de la protéine Cp183 est comprise entre $\sim 0,5$ et $\sim 0,9 \text{ mg.ml}^{-1}$.

Les concentrations déterminées selon la méthode BCA sont plus faibles que celles déterminées avec la méthode de Bradford ou la méthode de Porterfield & Zlotnick. D'ailleurs les concentrations non pondérées obtenues à l'aide de ces deux dernières semblent en accord ; selon ces méthodes :

- la concentration moyenne à la 1^{ère} campagne de Cp149 est de $\sim 1,2 \text{ mg.ml}^{-1}$,
- la concentration moyenne à la 2^{ème} campagne de Cp149 est de $\sim 1,1 \text{ mg.ml}^{-1}$,
- la concentration moyenne de Cp183 est de $\sim 0,9 \text{ mg.ml}^{-1}$.

Il faut garder à l'esprit que les méthodes de quantification utilisées sont caractérisées par des réactions chimiques ou des processus biophysiques différents. Les concentrations obtenues sont propres aux méthodes.

Méthodes de quantification	[Cp149] (mg.ml ⁻¹)	[Cp149] (µM)	[dimère Cp149] (µM)	[Cp183] (mg.ml ⁻¹)	[Cp183] (µM)	[dimère Cp183] (µM)
Porterfield & Zlotnick (260, 280, 340, 360 nm) 1 ^{ère} campagne	~1,30	~77,18	~38,59	NA	NA	NA
2 ^{ème} campagne	~1,08	~64,12	~32,06	~0,81	~38,36	~19,18
Absorbance (205 nm) 1 ^{ère} et 2 ^{ème} campagne	*	*	*	*	*	*
Bradford (595 nm) 1 ^{ère} campagne	~1,16	~68,87	~34,44	NA	NA	NA
1 ^{ère} campagne, arginines pondérées	~0,93	~55,21	~27,61	NA	NA	NA
2 ^{ème} campagne	~1,04	~61,74	~30,87	~0,89	~42,15	~21,07
2 ^{ème} campagne, arginines pondérées	~0,83	~49,27	~24,64	~0,29	~13,73	~6,87
BCA (562 nm) 1 ^{ère} campagne	~0,79	~46,90	~23,45	NA	NA	NA
2 ^{ème} campagne	~0,73	~43,34	~21,67	~0,46	~21,78	~10,89

Tableau 22 - Concentrations massiques et molaires de Cp149 et Cp183 déduites selon différentes méthodes. *, les concentrations pour l'absorbance à 205 nm sont aberrantes. NA, lors de la première campagne, l'échantillon de protéines de capsid Cp183 n'est pas utilisé. Les lignes grisées sont les méthodes qui n'ont pas fonctionnées ou pour lesquelles il y a un doute sur les concentrations estimées.

8.2.4 Quantification du zinc en interaction avec Core

Pour vérifier si Core interagit avec le zinc, 2 méthodes sont utilisées :

- La méthode colorimétrique basée sur l'absorbance du 4-(2-pyridylazo) resorcinol (PAR).
- L'analyse par spectrométrie de masse couplée à un plasma inductif (ICP-MS).

Le zinc, potentiellement en contact avec les protéines de capsid, est identifié et quantifié.

Des expériences préliminaires de colorimétrie, avec le PAR (principe : chapitre 8.2.4.1), révèlent que l'échantillon Cp183 contient une quantité considérable de zinc (Figure 97 et Tableau 23). Il y a jusqu'à ~25 µM de zinc dans l'échantillon. Les mesures sont effectuées au cours du temps, après avoir ajouté du PAR préparé dans un tampon contenant de l'urée.

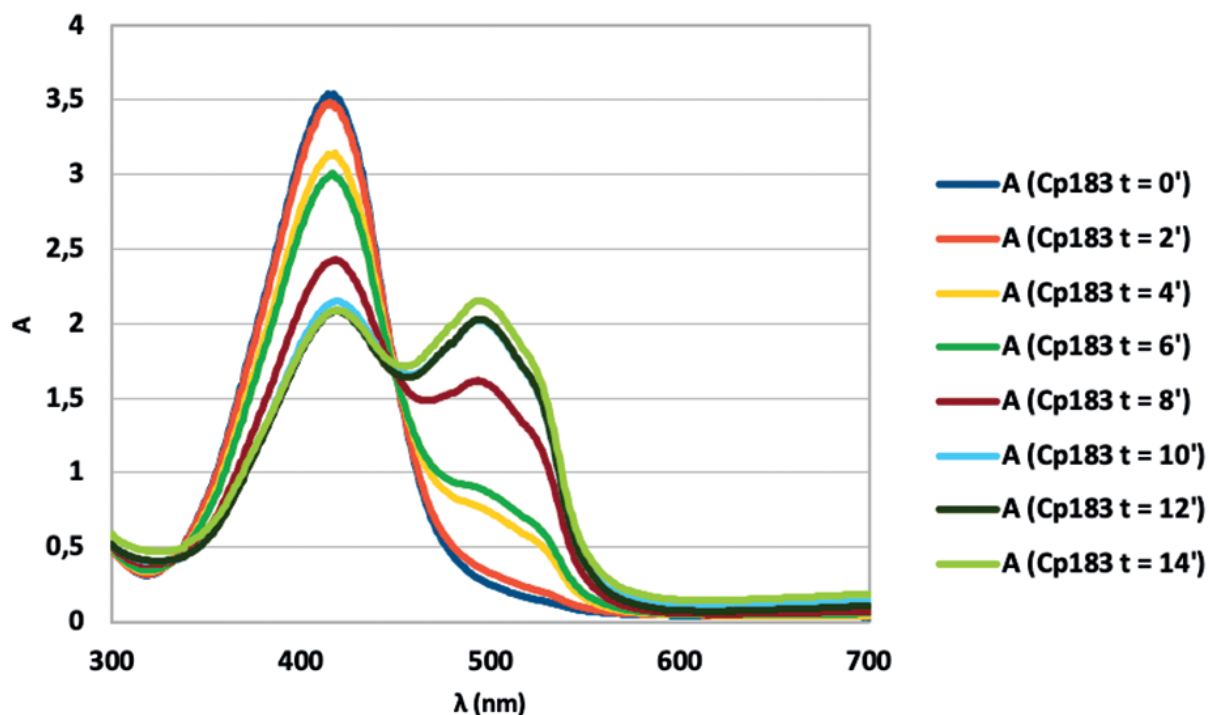


Figure 97 – Spectres de l'absorbance du PAR complexé avec le Zn^{2+} contenu dans l'échantillon Cp149.

Cp183 (min)	A (500 nm)	$[Zn^{2+}]$ lecture directe (μM)	$[Zn^{2+}]$ réelle (μM)
0	0,249	1,21	1,75
2	0,327	2,27	2,81
4	0,733	7,76	8,30
6	0,859	9,46	10,00
8	1,582	19,23	19,78
10	1,991	24,76	25,30
12	2,126	24,84	25,38
14	1,922	26,59	27,13

Tableau 23 - Mesures d'absorbance du PAR complexé et concentrations de zinc déduites.

Pour confirmer la présence d'ions zinc dans les échantillons, une quantification par ICP-MS (principe : chapitre 8.2.4.2) est réalisée à la suite (Tableau 24). Elle révèle, comme l'expérience de colorimétrie, que Cp149 contient du zinc mais que Cp183 en contient également. Elle montre que le zinc n'est pas issu du tampon de dialyse utilisé. Un tampon de dialyse contenant de l'EDTA met en évidence que les protéines de capsid interagissent fortement avec le zinc car des ions sont toujours quantifiés.

	Zn ²⁺ (ppb)	[Zn ²⁺] (µM)
Tampon de dialyse sans EDTA	<DL	<DL
Cp149	56,3	0,86
Cp183	138,3	2,12
Tampon de dialyse contenant de l'EDTA	<DL	<DL
8	12,6	0,19
10	15,4	0,24

Tableau 24 – Quantification du Zn²⁺ par ICP-MS.

Lors des 2 campagnes de mesures suivantes, la quantification du zinc est réalisée en parallèle de la quantification des protéines de capsid. Pour que les 2 méthodes soient comparables, les échantillons sont préparés dans les mêmes conditions. Ils sont minéralisés dans de l'HNO₃ à 2%.

8.2.4.1 Méthode colorimétrique de quantification du zinc - PAR

Une solution contenant du 4-(2-pyridylazo) resorcinol (PAR) forme en contact avec le zinc le complexe Zn(PAR)₂. Le PAR sous sa forme complexée absorbe au maximum à 500 nm. Le PAR seul a une absorbance maximale à 400 nm [120]. Selon ce procédé, des expériences de quantification du zinc dans nos échantillons ont été menées sous la co-supervision de Benoît d'Autreaux. À l'aide de gammes étalons de ZnSO₄, il est possible de déterminer la concentration de zinc dans nos échantillons (Figure 98).

Les spectres d'absorbance du PAR (Figure 98A et C) ont des points isobestiques (450 nm - 2 UA) sur la Figure 98A et (450 nm - 0,75 UA) sur la Figure 98C. Les deux espèces PAR complexé et PAR non complexé se croisent en ces points. Les 2 gammes étalons réalisées lors des deux campagnes ont un coefficient de détermination d'environ 0,99.

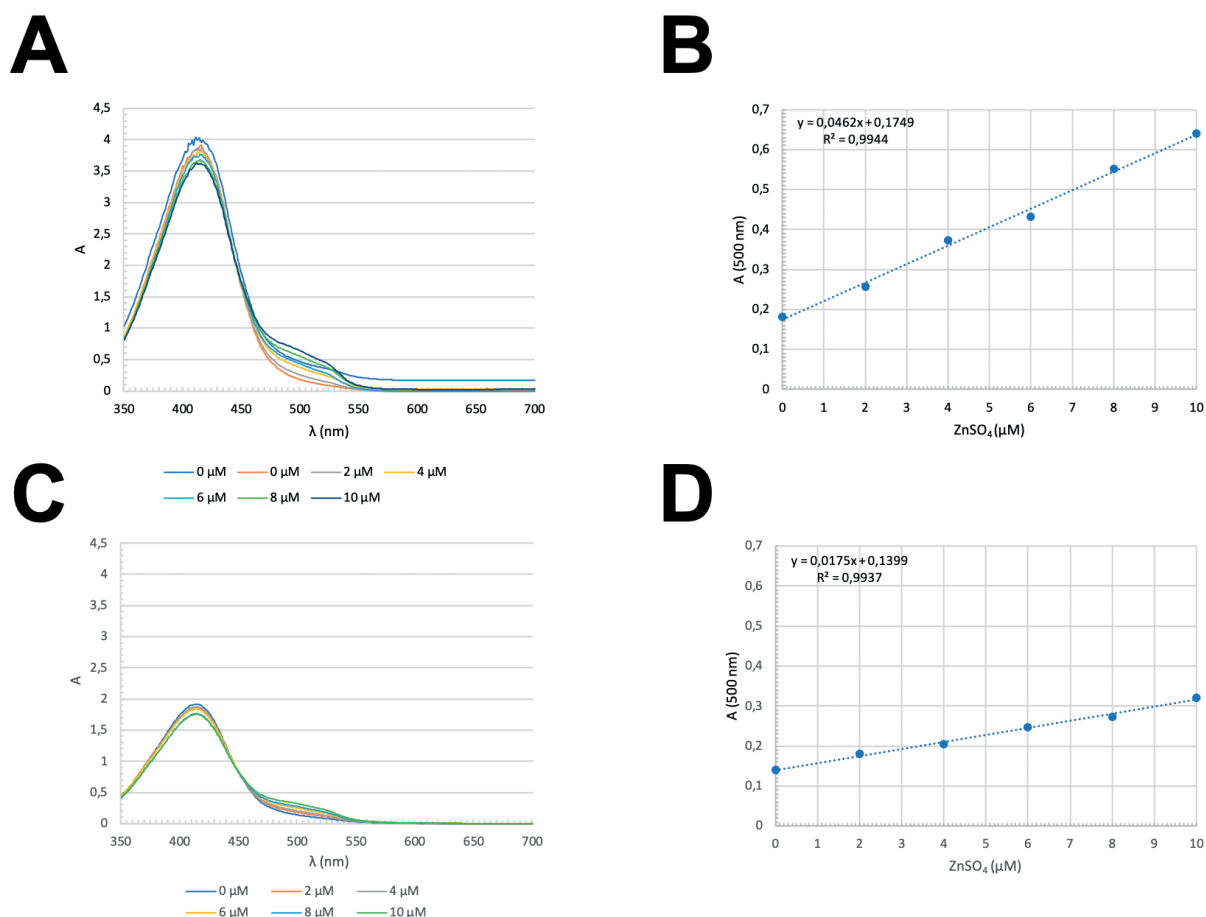


Figure 98 – Spectres et gammes d'étalonnages du PAR complexé ou non au Zn^{2+} provenant du $ZnSO_4$. A. et B. Première campagne de quantification du zinc sur l'échantillon Cp149. C. et D. Deuxième campagne de quantification du zinc sur les échantillons Cp149 et Cp183. A. et C. Spectres de l'absorbance du PAR. B. et D. Gammes d'étalonnages réalisées à partir des absorbances à 500 nm du PAR complexé au Zn^{2+} .

La pente de la gamme réalisée lors de la deuxième campagne (Figure 98D) a une pente 2,64 fois moins élevée que celle de la première campagne (Figure 98B). Ce qui aurait pu impacter la pente de la gamme d'étalonnage de la deuxième campagne sera abordé plus tard.

La concentration en zinc sur la gamme d'étalonnage peut être directement lu, sans se soucier de la concentration du PAR ; si et seulement si les échantillons sont préparés dans les mêmes conditions que les points de la gamme étalon. Nous sommes dans ce cas, il n'est donc pas nécessaire de corriger l'absorbance du PAR à 500 nm. Des spectres d'absorbance sont réalisés sur les échantillons dilués de la Cp149 et Cp183 (Figure 99).

Les spectres réalisés lors de la première campagne ont un point isobestique (Figure 99A). Les spectres comportent des oscillations au sommet du pic d'absorbance du PAR seul (bruit).

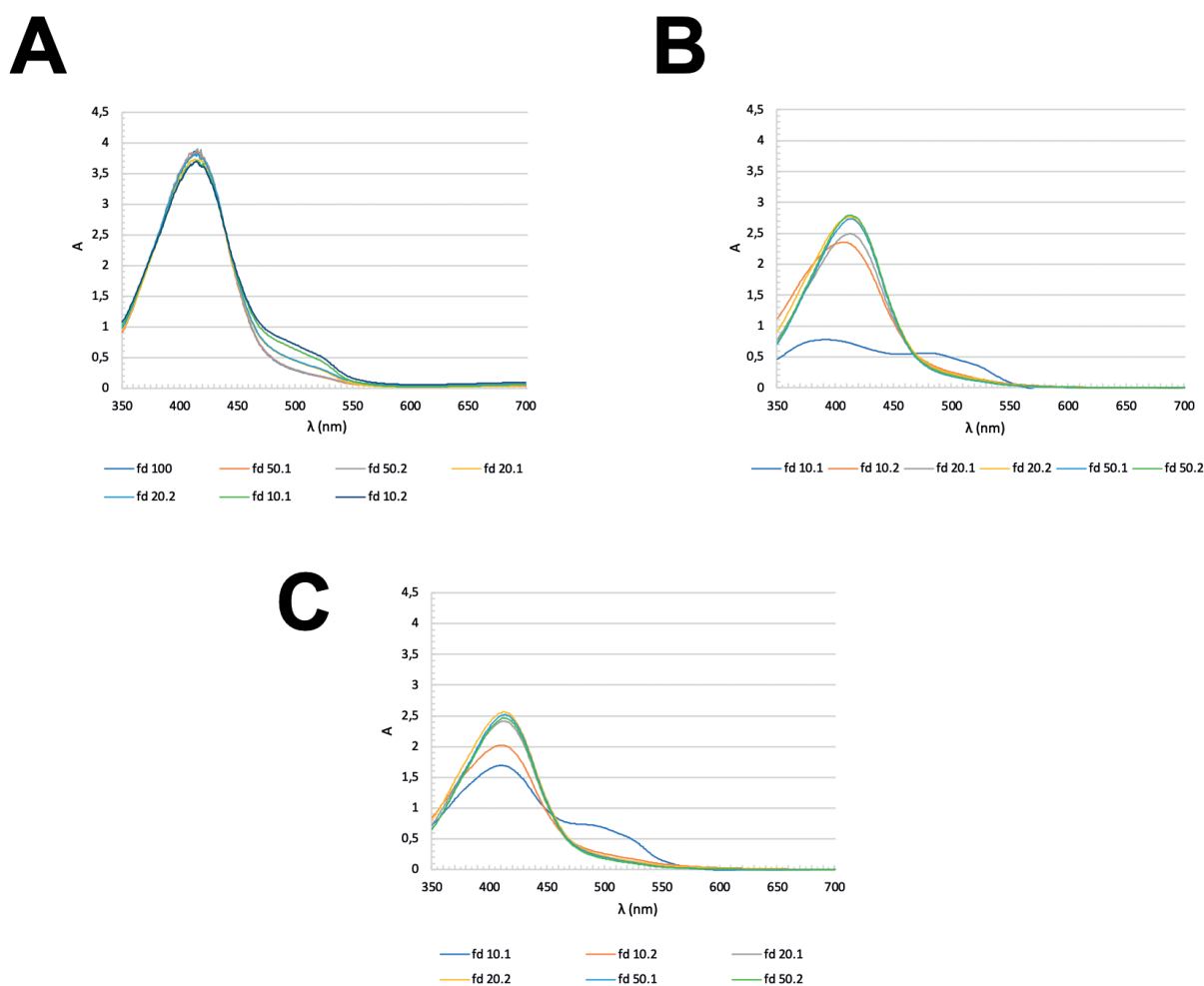


Figure 99 – Spectres de l’absorbance du PAR complexé ou non au Zn^{2+} contenu dans les échantillons.
 A. Première campagne de quantification du zinc sur l’échantillon Cp149. B et C. Deuxième campagne de quantification du zinc sur les échantillons Cp149 et Cp183. A et B Spectres d’absorbance du PAR en contact avec la protéine Cp149. C. Spectre de l’absorbance du PAR en contact avec la protéine Cp183. Dans les légendes “fd” correspond à facteur de dilution. “fd” est suivi du facteur de dilution entre 10 et 100 appliqué sur les échantillons.

Pour la deuxième campagne, le pic d’absorbance maximum du PAR non complexé (400 nm - PAR seul) se décale pour certaines dilutions d’échantillon Cp149 et Cp183 (Figure 99B et C respectivement). De plus, il n’y a pas de grande différence d’absorbance à 500 nm pour la plupart des dilutions.

Lors de la deuxième campagne, un problème a eu lieu lors des expériences de quantification du zinc avec le PAR. À la première campagne, le PAR a été préparé la veille des mesures. Pour la deuxième, le PAR a été utilisé le jour de sa préparation. C’est une molécule sensible dont la dissolution est difficile à mettre en œuvre. La concentration du PAR qui a été préparé à la deuxième campagne n’est sans doute pas correct car le PAR ne s’est pas solubilisé totalement.

L’absorbance du PAR à 500 nm est utilisée pour déterminer la concentration d’ions Zn^{2+} dans notre échantillon Cp149 (Tableau 25). À part la concentration calculée pour la dilution au 100^{ème}, les concentrations sont du même ordre de grandeur. La dilution à la première

ligne du tableau a été effectuée avec une pipette non adaptée au volume prélevé. Si on omet cette dilution, l'échantillon Cp149 lors de la première campagne a une concentration de $\sim 118,9 \mu\text{M}$ d'ions Zn^{2+} .

Cp149 (fd)	A (500 nm)	$[\text{Zn}^{2+}]$ lecture directe (μM)	$[\text{Zn}^{2+}]$ réelle (μM)
100*	0,289*	2,47*	$\sim 247,0^*$
50 (1)	0,296	2,62	$\sim 131,0$
50 (2)	0,298	2,67	$\sim 133,5$
20 (1)	0,444	5,83	$\sim 116,6$
20 (2)	0,448	5,91	$\sim 118,2$
10 (1)	0,631	9,87	$\sim 98,7$
10 (2)	0,708	11,54	$\sim 115,4$

Tableau 25 – Mesures d'absorbance du PAR complexé et concentrations calculées du zinc à la 1^{ère} campagne. * Erreur de dilution pour ce facteur de dilution (fd).

Le PAR préparé à la deuxième campagne n'est pas homogène (Tableau 26). Les absorbances à 500 nm sont donc probablement peu fiables. La concentration approximative en zinc de l'échantillon Cp149 serait de $151 \pm 17,5 \mu\text{M}$. En procédant de la même manière, la concentration imprécise de l'échantillon Cp183 serait de $135,25 \pm 29,75 \mu\text{M}$.

Cp149 (fd)	A (500 nm)	$[\text{Zn}^{2+}]$ lecture directe (μM) diluée	$[\text{Zn}^{2+}]$ (μM) corrigée pour la dilution
50 (1)	0,187	2,67	$\sim 133,5$
50 (2)	0,199	3,37	$\sim 168,5$
20 (1)	0,223	4,76	$\sim 95,2$
20 (2)	0,233	5,30	~ 106
10 (1)	0,487	19,78	$\sim 197,8$
10 (2)	0,253	6,46	$\sim 64,6$
Cp183 (fd)	A (500 nm)	$[\text{Zn}^{2+}]$ lecture directe (μM)	$[\text{Zn}^{2+}]$ réelle (μM)
50 (1)	0,198	3,31	$\sim 165,5$
50 (2)	0,177	2,10	$\sim 105,0$
20 (1)	0,206	3,75	$\sim 75,0$
20 (2)	0,218	4,43	$\sim 88,6$
10 (1)	0,674	30,46	$\sim 304,6$
10 (2)	0,256	6,63	$\sim 66,3$

Tableau 26 – Mesures d'absorbance du PAR complexé et concentrations calculées du zinc à la 2^{ème} campagne.

8.2.4.2 L'analyse par spectrométrie de masse couplée à un plasma inductif (ICP-MS)

L'ICP-MS est une méthode qui repose sur la séparation, l'identification et la quantification des éléments composant un échantillon en fonction de leur masse. Une torche plasma générant des ions. Ces ions sont séparés selon leur masse et leur charge par un spectromètre de masse quadripolaire afin de sélectionner très précisément les ions qui seront transmis au détecteur en fonction du rapport masse sur charge des ions (m/z). Le détecteur traduit le signal mesuré en nombre de coups. La calibration de l'appareil avec une solution étalon de l'ion d'intérêt ou de façon interne (spikes) pour convertir le nombre de coup en une concentration. L'incertitude des mesures effectuées par ICP-MS est inférieure à 3%.

Des expériences d'ICP-MS ont été réalisées sur nos échantillons par Valérie Geertsen (Tableau 27). Comme mentionné plus tôt, lors des 2 campagnes, les échantillons sont minéralisés au préalable dans une solution d' HNO_3 2%.

Échantillon mesuré Campagne n°1	Volume d'échantillon (μl)	Zn (ppb) (ng/g)	Zn (ng)	Zn réel (μM)
Cp149	50	5,87	617,90	~189,02
Échantillon mesuré Campagne n°2	Volume d'échantillon (μl)	Zn (ppb) (ng/g)	Zn (ng)	Zn réel (μM)
Cp149 50 mM Tris pH 7.5, 5% sucrose, 1 mM DTT (aliquot)	1000	7,20	763,80	~11,68
Cp183 50 mM Tris pH 7.5, 5% sucrose, 1 mM DTT (aliquot)	1000	9,92	1065,98	~16,30
Cp149, 50 mM Tris	300	6,54	706,74	~36,03
Cp183, 50 mM Tris	300	4,77	480,94	~24,52

Tableau 27 – Quantification du zinc à la 1^{ère} et 2^{ème} campagne de Cp149 et Cp183 E77K par ICP-MS.

Les résultats d'ICP-MS révèlent la présence de zinc dans tous les échantillons. La première campagne montre une quantité importante de zinc en contact avec la protéine Core recombinante tronquée (Cp149), ~189 μM . À la deuxième campagne, la concentration est de ~11,7 μM ou de ~36 μM . Le zinc contenu dans Cp183 a une concentration de l'ordre de 20 μM .

On s'attendait à ce que la quantité de zinc soit plus élevée pour les échantillons directement issus d'aliquots, car ils ne sont pas soumis à une étape de changement de tampon.

Le tampon contenu dans les aliquots a probablement impacté l'étape de minéralisation :

- Cp149 ; 50 mM Tris pH 7.5, 5% sucrose, 1mM DTT
- Cp183 ; 50 mM Tris pH 7.5, 5% sucrose, 1mM DTT

Le sucrose contenu dans le tampon s'est concentré et solidifié. Il interfère sans doute avec les mesures d'ICP-MS.

8.3 Conclusion sur la quantification du zinc en contact avec les protéines de capsid

Les méthodes Porterfield & Zlotnick et de Bradford non pondéré donnent les résultats de quantification de protéines de capsid les plus en accord. Les concentrations calculées avec ces 2 méthodes sont donc prises en compte pour la suite. La méthode de Porterfield & Zlotnick qui a été développée pour les protéines de capsid du VHB est la plus fiable.

Il est indéniable que qualitativement parlant, du zinc est présent dans les échantillons de la protéine Core du VHB. Le croisement des données de quantification approximatives de nos protéines et du zinc montre en général au moins un Zn^{2+} par dimère de Core (Tableaux 28 et 29).

Méthodes de quantification			1 ^{ère} campagne		2 ^{ème} campagne
			[Zn ²⁺] PAR (µM)	[Zn ²⁺] ICP-MS (µM)	[Zn ²⁺] ICP-MS (µM)
			~118,90	~189,02	~36,03
1 ^{ère} campagne	[dimère Cp149] Porterfield & Zlotnick 1 ^{ère} campagne (µM)	~38,59	3,08	4,90	
	[dimère Cp149] Bradford (595 nm) 1 ^{ère} campagne (µM)	~34,44	3,45	5,49	
2 ^{ème} campagne	[dimère Cp149] Porterfield & Zlotnick 2 ^{ème} campagne (µM)	~32,06			1,12
	[dimère Cp149] Bradford (595 nm) 2 ^{ème} campagne (µM)	~30,87			1,17

Tableau 28 – Ratio du zinc en fonction de la concentration de Core tronquée (Cp149). Les cellules grises claires correspondent à une estimation grossière du zinc lors de la 2^{ème} campagne avec la méthode colorimétrique de quantification au PAR.

La première campagne de quantification révèle qu'une protéine Core recombinante tronquée (Cp149) interagit avec plusieurs zincs, selon la méthode colorimétrique au PAR et l'ICP MS respectivement (Tableau 28). La deuxième campagne montre plutôt un zinc par dimère de Cp149.

Les résultats de la deuxième campagne montrent aussi environ un zinc par dimère de Core entière (Tableau 29).

Méthodes de quantification		[Zn ²⁺] ICP-MS (µM)
		~24,52
[dimère Cp183] Porterfield & Zlotnick (µM)	~19,18	1,28
[dimère Cp183] Bradford (595 nm) (µM)	~21,07	1,16

Tableau 29 – Ratio du zinc en fonction de la concentration de Core mutée (Cp183). Les cellules grises claires correspondent à une estimation grossière du zinc avec la méthode colorimétrique de quantification au PAR.

En conclusion et selon les résultats :

- Le tampon dans lequel se trouve la protéine Core est important lors de sa quantification ou de celle du zinc en interaction.
- Il y aurait environ un zinc par dimère de Core tronquée (Cp149) ou non (Cp183) dans des échantillons de capsid recombinante purifiée.
- 5 zincs, au maximum, peuvent interagir avec la protéine Core tronquée.

La protéine Core du VHB interagit donc avec le zinc. Le rôle que joue le zinc par rapport à Core et son intervention dans le cycle viral sont inconnus.

8.4 Chélation du Zn²⁺ par Core *in-silico*

En parallèle des expériences in-vitro de titration du zinc, la modélisation de Core en interaction avec du zinc a été effectuée. Cette étude computationnelle modélise les sites de chélation du zinc. Les contacts qui se forment entre les ions zinc et la protéine Core sont quantifiés. Ces expériences in-silico tiennent compte des données expérimentales.

8.4.1 Modélisation des sites de chélation de Core

En se servant des données structurales contenues dans la MetalPDB [121] et du champ de force Zinc AMBER (ZAFF) [122], les sites putatifs d'interaction de Core avec le zinc sont modélisés et simulés. Peters et al. se basent sur un champ de force classique (Amber) [109]. Ils génèrent les charges, ainsi que les constantes de force de liaison et d'angle, des sites métalliques tétravalents, ici pour le zinc. La modélisation repose sur les clusters d'histidines et de cystéines situés en dessous du NTD. 3 sites de chélation potentiels sont identifiés (①, ② et ③ Figure 100C). Les sites ① et ③ sont composés d'une cystéine (C107) et de 3 histidines (H51, H52 et H104). Ils correspondraient à un centre métallique de type Zn-CHHH (avec une cystéine déprotonée et 3 histidines neutres protonées sur le δ). Le site 2 est constitué de 2 cystéines (C48 des 2 monomères) et de 2 histidines (H47 des 2 monomères). Les résidus de ce dernier sont moins enfouis que ceux des 2 autres sites. Il correspondrait à un centre métallique de type Zn-CCHH (avec 2 cystéines déprotonées et 2 histidines neutres protonées sur le δ). Dans ce processus de modélisation, nous nous sommes principalement servis des structures 3D des 2 types de sites : Zn-CHHH et Zn-CCHH (PDB ID : 1CK7 et 1A1F respectivement) [123,124]. Le champ de force ZAFF [122] a été paramétré en utilisant en partie ces structures. Elles ont donc servi de "template" pour modéliser les sites de Core. Lors des étapes de modélisation, des liaisons sont établies entre les atomes SG des cystéines ou les atomes NE2 des histidines avec les ions zinc. Le modèle complet obtenu est ensuite simulé pendant 105 ns avec le champ de force ZAFF (Figure 100B). Cette simulation est réalisée sur un supercalculateur comportant des processeurs Intel(R) Xeon(R) CPU E5-2690 v4 de 2,60GHz et consomme 79 800 h. Le système solvaté et neutralisé comporte 598 142 atomes.

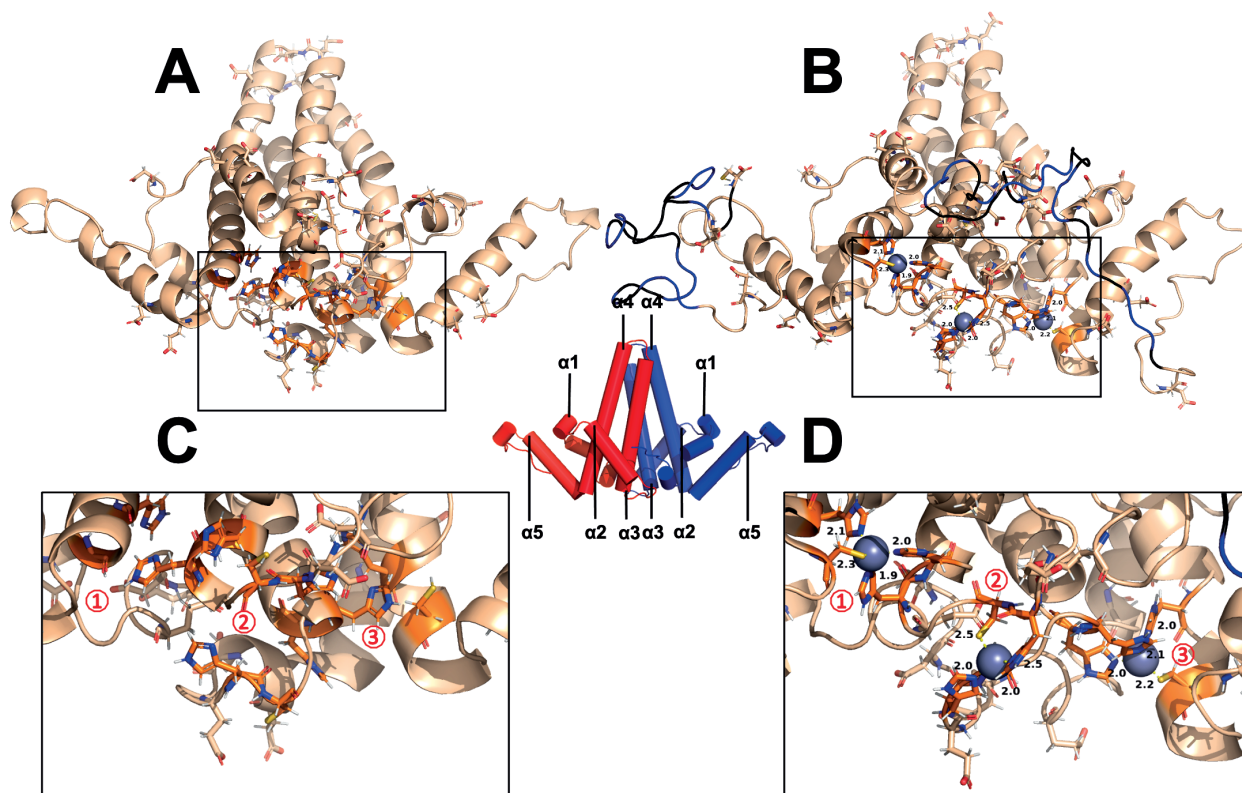


Figure 100 – Modèle de chélation du zinc par Core selon le champ de force ZAFF [122]. A. Dimère tronqué (sans CTD - Cp149). B. Dimère complet en interaction avec 3 ions zinc. Les régions composées d'arginines sont colorées en bleu. C. Zoom du dessous du NTD du dimère tronquée comportant 4 cystéines et 8 histidines. D. Cystéines et histidines du modèle complet en interaction avec 3 ions zinc. Les distances sont en angström. Les cystéines et histidines qui se trouvent en dessous du NTD sont colorées en orange, le NTD est représenté en beige et le CTD en noir. Les ions zinc sont représentés par des sphères de van der Waals en gris.

Le modèle simulé (Figure 100B) est superposé sur le NTD de Core cristallographique (Figure 100A) (PDB ID : 5E0I) [94], RMSD de 2,68 Å. La modélisation et la simulation des sites ont des effets sur la structure de Core. La dynamique des sites ① et ③ induit une déformation à la base des hélices $\alpha 4$ (Figure 100B et D). Les cystéines et histidines du site ② sont contenues dans des boucles. La conformation de ces boucles est très flexible et a également évolué (Figure 100C et D). Comme attendu, les zincs sont à une distance des cystéines et histidines comprise entre 1,9 et 2,5 Å tout au long de la simulation.

8.4.2 Interaction de Core avec le zinc libre

Des MD de dimères complets ou tronqués sont réalisées pour identifier les zones de contacts supplémentaires des ions zinc sur Core. Les simulations sont calculées par Amber20 [46] et selon le champ de force Amber 99SB-ILDN [109]. Pour chaque modèle, 105 ns de MD est produite. Une concentration de 20 mM de $ZnCl_2$ a été utilisée. En plus d'identifier et de quantifier les contacts entre les ions zinc et Core, nous avons également étudié l'effet de la déprotonation de toutes les cystéines situées en dessous du NTD (toutes sauf CYS61 et CYS183) sur l'interaction avec les ions zinc.

Acides aminés qui chélatent Zn ²⁺	Localisation	Durée d'interaction avec du Zn ²⁺ (%)	
		Cystéines protonées	Cystéines non protonées
D4(A), E14(A)	partie inférieure NTD ①	15,05	
E40(B), E43(B)	partie inférieure NTD ②	16,00	
E43(A), D2(B)	partie inférieure NTD ②	19,52	
D78(A), D78(B)	sommet spicule ③	27,62	
D78(A), E77(B)	sommet spicule ③	9,33	
D78(A), E77(A), D78(B)	sommet spicule ③	2,67	
E113(A), E117(A), E145(A)	base de l'hélice α5 + CTD ④	26,57	
E113(B), E117(B)	base de l'hélice α5 ④	12,48	
E40(A), E43(A)	partie inférieure NTD ②	11,81	1,71
E113(A), E117(A)	base de l'hélice α5 ④	63,52	1,14
E40(B), E46(B)	dessous NTD ⑤		1,14
E46(A), C48(A)	dessous NTD ⑤		3,24
E113(B), E145(B)	base de l'hélice α5 + CTD ④		2,29
E117(A), E145(A)	base de l'hélice α5 + CTD ④		4,19

Tableau 30 – Interaction du dimère tronqué (Cp149) avec le Zn²⁺. Dans la première colonne, les acides aminés qui interagissent avec un zinc sont référencés par le code à une lettre et la position du résidu en question, la chaîne du résidu est indiquée entre parenthèses. Dans la deuxième colonne, les contacts sont identifiés et localisés selon les zones définies Figure 101.

Les durées d'interaction des ions zincs avec la protéine Core tronquée ont été quantifiées avec CPPTRAJ [111] (Tableau 30). Dans ces modèles, Core interagit davantage avec les ions zinc lorsque les cystéines situées en dessous du NTD sont protonées. Ces résultats ne sont pas attendus. Les ions zincs interagissent avec les glutamates et aspartates au sommet de la spicule du NTD (③ sur la figure 101) pendant plus de 26% des 105 ns. Les glutamates et aspartates localisées en région N-terminale sont capable de fixer du zinc pendant plus de 20% des 105 ns (① et ② sur la figure 101). Enfin les glutamates aux positions 113 et 117, situées sur la partie inférieure du NTD sont énormément mis à contribution et établissent plusieurs formes de contacts avec le zinc. Pour plus de 63% des 105 ns les E113 et E117 interviennent dans un contact avec du zinc (④ et ⑤ sur la figure 101). Seuls les modèles dont les cystéines sont déprotonnées font apparaître des contacts entre une des cystéine (C48) et des ions zinc (3,24% des 105 ns).

Acides aminés qui chélatent Zn ²⁺	Localisation	Durée d'interaction avec du Zn ²⁺ (%)	
		Cystéines protonées	Cystéines non protonées
E77(B), D78(B)	sommet spicule ③	1,03	
E113(A), E117(A)	base hélice α5 ④	82,06	
E113(A), E46(B)	base hélice α5 + dessous NTD ⑤		2,40

Tableau 31 – Interaction du dimère complet (Cp183) avec le Zn²⁺.

La protéine Core complète forme moins d'interaction que la protéine Core tronquée (Tableau 31). Ce résultat est en accord avec les données expérimentales. Les glutamates 113 et 117 sont retrouvés dans des contacts avec les ions zinc (④ sur la figure 101). Ils sont responsables de plus de 82% des contacts au cours des 105 ns.

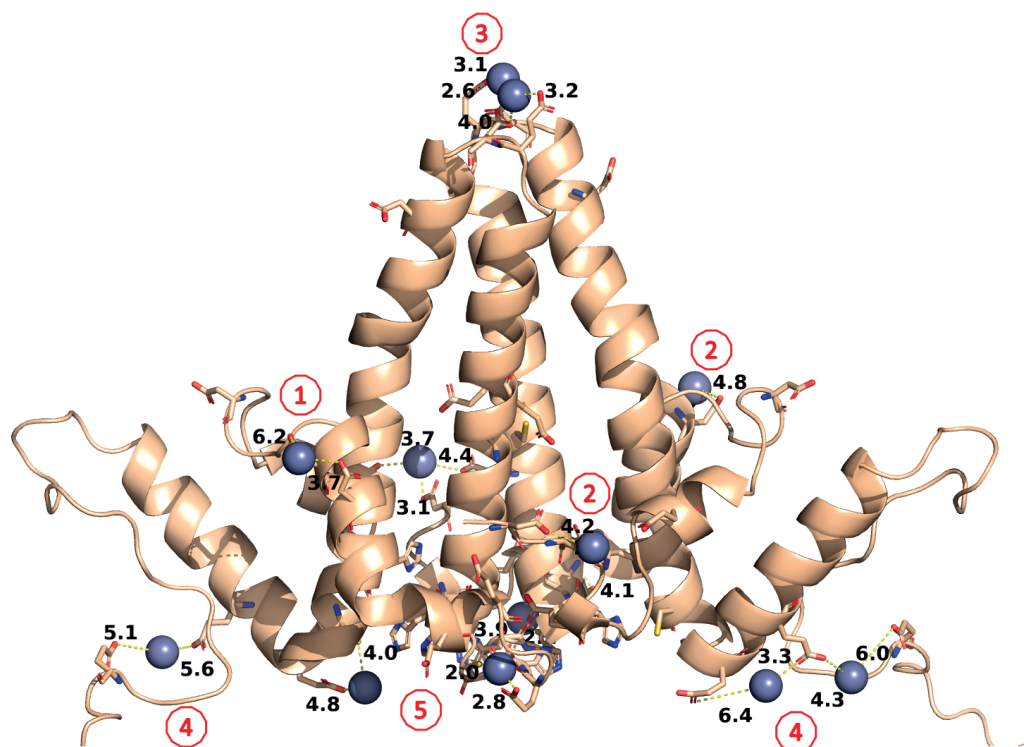
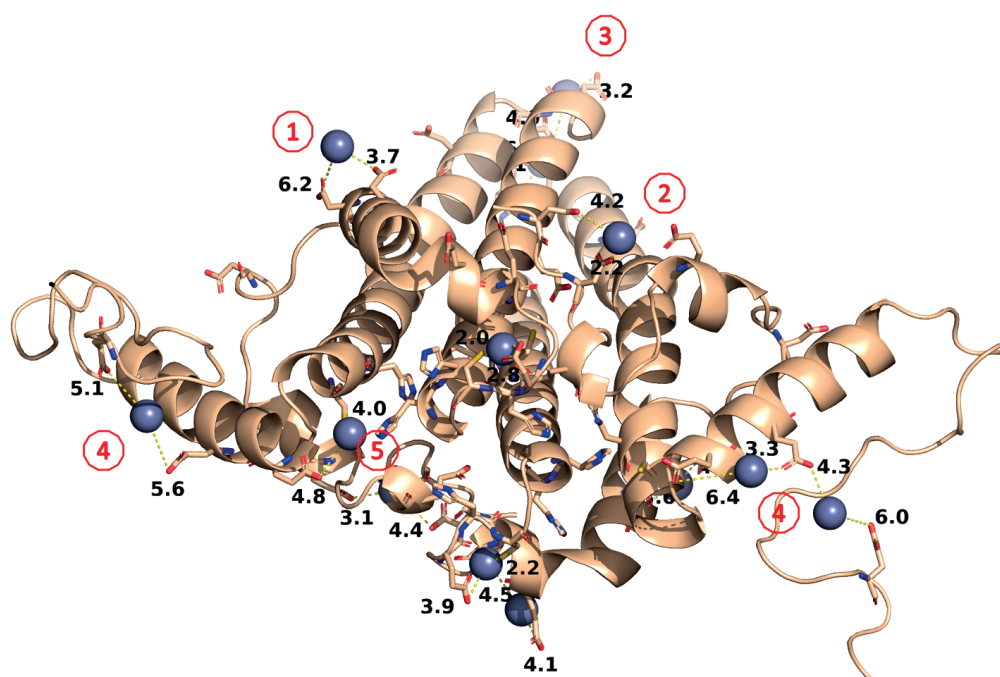
A**B**

Figure 101 – Localisation des points de contacts des ions zincs. A. Vue de face du dimère complet. B. Vue du dessous du dimère complet. Les parties des CTD qui n'interagissent pas avec le zinc ont été tronquées sur la figure. Les zones de contacts sont numérotées de 1 à 6 et correspondent aux données des tableaux 30 et 31. Les distances sont en angström.

De nombreux contacts entre les cystéines, histidines (en dessous du NTD) et les ions zinc étaient attendus. Ce n'est pas le cas mais il y a tout de même des zones sous le NTD ④ et ⑤ sur la figure 101) qui interagissent avec du zinc. Ces régions pourraient jouer un rôle lors de la fixation du zinc. Elles pourraient constituer des régions où les ions zinc

sont transitoirement fixés puis acheminés vers les cystéines et histidines à fort potentiel de chélation. Les temps de résidence des MD avec le champ de force classique (Amber 99SB-ILDN) [109] identifient les acides aminés de Core en interaction avec les ions zinc.

8.4.3 Accessibilité des sites putatifs

Une analyse de l'accessibilité des sites putatifs est réalisée pour déterminer pourquoi Core a davantage tendance à fixer, expérimentalement, un zinc plutôt que 3. L'accessibilité des résidus impliqués dans la chélation du zinc sous le NTD est donc mesurée au cours de 3 simulations : une MD de Core tronquée en présence de 20 mM de $ZnCl_2$ (28 Zn^{2+} dans la boîte d'eau) (Figure 102A) ; une MD de Core complète en présence de 20 mM de $ZnCl_2$ (28 Zn^{2+} dans la boîte d'eau) (Figure 102B) et une MD de Core complète comportant 3 zincs liés simulée selon le champ de force ZAFF [122] (Figure 102C).

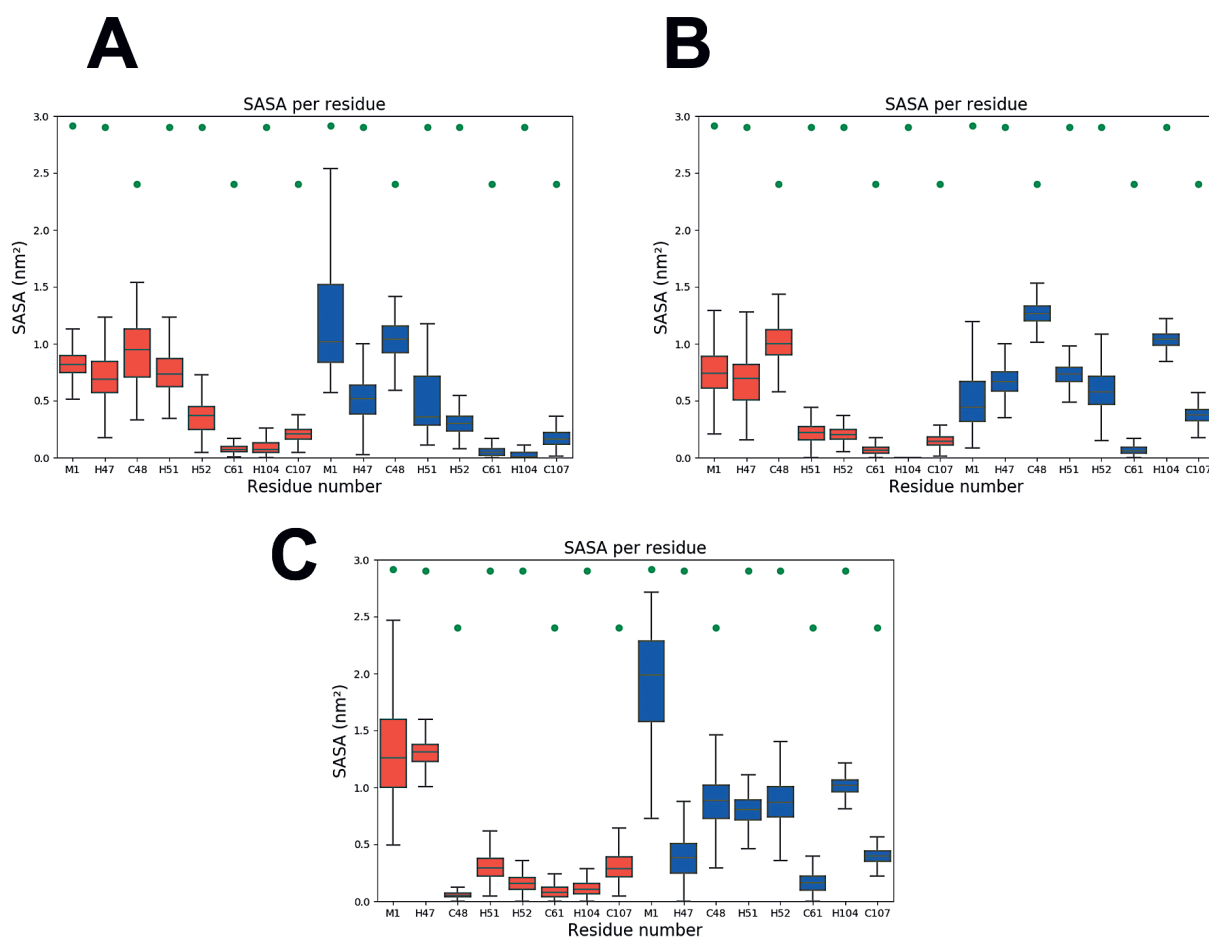


Figure 102 – Surface accessible au solvant (SASA) des résidus impliqués dans la chélation des ions zincs. Les SASA des résidus sont représentées dans des boîtes à moustache. Les boîtes à moustaches comportent la médiane des SASA pour un résidu ; le premier et troisième quartile aux extrémités de la boîte ; et les valeurs extrêmes en dehors de la boîte. Les boîtes à moustaches colorées en rouge correspondent à la chaîne A. Les boîtes à moustaches colorées en bleu correspondent à la chaîne B. Les points vert correspondent aux valeurs de SASA des acides aminés complètement exposés [125]. A. SASA par résidu pour la simulation du dimère tronqué, cystéines déprotonées (sans CTD). B. SASA par résidu pour la simulation du dimère complet, cystéines déprotonées C. SASA par résidu pour la simulation du dimère complet modélisé et simulé selon le ZAFF [122].

Dans les trois graphiques, les surfaces accessibles au solvant (SASA) de la méthionine en N-terminal (M1) et de la cystéine 61 (C61) sont calculées. Elles nous servent de référence. M1 est très accessible, tandis que C61 est très enfouie. Les SASA des résidus sont représentées dans des boîtes à moustache. De manière générale, les histidines et cystéines du site ② (H47 et C48) ont une SASA plus élevée, sauf pour Core selon le champ de force ZAFF. Les médianes des boîtes à moustaches de ces résidus sont comprises entre $\sim 0,75$ et ~ 1 nm². Les histidines et cystéines des sites ① et ③ (H51, H52, H104 et C107) sont plus enfouies. C'est surtout le cas pour H104 et C107 dont la médiane se situe en moyenne en dessous de $0,5$ nm², sauf pour Core selon le champ de force ZAFF. Les SASA des résidus de la chaîne B des dimères complet sont plus élevées que celles de la chaîne A (Figure 102B et C en bleu). Elles sont comprises entre $\sim 0,30$ et $\sim 1,30$ nm². La présence des 3 zincs dans le modèle de Core modélisé et simulé selon le champ de force ZAFF réduit considérablement l'accessibilité des résidus de la chaîne A (Figure 102C en rouge), la plupart des médianes sont en dessous de $0,5$ nm².

Cela montre que lorsqu'on prend les SASA des résidus 2 à 2 (chaîne contre chaîne), les chaînes A et B de Core tronqué ont une accessibilité quasi équivalente. Pour les protéines Core complète, les résidus de la chaîne B ont une accessibilité en moyenne plus élevée que la chaîne A. Le site ② est plus accessible que les sites ① et ③. Il constitue un bon site de chélation du zinc.

8.5 Conclusion

Les expériences de quantification in-vitro indiquent que Core est capable de chélater au moins un zinc. Core est donc qualitativement à même de fixer du Zn^{2+} . In-silico les nombreux glutamates et aspartates qui constituent, en partie la protéine Core, sont capables d'interagir transitoirement avec les ions zincs. Les glutamates (E40, E46, E113 et E117) situés à proximité des cystéines et histidines, et en dessous du NTD pourraient participer à la fixation des ions zinc sur les sites putatifs.

Dans le modèle de Core en interaction avec 3 ions zincs, les Zn^{2+} fixés sur les sites induisent une réorganisation du dessous du NTD (base des hélices $\alpha 4$ et boucles) au cours de la MD selon le champ de force ZAFF [122]. Le site au milieu de l'interface dimérique est le plus accessible. Il est donc plus probable que ce soit celui-ci qui soit mis en jeu dans l'interaction de Core avec le Zn^{2+} identifié expérimentalement.

La question est de savoir si le zinc a un rôle structural sur la protéine Core. Provoque-t-il le repliement du CTD pour former un motif d'interaction avec l'ADN ?

Les arginines et le zinc pourraient entrer en compétition pour les résidus chargés négativement sur Core : E40, E46 ...

9 DISCUSSION

L'étude du CTD, dans le contexte du dimère libre (Cp183), montre des propriétés intéressantes : (1) La partie terminale (7 derniers résidus) a tendance à se replier sur les domaines riches en arginine (ARD) du CTD, de sorte que le CTD comporte généralement une boucle. (2) Il forme plus de contacts intra-CTD qu'inter-domaines pour un même monomère. (3) Dans le cas des contacts inter-domaines, il interagit principalement sur le côté du NTD, défini comme la région allant de la base de l'hélice $\alpha 2$ à $\alpha 5$. Il peut aussi se fixer sur la partie supérieure de la spicule, le sommet des hélices $\alpha 3$ et $\alpha 4$. Lors de ces repliements, le CTD interagit avec les résidus chargés négativement du NTD (Figure 103).

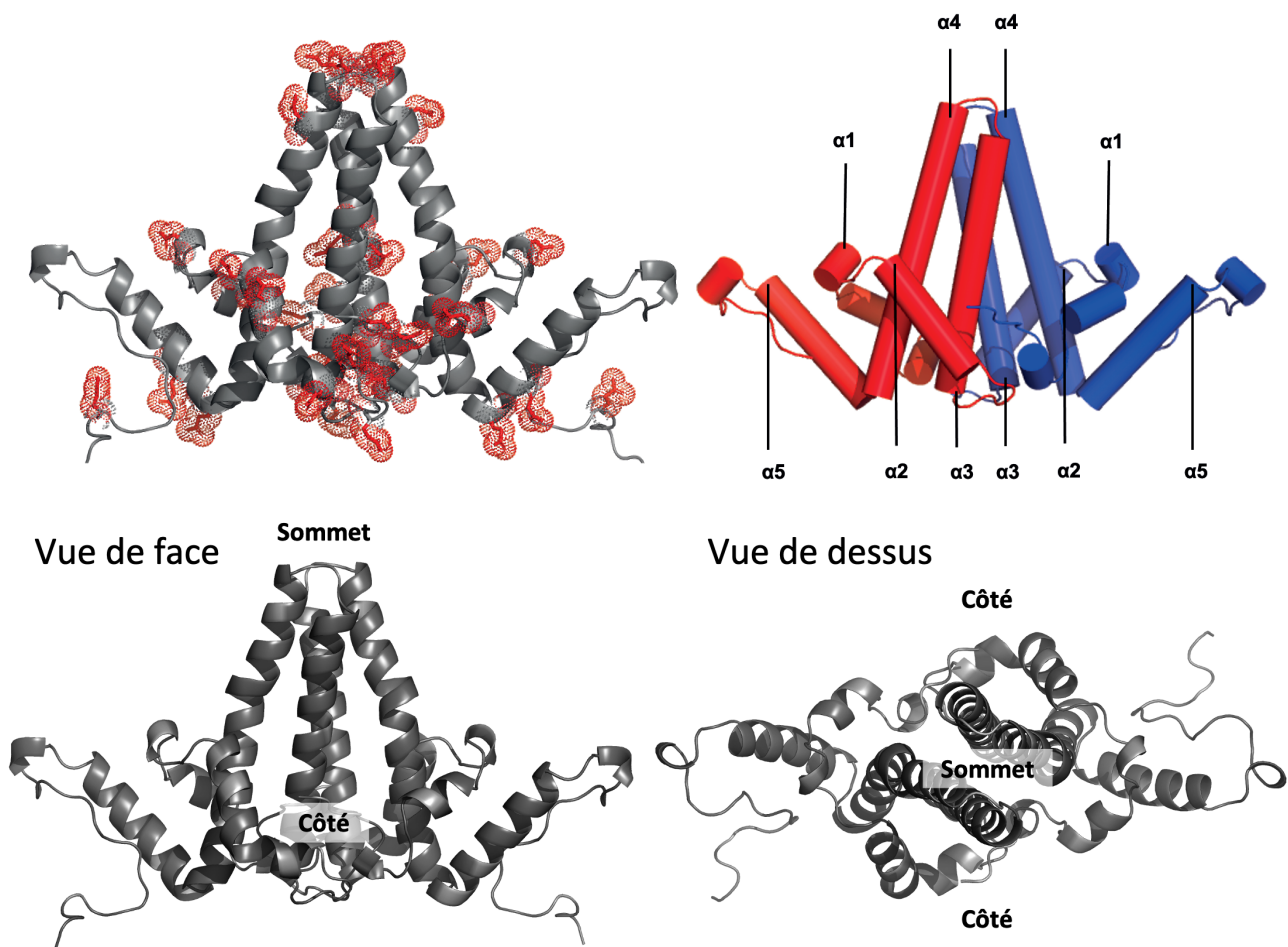


Figure 103 – Localisation des résidus chargés négativement sur le NTD de Core. Les chaînes latérales des aspartates et glutamates sont colorées en rouge. Ils sont représentés en bâtonnet avec des sphères de van der Waals des atomes des chaînes latérales figurent. Les positions des hélices sont montrées à droite, en mode cylindre. La chaîne A de Core est colorée en rouge et la B, en bleu.

Pour plus de 50% des MD, au moins l'un des 2 bras C-terminaux est localisé préférentiellement sur l'un des 2 côtés d'un dimère. Ce CTD est alors localisé en dessous d'un pore 3 ou q3 pour être exposé (Figure 74). Il est intéressant de noter que les conformations des dimères libres sont toutes favorables pour la formation de la capsid. Durant celle-ci, la partie terminale du CTD pourrait être exposée à l'extérieur de la capsid. Les repliements préférentiels des

bras C-terminaux déphosphorylés sont compatibles avec une sortie par q3. Ils suggèrent comment les signaux de localisation nucléaire pourraient se trouver exposés en dehors de la capsid au niveau des pores 3 et q3. La composition de l'intérieur des pores 3 et q3 est, globalement, chargée négativement (Figures 104 et 105).

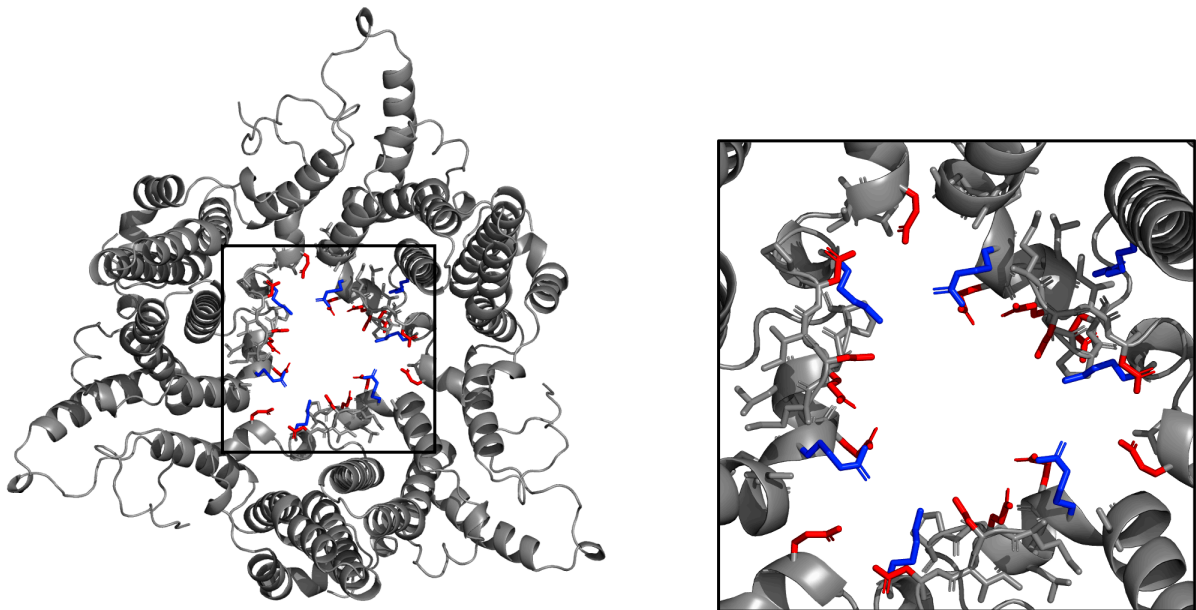


Figure 104 – Composition en acides aminés de l'intérieur du pore 3. Les résidus chargés négativement sont colorés en rouge et les résidus chargés positivement, en bleu.

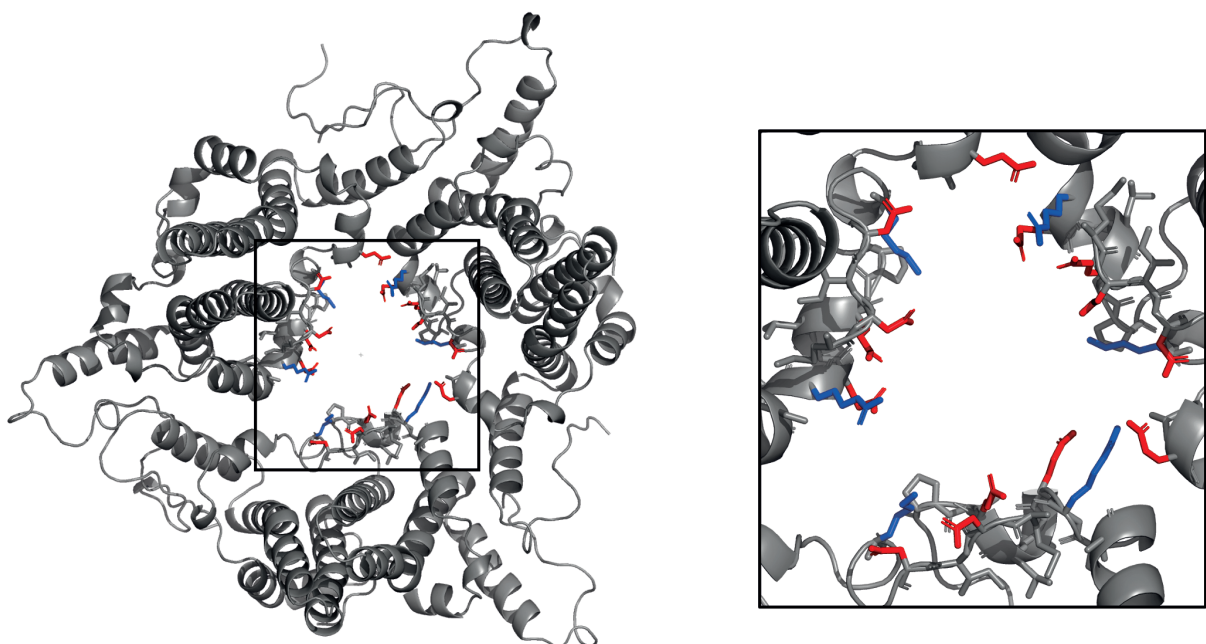


Figure 105 – Composition en acides aminés de l'intérieur du pore q3. Les résidus chargés négativement sont colorés en rouge et les résidus chargés positivement, en bleu.

En effet, l'intérieur des pores comporte 6 résidus chargés positivement (K7 et R39) et 14 résidus chargés négativement (notamment : D2, E14, E40, E43 et E46) pour le pore 3 (charge globale = -10 ; Figure 104) et q3 (charge globale = -8 ; Figure 105) [69]. Les CTD, globalement

chargés positivement (charge globale = 14), seraient donc attirés spontanément en dessous des pores 3 et q3. La chaîne latérale d'une arginine mesure entre ~5 et ~9 Å de longueur. Cette gamme de longueur est estimée à l'aide de PyMOL, à partir de la mesure de toutes les arginines des modèles de pores. Près de la moitié du CTD est composé d'arginines (16/34), résidu le plus grand du CTD. Étant donné les dimensions des pores (Figure 106), d'un point de vue stérique, une arginine peut passer au travers (3 et q3). Dans cette figure, les dimensions des pores sont surestimées car elles ne tiennent pas compte du van der Waals. Wynne et al. montrent que les pores mesurés, en réalité, 14 Å de diamètre [69].

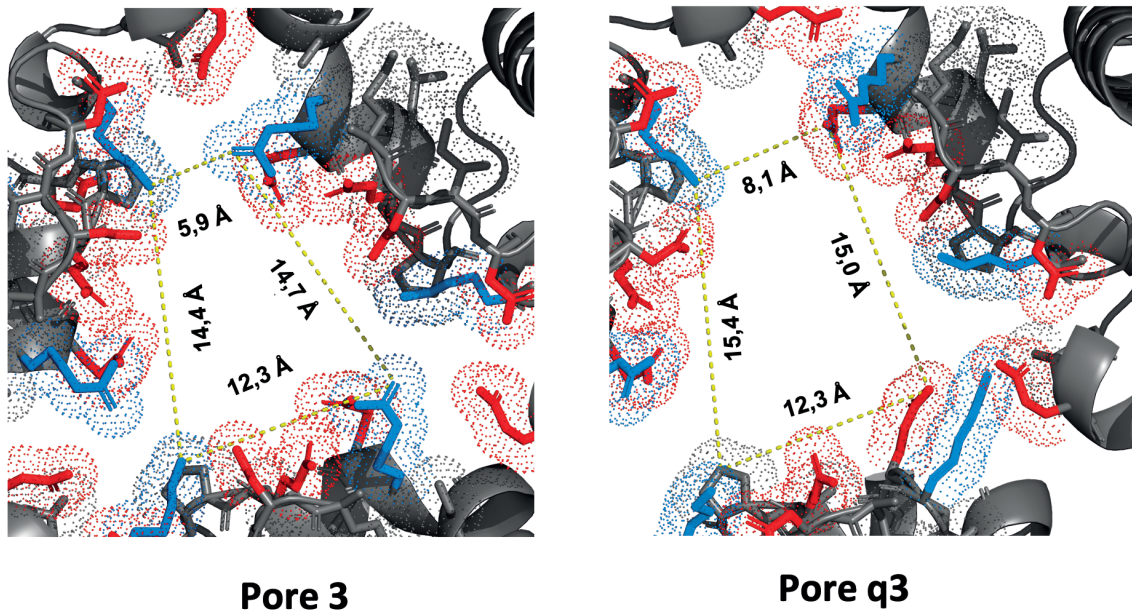


Figure 106 – Dimensions des pores 3 et q3. Les dimensions sont calculées avec PyMOL. Les résidus, à l'intérieur du pore, sont représentés en bâtonnet. Les acides aminés chargés négativement sont en rouge et les acides aminés chargés positivement, en bleu. Les arginines, en bleu, sont plus petites.

Si l'on ignore l'électrostatique, le "CTD-allongé" pourrait facilement traverser le pore. Les TMD du passage du "CTD-allongé", à travers les pores 3 et q3, le confirment (Figure 107). Lors du processus d'exposition d'un bras C terminal d'un des dimères du pore, des ponts-salins se forment entre celui-ci et l'intérieur du pore.

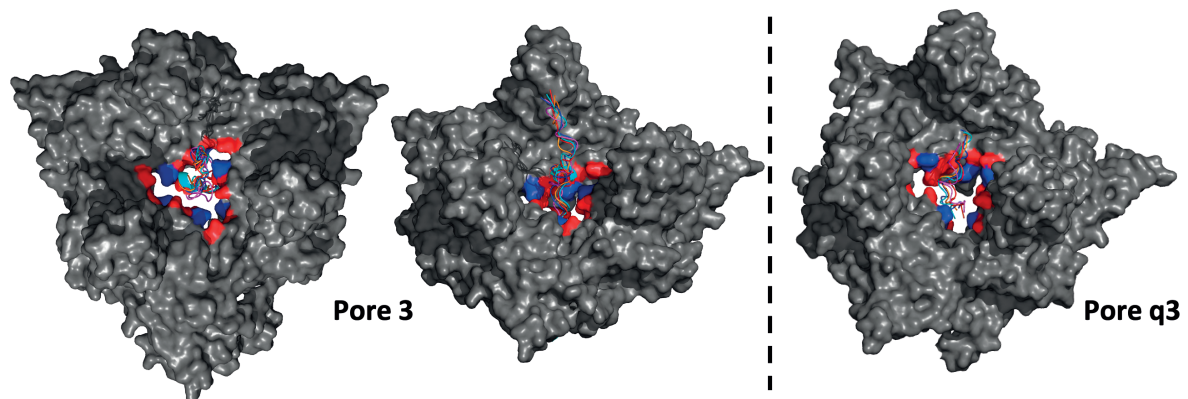


Figure 107 – Exposition du bras C-terminal allongé et non rattaché aux pores 3 et q3. Le pore est représenté en surface. Les acides aminés, chargés négativement, sont en rouge et les acides aminés chargés positivement, en bleu. Le bras C-terminal est représenté en "cartoon".

Les MD biaisées (Figures 107 et 108) montrent qu'un bras C-terminal est capable de passer à travers les pores (3 et q3) dans le contexte de la capside et, par conséquent, d'exposer les signaux d'importation nucléaire. Les CTD dépliés, dans le modèle de Core libre, ont des dynamiques intéressantes. Lorsqu'elles sont superposées sur l'un des dimères A-B ou C-D d'une capside Cp149, les conformations sont compatibles avec le passage. Les conformations d'un des 2 bras C-terminaux clash peu avec l'intérieur du pore q3. Les conformations finales des MD biaisées, dont le bras C-terminal se situe sur l'un des dimères du pore, sont tout à fait compatibles avec l'exposition du CTD. À l'issue de ces simulations, le domaine riche en arginines (ARDIV - résidus 172 à 175) et le ARDIII (résidus 164 à 167), sont exposés sous forme dépliée ou sous forme de boucle.

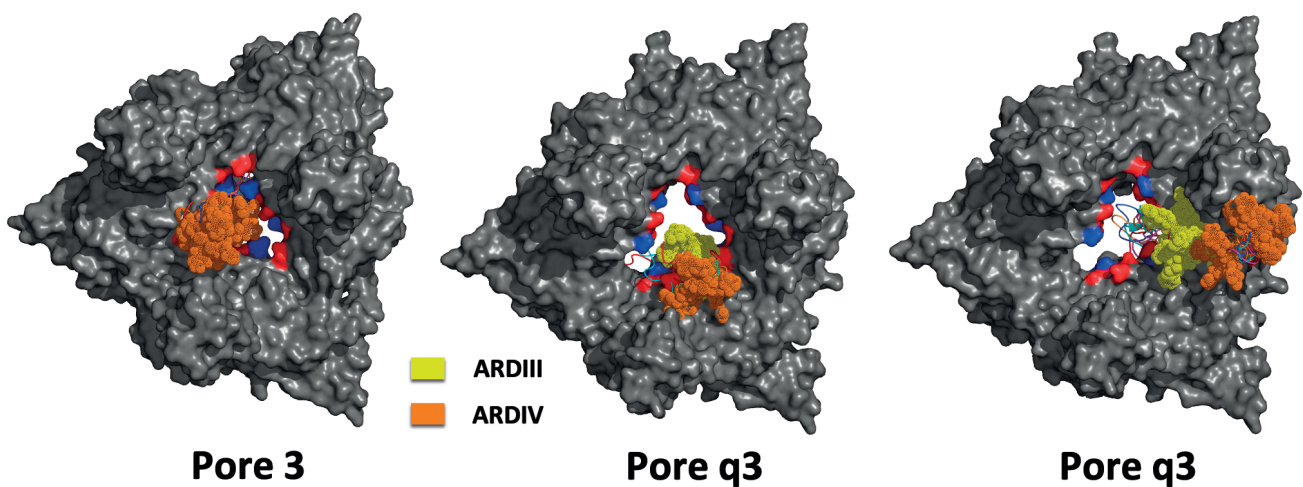


Figure 108 – Exposition des domaines riches en arginines (ARD). Au milieu, les CTD issus des TMD (bras C-terminal d'un des dimères A-B ou C-D du pore 3 ou q3) dont la constante harmonique (k) est égale à : 0,04 (en rouge) ou 0,06 (en cyan), sont représentés. Pour les 2 autres, tous les CTD sont représentés de $k = 0,02 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ à $k = 0,10 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ par pas de 0,02.

Le ARDIII est d'autant plus exposé lorsque le CTD est déplié. Il s'agit d'un signal de localisation nucléaire, tandis que le ARDIV correspond à un signal d'exportation du noyau [81]. Il est essentiel que le ARDIII soit exposé pour que le transport vers le noyau ait lieu. Les simulations sont en bon accord avec les expériences de protéolyse à la trypsine des capsides Cp183. Les résultats rapportés, notamment par Heger-Stevic et al. et Selzer et al. [82,83], ne sont pas complètement comparables. Néanmoins, il est clair que dans les capsides Cp183, une partie des CTD est clivée lentement, donc est accessible à la protéase au moins transitoirement et que ces phénomènes sont fortement influencés par l'état de phosphorylation des CTD. Dans les capsides Cp183 non phosphorylées, les coupures peuvent s'étendre au moins jusqu'aux arginines 157-159, c'est-à-dire l'ARDII (Figure 57), voire pour certains auteurs jusqu'à l'ARDI. Dans les TMD réalisées avec comme cible, un CTD sur la spicule, la plupart des CTD sont exposés à partir des résidus 159 ou 162 et l'ARDIII est très accessible au-dessus du pore. Ces simulations correspondent très bien à

l'une des interprétations d'Heger-Stevic et al. (Figure 58, gauche, modèle "extrude") [82]. Dans les simulations où le CTD vient sur un côté, une boucle interne du CTD sortirait par le pore 3 ou q3 (Figure 83), comme dans l'autre interprétation possible d'Heger-Stevic et al. (Figure 58, droite, modèle "loop out") [80]. Cependant, ces conformations 'CTD sur le côté', obtenues par simulation d'un dimère isolé, exposeraient dans le contexte de la capsid des résidus autour de S170, soit entre ARDIII et ARDIV. Les résultats des simulations sont aussi en accord avec la densité observée par cryo-microscopie électronique de capsides Cp183, au niveau des pores q3 [114]. Par contre, d'autres travaux de cryo-microscopie électronique de capsides Cp183 en complexe avec l'importine β , concluent à une exposition des CTD au niveau des axes quasi-6 (axes 2, Figure 83) [85].

L'étude de simulation de la capsid Cp149 de Hadden et al. a montré que la capsid Cp149 transfère le sodium à travers sa surface, plus rapidement que le chlorure [27]. Les auteurs suggèrent que, cette capacité à filtrer sélectivement les charges, pourrait faciliter l'exposition du CTD à travers les pores. [27]. Des ponts-salins se forment, en majorité, entre les ARD du CTD et les résidus chargés négativement des pores 3 et q3 durant toutes les TMD (Tableaux 32 et 33).

Acide aminé (intérieur du pore)	Domaine riche en arginines	Durée de contact maximum parmi toutes les TMD (en %)
E40	III	12,25
E40	IV	38
E43	IV	20,75
E46	III	20,25

Tableau 32 – Ponts-salins entre le pore 3 et les domaines riches en arginines du CTD. Les durées sont calculées à partir de toutes les TMD, dont la constante harmonique (k) est égale à 0,01. Il s'agit de la plus petite constante harmonique utilisée.

Acide aminé (intérieur du pore)	Domaine riche en arginines	Durée de contact maximum parmi toutes les TMD (en %)
D2	IV	54,25
E14	IV	20,5
E40	III	25,5
E40	IV	46,75
E43	IV	7,75
E46	III	67
E46	IV	74,5

Tableau 33 – Ponts-salins entre le pore q3 et les domaines riches en arginines du CTD. Les durées sont calculées à partir de toutes les TMD, dont la constante harmonique (k) est égale à 0,01.

Les CTD interagissent majoritairement avec quelques résidus du pore (E40, E43, E46...). Une nouvelle hypothèse est formulée, à partir de ces résultats, sur l'exposition du CTD. Un

échange de ponts-salins permettrait la progression du CTD au travers du pore (Figure 109). Des ions chargés positivement favoriseraient les échanges.

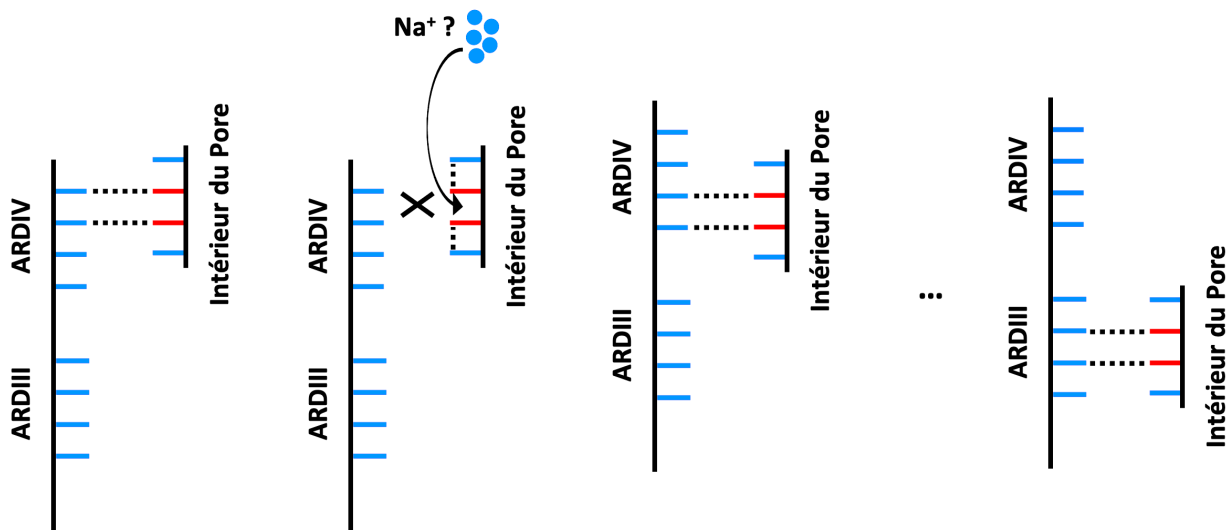


Figure 109 – Hypothèse du mécanisme d'exposition du CTD. Les bâtonnets bleus correspondent à des arginines et les rouges, à des résidus chargés négativement. Les pointillés symbolisent les ponts-salins.

Ces résultats confirment que les bras C-terminaux peuvent passer au travers des pores 3 et q3 et exposer les signaux de localisation. Pour rappel, dans plus de la moitié des MD de Core libre en solution, la cystéine terminale (C183) et le glutamate avoisinant (E180) forment des ponts-salins avec l'un des ARD. De cette manière, les charges négatives du CTD sont écrantées.

Les travaux *in vitro* de Core, vis-à-vis du Zn^{2+} , révèlent indéniablement l'association de zinc aux capsides recombinantes Cp183 et Cp149. L'étude *in-silico* localise des sites de chélation. L'alignement multiple des séquences protéiques de Core (<https://hbvdb.lyon.inserm.fr/HBVdb> - Figure 94) montre que tous les génotypes conservent strictement les cystéines et histidines (H47, C48, H52, H104 et C107) de ces sites putatifs. Ce n'est pas le cas du résidu H51, qui peut être remplacé par une asparagine. Ces sites sont localisés à la base de la protéine et sont, par conséquent, accessibles sur la face interne de la capside du VHB (Figure 110). L'étude de Kefalakes et al., sur les séquences du génotype A et D, montre que les cystéines et les histidines impliquées dans des sites de chélation sont très conservées tout au long de l'évolution [126]. Warner et al., dans le cadre d'une étude sur l'hépatite chronique B négative à l'antigène HBeAg, montrent également qu'elles ne sont pas mutées [127]. Ces résidus hautement conservés seraient alors importants pour le rôle de Core dans le cycle viral du VHB, au cours duquel l'interaction avec le zinc pourrait être essentielle (interaction avec le matériel génétique viral ou de l'hôte).

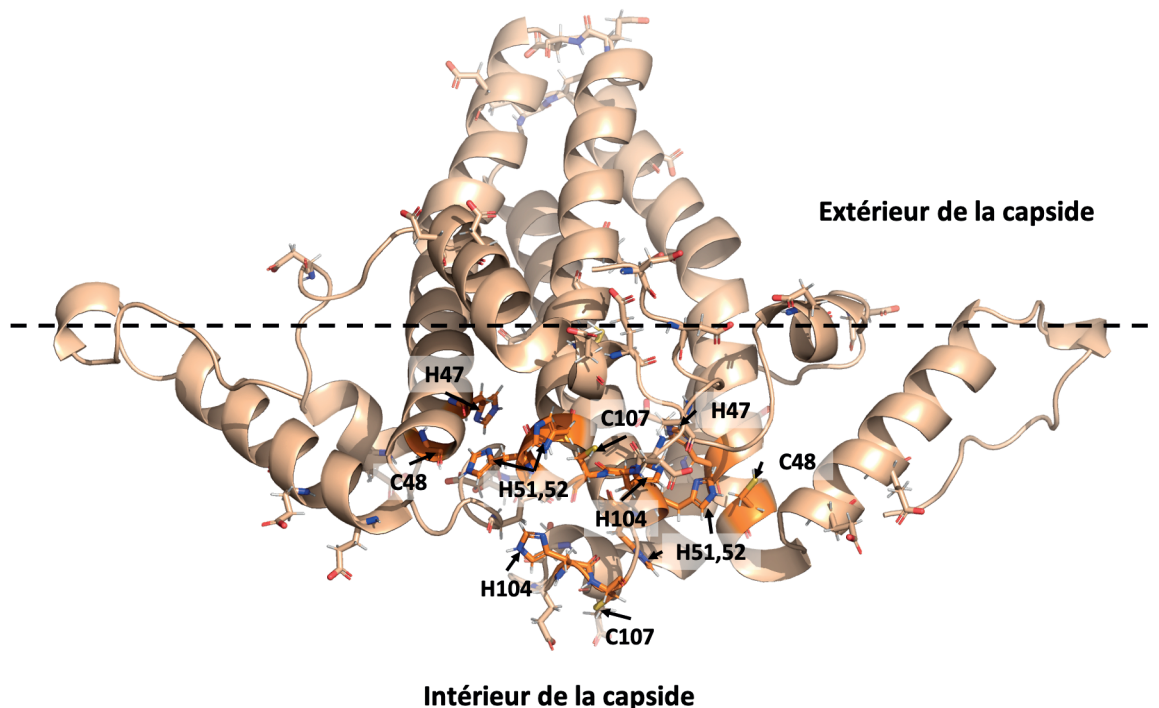


Figure 110 – Localisation des acides aminés potentiellement impliqués dans la chélation de Zn^{2+} . Les cystéines et histidines sont colorées en orange. Elles sont représentées en bâtonnet.

Les protéines, exprimées et purifiées, sont assemblées sous forme de capsid sans addition de Zn^{2+} . Les travaux expérimentaux d'ICP-MS et de colorimétrie, avec le 4-(2-pyridylazo) resorcinol (PAR), montrent que de l'ordre d'un ion zinc est fixé par dimère complet (Cp183) ou tronqué (Cp149). Ces résultats sont des estimations. Il est difficile de quantifier, avec précision, les protéines de capsid. Le rapport zinc/protéine dépend de leur concentration. L'identification du zinc est partiellement en accord avec l'étude de Stray et al. [77]. Dans leur étude, ils saturent en zinc des dimères dissociés de Cp149 à basse concentration et suivent un signal de fluorescence intrinsèque. Ils concluent que le dimère de Cp149 se lie à 3-4,5 Zn^{2+} , en fonction des conditions. Ils concluent de la même façon que Cp149, associée en capsid, ne capte pas Zn^{2+} (pas de variation de fluorescence), soit parce qu'elle n'en est plus capable, soit parce que les sites sont déjà saturés. Le modèle construit avec 3 sites de chélation est compatible avec ces données (Figure 111). À ce stade, l'intervention du CTD dans la fixation du zinc n'est pas étudiée. Elle n'est pas exclue car Böttcher et Nassal observent, par cryo-microscopie électronique, une densité qui s'apparente à C183 en dessous des dimères A-B et C-D [98].

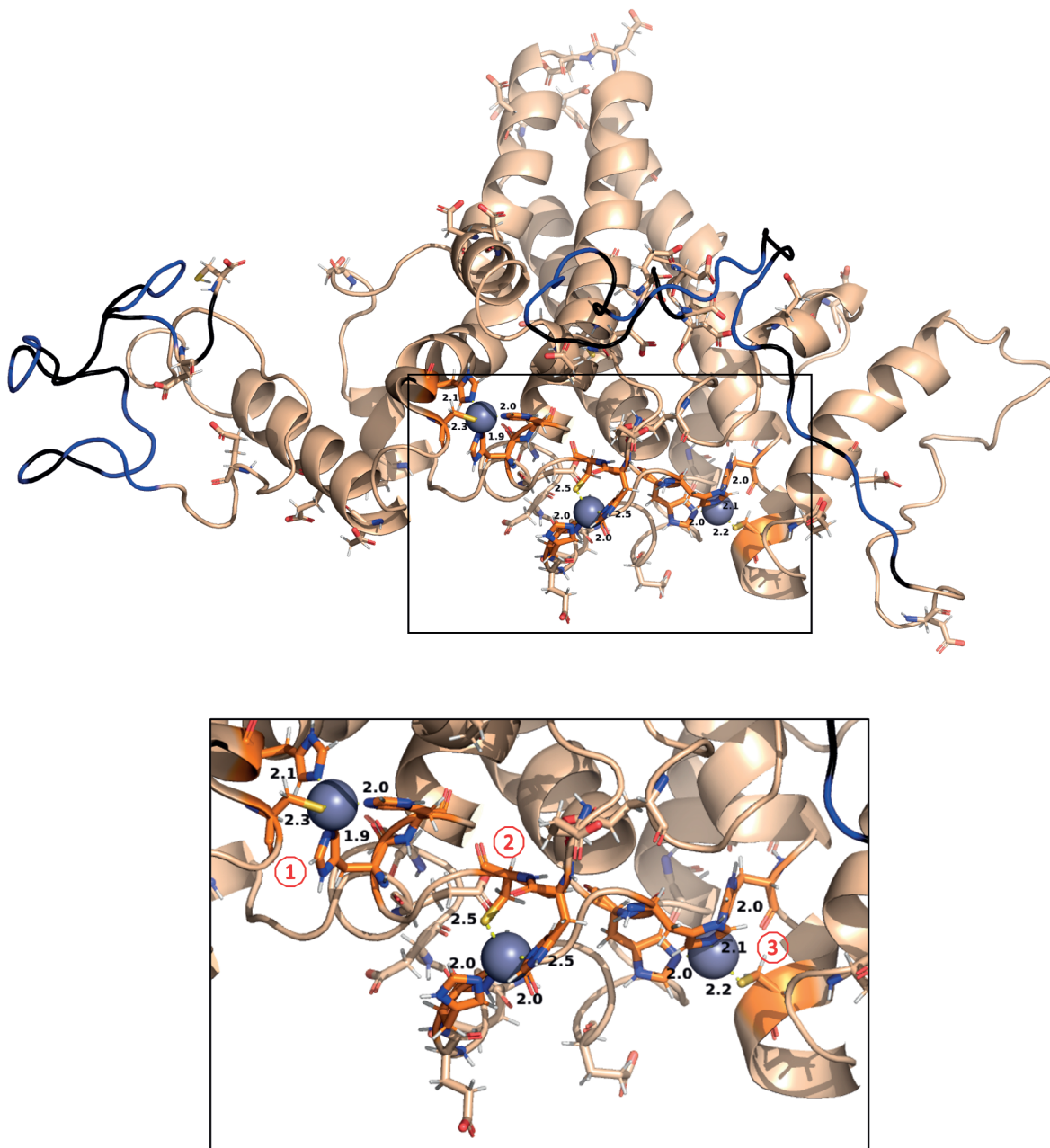


Figure 111 – Sites de chélation putatifs du zinc. Les cystéines et histidines sont représentées en bâtonnet orange. ① et ③ correspondent à des sites CHHH. ② correspond à un site CCHH. L'une des cystéines n'est pas visible sur ②.

L'accessibilité au solvant du site ② est plus élevée que celles des sites ① et ③. Le site ② serait donc plus favorable pour fixer un ion zinc. Les histidines 47 et 51 ainsi que la cystéine 48, sont plus enfouies que les histidines 104 et cystéines 107. Ceci est en accord avec l'identification de l'ordre d'un zinc par dimère complet ou tronqué (Cp149 ou Cp183). Core semble capable de fixer plus de 3 zincs. Les clusters de glutamates et d'aspartates interviendraient pour fixer les zincs supplémentaires (Figure 112). 5 autres sites d'interaction sont identifiés par les travaux in-silico. Il est possible que le zinc, lorsque les dimères sont associés en capsid, vienne interagir au niveau des pores 3 et q3 (E14, E40, E43 et E46 – Autres sites ④ et ⑤ sur la figure 112). Hadden et al. utilisent des ions Na⁺ pour neutraliser

la capside Cp149. Ils ont tendance à résider à l'intérieur des pores [27]. L'acheminement du zinc libre vers l'intérieur de la capside pourrait se faire grâce aux pores 3 et q3. Le sommet de la spicule de Core (E77, D78 – Autre site ⑥ sur la figure 112), chargé négativement, fixerait transitoirement du zinc. Il pourrait donc correspondre à un autre site potentiel non identifié *in vitro*. En effet, les mutants de Cp183 utilisés sont A80K et E77K.

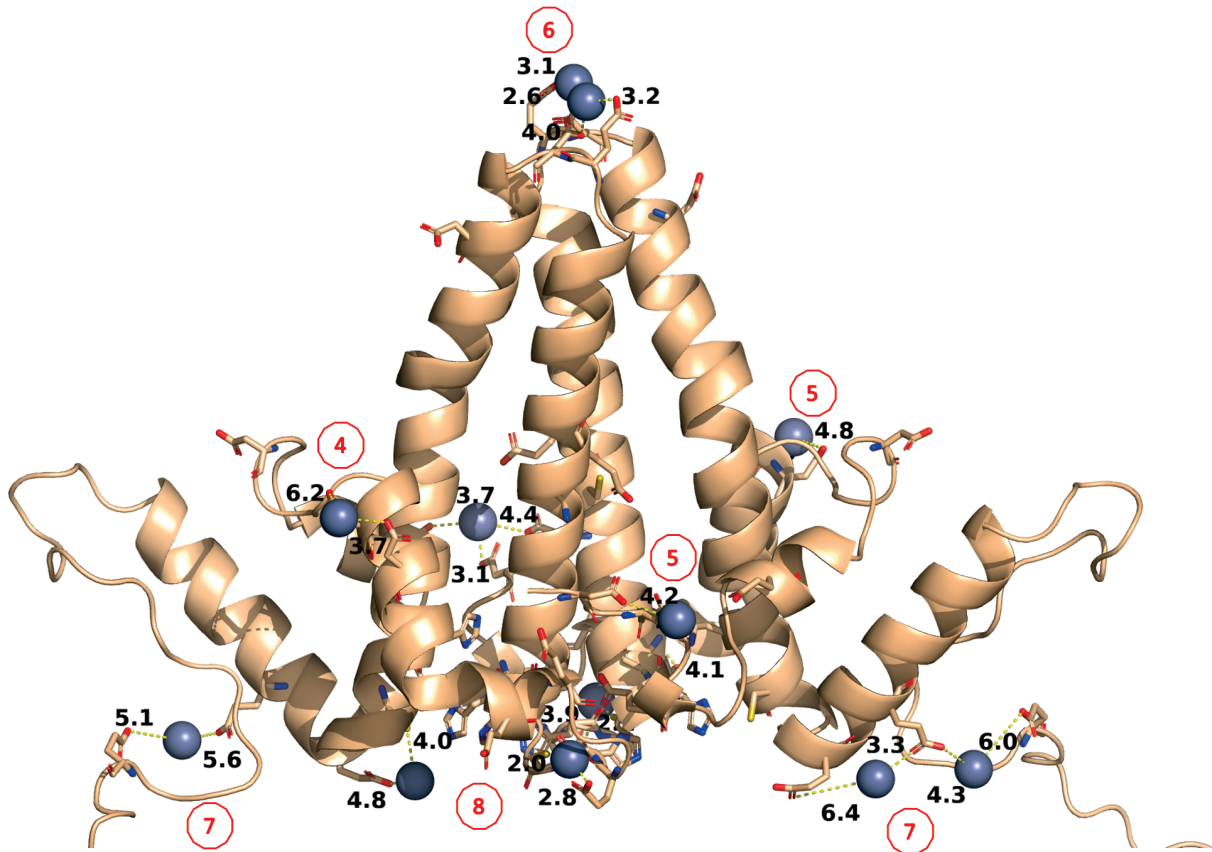


Figure 112 – Autres sites de chélation putatifs de Core.

Stray et al. mettent en évidence que l'assemblage de Cp149 *in vitro* est déclenché, à partir de dimères dissociés, par des concentrations de Zn^{2+} de $\sim 200 \mu M$, à plusieurs ordres de grandeur inférieures à Na^+ , K^+ , Mg^{2+} et Ca^{2+} dans les mêmes conditions [77]. Le zinc pourrait intervenir dans la formation des nucléocapsides. La concentration du zinc total, dans une cellule, est de l'ordre de 0,1 à 0,5 mM [128]. La quantité de Zn^{2+} libre pourrait atteindre 300 atomes par cellule (soit de l'ordre de 1 pM à 100 μM) [128]. Bien que la concentration exacte et la distribution dans la cellule soient encore controversées, il est possible que le zinc intervienne lors de l'assemblage de la capside. Cependant, Michael Nassal et son équipe n'ont obtenu que des agrégats de Core avec le zinc et non, un assemblage sous forme de capside (communication personnelle). Le zinc a donc très probablement un rôle structural sur Core. Chez les virus, il joue souvent un rôle de cofacteur de protéines [129]. Les protéines liant du zinc sont souvent décrites dans l'interaction avec de l'ARN ou de

l'ADN viral. Dans les cellules eucaryotes, le site de chélation le plus répandu est de type Cys2His2 (CCHH) [130]. L'un des sites putatifs modélisés ressemble au type Cys2His2 (Site ② sur la figure 111). D'ailleurs, tous les sites modélisés ont des similitudes avec des sites de chélation connus et s'apparentent à des doigts de zinc.

Le rôle de structuration du zinc sur Core est inconnu. Pour d'autres virus, des fonctions sont associées à des protéines virales [129]. L'association du zinc avec les structures en forme de doigts de zinc confère au moins 3 types d'interactions : (1) Les ions Zn^{2+} maintiennent de maintenir une protéine virale dans une conformation requise, nécessaire à son auto-agrégation. Ils participent à l'oligomérisation de la protéine virale [129]. Ce rôle pourrait être compatible avec Core. De fait, le zinc déclenche plus facilement l'auto-assemblage de la capsid tronquée (Cp149), à partir des dimères dissociés. (2) De manière générale, le Zn^{2+} peut être impliqué dans la formation de dimères. Il se lie simultanément à 2 conformations de la même protéine et sert de pont [129]. (3) Le zinc est un facteur critique dans les interactions entre les protéines virales et d'autres partenaires [129]. C'est le cas des facteurs de transcription et des nucléases qui font intervenir du zinc pour interagir avec du matériel génétique. Dans le cas de la protéine Core, le zinc pourrait induire une conformation conduisant à une interaction avec du matériel génétique. Une étude montre que l'utilisation d'une protéine antivirale, à doigt de zinc (ZAP), inhibe la réplication du VHB dans des cellules dérivées d'hépatocytes humains [131]. Son utilisation diminue la quantité d'ARN viral. Le ZAP est d'ailleurs capable d'interagir avec l'ARN viral. Il empêche donc l'interaction de l'ARNpg avec Core [131]. Il pourrait également rentrer en compétition avec la protéine Core pour fixer le zinc et inhiberait différemment l'interaction avec l'ARNpg.

Pour s'en assurer, il faudrait désassembler les capsides en dimères (jeux de tampon) et chélater tout le zinc (EDTA). Une fois cette opération réalisée, les capsides pourraient être réassemblées en présence d'ARNpg. Cette expérience démontrera si Core fixe du matériel génétique en l'absence totale de zinc.

10 PERSPECTIVES

Cette étude s'est heurtée à une IDR particulièrement difficile à étudier. Dans ces perspectives, un panel de méthodes in-silico, destinées à étudier les régions intrinsèquement désordonnées (IDR), sera abordé. Comme il n'est pas toujours possible de disposer du temps de calcul nécessaire sur un supercalculateur, une technique moins chronophage sera présentée.

Le CTD a une faible complexité de séquence. Il est essentiellement composé d'arginines (Figure 57) et prédit comme désordonné. Il se replie et s'enrichit transitoirement en structures secondaires, de type hélice, au cours des MD. Ces résultats sont obtenus avec plusieurs champs de force dont un spécialisé dans les régions intrinsèquement désordonnées, en utilisant une grande boîte d'eau. Les modèles de départ prédisent également des structures en hélices sur les CTD. Les bras C-terminaux ont toutes les propriétés d'une région intrinsèquement désordonnée (IDR). Les IDR ont généralement de nombreux partenaires. Dans le cas de Core, les partenaires principaux sont : (1) le matériel génétique viral (ARNpg, ADNrc, ADNccc) ainsi que le génome de l'hôte et (2) les protéines impliquées dans le transport du matériel génétique viral, vers le noyau des cellules hépatiques.

Déterminer comment le CTD peut être exposé, à travers un pore, est donc un problème d'autant plus difficile à appréhender. Les IDR sont caractérisées par un espace conformationnel très vaste. À notre échelle de temps (~100 ns), si l'on produisait des simulations non biaisées, elles ne seraient pas suffisantes pour observer la sortie du CTD par les pores. Même si l'on augmente la durée de simulation, il n'est pas sûr que l'exposition du CTD se produise. En effet, les champs de force classiques (Amber 99SB ILDN [109,132]) ne sont pas paramétrés de base pour les IDR. Les conformations et MD de Core donnent une bonne représentation de l'espace conformationnel du CTD mais sont peut-être issues d'un même paysage conformationnel (puit énergétique). Les structures finales après MD et les modèles correspondraient à des minimums énergétiques locaux (Figure 113). L'état replié d'une protéine est composé d'un ensemble d'états structurellement apparentés (Figure 113A) [133]. Le paysage énergétique d'une protéine intrinsèquement désordonnée est, quant à lui, beaucoup plus complexe (Figure 113B). Il montre plusieurs états désordonnés (dépliés), peuplés en solutions libres. Ces états sont séparés par des barrières énergétiques peu profondes et peuvent s'échanger rapidement, contrairement aux protéines repliées. Les puits énergétiques peu profonds sont synonyme d'une grande flexibilité structurale. Les modifications post-traductionnelles, telle que la phosphorylation, peuvent stabiliser l'un des états par rapport aux autres. Dans ce paysage figure également un état structuré mais il n'est stabilisé que lors de la liaison avec une protéine partenaire.

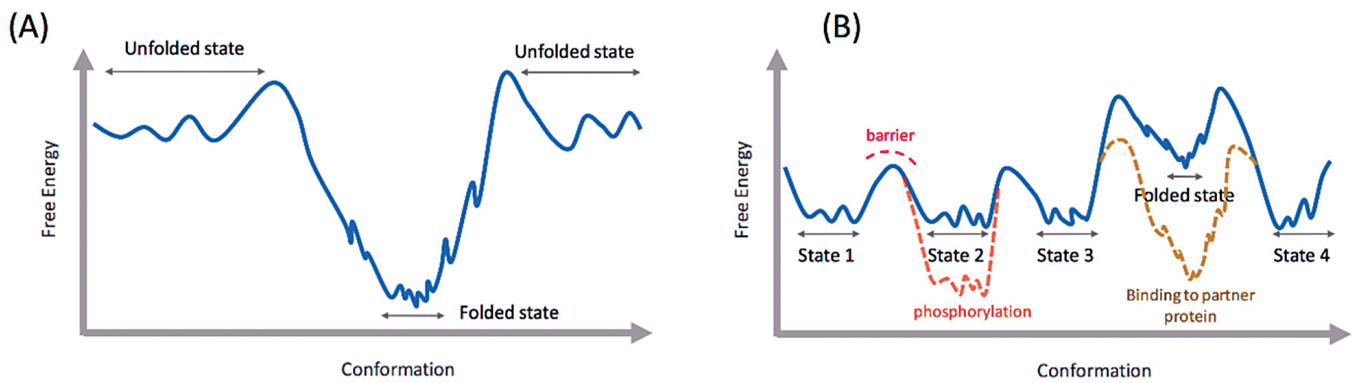


Figure 113 – Paysages énergétiques d'une protéine repliée et d'une protéine intrinsèquement désordonnée, issus de la revue [133]. A. Énergie libre d'une protéine repliée. B. Paysage hypothétique de l'énergie libre d'une protéine intrinsèquement désordonnée.

Toutes les caractéristiques de l'hypothétique paysage énergétique d'une protéine intrinsèquement désordonnée, décriraient toute la dynamique du CTD. Il possède des sérines qui peuvent être phosphorylées et qui modulent l'interaction avec ses partenaires (par exemple S155, S162 et S170 avec l'ARN viral) [78,79,82,134]. Core interagit avec de nombreuses molécules très différentes (ARN, ADN, protéines de transport...). Au cours des interactions avec celles-ci, le CTD adopte forcément des états de repliement qui diffèrent les uns des autres.

Des champs de force adaptés aux régions intrinsèquement désordonnées (IDR) ou protéines intrinsèquement désordonnées (IDP), sont développés [112]. Les phénomènes de repliement se font à des échelles de temps très larges (de la picoseconde à la milliseconde) et dépendent aussi de l'environnement (partenaire à proximité) [133]. Les simulations, sur une grande échelle de temps (microseconde à milliseconde), sont limitées par la puissance de calcul pour de gros systèmes (pore 3 ou q3). Ainsi, l'exploration de l'espace conformationnel de telles structures peut se faire en utilisant d'autres techniques in-silico. La MD avec échanges de répliques (REMD), est l'une des méthodes utilisées [135]. Des protéines de plus de 190 résidus et comportant des IDR, sont simulées à l'aide de celle-ci [136]. Pour étudier l'oligomérisation des β -amyloïdes ($A\beta$) de la maladie d'Alzheimer, la REMD est utilisée avec plusieurs champs de force tout-atomes [137]. L'un des systèmes de cette étude est un trimère du fragment $A\beta_{16-22}$ (7 x 3 résidus). Il faut 80 jours sur 40 processeurs pour simuler 40 répliques de 55 ns. La consommation correspond à ~80 000 d'heures de calcul. Des systèmes légèrement plus gros, comme l'amélogénine P173 qui est une protéine de 173 résidus de l'émail dentaire, sont simulés [132].

D'autres méthodes, comme les MD gros grains, sont un très bon moyen de réduire la complexité du système et donc, du coût computationnel [133]. Lorsque l'on ne dispose pas d'heures de calculs, à cette échelle granulométrique, des calculs peuvent être effectués sur une station de travail de laboratoire. C'est donc une excellente alternative pour étudier l'ensemble conformationnel des IDPs (et/ou IDRs). Dans ce cadre, le champ de force AWSEM est adapté pour étudier les IDPs et IDRs [135,138]. Wu et al. le testent, en simulant

la partie N-terminale de l'histone H4 et l'antitoxine 2 associée à ParE2 (PaaE2). AWSEM-IDP permet d'explorer, plus largement, l'espace conformationnel des IDP tout en maintenant une précision physico-chimique suffisante. Ce champ de force pourrait servir à étudier Core qui, comme on le sait, contient des régions ordonnées et désordonnées.

La métadynamique avec biais échangés (BEMD) est une autre alternative [133]. Cette méthode d'échantillonnage accéléré de dynamique moléculaire permet de prédire les états métastables ainsi que l'affinité de liaison d'un complexe protéine-protéine. Elle pourrait être adaptée dans l'objectif d'étudier l'interaction de Core avec le matériel génétique. De manière générale, les techniques de REMD, de BEMD et gros grains ont l'avantage d'échantillonner un espace conformationnel plus large, à des échelles de temps différentes (de la nanoseconde à la microseconde).

Les IDPs forment des complexes avec d'autres protéines et des polymères d'acide nucléique [133]. La question de l'interaction du CTD (IDR), avec le matériel génétique et les transporteurs nucléaires, se pose. Dans une étude à la fois théorique et expérimentale, Milles et al. étudient l'association entre les nucléoporines et l'importine β (transporteur nucléaire), en utilisant des MD tout-atomes [139]. Sur 10 simulations indépendantes, 4 événements de liaison sont observés en moins de 100 ns (Figure 114).

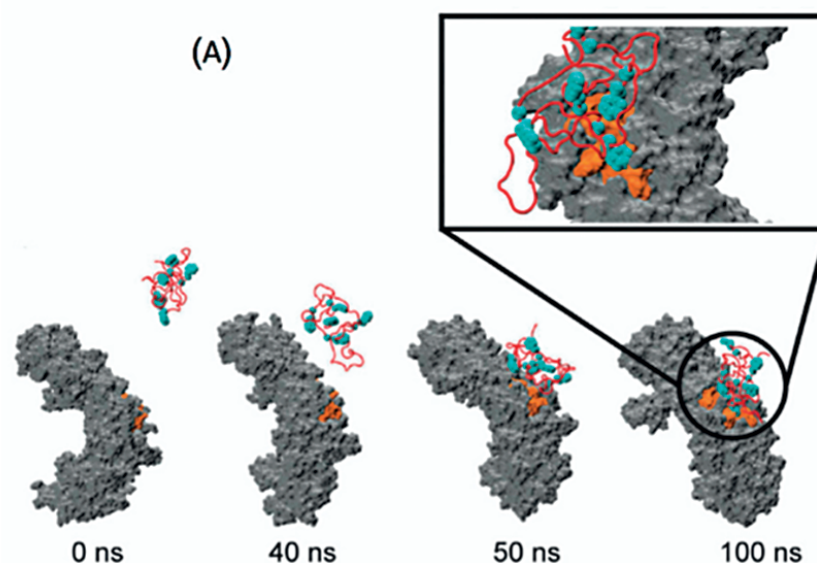


Figure 114 – Interaction de la nucléoporine avec l'importine β , issue de l'article [139]. La nucléoporine est représentée en "cartoon" rouge. L'importine β est représentée en surface grise. Les zones d'interaction sont respectivement en orange et cyan, pour la nucléoporine et l'importine β .

De manière générale, lors d'une liaison avec une molécule partenaire, les IDPs ou IDRs adoptent des structures repliées. 2 types de mécanismes sont proposés : (1) la sélection conformationnelle et (2) l'ajustement induit [133,139]. Dans la sélection conformationnelle, l'une des conformations repliées préexistantes se lie à la protéine partenaire. En ce qui

concerne l'ajustement induit, le repliement se produit après l'interaction avec le partenaire (Figure 114). Pour voir de quels mécanismes dépend l'interaction de Core avec l'ADN, le protocole de simulation de Milles et al. pourrait être adapté [139].

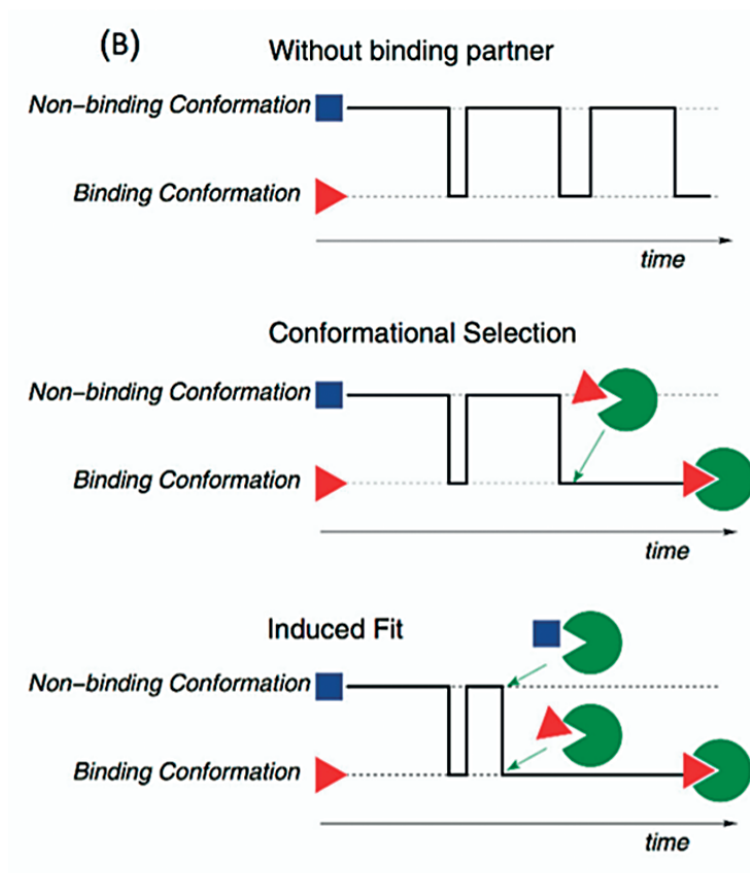


Figure 115 - Démonstration schématique des mécanismes de sélection conformationnelle et de liaison par ajustement induit, issue de l'article [140].

Les IDP suivraient une combinaison de ces 2 mécanismes. L'un des états métastables, partiellement replié, interagirait avec la protéine partenaire (sélection) et le repliement total, s'effectuerait à la suite (ajustement induit). Des barrières énergétiques doivent être franchies pour que l'interaction se produise. Les MD tout-atomes non biaisées en solvant explicite sont incapables de reproduire ces mécanismes. Les modèles gros grains, qui se basent sur les modèles de Gō [122,124], sont largement utilisés pour étudier les interactions des IDP [140,141,142]. Dans le champ de force du modèle gros grains Gō, les interactions sont paramétrées pour stabiliser l'état natif du système étudié. Le modèle Gō est adapté pour échantillonner plus d'une conformation, à la place d'un seul état natif. Pour ce faire, des termes énergétiques issus de plusieurs topologies sont utilisés (un terme par conformation) [140]. Si l'on adapte, une fois de plus, ces techniques à notre problématique, la question de l'interaction de Core avec ses partenaires pourrait être étudiée.

CONCLUSION GÉNÉRALE

L'objectif de cette thèse a été de comprendre la dynamique et les mécanismes d'assemblage de 2 protéines de capsid. Dans ce cadre, des études ont été effectuées sur le virus de Norwalk et sur le virus de l'hépatite B (VHB).

La première étude a donc porté sur l'assemblage de la capsid du Norovirus et plus particulièrement, sur la formation d'un intermédiaire allongé. Pour ce faire, un panel de techniques in-silico a été mis à contribution : la modélisation, la simulation, le regroupement de structures et l'amarrage. La combinaison de 2 échelles granulométriques, tout-atomes et gros grains MARTINI, a été un excellent compromis entre précision et coût de temps de calcul pour comprendre ce processus d'assemblage. L'ensemble des résultats a montré que l'intermédiaire allongé correspond à un pentamère de dimères et dans son prolongement, à la formation d'un morceau d'axe hexamérique.

La deuxième étude, sur le virus de l'hépatite B, s'est scindée en 2 parties. La première partie a consisté à étudier l'espace conformationnel d'une région intrinsèquement désordonnée (IDR). Cette région en question correspond au domaine C-terminal (CTD) de la protéine Core, protéine de capsid du VHB. Pour explorer la dynamique de ce domaine, dans le contexte du dimère libre et de la capsid, la simulation classique ou biaisée a été utilisée. Un champ de force traditionnel et un second, spécialisé dans l'étude d'IDR, ont également été utilisés dans ce cadre. Les résultats ont montré que le CTD est capable d'être exposé en dehors de la capsid par 2 types de pores sur 4. La seconde partie de l'étude a consisté à déterminer si Core interagit avec le zinc. Une modélisation de Core, en interaction avec le zinc, a donc été effectuée. Un champ de force destiné à l'étude de sites métalliques a été utilisé ainsi qu'un champ de force classique. Plusieurs modèles compatibles avec la chélation du Zn^{2+} ont été produits et simulés. Des expériences de titration du Zn^{2+} par colorimétrie et d'ICP-MS ont donc été menées et ont confirmés que Core fixe du zinc.

Ce manuscrit contribue donc à la compréhension de la dynamique et à l'assemblage des protéines de capsid icosaédriques. En effet, ces travaux décrivent des mécanismes majeurs d'auto-assemblage et des propriétés importantes des IDR des protéines de capsid. Grâce à l'ensemble des caractéristiques mises en évidence, ces travaux pourront participer à la lutte antivirale ainsi qu'à l'élaboration d'une thérapie génique.

BIBLIOGRAPHIE

1. Buchen-Osmond C. The universal virus database ICTVdB. *Computing in Science Engineering*. 2003;5: 16–25. doi:10.1109/MCISE.2003.1196303
2. Perlmutter JD, Hagan MF. Mechanisms of Virus Assembly. *Annual Review of Physical Chemistry*. 2015;66: 217–239. doi:10.1146/annurev-physchem-040214-121637
3. Tresset G, Castelnovo M, Leforestier A. Assemblage et désassemblage des virus : mode d'emploi. *Reflète phys.* 2017; 22–26. doi:10.1051/refdp/201752022
4. Ye Q, West AMV, Silletti S, Corbett KD. Architecture and self-assembly of the SARS-CoV-2 nucleocapsid protein. *Protein Sci*. 2020 [cited 5 Nov 2020]. doi:10.1002/pro.3909
5. Coscio F, Nadra AD, Ferreira DU. A structural model for the Coronavirus Nucleocapsid. arXiv:200512165 [q-bio]. 2020 [cited 20 Oct 2020]. Available: <http://arxiv.org/abs/2005.12165>
6. Caspar DL, Klug A. Physical principles in the construction of regular viruses. *Cold Spring Harb Symp Quant Biol*. 1962;27: 1–24. doi:10.1101/sqb.1962.027.001.005
7. Hulo C, de Castro E, Masson P, Bougueleret L, Bairoch A, Xenarios I, et al. ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res*. 2011;39: D576–D582. doi:10.1093/nar/gkq901
8. Berke JM, Dehertogh P, Vergauwen K, Damme EV, Mostmans W, Vandyck K, et al. Capsid Assembly Modulators Have a Dual Mechanism of Action in Primary Human Hepatocytes Infected with Hepatitis B Virus. *Antimicrobial Agents and Chemotherapy*. 2017;61. doi:10.1128/AAC.00560-17
9. Koromyslova AD, White PA, Hansman GS. Treatment of norovirus particles with citrate. *Virology*. 2015;485: 199–204. doi:10.1016/j.virol.2015.07.009
10. Koromyslova A, Tripathi S, Morozov V, Schroten H, Hansman GS. Human norovirus inhibition by a human milk oligosaccharide. *Virology*. 2017;508: 81–89. doi:10.1016/j.virol.2017.04.032
11. Netzler NE, Tuipulotu DE, White PA. Norovirus antivirals: Where are we now? *Medicinal Research Reviews*. 2019;39: 860–886. doi:https://doi.org/10.1002/med.21545
12. Wang M, Lu M, Fritz MP, Quinn CM, Byeon I-JL, Byeon C-H, et al. Fast Magic-Angle Spinning 19F NMR Spectroscopy of HIV-1 Capsid Protein Assemblies. *Angewandte Chemie International Edition*. 2018;57: 16375–16379. doi:https://doi.org/10.1002/anie.201809060
13. Chevreuil M, Law-Hine D, Chen J, Bressanelli S, Combet S, Constantin D, et al. Nonequilibrium self-assembly dynamics of icosahedral viral capsids packaging genome or polyelectrolyte. *Nature Communications*. 2018;9: 3071. doi:10.1038/s41467-018-05426-8
14. Zlotnick A, Johnson JM, Wingfield PW, Stahl SJ, Endres D. A Theoretical Model Successfully Identifies Features of Hepatitis B Virus Capsid Assembly. *Biochemistry*. 1999;38: 14644–14652. doi:10.1021/bi991611a
15. Hagan MF. Modeling Viral Capsid Assembly. *Adv Chem Phys*. 2014;155: 1–68. doi:10.1002/9781118755815.ch01
16. Zlotnick A. Theoretical aspects of virus capsid assembly. *Journal of Molecular Recognition*. 2005;18: 479–490. doi:10.1002/jmr.754
17. Matsudaira P, Bordas J, Koch MH. Synchrotron x-ray diffraction studies of actin structure during polymerization. *PNAS*. 1987;84: 3151–3155. doi:10.1073/pnas.84.10.3151
18. Leach AR, AR L. *Molecular Modelling: Principles and Applications*. Pearson Education; 2001.
19. Perkett MR, Hagan MF. Using Markov state models to study self-assembly. *J Chem Phys*. 2014;140: 214101. doi:10.1063/1.4878494
20. Wang B, Zhang J, Wu Y. A Multiscale Model for the Self-Assembly of Coat Proteins in Bacteriophage MS2. *J Chem Inf Model*. 2019;59: 3899–3909. doi:10.1021/acs.jcim.9b00514

21. Rapaport DC. Molecular dynamics study of T = 3 capsid assembly. *J Biol Phys.* 2018;44: 147–162. doi:10.1007/s10867-018-9486-7
22. Grime JMA, Dama JF, Ganser-Pornillos BK, Woodward CL, Jensen GJ, Yeager M, et al. Coarse-grained simulation reveals key features of HIV-1 capsid self-assembly. *Nature Communications.* 2016;7: 11568. doi:10.1038/ncomms11568
23. Li B, Zhao L, Qian H-J, Lu Z-Y. Coarse-grained simulation study on the self-assembly of miktoarm star-like block copolymers in various solvent conditions. *Soft Matter.* 2014;10: 2245–2252. doi:10.1039/C3SM52660G
24. Schindler T, Kröner D, Steinhauser MO. On the dynamics of molecular self-assembly and the structural analysis of bilayer membranes using coarse-grained molecular dynamics simulations. *Biochimica et Biophysica Acta (BBA) - Biomembranes.* 2016;1858: 1955–1963. doi:10.1016/j.bbamem.2016.05.014
25. Shih AY, Freddolino PL, Arkhipov A, Schulten K. Assembly of lipoprotein particles revealed by coarse-grained molecular dynamics simulations. *Journal of Structural Biology.* 2007;157: 579–592. doi:10.1016/j.jsb.2006.08.006
26. Vendruscolo M, Dobson CM. Protein Dynamics: Moore’s Law in Molecular Biology. *Current Biology.* 2011;21: R68–R70. doi:10.1016/j.cub.2010.11.062
27. Hadden JA, Perilla JR, Schlicksup CJ, Venkatakrishnan B, Zlotnick A, Schulten K. All-atom molecular dynamics of the HBV capsid reveals insights into biological function and cryo-EM resolution limits. *eLife.* 2018 [cited 19 Jul 2018]. doi:10.7554/eLife.32478
28. Perilla JR, Schulten K. Physical properties of the HIV-1 capsid from all-atom molecular dynamics simulations. *Nature Communications.* 2017;8: 15959. doi:10.1038/ncomms15959
29. Lopman BA, Steele D, Kirkwood CD, Parashar UD. The Vast and Varied Global Burden of Norovirus: Prospects for Prevention and Control. *PLoS Med.* 2016;13: e1001999. doi:10.1371/journal.pmed.1001999
30. Pires SM, Fischer-Walker CL, Lanata CF, Devleeschauwer B, Hall AJ, Kirk MD, et al. Aetiology-Specific Estimates of the Global and Regional Incidence and Mortality of Diarrhoeal Diseases Commonly Transmitted through Food. *PLoS One.* 2015;10: e0142927. doi:10.1371/journal.pone.0142927
31. Bartsch SM, Lopman BA, Ozawa S, Hall AJ, Lee BY. Global Economic Burden of Norovirus Gastroenteritis. *PLoS One.* 2016;11: e0151219. doi:10.1371/journal.pone.0151219
32. Vinjé J. Advances in laboratory methods for detection and typing of norovirus. *J Clin Microbiol.* 2015;53: 373–381. doi:10.1128/JCM.01535-14
33. Atmar RL, Estes MK. The epidemiologic and clinical importance of norovirus infection. *Gastroenterol Clin North Am.* 2006;35: 275–290, viii. doi:10.1016/j.gtc.2006.03.001
34. Prasad BVV, Hardy ME, Dokland T, Bella J, Rossmann MG, Estes MK. X-ray Crystallographic Structure of the Norwalk Virus Capsid. *Science.* 1999;286: 287–290.
35. Tubiana T, Boulard Y, Bressanelli S. Dynamics and asymmetry in the dimer of the norovirus major capsid protein. *PLOS ONE.* 2017;12: e0182056. doi:10.1371/journal.pone.0182056
36. Shoemaker GK, van Duijn E, Crawford SE, Uetrecht C, Baclayon M, Roos WH, et al. Norwalk Virus Assembly and Stability Monitored by Mass Spectrometry*. *Molecular & Cellular Proteomics.* 2010;9: 1742–1751. doi:10.1074/mcp.M900620-MCP200
37. Tresset G, Decouche V, Bryche J-F, Charpilienne A, Le Cœur C, Barbier C, et al. Unusual self-assembly properties of Norovirus Newbury2 virus-like particles. *Arch Biochem Biophys.* 2013;537: 144–152. doi:10.1016/j.abb.2013.07.003
38. Tresset G, Le Coeur C, Bryche J-F, Tatou M, Zeghal M, Charpilienne A, et al. Norovirus Capsid Proteins Self-Assemble through Biphasic Kinetics via Long-Lived Stave-like Intermediates. *J Am Chem Soc.* 2013;135: 15373–15381.
39. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics.* 2016;54: 5.6.1-5.6.37. doi:10.1002/cpbi.3

40. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, de Vries AH. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J Phys Chem B*. 2007;111: 7812–7824. doi:10.1021/jp071097f
41. Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, Marrink S-J. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J Chem Theory Comput*. 2008;4: 819–834.
42. Periole X, Cavalli M, Marrink S-J, Ceruso MA. Combining an Elastic Network With a Coarse-Grained Molecular Force Field: Structure, Dynamics, and Intermolecular Recognition. *J Chem Theory Comput*. 2009;5: 2531–2543.
43. Siuda I, Thøgersen L. Conformational flexibility of the leucine binding protein examined by protein domain coarse-grained molecular dynamics. *J Mol Model*. 2013;19: 4931–4945.
44. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*. 2015;1–2: 19–25.
45. Kozin MB, Svergun DI. Automated matching of high- and low-resolution structural models. *J Appl Cryst*. 2001;34: 33–41. doi:10.1107/S0021889800014126
46. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, et al. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*. 2005;26: 1668–1688. doi:10.1002/jcc.20290
47. Tubiana T, Carvaillo J-C, Boulard Y, Bressanelli S. TTClust: A Versatile Molecular Simulation Trajectory Clustering Program with Graphical Summaries. *J Chem Inf Model*. 2018;58: 2178–2182. doi:10.1021/acs.jcim.8b00512
48. Saporta G. Probabilités, analyse des données et statistique. Paris: Editions Technip; 2008.
49. Boyd KJ, Bansal P, Feng J, May ER. Stability of Norwalk Virus Capsid Protein Interfaces Evaluated by in Silico Nanoindentation. *Front Bioeng Biotechnol*. 2015;3. doi:10.3389/fbioe.2015.00103
50. Uetrecht C, Barbu IM, Shoemaker GK, van Duijn E, Heck AJR. Interrogating viral capsid assembly with ion mobility–mass spectrometry. *Nature Chemistry*. 2011;3: 126–132. doi:10.1038/nchem.947
51. Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, et al. The ClusPro web server for protein-protein docking. *Nat Protocols*. 2017;12: 255–278. doi:10.1038/nprot.2016.169
52. Kozakov D, Brenke R, Comeau SR, Vajda S. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*. 2006;65: 392–406. doi:10.1002/prot.21117
53. Poveda-Cuevas SA, Etchebest C, Barroso da Silva FL. Identification of Electrostatic Epitopes in Flavivirus by Computer Simulations: The PROCEEDpKa Method. *J Chem Inf Model*. 2020;60: 944–963. doi:10.1021/acs.jcim.9b00895
54. Lensink MF, Nadzirin N, Velankar S, Wodak SJ. Modeling protein-protein, protein-peptide, and protein-oligosaccharide complexes: CAPRI 7th edition. *Proteins: Structure, Function, and Bioinformatics*. 2020;88: 916–938. doi:https://doi.org/10.1002/prot.25870
55. Andrusier N, Mashlach E, Nussinov R, Wolfson HJ. Principles of Flexible Protein-Protein Docking. *Proteins*. 2008;73: 271–289. doi:10.1002/prot.22170
56. de Vries SJ, Schindler CEM, Chauvot de Beauchêne I, Zacharias M. A Web Interface for Easy Flexible Protein-Protein Docking with ATTRACT. *Biophys J*. 2015;108: 462–465. doi:10.1016/j.bpj.2014.12.015
57. May A, Zacharias M. Protein-protein docking in CAPRI using ATTRACT to account for global and local flexibility. *Proteins*. 2007;69: 774–780. doi:10.1002/prot.21735
58. Marze NA, Roy Burman SS, Sheffler W, Gray JJ. Efficient flexible backbone protein-protein docking for challenging targets. *Bioinformatics*. 2018;34: 3461–3469. doi:10.1093/bioinformatics/bty355
59. Schneidman-Duhovny D, Nussinov R, Wolfson HJ. Automatic prediction of protein interactions with large scale motion. *Proteins*. 2007;69: 764–773. doi:10.1002/prot.21759

60. Pipolo S, Salanne M, Ferlat G, Klotz S, Saitta AM, Pietrucci F. Navigating at Will on the Water Phase Diagram. *Phys Rev Lett*. 2017;119: 245701. doi:10.1103/PhysRevLett.119.245701
61. Saleh N, Ibrahim P, Saladino G, Gervasio FL, Clark T. An Efficient Metadynamics-Based Protocol To Model the Binding Affinity and the Transition State Ensemble of G-Protein-Coupled Receptor Ligands. *J Chem Inf Model*. 2017;57: 1210–1217. doi:10.1021/acs.jcim.6b00772
62. Limongelli V, Bonomi M, Parrinello M. Funnel metadynamics as accurate binding free-energy method. *PNAS*. 2013;110: 6358–6363. doi:10.1073/pnas.1303186110
63. Orellana L. Large-Scale Conformational Changes and Protein Function: Breaking the in silico Barrier. *Front Mol Biosci*. 2019;6. doi:10.3389/fmolb.2019.00117
64. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*. 1999;314: 141–151. doi:10.1016/S0009-2614(99)01123-9
65. Ostermeir K, Zacharias M. Advanced replica-exchange sampling to study the flexibility and plasticity of peptides and proteins. *Biochim Biophys Acta*. 2013;1834: 847–853. doi:10.1016/j.bbapap.2012.12.016
66. Pogan R, Dülfer J, Uetrecht C. Norovirus assembly and stability. *Current Opinion in Virology*. 2018;31: 59–65. doi:10.1016/j.coviro.2018.05.003
67. Zlotnick A, Aldrich R, Johnson JM, Ceres P, Young MJ. Mechanism of Capsid Assembly for an Icosahedral Plant Virus. *Virology*. 2000;277: 450–456. doi:10.1006/viro.2000.0619
68. Devant JM, Hofhaus G, Bhella D, Hansman GS. Heterologous expression of human norovirus GII.4 VP1 leads to assembly of T=4 virus-like particles. *Antiviral Research*. 2019;168: 175–182. doi:10.1016/j.antiviral.2019.05.010
69. Wynne SA, Crowther RA, Leslie AGW. The Crystal Structure of the Human Hepatitis B Virus Capsid. *Molecular Cell*. 1999;3: 771–780. doi:10.1016/S1097-2765(01)80009-5
70. Schreiner S, Nassal M. A Role for the Host DNA Damage Response in Hepatitis B Virus cccDNA Formation—and Beyond? *Viruses*. 2017;9: 125. doi:10.3390/v9050125
71. Durantel D, Zoulim F. New antiviral targets for innovative treatment concepts for hepatitis B virus and hepatitis delta virus. *Journal of Hepatology*. 2016;64: S117–S131. doi:10.1016/j.jhep.2016.02.016
72. Schlicksup CJ, Wang JC-Y, Francis S, Venkatakrishnan B, Turner WW, VanNieuwenhze M, et al. Hepatitis B virus core protein allosteric modulators can distort and disrupt intact capsids. *eLife*. 2018 [cited 19 Jul 2018]. doi:10.7554/eLife.31473
73. World Health Organization, World Health Organization, Global Hepatitis Programme. Global hepatitis report, 2017. 2017. Available: <http://apps.who.int/iris/bitstream/10665/255016/1/9789241565455-eng.pdf?ua=1>
74. Diab A, Foca A, Zoulim F, Durantel D, Andrisani O. The diverse functions of the hepatitis B core/capsid protein (HBc) in the viral life cycle: Implications for the development of HBc-targeting antivirals. *Antiviral Res*. 2018;149: 211–220. doi:10.1016/j.antiviral.2017.11.015
75. Billioud G, Ait-Goughoulte M, Zoulim F. Cycle de réplication du VHB et molécules antivirales. *Virologie*. 2010;14: 57–73. doi:10.1684/vir.2010.0301
76. Chevreuil M, Lecoq L, Wang S, Gargowitsch L, Nhiri N, Jacquet E, et al. Nonsymmetrical Dynamics of the HBV Capsid Assembly and Disassembly Evidenced by Their Transient Species. *J Phys Chem B*. 2020;124: 9987–9995. doi:10.1021/acs.jpcc.0c05024
77. Stray SJ, Ceres P, Zlotnick A. Zinc Ions Trigger Conformational Change and Oligomerization of Hepatitis B Virus Capsid Protein. *Biochemistry*. 2004;43: 9989–9998. doi:10.1021/bi049571k
78. Lan YT, Li J, Liao W, Ou J. Roles of the Three Major Phosphorylation Sites of Hepatitis B Virus Core Protein in Viral Replication. *Virology*. 1999;259: 342–348. doi:10.1006/viro.1999.9798
79. Gazina EV, Fielding JE, Lin B, Anderson DA. Core Protein Phosphorylation Modulates Pregenomic RNA Encapsidation to Different Extents in Human and Duck Hepatitis B Viruses. *J Virol*. 2000;74: 4721–4728. doi:10.1128/JVI.74.10.4721-4728.2000

80. Nassal M. Hepatitis B viruses: reverse transcription a different way. *Virus Res.* 2008;134: 235–249. doi:10.1016/j.virusres.2007.12.024
81. Li H-C, Huang E-Y, Su P-Y, Wu S-Y, Yang C-C, Lin Y-S, et al. Nuclear Export and Import of Human Hepatitis B Virus Capsid Protein and Particles. *Ou JJ, editor. PLoS Pathogens.* 2010;6: e1001162. doi:10.1371/journal.ppat.1001162
82. Heger-Stevic J, Zimmermann P, Lecoq L, Böttcher B, Nassal M. Hepatitis B virus core protein phosphorylation: Identification of the SRPK1 target sites and impact of their occupancy on RNA binding and capsid structure. *PLOS Pathogens.* 2018;14: e1007488. doi:10.1371/journal.ppat.1007488
83. Selzer L, Kant R, Wang JC-Y, Bothner B, Zlotnick A. Hepatitis B Virus Core Protein Phosphorylation Sites Affect Capsid Stability and Transient Exposure of the C-terminal Domain. *J Biol Chem.* 2015;290: 28584–28593. doi:10.1074/jbc.M115.678441
84. Meng D, Hjelm RP, Hu J, Wu J. A Theoretical Model for the Dynamic Structure of Hepatitis B Nucleocapsid. *Biophys J.* 2011;101: 2476–2484. doi:10.1016/j.bpj.2011.10.002
85. Chen C, Wang JC-Y, Pierson EE, Keifer DZ, Delaleau M, Gallucci L, et al. Importin β Can Bind Hepatitis B Virus Core Protein and Empty Core-Like Particles and Induce Structural Changes. *PLOS Pathogens.* 2016;12: e1005802. doi:10.1371/journal.ppat.1005802
86. Gabler F, Nam S-Z, Till S, Mirdita M, Steinegger M, Söding J, et al. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Current Protocols in Bioinformatics.* 2020;72: e108. doi:https://doi.org/10.1002/cpbi.108
87. Yu X, Jin L, Jih J, Shih C, Zhou ZH. 3.5Å cryoEM Structure of Hepatitis B Virus Core Assembled from Full-Length Core Protein. *PLOS ONE.* 2013;8: e69729. doi:10.1371/journal.pone.0069729
88. Bourne CR, Finn MG, Zlotnick A. Global Structural Changes in Hepatitis B Virus Capsids Induced by the Assembly Effector HAP1. *Journal of Virology.* 2006;80: 11055–11061. doi:10.1128/JVI.00933-06
89. Tan WS, McNae IW, Ho KL, Walkinshaw MD. Crystallization and X-ray analysis of the T = 4 particle of hepatitis B capsid protein with an N-terminal extension. *Acta Cryst F.* 2007;63: 642–647. doi:10.1107/S1744309107033726
90. Packianathan C, Katen SP, Dann CE, Zlotnick A. Conformational Changes in the Hepatitis B Virus Core Protein Are Consistent with a Role for Allostery in Virus Assembly. *Journal of Virology.* 2010;84: 1607–1615. doi:10.1128/JVI.02033-09
91. Alexander CG, Jürgens MC, Shepherd DA, Freund SMV, Ashcroft AE, Ferguson N. Thermodynamic origins of protein folding, allostery, and capsid formation in the human hepatitis B virus core protein. *PNAS.* 2013;110: E2782–E2791. doi:10.1073/pnas.1308846110
92. Katen SP, Tan Z, Chirapu SR, Finn MG, Zlotnick A. Assembly-Directed Antivirals Differentially Bind Quasiequivalent Pockets to Modify Hepatitis B Virus Capsid Tertiary and Quaternary Structure. *Structure.* 2013;21: 1406–1416. doi:10.1016/j.str.2013.06.013
93. Venkatakrishnan B, Katen SP, Francis S, Chirapu S, Finn MG, Zlotnick A. Hepatitis B Virus Capsids Have Diverse Structural Responses to Small-Molecule Ligands Bound to the Heteroaryldihydropyrimidine Pocket. *Journal of Virology.* 2016;90: 3994–4004. doi:10.1128/JVI.03058-15
94. Klumpp K, Lam AM, Lukacs C, Vogel R, Ren S, Espiritu C, et al. High-resolution crystal structure of a hepatitis B virus replication inhibitor bound to the viral core protein. *Proceedings of the National Academy of Sciences.* 2015;112: 15196–15201. doi:10.1073/pnas.1513803112
95. Qiu Z, Lin X, Zhou M, Liu Y, Zhu W, Chen W, et al. Design and Synthesis of Orally Bioavailable 4-Methyl Heteroaryldihydropyrimidine Based Hepatitis B Virus (HBV) Capsid Inhibitors. *J Med Chem.* 2016;59: 7651–7666. doi:10.1021/acs.jmedchem.6b00879
96. Zhou Z, Hu T, Zhou X, Wildum S, Garcia-Alcalde F, Xu Z, et al. Heteroaryldihydropyrimidine (HAP) and Sulfamoylbenzamide (SBA) Inhibit Hepatitis B Virus Replication by Different Molecular Mechanisms. *Scientific Reports.* 2017;7: 42374. doi:10.1038/srep42374

97. Zhao Z, Wang JC-Y, Gonzalez-Gutierrez G, Venkatakrishnan B, Asor R, Khaykelson D, et al. Structural Differences between the Woodchuck Hepatitis Virus Core Protein in the Dimer and Capsid States Are Consistent with Entropic and Conformational Regulation of Assembly. *Journal of Virology*. 2019;93. doi:10.1128/JVI.00141-19
98. Böttcher B, Nassal M. Structure of Mutant Hepatitis B Core Protein Capsids with Premature Secretion Phenotype. *Journal of Molecular Biology*. 2018;430: 4941–4954. doi:10.1016/j.jmb.2018.10.018
99. Kang J-A, Kim S, Park M, Park H-J, Kim J-H, Park S, et al. Ciclopirox inhibits Hepatitis B Virus secretion by blocking capsid assembly. *Nature Communications*. 2019;10: 2184. doi:10.1038/s41467-019-10200-5
100. Aston-Deaville S, Carlsson E, Saleem M, Thistlethwaite A, Chan H, Maharjan S, et al. An assessment of the use of Hepatitis B Virus core protein virus-like particles to display heterologous antigens from *Neisseria meningitidis*. *Vaccine*. 2020;38: 3201–3209. doi:10.1016/j.vaccine.2020.03.001
101. Wu W, Watts NR, Cheng N, Huang R, Steven AC, Wingfield PT. Expression of quasi-equivalence and capsid dimorphism in the Hepadnaviridae. *PLOS Computational Biology*. 2020;16: e1007782. doi:10.1371/journal.pcbi.1007782
102. Zhao Z, Wang JC-Y, Segura CP, Hadden-Perilla JA, Zlotnick A. The Integrity of the Intradimer Interface of the Hepatitis B Virus Capsid Protein Dimer Regulates Capsid Self-Assembly. *ACS Chem Biol*. 2020;15: 3124–3132. doi:10.1021/acscchembio.0c00277
103. Schlicksup CJ, Laughlin P, Dunkelbarger S, Wang JC-Y, Zlotnick A. Local Stabilization of Subunit–Subunit Contacts Causes Global Destabilization of Hepatitis B Virus Capsids. *ACS Chem Biol*. 2020;15: 1708–1717. doi:10.1021/acscchembio.0c00320
104. Makbul C, Nassal M, Böttcher B. Slowly folding surface extension in the prototypic avian hepatitis B virus capsid governs stability. Li W, Wolberger C, editors. *eLife*. 2020;9: e57277. doi:10.7554/eLife.57277
105. Hoffmann A, Grudinin S. NOLB: Nonlinear Rigid Block Normal-Mode Analysis Method. *J Chem Theory Comput*. 2017;13: 2123–2134. doi:10.1021/acs.jctc.7b00197
106. Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, et al. The Protein Model Portal--a comprehensive resource for protein structure and model information. *Database (Oxford)*. 2013;2013: bat031. doi:10.1093/database/bat031
107. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res*. 2015;43: W174–W181. doi:10.1093/nar/gkv342
108. Hiranuma N, Park H, Baek M, Anishchanka I, Dauparas J, Baker D. Improved protein structure refinement guided by deep learning based accuracy estimation. *bioRxiv*. 2020; 2020.07.17.209643. doi:10.1101/2020.07.17.209643
109. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics*. 2010;78: 1950–1958. doi:10.1002/prot.22711
110. Czaplewski C, Karczyńska A, Sieradzan AK, Liwo A. UNRES server for physics-based coarse-grained simulations and prediction of protein structure, dynamics and thermodynamics. *Nucleic Acids Research*. 2018;46: W304–W309. doi:10.1093/nar/gky328
111. Roe DR, Cheatham TE. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput*. 2013;9: 3084–3095. doi:10.1021/ct400341p
112. Robustelli P, Piana S, Shaw DE. Developing a molecular dynamics force field for both folded and disordered protein states. *PNAS*. 2018;115: E4758–E4766. doi:10.1073/pnas.1800690115
113. Piana S, Donchev AG, Robustelli P, Shaw DE. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J Phys Chem B*. 2015;119: 5113–5123. doi:10.1021/jp508971m
114. Yu X, Jin L, Jih J, Shih C, Zhou ZH. 3.5Å cryoEM Structure of Hepatitis B Virus Core Assembled from Full-Length Core Protein. *PLOS ONE*. 2013;8: e69729. doi:10.1371/journal.pone.0069729

115. Porterfield JZ, Zlotnick A. A simple and general method for determining the protein and nucleic acid content of viruses by UV absorbance. *Virology*. 2010;407: 281–288. doi:10.1016/j.virol.2010.08.015
116. Scopes RK. Measurement of protein by spectrophotometry at 205 nm. *Analytical Biochemistry*. 1974;59: 277–282. doi:10.1016/0003-2697(74)90034-7
117. Goldfarb AR, Sidel LJ. Ultraviolet Absorption Spectra of Proteins. *Science*. 1951;114: 156–157. doi:10.1126/science.114.2954.156
118. Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*. 1976;72: 248–254. doi:10.1016/0003-2697(76)90527-3
119. Smith PK, Krohn RI, Hermanson GT, Mallia AK, Gartner FH, Provenzano MD, et al. Measurement of protein using bicinchoninic acid. *Analytical Biochemistry*. 1985;150: 76–85. doi:10.1016/0003-2697(85)90442-7
120. Gervason S, Larkem D, Mansour AB, Botzanowski T, Müller CS, Pecqueur L, et al. Physiologically relevant reconstitution of iron-sulfur cluster biosynthesis uncovers persulfide-processing functions of ferredoxin-2 and frataxin. *Nature Communications*. 2019;10: 3566. doi:10.1038/s41467-019-11470-9
121. Putignano V, Rosato A, Banci L, Andreini C. MetalPDB in 2018: a database of metal sites in biological macromolecular structures. *Nucleic Acids Res*. 2018;46: D459–D464. doi:10.1093/nar/gkx989
122. Peters MB, Yang Y, Wang B, Füsti-Molnár L, Weaver MN, Merz KM. Structural Survey of Zinc-Containing Proteins and Development of the Zinc AMBER Force Field (ZAFF). *J Chem Theory Comput*. 2010;6: 2935–2947. doi:10.1021/ct1002626
123. Morgunova E, Tuuttila A, Bergmann U, Isupov M, Lindqvist Y, Schneider G, et al. Structure of Human Pro-Matrix Metalloproteinase-2: Activation Mechanism Revealed. *Science*. 1999;284: 1667–1670. doi:10.1126/science.284.5420.1667
124. Elrod-Erickson M, Benson TE, Pabo CO. High-resolution structures of variant Zif268–DNA complexes: implications for understanding zinc finger–DNA recognition. *Structure*. 1998;6: 451–464. doi:10.1016/S0969-2126(98)00047-1
125. Durham E, Dorr B, Woetzel N, Staritzbichler R, Meiler J. Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *J Mol Model*. 2009;15: 1093–1108. doi:10.1007/s00894-009-0454-9
126. Kefalakes H, Budeus B, Walker A, Jochum C, Hilgard G, Heinold A, et al. Adaptation of the hepatitis B virus core protein to CD8+ T-cell selection pressure. *Hepatology*. 2015;62: 47–56. doi:https://doi.org/10.1002/hep.27771
127. Warner BG, Tsai P, Rodrigo AG, ‘Ofanoa M, Gane EJ, Munn SR, et al. Evidence for reduced selection pressure on the hepatitis B virus core gene in hepatitis B e antigen-negative chronic hepatitis B. *Journal of General Virology*. 92: 1800–1808. doi:10.1099/vir.0.030478-0
128. Eide DJ. Zinc transporters and the cellular trafficking of zinc. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*. 2006;1763: 711–722. doi:10.1016/j.bbamcr.2006.03.005
129. Lazarczyk M, Favre M. Role of Zn²⁺ Ions in Host-Virus Interactions. *Journal of Virology*. 2008;82: 11486–11494. doi:10.1128/JVI.01314-08
130. Laity JH, Lee BM, Wright PE. Zinc finger proteins: new insights into structural and functional diversity. *Current Opinion in Structural Biology*. 2001;11: 39–46. doi:10.1016/S0959-440X(00)00167-6
131. Mao R, Nie H, Cai D, Zhang J, Liu H, Yan R, et al. Inhibition of Hepatitis B Virus Replication by the Host Zinc Finger Antiviral Protein. *PLOS Pathogens*. 2013;9: e1003494. doi:10.1371/journal.ppat.1003494
132. Homeyer N, Horn AHC, Lanig H, Sticht H. AMBER force-field parameters for phosphorylated amino acids in different protonation states: phosphoserine, phosphothreonine, phosphotyrosine, and phosphohistidine. *Journal of Molecular Modeling*. 2006;12: 281–289. doi:10.1007/s00894-005-0028-4


133. Bhattacharya S, Lin X. Recent Advances in Computational Protocols Addressing Intrinsically Disordered Proteins. *Biomolecules*. 2019;9. doi:10.3390/biom9040146
134. Su P-Y, Yang C-J, Chu T-H, Chang C-H, Chiang C, Tang F-M, et al. HBV maintains electrostatic homeostasis by modulating negative charges from phosphoserine and encapsidated nucleic acids. *Scientific Reports*. 2016;6. doi:10.1038/srep38959
135. Mu J, Liu H, Zhang J, Luo R, Chen H-F. Recent Force Field Strategies for Intrinsically Disordered Proteins. *J Chem Inf Model*. 2021;61: 1037–1047. doi:10.1021/acs.jcim.0c01175
136. Apicella A, Marascio M, Colangelo V, Soncini M, Gautieri A, Plummer CJG. Molecular dynamics simulations of the intrinsically disordered protein amelogenin. *J Biomol Struct Dyn*. 2017;35: 1813–1823. doi:10.1080/07391102.2016.1196151
137. Nguyen PH, Li MS, Derreumaux P. Effects of all-atom force fields on amyloid oligomerization: replica exchange molecular dynamics simulations of the A β 16–22 dimer and trimer. *Phys Chem Chem Phys*. 2011;13: 9778–9788. doi:10.1039/C1CP20323A
138. Wu H, Wolynes PG, Papoian GA. AWSEM-IDP: A Coarse-Grained Force Field for Intrinsically Disordered Proteins. *J Phys Chem B*. 2018;122: 11115–11125. doi:10.1021/acs.jpccb.8b05791
139. Milles S, Mercadante D, Aramburu IV, Jensen MR, Banterle N, Koehler C, et al. Plasticity of an Ultrafast Interaction between Nucleoporins and Nuclear Transport Receptors. *Cell*. 2015;163: 734–745. doi:10.1016/j.cell.2015.09.047
140. Knott M, Best RB. Discriminating binding mechanisms of an intrinsically disordered protein via a multi-state coarse-grained model. *J Chem Phys*. 2014;140: 175102. doi:10.1063/1.4873710
141. Takada S. Gō model revisited. *Biophysics and Physicobiology*. 2019;16: 248–255. doi:10.2142/biophysico.16.0_248
142. Lu Q, Lu HP, Wang J. Exploring the Mechanism of Flexible Biomolecular Recognition with Single Molecule Dynamics. *Phys Rev Lett*. 2007;98: 128105. doi:10.1103/PhysRevLett.98.128105

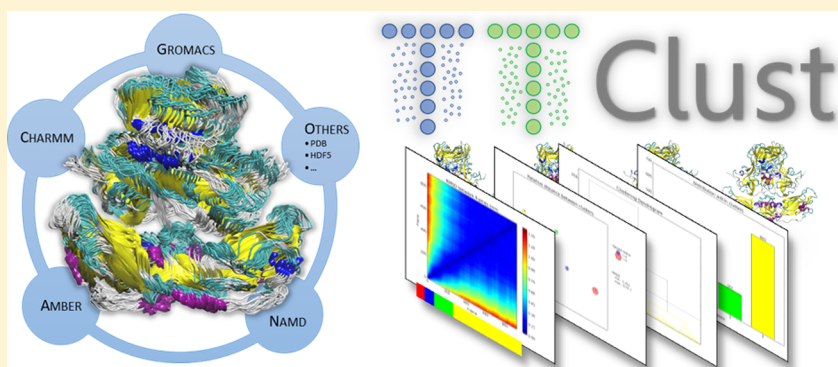
ANNEXES

TTClust: A Versatile Molecular Simulation Trajectory Clustering Program with Graphical Summaries

Thibault Tubiana,*¹ Jean-Charles Carvaillo, Yves Boulard, and Stéphane Bressanelli

Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, University Paris-Sud, Université Paris-Saclay, 91198 Gif sur Yvette cedex, France

 Supporting Information



ABSTRACT: It is extremely helpful to be able to partition the thousands of frames produced in molecular dynamics simulations into a limited number of most dissimilar conformations. While robust clustering algorithms are already available to do so, there is a distinct need for an easy-to-use clustering program with complete user control, taking as input a trajectory from any molecular dynamics (MD) package and outputting an intuitive display of results with plots allowing at-a-glance analysis. We present TTClust (for Trusty Trajectory Clustering), a python program that uses the MDTraj package to fill this need.

INTRODUCTION

Molecular dynamic simulations can produce a very large amount of data and routinely yield several thousand snapshots of large biomolecular systems. However, differences between these snapshots may be minimal, especially in a short time scale. It is almost always desirable to ascertain which differences are only fluctuations of the system and which are different conformations with regard to the question at hand. Hierarchical clustering methods have been developed to group similar conformations and thus allow partitioning of the studied system into its different states during a molecular dynamics simulation.

However, the ready-to-use clustering tools available need much user intervention, particularly in the analysis of results. The USCF Chimera Trajectory Cluster tool¹ is compatible with many molecular simulation packages and outputs a useful graphical timeline. However, clustering is based on the NMRclust clustering algorithm² and the user cannot control the clustering degree which can lead to the creation of very small clusters. A VMD plugin,³ Clustering Tool, can also be used to clusterize molecular dynamics trajectories. It groups all members of each cluster, allowing their simultaneous visualization but provides no other automated graphical analysis. It does not implement any automatic clustering function either and the “cutoff” notion may seem unclear for noninsiders. With the Gromacs clustering tool,⁴ the user can set a

parameter to adjust the clustering degree but it is only compatible with the gromacs trajectory format or PDB format. Finally, the CPPTRAJ⁵ analysis toolkit from the AMBER simulation package^{6,7} is powerful, but usage and results are not straightforward for users not already familiar with the AMBER suite. Generally these available tools require subsequent analysis of results before questions like “Does the system successively explore different conformational states?” or “How large are differences within clusters compared to differences between clusters?” can be answered.

We describe a new user-friendly clustering tool derived from a methodology described in the MDTraj package.⁸ It is immediately informative as it directly takes as input most of the molecular dynamics simulation data formats and has capabilities for automatically adjusting the clustering degree and quickly visualizing key properties of clusters, such as their sizes, spreads, separation, and timeline of appearance.

DESCRIPTION

TTClust is a Python program (compatible with Python 2.7 or 3.4) which can be used in the command line or with a simple graphical user interface (GUI) and is executable in the three major operating systems (Linux/Mac/Windows).

Received: July 27, 2018

Published: October 17, 2018

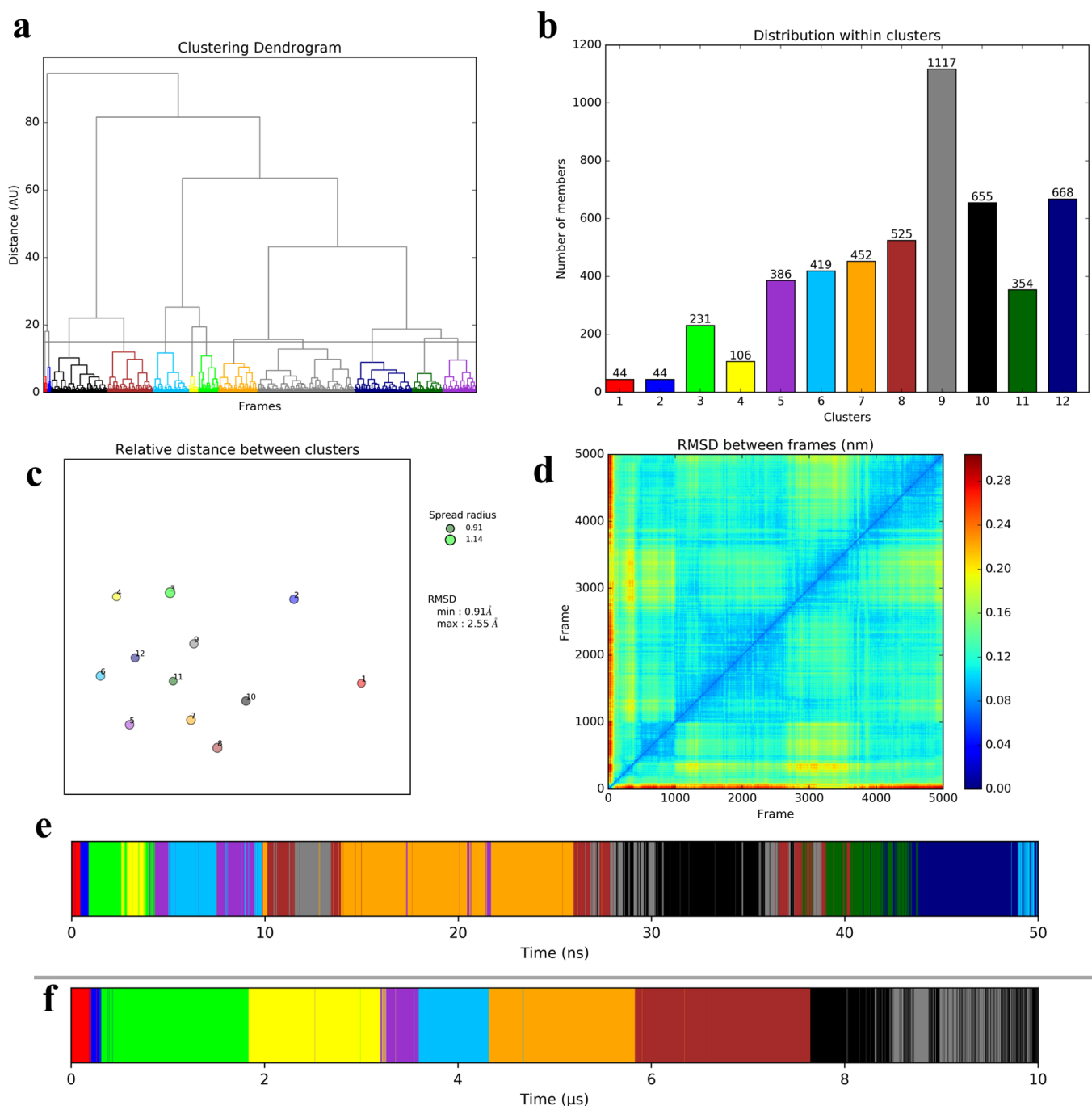


Figure 1. TTClust applied to two molecular dynamics simulations made with Gromacs 5.0. (a–e) Clustering for an all-atom protein simulation (50 ns) from PDB 4AEP. The cluster color code is the same between the dendrogram, histogram, 2D plot, and timeline (a–c, e). (a) If the automatic clustering mode is not chosen, the dendrogram helps the user to choose an appropriate clusterization level (here adjusted to yield 12 clusters). (b) Barplot numbers of frames within clusters. (c) 2D projection plot of the relative distances between clusters based on the RMSD between representative frames. The minimum and maximum distances are indicated, as well as the minimum and maximum spreads (average RMSD within clusters). (d) Distance matrix saved as a heatmap to allow a quick overview of the number of possible clusters. (e) Timeline barplot along the simulation. (f) Timeline barplot of clustering (10 clusters) for a 10 μ s coarse-grained simulation of a pentamer of Norovirus VP1 dimers (generated from PDB 1IHM¹³).

Thanks to the MDTraj package⁸ TTClust is able to deal with the trajectories of major simulation programs like Gromacs,⁴ AMBER,⁶ CHARMM,⁹ and NAMD¹⁰ and any trajectories in the PDB format. Furthermore, TTClust is fully compatible with the two major classes of biomolecules used in molecular dynamics simulations, namely proteins and nucleic acids (both DNA and RNA). The trajectory is first aligned on

the protein backbone (default) or any user-defined part of the structure (for instance in cases where there is no protein such as nucleic acid trajectories). The RMSD between all pairs of frames is calculated and stored into a matrix for the backbone (default) or for a user-defined part that can be different from the selection used to align the trajectory (e.g., to focus on the mobility in a particular part with respect to the rest of the

structure). This matrix is then used to calculate a linkage matrix by the hierarchical cluster linkage function of the SciPy package.¹¹ Several algorithms can be used during the linkage matrix calculation and can be defined by the user: single, complete, average, weighted, centroid, median, and ward.¹² The last one is the default method because it minimizes the variance within clusters and allows more demarcated clusters to be obtained.

A dendrogram is generated and then plotted with the Matplotlib package¹⁴ to have a graphical representation of the linkage matrix and see relations between all frames (Figure 1a). This allows the user to have an immediate view of the number of clusters at a given distance cutoff value for this particular trajectory, for instance to choose the cutoff value to create a required number of clusters. A number of desired clusters can also be defined directly by the user, and in this case the scipy maxclust algorithm is used on the linkage matrix to segregate at best all frames in the appropriate number of clusters (replaces the cutoff selection). Finally, the easiest way is to let the program compute the number of clusters automatically. The underlying algorithm is based on the “elbow” method with k-means,¹⁵ as outlined at <http://www.nbertagnolli.com/jekyll/update/2015/12/10/Elbow.html>: The average distance W_k between members inside a cluster decreases as the number of clusters k increases. Plotting W_k as a function of k , one looks for the number at which this curve flattens out. Practically, the trajectory is clusterized into $k = 2, 3, \dots, 15$ clusters, and W_k computed as a function of k . The normalized angle between consecutive segments of this curve is computed. The chosen number of clusters is the first value of k where this normalized angle is over 0.99.

Importantly, a color code is then assigned to each cluster and a histogram of the cluster sizes in order of appearance in the simulation (Figure 1b), as well as a timeline barplot of the trajectory, are generated with each successive frame assigned the color of its cluster (Figure 1e). Up to 12 clusters the colors are defined within the code so as to achieve good contrast and hence maximum readability. Beyond this value, colors are generated with a color gradient.

For each cluster, the spread (average RMSD between all pairs of frames in the cluster) is calculated and the representative frame (frame with the lowest average RMSD to all other frames in the cluster) is output. Distances between clusters are computed as RMSD between pairs of representative frames. These two crucial pieces of information are displayed in a 2D projection graph (Figure 1c) obtained with the multidimensional scaling algorithm¹⁶ within the scikit-learn Python module.¹⁷ The radius of each circle is relative to the cluster's spread and orientation of the graph is random.

All cluster data (members, representative frame, spread, separations between clusters) are written into a logfile as well as all parameters used in the program (cutoff, algorithm, files, topology, program version). Representative frames are also saved as aligned PDB files for ready visualization.

■ CASE STUDY

We applied this tool on two molecular dynamics simulations: one all-atom on the hepatitis C virus polymerase (NSSB) from PDB ID 4AEP¹⁸ and the other a Martini¹⁹ coarse grained system on a pentamer of dimers of the Norovirus capsid protein generated from PDB ID 1IHM.¹³ The clustering of both of them was made with the default parameters.

In the case of the NSSB simulation (all-atom, 100 ns), it can be seen at once that the first four clusters appear only at the beginning (Figure 1e), are most sparsely populated (Figure 1b) and lead to a conformational space segregated from the initial cluster (Figure 1c). Thus, the first few nanoseconds can be readily interpreted as a relaxation period, after which the protein explores a restricted conformational space. Indeed the final cluster 12 is close to cluster 6 (Figure 1c), explaining how the simulation can come back to cluster 6 at the end. The logfile (see the Supporting Information) gives a more precise and complete breakdown of this analysis.

In the case of the second system (coarse-grained, 10 μ s), the timeline barplot shows at a glance that there is a distinct direction all through the simulation (Figure 1f). The 10 clusters appear and disappear without coming back, except for the last two that alternate at the end of the simulation.

Thus, our tool allows quick and easy differentiation between two kinds of molecular dynamics simulations with (a) a (slowly evolving) system that after relaxation mainly cycles through previously explored conformations during the simulation and (b) a system which evolves directionally without returning to a conformation already seen. In all cases, major properties of the clustering's results are readily grasped thanks to the graphical summaries.

■ TTCLUST BASICS

TTclust can be used either from the command line or with a simple GUI thanks to the *goovey* python package based on the *wxpython* layout python package. There are three types of arguments and tuning parameters:

- (i) File arguments: trajectory and topology selection (required). The logfile name (optional, if given will also be the output folder).
- (ii) Selection arguments (optional): TTclust uses the MDTRAJ selection language (described on the MDTRAJ website, http://mdtraj.org/latest/atom_selection.html). Additionally, we add in TTclust keywords for DNA/RNA for easier use with nucleic acid and protein/nucleic acid complex simulations (“dna”, “rna”, “backbone_na”, “base_rna”, “base_dna”). These keywords are detected, replaced by more complex MDTRAJ selection syntax string and reinjected in the user selection string. Selection arguments allow selecting the atoms for alignment and for the distance matrix calculation (which is the basis of this clustering method) or extracting a part of the dynamics,
- (iii) Clustering arguments (optional): users can easily tune clustering methods (as described before, single, complete, average, weighted, centroid, median, and ward (default)) and the clustering cutoff method. It can be graphically based on the dendrogram of the hierarchical distance between frames, a number of groups desired, or the autoclustering algorithm (default with command line usage).

■ USAGE

Users can run TTclust with the GUI by executing `ttclust-GUI.py`. Arguments are categorized in tabs, and the standard output is redirected on the GUI (Figure S1). In the command line, the two minimal arguments are the trajectory file and the topology file and TTclust will be executed in automatic mode with autoclustering (eg: “`ttclust -f traj.nc -t topol.pdb`”).

The GUI, as well as the command line usage, generates a log file where all information shown in console is written (Figure S1). The users can find when it was generated, with which version, and the command line used (or equivalent command line if started from the GUI) which allows users to restart the calculation easily.

COMPUTATIONAL TIME

To compare resource consumption and computational times of some current clustering tools, clustering was performed with the same workstation (2 × 6-core Intel(R) Xeon(R) CPU E5-2630 v2@2.60 GHz) on the same 10 000 frame coarse-grained trajectory of 11 350 grains (see Table 1).

Table 1. Computational Times and Resource Consumption of Clustering Software

software	time	maximum memory allocated	maximum number of cores used
TTClust	1398 s	4.524 G	24 ^a
VMD clustering	1561 s	1.498 G	24
Gromacs gm_x_cluster	8318 s	1.529 G	1 ^b
Chimera	79596 s	22.058 G ^c	1

^aParallel processing by MDTRAJ during the computation of the distance matrix. ^bgm_x_cluster uses a single core for computation of the distance matrix, even in the MPI version of gm_x. ^cThe software crashed after computing clusters but before outputting results due to a memory allocation error.

The tools were all used with default parameters. In this case only TTClust yielded a usable result: VMD and gm_x_cluster do not perform automatic estimation of a suitable cutoff, and in this case their default of 1 Å proved too low and yielded 10 000 clusters. In this first run TTClust and VMD outperform other software tested, mostly because MDTRAJ (for TTClust) and “measure cluster” (for VMD) use multiple cores for computation of the distance matrix, that is by far the most demanding step. It is noteworthy that TTClust saves the distance matrix so that in a second run with the same atom selections but, e.g., a different number of clusters the calculation time decreases to less than a minute (Table 2).

Table 2. TTClust Computational Times and Resource Consumption by Trajectory Sizes and in Second Runs

frames	time		maximum memory allocated	maximum number of cores used	
	first run (auto)	second run ^a		first run	second run
1000	21 s	14 s	0.374 G	24	1
5000	320 s	35 s	1.591 G	24	1
10000	1430 s	50 s	3.058 G	24	1

^aA second cutoff was used with the previously saved distance matrix from the first run.

Another way to speed up calculations is to reduce the size of the trajectory, e.g., the number of frames. Thus, in separate runs TTClust clustered again the same 10 000-frame trajectory in 1430 s, in 320 s when it was reduced to 5000 frames, and in 21 s for a 1000 frame trajectory. Therefore, TTClust is useful for such trajectories but will become impractical for much larger ones.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.8b00512.

TTClust GUI description and log file produced by TTClust (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: tubiana.thibault@gmail.com (T.T.).

ORCID

Thibault Tubiana: 0000-0002-6490-4602

Author Contributions

T.T., Y.B., and S.B. conceived the research. T.T. wrote the program. T.T., J.-C.C., and S.B. completed testing and bug corrections. All authors analyzed the data. T.T. wrote a draft of the paper. T.T., Y.B., and S.B. wrote the final version.

Funding

This work was supported by a grant from Région Ile-de-France AAP DIM MALINF 2014 (maladies infectieuses) to S.B. including doctoral funding for T.T.

Notes

The authors declare no competing financial interest.

TTClust is freely available on github under the GPLv3 license at <https://github.com/tubiana/TTClust>.

ACKNOWLEDGMENTS

We thank Kaouther Ben Ouirane for extensive testing of TTClust. Simulations used in this work were performed through access to the high performance computing resources of TGCC under the allocation A0010707583 made by GENCI. The authors also thank the CEA-CCRT infrastructure for giving us access to the supercomputer COBALT.

REFERENCES

- Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605–1612.
- Kelley, L. A.; Gardner, S. P.; Sutcliffe, M. J. An Automated Approach for Clustering an Ensemble of NMR-Derived Protein Structures into Conformationally Related Subfamilies. *Protein Eng., Des. Sel.* **1996**, *9* (11), 1063.
- Gracia, L. *VMD Clustering Tool*, 2017.
- Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25.
- Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9* (7), 3084–3095.
- Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a Package of Computer Programs for Applying Molecular Mechanics, Normal Mode Analysis, Molecular Dynamics and Free Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. *Comput. Phys. Commun.* **1995**, *91* (1), 1–41.
- Case, D.; Babin, V.; Berryman, J.; Betz, R.; Cai, Q.; Cerutti, D.; Cheatham, T.; Darden, T.; Duke, R.; Gohlke, H.; Goetz, A.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A.; Lee, T.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K.; Paesani, F.; Roe, D.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.

Seabra, G.; Simmerling, C.; Smith, W.; Swails, J.; Walker, Wang, J.; Wolf, R.; Wu, X.; Kollman, P. *Amber 14*; 2014.

(8) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109* (8), 1528–1532.

(9) Brooks, B. R.; Brooks, C. L.; MacKerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Cafilisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30* (10), 1545–1614.

(10) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26* (16), 1781–1802.

(11) Jones, E.; Oliphant, T.; Peterson, P.; et al. *SciPy: Open Source Scientific Tools for Python*; 2001.

(12) Ward, J. H., Jr. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **1963**, *58*, 236.

(13) Prasad, B. V. V.; Hardy, M. E.; Dokland, T.; Bella, J.; Rossmann, M. G.; Estes, M. K. X-Ray Crystallographic Structure of the Norwalk Virus Capsid. *Science* **1999**, *286* (5438), 287–290.

(14) Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9* (3), 90–95.

(15) Tibshirani, R.; Walther, G.; Hastie, T. Estimating the Number of Clusters in a Data Set via the Gap Statistic. *J. R. Stat. Soc. Ser. B-Stat. Methodol.* **2001**, *63*, 411–423.

(16) Kruskal, J. B. Nonmetric Multidimensional Scaling: A Numerical Method. *Psychometrika* **1964**, *29* (2), 115–129.

(17) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D. Scikit-Learn: Machine Learning in Python. *J. Machine Learning Res.* **2011**, *12*, 2825–2830.

(18) Scrima, N.; Caillet-Saguy, C.; Ventura, M.; Harrus, D.; Astier-Gin, T.; Bressanelli, S. Two Crucial Early Steps in RNA Synthesis by the Hepatitis C Virus Polymerase Involve a Dual Role of Residue 405. *J. Virol.* **2012**, *86* (13), 7107–7117.

(19) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4* (5), 819–834.

Linking Bisphenol S to Adverse Outcome Pathways Using a Combined Text Mining and Systems Biology Approach

Jean-Charles Carvaillo,^{1,2} Robert Barouki,^{1,2,*} Xavier Coumoul,^{1,2,*} and Karine Audouze^{1,2,3,*}

¹University Paris Descartes, ComUE Sorbonne Paris Cité, Paris, France

²Institut national de la santé et de la recherche médicale (INSERM, National Institute of Health & Medical Research) UMR S-1124, Paris, France

³University Paris Diderot, ComUE Sorbonne Paris Cité, Paris, France

BACKGROUND: Available toxicity data can be optimally interpreted if they are integrated using computational approaches such as systems biology modeling. Such approaches are particularly warranted in cases where regulatory decisions have to be made rapidly.

OBJECTIVES: The study aims at developing and applying a new integrative computational strategy to identify associations between bisphenol S (BPS), a substitute for bisphenol A (BPA), and components of adverse outcome pathways (AOPs).

METHODS: The proposed approach combines a text mining (TM) procedure and integrative systems biology to comprehensively analyze the scientific literature to enrich AOPs related to environmental stressors. First, to identify relevant associations between BPS and different AOP components, a list of abstracts was screened using the developed text-mining tool AOP-helpFinder, which calculates scores based on the graph theory to prioritize the findings. Then, to fill gaps between BPS, biological events, and adverse outcomes (AOs), a systems biology approach was used to integrate information from the AOP-Wiki and ToxCast databases, followed by manual curation of the relevant publications.

RESULTS: Links between BPS and 48 AOP key events (KEs) were identified and scored via 31 references. The main outcomes were related to reproductive health, endocrine disruption, impairments of metabolism, and obesity. We then explicitly analyzed co-mention of the terms BPS and obesity by data integration and manual curation of the full text of the publications. Several molecular and cellular pathways were identified, which allowed the proposal of a biological explanation for the association between BPS and obesity.

CONCLUSIONS: By analyzing dispersed information from the literature and databases, our novel approach can identify links between stressors and AOP KEs. The findings associating BPS and obesity illustrate the use of computational tools in predictive toxicology and highlight the relevance of the approach to decision makers assessing substituents to toxic chemicals. <https://doi.org/10.1289/EHP4200>

Introduction

Integrative computational approaches that combine systems biology and toxicology can increase our understanding of the links between environmental chemical exposure and human health. Systems biology and advanced bioinformatics tools generate new hypotheses. They furnish new insights into and predictions of biological mechanisms induced by chemical substances, including drugs and environmental pollutants. Compelling evidence indicates that a number of chemical substances may play a causative role in diseases (Heindel and Blumberg 2018). Computational sciences, including systems toxicology, can speed up the identification of linkage between adverse outcome pathways (AOPs) and a chemical stressor as well as its effects on health (Ankley et al. 2010).

The concept of an AOP was originally proposed by Ankley et al. (2010). AOPs integrate various key events (KEs) to connect biological perturbations, at the molecular or cellular levels, to toxicity events [i.e., adverse outcomes (AOs)] at organismal and population levels. The use of clearly identified AOPs for decision-making is part of a global methodological initiative, which has, among its goals, the reduction of animal use in toxicity testing. AOPs are expected to be used more and more in regulatory frameworks since

they provide evidence-based mechanistic insights (Bopp et al. 2018). The AOPs, which have been identified, are stored in the AOP-Wiki online database (SAAOP 2016). The database is part of a collaborative program that involves the Organisation for Economic Co-operation and Development (OECD) and the European Commission. The AOP knowledge database (AOP-KB) is another tool from the OECD program for AOP development, to support and share information to the scientific community and harmonize the format of generated novel AOP (OECD). All the terms defined in the AOPs are standardized according to structured ontologies (Ives et al. 2017).

Although the development of AOPs has a great potential to address existing knowledge gaps, AOP development and assembly is laborious and time-consuming, since extensive toxicity data need to be gathered. Much of the information that is accumulating derives from omics technologies, high-throughput testing with robots [ToxCast (U.S. EPA) (Judson et al. 2010)], and novel databases derived from the compilation of heterogeneous information such as the Comparative Toxicogenomics database (CTD) (Davis et al. 2018). Therefore, the development of innovative computing methodologies that allow the prioritization of chemicals according to their inferred threats is highly relevant both for the research community and for health agencies (Richard et al. 2016; Thomas et al. 2013). Such *in silico* methods that use available data sources also can accelerate the description of new AOPs and provide integrated data to increase the information content of existing AOPs (Berggren et al. 2015; OECD 2014).

The breadth of the currently available scientific literature and diversity of synonyms for chemicals complicates meaningful integration of the information. Thus, it can be difficult to completely and accurately acquire the information on a selected topic, even if specific databases related to a given field have been compiled and information stored [e.g., the Developmental and Reproductive Toxicology database, DART (NIH and TOXNETb)]. In addition to enriching toxicological databases, there is a need for tools allowing better exploration of available databases (including available published literature), and to improve text mining (TM) is such a way to facilitate the establishment of links between a chemical and relevant AOP components. Such tools should be able to explore a wider range

*These authors contributed equally to this work.

Address correspondence to Karine Audouze, University Paris Descartes, INSERM UMR-S 1124, 45 rue des Saints-Pères 75006, Paris, France. Telephone: +33 142 864 010. Email: Karine.audouze@univ-paris-diderot.fr

Supplemental Material is available online (<https://doi.org/10.1289/EHP4200>).

The authors declare they have no actual or potential competing financial interests.

Received 19 July 2018; Revised 1 March 2019; Accepted 19 March 2019; Published 17 April 2019.

Note to readers with disabilities: *EHP* strives to ensure that all journal content is accessible to all readers. However, some figures and Supplemental Material published in *EHP* articles may not conform to 508 standards due to the complexity of the information being presented. If you need assistance accessing journal content, please contact ehponline@niehs.nih.gov. Our staff will work with you to assess and meet your accessibility needs within 3 working days.

of data and have the potential to prioritize the chemical–health outcome connections. We describe here a strategy that integrates a new tool called AOP-helpFinder version 1.0, downloadable on github (<https://github.com/jecarvaill/aop-helpFinder>). Using the available literature, AOP-helpFinder can automatically find, extract, and score links between chemical substances (i.e., stressors) and diverse biological elements, i.e., molecular initiating events (MIEs), KEs, and AOs, which are the components of AOPs (Villeneuve et al. 2014). The novelty of AOP-helpFinder is that it consists of a hybrid approach that combines TM and graph theory to explore the contents of abstracts for the identification of reliable associations between a chemical and AOPs. The main objective of AOP-helpFinder is to assist toxicologists and biologists in the identification of relevant associations between AOP components and small molecules through the analysis of large-scale, existing (published in peer-review journals and databases), text-based knowledge (*in vitro*, *in vivo*, and *in silico* data). As a proof of concept, we applied AOP-helpFinder to bisphenol S (BPS), a structural analog of bisphenol A (BPA) that is suspected to have endocrine-disrupting properties (Karrer et al. 2018). The biological mode of action (MoA) and potential toxicity of BPS are still poorly characterized. Using our computational method, links between BPS and AOP components have been uncovered.

Methods

A workflow of the strategy is shown in Figure 1.

Data Input: Data Description and Preprocessing

Development of the adverse outcome pathway dictionary. In order to explore associations between AOP-related terminologies (e.g., “obesity”) and a term of interest (here a chemical), an AOP dictionary that includes AOP events, i.e., MIEs, KEs, and AOs (Karrer et al. 2018), was generated. From the AOP-Wiki database, we downloaded the available .xml file that contains all AOP information (aop-wiki-xml-2017.gz, 3 July 2017). We extracted the AOP identifier (for example, “72”), the key event name (i.e., “obesity”), the key event identifier (“1447” for obesity), and the key event type (that is, whether the term is a MIE/KE or AO; for example, adipogenesis is defined as an AO). The AOP dictionary contained two files: one containing the AO names and one with the MIE/KE names. MIEs and KEs were combined into one unique table.

Development of the disease dictionary. Controlled disease vocabulary from the U.S. National Library of Medical Subject Headings (MeSH, <https://www.nlm.nih.gov/mesh/meshhome.html>) was used, which represented 11,850 disease terms (downloaded from the CTD database as of September 2017) (Davis et al. 2017).

Development of the bisphenol S dictionary. According to the data sources, BPS is identified by different terms for its name, synonyms, and Chemical Abstracts Service (CAS) registry number. In order to capture, as fully as possible, the existing biological and toxicological information related to BPS, chemical terms were retrieved using the PubChem database (NIH) (Table S1).

Text-based toxicological data. To compile abstracts linked to BPS, we used an integrative approach that consisted of manual searches of five specific toxicological databases; some from the U.S. Toxnet platform [the Chemical Carcinogenesis Research Information System (CCRIS) (NIH and TOXNETa), DART, Toxicology Literature Online (TOXLINE) (NIH and TOXNETd), and Hazardous Substances Data Bank (HSDB) (NIH and TOXNETc) databases, which are collections of publicly available information] and the Registry of Toxic Effects of Chemical Substances (RTECS) database (Biovia). For each database, using the BPS dictionary, we extracted information concerning abstracts,

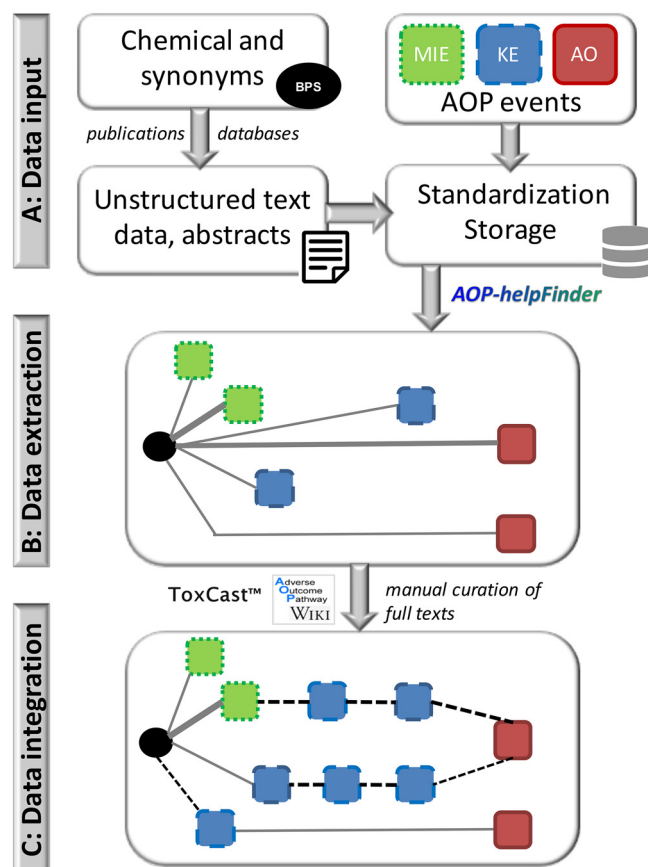


Figure 1. Workflow of the strategy for linking bisphenol S (BPS) to health effects. (A) Data input: we compiled *a*) published data from multiple sources related to BPS and its associated toxic effects, and *b*) adverse outcome pathway (AOP) events to prepare a dictionary of AOP events (standardization and storage in an in-house database), which includes molecular initiating events [MIEs (squares with small dots)], key events [KEs (squares with large dots)], and adverse outcomes [AOs (square with solid lines)] as defined in the AOP-Wiki database. (B) Data extraction: we performed preprocessing followed by text mining of the data using the developed AOP-helpFinder tool in order to identify links between BPS (black circle) and different AOP events (MIE, KE, and AO). Scores, represented by the width of the edges, were calculated to indicate the confidence of the association. (C) Data integration: To fill potential gaps between two nodes (such as a KE and an AO), existing AOP information was integrated into the network as well as information from the ToxCast database. Full-text manual curation of the identified publications (by text mining) was performed by experts, thus increasing the availability of knowledge (dashed nodes/edges lines).

authors, journal, year of publication, title, toxic effects, targets, and species. According to the databases, these data were mentioned under various fields. To avoid the duplication of information (since the same article can be present in several data sources) and subsequent overestimation of MIE/KE/AO, the parsing of extracted data was necessary since the toxicological data are not collected under the same field in the different data sources. For example, in the CCRIS format, the field “Target data” was found, and in the RTECS source, the field “Toxic Effects” was found. Therefore, to facilitate further analysis, three fields that contained toxicological information were retained to be screened against AOPs. These three fields were “Abstract,” “Toxic effects,” and “Target data,” since MIE, KE, and AO can be found in all these fields. The DART database provided more than >400,000 journal references related to teratology and other aspects of developmental and reproductive toxicology. The CCRIS database contained data curated from published studies that possess chemical records that have carcinogenicity,

mutagenicity, tumor promotion, and tumor inhibition test results. TOXLINE is a database based on bibliographic data, which includes specialized journals. It provided references from diverse fields, including toxicological effects of environmental chemicals. The HSDB is a database focusing on the toxicology of potentially hazardous chemicals. The RTECS database is a comprehensive collection of basic toxicity information that includes various types of chemicals such as drugs, food additives, and environmental pollutants. It covers six categories of toxicity data: acute toxicity, tumorigenicity, mutagenicity, skin and eye irritation, reproductive effects, and multiple dose effects.

Development of a relational database for storing data. All selected information were stored in an in-house database to facilitate further analyses. Data were entered once in a relational SQLite, version 3.27.2, (<https://www.sqlite.org/index.html>) database to avoid any redundancy. The architecture of the database consisted of twelve tables, in which the previously compiled information [database source, first author, journal, year of publication, title, data text (abstract, target text, and toxic effects information), related animal and sex information, and MIE, KE, and AO data linked to AOP] were stored and connected together.

Data Extraction: The AOP-helpFinder Tool

To identify reliable associations between MIE/KE/AO and BPS, we developed a new method called AOP-helpFinder, which is based on Natural Language Toolkit (NLTK, version 3.2.5) (<https://www.nltk.org/>) and on Dijkstra graph theory (Dijkstra 1959). NLTK is a leading platform for building programs that employ human language data. It contains a suite of libraries and small programs for symbolic and statistical natural language processing under the Python programming language. NLTK initially was created for fields such as cognitive science, artificial intelligence, and machine learning, to mention a few areas. Dijkstra's algorithm identifies the shortest path between two nodes in a graph, and in the AOP-helpFinder, it was used to determine the shortest path between words (by computing the distance between the terms, for example between a substance and an AOP event in an abstract). The AOP-helpFinder is a multistep TM procedure, consisting of the following:

1. data preprocessing

To maximize the chances of matching MIEs, KEs, and AOs in text data, it was necessary to clean and simplify them (Figure 2). This consisted of filtering out the noise and stemming the words remaining (the stem is the root or main part of a word to which inflections or formative elements are added), as well as considering spaces, conjunction, and punctuation. For example, for AOP 7, the text "Aromatase (Cyp19a1) reduction leading to impaired fertility in adult females" was simplified to "aromata cyp19a1 reduct lead impair fertil adult female." This multistep preprocessing consisted of *a*) dividing an abstract into sentences, *b*) splitting all sentences into words, *c*) removing sentences that contain a negation word (never, neither, no, not, did not, hasn't, should not, ...) to reduce the risk of false positives, *d*) deleting stop words (coordinating conjunction, punctuation, most common words) from sentences, and *e*) stemming the remaining words. As a result, we obtained stemmed data for information related to AOPs and text data related to the chemical of interest ready to be screened against the AOPs' data stored into the database.

2. identification of association and scoring function

Two different approaches were developed to screen for AO and MIE/KE terms. As AO is more related to a few words and MIE/KE to complex sentences, the approaches were based on the calculation of scoring functions that took into

consideration the complexity of the terms/sentences related to AO/KE/MIE.

Matching AO data. The stemmed AO were screened in the "abstracts," "targets," and "toxic effects" fields. Associations between chemicals and AO are likely to have different meanings if they are found in the beginning or at the end of an abstract. Generally, in published studies, a working hypothesis is found at the beginning of an abstract. However, chemical-AO associations co-mentioned at the end of an abstract are more likely to be considered as true positives because results and findings are often cited at the end of an abstract. Therefore, in order to differentiate working hypothesis from findings, we calculated a score based on the position (pS) using $pS = AO_{index} / L_{TMA}$. Both the position of the AO term relative to the other words (AO_{index}) and the total number of words in the text mined abstract (L_{TMA}) are taken into consideration. An optimal value is around 1.0, which corresponded to the last position in the abstract. The more the AO is placed toward the end of the abstract, the more it can be considered as a result and not a working hypothesis.

Matching molecular initiating event and key event data. On the average, a MIE/KE consisted of four terms in the AOP-Wiki database. Fifty-four percent of the MIE/KE contained at least four terms (for example, "allergic contact dermatitis challenge"), and among these, 21% were composed of exactly four terms. Moreover, 40% of the MIEs/KEs had between four and six terms. The minimum number of terms was one (such as "obesity"), and the maximum was 21 terms [KE 1,119: failure gamma glutamyl carboxylation glutamine residues clotting factors ii vii ix x under carboxylation clotting factors (gamma carboxy prothrombins)]. Based on these data, when a MIE/KE was composed of >three terms, we assumed that 25% of the information could be missing (for example, for "allergic contact dermatitis challenge," we retrieved "allergic contact dermatitis"), whereas for a MIE/KE having <four, we assumed that every term was important. Therefore, we applied a threshold of 75% of stemmed terms in order to identify relevant MIE/KE in the screened abstracts. The matching process continued under these conditions; otherwise, it was stopped.

The search for MIE/KE in an abstract was based on the creation of acyclic weighted graphs that use the positions of stemmed terms. The edges characterizing these graphs had a weight that was the distance between two nodes (positions of two different terms). Given the notion of weight on edges, we decided to use Dijkstra's algorithm because it supports finding the shortest path in a weighted acyclic graph (Dijkstra 1959). The acyclic property was composed of a step-by-step process, starting from the term n to the term $n + 1$, without returning to the n th term (unidirectional). Since the graphs that are formed are acyclics, the algorithmic complexity was solved by a maximum of $O(n^2 + m)$ calculations, where n represents the number of nodes, and m the number of edges in a graph O . Dijkstra's theory is represented by a quadratic complexity, i.e., $O(n^2)$. Considering all the existing MIE/KE, which have an average of four terms, the use of Dijkstra's algorithm resulted in a faster calculation compared with the use of a table. Therefore, the weight of the edges was used to find the shortest paths of the stemmed terms. The nearer the stemmed terms are to one another, the lower the total weight is. In order to identify a MIE/KE in texts and to classify the MIE/KE scores, we calculated a weighted score (wS) using $wS = (\min_{TW} + 1) / L_{MK}$. The wS is based on the lowest total weight calculated using graph theory (\min_{TW}) and the total number of words in a MIE or KE (L_{MK}). The ideal score is 1.0. At this score, all the stemmed terms are present and close to one another; that is, the expression/sentence found in the abstract is very similar to the MIE/KE of the reference (AOP-Wiki). The more the score varies from 1.0 (higher or lower), the less the MIE/KE is likely. Scores

the request's result can be saved in a non free format file. these files are the basis and input of aop-helpfinder module. unfortunately, information in these files are not always organized and structured similarly.

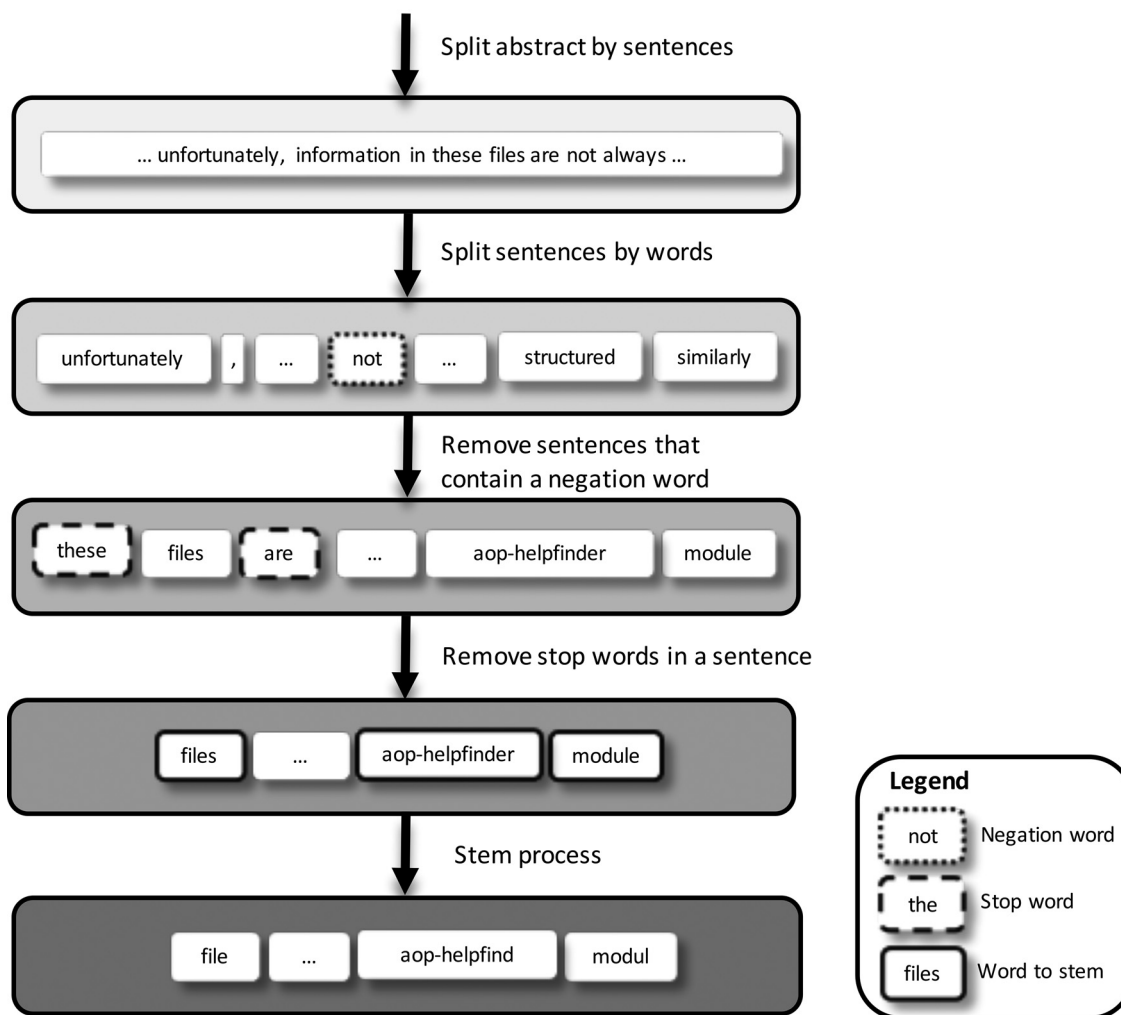


Figure 2. Data preprocessing in the AOP-helpFinder tool. Before identifying and scoring associations between a chemical and an adverse outcome pathway (AOP) event, text information collected from multiple sources (publications, databases) was preprocessed in order to obtain stemmed data, which were stored in a relational in-house database. Negation words (never, neither, no, not, did not, hasn't, should not, ...) were identified with squares surrounded with small dots, stop words (coordinating conjunctions, punctuations, most common words such as "the") with squares surrounded with large dots, and "words to stem" with solid lines (e.g., files → file or finder → find, considering that a word has a single stem, namely the part of the word that is common to all its inflected variants).

<1.0 indicate that there are missing MIE/KE stemmed terms in a sentence of an abstract. Scores >1.0 indicate that MIE/KE stemmed terms are far from each other. Scores <0.0 indicate that stemmed terms are missing (if a stemmed term is missing, a minus sign is added).

Performance of the AOP-helpFinder. Manual curation of the abstracts was used to assess the ability of the AOP-helpFinder to find relevant associations.

Data Integration: Linking a Chemical to Health Effects

Integrating known adverse outcome pathways and ToxCast information. A biological network of the key AOP event names for BPS was developed using the events (MIE, KE, and AO) identified with the AOP-helpFinder. To fill potential gaps between a MIE and a KE, or a KE and an AO, integration of an updated version of the AOP-Wiki (23 April 2018) was performed. We also integrated additional evidence of molecular events related

to BPS using the ToxCast database (April 2018). In this 2018 version, ToxCast provided data for 9,076 chemicals, tested on 359 assays, and information for 1,192 high-throughput endpoint components.

Integrating published data by manual curation. AOP-helpFinder identified relevant publications. To estimate their toxicological relevance as well as the importance of the experiments carried out for BPS, scientific experts manually analyzed the abstracts of the publications. For example, articles that used BPS to introduce a study or as a reference molecule in a study that involved other compounds were not taken into consideration for further analysis. Following this first manual curation, articles were classified according to the AOs to which they belong, and the corresponding full texts were read. This step highlighted the molecular events that were most often described, and particular attention was paid to those that are confirmed repeatedly by different articles. Based on the extracted results, AOP networks were suggested, and experts examined additional publications

(cited references in the selected publications) to identify potential indirect linkages between events (KEs).

Results

Data Preparation

We first had to retrieve the AOP terms on one hand and the BPS synonyms on the other hand. Annotations of AOPs were retrieved by downloading the full AOP-Wiki database. As of July 2017, there were 1,073 biological events (MIEs and KEs) and 61 defined AOs, some of which were under development, and others were included in the OECD work plan. These events were at different biological levels and ranged from the molecular (ID 18: Activation, AhR) to cellular (ID 52: decreased, calcium influx), tissue (ID 68: accumulation, collagen), organ (ID 161: increase, liver and splenic hemosiderosis), individual (ID 270: induction, sustained hepatotoxicity), and population (ID 417: skewed, sex ratio). To create the list of terms related to BPS (i.e., synonyms), we queried the PubChem database (Kim et al. 2019). We were able to retrieve 129 terms that include the MeSH vocabulary and depositors' names (see Table S1). For example, BPS has several synonyms such as bis(4-hydroxyphenyl)sulfone and 4,4'-sulfonyldiphenol, and has the CAS number of 80-09-1. Raw text data that mention BPS or synonyms were acquired from five toxicological databases (CCRIS, DART, TOXLINE, HSDB, and RTECS), and processed in an SQLite3 database. A total of 109 publications involving BPS and toxicological effects were identified (Table S2). Then, in order to identify co-occurrence between the BPS compound and the AOP terms in the 109 abstracts extracted from five databases, the BPS dictionary and the AOP dictionary were used.

Process of the AOP-helpFinder Tool

The central feature of the AOP-helpFinder was its ability to find associations between compounds (e.g., BPS) and AOP terms using a NLTK TM approach. After the calculation of scoring functions using the Djikstra graph theory, only the most relevant information to support linkage of a compound to an AOP was kept. As a result, the relationships between a chemical (i.e., a stressor) and a MIE, which is the initial component of an AOP (e.g., chemical binding to a receptor or a protein), were identified. Other types of associations also were uncovered between BPS, AOP KEs, and AOs, which are the ultimate component of an AOP that impacts health at an individual or a population level.

Performance of the AOP-helpFinder

To evaluate the performance of the method, we manually curated the 109 abstracts. We obtained a success rate of 76% with 85 true positive (existing MIE/KEs that were found) and 27 false positive (MIE/KEs found by the AOP-helpFinder, but not present in the 109 publications). The sensitivity of the method, i.e., the ability to find the right AOP related term, was 67%, with 41 false negatives (not found by AOP-helpFinder, but existing in at least one of the abstracts; for example, "reproductive defects" was not found but could be linked to the KE "decreased reproductive success").

The relevance of the position score was also evaluated. In other words, we evaluated the hypothesis that the position of the AO terms in the abstract relates to the context, that is, whether the statement was a working hypothesis or a finding. We manually checked the abstracts that co-mention BPS and at least one AO, and found a good correlation between the *pS* and the real position of the AO term. All the high *pS* scores (close to 1) were related to a result in an abstract, whereas low *pS* related to the

Table 1. List of the adverse outcomes (AOs) and molecular initiative and key events (MIEs and KEs) associated with bisphenol S (BPS).

Name	Score (<i>pS</i> or <i>wS</i>) ^a	AOP-Wiki ID
AO name		
Adipogenesis	0.96	16
Cancer	0.47	11
Hepatic steatosis	0.28	17
Steatosis	0.95	18
MIE/KE name		
Activated LXR	3	1,421
Activation androgen receptor	1.6	785
Activation estrogen receptor	4	1,181
Activation estrogen receptor alpha	-1	1,065
Activation glucocorticoid receptor	2	122
Activation hepatic nuclear receptor(s)	-2	1,157
Activation LXR	3	167
Activation oxidative stress pathway	-2.25	1,238
Allergic contact dermatitis challenge	-0.75	312
Alteration lipid metabolism	2.66	1,060
Altered gene expression	1.66	1,239
Apoptosis	1	1,262
Binding antagonist NMDA receptors	-1.5	201
Breast cancer	1	1,193
Chronic high-fat diet	-0.75	1,454
Decrease fertility	2	330
Decrease thyroid hormone synthesis	-0.75	277
Decreased androgen receptor activity	-1.25	742
Decreased body length	1.66	315
Decreased body weight	1.66	864
Decreased testosterone	1	808
Depletion of GSH	1	130
Estrogen receptor activation	4	1,180
Glucocorticoid receptor agonist activation	-1.5	494
Hippocampal gene expression altered	-1.25	756
Impaired development	1	577
Increase DNA damage	1	1,194
Increase lipid peroxidation	2.3	1,445
Increase plasma vitellogenin concentrations	-1.25	220
Increase reactive oxygen species production	-1	257
Increased DNA damage repair	-0.75	1,281
Increased lethality	8.5	342
Increased neuronal synaptic inhibition	-4	1,015
Increased reactive oxygen species	-0.75	1,115
Increased triglyceride	2	881
Liver fibrosis	5.5	344
Increased glucocorticoid receptor activity	-1.5	1,396
Necrosis	1	1,263
Obesity	1	1,447
Oxidative stress	1	210
Production reactive oxygen species	-0.75	249
Reduction testosterone level	2.3	446
ROS formation	1	1,278
Skewed sex ratio	1	417

Note: GSH, glutathione; LXR, liver X receptor; NMDA, N-methyl-D-aspartate; ROS, reactive oxygen species.

^aThe position score (*pS*) reflects the position of the AO term in the abstract. The weighted score (*wS*) allows the user to quantify the extent to which a MIE/KE expression is retrieved in an abstract.

introduction or a hypothesis in the abstract. For example, a *pS* of 0.95 was obtained for steatosis from the study of Héliers-Toussaint et al. (2014). Examination of the abstract revealed that BPS and steatosis were co-mentioned in the last sentence ("the findings suggest that both BPA and BPS could be involved in obesity and steatosis processes, but through two different metabolic pathways").

Linking Bisphenol S to Health Effects

The 109 publications that were obtained were filtered using AOP-helpFinder and only the ones mentioning associations between

AOP-related terms (MIEs, KEs, and AOs) and BPS were kept (see Table S2). Among the 109 publications, seven co-mention BPS and an AO term (e.g., “steatosis”), and 46 co-mention BPS and MIE/KE terms (for example, “activation estrogen receptor alpha” or “decrease body weight”). Very few AOP event duplicates were retrieved. Therefore, for further analysis, we kept a total of four unique AOs (for example, the AO “cancer” was retrieved in four of the seven publications that co-mention BPS and an AO term) and 45 unique MIE/KEs. We also removed the publications with the insignificant term “decrease” found among the MIE/KE, when it was alone, as it could not be mapped to a specific biological event. Thus, we ended up with four AOs and 44 MIEs/KEs (Table 1) present in 31 references (Table S2). To visualize these findings, we displayed bipartite networks using Cytoscape, version 3.5.1 (<https://cytoscape.org/>) (Figure 3). Among others, an association was found between BPS and “decrease thyroid hormone synthesis.” This corresponded to KE 277 in the AOP-Wiki database “Thyroid hormone synthesis, Decreased” with a score of 0.75. The most relevant health outcomes resulting from our study were reproductive effects (decreased testosterone, skewed sex ratio, decreased fertility), endocrine disruption (estrogen and androgen receptor activation, as observed biologically with BPA), and metabolism impairment (adipogenesis, increased triglycerides, obesity).

Among the latter, the most optimal score was associated with obesity (KE 1,447 in the AOP-Wiki database, belonging to the AOP 72) with a wS value of 1. Furthermore, obesity was mentioned in six of the 31 publications (Figure 3) (Table S2). Few published articles link directly BPS to obesity (Héliès-Toussaint et al. 2014); nevertheless, our approach allowed us to highlight this connection. Since BPS is a substitute for BPA, which is suspected to be an obesogen, the link between BPS and obesity that we found is of interest. The AOP-helpFinder tool allowed us to identify publications that mention a link between BPS and increased adipogenesis (ID 1,449 in the AOP-Wiki database). In order to have a more accurate assessment of the MoA leading to obesity, an integrative systems biology approach was used. We integrated information from current available sources such as the AOP-Wiki and ToxCast databases (as of April 2018). In the AOP-Wiki database, a MIE involving the peroxisome proliferator-activated receptor gamma (PPAR γ), “PPAR γ , activation” (ID 1028 in the AOP-Wiki database) was identified as being connected to other AOP terms such as “increase adipogenesis” and “obesity.” Indeed, PPAR γ activation is a known initiating event for increased adipogenesis (Lefterova et al. 2014). PPAR γ also is known to be involved in the regulation of adipocyte differentiation and has been implicated in several pathologies including obesity, diabetes, and cancer (Polvani et al. 2016). The investigation of

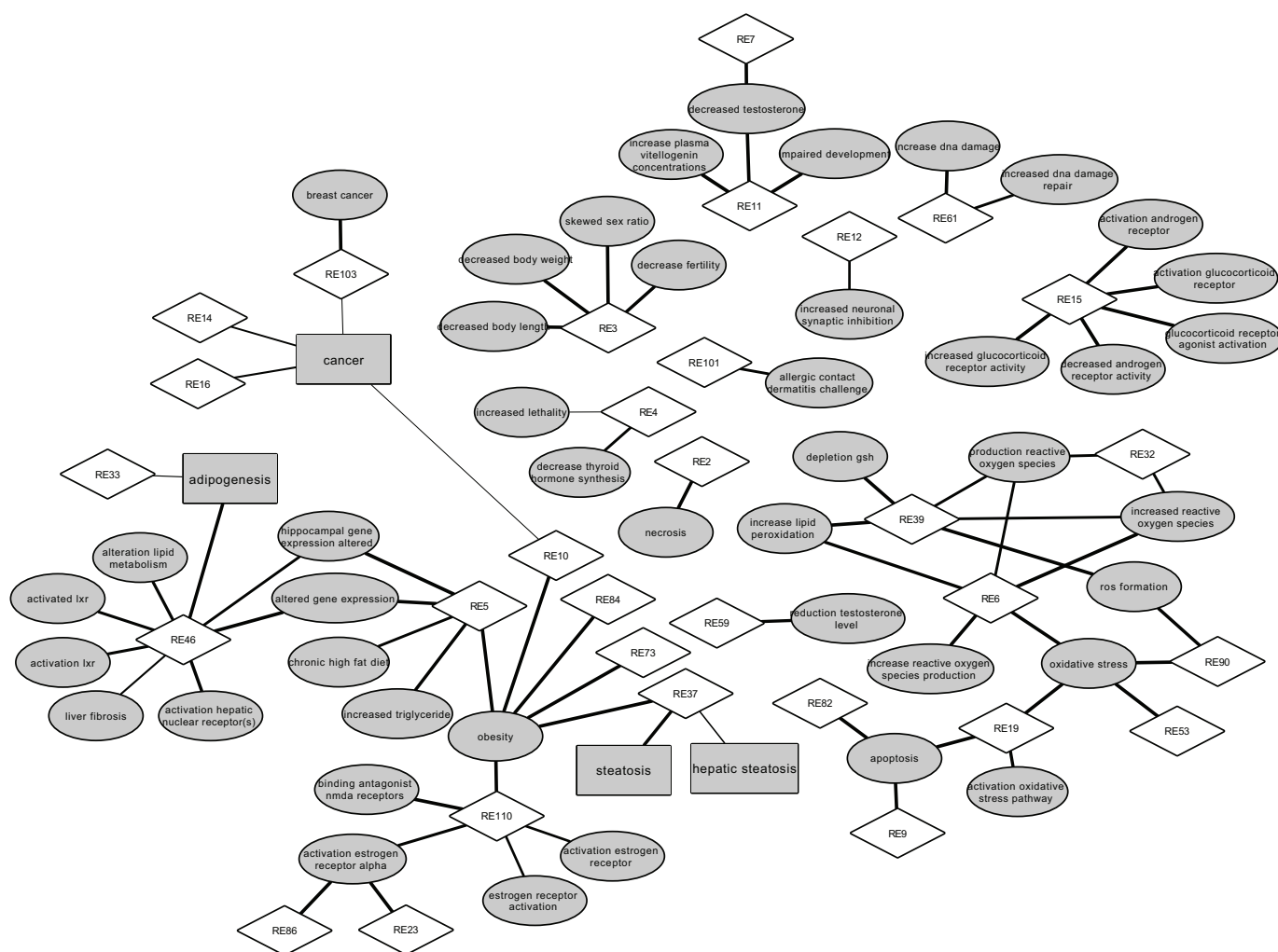


Figure 3. Bipartite networks of the relevant references and adverse outcome pathway (AOP) events for bisphenol S (BPS), identified by text mining. The bipartite networks consisting of 48 AOP terms: 4 AOs (rectangle), 44 molecular initiating events (MIEs)/key events (KEs) (ellipse), and 31 references (RE) (diamond) (see Table S2 for the corresponding references). The width of each AOP term–reference edge is proportional to the corresponding scores: position score (pS) for AO and weighted score (wS) for MIE/KE. Note: LXR, liver X receptor; NMDA, *N*-methyl-D-aspartate.

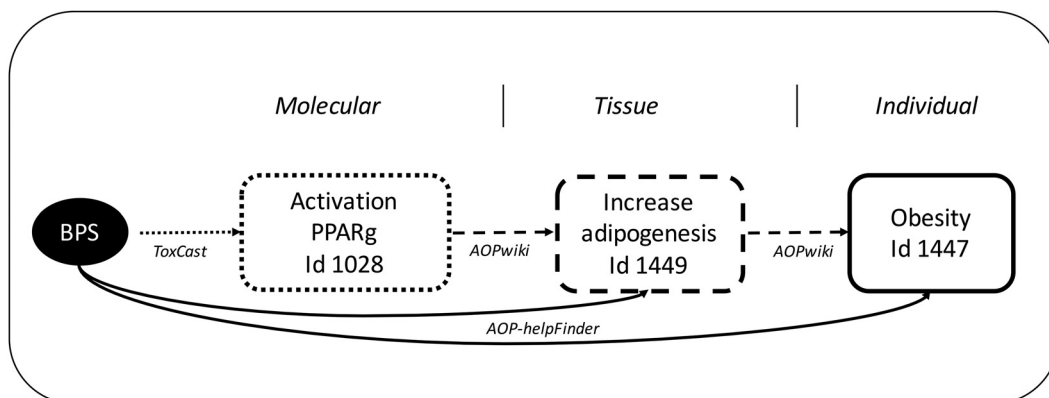


Figure 4. Potential mode of action for bisphenol S (BPS) leading to obesity resulting from text mining and integrated systems toxicology (data integration from the AOP-Wiki and ToxCast databases). Small dots (arrows and square lines) indicate information from the ToxCast database, large dots from the AOP-Wiki database, and solid lines, information identified using AOP-helpFinder.

the ToxCast database allowed us to identify positive associations between BPS and PPAR γ through two assays, i.e., ATG-PPAR γ _TRANS_up (mRNA activation in human HepG2 cell lines, 1.13 log 10 uM) and NVS_NR_hPPAR γ (active binding, 1.62 log 10 uM) (Figure 4).

We attempted to identify other biological pathways that link BPS to obesity. We sought to build patterns with coherent MoA using the full text of the publications that were identified by AOP-helpFinder through the screening of abstracts (Figure 5). Several molecular and cellular pathways, whose disruption could be linked to obesity, were identified: *a*) formation of adipocytes, *b*) increased lipogenesis, or *c*) decreased lipolysis (Table 2). Regarding adi-

pogenesis, the activation of estrogen receptor alpha (ER α) by BPS was shown to trigger the expression of several adipogenic markers in MCF-7 cells (Molina-Molina et al. 2013), HEK293T cells (Teng et al. 2013), and preadipocytes (Boucher et al. 2016a), such as Ap2 (adipocyte protein 2, a carrier protein for fatty acids). Since Ap2 was also shown to increase the expression of ER α in ER-negative MDA-MB-231 cells (McPherson and Weigel 1999), the existence of a positive feedback loop contributing to adipogenesis could be suggested. Interestingly, Ap2 was also shown to be a transcriptional target of PPAR γ in 3T3-L1 adipocytes (Rival et al. 2004), whose pathway was activated by BPS (Boucher et al. 2016a). Further, the expression of the PPAR γ coactivator 1 α , which is

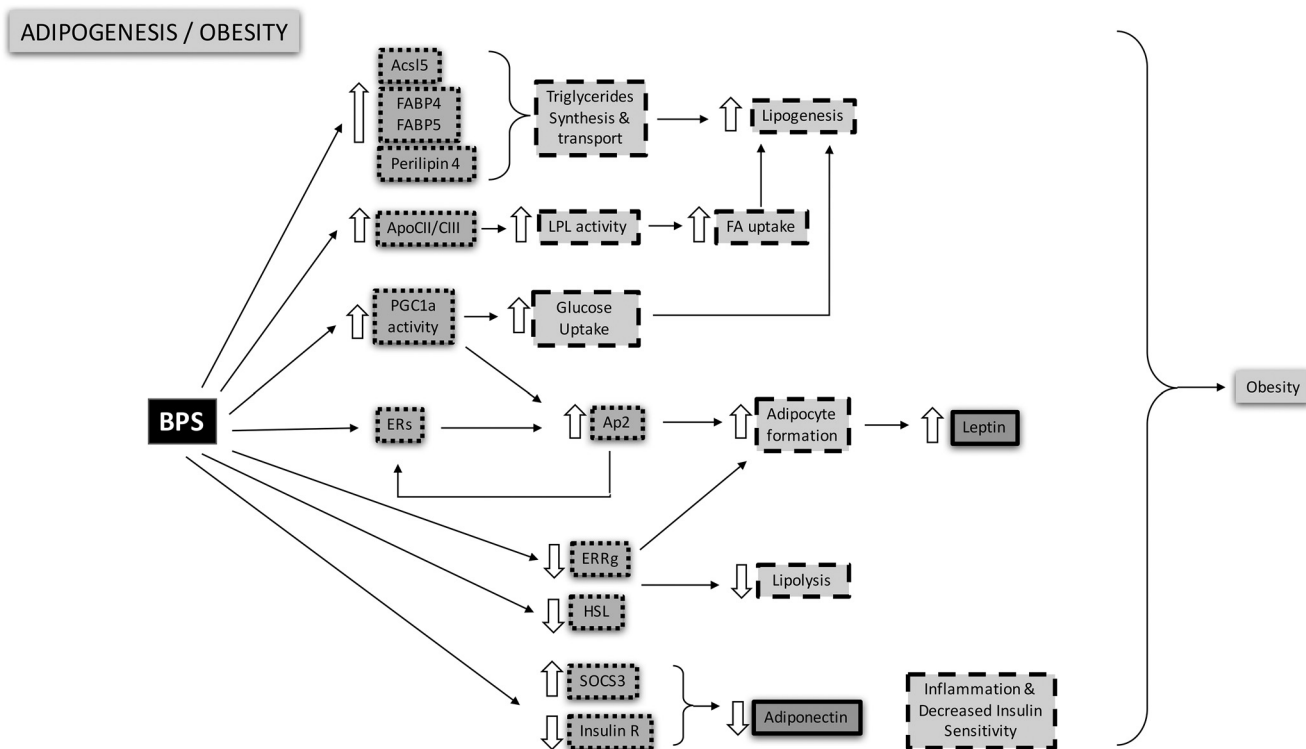


Figure 5. Identification of the mechanistic effects linking BPS to obesity. The AOP-helpFinder tool allowed us to select relevant publications that associated BPS, key events (KEs), and obesity. A manual curation of these studies by experts led to the identification of potential molecular targets of BPS (small dots surrounding squares), molecular processes (large dots surrounding squares), and hormone disruption (solid lines). Thick arrows: “increased” or “decreased.” Note: ApoC, apolipoprotein C-III; Ap2, adipocyte protein 2; Acsl5, Acyl-CoA synthetase long-chain family member 5; BPS, bisphenol S; ERRg, estrogen-related receptor gamma; ERs, estrogen receptors; FA, fatty acids; FABP, fatty acid-binding proteins; HSL, hormone-sensitive lipase; insulin R, insulin receptor; LPL, lipoprotein lipase; PGC1a: peroxisome proliferator-activated receptor gamma coactivator 1-alpha; SOCS3, suppressor of cytokine signaling 3.

Table 2. List of the seven publications, identified with AOP-helpFinder, fully explored to identify biological linkage between bisphenol S and obesity.

References	Detailed MIE-KE
Molina-Molina et al. 2013	Activation of human ER α and ER β by BPS
Boucher et al. 2016a	Activation of the expression of the adipogenic marker, adipocyte protein 2 (Ap2) (blocked by ER α antagonist)
Teng et al. 2013	Activation of ER α (EC ₅₀ of 2.2 μ M)
Héliès-Toussaint et al. 2014	BPS increased lipid content in the 3T3-L1 cell line and decreased the expression of ERR γ
Boucher et al. 2016a	BPS transactivates the expression of FABPs and perilipin 4 (microarray data, differentially expressed genes).
Crump et al. 2016	BPS transactivates the expression of ACSL5.
Ivry Del Moral et al. 2016	BPS decreased the expression of HSL.

Note: ACSL5, acyl-CoA synthetase long-chain family member 5; BPS, bisphenol S; EC₅₀, half-maximal effective concentration; ER, estrogen receptor; ERR γ ; estrogen-related receptor gamma; FABPs, fatty acids binding proteins; HSL, hormone-sensitive lipase; KE, key event; MIE, molecular initiating event.

recruited by PPAR γ to transactivate its target genes, was shown to be specifically up-regulated by BPS in the 3T3-L1 cell line (Héliès-Toussaint et al. 2014).

Additional observations collected from the retrieved articles could also contribute to the increased adipogenesis outcome, notably through metabolic disruption: BPS promotes the cellular uptake of glucose in two cell lines (3T3-L1 and HepG2), which could contribute to the production of glycerol-3-phosphate and then subsequently to the synthesis of triglycerides (Héliès-Toussaint et al. 2014). One study also has linked the activation of PPAR γ to an increased uptake of glucose (Zhang et al. 2017). Several key factors that regulate lipogenesis and lipolysis (including fatty acids binding proteins (FABP4 and FABP5) (Boucher et al. 2016b), perilipin 4 (Crump et al. 2016; Boucher et al. 2016b), or acyl-CoA synthetase long-chain family member 5 (ACSL5) (Crump et al. 2016) also were found to be associated with BPS exposure by AOP-helpFinder. Together, these proteins could contribute to an increased fatty acid uptake in preadipocytes (Boucher et al. 2016a) and to the production of triglycerides and, subsequently, of lipid droplets. Furthermore, BPS negatively regulates the expression of the hormone-sensitive lipase (HSL) in adipose tissue of C57BL/6 mice (Ivry Del Moral et al. 2016) and estrogen-related receptor gamma (ERR γ) in the 3T3-L1 cell line (Héliès-Toussaint et al. 2014). ERR γ regulates positively the expression of uncoupling protein 1 during the browning of 3T3-L1 adipocytes, which suggests that its down-regulation contributes to the increase in white adipose tissue, whereas down-regulation of HSL contributes to decreased lipolysis (Akter et al. 2008). In addition to metabolic disruption, AOP-helpFinder identified references that showed that BPS could be associated to the establishment of insulin resistance through both decreased expression of the insulin receptor and an increased inflammatory response in adipose tissue of C57BL/6 mice (Ivry Del Moral et al. 2016). Taken together, the studies retrieved through AOP-helpFinder suggested a link between BPS exposure and an AOP network related to adipogenesis and metabolic disruption.

In order to test additional applications of the AOP-helpFinder tool, we screened other data sources such as PubMed and included more disease terms than those in the AOP dictionary. Specifically, we used AOP-helpFinder to investigate BPS in the 109 selected publications, but instead of using AOP terms, the 11,850 MeSH disease terms were used (downloaded from the CTD database as of 28 September 2017). Obesity (MESH: D009765) was retrieved with a *pS* value of 0.93. Other disorders or phenotypes were also found, such as “diabetes mellitus” (*pS* of 0.84) or “body weight” (*pS* of 0.73) (see full list in Table S3).

Thus, our computational approach highlighted the need to integrate various data sources (publications, databases) in order to capture as much information as possible when studying links between a selected chemical and AOs.

Discussion

We have developed a novel computational approach using TM, graph theory, and systems biology to improve our capacity to link chemicals with AOPs. The AOP-helpFinder tool is particularly well suited to explore putative toxicity of chemicals to be used as replacements for toxic compounds, since studies exploring the human effects of the substitutes usually are scarce, whereas public decisions concerning the suitability of their use are urgently needed. The ability to identify chemical-biological associations is illustrated using BPS as an example. There is a large amount of published data for BPA. However, much less information is available for alternatives to BPA (BPS, for example), even though there has been a recent surge in the number of publications due to increased use of alternatives in consumer products (Rochester and Bolden 2015). We mined available information from multiple sources using the AOP-helpFinder tool. The ability to make novel observations or to confirm suspected effects using this tool, combined with integrative systems biology, is clearly illustrated by our observations on putative links between BPS and obesity. This method increases our knowledge, may give orientation for further experimental analysis, and supports the development of models that provide alternatives to animal testing [Replacement, Reduction and Refinement (3Rs)]. The advantages of our approach include the rapid exploration and the integration of existing information from multiple sources (databases and literature). It therefore has the potential to accelerate information gathering and is useful particularly when data are limited and present in diverse sources, which is the case for a large number of chemicals.

The strategy described here is complementary to existing systems toxicological models already developed for the identification and prediction of linkages between chemicals and human health. Some of these models are hybrid methods that combine multiple methods (Krysiak-Baltyn et al. 2014). Others are chemical structure based (Thomas et al. 2013; Ball et al. 2016), involving the physico-chemical and reactivity properties of the chemicals [read-across and quantitative structure-activity relationships (QSARs)]. Both are important for the interactions with specific biological targets and pathways, and therefore allow the prediction of toxic effects (Dang et al. 2017; Zang et al. 2017). Following the OECD recommendations for the construction of robust QSARs (OECD 2014), several models are now available such as the OPERA version 1.5 (Mansouri et al. 2018) or the VEGA platform (www.vegahub.eu). Recently, a new tool that integrates both chemical similarity and biosimilarity has been developed, the Chemical In vitro-In vivo Profiling (CIPro) (Russo et al. 2017). This tool profiles compounds of interest utilizing biological data from public resources. It uses the data for read-across assessment with the aim of predicting complex bioactivities. Several web-based tools, such as the REACHAcross™ tool (Hartung 2016) and the ChemProt database (using virtual screening) (Kim Kjærulff et al. 2013), are available.

Other methods are primarily based on the delineation of the MoA and on systems biology. Chemicals can disrupt a variety of pathways and lead to complex manifestation of toxicities. However, different chemicals can have similar MoA if they dysregulate the same target or targets that belong to the same signaling pathway (Bopp et al. 2018). Such pathway-based approach is recapitulated in AOPs and can be addressed by systems biology. Integrative systems biology models, which combine toxicogenomics, protein–chemical associations, protein–protein interactions (Audouze et al. 2010; Audouze and Grandjean 2011), disease phenotype information, genome-wide association, and genetic disease similarities (Audouze et al. 2013), have revealed molecular mechanisms of xenobiotics and have linked them to diseases such as obesity and endocrine disruption. Recently, data-driven approaches have been designed to generate and enrich AOP descriptions. For example, Nymark et al. (2017) proposed a multistep procedure based on an *in silico* pipeline to identify a network of functional elements for pulmonary fibrosis-associated genes. This generated novel AOP-linked molecular pathways (WP3624 in WikiPathways). Another recent study based on a systems toxicology profiling approach using information from the U.S. Environmental Protection Agency database (ToxRefDB) (U.S. EPA), the ToxCast database, and a comprehensive literature analysis, identified links between environmental chemicals, molecular targets, and AOs for male reproduction (Leung et al. 2016). Another type of data-driven approach is the use of TM and frequent item-set mining (FIM) to extract fragmented information in texts (e.g., abstracts or the full text of publications) (Jensen et al. 2012) or large datasets (clinical data, biobank, electronic patient records) (Jiang et al. 2019) and establish relevant associations. Currently, such methods are used widely in the biomedical area to identify information regarding biological entities (e.g., genes and proteins, metabolites, phenotypes, pathways) (Jensen et al. 2006; Krallinger et al. 2008). These approaches have been used less frequently in the toxicological field. However, in the case of U.S. biomonitoring surveys and Danish clinical studies, a study based on FIM has led to the successful identification and prioritization of connections between environmental chemicals, biomarkers, and human disorders (Krysiak-Baltyn et al. 2014). FIM has been also used to create computationally predicted AOPs, using the chemicals as the common aggregators between data (Oki and Edwards 2016), employing information from the *in vitro* ToxCast HTS assays and disease information from the CTD database (Davis et al. 2017). As a case study, they predicted an uncharacterized connection between the aryl hydrocarbon receptor and glaucoma resulting from changes in CYP1B1. This connection is in agreement with experimental data that shows an association between rare CYP1B1-activating mutations and congenital glaucoma (Alsaif et al. 2018; Stoilov et al. 1997).

As compared to the abovementioned structural or systems biology-based methods, the added value of the presented approach is that it combines the exploration of several databases (including ToxCast) with an improved method of TM that is based on the scoring of connections between chemicals and relevant AOP components. Therefore, it explores a wider range of data and has the potential to prioritize the chemical–health outcome connections, thus allowing further targeted studies. It also provides hints as to potential MoAs and, therefore, could be of use in risk assessment. We believe that such a proposed approach could be improved by screening more available data (the PubMed database, for example) and by including other disease terms in the disease dictionary that was initially limited to the AOP terms. As an example, when we used MeSH disease terms, the method was also efficient and confirmed the links between BPS and obesity.

In addition to providing a new tool to explore the putative toxicity of chemicals and chemical mixtures, our study highlights the need for methods improvements and for additional tools. Novel

computational systems toxicology models are needed to better characterize and predict the complex toxic effects of the chemicals to which humans are exposed. The current lack of high-quality data for some of the environmental chemicals and the current limitation of defined AOPs restricts approaches such as the one described here. An extension of AOP-helpFinder, which would implement a comprehensive dictionary of synonyms related to AOP terms, would increase the sensitivity of the method. Furthermore, the manual curation, used here, of the full texts of the selected publications could be automated by a TM approach. A recent comprehensive comparison of text mining of 15 million full-text articles vs. their corresponding abstracts from the period 1,823–2016 concluded that access to the full text improved findings (Westergaard et al. 2018).

We believe that the most relevant applications of the method described here are the delineation of chemical mixtures effects and a more rapid and efficient evaluation of the safety of substitutes to toxic chemicals. In both cases, it is critical to determine whether different chemicals exhibit similar MoA, and this can be accelerated by computational methods. Indeed, a critical issue in mixtures studies is to determine whether the compounds have similar modes of action (Kortenkamp and Faust 2018) because, in such a case, dose addition is the most appropriate method to assess the global effect of the mixture. Similarly, it is critical to determine whether substitutes to toxic chemicals exhibit an MoA that is similar to that of the parent compound. We have illustrated this aspect here by the study of BPS, which appears to be a putative obesogen, like BPA and other bisphenol compounds.

Conclusion

Exposure to chemical substances that can produce multiple health effects, such as endocrine disruption or metabolic disruption, represents one of the most critical public health threats at present. Novel, innovative computational methods, such as the AOP-helpFinder, are useful resources for exploring and predicting potential links among environmental chemicals and molecular targets, biological events, and AOs. We believe that the development and improvement of such *in silico* approaches has the potential to significantly impact relevant frameworks to assess chemical toxicity, such as the integrated approaches for testing and assessment, the reduction of animal testing, and the identification of safe chemical substitutes.

Acknowledgments

The authors would like to acknowledge HBM4EU (<https://www.hbm4eu.eu/>), a project funded by the European Union's Horizon 2020 research and innovation program under grant agreement no. 733032. This work was also supported by the University of Paris Descartes-USPC, INSERM, and Assistance Publique-Hôpitaux-de-Paris.


References

- Akter MH, Yamaguchi T, Hirose F, Osumi T. 2008. Perilipin, a critical regulator of fat storage and breakdown, is a target gene of estrogen receptor-related receptor alpha. *Biochem Biophys Res Commun* 368(3):563–568, PMID: 18243128, <https://doi.org/10.1016/j.bbrc.2008.01.102>.
- Alsaif HS, Khan AO, Patel N, Alkuraya H, Hashem M, Abdulwahab F, et al. 2018. Congenital glaucoma and CYP1B1: an old story revisited. *Hum Genet* 1–7.
- Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, et al. 2010. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. *Environ Toxicol Chem* 29(3):730–741, PMID: 20821501, <https://doi.org/10.1002/etc.34>.
- Audouze K, Brunak S, Grandjean P. 2013. A computational approach to chemical etiologies of diabetes. *Sci Rep* 3:2712, PMID: 24048418, <https://doi.org/10.1038/srep02712>.

- Audouze K, Grandjean P. 2011. Application of computational systems biology to explore environmental toxicity hazards. *Environ Health Perspect* 119(12):1754–1759, PMID: 21846611, <https://doi.org/10.1289/ehp.1103533>.
- Audouze K, Juncker AS, Roque F, Krysiak-Baltyn K, Weinhold N, Taboureau O, et al. 2010. Deciphering diseases and biological targets for environmental chemicals using toxicogenomics networks. *PLoS Comput Biol* 6(5):e1000788, PMID: 20502671, <https://doi.org/10.1371/journal.pcbi.1000788>.
- Ball N, Cronin MTD, Shen J, Blackburn K, Booth ED, Bouhifid M, et al. 2016. Toward Good Read-Across Practice (GRAP) guidance. *ALTEX* 33(2):149–166, PMID: 26863606, <https://doi.org/10.14573/altex.1601251>.
- Berggren E, Amcoff P, Benigni R, Blackburn K, Carney E, Cronin M, et al. 2015. Chemical safety assessment using read-across: assessing the use of novel testing methods to strengthen the evidence base for decision making. *Environ Health Perspect* 123(12):1232–1240, PMID: 25956009, <https://doi.org/10.1289/ehp.1409342>.
- Biovia. Registry of Toxic Effects of Chemical Substances (RTECS) database. <http://www.3dsbiovia.com/products/collaborative-science/databases/bioactivity-databases/rtecs.html> [accessed 25 August 2017].
- Bopp SK, Barouki R, Brack W, Dalla Costa S, Dorne JCM, Drakvik PE, et al. 2018. Current EU research activities on combined exposure to multiple chemicals. *Environ Int* 120:544–562, PMID: 30170309, <https://doi.org/10.1016/j.envint.2018.07.037>.
- Boucher JG, Ahmed S, Atlas E. 2016a. Bisphenol S induces adipogenesis in primary human preadipocytes from female donors. *Endocrinology* 157(4):1397–1407, PMID: 27003841, <https://doi.org/10.1210/en.2015-1872>.
- Boucher JG, Gagné R, Rowan-Carroll A, Boudreau A, Yauk CL, Atlas E. 2016b. Bisphenol A and bisphenol S induce distinct transcriptional profiles in differentiating human primary preadipocytes. *PLoS One* 11(9):e0163318, PMID: 27685785, <https://doi.org/10.1371/journal.pone.0163318>.
- Crump D, Chiu S, Williams KL. 2016. Bisphenol S alters embryonic viability, development, gallbladder size, and messenger RNA expression in chicken embryos exposed via egg injection. *Environ Toxicol Chem* 35(6):1541–1549, PMID: 26606162, <https://doi.org/10.1002/etc.3313>.
- Dang NL, Hughes TB, Miller GP, Swamidass SJ. 2017. Computational approach to structural alerts: furans, phenols, nitroaromatics, and thiophenes. *Chem Res Toxicol* 30(4):1046–1059, PMID: 28256829, <https://doi.org/10.1021/acs.chemrestox.6b00336>.
- Davis AP, Grondin CJ, Johnson RJ, Sciaky D, King BL, McMorran R, et al. 2017. The Comparative Toxicogenomics Database: update 2017. *Nucleic Acids Res* 45(D1):D972–D978, PMID: 27651457, <https://doi.org/10.1093/nar/gkw838>.
- Davis AP, Wiegiers TC, Wiegiers J, Johnson RJ, Sciaky D, Grondin CJ, et al. 2018. Chemical-induced phenotypes at CTD help inform the pre-disease state and construct adverse outcome pathways. *Toxicol Sci* 165(1):145–156.
- Dijkstra EW. 1959. A note on two problems in connexion with graphs. *Numer Math* 1(1):269–271, <https://doi.org/10.1007/BF01386390>.
- Hartung T. 2016. Making big sense from big data in toxicology by read-across. *ALTEX* 33(2):83–93, PMID: 27032088, <https://doi.org/10.14573/altex.1603091>.
- Heindel JJ, Blumberg B. 2018. Environmental obesogens: mechanisms and controversies. *Annu Rev Pharmacol Toxicol* 59:89–106.
- Héliès-Toussaint C, Peyre L, Costanzo C, Chagnon MC, Rahmani R. 2014. Is bisphenol S a safe substitute for bisphenol A in terms of metabolic function? An in vitro study. *Toxicol Appl Pharmacol* 280(2):224–235, PMID: 25111128, <https://doi.org/10.1016/j.taap.2014.07.025>.
- Ives C, Campia I, Wang RL, Wittwehr C, Edwards S. 2017. Creating a structured adverse outcome pathway knowledgebase via ontology-based annotations. *Appl In Vitro Toxicol* 3(4):298–311, PMID: 30057931, <https://doi.org/10.1089/avt.2017.0017>.
- Ivry Del Moral L, Le Corre L, Poirier H, Niot I, Truntzer T, Merlin JF, et al. 2016. Obesogen effects after perinatal exposure of 4,4'-sulfonyldiphenol (Bisphenol S) in C57BL/6 mice. *Toxicology* 357–358:11–20, PMID: 27241191, <https://doi.org/10.1016/j.tox.2016.05.023>.
- Jensen PB, Jensen LJ, Brunak S. 2012. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 13(6):395–405, PMID: 22549152, <https://doi.org/10.1038/nrg3208>.
- Jensen LJ, Saric J, Bork P. 2006. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet* 7(2):119–129, PMID: 16418747, <https://doi.org/10.1038/nrg1768>.
- Jiang L, Audouze K, Romero Herrera JA, Angquist LH, Kim-Kjaerulf S, Izarzugaza JMG, et al. 2019. Conflicting associations between dietary patterns and changes of anthropometric traits across subgroups of middle-aged women and men. *Clin Nutr*, <https://doi.org/10.1016/j.clnu.2019.02.003>.
- Judson RS, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Mortensen HM, et al. 2010. In vitro screening of environmental chemicals for targeted testing prioritization: the ToxCast project. *Environ Health Perspect* 118(4):485–492, PMID: 20368123, <https://doi.org/10.1289/ehp.0901392>.
- Karrer C, Roiss T, von Goetz N, Gramec Skledar D, Peterlin Mašič L, Hungerbühler K. 2018. Physiologically based pharmacokinetic (PBPK) modeling of the bisphenols BPA, BPS, BPF, and BPAF with new experimental metabolic parameters: comparing the pharmacokinetic behavior of BPA with its substitutes. *Environ Health Perspect* 126(7):077002, PMID: 29995627, <https://doi.org/10.1289/EHP2739>.
- Kim Kjaerulf S, Wich L, Kringelum J, Jacobsen UP, Kouskoumvekaki I, Audouze K, et al. 2013. ChemProt-2.0: visual navigation in a disease chemical biology database. *Nucleic Acids Res* 41(Database issue):D464–D469, PMID: 23185041, <https://doi.org/10.1093/nar/gks1166>.
- Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. 2019. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* 47(D1):D1102–D1109, PMID: 30371825, <https://doi.org/10.1093/nar/gky1033>.
- Kortenkamp A, Faust M. 2018. Regulate to reduce chemical mixture risk. *Science* 361(6399):224–226, PMID: 30026211, <https://doi.org/10.1126/science.aat9219>.
- Krallinger M, Valencia A, Hirschman L. 2008. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol* 9(Suppl 2):S8, PMID: 18834499, <https://doi.org/10.1186/gb-2008-9-s2-s8>.
- Krysiak-Baltyn K, Nordahl Petersen T, Audouze K, Jørgensen N, Angquist L, Brunak S. 2014. Compass: a hybrid method for clinical and biobank data mining. *J Biomed Inform* 47:160–170, PMID: 24513869, <https://doi.org/10.1016/j.jbi.2013.10.007>.
- Letterova MI, Haakonsson AK, Lazar MA, Mandrup S. 2014. PPAR γ and the global map of adipogenesis and beyond. *Trends Endocrinol Metab* 25(6):293–302, Jun, PMID: 24793638, <https://doi.org/10.1016/j.tem.2014.04.001>.
- Leung MCK, Phuon J, Baker NC, Sipes NS, Klinefelter GR, Martin MT, et al. 2016. Systems toxicology of male reproductive development: profiling 774 chemicals for molecular targets and adverse outcomes. *Environ Health Perspect* 124(7):1050–1061, PMID: 26662846, <https://doi.org/10.1289/ehp.1510385>.
- Mansouri K, Grulke CM, Judson RS, Williams AJ. 2018. OPERA models for predicting physicochemical properties and environmental fate endpoints. *J Cheminformatics* 10(1):10.
- McPherson LA, Weigel RJ. 1999. AP2alpha and AP2gamma: a comparison of binding site specificity and trans-activation of the estrogen receptor promoter and single site promoter constructs. *Nucleic Acids Res* 27(20):4040–4049, PMID: 10497269.
- Molina-Molina JM, Amaya E, Grimaldi M, Sáenz JM, Real M, Fernández MF, et al. 2013. In vitro study on the agonistic and antagonistic activities of bisphenol-S and other bisphenol-A congeners and derivatives via nuclear receptors. *Toxicol Appl Pharmacol* 272(1):127–136, PMID: 23714657, <https://doi.org/10.1016/j.taap.2013.05.015>.
- NIH (National Institutes of Health). PubChem Database. <https://pubchem.ncbi.nlm.nih.gov/> [accessed 28 July 2017].
- NIH, TOXNET (Toxicology Data Network)a. Chemical Carcinogenesis Research Information System (CCRIS). <https://toxnet.nlm.nih.gov/newtoxnet/ccris.htm> [accessed 28 July 2017].
- NIH, TOXNETb. Developmental and Reproductive Toxicology Database (DART). <https://toxnet.nlm.nih.gov/newtoxnet/dart.htm> [accessed 28 July 2017].
- NIH, TOXNETc. Hazardous Substances Data Bank (HSDB). <https://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB> [accessed 28 July 2017].
- NIH, TOXNETd. Toxicology Literature Online (TOXLINE). <https://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?TOXLINE> [accessed 28 July 2017].
- Nymark P, Rieswijk L, Ehrhart F, Jeliazkova N, Tsiliki G, Sarimveis H, et al. 2017. A data fusion pipeline for generating and enriching Adverse Outcome Pathway descriptions. *Toxicol Sci* 162(1):264–275, PMID: 29149350, <https://doi.org/10.1093/toxsci/kfx252>.
- OECD (Organisation for Economic Co-operation and Development). 2014. *The Guidance Document for Using the OECD (Q)SAR Application Toolbox to Develop Chemical Categories According to the OECD Guidance on Grouping Chemicals*. Paris, France:OECD Publishing.
- OECD. AOP Knowledge Base. <https://aopkb.oecd.org>.
- Oki NO, Edwards SW. 2016. An integrative data mining approach to identifying adverse outcome pathway signatures. *Toxicology* 350–352:49–61, PMID: 27108252, <https://doi.org/10.1016/j.tox.2016.04.004>.
- Polvani S, Tarocchi M, Tempesti S, Bencini L, Galli A. 2016. Peroxisome proliferator activated receptors at the crossroad of obesity, diabetes, and pancreatic cancer. *World J Gastroenterol* 22(8):2441–2459, PMID: 26937133, <https://doi.org/10.3748/wjg.v22.i8.2441>.
- Richard AM, Judson RS, Houck KA, Grulke CM, Volarath P, Thillainadarajah I, et al. 2016. ToxCast chemical landscape: paving the road to 21st century toxicology. *Chem Res Toxicol* 29(8):1225–1251, PMID: 27367298, <https://doi.org/10.1021/acs.chemrestox.6b00135>.
- Rival Y, Stenvenin A, Puech L, Rouquette A, Cathala C, Lestienne F, et al. 2004. Human adipocyte fatty acid-binding protein (aP2) gene promoter-driven reporter assay discriminates nonlipogenic peroxisome proliferator-activated receptor γ

- ligands. *J Pharmacol Exp Ther* 311(2):467–475, PMID: 15273253, <https://doi.org/10.1124/jpet.104.068254>.
- Rochester JR, Bolden AL. 2015. Bisphenol S and F: a systematic review and comparison of the hormonal activity of bisphenol A substitutes. *Environ Health Perspect* 123(7):643–650, PMID: 25775505, <https://doi.org/10.1289/ehp.1408989>.
- Russo DP, Kim MT, Wang W, Pinolini D, Shende S, Strickland J, et al. 2017. CIIPro: a new read-across portal to fill data gaps using public large-scale chemical and biological data. *Bioinformatics* 33(3):464–466, PMID: 28172359, <https://doi.org/10.1093/bioinformatics/btw640>.
- SAAOP (Society for the Advancement of Adverse Outcome Pathways). AOP-Wiki. Updated 4 December 2016. <https://aopwiki.org/>.
- Stoilov I, Akarsu AN, Sarfarazi M. 1997. Identification of three different truncating mutations in cytochrome P4501B1 (CYP1B1) as the principal cause of primary congenital glaucoma (Buphthalmos) in families linked to the GLC3A locus on chromosome 2p21. *Hum Mol Genet* 6(4):641–647, PMID: 9097971, <https://doi.org/10.1093/hmg/6.4.641>.
- Teng C, Goodwin B, Shockley K, Xia M, Huang R, Norris J, et al. 2013. Bisphenol A affects androgen receptor function via multiple mechanisms. *Chem Biol Interact* 203(3):556–564, PMID: 23562765, <https://doi.org/10.1016/j.cbi.2013.03.013>.
- Thomas RS, Philbert MA, Auerbach SS, Wetmore BA, Devito MJ, Cote I, et al. 2013. Incorporating new technologies into toxicity testing and risk assessment: moving from 21st century vision to a data-driven framework. *Toxicol Sci* 136(1):4–18, PMID: 23958734, <https://doi.org/10.1093/toxsci/kft178>.
- U.S. EPA (Environmental Protection Agency). ToxCast Dashboard. <https://www.epa.gov/chemical-research/toxcast-dashboard> [accessed 26 April 2018].
- U.S. EPA. ToxRefDB - Release user-friendly web-based tool for mining. Updated 8 May 2018. https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NCCT&dirEntryId=227139 [accessed 28 July 2017].
- Villeneuve DL, Crump D, Garcia-Reyero N, Hecker M, Hutchinson TH, LaLone CA, et al. 2014. Adverse outcome pathway (AOP) development I: strategies and principles. *Toxicol Sci* 142(2):312–320, PMID: 25466378, <https://doi.org/10.1093/toxsci/kfu199>.
- Westergaard D, Stærfeldt H-H, Tønsberg C, Jensen LJ, Brunak S. 2018. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. *PLoS Comput Biol* 14(2): e1005962, PMID: 29447159, <https://doi.org/10.1371/journal.pcbi.1005962>.
- Zang Q, Mansouri K, Williams AJ, Judson RS, Allen DG, Casey WM, et al. 2017. In silico prediction of physicochemical properties of environmental chemicals using molecular fingerprints and machine learning. *J Chem Inf Model* 57(1):36–49, PMID: 28006899, <https://doi.org/10.1021/acs.jcim.6b00625>.
- Zhang W, Xu Y, Xu Q, Shi H, Shi J, Hou Y. 2017. PPAR δ promotes tumor progression via activation of Glut1 and SLC1-A5 transcription. *Carcinogenesis* 38(7):748–755, PMID: 28419191, <https://doi.org/10.1093/carcin/bgx035>.

Deciphering Adverse Outcome Pathway Network Linked to Bisphenol F Using Text Mining and Systems Toxicology Approaches

Marylène Rugard, Xavier Coumoul, Jean-Charles Carvaillo, Robert Barouki, and Karine Audouze ¹

Université de Paris, Inserm UMR S-1124, 75006 Paris, France

¹To whom correspondence should be addressed at Université de Paris, Inserm UMR-S 1124, 45 rue des Saints-Pères, Paris 75006, France. Fax : 01 42 86 38 68; E-mail: karine.audouze@univ-paris-diderot.fr.

ABSTRACT

Bisphenol F (BPF) is one of several Bisphenol A (BPA) substituents that is increasingly used in manufacturing industry leading to detectable human exposure. Whereas a large number of studies have been devoted to decipher BPA effects, much less is known about its substituents. To support decision making on BPF's safety, we have developed a new computational approach to rapidly explore the available data on its toxicological effects, combining text mining and integrative systems biology, and aiming at connecting BPF to adverse outcome pathways (AOPs). We first extracted from different databases BPF-protein associations that were expanded to protein complexes using protein-protein interaction datasets. Over-representation analysis of the protein complexes allowed to identify the most relevant biological pathways putatively targeted by BPF. Then, automatic screening of scientific abstracts from literature using the text mining tool, AOP-helpFinder, combined with data integration from various sources (AOP-wiki, CompTox, etc.) and manual curation allowed us to link BPF to AOP events. Finally, we combined all the information gathered through those analyses and built a comprehensive complex framework linking BPF to an AOP network including, as adverse outcomes, various types of cancers such as breast and thyroid malignancies. These results which integrate different types of data can support regulatory assessment of the BPA substituent, BPF, and trigger new epidemiological and experimental studies.

Key words: artificial intelligence; AOP-helpFinder; integrative systems toxicology; adverse outcome pathway network; HBM4EU.

Bisphenol A (BPA) is a presumed endocrine disruptor due to its chemical similarity to natural estrogens and also a metabolic disruptor and neurotoxicant. Several studies have been carried out to understand its modes of action (MoA). Some of these studies revealed effects at lower doses than the no-observed-adverse-effect levels, mostly corresponding to the regulation of non-genomic pathways whereas the activation of nuclear receptors were involved at higher doses (FitzGerald and Wilks, 2014). Bisphenol A has been banned in some

countries (Canada, EU) for some specific uses (baby bottles, coating of infant formula). This led to replacement of BPA for the production of epoxy resins and polycarbonate to reduce putative adverse effects. As a consequence, there is an increasing use of substituents such as bisphenol S (BPS) in thermal paper, bisphenol B, and bisphenol F (BPF) and its isomers in canned foods and soft drinks. However, the MoA and potential toxicities of BPA analogs are still poorly characterized. Whether these substituents are safer than BPA remains a

matter of debate. Linking BPF and other BPS to adverse outcome pathways (AOPs) and therefore enhancing our knowledge on their putative toxicities is a major challenge for regulatory needs.

Evidence for environmental and health effects of chemical substances has increased considerably over the last years (certain hormone-dependent cancers, neurocognitive disorders, reproductive perturbations). Therefore, regulatory measures need to be taken in the EU and in member states for those chemicals. Thus, there is a need for assessment of hazards and risks of the thousands of existing substances we are exposed to; however, to avoid animal testing which has also its limitations, both *in vitro* and *in silico* methods were developed and recommended by Toxicity testing in the 21st Century in 2007 and Economic Co-operation and Development Organization (OECD) guidelines. Together with the U.S. Environmental Protection Agency (U.S. EPA), the National Toxicology Program has recommended to include, among others, *in silico* approaches in future assessments of toxicity as an inexpensive and efficient tool for screening purposes. Recently, the concept of new approach methodologies has been put forward; it refers to nonanimal technologies that can be used to provide information on chemical hazard and risk assessment (ICCVAM, 2018).

To carry out these assessments, the OECD has proposed 2 frameworks (Sakuratani et al., 2018). The first 1 is a science-based approach, the Integrated Approaches to Testing and Assessment (IATA) that combines various types of existing and new data (*in vitro*, *in vivo*, and computational) to study a specific question. The second 1 is the AOP, that can be used in the development of IATA (Tollefsen et al., 2014). Adverse outcome pathway consists in capturing and organizing key events (KEs), at different levels of biological organization (molecular, cellular, tissue, organ, organism, and population), that lead to toxic effects. Adverse outcome pathway starts with a molecular initiating event (MIE), which can be triggered by a stressor (eg: chemical). These MIEs are connected to a sequence of KEs linked together by KE relationships (KERs), which lead to an adverse outcome (AO) (Ankley et al., 2010; Villeneuve et al., 2014). Such AOP frameworks provide a scheme describing mechanistic knowledge from existing tests (Knapen et al., 2015). They are particularly relevant for putative endocrine disruptors (defined by their MoA).

To successfully incorporate AOPs into risk assessment multiple interacting pathways or networks of AOPs should be accounted for (Garcia-Reyero, 2015). Individual AOP could be considered as a linear description of biological events, and can be merged via their shared events (MIE or KE). Consequently, AOP networks can be constructed by connecting 2 or more individual AOPs, if they have at least 1 common KE or KER, to provide a better description of the biological complexity (Knapen et al., 2018) with considerable additional value for emerging toxicological knowledge (Pollesch et al., 2019). Adverse outcome pathway networks can be initiated by 1 or more stressors that can be environmental chemicals.

Computational approaches allow to explore and identify key information, and therefore can be used to develop AOP networks. In their study, Oki and Edwards, generated computationally predicted AOPs by integrating multiple data sources from HTS studies by using Frequent Itemset Mining (Oki and Edwards, 2016).

In this study, we describe a computational approach to establish linkages between environmental stressors and health effects, using available information from the literature and

databases. Different sources of information were considered such as PubMed, ToxCast, CompTox, and AOP-wiki, and integrated to develop individual AOP and AOP networks. We took advantage of existing tools, WebGestaltR and the recently hybrid method called AOP-helpFinder (Carvaillo et al., 2019; Liao et al., 2019), an artificial intelligence (AI) method that automatically screens and analyses abstracts from published articles to decipher relevant links between chemical substances (ie, stressors) and AOPs. The presented strategy demonstrates the ability to identify links between BPF, KEs, and toxic effects using existing sparse information. The integrative approach revealed a plausible complex AOP induced by BPF, leading to thyroid cancer, as well as an AOP network related to various types of malignancies.

MATERIALS AND METHODS

Overall strategy. Linkage between BPF (see §2), biological events and toxic effects were investigated using a systems toxicology approach (Figure 1). This multistep approach is based on integration of existing knowledge from various sources of information. In the first step, specific data on BPF-protein associations were extracted from chemical biology databases. Second, these BPF-protein associations were expanded to protein complexes. By using a high confidence interactome (Li et al., 2017), we were able to decipher protein complexes associated with BPF (§3). Then, protein-pathway information was integrated into these protein complexes to statistically order linkage between the BPF and biological pathways (§4). Finally, to have a better understanding of the mode of action of BPF leading to the identified pathways, literature searches were performed, and relevant information were integrated (§5 and 6). Therefore, these steps allowed to suggest a plausible AOP induced by BPF leading to thyroid cancer, that has been extended to an AOP network (§7).

Bisphenol F. To get the most complete biological picture of BPF, the 3 isomers for BPF were analyzed (Figure 2): the 2,2'-isomer of BPF (CAS rn 2467-02-9) (2,2-BPF), the 2,4'-isomer of BPF (CASRN 2467-03-0) (2,4-BPF), and 4,4'-isomer of BPF (CAS rn 620-92-8) (4,4-BPF). The commercially available BPF mixed isomers was also taken into consideration (CAS rn 1333-16-0). Information from the different sources were compiled using the common name "BPF," and various synonyms (Supplementary Table 1). These synonyms were extracted from the CompTox database, and only the valid ones were retained (19 for 4,4-BPF, 12 for 2,2-BPF, 11 for 2,4-BPF, and 13 for BPF mixture).

Linking BPF to protein complexes. To identify proteins known to be associated with BPF, we compiled information from 2 publicly available databases. First, we extracted information from the ToxCast database (<https://actor.epa.gov/dashboard>; dashboard accessed on December 2018) (Judson et al., 2010). The ToxCast database is based on high-throughput technologies, and aggregate information for thousands of chemical substances. This U.S. Environmental Protection Agency infrastructure contains information for 9076 chemicals that have been tested on 359 assays with a total of 1192 endpoints (Judson et al., 2010; Kavlock et al., 2012).

Then to collect more proteins associated with BPF, the Comparative Toxicogenomics database (CTD) (as of December 2018) (Davis et al., 2018) was used. CTD contained 1 898 228 chemical-protein curated associations mined from peer-

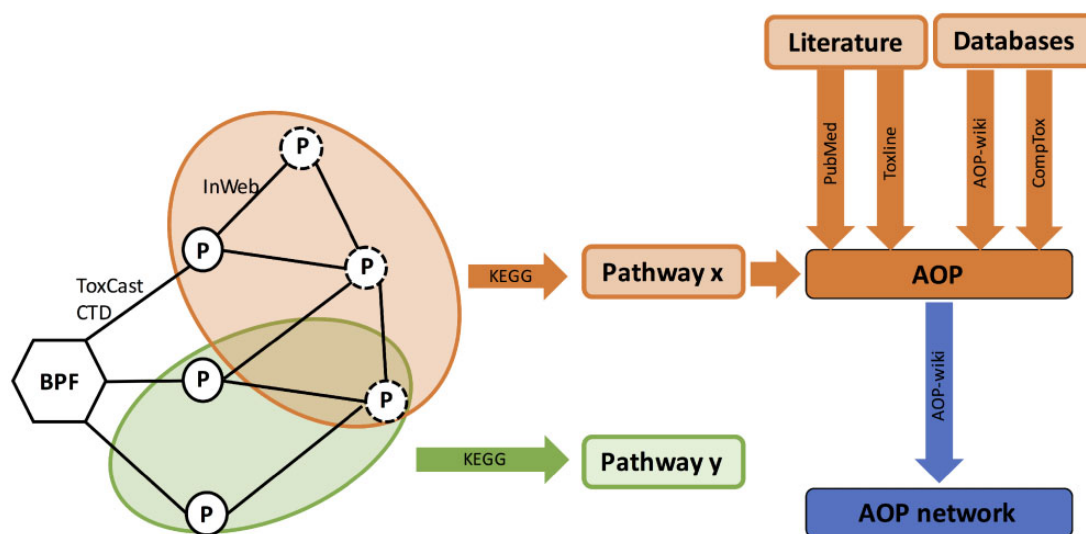


Figure 1. Overview of the systems toxicology strategy for developing adverse outcome pathway (AOP) networks for bisphenol F (BPF). In a first step, BPF-protein associations were extracted from the Comparative Toxicogenomics database and the ToxCast database. Then creation of protein complexes by integration of the first-order protein partners (P, dashed line) to the extracted proteins (P, full line) using a high confidence interactome based on experimental evidences (InWeb data source). The next step consisted of performing a biological enrichment of the protein complexes to statistically rank pathways linked to them. Finally, different data types were integrated from various sources (literature, databases) using AOP-helpFinder and by manual curation, to build comprehensive mechanisms between BPF and toxic effects, to develop an individual AOP and AOP network for which BPF may be a stressor.

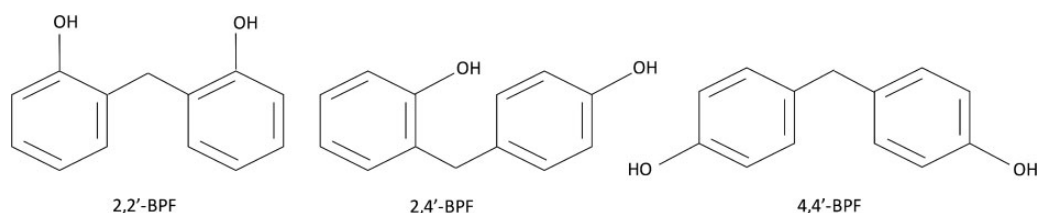


Figure 2. The 3 isoforms of bisphenol F.

reviewed scientific literature, between 13 108 chemicals and 47 581 proteins for 593 organisms. All proteins were mapped to HUGO name identifiers and Entrez gene id to facilitate further data integration.

Because in a biological system, proteins tend to function in groups or complexes, an important step was to enrich the list of compiled proteins using information from a high confidence human protein interactome that is protein-protein interactions (PPIs) based on experiments and inferred model organism data (Li et al., 2017). The InWeb 3.0 tool (www.cbs.dtu.dk/services/VirtualPullDown-1.1b/web/) was used to identify the first-order interacting proteins. Such strategy is based on a neighbor's pull down approach (Lage et al., 2007). The version that we used, contained a total of 507 142 unique PPIs involving 14 441 human proteins, as of December 2018. As a result, the list of relevant proteins associated with BPF was extended by inclusion of their first-order PPI partners, considering a significant pull down score threshold of 0.25 (Lage et al., 2007).

Biological enrichment of the protein complexes. To identify pathways and toxic effects related to BPF, information on biological pathways were integrated into the protein complex linked to BPF. To catch as much as possible information, the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways-based database was used as a source of information (Kanehisa et al., 2019).

To perform the over-representation analysis (ORA), the R package WebGestaltR (version 3), was used. The protein

complex was therefore tested for significant pathways associations using a test based on a hypergeometric distribution. A significance level of 0.05 after Benjamini-Hochberg correction for multiple testing of p values was used to select the most relevant associations.

Exploration of the literature to contribute to the development of AOP. The literature search was done in 2 ways with the aim to capture as much as possible information to link BPF to AOPs. As AOPs are in development, and limited number has been validated, we decided to combine our approach with a manual exploration of the PubMed database.

First, an automatic search was performed using the recently developed AOP-helpFinder tool (Carvalho et al., 2019), that is a hybrid approach that combined text mining procedure and graph theory, to identify linkage between BPF and biological events present in already defined AOPs. The TOXLINE database was screened to compile scientific publications mentioning BPF in a toxicological context (<https://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?TOXLINE>) (as of July 2017). To capture as much as possible of the existing information, several synonyms for BPF where used (Supplementary Table 1) that were extracted from the CompTox database (<https://comptox.epa.gov/dashboard>) (Williams et al., 2017). Then, to prepare the biological data, AOPs were considered. Based on the AOP-wiki database (as of August 2017) (<https://aopwiki.org>), a dictionary was developed. The AOP-wiki database contains information related to AOPs for which mechanistic representations of toxicological effects over various

level of the biological organization were reported. The developed dictionary contains information related to MIE/KE and AO. Both MIE and KE were integrated in the same “list” that contains 1318 events. The AO list contains 61 events. These 2 categories were analyzed separately using the text mining tool, mainly due to their wording composition. Biological events associated with MIE and KE were more complex sentences compared with AO that are more likely to be defined by 1 or few words. Finally, the literature screening approach was performed using the AOP-helpFinder tool. The text mining part is used to identify co-mentioned words (eg, a BPF isomer and a biological event) in an abstract from the scientific literature. The graph theory allowed to order the findings by calculating scores. Based on this tool, 2 different scores were calculated: a position score that determine the position of the co-occurred terms in an abstract. The more the AOP-related words are placed toward the end of the abstract, the more they can be considered as a result and not a working hypothesis. The second score, the weighted score takes into consideration the complexity of the AOP-related terms (from 1 word to 21 words) (Carvalho et al., 2019).

Then, the PubMed database (as of February 2019) was screened manually to decipher targeted information between identified proteins (through the biological enrichment of the protein complexes, as described above) involved in a disease of interest. More precisely, the research on PubMed was made using the keyword “thyroid cancer” associated with the name of each protein. The search results were analyzed by searching for the keywords in the title and abstracts of the articles. If links were mentioned in the summary, the article was read to find the relevant information and retrieve it. A second research was then conducted to find a link between proteins with a link with thyroid cancer and BPF. The results of this research were analyzed in the same way as the previous one. The results obtained from these 2 bibliographic researches also made it possible to reveal links between BPF or thyroid cancer and proteins that are not part of the initial protein complex. As a result, these links were further screened with a final bibliographical research.

Integrating biological data. To get a more comprehensive and complete mechanistic view of BPF associated with the identified AOPs, we included as much as possible information (molecular targets, biological events, toxic effects) at different levels of the biological organization by using several databases. The U.S. EPA’s ToxCast program has screened thousands of chemicals for biological activity, primarily using high-throughput *in vitro* bioassays. The ToxCast dashboard (<https://actor.epa.gov/dashboard/>) (as of March 2019) was used. Relevant data were also extracted from other sources of information such as the CompTox database (<https://comptox.epa.gov/dashboard/>) (as of March 2019) and the PubChem bioassays database (<https://pubchem.ncbi.nlm.nih.gov/>) (as of March 2019). This step allows us to build an individual AOP induced by BPF.

Generation of AOP network. To characterize an AOP network induced by BPF, we assembled shared events (MIE/KE/AO) between the putative AOP developed by the systems biology approach, and available information using the AOP-wiki database (as of April 2019). As a result, the structure of the AOP network is described by at least 2 or more AOPs that share at least 1 event. The development of the AOP network was done by screening the AOP-wiki database to extract events shared with the proposed individual AOP for BPF.

RESULTS

Generation of Protein Complexes for BPF Compounds

Using the ToxCast and CTDs, we extracted proteins for which the chemicals show biological activities. Therefore, we were able to compile information regarding 9 proteins for 2,4’-BPF, 12 proteins for 2,2’-BPF, and 111 proteins for 4,4’-BPF (Figure 3). As a result, 114 unique proteins were identified, and among them 6 proteins were common to the 3 compounds (AR, ESR1, POU2F1, VRD, NR1I2, and NFE2L2). No information was obtained for the BPF mixed isomer. Therefore, the data analyses were concentrated on these 114 unique proteins that are connected to at least 1 of 3 BPF isomers.

Using the list of identified proteins, BPF isomers were linked to protein complexes by determining first-order PPI partners for each of the 114 proteins. In the generated network, 307 proteins (including the 114 input proteins) were connected to their respective partners, with a total of 496 edges (see Supplementary Table 2).

Pathways-based Analysis of the Protein Complexes

To identify pathways associated with the 3 studied BPF, the protein complexes were enriched by ORA using the KEGG database. Among the 307 proteins present in the network, 288 were mapped to unique entrezGeneID, and therefore used for the ORA (Supplementary Table 3). Among these 288 proteins, 158 have annotations in the KEGG database. Consequently, the ORA of these 158 proteins revealed several statistically significant pathways (see Supplementary Table 4) and were mapped to a number of categories such as the function of the synapses or cell signaling. Some of these appear to be more specific and potentially connected in terms of mechanisms of action.

Several lipid metabolism-related pathways were retrieved with the KEGG analysis. For example, pathways such as “linoleic acid metabolism” (with a corrected p value of 8.51×10^{-13} , via 13 proteins) and “arachidonic acid metabolism” (with a corrected p value of 4.35×10^{-10} , through 13 proteins), were significantly connected to BPF. Interestingly, those 2 omega-6 fatty acids have been associated with inflammation, a process which is also linked to the development of several pathologies (including cancer).

Another finding was linked to endocrine-related pathways including signaling by cortisol (together with Cushing syndrome), estrogen, aldosterone, progesterone (together with 2 global steroidogenesis pathways: steroid hormone biosynthesis, ovarian steroidogenesis) and also parathyroid hormone and thyroid hormone pathways suggesting that BPF may act as an endocrine disruptor. Interestingly, changes in the expression of several adenylyl cyclases (involved in the production of cAMP and subsequently in signaling by the transcriptional factor, cAMP Responsive Element Binding protein) and of metabolizing enzymes (including cytochromes P450 and transferases) were observed suggesting that BPF could act on several processes related to these hormones (signaling and metabolism).

Moreover, among the identified pathways, 4 were associated with endocrine-related cancers (prostate, endometrial, breast, thyroid) (Table 1). According to the CompTox database (access as of March 2019), no cancer information was associated yet to the BPF compounds. Therefore, we decided to explore the most significant link (BPF-thyroid cancer) using an integrative approach including text mining, literature searches, and additional databases exploration.

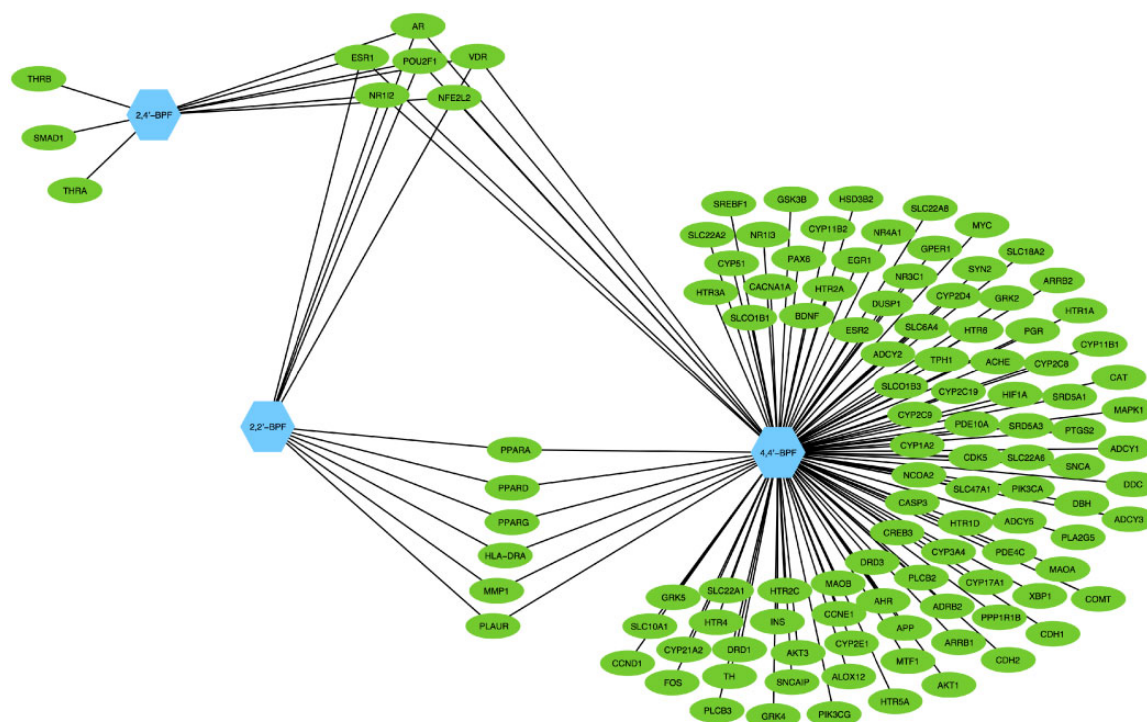


Figure 3. Bisphenol F -protein associations network. View of the proteins (green ovals) associated with the 3 bisphenol F isomers (blue hexagons). Data were extracted from the ToxCast and Comparative Toxicogenomics databases. Proteins are denoted by HUGO gene symbols to facilitate further analysis.

Table 1. List of the Statistically Significant Endocrine-Related Cancers Associated With the 3 Studied BPF Isomers

Pathways Name	<i>p</i> Value	FDR ^a	Gene Name
Thyroid cancer	1.00E-05	8.38E-05	CCND1; CDH1; MAPK1; MYC; NCOA4; PPARG; RXRG
Endometrial cancer	2.62E-05	1.90E-04	AKT1; AKT3; CCND1; CDH1; GSK3B; MAPK1; MYC; PIK3CA
Prostate cancer	3.49E-05	2.37E-04	AKT1; AKT3; AR; CCND1; CCNE1; CREB3; CREB3L4; GSK3B; MAPK1; PIK3CA
Breast cancer	1.25E-05	9.92E-05	AKT1; AKT3; CCND1; ESR1; ESR2; FOS; FRAT1; FRAT2; GSK3B; MAPK1; MYC; PGR; PIK3CA
Pathways in cancer	1.26E-07	2.29E-06	ADCY1; ADCY2; ADCY3; ADCY5; ADCY8; AKT1; AKT3; AR; CASP3; CCND1; CCNE1; CDH1; EGLN2; ESR1; ESR2; FOS; FRAT1; FRAT2; GSK3B; HIF1A; MAPK1; MYC; NCOA4; NFE2L2; PIK3CA; PLCB2; PLCB3; PPARG; PPARG; PTGS2; RXRG

^aFDR, corrected *p* value with Benjamini-Hochberg method.

Knowledge-Driven Analysis to Link BPF to Biological Events

As a next step, to be able to further support the association of BPF with AOPs, an unsupervised analysis of the existing literature was performed using the tool AOP-helpFinder, followed by a manual targeted literature analysis, and data integration. The main idea was to decipher potential linkages, based on existing knowledge, between BPF and the previously identified endocrine-related cancers by systems biology (Table 1).

First, an automatic screening of the literature was carried out using the newly developed AOP-helpFinder tool (Carvaillo et al., 2019). This tool was run on the 190 publications mentioning BPF extracted from the TOXLINE database, using the developed AOP dictionary (Carvaillo et al., 2019). This search led to the identification of 1 AO that is cancer, and 8 KEs (Supplementary Table 5). Allergic contact dermatitis challenge was the most common KE (KE 312), retrieved in 7 of 15 articles that mentioned a term related to AOPs, out of the 190 screened publications. Interestingly, BPA is associated with 63 dermatitis-related proteins in the CTD and BPF is also linked to dermatitis via 5 proteins (ASRGL1, BMP6, CXCL2, IFI30, also in CTD).

Among the other findings, the KEs (AOP-wiki KE 870 “Increase, Cell proliferation” and KE 1555 “Increase cell proliferation”) were linked to BPF (Perez et al., 1998), which is relevant since due to their potential endocrine-disrupting effects (eg, binding to estrogen receptors), bisphenols including BPF have been described as promoters of cell proliferation (Perez et al., 1998). Such a connection between BPF and cell proliferation events is in line with the known effect of estrogens on benign and malignant thyrocyte proliferation through transcriptional activation of various oncogenes including Bcl-2 or c-fos and stimulation of non-genomic pathways (including the ERK and Akt pathways which stimulate cell division) (Kumar et al., 2010; Manole et al., 2001).

Based on these text mining-based results and the ones from the systems biology analysis, we decided to further explore thyroid cancer, as this outcome was the most significant 1 identified by ORA. Therefore in a second literature-based step, associations between the stressor BPF, the biological events identified previously (cell proliferation and thyroid cancer), and the proteins found by ORA (CCND1; CDH1; MAPK1; MYC; NCOA4; PPARG; RXRG) (Table 1) were explored by manual

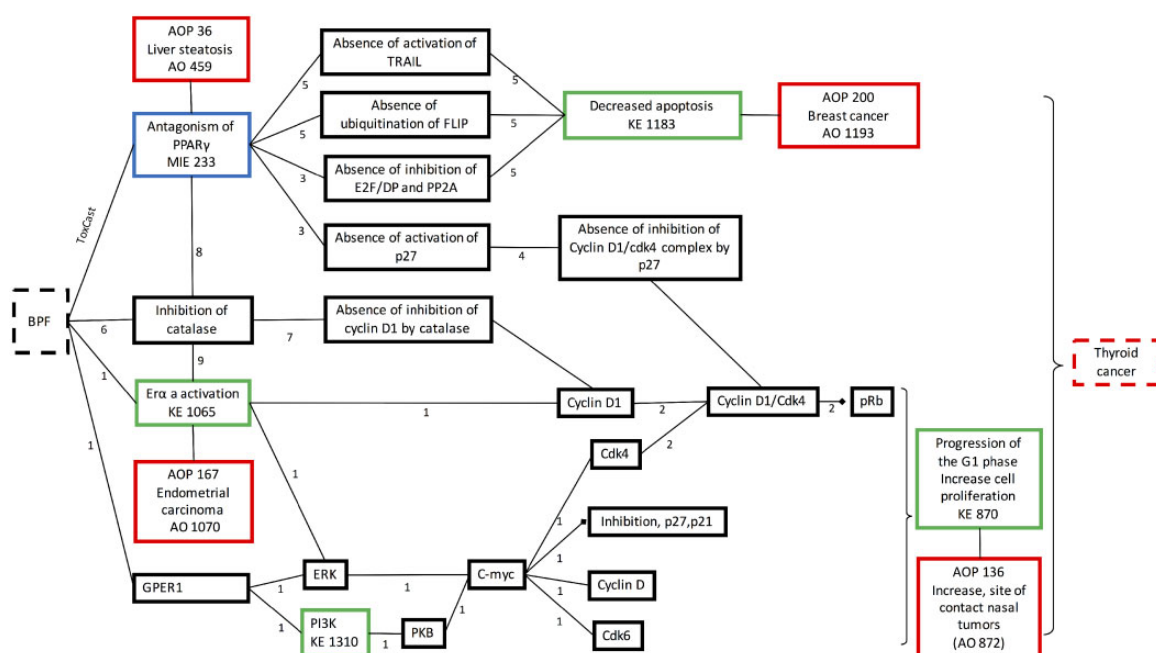


Figure 5. Representation of the adverse outcome pathway (AOP) network involving BPF. The primary AOP linked to thyroid cancer (dashed line) developed in Fig. 4 was enriched by querying the AOP-wiki database. Various events (MIEs, KEs and AOPs) were then added.

connected to BPF. By using such a higher-level combination of methods, we were not only able to link BPF to an individual AOP but we were also able to associate these isomers to a complex AOP network. These networks are particularly relevant to represent multiple outcomes of a substance and also of a mixture of substances. By combining different modules of text mining/scoring and of systems biology approaches, we provide a strategy to accelerate connection of a substance to AOs and identify potential critical steps of its putative mode of action.

One advantage of such approaches is the ability to gather disparate types of information (chemical-protein associations, PPIs, protein-signaling pathway annotations) from various sources (literatures, databases), and to integrate them with the aim to identify previously uncharacterized links (Audouze and Grandjean, 2011). The development of such models is now feasible due to recent advances in both experimental (via high-throughput technologies) and computational areas, eg, though advanced developed methods to identify association between chemicals and health disorders (Audouze et al., 2010). A recent study showed that AOP networks allowed to develop new AOPs, reflecting the power of generating AOP networks to better understand mechanistic pathways (Knapen et al., 2018). Adverse outcome pathway networks can also be used for assay development and refinement, as shown in a study that created an AOP network for reproductive and developmental toxicity in fish, based on 5 relevant existing AOPs (Knapen et al., 2015). A recently developed stressor-AOP network webserver, integrating information from the ToxCast and AOP-wiki databases, revealed that many chemical stressors can putatively interfere with 1 or several AOPs (Aguayo-Orozco et al., 2019). Due to the rapidly expanding available toxicology data (including omics), data-driven and computer-based tools are now believed to be a realistic option for the development of nonanimal-based hazard and risk assessment. Pipelines for generating and enriching AOP descriptions including literature mining and integration of

diverse data sources have been proposed recently (Carvaillo et al., 2019; Nymark et al., 2018).

Such integrative approach is essentially qualitative, and will be improved in the future to generate more quantitative models, that will take into consideration dose and time effects. A next step would be to include, for example, the dose-dependent activation of MIEs or detailed pathways of toxicity. Another aspect will be to establish quantitative AOP networks. A recent study examined how AOPs can be used to develop computational pathway-based quantitative models that will be useful for regulatory chemical safety assessment (Perkins et al., 2019). All such improvements could be envisaged by considering other data sources (eg, joint pathways analysis from cross-omics studies and databases such as Effectopedia and the human toxicology knowledgebase).

Compared with BPA, the effects of BPF on human health are poorly characterized. Most studies focused on its physiological and endocrine activities. Few models have been used to characterize such effects in humans; for example, a fetal testis assay developed in 3 different species (mouse, rat, human) showed that nanomolar concentrations of BPA, BPS, and BPF are able to reduce basal testosterone secretion in the human *ex vivo* model (Eladak et al., 2015). This *ex vivo* study suggested that as for BPA, BPF acts as an endocrine disruptor in humans. Such results are supported by *in vitro* studies using human cell lines which indicate estrogenic and antiandrogenic effects of BPF (Cabaton et al., 2009; Molina-Molina et al., 2013; Satoh et al., 2004). In addition to such endocrine-disrupting effects, our approach has suggested a link between BPF and cell proliferation. The putative effect on thyroid cancer appeared to be the most significant 1 and this was highlighted in the AOP model that is presented. Relevant initiating events were identified including inhibition of PPAR γ and activation of ER- α . Overall, these results could be used as the basis for further epidemiological and experimental studies, thus providing additional evidence for causal links between BPF

and tumor development. However, in addition to the potential influence of BPF on tumorigenesis, the linkage of 1 MIE (antagonism of PPAR γ) with liver steatosis (AOP36, A0459) is coherent with one of our recent findings suggesting that bisphenols could be associated with obesity and metabolic disruptions (Carvaillo *et al.*, 2019). Decreased PPAR γ activity leads to decreased expression of Hydroxysteroid 17-Beta Dehydrogenase 10 (or 3-hydroxyacyl-CoA dehydrogenase type-2) (KER260), leading to impaired β -oxidation and mitochondrial dysfunction (Yang *et al.*, 2011), and therefore to lipid accumulation. Interestingly, a direct relationship between obesity and thyroid diseases has been suggested (Santini *et al.*, 2014).

New and innovative computational strategies are needed to help in the identification of health effects of chemical substituents and mixture, and to link them to AOPs that could be used for regulatory risk assessment. We believe that the development of systems toxicology approaches that relies on existing data, as the one proposed here, is extremely relevant in the area of IATA that contribute to the reduction of animal testing. These approaches also contribute to associate exposure to chemicals of concern with actual hazards and health outcomes.

SUPPLEMENTARY DATA

Supplementary data are available at Toxicological Sciences online.

DECLARATION OF CONFLICTING INTERESTS

The author/authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ACKNOWLEDGMENT

The authors would like to acknowledge HBM4EU (<https://www.hbm4eu.eu/>).

FUNDING

This project has received funding from the European Union's Horizon 2020 research and innovation programme (733032 HBM4EU). This work was also supported by Université de Paris, Assistance Publique-Hôpitaux-de-Paris, and INSERM.

AUTHOR CONTRIBUTIONS

M.R. performed the text mining experiments. J.C.C. developed the AOP-helpFinder tool. M.R., X.C., and K.A. developed the AOP and AOP network. K.A. designed the study and performed the integrative systems biology experiments. All authors have contributed in discussing the study results and writing the manuscript.

REFERENCES

Aguayo-Orozco, A., Audouze, K., Siggaard, T., Barouki, R., Brunak, S., Taboureau, O. (2019). sAOP: Linking chemical stressors to adverse outcomes pathway networks. *Bioinformatics*. pii: btz570.

Ankley, G. T., Bennett, R. S., Erickson, R. J., Hoff, D. J., Hornung, M. W., Johnson, R. D., Mount, D. R., Nichols, J. W., Russom, C. L.,

Schmieder, P. K., *et al.* (2010). Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environ. Toxicol. Chem.* **29**, 730–741.

Audouze, K., and Grandjean, P. (2011). Application of computational systems biology to explore environmental toxicity hazards. *Environ. Health Perspect.* **119**, 1754–1759.

Audouze, K., Juncker, A. S., Roque, F. J. S. S. A., Krysiak-Baltyn, K., Weinhold, N., Taboureau, O., Jensen, T. S., and Brunak, S. (2010). Deciphering diseases and biological targets for environmental chemicals using toxicogenomics networks. *PLoS Comput. Biol.* **6**, e1000788.

Audouze, K., Taboureau, O., Grandjean, P. (2018). A systems biology approach to predictive developmental neurotoxicity of a larvicide used in the prevention of Zika virus transmission. *Toxicol. Appl. Pharmacol.* **354**, 56–63.

Bajard, L., Melymuk, L., and Blaha, L. (2019). Prioritization of hazards of novel flame retardants using the mechanistic toxicology information from ToxCast and adverse outcome pathways. *Environ. Sci. Eur.* **31**, 14.

Cabaton, N., Dumont, C., Severin, I., Perdu, E., Zalko, D., Cherkaoui-Malki, M., and Chagnon, M.-C. (2009). Genotoxic and endocrine activities of bis(hydroxyphenyl)methane (bisphenol F) and its derivatives in the HepG2 cell line. *Toxicology* **255**, 15–24.

Carvaillo, J.-C., Barouki, R., Coumoul, X., and Audouze, K. (2019). Linking bisphenol S to adverse outcome pathways using a combined text mining and systems biology approach. *Environ. Health Perspect.* **127**, 47005.

Davis, A. P., Wieggers, T. C., Wieggers, J., Johnson, R. J., Sciaky, D., Grondin, C. J., Mattingly, C. J. (2018). Chemical-induced phenotypes at CTD help inform the pre-disease state and construct adverse outcome pathways. *Toxicol. Sci.* **165**, 145–156.

Eladak, S., Grisin, T., Moison, D., Guerquin, M.-J., N'Tumba-Byn, T., Pozzi-Gaudin, S., Benachi, A., Livera, G., Rouiller-Fabre, V., Habert, R., *et al.* (2015). A new chapter in the bisphenol A story: Bisphenol S and bisphenol F are not safe alternatives to this compound. *Fertil. Steril.* **103**, 11–21.

FitzGerald, R. E., and Wilks, M. F. (2014). Bisphenol A—Why an adverse outcome pathway framework needs to be applied. *Toxicol. Lett.* **230**, 368–374.

Garcia-Reyero, N. (2015). Are adverse outcome pathways here to stay? *Environ. Sci. Technol.* **49**, 3–9.

ICCVAM (2018). A strategic roadmap for establishing new approaches to evaluate the safety of chemicals and medical products in the United States (<https://ntp.niehs.nih.gov/go/iccvam-rdmp>).

Judson, R. S., Houck, K. A., Kavlock, R. J., Knudsen, T. B., Martin, M. T., Mortensen, H. M., Reif, D. M., Rotroff, D. M., Shah, I., Richard, A. M., *et al.* (2010). *In vitro* screening of environmental chemicals for targeted testing prioritization: The ToxCast project. *Environ. Health Perspect.* **118**, 485–492.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2019). New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **47**, D590–595.

Kavlock, R., Chandler, K., Houck, K., Hunter, S., Judson, R., Kleinstreuer, N., Knudsen, T., Martin, M., Padilla, S., Reif, D., *et al.* (2012). Update on EPA's ToxCast program: Providing high throughput decision support tools for chemical risk management. *Chem. Res. Toxicol.* **25**, 1287–1302.

Kisková, T., Mungenast, F., Suváková, M., Jäger, W., Thalhammer, T. (2019). Future Aspects for cannabinoids in breast cancer therapy. *Int. J. Mol. Sci.* **20**, E1673.

Klinge, C. M., Blankenship, K. A., Risinger, K. E., Bhatnagar, S., Noisin, E. L., Sumanasekera, W. K., Zhao, L., Brey, D. M., and

- Keynton, R. S. (2005). Resveratrol and estradiol rapidly activate MAPK signaling through estrogen receptors alpha and beta in endothelial cells. *J. Biol. Chem.* **280**, 7460–7468.
- Knapen, D., Angrish, M. M., Fortin, M. C., Katsiadaki, I., Leonard, M., Margiotta-Casaluci, L., Munn, S., O'Brien, J. M., Pollesch, N., Smith, L. C., et al. (2018). Adverse outcome pathway networks I: Development and applications. *Environ. Toxicol. Chem.* **37**, 1723–1733.
- Knapen, D., Vergauwen, L., Villeneuve, D. L., and Ankley, G. T. (2015). The potential of AOP networks for reproductive and developmental toxicity assay development. *Reprod. Toxicol.* **56**, 52–55.
- Ko, H., Lee, J. H., Kim, H. S., Kim, T., Han, Y. T., Suh, Y.-G., Chun, J., Kim, Y. S., and Ahn, K. S. (2019). Novel galiellactone analogues can target STAT3 phosphorylation and cause apoptosis in triple-negative breast cancer. *Biomolecules* **9**, 170.
- Kongsbak, K., Vinggaard, A. M., Hadrup, N., and Audouze, K. (2014). A computational approach to mechanistic and predictive toxicology of pesticides. *ALTEX* **31**, 11–22.
- Kumar, A., Klinge, C. M., Goldstein, R. E. (2010). Estradiol-induced proliferation of papillary and follicular thyroid cancer cells is mediated by estrogen receptors alpha and beta. *Int. J. Oncol.* **36**, 1067–1080.
- Lage, K., Karlberg, E. O., Størling, Z. M., Ólason, P. Í., Pedersen, A. G., Rigina, O., Hinsby, A. M., Tümer, Z., Pociot, F., Tommerup, N., et al. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **25**, 309–316.
- Li, T., Wernersson, R., Hansen, R. B., Horn, H., Mercer, J., Slodkowitz, G., Workman, C. T., Rigina, O., Rapacki, K., Stærfeldt, H. H., et al. (2017). A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat. Methods* **14**, 61–64.
- Liao, Y., Wang, J., Jaehnic, E. J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: Gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–205.
- Manole, D., Schildknecht, B., Gosnell, B., Adams, E., Derwahl, M. (2001). Estrogen promotes growth of human thyroid tumor cells by different molecular mechanisms. *J. Clin. Endocrinol. Metab.* **86**, 1072–1077.
- Molina-Molina, J.-M., Amaya, E., Grimaldi, M., Sáenz, J.-M., Real, M., Fernández, M. F., Balaguer, P., and Olea, N. (2013). *In vitro* study on the agonistic and antagonistic activities of bisphenol-S and other bisphenol-A congeners and derivatives via nuclear receptors. *Toxicol. Appl. Pharmacol.* **272**, 127–136.
- Nymark, P., Rieswijk, L., Ehrhart, F., Jeliaskova, N., Tsiliki, G., Sarimveis, H., Evelo, C. T., Hongisto, V., Kohonen, P., Willighagen, E., et al. (2018). A data fusion pipeline for generating and enriching adverse outcome pathway descriptions. *Toxicol. Sci.* **162**, 264–275.
- Oki, N. O., and Edwards, S. W. (2016). An integrative data mining approach to identifying adverse outcome pathway signatures. *Toxicology* **350–352**, 49–61.
- Perez, P., Pulgar, R., Olea-Serrano, F., Villalobos, M., Rivas, A., Metzler, M., Pedraza, V., and Olea, N. (1998). The estrogenicity of bisphenol A-related diphenylalkanes with various substituents at the central carbon and the hydroxy groups. *Environ. Health Perspect.* **106**, 167–174.
- Perkins, E. J., Ashauer, R., Burgoon, L., Conolly, R., Landesmann, B., Mackay, C., Murphy, C. A., Pollesch, N., Wheeler, J. R., Zupanic, A., et al. (2019). Building and applying quantitative adverse outcome pathway models for chemical hazard and risk assessment. *Environ. Toxicol. Chem.* **38**, 1850–1865.
- Pollesch, N. L., Villeneuve, D. L., O'Brien, J. M. (2019). Extracting and benchmarking emerging adverse outcome pathway knowledge. *Toxicol. Sci.* **168**, 349–364.
- Sakuratani, Y., Horie, M., and Leinala, E. (2018). Integrated approaches to testing and assessment (IATA): OECD activities on the development and use of adverse outcome pathways and case studies. *Basic Clin. Pharmacol. Toxicol.* **123**, 20.
- Santini, F., Marzullo, P., Rotondi, M., Ceccarini, G., Pagano, L., Ippolito, S., Chiovato, L., and Biondi, B. (2014). Mechanisms in endocrinology: The crosstalk between thyroid gland and adipose tissue: Signal integration in health and disease. *Eur. J. Endocrinol.* **171**, R137–152.
- Sato, K., Feng, X., Chen, J., Li, J., Muranski, P., Desierto, M. J., Keyvanfar, K., Malide, D., Kajigaya, S., Young, N. S., et al. (2016). PPAR γ antagonist attenuates mouse immune-mediated bone marrow failure by inhibition of T cell function. *Haematologica* **101**, 57–67.
- Satoh, K., Ohyama, K., Aoki, N., Iida, M., and Nagai, F. (2004). Study on anti-androgenic effects of bisphenol a diglycidyl ether (BADGE), bisphenol F diglycidyl ether (BFDGE) and their derivatives using cells stably transfected with human androgen receptor, AR-EcoScreen. *Food Chem. Toxicol. Int. J. Publ. Br. Ind. Biol. Res. Assoc.* **42**, 983–993.
- Sutherland, R. L., Prall, O. W., Watts, C. K., and Musgrove, E. A. (1998). Estrogen and progesterin regulation of cell cycle progression. *J. Mammary Gland Biol. Neoplasia* **3**, 63–72.
- Tollefsen, K. E., Scholz, S., Cronin, M. T., Edwards, S. W., de Knecht, J., Crofton, K., Garcia-Reyero, N., Hartung, T., Worth, A., Patlewicz, G., et al. (2014). Applying adverse outcome pathways (AOPs) to support integrated approaches to testing and assessment (IATA). *Regul. Toxicol. Pharmacol.* **70**, 629–640.
- Villeneuve, D. L., Crump, D., Garcia-Reyero, N., Hecker, M., Hutchinson, T. H., LaLone, C. A., Landesmann, B., Lettieri, T., Munn, S., Nepelska, M., et al. (2014). Adverse outcome pathway (AOP) development I: Strategies and principles. *Toxicol. Sci.* **142**, 312–320.
- Williams, A. J., Grulke, C. M., Edwards, J., McEachran, A. D., Mansouri, K., Baker, N. C., Patlewicz, G., Shah, I., Wambaugh, J. F., Judson, R. S., et al. (2017). The CompTox chemistry dashboard: A community data resource for environmental chemistry. *J. Cheminform.* **9**, 61.
- Yang, S.-Y., He, X.-Y., and Miller, D. (2011). Hydroxysteroid (17 β) dehydrogenase X in human health and disease. *Mol. Cell. Endocrinol.* **343**, 1–6.

Titre : De l'unité d'assemblage à la capsid : application in silico au norovirus et au virus de l'hépatite B

Mots clés : modélisation moléculaire, amarrage moléculaire, assemblage moléculaire, région intrinsèquement désordonnée

Résumé : Les virus comportent une coque permettant de protéger et de transporter l'information virale. Cette coque est composée de plusieurs copies de protéines capables de s'assembler d'elles-mêmes, autour ou indépendamment de l'information virale. Au cours de la vie d'un virus, les protéines impliquées dans l'assemblage sont dotées de multiples rôles. En plus du transport et de la protection, elles peuvent réguler ou détourner la machinerie cellulaire de l'hôte infecté. Elles peuvent aussi servir de leurre face aux défenses immunitaires de l'organisme. Dans cette thèse, à l'aide de méthodes calculatoires sur des superordinateurs, les mécanismes d'assemblage et d'interaction de ces protéines sont étudiés.

Ces travaux permettent d'avoir une meilleure idée de l'assemblage de la coque vide du virus de la gastroentérite de l'Homme. En ce qui concerne le virus de l'hépatite B (VHB), ces études mettent en évidence la fixation d'ions zinc par la protéine de coque du VHB. Le zinc pourrait jouer un rôle essentiel dans l'interaction avec le matériel génétique. Ces travaux aident aussi à comprendre comment cette protéine transporte l'information virale vers le noyau des cellules du foie.

Title : From assembly unit to capsid: in silico application to norovirus and hepatitis B virus

Keywords : molecular modeling, docking, molecular assembly, intrinsically disordered region

Abstract : Viruses are characterized by a shell that protects and transports viral information. This shell is composed of several copies of proteins capable of self-assembling, around or independently of the viral information. During the life of a virus, proteins involved in this assembly have multiple roles. In addition to transport and protection, they can regulate or hijack the cellular machinery of the infected host. They can also serve as decoys against the organism's immune defenses. In this thesis, using computational methods on supercomputers, the mechanisms of assembly and interaction of these proteins are studied. This work allows to have a better idea of the assembly of the empty shell of the human gastroenteritis virus.

For hepatitis B virus (HBV), these studies show the binding of zinc ions by the HBV shell protein. Zinc may play an essential role in the interaction with genetic material. This work also helps to understand how this protein transports viral information to the nucleus of liver cells.