Identification de nouvelles voies d'inhibition ciblant les mouvements fonctionnels de protéines : application à la transition allostérique du récepteur nicotinique de l'acétylcholine

Damien Monet

Université Pierre et Marie Curie
ED515 - Complexité du Vivant

Unité de Bioinformatique Structurale
Institut Pasteur

Thèse de doctorat en Sciences de la Vie

Identification de nouvelles voies d'inhibition ciblant les mouvements fonctionnels de protéines : application à la transition allostérique du récepteur nicotinique de l'acétylcholine

Damien Monet

Directeur de thèse : **Pr. Michael Nilges**
Co-encadrant de thèse : **Dr. Arnaud Blondel**

Présentée et soutenue publiquement le 6 décembre 2018

Jury

Pr. Anne-Claude Camproux	Rapporteur
Pr. Annick Dejaegere	Rapporteur
Pr. Catherine Vénien-Bryan	Examinateur
Dr. Bogdan I. Iorga	Examinateur
Pr. Michael Nilges	Examinateur
Dr. Arnaud Blondel	Examinateur

Identification de nouvelles voies d'inhibition ciblant les mouvements fonctionnels de protéines : application à la transition allostérique du récepteur nicotinique de l'acétylcholine

Thèse de doctorat en Sciences de la Vie

Rapporteurs : Pr. Anne-Claude Camproux et Pr. Annick Dejaegere

Examinateurs : Pr. Catherine Vénien-Bryanet Dr. Bogdan I. Iorga

Directeur de thèse : Pr. Michael Nilges

Co-Directeur de thèse : Dr. Arnaud Blondel

Université Pierre et Marie Curie

ED515 - Complexité du Vivant

Unité de Bioinformatique Structurale

Institut Pasteur

25-28 Rue du Docteur Roux

75724 Paris Cedex 15

Résumé

L'étude de la dynamique fonctionnelle de protéines impliquées dans des processus pathologiques et en corollaire des fluctuations associées de leurs cavités et poches offre de nouvelles stratégies pour l'identification de molécules effectrices. Ainsi, durant ma thèse, j'ai décrit, grâce au développement de méthodes *in silico* novatrices, la transition d'activation d'un récepteur nicotinique de l'acétylcholine. L'analyse dynamique des cavités apporte un modèle des mécanismes allostériques en jeu au niveau atomique et supporte la recherche de sites présentant un levier sur la fonction du récepteur.

Les récepteurs nicotiniques sont des canaux exprimés à la surface des cellules nerveuses et musculaires. La fixation de l'acétylcholine induit l'ouverture du canal et ainsi, l'entrée de cations dans la cellule. Le sous-type $(\alpha 7)_5$ est impliqué dans des processus cognitifs et certains désordres neurodégénératifs, ce qui en fait une cible thérapeutique de choix. Le processus d'activation du récepteur a été modélisé par une série de conformations intermédiaires reliant les états de repos et actif. J'ai pu coupler une méthode de calcul de chemins de transition adiabatique développée au laboratoire avec la méthode dynamique *String of Swarms*, adaptée à cette fin, pour faire évoluer les intermédiaires dans le champ d'énergie libre tout en maintenant une description essentielle du mécanisme fonctionnel. Notre modèle de transition reproduit correctement les mouvements quaternaires connus, le *blooming* et le *twisting*. Parallèlement, nous avons mis au point, un algorithme robuste permettant de donner une vision unitaire des cavités issues de conformations structurales différentes et situées dans des régions similaires de la protéine. Ces groupes cohérents de cavités définissent des poches, sites potentiels pour la liaison de ligands. La méthode, sa mise au point et sa validation sur des dynamiques d'un ensemble de 15 sites de référence sont décrites. Un programme, *mkgridXf*, implémente le suivi des cavités et l'identification cohérente de sites sur les trajectoires de protéines. La cartographie des cavités de la transition $(\alpha 7)_5$ a révélé 68 sites distincts. 6 ont un volume variant de façon significative avec l'état conformationnel. Parmi eux, nous retrouvons le site orthostérique, le site modulateur Ca^{2+} ainsi que 2 sites allostériques précédemment décrits. En complément, le *docking* de modulateurs allostériques sur l'ensemble des sites de la transition permet de proposer l'existence d'un site de liaison effecteur à un *locus* proche du site de l'ivermectine dans les structures publiées des récepteurs GluCl et Glycine. Ces données suggèrent de nouvelles routes de dessins d'effecteurs ayant des activités ciblées.

Abstract

The analysis of the functional motion of proteins involved in various diseases and the associated evolution of cavities and grooves offers novel strategies to identify effector molecules. During my Ph.D. thesis, I modeled the nicotinic acetylcholine receptor gating mechanism, thanks to the development of innovative *in silico* methods. The cavity analysis in the dynamical model brings a description of the allosteric mechanism and support the research of protein sites presenting a lever on the receptor function.

Nicotinic receptors are ligand gated ion channels found on nervous and muscular cells membrane. The binding of acetylcholine induces the channel gating and an ion flow in the cell. The $(\alpha 7)_5$ subtype is involved in cognitive processes and various neurological disorders. The activation mechanism has been modeled by a series of intermediate conformations linking the resting and the active states of the receptor. I coupled and adapted the *String of Swarms* dynamic method with an adiabatic transition path calculation method *Path Optimization and Exploration* to let the intermediates evolve in the free energy landscape while preserving an essential description of the functional mechanism. Our transition model showed the two well defined quaternary motion, the blooming and twisting, as already highlighted on homologous receptors. We also developed a robust algorithm to consistently track cavities in protein dynamics. Groups of protein cavities define pockets, potential binding sites for small molecules. The approach and its validation on the molecular dynamics of 15 reference sites are described. A practical implementation, *mkgridXf*, is given to automatically track and identify sites in protein trajectories. The complete mapping of cavities during the $(\alpha 7)_5$ transition revealed 68 distinct sites. 6 had a size and shape varying significantly with the conformational state of the protein. Interestingly, among these selected pockets, we found the orthosteric site, the Ca^{2+} modulatory site and 2 previously described allosteric sites. Additionally, the molecular docking of allosteric modulators in all the pockets identified along the transition suggested the existence of an effector binding site close to the ivermectin site found in the previously published structures of GluCl and Glycine receptors homologs.

Remerciements

Tout d'abord, je remercie Michael Nilges de m'avoir offert le privilège d'intégrer l'Unité de Bioinformatique Structurale ainsi que d'avoir accepté la responsabilité de diriger ma thèse. Je remercie chaleureusement Arnaud Blondel pour son encadrement durant ces quelques années de stage de master puis de thèse dans le laboratoire. Ton expertise et ta rigueur scientifique auront été des exemples. Plus que tout je suis reconnaissant de la confiance que tu m'as accordée et qui m'a sans doute permis d'aller jusqu'au bout de ce projet.

Je remercie également l'ensemble de mes collègues, passés et présent avec qui j'ai partagé tant de moments, dans la science et dans la bonne humeur. Merci Laura, Irène, Fabrice, Nathan compagnons de thèse, Guillaume, Benjamin, Mathias partenaires de tennis de table, Nathalie et Maya pour votre soutien de tous les jours, et Tru pour tes conseils techniques avisés.

Un grand merci à ma famille, à mes amis, Théo, Baptiste, Romain, Clara et à Lucie pour sa relecture méticuleuse du manuscrit. Merci à Corinne et Bruno pour ces week-end ressourçant à Viglain. Et merci à toi, Anna, pour m'avoir apporté ton réconfort durant toutes ces années.

Enfin, j'exprime ma profonde gratitude à Anne-Claude Camproux, Annick Dejaegere, Catherine Vénien-Bryan et Bogdan Iorga pour avoir accepté mon invitation à rejoindre le jury de thèse.

Cette thèse a bénéficié d'un contrat doctoral de l'Ecole Doctorale Complexité du Vivant (ED-515), d'un financementde l'Union Européenne Horizon 2020 HBP SGA1 (Grant Agreement No. 785907) et de l'ANR Nicofive (ANR-17-CE11-0030).

Table des matières

Table des figures

Liste des tableaux

I

Introduction

1 Stratégie de recherche

Les premiers médicaments remontent à la nuit des temps. Ces découvertes ont été faites de manière empirique, sans souci du mécanisme en jeu. Par exemple, dans l'Égypte ancienne, des extraits de plantes étaient connus pour faire baisser la fièvre [1] (dont la molécule active, l'acide acétylsalicylique, ou *aspirine*, a été révélée plus tard). D'autres substances, comme les opiacés, ont toujours une place importante dans la médecine moderne [2]. À partir du XIXe siècle, grâce à l'augmentation des connaissances en physiologie, les découvertes sont *fonctionnelles*. Les effets thérapeutiques s'observent alors au travers de leur impact sur la physiologie des organes. Le processus de découverte est encore long et aléatoire. Ces 40 dernières années, les techniques de conception de médicaments se sont progressivement orientées sur l'analyse des pathologies au niveau moléculaire. L'identification des acteurs moléculaires impliqués dans la maladie supporte la découverte de nouvelles molécules et la validation de leurs mécanismes d'action [3, 4]. De nos jours les nouvelles entités chimiques sont identifiées par chimie combinatoire, criblage haut débit de librairies de composés chimiques ou par des approches rationnelles qui utilisent la connaissance détaillée du mécanisme pathologique. Le dessin de molécules thérapeutiques assisté par ordinateurs, et plus spécifiquement, par criblage virtuel, peut accélérer significativement le processus de découverte [5]. Cependant, ces approches requièrent la connaissance préalable d'inhibiteurs ou de la structure de la protéine cible avec un site de liaison principal. Ainsi, la découverte de sites de liaison potentiels sur des protéines cibles nouvelles ou déjà connues est essentielle pour créer de nouvelles opportunités et trouver de nouvelles familles d'inhibiteurs. Les avancées techniques et méthodologiques obtenues pendant cette thèse s'inscrivent dans les recherches menées ces dernières années dans le laboratoire de Bioinformatique Structurale de l'Institut Pasteur pour identifier de nouvelles molécules effectives contre diverses pathologies. La stratégie choisie, décrite dans la Figure I.1/p.4, passe par le couplage de méthodes décrivant le mécanisme fonctionnel du facteur pathologique avec l'analyse de l'évolution des cavités et poches pour identifier de nouveaux sites de liaisons, suivi de la recherche de ligands par criblage virtuel et le test *in vitro* des molécules sélectionnées en collaboration avec les Unités de recherche expérimentales de l'Institut Pasteur. En utilisant des calculs de chemins de transition, des recherches au laboratoire ont pu identifier un site allostérique précédemment inconnu sur la toxine de l'anthrax et en dessiner des inhibiteurs actifs [6]. Cette approche a aussi été utilisée pour élargir l'espace chimique exploré en exploitant la flexibilité de la protéine cible comme illustré avec l'identification de nouvelles familles d'inhibiteurs ciblant la Proline Racemase *T.cruzi* [7, 8]. Des résultats théoriques ont aussi démontré la corrélation entre l'évolution géométrique des cavités et les mouvements fonctionnels

de protéines [9]. Ces réussites passées justifient et ouvrent la voie à l'analyse de systèmes moléculaires plus complexes et intriqués, comme présenté dans cette thèse avec l'étude de la transition conformationnelle du récepteur nicotinique.

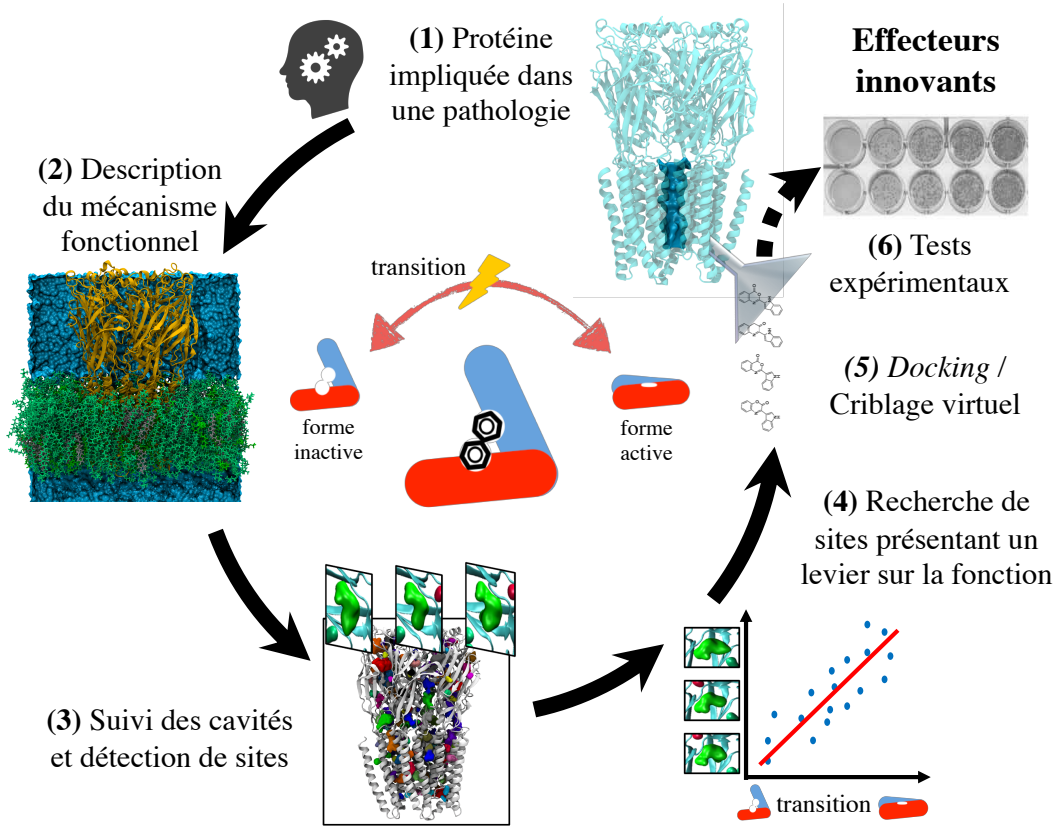


Figure I.1 Stratégie pour la recherche d'effecteurs ciblant les mouvements fonctionnels de protéines. **(1)** L'approche vise des protéines cibles impliquées dans des processus pathologiques graves pour lesquelles des traitements thérapeutiques sont inexistants ou insuffisants. **(2)** La description *in silico* du mécanisme de la cible, grâce à des algorithmes de dynamique moléculaire et de calculs de chemins de transition, est une aide à la compréhension des déterminants structuraux de l'activité de la protéine et supporte la recherche de leviers mécanistiques. **(3)** Les cavités forment une interface privilégiée entre le solvant et la protéine. Leur suivi cohérent le long des dynamiques de protéine nous aide à isoler les sites de la protéine susceptibles de fixer de petites molécules. **(4)** L'analyse systématique de l'évolution des sites (forme, volume) permet d'identifier les sites significativement impactés par la dynamique fonctionnelle de la cible. **(5)** Des calculs d'amarrage et de criblage *in silico* identifient de petites molécules dont les interactions avec les sites choisis sont favorables dans l'état fonctionnel à privilégier. **(6)** Des tests expérimentaux peuvent être menés en collaboration avec d'autres laboratoires pour vérifier la pertinence des prédictions.

2 Objet de l'étude

Certains processus pathologiques s'expriment à travers l'action de protéines impliquant un mécanisme moléculaire complexe. L'activation d'une enzyme, de récepteurs *via* l'ouverture d'un canal, des processus de fusion des membranes, en sont des exemples. Le récepteur nicotinique est considéré comme "le père fondateur" pour la compréhension des mécanismes de transition allostériques subtils qui gouvernent la fonction des récepteurs ionotropes pentamériques [10]. Ces récepteurs ont leur importance dans des processus physiologiques cognitifs et sensoriels et se trouvent impliqués dans de nombreuses pathologies : désordres neurodégénératifs (par exemple les maladies d'Alzheimer et de Parkinson), la schizophrénie, l'addiction aux drogues et les traitements de la douleur. De plus, des effecteurs existants, mais dont les modes d'action sont parfois peu connus, peuvent agir avec des différences subtiles : antagonistes, agonistes, agonistes partiels, modulateurs allostériques positifs, négatifs, etc., et ciblent des sous-types de récepteurs spécifiques ou avoir un spectre d'actions plus large.

La fonction de ces récepteurs s'exprime par un mouvement conformationnel étendu caractérisé par l'ouverture du canal ionique situé à près de 50 Å de distance du site de fixation principale. Grâce à l'expérience acquise du laboratoire dans l'étude des mouvements fonctionnels de protéines, ce mécanisme allostérique nous a semblé être particulièrement intéressant à étudier pour les défis scientifiques et thérapeutiques qu'il pose.

Les différents challenges auxquels j'ai dû faire face durant la thèse ont nécessité des développements méthodologiques profonds en lien avec les spécificités de la cible :

- *L'absence de structures expérimentales haute résolution du pentamère* $\alpha 7$ fait obstacle à l'analyse du mécanisme au niveau atomique. Une méthode originale de modélisation a été mise au point pour créer des modèles des états actif et de repos du récepteur en intégrant les données structurales hétérogènes de récepteurs homologues (Partie II/p.21).

- *La sensibilité du récepteur aux conditions physiologiques* (présence et composition de la membrane, thermalisation du système) appuie une représentation complète de l'environnement de la protéine lors des modélisations. Des méthodes explorant la surface d'énergie libre grâce à la dynamique moléculaire offrent un cadre théorique adapté pour décrire des états intermédiaires plau-

sibles de la transition allostérique du récepteur. Cependant, l'aspect chaotique de la dynamique peut altérer l'interprétabilité des résultats. La combinaison des méthodes de calcul de chemins de transition dynamiques et adiabatiques nous a aidés à garder une description claire du processus (Partie III/p.47).

- *Des réorganisations structurales globales* lors de l'activation du récepteur impactent la dynamique interne des cavités et rendent particulièrement difficile leur suivi cohérent au cours de la transition. Des développements techniques et conceptuels ont permis d'élargir l'applicabilité du suivi dynamique des cavités à des trajectoires moléculaires de grande taille (nombre d'atomes et de conformations) et aux fluctuations variées (Partie IV/p.85).

- *La grande diversité des interactions entre la protéine et le solvant* complexifie la prédiction correcte des modes de liaison de petites molécules sur le récepteur. L'ajout explicite de l'interaction cation-π à un programme d'amarrage moléculaire et l'évaluation comparée des méthodes existantes de mesure *in silico* d'affinité ligand/protéine nous ont permis d'établir un protocole pour la prédiction du site de liaison de molécules effectrices connues (Partie V/p.131).

Mises bout à bout ces avancées méthodologiques ont mis en lumière des déterminants structuraux de l'activation du récepteur. Deux mouvements quaternaires principaux, déjà mis en valeur pour des récepteurs homologues, sont apparus lors des raffinements successifs des chemins de transition (Partie III/p.47). L'analyse dynamique et systématique de l'évolution des cavités nous a ensuite aidés à cartographier la multitude de sites apparaissant et disparaissant le long de la transition (Partie IV/p.85). De façon intéressante, un sous-ensemble des sites dont la géométrie est impactée par le changement d'états du récepteur représente des sites effecteurs déjà caractérisés du récepteur nicotinique, ou semblables à certains sites allostériques de récepteurs canaux proches (Partie V/p.131). Cela renforce l'importance de l'étude des cavités dans la description des mécanismes de modulation allostériques et pose les bases pour le dessin de nouveaux effecteurs à l'activité ciblée.

3 Récepteurs nicotiniques de l'acétylcholine

3.1 Contexte biologique

Le récepteur nicotinique de l'acétylcholine (nAChR) est une protéine transmembranaire faisant partie de la famille des canaux ioniques pentamériques modulés par des ligands (*pentameric Ligand Gated Ion Channel* ou pLGIC). Les pLGICs ont principalement un rôle de transmission et de régulation de l'information entre les cellules. Ces récepteurs convertissent un signal chimique, la liaison avec une molécule spécifique, en un signal électrique, grâce à l'ouverture du canal et l'entrée régulée d'ions à l'intérieur de la cellule. Certains récepteurs sont spécifiquement perméables aux cations - récepteurs à acétylcholine (nAChR) ou à la sérotonine (5-HT_3R) - ou aux anions - récepteurs à l'acide γ-aminobutyrique ($GABA_A$R), à la glycine (GlyR) ou au glutamate (GluClR). Leur rôle est décisif pour la progression des impulsions nerveuses entre les neurones. À l'extrémité terminale d'un neurone (membrane présynaptique), le signal codé électriquement est converti en signal chimique dans la synapse sous forme de neurotransmetteurs (par exemple l'acétylcholine). La liaison de ces signaux chimiques avec leurs récepteurs respectifs à la surface de la membrane post-synaptique du neurone suivant provoque l'entrée d'ions dans la cellule, et dans le cas d'une entrée de cations la dépolarisation de la cellule neuronale. À l'inverse, le passage d'anion réduit le potentiel d'action en hyperpolarisant la cellule et inhibe la transmission du signal. Une alternance du gradient d'ions le long de l'axone du neurone par des canaux tensiodépendants *via* une onde de dépolarisation qui permet la transmission du signal jusqu'à sa membrane présynaptique, et la libération de neurotransmetteurs dans la synapse suivante. À la jonction neuromusculaire, le potentiel d'action créé par l'ouverture des canaux cationiques conduit à la contraction du muscle.

Le récepteur nicotinique se présente sous la forme d'un pentamère dont les sous-unités peuvent être similaires (homopentamère), ou différentes (hétéropentamère). Depuis l'isolation de la première séquence d'acide aminé d'un récepteur nicotinique par Jean-Pierre Changeux en 1970 [11], 17 types de sous-unités ont été découverts chez les vertébrés [12]. Leur numérotation, basée sur leur homologie de séquence [13], comprend les sous-types $\alpha 1$ à $\alpha 10$, $\beta 1$ à $\beta 4$, δ, ϵ et γ. La combinatoire de pentamères est donc potentiellement énorme ($> 10^6$). Tous les assemblages n'étant pas compatibles, une trentaine de combinaisons différentes ont tout de même été identifiées chez les vertébrés [12]. Les récepteurs nicotiniques s'expriment dans le système nerveux central, périphérique, aux jonctions neuromusculaires, ainsi que sur

une variété d'autres tissus [14]. De façon intéressante, leur localisation, par exemple dans les différentes aires du cerveau, conditionne des compositions spécifiques de sous-types [12]. Cela suggère des réponses fonctionnelles différentes en fonction de la stœchiométrie du pentamère, confirmées par des expériences d'électrophysiologie et de pharmacologie : des nAChRs de compositions différentes présentes des vitesses d'activation et des affinités aux ligands différentes [15]. L'application prolongée de molécules activatrices (agonistes) résulte aussi en la *désensibilisation* des récepteurs, c'est-à-dire l'arrêt de l'entrée des ions malgré la présence d'agonistes. Cet état intermédiaire, réversible après arrêt de l'application d'agonistes permettrait la prévention de l'excitotoxicité et aurait un rôle important dans la régulation de l'efficacité synaptique [16]. La pharmacologie de ces récepteurs est large, diverse, et parfois spécifique à certains sous-types du récepteur. Il existe des agonistes (l'acétylcholine étant l'activateur endogène), des antagonistes compétitifs et non compétitifs, des modulateurs allostériques positifs, négatifs, etc., permettant la modulation fine des réponses fonctionnelles associées (voir l'introduction de la Partie 1/p.132 ou la référence [17]). Pourtant, les modes d'action de ces molécules restent souvent obscures et appellent à une meilleure compréhension des mécanismes allostériques en jeu. Lors de cette thèse je me suis intéressé au récepteur nicotinique neuronal de sous-unité $(\alpha 7)_5$ (5 chaînes de type $\alpha 7$).

3.2 Intérêt thérapeutique

Les récepteurs nicotiniques de sous-unités $\alpha 7$, de par leur prépondérance dans le système nerveux central, sont potentiellement impliqués dans différentes pathologies neurologiques et psychiatriques tel que la schizophrénie [18–20], l'épilepsie [21], l'autisme [21, 22], la maladie de Parkinson [23] et autres désordres neuropsychiatriques [21]. Le déficit cholinergique observé dans les premiers stades de la maladie d'Alzheimer pourrait s'expliquer par des interactions directes entre le peptide amyloid-β et le récepteur $\alpha 7$ [24, 25]. De plus, l'activation du récepteur $\alpha 7$ aurait un effet protecteur sur certaines blessures ou pathologies sévères (brûlures, trauma, septicémies, pancréatites aiguë) par modulation de la réponse immunitaire [26], grâce à son expression dans la voie cholinergique anti-inflammatoire,. Malgré les nombreux efforts entrepris pour la recherche de nouveaux médicaments ciblant le récepteur, aucune molécule thérapeutique disponible sur le marché ne cible directement le récepteur $\alpha 7$. Pour l'ensemble des récepteurs nicotinique, seule la Varénicline, agoniste partiel du récepteur $\alpha 4\beta 2$, est approuvée par la *Food And Drug Administration* pour lutter contre l'addiction au tabac, avec des effets indésirables notables [27, 28].

3.3 Aspects structuraux

Les premières observations expérimentales de récepteurs nicotiniques ont été publiées dans les années 70, avec des images de microscopie électronique issues des organes électriques riches en nAChRs provenant de la torpille marbrée *Torpedo Marmorata* et de l'anguille électrique (*Electrophorus Electricus*) [29]. Il apparaît clair que le récepteur est constitué de multiples sous-unités agencées en anneaux autour d'un axe de symétrie centrale formant un pore. Des études complémentaires ont ensuite démontré l'organisation pentamérique du récepteur [30–32]. Le premier modèle atomique expérimental d'un récepteur nicotinique (type musculaire $\alpha\gamma\alpha\beta\delta$ de l'organisme *Torpedo Marmorata*) a été dévoilé par microscopie électronique en 2005 [33]. En 2016, la structure cristallographique du récepteur $\alpha4\beta2$ a permis d'identifier un état potentiellement désensibilisé du récepteur [34]. Cependant, aucune structure expérimentale n'est encore disponible pour l'homopentamère $\alpha7$, et les études structurales ne sont permises que par comparaison avec des récepteurs homologues.

Le récepteur nicotinique est constitué de 5 sous-unités symétriquement réparties autour d'un pore central traversant la membrane. 3 domaines distincts sont ainsi définis : la partie extracellulaire qui porte les sites effecteurs principaux (orthostériques), la partie transmembranaire et un domaine intracellulaire (cf. Figure I.2/p.suiv.). Le domaine extracellulaire, considéré chaîne par chaîne, est composé de structures secondaires en feuillets bêtas (numérotés de 1 à 10) et d'une hélice alpha en N-terminale. Plusieurs boucles, chaînes polypeptidiques reliant les feuillets β consécutifs dans la séquence, sont fonctionnellement importantes : les boucles A ($\beta4$-$\beta5$), Cys ($\beta6$-$\beta7$), B ($\beta8$-$\beta9$) et C ($\beta9$-$\beta10$). La boucle Cys englobe 13 résidus très conservés entre deux résidus cystéines reliés par un pont disulfure caractéristique des récepteurs à boucle Cys. Le site orthostérique est formé par les résidus des boucles A, B et C d'une sous-unité α, et les boucles D, E, F d'une autre sous-unité. Sa configuration atomique est relativement bien connue grâce aux structures cristallographiques d'*acetylcholine binding protein*, des protéines de mollusques qui n'expriment qu'une partie extracellulaire et dont des morceaux de la séquence peuvent être mutés pour correspondre à l'$\alpha7$ [35, 36]. De façon intéressante, la poche orthostérique contient plusieurs résidus aromatiques, responsables de la reconnaissance des ligands au travers d'interactions cation-π [37]. Le domaine transmembranaire est constitué de 4 hélices α par sous-unités, numérotées M1, M2, M3 et M4 suivant l'ordre des hélices dans la séquence. Le pore central du récepteur est délimité par la succession des 5 hélices M2. Un anneau de constriction composé de résidus hydrophobes au centre du pore serait responsable de l'imperméabilité aux ions quand le récepteur est en état de repos [38]. Le domaine intracellulaire est formé par une extension entre les hélices M3 et M4. Il serait responsable de l'assemblage et de la localisation du récepteur à la

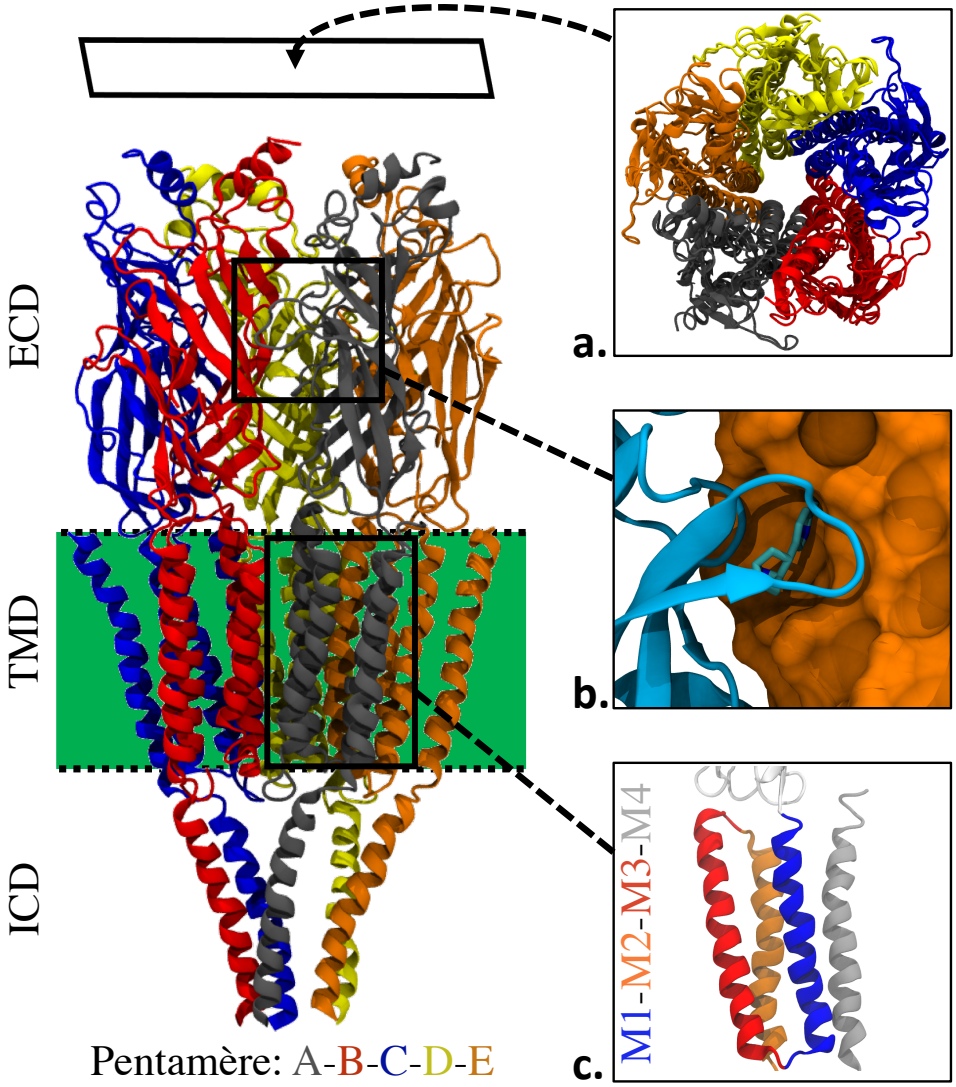


Figure I.2 Structure tridimensionnelle du récepteur nicotinique. À gauche, vu transversale de la structure du récepteur nicotinique de type musculaire de *Torpedo Marmorata* (PDB : 2BG9), colorée selon ses différentes sous-unités. ECD, TMD, ICD identifient respectivement les domaines extracellulaire, transmembranaire et intracellulaire. **a.** vu de dessus, le pore perméable aux ions est au centre. **b.** poche orthostérique du récepteur $\alpha 7$ (chimère AchBP *Lymnaea stagnalis*, PDB : 3SQ6) liée à l'agoniste épibatidine. La sous-unité principale (boucle C), est en représentation Cartoon, et la sous-unité complémentaire en surface orange. **c.** Zoom sur les 4 hélices transmembranaires d'une des sous-unités du récepteur, colorées selon les différentes hélices.

surface des cellules [39, 40]. La chaîne polypeptidique intracellulaire est supposée être dans un état peu ordonné, et est difficilement observable dans les structures expérimentales [33]. La troncation de la majeure partie de la région intracellulaire permet tout de même d'obtenir des récepteurs fonctionnels [34, 41].

3.4 Aspects dynamiques

Des structures cristallographiques de récepteurs homologues ont permis la caractérisation au niveau atomique d'états divers des récepteurs canaux. Les récepteurs bactériens GLIC [42] et eucaryotes GluCl [43, 44] et Glycine [45] sont visibles dans leurs états actifs/canal ouvert et de repos/canal fermé. D'autres états intermédiaires démontrent une grande diversité structurale de ces récepteurs : GLIC en état localement fermé [46], $\alpha 4\beta 2$ [34] et GABAA [34] dans des états vraisemblablement désensibilisés. Bien que les états conformationnels ne soient pas toujours consensuels au niveau atomique entre les différents récepteurs homologues, leur organisation structurale globale laisse transparaître deux mouvements distincts de grande amplitude lors de la transition des récepteurs [47] : Une contraction concertée de la partie extracellulaire (*unblooming*) ainsi qu'un mouvement de torsion entre les domaines extracellulaires et transmembranaires (*twisting*). Ces mouvements de la structure quaternaire ont aussi été observés dans des simulations numériques. Le *twist* de la structure quaternaire est apparu dans les premiers modes de fréquences d'analyses en modes normaux de modèle α7-nAChR [48, 49]. Plusieurs simulations de dynamique moléculaires de l'ordre de la microseconde de la désactivation des récepteurs (GLIC [50] et GluCl [51]) supportent un processus en deux temps : le *twisting* suivi du *blooming*. Le modèle d'activation admis implique ces deux mouvements quaternaires qui se traduisent par un réarrangement des feuillets β ECD suivant la liaison de l'agoniste dans la poche orthostérique, jusqu'à l'interface ECD-TMD (boucles $\beta 1 - \beta 2$ et Cys), et un déplacement des hélices transmembranaire permettant l'ouverture du canal une cinquantaine d'angströms plus loin [47].

4 Approches méthodologiques

4.1 Modélisation comparative

Les méthodes d'acquisitions expérimentales de la structure de protéines les plus courantes sont la cristallographie par rayon X, la résonance magnétique nucléaire et la cryo-microscopie électronique. Ces processus expérimentaux sont souvent longs, coûteux et l'obtention de la structure n'est pas garantie. Lorsque la structure cible n'est expérimentalement pas connue, mais qu'il existe au moins une protéine dont la structure est connue et partageant une homologie détectable avec la séquence cible, il est possible d'utiliser des algorithmes de modélisation comparative pour reconstruire un modèle de la structure cible. Ce type de modélisation met à profit le fait que, dans le contexte de l'évolution, la fonction d'une protéine est héritée de son repliement tridimensionnel. En effet, l'agencement tridimensionnel d'une protéine est généralement plus conservé à travers les mécanismes évolutifs que sa séquence en acides aminés. Par exemple, on observe que deux séquences de protéines dont la similarité est faible mais existante ($\approx$ 20 % d'identité) auront vraisemblablement une conformation globale semblable [52].

À partir, de la séquence d'une protéine cible, la génération d'un modèle par homologie se décompose classiquement en 4 étapes :

1. *Sélection des structures référentes (templates)*. Les structures expérimentales les plus pertinentes sont celles présentant la plus grande homologie de séquence avec la protéine à modéliser. Pour cela, une recherche peut être réalisée en utilisant un BLAST [53] de la séquence requête sur la base de données de structures expérimentales de la *RCSB Protein Data Bank* [54]. La ou les structures dont l'homologie est la plus forte sont récupérées.

2. *Alignement entre la séquence requête avec les séquences templates.* Cette étape détermine quelles régions des structures *templates* seront utilisées pour modéliser la position de chacun des résidus de la protéine cible. Elle est donc critique. Cet alignement doit privilégier l'appariement des résidus fonctionnels (conservés) entre la séquence cible et la séquence des *templates*. À cette fin, des programmes d'alignement multiple de séquence peuvent être utilisés (MUSCLE [55], T-COFFEE [56]).

3. *Construction du modèle.* L'implémentation de cette étape diffère en fonction du programme de modélisation utilisé. Ici la routine réalisée par le logiciel MODELLER [57] est décrite. Des distances entre atomes, des angles, des angles dièdres, etc. sont mesurées dans les *templates* et transposées sous forme de contraintes spatiales pour la protéine cible. Ces contraintes sont des fonctions de densité de probabilité conditionnelles dont la forme dépend de l'homologie entre les *templates* et la séquence requête ainsi que de l'analyse statistique d'une base de données de structures expérimentales. MODELLER optimise ensuite la position des atomes de la protéine cible de sorte à maximiser la satisfaction des contraintes. Pour cela, le programme réalise une succession de gradients conjugués et de dynamique moléculaire avec recuit simulé sur les coordonnées cartésiennes de la protéine. La Figure I.3/p.suiv. illustre cette étape de construction des modèles.

4. *Évaluation et sélection d'un modèle* MODELLER produit en sortie un ensemble de modèles de la protéine cible. Une sélection du modèle le plus pertinent peut ensuite être réalisée grâce à des fonctions de score capables de mesurer la vraisemblance des structures indépendamment du processus de création des modèles. On relèvera les scores zDOPE [58] et GA341 [59] inclus dans MODELLER ou encore les scores PROCHECK [59], QMEAN [59] ou proQM [60]. Ce dernier répond à la particularité des nAChRs de comporter une partie transmembranaire.

Si le modèle obtenu n'est pas satisfaisant, l'ensemble de ces étapes peut être répété en corrigeant l'alignement multiple de séquence ou en ajoutant des contraintes spatiales supplémentaires lors de la création des modèles. À la fin du processus de modélisation comparative, et en fonction de la qualité des modèles obtenus, la structure de la protéine cible peut être par exemple soumise à de la dynamique moléculaire ou utilisée pour l'amarrage *in silico* de molécules.

Problématique Dans notre étude du récepteur nicotinique une dimension supplémentaire s'ajoute au processus de modélisation comparative. Pour décrire le mécanisme d'activation du récepteur, nous avons besoin d'en modéliser des conformations fonctionnelles distinctes (état de repos et état actif) à partir d'une même séquence d'acides aminés. Cela implique un tri préalable des *templates* structuraux en rapport à la conformation ciblée. Des *templates* dont la conformation globale n'est pas clairement identifiée ne peuvent alors pas être utilisés, ce qui peu correspondre à une perte d'information structurale lors de la modélisation. Dans la Partie II/p.21 nous présentons une méthode alternative de modélisation comparative autorisant l'utilisation d'un maximum de structures homologues, peu importe leur état conformationnel, en dissociant les contraintes structurales locales (p.ex. liaisons covalentes,

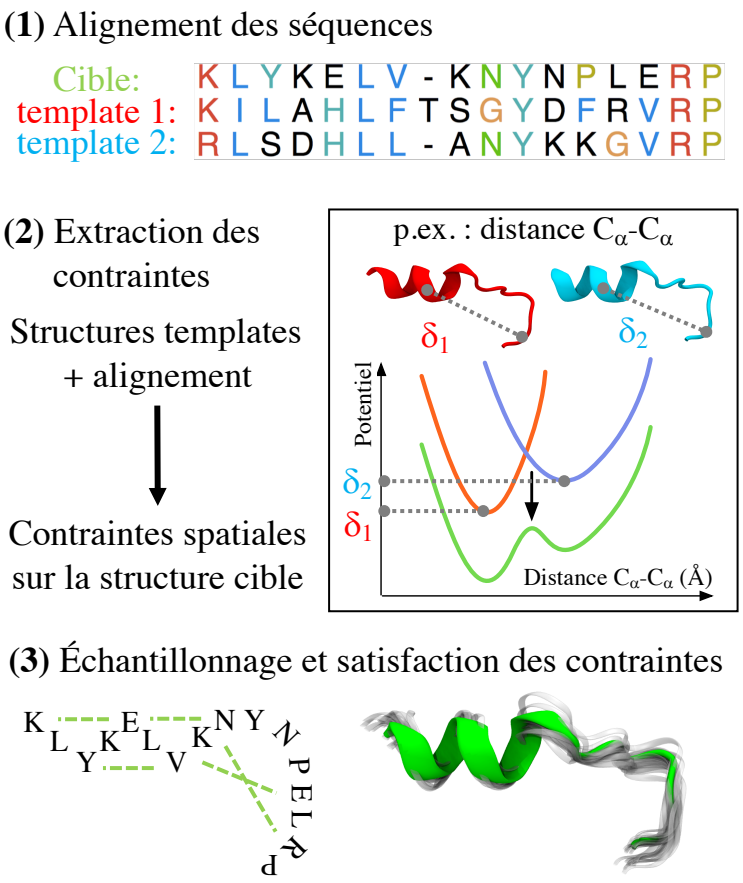


Figure I.3 Création de modèles par homologie. Le programme MODELLER prend en entrée la séquence de la protéine cible et son alignement multiple avec les séquences *templates* (**(1)**). Des mesures spatiales, par exemple la distance entre deux atomes de type C_α sont extraites des *templates* et traduites en contraintes structurales sur les atomes de la protéine à modéliser (**(2)**). La combinaison de ces contraintes est ensuite décrite par une fonction objectif qui est minimisée par des itérations de minimisations et de dynamiques moléculaires avec recuit simulé. Un ensemble de structures maximisant la satisfaction des contraintes structurale est alors produit. Des critères d'évaluation externes permettent finalement de choisir la conformation de la protéine cible la plus vraisemblable (**(3)**.

angles entre atomes, etc.) des contraintes globales responsables de l'état général de la protéine (p.ex. entre atomes distants). L'intégration de ces contraintes nous a permis de créer des modèles plus pertinents que ceux obtenus par une approche classique (avec tri des *templates*).

4.2 Dynamique moléculaire

La dynamique moléculaire (*Molecular Dynamics* ou MD) est une technique de modélisation de l'évolution temporelle des coordonnées atomiques de systèmes moléculaires complexes. En biologie, la dynamique moléculaire est principalement appliquée pour décrire le comportement dynamique des protéines (par exemple au travers de chan-

gement de conformations) ou des interactions entre biomolécules, tout en autorisant un contrôle fin de l'environnement modélisé (eau, membranes) et des propriétés thermodynamiques du système (température, pression). La MD est aussi utilisée en science des matériaux (études de la dynamique des polymères, des surfaces et interfaces entre solides, etc.). Par le calcul informatique, la MD permet d'explorer des propriétés moléculaires non accessibles à l'expérimentation (mouvements subtils des atomes, simulation de conditions extrêmes, substances dangereuses ou extrêmement coûteuses). Par ailleurs, l'étude statistique d'ensembles d'états correctement simulés offre la possibilité de calculer des propriétés macroscopiques fondamentales des systèmes (pression, potentiel chimique, capacité thermique, énergie libre, entropie, enthalpie).

Contrairement à la mécanique quantique relativiste qui offre un cadre de modélisation rigoureux mais qui est inaccessible pour des systèmes de grande taille, la MD fait plusieurs approximations qui rendent possible la simulation de systèmes de plusieurs centaines de milliers d'atomes. L'atome est considéré comme le niveau de représentation le plus petit des systèmes moléculaires. La MD ne considère pas les électrons explicitement, mais inclut leurs effets moyens dans des interactions spécifiques. Les champs de force polarisables, encore en développement, considèrent le couple noyau - nuage électronique. Ils n'ont pas été considérés ici en raison du manque de recul sur leur utilisation et du surcoût à l'utilisation. Enfin, la dynamique moléculaire utilise la mécanique classique pour calculer la trajectoire des atomes.

Les différentes composantes d'un programme de dynamique moléculaires sont les suivantes :

- **La topologie du système** Définition de l'ensemble des atomes du système ainsi que des liaisons entre les atomes.

- **Le champ de force.** Fonction des coordonnées atomiques décrivant l'énergie potentielle du système à partir de la topologie. Le champ de force est généralement constitué de l'expansion de potentiels interatomiques et doit être différentiable pour le calcul des forces exercées sur chaque atome par l'intégrateur. La paramétrisation du champ de force est choisie pour reproduire des résultats expérimentaux et des calculs de chimie quantique incluant des résultats *ab initio* [61, 62].

- **L'intégrateur.** À partir des positions et/ou vitesses passées de chacun des atomes du système, et du champ de force, un algorithme calcule la position future, à un temps $t + \Delta t$, de chacune des particules. Dans le cadre de la mécanique classique, la position des atomes est mise à jour par intégration de l'équation du mouvement de Newton.

- **Les thermostats et barostats** Se rajoutent à l'intégrateur de façon plus ou moins complexe et offrent un contrôle sur la température et la pression du système pendant la simulation.

Le champ de force utilisé lors de cette thèse est le champ de force tout-atome CHARMM36 [61–63], implémenté dans le programme de dynamique moléculaire CHARMM [64]. Comme d'autres champs de force populaires pour la modélisation de macromolécules biologiques (par exemple GROMOS [65] ou AMBER [66]), la fonction CHARMM36 est une somme de potentiels décrivant les interactions entre atomes liés et non-liés de tout état $\mathbf{r}$ du système moléculaire :

$$\begin{aligned} V(\mathbf{r}) &= V_{\text{liées}}(\mathbf{r}) + V_{\text{non-liées}}(\mathbf{r}) \\ V_{\text{liées}}(\mathbf{r}) &= V_{\text{liaison}}(\mathbf{r}) + V_{\text{ang.}}(\mathbf{r}) + V_{\text{tors.}}(\mathbf{r}) + V_{\text{imp.}}(\mathbf{r}) + V_{\text{corr.}}(\mathbf{r}) \\ V_{\text{non-liées}}(\mathbf{r}) &= V_{\text{LJ}}(\mathbf{r}) + V_{\text{élec.}}(\mathbf{r}) \end{aligned}$$

Le terme des interactions liées correspond aux liaisons covalentes définies par la topologie du système, et comprend des sous-potentiels pour décrire les distances entre atomes (V_{liaison}), les angles ($V_{\text{ang.}}$), les angles de torsion ($V_{\text{tors.}}$), les angles impropres ($V_{\text{imp.}}$), p.ex. pour forcer la planéité des cycles aromatiques), ainsi que des termes de corrections annexes ($V_{\text{corr.}}$, p.ex. Urey-Bradley [61] - couplage liaison-angle ou CMAP [62] - couplage des angles ϕ et ψ basé sur la carte de Ramachandran).

Le terme des interactions non-liées est fonction des distances entre atomes non-liés. Un potentiel de Lennard-Jones (V_{LJ}) prend compte des forces de Van der Waals et du principe d'exclusion de Pauli (répulsion interatomique). Finalement, un terme d'électrostatique ($V_{\text{élec.}}$) ajoute un potentiel de Coulomb à la formule.

Le système à simuler en solvant explicite est généralement placé dans une boîte aux limites périodiques pour simuler un système de dimension infinie. Le calcul du terme électrostatique est alors partitionné en deux termes avec une zone de transition entre les deux : une sommation directe pour les interactions de courte portée, et une sommation d'Ewald [67, 68], dans l'espace réciproque, pour les interactions au-delà. Le solvant, représenté de façon explicite est placé dans la boîte de simulation remplie de molécules le représentant, le plus souvent des molécules d'eau, avec éventuellement des ions et des lipides. Cet ajout augmente considérablement la taille du système ce qui ralentit le calcul. Des modèles implicites peuvent être incorporés au champ de force pour reproduire les propriétés diélectriques du solvant en l'absence de molécules explicites (p.ex. *ACE* [69], *Generalized Born* [70]). Il n'est alors pas nécessaire de travailler en conditions périodiques. Le pas d'intégration (Δt) doit être court (<1 fs) pour tenir compte des dynamiques de vibration de tous les atomes. L'utilisation de l'algorithme SHAKE, permet d'augmenter ce temps à 2 fs en fixant la longueur des liaisons impliquant des atomes d'hydrogènes.

4.3 Calcul de chemins de transition

Les processus de repliement et de changement de conformation des protéines ne sont pas seulement aléatoires mais sont conduits par des mécanismes énergétiques complexes. Ces mécanismes peuvent être décrits par la théorie du paysage énergétique [71]. Supposons qu'il existe une fonction G capable d'attribuer un réel, une énergie, à tout état du système spécifié par la position tridimensionnelle de chacun de ses atomes. En cela, G décrit une hypersurface dans un espace dont la dimension, très grande, est égale au nombre d'atomes du système multiplié par 3. Cette hypersurface est aussi appelée surface, ou paysage énergétique. Sur cette surface, le repliement d'une protéine se traduit par le déplacement d'un point A de l'espace des conformations (p.ex. : chaîne polypeptidique dépliée) vers un point B (protéine repliée). Si le paysage énergétique était tout à fait plat et que la protéine s'y déplaçait par une marche aléatoire le processus de repliement prendrait un temps quasiment infini au vu du nombre de degrés de liberté considéré (paradoxe d'Anfinsen [72]). En fait, la surface énergétique pour un système comprenant plusieurs centaines d'atomes est hautement rugueuse, comprenant des pics (hautes énergies : collisions entre atomes, interactions non favorables, etc.) et des bassins (énergie faible, état du système favorable). Dans ce contexte la transition de l'état A à l'état B, peut se faire spontanément si l'état B est plus favorable (énergie plus basse) que l'état A. Telle une bille posée sur le rebord d'un bol et qui roulerait en son centre, la protéine converge vers son état natif de plus basse énergie. La physique statistique nous donne les outils pour étudier ces phénomènes grâce à la notion d'énergie libre. L'énergie libre est un indicateur de la stabilité du système et gouverne la direction de ses changements spontanés en prenant en compte des sous populations d'états assimilées à un état donné (par exemple une conformation de la protéine et un ensemble pertinent de configurations du solvant associé à cette conformation). Les variations de l'énergie libre (ΔG) sont dictées par les variations de l'enthalpie (ΔH) et de l'entropie (ΔS) par la formule :

$$\Delta G = \Delta H - T\Delta S$$

où T est la température du système.

Dans un premier temps nous pouvons considérer le système à une température nulle ($T = 0$). Dans ce cas l'énergie libre du système n'est plus dictée que par le terme enthalpique. À $0\ K$ et en négligeant les énergies vibrationnelles de point 0, l'enthalpie correspond à l'énergie potentielle du système, elle-même décrite par les champs de force de dynamique moléculaire. Un premier cadre théorique pour étudier les transitions entre états alternatifs de protéines est donc d'analyser la surface d'énergie potentielle dans un contexte adiabatique. Profitant de certaines propriétés des champs de force (continuité, dérivabilité), des méthodes d'optimisation numériques

ont été mises au point pour étudier ce contexte enthalpique (ΔH). Les méthodes *Conjugate Peak Refinment* et *Nudged Elastic Band* en sont deux exemples.

Le paysage d'énergie potentielle peut être décrit comme un relief terrestre. Les points d'énergies défavorables sont des sommets de montagnes, les bassins d'énergie favorable sont des lacs, et les points de selles des passages entre les montagnes. Un point de selle, est un point de l'hyperespace minimum local sur toutes les dimensions sauf une pour laquelle c'est un maximum local. Un bassin énergétique est dans un minimum local sur toutes ses dimensions. Un chemin de transition sur la surface d'énergie potentielle est caractérisé par un ensemble de points où l'on est dans un minimum pour toutes les directions transverses au chemin et qui peu contenir une série de points de selles délimitant des minima énergétiques locaux et joignant les états stables ou métastables d'une protéine. Le *Conjugate Peak Refinment* est une méthode de calcul de chemin d'énergie minimal par recherche de points de selles. L'algorithme prend en entrée deux conformations distinctes d'une protéine. Dans un premier temps, il mesure l'énergie (grâce au champ de force) des structures placées sur une interpolation linéaire entre les deux points initiaux (P et R). La structure d'énergie maximale est un point qu'il faut contourner (tel un randonneur qui souhaiterait aller d'une ville à une autre en évitant de gravir des sommets). Pour cela, l'algorithme minimise la structure dans une direction conjuguée au segment interpolé pour s'approcher de la vallée énergétique. La structure ainsi obtenue est un nouveau point s_1 du chemin de transition. La procédure est ensuite répétée sur les segments interpolés $P - s_1$ et $s_1 - R$, avec recherche de maxima, et minimisations conjuguées. L'ajout des points est entrelacé d'optimisations et suppressions des points existants. À la fin des itérations tous les points sont à des minima dans les directions transverses au chemin. Par ailleurs, l'algorithme supprime les barrières énergétiques entre les points intermédiaires (connexité du chemin). Les *Nudged Elastic Band* sont un autre exemple de méthode de calcul de chemins de transition. L'algorithme est basé sur la relaxation successive d'une chaîne d'états de la protéine. Le nombre de points intermédiaires est fixé à l'avance et un premier chemin est créé (p.ex par interpolation initiale des 2 points extrémaux). Des potentiels élastiques sont placés entre chacun des points adjacents. Chaque point ressent deux forces. La première s'exerce en direction des vallées du potentiel (par minimisation), et la seconde vient du potentiel élastique. En pratique, chaque point intermédiaire est mis à jour de sorte à suivre la contribution des forces résultantes perpendiculaires à la force élastique (*nudging*). Ainsi, la chaîne d'état converge petit à petit pour épouser uniformément la vallée d'énergie minimale. Contrairement, au *Conjugate Peak Rafinment* cette méthode ne garantit pas la connexité du chemin.

À température physiologique ($\approx$ 300 K), les macromolécules en solution sont en constante agitation due à la thermalisation du système. Leur évolution doit être considérée dans le contexte d'énergie libre (ΔG). La surface d'énergie libre est

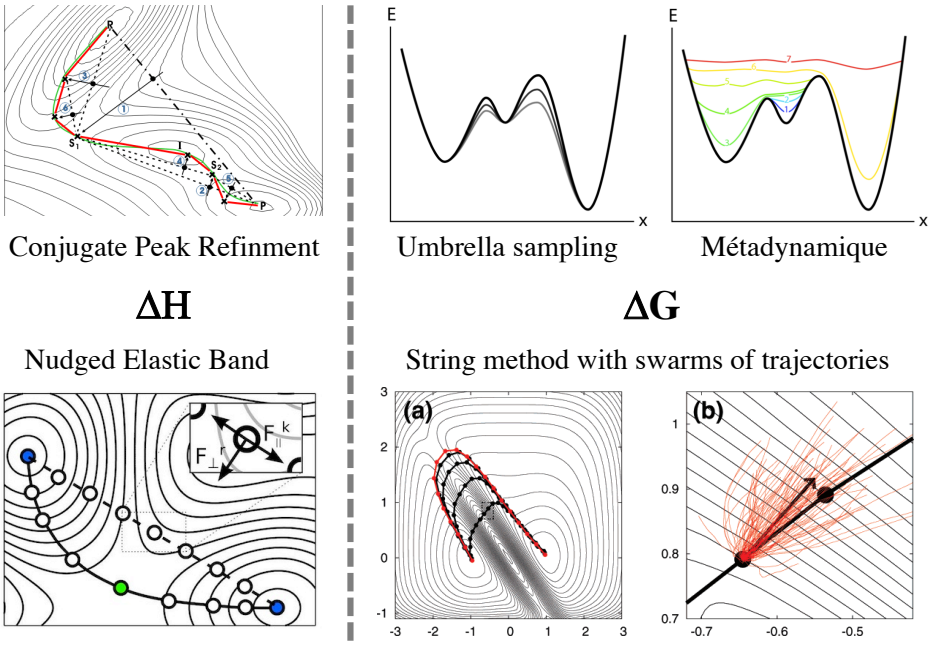


Figure I.4 Exemples de méthodes pour le calcul de chemins de transitions. À gauche, méthodes adiabatiques pour le calcul de chemin d'énergie minimum : *Conjugate Peak Refinment* et *Nudged Elastic Band*. À droite, méthodes dynamiques pour le calcul de chemin d'énergie libre : *Umbrella Sampling*, *Métadynamique*, et *String method with swarms of trajectories*.

une version de la surface d'énergie potentielle repondérée par la distribution de Boltzmann en tenant compte des micros états associés. Au terme enthalpique (H) se soustrait un terme entropique (S) qui lui-même dépend de la connaissance de ses micros états associés. À cause de cela, la fonction G n'est jamais connue mais approximée en échantillonnant les conformations de l'espace des phases. La recherche de chemins d'énergie libre minimale est ainsi plus laborieuse, car elle nécessite l'échantillonnage exhaustif des conformations de l'espace des phases. Les méthodes de *Métadynamique* [73] et de l'*Umbrella Sampling* [74] facilitent l'exploration en modifiant le paysage énergétique et/ou en exploitant des dimensions particulières définies par des Variables Collectives. Les Variables Collectives, ou Coordonnées de Réaction sont fonctions des coordonnées du système et choisies, généralement *a priori*, pour décrire à elle seule le mécanisme fonctionnel. En pratique, la sélection de Variables Collectives pertinentes est difficile, surtout quand le mécanisme étudié est encore peu connu. La *Métadynamique* force la dynamique à explorer exhaustivement les conformations du paysage d'énergie libre en ajoutant un terme énergétique défavorable aux endroits déjà visités. L'*Umbrella Sampling*, ajoute un biais artificiel au potentiel le long des Variables Collectives pour diminuer et effacer les barrières d'énergie libre. Dans les deux cas, la surface d'énergie libre peut être reconstruite *a posteriori*. Lorsque, les coordonnées réactionnelles ne sont pas précisément connues, la méthode des *String of Swarms* (SoS) est un autre choix possible. À partir d'un chemin de transition initial, SoS met à jour chacune des structures de la transition de

sorte à ce qu'elles suivent le gradient de diffusion d'un essaim de courtes trajectoires moléculaires. Après une reparamétrisation des structures de sorte à ce que chacun des intermédiaires reste au même stade de la réaction, le chemin produit est passé en entrée d'une nouvelle itération, jusqu'à convergence des chemins dans une vallée d'énergie libre. Plutôt que d'explorer l'hypersurface d'énergie libre sur tout son volume, la méthode des *String of Swarms* n'échantillonne que les conformations du *tube* réactionnel local au chemin de transition et suffisant pour "sentir" la direction des vallées énergétique. Grâce à cette astuce, SoS peut prendre en compte un nombre beaucoup plus important de variables collectives (potentiellement tous les atomes du système).

Problématique Le récepteur nicotinique est une protéine transmembranaire qui pose un défi majeur aux méthodes de calcul de chemins de transition. La description complète de son mécanisme d'activation doit tenir compte de ses interactions avec le solvant : eau, lipides, ions. De plus, le nombre de degrés de libertés d'une simulation comprenant le récepteur entouré de son solvant est gigantesque (>600 000). Nous observerons qu'une méthode comme les *String of Swarms* a tendance à créer des chemins de transition "bruités" par les mouvements Browniens issus de la thermalisation du système. Pour corriger ces défauts, nous avons voulu coupler les *String of Swarms* avec une méthode de calcul de chemin d'énergie minimal développée au laboratoire (*Path Optimization and Exploration* - POE). POE régularise des chemins construits sur la surface d'énergie potentielle grâce à la recherche systématique de raccourcis topologiques pertinents entre les conformations de la transition. Dans la Partie III/p.47 un chemin de transition pour le récepteur nicotinique est construit, et alternativement raffiné dans un contexte enthalpique (ΔH à 0 K) pour obtenir une description essentielle de la réaction et dans un contexte d'énergie libre (ΔG) pour une description physiquement plus juste du mécanisme (thermalisation, interactions avec le solvant).

II

Modélisation comparative d'états conformationnels multiples de protéine

Au début de ce projet, aucune structure expérimentale du récepteur nicotinique de sous-unités $\alpha 7$ n'était disponible. Les stratégies d'exploration des mécanismes fonctionnels de protéines développées lors de cette thèse sont basées sur les structures, et nécessitent, en première instance, de connaître les différents états conformationnels de la cible. De nombreux récepteurs homologues ont été cristallisés dans des états fonctionnels multiples. Leur identité de séquence avec l'$\alpha 7$ est cependant souvent faible. Dans cette Partie, nous proposons une méthode de modélisation comparative qui, à partir d'une même séquence de protéine, permet d'en générer des états conformationnels multiples. La structure locale des conformations (p.ex. l'orientation et la conformation des chaînes latérales) est consensuellement extraite d'un maximum de *templates* structuraux, alors que la structure globale est dictée par un sous-ensemble de *templates* dont la conformation correspond à celle modélisée. Cette approche a été utilisée pour modéliser les états actifs et de repos du récepteur $\alpha 7$, puis comparée à l'approche classique de modélisation communément employée dans la littérature pour modéliser ces récepteurs.

1 Introduction et objectifs

La détermination de l'agencement tridimensionnel des atomes du récepteur nicotinique est un défi pour la communauté scientifique depuis leur découverte, il y a près de 50 ans [75]. Des données structurelles précises faciliteraient la compréhension des mécanismes complexes qui régissent ce canal ionique, et accéléreraient la conception de molécules thérapeutiques ciblant spécifiquement le récepteur. Cependant, les méthodes expérimentales de détermination des structures de protéine font face à de nombreux obstacles spécifiques aux récepteurs nicotiniques. Ils sont difficiles à purifier, ce qui rend complexe l'obtention de grandes quantités de nAChR à une concentration suffisante pour qu'ils soient utilisés en cristallographie aux rayons X [76]. La séquence polypeptidique d'une sous-unité $\alpha 7$ neuronale humaine contient 552 acides aminés (domaine intracellulaire inclus) soit 2 760 résidus et un poids moléculaire total de plus de 250 kilodalton pour un pentamère complet. De plus, c'est une protéine transmembranaire dont les propriétés fonctionnelles diffèrent en fonction de la composition de la bicouche lipidique [77]. De par sa fonction intrinsèque, cette protéine est supposée pouvoir se mouvoir entre des états multiples et subtils (canal ouvert, fermé, localement fermé, état désensibilisé). Ainsi l'acquisition expérimentale de structures est rendue particulièrement compliquée, ce qui explique qu'aucune structure du récepteur $\alpha 7$, comprenant au moins sa partie extracellulaire et transmembranaire, n'ait encore été publiée.

Nous nous basons dans ce chapitre sur des structures cristallographiques obtenues à partir de canaux ioniques dont la séquence en acides aminés est homologue au pentamère de sous-unités $\alpha 7$. Cela nous donne la possibilité de créer des modèles par *homologie*, ou par *comparaison*. Le postulat fondateur de la modélisation comparative est que deux protéines dont la séquence en acides aminés est proche, sont supposées partager une structure tridimensionnelle similaire. Cette assertion devient ambiguë lorsque de larges changements conformationnels interviennent dans les mécanismes de la cible. Deux chaînes polypeptidiques strictement identiques peuvent alors avoir un repliement en structures distinctes. L'approche de modélisation classique implique alors de ne sélectionner que des *templates* structuraux dont la conformation correspond à l'état fonctionnel à modéliser (voir Figure II.2/p.suiv.). De nombreux exemples de modélisation de récepteurs canaux utilisant cette approche sont disponibles dans la littérature [34, 78–82].

Cependant, le fait d'écarter certaines structures homologues lors de la modélisation peut être vu comme une perte d'information. Par exemple, en octobre 2016, la structure cristallographique d'un récepteur nicotinique, de sous-types $\alpha 4\beta 2$ a été

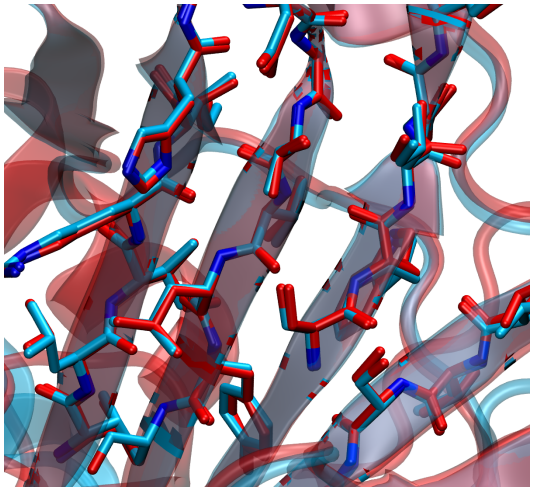


Figure II.1 Similarités structurales locales entre deux états du récepteur GluCl. Comparaison des structures cristallographiques du récepteur GluCl en état actif (bleu) et à l'état de repos (rouge). Vu prise dans la partie extracellulaire du récepteur (feuillets β du vestibule).

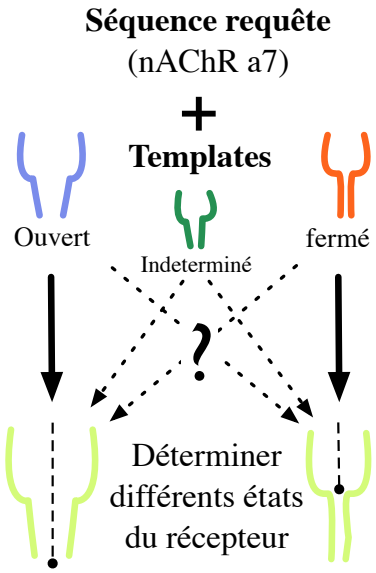


Figure II.2 Problématique de la modélisation d'états conformationnels. L'approche classique de modélisation par homologie n'utilise que les *templates* dont la conformation correspond à celle à modéliser (flèches pleines). L'approche hybride (flèches pointillés) cherche à utiliser l'ensemble des *templates* disponibles en triant lors de la modélisation les contraintes dictant la structure globale du récepteur.

publiée [34]. La comparaison de sa structure avec des conformations homologues semble démontrer que le récepteur a été cristallisé dans un état désensibilisé [34]. Alors que sa séquence est très proche du récepteur $\alpha 7$, et donc d'autant plus pertinente pour de la modélisation comparative, cette structure ne pourrait être utilisée avec l'approche classique pour créer des modèles du récepteur dans ses états actifs et de repos. Or, il semble probable que certaines régions de la protéine ne soient pas altérées par les transconformations du récepteur, comme illustré par la superposition de deux structures en états actif et de repos du récepteur GluCl visible dans la Figure II.1. Ces régions consensus pourraient profiter d'un plus grand nombre de *templates* structuraux, peu importe leur état conformationnel global.

Nous proposons dans ce chapitre une approche de modélisation entrelaçant des contraintes structurales hybrides (issues de l'ensemble des *templates* choisis - en état ouvert, de repos ou indéterminé) et les contraintes structurales des *templates* en état actif ou de repos. Les contraintes locales (distances entre atomes liés, angles entre triplets d'atomes, angles dièdres...) sont systématiquement hybrides. Les contraintes agissant sur la conformation globale du récepteur (typiquement les distances entre couples de carbones α) peuvent être hybrides ou spécifiques en fonction de l'importance du changement conformationnel dans la région considérée. De ce fait,

les régions de la protéine similaires entre les états de repos et actif profitent des meilleurs *templates* structuraux indépendamment de leur état conformationnel. À l'inverse, les régions impactées par la transconformation sont modélisées par l'ajout de contraintes spécifiques qui permettent d'obtenir l'état fonctionnel voulu.

Cette méthode a été appliquée à la modélisation du récepteur nicotinique dans son état actif (canal ouvert au passage des ions) et de repos (canal fermé). Une comparaison a été réalisée avec la méthode de modélisation classique usuellement employée. Un ensemble divers de fonctions de score de la qualité des structures de protéine a été utilisé pour mesurer la pertinence des structures prédites. D'autre part, l'environnement transmembranaire du récepteur a pu être pris en compte lors de la modélisation (protocole Rosetta Membrane) et de l'évaluation des structures (score ProQM).

2 Matériels et Méthodes

2.1 Définitions

Template Le terme *template* est abusivement conservé en français pour désigner la ou les structures homologues à la structure cible qui servent de base de comparaison lors de la modélisation par homologie. *Template* a été préféré à *modèle*, pour ne pas interférer avec la définition suivante communément employée lors de cette thèse.

Modèle Un modèle structural fait ici référence à la conformation d'une macromolécule dont les coordonnées ont été obtenues à l'aide d'un processus de modélisation informatique (simulation, optimisation...).

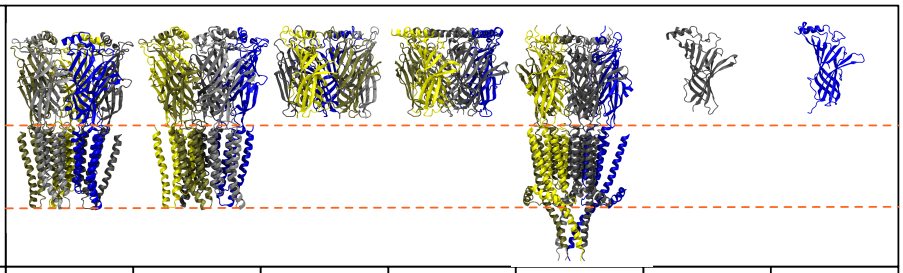
Contrainte Une contrainte, au sens *restraint* en anglais, décrit une fonction de densité de probabilité sur le placement spatial de groupes d'atomes de la protéine cible (p.ex. distance entre atomes, angles, angles dièdres, etc...). Les contraintes sont calculées à partir des *templates* structuraux et de leur alignement multiple avec la séquence.

Actif/Ouvert - Repos/Fermé L'état conformationnel du récepteur nicotinique est parfois appelé *fermé* dans sa forme apo de repos et *ouvert* dans un état holo *actif*, en référence à l'état du canal ionique.

2.2 Modélisation comparative

2.2.1 Séquence de la protéine cible

La séquence requête correspond au récepteur nicotinique à acétylcholine neuronal Humain de sous-unité $\alpha 7$ (gène CHRNA7, entrée UniProtKB : P365344). Plusieurs modifications ont été apportées à la séquence pour exclure de la modélisation les portions du récepteur non couvertes par les structures homologues disponibles : d'une part, les 25 premiers résidus en N-terminale et les 11 derniers en extrémité C-terminale et, d'autre part, 136 résidus correspondant à la partie intracellulaire du récepteur. Cette dernière troncature, située entre les hélices transmembranaires M3 et M4 est basée sur les travaux de Kouvatsos et. al. 2014 [41]. Elle préserve le triplet de résidus "MKR" indispensable à l'expression du récepteur [83] (pour



ID. PDB	3RIF	4TNV	3SQ6	3SH1	4PIR	2QC1	4D01
Organisme	souris	souris	lymnae (chimère α7)	aplysia (chimère α7)	souris	souris	humain
Protéine	GluCl	GluCl	AChBP	AChBP	5-HT3	nAChR (α1)	nAChR (α9)
Résolution	3,35	3,60	2,80	2,90	3,50	1,94	1,79
%id. α7	20,2	20,2	62,3	35,7	32,6	37,3	39,0
État	actif	repos	actif	repos	indéter.	indéter.	indéter.

Figure II.3 Structures *templates* sélectionnées pour la modélisation. En haut, les modèles cristallographiques en représentation *cartoon*. Les deux lignes en pointillé séparent le domaine extracellulaire (ECD), du domaine transmembranaire (TMD) et du domaine intracellulaire (ICD).

l'$\alpha 7$ nAChR du rat), engendre des récepteurs fonctionnels et dont les profils de liaison aux agonistes classiques sont similaires aux récepteurs sauvages [41] (pour l'$\alpha 4\beta 2$). La séquence ainsi tronquée, est rapportée en Annexe VII.1/p.193. Pour modéliser le récepteur nicotinique, cette séquence est concaténée 5 fois pour générer un pentamère de stœchiométrie $(\alpha 7)_5$.

2.2.2 Sélection et annotation des structures homologues

Une recherche BLAST [53] de la séquence d'$\alpha 7$ contre la Protein Data Bank (NCBI [84] blastp, paramètres par défaut et e-value <10e-5) détecte 72 séquences homologues. Parmi ces structures, on trouve des récepteurs procaryotes (GLIC, ELIC), les récepteurs eucaryotes 5-HT_3, Glycine, GluCl, et des *AcetylCholine Binding Protein* (AChBP), en conformations diverses. Le choix des structures *templates* a été fait en collaboration avec l'unité des Récepteurs Canaux de l'Institut Pasteur (Pierre-Jean Corringer, Marc Gielen). Ces structures, appelées par leur identifiant à 4 lettre de la RCSB Protein Data Bank [54] (PDB), sont présentées en Figure II.3. Les états spécifiques actif/ouvert et repos/fermé sont déterminés par les structures cristallographiques du récepteur GluCl (ouvert 3RIF cristallisé et apo/fermé 4TNV [44]). Ce choix par rapport à des structures ouvert/fermé GLIC sera explicité en Discussion 4.2/p.40. Nous avons aussi sélectionné deux chimères AChBP partiellement mutées en sous-unités $\alpha 7$. La première structure, 3SQ6 [35] lie l'agoniste épibatidine et est classée en état "ouvert" (boucle C refermée). La seconde, 3SH1 [85] lie l'antagoniste Methyllycaconitine et caractérise l'état "fermé" (boucle C ouverte). La modélisation hybride nous

permet de choisir des *templates* structuraux dont l'état est indéterminé. Nous avons utilisé le récepteur $5HT_3$ 4PIR [86] cristallisé avec un inhibiteur (VHH15) mais dont la conformation n'est pas clairement dans un état conducteur d'après les données expérimentales, et les structures $\alpha 1$ (2QC1 [87]) et $\alpha 9$ (4D01 [88]) dont l'homologie à la séquence requête et la résolution sont avantageuses mais dont la conformation est indifférenciée (une seule sous-unité par unité asymétrique cristallographique).

2.2.3 Alignement multiple des séquences

Un premier alignement multiple de séquence des 7 *templates* contre la séquence $\alpha 7$ a été réalisé en utilisant le serveur web TCOFFEE [56] (paramètres par défaut). L'alignement a ensuite été raffiné manuellement, en particulier aux extrémités des structures secondaires - les boucles - pour lesquelles la conservation de séquence est souvent plus faible et donc plus difficile à apparier avec la séquence cible. Ce raffinement a été réalisé d'une part par la superposition et la visualisation directe des structures *templates* et d'autre part par le contrôle systématique des scores de qualité de modèles construits à partir de nouveaux alignements. L'étape de contrôle des scores se déroule comme suit (cf. Figure II.4/p.suiv.) :

1. La modification ponctuelle de l'alignement. Typiquement, le décalage d'une colonne contenant un ou plusieurs résidus des séquences *templates* par rapport à la séquence requête.
2. Le nouvel alignement multiple et les 7 *templates* structuraux sont utilisés pour générer 1 000 modèles du récepteur $\alpha 7$ avec le programme MODELLER [57] (configuration décrite dans la sous-section suivante).
3. Le score de qualité zDOPE est calculé pour chacune de ces structures et moyenné.
4. Le nouvel alignement est accepté si le zDOPE score moyen obtenu est supérieur à celui de l'alignement précédent.

Cette méthode permet d'améliorer le score de qualité des structures produites (Tableau II.1/p.suiv.), en complément de l'analyse visuelle des structures *templates*. L'alignement utilisé pour la suite de ces travaux est rapporté en Annexe VII.2/p.194.

2.2.4 Configuration de MODELLER

Nous avons utilisé le logiciel MODELLER [57] (version 9.14). Usuellement, cet outil prend en entrée une séquence requête d'acides aminés à modéliser, des structures de protéines homologues (en coordonnées cartésiennes, fichier PDB) et l'alignement multiple de la séquence de chacun des *templates* contre la séquence requête.

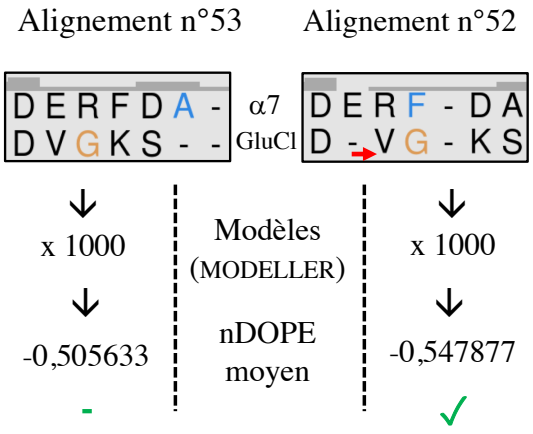


Figure II.4 Comparaison de deux alignements multiples de séquence. Pour chaque nouvel alignement multiple de la séquence requête contre la séquence des *templates* structuraux, 1 000 modèles du récepteur nicotinique sont construits. Un score de qualité est calculé pour chacune de ces structures (zDOPE). La moyenne de ces scores est utilisée pour favoriser un alignement sur un autre.

N°alignement	zDOPE moyen
tcoffee	-0,477615
45	-0,493862
79	-0,494921
41	-0,500905
...	...
91	-0,563301
67	-0,564282
72	-0,564288
92	-0,597697
92_1	-0,598041
92_2	-0,598206

Tableau II.1 Score (zDOPE moyen) des variations de l'alignement multiple. Chacun des alignements est issu d'une modification ponctuelle d'un alignement précédent. En bas, l'alignement n°92, est le meilleur candidat sélectionné. Les scores des alignements 92_1 et 92_2 correspondent à la génération de 1 000 nouvelles structures et montrent les variations des scores moyens produits.

Dans un premier temps MODELLER dérive de l'alignement multiple de séquences une liste de contraintes structurales sur les atomes de la protéine requête et les sauvegarde dans un fichier d'extension .rsr. Des contraintes additionnelles sont ajoutées dans le cadre de la modélisation du récepteur $\alpha 7$.

- Dix contraintes de symétrie entre les carbones alpha (C_α) de tous les couples des 5 chaînes du récepteur (fonction "restraints.symmetry", paramètre "weight" à 1).
- Une contrainte dans chacune des chaînes, imposant la configuration *cis* pour la proline n°133.

Dans un second temps, MODELLER optimise le placement dans l'espace des atomes de la structure requête de façon à limiter au maximum la violation des contraintes structurales. Ici les paramètres utilisés sont "library_schedule = autosched.slow" ainsi que "md_level = refine.slow". Finalement, MODELLER produit un ensemble de structures, triées suivant différents scores internes. Lors de chaque exécution complète de MODELLER, nous avons utilisé les paramètres "starting_model = 1" et "ending_model = 1000" pour créer indépendamment 1 000 modèles différents. Grâce à la fonction "sge_qsub_job()" de MODELLER, la production de modèles est distribuée sur 260 nœuds du *cluster* de calcul pour environ 45 minutes de temps réel d'exécution.

2.3 Approche de modélisation classique

Lors de la modélisation par homologie classique, seule les *templates* dont la conformation a été annotée sont utilisés, soit les structures GluCl (3RIF) et AChBP (3SQ6) pour l'état actif/ouvert et GluCl (pdb :4TNV) et AChBP (3SH1) pour l'état repos/-fermé. L'alignement de séquences entre les *templates* et la séquence $\alpha 7$ est présenté dans la sous-section 2.2.3/p.27. Un millier de modèles sont ensuite créés avec le logiciel MODELLER pour chacun des deux états conformationnels.

2.4 Approche de modélisation hybride

2.4.1 Génération des contraintes hybrides

MODELLER est utilisé pour calculer le fichier d'extension .rsr contenant l'ensemble des contraintes structurales dérivées des 7 *templates*. Les fichiers d'entrées correspondent à la séquence requête, l'alignement multiple des *templates* contre la requête et les structures PDB des 7 *templates*. En sortie, chacune des lignes du fichier .rsr renseigne la définition d'une contrainte structurale agissant sur un groupe d'atomes particuliers du récepteur $\alpha 7$.

2.4.2 Sélection des contraintes spécifiques

Les structures *templates* du récepteur GluCl servent à récupérer les contraintes spécifiques discriminantes de l'état du canal. Pour l'état ouvert (ou fermé), MODELLER est utilisé pour calculer l'ensemble des contraintes entre C_α issues du *template* 3RIF (respectivement 4TNV). La fonction MODELLER utilisée est "restraints.make_distance" avec les paramètres "restraint_group=physical.ca_distance", "distance_rsr_model=5", "maximal_distance=14" et "restraint_stdev=(0,1.0)". Le fichier de contraintes (extension .rsr) ainsi produit contient pour chaque ligne la description de la contrainte gaussienne exercée entre les atomes *i* et *j*, avec pour moyenne une distance μ et un écart type σ. Comme par exemple :

```
                                    i     j      μ          σ
R   3   1   1   9   2   2   1   11   461   21.9610   1.3689
```

Ici l'écart type σ est fonction de la similarité entre la séquence requête $\alpha 7$ et la séquence *template* GluCl. Ces contraintes C_α-C_α sont ainsi calculées pour les deux *templates* 3RIF et 4TNV.

Pour chacune des paires d'atomes (*i*,*j*), s'il existe une contrainte $C_o(i,j) = (\mu_o,\sigma_o)$ dans le fichier de contraintes du *template* ouvert (3RIF) et une contrainte $C_f(i,j) = (\mu_f,\sigma_f)$ dans le fichier de contraintes du *template* fermé (4TNV), alors il est possible de les comparer directement avec la distance dite de Bhattacharyya [89] :

$$bha(C_o(i,j),C_f(i,j)) = \frac{1}{4} ln \left(\frac{1}{4} \left(\frac{{\sigma_o}^2}{{\sigma_f}^2} + \frac{{\sigma_f}^2}{{\sigma_o}^2} + 2 \right) \right) + \frac{1}{4} \left(\frac{(\mu_o - \mu_f)^2}{{\sigma_o}^2 + {\sigma_f}^2} \right)$$

Pour chacune des paires d'atomes comparées, si cette distance $bha(C_o(i,j),C_f(i,j))$ est supérieure à un certain seuil $S_{\text{spé}}$, la contrainte est considérée comme suffisamment différente entre les états ouverts et fermés. La paire d'atomes (*i*,*j*) est alors gardée en mémoire. Le choix de ce seuil $S_{\text{spé}}$ permet de faire varier le nombre de contraintes spécifiques sélectionnées, et à terme de créer des modèles plus ou moins ouverts ou fermés. Ce paramètre a été itérativement optimisé (cf. Résultats 3.4/p.37). Après optimisation, une valeur de 0,015 a été utilisée pour créer les modèles ouverts et une valeur de 0,005 pour les modèles fermés.

2.4.3 Renforcement des contraintes spécifiques

Les contraintes précédemment sélectionnées ne peuvent pas être directement intégrées parmi l'ensemble de contraintes *hybrides*. En effet, en raison de la faible homologie de séquence entre le récepteur GluCl et le récepteur nicotinique $\alpha 7$, les contraintes de distances entre C_α ont le plus souvent un écart type σ élevé. Ces contraintes "lâches" ne permettent pas de tirer suffisamment les modèles produits dans des conformations ouvertes ou fermées. Pour remédier à cela, les contraintes de distance C_α-C_α sont calculées sur les modèles ouverts et fermés $\alpha 7$ générés avec l'approche classique. En prenant comme séquence requête un $\alpha 7$ et comme *template* une structure $\alpha 7$, nous avons la garantie que les contraintes structurales ont un écart type faible. Les paires d'atomes sélectionnées dans la sous-section précédente permettent de récupérer les contraintes spécifiques ouvertes et fermés.

2.4.4 Intégration des contraintes spécifiques parmi les contraintes hybrides

Les contraintes *hybrides* sont passées en revue une par une. Lorsque la contrainte hybride correspond à une contrainte de distance C_α-C_α et qu'elle agit sur des atomes (*i*,*j*) auxquelles correspond une contrainte spécifique, alors la contrainte spécifique ouverte ou fermée vient remplacer la contrainte hybride. Les deux nouveaux ensembles de contraintes *hybride-ouvert* et *hybride-fermé* sont ensuite utilisés pour créer des modèles du récepteur $\alpha 7$ dont l'état du canal est ouvert ou fermé.

2.5 Relaxation des structures avec Rosetta

Un modèle créé par MODELLER peut-être individuellement optimisé avec le protocole “Membrane Protein Relax” [90–92] du logiciel Rosetta. Le placement de la partie transmembranaire est similaire à celui décrit pour la fonction de score ProQM. La fonction de score utilisée est “membrane_highres_Menv_smooth.wts”. Les paramètres autres que ceux par défaut sont : “-relax :default_repeats 8”, “-relax :ramp_constraints false”. 1 000 nouveaux modèles, relaxés dans un environnement membranaire implicite, peuvent ainsi être créés.

2.6 Évaluation de la qualité des prédictions

Trois fonctions de score ont été utilisées pour mesurer la vraisemblance des structures modélisées.

- Normalized zDOPE [58] score. Fonction de score interne au programme MODELLER (fonction assess.normalized_dope).
- ProQM [60] score. Le ProQM score est incorporé dans le logiciel Rosetta [93] (version 2015.05.57576 utilisée). Un ensemble de descripteurs associés à la séquence du récepteur $\alpha 7$ a été précalculé en utilisant les scripts mis à disposition par les développeurs de l’outil de d’évaluation (`https://github.com/bjornwallner/ProQ_scripts`). En complément, la localisation des domaines transmembranaires a été prédite grâce au serveur web TOPCONS [94] (paramètres par défaut et vérification manuelle de la prédiction) et les descripteurs de conservation de séquence avec PSI-BLAST [53] (*blastpgp* v. 2.2.18) contre la base de données Uniref90 [95] (v2015_05).
- PROCHECK G-factor [96]. Inclus dans l’exécutable PROCHECK (v3.5.4).

En complément, les modèles sélectionnés pour les états ouvert/fermé du récepteur $\alpha 7$ ainsi que les structures cristallographiques *templates* ont été évalués sur les serveurs en ligne SWISS-MODEL QMEAN [97] (score QMEAN6 normalisé), MOLProbity [98] (MOLProbity score), Verify3D [99] (pourcentage de résidus dont le score est supérieur ou égale à 0,2) et ProSA-web [100] (Z-Score global).

2.7 Sélection d’une *meilleure* structure

Chacune des exécutions de MODELLER ou de Rosetta s’accompagne de la création d’un millier de modèles. La sélection d’un *meilleur* représentant se fait par compa-

raison des scores précédemment décrits. Le zDOPE, le score ProQM et le G-factor sont calculés pour chacune des 1 000 structures. Le *meilleur* modèle est celui qui maximise la somme des rangs de chaque modèle trié selon les 3 scores.

3 Résultats

3.1 Résumé de l'approche de modélisation hybride

Le protocole de modélisation hybride profite du fonctionnement interne du programme MODELLER. En pratique, MODELLER mesure les motifs structuraux présents dans les structures *templates* et les traduit en contraintes structurales sur les atomes de la structure à modéliser. Par exemple, la valeur d'un angle entre 3 atomes dans les différents *templates* est traduite en une contrainte représentant la distribution de cet angle et appliquée aux 3 atomes correspondant dans la structure requête. Ici la correspondance entre des atomes des structures *templates* et de la structure à modéliser est directement issue de l'alignement multiple (typiquement les atomes de résidus d'une même colonne de l'alignement). Lorsque ces contraintes structurales ont été calculées, MODELLER optimise la position des atomes de la structure requête pour maximiser la satisfaction des contraintes structurales.

Afin d'enrichir nos modèles (α7)nAChR de l'information structurale provenant de l'ensemble des *templates*, nous commençons par calculer un ensemble de contraintes *hybrides*, c'est-à-dire issues de récepteurs homologues en états actifs, de repos, ou indéterminés (voir Figure II.5/p.suiv.). Cet ensemble hybride ne peut être utilisé directement par MODELLER pour créer des structures d'α7 au risque de créer des récepteurs mi-ouverts mi-fermés invraisemblables. L'idée est de déterminer un sous-ensemble de ces contraintes *hybrides* et de les remplacer par des contraintes dites *spécifiques* de l'état conformationnel du récepteur. Parmi l'ensemble des motifs structuraux calculés par MODELLER, les contraintes de distance entre carbones alpha (C_α) ont semblé être les plus naturelles pour dicter la conformation globale du récepteur. Deux nouveaux ensembles de contraintes, restreintes aux distances C_α-C_α, sont calculés à partir des *templates* ouverts et fermés. Ces deux jeux de contraintes sont ensuite comparés pour discerner ces contraintes significativement différentes entre les états actifs et de repos du récepteur. Ces contraintes *spécifiques* sont supposées être déterminantes pour la conformation finale du récepteur et sont alors réinjectées dans l'ensemble hybride. MODELLER est finalement chargé de construire des modèles structuraux à partir de ces ensembles de contraintes *hybrides-ouvertes* ou *hybrides-fermées*.

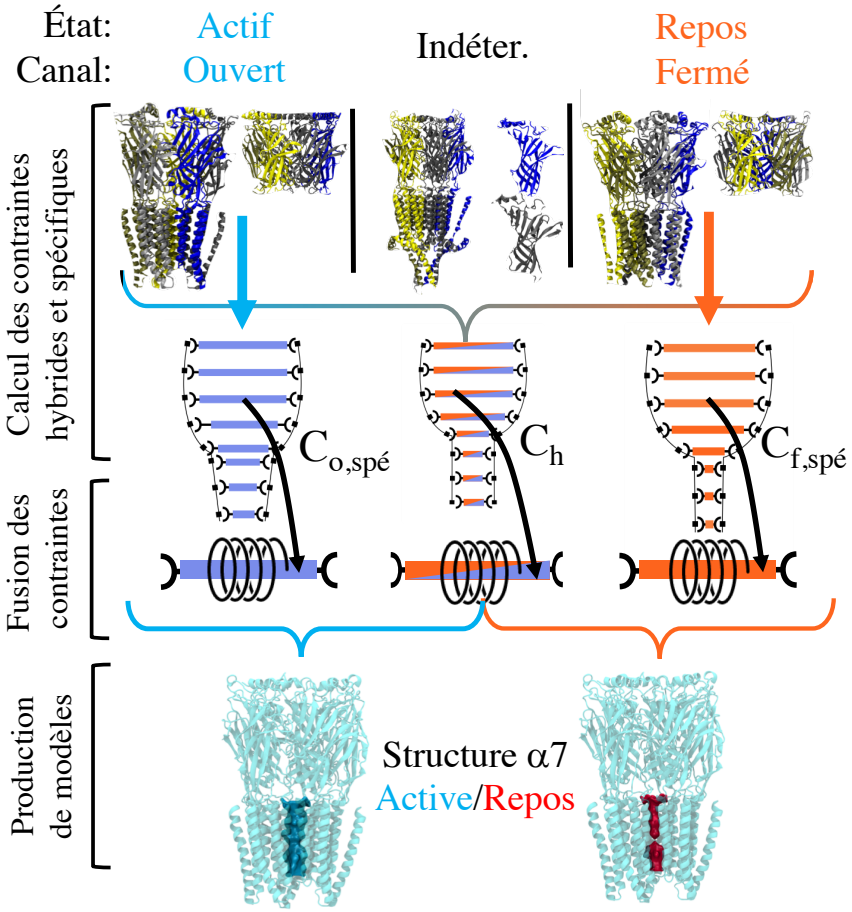


Figure II.5 Approche de modélisation hybride. Les contraintes structurales sur les atomes du récepteur $\alpha 7$ sont calculées à partir de l'ensemble des *templates*. Ces contraintes sont dites *hybrides* (C_h). D'autre part, deux autres ensembles de contraintes sont calculés à partir des *templates* ouverts (C_o) et fermés (C_f). Ces contraintes C_o et C_f sont alors comparées et triées pour déterminer un sous-ensemble de contraintes spécifiques à la conformation ouverte $C_{o,\text{spé}}$ ou fermée $C_{f,\text{spé}}$ du récepteur. Finalement, la fusion des contraintes *hybrides* et des contraintes *spécifiques* ouvertes/fermées permettent de générer des modèles du récepteur dans ses états actifs ou de repos.

3.2 Comparaison avec l'approche de modélisation classique

L'approche de modélisation classique consiste à n'utiliser que des *templates* dont l'état conformationnel est jugé comme étant celui à modéliser. À l'opposé, l'approche hybride autorise l'utilisation conjointe de structures de récepteurs homologues dont l'état fonctionnel diffère. La Figure II.6/p.suiv. démontre l'intérêt de la méthode hybride au regard de deux scores de qualités de structures, le score zDOPE et le score ProQM. Les scores d'un millier de modèles en état actif ou de repos sont produits avec l'approche classique et projetés (croix rouges) au côté du même nombre de modèles produits avec l'approche hybride (croix bleues). Les scores moyens entre approche classique et hybride sont respectivement de -0,61 et -0,71 (ouvert/zDOPE),

1 178 et 1 183 (ouvert/ProQM), -0,40 et -0,78 (fermé/zDOPE), 1 175 et 1 190 (fermé/proQM). Dans les deux cas, canal ouvert ou fermé, les modèles créés sont en moyenne plus pertinents avec l'approche hybride. Dans un second temps, une unique structure est sélectionnée dans chacun des ensembles de modèles précédemment créés (*meilleure* structure dont le choix est basé sur la somme des rangs des scores de modèles, cf. Section 2.7/p.31).

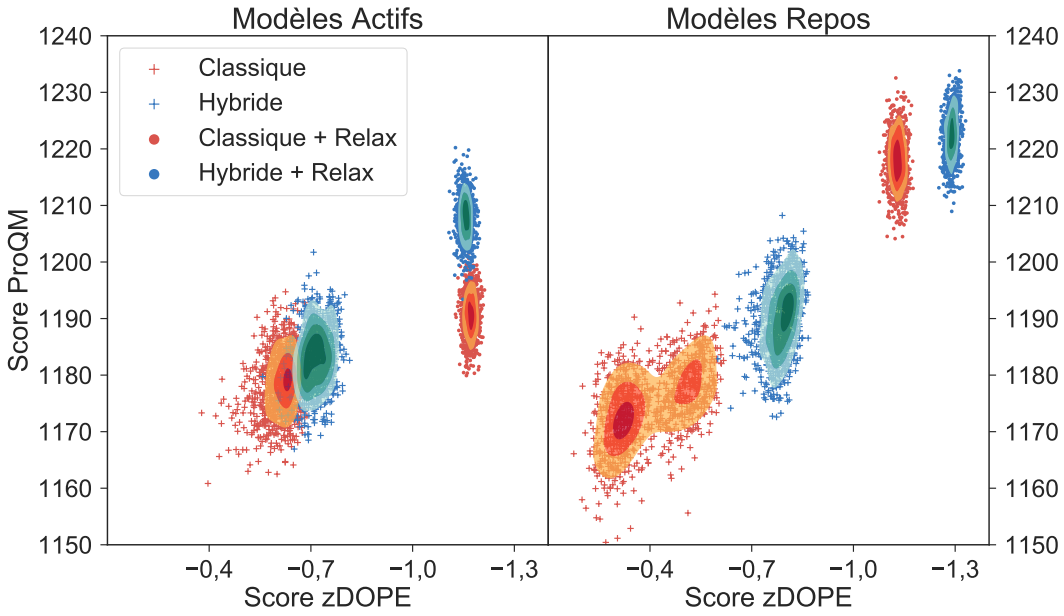


Figure II.6 Distribution des scores ProQM et zDOPE de modèles classiques et de modèles utilisant l'approche hybride. Les meilleurs scores de qualité sont vers le haut pour le score ProQM et vers la droite pour le score zDOPE. Les deux ensembles représentés par des points correspondent aux raffinements des "meilleurs" candidats des structures représentées par des croix. Le raffinement est réalisé par le script "Rosetta Membrane Relax" [92] et produit un ensemble de 1 000 nouveaux modèles.

Ces quatre modèles sont ensuite optimisés avec le protocole "Rosetta Membrane Relax" [92]. Ce protocole explore l'espace conformationnel proche de la structure initiale par la succession d'étapes de minimisations et de repositionnements des chaînes latérales (recuit simulé) tout en prenant en compte, implicitement, la membrane entourant le récepteur. Après cette optimisation, les scores moyens passent alors aux valeurs suivantes entre l'approche classique et hybride : -1,17 contre -1,15 (actif/zDOPE), 1 191 contre 1 207 (actif/proQM), -1,13 contre -1,29 (repos/zDOPE) et 1 218 contre 1 222 (repos/proQM). On remarque d'une part que le protocole "Rosetta Membrane Relax" a pour effet d'augmenter significativement le score des structures. En particulier, pour les modèles fermés, la totalité des modèles créés par Rosetta ont des scores zDOPE et ProQM supérieurs au score de la structure initiale utilisée pour l'optimisation. D'autre part, dans trois cas sur quatre (actif/proQM, repos/zDOPE et repos/proQM), les scores moyens de modèles créés à partir d'une structure initiale issue de l'approche classique sont inférieurs à ceux de l'approche hybride. Pour le cas de récepteurs actifs avec score zDOPE (-1,17 en Classique+Relax

et -1,15 en Hybride+Relax), les scores sont suffisamment proches pour ne pas remettre en cause l'approche hybride. Pour conclure, on observe que pour construire des modèles du récepteur nicotinique dans un état conformationnel actif ou de repos, l'approche hybride semble être plus efficace, du moins selon les deux fonctions de contrôle de qualités de structures utilisées. La Section 3.4/p.suiv. s'attachera à vérifier que les états conformationnels construits sont suffisamment proches des états fonctionnels attendus.

3.3 Qualité des structures produites avec l'approche hybride

La Figure II.7 répertorie les scores *in silico* des modèles ouverts et fermés créés à partir de l'approche hybride suivie d'un raffinement avec le protocole Rosetta Membrane Relax. Parce qu'il est difficile de juger un score absolu de qualité de structure, les différents scores associés aux *templates* cristallographiques ont aussi été retranscrits. Il s'agit alors non pas d'affirmer qu'un QMEAN de 0,720 est un bon ou un mauvais score, mais de le comparer aux scores que l'on peut attendre de structures cristallographiques. Les modèles ouverts et fermés ont de meilleurs scores que les 3 *templates* PDB : 4PIR, 4TNV et 3RIF pour les fonctions de score ProQM, zDOPE, QMEAN et molprobity.

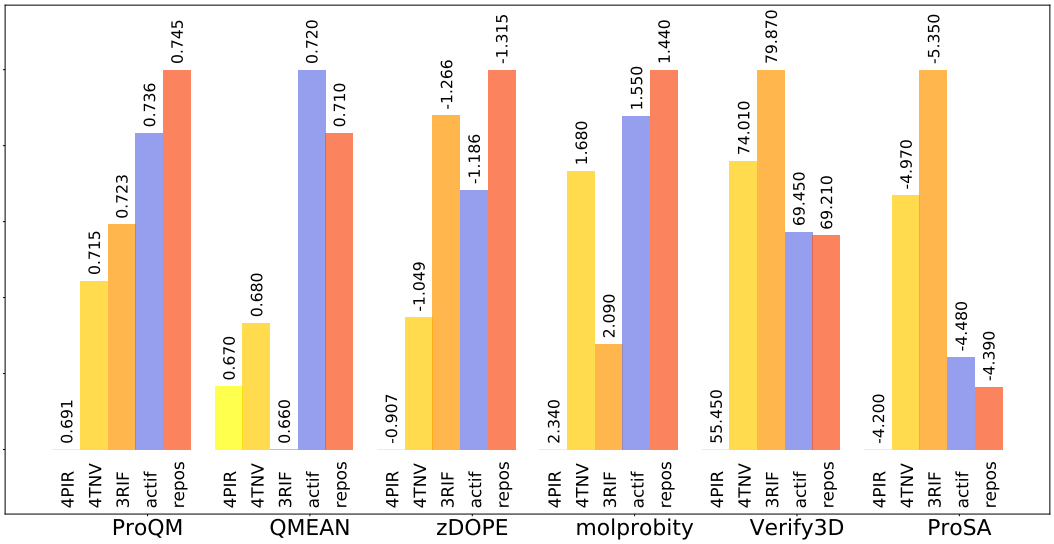


Figure II.7 Évaluation des structures selon diverses fonctions de score. Les modèles ouverts et fermés issus de l'approche hybride sont comparés aux structures cristallographiques utilisées comme *templates*. Pour chaque fonction de score utilisée, le moins bon score parmi les structures évaluées a une hauteur d'histogramme nulle alors que le meilleur score à une hauteur maximale.

Ces deux dernières fonctions de score, disponibles sous forme de serveurs web, sont intéressantes du fait qu'elles n'ont été utilisées à aucun moment de la construction ni

de la sélection des modèles. À l'inverse, il est moins surprenant que le score ProQM soit meilleur pour les modèles que pour les *templates*, le ProQM intervenant dans la sélection des *meilleurs* modèles ainsi que dans le raffinement des structures avec "Rosetta Membrane Relax" [92]. Le zDOPE score montre quant à lui un modèle ouvert plus défavorable que son homologue GluCl ouvert (id. PDB : 3RIF), mais plus pertinent pour l'état fermé (comparé à 4TNV). Les deux derniers programmes, Verify3D et ProSA, attribuent un score moins bon aux modèles hybrides qu'aux récepteurs *templates* GluCl, mais plus favorable que le score de la structure 4PIR. Finalement, il apparaît que les modèles créés avec l'approche hybride ont des scores de qualités comparables à ceux des *templates* utilisés pour la modélisation pour les six fonctions de score testées.

3.4 Choix du seuil de sélection des contraintes à remplacer

Le seuil de sélection des contraintes est utilisé pour segmenter les contraintes consensus entre *templates* ouverts/fermés des contraintes spécifiques de l'état actif ou de repos du récepteur. Lorsqu'une contrainte spécifique est repérée (c'est-à-dire une distance entre contraintes C_α-C_α actif et de repos supérieure à $S_{\text{spé}}$), elle vient remplacer la contrainte hybride correspondante. Ainsi, plus le seuil $S_{\text{spé}}$ est grand, moins nombreuses sont les contraintes remplacées, et plus les modèles produits seront éloignés d'une structure fonctionnellement en état actif ou de repos. En ne remplaçant aucune contrainte dans l'ensemble de contraintes hybrides, les modèles produits sont en moyenne à une distance RMSD (restreinte à la partie transmembranaire) de 2,01 Å de l'état actif, et à 3,12 Å de l'état de repos (comparaison avec les structures calculées avec l'approche classique). On peut observer sur la Figure II.8/p.suiv. que lorsque le seuil $S_{\text{spé}}$ décroît, le nombre de contraintes remplacées augmente alors mécaniquement, et les structures produites ont tendance à être de plus en plus proches d'une structure ouverte ou fermée. En l'occurrence, lorsque le seuil de sélection est inférieur ou égal à 0,025 les structures produites ont une distance RMSD inférieure à 1 Å de leur état structural ciblé. Au-delà de ce seuil un plateau est ensuite atteint. En utilisant un seuil égal à 0, 31 872 contraintes hybrides correspondant aux contraintes de distance entre C_α sont remplacées.

Les modèles produits ont ainsi un état du canal ouvert/fermé maximum, avec un RMSD moyen de 0,63 et de 0,56 entre les modèles Hybride-Actif/Repos et les 2 meilleurs modèles Classique-Actif/Repos. De façon intéressante, l'augmentation du nombre de contraintes remplacées dégrade le score G-Factor alors qu'elle améliore le score zDOPE (la tendance étant moins claire pour le score proQM). Ceci peut s'expliquer par le fait que le G-Factor est une mesure de qualité locale (mesure

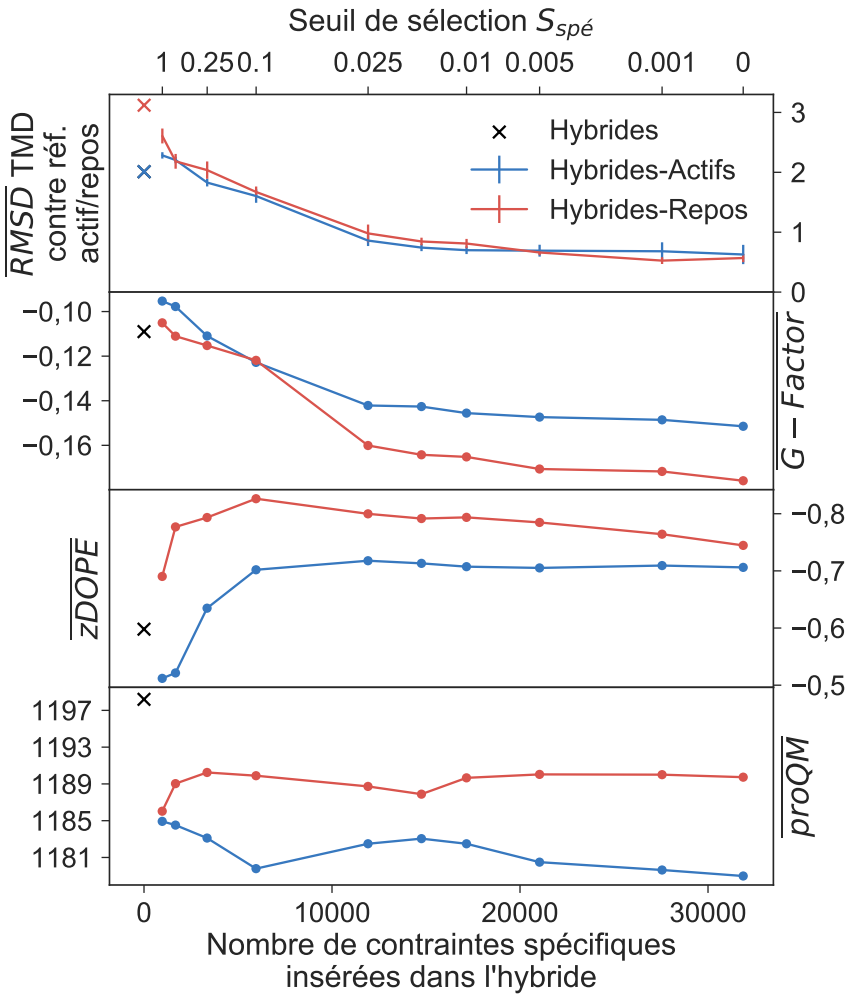


Figure II.8 Choix du seuil $S_{spé}$. Évolution des propriétés structurales de modèles $\alpha 7$ hybrides - actifs (courbes bleu) ou - repos (courbe rouges) en fonction du seuil $S_{spé}$ de sélection des contraintes spécifiques (axe des abscisses du haut) et nombre de contraintes remplacées dans les contraintes hybrides correspondant (axe des abscisses du bas). Chaque propriété est moyennée à partir de 1 000 modèles crées en utilisant le seuil $S_{spé}$. $\overline{RMSD}$ est le RMSD moyen entre les domaines transmembranaires (sélection des C_α et index de résidus supérieur à 204) des modèles Hybride-Repos (respectivement Hybrides-Actif) contre le *meilleur* modèle Classique-Repos (Classique-Actif) et après alignement des structures. Les scores G-Factor, proQM et zDOPE sont orientés par ordre croissant de qualité des structures (meilleures qualités en haut).

de la normalité des propriétés stéréochimiques des résidus), alors que le zDOPE est une mesure plus globale (basée sur des histogrammes de distances entre types d'atomes). Échanger une contrainte hybride par une contrainte ouverte/fermée a probablement pour conséquence de restreindre l'espace conformationnel accessible à MODELLER et rend plus difficile l'optimisation des angles de torsion et dièdres. La conformation globale est quant à elle plus pertinente proche de ses états ouverts et fermés fonctionnels. Finalement, les seuils de sélection choisis pour les états ouverts et fermés sont respectivement de 0,015 et 0,005, soit 14 757 et 21 040 contraintes remplacées dans l'ensemble de contraintes hybrides. Ces seuils ont été sélectionnés de sorte à produire des états du canal proches de ceux obtenus avec l'approche

classique (moyenne de 0,74 Å RMSD pour les ouverts et 0,66 Å RMSD les fermés) tout en garantissant des scores moyens pertinents.

4 Discussion

4.1 Intérêts de la méthode hybride

La méthode de modélisation comparative présentée dans cette partie peut être utilisée pour générer des états conformationnels multiples d'une protéine en intégrant un même ensemble hétérogène de *templates* structuraux. La modélisation des états actifs et de repos du récepteur nicotinique se prête particulièrement à la modélisation hybride pour les raisons suivantes : (1) les mécanismes fonctionnels du récepteur impliquent des conformations diverses, (2) les seules structures comparatives disponibles pour les états conformationnels ciblés proviennent d'homologues distants (identité $\approx$ 20 %) et (3) des structures plus proches ($\approx$ 60 % id.) existent mais ne correspondent pas à la conformation ciblée. La modélisation hybride récupère l'information structurale des *template* (sous forme de contraintes), peu importe leur état conformationnel, et va subtilement la biaiser vers l'état conformationnel ciblé. Pour cela, des contraintes à longue distance (distances entre C_α) spécifiques de l'état conformationnel ciblé sont sélectionnées. Cette sélection de contraintes spécifiques longue distance ne doit pas trop souffrir de la faible homologie des structures *template* choisies (raison n°2), puisque la structure tertiaire des protéines est plus robuste aux évolutions de séquence. À l'inverse, l'utilisation des contraintes hybrides, issues de tous les *templates*, permet de profiter de l'information structurale locale de structures homologues plus proches (raison n°3). Dans l'étude présente, l'approche hybride nous a permis d'obtenir des structures plus pertinentes (selon des fonctions de score variées) qu'avec l'approche classique de modélisation comparative qui ne peut, dans ce cas particulier, profiter que des *templates* de faible homologie.

4.2 Choix des structures *templates*

Le choix des structures *templates* a été établi en collaboration avec l'unité des Récepteurs Canaux de l'Institut Pasteur (Pierre-Jean Corringer, Marc Gielen). Premièrement, les structures cryo-EM du récepteur nicotinique *torpedo marmorata* [33, 101] n'ont pas été utilisées, bien que publiées pour le récepteur entier (ECD, TMD et une partie ICD) et en états canal ouvert et au repos. Un certain nombre d'études ultérieures ont remis en cause l'attribution de la structure dans la densité électronique [102–104] On notera tout de même que de nouveaux modèles en état ouvert et fermé ont récemment été publiés à partir des données existantes de densités [80]. Parmi la variété de structures homologues publiées au moment de la réalisation de

ce projet, deux autres récepteurs étaient connus dans leur conformation active et de repos. Les récepteurs GluCl [42, 44] et les récepteurs GLIC. Il est difficile de dire quel système est *a priori* plus pertinent pour modéliser le récepteur nicotinique. L'homologie de séquence avec la sous-unité $\alpha 7$ est faible pour les deux systèmes ($\approx$ 20 % pour GluCl et $\approx$ 19 % pour GLIC). Leur organisation structurale est relativement similaire (observation des mouvements de *blooming* et de *twisting*) mais diffère sur certains points (différences locales entre feuillets β de la partie extracellulaire et globales dans l'inclinaison du domaine EC lors de l'extension EC) [44, 105]. Pour ce travail nous avons choisi le récepteur GluCl comme *template* de la structure quaternaire du récepteur nicotinique. Des modèles utilisant GLIC ont aussi été testés lors de l'optimisation de l'alignement multiple, mais ne permettaient pas d'obtenir des scores de qualité comparables à ceux obtenus avec les structures GluCl. On notera que GLIC n'est pas activé par des ligands mais par les protons, ce qui change probablement son mécanisme d'activation par rapport au récepteur nicotinique et GluCl (à décharge d'un choix de GLIC, comme les nAChR ils sont perméables aux cations, contrairement au GluCl). Ce choix est débattable, et idéalement, on pourrait utiliser l'approche hybride avec les structures ouvertes et de repos de GLIC plutôt que GluCl et en comparer les différences.

4.3 Métriques d'évaluation

L'évaluation de la méthode de modélisation hybride présentée dans cette partie reste minimale. Nous avons montré que les modèles produits avec l'approche hybride ont tendance à être plus pertinents que ceux permis par l'approche classique de modélisation, grâce à l'évaluation *in silico* des structures par une sélection diverse de fonctions de score de qualité des protéines. Cependant, rien ne garantit que les structures prédites soient biologiquement vraisemblables (au-delà de la précision de ces fonctions de score). Rigoureusement, il faudrait pouvoir comparer le résultat de la modélisation avec des structures expérimentales du récepteur nAChR en état actif et de repos, qui ne sont pas encore disponibles (bien que des efforts considérables aient été faits dans ce domaine, avec la structure homologue proche $\alpha 4\beta 2$, récemment cristallisés en état désensibilisé). Une autre possibilité d'évaluation de la méthode profiterait des structures expérimentales ouvertes, fermées et en état alternatif de GLIC. En utilisant la méthode hybride, nous pouvons prédire les états conformationnels actifs et de repos avec la configuration minimale de la méthode hybride suivante : Séquence cible : GLIC, *template* ouvert : GluCl (3RIF), *template* fermé : GluCl (4TNV), *template* indéterminé : GLIC (p.ex. 3TLW, état "localy closed"). En sortie de la modélisation hybride, nous obtiendrions des modèles en état actif et de repos de GLIC, directement comparables avec les structures cristallographiques existantes de GLIC (p.ex. GLIC PH4 - 4HFI, pour l'état actif, et GLIC PH7 - 4NPQ,

pour l'état de repos). Cette comparaison plus rigoureuse, permettrait d'éprouver la méthode par rapport à des données expérimentales, plutôt que des fonctions de score dont les limites sont réelles.

4.4 Critiques *a posteriori* des modèles

Au début de la thèse, l'obtention de structures pour les états actifs et de repos constituait un prérequis nécessaire à l'avancement du projet : calcul de la transition conformationnelle, étude des sites effecteurs, etc.. Le développement de la méthode a occupé les 6 premiers mois du travail de thèse. Les choix et résultats obtenus pendant cette période ont donc nécessairement eu des conséquences sur le reste des travaux qui ont suivi. Voici plusieurs remarques concernant les modèles produits.

La première faiblesse des modèles s'est manifestée lors des calculs de chemins de transition et du suivi simultané des mouvements quaternaires supposés indispensables à l'ouverture du canal. Une extension radiale de l'ensemble de la partie extracellulaire est clairement visible dans les structures cristallographiques de canaux ioniques homologues à l'état de repos (récepteurs GLIC [42] et GluCl [44]). Cette extension s'est trouvé être très réduite dans les premiers chemins de transition calculés à partir des deux structures obtenues par modélisation hybride (cf. Figure III.10/p.72). L'extension ECD a par la suite été analysée pour chacun des *templates* utilisés pendant la modélisation. Il s'est avéré que les structures chimériques ($\alpha 7$) d'AChBP utilisées ne présentent pas d'extension ECD significative entre le *template* de la forme au repos (3SH1) et active (3SQ6). Du fait de leur plus forte homologie avec la séquence $\alpha 7$, la conformation quaternaire des AChBP a eu plus de poids lors de la modélisation que les structures *templates* des récepteurs GluCl (qui quant à eux possèdent cette propriété d'extension). Du fait de la faible homologie de séquence de la majorité des *templates* et de moins bonnes résolutions cristallographiques lorsque la structure inclut la partie transmembranaire, il aurait été difficile de se passer des AChBP pour modéliser le récepteur. On notera tout de même l'existence de structures d'AChBP présentant une extension ECD à l'état de repos [106]. Bien que n'étant pas des chimères $\alpha 7$, ces structures auraient pu éventuellement nous aider à obtenir des conformations plus satisfaisantes.

Un autre défaut des modèles est apparu plus tard, lorsque le récepteur humain $\alpha 4\beta 2$ a été publié. La comparaison attentive des résidus conservés de la poche orthostérique de l'$\alpha 4\beta 2$ a laissé apparaître des inconsistances dans l'orientation de certains résidus aromatiques. La Figure II.9/p.suiv. illustre cette problématique. Le tryptophane 146 du modèle $\alpha 7$ à l'état de repos a un positionnement similaire à celui observé dans la structure 5HT3, alors que la comparaison avec l'$\alpha 4\beta 2$ nous aurait

fait privilégier l'orientation de l'AChBP, dont la poche orthostérique partage d'ailleurs une plus grande identité avec la séquence $\alpha 7$. En pratique, cela met en évidence des ambiguïtés d'orientation des chaînes latérales identiques (et très conservées) entre les *templates*. Pour ces cas particuliers, nous laissons implicitement le programme MODELLER choisir la meilleure orientation du résidu, non plus sur un critère de conservation de séquence, mais par rapport à ses métriques internes (champs de force, fonctions de score), ce qui semble dommageable pour la modélisation. Il est probable que le nombre significatif de structures *templates* utilisées pendant la modélisation (x7) soit la cause de ces ambiguïtés.

Bien que les deux remarques précédentes aient eu un impact sur les travaux qui ont suivi la modélisation (on notera que ces problèmes ont pu être corrigés), elles ne remettent pas directement en cause la modélisation hybride mais posent plutôt des alertes sur la modélisation par homologie classique (choix des *templates*, nombre de *templates*...).

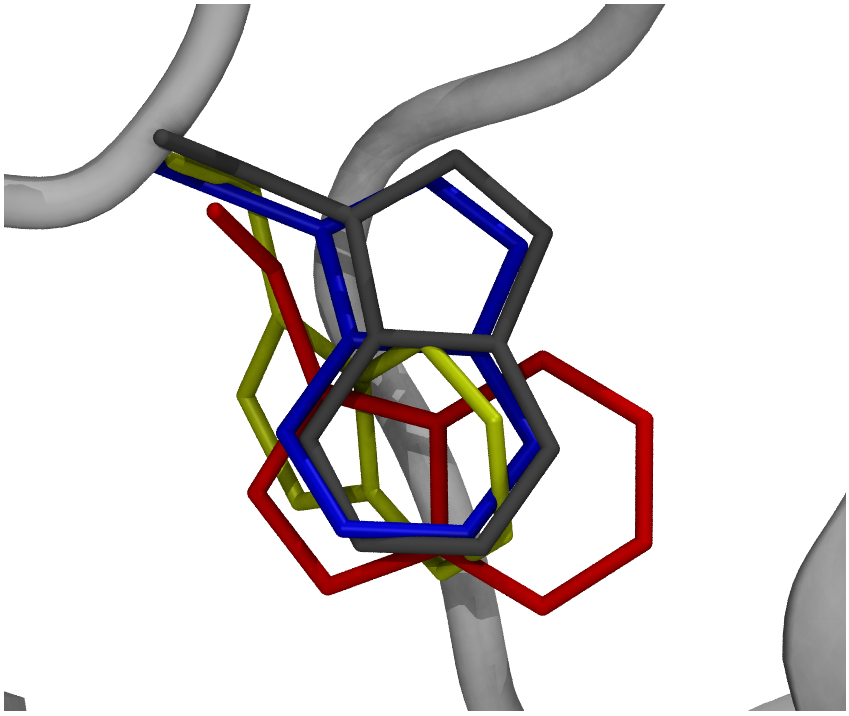


Figure II.9 Ambiguïté locale des *templates*. Orientation du résidu TRP146 dans la poche orthostérique du modèle $\alpha 7$ en état actif (résidu jaune). Les tryptophanes du *template* AChBP (3SQ6-TRP145) sont en bleu et $5HT_3$ (PDB :4PIR-TRP156) en rouge et d'$\alpha 4\beta 2$ (5KXI-($\alpha 4$)TRP156) en gris. Le squelette de la structure 3SQ6 est représenté en *cartoon* blanc. L'alignement des résidus a été réalisé par alignement structural de la chaîne A des 3 structures grâce au module MultiSeq de VMD.

5 Conclusions et perspectives

L'approche de modélisation hybride présentée dans cette partie nous a permis d'établir des modèles structuraux pour les états actifs et de repos du récepteur nicotinique de sous-unités $\alpha 7$. Cette étape a été un préalable important au développement des méthodes de calcul de chemins de transition. On notera que la philosophie de la méthode hybride est en accord avec un prérequis du *Conjugate Peak Refinment*, la méthode d'optimisation de chemins adiabatiques utilisée dans la partie suivante :

> *Make sure that the internal coordinates of atoms that are not significantly involved in the reaction are the same for the reactant and the product.* [107]

En ne plaçant pas de contraintes spécifiques dans les régions consensuelles des récepteurs actifs et au repos, la méthode hybride garantit qu'un ensemble de contraintes hybrides identiques sera utilisé pour modéliser ces régions non impactées par la réaction.

La publication récente d'une structure expérimentale du récepteur nicotinique $\alpha 4\beta 2$ humain [34] nous permettra sans doute de raffiner la structure de l'$\alpha 7$, et justifie d'autant plus l'utilisation de la méthode hybride. La conformation saisie par cristallographie est prétendument dans un état désensibilisé, non conducteur. Il est donc impossible d'utiliser cette structure, pourtant évolutivement proche, pour modéliser le récepteur nicotinique dans ses états actifs ou de repos en utilisant l'approche de modélisation comparative classique. La méthode hybride pourrait être configurée pour utiliser les structures GluCl ou GLIC comme *template* des états actifs et de repos, et le *template* $\alpha 4\beta 2$ comme *template* d'état indéterminé. Cela aurait l'avantage de limiter les ambiguïtés mises en évidence en Discussion (cf. Section 4.4/p.42) du trop grand nombre de *templates* (passant de 7 à 3), et de l'utilisation des AChBP pour modéliser les états actifs et inactifs (remplacés par l'$\alpha 4\beta 2$ dont la séquence est proche de l'$\alpha 7$).

Cette méthode ouvre la voie à la modélisation de stœchiométries plus complexes des sous unités du récepteur nicotinique. Vis-à-vis de l'approche hybride peu de changements techniques sont nécessaires. Il faut d'une part préparer la séquence cible comme la concaténation des séquences de chacune des sous-unités et d'autre part modifier la contrainte de symétrie appliquée au récepteur entier (n'appliquer la symétrie que sur les chaînes identiques). La génération de structures de sous-types différents pourra par exemple permettre une meilleure évaluation par amarrage

moléculaire de la spécificité de petites molécules, pour des sites de liaisons différents en fonction de la sous-unité exprimée.

III

Calcul de chemins de transition : couplage POE/SoS

Nous cherchons à calculer une série d'intermédiaires structuraux décrivant le mécanisme d'activation du récepteur nicotinique. La caractérisation précise d'états transitoires de protéine fait l'objet d'actives recherches, en particulier pour ses applications dans le développement de nouvelles molécules thérapeutiques interférant les mécanismes fonctionnels de la cible [108–110]. L'environnement hétérogène avec lequel interagit le récepteur nicotinique (eau, membrane, ions), ainsi que la subtilité des mécanismes énergétiques mis en jeu, tend à nous faire privilégier l'utilisation de méthodes dynamiques calculant des chemins d'énergie libre minimaux. Cependant, ces méthodes sont souvent limitées à un nombre de Variables Collectives restreint dont le choix pose question, et exposées aux bruits browniens présents dans les dynamiques moléculaires, ce qui peut nuire à l'interprétabilité des résultats. L'Unité de Bioinformatique Structurale a développé des méthodes de simplification de chemins de transition, dans le contexte plus rigide décrit par le paysage d'énergie potentiel. Cette partie propose une approche nouvelle, le couplage des méthodes *Path Optimization and Exploration* [109] pour le calcul de chemins de transition adiabatiques, avec la *String Method with Swarms of Trajectories* [111] pour le raffinement des structures intermédiaires dans un contexte d'énergie libre.

1 Introduction

Des mouvements conformationnels sont impliqués dans de nombreux processus biologiques, allant du repliement des protéines, à la reconnaissance de ligands et à la catalyse des enzymes [112–114], en passant par des phénomènes complexes de régulations allostériques où de larges réorganisations conformationnelles peuvent rentrer en jeu [115]. La description fine de ces mouvements s'avère ainsi d'une importance cruciale pour mieux comprendre les mécanismes fonctionnels qui régissent ces systèmes biologiques. L'analyse de la dynamique des protéines peut alors ouvrir la voie vers le développement de molécules thérapeutiques spécifiquement conçues pour interférer de tels mouvements et moduler le comportement de la cible étudiée [116, 117].

Les protéines sont des macromolécules intrinsèquement dynamiques. La flexibilité donnée par le nombre important de degrés de liberté de rotation par résidu implique une immensité de conformations structurales théorique [118]. Pourtant, le plus souvent, seule un petit nombre de conformations stables sont privilégiées. La description intelligible des mouvements fonctionnels d'une protéine implique la connaissance de ses états stables, énergétiquement favorables, mais aussi de l'essentiel de ses conformations intermédiaires, énergétiquement moins favorables mais accessibles. Alors que les états stables sont dans un grand nombre de cas accessibles expérimentalement par des méthodes biophysiques reconnues (cristallographie aux rayons X, spectroscopie RMN, cryo-microscopie électronique), la détermination d'états "rares" de la protéine fait encore l'objet d'actives recherches (par exemple : *Room temperature X-Ray Crystallography* [119, 120]). De nombreuses méthodes *in silico* ont été établies pour dévoiler ces états transitoires.

Des mesures *in vitro* réalisées sur l'activation de récepteurs nicotiniques individuels montrent qu'une ouverture spontanée du canal (en absence d'agonistes) s'observe en moyenne une fois par seconde (1,3±1,4/sec [121]). En comparaison, les simulations de dynamique moléculaire sur des récepteurs canaux se limitent à la microseconde et peinent à observer des évènements de fermeture complets [50]. Des méthodes de simulation ont été développées pour accélérer l'exploration de la dynamique moléculaire séparant des états stables de protéine [122, 123]. La plupart de ces méthodes simplifient l'hyperespace à explorer en ne considérant qu'un petit nombre de Variables Collectives (ou coordonnées réactionnelles et abrégées "CV"), fonctions des coordonnées cartésiennes du système, et supposées suffisantes pour décrire pleinement la réaction. Par exemple, certaines approches échantillonnent le gradient d'énergie libre (la force moyenne) autour de points particuliers de l'espace définit

par Variables Collectives (TI [124], ABF [125], ou TAMD [126]). D'autres méthodes (*Umbrella Sampling* [74], *Métadynamique* [73]) biaisent le potentiel énergétique pour forcer la simulation à sortir des bassins énergétiques favorables déjà explorés et projettent alors la distribution d'états réduite sur les CV. Cependant, une mauvaise sélection des Variables Collectives peut mener à des inconsistances lors de l'exploration du paysage énergétique (typiquement, des barrières énergétiques mal prises en compte dans l'espace restreint des CV [127, 128]). Cette sélection est d'autant plus difficile à surmonter que beaucoup de ces méthodes n'acceptent qu'un petit nombre de Variable Collective (p.ex. max. ≈ 3 pour la Métadynamique) alors que la dynamique d'une protéine peut avoir des dizaines de milliers de degrés de liberté. De plus, les déterminants mécanistiques du système étudié sont généralement inconnus *a priori*.

La *String Method with Swarms of Trajectories* (ou *String of Swarms*, SoS, dont le principe est décrit Figure III.1/p.suiv.), initialement basée sur la "string method" de Maragliano *et al.* [129], simplifie l'exploration du paysage énergétique entre deux états stables en raffinant un chemin d'énergie libre minimale. Alors que les méthodes décrites précédemment nécessitent l'exploration exhaustive de l'hypervolume décrit par les CV, les *String of Swarms* se contentent d'optimiser un chemin, c'est-à-dire une construction unidimensionnelle dans l'espace des CV. Ainsi l'ensemble des Variables Collectives choisi pour décrire le système peut être beaucoup plus large, voire comprendre la majeure partie du système (par exemple l'ensemble des atomes du squelette de la protéine [130, 131]), sans pour autant pénaliser la convergence de l'algorithme. Le chemin réactionnel est constitué d'une chaîne de structures intermédiaires (aussi appelés états, ou billes) de la protéine. À chaque itération, un essaim de courtes dynamiques moléculaires non contraintes est lancé à partir de chacun des intermédiaires structuraux. Ces dynamiques moléculaires, une fois moyennées, servent à apprécier la direction du gradient de diffusion. Le chemin est ensuite mis à jour pour s'aligner le long du gradient de diffusion et ainsi glisser naturellement dans une vallée d'énergie libre. SoS a été utilisé dans de nombreuses études pour décrire le chemin réactionnel de systèmes complexes impliquant de larges changements de conformations [110, 111, 130–140] et même plus récemment sur un homologue procaryote du récepteur nicotinique [141, 142]. Du fait des mouvements aléatoires inhérents à la dynamique moléculaire, SoS ne converge pas nécessairement en un chemin précis, mais en un ensemble de chemins bruités (contenus dans un tube réactionnel), ce qui rend difficile l'appréciation de la convergence de la méthode ainsi que l'analyse mécanistique de la transition [143]. Des astuces permettent de contourner cette difficulté (p.ex. moyenner les derniers chemins SoS [139]) mais ne nous semblent pas satisfaisantes. Pour surmonter cette difficulté, nous proposons dans ce travail de coupler la méthode dynamique des *String of Swarms*, avec la méthode de calcul de chemins adiabatiques *Path Optimization and Exploration* (POE), développée au laboratoire et spécialisée dans la

recherche de raccourcis topologiques pertinents. Alors que les méthodes dynamiques

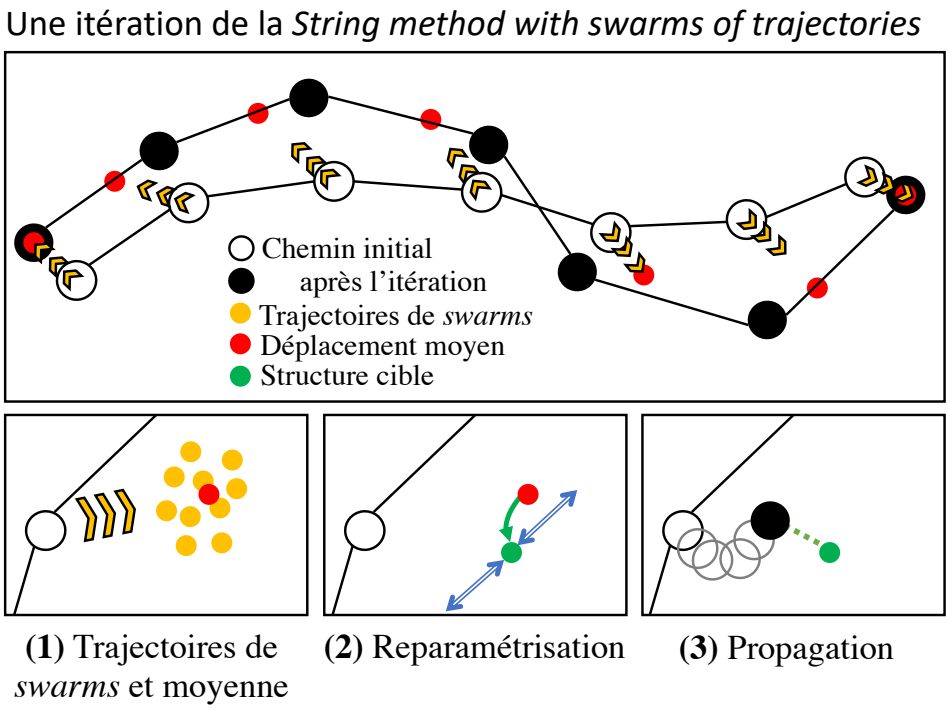


Figure III.1 Principe de la *String method with swarms of trajectories*. À partir d'une série de structures intermédiaires séparant deux états distincts de la protéine, le chemin est relaxé dans le paysage d'énergie libre par la procédure suivante réalisée sur chacune des structures du chemin : **(1)** Une multitude de courtes trajectoires de dynamique moléculaires indépendantes et sans contraintes sont lancées à partir de la structure initiale (bille blanche). Les conformations de fin de trajectoires (billes jaunes) sont converties en une structure moyenne (bille rouge), qui pointe la direction du gradient de diffusion. **(2)** la structure moyenne est reparamétrisée pour limiter une dérive des structures le long du chemin. La structure cible (bille verte) est ainsi créée pour préserver une équidistance entre les structures. **(3)** une trajectoire de propagation est lancée à partir de la structure initiale et une contrainte harmonique sur la bille cible. La dernière structure de la trajectoire de propagation (bille noire) devient un nouveau point intermédiaire du chemin de transition. Après plusieurs itérations de ce protocole, les chemins produits convergent dans une vallée du paysage d'énergie libre.

précédemment décrites sont contraintes d'échantillonner le paysage d'énergie libre exhaustivement pour évaluer les contributions entropiques, d'autres méthodes sont spécialisées dans la recherche de chemins d'énergie minimale sur la surface d'énergie potentielle. L'exploration de la surface d'énergie potentielle revient à décrire le mécanisme de transition dans un contexte adiabatique. Les propriétés avantageuses des champs de force de mécanique moléculaire (continuité, dérivabilité) permettent l'optimisation numérique de chemins d'énergie minimale sans aucun *a priori* sur les coordonnées réactionnelles conduisant la réaction. On citera parmi les méthodes les plus connues, les *Nudge Elastic Band* [144], la *Zero Temperature String Method* [145, 146], ou le *Conjugate Peak Refinment* (CPR) [147]. La méthode CPR a été privilégiée au laboratoire de par ses propriétés intéressantes pour l'interprétation des transitions. Les structures intermédiaires calculées par CPR sont ordonnées et constituent un

chemin énergétiquement valide (pas de barrières énergétiques entre les points du chemin). Chaque point de la transition est ainsi (théoriquement) cinétiquement accessible [148]. Cependant, à très haute dimension, la surface d'énergie potentielle est un monstre topologique, et les chemins optimisés deviennent rapidement très complexes (observation faites au laboratoire, et identifiable dans la référence [149]) ce qui a mené à l'élaboration de la méthode POE. POE recherche des raccourcis topologiques le long de la transition et construit puis sélectionne des chemins alternatifs plus pertinents sur des critères énergétiques (abaissement du maximum énergétique) et topologiques (chemin plus court, moins de points intermédiaires). POE a en particulier permis de calculer des conformations intermédiaires de l'Adenylyl Cyclase du facteur œdémateux de l'Anthrax à l'origine de la découverte du premier inhibiteur allostérique identifié *in silico* [150, 151].

L'objectif du travail présenté dans cette partie est d'étudier le mécanisme d'activation du récepteur nicotinique de sous-unité $\alpha 7$, grâce au couplage avantageux de la méthode dynamique des *String of Swarms* avec la méthode d'optimisation de chemins Path Optimization and Exploration. Nous souhaitons profiter des simulations réalistes de SoS (chemin d'énergie libre, solvatation comprenant eau, ions et membrane (phospholipides et cholestérols) explicites, système thermalisé) avec les bonnes propriétés de POE (absence de barrières enthalpiques, simplicité des chemins).

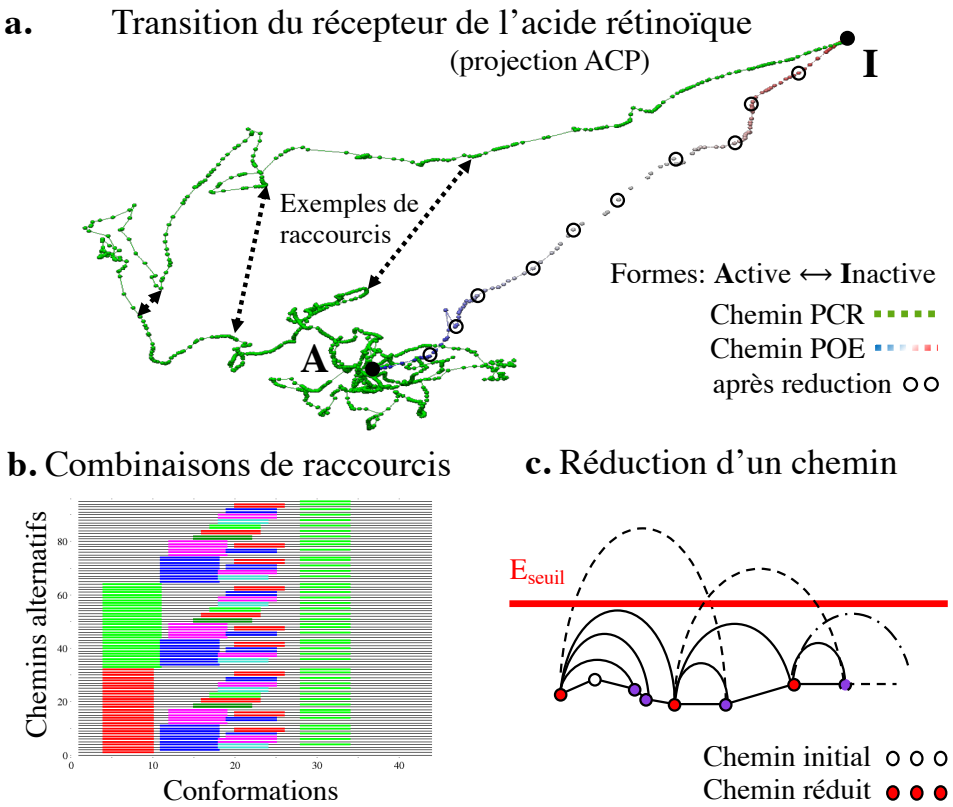


Figure III.2 **Optimisation de chemins de transition à l'aide de POE. a.** Projection sur les deux premiers modes d'une analyse en composante principale des chemins PCR et POE d'une transition du récepteur de l'acide rétinoïque. Le chemin PCR est topologiquement très complexe (points verts). POE recherche les raccourcis topologiques entre les structures intermédiaires du chemin PCR (flèches noirs). Après plusieurs itérations, le chemin obtenu avec POE est beaucoup plus court (points en dégradé du bleu au rouge). Une procédure de réduction du chemin POE permet de supprimer les structures intermédiaires superflues à une échelle plus locale à différentes étapes de la méthode (ronds noirs, voir **c.**). **b.** Combinatoire des combinaisons de raccourcis. Les lignes correspondent à des chemins de transition alternatifs qui seront confronté au chemin initial grâce à des critères énergétiques (énergie maximale le long du chemin) et topologiques (nombre de points intermédiaires et longueur totale du chemin). Les lignes noires réfèrent à des portions du chemin initial. Les lignes en couleurs sont des raccourcis topologiques trouvés par POE. **c.** Procédure de réduction d'un chemin de transition à l'échelle plus locale. La réduction préserve la continuité énergétique du chemin. En d'autres termes, les structures calculées sur une interpolation entre chacune des paires consécutives de structures du chemin ne dépassent pas un plateau d'énergie $E_{\text{plat.}}$. Sous cette contrainte de continuité, la réduction supprime un maximum de structures intermédiaires.

2 Matériels et Méthodes

2.1 Couplage POE-SoS

2.1.1 Solvatation et équilibration des structures de départ

Les deux structures issues de la modélisation comparative, canal ouvert et canal en état de repos ont indépendamment été solvatées à l'aide du programme "Membrane Builder" du serveur web CHARMM-GUI [152]. Le format de coordonnées PDB des modèles MODELLER est traduit en format *CHARMM coordinates* et PSF. Deux ponts disulfures sont manuellement ajoutés au PSF pour chacune des 5 chaînes du récepteur entre les cystéines 125-139 et 187-188. Le récepteur est orienté sur l'axe des Z et la partie transmembranaire centrée en Z=0. La protéine est placée dans une boîte tétragonale de taille 115x115x153 $Å^3$. Une bicouche lipidique hétérogène est constituée autour du récepteur : 213 1-palmitoyl-2-oleoyl-phosphatidylcholine (POPC) et 71 1-palmitoyl-2-oleoyl-phosphatidic-acid (POPA) ainsi que 71 molécules de cholestérols (CHL). 180 lipides sont placés sur la couche inférieure et 175 sur la couche supérieure. La stœchiométrie choisie de rapport 3xPOPC/1xPOPA/1xCHL est connue pour stabiliser les nAChRs dans un état fonctionnel [153–156]. La neutralité électrique à pH 7 a été obtenue avec 41 ions Cl^- et 142 Na^+ pour une concentration de NaCl d'environ 150mM. Lipides et ions ont été ajoutés autour de la protéine grâce aux scripts du CHARMM-GUI. La gestion du placement des molécules d'eau a quant à elle a été gérée par des immersions et déletions itératives, à l'aide d'un script CHARMM écrit au laboratoire. Le nombre de molécules d'eau inclues dans la boîte de solvatation est ainsi ajusté pour une densité d'eau de 0,033 molécule d'H_2O (type CHARMM TIP3P [157]) par $Å^3$, soit :

$$(Volume_{\text{boîte}} - Volume_{\text{protéine}} - Volume_{\text{lipides}}) \times 0{,}033$$

Le volume de la protéine a été calculé à partir des données biophysiques du volume individuel des acides aminés, soit 232 456 $Å^3$. Le volume de la membrane a été obtenu avec le programme MSMS (v2.6.1), et est de 386 311 $Å^3$ pour la forme au repos du récepteur et de 381 617 $Å^3$ pour la forme active. Ce qui correspond à 46 954 molécules d'eau calculées pour la forme canal fermé et 47 111 pour la forme ouverte. Un nombre de molécules consensus de 47 033 a été utilisé. Le placement des molécules d'eau est réalisé comme suit :

1. Saturation de la boîte. Une boîte d'eau pré-équilibrée est superposée à la boîte contenant protéine, ions et lipides. Seules les molécules d'eau ne chevauchant aucun atome du système sont conservées. Cette opération est ensuite réitérée, après légère translation aléatoire de la boîte d'eau pré-équilibrée sur ses 3 axes, jusqu'à l'ajout de 48 000 molécules.

2. Suppression itérative des molécules d'eau. L'énergie de chacune des molécules d'eau est calculée. Les 25 molécules les moins favorables sont supprimées (plus hautes énergies et en excluant les plus proches voisines des molécules déjà supprimées). Après réitérations successives, la boîte d'eau contenant 47033 molécules est sauvegardée.

La boîte solvatée est ensuite minimisée avec l'algorithme de *steepest descent* combiné à la détection et au déplacement systématique des molécules d'eau "bloquées" (hautes énergies) dans la protéine ou la membrane. Deux premières équilibrations sont réalisées avec la protéine gardée fixe (dynamique de Langevin avec coefficient de friction à 3,0 ps^{-1} de 50 ps, et seconde avec coefficient à 0,1 ps^{-1} pour 500 ps de simulation avec un pas temporel d'une femtoseconde) suivie d'une seconde équilibration de 500 ps où la protéine est soumise à une contrainte harmonique sur la structure initiale (paramètre de force à 2,5) ainsi qu'à une contrainte hyperplane (force 10^6 kcal/mol.$Å^2$).

Finalement, la boîte de simulation contient 209 709 atomes dont 27 395 atomes pour la protéine seule. Cette solvatation initiale, coûteuse en temps de calculs et en contrôles humains, a seulement été calculée pour les états initiaux, ouverts et fermés, du récepteur.

2.1.2 *Path Optimization and Exploration*

Path Optimization and Exploration (POE) est un ensemble de routines Shell, CHARMM et en Fortran, permettant le calcul et le raffinement de chemins de transition.

POE a été exécuté avec CHARMM [64] version 35b2 et la paramétrisation du champ de force tout-atome CHARMM36 [61–63]. Un modèle de solvatation approximatif, qui mime au mieux un diélectrique sigmoïdal [158], est utilisé pour l'ensemble des optimisations POE : diélectrique dépendant de la distance (RDIElec EPS 1,416) avec décalage du potentiel à 10 Å (SHIFt CUTN 10,0), les interactions Van der Waals sont décalées entre 7 et 8 Å (VSWItch CTON 7,0 CTOF 8,0). Des contraintes sont ajoutées sur les prolines du récepteur pour éviter les isomérisations *cis-trans* (force 1 kcal/mol.rad^2). La proline 133 est placée en conformation *cis* [159] (Ω=0 °), les autres prolines en *trans* (Ω=180 °). L'algorithme CPR utilisé dans POE est implé-

menté par le module TReK de CHARMM [107] (incluant une implémentation de la procédure de réduction (décrite dans la référence [6])) et autorise un déplacement maximum de 0,5 Å entre les atomes.

Pour la première itération de POE, un premier chemin est créé par interpolation entre les deux structures initiales désolvatées (26 395 atomes). Pour cela, le script CHARMM *Hammer Drill* (développé au laboratoire) combine une interpolation en coordonnées cartésiennes pour le squelette de la protéine et une interpolation en coordonnées internes pour les chaînes latérales (vector.pl fourni dans la distribution CHARMM [107]).

Pour les itérations de POE dont le chemin provient d'une série de *String of Swarms*, les billes extrémales sont extraites de la transition et minimisées par un enchaînement des algorithmes *Steepest Descent* et *Adopted Basis Newton Raphson* (avec la contrainte hyperplane - C_{HP} - dont la définition est donnée plus loin 2.1.4/p.58). Finalement, elles sont ajoutées en première et dernière position du chemin et transmises à POE.

Une itération POE est composée de deux procédures : *PathFinder* identifie les raccourcis plausibles le long de la transition, et *Assemble* reconstruit des chemins de transition alternatifs à partir de ces combinaisons de raccourcis et du chemin de départ.

Pour un chemin de transition composé de M structures intermédiaires, l'ensemble des paires de conformations $X^{(k)}$ et $X^{(k')}$ est évalué par *PathFinder*. Bien que la combinatoire des raccourcis soit grande ($M(M-1)/2$) la taille du *cluster* de calcul ($\approx$ 500 cœurs) nous a permis d'exécuter *PathFinder* sur la totalité des raccourcis possibles. Pour chaque paire de conformation *PathFinder* exécute la procédure *Hammer Drill* pour construire le chemin de transition séparant les deux points. Un sous-ensemble de raccourcis est sélectionné sur des critères énergétiques (énergie maximale le long du chemin) et topologiques (nombres de points intermédiaires, longueur du chemin).

Le script *Assemble* combine les raccourcis pour former de nouvelles transitions de la structure de départ à celle d'arrivée (combinaisons de raccourcis non chevauchants, complétées par les portions du chemin initial). Le chemin initial, ainsi que les chemins alternatifs sont optimisés par itérations successives de CPR et de réduction des chemins (15x). Finalement, un chemin alternatif est choisi si son énergie maximale et sa longueur curvilinéaire (somme des rmsd entre conformations adjacentes) sont plus faibles que celles du chemin initial. Le chemin sélectionné est réduit avant d'être soumi à une autre itération de POE ou à une série de SoS.

La réduction des chemins est une procédure visant à supprimer un maximum de points de la transition tant qu'un plateau énergétique $E_{\text{plat.}}$ n'est pas dépassé le long du chemin. Soit $X^{(k)}$ et $X^{(k+i)}$ deux conformations du chemin. Les (i-1) intermédiaires sont supprimés si l'énergie de structures interpolées entre $X^{(k)}$ et $X^{(k+i)}$ ne dépasse pas $E_{\text{plat.}}$. En pratique, la recherche se fait en arrière, c'est à dire pour k allant de 1 à M et i de $M - k$ à 1. La valeur de k suivante est alors incrémentée de la plus grande valeur de i pour laquelle le critère d'énergie est vérifié.

Deux itérations de POE ont été exécutées entre chacune des séries de *String of Swarms*.

2.1.3 Solvatation des structures d'un chemin POE

À chaque passage d'un chemin de transition issu de POE vers la méthode des *String of Swarms*, chacune des structures de la trajectoire de transition doit être convenablement solvatée. Au fil des itérations POE-SoS, 159 solvatations de conformations du récepteur ont été réalisées. Ce nombre justifie la mise au point d'une stratégie de solvatation plus rapide que celle proposée pour les structures de départ. La Figure III.3/p.suiv. illustre la solution proposée. Chacune des nouvelles structures à solvater est comparée aux structures de la trajectoire de transition obtenues par la série SoS précédente. La boîte solvatée contenant la structure du récepteur la plus proche de la structure à solvater est sélectionnée. Les coordonnées atomiques des molécules d'eau, des lipides et des ions sont ensuite ajoutées aux coordonnées de la protéine à solvater. Afin de résoudre les possibles collisions entre atomes, en particulier à l'interface entre la protéine et le solvant, les mêmes étapes de minimisation et d'équilibration (500 ps avec protéine fixe et 100 ps avec contraintes harmoniques et hyperplanes, voir 2.1.4/p.58) que celle réalisées lors de la solvatation initiale sont ajoutées. Les molécules d'eau dont l'énergie reste trop forte après minimisation sont repoussées radicalement (coordonnées x et y multipliées par $\sqrt{2}$) puis reminimisées. La procédure est reconduite jusqu'à disparition de toutes les alertes. Ces simulations sont réalisées indépendamment les unes des autres pour chacune des structures du chemin et elles sont donc traitées en parallèle sur le *cluster* de calcul.

2.1.4 *String of Swarms*

Un chemin de transition est représenté par une chaîne de M conformations X de la protéine, $\{X^{(1)}, X^{(2)}, ..., X^{(M)}\}$. En l'occurrence, le nombre d'états M correspond au nombre de structures intermédiaires obtenues après optimisation par POE, et varie donc au cours des itérations POE-SoS ($M_1 = 36$, $M_2 = 33$, $M_3 = 32$, $M_4 = 30$, $M_5 = 28$). Chaque itération de la méthode *String of Swarms* a pour objectif de

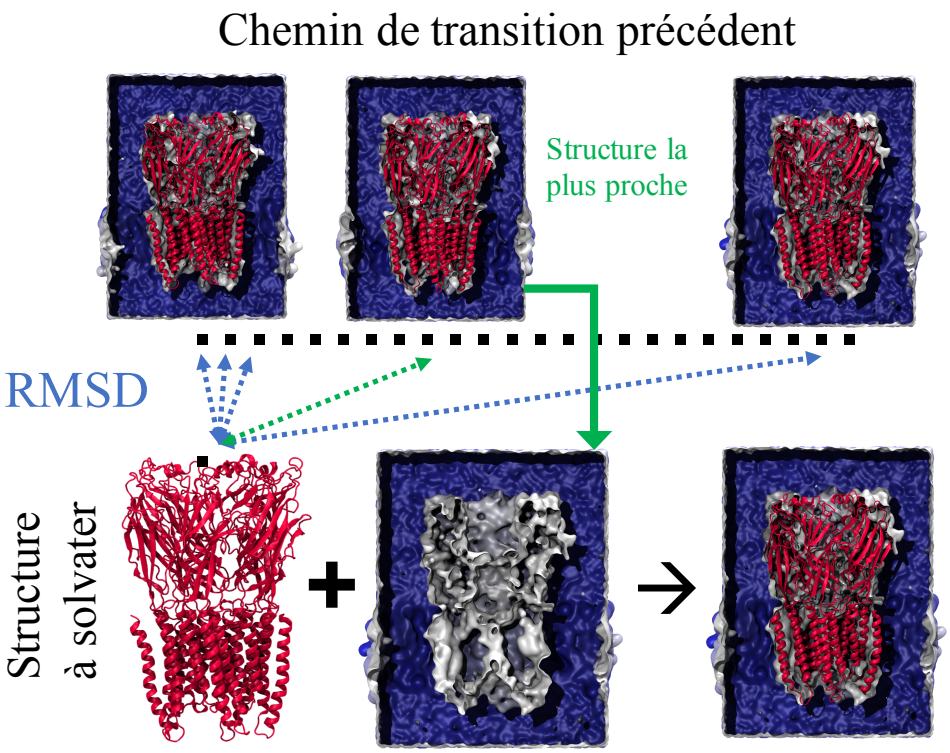


Figure III.3 Réutilisation de boîtes solvatées pré-équilibrées. La structure à solvater est insérée (copie des coordonnées atomiques) dans la boîte solvatée contenant la conformation la plus proche (RMSD minimal). Après minimisation et équilibration suffisante, le système est prêt à être utilisé par les *String of Swarms*.

mettre à jour chaque structure intermédiaire $X^{(i)}$ de sorte à ce qu'elle s'adapte au gradient de diffusion d'un essaim de courtes dynamiques moléculaires non contraintes. La procédure de référence des *String of Swarms* est décrite dans la référence [111]. Des adaptations spécifiques de la méthode ont été nécessaires en vue de sa complémentarité avec *Path Optimization and Exploration*. En particulier, la diffusion des structures intermédiaires est contrainte dans des hyperplans conjugués au chemin de la réaction dans le but de limiter la diffusion spontanée des structures vers les bassins énergétiques stables aux extrémités du chemin.

Remarques générales sur les simulations moléculaires

Dynamique moléculaire L'ensemble des simulations de dynamique moléculaire concernant les *String of Swarms* a été réalisé avec le programme CHARMM [64] version 39b1 et le champ de force tout-atome CHARMM36 [61–63] pour la topologie et les paramètres de la protéine, des lipides [160, 161], du cholestérol (généré avec CGenFF [162, 163]), des ions [164] et des molécules d'eau (TIP3P [157]). Le système est simulé dans l'ensemble canonique (NVT). Un thermostat de Langevin est utilisé pour chauffer le système à 303,15 °K, avec un coefficient de friction à 0,2 ps^{-1}. Les interactions électrostatiques à longue distance sont calculées dans l'espace réciproque à l'aide de la sommation d'Ewald (Particle Mesh Ewald [68]). Les

conditions périodiques sont appliquées à la boîte solvatée avec le module CRYSTAL de CHARMM. Les liaisons hydrogènes sont contraintes avec l'algorithme SHAKE [165]. Pour accélérer les temps de calcul la décomposition en domaines de la boîte de simulation est gérée par DOMDEC [166], ce qui permet l'utilisation conjointe de plusieurs CPUs et du GPU pour une même simulation.

Hyperplans et contraintes hyperplanes Soit $\{X^{(1)}, X^{(2)}, ..., X^{(M)}\}$ la chaîne ordonnée de conformations de la protéine décrivant le chemin de transition.

L'hyperplan $HP^{(i)}$ est défini pour une structure $X^{(i)}$ comme l'espace à 3N-1 dimensions, bissecteur des deux vecteurs formés par les conformations $X^{(i-1)}$ et $X^{(i+1)}$ avec $X^{(i)}$. Autrement dit, $HP^{(i)}$ contient l'ensemble des dimensions orthogonales au vecteur normal $\overrightarrow{N}^{(i)}$ ainsi défini :

$$\overrightarrow{N}^{(i)} = \frac{\overrightarrow{V}^{+} + \overrightarrow{V}^{-}}{2}$$

avec $\overrightarrow{V}^{-} = \frac{X^{(i)} - X^{(i-1)}}{\|X^{(i)} - X^{(i-1)}\|}$ et $\overrightarrow{V}^{+} = \frac{X^{(i+1)} - X^{(i)}}{\|X^{(i+1)} - X^{(i)}\|}$. Les cas extrémaux $HP^{(1)}$ et $HP^{(M)}$ sont définis en considérant respectivement $\overrightarrow{V}^{-} = 0$ et $\overrightarrow{V}^{+} = 0$.

La contrainte hyperplane, notée $C_{HP}^{(i)}$, est sous la forme d'une contrainte harmonique entre une conformation courante X', et sa projection dans l'hyperplan $HP^{(i)}$.

$$C_{HP}^{(i)} = k \left[\left(X^{(i)-X'} \right) \cdot \overrightarrow{N}^{(i)} \right]^2$$

Tous les atomes de la protéine sont ici concernés mais pas le solvant, ni la membrane et ni les ions. La constante de force k a une valeur de 10^6 kcal/mol/Å^2 jugée suffisante pour que les structures produites soient suffisamment contraintes dans l'hyperplan (voir Annexe VII.7/p.199). La contrainte hyperplane a été implémentée au laboratoire et distribuée à partir de la version c38b2 dans le module CONS de CHARMM (sous-modules "RMSD RELA HPLA").

Protocole d'une itération SoS

Trajectoires de *Swarms* Chacune des structures initiales $X^{(i)}$ est utilisée comme structure de départ de 32 dynamiques moléculaires indépendantes et sans contrainte. L'assignation initiale des vélocités est réalisée par CHARMM avec des graines aléatoires différentes. Le temps de simulation de chaque dynamique est de 20 ps (10 000 itérations avec un pas de temps de 2 fs). Seule la dernière structure de chacune des simulations, $D_j^{(i)}$ avec $1 \leq j \leq 32$, est gardée en mémoire. Les 32 structures sont alors moyennées :

$$\overline{D}^{(i)} = \frac{1}{32} \sum_{i=1}^{M} D_j^{(i)}$$

Équilibration supplémentaire du solvant Une dynamique supplémentaire d'équilibration du solvant (lipides, ions, eau) de 200 ps est réalisée avec la protéine fixe (les $X^{(i)}$ restent inchangés). En pratique, cette étape est réalisée en même temps que le calcul des trajectoires de *Swarms*.

Reparamétrisation Soit $\overline{D}^{(i)}$, la structure moyenne des trajectoires de *Swarms*. La structure cible $T^{(i)}$ pour l'étape de propagation, est définie comme la structure résultant de la projection de $\overline{D}^{(i)}$ dans l'hyperplan $HP^{(i)}$. En d'autres termes,

$$T^{(i)} = \overline{D}^{(i)} - \left(\left(\overline{D}^{(i)} - X^{(i)}\right) \cdot \overrightarrow{N}^{(i)}\right) \times \overrightarrow{N}^{(i)}$$

Propagation Durant l'étape de propagation, une courte dynamique moléculaire (4 ps, pas de trajectoire à 2fs) démarre avec $X^{(i)}$ comme structure de départ. Une contrainte $C_{propa.}$ est ajoutée au champ de force pour guider les structures de la dynamique vers la bille cible.

$$C^{(i)}_{propa.} = C^{(i)}_{harm.} + C^{(i)}_{HP}$$

où $C_{harm.}$ est une contrainte harmonique sur la structure cible $T^{(i)}$ (contrainte sur les carbones alpha (C_α) du récepteur avec constante de force à 2,5 kcal/mol/Å^2), et la contrainte hyperplane C_{HP} précédemment définie. Finalement, les conformations atomiques obtenues à la fin des trajectoires de propagation forment une nouvelle chaîne d'état de la protéine, $\{X^{(1)}, X^{(2)}, ..., X^{(M)}\}$, qui peut-être soit soumise à une nouvelle itération de *String of Swarms*, soit, si elles sont arrivées à convergence, désolvatées et réintroduites à une itération de la méthode *Path Optimization and Exploration*.

2.2 Modifications ponctuelles des transitions

2.2.1 Correction locale de la poche orthostérique

La comparaison de nos modèles de transition avec la structure $\alpha4\beta2$ (PDB : 5KXI [34]) publiée fin 2016 nous a permis d'identifier des incohérences dans l'orientation de résidus aromatiques de la poche orthostérique (voir Discussion en Partie 4.4/p.42). Pour tenir compte de ces nouvelles données structurales, les angles dièdres chi1 et chi2 de 5 résidus de la poche (voir le Tableau III.1/p.suiv.) ont été transposés de la chaîne A ($\alpha4$) aux 5 chaînes de toutes les conformations de la transition de fin de série Swarms S3, avant le POE S4. La correction est faite par édition des coordonnées

internes avec CHARMM (“ic fill”, “ic edit”, “coor init”, et “ic build”) et minimisation (“steepest descent”) locale des atomes à moins de 5 Å des résidus concernés.

	$\alpha 4$ (resid)	chi1 (°)	chi2 (°)	$\alpha 7$ (resid)
TRP	62	-51,62	147,66	52
TYR	100	-77,81	-60,42	90
TRP	156	-156,85	81,59	146
TYR	197	-62,07	-64,95	185
TYR	204	-67,13	-88,41	192

Tableau III.1 Correction locale des modèles. Angles chi1 et chi2 reportés de la sous-unité $alpha4$ du récepteur $\alpha 4\beta 2$ humain (PDB : 5KXI) vers les 5 sous-unités $\alpha 7$ par translation des coordonnées internes. Cette étape a eu lieu entre SoS-S3 et POE-S4.

2.2.2 Extrapolation des bornes

Cette opération a été une tentative de correction des inconsistances énergétiques observées aux bornes des profils d'énergie libre et tient compte du fait que lors de l'optimisation d'un chemin avec POE, seul les points intermédiaires du chemin sont raffinés. Le chemin provenant de la dernière itération des *String Of Swarms* de la série S3 est une série de conformations $\{X^{(1)}, X^{(2)}, ..., X^{(33)}\}$. Des structures supplémentaires $X^{(0)}$ et $X^{(34)}$ ont été construites par extrapolation à partir des couples de structures $\left(X^{(1)},X^{(2)}\right)$ et $\left(X^{(32)},X^{(33)}\right)$. POE a ensuite été exécuté deux fois (a7m9 et a7m10, voir le Tableau III.3/p.78) sur le chemin $\{X^{(0)}, X^{(1)}, ..., X^{(34)}\}$. Le chemin a7m10 a ensuite été tronqué avant réduction en utilisant comme nouvelles bornes les 2 structures les plus proches (RMS) des structures $X^{(1)}$ et $X^{(33)}$ à l'étape précédant les extrapolations. Cette modification n'a pas été en mesure de résoudre les inconsistances relevées.

2.2.3 Régularisation des chaînes latérales symétriques

Avant raffinement par POE du dernier chemin SoS de la série S4, les résidus possédant des chaînes latérales symétriques (aspartate, glutamate, arginine, phénylalanine et tyrosine) ont été régularisés pour éviter les rotations superflus de chaînes latérales. Le chemin résultant est raffiné par le POE de la série S5.

2.3 Analyse des simulations

2.3.1 Profils d'énergie libre

Les profils d'énergie libre sont évalués par intégration thermodynamique en intégrant les forces moyennes le long du chemin de transition :

$$(PMF)_{0,\alpha} = F(X^{(\alpha)}) - F(X^{(0)}) = \int_0^\alpha \frac{dF}{d\alpha'} d\alpha' = \int_0^\alpha \frac{\delta F(X^{(\alpha')})}{\delta X_i} d\alpha' \cdot \frac{dX^{(\alpha')}}{d\alpha'}$$

où $\frac{dX_i^{(\alpha')}}{d\alpha'} = \overrightarrow{N}^{(\alpha')}$ le vecteur normal à l'hyperplan pour la structure intermédiaire $X^{(\alpha')}$ et

$$\frac{\delta F(X^{(\alpha')})}{\delta X} = -\langle \overrightarrow{F}_{C_{HP}} + \overrightarrow{F}_{C_{harm}} \rangle$$

La moyenne des forces résultantes des contraintes harmoniques et hyperplanes est échantillonnée toutes les 10 ps d'une trajectoire totale d'une nanoseconde, avec contraintes C_{HP} et C_{harm} sur la structure courante $X^{(i)}$. L'intégrale est évaluée par une somme pondérée où les w_k sont calculés par la méthode des trapèzes :

$$(PMF)_{0,\alpha} = \sum_{k=1}^{M} w_k \left(\frac{dF}{d\alpha}\right)_k$$

où $w_k = 1$ sauf pour $k = 1$ et M pour lesquels $w_k = \frac{1}{2}$. De plus l'erreur statistique est évaluée au fil de la réaction :

$$err((PMF)_{0,\alpha}) = \sqrt{var((PMF)_{0,\alpha})} = \sqrt{\sum_{k=1}^{M} w_k^2 var\left(\frac{dF}{d\alpha}\right)_k}$$

2.3.2 Distance entre deux chemins de transition

Soit C_1 et C_2 deux chemins de transition respectivement composés de N et M structures intermédiaires. Nous définissons la distance entre les deux chemins $\delta(C_1,C_2)$, exprimée en angström, tels que :

$$\delta(C_1,C_2) = \frac{1}{2}\left(\frac{1}{N}\sum_{i=1}^{N} \min_{j=1\ldots M}\left(rms(C_1^{(i)},C_2^{(j)})\right) + \frac{1}{M}\sum_{j=1}^{M} \min_{i=1\ldots N}\left(rms(C_1^{(i)},C_2^{(j)})\right)\right)$$

Où $rms(C_1^{(i)},C_2^{(j)})$ correspond à la distance RMSD tout atome entre les deux structures $C_1^{(i)}$ et $C_2^{(j)}$ après alignement structural.

2.3.3 Propriétés du récepteur nicotinique

Extension ECD

La propriété d'extension du domaine extracellulaire est calculée comme le rayon de giration $(1/N \sum_i (x_{mean} - x_i)^2 + (y_{mean} - y_i)^2 + (z_{mean} - z_i)^2)$ des atomes de type C_α de la partie extracellulaire définie par la plage de résidus 15 à 205. Des plages de résidus équivalentes ont visuellement été sélectionnées pour chacune des structures cristallographiques afin de garantir une comparaison juste de la propriété étudiée : résidus 17 à 213 pour 3RIF (id. PDB) et 4TNV, 5 à 194 pour 4HFI et 4NPQ, 19 à 209 pour 3SH1 et 17 à 204 pour 3SQ6, sélectionnées par superposition des structures cristallographiques avec les modèles du récepteur $\alpha 7$.

Torsion ECD-TMD

La propriété de torsion entre la partie extracellulaire et la partie transmembranaire du récepteur est calculée comme la moyenne des angles dièdres obtenus pour chacune des chaînes ($i = \{a,b,c,d,e\}$) à partir des trois vecteurs consécutifs $\overrightarrow{ECD_i,ECD_{abcde}}$, $\overrightarrow{ECD_{abcde},TMD_{abcde}}$ et $\overrightarrow{TMD_{abcde},TMD_i}$ où ECD_i est la moyenne des coordonnées atomiques de type C_α de la partie extracellulaire de la chaîne *i* , ECD_{abcde} de la partie extracellulaire entière, TMD_{abcde} de la partie transmembranaire entière et TMD_i de la partie transmembranaire de la chaîne *i*. La partie extracellulaire est délimitée par les résidus précédemment cités dans la sous-section précédente, la partie transmembranaire par tous les résidus d'index supérieurs. Pour des raisons pratiques, le signe des angles calculés a été inversé.

3 Résultats

3.1 Couplage POE/SoS

Ce travail décrit le couplage de deux méthodes de calcul de chemins de transition : *Path Optimization and Exploration* (POE) optimise un chemin sur la surface d'énergie potentielle (ΔH), et la *String method with Swarms of Trajectories* (SoS) sur le paysage d'énergie libre (ΔG). Pour montrer l'intérêt de cette approche, la combinaison des deux méthodes a été appliquée à la description du mécanisme d'activation/fermeture du récepteur nicotinique de l'acétylcholine de sous-unités $\alpha 7$. Les conformations active et de repos sont celles modélisées dans la Partie II/p.21. En pratique, le couplage est réalisé par l'optimisation successive de chemins de transition au travers de POE et des *String of Swarms*, comme illustré dans la Figure III.4.

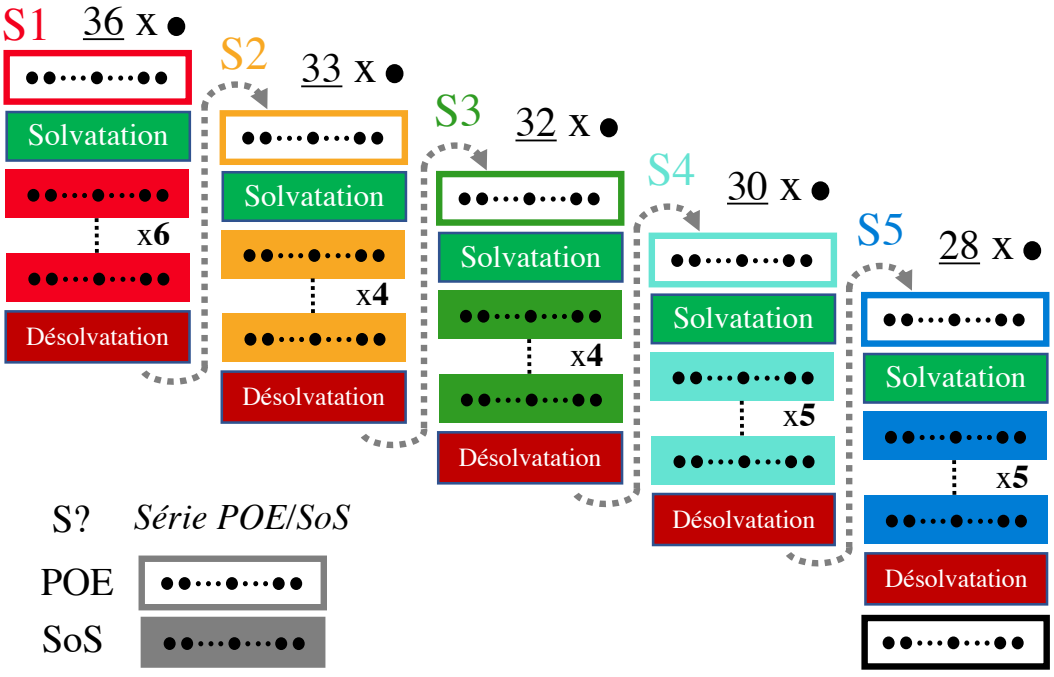


Figure III.4 Plan de simulation pour le couplage POE-SoS. Chacune des 5 colonnes schématise le passage d'un chemin de transition par "Path Optimization and Exploration" (POE) suivie de plusieurs itérations (nombres en gras) de la "String method with Swarms of Trajectories" (SoS). Le nombre de structures intermédiaires des chemins est déterminé par POE (nombres soulignés) et ne varie pas lors des itérations de SoS. Une étape de solvatation et de désolvatation (eau/ions/lipides) encadre chacune des itérations des *String of Swarms*.

Un chemin de transition est ici défini par un ensemble ordonné de structures (aussi appelées conformations ou billes). Le premier chemin est créé entre les deux modèles du récepteur par interpolation linéaire (interpolation en coordonnées cartésienne pour le squelette de la protéine et interpolation en coordonnées internes pour

Série SoS	équili. init.	équili. supp.	Swarms	propagations	total
S1	5 964	8 134	17 724	6 448	38 270
S2	5 072	3 897	23 171	3 947	36 087
S3	4 305	2 622	10 874	3 567	21 368
S4	4 152	5 477	10 398	4 032	24 059
S5	4 081	5 422	10 211	3 755	23 469
Total	23 574	25 552	72 378	21 749	143 253 $\approx$ 16ans

Tableau III.2 Temps de calcul CPU (en heures) alloués pour les différentes séries de *String of Swarms*. On notera que ces chiffres sont hautement dépendants de l'utilisation ou non de GPU lors des dynamiques moléculaires. En particulier, l'accélération de l'étape S3 par rapport à S2 est non seulement dûe à un nombre de billes intermédiaires plus faibles mais aussi à un meilleur adressage des simulations vers des machines avec GPU. L'équilibration initiale (équili. init.) succède à la solvatation des chemins de POE. L'équilibration supplémentaire (équili. supp.) est effectuée en parallèle de l'étape des *Swarms*.

les chaînes latérales, voir documentation TreK de CHARMM [107]). La série de structures obtenues est ensuite passée en entrée à POE. Ce premier chemin est ainsi optimisé une première fois sur la surface d'énergie potentielle décrite par le champ de force CHARMM36. L'optimisation s'y fait à 0° Kelvin degrés sans échange de chaleurs (adiabatique), ce qui rend nécessaire l'utilisation d'une représentation implicite (risque de *gel* des membranes). Chacune des structures construites par POE doit ensuite être solvatée (eau, ions, POPC-POPA-Cholestérol) en vue du transfert du chemin à la méthode des *String of Swarms*. Cette solvatation des chemins, rencontrée à chaque passage entre une itération de POE vers une itération de SoS a été spécifiquement élaborée pour profiter des temps d'équilibration accumulés dans les itérations précédentes (voir Matériels et Méthodes). Les structures solvatées sont alors soumises à plusieurs itérations des *String of Swarms*, de telle sorte que chacune des billes intermédiaires soit déplacée en direction du gradient de diffusion d'un essaim de courtes dynamiques moléculaire sans contrainte et à 303 degrés Kelvin. Dans ce contexte, la chaîne de billes décrivant la transition évolue progressivement en direction d'une vallée d'énergie libre. Lorsque deux chemins consécutifs des *String of Swarms* sont suffisamment proches, la méthode est supposée avoir convergé et les itérations s'arrêtent. Entre 4 et 6 itérations ont été réalisées pour chacune des séries des *String of Swarms*. Une fois le chemin ayant convergé dans le contexte d'énergie libre, les structures sont une à une désolvatées, et repassées en entrée à une itération de Path Optimization and Exploration.

5 séries successives de passages POE-SoS ont été réalisées. Rien que pour les itérations de SoS, cela représente environ 16 ans d'utilisation sur un CPU (Figure III.2). Grâce à la grille de calcul du laboratoire, il faut environ une semaine de calculs pour chacune des séries de Strings of Swarms et 4 à 5 jours pour une exécution complète de POE.

Les deux méthodes, POE et SoS, optimisent les structures intermédiaires du chemin dans des contextes différents (ΔH et ΔG). Nous devons donc nous assurer, que le passage d'une méthode à l'autre se fait de façon conciliante pour chacune des structures du chemin. À la fin d'un cycle POE, chacune des structures intermédiaires du chemin est un minimum local sur les $3N - 1$ dimensions conjuguées. Dans l'algorithme CPR utilisé par POE, ces dimensions conjuguées sont matérialisées par un hyperplan, bissecteur des deux vecteurs normalisés formés avec les structures adjacentes du chemin [147, 167] (voir Figure III.5). Le protocole des *String of Swarms* a été adapté en conséquence pour garantir que la diffusion des structures du chemin dans le paysage d'énergie libre tienne compte de ces hyperplans au cours de l'optimisation. En pratique cela passe par une modification de l'étape de reparamétrisation [111]. La reparamétrisation des billes moyennes de trajectoires de Swarms est indispensable, pour éviter que les structures intermédiaires ne diffusent naturellement vers les états extrémaux stables de la transition.

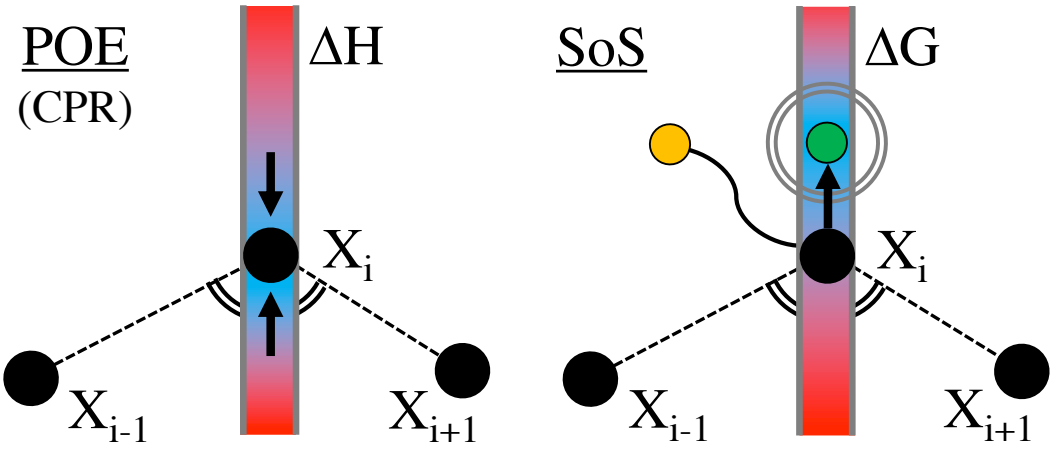


Figure III.5 Compatibilité technique entre POE et SoS. En gris, l'hyperplan bissecteur des vecteurs formés entre la structure courante (X_i) et ses 2 structures adjacentes (X_{i-1} et X_{i+1}). À la fin de POE, chacune des structures du chemin est dans un minimum énergétique local dans son hyperplan (contexte ΔH). Lorsque le chemin est plongé dans le contexte d'énergie libre avec la méthode des *String of Swarms*, ce même hyperplan est utilisé pour laisser diffuser la structure dans un minimum d'énergie libre (ΔG) et garantir que les structures intermédiaires restent au même stade d'avancement de la réaction. En bille jaune, la structure moyenne des trajectoires de Swarms et en vert la structure cible, projection de la bille moyenne dans l'hyperplan.

Dans l'étape de reparamétrisation décrite par Pan et al. [111], les billes moyennes sont reconstruites de telle sorte qu'elles soient géométriquement équidistantes les unes par rapport aux autres. Ce déplacement artificiel pousse les structures hors des hyperplans dans lesquelles elles ont été optimisées par POE, et par conséquent risque de nuire à une optimisation consécutive par POE en éloignant la structure des pivots structuraux contournant les pics d'énergie. Ainsi, en suivant la procédure originelle on brise la continuité énergétique du chemin. Nous proposons de modifier la reparamétrisation en projetant la bille moyenne des Swarms directement dans l'hyperplan bissecteur (Figure III.5) et en ajoutant une contrainte harmonique sur

l'hyperplan lors de l'étape de propagation. Cette mise au point méthodologique permet en théorie la compatibilité des chemins lors du transfert de la forme adiabatique vers le champ d'énergie libre et la garantie du maintien des billes bien réparties tout au long du chemin lors de la propagation des *Swarms*.

3.1.1 Espace conformationnel exploré

L'ensemble des conformations des chemins de transition calculés au fil des itérations POE-SoS a été projeté sur les deux premiers modes d'une Analyse en Composante Principale (ACP) réalisée sur ces mêmes structures (Figure III.6). Cet ensemble contient 925 structures résultant de 29 chemins. Les deux premières composantes de l'ACP rendent compte de 63 % de la variance des données. Les deux axes ont été redimensionnés en angström rms (divisé par $sqrt(3N)$ où N=26 395), ce qui nous donne une indication (partielle) de l'amplitude des mouvements principaux. Le premier mode de l'ACP correspond ici au mouvement de fermeture/ouverture du récepteur. Le RMSD (réel, tout atome) entre la première et la dernière structure du POE initiale est de 3,39 Å (ligne en pointillé rouge). À la fin de l'ensemble

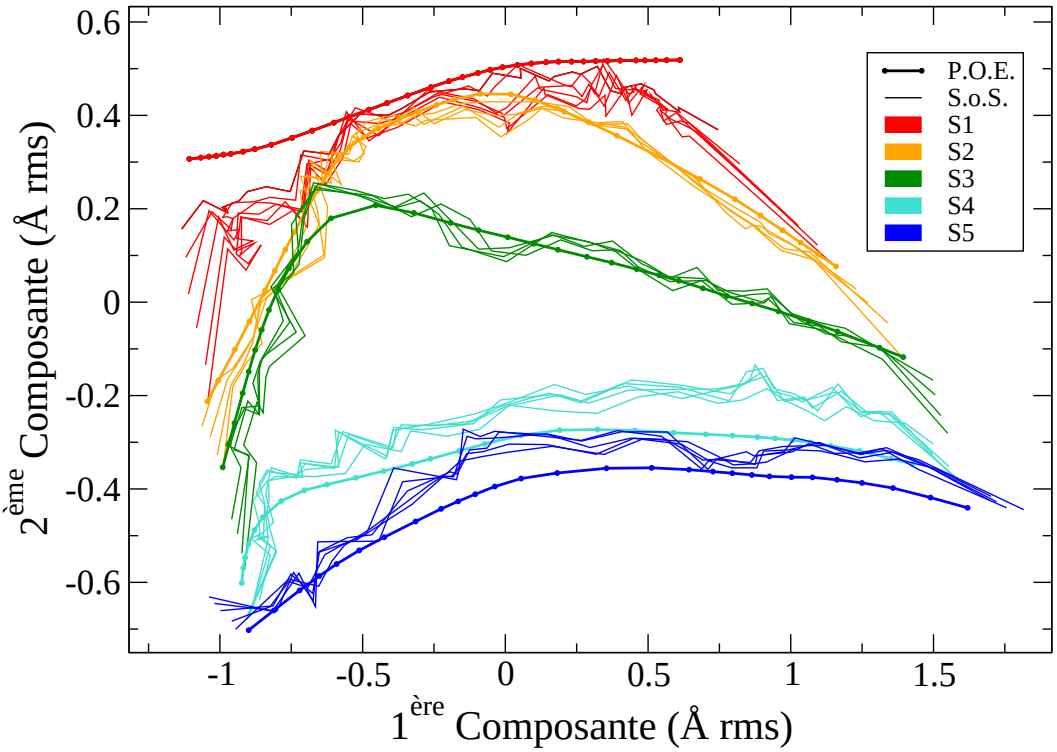


Figure III.6 Projection ACP des structures du récepteur nicotinique pour l'ensemble des 5 séries POE-SoS. Les chemins de transition individuels sont représentés par les courbes en pointillé (POE) et les traits fins (SoS). L'orientation de la première composante place les structures actives à gauche de la figure et les structures de repos à droite. Les deux composantes normalisées permettent d'avoir une estimation des distances entre les structures projetées (en angström).

des séries POE-SoS, cette distance passe à 4,83 Å, ce qui montre une élongation

du chemin de transition autorisée par la méthode des *String of Swarms* qui laisse évoluer les billes extrémales de la même manière que les structures intermédiaires. La seconde composante principale met en exergue une dérive des chemins dans l'espace des conformations. Le RMSD (tout atome) entre la première bille du POE de la Série S1 et le POE de la Série S5 est de 2,65 Å, et de 3,01 Å pour les deux dernières billes de POE S1 et S5. Ce mouvement de large amplitude, transverse à l'ouverture du récepteur, montre à quel point les chemins successifs s'éloignent du chemin initial dont les points extrémaux correspondent aux modèles construits dans la Partie II/p.21. Cette évolution sera étudiée dans la section suivante.

Il est intéressant d'observer la direction de diffusion des chemins Swarms succédant à une itération POE, en particulier sur la seconde composante. À la première itération (courbes rouges), les chemins SoS diffusent fortement vers le bas du graphe. Lors des 4^{e} et 5^{e} séries POE-SoS (courbes en turquoise et en bleu), le déplacement des points correspondant aux Swarms a tendance à aller dans le sens contraire, et les *String of Swarms* S5 à remonter en direction du POE S4. Bien que la projection ACP soit incomplète (elle efface une grande partie des mouvements subtils des structures), cette observation peut suggérer que l'espace conformationnel exploré par le couplage POE-SoS englobe ce mouvement de grande amplitude visible sur la seconde composante de l'ACP.

Graphiquement, on peut apprécier sur la projection PCA des différences de tracés entre les chemins POE (traits avec points) et les chemins issus des Strings of Swarms (en traits fins). Les chemins POE sont beaucoup plus "lisses" que les chemins SoS qui ont tendance à zigzaguer et s'entrecroiser. On observe d'ailleurs des raccourcis topologiques imposés par POE entre les SoS S2 et POE S3 et les SoS S3 et POE S4. Cette notion de "complexité" topologique des chemins est mesurée dans la Figure III.7/p.suiv.. La complexité d'un chemin est mesurée comme le ratio entre sa longueur curvilinéaire (la somme des distances RMSD entre structures consécutives) et sa longueur transverse (distance entre les structures extrémales). Un chemin trop complexe peut résulter de l'apparition de bruits Browniens (dynamique) ou mettre en valeur la rugosité du potentiel (statique). Dans tous les cas, cela ne reflète pas des mouvements indispensables à la réaction et pénalise l'analyse mécanistique de la transition. Comme attendu, les chemins POE ont une complexité relativement faible ($<$1,5). Lorsque les transitions sont thermalisées et simulées avec la méthode des *String of Swarms* on observe une explosion de cette complexité. Lors des itérations successives de SoS, le bruit causé par les mouvements browniens aléatoires s'accumule, en plus de la diffusion du chemin sur la surface d'énergie libre. À la fin de la série S1, la dernière transition SoS a une longueur curvilinéaire de 28 Å, pour une longueur transverse de seulement 3,6 Å. Il est intéressant de voir que lorsque le chemin est renormalisé par POE, sa complexité est fortement réduite, et cela pour chacune des séries. Au fil des itérations, les *String of Swarms* produisent

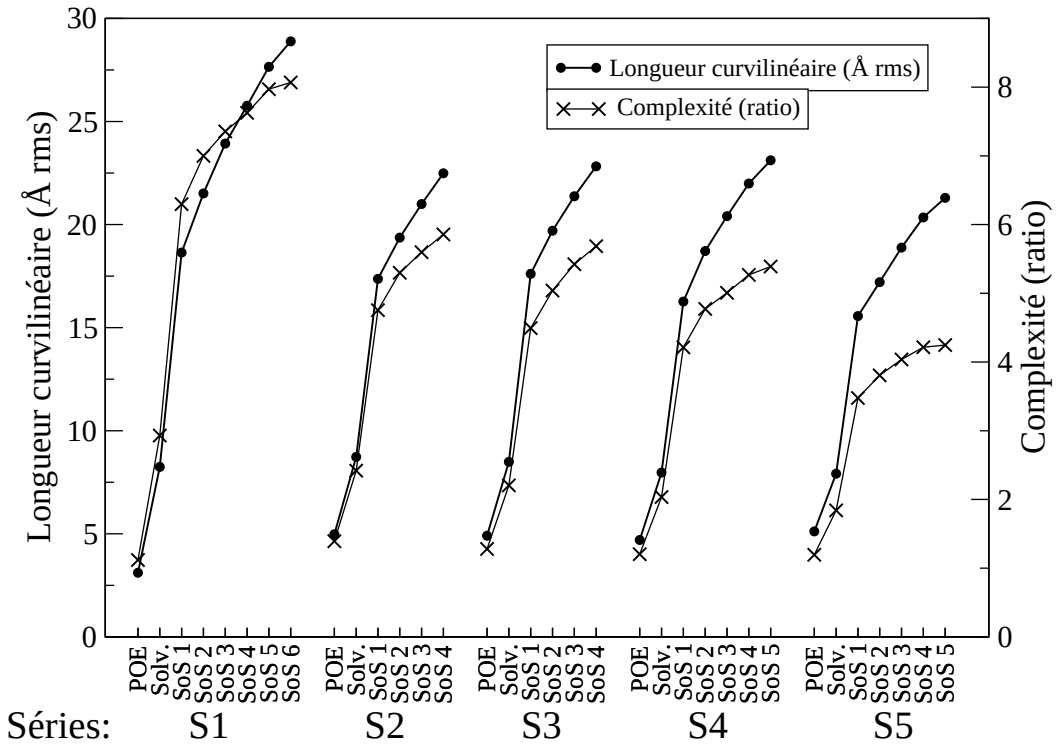


Figure III.7 Longueur curvilinéaire et complexité des chemins POE-SoS. La complexité des chemins après solvatation et équilibration (Solv.) est donnée à titre indicatif. La complexité est calculée comme le ratio entre la longueur curvilinéaire et la longueur transverse du chemin : $\left(\sum_{i=1}^{N-1} RMSD(X_i; X_{i+1})/RMSD(X_1,X_N)\right)$.

des transitions de moins en moins complexes. POE réussi à maintenir la complexité des chemins produits. La simplification topologique des chemins successifs se voit aussi sur le nombre de points nécessaires pour contourner les obstacles et décrire la transition qui décroît de 36 à 28 III.4/p.63.

3.2 Vraisemblance des chemins

3.2.1 Profils d'énergie libre

Une estimation des profils d'énergie libre le long des chemins de transition a été calculée à partir des dernières itérations de chacune des séries de Swarms et est reportée dans la Figure III.8/p.suiv.. Ils correspondent à l'intégration thermodynamique des forces moyennes exercées sur chacune des billes des chemins obtenus par de longues dynamiques moléculaires contraintes (1 ns pour chaque conformation). Ces profils sont donnés à titre indicatif et ne doivent pas être considérés comme une estimation robuste des différences d'énergie libre entre les états du récepteur. En particulier, l'indépendance des profils au nombre d'états de la transition ou encore l'échantillonnage suffisant lors des trajectoires n'ont pas été évalués. On notera aussi

que le choix de l'ensemble des coordonnées cartésiennes projetées sur la normale aux hyperplans comme variables collectives, elles-mêmes non invariantes par rotation et translation du système peut engendrer des artefacts lors du calcul des profils [132, 168]. Cependant, plusieurs points sont à souligner. Pour l'ensemble des séries l'amplitude des profils est tout à fait raisonnable (inférieures à 5 kcal/-mol sans compter les structures extrémales). Ils sont continus et lisses, ce qui nous

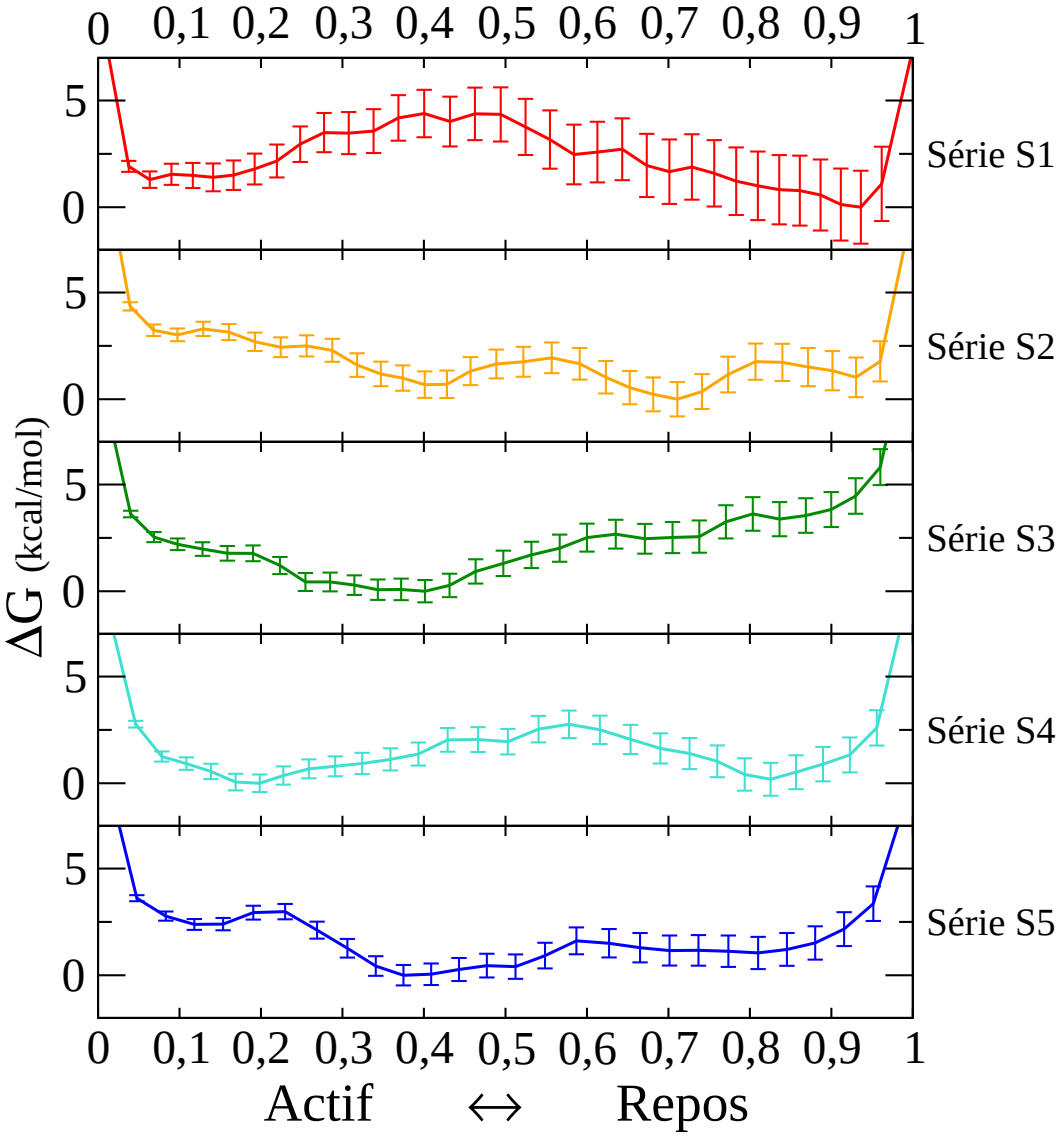


Figure III.8 Potentiel de force moyenne le long des chemins réactionnels. Dernières itérations de *Swarms*.

conforte dans l'idée qu'il n'y a pas de barrières énergétiques majeures le long des transitions calculées. Les barres d'erreur de bouts d'intégration s'avèrent diminuer au fil des séries ($\sigma_{S1} = 1{,}78$, $\sigma_{S1} = 0{,}96$, $\sigma_{S1} = 0{,}87$, $\sigma_{S1} = 0{,}85$, $\sigma_{S1} = 0{,}82$). Dans la série S4 (turquoise), il est intéressant de remarquer deux bassins énergétiquement favorables à 20 % et 80 % de la transition. Nous verrons que les structures associées

sont de bons candidats pour représenter les états actifs et de repos du récepteur nicotinique.

3.2.2 Solvatation du canal

Au fil des multiples itérations de Swarms plusieurs contrôles ont été systématiquement réalisés pour surveiller la stabilité des simulations le long des transitions, par exemple, pour contrôler l'intégrité des phospholipides dans la bicouche lipidique (Annexe : Figure VII.6/p.198). Une autre propriété primordiale est l'ouverture du canal. Notre adaptation des *String of Swarms* contraint les structures intermédiaires à évoluer dans un hyperplan bissecteur. Il est ainsi attendu que série après série de POE-SoS les chemins de transition présentent une ouverture du canal similaire en fonction de l'avancement de la transition. La Figure VII.3/p.196 donnée en annexe montre l'évolution du nombre de molécules d'eau dans le pore du récepteur le long du chemin de réaction et pour chacun des chemins arrivés à convergence de *Swarms*. Il est clair que pour les 5 séries de Swarms, le canal passe d'un état hydraté (entre 80 et 130 molécules d'eau) à un état déshydraté (entre 10 et 20 molécules d'eau). On observe aussi une tendance des états actifs à accueillir de plus en plus de molécules d'eau fil des séries (élargissement du pore), alors que les états de repos restent très similaires à partir de 80 % de la réaction. Ceci montre l'efficacité des hyperplans à contraindre les billes à un avancement réactionnel donné. Pour ce qui est de la transition S4, si l'on reprend les structures supposées stables à 20 % et à 80 % de la réaction (Figure III.8/p.préc.), elles correspondent à des états du récepteur dont le pore est pleinement ouvert et fermé.

La Figure III.9/p.suiv. donne plus de détails sur la constriction du pore, spécifiquement pour la dernière transition des *String of Swarms* de la série S4. L'image de gauche, nous montre l'hélice transmembranaire M2 des chaînes A et D, longeant le pore et correspondant à la structure visible à 80 % de la transition, c'est-à-dire une structure en état supposément fermé. Le pore, visible en cavité rouge, est clairement réduit par la Leucine 9', ce qui est en accord avec les données expérimentales de mutagenèse du récepteur nicotinique [170–172] relevant un rôle crucial de ce résidu sur la perméation du récepteur. Le diamètre du pore y est d'environ 1 Å (atomes d'hydrogènes inclus) suffisamment étroit pour bloquer des ions sodium et potassium. Les résidus de l'hélice M2 faisant face au canal sont les mêmes (à conservation de résidus près) que ceux identifiés dans les structures expérimentalement identifiées de récepteurs homologues ($\alpha4\beta2$ [34] et GluClR [44]). Les profils de diamètre du pore nous montrent que lorsque le récepteur est supposément dans des états actifs (courbes bleues) le canal est largement ouvert sur toute sa longueur (>5 Å).

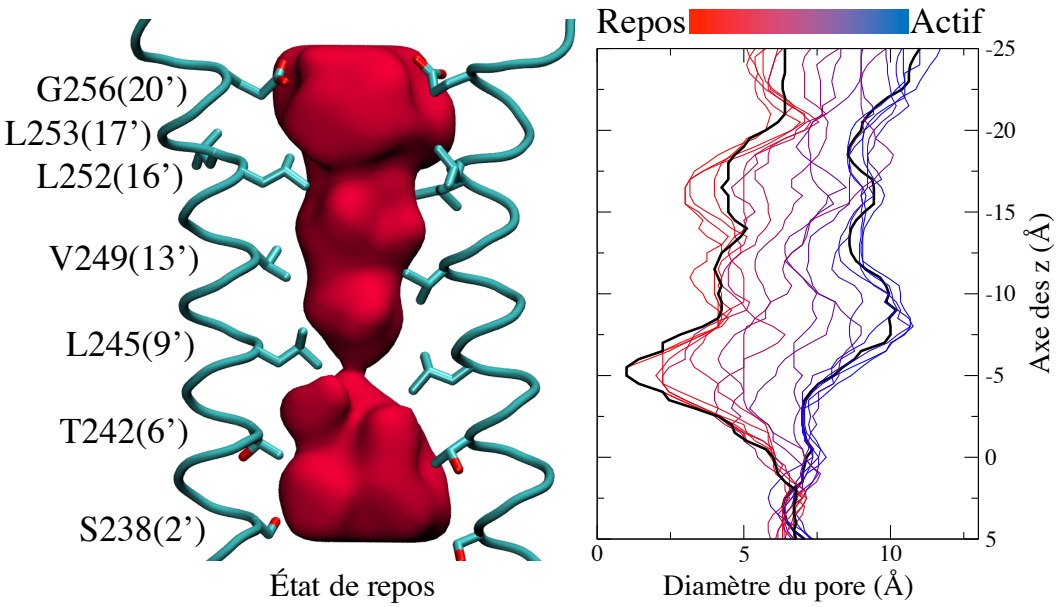


Figure III.9 Constriction du pore observée pour la dernière itération des Swarms de la série S4. Le volume du canal de la structure stable au repos (80 % de la réaction) est représenté en surface rouge à gauche. La numérotation des résidus est propre à nos modèles $\alpha 7$ et une numérotation consensuelle est donnée entre parenthèses (numérotation de l'hélice M2 de C.Miller [169]). À droite, les profils de diamètre du pore en fonction de l'axe des z, calculées avec mkgridXf, et tiennent compte du rayon de Van der Waals des atomes. Les structures stables repos et actifs (20 % et 80 % de la réaction) sont surlignées en noire.

3.2.3 Mouvements quaternaires

Depuis la détermination expérimentale de structures de canaux ioniques en états diverses, il est communément admis dans la littérature que deux mouvements de grande amplitude accompagnent l'ouverture du récepteur [105, 173]. Ces mouvements quaternaires correspondent à une contraction du domaine extracellulaire (ou inversement *blooming*, éclosion ou extension EC lors de la fermeture) induite par la fixation d'un ligand agoniste dans les sites orthostériques suivi par un mouvement de torsion (*twisting*) entre le domaine extracellulaire et transmembranaire. Il est intéressant de noter que ces mouvements ont spontanément été observés lors de longues simulations moléculaires : GLIC en état ouvert simulé à PH7 [50], le récepteur Glutamate après délétion de la molécule agoniste ivermectine [174] et le récepteur Glycine - GlyR$\alpha 1$ (travail de nos collaborateurs M.Ceccini et A.Cerdan). Nous avons analysé l'évolution de ces deux propriétés le long des chemins de transition obtenus par le couplage POE-SoS.

L'état d'extension du récepteur est mesuré grâce au rayon de giration des carbones α du domaine extracellulaire (définition empruntée à [142]). La Figure III.10/p.suiv. décrit ce rayon en angström, pour l'ensemble des chemins POE et des dernières itérations des séries de *Swarms*.

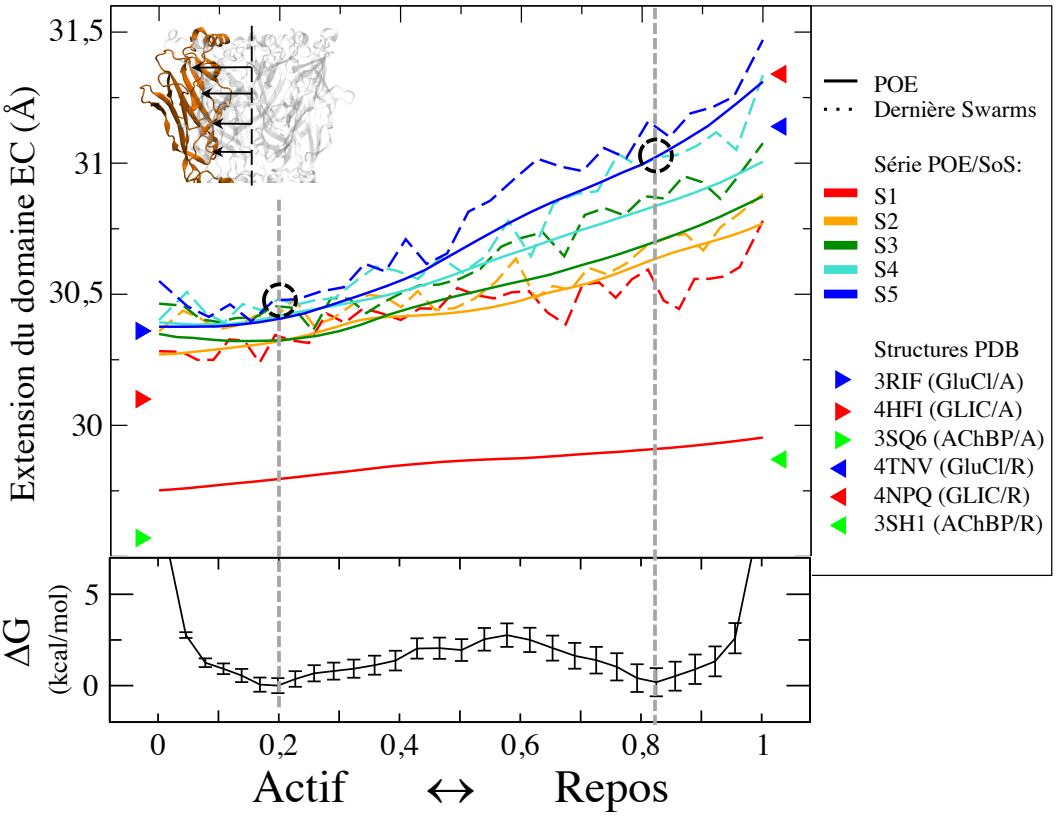


Figure III.10 Extension du domaine extracellulaire le long des modèles de transition. Les transitions POE sont représentées en traits pleins, les dernières itérations de SoS en traits pointillés. Les valeurs comparatives observées dans les structures cristallographiques de récepteurs homologues sont reportées en triangles. Les structures stables de la série S4 (20 % et 80 % de la réaction) sont pointées par des cercles pointillés noirs. Le profil d'énergie libre pour la transition S4 est aligné avec les profils d'extension.

Pour faciliter la comparaison avec les données expérimentales connues, nous y avons consigné les valeurs observées dans les structures cristallographiques de récepteurs supposés être en état actif et de repos : GLIC, GluCl et AChBP. Il est frappant de remarquer que les structures AChBP, 3SH1 et 3SQ6, ont des valeurs d'extension du domaine extracellulaire ainsi qu'une différence d'extension entre états actif/repos nettement inférieure à celles des structures GluCl et GLIC (0,3 Å contre 0,78 Å et 1,24 Å). Le fait que le mouvement de *blooming* ne soit pas observé dans les structures d'Acétylcholine Binding Protein (AChBP) est relevé et commenté dans plusieurs publications [175–177] (bien que des exceptions notables existent [106]). On rappellera ici que les AChBPs ont été utilisés comme *templates* homologues pour la construction des modèles de départ actif et de repos du récepteur $\alpha 7$. À la première itération de POE (trait plein rouge) la transition ne présente qu'une très faible différence d'extension ECD (0,20 Å entre la dernière et la première bille) et dont les valeurs sont étroitement proches de celles associées aux structures AChBPs. Dès la fin de la première série de Swarms (trait rouge pointillé) les valeurs d'extension du domaine EC augmentent significativement. L'extension de début de transition (état ouvert) se place entre les valeurs trouvées pour les structures

GluCl et GLIC en état ouvert. La différence entre état repos/actif est alors de 0,50 Å. Au fil des séries POE-SoS cette différence d'extension ne fait qu'augmenter pour se stabiliser autour 0,90 Å, en comparaison des 0,78 Å le GluCl et 1,24 Å pour GLIC. Une convergence de l'extension EC apparaît clairement être atteinte pour les structures actives, se stabilisant autour de 30,5 Å. Le profil d'énergie libre de la série de Swarms S4 est aligné le long du chemin réactionnel du graphe d'extension. Il est intéressant de remarquer que les structures correspondant à des bassins énergétiques favorables dans la transition Swarms S4 (courbe turquoise pointillée) s'alignent nettement avec les valeurs mesurées pour le récepteur GluCl (ronds pointillés noirs et triangles bleus). L'angle de torsion entre le domaine extracellulaire

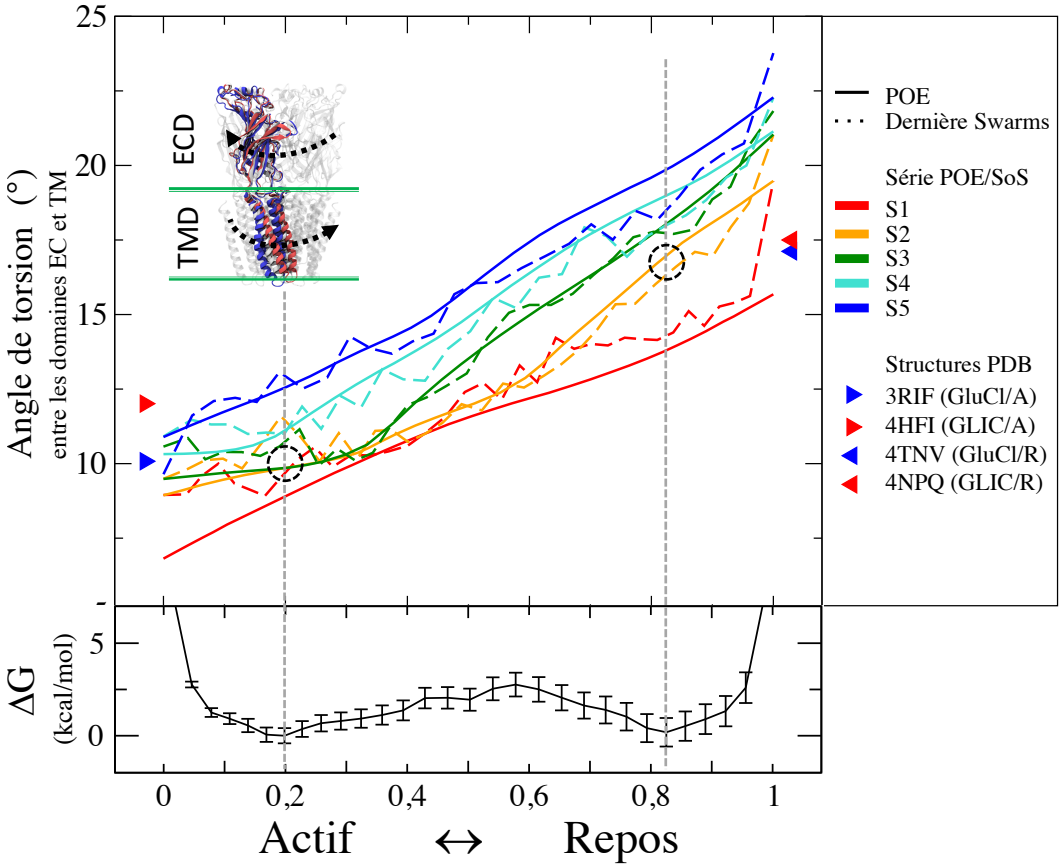


Figure III.11 Angle de torsion ECD-TMD le long des modèles de transition. Les transitions POE sont représentées en traits pleins, les dernières itérations de SoS en traits pointillés. Les valeurs comparatives observées dans les structures cristallographiques de récepteurs homologues sont reportées en triangles. Les structures stables de la série S4 (20 % et 80 % de la réaction) sont pointées par des cercles pointillés noirs. Le profil d'énergie libre pour la transition S4 est aligné avec les profils de torsion ECD-TMD.

du récepteur et sa partie transmembranaire est rapporté dans la Figure III.11. Comme précédemment, les valeurs trouvées dans les structures homologues sont indiquées par des triangles rouges (GLIC) et bleus (GluCl). Les AChBP n'exprimant pas de partie transmembranaire, la torsion ECD-TMD n'y est pas mesurée. De la même façon que la propriété d'extension ECD, les structures du chemin initial POE

montrent une torsion légèrement réduite par rapport aux modèles cristallographiques homologues. Les itérations successives de séries POE-SoS accentuent ce mouvement de torsion. Elle se stabilise entre les valeurs de GLIC actif et GluCl actif pour le début de la transition. En fin de transition (vers les états de repos) les structures ont d'ailleurs tendance à sur-pivoter par rapport aux angles identifiés sur GLIC/repos et GluCl/repos. Cependant, en reprenant les structures minimisant le profil d'énergie des Swarms S4 on retrouve des valeurs comparables : 11,91 et 17,64 contre 10,08 et 17,12 pour GluCl (actif/repos) et 12,01 et 17,50 pour GLIC (actif/repos). Pour cette propriété de torsion, les chemins de transition successifs laissent apparaître une convergence : la courbe des derniers chemins SoS bascule sous la courbe de la transition du POE précédent.

3.3 Convergence des *String of Swarms*

La méthode des *String of Swarms* comprend un ensemble de 4 étapes (diffusion des Swarms, moyenne des Swarms, reparamétrisation et propagation) qui doivent être cycliquement réitérées jusqu'à la convergence des chemins produits. En d'autres termes, les transitions glissent au fond d'une vallée d'énergie minimum et, mis à part les mouvements browniens intrinsèques à la dynamique moléculaire, les structures ne doivent plus évoluer significativement. Pour vérifier la convergence des 5 séries de Swarms nous nous sommes basés sur des critères énergétiques (contraintes de propagation) et géométrique (distance RMSD entre les structures successives).

3.3.1 Critère énergétique

Une façon d'évaluer la convergence consiste à étudier l'énergie de la contrainte de propagation $C_{propa.}$ utilisée lors de l'étape de même nom. Cette contrainte guide la trajectoire de propagation vers la structure cible, incarnation géométrique de la position de l'espace conformationnel vers laquelle a diffusé l'essaim de dynamique des Swarms (et projeté sur l'hyperplan bissecteur). Elle est la somme d'une contrainte harmonique sur la cible et de la contrainte C_{HP} sur l'hyperplan bissecteur. La Figure III.12/p.suiv. résume pour chacune des itérations de *String of Swarms* et l'ensemble des structures intermédiaires, l'énergie de la contrainte $C_{propa.}$ entre la structure courante et la bille cible au début (t=0 ps), au milieu (t=2 ps) et à la fin (=4 ps) de la trajectoire de propagation. Au départ de la trajectoire de propagation, la structure courante correspond à la bille du chemin de transition précédent. Par construction cette structure est incluse dans l'hyperplan bissecteur. Ainsi à t=0 on a $C_{HP} = 0$ et la contrainte $C_{propa.}$ reflète à quel point la bille initiale est éloignée de la bille cible. Les énergies de la contrainte à t=0 sont visibles dans la Figure III.12/p.suiv. en courbes colorées. Lors de la première série SoS (S1, courbe

rouge) et à la première itération, la totalité des 33 structures ont une énergie très grande (>2 000 kcal/mol). Il faut au moins 3 itérations de SoS dans la série S1 pour que ces contraintes atteignent un plateau énergétique autour de 500 kcal/mol. Dans les séries suivantes il ne faut qu'une seule itération pour arriver à ces valeurs d'énergie. Dès 2 picosecondes de simulation la contrainte hyperplane atteint ce qui

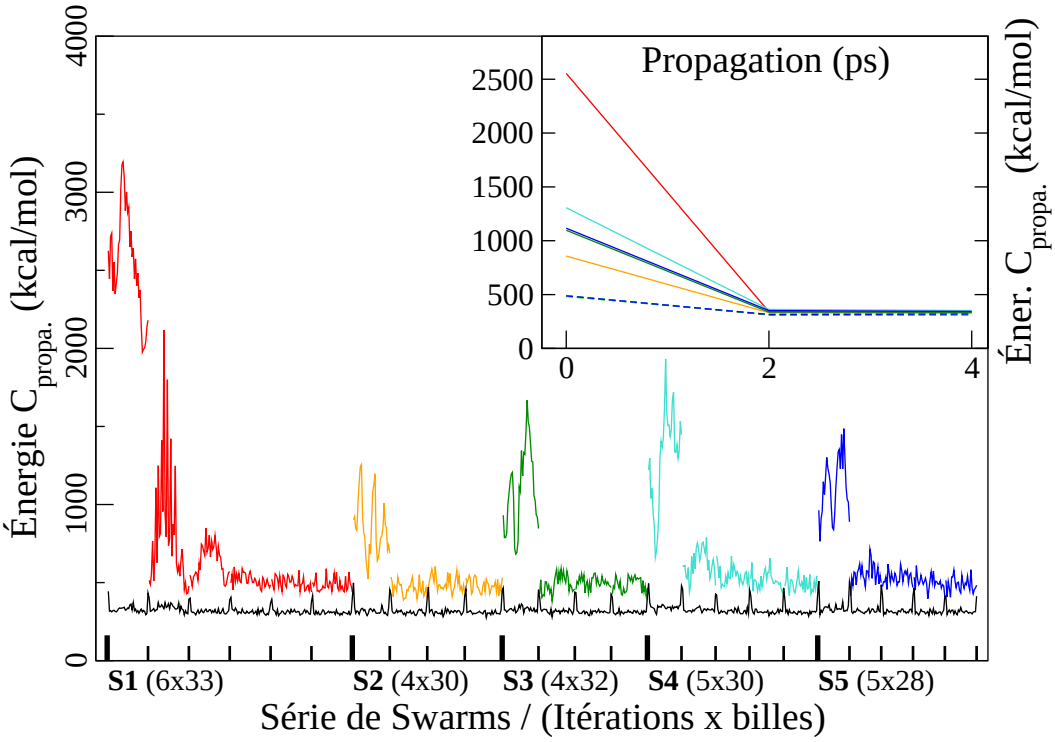


Figure III.12 Évolution de la contrainte de propagation. Le graphe global spécifie pour chacune des billes des chemins SoS, l'énergie de la contrainte à t=0 ps (en couleur) et t=4 ps (en noir). L'encart en haut à droite renseigne la moyenne des énergies sur toutes les billes d'une itération SoS à t=0 ps, t=2 ps et t=4 ps, pour la première (traits plein) et la dernière (traits pointillés) itération de chaque série de Swarms (différentes couleurs).

semble être sa valeur énergétique minimale (environ 300 kcal/mol). Ainsi, l'étape de propagation converge rapidement vers la cible. Cette valeur résiduelle (elle ne tombe jamais à 0, la trajectoire n'atteint jamais *exactement* la structure cible, indépendamment du temps de simulation) est exemplifiée dans l'Annexe VII.4/p.197. Elle est causée par les fluctuations aléatoires et les vitesses instantanées causées par la thermalisation du système qui interdisent de ramener le système à un point unique et couvrent le volume d'une petite hypersphere. Il apparaît ainsi qu'un nombre d'itérations de Swarms entre 4-5 est satisfaisant pour atteindre la convergence des chemins à chacune des séries des *Swarms*.

3.3.2 Critère géométrique

La Figure III.13 nous donne un ordre de grandeur de l'éloignement successif des chemins relaxés par une nouvelle itération de *String of Swarms* par rapport à la transition POE de la série POE-SoS correspondante. Les remarques sont ici très similaires à celles relevées par le critère énergétique (qui capture d'ailleurs indirectement la même donnée).

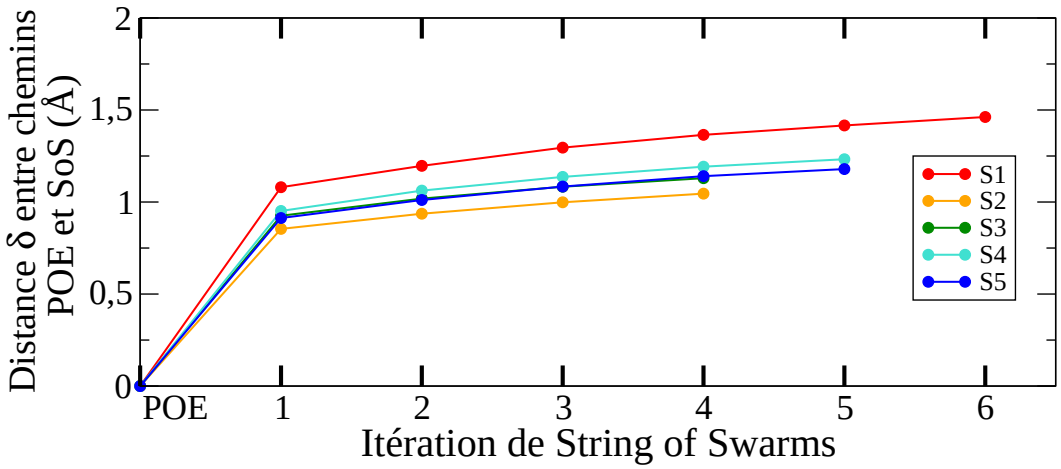


Figure III.13 Convergence géométrique des itérations de *String of Swarms*. Distance δ (voir Matériels et Méthodes) entre le chemin POE et chacune des itérations de *String of Swarms* pour chaque série POE-SoS. Note : la trace du chemin S3 est très proche, mais distincte, de celle du chemin S5.

Pour les 5 séries la plus grosse partie de l'évolution géométrique des chemins se retrouve dans la première itération de SoS, avec des distances à la transition POE comprises entre 0,85 Å et 1,08 Å. Les transitions créées par les itérations suivantes diffusent de manière beaucoup plus restreinte. Entre la dernière et l'avant dernière itération SoS les chemins n'évoluent quasiment plus par rapport au chemin POE (différence strictement inférieure à 0,05 Å pour les 5 séries). Sur ce critère géométrique la convergence des chemins SoS apparaît satisfaisante pour l'ensemble des séries calculées.

3.4 Compatibilité POE-SoS

Dans cette section nous évaluons comment les transitions issues d'une optimisation adiabatique (POE) arrivent à s'adapter au contexte d'énergie libre des *String of Swarms* et vice-versa.

3.4.1 Passage POE vers SoS

La première itération de chacune des séries de Swarms correspond à la relaxation dans le paysage d'énergie libre de structures optimisées par POE (après une brève équilibration - sous contraintes - des structures). La Figure III.12/p.75 nous montre une forte décroissance des énergies de contrainte de propagation entre les séries S1 et S2 qui pourrait signifier que la transition POE/S2 est plus proche d'une vallée d'énergie libre que la série S1. Lors des séries suivantes, S3, S4, S5, les énergies de contrainte de propagation à la première itération restent élevées (entre 500 et 2 000 kcal/mol). Il est intéressant de rappeler ici que les différentes séries de *String of Swarms* se situent dans des régions différentes de l'espace conformationnel, comme élicité par la dérive des chemins visibles dans la projection PCA (Figure III.6/p.66). Il y a donc un coût non négligeable au passage d'un chemin de POE vers les *String of Swarms*, mais ce coût est amorti par la convergence somme toute rapide des *Swarms*.

3.4.2 Passage SoS vers POE

Après leur relaxation par les *String of Swarms* les chemins de transition sont désolvatés puis réinjectés à la méthode POE. POE cherche alors des raccourcis topologiques dans la transition, les réassemble et les optimise par itérations successives de l'algorithme CPR. Le Tableau III.3/p.suiv. nous montre les énergies maximales, la longueur et le nombre de structures intermédiaires de ces chemins optimisés. La colonne "Chemin initial" correspond à l'optimisation des chemins directement issus de l'itération précédente, c'est à dire soit des *String of Swarms*, soit d'une optimisation préalable de POE. En l'occurrence, a7m3, a7m6, a7m9 et a7m12 raffinent des transitions sorties des séries S1, S2, S3 et S4 des *String of Swarms*.

L'énergie des structures les moins favorables du chemin (Ener. Max) y est irréalistement grande (>1E10 kcal/mol). Pour comparaison, les chemins optimisés à partir d'une transition POE ont des énergies maximales plus raisonnables ($E_{\max}$(a7m4)=3 146, $E_{\max}$(a7m7)=3 125, $E_{\max}$(a7m10)=2 933, excepté a7m13 qui a été traité spécifiquement). Deux raisons peuvent expliquer la mauvaise appréciation par POE des chemins directement issus des *Swarms*. Premièrement, les contextes d'optimisation sont particulièrement différents entre SoS et POE. En particulier, lors des simulations de dynamique moléculaire thermalisées, les degrés de liberté structuraux captent approximativement $\frac{1}{2}kT$ chacun (approximation car le potentiel est non harmonique). Deuxièmement, durant les multiples itérations des *String of Swarms*, les structures ont évolué dans l'espace conformationnel. Or, les chemins de transition calculés par SoS ne garantissent pas la continuité entre les billes du chemin (il n'y a pas de barrières énergétiques séparant les structures intermédiaires).

Nom	**Chemin initial***		#Chemins	**Chemin sélectionné**			**Chemin réduit**		
	E_{max}	Long.	alternatifs	E_{max}	Long.	#pts	$E_{plat.}$	Long.	#pts
a7m1	2 961,1	6,1	89	2 722,4	5,9	288	4 500,0	5,3	36
				SoS 1					
a7m3	5E+15	46,0	508	3 347,8	9,1	311	7 347,8	8,1	35
a7m4	3 145,7	4,7	930	3 042,1	8,9	342	7 042,1	7,7	33
				SoS 2					
a7m6	2E+11	39,8	25	3 432,9	9,3	290	7 432,9	8,1	33
a7m7	3 125,0	9,3	533	3 038,5	8,6	278	7 038,5	7,6	32
				SoS 3					
a7m9	1E+16	42,0	3	3 254,6	10,9	283	7 254,6	5,5	35
a7m10	2 932,9	11,1	94	2 789,6	10,3	290	6 789,7	7,1	30
				SoS 4					
a7m12	7E+22	40,3	186	2 927,3	9,7	241	6 927,3	7,8	33
a7m13	8E+20	366,8	75	2 340,6	8,4	215	6 340,6	7,49	28
				SoS 5					

Tableau III.3 Résumé des chemins analysés et optimisés par POE. Les chemins initiaux, passés en entrée de POE, sont comparés avec les chemins alternatifs sélectionnés et raffinés par POE. Pour comparaison, les chemins initiaux sont optimisés de la même façon que les chemins sélectionnés (15 itérations “Hammer Drill”). Les chemins réduits correspondent aux chemins sélectionnés nettoyés des structures intermédiaires non indispensables pour former un chemin de transition dont l’énergie maximale est inférieure au plateau énergétique ($E_{plat.}$). Les Énergies Maximales (E_{max}) sont en kcal/mol. Les Longueurs (“Long.”) sont données en angström rms. #pts est le nombre de structures intermédiaires du chemin. En pratique, 2 itérations de POE sont intercalées entre chacune des séries de *Swarms*. L’énergie maximale du chemin a7m13 est irrégulière et due à une correction ponctuelle du chemin décrite en Matériels et Méthodes.

Cette moins bonne continuité contrarie l'optimisation de CPR. Les chemins de transition alternatifs qui exploitent les raccourcis contournent ces barrières énergétiques, même si l'utilisation des hyperplans devrait limiter cette tendance, et permet de récupérer la transition. C'est ce que fait POE en testant et en raffinant un ensemble de chemins alternatifs qui empruntent un ou plusieurs raccourcis topologiques (colonne "# Chemins alternatifs"). Pour l'ensemble de ces cas pathologiques, POE trouve (colonne "Chemin sélectionné") un chemin de transition dont l'énergie maximale est inférieure à 3 500 kcal/mol, et dont la longueur curvilinéaire est beaucoup plus courte que le chemin initial. POE s'adapte de façon efficace aux chemins de transition trouvés par les *String of Swarms* grâce à la présence de la recherche de raccourcis.

4 Discussion

4.1 Couplage POE-SoS

Ce travail a révélé l'émergence de deux mouvements quaternaires distincts lors de la modélisation du mécanisme d'activation/désactivation du récepteur nicotinique de sous-unités $\alpha 7$: l'apparition spontanée au fil des transitions d'une extension du domaine extracellulaire accompagné de l'accentuation d'un mouvement de torsion entre les domaines extracellulaire et transmembranaire, en accord avec les structures connues de récepteurs homologues. Plusieurs éléments nous indiquent qu'un tel résultat n'aurait pas été observé à partir des données structurales de départ sans un couplage des deux méthodes *Path Optimization and Exploration* et *String of Swarms*. Les *String of Swarms* convergent pour chacune des itérations POE-SoS comme décrit dans la Section 3.3/p.74 et se bloquent, probablement, dans des minima énergétiques du paysage d'énergie libre. Un plus grand nombre d'itérations dans la première série SoS ne garantirait donc pas une exploration suffisante de l'espace conformationnel permettant d'atteindre les transitions plus "vraisemblables" (Section 3.2/p.68) des derniers cycles POE-SoS. La méthode POE s'est ici montrée pertinente pour régulariser les transitions hautement complexifiées par le bruit conformationnel généré lors des dynamiques moléculaires des *String of Swarms* (Figure III.7/p.68). De façon incidente, la recherche et l'insertion de raccourcis topologiques dans les transitions POE permettent aux *String of Swarms* d'explorer de nouvelles régions de l'espace des conformations (Figure III.6/p.66).

4.2 Convergence du couplage

Au-delà de la complémentarité des deux méthodes, la question de la convergence des chemins itérativement construits par le couplage POE-SoS reste difficile à établir. Idéalement, une convergence claire aurait été démontrée si la transition POE de la série S5 se confondait avec celle de la série précédente S4. Certains éléments sont cependant en faveur d'un accord entre les dernières séries. Selon la projection ACP (Figure III.6/p.66), les trajectoires de *Swarms* de l'itération S5 ont tendance à diffuser vers des conformations déjà explorées par la transition POE de l'itération S4. De plus les structures extrémales, dont la divergence est forte en début de cycles POE-SoS apparaissent converger autour des deux premiers modes de la PCA : (-1 ; -0,5) pour les états canal ouvert et (1,7 ; -0,4) pour les états canal fermé. Nous remarquerons tout de même que la projection ACP ne résume que les mouvements principaux

des transitions. En termes de distance δ entre chemins, le dernier chemin SoS de la série S5 est plus éloigné du POE-S4 (δ = 1,6 Å) que le dernier SoS-S4 (δ = 1,2 Å) et le POE-S5 (δ = 1,3 Å). Les chemins POE et SoS se stabilisent pour la propriété d'extension (Figure III.10/p.72) dans leur état canal ouvert, et sur l'ensemble de la transition pour la propriété de torsion ECD/TMD (Figure III.11/p.73) pour laquelle les Swarms S5 diffusent vers des structures aux valeurs inférieures à celles trouvées dans la transition POE-S5 et plus proches de celles du POE-S4 (courbes bleues et turquoises).

4.2.1 Temps de simulation des trajectoires de *Swarms* et de propagation

Il est intéressant d'observer que les *String of Swarms* convergent significativement plus rapidement que les implémentations décrites dans la littérature. La convergence des chemins est atteinte en 4 à 6 itérations alors qu'une centaine est requise pour la transition du domaine allostérique $NtrC^r$ [111] et la Kinase Src [110], et près de 400 pour le récepteur GLIC [142]. Ce travail nous a permis d'expérimenter des trajectoires de *Swarms* non contraintes plus longues (20 ps par trajectoire), et des temps de propagation plus courts (4 ps) que ce qui est habituellement réalisé. Pour comparaison, l'implémentation originale de Pan et al. [111] génère des dynamiques de 500fs pour les trajectoires de *Swarms* et de 125 ps pour la propagation. Des données expérimentales montrent que les temps de relaxation de protéines globulaires en solution sont de l'ordre de $\approx$ 4,5 ps [178]. Cela suggère qu'une durée de simulation plus longue est préférable pour que les *Swarms* diffusent convenablement. L'analyse de ces simulations, montrée dans l'Annexe VII.5/p.198, nous a orientés sur un choix de 20 ps, satisfaisant pour atteindre une stabilité énergétique moyenne des trajectoires. Concernant l'étape de propagation, l'analyse géométrique de l'espace des conformations visitées (Figure VII.4/p.197) nous indique qu'une dynamique contrainte de 4 ps est suffisante pour arriver au plus près de la structure cible.

4.3 Choix de la transition représentative de l'activation du récepteur

Plusieurs arguments nous font privilégier la dernière itération de la 4^e série de *String of Swarms* pour modéliser le mécanisme d'activation/désactivation du récepteur nicotinique. Le profil d'énergie libre de la série S4, bien que probablement approximatif, affiche deux états stables, séparés par une barrière énergétique raisonnable d'environ 2,5 kcal/mol. Le pore de ces deux états est convenablement solvaté/désol-

vaté et présente un diamètre d'ouverture compatible avec le passage (ou non) de cations. Leur structure quaternaire est aussi très proche de celle observée chez des récepteurs homologues (GluCl, GLIC). Il est difficile d'apporter plus de précisions quant au mécanisme d'activation du récepteur nicotinique de sous-unité $\alpha 7$ à partir de cette transition. En particulier, l'utilisation de structures issues d'une modélisation comparative pour initialiser la transition fait obstacle à l'analyse du récepteur au niveau atomique. Il est par exemple très compliqué d'évaluer si une interaction atomique non présente dans les structures homologues est un phénomène spécifique au mécanisme du récepteur $\alpha 7$ ou un artefact de modélisation/simulation. À l'inverse, un motif atomique en correspondance avec l'un des récepteurs *templates* est très probablement directement hérité de la modélisation comparative. La pertinence à l'échelle atomique des modèles de la transition S4 sera sondée dans la partie suivante, avec pour objectif de tester la propension du récepteur à s'apparier avec de petites molécules connues pour moduler sa fonction.

5 Conclusions et perspectives

Le couplage entre l'approche *Path Optimization and Exploration* (POE) et la méthode des *String of Swarms* (SoS) a permis de construire un modèle d'activation du récepteur nicotinique $\alpha 7$ dont les propriétés dynamiques (*blooming*, *twisting*, diamètre et constriction du pore) sont similaires à celles observées dans les structures expérimentales de récepteurs homologues. La difficulté principale s'est trouvée dans la régularisation de modèles initiaux imparfaits. L'alternance d'itérations POE et SoS a rendu possible l'exploration d'un espace conformationnel large et la relaxation des chemins de transition vers des états plus vraisemblables. Malgré la haute dimensionalité de la protéine et les bruits browniens aléatoires qui agitent les simulations, la méthode POE a su être efficace pour nettoyer les trajectoires issues des *String of Swarms*. La simplicité des dynamiques est une caractéristique qui simplifie le suivi dynamique des cavités dans les trajectoires de protéines, comme nous le verrons dans la Partie IV/p.85.

La robustesse du couplage a pu être testée sur un système de très grande taille (>200 000 atomes). Bien que les temps de calculs restent encore conséquents, l'automatisation des procédures et la parallélisation des calculs rendent le couplage raisonnable. Un paquet comprenant l'ensemble des scripts utilisés pour réaliser les différents cycles de POE et des SoS sera mis à la disposition de la communauté.

Le chemin de transition de la dernière itération des *String of Swarms* de la série S4 sera brièvement utilisé au sein de la Partie IV/p.85 dans laquelle nous appliquerons des méthodes d'identification dynamique de sites de protéines pour cartographier de potentiels sites effecteurs du récepteur nicotinique. Dans la Partie V/p.131 la même transition sera le support pour l'amarrage moléculaire différentiel de petites molécules sur l'ensemble des sites et conformations du récepteur.

Les méthodes développées seraient facilement transposables à des récepteurs nicotiniques de sous-unités différentes, voire à des récepteurs homologues. Du fait de la diversité des états structuraux des récepteurs canaux, il serait intéressant d'étudier des transitions collatérales, par exemple le mécanisme de désensibilisation, à partir de structures cristallographiques de GLIC, en état actif (A), de repos (R) et désensibilisé (D). La combinatoire des chemins de transition (A-R, A-D, D-R) pourrait laisser transparaître des intersections topologiques ainsi que des états intermédiaires connus (p.ex. localement fermé [46]) ou encore inconnus.

IV

Suivi des cavités dans les trajectoires de protéines et détection de sites

Dans de nombreux cas, les protéines expriment leur fonction grâce au mouvement concerté de leurs atomes. Qu'il soit subtil ou de grande amplitude, ce déplacement stérique des atomes a un impact sur les "espaces libres", ou cavités, autour et à l'intérieur de la protéine, potentiels sites de liaison pour de petites molécules. Nous proposons dans cette partie une méthode de suivi des cavités et de détection de sites dans le contexte dynamique des trajectoires de protéines (p.ex. dynamique moléculaire, chemins de transition, etc.). La méthode, sa mise au point et son implémentation, *mkgridXf*, sont décrites. Pour valider l'approche, un ensemble de 15 sites choisis sur 4 protéines types ont servi de données de référence : la myoglobine du cachalot, le dimère de la protéine d'enveloppe du virus de la dengue, le facteur œdématogène de la toxine de l'anthrax (EF) et la tyrosine kinase Abl (abl1). Étudiés en dynamique moléculaire, ces sites se sont montrés particulièrement mobiles et changeants. Malgré cela, notre algorithme s'est montré robuste pour suivre et caractériser leur évolution au cours de la trajectoire. Finalement, *mkgridXf* a été utilisé pour cartographier l'ensemble des sites de la transition d'activation du récepteur nicotinique, rendant possible l'analyse des mécanismes allostériques décrite dans la partie V/p.131.

1 Introduction

La fonction d'une protéine est étroitement liée aux interactions qu'elle peut former avec ses substrats et/ou partenaires [179]. Sa conformation, c'est-à-dire l'arrangement spatial de ses atomes, est essentielle pour établir ces interactions. Les cavités, espaces accessibles au solvant entre les atomes de la protéine, forment une interface privilégiée entre le milieu extérieur et les résidus de la protéine. L'explosion du nombre de structures expérimentales, parfois co-cristallisées avec leur substrat, a rapidement laissé transparaître l'ubiquité des cavités dans les conformations de protéines. Les formes des sites de liaison sont diverses (globulaires, ramifiées, petites, volumineuses) et réparties sur l'ensemble du volume des protéines (parfois enfouis, ou en surfaces) [180, 181]. La forme et la taille des cavités influencent la sélectivité et la spécificité des ligands pour leur cible [182] et sont maintenant des descripteurs communément utilisés lors de la recherche et l'optimisation *in silico* de nouvelles molécules effectrices [150, 182–185].

Parallèlement, l'identification des sites fonctionnels est cruciale pour comprendre les mécanismes d'action de la protéine cible, évaluer les interactions moléculaires possibles et trouver de nouvelles stratégies de modulation [186, 187]. De tels sites sont classiquement divisés en deux catégories : orthostérique ou allostérique. Le site orthostérique, ou site primaire, correspond à la région de la protéine dans laquelle se lient un ou plusieurs ligands endogènes. Par exemple, le site catalytique des enzymes. Ces lieux privilégiés d'interaction sont relativement bien caractérisés par la forme de leur cavité (souvent parmi les plus volumineuses [180]), leur empreinte sur la séquence (conservation des résidus impliqués) et la signature des groupes de résidus nécessaires à la reconnaissance ou à la catalyse de ligand [188–190]. À l'inverse, les sites allostériques, se trouvent à des endroits différents du site actif et la fixation d'un ligand, souvent par le biais d'un changement conformationnel, altère la fonction [191–193]. Les avantages de ces sites modulateurs sont multiples. Une pression évolutive plus faible rend l'interface de ces sites plus diverse que les sites actifs [194], ce qui facilite le dessin de molécules spécifiques et limite les effets de bords sur d'autres protéines de l'organisme [195–197]. Certains modulateurs allostériques ne modifient pas la concentration du ligand endogène, ce qui peut limiter les risques de toxicité par surstimulation. Cependant la recherche de sites allostériques ne peut s'appuyer sur les méthodes utilisées pour les sites orthostériques [198] et représente un domaine d'actives recherches [199, 200].

À cause des mouvements browniens et/ou des libérations transitoires d'énergie, la conformation de la protéine fluctue et évolue au cours du temps. Ces évolutions sont

essentielles pour la fonction. Les enzymes réalisent la catalyse de substrats grâce à des changements subtils dans leurs modes de vibration [201, 202] et des déplacements de résidus sont souvent observés dans les sites actifs [203]. Des mouvements de "respiration" facilitent la pénétration de l'oxygène dans les globines (myoglobine, hémoglobine...) [9, 204–210]. D'autres protéines expriment leur fonction au travers de larges changements conformationnels, pour lesquels des domaines entiers peuvent se déplacer les uns par rapport aux autres (protéine d'enveloppe de la dengue [211], Facteur Œdémateux de l'anthrax [150, 212] et beaucoup d'autres [213]).

Une publication récente au laboratoire [9] a montré le lien direct entre les mouvements fonctionnels des protéines et l'évolution associée des cavités. Dans ce cas, il est pertinent de pouvoir évaluer l'impact d'un changement conformationnel sur un site en particulier (changement du volume de la cavité, modification des interactions). Pour cela, une délimitation précise du site est nécessaire pour mesurer la corrélation entre la géométrie de la cavité et la transconformation au cours des différents états fonctionnels. L'analyse dynamique des cavités peut aussi dévoiler des cavités transitoires, absentes des structures cristallographiques, mais qui peuvent se révéler importante pour la fonction de la protéine [214]. La géométrie des cavités a un impact crucial sur la liaison des ligands [180, 215, 216] et les techniques d'amarrage moléculaire et criblages virtuelles sur des conformations multiples de récepteurs sont activement développées pour améliorer les résultats [217–219].

Les approches computationnelles de détection de sites et de calcul de cavités ont longtemps été restreintes à l'analyse de systèmes statiques (structures X-ray, Cryo-EM, RMN...). La démocratisation des méthodes de simulation moléculaire facilite l'acquisition de longues trajectoires de conformations vraisemblables de protéines. Cependant, à cause de sa difficulté intrinsèque liée à ses ambiguïtés, l'analyse dynamique des cavités est le plus souvent réalisée sur quelques structures, ou appliquée à des locus prédéfinis (voir [220, 221] pour une revue récente des outils de détection de cavités). Par conséquent, des méthodes cohérentes, avec une applicabilité large, sont nécessaires.

1.1 Nature ambiguë de la définition de sites

Il y a une ambiguïté intrinsèque à la définition d'un site. Les sites sont classiquement définis par la poche entourant un ligand. Cependant beaucoup d'exemples existent de sites pouvant lier des molécules de tailles et de formes diverses dans un même site [222], ce qui rend difficile la définition de la poche. Cette difficulté est d'autant plus fondamentale lorsque aucun ligand n'est connu ou que la structure du complexe ligand/protéine n'est pas déterminée [223–225].

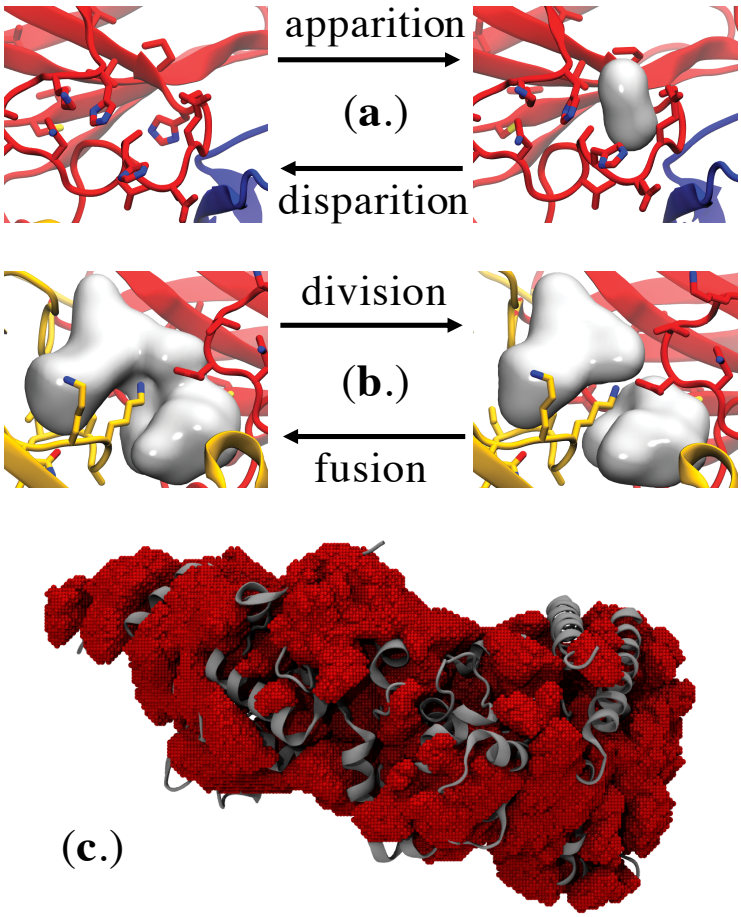


Figure IV.1 Difficultés d'attribution d'une définition consensus pour un site.. **a.** Apparition ou disparition d'une cavité lorsque sa géométrie atteint un seuil, par exemple, un volume minimum ou le solvant extérieur (*bulk solvant*). **b.** Fusion ou division de cavités induite par le mouvement subtil d'un acide aminé. **c.** Domaine de définition des cavités. Les cavités apparaissent labiles et omniprésentes dans les dynamiques de protéines.

De plus, les sites, en particulier allostériques ou cryptiques ont une nature intrinsèquement dynamique. Dans les systèmes dynamiques, les conformations peuvent largement varier, et l'identification des sites de liaisons devient encore plus intriquée et ambiguë à cause de phénomènes d'apparition et disparition (Figure IV.1 - **a.**), de fusion et de division (Figure IV.1 - **b.**). De larges transconformations de la protéine peuvent aussi faire apparaître de nouvelles cavités ou en découper d'autres.

La grande mobilité des cavités crée de nouvelles difficultés. Elles peuvent apparaître presque partout dans le volume de la protéine au cours de l'évolution de la dynamique moléculaire (Figure IV.1 **c.**). À cause des mouvements de la protéine, l'identification va probablement échouer si elle n'est basée que sur un chevauchement des cavités d'une conformation à une autre [226].

Par conséquent, les larges changements de conformation apparaissent être une source évidente d'échecs, indépendamment des descripteurs utilisés pour représenter la géométrie des cavités, cubes, sphères, cylindres ou autre.

Le nombre total de cavités observées au cours de longues trajectoires peut être grand, plusieurs millions, bien au-delà de ce qui est manuellement analysable. Cela obscurcit le processus d'identification des cavités ou des sites, et suivre une cavité au travers de multiples conformations devient un vrai défi.

1.2 Identification des sites dans les trajectoires

Pour surmonter le manque de référence spatiale globale, défavorisant les approches basées sur la localisation géométrique des cavités, nous avons considéré les groupes de la protéine délimitant les sites : la poche.

Cela atténue la nécessité d'aligner spatialement les structures, et permet de traiter des cibles présentant de larges mouvements conformationnels. Néanmoins, l'environnement des cavités apparaît être d'une grande variabilité au cours des simulations de dynamique moléculaire de protéines, et l'identification d'une cavité ou d'un site requiert une classification appropriée des poches délimitant les cavités.

1.3 Identification cohérente des cavités : approche et objectifs

Nous proposons une approche pour identifier de façon robuste les sites présents dans les trajectoires de protéines, en partitionnant les cavités selon les "poches" qui les entourent. Chaque groupe de cavités définit par construction une *poche transverse* consensus. L'assemblage de ce dernier avec les cavités transverses associées forme un *site*.

La nature complexe de la définition d'un site requiert l'évaluation de la pertinence de la méthode et l'identification des options et paramètres les plus appropriés.

Par conséquent, nous avons systématiquement mis à l'épreuve les différentes définitions, méthodes et paramètres utilisés. Pour surmonter l'ambiguïté et la subjectivité du suivi des cavités, nous avons sélectionné un ensemble de 15 sites références précédemment décrits dans la littérature pour éprouver notre méthode.

Un ensemble consensus de paramètres s'est trouvé applicable à la plupart des systèmes testés. D'autres combinaisons de paramètres, avec des performances similaires sont aussi documentées et listées.

2 Matériels et Méthodes

2.1 Méthodes de conception

2.1.1 Définitions

Cavité Instantanée Partie connexe (un seul morceau) du volume accessible au solvant situé entre les atomes d'une conformation de protéine. Une cavité instantanée peut-être enfouie ou à la surface de la protéine.

Poche et empreinte de poche La poche correspond à la liste de groupes d'atomes de la protéine entourant une cavité instantanée. Ces groupes peuvent être des résidus, les atomes individuels ou d'autres groupements d'atomes. L'empreinte d'une poche est un descripteur numérique de la cavité instantanée, un vecteur (de la dimension du nombre de groupes dans la protéine) indiquant la proximité (booléenne ou réelle) de chaque groupe à la cavité instantanée.

Poche Transverse Ensemble de poches qui partagent des empreintes semblables dans des conformations différentes (ou identiques) de la trajectoire.

Empreinte Moyenne et Poche Consensus L'empreinte moyenne est la moyenne numérique des empreintes associées à une même poche transverse. La poche consensus correspond alors à la liste de groupes d'atomes les plus représentés dans l'empreinte moyenne.

Cavité transverse Ensemble de cavités définies associées aux poches d'une même poche transverse.

Site Un site est une région de la protéine où une cavité instantanée est observée de façon régulière au cours de la trajectoire moléculaire. Un site est pleinement défini par une poche transverse et sa cavité transverse associées.

2.1.2 Suivi des cavités : schéma et méthodes de calcul

L'enchaînement des différentes étapes établies pour le suivi des cavités et la définition des sites est exposé dans la Figure IV.2/p.suiv..

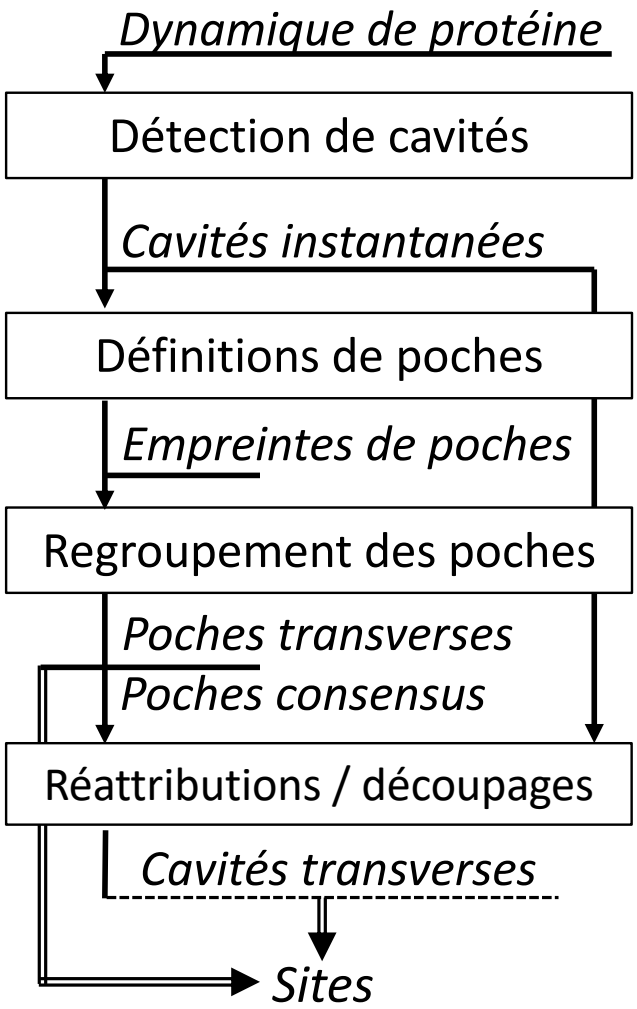


Figure IV.2 Étapes du suivi des cavités pour la détection de sites. Les routines sont représentées dans des boîtes, les données en *italiques*.

Une trajectoire moléculaire est utilisée en entrée de la procédure. L'implémentation ici proposée est optimisée pour calculer les cavités de protéines représentées par les coordonnées cartésiennes de leurs atomes. Cependant, la procédure est tout aussi applicable à des représentations moléculaires à plus gros grains (p.ex. format MARTINI), ou à des macromolécules diverses (p.ex. lipides, ADN, nanostructures, etc.). La trajectoire doit préalablement être désolvatée puis éventuellement alignée pour faciliter l'analyse et la visualisation des cavités transverses.

L'algorithme calcule les cavités instantanées sur l'ensemble des conformations de la trajectoire ainsi que l'empreinte de leur poche. Si elles sont en trop grand nombre, les empreintes peuvent être échantillonnées dans un sous-ensemble de conformations. Les poches sont ensuite partitionnées en différents groupes par similarité de leurs empreintes. Ces groupes forment des poches transverses, pour lesquelles une empreinte moyenne est ensuite utilisée pour réassigner les cavités instantanées de départ à une (ou plusieurs, si découpage) cavités transverses. Cet algorithme

réalise ainsi l'énumération exhaustive de l'ensemble des cavités de la trajectoire pour identifier les sites le long de la trajectoire.

Détection des cavités instantanées Dans l'implémentation présentée conjointement avec ce travail nous utilisons un outil de calcul développé au laboratoire représentant les cavités par des points de grilles (voxels). Cependant, n'importe quel outil de calcul de cavité peut en théorie être associé à ce protocole pour peu : (1) qu'il fournisse une représentation spatiale de la cavité à même d'identifier sans ambiguïté la poche l'entourant et (2) que les cavités instantanées d'une même conformation de protéine soient individuellement identifiées.

Définition de poches : calcul des empreintes La proximité des groupes d'atomes délimitant une cavité instantanée, sa poche, est décrite par un vecteur, appelé "empreinte". 3 définitions pour les groupes d'atomes ont été envisagées et testées : atomes (*byatom*), résidus (*byres*), ou B.S. (Backbone-Sidechain, 2 groupes par résidu). La taille du vecteur d'empreinte est identique au nombre de groupes dans la protéine.

Les définitions données ci-dessous sont spécifiques à une représentation des cavités instantanées sous forme d'ensembles de points de grilles. Une représentation différente demanderait une reformulation de ces concepts.

Soit c une cavité instantanée composée de voxels v ; G l'ensemble des groupes g de la protéine et composé des atomes a. La distance entre c et g est donnée par :

$$\delta(c,g) = \min_{v \in c, a \in g} (d(v,a) - rad(a)),$$

où $d(v,a)$ est la distance euclidienne entre v et a, et $rad(a)$ le rayon de Van der Waals de l'atome a.

3 définitions d'empreintes, *fp* (de l'anglais $footprint$), sont ici considérées :

- *Réelle*,

$$fp_g^{real}(c) = \begin{cases} \sigma - \delta(c,g) & \text{if } \delta(c,g) < \sigma \\ 0 & \text{otherwise,} \end{cases}$$

où le seuil, σ, est de 5 Å.

- *Booléenne*,

$$fp_g^{bool}(c) = (\delta(c,g) < \sigma)$$

- *Booléenne locale*,

$$fp_g^{lc-bool}(c) = (\delta'(c,g) < \sigma')$$

où σ' est à 3 Å et δ' la distance locale :

$$\delta'(c,g) = \min_{v \in c, a \in \{a_v\} \cap g} (+\infty, d(v,a) - rad(a)),$$

avec a_v, l'atome le plus proche du voxel v (avec prise en compte du rayon de Van der Waals). En d'autres termes, l'empreinte booléenne locale correspond à l'ensemble des groupes d'atomes de la protéine plus proche d'au moins un voxel de la cavité que tous les autres groupes. En cela, elle représente la définition la plus minimale et essentielle de la poche délimitant la cavité.

Plusieurs remarques sont à noter concernant ces définitions : D'une part, les groupes trop "éloignés" (distance supérieure au seuil σ) ne sont pas pris en compte. Ceci à pour but d'éviter qu'un mouvement conformationnel distant (par exemple le mouvement d'un domaine à l'autre extrémité de la protéine) n'influe sur l'attribution de l'empreinte. D'autre part, les empreintes sont invariantes par rotation et translation de la protéine (à erreurs d'arrondis près dans le calcul des voxels).

Distance entre empreintes Selon la nature de l'empreinte, la distance entre deux empreintes a et b est calculée avec différentes métriques.

- *Distance Euclidienne, pour empreintes Réelles :*

$$d(a,b) = \| a - b \| = \sqrt{(a-b)^2}$$

- *Distance Cosine, pour empreintes Réelles :*

$$d^{cos}(a,b) = 1 - \frac{a \cdot b}{\|a\|\|b\|}$$

- *Distance de Jaccard, pour empreintes Booléennes*

$$d^{jac}(a,b) = 1 - \frac{|a \cap b|}{|a \cup b|},$$

et notée $d^{jac-loc}$ lorsqu'elle est appliquée sur des empreintes booléennes locales.

Partitionnement des empreintes Différentes méthodes ont été implémentées et évaluées : les méthodes de regroupement hiérarchique avec liens maximum (*Complete* linkage) ou moyen (*UPGMA*), ainsi que les algorithmes de partitionnement *Spectral*, *DBSCAN* et *MeanShift*.

Ces méthodes ont été choisies parce qu'elles autorisent une modulation directe du nombre de groupes calculés en ajustant un unique paramètre d'entrée (par exemple le seuil de coupe dans l'arbre de partitionnement hiérarchique). Ce paramètre est compris entre 0 et 1 lorsque les distances Cosine et Jaccard sont utilisées pour comparer les empreintes. Il est Réel pour la distance euclidienne.

Empreinte moyenne / Poche consensus Pour chacun des groupes d'empreintes (appartenant à la même poche transverse), une empreinte moyenne peut-être calculée. Soit L l'ensemble des empreintes associées à un label donné.

$$E_{mean}(L) = \frac{1}{\#L} \sum_{l \in L} l$$

La poche consensus correspond à l'ensemble des groupes dont la valeur de l'empreinte moyenne est supérieure à un certain seuil. Cette définition peut par exemple servir à caractériser un site de la protéine sous la forme de ses résidus pour réaliser un amarrage moléculaire dans des conformations multiples de la même poche.

Réattribution/découpage des cavités Lors de l'étape de réattribution, les empreintes de l'ensemble des cavités instantanées sont comparées et reclassées selon la plus proche empreinte moyenne calculée après partitionnement. Cette étape permet en particulier de reclasser les empreintes des cavités instantanées qui n'ont pas été échantillonnées lors du partitionnement.

Un niveau plus fin de réattribution des cavités est établi lorsqu'un découpage des cavités instantanées est nécessaire. Dans ce cas, une empreinte spécifique est attribuée à chaque voxel de la cavité instantanée. Le voxel prend ensuite l'identifiant de l'empreinte consensus la plus proche. Ce réassignement des voxels permet le découpage des cavités instantanées dans une ou plusieurs cavités transverses, et maintient une définition cohérente des sites lorsqu'une ou plusieurs cavités fusionnent au cours de la dynamique.

2.1.3 Implémentation pratique

Deux implémentations de l'approche de suivi des cavités ont été codées pour permettre la réalisation de ce travail. Une première version sous forme d'un module Python, appelé *PyCAV*, a été développée par Nathan Desdouits, doctorant au laboratoire pendant mes deux premières années de thèse. *PyCAV* a permis de réaliser l'ensemble des conditions de suivi de cavités testées et analysées dans ce manuscrit. C'est aussi un outil puissant de manipulation et d'analyse de la géométrie (Analyse

en Composantes Principales sur les cavités) et des trajectoires de cavités [9, 227]. J'ai pris part au développement de cet outil pendant et après son départ du laboratoire.

Un autre outil développé au laboratoire, *mkgridXf*, était historiquement dédié à la détection des cavités dans les dynamiques de protéine. En début de thèse, j'ai commencé à porter ce programme en langage C avec l'objectif de pouvoir le faire interagir avec le module python *PyCAV*. Puis, faisant face à des limitations de *PyCAV* quant au suivi des cavités sur le récepteur nicotinique (temps de calculs trop long et empreinte mémoire trop lourde), j'ai progressivement ajouté la fonctionnalité du partitionnement des cavités dans *mkgridXf*. À l'inverse de *PyCAV*, *mkgridXf* implémente seulement les options qui ont été reconnues (dans ce travail) comme efficaces sur un ensemble de cibles de référence et la possibilité de réassigner les cavités au niveau des voxels pour le découpage des cavités instantanées ambiguës.

Sortie des programmes Toutes les structures de données calculées par *PyCAV* peuvent être exportées sous forme d'un format binaire facilitant la réimportation des données pour analyses complémentaires. Le logiciel *mkgridXf* peut exporter les cavités dans un format structuré contenant les points de grille de chacune des cavités instantanées pour chacune des conformations de la trajectoire, ainsi que l'identifiant de cavité transverse associé si l'identification des sites a été demandée. Le cas échéant l'empreinte moyenne de chacun des sites est aussi donnée. Les fichiers binaires produits par *mkgridXf* peuvent être lus par un programme complémentaire, *mkread*, et par un *plugin* VMD pour la visualisation directe des trajectoires de cavités.

2.1.4 Évaluation du suivi des cavités et de la détection de sites

Pour surmonter l'ambiguïté fondamentale de la définition d'un site, particulièrement dans un contexte dynamique, nous avons sélectionné un ensemble de "cavités de références" pour éprouver les différentes combinaisons d'options de partitionnement possibles et calibrer la méthode.

Ce choix s'est porté sur des protéines de tailles différentes, dont les dynamiques sont variées et pour lesquelles des sites de liaison effecteurs sont connus et documentés dans la littérature (la myoglobine, la protéine E du virus de la dengue, la toxin EF de l'anthrax et abl1).

2.2 Matériels et Méthodes

2.2.1 Dynamiques Moléculaires

4 trajectoires de dynamique moléculaire, une pour chaque système étudié. La trajectoire de la myoglobine est extraite d'une trajectoire de 120 ns (T=300 K et solvant explicite) précédemment publiée [9] et initialisée à partir du test case CHARMM "mbco4958" dérivée de la publication [204]. L'histidine proximale n° 93 y est liée à l'Hème et le CO laissé libre. La trajectoire de l'abl1 tyrosine-kinase (structure initiale de code PDB 2HZI [228]) provient d'une dynamique moléculaire de 200 ns en solvant explicite précédée de 2 ns de prééquilibration (6 395 H2O, 18 Cl^-, 27 Na^+, boîte de taille 76x52x62 $Å^3$, T = 298K avec une constante de couplage de Langevin de 1 ps^{-1}). La protéine d'enveloppe E du virus de la dengue (PDB 1OKE [229]), a été soumise à 10 ns de dynamique moléculaire, en solvant explicite (39 852 H2O dans une boîte de 180x90x85 $Å^3$) à 300 K avec une constante de couplage à 1 ps^{-1}. 1 000 conformations ont uniformément été sélectionnées pour chacune des 3 trajectoires précédemment décrites. Dans le cas de la trajectoire de la toxine d'anthrax (EF), deux dynamiques issues de travaux antérieurs [6, 109] ont été utilisées : la première de la forme apo/inactive (PDB : 1K8T [230]) et la seconde à partir de la conformation active en complexe avec la calmoduline et 2 ions Ca^{2+} (PDB : 1K93 [230]). Ces deux trajectoires ont été désolvatées (dont suppression de la calmoduline), orientées et 1 000 conformations ont été uniformément sélectionnées et concaténées résultant en une trajectoire de 2 000 conformations.

2.2.2 Détection des cavités instantanées

La détection des cavités a été effectuée avec le programme *mkgridXf*. Celui-ci utilise la Surface Moléculaire pour délimiter les cavités, comme initialement introduit par Richards in 1971 [231], et implémenté en 1983 par Connolly [232].

Le calcul est réalisé sur une grille orthogonale tridimensionnelle, avec un espacement unitaire grd (0,5 Å par défaut). Une petite sonde sphérique (rayon rin = 1,4 Å) est itérativement déplacée sur la grille et son centre est ajouté au volume interne, V_{in}, si aucun atome de la protéine ne la chevauche. Les rayons de Van der Waals sont soit récupérés de Bondi et al. [233] (choix par défaut quand les hydrogènes sont spécifiés) ou du fichier de topologie du champ de force CHARMM19 (lorsque les hydrogènes ne sont pas spécifiés). Une procédure identique est réalisée avec une sphère plus large (rou = 8 Å) pour définir le volume V_{out} accessible au solvant extérieur ("bulk solvent"). Le volume V_{rou} longe la surface accessible au centre de la sphère et non la surface moléculaire. Pour arriver à un tel résultat les points de cavités sont étendus de $rou + srou$ = 8+3.3 Å par dilatation morphologique des points de grille sélectionnés (troncation de la carte de distance - *Distance Transform*). Le paramètre *srou* sert à garantir un enfouissement suffisant des cavités en "érodant"

les cavités de surface. Un point de grille est alors considéré comme un voxel de cavité s'il se trouve dans le volume V_{rin}, mais pas dans V_{rou}. Les points de cavité sont ensuite regroupés en cavités instantanées en utilisant l'algorithme "En deux passes" de Shapiro et al. [234] permettant d'identifier les différentes régions connexes d'une grille de points (connexités avec les 26 cases voisines). Finalement, le volume des cavités instantanées restantes est dilaté de rin = 1,4 Å, de la même manière que le volume V_{rou} précédemment. Le fait de décaler cette étape de dilatation après l'attribution des identifiants de cavités instantanées permet de définir des cavités instantanées qui se touchent mais qui ont un identifiant différent du fait qu'une petite sphère de 1,4 Å ne peut pas passer continûment de l'une à l'autre.

Le volume d'une cavité instantanée est calculé comme la somme des volumes de ses voxels constitutifs (cubes unitaires de grd^3 = 0,125 Å^3).

Le volume de l'enveloppe de la protéine est défini comme le complément du volume accessible au solvant. Le domaine de définition des cavités ou de l'enveloppe de la protéine est l'union des points de grilles de cavités sur toutes les conformations de la trajectoire. Le domaine de définition est calculé sur les trajectoires alignées.

2.2.3 Classification des cavités

Par chevauchement spatial

Une approche intuitive pour regrouper des cavités est de regarder si elles se chevauchent dans des conformations différentes de la protéine. Les trajectoires sont préalablement alignées sur la première conformation de la trajectoire et les cavités détectées avec *mkgridXf*. Nous pouvons définir un graphe dont chaque nœud est associé à une cavité instantanée. Deux nœuds sont alors reliés par une arête si l'intersection spatiale de leurs cavités instantanées respectives a un volume supérieur à 12 Å^3. Pour le *graphe séquentiel*, seuls les liens entre conformations consécutives de la trajectoire sont considérés. À l'inverse, le *graphe complet* considère l'ensemble des pairs de conformations possible. Les composantes connexes de ces deux graphes définissent les *cavités transverses*. Les cavités *orphelines* sont des nœuds sans liens vers d'autres nœuds. En orientant le *graphe séquentiel* (une arête est considérée comme un arc entre les conformations n et $n+1$) on peut définir un événement de *fusion* (respectivement division) quand un nœud a plus d'un prédécesseur direct (resp. successeur directe). Un nœud sans successeur (ou prédécesseur) définit un événement d'apparition (disparition) d'une cavité instantanée.

Options de partitionnement des empreintes

Le seuil ajustable, d_{th}, est utilisé pour faire varier le nombre de groupes lors du partitionnement par le biais d'un paramètre en entrée de l'algorithme. L'idée est de trouver un seuil consensus satisfaisant pour l'ensemble des systèmes de référence testés. Une multiplicité de seuils a été évaluée d_{th} :
- des seuils fixes, de 0,05 à 0,95 par incréments de 0,05,
- des seuils par ratio de r_{th} (0,01, 0,05, 0,1, 0,3, 0,5, 0,7, 0,9, 0,95, 0,99).

Dans ce dernier cas, d_{th} est identifié comme la première valeur, d, pour laquelle le ratio $r_{th} < H_i(d)/H_w(d)$, avec $H_i(d)$, l'histogramme normalisé ($\Sigma_d H(d)$=1) des distances entre empreintes intraconformation (où toutes les poches sont distinctes) et $H_w(d)$, l'histogramme normalisé pour toutes les distances. Finalement, le seuil semi-automatique, d_{sat}, est la valeur la plus grande tel que le ratio $H_i(d)/H_w(d) < th$ pour $d \in [0,d_{sat}]$. th a été échantillonné entre 0,01 et 0,99 par incréments de 0,01. Pour les paramètres utilisant la distance euclidienne, seule la méthode semi-automatique a pu être utilisée.

Pour le partitionnement hiérarchique avec lien moyen (*UPGMA*) ou maximum (*complete*), l'implémentation utilisée est celle de *Scipy* [235]. Le paramètre ajustable d_{th} est identifié par le paramètre de coupe dans l'arbre hiérarchique.

Pour *DBSCAN*, l'implémentation est celle de *Scikit-learn* [236]. d_{th} est associé au paramètre de rayon maximum entre deux points voisins (ϵ), et le nombre minimum de voisins (*minPts*) est fixé à 10.

Pour le partitionnement *Spectral*, d_{th} est utilisé pour ajuster la matrice d'adjacence A, où $A_{i,j} = max(0,d_{th} - d_{i,j})/d_{th}$ pour chaque pair d'empreinte (*i,j*). La normalisation/diagonalisation et projection des vecteurs de distance sur les vecteurs propres est réalisée de façon classique (description et implémentation détaillée dans la thèse de N.Desdouits [227]).

Le partitionnement *Mean-shift* de *Scikit-learn* [236] est utilisé avec les paramètres par défaut. Seule la distance euclidienne a pu être évaluée avec cette méthode de partitionnement. Pour cela le seuil d_{th} correspond au paramètre de "bande passante" (*bandwidth* dans *sklearn*).

Échantillonnage des empreintes et réassignement

Lorsque le nombre de cavités détectées au cours de la trajectoire est trop grand, le temps de calcul du partitionnement des empreintes peut devenir ingérable. Typiquement, pour un partitionnement hiérarchique, dont la complexité algorithmique est en $\mathcal{O}(N^3)$, il faut une dizaine d'heures pour partitionner 30 000 empreintes

sur une machine standard et 1CPU. Pour surmonter cela, un sous-ensemble des empreintes peut être échantillonné. C'est le cas de la protéine E de la dengue et de la protéine EF de l'anthrax pour lesquelles seule 1 empreinte sur 10 a été choisie. Pour classer les empreintes non échantillonnées, deux méthodes de réassignement ont été envisagées :
- assignement par distance *min*-imum : classe l'empreinte suivant le groupe de l'empreinte déjà partitionnée la plus proche.
- assignement par empreinte moyenne (*mean*) : classe l'empreinte suivant le groupe de l'empreinte *consensus* la plus proche. Pour reclasser les empreintes booléennes, la variante min/max de la distance Jaccard est calculée :

$$d^{jac-reel}(a,b) = 1 - \frac{\sum_i min(a_i,b_i)}{\sum_i max(a_i,b_i)}.$$

2.2.4 Sites de référence

Parmi les 15 sites sélectionnés dans les 4 protéines étudiées, 10 sont définis par une structure co-cristallisée avec un ligand : (id. PDB) 1J52 pour la myoglobine ; 3K5V pour la kinase abl1 ; 1OKE pour la protéine E de la dengue et 1K90 pour la toxine EF de l'anthrax. Les autres sites sont identifiés à partir de listes de résidus trouvées dans la littérature (voir Résultats 3.4/p.108).

La détermination de la poche de référence définissant chacun des sites est ensuite calculée par l'empreinte "Locale Booléenne, par résidus" de la (ou des) cavité(s) instantanée(s) identifiée(s) dans le site.

Assignements de référence

Pour chaque conformation f de la trajectoire, le centre géométrique de la poche de référence, $g_{pocket}(f)$ est utilisé pour déterminer l'ensemble des cavités instantanées de la trajectoire associée à un site de référence. Si $\delta(c,g_{pocket}(f))$ est inférieure à 3 Å (distance δ décrite dans la section précédente), alors la cavité c est assignée au site de référence correspondant. Lorsqu'une cavité est assignable à deux sites de référence, le site le plus proche est privilégié. En pratique, cela signifie qu'un suivi de cavités sera satisfaisant pour un site de référence, s'il classe dans un même groupe les cavités assignées à ce site.

Trajectoire de cavité de référence

La trajectoire de cavité de référence, pour un site s, est utilisée pour évaluer la qualité de la géométrie et des volumes des cavités transverses prédites. Elle correspond

à l'ensemble des points de grilles de cavité qui seraient détectés le long de la trajectoire, si le site était connu *a priori*. 3 ensembles de cavités sont calculés pour chaque conformation, f, de la trajectoire :

- C_{all}, les cavités du système entier,
- C_L, les cavités détectées dans la poche de référence seulement, avec $srou = 0$
- C_p, les cavités détectées dans la poche de référence seulement, avec $vex = 0$ et $srou = 3{,}1$.

Alors, $C_{s,f}$ est donnée par les composantes connexes (comme définis dans C_{all}) de $C_{all} \cap C_L$ qui chevauchent C_p et ont un volume supérieur à 12 Å^3 (voir Figure IV.3). La prise en compte de C_p et de C_L vise à confiner les cavités à leur poche de référence et à supprimer les extensions ambiguës de cavités qui peuvent résulter d'une fusion de cavités issues de plusieurs sites dans certaines conformations de la trajectoire.

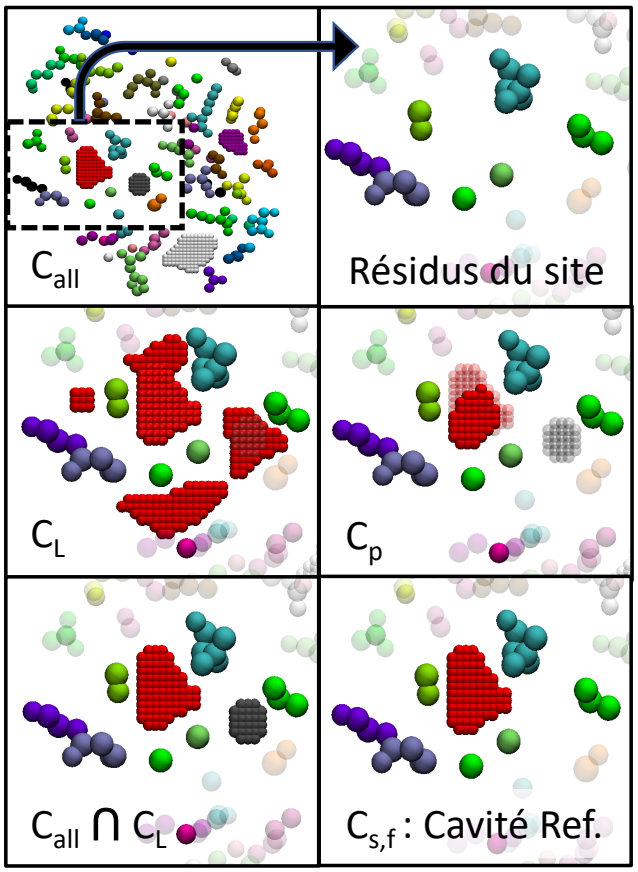


Figure IV.3 Détermination des trajectoires de cavités de référence. Calcule des cavités de référence $C_{s,f}$ pour une conformation f du site de référence s. C_{all} montre les cavités instantanées du système entier. Les cavités sont ensuite calculées avec *mkgridXf* en ne tenant compte que des résidus du site de référence. C_L correspond à ce résultat, mais doit être corrigé de 2 façons : (1) $C_{all} \cap C_L$ permet de supprimer les points de C_L qui ne sont pas dans C_{all} et qui ne seront donc jamais observés dans une détection de cavités sur la protéine entière. (2) La suppression des cavités périphériques ou trop petites, en ne maintenant que les cavités dont au moins un point de grille est dans C_p.

2.2.5 Évaluation des assignements

Pour attribuer un score d'assignement, le résultat du partitionnement des cavités instantanées et l'assignement de référence sont encodés dans un vecteur dans lequel chaque élément correspond à une cavité instantanée c de la trajectoire. L'assignement de référence, pour un $site$, est donné par un vecteur booléen, P^{site}, égale à 1 si la cavité c est dans le $site$ et 0 sinon.

Score d'un assignement prédit

Pour une combinaison d'options données, le résultat d'un partitionnement des cavités instantanées est encodé dans un vecteur P, défini par $P(c) = k$, où k est le numéro de groupe assigné à chacune des cavité c (le nombre total de groupes est K).

Pour chaque groupe k, nous définissons le vecteur booléen P_k par $P_k(c) = 1$ lorsque $P(c) = k$, et 0 autrement. P_k et P^{site} sont comparés avec un $F1$-score :

$$F1(P^{site},P_k) = \frac{2.prec.rec}{prec + rec}, \quad \text{où,}$$

$$prec = \frac{TP}{TP + FP} \text{ et } rec = \frac{TP}{TP + FN}$$

$$\text{avec} \quad TP = P^{site}.P_k \ , \quad FP = |P_k| - TP \quad \text{et} \quad FN = |P^{site}| - TP$$

Le score d'assignement attribué à un site $site$ est donné par le score du groupe k qui maximise $F1(P^{site},P_k)$ (inspiré du "overall-$F1$_score" décrit par Larsen et al. [237]) :

$$F1_{site}(P^{site},P) = max_{k\in[1;K]}F1(P^{site},P_k)$$

Finalement, nous calculons un $F1$ score moyennant les scores des différents $sites \in Sites(prot)$ étudiés dans le système courant :

$$F1_{prot} = \frac{\sum_{site \in Sites(prot)} |P^{site}| \times F1_{site}(P^{site},P)}{\sum_{site \in Sites(prot)} |P^{site}|}$$

Score de couverture

Pour différentier les assignements faux, incomplets et ceux dont au moins une partie de la cavité est correctement assignée dans chacune des conformations, les valeurs suivantes sont calculées :

$$\text{valide} = \frac{1}{F}\sum_{f=1}^{F}(N_{TP}(f) > 0 \ \vee \ N_R(f) = 0)$$

$$\text{indéterminé} = \frac{1}{F}\sum_{f=1}^{F}(N_k(f) = 0 \ \wedge \ N_R(f) > 0)$$

$$\text{erreur} = \frac{1}{F}\sum_{f=1}^{F}(N_{TP}(f) = 0 \ \wedge \ N_k(f) > 0 \ \wedge \ N_R(f) > 0)$$

où F est le nombre de conformations de la trajectoire ; k est le groupe maximisant $F1(P^{site},P_k)$; $N_{TP}(f) = P_{k,f}.P_f^{site}$; $N_k(f) = |P_{k,f}|$; et $N_R(f) = |P_f^{site}|$; pour P_f^{site} et $P_{k,f}$ restreints aux assignements de cavités instantanées de la conformation f.

Évaluation de la géométrie des cavités transverses

Comme nous l'observerons, il arrive que de larges cavités instantanées recouvrent plusieurs sites distincts de la protéine. Le découpage des cavités instantanées permet de réattribuer des morceaux de la cavité à des sites différents. Dans le but de contrôler la pertinence du découpage des cavités instantanées, nous proposons de confronter les cavités prédites aux trajectoires de cavités de référence. L'évaluation la plus triviale est de comparer leur volume instantané dans chacune des conformations de la trajectoire. Une mesure plus précise de la similarité entre deux géométries de cavités est donnée par la distance d_{Geo}, inspirée par Hausdorff et al. [238]. Pour deux ensembles non vides de cavités C_1 et C_2 constitués de voxels v de rayon nul ($rad(v) = 0$), on a :

$$d_{Geo}(C_1,C_2) = \sqrt{\frac{\sum_{v\in C_1}\delta^2(v,C_2) + \sum_{v\in C_2}\delta^2(v,C_1)}{|C_1 \cup C_2|}} \ ,$$

$$d_{Geo}(C_1,\emptyset) = d_{Geo}(\emptyset,C_1) = \sqrt{\frac{(2\delta(g_1,C_1))^2}{|C_1|}} \ ,$$

où $\emptyset$ est la cavité nulle, et g_1 le centre géométrique de C_1 avec $rad(g_1) = 0$, et finalement

$$d_{Geo}(\emptyset,\emptyset) = 0.$$

2.2.6 Cavité moyenne et domaine de définition

Soit C un ensemble de cavités réparties dans les différentes conformations d'une trajectoire moléculaire. Nous définissons la *cavité moyenne tronquée à p %*, $\overline{C}^{p\%}$, comme l'ensemble des voxels apparaissant dans plus de p % des conformations de la trajectoire dans lesquelles au moins une des cavités de C est présente.

Le *domaine de définition* de C correspond à l'ensemble des voxels appartenant à une des cavités à au moins un moment de la trajectoire (c'est-à-dire $\overline{C}^{0\%}$).

Cette représentation n'est pertinente que si l'ensemble des conformations de la trajectoire ont été préalablement alignées les une par rapport aux autres, ce qui est le cas pour l'ensemble des trajectoires étudiées dans ce travail.

3 Résultats

3.1 Cavités instantanées dans les dynamiques de protéines

Nous avons utilisé l'outil de détection *mkgridXf* pour identifier les cavités apparaissant au cours des 4 trajectoires de tests. Le nombre total de cavités instantanées varie entre 11 863 pour la myoglobine et 72 507 pour la toxine EF de l'Anthrax. En accord avec des études précédentes [239–241], le nombre de cavités par conformation, ainsi que leurs volumes cumulés, est environ proportionnel au volume moyen de l'enveloppe de la protéine (voir Figure IV.4).

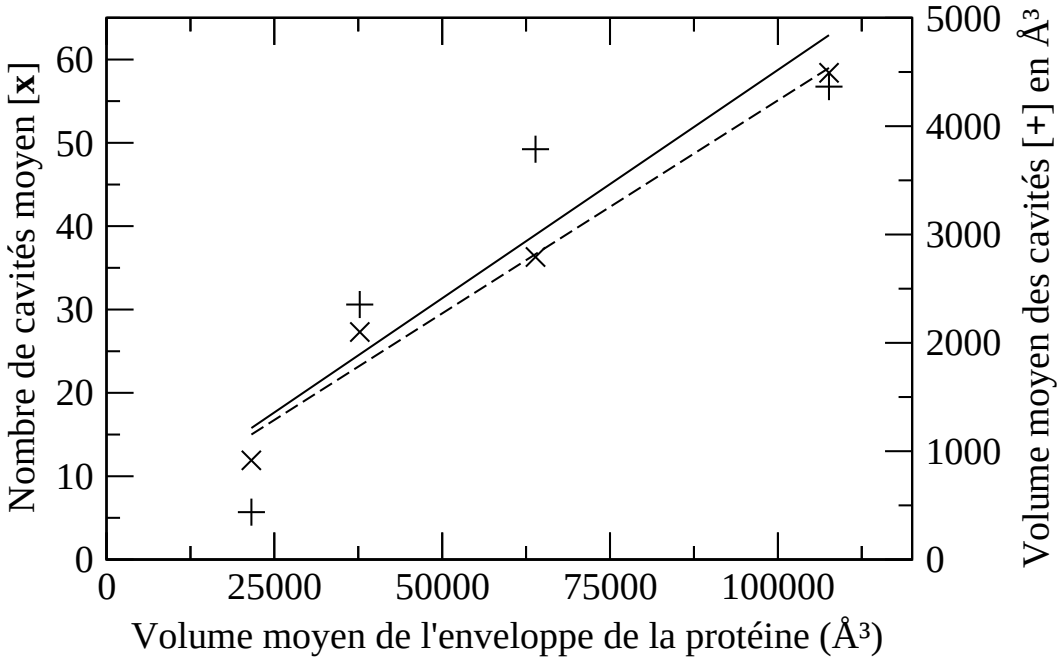


Figure IV.4 Statistiques de la dynamique des cavités. Nombre de cavités moyen par conformation (×) et volume moyen des cavités (+) en fonction du volume moyen de la protéine. Les régressions linéaires sont respectivement tracées en traits hachurés et pleins.

Le ratio entre le volume moyen des cavités et de la protéine est faible, entre 2 % et 6 % (tableau en Annexe VII.2/p.203). Par contraste, le volume couvert par les cavités (leur domaine de définition) est 10 fois supérieur à leur volume cumulé instantané (par conformation), représentant jusqu'à 34 % du domaine de définition de l'enveloppe de la protéine et plus de 89 % du volume moyen de la protéine (abl1). Par conséquent, les cavités apparaissent nombreuses et labiles au cours des dynamiques, s'étalant sur une large portion du volume des protéines.

3.2 Suivi des cavités par chevauchement spatial

La première tentative pour identifier des cavités partageant un site en commun dans des conformations différentes de la trajectoire utilise l'approche du graphe *séquentiel*, décrite en Matériels et Méthodes. Les cavités se chevauchant spatialement un minimum (12 Å^3, approx. le volume d'une molécule d'eau) dans des conformations consécutives de la trajectoire sont considérées comme ayant une même identité transverse. Ainsi définis, les évènements de division/fusion de cavités apparaissent communs, tandis que les apparitions/disparitions sont très fréquentes (Tableau IV.1).

Système	Myoglob.	Abl1	Dengue	EF
Nombre de cavités	11 863	27 308	58 376	72 507
		Graphe séquentiel		
[a] # Comp. Conn.	5 914	12 002	21 323	25 988
[b] ++Comp. (%)	6,3	67,1	31,6	46,5
# fusion	174	1 859	3 420	4 724
# division	160	1 831	3 413	4 718
# apparition	6 053	12 982	22 784	27 704
# disparition	6 036	12 931	22 755	27 731
# noeuds orphelins	4 602	10 044	17 925	20 859
		Graphe complet		
[a] # Comp. Conn.	343	848	2 637	2 532
[b] ++Comp. (%)	58,6	93,96	49,35	63,95
# noeuds orphelins	0	0	0	0
# fusion+division	98 069	1 415 133	3 020 607	6 353 479

Tableau IV.1 Statistiques des graphes de cavités *Séquentiel* et *Complet*. [a]Nombre de composantes connexes du graphe, [b] Domaine de définition des cavités associées à la plus large composante connexe du graphe.

La trajectoire de la myoglobine est stable (RMSD : 1,29 Å, tableau en Annexe VII.2/p.203), sa composante connexe la plus large couvre 6,3 % du domaine de définition de l'ensemble des cavités. À l'inverse, pour abl1, ce pourcentage atteint 67 % (voir Figure IV.5/p.suiv.), reflétant la plus grande flexibilité des structures (RMSD : 3,32 Å), et l'incapacité de la méthode à identifier des cavités appartenant à un site localisé de la protéine. Une conclusion similaire peut être formulée pour la trajectoire de la dengue. Près de 30 % des cavités ne chevauchent aucune autre cavité dans sa conformation consécutive (nœuds orphelins), et cela pour tous les systèmes.

Nous pouvons aussi considérer les conformations de la trajectoire comme un ensemble désordonné, ce qui correspond à l'analyse du graphe *complet*. La connectivité

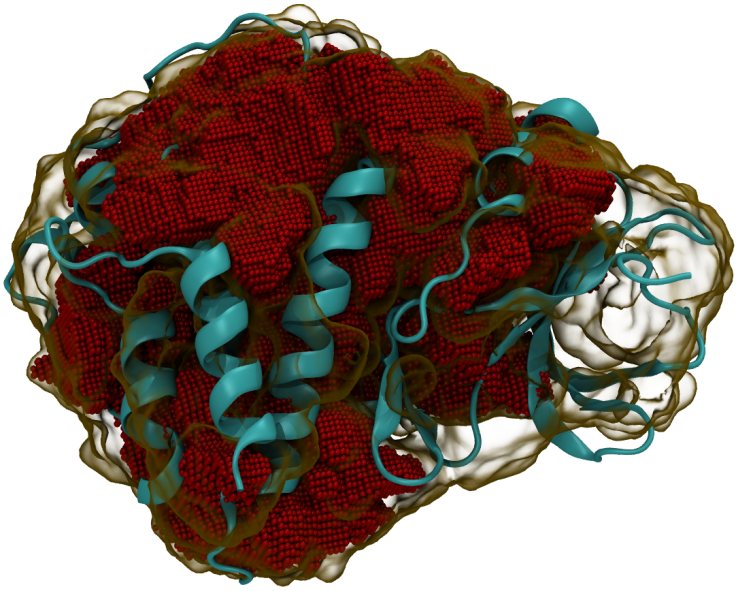


Figure IV.5 Domaine de définition des cavités de la plus large composante connexe du graphe *séquentiel*. Pour la trajectoire abl1 (points rouges) le volume couvert représente 67,1 % du domaine de définition de l'ensemble des cavités qui apparaissent au cours de la trajectoire (surface transparente).

des graphes augmente considérablement, couvrant de 49 % à 94 % du volume des protéines étudiées. Ces résultats suggèrent que l'identification par chevauchement de cavités n'est en pratique pas réalisable, en particulier pour des trajectoires trop bruitées, ou décrivant de larges changements conformationnels. De telles méthodes adaptées au suivi local des cavités ne sont alors envisageables que pour des trajectoires continues, dont les intervalles de temps de simulation entre deux conformations sont très courts (< 10 ps, voir référence [226]).

3.3 Identification des cavités selon leur environnement protéique

Nous utilisons l'empreinte de la cavité sur la protéine (p.ex. une liste de groupes d'atomes formant la poche) dans le but de réduire la dépendance de la méthode envers les mouvements conformationnels. Pour tester cela, nous avons testé cette configuration simple :

- Groupe d'atomes : les résidus de la protéine
- Empreinte : booléenne (Vrai si plus proche que 5 Å)
- Distance : mesure de dissimilarité cosine
- Partitionnement : hiérarchique, lien *moyen*
- Seuil de partitionnement : réel $\in$ [0-1]

La dissimilarité cosine est ici privilégiée à la distance euclidienne à cause des dif-

ficultés parfois observées de la norme L2 à comparer des vecteurs de grandes dimensions [242]. Cependant, la majorité des autres options de configuration est choisie arbitrairement et il est difficile d'en justifier le choix. De même pour le choix du seuil de partitionnement, la Figure IV.6 montre que le nombre de cavités transverses varie quasiment continûment avec la valeur de ce seuil, ce qui rend difficile le choix *a priori* d'un seuil plutôt qu'un autre.

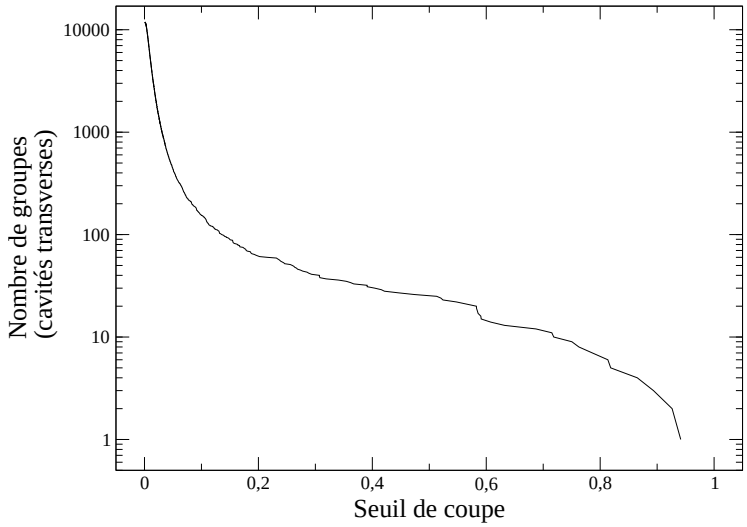


Figure IV.6 Nombre de groupes obtenus en faisant varier le paramètre de coupe lors du partitionnement. Obtenu par un partitionnement hiérarchique avec lien *moyen* des empreintes de cavité de la myoglobine.

Par conséquent, il apparaît essentiel de mettre au point un ensemble de données de référence afin de pouvoir juger objectivement de la qualité d'un suivi global des cavités de protéine et d'en établir une configuration pertinente.

3.4 Choix des sites de référence

Les sites de référence ont été sélectionnés parmi un ensemble de cibles protéiques pour lesquelles un ou plusieurs sites de liaison ont été documentés dans la littérature. Ces systèmes ont été choisis de par leurs tailles et fonctions diverses, ainsi qu'une variété de mouvements structuraux. En l'occurrence, la myoglobine est relativement statique (Annexe VII.2/p.203) au cours de la trajectoire, alors que la protéine d'enveloppe du virus de la dengue et la kinase abl1 présentent des fluctuations plus larges. La composante EF de la toxine d'anthrax a été étudiée sur deux ensembles de conformations, le premier dans un état inactif et l'autre dans un état actif, avec une moyenne de dissimilarité structurale de plus de 10 Å RMSD entre les deux jeux de dynamiques.

Les sites de référence sont retranscrits en Annexe VII.8/p.201, en partant soit du ligand co-cristallisé avec la protéine, soit de la délimitation de la poche comme décrite dans la littérature. Pour la myoglobine, les poches sont localisées dans la structure par le CO (poche distale) et les atomes de Xenon (Figure VII.8/p.201 : myoglobine). Le β-Octyl-Glucoside marque le site βOG dans la protéine E de la dengue [211, 229] (Figure VII.8/p.201 :Dengue). Les molécules Imatinib [243] et GNF-2 [244] sont respectivement liées au site catalytique et allostérique de l'abl1 tyrosine-kinase (Figure VII.8/p.201 : abl1). Le 3'dATP marque le site catalytique de EF [230]. Nous avons aussi utilisé les poches dénommées Site 1 et Site 2 présentes dans la protéine d'enveloppe de la dengue et décrites dans les références [245] et [224]. Ce sont des sites de liaison potentiels découverts grâce à l'analyse des cavités entre les formes cristallographiques des états pré/post fusion et par simulation de dynamique moléculaire. La définition de leur poche (liste de résidus) diffère dans les deux publications. Le Site 1 est formé par deux cavités adjacentes visibles dans la structure cristallographique (PDB : 1OKE) entre les résidus désignés par Fuzo et al. [224] alors qu'il n'en couvre qu'une seule avec la définition de Yennamalli et al. [245]. Pour les sites 1 et 2, nous avons finalement privilégié la définition de [245] puisqu'elle délimite un site dont la liste de résidus est plus spécifique à la cavité présente dans la structure expérimentale. Pour le site SABC de EF, découvert par calcul de chemins de transitions entre les états actifs et inactifs, nous avons utilisé la définition transcrite dans la référence E.Laine et al. [6].

3.5 Combinaisons d'options de partitionnement pour l'identification des sites

Les combinaisons d'options utilisées impactent fortement le suivi des cavités (Figure IV.6/p.préc.). Par conséquent, pour en sélectionner la configuration la plus pertinente, nous avons exhaustivement exécuté les 2 286 combinaisons impliquées par la Figure IV.7/p.suiv. sur les 4 systèmes de référence précédemment décrits. 2 189 exécutions de ces calculs se sont terminées avec succès pour les 4 trajectoires et ont pu être évaluées.

La qualité d'un assignement de cavités est évaluée avec le score $F1_{prot}$. En fonction de l'objectif recherché (meilleure performance, plus robuste, etc.), les scores obtenus sur chacune des cibles peuvent être combinés de façons différentes : la moyenne des scores individuels, la moyenne des rangs, le pire score par site de référence ou par moyenne sur les systèmes, etc.. Pour mitiger l'impact des sites "faciles à identifier" et des sites plus "difficiles", nous avons initialement utilisé la somme des rangs des sites moyennés par systèmes ($\sum rang(F1_{prot})$: *RSum*, Tableau IV.2/p.111).

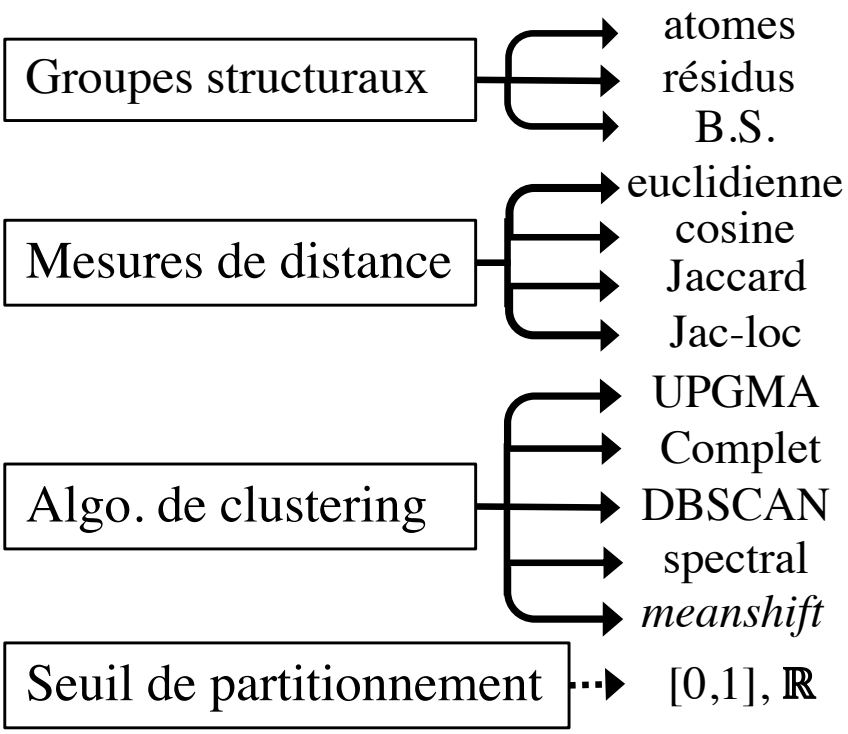


Figure IV.7 Combinatoire des combinaisons d'options pour le suivi des cavités. B.S. : *Backbone-Sidechain*, UPGMA ou Complet : Partitionnement hiérarchique avec lien moyen ou maximum

La meilleure combinaison de paramètres utilise un partitionnement hiérarchique de type *UPGMA*. La valeur élevée de *RSum* indique un grand nombre de bonnes méthodes sur certains systèmes et une certaine variation des performances entre les systèmes testés. Néanmoins, les deux meilleures méthodes, qui ne diffèrent que pour leur solution de réassignement (min ou mean) ont une somme des rangs près de deux fois inférieure à la 3e meilleure méthode. La meilleure combinaison d'options semble être *UPGMA*/cosine dist./real empreinte/*par atomes*/d_{th}=0,5/*min*. C'est aussi la combinaison qui maximise la moyenne des scores entre systèmes ($^1/_4 \sum F1_{prot}$; "moy." dans le Tableau IV.2/p.suiv.).

Parmi les meilleures combinaisons d'options, on remarque aussi que la myoglobine et la protéine E de la dengue ont constamment des scores élevés, tandis que EF et abl1 ont des scores plus modérés, autour de 0,8. Pour comprendre pourquoi ces assignements sont moins satisfaisants, nous les avons analysés site par site.

3.6 Qualité du suivi des sites individuels

La distribution des 100 meilleurs scores $F1_{site}$ pour chacun des sites étudiés est présentée dans la Figure IV.8/p.112. Les sites de la myoglobine sont correctement assignés (scores moyens $>$ 0,95 pour tous les sites, excepté pour Xe2, $\approx$ 0,92).

Rang	RSum	Groupe	Dist.	Partit.	Seuil	Réass.	Myo	EF	DENV	Abl1	moy.	min
1	112	byatom	cosine	upgma	50	min	0,958	0,797	0,901	0,735	**0,848**	0,735
2	124	byatom	cosine	upgma	50	mean	0,958	0,779	0,902	0,735	0,843	0,735
3	213	byatom	cosine	upgma	55	min	0,958	0,779	0,872	0,728	0,834	0,728
4	225	byatom	cosine	upgma	55	mean	0,958	0,762	0,878	0,728	0,831	0,728
5	269	byatom	cosine	upgma	60	mean	0,923	0,764	0,867	0,728	0,820	0,728
6	270	byatom	cosine	upgma	60	min	0,923	0,779	0,860	0,728	0,822	0,728
7	272	byatom	jaccard	upgma	65	mean	0,955	0,665	0,889	0,754	0,816	0,665
8	285	byatom	jaccard	upgma	75	mean	0,895	0,668	0,918	0,744	0,806	0,668
9	308	byatom	jaccard	upgma	65	min	0,955	0,659	0,902	0,754	0,818	0,659
10	322	byatom	jaccard	upgma	60	min	0,957	0,659	0,907	0,739	0,815	0,659
11	322	byres	jaccard	upgma	60	mean	0,931	0,659	0,921	0,738	0,813	0,659
...14	337	byres	cosine	upgma	45	mean	0,899	0,662	0,913	0,745	0,805	0,662
15	339	B.S.	jac-loc	upgma	75	min	0,914	0,799	0,851	0,720	0,821	0,720
...20	397	byres	cosine	upgma	50	mean	0,875	0,749	0,885	0,721	0,808	0,721
...30	443	B.S.	cosine	upgma	55	mean	0,885	0,739	0,858	0,727	0,802	0,727
...47	507	byres	cosine	upgma	35	min	**0,963**	0,657	0,834	0,745	0,800	0,657
„61	567	B.S.	cosine	upgma	50	mean	0,895	0,645	0,894	0,725	0,790	0,645
...66	590	B.S.	cosine	complet	90	min	0,845	0,751	0,823	0,829	0,812	0,751
...104	746	B.S.	jaccard	*spect.	70	mean	0,867	0,641	0,846	0,733	0,772	0,641
...124	860	byres	jaccard	upgma	75	mean	0,797	0,764	0,821	0,771	0,788	**0,764**
...198	1 127	byres	jac-loc	complet	95	min	0,815	**0,825**	0,756	0,706	0,776	0,706
...248	1 311	byres	euclid.	m-shift	a05	min	0,921	0,514	0,879	0,558	0,718	0,514
...270	1 418	B.S.	jaccard	dbscan	20	mean	0,840	0,573	0,733	0,736	0,721	0,573
...764	3 184	byres	jaccard	dbscan	25	min	0,368	0,665	**0,924**	0,105	0,515	0,105
...779	3 232	byres	jaccard	upgma	a01	min	0,903	0,110	0,150	**0,865**	0,507	0,110
...	...	...	...	...	...	...	...	...	...	...	...	...
2 189	8 694	byres	jaccard	complet	a09	min	0,055	0,061	0,044	0,008	0,042	0,008

Tableau IV.2 Performances de partitionnement suivant les options utilisées. Les combinaisons sont ordonnées par $F1_{prot}$ pour chaque protéine. Leur rang pour les 4 systèmes est ensuite sommé (*RSum*). La combinaison ayant le plus faible rang cumulé *RSum* est jugée la meilleure. Les combinaisons de partitionnement, Groupe, Dist., Partit., Seuil, sont décrites dans la Figure IV.7/p.préc.. Les seuils sont donnés en pourcentage, et sont semi-automatiques lorsqu'un *a* précède le seuil. Un réassignement, Réass. est appliqué si nécessaire. Les scores $F1_{prot}$ sont donnés pour chacune des protéines. La moyenne (moy.) et le pire (min) scores sont aussi renseignés. *Le partitionnement Spectral pour abl1 a spécifiquement été réalisé avec un échantillonnage 1/10 pour cause de dépassement de mémoires incombant à la méthode de partitionnement.

Les deux copies du Site 2 et de la poche βOG de la protéine de la dengue ont des scores supérieurs au site 1. De façon intéressante, des scores similaires sont obtenus pour chaque pair symétrique des sites de la dengue. Aucune méthode n'a pu prédire correctement le site SABC de EF avec un meilleur score que 0,8 et le top-100 moyen est d'ailleurs assez faible (0,70). Le site GNF-2 d'abl1 a les plus mauvais scores, avec un top-100 moyenné à 0,63. Ainsi, ces deux derniers sites apparaissent intrinsèquement difficiles à identifier correctement.

Pour rappel, la trajectoire EF est composée de deux états de la protéine. Inactive dans les 1 000 premières conformations, et en état actif dans les 1 000 suivantes. Nous avons analysé le score $F1_{site}$ dans ces 2 phases indépendamment (Figure IV.8/p.suiv., à droite). En suivant le mécanisme d'activation, le site catalytique est mieux formé

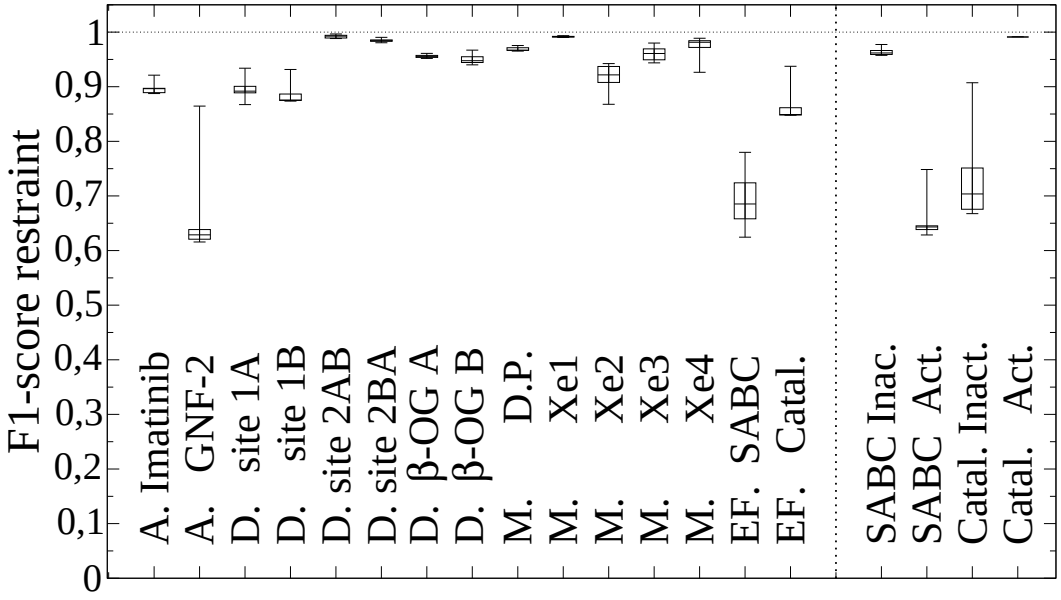


Figure IV.8 Distribution des 100 meilleurs scores $F1_{site}$. Les systèmes étudiés sont abréviés par A., D., M. et EF. pour abl1, protéine E de la dengue, myoglobine, et EF respectivement. Les diagrammes en boîte retranscrivent : minimum, 1er quartile, médiane, 3e quartile et valeur maximale. Sur la droite, $F1_{site}$ est calculé soit sur les 1 000 premières conformations "inactives", ou sur les conformations 1 001 à 2 000 "actives" de la trajectoire EF pour les sites catalytiques et SABC.

dans la seconde moitié de la trajectoire. À l'inverse, le site SABC est mieux formé dans la première moitié, mais déstructuré dans la seconde moitié comme supposé dans de précédents travaux [6]. De la même façon, SABC est bien identifié dans la première phase et le site catalytique dans la seconde ($\approx$ 0,96 or more), alors qu'elles sont mal définies dans leurs parties complémentaires de la trajectoire ($\approx$ 0,65).

3.7 Assignement des sites non ambigus

Afin de vérifier que la paramétrisation "optimale" de la méthode pour tous les sites est aussi capable de prédire les sites consensus, nous avons calculé les rangs sans tenir compte de ces sites dits "difficiles". Par conséquent le site allostérique d'abl1 a été supprimé, ainsi que le site SABC d'EF sur la seconde partie de la trajectoire, et le site catalytique d'EF sur la première (voir le Tableau IV.3/p.suiv.). Le résultat, quelque peu attendu, est une amélioration notable des scores des systèmes abl1 et EF. Ce calcul confirme aussi que la meilleure combinaison sélectionnée sur l'ensemble des sites donne aussi des scores élevés sur les sites consensus ($\gtrsim$ 0,9). La somme des rangs plus élevée, est indicative d'un plus grand nombre de combinaisons dont le score est élevé et donc plus difficiles à discriminer. La meilleure combinaison selon le *RSum* apparaît être *byatom-jaccard-UGPMA-55* conséquence d'une légère amélioration du score EF au-dessus de 0,96, et montrant les limites de la somme des rangs pour évaluer des systèmes proches du score maximum 1. Remarquablement, le

Rang	RSum	Groupe	Dist.	Partit.	Seuil	Réass.	Myo	EF	DENV	Abl1	moy.	min
1	200	byatom	jaccard	upgma	55	min	0,890	0,979	0,904	0,896	0,917	0,890
2	258	byatom	cosine	upgma	50	mean	0,958	0,966	0,902	0,898	0,931	0,898
3	280	byatom	cosine	upgma	50	min	0,958	0,965	0,901	0,898	0,930	0,898
4	294	byatom	cosine	upgma	45	mean	0,926	0,965	0,896	0,898	0,921	0,896
5	300	byres	cosine	upgma	30	min	0,925	0,977	0,835	0,897	0,908	0,835
6	305	byatom	cosine	upgma	45	min	0,926	0,963	0,903	0,898	0,922	0,898
7	355	byatom	jaccard	upgma	55	mean	0,890	0,974	0,881	0,896	0,910	0,881
8	355	byatom	jaccard	upgma	65	mean	0,955	0,959	0,889	0,906	0,927	0,889
9	363	byres	cosine	upgma	30	mean	0,925	0,976	0,828	0,897	0,906	0,828
10	397	byatom	jaccard	upgma	60	min	0,957	0,954	0,907	0,906	0,931	0,906
... 16	432	B.S.	cosine	upgma	30	min	0,891	0,975	0,858	0,886	0,902	0,858

Tableau IV.3 Performances de partitionnement sans les sites ambigus.. Les scores abl1 correspondent à ceux du site GNF-2 seul, et les scores EF à la combinaison du site SABC pour les conformations de 1 à 1 000 et du site catalytique pour les conformations de 1 001 à 2 000. Les abréviations sont similaires à celles décrites dans le Tableau IV.2/p.111.

partitionnement *byatom-cosine-UPGMA-50* reste un très bon candidat avec le meilleur score moyen (0,931) et un très bon score minimum (0,898) sur toutes les autres combinaisons. Ces deux valeurs sont meilleures que celles obtenues avec les options du rang 1, et ne sont surpassées que par la 10e méthode *byatom-jaccard-ugpma-60* (0,931/0,906). Cependant cette dernière est beaucoup moins robuste sur les sites "difficiles". Les empreintes par résidus (*byres*) et *backbone-sidechain B.S.* sont placés dans le top 10 des options avec méthodes et seuils similaires, ce qui démontre leur intérêt pour le suivi des cavités.

3.8 Localisation correcte des sites de liaison prédits

Un score $F1_{site}$ modeste (p.ex. abl1, site GNF-2, Figure IV.8/p.préc.) peut être indicatif de cavités instantanées dont la géométrie ne correspond pas exactement à la cavité de référence dans certaines conformations (par exemple un site qui se découpe en de multiples cavités), ce qui est normal. Ou bien, cela peut indiquer que l'assignement n'est pas à la bonne position dans certaines conformations, ce qui est plus problématique. Pour discriminer ces situations, nous avons testé, pour la combinaison d'options optimales, si dans chaque conformation, au moins une des cavités instantanées prédite pour ce site correspond à au moins une cavité instantanée de référence (Tableau IV.4/p.suiv.).

Les erreurs d'assignement (aucune prédiction satisfaisante dans une même conformation) sont en fait extrêmement rares, et apparaissent uniquement dans les cas difficiles, 1 % du temps pour le site GNF-2 d'abl1, qui est un site instable et plus de

Site Ref.	Val.	Ind.	Err.	Site Ref.	Val.	Ind.	Err.
D.Poc. (Myo)	1,00	0,00	0,00	βOG (D_A)	0,98	0,02	0,00
Xe1 (Myo)	1,00	0,00	0,00	βOG (D_B)	1,00	0,00	0,00
Xe2 (Myo)	0,98	0,02	0,00	Imatinib (Abl1)	0,95	0,05	0,00
Xe3 (Myo)	0,99	0,01	0,00	GNF-2 (Abl1)	0,72	0,27	0,01
Xe4 (Myo)	0,99	0,01	0,00	SABC (EF)	0,89	0,06	0,05
site1 (D_A)	0,98	0,02	0,00	(0-1000)	1,00	0,00	0,00
site1 (D_B)	0,98	0,01	0,00	(1000-2000)	0,79	0,12	0,09
site2 (D_AB)	1,00	0,00	0,00	Catalytique (EF)	0,76	0,24	0,00
site2 (D_BA)	1,00	0,00	0,00	(0-1000)	0,52	0,48	0,00
				(1000-2000)	1,00	0,00	0,00

Tableau IV.4 Scores de couverture pour la combinaison d'options *byatom-cosine-upgma-50-min*.. *Valide*, "Val.", *Indéterminé*, "Ind", and *Erreur*, "Err." sont calculés comme décrit en Matériels et Méthodes. D_A vaut pour DENV chaîne A, de même que B, AB (site à l'interface), et BA.

9 % du temps pour le site SABC dans la partie active de la trajectoire où le site est dispersé en plusieurs domaines distincts.

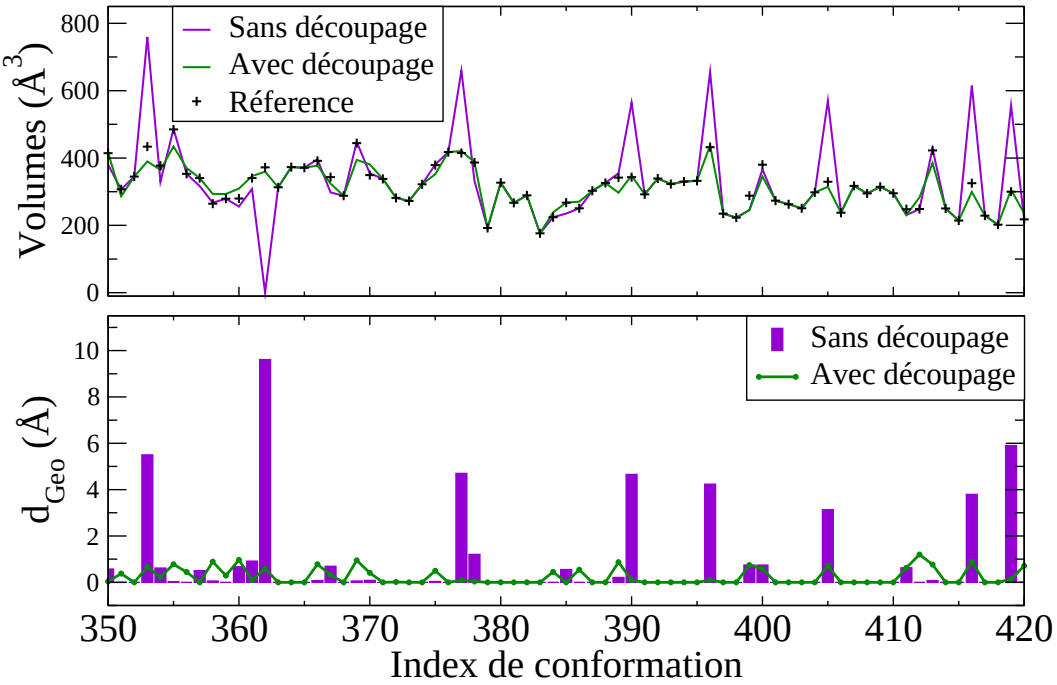


Figure IV.9 Accord entre les cavités de référence et les cavités prédites et effet du découpage des cavités. Comparaison du volume et de la géométrie des cavités de référence et prédites pour les sites βOG de la dengue, chaîne B. En haut : suivi du volume de la cavité de référence (+), de la cavité prédite (violet) et prédite après découpage des cavités instantanées (vert). En bas : distance géométrique d_{Geo} entre la cavité de référence et la cavité prédite, sans découpage (violet) et avec découpage (vert).

3.9 Nécessité du redécoupage des cavités instantanées

La cavité trouvée dans le site de référence ne correspond pas toujours avec la poche de référence à cause de fusions avec les cavités voisines. La Figure IV.10/p.suiv. montre l'exemple d'une cavité instantanée dont le volume sort grossièrement de son site de référence (volume bleu transparent). Ces cavités sont doublement problématiques. D'une part elles couvrent plusieurs sites consensus, ce qui rend leur attribution à une unique cavité transverse ambiguë. D'autre part, elles créent des discontinuités dans le suivi de la géométrie des cavités et brouillent l'analyse cohérente du volume accessible aux résidus du site. Dans cette section, nous détectons ces cavités problématiques à l'aide de cavités de référence construites à partir des sites de référence. Nous proposons alors le découpage de la cavité et la réattribution de ses morceaux à leurs sites consensus respectifs.

3.9.1 Trajectoire de cavité de référence

Site de réf.	# Résidus	Volume moyen (std.dev./ min/max)	Moy. (max) # de morceaux par confs.	# de confs. vides
Poche D. (Myo)	7	69 (33/13/233)	1,35 (3)	6
Xe1 (Myo)	8	36 (24/13/155)	1,03 (3)	247
Xe2 (Myo)	8	44 (30/12/208)	1,17 (3)	395
Xe3 (Myo)	10	82 (40/13/254)	1,11 (4)	60
Xe4 (Myo)	8	30 (17/13/97)	1,06 (2)	600
site1 (DENV :A)	17	210 (53/69/381)	2,12 (5)	0
site1 (DENV :B)	17	202 (62/30/404)	1,96 (5)	0
site2 (DENV :AB)	12	320 (79/110/527)	1,02 (3)	0
site2 (DENV :BA)	12	282 (67/76/547)	1,05 (3)	0
βOG (DENV :A)	21	224 (85/13/572)	1,68 (5)	0
βOG (DENV :B)	21	274 (92/66/589)	1,30 (4)	0
Imatinib (Abl1)	21	373 (158/67/1 062)	2,15 (6)	0
GNF-2 (Abl1)	14	158 (82/13/424)	1,67 (4)	61
SABC (EF)	16	261 (167/14/694)	1,20 (4)	3
Catalytic (EF)	9	480 (177/15/867)	1,08 (4)	238

Tableau IV.5 Statistiques des trajectoires de cavités de référence. Les conformations où aucune cavité n'est présente ne sont pas prises en compte dans le calcul des volumes ou du nombre de morceaux. Les volumes sont exprimés en angström.

Les trajectoires de cavités de référence ont été calculées pour chacun des sites de référence comme explicité en Matériels et Méthodes. Les statistiques de leurs volumes, du nombre de morceaux et de leur présence dans les conformations sont données dans le Tableau IV.5. La variance des volumes peut être relativement large :

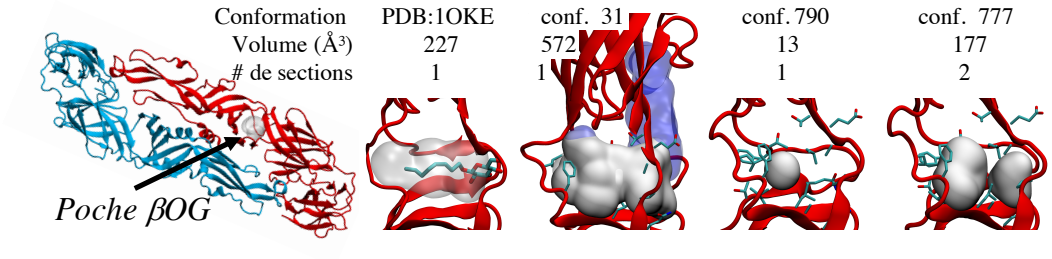


Figure IV.10 Diversité des géométries de cavités. Site βOG (volume gris, chaîne A de la protéine E de la dengue). La molécule βOG est montrée en sticks dans la structure 1OKE (en transparence de la cavité). Les volumes et nombre de morceaux des cavités de référence sont exprimés. Un exemple de fusion d'une cavité instantanée entre plusieurs sites est montré dans la conformation 31 (bleu transparent). En comparaison, la cavité de référence est confinée dans son site de référence.

de 17 Å^3 pour Xe2 dans la myoglobine à 158 Å^3 pour le site Imatinib d'abl1 et 177 Å^3 pour le site catalytique de EF (pour ce dernier cas la trajectoire est composée de la concaténation des formes actives et inactives de l'enzyme). Remarquablement, la géométrie des cavités varie largement comme cela peut être observé dans le site βOG de la protéine E (Figure IV.10). Indépendamment de leur taille et du nombre de résidus de la poche, les cavités apparaissent généralement en un seul morceau. Néanmoins, une segmentation en deux morceaux ou plus est régulièrement observée, en particulier pour la dengue/site1, βOG et abl1. Les sites de liaison larges, comme la poche βOG peuvent se scinder en 4 ou 5 morceaux grâce aux mouvements subtils des résidus de la poche, comme illustré dans le Tableau IV.5/p.préc. et la Figure IV.10. Les petites poches trouvées dans la myoglobine disparaissent régulièrement. Bien que plus large, un évènement surprenant de disparition complet de cavité est observé dans la poche βOG, une unique fois le long de la trajectoire.

3.9.2 Évaluation de la géométrie des cavités identifiées

Nous avons comparé les cavités transverses prédites par l'algorithme de suivi avec les trajectoires de cavités de référence. Pour cela, nous analysons leurs volumes au cours de la trajectoire, ainsi que leur distance d_{Geo} mutuelles. Un exemple de ces mesures est donné en Figure IV.9/p.114 pour le site βOG (chaîne B), sur un intervalle de temps restreint. Les volumes sont proches la plupart du temps, mais pour certains cas particuliers, le volume de la cavité prédite devient nul (c'est le cas de 31 intermédiaires sur 1000) ou devient invraisemblablement large (de plus de 200 Å^3 dans 24 conformations). De la même façon, d_{Geo} augmente sur ces cas pathologiques (à 11 Å au maximum pour cette trajectoire). La visualisation concrète de ces exemples révèle de larges cavités instantanées qui couvrent de multiples sites de la protéine. Un exemple en est donné dans la Figure IV.10. Pour ces cas, un assignement de la cavité instantanée à un unique site ne semble pas pertinent. Cela mène à l'introduction du concept de découpage de cavités instantanées.

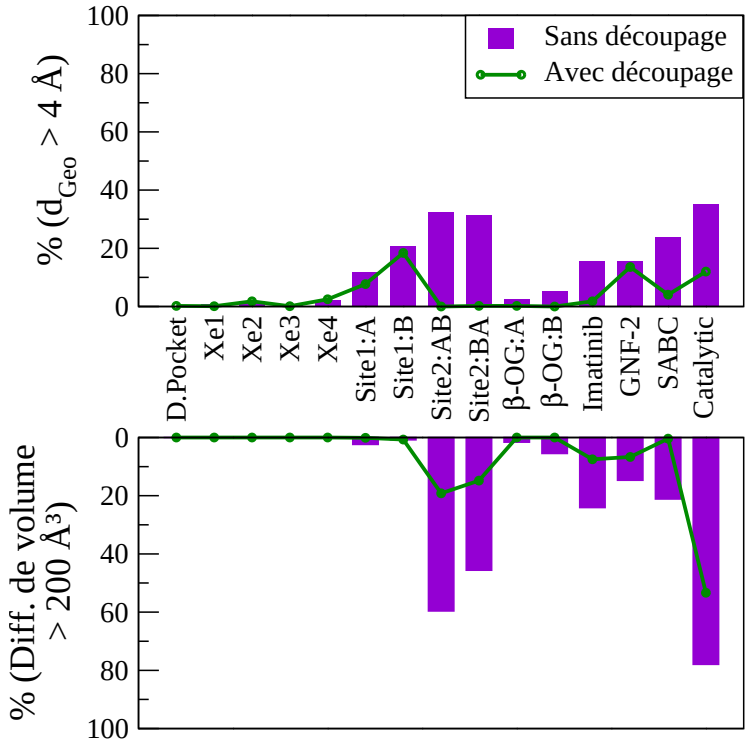


Figure IV.11 Pourcentage de conformations où les cavités présentent de larges déviations par rapport à la cavité de référence. Des exemples de grande dissimilarité géométriques (d_{Geo} > 4 Å, en haut) et de larges différences de volumes (>200 Å^3, en bas) sont rapportés. Les valeurs pour les cavités prédites sans découpage des cavités sont illustrées en violet, et en vert pour les cavités découpées.

Le découpage réduit largement les incohérences (Figure IV.9/p.114). La plus mauvaise différence de volumes pour le site βOG (chaîne B), 670 Å^3, est réduit à 21 Å^3. De plus, la différence moyenne de volume sur l'ensemble de la trajectoire passe de 31 Å^3 à 9 Å^3. Globalement, le volume des cavités s'accorde plus rigoureusement aux cavités de référence et cela pour la majorité des systèmes (voir la Figure IV.11). Les sites ambigus (Imatinib/abl1, catalytique/EF, et Site 2 de la dengue) gardent une délimitation approximative. Cela suggère une différence intrinsèque entre les sites de référence et la cavité trouvée dans la trajectoire.

3.9.3 Pertinence des sites prédits

Pour essayer d'évaluer à quel point les cavités identifiées par la méthode correspondent aux sites de référence, nous avons évalué, en utilisant le $F1$ score, la similarité entre la poche consensus (issus de la méthode) et la poche de référence. Puisque cette dernière est calculée par une empreinte *booléenne-par-résidus*, les empreintes *réelles-par-atomes* des poches prédites ont été converties dans ce format. Cela requiert la combinaison des empreintes par atomes en empreintes par résidus en sélectionnant un seuil de valeur. Nous avons choisi de combiner les empreintes

par union (voir Tableau IV.6). En testant différent seuils nous avons observé que les $F1$ optimaux sont obtenus autour de 2,5 ($\sim \sigma/2$) pour l'ensemble des sites. Par conséquent nous avons fixé ce seuil à $\sigma/2$.

Site (Cav./Ref.)	$F1$-Score	TP	FP	FN
Poche D. (Myo)	0,71	5	2	2
Xe1 (Myo)	1,00	8	0	0
Xe2 (Myo)	0,89	8	1	1
Xe3 (Myo)	0,86	9	1	2
Xe4 (Myo)	0,95	9	1	0
site1 (DENV :A)	0,71	12	5	5
site1 (DENV :B)	0,87	14	1	3
site2 (DENV :AB)	0,60	12	16	0
site2 (DENV :BA)	0,60	12	16	0
βOG (DENV :A)	0,92	18	0	3
βOG (DENV :B)	0,93	19	1	2
Imatinib (Abl1)	0,76	16	5	5
GNF-2 (Abl1)	0,62	9	6	5
SABC (EF)	0,67	10	4	6
Catalytique (EF)	0,36	9	32	0

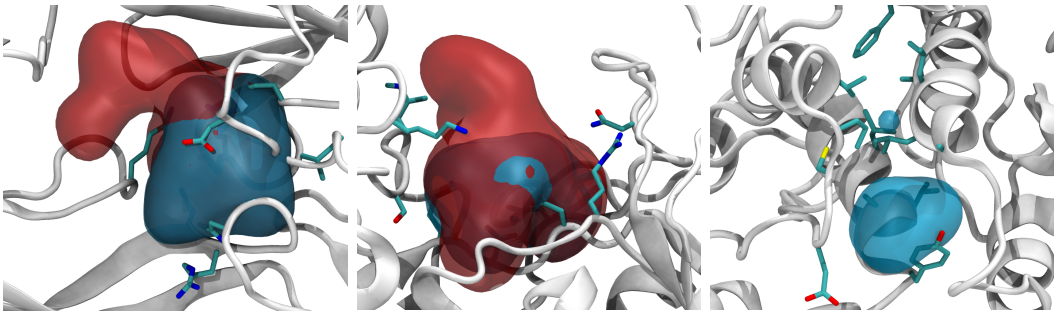
Tableau IV.6 Comparaison (F1-score) entre les sites de référence et les sites prédits. Pour chaque site, l'empreinte moyenne a été convertie en "résidus" booléen, le format des poches de références (fp_{ref}) : $fp^{bool}_{res} = \bigcup_{a \in res}(fp^{real}_{a} > \sigma/2)$ est calculé comme défini en Matériels et Méthodes pour comparer ces empreintes. TP, FP et FN sont donnés en nombre de résidus.

Il est intéressant de voir que 10 poches consensus sur les 15 étudiées sont très similaires à la poche de référence (sites de la myoglobine, Site 1 et poche βOG de la dengue, Imatinib d'abl1 ont un F1-score supérieur à 0,70). Cela signifie que la méthode de suivi des cavités, sans aucun *a priori* sur la localisation ni la composition des sites, a réussi à redessiner les poches, en ne prenant appui que sur les cavités instantanées observées au cours de la trajectoire. Pour d'autres sites les résultats sont plus mitigés. En particulier, pour le site 2 (dengue), le site catalytique de EF, et un grand nombre de Faux Positifs sont présents. En d'autres termes, la poche de consensus décrit une poche plus grande que la poche de référence. À l'inverse, le site GNF-2 d'abl1 et SABC de EF ont un plus grand nombre de Faux Positifs : la poche consensus "manque" des résidus de la poche de référence. Ces résultats, discutés plus loin (Section 4.1/p.suiv.), suggèrent que la considération de la dynamique des protéines appelle à une définition étendue des sites.

4 Discussion

4.1 Analyse des sites difficiles

La nature élusive des cavités dans les dynamiques de protéine et la difficulté à les identifier dans certains cas ont été une surprise notable. L'approche développée dans ce travail a réussi à annoter les cavités instantanées de sorte à suivre continûment leur évolution. Dans certains cas les sites ont été jugés difficiles, en particulier les Sites 1 & 2 de la dengue, GNF-2 (abl1), sites EF. Nous proposons de visualiser ces sites par le biais de la cavité moyenne de la trajectoire de cavité de référence (calculée en ne considérant que les résidus de la poche) et l'assignement de cavités instantanées de référence (calculé en sélectionnant les cavités instantanées se trouvant dans le site). Les cavités moyennes, représentées dans la Figure IV.4/p.105, ont été tronquées à 50 % d'apparition (définition en Matériels et Méthodes 2.2.6/p.104).



Site 2 (DENV : AB) Catalytique (EF) GNF-2 (Abl1)

Figure IV.12 Comparaison visuelle entre la cavité moyenne des cavités de référence $\overline{C}^{50\%}$ (en bleu) et la cavité moyenne des cavités instantanées assignées à la poche de référence pour chaque site $\overline{A}^{50\%}$ (en rouge). L'ensemble des cavités considérées pour A ne prend en compte que la plus petite cavité instantanée lorsque plusieurs sont assignées à une même conformation. Les cavités moyennes sont tronquées à 50 % d'apparition.

En d'autres termes, les points de grilles mis en valeur par les volumes rouges et bleus sont présents dans la trajectoire plus de la moitié du temps. Les Site 2 (dengue_AB) et catalytique (EF) sont frappants. On distingue clairement que les deux cavités moyennes ne se chevauchent pas totalement. En pratique, le lobe rouge qui déborde du volume bleu a un volume de 119 Å^3 pour le Site 2 (dengue_AB), et de 223 Å^3 pour le site catalytique (EF). Ainsi, dans plus de la moitié des conformations, le site contient une cavité instantanée significativement plus large que le site lui-même. Dans la trajectoire de EF, cela s'explique par la présence de la forme inactive dans la moitié des conformations de trajectoire dans laquelle le site catalytique est déstructuré (présence hétérogène de grosses cavités). Pour ces cas particuliers, ni le changement du seuil pour le partitionnement, ni le découpage des cavités (qui se

base sur la poche de référence) ne permettraient de retrouver les sites, car il n'y a pas consensus entre la dynamique des cavités instantanées et le volume recouvert par la poche. Cette observation est spécifique à ces 2 sites (voir Annexe VII.1/p.202). Le cas de la poche GNF-2 (abl1) est différent. On notera que dans la structure cristallographique (PDB : 3K5V) la molécule co-cristallisée dans le site, GNF-2, forme une cavité unique et que les résidus de la poche de référence ont été sélectionnés à partir de cette cavité. Le site est profond dans la protéine et la cavité moyenne de la trajectoire de cavité de référence (en bleu) montre deux volumes, autrement dit, la poche est en moyenne dans un état "écrasé", scindée en plusieurs morceaux au long de la trajectoire. Cela est confirmé par le Tableau IV.5/p.115 : 1,67 cavités de la trajectoire de cavités de référence sont présentes en moyenne dans le site. Nous avons calculé que dans les 1 000 conformations de la trajectoire d'abl1, seulement 77 cavités instantanées recouvrent les deux volumes simultanément. Le site prêt à accueillir GNF-2 n'est donc "ouvert" que dans à peine 10 % de la trajectoire. C'est insuffisant pour que le partitionnement attribue à ces 77 cavités une poche consensus couvrant entièrement le site. Cela explique pourquoi la poche consensus de la GNF-2 a beaucoup de faux négatifs (5) par rapport au site de référence (cf. Tableau IV.6/p.118 : les résidus prédits ne recouvrent que la moitié du site de référence et n'englobent qu'un des deux volumes de la cavité moyenne). Cela met en valeur un autre type de difficulté due au fait que les simulations ont été réalisées sans ligand (forme *apo*) alors que le site de référence est calculé dans sa forme *holo*. Dans le cas de la GNF-2 le site est déstabilisé et plus difficile à caractériser. Cela n'est pas le cas de la poche βOG par exemple, qui reste dynamiquement bien formée malgré l'absence de la molécule. En pratique, le site de la GNF-2 peut-être correctement caractérisé en diminuant artificiellement le nombre de groupes obtenu lors du partitionnement (modification du seuil de partitionnement) mais au détriment du suivi correct des autres sites : la meilleure méthode de partitionnement pour le site de la GNF-2 correspond à celle de Rang 779 dans le Tableau IV.2/p.111 et a un nombre de groupes deux fois inférieur à la méthode de Rang 1 (59 contre 127). Ces sites "difficiles" à caractériser soulignent les discordances entre une définition statique des sites, extraite d'une structure expérimentale fixe, et ce qui est observé dans un contexte dynamique.

4.2 Données de référence

Comme indiqué dans une revue récente (Krone et al. [220]), contrairement à d'autres domaines de recherche en bioinformatique (typiquement l'amarrage moléculaire et la mesure d'affinité de liaison ligand/protéine) il n'existe pas de données de référence consensuelles pour le calcul dynamique des cavités dans les protéines. En cela la publication de notre jeu de données de trajectoires de cavités de référence pourrait

permettre la comparaison objective de méthodes de suivi diverses. Pour un site donné, la trajectoire des cavités de référence contient les différentes géométries du volume inclus dans le site, sous forme d'ensemble de points de grille. Une méthode de suivi des cavités analogue qui utiliserait une représentation différente pour les cavités, par exemple des ensembles de sphères, pourrait utiliser nos métriques (mesure du volume, distance d_{Geo}) en traduisant leur représentation en point de grilles *mkgridXf* (pour les sphères, en ne considérant que les points de grilles *mkgridXf* présents dans au moins une sphère). Une comparaison juste avec la trajectoire de cavités de référence est alors possible (restreinte à des ensembles de points de grille).

Les limites de ces données de référence peuvent être discutées. D'une part, bien que dynamiques, ces données de référence sont construites à partir de sites de référence statiques. Les résidus sélectionnés pour les sites de référence sont pointés grâce aux ligands présents dans des structures cristallographiques, alors que les simulations sont faites sans la molécule d'intérêt (Xeon1-4 pour myoglobine, β-Octyl-Glucoside pour la dengue, Imatinib et GNF-2 pour abl1). L'influence de la molécule sur la conformation de la poche par rapport à sa forme libre est non négligeable, comme discuté plus haut (Section 4.1/p.119). Un autre facteur d'ambiguïtés se trouve dans la représentation sous forme de points de grilles des cavités instantanées. Le pas de la grille (0,5 Å) ne garantit pas une grande précision dans l'estimation du volume des cavités. Les analyses communément réalisées à partir d'un suivi de cavités tendent à évaluer des tendances (moyennes) d'évolution et en pratique le pas de $0{,}5$ est satisfaisant pour la plupart de nos applications (pour des analyses poussées de la géométrie des cavités, *mkgridXf* reste gérable en temps de calculs à $grd = 0{,}2$ Å). Une autre difficulté, propre à toutes les méthodes de détection de cavités, se trouve dans la délimitation entre la cavité et le solvant extérieur (*bulk solvent*), d'autant plus lorsque le système est tronqué aux uniques résidus de la poche. Dans *mkgridXf*, le paramètre *srou* est utilisé pour ronger les cavités en surface et s'assurer qu'elles sont suffisamment enfouies dans la protéine. Ce comportement a été neutralisé lors du calcul des trajectoires de cavités de référence lorsque la poche est enfouie grâce à la paramétrisation du $srou$ à 0 et la suppression des cavités satellites de la poche par comparaison avec une autre exécution où le $srou$ a une valeur standard.

4.3 Opportunités pour la recherche de nouveaux sites effecteurs

Dans ce travail, nous nous sommes focalisés sur la capacité de l'algorithme à identifier un petit nombre de cavités présentes dans les sites de référence. L'exemple le plus probant est celui de la myoglobine pour laquelle les 5 sites de référence sont clairement identifiés en même temps. Pour rappel, la classification des cavités

instantanées est globale, c'est-à-dire qu'en une seule itération du programme, ce ne sont pas seulement les cavités de ces 5 sites qui sont partitionnées, mais l'ensemble des cavités instantanées présentes dans les conformations de la trajectoire. En l'occurrence, 34 sites sont détectés dans la trajectoire de myoglobine, dont 18 cavités transverses *transitoires* qui apparaissent dans moins de 25 % des conformations. Le constat est similaire pour les 3 autres systèmes étudiés, avec 89, 127 et 107 apparitions transitoires pour abl1, la protéine E du virus de la dengue et la protéine EF de l'anthrax. Cette observation est en accord avec les résultats obtenus précédemment : au cours de la dynamique, des cavités peuvent spontanément émerger dans de nombreux *locus* de la protéine. Couplées à l'analyse mécanistique de la cible, ces cavités transitoires pourraient supporter de nouvelles stratégies de conception de modulateurs, par le biais de la caractérisation de sites allostériques. Par exemple, une molécule pourrait cibler une cavité apparaissant spécifiquement dans un état conformationnel de la protéine et en favoriser son état actif ou inactif.

Trajectoire	# Sites	Fréquence d'apparition de la cavité			
		# <= 1 %	# <= 25 %	# >= 75 %	# >= 99 %
Myoglobine	34	3	18	5	1
Abl1	123	12	89	6	0
Dengue E.	193	30	127	27	3
EF Anthrax	159	36	107	8	0

Tableau IV.7 Nombre de sites détectés et fréquence d'apparition d'au moins une cavité (> 12 Å^3) dans le site. Pour les 3 premiers systèmes 1 % d'apparition correspond à 10 conformations (sur 1 000 conformations) et 20 pour la protéine EF de la toxine de l'anthrax (sur 2 000 conformations).

Il est important de rappeler que l'algorithme présenté ne garantit pas que les sites détectés soient prêt à accueillir une molécule thérapeutique (sites de liaison dits *druggable*). Pour cela l'étude des propriétés physico-chimiques du site est un complément indispensable. On remarquera que notre spécification du site combine la cavité transverse et sa poche consensus associée. Les atomes/résidus à l'interface des sites sont donc connus et peuvent être une aide pour l'évaluation de propriétés déterminantes pour la liaison de petites molécules (hydrophobicité, aromaticité, conservation de la séquence, etc.). Une évaluation de la *druggabilité* est aussi envisageable par amarrage moléculaire et mesure d'affinité *in silico* des sites ainsi prédits, comme réalisé dans la Partie V/p.131.

4.4 Remarques sur l'implémentation *mkgridXf*

Le programme *mkgridXf* implémente la détection des cavités ainsi que les options de suivi des cavités optimisées dans le travail présenté ici. Des paramètres par défaut facilitent la prise en main de l'outil. En l'occurrence, les options "–pdb FILENAME",

"–dcd FILENAME" et "–tracking" sont suffisantes pour obtenir un premier résultat. Un *plugin* VMD permet de visualiser rapidement les trajectoires des cavités produites en choisissant la représentation VMD "resid" pour discerner les cavités transverses par des couleurs différentes. Un préalable au suivi des cavités et à la détection des sites est de vérifier que les paramètres de détection des cavités instantanées conviennent au système étudié. Typiquement, une réduction du rayon de la grande sphère représentant le solvant extérieur (option "–rou") va augmenter la segmentation des cavités en surface, et faciliter leur suivi, au risque de supprimer des cavités instantanées peu enfouies (exemple donné dans le chapitre suivant). Grâce à l'échantillonnage des cavités, *mkgridXf* peut prendre en charge des trajectoires de taille quasiment illimitée (le temps de détection des cavités est linéaire en le nombre de conformations à analyser). Le nombre de cavités échantillonnées fait augmenter les temps de partitionnement au cube, il est donc conseillé de tester une exécution avec un petit nombre, 3 000, 6 000, etc. (option "–msample 3000" par défaut) et d'interpoler un échantillonnage maximal, raisonnable en temps de calculs. L'étape de partitionnement hiérarchique est distribuée grâce à la libraire OpenMP, ce qui facilite l'augmentation du nombre de cavités échantillonnées (p.ex. 110 000 avec 20 CPUs au laboratoire et option "–nthreads"). Le principal risque d'un sous-échantillonnage est de manquer des cavités transitoires, qui seront alors réattribuées à des cavités transverses potentiellement éloignées, mais ce cas est assez rare sur des systèmes de tailles modérées. Le seuil de partitionnement est à 0,5 par défaut ("–thresh") mais peut-être modulé lorsque le suivi n'apparaît pas satisfaisant. Un seuil plus grand diminue le nombre de cavités transverses et forme des sites plus larges. Le rôle du découpage est de segmenter les cavités en morceaux réassignés à la poche consensus la plus proche. Le découpage des cavités est activé par défaut. Dans les cas extrêmes (protéines avec de longues cavités ramifiées) le découpage peut nuire à l'analyse de la trajectoire, auquel cas il peut être enlevé avec l'option "–nosplit" pour revenir à un assignement standard par cavités instantanées.

5 Suivi des cavités de la transition du récepteur nicotinique

5.1 Choix de la trajectoire de transition

Nous profitons ici du travail réalisé dans la Partie III/p.47 sur la description du mécanisme d'ouverture du canal ionique. La transition calculée lors de la série de *String of Swarms* S4 a été jugée comme la plus pertinente selon un certain nombre de critères (voir la Discussion 4.3/p.81). La trajectoire de transition entre l'état actif - canal ouvert, et l'état de repos - canal fermé contient 30 intermédiaires structuraux. Afin d'élargir la diversité conformationnelle de la transition, nous avons concaténé les groupes de structures de fin de trajectoires de *Swarms* de la 5^{e} itération SoS, soit un total de 30x32=960 conformations décrivant l'activation du récepteur. La trajectoire est ensuite désolvatée pour la détection des cavités. Le RMSD (tout atome) moyen après alignement sur la première est de 2,79 Å. Le RMSD entre la première et la dernière conformation est de 4,03 Å.

5.2 Ajustement des paramètres de détection de cavités

Les paramètres de *mkgridXf* ont été ajustés pour garantir une délimitation convenable des cavités. Un exemple est donné dans la Figure IV.13/p.suiv. avec la sélection du rayon de la grande sphère permettant de supprimer les points de grilles de cavités situés dans le solvant extérieur (*bulk solvent*). Lorsque la sphère est trop grande, une grosse cavité est détectée à l'intérieur du pore et agrège l'ensemble des petites cavités à la surface du pore, ce qui rend leur discrimination beaucoup plus difficile. Les paramètres de détection de cavités ont été choisis pour limiter cet artéfact tout en garantissant la présence de cavités dans les sites allostériques répertoriés dans une publication récente [36] (3 sites situés dans la partie extracellulaire, pointés dans des co-cristaux d'une chimère $\alpha 7$ d'AChBP). Ce travail a été réalisé par Noëlie Debs lors d'un stage de 3 mois dans le laboratoire. Le paramètre du rayon de la grande sphère rou a été choisi à 6 Å. Un effet de bord est que lorsque la conformation du canal est dans un état ouvert la cavité du pore n'y est plus détectée,. L'évolution du pore centrale n'est pas prise en compte dans le reste des analyses présentées pour ce travail.

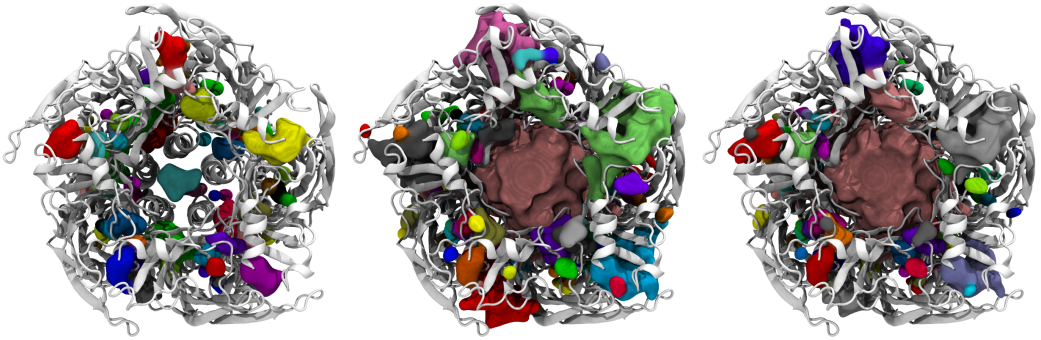


Figure IV.13 Effet du paramètre rou lors de la détection des cavités. Les cavités détectées dans cette conformation (canal en état fermé) ont été calculées avec une valeur du rou égale à 6, 8 et 10 Å. La variation du rayon de la grande sphère de détection de cavités ronge plus ou moins la cavité centrale et segmente les cavités à la surface du pore.

5.3 Prise en compte de la symétrie de séquence

Nous avons tenu compte de la symétrie de séquence du récepteur nicotinique (5 sous-unités $\alpha 7$) lors du partitionnement des empreintes. Du fait de la symétrie, des cavités peuvent apparaître dans des chaînes différentes (ou à l'interface entre deux chaînes) mais dans des poches chimiquement similaires. L'implémentation de la symétrie de séquence dans la méthode de suivi des cavités a été réalisée en modifiant la formule de calcul des distances entre empreintes. Pour deux empreintes *byres* (ou respectivement *byatom*) de cavités a et b, la distance $d(a,b)$ devient :

$$d_{sym}(a,b) = \min_{i=1..N} d(a, \text{shift}(b,i))$$

où N est le nombre de chaînes, et $shift(b,i)$ est obtenu en décalant les éléments du vecteur d'empreinte de i fois le nombre de résidus (ou d'atomes) par chaîne. Cette définition impose la contiguïté des atomes de chacune des chaînes dans le fichier PDB passé en entrée ainsi qu'une stricte identité entre les topologies des différentes chaînes (mêmes type d'atomes, même ordres des atomes), ce qui ne la rend pas applicable, par exemple, à des hétéromères. Lors du partitionnement, deux cavités dans des sites similaires mais dans des chaînes différentes auront une distance proche et pourront être regroupées ensemble. La Figure IV.14/p.suiv. montre un exemple de deux cavités transverses regroupant des cavités provenant de chaînes différentes. Après le partitionnement, cela implique aussi une modification du calcul de l'empreinte consensus. Pour chaque groupe d'empreintes G, l'empreinte $fp^{ref} \in G$ dont la contribution est la plus forte dans la première chaîne sert de référence. Alors :

$$fp^{cons}(G) = \frac{1}{\#G} \sum_{a \in G} \text{shift}(a,sel) \quad \text{avec} \quad sel = \arg\min_{i=1..N} d(fp^{ref}, \text{shift}(a,i))$$

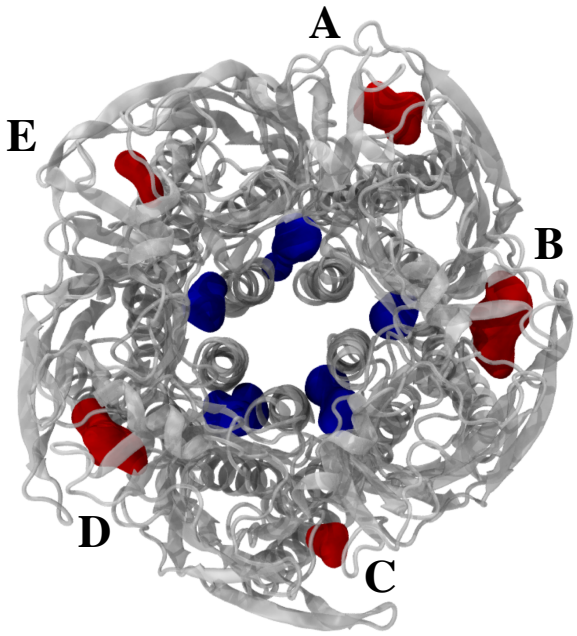


Figure IV.14 Suivi des cavités avec prise en compte de la symétrie de séquence. Bien que spatialement situées à des positions différentes du récepteur, ces cavités instantanées (rouge ou bleu) sont placées dans des sites similaires.

où $\#G$ est le nombre d'empreintes assignées au groupe G. Le réassignement des cavités non classées se fait par recherche de la cavité consensus la plus proche avec la distance d_{sym}.

5.4 Suivi des cavités et détection de sites

La structure du récepteur nicotinique possède 26 395 atomes. Pour accélérer le calcul de la matrice de distance lors du partitionnement des empreintes de cavités nous avons privilégié une définition des empreintes "par résidus" (*byres*) qui correspond à la méthode de Rang 14 dans le Tableau IV.2/p.111 (partitionnement "UPGMA", distance "cosine", seuil à 0,45). Toutes les empreintes présentes dans la trajectoire ont été considérées pour le partitionnement. En pratique *mkgridXf* s'exécute en 7,05 heures en étant distribué sur 18 CPUs.

88 541 cavités instantanées sont détectées le long de la transition. À la fin du partitionnement avec symétrie de chaînes, 68 sites ont été détectés, soit 340 (68x5) cavités transverses. Le découpage des cavités instantanées augmente légèrement le nombre de cavités à 90 395, soit en moyenne 94,2 cavités par conformations (minimum : 75, maximum : 116). La Figure IV.15/p.suiv. montre que les cavités apparaissent sur l'ensemble des régions de la protéine, que certaines cavités sont transitoires (colorées en rouge) et que d'autres sont persistantes le long de la

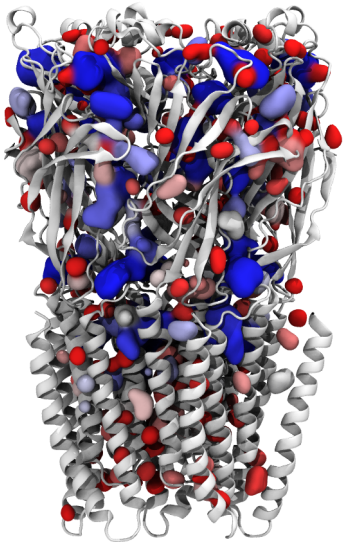


Figure IV.15 Cavités moyennes et leur pourcentage d'apparition. Le pourcentage d'apparition des cavités moyennes (seuil à 25 %) des 68x5 cavités transverses détectées le long de la transition est renseigné par un gradient de couleur de rouge (0 %) à bleu (100 %).

transition (colorées en bleu). Seule 51 cavités transverses apparaissent dans plus de 75 % des conformations de la trajectoire et à l'inverse, 179 moins de 25 % du temps, et 55 moins de 1 %. Cette mobilité des cavités s'accompagne de larges variations de leur géométrie. En fonction de l'état du récepteur, le volume accessible au solvant de certains sites peut varier de près de 100 $Å^3$, comme visible dans la Figure IV.16/p.suiv.. Sur les 68 sites, 10 ont un volume moyen supérieur à 100 $Å^3$ à au moins un moment de la transition.

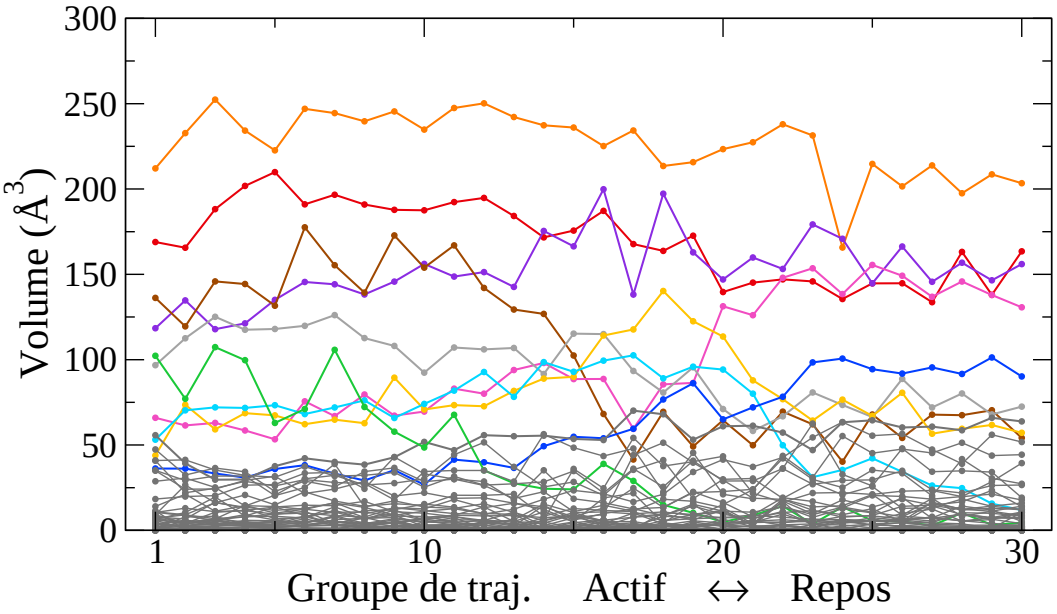


Figure IV.16 Volumes moyens des sites par groupe de conformations décrivant la transition. Les 10 sites dont le volume moyen est supérieur à 100 Å^3 sont tracés en couleur, les autres en gris. Chaque volume est moyenné sur 32x5 cavités localisées dans un même site (groupe de *Swarms* x symétrie).

6 Conclusions et perspectives

Dans cette partie, nous avons introduit une méthode de suivi dynamique des cavités dans les trajectoires de protéines pour la détection de sites. En répertoriant et en classant l'ensemble des cavités apparaissant dans les différentes conformations de la trajectoire en fonction des groupes d'atomes qui les entoure, il est possible de définir le contour des régions de la protéine présentant une interface cohérente avec le solvant. Dans 10 cas sur les 15 sites testés, nous avons pu retrouver grâce au suivi dynamique des cavités une description du site proche de celle spécifiée par des complexes cristallographiques ligand/protéine ou décrite dans la littérature. Les cas les plus difficiles ont mis en valeur des différences intrinsèques entre la vision statique des sites définis par leur structure expérimentale et la vision dynamique observée dans la dynamique moléculaire.

Ce travail souligne aussi l'extrême variabilité des cavités au cours des simulations moléculaires. Des espaces de volumes significatifs se créent et évoluent suivant le déplacement fin des atomes de la protéine. Comme montré dans une publication récente au laboratoire [9], l'évolution des cavités est étroitement liée aux mouvements fonctionnels des protéines. Leur analyse détaillée ouvre de nouvelles voies pour mieux comprendre les mécanismes moléculaires en jeu et pourrait faciliter la recherche de nouvelles molécules thérapeutiques. Deux exemples concrets d'études de la dynamique des cavités sont donnés en Figure IV.17/p.suiv. : la sélection de conformations diverses de poches préalable à un *docking* moléculaire et l'étude des réseaux de cavités. Deux délivrables conséquents ont été produits durant cette thèse. D'une part, le logiciel *mkgridXf* qui implémente le protocole de suivi des cavités décrit dans les chapitres précédents. L'outil se charge de la détection et du partitionnement des cavités. Lors de son développement, l'accent a été mis sur les performances et la simplicité d'utilisation. Des paramètres par défaut (dérivés de ces travaux) ont été mis en place pour obtenir simplement un résultat de suivi pertinent sur des trajectoires relativement larges telles que celles utilisées pour décrire la transition du récepteur nicotinique (plus de 26 000 atomes et environ un 1 000 conformations). Un programme compagnon, *mkread*, et un *plugin* Python, *PyMkgrid*, simplifient la manipulation et l'analyse des fichiers de cavités. De plus, un *plugin* VMD permet la visualisation directe des trajectoires de cavité. Ce programme sera prochainement mis à disposition de la communauté scientifique. D'autre part, nous proposons un ensemble de sites et de trajectoires de cavités de références pour 4 systèmes de tailles variées : la myoglobine, le facteur œdématogène de l'anthrax, la protéine d'enveloppe E de la dengue ainsi que la tyrosine kinase abl1. Indispensables

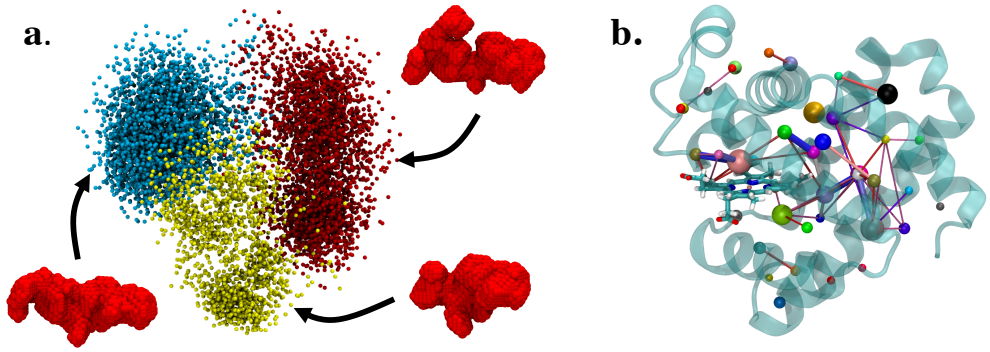


Figure IV.17 Exemples d'analyse de la dynamique des cavités. **a.** Sélection de poches à la géométrie diverse. L'ensemble des cavités d'un même site est extrait d'une trajectoire moléculaire. Une analyse en composantes principales sur la géométrie des cavités [9] permet d'extraire les principaux modes de variation des volumes considérés. Après projection (ici sur les 3 premiers modes d'une analyse en composantes principales), les cavités peuvent être triées en différents groupes par similarité de forme. La sous-sélection de cavités représentatives de chacun de ces groupes (critères énergétiques et autres) simplifie la prise en compte exhaustive de la dynamique du site, par exemple, pour réaliser une campagne de criblage virtuel par *docking* moléculaire. **b.** Analyse du réseau de cavités de la myoglobine. Un suivi des cavités avec découpage permet de discerner deux cavités présentes dans des sites distincts mais spatialement jointes. Un graphe dont les nœuds représentent chaque site et dont les arrêtes sont pondérées par le nombre d'évènements de fusion entre les sites peut alors être construit (cf. les billes et les tubes dans la figure). Sur de longues dynamiques de protéine il devient possible, par exemple, de mesurer la probabilité d'accessibilité d'une poche enfouie avec le solvant extérieur.

à l'évaluation de notre méthode, ces données de référence pourraient être utilisées pour comparer la pertinence des futurs outils de suivi de cavités.

Finalement, dans le chapitre précédent, nous avons pu appliquer la méthode de suivi des cavités à la transition du récepteur nicotinique. Comme pressenti, les cavités y sont nombreuses et labiles. 68 sites ont été répertoriés et seront utilisés comme support pour l'amarrage moléculaire de molécules effectrices. Nous verrons de plus que les variations moyennes de la géométrie des poches au fil du mouvement fonctionnel peuvent nous aider à identifier des sites potentiellement effecteurs du récepteur.

V

Recherche de sites effecteurs du récepteur nicotinique

Les mécanismes de modulation du récepteur nicotinique sont encore peu connus. Bien qu'un nombre important de molécules effectrices ait été découvert, leurs modes d'action, en particulier lorsqu'elles se situent hors du site principal orthostérique, sont obscurcis par le manque de structures expérimentales en complexe avec la protéine. Nous proposons dans cette partie d'utiliser les modèles de transition construits dans les parties précédentes comme support pour l'amarrage moléculaire de petites molécules effectrices. L'estimation *in silico* de l'affinité ligand/récepteur pourrait alors nous aider à discerner un site de liaison probable. Cette approche a été calibrée sur un ensemble divers de molécules de référence pour lesquelles le site de liaison est déjà connu. L'analyse de l'évolution des cavités lors de la transition allostérique $\alpha 7$ nous a permis d'identifier 6 sites potentiellement impactés par le changement conformationnel, parmi lesquels 4 sont déjà la cible de modulateurs allostériques dans des récepteurs homologues. Une poche située dans la partie transmembranaire du récepteur, à l'interface des sous-unités adjacentes, est prédite comme site de fixation de 4 modulateurs allostériques. Ces résultats, en accord avec les données expérimentales disponibles, ouvrent la voie au dessin et à l'optimisation de nouvelles molécules thérapeutiques ciblant le mécanisme d'activation du récepteur.

1 Introduction

Les récepteurs nicotiniques peuvent être modulés par de nombreuses molécules associées à une multiplicité de réponses fonctionnelles. Les agonistes activent le récepteur, les antagonistes bloquent le passage des ions, les modulateurs allostériques altèrent subtilement les mécanismes de transition entre états ouvert, de repos ou désensibilisé. Nous considérerons ici les modulateurs du récepteur de sous-unités $\alpha 7$.

L'acétylcholine est l'agoniste endogène de tous les sous-types de récepteurs nicotiniques. La fixation de l'acétylcholine dans un ou plusieurs sites orthostériques provoque l'activation et l'ouverture associée du canal du récepteur. D'autres molécules naturelles ont de telles propriétés, comme la cytisine, la nicotine et l'épibatidine. Des agonistes synthétiques ont aussi été découverts : le PNU-282987 est un agoniste spécifique du récepteur $\alpha 7$ [246, 247], A-844606 agoniste partiel [248], tout comme GTS-21 qui est aussi antagoniste faible des récepteurs $\alpha 4\beta 2$ [249]. Les antagonistes compétitifs bloquent l'activation du récepteur en se fixant dans le site orthostérique en lieu et place des molécules agonistes. L'inhibition produite est alors plus ou moins prononcée en fonction de la concentration d'agonistes présente simultanément lors la mesure d'activité du récepteur. La méthyllycaconitine (MLA) est un antagoniste compétitif réversible et spécifique du récepteur $\alpha 7$ [250]. À l'inverse, l'α-bungarotoxin a une très forte affinité pour les nAChRs et met ainsi plus de temps à se dissocier du site orthostérique (antagoniste quasi irréversible) [251]. Une autre catégorie de molécules effectrices comprend les modulateurs allostériques. Les PAM (*Positive Allostérique Modulateurs*) augmentent la réponse du récepteur aux agonistes. Les PAM dits de type I, tels que NS-1738, CCMI, ou encore l'ivermectine facilitent l'activation du récepteur mais n'ont pas d'impact sur la vitesse de désensibilisation. Les PAM de type II comme le PNU-120596 ou le TQS inhibent la désensibilisation du récepteur ce qui a pour conséquence une augmentation prolongée de l'efficacité des agonistes. On notera aussi l'existence d'antagonistes non-compétitifs (NAM, *Negative Allosteric Modulators*) qui permettent le blocage des ions indépendamment de la concentration de l'agoniste en se fixant dans des sites distincts du site orthostérique comme la mecamylamine qui se fixe probablement dans le pore [252]. Des molécules se liant dans des sites allostériques mais dénuées d'effets modulateurs (SAM, *silent allosteric modulator*) ont aussi été découvertes et peuvent modifier les cinétiques de liaison d'autres modulateurs allostériques [253]. Une revue complète des modulateurs du récepteur nicotinique est donnée dans les références [254] et [17].

Malgré cette large palette de molécules effectrices, très peu des molécules ayant pour cible principale les récepteurs nicotiniques ont jusqu'ici été approuvées par les autorités de régulation de médicaments [255]. La place déterminante du récepteur dans la fine régulation de la transmission synaptique rend difficile le développement de nouvelles molécules non toxiques pour l'organisme [256, 257]. L'absence de données structurales et le manque de connaissance du mode d'action des molécules connues sont un frein à la découverte et à l'optimisation de nouveaux effecteurs.

Nous explorons dans cette partie la propension de nos modèles de transition à pouvoir lier des molécules effectrices du sous-type $\alpha 7$. Nous utilisons les sites précédemment définis par le suivi dynamique des cavités comme support pour placer et mesurer l'affinité *in silico* de ces molécules (algorithmes de *docking*) dans l'ensemble des conformations de poches définies par les intermédiaires structuraux de la transition. L'analyse différentielle des poses devrait permettre de localiser le ou les sites d'interaction les plus probables.

Ce travail doit nécessairement prendre en compte les spécificités du récepteur et des modèles de transition :

1. 90 395 cavités (réparties dans 68 sites) ont été détectées parmi les 960 conformations du chemin de transition, soit autant de *docking* à exécuter par molécule.
2. L'interaction principale permettant la liaison des effecteurs dans la poche orthostérique est assurée par une somme d'interactions cation-π [37].
3. Les modèles conformationnels sont issus de simulations moléculaires.
4. Des environnements moléculaires très différents, aqueux et lipidiques, entourent les différents domaines du récepteur.

Le premier point nous contraint à choisir des algorithmes de *docking* rapides faisant nécessairement des concessions sur la précision des fonctions de score de l'affinité ligand/protéine. En particulier, les interactions cation-π, soulevées par le point (**2.**) sont très souvent omises par ces algorithmes. Le chapitre 2 décrit la prise en compte de l'interaction cation-π dans le programme de *docking* LeadIT-FlexX [258]. Dans le chapitre 3, une approche de *docking* global, sur l'ensemble des conformations et des sites de la transition, est mise à l'épreuve sur un ensemble d'effecteurs dont le site de liaison est suggéré par l'existence de poses cristallographiques dans des récepteurs homologues. Beaucoup de méthodes de mesure d'affinité de liaison sont calibrées à partir de structures obtenues par co-cristallisation (**3.**) et pour des complexes solvatés en milieux aqueux (**4.**). Pour identifier la configuration optimale pour le récepteur nicotinique, 4 programmes de *docking*, ainsi que 9 fonctions de score d'affinité ont été confrontées aux données de référence. Finalement, dans le chapitre 4, nous analysons l'ensemble des sites observés dans la dynamique de transition $\alpha 7$. Les sites

significativement impactés par le changement conformationnel pourraient avoir un effet de levier allostérique lorsqu'ils sont liés à une petite molécule. De plus, le chapitre 4 rend compte des sites de liaisons prédits pour des modulateurs allostériques dont la localisation est largement débattue par la communauté scientifique.

2 Ajout de l'interaction cation-π au programme FlexX

Le site orthostérique des récepteurs nicotiniques, comme l'ensemble des récepteurs à boucle Cys, est riche en résidus aromatiques [259]. La configuration de la poche spécifique au pentamère de sous-unité $\alpha 7$ est relativement bien connue grâce aux structures cristallographiques d'Acétylcholine Binding Proteins (AChBP) dont la partie extracellulaire est partiellement mutée pour ressembler à l'$\alpha 7$ (p.ex. [35, 36, 85]) : deux tyrosines sur la boucle C, une autre sur la boucle A, un tryptophane sur la boucle B et un deuxième sur la boucle D de la sous-unité complémentaire. Cette "boîte aromatique" joue un rôle critique dans la reconnaissance des ligands par le site [37, 260–262]. Typiquement, un groupe ammonium quaternaire chargé positivement porté par le ligand interagit de façon non covalente avec les électrons délocalisés des groupements aromatiques. La contribution énergétique d'une telle interaction se situerait entre 1 et 4 kcal/mol, ce qui la rend non négligeable et comparable à des liaisons hydrogènes ou à un pont salin [263].

L'estimation rigoureuse *in silico* de la contribution associée à l'interaction cation-π nécessite des calculs de mécanique quantique [261]. Cependant la complexité des calculs rend impossible l'évaluation énergétique des systèmes moléculaires complexes. Dans les champs de force de dynamique moléculaire non polarisables (charges fixées sur le centre des atomes), la contribution des interactions cation-π est généralement prise en compte par le potentiel électrostatique. Par exemple, dans le fichier de topologie du champ de force CHARMM36, le noyau aromatique des tyrosines est constitué d'un anneau de carbones chargés négativement entouré par des hydrogènes chargés positivement. Cette paramétrisation produit un moment quadrupole qui approxime l'interaction avec les cations [264, 265]. Lorsqu'il s'agit d'évaluer rapidement l'affinité de petites molécules dans des dizaines de milliers de poches, des fonctions de score empiriques sont plus communément employées. Ces fonctions sont usuellement représentées comme la somme pondérée des contributions majeures de l'affinité (nombre de liaisons hydrogènes, d'interactions lipophiles, décompte des collisions entre les atomes, etc.) et ignorent le plus souvent les interactions sous-représentées dans les données structurales expérimentales comme les cation-π [266]. Ainsi, de tous les outils de *docking* testés (Autodock Vina, FlexX, DOCK6), aucun ne prend en compte explicitement les interactions cation-π.

Il s'agit dans cette section d'incorporer l'interaction cation-π dans le programme de *docking* FlexX. Pour simplifier cette paramétrisation, l'interaction est considérée comme étant unidirectionnelle : le cation est toujours porté par le ligand et l'anneau

aromatique par les résidus de la protéine cible. L'objectif est ici de proposer une paramétrisation basique adaptée à la boîte aromatique du récepteur nicotinique.

Ce travail (implémentation, paramétrisation et évaluation) a été réalisé avec Laura Ortega-Varga, doctorante au laboratoire.

2.1 Matériels et Méthodes

2.1.1 Implémentation technique

Les modifications apportées aux fichiers de configuration de FlexX sont rapportées en Annexe D.1/p.204. En résumé, un premier motif d'interaction est créé pour correspondre à un ion ammonium. Pour rendre l'interaction unidirectionnelle, les groupements ammoniums des acides aminés lysine et arginine sont exclus par le biais d'une sélection en notation SMARTS [267]. Les points d'interaction avec les motifs structuraux de la protéine sont représentés sur la surface d'une sphère de 4,5 Å de rayon centré sur l'atome d'azote. Le second motif d'interaction identifie les anneaux de 6 carbones des résidus phénylalanine, tyrosine et tryptophane ainsi que la fonction pyrrole de ce dernier. Les points d'interaction avec le ligand sont placés sur l'intersection d'une sphère de 4,5 Å de rayon et d'un cône d'ouverture égale à $2 \times 35°$ (voir Figure V.1/p.suiv.).

Lorsqu'un point d'interaction du cation est occupé par le centre du motif d'interaction des anneaux de carbone et réciproquement, alors une interaction cation-π est comptée et contribue pour 0,75 kcal/mol à l'affinité ligand/protéine (énergie pouvant varier en fonction de la distance entre les deux centres d'interaction). Dans les fichiers de configuration cette énergie est négative, et donnée en kJ/mol, soit -3,138 kJ/mol.

2.1.2 Jeu de données de référence

Le placement de molécules dans la poche orthostérique de canaux ioniques a été évalué à l'aide des structures de complexes ligand/protéine cristallographiques : des agonistes, la nicotine (NCT - nAChR $\alpha4\beta2$, PDB : 5KXI), la cocaïne (COC - AChBP *Aplysia Californica*, PDB : 2PGZ) et l'antagoniste MLA (MLA1 et MLA2 respectivement issus d'une chimère AChBP-$\alpha4\alpha5$ - données internes, et une chimère AChBP-$\alpha7$, PDB : 3SIO). On notera que chacune de ces 3 molécules possède l'ammonium chargé nécessaire à l'établissement d'une liaison cation-π. Deux interfaces ont été sélectionnées pour chaque molécule : les chaînes AB et DE pour NCT, AB et DE pour COC, ED

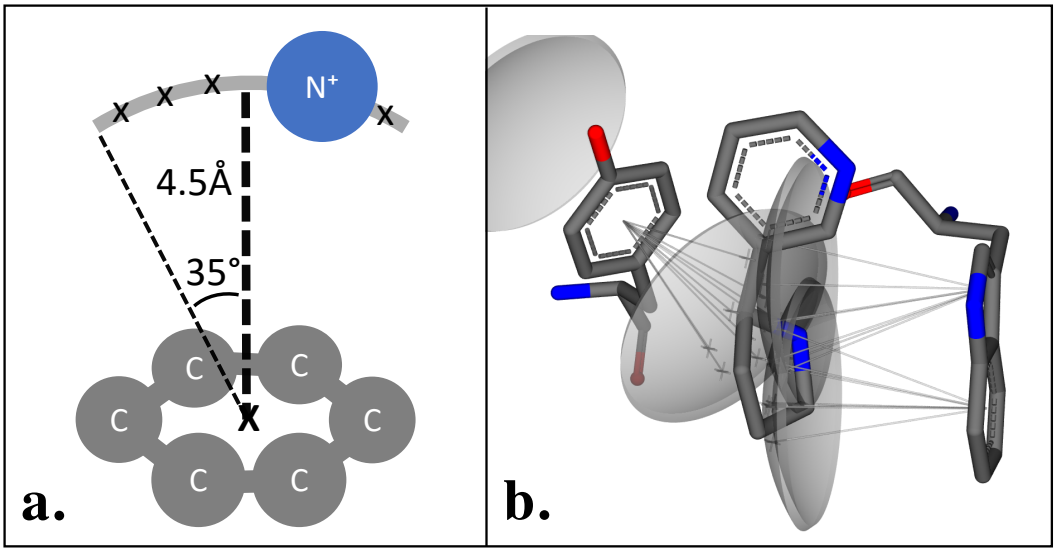


Figure V.1 Incorporation de l'interaction cation-π dans FlexX. a. Motifs structuraux définissant l'interaction cation-π. Lorsque le centre de la sphère d'interaction d'un cation du ligand s'aligne sur la section de sphère définie pour le groupement aromatique d'un résidu de la protéine, une contribution cation-π est ajoutée à la fonction de score FlexX. **b.** Cette pose de la nicotine dans la boîte aromatique d'une chimère $\alpha 7$ d'AChBP montre une double contribution cation-π du ligand avec un tryptophane et une tyrosine. Les petites croix placées dans les tranches de sphères définissent les points d'ancrage pour le positionnement de la molécule par l'algorithme de *docking*.

et BA pour MLA1 et JI et GF pour MLA2. Nous avons ainsi à disposition 8 conformations de la poche orthostérique ainsi que les poses produites cristallographiques associées. Parallèlement, la structure 3D des ligands a été extraite d'une base de données de fragments de modulateurs nicotiniques conçue au laboratoire par Laura Ortega-Varga. Nicotine, cocaïne et MLA y sont présentes en deux énantiomères différentes.

2.1.3 *Docking* avec FlexX

FlexX a été paramétrisé pour placer la nicotine, la cocaïne et le MLA dans leurs poches orthostériques cristallographiques. Les conformations réceptrices sont préparées avec la suite Leadit (v2.2.0), paramètres par défaut. La définition du site de liaison, en l'occurrence le site orthostérique, a été établie manuellement pour chaque système. Cette sélection recouvre au minimum l'ensemble des résidus conservés et déterminants pour la liaison dans le site (cf. figure 15.2 de la référence [268]). FlexX est utilisé avec sélection automatique des fragments de base ("selbas a"), placement standard des fragments de base ("placebas 3"), reconstruction de tous les fragments ("complex all") et ré-optimisation des 50 premières poses avec ré-ordonnement en fonction des scores ("optimize 1-50 5 1000 y"). Seules les 10 meilleures poses sont sauvegardées sur disque. Chacune des poses peut ensuite être comparée au ligand

cristallographique de référence par calcul de RMSD entre coordonnées atomiques, sans prise en compte des atomes d'hydrogènes.

2.2 Résultats

2.2.1 Calibration de la contribution

La paramétrisation d'une fonction de score de *docking* est un travail minutieux et complexe du fait de l'intrication des termes pris en compte. L'ajout d'un terme supplémentaire nécessiterait idéalement une repondération de l'ensemble des termes de la fonction empirique, ce qui n'a pas pu être fait ici. La philosophie privilégiée a été d'attribuer un poids minimal à cette nouvelle interaction mais suffisant pour obtenir des poses correctes. La Figure V.2 montre l'évolution de la qualité des poses

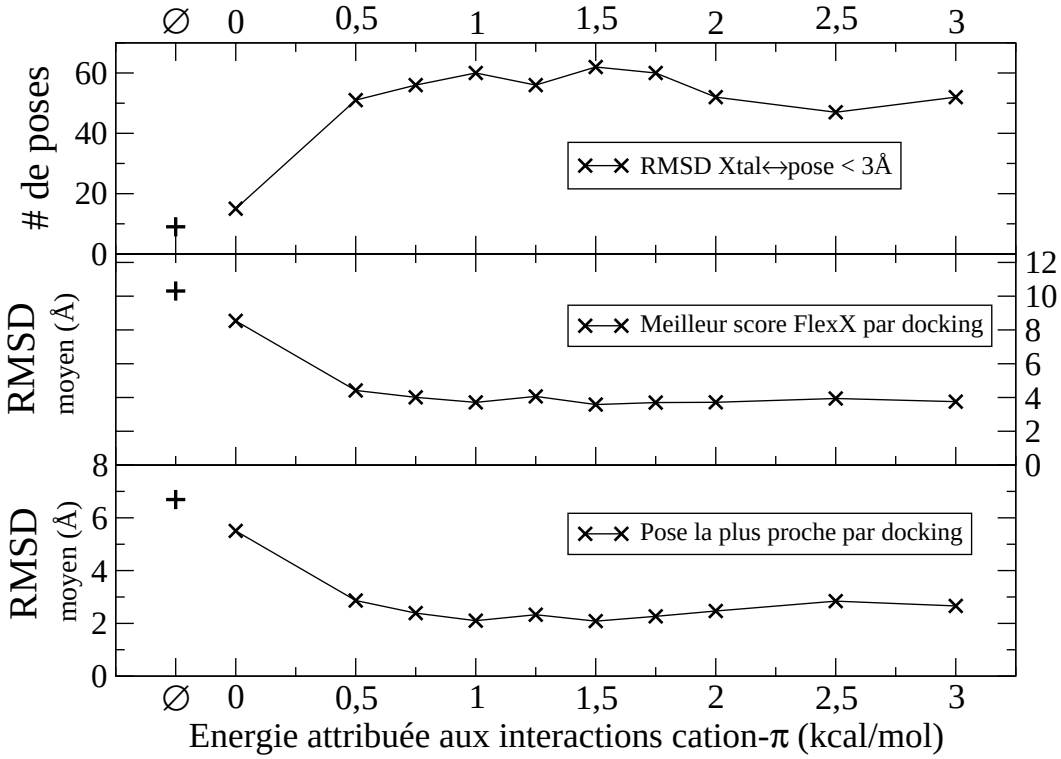


Figure V.2 Qualité des poses de *docking* en fonction de la contribution énergétique associée à l'interaction cation-π. L'abscisse $\emptyset$ est associée à une configuration de FlexX dépourvue de l'interaction cation-π alors que pour l'abscisse 0 l'interaction existe et a une énergie associée nulle. **Graphe 1 :** Nombre de poses (/160) dont le RMSD à la pose cristallographique associée est inférieur à 3 Å. **Graphe 2 (resp. 3) :** RMSD moyen sur les 16 *dockings*, de la pose de meilleur score FlexX (resp. de la pose la plus proche de la pose cristallographique) par rapport à la pose cristallographique.

obtenues en fonction de l'énergie attribuée à l'interaction cation-π. 160 poses sont récupérées pour chacune des 16 séries de *docking* (4 structures x 2 interfaces x 2 énantiomères x 10 poses). La première remarque concerne la différence observée

entre un *docking* avec la version originale de FlexX et FlexX avec cation-π mais dont la contribution énergétique est nulle (abscisses $\emptyset$ et 0 sur le graphe). Le nombre de poses à une distance inférieure à 3 Å de la pose cristallographique passe de 9 à 15 par la simple définition des motifs d'interaction cation-π. Ceci s'explique par l'ajout de nouveaux points d'interactions utilisés lors du placement du ligand. L'espace conformationnel exploré par les molécules est alors élargi et permet à FlexX de sélectionner des poses précédemment inaccessibles. L'augmentation progressive de la contribution énergétique s'accompagne d'une augmentation nette du nombre de poses satisfaisantes, passant de 15 à 62 poses (sur 160) à 1,5 kcal/mol. À 1 kcal/mol les poses les plus proches dans chacun des 16 *dockings* sont en moyenne à 2,1 Å, ce qui reste satisfaisant sachant que la moitié des poses proviennent du placement du MLA qui est une molécule relativement grosse (49 atomes lourds). À l'inverse, le RMSD moyen des poses de meilleur score est près de deux fois plus éloigné (3,7 Å à 1 kcal/mol), peu importe la valeur de l'énergie apportée par les cation-π. Cela remet en cause la capacité du programme de *docking* à trouver LA meilleure pose en se basant sur son score d'affinité FlexX et encourage l'utilisation complémentaire d'outils spécialisés pour réévaluer les poses de *docking*. Il est intéressant d'observer que lorsque la valeur associée à l'interaction est trop forte ($>$ 1,5 kcal/mol) le nombre de poses inférieures à 3 Å des poses cristallographiques tend à diminuer (passant de 62 à 1,5 kcal/mol à 47 pour 2,5 kcal/mol). On peut faire l'hypothèse qu'un poids trop fort placé sur l'interaction cation-π force le cation ammonium à être dans une position idéale et perturbe le placement du reste de la molécule. Nous avons choisi une énergie d'interaction à 0,75 kcal/mol, comme choix le plus neutre vis-à-vis du reste de la fonction de score, et garantissant un enrichissement satisfaisant des poses trouvées par FlexX dans la poche orthostérique du récepteur.

2.2.2 Enrichissement des poses satisfaisantes

La Figure V.3/p.suiv. montre la distance RMSD de chacune des 160 poses prédites à sa pose cristallographique, obtenues avec la version originale de FlexX et FlexX avec cation-π. Lorsque l'interaction n'est pas définie seul le système nicotine/$\alpha 4\beta 2$ montre des poses inférieures de 1 Å. Les autres systèmes obtiennent des poses extrêmement éloignées ($>$ 6 Å RMSD) ce qui montre l'incapacité de FlexX à explorer des régions de l'espace des poses compatibles avec les poses cristallographiques. Lorsque l'interaction cation-π est ajoutée, des poses très proches des co-cristaux ($<$ 1 Å RMSD) sont trouvées pour chacun des 4 systèmes, même pour la molécule MLA, dont le nombre de degrés de libertés est plus grand. D'autre part, une certaine diversité conformationnelle semble préservée puisque des poses éloignées sont encore présentes dans les *docking* MLA 1 et 2 et cocaïne.

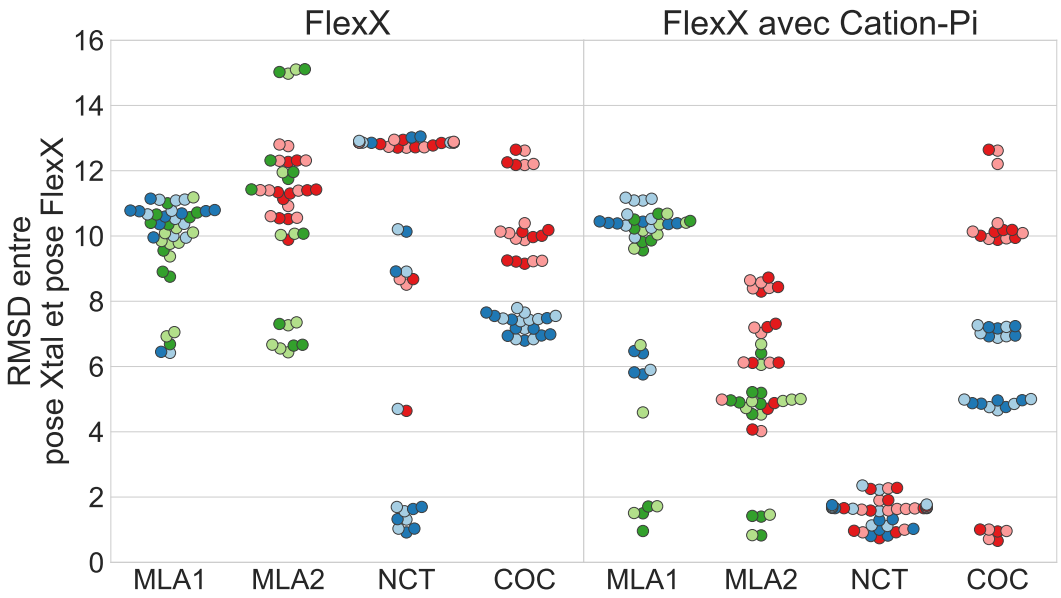


Figure V.3 Distance RMSD à la pose cristallographique pour chacune des 160 poses générées par les *dockings*. À gauche avec la version originale de FlexX, et à droite après incorporation de l'interaction cation-π (avec contribution énergétique à 0,75 kcal/mol). La couleur des points met en valeur les 10 poses d'un même *docking*, soit 4 couleurs différentes (2 interfaces x 2 énantiomères) pour chaque molécule.

2.3 Discussion

L'ajout de l'interaction cation-π au programme FlexX est prometteur puisqu'il permet de prédire des poses pertinentes qui n'étaient pas accessibles auparavant. Au-delà de l'addition du terme de l'interaction dans la fonction de score, c'est l'insertion de nouveaux points d'ancrage dans l'algorithme de recherche de FlexX qui permet d'atteindre des régions précédemment inexplorées de la boîte aromatique du récepteur nicotinique. Il faut rester prudent sur la généralisation de cette méthode à d'autres protéines puisque les tests présentés sont spécifiques à la poche orthostérique des récepteurs nicotiniques, ne comptent pas de contrôles négatifs (c'est-à-dire tester qu'une molécule connue pour ne pas se lier à la poche orthostérique ne s'y place pas) et n'incluent pas un nombre suffisant et divers de molécules. Ce travail sera utilisé dans le chapitre suivant pour le *docking* global sur la transition de l'α7. En particulier, d'autres analyses seront faites montrant la pertinence de FlexX avec cation-π pour le placement de ligand sur des conformations du récepteur non cristallographiques.

3 *Docking* global sur la transition

De petites molécules effectrices sont choisies et positionnées sur la totalité des conformations et des poches détectées dans la transition $\alpha 7$ par le suivi des cavités décrit dans la Partie IV/p.85. L'analyse des poses de meilleures affinités pourrait alors être une aide à l'identification d'un site de liaison probable, ainsi que des modes de positionnement privilégiés par la molécule. Le travail présenté dans ce chapitre dessine les contours de cette approche et tente d'en définir les limites actuelles et les améliorations futures nécessaires.

Le premier objectif est de définir un protocole de *docking* des molécules qui soit acceptable en temps de calculs. Ce *docking* s'applique aux 960 conformations des 68 sites détectés dans chacune des 5 chaînes du récepteur. En pratique, cela représente 90 395 poches du récepteur à traiter, et cela pour chacune des molécules testées. De telles quantités de calculs nécessitent de faire des concessions sur la précision des algorithmes d'amarrage moléculaire ainsi que sur les fonctions de score utilisées. Nous verrons en particulier, que la ré-optimisation et le *rescoring* des poses de molécules permettent d'améliorer la performance des résultats. FlexX, Autodock Vina, UCSF DOCK6 sont comparés pour la partie *docking*, et les fonctions de score Hyde, RF-score-vs, Cyscore, Smina, et XedMin pour le *rescoring* des molécules.

Le deuxième objectif est d'évaluer la pertinence de l'approche au vu des données structurales et outils disponibles actuellement de *docking/scoring*. Pour estimer la pertinence des *docking*, nous avons préparé une sélection diverse de molécules connues pour être effectrices sur le récepteur de sous-unités $\alpha 7$. Des agonistes (lobéline, acétylcholine, nicotine, épibatidine), un antagoniste (MLA), et des modulateurs allostériques (5 fragments [36], ivermectine, NS1738, PNU-120596, LY-2087101, CCMI). Les données structurales concernant ces molécules sont clairsemées. Pour notre recherche de sites de liaisons, la donnée la plus significative est le complexe ligand/protéine obtenu expérimentalement par cristallographie aux rayons X sur un récepteur homologue. Comme aucune structure cristallographique du récepteur de sous-unité $\alpha 7$ n'a été publiée, nous réalisons des comparaisons avec des récepteurs homologues. La lobéline ainsi que les 5 fragments allostériques ont été co-cristallisés sur une chimère AChBP (*lymnae stagnalis*) mutée à 71 % en récepteur $\alpha 7$ [36]. L'épibatidine est présente avec une chimère $\alpha 7$ AChBP (*lymnae stagnalis*) qui partage 64 % d'identité avec le récepteur $\alpha 7$ [35]. La molécule MLA est présente sur une chimère AChBP (*Aplysia Californica*) dont le site agoniste est muté en $\alpha 7$ [85]. La nicotine est disponible dans un cocristal avec le récepteur $\alpha 4\beta 2$ (nAChR Humain) [34]. Enfin, l'ivermectine est visible dans le récepteur GluCl (*spodoptera frugiperda*) [43].

Alors que le positionnement de l'agoniste lobeline sur la chimère $\alpha 7$ est probablement proche de celui d'un $\alpha 7$ humain, il faut être plus prudent pour les structures plus distantes comme le récepteur GluCl pour lesquelles ni le site de liaison, ni la pose ne peuvent être assurément similaires. Pour des modulateurs allostériques comme NS1738, PNU-120596 et LY-2087101 la localisation du site de liaison est encore floue, et reste l'objet de débats au sein de la communauté scientifique [78, 80, 269] bien que de nombreuses études supportent une localisation dans la partie transmembranaire [270–273].

3.1 Matériels et méthodes

3.1.1 Préparation des molécules

Les molécules proviennent d'une chimiothèque virtuelle développée dans le laboratoire par Laura Ortega-Varga en callaboration avec D.Joseph (Université Paris-Sud) et P.J. Corringer (Institut Pasteur, Paris). Chaque molécule de la base de données a été préalablement traitée par la succession des opérations suivantes :

- Génération des stéréoisomères (Corina Classic [274])
- Attribution de l'état de protonation, à PH=7 (MarvIn [275], Open Babel [276])
- Génération de la structure 3D (Corina Classic [274])
- Distribution des tautomères (Indigo [277], Marvin [275])
- Optimisation de la géométrie par minimisation énergétique (Chimera [278])

Chacune des molécules est convertie dans l'un des 3 formats utilisés par les outils de *docking* testés : format sdf pour FlexX (Open Babel), format PDBQT pour Autodock Vina (script "prepare_ligand4.py", MGLTools [279]), et mol2 pour UCSF DOCK6 (Chimera).

3.1.2 *Docking*

90 395 cavités sont détectées dans l'ensemble des intermédiaires structuraux de la transition $\alpha 7$ et répartis dans 68 sites. Les cavités sont utilisées pour guider le *docking* des molécules. Ainsi, 90 395 *dockings* sont nécessaires. Comme chaque *docking* peut-être réalisé indépendamment des autres, les calculs sont systématiquement distribués sur le *cluster* de calcul du laboratoire.

Pour chacune des poches, le centre et le rayon de 10 sphères maximisant la circonscription des points de grilles de la cavité associée sont calculés (centres déterminés

par partitionnement des points de grille avec la méthode des k-moyennes). Ces sphères sont utilisées lors des amarrages pour garantir que les poses trouvées chevauchent un minimum la cavité, ce qui facilite l'analyse des résultats site par site.

La paramétrisation de chacun des outils d'amarrage moléculaire est maintenant décrite et correspond au *docking* d'une molécule dans une des conformations d'un site, pour laquelle nous souhaitons générer et sauvegarder 10 poses.

FlexX [258] La poche du site actif (les résidus autorisés à former des interactions avec les ligands) est composée des résidus de la poche consensus (seuil "byres" > 3) et étendue aux résidus à moins de 3 Å d'un de ces résidus à au moins un moment de la trajectoire. La conformation globale du récepteur est aussi tronquée, et ne contient que les résidus à moins de 12 Å de la poche consensus à au moins un moment de la trajectoire. FlexX est exécuté avec les paramètres "selbas a", "placebas 3", "complex all". Une règle *FlexX-Pharm* est utilisée pour supprimer les molécules dont au moins un atome ne se trouve pas dans l'une des 10 sphères circonconscrivant la cavité. Les molécules restantes sont réoptimisées individuellement avec l'option "optimize 1-100 3 1000 n" et triées. Finalement, seules les 10 poses de meilleurs scores sont sauvegardées.

FlexX+cπ Paramètres identiques à ceux décrits pour FlexX et ajout des fichiers de configuration comprenant l'interaction cation-π définie dans le chapitre précédent.

Autodock Vina (x1 et x2) [280] La conformation entière du récepteur est préalablement convertie au format pdbqt grâce au script "prepare_receptor4.py" des MGLTools [279]. L'espace de recherche est inclut dans la plus petite boîte rectangulaire pouvant contenir la cavité et élargie de 5 Å sur chaque côté. L'exhaustivité de recherche (option "exhaustiveness") est paramétrée à 1 (x1) ou à 2 (x2), le nombre maximal de molécules à placer ("num_modes") à 100, et la graine aléatoire ("seed") est arbitrairement choisi à 42. Après *docking*, seule les 10 meilleures poses interceptant au moins une des 10 sphères circonscrivant la cavité sont sauvegardées.

UCSF DOCK6 [281] Comme pour FlexX, le récepteur est tronqué à 12 Å de la poche consensus de la cavité. Un script UCSF Chimera prépare le récepteur (ajout des hydrogènes, assignation des charges partielles, etc. par DockPrep [282]), et calcule sa surface moléculaire (fonction MSMSModel, après suppression des hydrogènes). Les sphères chevauchantes représentant l'image négative du récepteur sont classiquement calculées avec "sphgen", puis filtrées de telle sorte que seul les sphères chevauchant l'une des 10 sphères circonscrivant la cavité soient gardées. Une boîte rectangulaire contenant les sphères précédemment sélectionnées est calculée avec l'utilitaire "showbox" (avec marge supplémentaire de 5 Å). Ensuite "grid" calcule

la grille utilisée pour la fonction de score DOCK. Finalement “dock6” est exécuté pour placer de façon flexible le ligand dans le récepteur rigide et retenir les 10 meilleures poses selon le GridScore. Le GridScore est ici utilisé pour évaluer l’affinité des poses.

3.1.3 *Rescoring* des poses

Une fois que les 10 poses de *docking* ont été calculées pour chaque conformation de poche, nous réévaluons individuellement leur score d’affinité pour le récepteur grâce à des programmes de *rescoring*. Les noms de programmes précédés du signe “*” réoptimisent localement la géométrie des poses. Dans tous les cas de figure, la conformation du récepteur nicotinique considérée est celle utilisée pour le placement initial de la pose à évaluer. Pour une méthode de *rescoring* donnée et une molécule, la pose de “meilleur score” correspond à la pose dont l’affinité mesurée pour le récepteur nicotinique est la plus forte selon l’unité de mesure considérée.

*** Hyde [283]** Le programme Hyde est disponible dans la suite SeeSAR (en version 7 datée du 19/10/2017) et peut être utilisé en ligne de commande. La conformation du récepteur est tronquée comme expliqué pour les amarrages FlexX. Le score d’affinité est donné par un intervalle logarithmique entre deux K_i (en nanomolaires). Le milieu de l’intervalle a été choisi comme score.

RF-score-vs [284] L’exécutable RF-Score-VS 1.0 est disponible sur le dépôt Github du projet [285]. La conformation du récepteur entier est passée en paramètre. Le score “RFScoreVS_v2” est récupéré en sortie du programme.

Cyscore [286] Le binaire Cyscore a été utilisé en version 2.0. La conformation du récepteur entier est passée en paramètre.

*** Smina [287] - Vinardo [288] / Vina [280] / Autodock4 [279]** Smina est une réécriture du logiciel Autodock Vina 1.1.2. La version utilisée est celle publiée le 9 Novembre 2017 sur le site SourceForge à l’adresse renseignée en référence [289]. Plusieurs fonctions de score y sont implémentées : la fonction de score Vinardo [288], la fonction de score d’Autodock Vina [280], ainsi que la fonction de score d’Autodock 4 [279]. La conformation du récepteur entier est passée en paramètre avec l’option “–minimize” pour réoptimiser la géométrie du ligand dans la poche du récepteur.

*** XedMin [290]** Le programme XedMin (v3.2.0.24585) fait partie de la suite d’outils XedTools3.0 fournie par la société Cresset en version académique. Une conformation tronquée du récepteur a été préparée comme spécifié pour la configuration du

docking FlexX, suivi du calcul des charges formelles et protonation du récepteur et du ligand avec le binaire XedConvert. Le score fourni par XedMin correspond à l'énergie du complexe après convergence de l'algorithme de minimisation.

*** Xedmin repondéré (ρ)** Des pondérations alternatives du terme électrostatique du champ de force XED ont été testées. L'énergie totale du complexe ligand/protéine est donnée par la formule :

$$E_{\text{totale}} = E_{\text{élec.}} + E_{\text{VdW}} + E_{\text{liaisons}} + E_{\text{angles}}$$

où $E_{\text{élec.}}$ est un terme électrostatique (énergie de Coulomb), E_{VdW} l'énergie des interactions de Van Der Waals, E_{liaisons} et E_{angles} les termes d'interaction d'atomes liés, i.e énergie des liaisons et énergie des angles entre liaisons. Le détail du calcul de ces termes est peu documenté et ne sera pas décrit dans ce travail. Nous étudions l'influence d'un paramètre l_ρ sur l'énergie totale E_{totale} :

$$E_{\text{totale}} = E_{\text{élec.}} \times l_\rho + E_{\text{VdW}} + E_{\text{liaisons}} + E_{\text{angles}}$$

$$\textit{avec : } l_\rho = \begin{cases} 1, & \text{Si } \overline{P_z} \text{ n'est pas dans la membrane} \\ \rho, & \text{Si } \overline{P_z} \text{ est dans la membrane} \end{cases}$$

où ρ est un réel positif, et $\overline{P_z}$ la moyenne arithmétique de la coordonnée z des atomes de la pose P. $\overline{P_z}$ est située dans la membrane si sa valeur est comprise entre -15 et 10 Å sur l'axe des z, ce qui correspond à la région transmembranaire dans la boîte de simulation des conformations du récepteur tronqué à 10 Å côté intracellulaire en raison de la présence de molécules d'eau à la proximité du pore. Bien que la valeur absolue E_{totale} ne soit plus cohérente avec la paramétrisation initiale du champ de force, nous n'étudions que sa valeur relative lors de la sélection de la pose de "meilleure score".

3.1.4 Molécules de références

Les co-cristaux utilisés pour calculer les poses de référence sont rapportés dans le Tableau V.1/p.suiv.. La pose cristallographique et les sous-unités spécifiées sont extraites du fichier PDB téléchargé de la Protein Data Bank. Si la molécule est à l'interface de deux chaînes, les deux sous-unités sont considérées. La structure obtenue est ensuite itérativement alignée sur les 5 sous-unités (ou doublets de sous-unités) des 960 conformations du récepteur $\alpha 7$ avec la fonction "align" de Pymol [291]. Pour chacune des 960 conformations de la transition, les 5 poses R_1,R_2,R_3,R_4,R_5 sont sauvegardées et servent de poses de référence.

Molécule	Id. MOL	Id. PDB	Chaîne(s)	Id. MOL	MOL Chaîne
lobéline	LOB	5AFH	A-B	L0B	A
methyllycaconitine	MLA	3SIO	A-B	MLK	A
acétylcholine	ACH	3WIP	A-B	ACH	A
nicotine	NCT	5KXI	A-B	NCT	A
épibatidine	EPJ	3SQ6	A-B	NCT	A
ivermactine	IVM	3RHW	A-B	IVM	A
42R	42R	5AFJ	A-B	42R	A
5VU	5VU	5AFK	A-B	5VU	A
9Z0	9Z0	5AFM	A	9Z0	A
FHV	FHV	5AFL	E	FHV	E
OJD	OJD	5AFN	A	OJD	A

Tableau V.1 Molécules de référence. L'identifiant des co-cristaux est donné, ainsi que les chaînes de la structure utilisée pour alignement de la pose cristallographique sur les conformations de la transition $\alpha 7$. L'identifiant Id. MOL est celui spécifié dans les figures.

La distance d'une pose de *docking* P à sa molécule de référence est égale à :

$$D_{ref}(P) = \min_{i=1,..,5} \delta(P,R_i)$$

où δ est le RMSD entre les coordonnées cartésiennes des deux poses.

Une distance aux molécules de référence a été calculée pour chacune des poses générées par les 4 programmes d'amarrage moléculaire. Elles sont recalculées après *rescoring* des poses si la fonction de score permet l'optimisation de la géométrie de la molécule.

3.2 Résultats

3.2.1 Poses et sites de référence

10 molécules ont été choisies afin d'évaluer l'efficacité des outils d'amarrage moléculaire pour replacer les molécules dans leurs sites de liaisons respectifs. Pour chacune d'entre elles, des informations structurales expérimentales (co-cristaux) viennent supporter un site de liaison, ainsi qu'un mode de fixation, sur des structures de récepteurs canaux homologues du nAChR $\alpha 7$. Ces 10 molécules couvrent un éventail varié de modes d'action, de sites de liaison et de poids moléculaires. La poche orthostérique est représentée par 5 molécules : 3 agonistes classiques, l'épibatidine (EPJ), la nicotine (NCT) et l'acétylcholine (ACH), ainsi qu'un agoniste partiel, la lobéline (LOB), et l'antagoniste Methyllycaconitine (MLA). Les 5 fragments publiés

par Spurny et al. [36], nous permettent d'incorporer des modulateurs allostériques. 42R est positionné dans une poche localisée légèrement en dessous du site orthostérique. 5VU et FHV sont dans une poche au niveau de l'hélice *alpha* N-terminale. Et finalement, 9Z0 et OJD se trouvent dans la poche vestibule. Les chimères $\alpha 7$ de Spurny et al. sont très proches de l'$\alpha 7$ sauvage. Les molécules précédemment citées ont été co-cristallisées sur des chimères $\alpha 7$ d'AChBP dont l'identité de séquence est très proche des récepteurs sauvages, bien que seule la partie extracellulaire du récepteur soit accessible. Ces 5 fragments ont été caractérisés comme étant des modulateurs allostériques négatifs des récepteurs $\alpha 7$ [36]. Enfin, le modulateur allostérique positif de type I, l'ivermectine (IVM), est situé dans une poche entre deux sous-unités dans la partie transmembranaire entre les hélices M3(+) et M1(-) chez le récepteur GluCl [43] et a une position similaire dans le récepteur Glycine [45]. En résumé, ce sont donc 5 poches qui peuvent être sondées avec ces molécules : la poche orthostérique (NCT, ACH, LOB, EPJ), la sous-poche agoniste (42R), la poche vestibule (9Z0, OJD), la poche N-terminale (5VU, FHV) et la poche présumée de l'ivermectine (IVM). Pour avoir une idée de la pertinence des amarrages moléculaires sur la transition $\alpha 7$, nous supposons que ces poses cristallographiques se transposent par homologie de séquence au récepteur $\alpha 7$ et que d'autre part leurs alignements sur chacune des conformations constituent des poses correctes de référence.

3.2.2 *Docking* restreint aux sites étudiés

3 programmes d'amarrage moléculaire ont été testés : UCSF DOCK6, Autodock Vina, Leadit/FlexX. Bien que cela ne soit pas une liste exhaustive des outils existants, ces programmes ont déjà été éprouvés par le passé au laboratoire ce qui a facilité leur mise en place. Chacune des molécules a été amarrée dans chacune des conformations des 5 sites pour lesquelles une pose de référence existe, soit 5(site) x 5(symétrie) x 960(conformations) amarrages moléculaires possibles. En ne comptant que les poches dont le volume de cavité est non nul, cela correspond à 18 438 exécutions indépendantes par molécule, soit environ 20 % du nombre de *docking* nécessaire pour cibler les 68 sites du récepteur (90 395 conformations de poches). Les temps cumulés d'utilisation CPU sur le *cluster* de calcul pour chacune des méthodes sont de 89 jours pour Vina(x1), 168 jours pour Vina(x2), 55 jours pour FlexX, 49 jours pour FlexX+cπ et $\approx$ 600 jours pour DOCK6. Ces temps de calcul incluent la préparation de la poche réceptrice, le *docking* des molécules dans la poche, et la conversion des 10 meilleures poses de chaque *docking* vers un format commun. Ils varient de 10 à 50 % selon la vitesse des machines utilisées. On observe que le paramètre d'exhaustivité d'Autodock Vina augmente le temps de calcul quasi linéairement, comme annoncé par la documentation du programme. Nous en concluons aussi que la valeur par défaut du paramètre, 20 (au lieu de 1 ou 2), est quasiment inaccessible en un temps raisonnable. DOCK6 est handicapé

par les étapes de préparation du récepteur, qui sont nombreuses et coûteuses (p.ex. calcul de la surface moléculaire, de la grille de score). FlexX est ici la solution la plus rapide avec 49 ou 55 jours en fonction de l'ajout ou non de l'interaction cation-π dans la fonction de score. Il semble étonnant que la version incluant l'interaction cation-π soit plus rapide que celle sans. Il s'avère qu'en pratique FlexX+cπ a été adressé sur des machines plus rapides du *cluster* de calcul, ce qui donne un temps de calcul plus court. Ainsi, nous pouvons en conclure que l'ajout de l'interaction n'a pas un impact majeur sur les performances du programme.

3.2.3 Accessibilité des poses de référence

Plusieurs dizaines de milliers de poses sont ainsi enregistrées pour chaque molécule dans leur site de liaison associé. La question qui se pose naturellement est de regarder si au moins un certain nombre de ces poses sont proches des poses de référence. Si la réponse est non, cela signifie que l'outil de *docking* n'explore pas suffisamment l'espace des différentes poses de la molécule (manque de temps de calcul, problème de *scoring* des poses), ou bien qu'aucune des structures du modèle de transition $\alpha 7$ ne présente une conformation compatible avec la liaison de la molécule de référence. Cette information est donc particulièrement intéressante, avant même d'étudier la pertinence des scores d'affinité.

La Figure V.4/p.suiv. présente un résumé du décompte de ces poses dites "satisfaisantes" pour chacune des 5 séries de *docking*. Le critère "pose satisfaisante" est ici attribué aux poses dont la distance à la pose de référence est inférieure à 3 ou à 1,5 Å. Des molécules semblent plus faciles à replacer dans leur site de liaison que d'autres. Plus d'un millier de poses satisfaisantes (< 3 Å) sont trouvées pour l'acétylcholine (ACH) et la nicotine (NCT) avec les 5 outils de *docking* testés. Les poses les plus proches parmi tous les amarrages moléculaires sont respectivement à 0,77 Å (avec DOCK6) et 0,71 Å (avec FlexX) pour ACH et NCT (Figure VII.9/p.206 en annexe). Seul Autodock Vina ne trouve aucune pose ACH à moins de 1,5 Å. D'autres molécules sont logiquement plus difficiles à placer de par leur taille significativement plus large en nombre d'atomes, comme la lobéline et le MLA dont des poses sont trouvées à moins de 3 Å mais pas à moins de 1,5 Å. 3 poses de références semblent inaccessibles : l'ivermectine, 5VU et FHV n'ont aucune pose satisfaisante parmi les 5 *dockings*, et les poses les plus proches sont très éloignées de la référence (6,36 Å pour IVM et 3,97 Å e 3,40 Å pour 5VU et FHV). On ajoutera que DOCK6 ne réussit à placer aucune pose dans le site associé à la poche de l'ivermectine. De façon intéressante, un grand nombre de poses satisfaisantes (<3 Å) est trouvé pour les modulateurs allostériques 42R, 9Z0 et OJD, bien que les poses proches (<1,5 Å) soient plus rares. OJD est trouvée à 0,64 Å (DOCK6), 0,69 Å (FlexX) et 0,72 Å (FlexX+Cπ) de sa pose de référence dans les meilleurs des cas (Annexe VII.9/p.206).

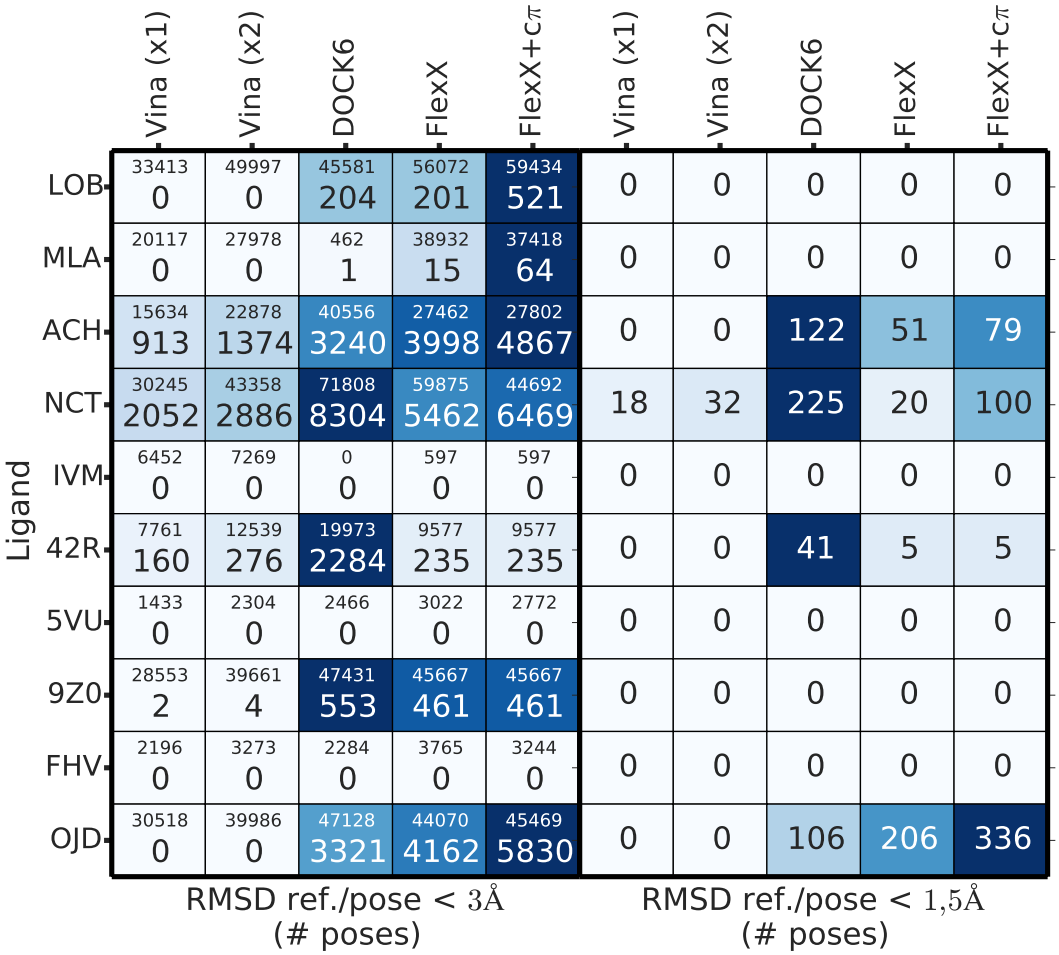


Figure V.4 Décompte du nombre de poses satisfaisantes pour chacun des outils d'amarrage moléculaire testé. En petit caractère, le nombre total de poses dans les sites considérés. Le tableau de gauche décompte les poses proches d'au moins 3 Å de la pose de référence, et le tableau de droite est restreint aux poses à moins de 1,5 Å.

FlexX+cπ démontre clairement son utilité pour placer les molécules dans le site orthostérique, avec une augmentation systématique du nombre de poses satisfaisantes par rapport à son homologue FlexX standard. L'exemple de la nicotine est intéressant. Le nombre total de poses passe de 59 875 pour FlexX à 44 692 pour FlexX+cπ, alors que le nombre de poses satisfaisantes augmente de 5 462 à 6 469. Cette tendance facilite potentiellement le travail des fonctions de score qui auront à identifier un plus grand nombre de *bonnes* poses parmi un moins grand nombre de *mauvaises*. De façon surprenante, nous observons une différence notable de poses satisfaisantes à 3 Å entre les deux méthodes pour la molécule OJD. Cette molécule ne possède pas de cation et aucune contribution ne peut être attribuée à une interaction cation-π dans la fonction de score de FlexX+cπ. La différence est en fait due aux points d'interaction cation-π ajoutés aux résidus tryptophanes, phénylalanine et tyrosine. Ils servent de points d'ancrage supplémentaires pour de nouvelles poses qui bien qu'elles ne contribuent pas à l'interaction cation-π, élargissent l'espace des poses explorées par FlexX.

La comparaison entre toutes les méthodes de *dockings* fait ressortir les programmes DOCK6 et FlexX+cπ qui obtiennent systématiquement de meilleurs résultats que les 3 autres méthodes. Leurs décomptes de poses satisfaisantes ainsi que les distances à la pose la plus proche de la référence sont très similaires. Pour la suite de l'étude, nous privilégierons FlexX+cπ qui est plus pertinent sur les poses de la poche orthostérique, et près de 10 fois plus rapide à exécuter que DOCK6.

Pour finir, il est percutant de confronter le nombre de poses totales, avec le nombre de poses satisfaisantes. Avec un *docking* de l'acétylcholine dans la poche orthostérique par FlexX-cπ, seules 4 867 molécules ont une distance à la référence inférieure à 3 Å parmi 27 802 poses dans l'ensemble restreint de site et 389 879 sur l'ensemble du récepteur. Cela montre l'ampleur du défi posé aux fonctions de score d'affinité ligand/récepteur pour discerner les quelques poses satisfaisantes parmi le grand nombre de poses amarrées sur la totalité du récepteur.

3.2.4 *Docking* global et *rescoring* des poses

Le *docking* global consiste à amarrer une molécule sur l'ensemble du récepteur, pour l'ensemble des conformations du modèle de transition. Dans un cas idéal, la ou les poses de meilleures affinités devraient nous aider à discerner d'une part le site de liaison et d'autre part le mode de fixation de la molécule sur le récepteur. Dans cette section, nous explorons la possibilité d'une telle approche en nous appuyant sur les molécules de référence précédemment décrites.

Un *docking* FlexX-cπ a été exécuté sur chacune des 90 395 conformations de poches identifiées dans la transition du récepteur α7. Cela représente un total de 550 jours CPU distribués sur les nœuds du *cluster* du laboratoire. Entre 200 000 et un million de poses sont stockées sur disque pour chacune des molécules. Chacune des poses s'est vue attribuer un score par la fonction de score FlexX-cπ. Nous pouvons alors étudier pour chaque molécule la localisation sur le récepteur nicotinique de la pose de meilleure affinité ainsi que sa distance à la pose de référence. En première approximation, seules les poses situées dans la partie extracellulaire du récepteur sont considérées. La Figure V.5/p.suiv. détaille les résultats obtenus pour la fonction de score FlexX-cπ (première colonne). Sur l'ensemble des poses de la lobéline dans la partie extracellulaire du récepteur, la pose de meilleure affinité est à 4,7 Å de la pose de référence. Bien que dans une position différente de la pose de référence, la pose de meilleure affinité se trouve bien dans la poche orthostérique, ce qui est un résultat positif. De la même façon, les meilleures poses de l'acétylcholine (ACH), la nicotine (NCT) et l'épibatidine (EPJ) se situent dans le site orthostérique, bien qu'à une distance supérieure de 5 Å de leur molécule de référence. À l'inverse, aucun des modulateurs allostériques étudiés n'est prédit dans son site extracellulaire respectif.

Plusieurs éléments viennent obscurcir l'analyse de ces résultats. Premièrement, la fonction de score FlexX-cπ a été optimisée pour pondérer la proximité du cation des agonistes testés dans la poche orthostérique. Il est difficile d'affirmer que la préférence de ces molécules pour cette poche n'est pas artificiellement produite par une surpondération de l'interaction. Deuxièmement, la fonction de score FlexX est construite pour pouvoir évaluer l'affinité de dizaines de molécules par seconde et fait nécessairement des concessions sur la pertinence physique des modèles d'interactions utilisés. Pour pallier ces défauts, 8 fonctions de score d'affinité ont indépendamment été utilisées pour évaluer l'affinité de l'ensemble des poses FlexX. Les fonctions de

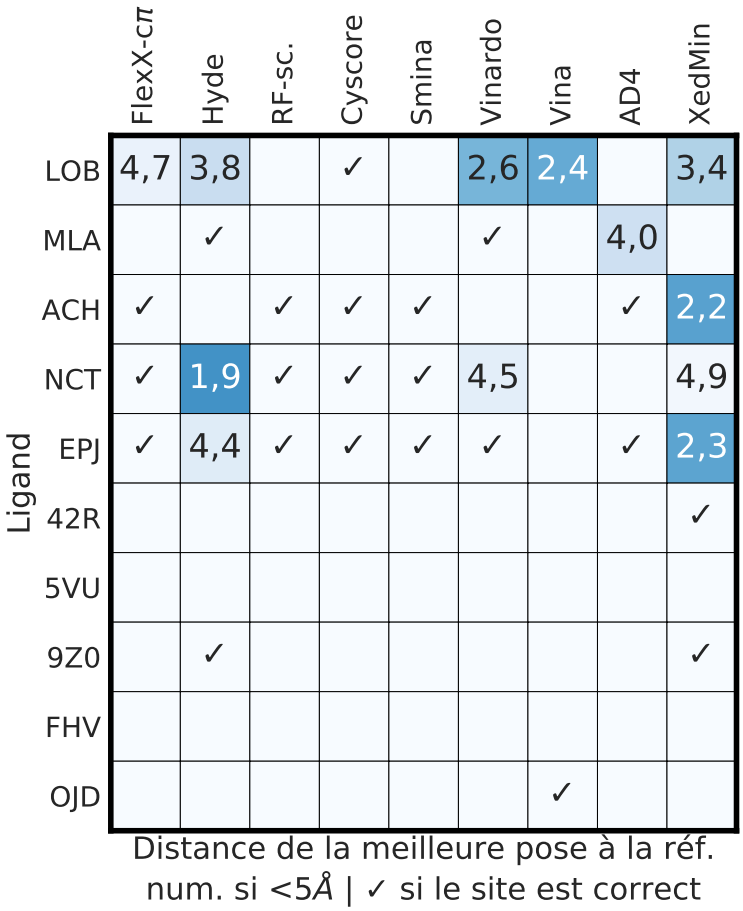


Figure V.5 Localisation de la pose de meilleure affinité selon les différents programmes de *rescoring*. Présence d'une valeur réelle (RMSD en angström) si la meilleure pose est à moins de 5 Å de la pose de référence, et ✓ si la pose est dans le bon site. La case reste vide dans tous les autres cas. L'analyse est ici restreinte aux poses situées dans la partie extracellulaire du récepteur nicotinique.

score ligand/protéine ont été choisies de par leur variété méthodologique, ainsi que leur performance supposée décrite dans la littérature. Les scores XEDmin sont basés sur un champ de force, Hyde, Vina, Smina et Vinardo sont des fonctions empiriques et le RF-score utilise des algorithmes d'apprentissage automatique. Autodock 4 et Vinardo sont décrits comme des champs de forces semi-empiriques. La Figure V.5 détaille les performances de ces 8 fonctions de score en rapportant les distances à la

pose de référence des poses de meilleure énergie après le *rescoring*. Clairement, les molécules se liant à la poche orthostérique sont plus facilement identifiées. FlexX-cπ, Hyde, Cyscore, Vinardo et XedMin attribuent le meilleur score d'affinité à une pose du site orthostérique pour 4 molécules sur 5. Hyde et XedMin trouvent des poses proches de leur référence pour la nicotine (1,9 Å) et pour l'acétylcholine (2,2 Å). Le constat est différent pour les modulateurs allostériques. La plupart des méthodes échouent à attribuer la meilleure pose à son site correspondant. Seul le programme XedMin replace de façon satisfaisante 2 des 5 molécules non orthostériques : 42R dans la poche sous-agoniste et 9Z0 dans la poche du vestibule. Cependant, bien que situées dans le bon site, ces poses sont toutes relativement éloignées de la pose de référence, à plus de 5 Å. On notera tout de même que les poses de référence 5VU et FHV (comme l'ivermectine) ne sont quasiment pas accessibles par le *docking* FlexX (Figure V.4/p.149). Il était donc prévisible qu'aucune des méthodes de *rescoring* ne puisse prédire la localisation de ces sites.

3.2.5 Paramétrisation du *scoring* XED

Dans la section précédente, l'analyse des molécules est restreinte aux poses localisées dans la partie extracellulaire du récepteur. Nous traitons maintenant l'ensemble des poses placées sur les 960 conformations du récepteur nicotinique, après *rescoring* par la fonction de score XedMin. Cela représente pour la lobéline 1 029 143 poses, 389 879 pour l'acétylcholine, 604 389 pour la nicotine et des ordres de grandeur semblable pour les autres molécules. Comme précédemment, nous observons pour chacune des molécules la pose de meilleur score. La première colonne ("XedMin") de la Figure V.6/p.suiv. expose les résultats ainsi obtenus. Il est frappant de voir que sur plus d'un million de positionnements, la pose de meilleur score de la lobéline est à seulement 3,4 Å de sa pose de référence. Mis à part l'épibatidine (EPJ), l'ensemble des poses correctement placées en se restreignant aux poses de la partie extracellulaire le sont aussi en considérant l'ensemble du récepteur. Nous avons ajouté à l'ensemble des molécules testées 4 modulateurs allostériques : NS-1738, PNU-120596, LY-2087101 et CCMI. Grâce à des expérimentations de mutagenèse dirigées ainsi qu'à des études d'amarrage moléculaire *in silico*, il est supposé que ces molécules se lient dans la partie transmembranaire du récepteur, bien que leur localisation exacte soit encore débattue. Nous pouvons alors évaluer si le programme XedMin est capable d'attribuer un score d'affinité maximale à une pose située dans la région transmembranaire du récepteur. La première colonne de la Figure V.6/p.suiv. nous montre que ce n'est le cas pour aucune de ces molécules. En fait, c'est une constante pour la plupart des méthodes de *rescoring*. En évaluant l'ensemble des poses du récepteur, soit les meilleures poses se situent systématiquement dans la partie extracellulaire (FlexX, Vina, Vinardo, Smina, AD4, XedMin), soit elles sont toutes trouvées dans la partie transmembranaire (Hyde). Cela met en valeur un biais

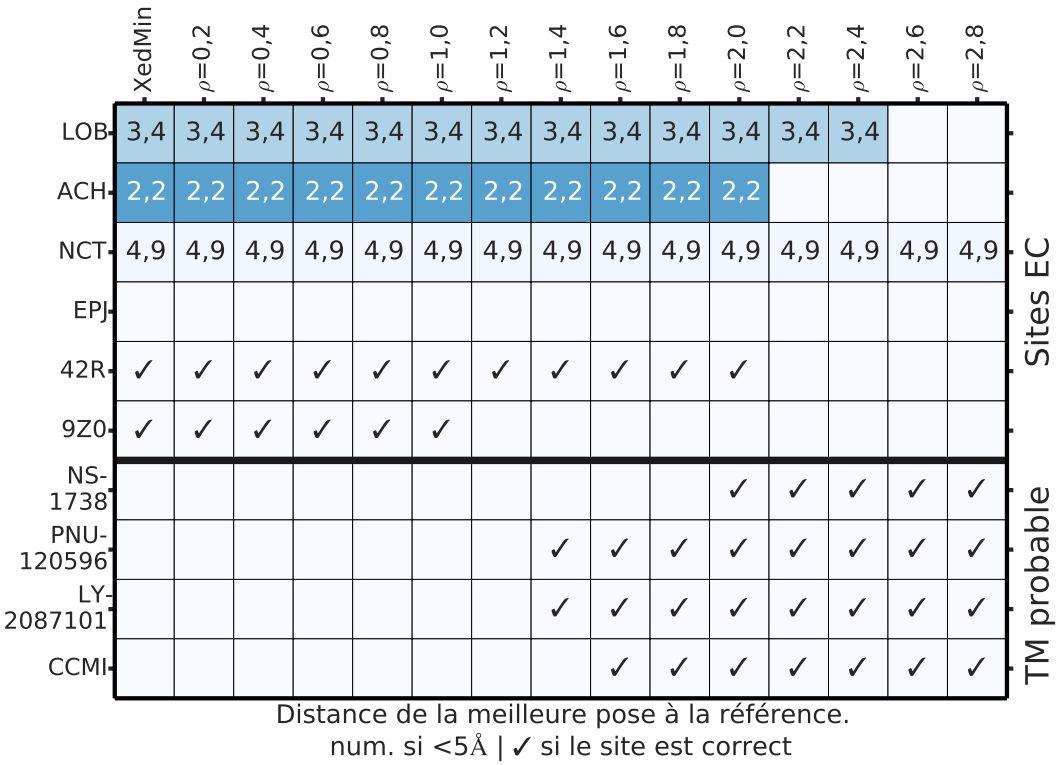


Figure V.6 Comparaison des prédictions XedMin et XedMin$_\rho$. Localisation des poses de meilleure affinité selon le programme XedMin, version par défaut (première colonne) ou repondérée avec le paramètre ρ. Présence d'une valeur réelle (RMSD en Å) si la meilleure pose est à moins de 5 Å de la pose de référence, ✓ si la pose est dans le bon site (partie haute du tableau) ou dans la partie transmembranaire (partie basse). La case reste vide dans tous les autres cas.

dans les fonctions de score pour discriminer des poses de la même molécule situées dans des environnements différents. Une tentative de correction de ce biais a été réalisée pour l'outil XedMin en pondérant le terme électrostatique de la fonction de score lorsque la pose se trouve au niveau de la partie transmembranaire. En modifiant le paramètre ρ, l'ensemble des poses transmembranaires ont une affinité relative qui augmente si $\rho > 1$ ou qui décroît si $\rho < 1$. Sans surprise, lorsque $\rho < 1$ pénalise l'affinité, les meilleures poses restent toutes situées dans la partie extracellulaire. En augmentant ρ à 1,4 on observe un placement correct du PNU-120596 et de LY-2087101 dans la partie transmembranaire, tout en préservant les poses extracellulaires de la lobéline, de l'acétylcholine, de la nicotine et de l'effecteur 42R. À l'inverse, lorsque $\rho > 2{,}4$ on perd le bon positionnement de la plupart des molécules dans la partie extracellulaire. $\rho = 2{,}0$ semble ici être un compromis efficace pour placer correctement un maximum de molécules. Les valeurs absolues des affinités prédites ne sont plus physiquement vraisemblables, mais leurs valeurs relatives pour sélectionner la pose la plus pertinente apparaissent être capables de discriminer des ligands se liant dans la partie extracellulaire ou transmembranaire du récepteur. La remarque principale de cette analyse est un appel à une meilleure prise en compte de l'effet de la solvatation de la molécule lors de l'évaluation des

scores d'affinité. À partir de ces résultats, nous proposerons une hypothèse de sites de liaison pour les 4 modulateurs allostériques dans le chapitre suivant.

3.3 Discussion

Dans ce chapitre, de petites molécules connues pour être des modulateurs du récepteur $\alpha 7$ nAChR ont été fixées sur l'ensemble des sites et des conformations de la transition d'activation du récepteur. Sur la base d'un ensemble de molécules de référence, pour lesquelles le site de liaison est connu, nous avons comparé 3 outils d'amarrage moléculaire (Autodock Vina, DOCK6, Leadit/FlexX), ainsi que 9 fonctions d'évaluation de l'affinité des poses pour la protéine (FlexX, Hyde, RF-score, Cyscore, Smina, Vinardo, Vina, AD4, XedMin). La combinaison d'outils la plus pertinente sur notre jeu de données correspond à FlexX avec ajout de l'interaction cation-π puis la réévaluation de l'affinité des poses avec le programme XedMin. Cette configuration replace correctement des agonistes se liant à la poche orthostérique (p.ex. la meilleure pose de l'acétylcholine est à moins de 2,2 Å de sa pose de référence sur 389 879 poses réparties sur les conformations du récepteur), ainsi que deux modulateurs allostériques dans la partie extracellulaire (42R et 9Z0).

Cependant, les limites de l'approche se sont fait ressentir. D'une part, le *docking* initial des molécules est crucial. Si aucune pose n'est initialement placée suffisamment proche de la pose de référence, il devient impossible de discerner le site de liaison, même lorsque le programme de *rescoring* permet l'optimisation locale des poses. C'est le cas pour le MLA, et les modulateurs allostériques ivermectine, 5VU et FHV. D'où l'importance d'un échantillonnage large et diverse des poses, quitte à générer un grand nombre de poses invraisemblables qui pourront être pénalisées dans un second temps par la fonction de *rescoring*. Un deuxième facteur d'échec vient de l'évaluation de l'affinité ligand/protéine. Une sous- ou surévaluation de l'affinité de certaines poses, provoque l'apparition de faux positifs (molécule avec un bon score, mais éloignée de la pose de référence). Les molécules FHV, OJD et l'épibatidine en sont des exemples lorsque XedMin a été utilisé. Finalement, nous avons observé la nécessité d'une prise en compte de l'environnement directe de la protéine. L'évaluation de l'affinité des poses placées dans la région transmembranaire de la protéine, généralement plus hydrophobe, devrait être considérée de façon plus rigoureuse et en consistance avec l'évaluation des poses situées hors de la zone membranaire. Le choix de la pose de meilleur score pour conduire la détection du site de liaison remet en question la robustesse de la méthode. Il suffit d'une seule surévaluation d'affinité sur le million de poses à analyser pour que la prédiction de site soit fausse. En pratique, il pourrait être avantageux de considérer les groupes de poses similaires entre elles et d'affinité, en moyenne, favorables, ou de combiner

les prédictions des différents programmes de scoring. Ces options n'ont pas été explorées à cause du nombre trop limité de poses cristallographique de référence (risque élevé de surajustement, *overfitting* aux données de référence).

4 Recherche de sites à effet allostérique

L'existence de molécules à effet allostérique ciblant le récepteur nicotinique implique la présence de sites de liaison spécifiques. La localisation de ces sites est déterminante pour la découverte et l'optimisation de nouvelles familles de molécules thérapeutiques, par exemple, au travers de *drug design* rationnel. La recherche de sites allostériques pour le récepteur $\alpha 7$ est soutenue par des études de cristallographie [36] et de RMN [292]. Des mutations systématiques sur la séquence du récepteur permettent d'identifier les résidus impactés par la liaison d'un ligand et l'effet associé sur le récepteur grâce à des analyses électrophysiologiques [78, 270, 273, 293]. Ces études sont souvent complétées par des campagnes *in silico* d'amarrages moléculaires [78, 80, 269].

Malgré cet éventail de techniques, relativement peu de sites allostériques sont connus, en particulier dans la région transmembranaire du récepteur, et les sites d'interaction proposés sont souvent sujets à discussions. En 2008, Young et al. [78] ont déterminé un ensemble de 5 mutations réparties sur les 4 hélices transmembranaires du récepteur ayant un effet significatif sur la potentialisation du récepteur par le modulateur allostérique PNU-120596. Par comparaison avec les structures cryo-EM [101] du récepteur nicotinique *Torpedo Marmarata*, les résidus en question pointent une poche intra sous-unités, dont des interactions possibles avec le ligand sont confirmées par simulation de *docking*. Des conclusions similaires sont obtenues deux ans plus tard, avec l'ivermectine, qui se lierait dans un site similaire entre les 4 hélices d'une même sous-unité [294], contrairement à la poche inter sous-unités observée dans les structures des récepteurs homologues GluCl [43] et Glycine [45]. Plus tard, une erreur d'assignement des acides aminés de la structure nAChR *Torpedo* est relevée dans la densité de cryo-microscopie et remet en question les modèles utilisés pour la localisation des poches transmembranaires [43, 102, 103]. Début 2018, une nouvelle publication basée sur des structures *Torpedo* corrigées, privilégie une localisation du PNU-120596 entre deux sous-unités adjacentes [80]. Cela montre l'extrême difficulté à identifier la localisation de sites d'interaction lorsque les structures expérimentales du complexe ne sont pas encore disponibles.

Dans ce chapitre nous analysons les sites détectés dans la transition $\alpha 7$ grâce au suivi des cavités. L'évolution de la géométrie des cavités le long du mécanisme d'activation est utilisée comme descripteur pour déterminer des sites potentiellement effecteurs. Dans un second temps les meilleures poses obtenues dans le chapitre précédent sont observées et comparées aux modes de liaison prédits dans les publications récentes.

4.1 Sites impactés par la transition

Le suivi global des cavités le long de la transition $\alpha 7$ nous a permis d'identifier 68 sites englobant régulièrement une cavité de volume non nul (cf. Partie IV/p.85). Nous cherchons à déterminer un sous-ensemble de ces sites dont la géométrie des cavités évolue en corrélation avec le changement conformationnel. Par effet de levier sur le mécanisme d'activation du récepteur, la liaison d'une molécule dans un tel site pourrait stabiliser un état privilégié du canal ionique. Cette approche a été appliquée avec succès pour la découverte d'un inhibiteur du facteur œdématogène de l'anthrax [6, 109]. Le descripteur géométrique le plus intuitif est la mesure du volume instantané de la cavité. En comptant le critère de symétrie de séquence, nous disposons pour chaque site de 5 évolutions distinctes du volume de la cavité. Une régression linéaire sur les volumes nous donne une idée de la tendance de la cavité à se dilater ou à se contracter au cours de la trajectoire, grâce à la pente de la régression. Pour chaque site cette pente est moyennée sur les 5 cavités transverses symétriques. La Figure V.7 présente l'évolution du volume des 6 sites (parmi les 68)

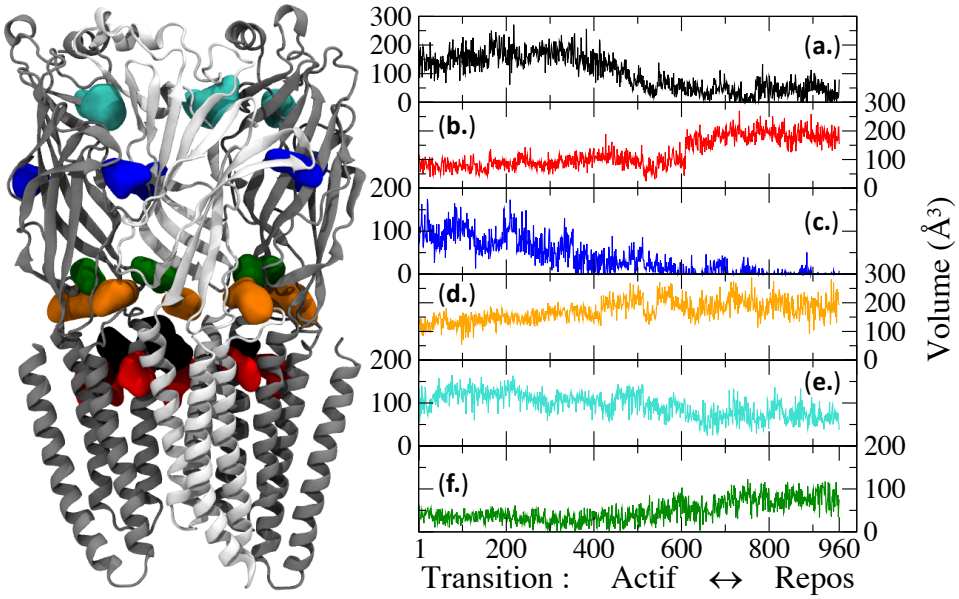


Figure V.7 Sous-ensemble des sites évoluant le plus avec le changement conformationnel. À gauche, cavité moyenne (tronquée à 0,25) de chacun des sites sélectionnés. À droite, volumes des différents sites (moyenné sur les 5 poches symétriques). en fonction du pas de trajectoire (fermeture du canal). Les sites, notés de **a.** à **f.**, sont triés en fonction de la valeur absolue de la pente de la régression linéaire sur les volumes de la transition.

dont la valeur absolue de la pente moyenne est la plus forte. Une vue globale permet de situer la localisation de ces sites sur le récepteur. La transition est orientée d'un état actif/canal ouvert en début de trajectoire, à un état de repos/canal fermé à droite. Les commentaires proposés ici sont fonction de la fermeture du récepteur, mais l'interprétation peut éventuellement être inversée. La plupart des sites sélectionnés

ont des volumes significatifs à au moins un moment de la transition, ce qui en fait des sites capables de contenir de petites molécules (p.ex. 156,67 $Å^3$ pour une molécule d'acétylcholine). Les sites **a.**, **c.** et **e.** ont une tendance à se contracter lors de la fermeture, alors que les sites **b.**, **d.** et **f.** ont des volumes qui augmentent. On remarque que le site **c.** se situe dans la poche orthostérique. Cette assertion est vérifiée en alignant la structure cristallographique d'un AChBP lié à un agoniste (la lobéline) sur une des conformations de la trajectoire, comme présenté dans la Figure V.7/p.préc.. La lobéline cristallographique y chevauche la cavité moyenne du site **c.**. Les autres sites sont distincts du site orthostérique : ce sont des sites allostériques potentiels. Nous avons recherché si ces sites de l'$\alpha 7$ correspondent à des sites connus chez des récepteurs homologues. Le site **b.** se trouve à une localisation similaire du site de liaison du modulateur allostérique positif ivermectine chez le récepteur GluCl (p.ex. PDB : 3RIF) et récepteur Glycine (p.ex. PDB : 3JAF ou PDB : 5VDI). Le site **e.** recouvre le site de liaison du fragment 5VU qui est

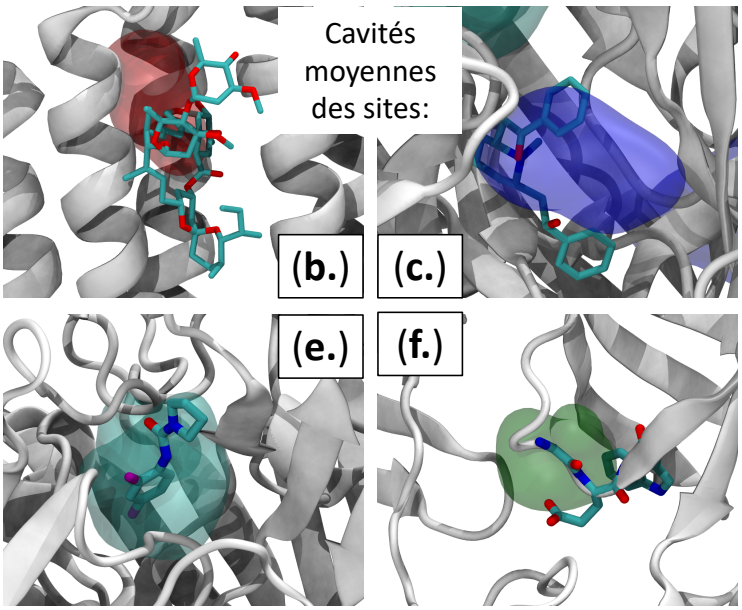


Figure V.8 Alignement de co-cristaux de récepteurs homologues sur la première conformation de la transition. En haut à gauche, cavité moyenne du site **b.** et structure du récepteur GluCl (PDB : 3RIF) liée avec l'ivermectine (IVM). En haut à droite, cavité moyenne du site orthostérique (**c.**) et structure de la chimère $\alpha 7$ AChBP (PDB : 5AFH) liée avec la lobéline (L0B). En bas à gauche, cavité moyenne du site **e.** et structure de la chimère $\alpha 7$ AChBP (PDB : 5AFK) liée au modulateur allostérique 5VU. En bas à droite, cavité moyenne du site **f.** et le triplet de résidus conservés GEW définissant le site de régulation par le Ca^{2+} [295, 296].

un modulateur allostérique négatif du récepteur $\alpha 7$ visible dans le complexe co-cristallisé d'une chimère $\alpha 7$ AChBP [36]. Le site **f.** est en contact direct avec le motif conservé de résidus G-E-W impliqués dans la liaison régulatrice avec les ions Ca^{2+} [42, 295], et chevauche le site de la chlorpromazine et la bromopromazine liée au récepteur ELIC [296]. Les sites **a.** et **d.** sont moins documentés dans les récepteurs homologues. On notera que par superposition des structures $\alpha 7$ et GLIC

PDB : 4HFD, la cavité moyenne **a.** contient une molécule de bromoforme. Il est ainsi particulièrement intéressant d'observer que parmi les 68 sites détectés lors de la transition $\alpha 7$, 4 des 6 sites dont le volume évolue significativement avec la transition sont localisés à des positions équivalentes de sites allostériques déjà connus dans des récepteurs homologues. Bien que cela ne constitue pas une preuve de la liaison de ces modulateurs allostériques sur le récepteur $\alpha 7$, cela met en valeur des mécanismes d'activation et de modulation partagés entre les différentes formes de récepteur canaux. Ces sites, déjà connus (**b.**, **e.**, **f.**) ou encore peu explorés (**a.**, **d.**) sont probablement des cibles à privilégier pour la recherche de nouveaux effecteurs spécifiques du récepteur nicotinique.

4.2 Analyse des poses de *docking*

En plus des effecteurs de référence (dont le site de liaison est connu) nous avons amarré et évalué l'affinité d'un petit ensemble de modulateurs allostériques (NS-1738, LY-2087101, CCMI et PNU-120596) sur l'ensemble des sites et conformations de la transition $\alpha 7$. Les 3 premières molécules sont des modulateurs allostériques positifs (PAM) de type I, et le PNU-120596 est un PAM de type II. Des études d'électrophysiologie couplées à la mutagenèse dirigée du récepteur $\alpha 7$ tendent à privilégier une liaison des PAMs dans la partie transmembranaire du récepteur [78, 253, 273]. Les figures V.10/p.162 et V.9/p.suiv. montrent les poses de meilleures affinités ainsi que leur localisation dans la partie transmembranaire du récepteur. Le *docking* global des molécules a été réalisé avec le programme FlexX/Leadit, suivi d'un *rescoring* de chacune des poses par XedMin avec restriction des poses de la partie transmembranaire (XedMin_{TMD}), ou utilisation du score XedMin reponderé ($\text{XedMin}_{\rho=2,0}$, voir 3.1.3/p.145). Pour l'évaluation avec $\text{XedMin}_{\rho=2,0}$, la pose de meilleure affinité est choisie parmi 1 137 080 poses pour NS-1738 et plus de 500 000 poses pour chacune des 3 autres molécules. Un site de liaison est largement privilégié par les deux fonctions de score : le site **1** visible dans la Figure V.9/p.suiv. est la cible de 7 des 8 poses prédites, à l'interface des sous-unités adjacentes et à l'exact emplacement du site **b.** de la Figure V.7/p.157. Une seule meilleure pose (CCMI avec score $\text{XedMin}_{\rho=2,0}$) se trouve dans le pore central du récepteur (site **2**). Le site **1** correspond au site de liaison de l'ivermectine pour les récepteurs homologues GluCl et Glycine. La Figure V.9/p.suiv. montre la position de liaison de l'ivermectine dans la poche du récepteur GluCl. On remarquera que la Ser260 (numérotation PDB : 3RIF) forme une liaison hydrogène avec un groupe hydroxyle de la molécule [43]. Une serine est aussi présente à la même position de l'hélice M2 pour les récepteurs Glycine et GABAA_A, tous deux activés par la liaison de la molécule. L'ivermectine n'est qu'un potentialisateur du récepteur nicotinique [297] et dans la séquence de la sous-unité $\alpha 7$ cette serine est remplacée par la méthionine

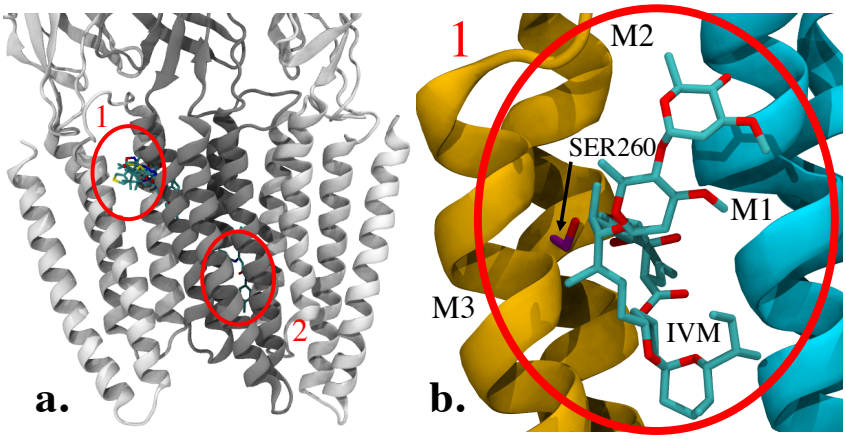


Figure V.9 Localisation des poses de meilleures affinités. **a.** Localisation des 8 meilleures poses obtenues par le protocole de *docking* global. La région **1**, la plus représentée, se trouve à l'interface des sous-unités, entre les hélices M3(+), M2(+) et M1(-), et est similaire aux sites de liaison de l'ivermectine des récepteurs GluCl et Glycine. Le site **2** se trouve dans la partie basse du pore, entre les hélices M2. 3 chaînes du récepteur $\alpha 7$ sont représentées, avec la seconde sous-unité en gris foncé. **b.** Vue du site de liaison de l'ivermectine, structure PDB : 3RIF avec sous-unités portant la boucle C (+) en orange, et la sous-unité complémentaire (-) en bleu. La serine 260 est visible en licorice violet.

253 (position 272 de l'alignement multiple présenté en Annexe VII.2/p.194). Il est cependant intéressant de noter que cette méthionine, une fois mutée en leucine, convertit l'ivermectine en un inhibiteur (antagoniste non-compétitif) [294], ce qui présume un rôle important de ce résidu et de ce site dans le mécanisme allostérique. Cette méthionine est aussi un déterminant structural crucial dans la potentialisation du récepteur $\alpha 7$ nAChR par les modulateurs allostériques PNU-120596 [78, 273], LY-2087101 [78] et NS-1738 [273]. Il est donc particulièrement intéressant d'observer que les meilleures poses trouvées par notre protocole de *docking* sont toutes très proches de ce résidu clé (voir Figure V.10/p.162). On remarque par ailleurs des poses consensuelles pour les deux scores XedMin_{TMD} et $\text{XedMin}_{\rho=2,0}$, avec un choix de poses strictement identiques pour NS-1738 et PNU-120596, et légèrement dissimilaires pour LY-2087101. La présence d'un site allostérique entre les sous-unités dans la partie transmembranaire est en accord avec les études de *docking* récentes sur les interfaces d'$\alpha 7 - \alpha 7$ (avec LY-2087107, NS-1738 et PNU-120596) [80], et d'$\alpha 4 - \alpha 4$ (LY-2087107) [298]. Concernant la pose CCMI de $\text{XedMin}_{\rho=2,0}$, aucun élément expérimental ne vient supporter la liaison de la molécule dans le canal du récepteur.

4.3 Discussion

Deux résultats principaux sont suggérés par cette étude. Premièrement, l'analyse globale du volume des cavités présentes dans la trajectoire de transition du récepteur $\alpha 7$ a pointé un sous-ensemble de 6 sites (sur 68) dont la géométrie évolue de concert

avec le mouvement conformationnel induit par l'activation du récepteur. Le fait que 4 de ces sites correspondent à des sites effecteurs connus chez des récepteurs homologues suggère la présence de mécanismes de régulations allostériques partagés avec le récepteur nicotinique. Par transitivité, il pourrait être intéressant d'étudier les 2 autres sites identifiés, bien que leur capacité à lier de petites molécules n'ait pas encore été explorée. Un second résultat met en valeur une poche située dans la partie transmembranaire du récepteur à l'interface des sous-unités $\alpha 7$, qui agrège les meilleures poses de *docking* de 4 modulateurs allostériques du récepteur nicotinique. Du fait de l'enchaînement des hypothèses qui ont été faites au cours de ce travail, ce résultat reste largement spéculatif (incertitudes des modèles de transition, des fonctions de *docking* et de *rescoring*). Cependant, un certain nombre de points viennent étayer ces hypothèses :

- Le modèle de transition utilisé reproduit les mouvements quaternaires de *blooming* et de *twisting* considérés comme cruciaux pour décrire le mécanisme d'activation (Partie III/p.47).
- Les poches utilisées pour le *docking* couvrent quasi uniformément l'ensemble du récepteur (Figure IV.15/p.127).
- Le protocole de *docking/rescoring* établi est capable de replacer simultanément et correctement 3 agonistes dans la poche orthostérique (lobéline, acétylcholine, nicotine) et 1 modulateur allostérique dans la partie extracellulaire (42R) (colonne $\text{XedMin}_{\rho=2,0}$ de la Figure V.6/p.153).
- Un des sites de liaison trouvé est l'un des plus impacté par la transconformation du récepteur et est situé à la même position qu'un site allostérique déjà trouvé chez les récepteurs Glycine et GluCl (site **b.** Figure V.7/p.157 et site **1.** Figure V.9/p.préc.).
- Les meilleures poses des modulateurs PNU-120596, LY-2087101 et NS-1738 sont très proches du résidu M253 qui est un déterminant structural de la liaison de la molécule avec le récepteur (Figure V.10/p.suiv.).

Bien que les arguments soient nombreux, une démonstration rigoureuse de la localisation des sites de liaison associés à ces molécules allostériques ne saura se passer que difficilement de la résolution expérimentale des complexes ligand/protéine.

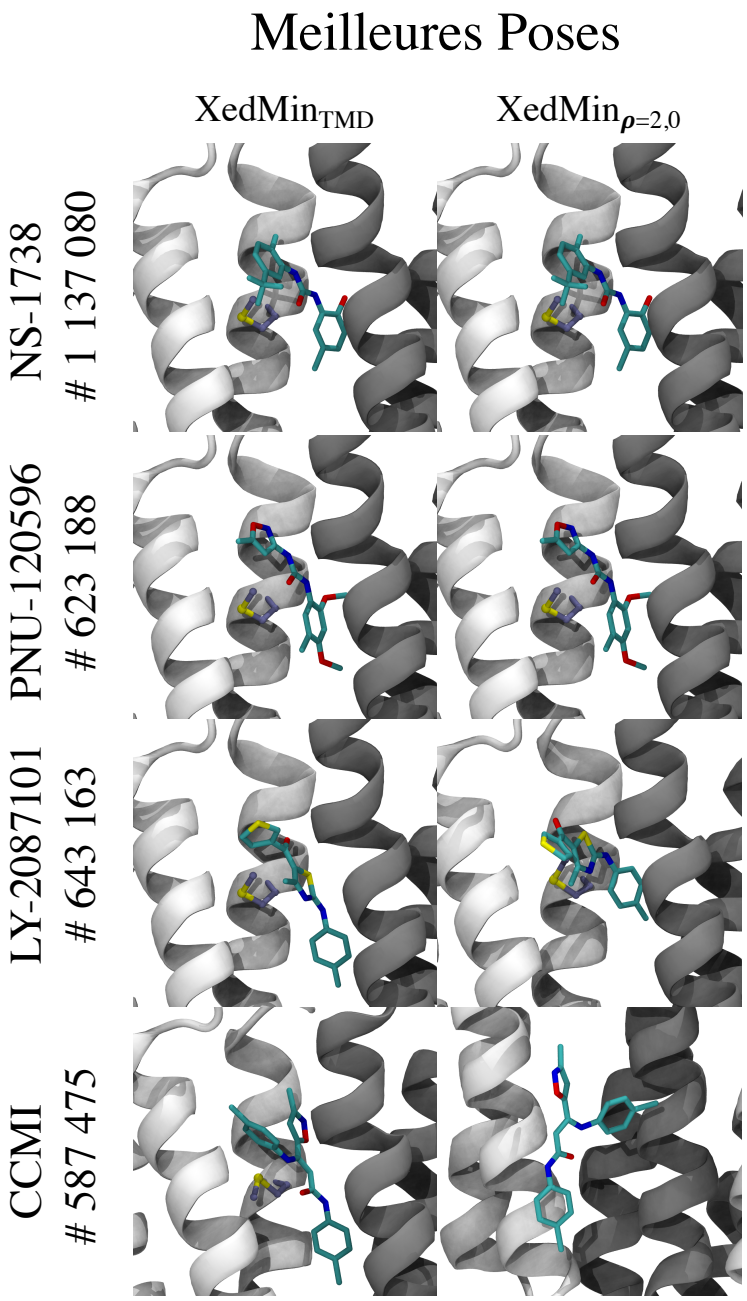


Figure V.10 Poses de meilleures affinités par les scores XedMin$_{TMD}$ et XedMin$_{\rho=2,0}$ pour chacun des modulateurs allostériques amarrés sur l'ensemble du récepteur $\alpha 7$. Meilleures poses XedMin$_{TMD}$: pose de meilleur score XedMin sur l'ensemble des poses situées dans la partie transmembranaire du récepteur. Meilleures poses XedMin$_{\rho=2,0}$: pose de meilleur score XedMin sur l'ensemble des poses placées sur le récepteur et après repondération à $\rho = 2{,}0$ des termes d'interaction. Le nombre total de poses considérées est indiqué sous le nom de chaque molécule. Lorsque la pose se situe dans le site **1**, la méthionine 253 est représentée en licorice violet.

5 Conclusions

Cette partie démontre la faisabilité technique du *docking* global de petites molécules sur l'ensemble des sites et des conformations de la transition conformationnelle du récepteur nicotinique. L'utilisation d'un programme d'amarrage moléculaire rapide associé au *rescoring* de l'affinité des poses obtenues rend la méthode raisonnable en temps de calcul tout en préservant une pertinence acceptable des résultats. Nous avons pu observer l'importance de la prise en compte des spécificités structurales de la cible avec l'ajout explicite de l'interaction cation-π au programme de *docking* FlexX puis en choisissant la combinaison d'outils la plus efficace sur un ensemble de données de test. L'approche utilisée a pu prédire correctement la localisation de 3 molécules agonistes (site orthostérique), d'un modulateur allostérique négatif (site extracellulaire), et la localisation transmembranaire de 4 modulateurs allostériques (NS1738, PNU-120596, LY-2087101 et CCMI) en accord avec les données expérimentales existantes. À l'inverse, le protocole échoue dans le replacement de 7 molécules de références. Les raisons de ces échecs sont difficiles à identifier du fait de l'incertitude qui pèse sur la vraisemblance au niveau atomique des modèles de transition. La méthyllycaconitine, et 3 modulateurs allostériques (dont l'ivermectine) ne sont pas accessibles dans leurs sites respectifs dès l'étape de *docking*. Les poses de l'épibatidine et de deux modulateurs allostériques, bien que parfois proches de leurs sites de référence ne sont pas correctement évaluées par la fonction de *rescoring*. Ces derniers cas appellent à une prise en compte plus rigoureuse de l'environnement membranaire du récepteur lors de l'évaluation des scores d'affinité ligand/protéine.

Ce travail nous a aussi permis de cartographier les régions du récepteur impactées par la transition conformationnelle d'activation du récepteur. De façon intéressante, ces régions correspondent pour partie à des sites allostériques précédemment isolés dans des récepteurs homologues et soulignent la pertinence de l'analyse systématique des cavités dans les trajectoires de dynamiques de protéine. Les résultats du *docking* global pour les 3 effecteurs allostériques NS1738, PNU-120596, LY-2087101, dont les modes de fixation sont encore inconnus, suggèrent un site de liaison possible dans l'un de ces sites impactés par la transition du récepteur. En particulier, ce site à effet potentiellement allostérique se trouve dans la partie transmembranaire à l'interface des sous-unités $\alpha 7$ et à un *locus* similaire au site de l'ivermectine dans les structures connues du récepteur Glycine et GluCl.

VI

Conclusions générales

Modèle de transition du récepteur nicotinique

Nous avons présenté dans ces pages un modèle de transition décrivant le mécanisme d'activation du récepteur nicotinique de sous-unités $\alpha 7$. Plusieurs résultats clés en sont ressortis. Au cours des raffinements successifs de notre modèle de transition, les différentes conformations du récepteur ont progressivement évolué vers des états plus vraisemblables qui se sont traduits par une accentuation de la torsion entre les domaines extracellulaire et transmembranaire (*twisting*) et l'apparition d'une extension du domaine extracellulaire (*blooming*) entre les états possiblement de repos et actif de la protéine. Ces données sont en accord avec les mouvements quaternaires de grande amplitude observés dans les structures cristallographiques de canaux ioniques homologues. Ce résultat justifie la méthode d'optimisation de chemins de transition développée : le couplage avantageux des méthodes *String of Swarms* (SoS) et *Path Optimisation and Exploration* (POE). Le chemin bénéficie de la relaxation des chemins avec la méthode SoS dans le champ d'énergie libre, avec prise en compte des interactions explicites avec le solvant dont la prise en compte des phospholipides et de molécules de cholestérol dans la membrane. Cela suggère que l'approche pourrait améliorer la qualité des structures dans leur environnement. De son côté, POE affine les chemins SoS hors des minima locaux et efface la complexité topologique générée par les dynamiques moléculaires propres à SoS. Ce protocole de calcul de chemins de transition pourra aisément être adapté pour d'autres protéines cibles. Une autre observation encourageante concerne le lien étroit entre le changement d'état du récepteur et l'évolution associée des cavités. Nous avons pu identifier les principaux sites de la protéine impactés par l'activation du récepteur. Parmi ceux-ci, 4 sites sont des sites effecteurs du récepteur $\alpha 7$ ou des sites allostériques déjà identifiés chez des récepteurs homologues. Ces exemples démontrent l'intérêt de l'analyse des cavités dans la recherche de nouveaux leviers allostériques. Le *docking* de petites molécules sur l'ensemble des conformations et sites de la transition, bien que perfectible, peut suggérer des sites de liaisons potentielles pour des molécules connus dont le siège d'action est encore inconnu. La liaison de 3 modulateurs allostériques, NS1738, PNU-120596 et LY-2087101 a été prédite dans une poche placée dans la partie transmembranaire entre les sous-unités $\alpha 7$, et déjà caractérisée expérimentalement comme un site allostérique dans des structures de récepteurs homologues. Bien qu'une démonstration rigoureuse des sites de fixation de ces molécules nécessite la résolution des co-complexes expérimentalement, ces résultats proposent un modèle de liaison qui pourrait aider les recherches expérimentales futures, par exemple, en proposant des choix d'acides aminés à muter.

Programme pour le suivi des cavités dans les dynamiques de protéines : *mkgridXf*

L'Unité de Bioinformatique Structurale a une longue expérience dans l'étude des cavités de protéine. Lorsque je suis arrivé au laboratoire, les concepts clés pour la détection et le suivi des cavités dans les trajectoires de protéine étaient déjà à un stade de développement avancé dans le laboratoire, en particulier grâce aux travaux initiateurs d'Arnaud Blondel et de Nathan Desdouits [227]. Pendant 6 mois (stage de fin d'étude), j'ai utilisé ces outils pour l'étude du changement conformationnel d'une protéine d'enveloppe du Virus de la Fièvre de la Vallée du Rift. J'ai rencontré trois difficultés principales lors des analyses de cavités : (1) les fusions de cavités, (2) le choix du paramètre de partitionnement et (3) une dégradation des performances de la méthode pour de longues trajectoires. Le point (1) a été résolu avec l'idée de Nathan de découper des cavités en plusieurs morceaux et de les réassocier à leurs sites respectifs. Bien que son implémentation pratique ait évolué ces dernières années, le concept de "découpage" n'a pas changé et donne maintenant des résultats très satisfaisants pour gérer les fusions de cavités. Le choix du paramètre de partitionnement, point (2), a été plus délicat, car son changement induit des changements subtils dans le suivi des cavités. Il a aussi ouvert la porte à des questionnements plus larges : quelle méthode de partitionnement est la plus pertinente, quelles mesures de distance entre les empreintes ou définitions d'empreintes privilégier ? L'appréciation du suivi correcte des cavités dans les dynamiques de protéines se faisait alors principalement à l'œil, par visualisation sur ordinateur des trajectoires. Or, comme nous avons pu le voir, les cavités sont hautement fluctuantes, mobiles et omniprésentes dans les trajectoires de dynamique moléculaire, ce qui rend ce mode d'évaluation particulièrement subjectif. À un certain stade de raffinement de la méthode, la nécessité de mise en place d'un système de référence pour évaluer le suivi des cavités est devenue évidente. Ces travaux, entrepris pendant les derniers mois de thèse de Nathan, ont produit les résultats présentés dans la Partie IV/p.85, et nous donnent aujourd'hui une mesure claire de la pertinence et de la robustesse de la méthode. Le point (3), la vitesse de calcul et la taille de l'empreinte mémoire utilisée par le programme de suivi, est devenu critique lorsque j'ai commencé à travailler sur le récepteur nicotinique. Le grand nombre de cavités (parfois supérieur au million), était quasi inaccessible à la précédente implémentation. Durant la thèse, j'ai écrit le programme *mkgridXf*, dont la gestion de la mémoire et du CPU est optimisée pour le suivi des cavités (lire l'Annexe C.1/p.200 pour plus de précisions). *mkgridXf* réimplémente le programme Fortran *mkgrid* développé au laboratoire il y a plusieurs années pour la détection des cavités [6–9] et incorpore les routines nécessaires au partitionnement et à la reconstruction des trajectoires de cavités. Le programme est écrit en C et correspond à environ 7000 lignes de code. Nous avons montré que l'applicabilité de l'analyse

des cavités est large : l'analyse de l'évolution des volumes (Section 4.1/p.157) de cavités mais aussi leur caractérisation géométrique précise [9], la cartographie des sites de la protéine (Partie IV/p.85), etc. *mkgridXf* qui sera prochainement distribué, est déjà utilisé par d'autres groupes dans l'Unité, ainsi qu'une collaboration avec l'équipe de M.Baaden à l'Institut de Biologie de physico-chimique à Paris.

Recherches en cours et futures

Lors de ma dernière année de thèse, j'ai eu la chance de pouvoir travailler en collaboration avec le Professeur J.P. Changeux, grâce au financement européen *Human Brain Project* - SGA1. L'objectif de recherche de la composante CDP6 étant étroitement lié à mon domaine d'expertise (conception d'effecteurs allostériques de cibles impliqués dans des neuropathologies), j'ai pu approfondir les résultats déjà acquis dans ce qui est devenu la Partie V/p.131 de ce manuscrit. Ces résultats, ainsi que des données complémentaires, ont été incorporés au rapport de la composante CDP6 [299]. Une partie significative de mes travaux de thèse a été utilisée comme données préliminaires pour des demandes de financements majeurs. Les deux projets suivants ont par la suite été financés. Une bourse européenne à 5 ans *ERC-ADG - Advanced Grant* pour le projet DYNACOTINE (J.P. Corringer) qui cherchera dès 2019 à enregistrer en temps réel simultanément l'activation des récepteurs nicotiniques $\alpha 7$ et $\alpha 4\beta 2$ ainsi que leurs mouvements grâce à la mise au point de nouveaux désactivateurs fluorescents (*quencher*). Le projet comprend aussi l'étude des interactions entre modulateurs allostériques et les lipides de la membrane et leur co-cristallisation inédite avec le récepteur. Notre modèle de transition du récepteur $\alpha 7$ pourra être d'une aide importante pour la sélection d'acides aminés pertinents pour la mise au point des méthodes de fluorescences. De plus, les résultats expérimentaux pourront être réutilisés pour raffiner le modèle de transition existant. Le projet NICOFIVE, financé par l'Agence Nationale de la Recherche (ANR) depuis 2017, vise à dessiner de nouveaux modulateurs allostériques ciblant spécifiquement les récepteurs nicotiniques comprenant une sous-unité $\alpha 5$, pour lutter contre les addictions. Ce travail est mené en collaboration avec l'équipe de D. Joseph du laboratoire de Biomolécules : Conception, Isolement, Synthèse de l'Université Paris-Sud, et de P.J. Corringer du département de l'Unité Récepteurs Canaux de l'Institut Pasteur. La stratégie utilisée est basée sur la recherche *in silico* de fragments de molécules affins pour l'interface $\alpha 5 - \alpha 4$ et leur reconstruction par des techniques expérimentales de liaison, fusion chimique et par *click-chemistry*. J'ai pris part à ce projet (financement de 8 mois) au côté de Laura Ortega-Varga pour la partie *in silico*. Pour ce travail, la modélisation d'états conformationnels diverses de l'interface $\alpha 4 - \alpha 5$ vient enrichir des modèles cristallographiques AChBP-5/4 pour le *docking* de petites molécules.

Bibliographie

[1] Diarmuid Jeffreys. *Aspirin : The Extraordinary Story of a Wonder Drug*. Bloomsbury Publishing, 2010 (cf. p. 2).

[2] Zanders E.D. *Laying the Foundations : Drug Discovery from Antiquity to the Twenty-First Century. In : The Science and Business of Drug Discovery.* Springer, Boston, MA, 2011 (cf. p. 2).

[3] Elisabet Gregori-Puigjané, Vincent Setola, Jérôme Hert et al. „Identifying mechanism-of-action targets for drugs and probes“. In : *Proceedings of the National Academy of Sciences* 109.28 (2012), p. 11178–11183 (cf. p. 2).

[4] Monica Schenone, Vlado Dančík, Bridget K Wagner et Paul A Clemons. „Target identification and mechanism of action in chemical biology and drug discovery“. In : *Nature chemical biology* 9.4 (2013), p. 232 (cf. p. 2).

[5] A Wadood, N Ahmed, L Shah et al. „In-silico drug design : An approach which revolutionarised the drug discovery process“. In : *OA drug design & delivery* 1.1 (2013), p. 3–7 (cf. p. 2).

[6] Elodie Laine, Christophe Goncalves, Johanna C Karst et al. „Use of allostery to identify inhibitors of calmodulin-induced activation of Bacillus anthracis edema factor.“ In : *Proceedings of the National Academy of Sciences of the United States of America* 107.25 (juin 2010), p. 11277–82 (cf. p. 2, 55, 97, 109, 112, 157, 167).

[7] Armand Berneman, Lory Montout, Sophie Goyard et al. „Combined approaches for drug design points the way to novel proline racemase inhibitor candidates to fight Chagas′ disease“. In : *PLoS One* 8.4 (2013), e60955 (cf. p. 2, 167).

[8] Partricia de Aguiar Amaral, Delphine Autheman, Guilherme Dias de Melo et al. „Designed mono- and di-covalent inhibitors trap modeled functional motions for Trypanosoma cruzi proline racemase in crystallography“. In : *PLOS Neglected Tropical Diseases* (2018) (cf. p. 2, 167).

[9] N. Desdouits, M. Nilges et A. Blondel. „Principal Component Analysis Reveals Correlation of Cavities Evolutionand Functional Motions in Proteins“. In : *Journal of Molecular Graphics and Modelling* (oct. 2015) (cf. p. 3, 87, 96, 97, 129, 130, 167, 168).

[10] Jean-Pierre Changeux. „The nicotinic acetylcholine receptor : the founding father of the pentameric ligand-gated ion channel superfamily“. In : *Journal of Biological Chemistry* 287.48 (2012), p. 40207–40215 (cf. p. 5).

[11] JP Changeux, M Kasai, M Huchet et JC Meunier. „Extraction from electric tissue of gymnotus of a protein presenting several typical properties characteristic of the physiological receptor of acetylcholine“. In : *Comptes rendus hebdomadaires des seances de l'Academie des sciences. Serie D : Sciences naturelles* 270.23 (1970), p. 2864 (cf. p. 7).

[12] Neil S Millar et Cecilia Gotti. „Diversity of vertebrate nicotinic acetylcholine receptors“. In : *Neuropharmacology* 56.1 (2009), p. 237–246 (cf. p. 7, 8).

[13] Nicolas Le Novere et Jean-Pierre Changeux. „Molecular evolution of the nicotinic acetylcholine receptor : an example of multigene family in excitable cells“. In : *Journal of Molecular Evolution* 40.2 (1995), p. 155–172 (cf. p. 7).

[14] I Wessler et CJ Kirkpatrick. „Acetylcholine beyond neurons : the non-neuronal cholinergic system in humans“. In : *British journal of pharmacology* 154.8 (2008), p. 1558–1571 (cf. p. 8).

[15] Clément Léna, Alban de Kerchove d'Exaerde, Matilde Cordero-Erausquin et al. „Diversity and distribution of nicotinic acetylcholine receptors in the locus ceruleus neurons“. In : *Proceedings of the National Academy of Sciences* 96.21 (1999), p. 12126–12131 (cf. p. 8).

[16] Michael W Quick et Robin AJ Lester. „Desensitization of neuronal nicotinic receptors“. In : *Journal of neurobiology* 53.4 (2002), p. 457–478 (cf. p. 8).

[17] University of Bath Susan Wonnacott. *Nicotinic ACh Receptors Scientific Review*. https://www.tocris.com/literature/scientific-reviews/nicotinic-ach-receptors (cf. p. 8, 132).

[18] Sherry Leonard, Catherine Adams, Charles R Breese et al. „Nicotinic receptor function in schizophrenia“. In : *Schizophrenia bulletin* 22.3 (1996), p. 431–446 (cf. p. 8).

[19] Tanya L Wallace et Daniel Bertrand. „Alpha7 neuronal nicotinic receptors as a drug target in schizophrenia“. In : *Expert opinion on therapeutic targets* 17.2 (2013), p. 139–155 (cf. p. 8).

[20] Claudia Soler-Alfonso, Claudia MB Carvalho, Jun Ge et al. „CHRNA7 triplication associated with cognitive impairment and neuropsychiatric phenotypes in a three-generation pedigree“. In : *European Journal of Human Genetics* 22.9 (2014), p. 1071 (cf. p. 8).

[21] Agata Rozycka, Jolanta Dorszewska, Barbara Steinborn et al. „A transcript coding for a partially duplicated form of $\alpha 7$ nicotinic acetylcholine receptor is absent from the CD4+ T-lymphocytes of patients with autosomal dominant nocturnal frontal lobe epilepsy (ADNFLE)“. In : *Folia neuropathologica* 51.1 (2013), p. 65–75 (cf. p. 8).

[22] Hilary Coon, Michele E Villalobos, Reid J Robison et al. „Genome-wide linkage using the Social Responsiveness Scale in Utah autism pedigrees“. In : *Molecular autism* 1.1 (2010), p. 8 (cf. p. 8).

[23] Maryka Quik, Danhui Zhang, Matthew McGregor et Tanuja Bordia. „Alpha7 nicotinic receptors as therapeutic targets for Parkinson's disease“. In : *Biochemical pharmacology* 97.4 (2015), p. 399–407 (cf. p. 8).

[24] Michael R D'Andrea et Robert G Nagele. „Targeting the alpha 7 nicotinic acetylcholine receptor to reduce amyloid accumulation in Alzheimer's disease pyramidal neurons“. In : *Current pharmaceutical design* 12.6 (2006), p. 677–684 (cf. p. 8).

[25] Murat Oz, Georg Petroianu et Dietrich E Lorke. „$\alpha 7$-nicotinic acetylcholine receptors : new therapeutic avenues in Alzheimer's disease“. In : *Nicotinic Acetylcholine Receptor Technologies*. Springer, 2016, p. 149–169 (cf. p. 8).

[26] REN Chao, Ya-lin Tong, Jun-cong Li, Zhong-qiu Lu et Yong-ming Yao. „The protective effect of alpha 7 nicotinic acetylcholine receptor activation on critical illness and its mechanism“. In : *International journal of biological sciences* 13.1 (2017), p. 46 (cf. p. 8).

[27] Lawrence K Leung, Francis M Patafio et Walter W Rosser. „Gastrointestinal adverse effects of varenicline at maintenance dose : a meta-analysis“. In : *BMC clinical pharmacology* 11.1 (2011), p. 15 (cf. p. 8).

[28] Muhamad Y Elrashidi et Jon O Ebbert. „Emerging drugs for the treatment of tobacco dependence : 2014 update“. In : *Expert opinion on emerging drugs* 19.2 (2014), p. 243–260 (cf. p. 8).

[29] Jean Cartaud, E Lucio Benedetti, Jonathan B Cohen, Jean-Claude Meunier et Jean-Pierre Changeux. „Presence of a lattice structure in membrane fragments rich in nicotinic receptor protein from the electric organ of Torpedo marmorata“. In : *FEBS letters* 33.1 (1973), p. 109–113 (cf. p. 9).

[30] Ferdinand Hucho. „Molecular weight and quaternary structure of the cholinergic receptor protein extracted by detergents from Electrophorus electricus electric tissue“. In : *FEBS letters* 38.1 (1973), p. 11–15 (cf. p. 9).

[31] Cheryl L Weill, Mark G McNamee et Arthur Karlin. „Affinity-labeling of purified acetylcholine receptor from Torpedo californica“. In : *Biochemical and biophysical research communications* 61.3 (1974), p. 997–1003 (cf. p. 9).

[32] MA Raftery, R Vandlen, D Michaelson et al. „The biochemistry of an acetylcholine receptor“. In : *Journal of supramolecular structure* 2.5-6 (1974), p. 582–592 (cf. p. 9).

[33] Nigel Unwin. „Refined structure of the nicotinic acetylcholine receptor at 4 Å resolution“. In : *Journal of molecular biology* 346.4 (2005), p. 967–989 (cf. p. 9, 11, 40).

[34] Claudio L Morales-Perez, Colleen M Noviello et Ryan E Hibbs. „X-ray structure of the human $\alpha 4\beta 2$ nicotinic receptor“. In : *Nature* 538.7625 (2016), p. 411 (cf. p. 9, 11, 22, 23, 44, 59, 70, 141).

[35] Shu-Xing Li, Sun Huang, Nina Bren et al. „Ligand-binding domain of an α 7-nicotinic receptor chimera and its complex with agonist“. In : *Nature neuroscience* 14.10 (2011), p. 1253 (cf. p. 9, 26, 135, 141).

[36] Radovan Spurny, Sarah Debaveye, Ana Farinha et al. „Molecular blueprint of allosteric binding sites in a homologue of the agonist-binding domain of the $\alpha 7$ nicotinic acetylcholine receptor“. In : *Proceedings of the National Academy of Sciences* 112.19 (2015), E2543–E2552 (cf. p. 9, 124, 135, 141, 147, 156, 158).

[37] Xinan Xiu, Nyssa L Puskar, Jai AP Shanata, Henry A Lester et Dennis A Dougherty. „Nicotine binding to brain receptors requires a strong cation–π interaction“. In : *Nature* 458.7237 (2009), p. 534 (cf. p. 9, 133, 135).

[38] Oliver Beckstein et Mark SP Sansom. „A hydrophobic gate in an ion channel : the closed state of the nicotinic acetylcholine receptor“. In : *Physical biology* 3.2 (2006), p. 147 (cf. p. 9).

[39] Nicolas Le Novere, Pierre-Jean Corringer et Jean-Pierre Changeux. „The diversity of subunit composition in nAChRs : evolutionary origins, physiologic and pharmacologic consequences“. In : *Journal of neurobiology* 53.4 (2002), p. 447–456 (cf. p. 11).

[40] S Kracun, PC Harkness, AJ Gibb et NS Millar. „Influence of the M3–M4 intracellular domain upon nicotinic acetylcholine receptor assembly, targeting and function“. In : *British journal of pharmacology* 153.7 (2008), p. 1474–1484 (cf. p. 11).

[41] Nikolaos Kouvatsos, Athanasios Niarchos, Paraskevi Zisimopoulou et al. „Purification and functional characterization of a truncated human $\alpha 4\beta 2$ nicotinic acetylcholine receptor“. In : *International journal of biological macromolecules* 70 (2014), p. 320–326 (cf. p. 11, 25, 26).

[42] Ludovic Sauguet, Azadeh Shahsavar, Frédéric Poitevin et al. „Crystal structures of a pentameric ligand-gated ion channel provide a mechanism for activation“. In : *Proceedings of the National Academy of Sciences* 111.3 (2014), p. 966–971 (cf. p. 11, 41, 42, 158).

[43] Ryan E Hibbs et Eric Gouaux. „Principles of activation and permeation in an anion-selective Cys-loop receptor“. In : *Nature* 474.7349 (2011), p. 54 (cf. p. 11, 141, 147, 156, 159).

[44] Thorsten Althoff, Ryan E Hibbs, Surajit Banerjee et Eric Gouaux. „X-ray structures of GluCl in apo states reveal a gating mechanism of Cys-loop receptors“. In : *Nature* 512.7514 (2014), p. 333 (cf. p. 11, 26, 41, 42, 70).

[45] Juan Du, Wei Lü, Shenping Wu, Yifan Cheng et Eric Gouaux. „Glycine receptor mechanism elucidated by electron cryo-microscopy“. In : *Nature* 526.7572 (2015), p. 224 (cf. p. 11, 147, 156).

[46] Marie S Prevost, Ludovic Sauguet, Hugues Nury et al. „A locally closed conformation of a bacterial pentameric proton-gated ion channel“. In : *Nature structural & molecular biology* 19.6 (2012), p. 642 (cf. p. 11, 83).

[47] Jean-Pierre Changeux. „The nicotinic acetylcholine receptor : a typical ‘allosteric machine’“. In : *Phil. Trans. R. Soc. B* 373.1749 (2018), p. 20170174 (cf. p. 11).

[48] Antoine Taly, Marc Delarue, Thomas Grutter et al. „Normal mode analysis suggests a quaternary twist model for the nicotinic receptor gating mechanism“. In : *Biophysical journal* 88.6 (2005), p. 3954–3965 (cf. p. 11).

[49] Xiaolin Cheng, Benzhuo Lu, Barry Grant, Richard J Law et J Andrew McCammon. „Channel opening motion of $\alpha 7$ nicotinic acetylcholine receptor as suggested by normal mode analysis“. In : *Journal of molecular biology* 355.2 (2006), p. 310–324 (cf. p. 11).

[50] Hugues Nury, Frédéric Poitevin, Catherine Van Renterghem et al. „One-microsecond molecular dynamics simulation of channel gating in a nicotinic receptor homologue“. In : *Proceedings of the National Academy of Sciences* 107.14 (2010), p. 6275–6280 (cf. p. 11, 48, 71).

[51] Nicolas E Martin, Siddharth Malik, Nicolas Calimet, Jean-Pierre Changeux et Marco Cecchini. „Un-gating and allosteric modulation of a pentameric ligand-gated ion channel captured by molecular dynamics“. In : *PLoS computational biology* 13.10 (2017), e1005784 (cf. p. 11).

[52] Cyrus Chothia et Arthur M Lesk. „The relation between the divergence of sequence and structure in proteins.“ In : *The EMBO journal* 5.4 (1986), p. 823–826 (cf. p. 12).

[53] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer et al. „Gapped BLAST and PSI-BLAST : a new generation of protein database search programs“. In : *Nucleic acids research* 25.17 (1997), p. 3389–3402 (cf. p. 12, 26, 31).

[54] Helen M Berman, John Westbrook, Zukang Feng et al. „The protein data bank“. In : *Nucleic acids research* 28.1 (2000), p. 235–242 (cf. p. 12, 26).

[55] Robert C Edgar. „MUSCLE : multiple sequence alignment with high accuracy and high throughput“. In : *Nucleic acids research* 32.5 (2004), p. 1792–1797 (cf. p. 12).

[56] Cédric Notredame, Desmond G Higgins et Jaap Heringa. „T-coffee : a novel method for fast and accurate multiple sequence alignment1“. In : *Journal of molecular biology* 302.1 (2000), p. 205–217 (cf. p. 12, 27).

[57] Benjamin Webb et Andrej Sali. „Comparative protein structure modeling using MODELLER“. In : *Current protocols in protein science* 86.1 (2016), p. 2–9 (cf. p. 13, 27).

[58] Min-yi Shen et Andrej Sali. „Statistical potential for assessment and prediction of protein structures“. In : *Protein science* 15.11 (2006), p. 2507–2524 (cf. p. 13, 31).

[59] Francisco Melo, Roberto Sánchez et Andrej Sali. „Statistical potentials for fold assessment“. In : *Protein science* 11.2 (2002), p. 430–448 (cf. p. 13).

[60] Arjun Ray, Erik Lindahl et Björn Wallner. „Model quality assessment for membrane proteins“. In : *Bioinformatics* 26.24 (2010), p. 3067–3074 (cf. p. 13, 31).

[61] Alex D MacKerell Jr, Donald Bashford, MLDR Bellott et al. „All-atom empirical potential for molecular modeling and dynamics studies of proteins“. In : *The journal of physical chemistry B* 102.18 (1998), p. 3586–3616 (cf. p. 15, 16, 54, 57).

[62] Robert B Best, Xiao Zhu, Jihyun Shim et al. „Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles“. In : *Journal of chemical theory and computation* 8.9 (2012), p. 3257–3273 (cf. p. 15, 16, 54, 57).

[63] Alexander D MacKerell Jr, Michael Feig et Charles L Brooks. „Improved treatment of the protein backbone in empirical force fields“. In : *Journal of the American Chemical Society* 126.3 (2003), p. 698–699 (cf. p. 16, 54, 57).

[64] Bernard R Brooks, Charles L Brooks, Alexander D MacKerell et al. „CHARMM : the biomolecular simulation program“. In : *Journal of computational chemistry* 30.10 (2009), p. 1545–1614 (cf. p. 16, 54, 57).

[65] Nathan Schmid, Andreas P Eichenberger, Alexandra Choutko et al. „Definition and testing of the GROMOS force-field versions 54A7 and 54B7“. In : *European biophysics journal* 40.7 (2011), p. 843 (cf. p. 16).

[66] David A Case, Thomas E Cheatham, Tom Darden et al. „The Amber biomolecular simulation programs“. In : *Journal of computational chemistry* 26.16 (2005), p. 1668–1688 (cf. p. 16).

[67] Paul P Ewald. „Die Berechnung optischer und elektrostatischer Gitterpotentiale“. In : *Annalen der physik* 369.3 (1921), p. 253–287 (cf. p. 16).

[68] Ulrich Essmann, Lalith Perera, Max L Berkowitz et al. „A smooth particle mesh Ewald method“. In : *The Journal of chemical physics* 103.19 (1995), p. 8577–8593 (cf. p. 16, 57).

[69] Michael Schaefer et Martin Karplus. „A comprehensive analytical treatment of continuum electrostatics“. In : *The Journal of Physical Chemistry* 100.5 (1996), p. 1578–1599 (cf. p. 16).

[70] Raymond Constanciel et Renato Contreras. „Self consistent field theory of solvent effects representation by continuum models : Introduction of desolvation contribution“. In : *Theoretica chimica acta* 65.1 (1984), p. 1–11 (cf. p. 16).

[71] Joseph D Bryngelson, Jose Nelson Onuchic, Nicholas D Socci et Peter G Wolynes. „Funnels, pathways, and the energy landscape of protein folding : a synthesis“. In : *Proteins : Structure, Function, and Bioinformatics* 21.3 (1995), p. 167–195 (cf. p. 17).

[72] Christian B Anfinsen. „Principles that govern the folding of protein chains“. In : *Science* 181.4096 (1973), p. 223–230 (cf. p. 17).

[73] Alessandro Laio et Michele Parrinello. „Escaping free-energy minima“. In : *Proceedings of the National Academy of Sciences* 99.20 (2002), p. 12562–12566 (cf. p. 19, 49).

[74] Glenn M Torrie et John P Valleau. „Nonphysical sampling distributions in Monte Carlo free-energy estimation : Umbrella sampling“. In : *Journal of Computational Physics* 23.2 (1977), p. 187–199 (cf. p. 19, 49).

[75] Jean-Pierre Changeux, Michiki Kasai et Chen-Yuan Lee. „Use of a snake venom toxin to characterize the cholinergic receptor protein“. In : *Proceedings of the National Academy of Sciences* 67.3 (1970), p. 1241–1247 (cf. p. 22).

[76] Hao Cheng, Chen Fan, Si-wei Zhang et al. „Crystallization scale purification of α7 nicotinic acetylcholine receptor from mammalian cells using a BacMam expression system“. In : *Acta Pharmacologica Sinica* 36.8 (2015), p. 1013 (cf. p. 22).

[77] John E Baenziger, Stephen E Ryan, Michael M Goodreid et al. „Lipid composition alters drug action at the nicotinic acetylcholine receptor“. In : *Molecular pharmacology* 73.3 (2008), p. 880–890 (cf. p. 22).

[78] Gareth T Young, Ruud Zwart, Alison S Walker, Emanuele Sher et Neil S Millar. „Potentiation of α7 nicotinic acetylcholine receptors via an allosteric transmembrane site“. In : *Proceedings of the National Academy of Sciences* 105.38 (2008), p. 14686–14691 (cf. p. 22, 142, 156, 159, 160).

[79] Jaimee Allison Domville. „Mapping the Allosteric Pathway Leading from a Mutation in the Nicotinic Acetylcholine Receptor to a Congenital Myasthenic Syndrome“. Thèse de doct. Université d’Ottawa/University of Ottawa, 2017 (cf. p. 22).

[80] Joseph Newcombe, Anna Chatzidaki, Tom D Sheppard, Maya Topf et Neil S Millar. „Diversity of nicotinic acetylcholine receptor positive allosteric modulators revealed by mutagenesis and a revised structural model“. In : *Molecular pharmacology* (2017), mol–117 (cf. p. 22, 40, 142, 156, 160).

[81] Richard W Olsen, Guo-Dong Li, Martin Wallner et al. „Structural models of ligand-gated ion channels : sites of action for anesthetics and ethanol“. In : *Alcoholism : Clinical and Experimental Research* 38.3 (2014), p. 595–603 (cf. p. 22).

[82] Megan O'Mara, Brett Cromer, Michael Parker et Shin-Ho Chung. „Homology model of the GABA A receptor examined using Brownian dynamics". In : *Biophysical journal* 88.5 (2005), p. 3286–3299 (cf. p. 22).

[83] Luis M Valor, José Mulet, Francisco Sala et al. „Role of the large cytoplasmic loop of the $\alpha 7$ neuronal nicotinic acetylcholine receptor subunit in receptor expression and function". In : *Biochemistry* 41.25 (2002), p. 7931–7938 (cf. p. 25).

[84] Thomas Madden. „The BLAST sequence analysis tool". In : (2013) (cf. p. 26).

[85] Ákos Nemecz et Palmer Taylor. „Creating an $\alpha 7$ Nicotinic Acetylcholine Recognition Domain from the Acetylcholine-binding Protein CRYSTALLOGRAPHIC AND LIGAND SELECTIVITY ANALYSES". In : *Journal of Biological Chemistry* 286.49 (2011), p. 42555–42565 (cf. p. 26, 135, 141).

[86] Ghérici Hassaine, Cédric Deluz, Luigino Grasso et al. „X-ray structure of the mouse serotonin 5-HT 3 receptor". In : *Nature* 512.7514 (2014), p. 276 (cf. p. 27).

[87] L Chen, CD Dellisanti, Y Yao, CJ Stroud et ZZ Wang. „Crystal structure of the extracellular domain of nAChRa1 bound to alpha-bungarotoxin at 1.94 Å resolution". In : *Nature Neurosci* 10 (2007), p. 953–962 (cf. p. 27).

[88] Marios Zouridakis, Petros Giastas, Eleftherios Zarkadas et al. „Crystal structures of free and antagonist-bound states of human $\alpha 9$ nicotinic receptor extracellular domain". In : *Nature Structural and Molecular Biology* 21.11 (2014), p. 976 (cf. p. 27).

[89] Anil Bhattacharyya. „On a measure of divergence between two statistical populations defined by their probability distributions". In : *Bull. Calcutta Math. Soc.* 35 (1943), p. 99–109 (cf. p. 30).

[90] Sarel J Fleishman, Andrew Leaver-Fay, Jacob E Corn et al. „RosettaScripts : a scripting language interface to the Rosetta macromolecular modeling suite". In : *PloS one* 6.6 (2011), e20161 (cf. p. 31).

[91] Michael D Tyka, Daniel A Keedy, Ingemar André et al. „Alternate states of proteins revealed by detailed energy landscape mapping". In : *Journal of molecular biology* 405.2 (2011), p. 607–618 (cf. p. 31).

[92] P Barth, Jack Schonbrun et David Baker. „Toward high-resolution prediction and design of transmembrane helical protein structures". In : *Proceedings of the National Academy of Sciences* 104.40 (2007), p. 15682–15687 (cf. p. 31, 35, 37).

[93] Carol A Rohl, Charlie EM Strauss, Kira MS Misura et David Baker. „Protein structure prediction using Rosetta". In : *Methods in enzymology*. T. 383. Elsevier, 2004, p. 66–93 (cf. p. 31).

[94] Konstantinos D Tsirigos, Christoph Peters, Nanjiang Shu, Lukas Käll et Arne Elofsson. „The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides". In : *Nucleic acids research* 43.W1 (2015), W401–W407 (cf. p. 31).

[95] Baris E Suzek, Yuqi Wang, Hongzhan Huang et al. „UniRef clusters : a comprehensive and scalable alternative for improving sequence similarity searches". In : *Bioinformatics* 31.6 (2014), p. 926–932 (cf. p. 31).

[96] Roman A Laskowski, Malcolm W MacArthur, David S Moss et Janet M Thornton. „PROCHECK : a program to check the stereochemical quality of protein structures“. In : *Journal of applied crystallography* 26.2 (1993), p. 283–291 (cf. p. 31).

[97] Pascal Benkert, Marco Biasini et Torsten Schwede. „Toward the estimation of the absolute quality of individual protein structure models“. In : *Bioinformatics* 27.3 (2010), p. 343–350 (cf. p. 31).

[98] Vincent B Chen, W Bryan Arendall, Jeffrey J Headd et al. „MolProbity : all-atom structure validation for macromolecular crystallography“. In : *Acta Crystallographica Section D : Biological Crystallography* 66.1 (2010), p. 12–21 (cf. p. 31).

[99] David Eisenberg, Roland Lüthy et James U Bowie. „[20] VERIFY3D : Assessment of protein models with three-dimensional profiles“. In : *Methods in enzymology*. T. 277. Elsevier, 1997, p. 396–404 (cf. p. 31).

[100] Markus Wiederstein et Manfred J Sippl. „ProSA-web : interactive web service for the recognition of errors in three-dimensional structures of proteins“. In : *Nucleic acids research* 35.suppl_2 (2007), W407–W410 (cf. p. 31).

[101] Nigel Unwin et Yoshinori Fujiyoshi. „Gating movement of acetylcholine receptor caught by plunge-freezing“. In : *Journal of molecular biology* 422.5 (2012), p. 617–634 (cf. p. 40, 156).

[102] Nelli Mnatsakanyan et Michaela Jansen. „Experimental determination of the vertical alignment between the second and third transmembrane segments of muscle nicotinic acetylcholine receptors“. In : *Journal of neurochemistry* 125.6 (2013), p. 843–854 (cf. p. 40, 156).

[103] Pierre-Jean Corringer, Marc Baaden, Nicolas Bocquet et al. „Atomic structure and dynamics of pentameric ligand-gated ion channels : new insight from bacterial homologues“. In : *The Journal of physiology* 588.4 (2010), p. 565–572 (cf. p. 40, 156).

[104] Paul S Miller et A Radu Aricescu. „Crystal structure of a human GABA A receptor“. In : *Nature* 512.7514 (2014), p. 270 (cf. p. 40).

[105] Marco Cecchini et Jean-Pierre Changeux. „The nicotinic acetylcholine receptor and its prokaryotic homologues : Structure, conformational transitions & allosteric modulation“. In : *Neuropharmacology* 96 (2015), p. 137–149 (cf. p. 41, 71).

[106] Katarzyna Kaczanowska, Michal Harel, Zoran Radić et al. „Structural basis for cooperative interactions of substituted 2-aminopyrimidines with the acetylcholine binding protein“. In : *Proceedings of the National Academy of Sciences* (2014), p. 201410992 (cf. p. 42, 72).

[107] *TReK : a program for Trajectory REfinement and Kinematics*. https://www.charmm.org/charmm/documentation/by-version/c42b1/params/doc/trek/. CHARMM Documentation : Version 2.10 , July 5-2003. (cf. p. 44, 55, 64).

[108] Dongxiang Liu, Yechun Xu, Yu Feng et al. „Inhibitor discovery targeting the intermediate structure of β-amyloid peptide on the conformational transition pathway : implications in the aggregation mechanism of β-amyloid peptide“. In : *Biochemistry* 45.36 (2006), p. 10963–10972 (cf. p. 47).

[109] Elodie Laine, Julliane D Yoneda, Arnaud Blondel et Thérese E Malliavin. „The conformational plasticity of calmodulin upon calcium complexation gives a model of its interaction with the oedema factor of Bacillus anthracis“. In : *Proteins : Structure, Function, and Bioinformatics* 71.4 (2008), p. 1813–1829 (cf. p. 47, 97, 157).

[110] Wenxun Gan, Sichun Yang et Benoît Roux. „Atomistic view of the conformational activation of Src kinase using the string method with swarms-of-trajectories“. In : *Biophysical journal* 97.4 (2009), p. L8–L10 (cf. p. 47, 49, 81).

[111] Albert C Pan, Deniz Sezer et Benoît Roux. „Finding transition pathways using the string method with swarms of trajectories“. In : *The journal of physical chemistry B* 112.11 (2008), p. 3432–3440 (cf. p. 47, 49, 57, 65, 81).

[112] Nina M Goodey et Stephen J Benkovic. „Allosteric regulation and catalysis emerge via a common route“. In : *Nature chemical biology* 4.8 (2008), p. 474 (cf. p. 48).

[113] Rani P Venkitakrishnan, Eduardo Zaborowski, Dan McElheny et al. „Conformational changes in the active site loops of dihydrofolate reductase during the catalytic cycle“. In : *Biochemistry* 43.51 (2004), p. 16046–16055 (cf. p. 48).

[114] Ruth Nussinov, Chung-Jung Tsai et Buyong Ma. „The underappreciated role of allostery in the cellular network“. In : *Annual review of biophysics* 42 (2013), p. 169–189 (cf. p. 48).

[115] Dorothee Kern et Erik RP Zuiderweg. „The role of dynamics in allosteric regulation“. In : *Current opinion in structural biology* 13.6 (2003), p. 748–757 (cf. p. 48).

[116] Ahmed F Abdel-Magid. *Allosteric Modulators : An Emerging Concept in Drug Discovery*. 2015 (cf. p. 48).

[117] Shaoyong Lu, Mingfei Ji, Duan Ni et Jian Zhang. „Discovery of hidden allosteric sites as novel targets for allosteric drug design“. In : *Drug discovery today* (2017) (cf. p. 48).

[118] C Levinthal. *How to Fold Graciously. Mossbauer Spectroscopy in Biological Systems, Allerton House*. 1969 (cf. p. 48).

[119] Marcus Fischer, Brian K Shoichet et James S Fraser. „One crystal, two temperatures : cryocooling penalties alter ligand binding to transient protein sites“. In : *ChemBioChem* 16.11 (2015), p. 1560–1564 (cf. p. 48).

[120] Thomas Szyperski. „Room Temperature X-Ray Crystallography Reveals Conformational Heterogeneity of Engineered Proteins“. In : *Structure* 25.5 (2017), p. 691–692 (cf. p. 48).

[121] Meyer B Jackson. „Spontaneous openings of the acetylcholine receptor channel“. In : *Proceedings of the National Academy of Sciences* 81.12 (1984), p. 3901–3904 (cf. p. 48).

[122] Cameron Abrams et Giovanni Bussi. „Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration“. In : *Entropy* 16.1 (2013), p. 163–199 (cf. p. 48).

[123] Vojtech Spiwok, Zoran Sucur et Petr Hosek. „Enhanced sampling techniques in biomolecular simulations“. In : *Biotechnology advances* 33.6 (2015), p. 1130–1140 (cf. p. 48).

[124] John G Kirkwood. „Statistical mechanics of fluid mixtures“. In : *The Journal of Chemical Physics* 3.5 (1935), p. 300–313 (cf. p. 49).

[125] Jérôme Hénin et Christophe Chipot. „Overcoming free energy barriers using unconstrained molecular dynamics simulations“. In : *The Journal of chemical physics* 121.7 (2004), p. 2904–2914 (cf. p. 49).

[126] Luca Maragliano et Eric Vanden-Eijnden. „A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations“. In : *Chemical physics letters* 426.1-3 (2006), p. 168–175 (cf. p. 49).

[127] Peter G Bolhuis, David Chandler, Christoph Dellago et Phillip L Geissler. „Transition path sampling : Throwing ropes over rough mountain passes, in the dark“. In : *Annual review of physical chemistry* 53.1 (2002), p. 291–318 (cf. p. 49).

[128] Sergei V Krivov et Martin Karplus. „Hidden complexity of free energy surfaces for peptide (protein) folding“. In : *Proceedings of the National Academy of Sciences of the United States of America* 101.41 (2004), p. 14766–14770 (cf. p. 49).

[129] Luca Maragliano, Alexander Fischer, Eric Vanden-Eijnden et Giovanni Ciccotti. „String method in collective variables : Minimum free energy paths and isocommittor surfaces“. In : *The Journal of chemical physics* 125.2 (2006), p. 024106 (cf. p. 49).

[130] Chunfeng Zhao et Sergei Yu Noskov. „The molecular mechanism of ion-dependent gating in secondary transporters“. In : *PLoS computational biology* 9.10 (2013), e1003296 (cf. p. 49).

[131] Adam Chamberlin, Feng Qiu, Yibo Wang, Sergei Y Noskov et H Peter Larsson. „Mapping the gating and permeation pathways in the voltage-gated proton channel Hv1“. In : *Journal of molecular biology* 427.1 (2015), p. 131–145 (cf. p. 49).

[132] Victor Ovchinnikov, Martin Karplus et Eric Vanden-Eijnden. „Free energy of conformational transition paths in biomolecules : The string method and its application to myosin VI“. In : *The Journal of chemical physics* 134.8 (2011), 02B631 (cf. p. 49, 69).

[133] Jérôme J Lacroix, Stephan A Pless, Luca Maragliano et al. „Intermediate state trapping of a voltage sensor“. In : *The Journal of general physiology* 140.6 (2012), p. 635–652 (cf. p. 49).

[134] Melchor Sanchez-Martinez, Martin Field et Ramon Crehuet. „Enzymatic minimum free energy path calculations using swarms of trajectories“. In : *The Journal of Physical Chemistry B* 119.3 (2014), p. 1103–1113 (cf. p. 49).

[135] F Pontiggia, DV Pachov, MW Clarkson et al. „Free energy landscape of activation in a signalling protein at atomic resolution“. In : *Nature communications* 6 (2015), p. 7284 (cf. p. 49).

[136] Yasuhiro Matsunaga, Hiroshi Fujisaki, Tohru Terada et al. „Minimum free energy path of ligand-induced transition in adenylate kinase“. In : *PLoS computational biology* 8.6 (2012), e1002555 (cf. p. 49).

[137] Abhishek Singharoy, Christophe Chipot, Mahmoud Moradi et Klaus Schulten. „Chemomechanical coupling in hexameric protein–protein interfaces harnesses energy within V-Type ATPases“. In : *Journal of the American Chemical Society* 139.1 (2016), p. 293–310 (cf. p. 49).

[138] Yasuhiro Matsunaga, Tsutomu Yamane, Tohru Terada et al. „Energetics and conformational pathways of functional rotation in the multidrug transporter AcrB“. In : *Elife* 7 (2018), e31715 (cf. p. 49).

[139] Avisek Das, Huan Rui, Robert Nakamoto et Benoît Roux. „Conformational transitions and alternating-access mechanism in the sarcoplasmic reticulum calcium pump“. In : *Journal of molecular biology* 429.5 (2017), p. 647–666 (cf. p. 49).

[140] Mahmoud Moradi, Giray Enkavi et Emad Tajkhorshid. „Atomic-level characterization of transport cycle thermodynamics in the glycerol-3-phosphate : phosphate antiporter“. In : *Nature communications* 6 (2015), p. 8393 (cf. p. 49).

[141] Fangqiang Zhu et Gerhard Hummer. „Pore opening and closing of a pentameric ligand-gated ion channel“. In : *Proceedings of the National Academy of Sciences* 107.46 (2010), p. 19814–19819 (cf. p. 49).

[142] Bogdan Lev, Samuel Murail, Frédéric Poitevin et al. „String method solution of the gating pathways for a pentameric ligand-gated ion channel“. In : *Proceedings of the National Academy of Sciences* 114.21 (2017), E4158–E4167 (cf. p. 49, 71, 81).

[143] Yilin Meng, Diwakar Shukla, Vijay S Pande et Benoît Roux. „Transition path theory analysis of c-Src kinase activation“. In : *Proceedings of the National Academy of Sciences* 113.33 (2016), p. 9193–9198 (cf. p. 49).

[144] Hannes Jónsson, Greg Mills et Karsten W Jacobsen. „Nudged elastic band method for finding minimum energy paths of transitions“. In : *Classical and quantum dynamics in condensed phase simulations*. World Scientific, 1998, p. 385–404 (cf. p. 50).

[145] E Weinan, Weiqing Ren et Eric Vanden-Eijnden. „String method for the study of rare events“. In : *Physical Review B* 66.5 (2002), p. 052301 (cf. p. 50).

[146] E Weinan, Weiqing Ren et Eric Vanden-Eijnden. „Simplified and improved string method for computing the minimum energy paths in barrier-crossing events“. In : *Journal of Chemical Physics* 126.16 (2007), p. 164103 (cf. p. 50).

[147] Stefan Fischer et Martin Karplus. „Conjugate peak refinement : an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom“. In : *Chemical physics letters* 194.3 (1992), p. 252–261 (cf. p. 50, 65).

[148] Stefan Fischer, Björn Windshügel, Daniel Horak, Kenneth C Holmes et Jeremy C Smith. „Structural mechanism of the recovery stroke in the myosin molecular motor“. In : *Proceedings of the National Academy of Sciences of the United States of America* 102.19 (2005), p. 6873–6878 (cf. p. 51).

[149] Sidonia Mesentean, Stefan Fischer et Jeremy C Smith. „Analyzing large-scale structural change in proteins : Comparison of principal component projection and sammon mapping“. In : *Proteins : Structure, Function, and Bioinformatics* 64.1 (2006), p. 210–218 (cf. p. 51).

[150] Elodie Laine, Christophe Goncalves, Johanna C Karst et al. „Use of allostery to identify inhibitors of calmodulin-induced activation of Bacillus anthracis edema factor“. In : *Proceedings of the National Academy of Sciences* 107.25 (2010), p. 11277–11282 (cf. p. 51, 86, 87).

[151] Paola Conti, Lucia Tamborini, Andrea Pinto et al. „Drug discovery targeting amino acid racemases“. In : *Chemical reviews* 111.11 (2011), p. 6919–6946 (cf. p. 51).

[152] Sunhwan Jo, Taehoon Kim, Vidyashankara G Iyer et Wonpil Im. „CHARMM-GUI : a web-based graphical user interface for CHARMM“. In : *Journal of computational chemistry* 29.11 (2008), p. 1859–1865 (cf. p. 53).

[153] Tung Ming Fong et Mark G McNamee. „Correlation between acetylcholine receptor function and structural properties of membranes“. In : *Biochemistry* 25.4 (1986), p. 830–840 (cf. p. 53).

[154] MP McCarthy et Marjorie A Moore. „Effects of lipids and detergents on the conformation of the nicotinic acetylcholine receptor from Torpedo californica.“ In : *Journal of Biological Chemistry* 267.11 (1992), p. 7655–7663 (cf. p. 53).

[155] Saffron E Rankin, George H Addona, Marek A Kloczewiak, Birgitte Bugge et Keith W Miller. „The cholesterol dependence of activation and fast desensitization of the nicotinic acetylcholine receptor“. In : *Biophysical journal* 73.5 (1997), p. 2446–2455 (cf. p. 53).

[156] JB Corrie, Andrei A Ogrel, Elizabeth A McCardy, Michael P Blanton et John E Baenziger. „Lipid-protein interactions at the nicotinic acetylcholine receptor A functional coupling between nicotinic receptors and phosphatidic acid-containing lipid bilayers“. In : *Journal of Biological Chemistry* 277.1 (2002), p. 201–208 (cf. p. 53).

[157] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey et Michael L Klein. „Comparison of simple potential functions for simulating liquid water“. In : *The Journal of chemical physics* 79.2 (1983), p. 926–935 (cf. p. 53, 57).

[158] M Toulouse, V Fritsch et E Westhof. „Rapid Calculation of Any Dielectric Function for Molecular Dynamics Simulations of Biological Macromolecules“. In : *OA drug design & delivery* 9.3 (1992), p. 193–200 (cf. p. 54).

[159] Mariama Jaiteh, Antoine Taly et Jérôme Hénin. „Evolution of pentameric ligand-gated ion channels : Pro-loop receptors“. In : *PloS one* 11.3 (2016), e0151934 (cf. p. 54).

[160] Jeffery B Klauda, Richard M Venable, J Alfredo Freites et al. „Update of the CHARMM all-atom additive force field for lipids : validation on six lipid types“. In : *The journal of physical chemistry B* 114.23 (2010), p. 7830–7843 (cf. p. 57).

[161] Jeffery B Klauda, Viviana Monje, Taehoon Kim et Wonpil Im. „Improving the CHARMM force field for polyunsaturated fatty acid chains“. In : *The journal of physical chemistry B* 116.31 (2012), p. 9424–9431 (cf. p. 57).

[162] Kenno Vanommeslaeghe, Elizabeth Hatcher, Chayan Acharya et al. „CHARMM general force field : A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields“. In : *Journal of computational chemistry* 31.4 (2010), p. 671–690 (cf. p. 57).

[163] Wenbo Yu, Xibing He, Kenno Vanommeslaeghe et Alexander D MacKerell. „Extension of the CHARMM general force field to sulfonyl-containing compounds and its utility in biomolecular simulations“. In : *Journal of computational chemistry* 33.31 (2012), p. 2451–2468 (cf. p. 57).

[164] Dmitrii Beglov et Benoit Roux. „Finite representation of an infinite bulk system : solvent boundary potential for computer simulations“. In : *The Journal of chemical physics* 100.12 (1994), p. 9050–9063 (cf. p. 57).

[165] Jean-Paul Ryckaert, Giovanni Ciccotti et Herman JC Berendsen. „Numerical integration of the cartesian equations of motion of a system with constraints : molecular dynamics of n-alkanes“. In : *Journal of Computational Physics* 23.3 (1977), p. 327–341 (cf. p. 58).

[166] Antti-Pekka Hynninen et Michael F Crowley. „New faster CHARMM molecular dynamics engine“. In : *Journal of computational chemistry* 35.5 (2014), p. 406–413 (cf. p. 58).

[167] Florian J Gisdon, Martin Culka et G Matthias Ullmann. „PyCPR–a python-based implementation of the Conjugate Peak Refinement (CPR) algorithm for finding transition state structures“. In : *Journal of molecular modeling* 22.10 (2016), p. 242 (cf. p. 65).

[168] Davide Branduardi et José D Faraldo-Gómez. „String Method for Calculation of Minimum Free-Energy Paths in Cartesian Space in Freely Tumbling Systems“. In : *Journal of chemical theory and computation* 9.9 (2013), p. 4140–4154 (cf. p. 69).

[169] Christopher Miller. „Genetic manipulation of ion channels : a new approach to structure and mechanism“. In : *Neuron* 2.3 (1989), p. 1195–1205 (cf. p. 71).

[170] F Revah, D Bertrand, J-L Galzi et al. „Mutations in the channel domain alter desensitization of a neuronal nicotinic receptor“. In : *Nature* 353.6347 (1991), p. 846 (cf. p. 70).

[171] Cesar Labarca, Mark W Nowak, Haiyun Zhang et al. „Channel gating governed symmetrically by conserved leucine residues in the M2 domain of nicotinic receptors“. In : *Nature* 376.6540 (1995), p. 514 (cf. p. 70).

[172] Gregory N Filatov et Michael M White. „The role of conserved leucines in the M2 domain of the acetylcholine receptor in channel gating.“ In : *Molecular Pharmacology* 48.3 (1995), p. 379–384 (cf. p. 70).

[173] Jean-Pierre Changeux et Arthur Christopoulos. „Allosteric modulation as a unifying mechanism for receptor function and regulation“. In : *Cell* 166.5 (2016), p. 1084–1102 (cf. p. 71).

[174] Nicolas Calimet, Manuel Simoes, Jean-Pierre Changeux et al. „A gating mechanism of pentameric ligand-gated ion channels“. In : *Proceedings of the National Academy of Sciences* (2013), p. 201313785 (cf. p. 71).

[175] August B Smit, Katjuša Brejc, Naweed Syed et Titia K Sixma. „Structure and function of AChBP, homologue of the ligand-binding domain of the nicotinic acetylcholine receptor“. In : *Annals of the New York Academy of Sciences* 998.1 (2003), p. 81–92 (cf. p. 72).

[176] Patrick HN Celie, Sarah E van Rossum-Fikkert, Willem J van Dijk et al. „Nicotine and carbamylcholine binding to nicotinic acetylcholine receptors as studied in AChBP crystal structures“. In : *Neuron* 41.6 (2004), p. 907–914 (cf. p. 72).

[177] Jean-Pierre Changeux et Stuart J Edelstein. *The brain as a chemical machine : nicotinic receptors and neuronal communication*. Odile Jacob, 2012 (cf. p. 72).

[178] Javier Pérez, Jean-Marc Zanotti et Dominique Durand. „Evolution of the internal dynamics of two globular proteins from dry powder to solution“. In : *Biophysical journal* 77.1 (1999), p. 454–469 (cf. p. 81).

[179] Remo Perozzo, Gerd Folkers et Leonardo Scapozza. „Thermodynamics of protein–ligand interactions : history, presence, and future aspects“. In : *Journal of Receptors and Signal Transduction* 24.1-2 (2004), p. 1–52 (cf. p. 86).

[180] Jie Liang, Clare Woodward et Herbert Edelsbrunner. „Anatomy of protein pockets and cavities : measurement of binding site geometry and implications for ligand design“. In : *Protein science* 7.9 (1998), p. 1884–1897 (cf. p. 86, 87).

[181] Mark A Williams, Julia M Goodfellow et Janet M Thornton. „Buried waters and internal cavities in monomeric proteins“. In : *Protein Science* 3.8 (1994), p. 1224–1235 (cf. p. 86).

[182] Sandhya Kortagere, Matthew D Krasowski et Sean Ekins. „The importance of discerning shape in molecular pharmacology“. In : *Trends in pharmacological sciences* 30.3 (2009), p. 138–147 (cf. p. 86).

[183] Juan Alvarez et Brian Shoichet. *Virtual screening in drug discovery*. CRC press, 2005 (cf. p. 86).

[184] Andy Jennings. „Chemical Informatics : Using Molecular Shape Descriptors in Structure-Based Drug Design“. In : *Structure-Based Drug Discovery*. Springer, 2012, p. 235–250 (cf. p. 86).

[185] Nataraj S Pagadala, Khajamohiddin Syed et Jack Tuszynski. „Software for molecular docking : a review“. In : *Biophysical reviews* 9.2 (2017), p. 91–102 (cf. p. 86).

[186] Amy C Anderson. „The process of structure-based drug design“. In : *Chemistry & biology* 10.9 (2003), p. 787–797 (cf. p. 86).

[187] Miles Congreve, Christopher W Murray et Tom L Blundell. „Keynote review : Structural biology and drug discovery“. In : *Drug discovery today* 10.13 (2005), p. 895–907 (cf. p. 86).

[188] Jaeju Ko, Leonel F Murga, Ying Wei et Mary Jo Ondrechen. „Prediction of active sites for protein structures from computed chemical properties“. In : *Bioinformatics* 21.suppl_1 (2005), p. i258–i265 (cf. p. 86).

[189] Sriram Sankararaman, Fei Sha, Jack F Kirsch, Michael I Jordan et Kimmen Sjölander. „Active site prediction using evolutionary and structural information“. In : *Bioinformatics* 26.5 (2010), p. 617–624 (cf. p. 86).

[190] Craig T Porter, Gail J Bartlett et Janet M Thornton. „The Catalytic Site Atlas : a resource of catalytic sites and residues identified in enzymes using structural data“. In : *Nucleic acids research* 32.suppl_1 (2004), p. D129–D133 (cf. p. 86).

[191] Arthur Christopoulos. „Allosteric binding sites on cell-surface receptors : novel targets for drug discovery“. In : *Nature reviews Drug discovery* 1.3 (2002), p. 198–210 (cf. p. 86).

[192] P Jeffrey Conn, Arthur Christopoulos et Craig W Lindsley. „Allosteric modulators of GPCRs : a novel approach for the treatment of CNS disorders“. In : *Nature reviews Drug discovery* 8.1 (2009), p. 41–54 (cf. p. 86).

[193] Lauren T May, Katie Leach, Patrick M Sexton et Arthur Christopoulos. „Allosteric modulation of G protein–coupled receptors“. In : *Annu. Rev. Pharmacol. Toxicol.* 47 (2007), p. 1–51 (cf. p. 86).

[194] Jae-Seong Yang, Sang Woo Seo, Sungho Jang, Gyoo Yeol Jung et Sanguk Kim. „Rational engineering of enzyme allosteric regulation through sequence evolution analysis“. In : *PLoS computational biology* 8.7 (2012), e1002612 (cf. p. 86).

[195] A Christopoulos, LT May, VA Avlani et PM Sexton. *G-protein-coupled receptor allosterism : the promise and the problem (s)*. 2004 (cf. p. 86).

[196] Terry Kenakin et Laurence J Miller. „Seven transmembrane receptors as shapeshifting proteins : the impact of allosteric modulation and functional selectivity on new drug discovery“. In : *Pharmacological reviews* (2010), pr–108 (cf. p. 86).

[197] Zhizhou Fang, Christian Grütter et Daniel Rauh. „Strategies for the selective regulation of kinases with allosteric modulators : exploiting exclusive structural features“. In : *ACS chemical biology* 8.1 (2012), p. 58–70 (cf. p. 86).

[198] Shaoyong Lu, Wenkang Huang et Jian Zhang. „Recent computational advances in the identification of allosteric sites in proteins“. In : *Drug discovery today* 19.10 (2014), p. 1595–1600 (cf. p. 86).

[199] Wenkang Huang, Ruth Nussinov et Jian Zhang. „Computational Tools for Allosteric Drug Discovery : Site Identification and Focus Library Design“. In : *Computational Protein Design*. Springer, 2017, p. 439–446 (cf. p. 86).

[200] Joe G Greener et Michael JE Sternberg. „Structure-based prediction of protein allostery“. In : *Current opinion in structural biology* 50 (2018), p. 1–8 (cf. p. 86).

[201] Stephen J Benkovic et Sharon Hammes-Schiffer. „A perspective on enzyme catalysis“. In : *Science* 301.5637 (2003), p. 1196–1202 (cf. p. 87).

[202] Dimitri Antoniou et Steven D Schwartz. „Internal enzyme motions as a source of catalytic activity : rate-promoting vibrations and hydrogen tunneling“. In : *The Journal of Physical Chemistry B* 105.23 (2001), p. 5553–5558 (cf. p. 87).

[203] Vishal C Nashine, Sharon Hammes-Schiffer et Stephen J Benkovic. „Coupled motions in enzyme catalysis“. In : *Current opinion in chemical biology* 14.5 (2010), p. 644–651 (cf. p. 87).

[204] Robert F. Tilton, Irwin D. Kuntz et Gregory A. Petsko. „Cavities in proteins : structure of a metmyoglobin xenon complex solved to 1.9 Å“. In : *Biochemistry* 23.13 (juin 1984), p. 2849–2857 (cf. p. 87, 97).

[205] Maurizio Brunori et Quentin H Gibson. „Cavities and packing defects in the structural dynamics of myoglobin“. In : *EMBO reports* 2.8 (2001), p. 674–679 (cf. p. 87).

[206] Jory Z Ruscio, Deept Kumar, Maulik Shukla et al. „Atomic level computational identification of ligand migration pathways between solvent and binding site in myoglobin“. In : *Proceedings of the National Academy of Sciences* 105.27 (2008), p. 9204–9209 (cf. p. 87).

[207] Cecilia Bossa, Andrea Amadei, Isabella Daidone et al. „Molecular dynamics simulation of sperm whale myoglobin : effects of mutations and trapped CO on the structure and dynamics of cavities“. In : *Biophysical journal* 89.1 (2005), p. 465–474 (cf. p. 87).

[208] Ayana Tomita, Tokushi Sato, Kouhei Ichiyanagi et al. „Visualizing breathing motion of internal cavities in concert with ligand migration in myoglobin“. In : *Proceedings of the National Academy of Sciences* 106.8 (2009), p. 2612–2616 (cf. p. 87).

[209] Mariano Andrea Scorciapino, Arturo Robertazzi, Mariano Casu, Paolo Ruggerone et Matteo Ceccarelli. „Breathing motions of a respiratory protein revealed by molecular dynamics simulations“. In : *Journal of the American Chemical Society* 131.33 (2009), p. 11825–11832 (cf. p. 87).

[210] Matteo Gabba, Stefania Abbruzzetti, Francesca Spyrakis et al. „CO rebinding kinetics and molecular dynamics simulations highlight dynamic regulation of internal cavities in human cytoglobin“. In : *PLoS One* 8.1 (2013), e49770 (cf. p. 87).

[211] Mee Kian Poh, Andy Yip, Summer Zhang et al. „A small molecule fusion inhibitor of dengue virus.“ In : *Antiviral Res.* 84.3 (déc. 2009), p. 260–6 (cf. p. 87, 109).

[212] Wei-Jen Tang et Qing Guo. „The adenylyl cyclase activity of anthrax edema factor“. In : *Molecular aspects of medicine* 30.6 (2009), p. 423–430 (cf. p. 87).

[213] Zhimin Huang, Liang Zhu, Yan Cao et al. „ASD : a comprehensive database of allosteric proteins and modulators“. In : *Nucleic acids research* 39.suppl_1 (2010), p. D663–D669 (cf. p. 87).

[214] Susanne Eyrisch et Volkhard Helms. „Transient pockets on protein surfaces involved in protein- protein interaction“. In : *Journal of medicinal chemistry* 50.15 (2007), p. 3457–3464 (cf. p. 87).

[215] Murad Nayal et Barry Honig. „On the nature of cavities on protein surfaces : application to the identification of drug-binding sites“. In : *Proteins : Structure, Function, and Bioinformatics* 63.4 (2006), p. 892–906 (cf. p. 87).

[216] Stéphanie Pérot, Olivier Sperandio, Maria A Miteva, Anne-Claude Camproux et Bruno O Villoutreix. „Druggable pockets and binding site centric chemical space : a paradigm shift in drug discovery“. In : *Drug discovery today* 15.15 (2010), p. 656–667 (cf. p. 87).

[217] Holger Claußen, Christian Buning, Matthias Rarey et Thomas Lengauer. „FlexE : efficient molecular docking considering protein structure variations“. In : *Journal of molecular biology* 308.2 (2001), p. 377–395 (cf. p. 87).

[218] Anna Maria Ferrari, Binqing Q Wei, Luca Costantino et Brian K Shoichet. „Soft docking and multiple receptor conformations in virtual screening“. In : *Journal of medicinal chemistry* 47.21 (2004), p. 5076–5084 (cf. p. 87).

[219] Sandro Cosconati, Luciana Marinelli, Francesco Saverio Di Leva et al. „Protein flexibility in virtual screening : the BACE-1 case study“. In : *Journal of chemical information and modeling* 52.10 (2012), p. 2697–2704 (cf. p. 87).

[220] M. Krone, B. Kozlíková, N. Lindow et al. „Visual Analysis of Biomolecular Cavities : State of the Art“. In : *Comput. Graph. Forum* 35.3 (juin 2016), p. 527–551 (cf. p. 87, 120).

[221] Tiago Simões, Daniel Lopes, Sérgio Dias et al. „Geometric detection algorithms for cavities on protein surfaces in molecular graphics : a survey“. In : *Computer Graphics Forum*. T. 36. 8. Wiley Online Library. 2017, p. 643–683 (cf. p. 87).

[222] Buyong Ma, Maxim Shatsky, Haim J Wolfson et Ruth Nussinov. „Multiple diverse ligands binding at a single protein site : A matter of pre-existing populations“. In : *Protein science* 11.2 (2002), p. 184–197 (cf. p. 87).

[223] Christoph Globisch, Ilza K Pajeva et Michael Wiese. „Identification of Putative Binding Sites of P-glycoprotein Based on its Homology Model“. In : *ChemMedChem* 3.2 (2008), p. 280–295 (cf. p. 87).

[224] Carlos A Fuzo et Léo Degrève. „New pockets in dengue virus 2 surface identified by molecular dynamics simulation.“ In : *J. Mol. Model.* 19.3 (mar. 2013), p. 1369–77 (cf. p. 87, 109).

[225] Ron Diskin, David Engelberg et Oded Livnah. „A novel lipid binding site formed by the MAP kinase insert in p38α“. In : *Journal of molecular biology* 375.1 (2008), p. 70–79 (cf. p. 87).

[226] Norbert Lindow, Daniel Baum, Ana-Nicoleta Bondar et Hans-Christian Hege. „Exploring cavity dynamics in biomolecular systems“. In : *BMC bioinformatics* 14.19 (2013), S5 (cf. p. 88, 107).

[227] Nathan Desdouits. „Concepts et méthodes d'analyse numérique de la dynamique des cavités au sein des protéines et applications à l'élaboration de stratégies novatrices d'inhibition“. In : (2015). https://www.theses.fr/2015PA066250, https://tel.archives-ouvertes.fr/tel-01316546/document (cf. p. 96, 99, 167).

[228] Sandra W Cowan-Jacob, Gabriele Fendrich, Andreas Floersheimer et al. „Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia“. In : *Acta Crystallographica Section D : Biological Crystallography* 63.1 (2007), p. 80–93 (cf. p. 97).

[229] Yorgo Modis, Steven Ogata, David Clements et Stephen C Harrison. „A ligand-binding pocket in the dengue virus envelope glycoprotein.“ In : *Proc. Natl. Acad. Sci. U. S. A.* 100.12 (juin 2003), p. 6986–91 (cf. p. 97, 109).

[230] Chester L Drum, Shui-Zhong Yan, Joel Bard et al. „Structural basis for the activation of anthrax adenylyl cyclase exotoxin by calmodulin.“ In : *Nature* 415.6870 (jan. 2002), p. 396–402 (cf. p. 97, 109).

[231] Byungkook Lee et Frederic M. Richards. „The interpretation of protein structures : Estimation of static accessibility“. In : *Journal of Molecular Biology* 55.3 (1971), p. 379–400 (cf. p. 97).

[232] Michael L Connolly. „Solvent-accessible surfaces of proteins and nucleic acids“. In : *Science* 221.4612 (1983), p. 709–713 (cf. p. 97).

[233] A_ Bondi. „van der Waals volumes and radii“. In : *The Journal of physical chemistry* 68.3 (1964), p. 441–451 (cf. p. 97).

[234] Linda G Shapiro et G Linda. „stockman, George C“. In : *Computer Vision, Prentice hall. ISBN 0-13-030796-3* (2002) (cf. p. 98, 200).

[235] Eric Jones, Travis Oliphant, Pearu Peterson et al. *SciPy : Open source scientific tools for Python*. [Online ; accessed <today>] ; http://www.scipy.org/. 2001– (cf. p. 99).

[236] F. Pedregosa, G. Varoquaux, A. Gramfort et al. „Scikit-learn : Machine Learning in Python“. In : *Journal of Machine Learning Research* 12 (2011), p. 2825–2830 (cf. p. 99).

[237] Bjornar Larsen et Chinatsu Aone. „Fast and effective text mining using linear-time document clustering“. In : *Proc. fifth ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '99* (1999), p. 16–22 (cf. p. 102).

[238] R Tyrrell Rockafellar et Roger J-B Wets. *Variational analysis*. T. 317. Springer Science & Business Media, 2009 (cf. p. 103).

[239] S J Hubbard et P Argos. „Cavities and packing at protein interfaces.“ In : *Protein science : a publication of the Protein Society* 3.12 (déc. 1994), p. 2194–206 (cf. p. 105).

[240] Simon J. Hubbard, Karl-Heinz Gross et Patrick Argos. „Intramolecular cavities in globular proteins“. In : "*Protein Engineering, Design and Selection*" 7.5 (mai 1994), p. 613–626 (cf. p. 105).

[241] Shrihari Sonavane et Pinak Chakrabarti. „Cavities and Atomic Packing in Protein Structures and Interfaces“. In : *PLoS Computational Biology* 4.9 (sept. 2008). Sous la dir. de Cyrus Chothia, e1000188 (cf. p. 105).

[242] Charu C. Aggarwal, Alexander Hinneburg et Daniel A. Keim. „On the Surprising Behavior of Distance Metrics in High Dimensional Space“. In : *Springer Berlin Heidelberg* (2001), p. 420–434 (cf. p. 108).

[243] T. Schindler. „Structural Mechanism for STI-571 Inhibition of Abelson Tyrosine Kinase“. en. In : *Science* 289.5486 (sept. 2000), p. 1938–1942 (cf. p. 109).

[244] Jianming Zhang, Francisco J Adrián, Wolfgang Jahnke et al. „Targeting Bcr-Abl by combining allosteric with ATP-binding-site inhibitors.“ In : *Nature* 463.7280 (jan. 2010), p. 501–6 (cf. p. 109).

[245] Ragothaman Yennamalli, Naidu Subbarao, Thorsten Kampmann et al. „Identification of novel target sites and an inhibitor of the dengue virus E protein.“ In : *J. Comput. Aided. Mol. Des.* 23.6 (juin 2009), p. 333–41 (cf. p. 109).

[246] Alice L Bodnar, Luz A Cortes-Burgos, Karen K Cook et al. „Discovery and structure-activity relationship of quinuclidine benzamides as agonists of α7 nicotinic acetylcholine receptors“. In : *Journal of medicinal chemistry* 48.4 (2005), p. 905–908 (cf. p. 132).

[247] M Hajos, RS Hurst, WE Hoffmann et al. „The selective α7 nicotinic acetylcholine receptor agonist PNU-282987 [N-[(3R)-1-azabicyclo [2.2. 2] oct-3-yl]-4-chlorobenzamide hydrochloride] enhances GABAergic synaptic activity in brain slices and restores auditory gating deficits in anesthetized rats“. In : *Journal of Pharmacology and Experimental Therapeutics* 312.3 (2005), p. 1213–1222 (cf. p. 132).

[248] CA Briggs, MR Schrimpf, DJ Anderson et al. „α7 nicotinic acetylcholine receptor agonist properties of tilorone and related tricyclic analogues“. In : *British journal of pharmacology* 153.5 (2008), p. 1054–1061 (cf. p. 132).

[249] William R Kem. „The brain α7 nicotinic receptor may be an important therapeutic target for the treatment of Alzheimer's disease : studies with DMXBA (GTS-21)“. In : *Behavioural brain research* 113.1-2 (2000), p. 169–181 (cf. p. 132).

[250] JM Ward, VB Cockcroft, GG Lunt, FS Smillie et S Wonnacott. „Methyllycaconitine : a selective probe for neuronal α-bungarotoxin binding sites“. In : *FEBS letters* 270.1-2 (1990), p. 45–48 (cf. p. 132).

[251] MICHAEL J Marks, JERRY A Stitzel, ELENA Romm, JEANNE M Wehner et ALLAN C Collins. „Nicotinic binding sites in rat and mouse brain : comparison of acetylcholine, nicotine, and alpha-bungarotoxin.“ In : *Molecular pharmacology* 30.5 (1986), p. 427–436 (cf. p. 132).

[252] Hugo R Arias, Avraham Rosenberg, Katarzyna M Targowska-Duda et al. „Tricyclic antidepressants and mecamylamine bind to different sites in the human $\alpha4\beta2$ nicotinic receptor ion channel“. In : *The international journal of biochemistry & cell biology* 42.6 (2010), p. 1007–1018 (cf. p. 132).

[253] JasKiran K Gill-Thind, Persis Dhankher, Jarryl M D'Oyley, Tom D Sheppard et Neil S Millar. „Structurally similar allosteric modulators of $\alpha7$ nicotinic acetylcholine receptors exhibit five distinct pharmacological effects“. In : *Journal of Biological Chemistry* (2014), jbc–M114 (cf. p. 132, 159).

[254] Anna Chatzidaki et Neil S Millar. „Allosteric modulation of nicotinic acetylcholine receptors“. In : *Biochemical pharmacology* 97.4 (2015), p. 408–417 (cf. p. 132).

[255] Emilie Pihan, Lionel Colliandre, Jean-François Guichou et Dominique Douguet. „e-Drug3D : 3D structure collections dedicated to drug repurposing and fragment-based drug design“. In : *Bioinformatics* 28.11 (2012), p. 1540–1541 (cf. p. 133).

[256] Dustin K Williams, Can Peng, Matthew R Kimbrell et Roger L Papke. „Intrinsically low open probability of $\alpha7$ nicotinic acetylcholine receptors can be overcome by positive allosteric modulation and serum factors leading to the generation of excitotoxic currents at physiological temperatures“. In : *Molecular pharmacology* 82.4 (2012), p. 746–759 (cf. p. 133).

[257] María Guerra-Álvarez, Ana J Moreno-Ortega, Elisa Navarro et al. „Positive allosteric modulation of alpha-7 nicotinic receptors promotes cell death by inducing Ca2+ release from the endoplasmic reticulum“. In : *Journal of neurochemistry* 133.3 (2015), p. 309–319 (cf. p. 133).

[258] Matthias Rarey, Bernd Kramer, Thomas Lengauer et Gerhard Klebe. „A fast flexible docking method using an incremental construction algorithm“. In : *Journal of molecular biology* 261.3 (1996), p. 470–489 (cf. p. 133, 143).

[259] Dennis A Dougherty. „Cys-loop neuroreceptors : structure to the rescue ?“ In : *Chemical reviews* 108.5 (2008), p. 1642–1653 (cf. p. 135).

[260] Jean-Luc Galzi, Daniel Bertrand, Anne Devillers-Thiéry et al. „Functional significance of aromatic amino acids from three peptide loops of the $\alpha7$ neuronal nicotinic receptor site investigated by site-directed mutagenesis“. In : *FEBS letters* 294.3 (1991), p. 198–202 (cf. p. 135).

[261] Wenge Zhong, Justin P Gallivan, Yinong Zhang et al. „From ab initio quantum mechanics to molecular neurobiology : a cation–π binding site in the nicotinic receptor“. In : *Proceedings of the National Academy of Sciences* 95.21 (1998), p. 12088–12093 (cf. p. 135).

[262] Pierre-Jean Corringer, Nicolas Le Novère et Jean-Pierre Changeux. „Nicotinic receptors at the amino acid level“. In : *Annual review of pharmacology and toxicology* 40.1 (2000), p. 431–458 (cf. p. 135).

[263] Jennifer C Ma et Dennis A Dougherty. „The cation- π interaction“. In : *Chemical reviews* 97.5 (1997), p. 1303–1324 (cf. p. 135).

[264] Sandro Mecozzi, Anthony P West et Dennis A Dougherty. „Cation- π interactions in simple aromatics : electrostatics provide a predictive tool“. In : *Journal of the American Chemical Society* 118.9 (1996), p. 2307–2308 (cf. p. 135).

[265] Dennis A Dougherty. „The cation- π interaction“. In : *Accounts of chemical research* 46.4 (2012), p. 885–893 (cf. p. 135).

[266] Jie Liu et Renxiao Wang. „Classification of current scoring functions“. In : *Journal of chemical information and modeling* 55.3 (2015), p. 475–482 (cf. p. 135).

[267] Roger Sayle. „1st-class SMARTS patterns“. In : *EuroMUG 97*. 1997 (cf. p. 136).

[268] Mark M Levandoski et Sivaramakrishna Koganti. „Allosteric Modulation of Neuronal Nicotinic Acetylcholine Receptors“. In : *Allosterism in Drug Discovery*. 2016, p. 334–359 (cf. p. 137).

[269] Raja Dey et Lin Chen. „In search of allosteric modulators of α7-nAChR by solvent density guided virtual screening“. In : *Journal of Biomolecular Structure and Dynamics* 28.5 (2011), p. 695–715 (cf. p. 142, 156).

[270] Daniel Bertrand, Sonia Bertrand, Steven Cassar et al. „Positive allosteric modulation of the α7 nicotinic acetylcholine receptor : ligand interactions with distinct binding sites and evidence for a prominent role of the M2-M3 segment“. In : *Molecular pharmacology* 74.5 (2008), p. 1407–1416 (cf. p. 142, 156).

[271] Xiang-Qun Hu et David M Lovinger. „The L293 residue in transmembrane domain 2 of the 5-HT3A receptor is a molecular determinant of allosteric modulation by 5-hydroxyindole“. In : *Neuropharmacology* 54.8 (2008), p. 1153–1165 (cf. p. 142).

[272] Clark A Briggs, Jens Halvard Grønlien, Peter Curzon et al. „Role of channel activation in cognitive enhancement mediated by α7 nicotinic acetylcholine receptors“. In : *British journal of pharmacology* 158.6 (2009), p. 1486–1494 (cf. p. 142).

[273] Toby Collins, Gareth T Young et Neil S Millar. „Competitive binding at a nicotinic receptor transmembrane site of two α7-selective positive allosteric modulators with differing effects on agonist-evoked desensitization“. In : *Neuropharmacology* 61.8 (2011), p. 1306–1313 (cf. p. 142, 156, 159, 160).

[274] Christof H Schwab. „Conformations and 3D pharmacophore searching“. In : *Drug Discovery Today : Technologies* 7.4 (2010), e245–e253 (cf. p. 142).

[275] *Marvin - ChemAxon - Software Solutions and Services for Chemistry*. `https://http://www.chemaxon.com`. Marvin was used for drawing, displaying and characterizing chemical structures, substructures and reactions (cf. p. 142).

[276] Noel M O'Boyle, Michael Banck, Craig A James et al. „Open Babel : An open chemical toolbox“. In : *Journal of cheminformatics* 3.1 (2011), p. 33 (cf. p. 142).

[277] Joos Kiener. „Molecule database framework : a framework for creating database applications with chemical structure search capability“. In : *Journal of cheminformatics* 5.1 (2013), p. 48 (cf. p. 142).

[278] Eric F Pettersen, Thomas D Goddard, Conrad C Huang et al. „UCSF Chimera—a visualization system for exploratory research and analysis“. In : *Journal of computational chemistry* 25.13 (2004), p. 1605–1612 (cf. p. 142).

[279] Garrett M Morris, Ruth Huey, William Lindstrom et al. „AutoDock4 and AutoDockTools4 : Automated docking with selective receptor flexibility“. In : *Journal of computational chemistry* 30.16 (2009), p. 2785–2791 (cf. p. 142–144).

[280] Oleg Trott et Arthur J Olson. „AutoDock Vina : improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading“. In : *Journal of computational chemistry* 31.2 (2010), p. 455–461 (cf. p. 143, 144).

[281] William J Allen, Trent E Balius, Sudipto Mukherjee et al. „DOCK 6 : impact of new features and current docking performance“. In : *Journal of computational chemistry* 36.15 (2015), p. 1132–1156 (cf. p. 143).

[282] P Therese Lang, Scott R Brozell, Sudipto Mukherjee et al. „DOCK 6 : Combining techniques to model RNA–small molecule complexes“. In : *Rna* (2009) (cf. p. 143).

[283] Nadine Schneider, Gudrun Lange, Sally Hindle, Robert Klein et Matthias Rarey. „A consistent description of HYdrogen bond and DEhydration energies in protein–ligand complexes : methods behind the HYDE scoring function“. In : *Journal of computer-aided molecular design* 27.1 (2013), p. 15–29 (cf. p. 144).

[284] Maciej Wójcikowski, Pedro J Ballester et Pawel Siedlecki. „Performance of machine-learning scoring functions in structure-based virtual screening“. In : *Scientific Reports* 7 (2017), p. 46710 (cf. p. 144).

[285] *RF-Score-VS - Random forest based protein-ligand scoring function for Virtual Screening*. `https://github.com/oddt/rfscorevs` (cf. p. 144).

[286] Yang Cao et Lei Li. „Improved protein–ligand binding affinity prediction by using a curvature-dependent surface-area model“. In : *Bioinformatics* 30.12 (2014), p. 1674–1680 (cf. p. 144).

[287] David Ryan Koes, Matthew P Baumgartner et Carlos J Camacho. „Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise“. In : *Journal of chemical information and modeling* 53.8 (2013), p. 1893–1904 (cf. p. 144).

[288] Rodrigo Quiroga et Marcos A Villarreal. „Vinardo : A scoring function based on autodock vina improves scoring, docking, and virtual screening“. In : *PloS one* 11.5 (2016), e0155183 (cf. p. 144).

[289] *Scoring and Minimization with AutoDock Vina*. `https://sourceforge.net/projects/smina/files/` (cf. p. 144).

[290] Gianni Chessari, Christopher A Hunter, Caroline MR Low et al. „An evaluation of force-field treatments of aromatic interactions“. In : *Chemistry–A European Journal* 8.13 (2002), p. 2860–2867 (cf. p. 144).

[291] T Pymol. *The PyMOL molecular graphics system*. 2010 (cf. p. 145).

[292] Vasyl Bondarenko, David D Mowrey, Tommy S Tillman et al. „NMR structures of the human α7 nAChR transmembrane domain and associated anesthetic binding sites“. In : *Biochimica Et Biophysica Acta (BBA)-Biomembranes* 1838.5 (2014), p. 1389–1395 (cf. p. 156).

[293] JB Corrie et Steven M Sine. „Stoichiometry for drug potentiation of a pentameric ion channel“. In : *Proceedings of the National Academy of Sciences* (2013), p. 201301909 (cf. p. 156).

[294] Toby Collins et Neil S Millar. „Nicotinic acetylcholine receptor transmembrane mutations convert ivermectin from a positive to a negative allosteric modulator“. In : *Molecular pharmacology* (2010), mol–110 (cf. p. 156, 160).

[295] Jean-Luc Galzi, Sonia Bertrand, Pierre-Jean Corringer, Jean-Pierre Changeux et Daniel Bertrand. „Identification of calcium binding sites that regulate potentiation of a neuronal nicotinic acetylcholine receptor.“ In : *The EMBO Journal* 15.21 (1996), p. 5824–5832 (cf. p. 158).

[296] Mieke Nys, Eveline Wijckmans, Ana Farinha et al. „Allosteric binding site in a Cys-loop receptor ligand-binding domain unveiled in the crystal structure of ELIC in complex with chlorpromazine“. In : *Proceedings of the National Academy of Sciences* 113.43 (2016), E6696–E6703 (cf. p. 158).

[297] Ryoko M Krause, Bruno Buisson, Sonia Bertrand et al. „Ivermectin : a positive allosteric effector of the α7 neuronal nicotinic acetylcholine receptor“. In : *Molecular pharmacology* 53.2 (1998), p. 283–294 (cf. p. 159).

[298] Farah Deba, Hamed I Ali, Abisola Tairu et al. „LY2087101 and dFBr share transmembrane binding sites in the (α4) 3 (β2) 2 Nicotinic Acetylcholine Receptor“. In : *Scientific reports* 8.1 (2018), p. 1249 (cf. p. 160).

[299] *CDP6 Components Report for SGA1 M13-M24, PDF format*. https://sos-ch-dk-2.exo.io/public-website-production/filer_public/51/84/5184516b-5fba-4a23-ac57-9dbab8f0769b/d275_d125_d57_sga1_m24_accepted_180709.pdf (cf. p. 168).

[300] Arnold Meijster, Jos BTM Roerdink et Wim H Hesselink. „A general algorithm for computing distance transforms in linear time“. In : *Mathematical Morphology and its applications to image and signal processing*. Springer, 2002, p. 331–340 (cf. p. 200).

[301] Stephen M Omohundro. *Five balltree construction algorithms*. International Computer Science Institute Berkeley, 1989 (cf. p. 200).

[302] Leonardo Dagum et Ramesh Menon. „OpenMP : an industry standard API for shared-memory programming“. In : *Computational Science & Engineering, IEEE* 5.1 (1998). http://www.openmp.org, p. 46–55 (cf. p. 200).

[303] OpenMP Architecture Review Board. „OpenMP Application Program Interface Version 3.0“. In : (mai 2008). http://www.openmp.org/mp-documents/spec30.pdf (cf. p. 200).

[304] Michiel JL de Hoon, Seiya Imoto, John Nolan et Satoru Miyano. „Open source clustering software“. In : *Bioinformatics* 20.9 (2004). http://bonsai.hgc.jp/$\sim$mdehoon/software/cluster/, p. 1453–1454 (cf. p. 200).

VII

Annexes

A Modélisation comparative d'états conformationnels multiples de protéine

```
> Neuronal acetylcholine receptor subunit alpha-7 troncated
QRKLYKELVKNYNPLERPVANDSQPLTVYFSLSLLQIMDVDEKNQVLTTNIWLQMSWTDH
YLQWNVSEYPGVKTVRFPDGQIWKPDILLYNSADERFDATFHTNVLVNSSGHCQYLPPGI
FKSSCYIDVRWFPFDVQHCKLKFGSWSYGGWSLDLQMQEADISGYIPNGEWDLVGIPGKR
SERFYECCKEPYPDVTFTVTMRRRTLYYGLNLLIPCVLISALALLVFLLPADSGEKISLG
ITVLLSLTVFMLLVAEIMPATSDSVPLIAQYFASTMIIVGLSVVVTVIVLQYHHHDPDGG
MKRCVVDRLCLMAFSVFTIICTIGILMSAP
```

Figure VII.1 **Séquence de la sous-unité $\alpha 7$ du récepteur modélisé.**

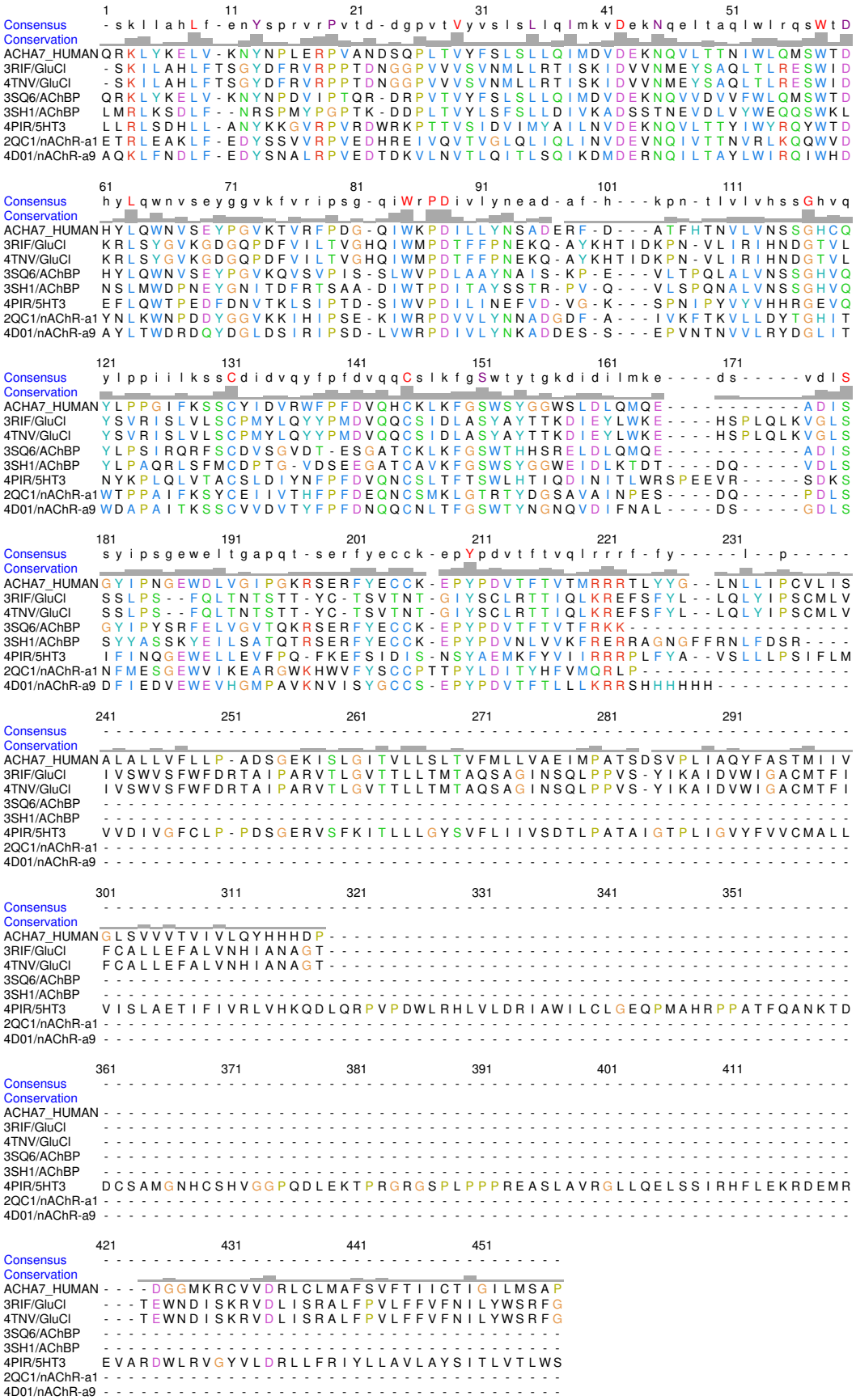


Figure VII.2 Alignement multiple des séquences *templates* avec la séquence $\alpha 7$ cible.

B

Calcul de chemins de transition : couplage POE/SoS

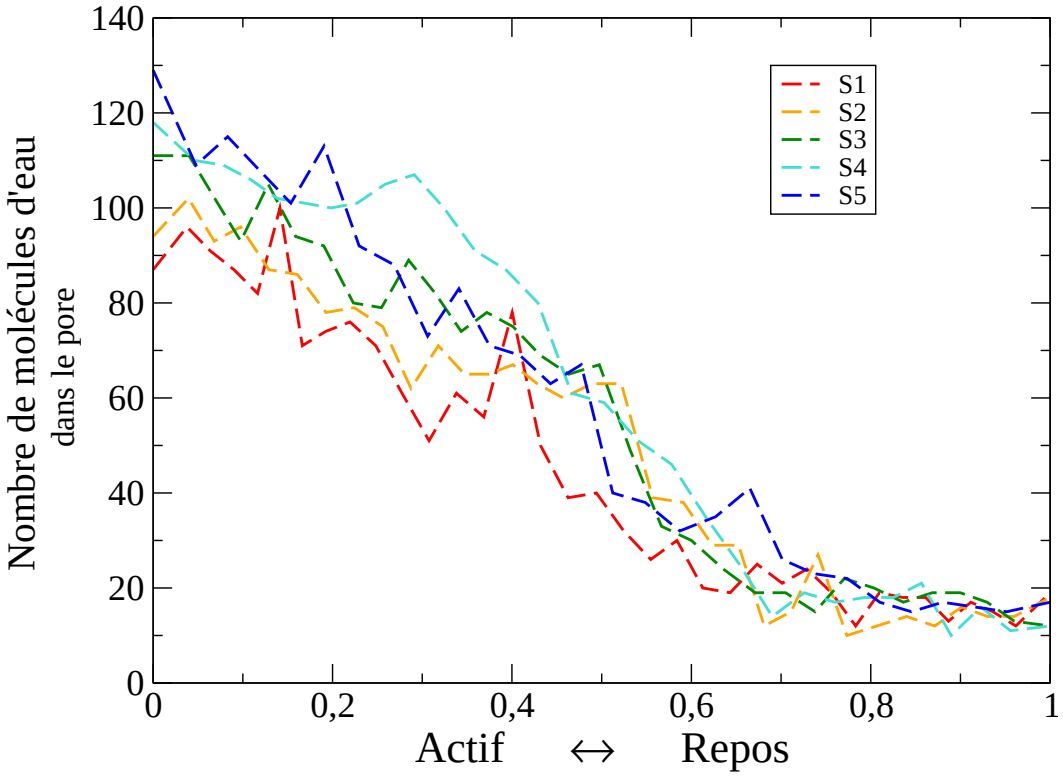


Figure VII.3 Hydratation du canal pour les dernières itérations de String of Swarms. L'hydratation du canal est calculée en comptant les molécules d'eau présentes dans le pore du récepteur. La région considérée pour le comptage correspond à un cylindre de 10 Å de rayon et de 25 Å de longue, centré au milieu des 5 hélices transmembranaires M2 du récepteur ((x,y,z)=(0 ;0 ;-7,5)).

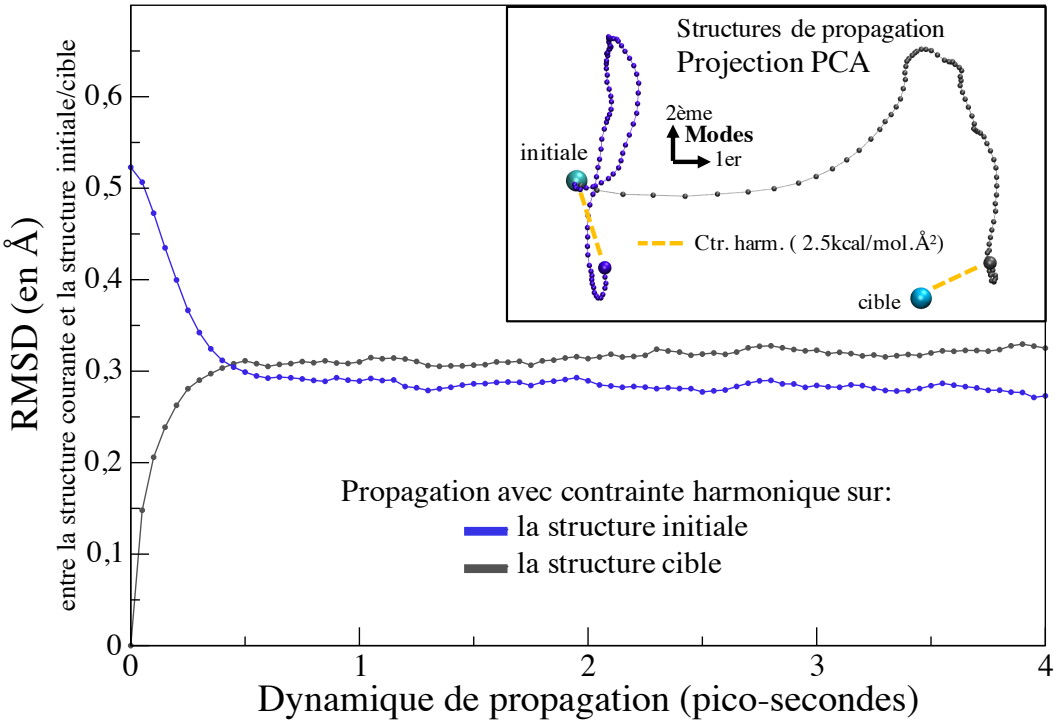


Figure VII.4 Accessibilité de la structure cible lors des trajectoires de propagation. Suivi de la distance (RMSD) à la bille cible lors de la trajectoire de propagation, avec contrainte harmonique sur la cible et sur l'hyperplan bissecteur. Pour comparaison, la bille cible est ensuite remplacée par la bille initiale lors d'une seconde propagation. La projection de l'ensemble des structures sauvegardées sur les deux premiers modes d'une ACP (sur les mêmes structures) permet de visualiser l'évolution des propagations vers leurs cibles respectives. Dès la première picoseconde de simulation, la structure converge très proche de sa cible ($\approx$ 0,3 Å) et ne peut s'en rapprocher plus près malgré les 3 ps de simulation suivant. Cette distance minimale de 0,3 Å est incompressible du fait des mouvements aléatoires qui agitent la dynamique moléculaire thermalisée.

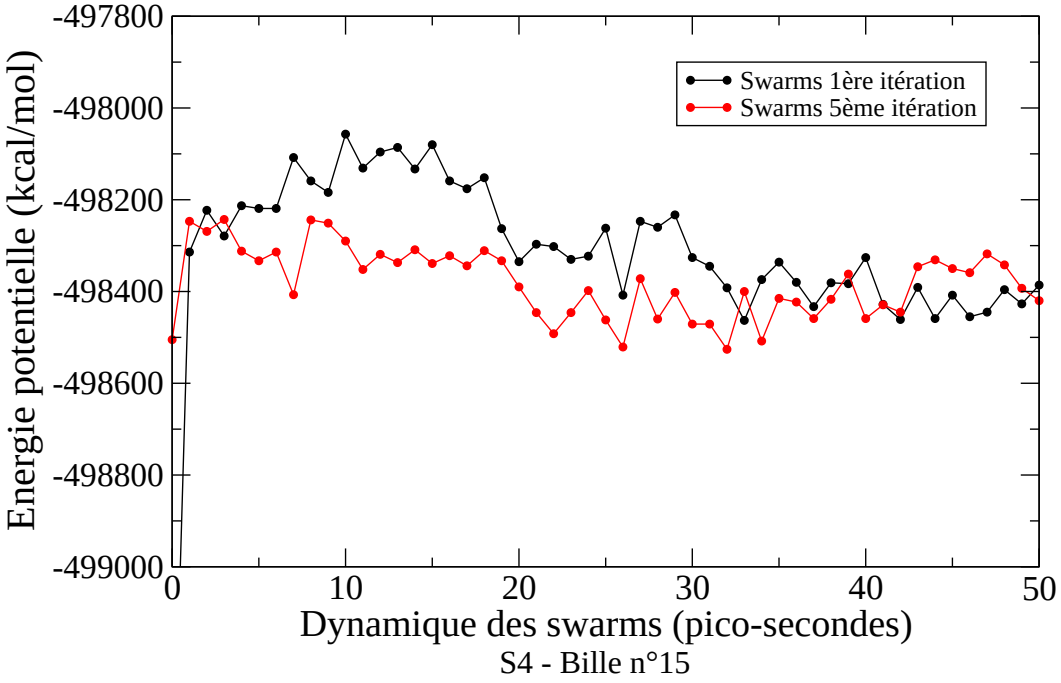


Figure VII.5 Détermination de la longueur des trajectoires de Swarms. Moyenne de l'énergie potentielle des 32 trajectoires non contraintes de Swarms mesurée toutes les pico-secondes (Série S4, bille 15, pour la 1re et la 5e itération de SoS). Les dynamiques de Swarms ont été rallongées à 50 ps pour évaluer un temps suffisant à la relaxation des billes. Les trajectoires atteignent un plateau énergétique entre 20 et 30 ps. En pratique, cette durée a été placée à 20 ps. En début de trajectoire la courbe rouge (5e itération SoS) apparaît être plus stable énergétiquement que la courbe noire (1re itération SoS, sortie de POE), ce qui en accord avec une diffusion de la bille 15 dans un minimum du paysage d'énergie libre.

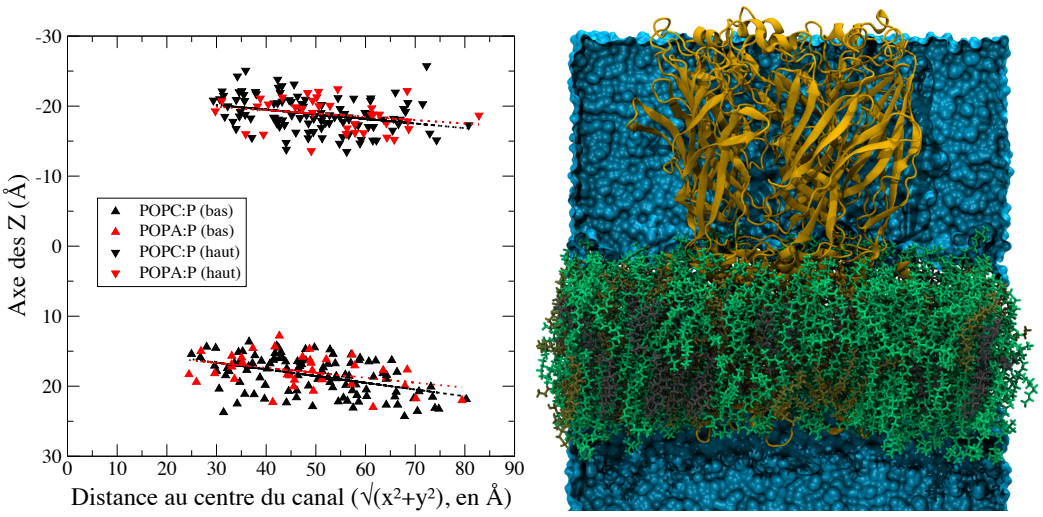


Figure VII.6 Mesure de la stabilité des lipides de la membrane. La hauteur (axe des z) des atomes de phosphore des têtes de lipides en fonction de leur distance au centre du canal ionique ((x,y)=(0,0)). Cette mesure permet de surveiller que la bicouche lipidique ne se décompose pas au cours des simulations.

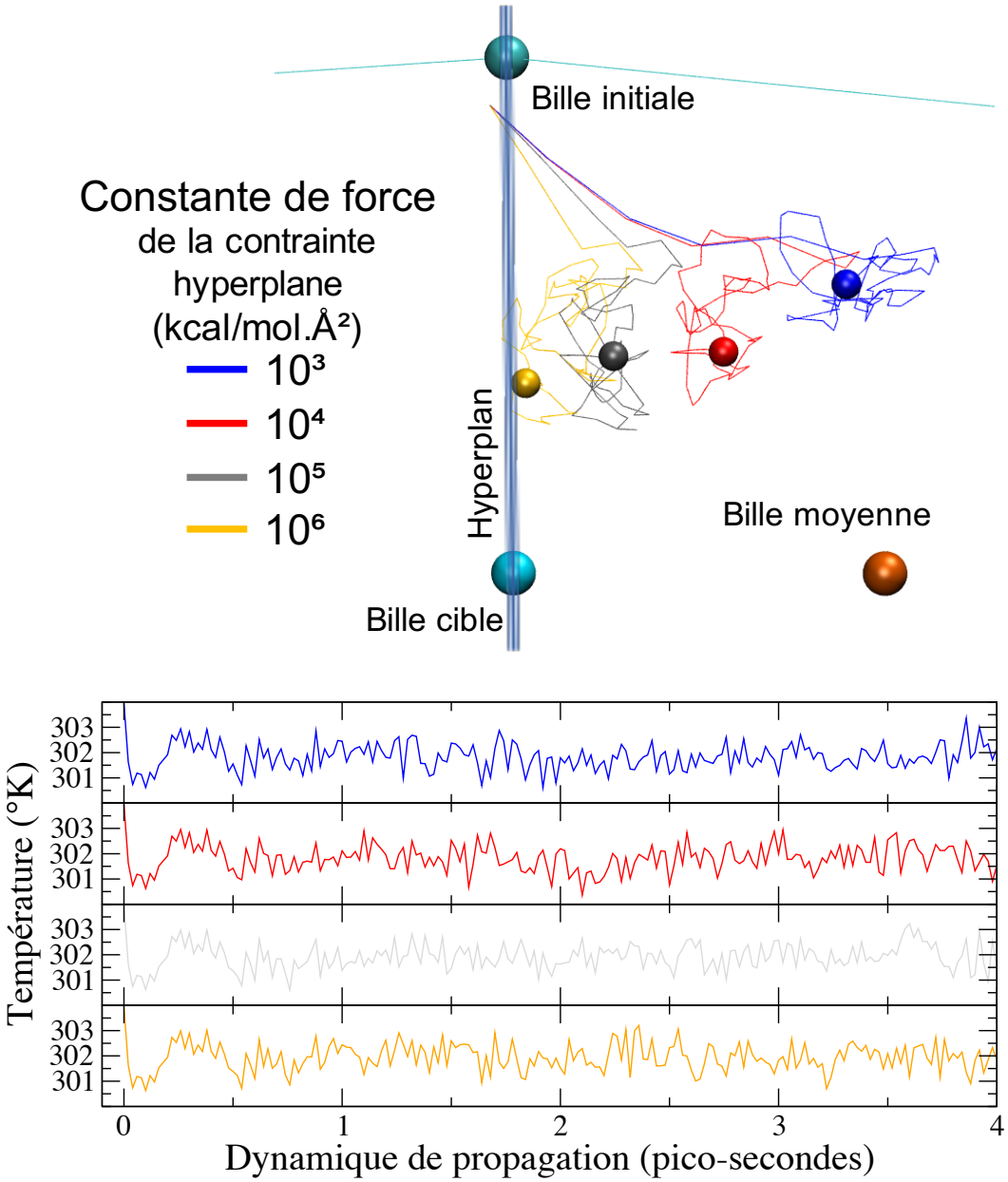
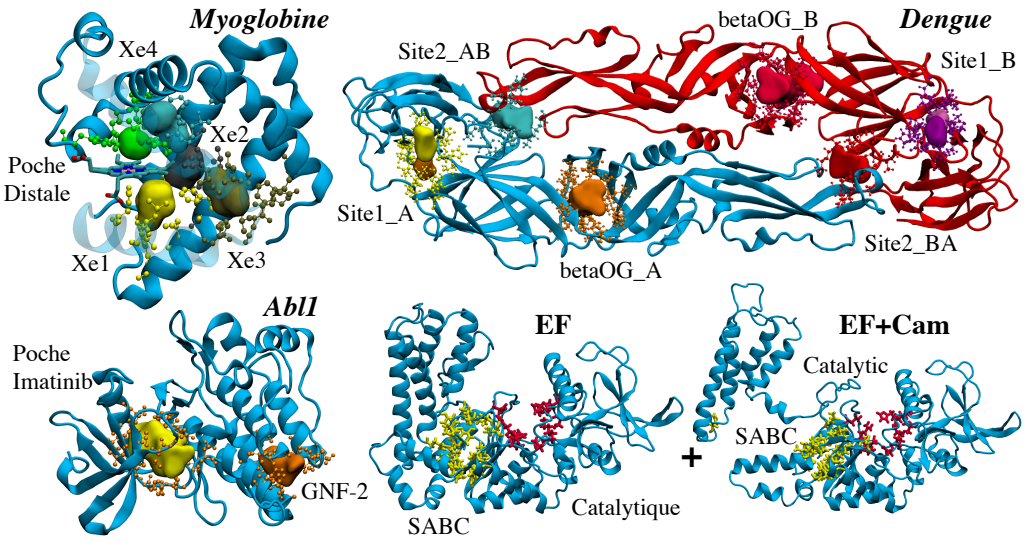


Figure VII.7 Paramétrisation de la contrainte hyperplane. Plusieurs constantes de force ont été testées pour évaluer sa capacité à contraindre les structures de propagation dans l'Hyperplan bissecteur. En haut, la projection de l'ensemble des structures de trajectoires de propagation, sur une analyse en composantes principales calculée sur le triplet de structures (bille initiale, bille cible, bille moyenne). L'hyperplan (de dimension 26 294) est directement visualisable, dans le plan orthogonal à la page et passant par la bille initiale et la bille cible. Ainsi sur la figure, l'éloignement des billes colorées de la ligne représentant l'hyperplan est directement relié au relâchement de la contrainte hyperplane. En pratique, nous avons choisi la contrainte la plus forte ($10^6\ kcal/mol.\text{Å}^2$), en vérifiant qu'elle ne déstabilisait pas la dynamique de propagation. Les 4 dynamiques de propagations ont été lancées avec la même graine aléatoire, et un coefficient de friction nulle sur les atomes lourds.

C Suivi des cavités dans les trajectoires de protéines et détection de sites

C.1 *mkgridXf* : Détails techniques

mkgridXf est un utilitaire écrit en C (standard C99) et utilisable en ligne de commande dans un terminal UNIX. La détection de cavités a été optimisée en utilisant des algorithmes spécialisés dans le traitement d'images (*Distance Transform*, codé à partir de la référence [300], *Two-pass Connecting Component Labelling* [234]) et dans le partitionnement spatial (*Ball Tree* [301]). *mkgridXf* profite de la parallélisation triviale du calcul des cavités et des empreintes pour chacune des conformations de la trajectoire, et utilise à profit la librarie open source OpenMP [302, 303]. Lors de longues trajectoires ($> 10^4$ conformations), le nombre de cavités total peut facilement dépasser les millions, pour plus de $\sim 10^{10}$ points de grilles. Comme la détection de cavités est rapide, *mkgridXf* n'a pas besoin de garder les points de grille en mémoire. Les points de grilles sont recalculés lors du réassignement et sauvegardés sur disque le cas échéant. Par conséquent, l'empreinte mémoire de *mkgridXf* est proportionnelle au nombre de processus utilisés (et nom au nombre de conformation de la trajectoire). La librarie open source *The C clustering Library* [304] est utilisée pour le partitionnement hiérarchique des empreintes (partitionnement UPGMA). *mkgridXf* supporte les définitions d'empreintes par atomes et par résidus, les empreintes réelles (et local-booléennes en interne pour le découpage des poches) ainsi que la distance cosine. L'ensemble des résultats peuvent être exportés au format binaire HDF5, librement disponible sur internet. Un programme compagnon, *mkread* permet l'extraction des cavités transverses au format texte, ainsi que la série temporelle de leur volume, et des empreintes moyennes correspondantes. Une sortie des cavités au format texte est également disponible. Un *plugin* VMD a été écrit en langage C pour permettre la visualisation dynamique des cavités au côté des structures de la trajectoire moléculaire.



Reference Site	Residues
Myoglobin	
Distal Pocket	L29 L32 F43 H64 V68 I107
Xe1	L89 A90 H93 L104 F138 I142 Y146
Xe2	L72 L104 I107 S108 L135 F138 R139
Xe3	W7 I75 L76 K79 G80 H82 A134 L135 L137 F138
Xe4	G25 I28 L29 G65 V68 L69 L72 I107 I111
Dengue	
Site1	P39 T40 H144 S145 G146 E147 Y178 L294 K295 T353 V354 N355 P356 I357 T359 S363 V365
Site2	chaîneA : D98 R99 G100 G102 N103 K246
	chaîneB : R2 I4 G5 I6 G152 D154
βOG	T48 E49 A50 P53 K128 V130 L135 G190 L191 F193 L198 Q200 A205 L207 T268 I270 Q271 L277 F279 T280 G281
Abl1	
Imatinib	L247 G248 Y252 V255 A268 V269 K270 E285 M289 V298 I312 T314 E315 F316 M317 G320 N321 L369 A379 D380 F381
GNF-2	A336 L339 L340 A343 L428 I431 A432 Y434 E461 G462 C463 P464 V467 F492
EF	
SABC	A496 P499 I538 E539 P542 S544 S550 W552 Q553 T579 Q581 L625 Y626 Y627 N629 N709
Catalytic	R329 K346 H351 S354 K372 D491 D493 H577 N583

Figure VII.8 Définition des sites de référence. Les poches des sites de référence sont illustrées pour les structures cristallographiques suivantes : myoglobine, PDB : 1J52 ; dengue, PDB : 1OKE ; abl1, PDB : 2HZI ; EF, PDB : 1K8T ; EF+Cam, PDB : 1K93 (calmoduline non montrée). Le volume des cavités est indiqué est représenté par des surfaces solides pour les 3 premiers systèmes. La structure de la protéine E de la dengue est symétrique et chacun des sites est présent en deux locus. Leur poche de référence, aussi symétrique, n'est listée qu'une seule fois. Le Site 2 de la dengue se trouve à l'interface entre les chaînes A et B.

Site (Cav./Ref.)	$\#\overline{C}^{50\%}$	$\#\overline{A}^{50\%}$	$\#\left(\overline{C}^{50\%} \setminus \overline{A}^{50\%}\right)$
Poche D. (Myo)	342	172	0
Xe1 (Myo)	186	177	0
Xe2 (Myo)	103	2	0
Xe3 (Myo)	486	404	0
Xe4 (Myo)	170	106	0
site1 (DENV_A)	1236	178	0
site1 (DENV_B)	1145	176	0
site2 (DENV_AB)	2209	3109	948
site2 (DENV_BA)	1939	2292	386
βOG (DENV_A)	1140	1056	0
βOG (DENV_B)	1380	1245	0
Imatinib (Abl1)	1761	1305	0
GNF-2 (Abl1)	406	0	0
SABC (EF)	323	131	25
Catalytique (EF)	2345	3942	1787

Tableau VII.1 Comparaison entre la cavité moyenne des cavités de référence $\overline{C}^{50\%}$ et la cavité moyenne des cavités instantanées assignées à la poche de référence pour chaque site $\overline{A}^{50\%}$. L'ensemble des cavités considérées pour A ne prend en compte que la plus petite cavité instantanée lorsque plusieurs sont assignées à une même conformation. Les cavités moyennes sont tronquées à 50 % d'apparition. Chaque colonne rapporte le nombre de voxels constituant la cavité ainsi produite. La colonne la plus à droite indique le nombre de voxels présents dans $\overline{C}^{50\%}$ mais pas dans $\overline{A}^{50\%}$. On remarque ainsi que la majorité des sites considérés comme étant *difficiles* à suivre correspondent au cas où les cavités instantanées "débordent" du volume délimité par le site de référence.

Système	# Confs.	$\overline{RMSD}$ (Å)	$\overline{\#cavités}$/conf. (std/min/max)	Vol. moyen par cavité (Å^3)
Myoglobin	1 000	1,29	11,9 (2,4/6/20)	37
Abl1	1 000	3,32	27,3 (4,3/12/40)	86
EF-CaM	2 000	*	36,3 (4,8/22/54)	104
Dengue	1 000	2,63	58,4 (5,3/42/83)	75
		Volume moyen par conformation (Å^3)		
Système	Cavités	Env. protéine[a]	(%)	
Myoglobin	438	21 576	(2,0)	
Abl1	2 354	37 725	(6,2)	
EF-CaM	3 788	63 903	(5,9)	
Dengue	4 365	107 622	(4,1)	
		Domaine de définition (Å^3)		
Système	Cavités	Env. protéine[a]	(%)	(% cd/mpe[b])
Myoglobin	7 803	43 589	(17,9)	(36,2)
Abl1	33 462	98 981	(33,8)	(88,7)
EF-CaM	56 872	188 303	(30,2)	(89,0)
Dengue	51 630	186 858	(27,6)	(48,0)

Tableau VII.2 Analyse des cavités détectées dans les trajectoires. Volumes, et domaines de définitions sont calculés comme explicités en Matériels et Méthodes 2/p.91). *La trajectoire EF inclut 1 000 conformations de la forme inactive (2,07 Å RMSD), 1 000 conformations de la forme active (2,3 Å RMSD) pour une variation moyenne totale de 6,17 Å RMSD. [a] Enveloppe de la protéine, [b] volume du domaine de définition des cavités en pourcentage du volume moyen de l'enveloppe de la protéine.

D Recherche de sites effecteurs du récepteur nicotinique

D.1 FlexX+cπ : implémentation

```
### Fichier: contact.dat
@subgraph 1 6 ammonium
  SMARTS [!$([NH3X4+][CH2X4][CH2X4][CH2X4][CH2X4][CHX4]);N+,N+R]
data
iact    1       -       -       -       cation          sphere_cat_N
end

@subgraph 1 1 phenyl_center
  atom  1 C2ar
  ...
  bond  5 4 un 6 un
data
  iact  1       4       2       -       phenyl_center   phenyl_center
  iact  1       4       2       -       phenyl_pi       phenyl_pi
end

@subgraph 2 2 aro_ring5_trplike
  atom  1 N2ar
  ...
  bond  6 3  un
data
  iact  3       1       4       -       phenyl_center   phenyl_center
  iact  3       1       4       -       phenyl_pi       phenyl_pi
end

### Fichier: geometry.dat
@geometry sphere_cat_N
radius  4.5
delta  1.6 1.6
sphere
surface_mode 0
energy  0
distance_scaling 0.5 1.3
```

```
@geometry phenyl_pi
center  0.5     0.0     0.0     # center of sphere coincides with ring
center
radius 4.5
delta  1 1
cone    b 0 0 1 0 35
cone    b 0 0 -1 0 35
surface_mode 0
energy  -3.138   # en kJ/mol soit -0.75 kcal/mol
distance_scaling 0.5 1.3

### Fichier: contype.dat
@contact_types
...
phenyl_pi | cation
...

@placement_level
...
cation                  3
phenyl_pi               3
...

@min_vdw_radius
...
cation                  2.5
phenyl_pi               2.5
...
```

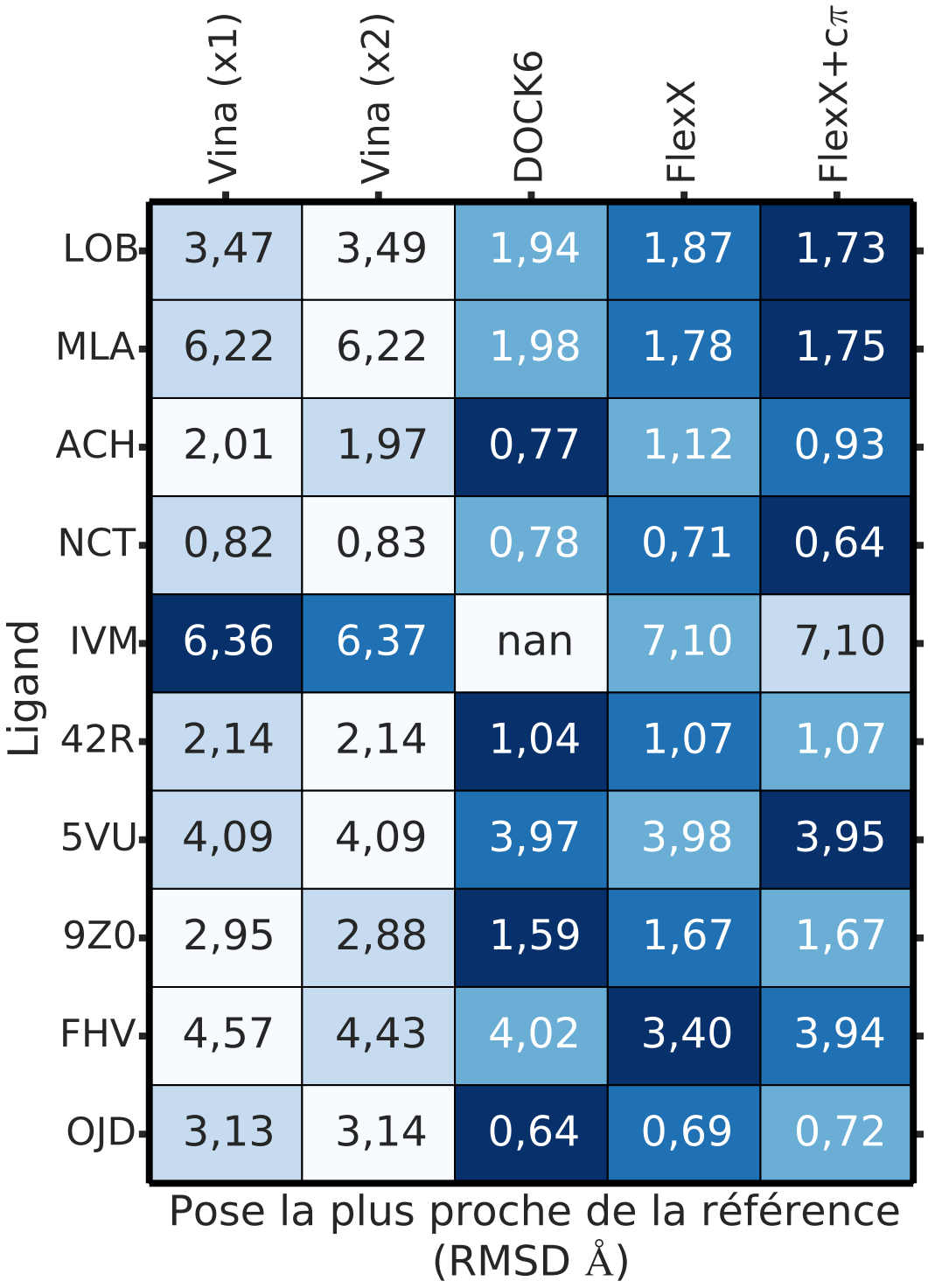


Figure VII.9 **Distance entre la molécule de référence et sa pose la plus proche suivant l'outil de *docking* utilisé.**

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at http://cleanthesis.der-ric.de/.

Declaration

Je soussigné Damien Monet certifie que le manuscrit présenté en vue de la soutenance est le fruit d'un travail original et que toutes les sources utilisées ont été clairement indiquées.

Je certifie, de surcroît, que je n'ai ni copié ni utilisé des idées ou des formulations tirées d'un ouvrage, article ou mémoire, en version imprimée ou électronique, sans mentionner précisément leur origine et que les citations sont expressément signalées entre guillemets (ou par une autre disposition graphique sans ambiguïté).

Conformément à la loi, le non-respect de ces dispositions me rend passible de poursuites devant la commission disciplinaire et les tribunaux de la République française pour plagiat universitaire.

Fait à Paris le 6 décembre 2018

Damien Monet