Université de Montréal

Simulations numériques de la dynamique des protéines: translation de ligands, flexibilité et dynamique des boucles.

par
Jean-François St-Pierre

Département de biochimie
Faculté de médecine

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Philosophiæ Doctor (Ph.D.)
en Bio-informatique

Mars, 2012

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée:

Simulations numériques de la dynamique des protéines: translation de ligands, flexibilité et dynamique des boucles.

présentée par:

Jean-François St-Pierre

a été évaluée par un jury composé des personnes suivantes:

François Major,	président-rapporteur
Mousseau Normand,	directeur de recherche
Radu Iftimie,	membre du jury
Pierre Tuffery,	examinateur externe
Nicolas Lartillot,	représentant du doyen de la FES

Thèse acceptée le: . . . . . .

RÉSUMÉ

La flexibilité est une caractéristique intrinsèque des protéines qui doivent, dès le moment de leur synthèse, passer d'un état de chaîne linéaire à un état de structure tridimensionnelle repliée et enzymatiquement active. Certaines protéines restent flexibles une fois repliées et subissent des changements de conformation de grande amplitude lors de leur cycle enzymatique. D'autres contiennent des segments si flexibles que leur structure ne peut être résolue par des méthodes expérimentales. Dans cette thèse, nous présentons notre application de méthodes *in silico* d'analyse de la flexibilité des protéines :

- À l'aide des méthodes de dynamique moléculaire dirigée et d'échantillonnage parapluie, nous avons caractérisé les trajectoires de liaison de l'inhibiteur Z-pro-prolinal à la protéine Prolyl oligopeptidase et identifié la trajectoire la plus probable. Nos simulations ont aussi identifié un mode probable de recrutement des ligands utilisant une boucle flexible de 19 acides aminés à l'interface des deux domaines de la protéine.

- En utilisant les méthodes de dynamique moléculaire traditionnelle et dirigée, nous avons examiné la stabilité de la protéine SAV1866 dans sa forme fermée insérée dans une membrane lipidique et étudié un des modes d'ouverture possibles par la séparation de ses domaines liant le nucléotide.

- Nous avons adapté au problème de la prédiction de la structure des longues boucles flexibles la méthode d'activation et de relaxation ART-nouveau précédemment utilisée dans l'étude du repliement et de l'agrégation de protéines. Appliqué au repliement de boucles de 8 à 20 acides aminés, la méthode démontre une dépendance quadratique du temps d'exécution sur la longueur des boucles, rendant possible l'étude de boucles encore plus longues.

Mots clés: Dynamique moléculaire, Échantillonnage parapluie, Activation-relaxation technique, SAV1866, Prolyl oligopeptidase, Prédiction de structure de boucles.

ABSTRACT

Flexibility is an intrinsic characteristic of proteins who from the moment of synthesis into a linear chain of amino acids, have to adopt an enzymatically active tridimensionnel structure. Some proteins stay flexible once folded and display large amplitude conformational changes during their enzymatic cycles. Others contain parts that are so flexible that their structure can't be resolved using experimental methods. In this thesis, we present our application of *in silico* methods to the study of protein flexibility.

- Using steered molecular dynamics and umbrella sampling, we characterized the binding trajectories of the Z-pro-prolinal inhibiter to the Prolyl oligopeptidase protein and we identified the most probable trajectory. Our simulations also found a possible ligand recrutement mechanism that involves a 19 amino acids flexible loop at the interface of the two domains of the protein.

- Using traditional and steered molecular dynamics, we examined the stability of the SAV1866 protein in its closed conformation in a lipid membrane and we studied one of its proposed opening modes by separating its nucleotide binding domains.

- We also adapted the activation-relaxation technique ART-nouveau which was previously used to study protein folding and aggregation to the problem of structure prediction of large flexible loops. When tested on loops of 8 to 20 amino acids, the method demonstrate a quadratic execution time dependance on the loop length, which makes it possible to use the method on even larger loops.

Keywords: Molecular dynamics, Umbrella sampling, Activation-relaxation technique, SAV1866, Prolyl oligopeptidase, Loop structure prediction

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES

a.a.	Acide aminé
ABC	ATP binding cassette
ADP	Adénosine diphosphate
AMPPNP	Adénosine 5'-(β,γ-imido)triphosphate
ATP	Adénosine triphosphate
C-α	Carbone alpha
DC	Domaine catalytique
DLN	Domaine liant les nucléotides
DPβ	Domaine en propulseur-β
DM	Dynamique moléculaire
DMD	Dynamique moléculaire dirigée
DMER	Dynamique moléculaire avec échange de répliques
DMV	Dynamique moléculaire visée
DTM	Domaine transmembranaire
EP	Échantillonnage parapluie
MC	Monte-Carlo
MDR	Protéine résistante à de multiples drogues
MM	Mécanique moléculaire
PFM	Potentiel de force moyen
PLA	Protéine périplasmique liant l'allocrite

POP	Prolyl oligopeptidase
RMN	Résonance magnétique nucléaire
RMSD	Racine carrée de la moyenne des distances au carré
WHAM	Méthode d'analyse des histogrammes par assignation de poids
ZPP	Z-pro-prolinal

Nikol, mé milované manželce.

REMERCIEMENTS

Tout d'abord et avant tout, j'aimerais remercier mon directeur de recherche, Normand Mousseau, pour sa patience, ses conseils, sa supervision, son aide financière, son optimisme face à des résultats souvent décevants et son entêtement à me faire voir sous un autre oeil ces résultats. Je resterai à toujours reconnaissant pour toutes les opportunités qui m'ont été offertes sous sa supervision et pour son aide à rendre à terme ce grand projet. J'aimerais aussi remercier Alex Bunker pour sa supervision en Finlande et pour m'avoir offert ce stage d'un an qui fut aussi bien formateur sur le plan professionnel que personnel. Je remercie mes autres principaux collaborateurs, Tomasz Róg et Mikko Karttunen, pour leur contributions et leur conseils.

Ensuite, j'aimerais remercier les membres de mon entourage à l'université et à la maison. Je remercie tous les membres de ma famille et particulièrement mon épouse Nikol pour avoir cru en moi tout au long de ces années et pour leur support moral lors des moments difficiles. Je garderai un bon souvenir de mes collègues de laboratoire Lilianne Dupuis, Jean-François Joly, Sébastien Côté, Jessica Nasica, Rozita Laghaei, Saïd Bouzakraoui, Peter Brommer et bien d'autres qui sont passés par notre laboratoire et repartis et que j'espère recroiser au cours de nos carrières.

Il n'y a pas de science sans financement, et pour cette raison j'aimerais remercier tous les organismes subventionnaires qui m'ont donné les moyens d'effectuer mes recherches, soit le réseau européen Galenos m'ayant accordé la bourse de stage Marie Curie, le programme stratégique des Instituts de recherche en santé du Canada pour la bourse d'excellence biT et le Fond de recherche du Québec en nature et technologies pour leur bourse de recherche, mais aussi les réseaux Calcul Québec, Finnish IT Centre for Science (CSC) et SharcNet pour leur octrois de ressources informatiques.

Enfin, j'aimerais remercier ma correctrice dédiée, Myrian Grondin, grâce à qui tous ces remerciements vous sont livrés sans faute ni erreur.

INTRODUCTION

La flexibilité structurale intrinsèque des protéines est une condition indispensable à leur fonctionnement. On n'a qu'à penser au fait que toute protéine commence son existence sous la forme d'une chaîne linéaire d'acides aminés (a.a.) qui a vite fait de se replier en une forme compacte et fonctionnelle. Une fois repliées, les protéines sont dotées d'éléments de structure secondaire, assemblés en structures tertiaires, puis parfois en structures quaternaires lorsque plusieurs protéines sont impliquées dans un complexe.

Les protéines repliées n'ont pas pour autant atteint une conformation statique. Lorsqu'on observe la structure cristalline des protéines par spectre de diffraction à rayon X, certaines zones sont visiblement plus agitées, dotées de facteurs de Debye-Waller, ou facteur B, plus élevés indiquant des régions plus affectées par l'agitation thermique [246]. D'autres régions fluctuent au point qu'elles ne peuvent être résolues et sont tout simplement manquantes dans l'image tridimensionnelle de la protéine. Ces désordres peuvent ne pas avoir d'impact sur les qualités enzymatiques d'une protéine, mais ils peuvent aussi servir de site de régulation pour des protéases ou pour les diverses modifications post-traductionnelles [36].

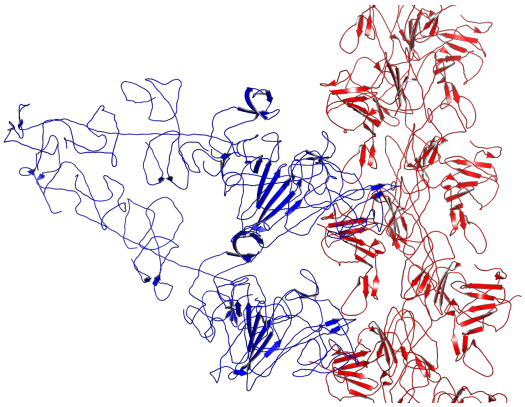

Figure 1 – Protéine de myosine (bleue) se déplaçant sur un filament d'actine (rouge) (PDB : 1M8Q) [29].

Des exemples de mouvement dans les protéines existent sur toutes les échelles de grandeur. Lors de la contraction musculaire, le myofilament protéinique de myosine se

déplace sur le myofilament d'actine (figure 1) dans un mouvement similaire à la marche humaine [183, 277]. Sur de plus petites échelles, il a été postulé que le co-transport de NH^3 et de son proton dans les transporteurs d'ammonium est réalisé par la simple rotation de la chaîne latérale de deux histidines, se passant le proton *à la chaîne* [139].

Plusieurs techniques expérimentales permettent d'étudier directement ou non les mouvements dans les protéines. En mutant certains a.a. pour des cystéines qui formeront entre elles des liaisons covalentes restreignant les changements de conformation, on peut réduire l'amplitude du mouvement interne d'une protéine et mesurer l'impact sur les capacités enzymatiques de celle-ci [2]. Aussi, en exposant la protéine à de l'eau lourde composée de D_2O, on peut marquer les parties internes de la protéine qui sont exposées au solvant pendant leur repliement [61]. Ou encore, par la technique de microscopie à force atomique, il est possible de retirer un substrat d'une protéine en attachant ce ligand à un levier qui est lentement éloigné de la protéine et ainsi calculer l'énergie nécessaire au dépliement [181] ou à la dissociation du ligand [169].

Les méthodes *in silico* ont le désavantage marqué de ne pouvoir atteindre les échelles de temps que couvrent les méthodes expérimentales. Puisque les simulations sur des protéines de taille moyenne atteignent rarement des temps de simulation dépassant quelques microsecondes [69], il est encore impossible à ce jour de modéliser un système sans biais dans lequel un enzyme recruterait un substrat, catalyserait la transformation de ce dernier en produits, puis relâcherait le ou les produits dans le solvant. Il serait encore plus difficile de simuler un système du genre suffisamment longtemps ou souvent pour pouvoir calculer des moyennes d'ensemble et dériver les constantes de réaction à l'aide de ces dernières.

Cependant, les méthodes *in silico* ont l'avantage de pouvoir tester des hypothèses difficiles à examiner expérimentalement. Lorsqu'on pense à des méthodes expérimentales tel que le transfert d'énergie par résonance de type Förster pour calculer des distances inter-atomiques dans des systèmes dynamiques [91] ou la formation d'éléments de structure secondaire par dichroïsme circulaire [109], des conclusions microscopiques sont portées en observant des propriétés macroscopiques. En se basant sur une description empirique des interactions atomiques, des méthodes *in silico* comme la dynamique

moléculaire (DM) ou l'échantillonnage Monte-Carlo (MC) permettent d'étudier finement des processus complexes tels que le repliement des petites protéines ou l'agrégation de peptides dans tous leurs états intermédiaires instables. De plus, puisque les méthodes *in silico* sont basées sur une description des lois physiques décrivant l'énergie d'un système, il est aussi possible de modifier ces lois pour forcer des changements de conformations difficiles à observer expérimentalement.

Cette thèse porte sur la flexibilité et les mouvements au sein des protéines. Plus précisément, nous nous intéressons méthodes *in silico* permettant d'échantillonner les mouvements de grande amplitude que certaines protéines subissent lors de leur cycle enzymatique. À l'aide de deux protéines modèles, nous avons évalué divers outils numériques découlant de la dynamique moléculaire dans leur performance à identifier la trajectoire de liaison d'un ligand à sa protéine cible et à échantillonner les changements de conformation de grande amplitude d'une protéine membranaire. Puis en utilisant une méthode d'échantillonnage de surface énergétique, nous avons développé une méthode de prédiction des structures en boucle des protéines basée sur le repliement protéinique qui est mieux adapté à la prédiction de longues boucles.

Spécifiquement, nous nous sommes penchés sur une protéine de grand intérêt pharmacologique, la protéine P-glycoprotéine (Pgp, ABCB1 ou MDR1), membre de la famille des transporteurs ABC, de l'anglais *ATP binding cassette transporteurs*. Cette protéine membranaire dont le rôle est l'expulsion de toxines vers l'extérieur de la cellule est impliquée dans une grande partie des cancers résistants à la chimiothérapie [58], d'où le nom de *Multidrug resistance protein* (MDR).

Jusqu'à récemment, aucune structure tridimensionnelle de Pgp n'était disponible bien que plusieurs structures de transporteurs ABC provenant de procaryotes avaient été obtenues par cristallographie aux rayons X, autant pour des importateurs de ligands qui n'ont pas d'équivalent eucaryote [94, 100, 161, 184, 201], que pour des exportateurs de ligands [44, 45, 264]. Les structures de la forme ouverte de MsbA ayant été rétractées [25], la seule structure d'un exportateur restante était celle de SAV1866 dans sa forme fermée [44]. C'est dans ce contexte que nous avons décidé de tenter de répondre à trois questions concernant Pgp :

1. Est-il possible d'obtenir la forme ouverte d'un transporteur ABC à partir de sa forme fermée et d'utiliser la forme ouverte pour bâtir un modèle par homologie de Pgp,

2. Quels sont les mécanismes structurels permettant de passer de la forme fermée à la forme ouverte d'un transporteur ABC, et

3. Quelle est la structure de la boucle de 60 a.a. reliant les deux moitiés de Pgp et qui est absente des autres structures publiées ?

Depuis la publication de la première structure de la protéine Pgp chez la souris [3] dans sa forme ouverte, nous avons une structure de haute résolution pouvant servir de cible dans les simulations visant l'ouverture de la structure fermée de SAV1866. Puisque la structure de la protéine Pgp publiée ne comporte pas la boucle de 60 a.a., les deux autres objectifs sont inchangés et pour les atteindre, des outils devaient être développés.

Pour caractériser la trajectoire d'ouverture de la structure transporteur ABC, nous avons choisi d'utiliser des techniques de simulation, telle que la dynamique moléculaire (DM) et la dynamique moléculaire dirigée (DMD) présentées au chapitre 1. Les résultats de ces simulations sur la protéine SAV1866 dans sa forme fermée vous sont présentés dans un premier article au chapitre 3.

Les méthodes de DMD, ou encore d'échantillonnage parapluie (EP), peuvent aussi être utilisées pour déterminer le profil d'énergie libre d'une transition et permettre ainsi d'assigner des probabilités à différentes trajectoires échantillonnées. Puisque ces méthodes n'ont été utilisées que sur des systèmes de petite taille, nous avons décidé de les tester sur un système de taille intermédiaire, soit le chemin de liaison du ligand z-prolinal à la protéine prolyl oligopeptidase présentée au chapitre 5. Les résultats de ces simulations ont fait l'objet d'un second article présenté au chapitre 7.

Puis finalement, le problème du repliement de la boucle de 60 a.a. nécessite le développement d'une nouvelle gamme de logiciels de prédiction des structures de boucle. Les méthodes existantes basées sur des bases de données de conformations de protéines ne donnent généralement pas de bons résultats pour de longues boucles alors que les

méthodes *ab initio* ont des exigences en temps de calcul qui augmentent exponentiellement en fonction de la longueur de la séquence. Nous avons donc adapté notre méthode d'échantillonnage de surface énergétique ART-nouveau [163, 176] au problème de prédiction des boucles. L'hypothèse derrière ce choix est que dans le cas d'une longue boucle, une méthode habilitée à traiter le problème du repliement des protéines tel que ART-nouveau peut être mieux adaptée qu'une méthode basée sur l'échantillonnage exhaustif de l'espace des conformations. La méthode modifiée est présentée au chapitre 9 et les résultats de nos comparaisons avec d'autres méthodes dans l'article présenté au chapitre 11.

Les méthodes de DMD, d'EP et de MC telle que Art-nouveau ont toutes leur application dans l'échantillonnage de trajectoires de basse énergie. Dans les cas de la DMD et de l'EP, nous sommes intéressés à trouver des trajectoires de basse énergie libre reliant deux conformations du système d'intérêt. De son côté, la méthode ART-nouveau qui a déjà été utilisée pour trouver des trajectoires de repliement de protéines [232] est ici utilisée pour trouver des conformations de basse énergie potentielle représentatives de la forme repliée des boucles.

CHAPITRE 1

DYNAMIQUE MOLÉCULAIRE DE SAV1866

Ce chapitre introduit la protéine qui a motivé ce projet, la P-glycoprotéine, et la protéine SAV1866 nous servant à étudier Pgp, ainsi que la famille des transporteurs ABC dont elles sont toutes deux membres. Enfin, nous décrivons les méthodes de simulations utilisées pour étudier les mouvements de grande amplitude dans SAV1866, soit la dynamique moléculaire et la dynamique moléculaire dirigée.

1.1 La famille des ABC transporteurs

Les transporteurs ABC sont exprimés dans tous les domaines du monde vivant [92] et sont impliqués dans l'importation ou l'exportation active de molécules que l'on nomme allocrites du fait qu'ils ne sont pas altérés par le processus de transport [279]. Certains représentants de la famille sont impliqués dans la régulation de la pression osmotique, le trafic du cholestérol et des lipides, le transport de nutriments, ainsi que la résistance cellulaire aux drogues et toxines [47].

Dans les procaryotes, ces protéines peuvent jouer le rôle d'importateur ou d'exportateur d'allocrites, mais seuls les exportateurs sont exprimés chez les eucaryotes. Chez l'homme, on connaît 48 protéines regroupées en 7 familles (ABCA à ABCG).

1.1.0.1 Intérêt pharmacologique

Au moins 11 et jusqu'à 19 maladies héréditaires sont causées par des dysfonctionnements de transporteurs ABC [255], la plus connue étant la fibrose kystique [74], mais aussi d'autres comme la dégénérescence maculaire de Stargardt [174], la maladie de Tangier [8] et l'adrénoleucodystrophie [15].

Alors que la majorité des transporteurs ABC sont impliqués dans le transport actif d'un seul allocrite, ceux spécialisés dans l'expulsion de composés xénobiotiques ont la capacité de transporter plusieurs types d'allocrites dont des médicaments. Au nombre

des drogues interagissant avec Pgp, on trouve des agents anti-cancer, des inhibiteurs de la protéase du VIH, des agents immunosuppresseurs, des agents antiarythmiques, des antiépileptiques et bien d'autres [226]. Fait surprenant, la taille des allocrites expulsés par Pgp est aussi disparate, allant de 293 Da pour l'odansetron à 1200 Da pour la cyclosporine [85].

Chez l'homme, la sur-expression d'une ou plusieurs des trois protéines Pgp, MRP1 et BCRP de la famille des transporteurs ABC est responsable de presque la totalité des cas de résistance à la chimiothérapie des tumeurs cancéreuses [114]. De plus, la forte expression de ces protéines dans les intestins et la barrière hémato-encéphalique rendent difficile l'absorption de certains médicaments, alors que leur expression dans les intestins, les reins et le foie accélère l'excrétion de ces mêmes médicaments. Il y a donc un intérêt à développer des inhibiteurs pour cette classe de protéines. Pour plus de détails, la revue de littérature faite par le *International Transporter Consortium* énumère la liste des transporteurs ABC d'intérêt émergeant [78].

1.1.0.2 Relation structure-fonction

Chaque transporteur ABC est constitué d'un ou deux domaines transmembranaires (DTM) et d'un ou deux domaines liant les nucléotides ATP (DLN). La structure quaternaire des transporteurs ABC qui ont été cristallisés peut être composée soit de deux protéines, chacune contenant un DTM suivi d'un DLN, soit d'une seule protéine contenant en alternance les domaines : DTM – DLN – DTM – DLN. Les transporteurs dotés de deux protéines peuvent être observés en agencement homodimères (SAV1866) ou hétérodimères (BtuCD,MalFKG_2), ou sur une seule protéine (Pgp).

Les DTM jouent le rôle de pompe à ligands. Habituellement composés de 6 hélices transmembranaires, les DTMs peuvent compter de 5 à 10 hélices par domaine (BtuC). Les DTMs s'assemblent en paire pour exposer une cavité qui est, en alternance, exposée au milieu extracellulaire dans la forme fermée de la protéine, ou encore exposée au cytoplasme et au feuillet cytoplasmique de la membrane cellulaire dans la forme ouverte de la protéine. Pour les importateurs tout comme les exportateurs, le DTM possède le ou les sites de liaison d'allocrites. Les hélices des DTM dans les structures qui ont

été cristallographiées démontrent un entrelacement des domaines. Pour Pgp, les hélices transmembranaires H4 et H5 du premier DTM traversent dans le deuxième DTM alors que les hélices H10 et H11 du deuxième DTM traversent vers le premier.

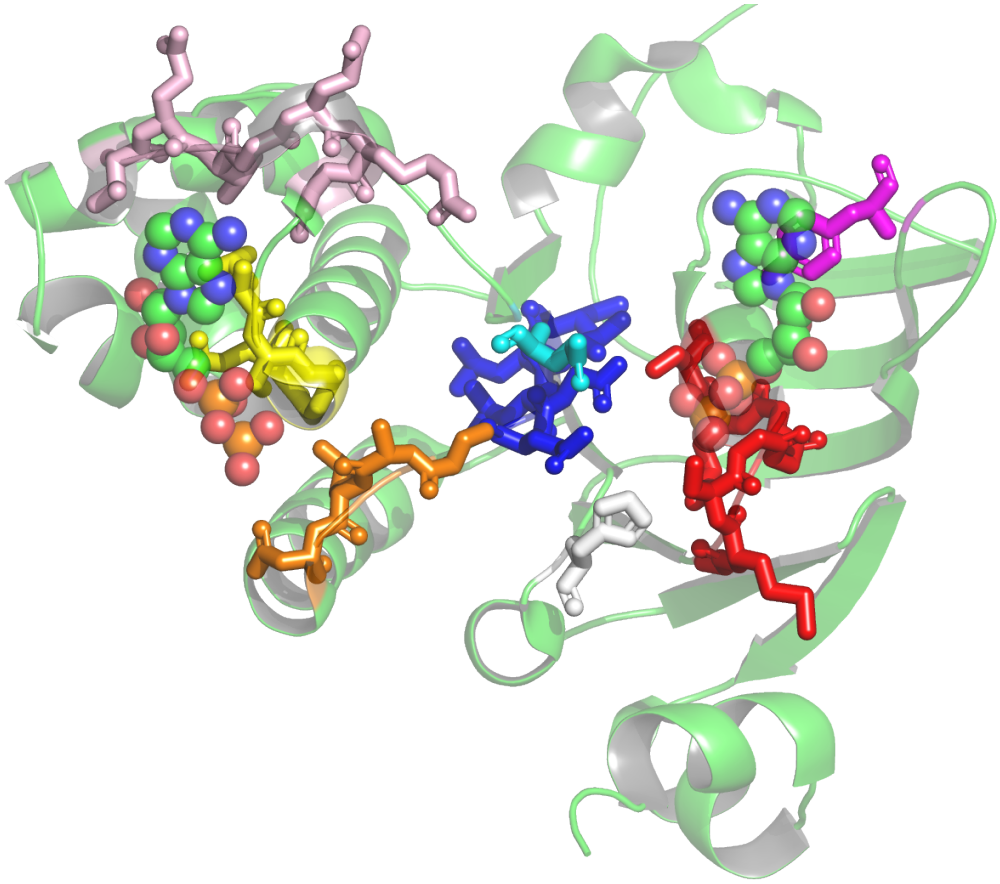

Figure 1.1 – Vue d'un domaine liant les nucléotides de SAV1866 en présence d'ADP (représentation bille). En représentation bâtonnets dans la section de droite, on retrouve les motifs Walker A (rouge) et B (bleu), les boucles Q (cyan) et H (blanc). Dans la section de gauche, on a le motif signature LSGGQ (jaune), la boucle D (orange) et la boucle X (rose).

Les domaines liant les nucléotides sont hautement conservés entre tous les transporteurs ABC. Leur structure peut être divisée en deux sections contenant les motifs de liaison à l'ATP. La première section contient le motif Walker A [259] de séquence *GxxGxGKS/T* (*x* étant n'importe quel a.a.), la boucle A contenant un a.a. aromatique interagissant avec l'adénine du nucléotide, le motif Walker B de séquence $\phi\phi\phi\phi DE$ (ϕ

étant un a.a. aliphatique), la boucle D de séquence *SALD*. La deuxième section contient le motif *LSGGQ* caractéristique des transporteurs ABC. Tous ces motifs se retrouvent sur un même plan formant le site d'interaction entre les DLNs. Lors de l'assemblage de deux DLNs, les 2 sections de chaque DLN interagissent avec la section de type opposée de l'autre DLN, créant effectivement deux sites de liaisons à l'ATP symétriques.

Le couplage entre les DTMs et les DLNs diffère entre les exportateurs et les importateurs. Deux hélices situées à l'extrémité cytosolique des DTM forment l'ensemble des contacts avec les DLN. Dans Pgp [3], l'hélice de couplage CH1 est située entre H2 et H3 du premier DTM et interagit avec le premier DLN, mais l'hélice de couplage CH2 située entre H5 et H6 se retrouve en contact avec le deuxième DLN dû à l'entrecroisement des deux DTMs.

1.1.1 Structures cristallographiques connues

Le premier système de transporteurs ABC à avoir été cristallisé est celui de l'importateur de vitamine B12 BtuCD [161] (PDB : 1L7V, 3.2 Å de résolution) présenté en Figure 1.2. Plus tard, le même groupe a réussi à obtenir la structure de la protéine liée à sa protéine périplasmique (PLA) BtuF liant la vitamine B12 [100] (PDB : 2QI9, 2.6 Å). Les deux structures sont en conformation ouverte avec les DLNs partiellement séparés. Les DTMs sont dotés de 10 hélices transmembranaires, mais la cavité centrale est de petite taille et ouverte vers l'extérieur. L'orientation de la structure ressemble à celle de la protéine moins volumineuse de l'importateur de molbidate $ModB_2C_2$ de *Archaeoglobus fulgidus* [94] (PDB : 2ONK 3.1 Å) et à la structure de HI1470/1 de *Haemophilus influenzae* [201] (PDB : 2NQ2, 2.4 Å). Cette dernière possède une cavité dans les DTM aussi petite que BtuCD, mais pointant vers le cytoplasme.

Un autre importateur, le transporteur du maltose $MalFGK_2$ d'*Escherichia coli* a été cristallisé en conjonction avec sa protéine périplasmique recrutant l'allocrite [184] (PDB : 2R6G, 2.8 Å). Tout comme BtuCD, la section transmembranaire du transporteur du maltose est un hétérodimère de DTMs qui n'ont pas de lien covalent avec les DLNs.

La première structure d'un exportateur nous vient aussi du groupe de Locher, celle de la protéine SAV1866 de *Staphylococcus aureus* dans sa forme fermée et comptant deux

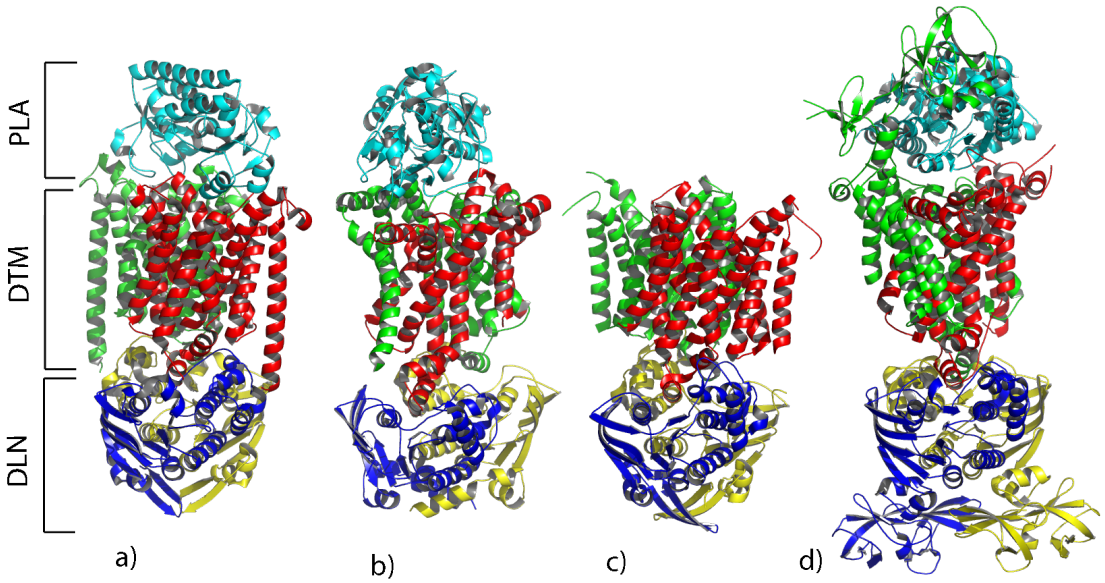

Figure 1.2 – Structures de 4 importateurs de type ABC : a) BtuCD (PDB : 2QIA), b) $ModB_2C_2$ (PDB : 2ONK), c) HI1470/1 (PDB : 2NQ2) et d) $MalFGK_2$ (PDB : 2R6G). Les couleurs vert et rouge sont associées aux domaines transmembranaires, le bleu et le jaune aux domaines liant le nucléotide et le cyan à la protéine périplasmique liant l'allocrite.

nucléotides ADP dans ses DLNs [44] (PDB : 2HYD,3.0 Å). Une structure de plus basse résolution est aussi disponible liée à l'AMPPNP , un inhibiteur similaire à l'ATP [45] (PDB : 2ONJ,3.4 Å). Dans cette forme fermée de l'homodimère, le site de liaison à l'allocrite est exposé au solvant extracellulaire dans une cavité en forme de V.

L'exportateur et flippase de lipide MsbA de *E. coli* a été cristallisé dans 3 formes différentes [264] : La forme ouverte sans nucléotide (apo) (PDB : 3B5W,5.3 Å), une forme fermée apo (PDB : 3B5X,5.5 Å), et une forme fermée avec le nucléotide AMPPNP (PDB : 3B60,3.7 Å). Malgré le fait que les formes apo de cette protéine soient de résolution assez faible, ce jeu de structure permet de visualiser une étape intermédiaire possible du mécanisme des exportateurs ABC : dans sa forme fermée apo, on peut voir que les deux DLNs sont décalés de façon à exposer au solvant les parties des DLNs contenant le motif signature *LSGGQ*. Aussi, la structure fermée liée à l'AMPPNP est en bon accord

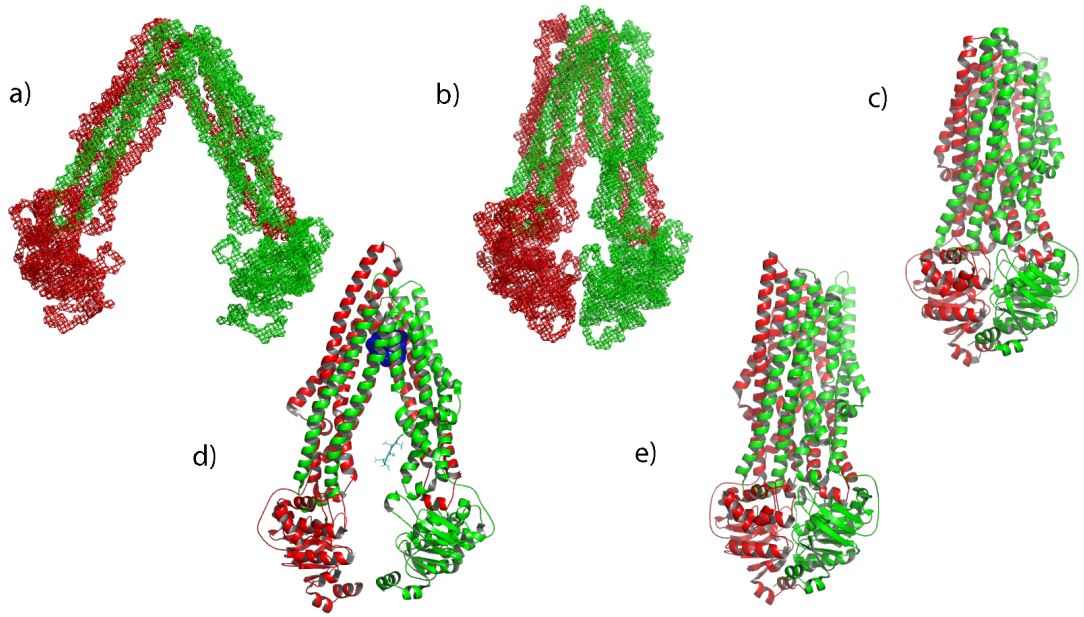

Figure 1.3 – Structures de 5 exportateurs de type ABC : MsbA dans la forme a) ouverte (PDB : 3B5W), b) fermée apo(PDB : 3B5X), c) fermée et liée à l'AMPPNP (PDB : 3B60), d) Pgp en forme ouverte lié à un inhibiteur (PDB : 3G60) et e) SAV1866 dans la forme fermée liant l'ADP. Les couleurs vert et rouge sont associées aux deux moities de la protéines, qu'elles soient monomèriques (Pgp) ou dimèriques (MsbA et SAV1866).

avec celle de SAV1866. Le lecteur devra porter attention au fait que la première structure de MsbA publiée par le groupe de Chang en 2001 a été rétractée [25], mais pas les publications y faisant référence.

P-glycoprotéine de *Mus musculus* est la toute dernière structure d'un transporteur ABC publiée par le groupe de Chang [3] (PDB : 3G5U,3.8 Å). Deux structures supplémentaires liées à un inhibiteur situé dans la région des DTMs et démontrant des modes de liaisons différents sont aussi disponibles (PDB : 3G60 et 3G61, 4.4 Å). Les structures ouvertes présentées sont en accord avec celle de MsbA ouverte avec des DLNs distants de 30 Å. La liaison des inhibiteurs n'a pas d'effet sur la séparation des deux moitiés de Pgp, ce qui semble indiquer que la forme ouverte est la forme liant les allocrites. Les auteurs suggèrent que l'extension de la séparation entre les deux moitiés de Pgp pourrait être modulée par la taille de l'allocrite expulsé.

La toute dernière structure à avoir été déposée à la base de données du RCSB est celle de la protéine ABCB10 dans sa forme ouverte et exprimée dans les mitochondries humaines (PDB : 2YL4). La structure n'a toujours pas d'article rattaché, mais avenant une publication acceptée, cette structure serait la toute première pour un transporteur ABC humain.

1.1.1.1 Cycle d'expulsion des ligands de Pgp

Le modèle le plus courant du cycle d'expulsion des transporteurs ABC est le modèle interrupteur [89, 156, 252]. Le cycle commence avec un transporteur en conformation ouverte apo tel qu'observé dans la structure cristalline de Pgp. La liaison d'un allocrite dans les DTMs altère l'affinité des DLNs qui recrutent deux ATPs. Une fois les ATPs liés à un DLN chaque, la protéine se referme dans un mouvement en pincette pour les DLNs et en épingle à linge pour les DTMs [28]. La conformation serait similaire à celle observée dans la structure cristalline de SAV1866. Le ligand, maintenant exposé au solvant ou au feuillet extracellulaire de la membrane, serait libéré. Il s'en suit une hydrolyse simultanée ou alternative des deux ATPs conférant l'énergie nécessaire à la protéine pour rétablir la conformation ouverte initiale.

Le modèle a été remis en question suite à l'observation de la distance séparant les DLNs dans des structures telles que celles de MsbA ouverte ou de Pgp. Il serait peu probable que la liaison de deux ATPs favorise l'attraction d'un domaine 20 Å plus loin mais pas l'ADP qui ne diffère que par une charge négative de moins [114]. Aussi, quelques expériences ont démontré que les sites de liaison à l'ATP sont faiblement exposés au solvant pendant le cycle d'hydrolyse [20] alors que l'inhibition d'un site actif semble exposer l'autre site au solvant [205, 287]. Le modèle à contact constant a donc été proposé [111, 112]. Celui-ci diffère par le fait que les DLNs restent en contact tout au long du cycle, mais seulement un site actif à la fois, et alternent de site actif au moment de l'hydrolyse, recrutant un ATP avant de libérer l'ADP.

1.1.2 Homologie de séquences entre Pgp les autres exportateurs ABC

Un alignement de séquences entre SAV1866 et Pgp humain est présenté en Figure 1.4, la séquence de SAV1866 est alignée à deux reprises sur les deux moitiés de Pgp. On y voit que les DLNs sont mieux conservés que les DTMs avec une l'identité de séquence de 50% et 20% respectivement. L'identité de séquence totale de SAV1866 et Pgp est de 29% alors qu'elle atteint les 31% entre MsbA et Pgp. Les raisons de notre choix de poursuivre nos simulations sur SAV1866 sont que la résolution des deux structures de SAV1866 est supérieure à celles de MsbA et qu'il a été démontré que certaines des toxines expulsées par SAV1866 le sont aussi par Pgp [256].

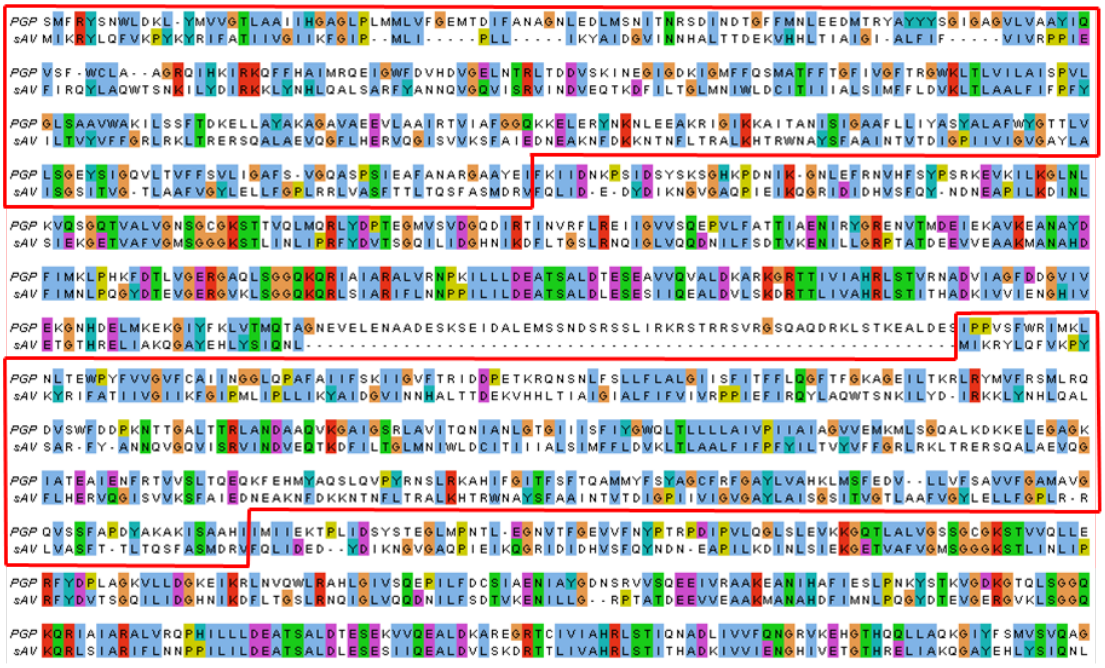

Figure 1.4 – Alignement de séquences entre Pgp et SAV1866. Les sections encadrées en rouge correspondent aux DTMs et le reste aux DLN. La séquence de SAV1866 est alignée deux fois pour permettre de comparer le dimère SAV1866 au monomère Pgp.

La structure de SAV1866 a déjà été utilisée pour générer un bon nombre de modèles par homologie de transporteurs ABC [49, 62, 84] dont certaines de Pgp [79, 143, 190].

1.1.2.1 Boucle manquante

L'alignement de séquences entre SAV1866 et Pgp en Figure 1.4 affiche une boucle de 60 a.a. manquante dans la séquence de SAV1866. Cette boucle est à l'origine du projet de développement de méthode présenté au chapitre 9. Elle relie l'extrémité C-terminale du premier DLN au N-terminale du second DTM de Pgp. Dans la structure de SAV1866 dans la forme fermée, ces deux extrémités sont distantes de 69 Å alors que dans celle de Pgp dans la forme ouverte, cette distance augmente proportionnellement avec l'écart entre les DNLs. Cependant, du fait que la structure de Pgp est ouverte, le chemin le plus court pour fermer cette boucle passe par le sillon entre les deux moitiés de la protéine. Un fragment de cette boucle d'environ 10 a.a. attaché au N-terminale du deuxième DTM est le seul vestige de la boucle visible dans la structure cristallographique de Pgp, mais ce dernier est inséré entre les deux TMD, une position impossible dans la structure fermée. La dynamique de la boucle lors de l'ouverture de Pgp est un mystère qui reste à expliquer. Puisqu'en théorie, l'ossature d'un peptide de 60 a.a. peut s'étirer à une longueur de 207 Å, l'ellipsoïde décrit par la boucle de Pgp a un volume 4.1×10^6 Å^3, ce qui serait suffisant pour que la boucle puisse interagir avec la membrane ou n'importe quelle partie cytosolique de la protéine. Une illustration d'une boucle de 60 a.a. artificiellement attachée à SAV1866 est présentée à titre d'exemple à la Figure 1.5

Des études ont tout de même été faites pour tenter de comprendre le rôle de cette boucle. À l'aide d'enzyme de clivage peptidique tel que la trypsine, Sato *et al.* ont comparé l'activité de transport de Pgp avec une variante sans la boucle et ont démontré que Pgp sans boucle était tout aussi stable thermiquement que Pgp normal [217]. Cependant, l'activité d'hydrolyse de l'ATP basale k_{basal} et l'activité d'hydrolyse en présence d'allocrites $k_{substrat}$ ont toutes deux augmenté après clivage de la boucle. Toutefois, le ratio $k_{substrat}/k_{basal}$ a diminué indiquant que sans la boucle le couplage entre la présence d'un allocrite dans les TMD et l'hydrolyse de l'ATP perd en efficacité.

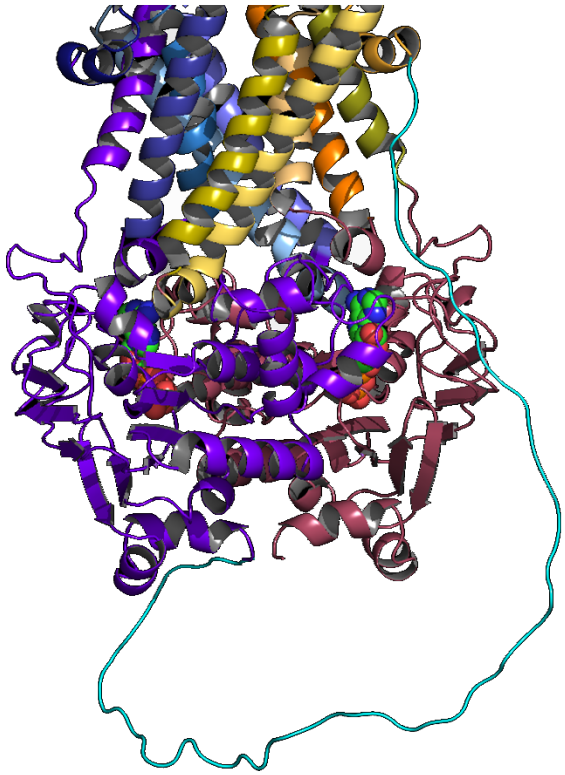

Figure 1.5 – Exemple d'une boucle de 60 a.a. dans une forme étendue (cyan) lié au C-terminale du premier DLN et au N-terminale du deuxième DTM de SAV1866.

1.2 Méthodes de simulation

L'objectif de cette partie du projet est de caractériser la dynamique de l'ouverture d'un transporteur ABC homologue de Pgp. Notre choix a donc basculé vers SAV1866 qui a des propriétés physiologiques de Pgp [256] et qui a été cristallisée avec une haute résolution à 3.0 Å. Pour ce faire, nous avons construit un système modèle comprenant une membrane lipidique et 55,100 molécules d'eau pour un total cd 235,500 atomes. Deux copies du système sont créées, l'une avec l'ADP co-cristallisé, l'autre apo. Les systèmes sont ensuite équilibrés et étudiés par une simulation de dynamique moléculaire de 100 ns, puis soumis à plusieurs simulations de dynamique moléculaire dirigée pour étudier l'ouverture de la protéine. Les méthodes et composantes de cette étude vous sont maintenant présentées.

1.2.1 Dynamique moléculaire

La dynamique moléculaire est une méthode d'échantillonnage de l'espace qui repose sur l'intégration numérique itérative des équations du mouvement de Newton définissant les vitesses V et les positions atomiques X d'un système au temps $t_i + \Delta t$ en fonction de la position, de la vitesse et du gradient de l'énergie potentielle ∇E (i.e. : le vecteur de forces F). Au temps t_i, l'accélération est donnée par :

$$a(t_i) = -\nabla E(X(t_i))/M(X) = F(X(t_i))/M(X), \tag{1.1}$$

où $M(X)$ est la matrice diagonale des masses atomiques. Le vecteur d'accélération du système $a(t_i)$ est ensuite utilisé pour calculer le vecteur vitesse V au temps $t_i + \Delta t$:

$$V(t_i + \Delta t) = V(t_i) + a(t_i)\Delta t, \tag{1.2}$$

puis finalement le vecteur des postitions X au temps $t_i + \Delta t$:

$$X(t_i + \Delta t) = X(t_i) + V(t_i)\Delta t + a(t_i)\Delta t^2, \tag{1.3}$$

où Δt est le pas d'intégration.

1.2.1.1 Gromacs

Le groupe d'Alex Bunker en Finlande utilise depuis quelques années avec succès la suite de logiciel de simulation Gromacs [153] pour l'étude de la protéine COMT [21] et Prolyl oligopeptidase [128, 129]. Gromacs est distribué avec plusieurs outils d'analyse et est généralement regardé comme étant le plus rapide des logiciels de dynamique moléculaire lorsque faiblement parallélisé, et les performances en exécution parallèle ont fortement augmenté depuis la dernière version [87].

Au coeur du logiciel, l'intégrateur par défaut des équations de Newton est l'algorithme "saut de grenouille" (Leapfrog) dans lequel la position $X(t_i + \Delta t)$ est déterminée par la vitesse au temps $t_i + \frac{\Delta t}{2}$ et la position au temps $X(t_i)$:

$$V(t_i + \frac{\Delta t}{2}) = V(t_i - \frac{\Delta t}{2}) + a(t_i)\Delta t, \tag{1.4}$$

$$X(t_i + \Delta t) = X(t_i) + V(t_i + \frac{\Delta t}{2})\Delta t, \tag{1.5}$$

$$V(t_i + \Delta t) = V(t_i + \frac{\Delta t}{2}) + a(t_i + \Delta t)\frac{\Delta t}{2} \tag{1.6}$$

La méthode est stable pour les mouvements oscillatoires pour des temps d'intégration Δt plus petits que la période du système harmonique. Le choix d'une taille de pas d'intégration dépend donc des plus petites périodes d'oscillation de notre système d'atomes qui sont les vibrations dans les liens covalents impliquant un atome d'hydrogène ($\sim$11 fs pour un lien C-H), nécessitant un temps d'intégration de 1 fs. En utilisant des méthodes itératives d'application de contraintes sur les liaisons covalentes telles que SHAKE [213] ou encore LINCS [86], on peut éliminer ces modes oscillatoires et pousser le pas d'intégration à 2 fs.

1.2.1.2 Ensemble isothermique-isobarique

Nos simulations sont exécutées à température et pression constantes (ensemble NPT). Ces conditions sont maintenues par un bain de température et bain de pression. Dans les simulations de DM sur SAV1866, nous avons opté pour un bain de Berendsen [14] pour la température et pour la pression. Dans l'équation qui suit, l'algorithme de Berendsen diminue progressivement l'écart entre la température observée $T(t_i)$ et la température désirée T_0 à l'aide d'une constante de couplage τ :

$$T(t_i + \Delta t) = T(t_i) + \frac{\Delta t(T_0 - T(t_i))}{\tau}, \tag{1.7}$$

où Δt est le temps d'intégration. La température instantanée est obtenue à l'aide de la relation :

$$T(t_i) = \frac{1}{3Nk_{\mathrm{B}}} \sum_{j}^{N} m_j v_{ij}^2, \tag{1.8}$$

où N est le nombre d'atomes, m_j et v_{ij} sont la masse et la vitesse de chaque atome au temps t_i et k_{B} est la constante de Boltzmann. La variation de température ainsi calculée est rappliquer au système à chaque pas d'intégration en rajustant les vitesses de chaque atome proportionnellement.

Le même algorithme peut aussi être utilisé pour réguler la pression. L'inconvénient avec les bains de Berendsen est qu'ils ne génèrent pas parfaitement l'ensemble canonique désiré. Les systèmes à membrane lipidique sont peu sensibles au choix de l'algorithme de bain de pression tel que démontré par des simulations comparatives entre l'algorithme de Berendsen et celui de Parrinello-Rahman qui démontrent que le volume moyen et la surface moyenne par lipide ainsi que le module de compressibilité ne sont pas affectés par le choix de l'algorithme [6]. L'algorithme est encore utilisé à ce jour [132, 254] et nous avons donc utilisé le bain de pression de Berendsen pour les simulations de MD. Notre usage de ces bains était purement pour des raisons techniques et n'avait pas de justification scientifique.

Pour nos simulations de DMD, nous avons opté pour le bain thermique de Nosé-Hoover [180] et le bain de pression de Parrinello-Rahman [196]. Le bain de Nosé-Hoover a la propriété de respecter l'ensemble canonique d'intérêt. La méthode utilise un réservoir virtuel de masse prédéterminée pour équilibrer la température du système étudié. Contrairement à la méthode de Berendsen qui relaxe la température suivant une courbe exponentielle en τ, l'algorithme de Nosé-Hoover affiche une relaxation oscillatoire. L'algorithme analogue pour les bains de pression est celui de Parrinello-Rahman dont le fonctionnement est similaire et qui respecte l'ensemble NPT.

1.2.1.3 Potentiel énergétique

Des grands potentiels énergétiques couramment utilisés en dynamique moléculaire sur des systèmes biochimiques, CHARMM [19] et OPLS [115] sont bien adaptés aux simulations de systèmes contenant des membranes lipidiques [123] et des potentiels éner-

gétiques de lipides ont été construits pour interagir avec ces derniers. Dans le cas de CHARMM, des définitions tout-atomes des lipides ont été calibrées à l'aide de données expérimentales et *ab initio* [63, 64], alors que pour OPLS l'approche privilégiée emploie une représentation gros-grain en combinant certains termes d'OPLS avec la définition d'atomes unifiés de Berger [16]. Plus récemment, des définitions tout-atome de lipides [210] ont été calibrées à l'aide du potentiel OPLS dans sa forme tout-atome (OPLS-aa) [117] et nous avons opté pour ce potentiel étant donné sa disponibilité avec le logiciel de simulation Gromacs.

Le traitement des interactions électrostatiques est une des composantes les plus coûteuses des simulations de DM puisque pour être exact, ce terme énergétique doit être évalué entre toutes les paires d'atomes du système en temps O(N^2). Traditionnellement, pour économiser du temps de calcul, les forces n'étaient pas calculées entre les atomes dont la distance les séparant dépassait un certain seuil entre 10 et 20 Å. Cependant, pour limiter les artefacts du potentiel électrostatique tronqué [198], une alternative est d'utiliser la méthode du maillage particulier d'Ewald [43] qui divise le calcul en deux parties : Pour les paires d'atomes rapprochées, le calcul des forces électrostatiques est fait dans l'espace réel, mais pour les interactions dépassant un seuil donné, ce calcul se fait dans l'espace de Fourier.

1.2.2 Méthodes de dynamique moléculaire dirigée

Puisque les échelles de temps accessibles aux simulations de DM sont plusieurs ordres de grandeurs inférieures aux temps de réaction biochimique, on peut introduire un biais pour augmenter la probabilité d'échantillonner des événements rares ou improbables. Trois méthodes sont ici présentées. La plus intuitive des trois est la dynamique moléculaire dirigée. En DMD, on introduit un potentiel harmonique entre deux points du système $U_{dirigee}$:

$$U_{dirigee}(X) = \frac{k}{2}(|X - X_{dirigee}(t)|)^2, \tag{1.9}$$

où k est une constante de ressort suffisamment élevée, X est le vecteur multidimensionnel

de la position du système dirigé, $X_{dirigee}$ est la position d'équilibre du système dirigé au temps t. Pour diriger notre système, on n'a qu'à définir une trajectoire pour $X_{dirigee}$ en fonction du temps de simulation. Bien que $X_{dirigee}$ peut suivre toute forme de trajectoire, la plus courante est la trajectoire linéaire amenant la structure du point X_{init} au point X_{final} dans un temps T :

$$X_{dirigee}(t) = X_{init} + \frac{t}{T}(X_{finale} - X_{init}) \tag{1.10}$$

La méthode de dynamique moléculaire biaisée (DMB) [167, 193] est une adaptation adiabatique de la DMD. Au lieu de forcer le système vers une position voulue, on définit un potentiel U_{biais} qui pousse le système seulement dans la direction s'éloignant de la destination :

$$U_{biais}(X) = \begin{cases} \frac{k}{2}(X - X_{biais})^2 & (X < X_{biais}) \\ 0 & (X >= X_{biais}) \end{cases} \tag{1.11}$$

$$X_{biais}(t) = \begin{cases} X_{init} + \frac{t}{T}(X_{finale} - X_{init}) & (X < X_{biais}) \\ X & (X >= X_{biais}) \end{cases} \tag{1.12}$$

Finalement, la méthode de la dynamique moléculaire visée [134, 220] (DMV) définit un potentiel basé sur la racine carrée de la moyenne des distances atomiques au carré (RMSD) U_{vise} :

$$U_{vise}(X) = \frac{k}{2N}(RMSD(X) - RMSD_{vise}(t))^2, \tag{1.13}$$

où N est le nombre d'atomes et $RMSD_{vise}(t)$ est le RMSD désiré au temps T.

1.2.2.1 Choix d'une méthode de tir

Des trois méthodes présentées ci-haut, la DMV est la plus contraignante forçant un système à suivre une trajectoire donnée sans aucune liberté de conformation au niveau atomique. Si un mouvement de torsion ou de rotation temporaire est nécessaire lors de

l'échantillonnage d'une trajectoire, la DMV risque fortement de forcer la structure du système poussé dans des conformations non-naturelles.

Pour un système tel que SAV1866, la DMD et la DMA offrent des avantages de flexibilité. Pour étudier la séparation des domaines de SAV1866, nous avons opté pour la DMD parce que la méthode était déjà implémentée dans Gromacs et parce qu'il est plus facile d'extraire un travail avec la DMD en intégrant la force ressentie par le ressort. Les méthodes de calcul de travail et d'énergie libre sont présentées en plus grands détails à la section 5.2.

1.2.2.2 Choix de points d'ancrage

Deux types de points d'ancrage sont habituellement utilisés. Pour les systèmes tentant de simuler ou de reproduire des résultats de microscopie par force atomique [68], le point d'ancrage est souvent un seul atome de la molécule poussée ou tirée [83, 103]. Ce choix peut par contre déformer la molécule si la vitesse de tir est trop élevée, ce qui est le cas pour les simulations de DMD qui ne peuvent atteindre les échelles de temps expérimentales de microscopie. Même si la molécule n'est pas déformable, l'apposition d'une force sur un atome loin du centre de la molécule affectera l'orientation de la molécule lors du tir.

Un autre choix consiste à appliquer une force sur le centre de masse du système poussé. Cette méthode a l'avantage de permettre des rotations libres du système poussé, mais possède aussi le défaut d'être susceptible de développer un moment angulaire pour de gros systèmes. Nous avons opté pour cette approche avec SAV1866 en utilisant le centre de masse d'un groupe d'atomes entourant le motif de Walker B dans chaque domaine liant le nucléotide. Ainsi, nous diminuons le risque de déformation et nous permettons aux DLNs de pouvoir se déformer à gré.

1.3 Composantes du système simulé

Notre système est composé du dimère de la protéine SAV1866 insérée dans une membrane composée de dilinoleoylphosphatidylcholine (DLPC) et de dilinoleoylphos-

phatidylethanolamine (DLPE) dans un rapport 55% - 45% exemplifié par la figure 1.6.

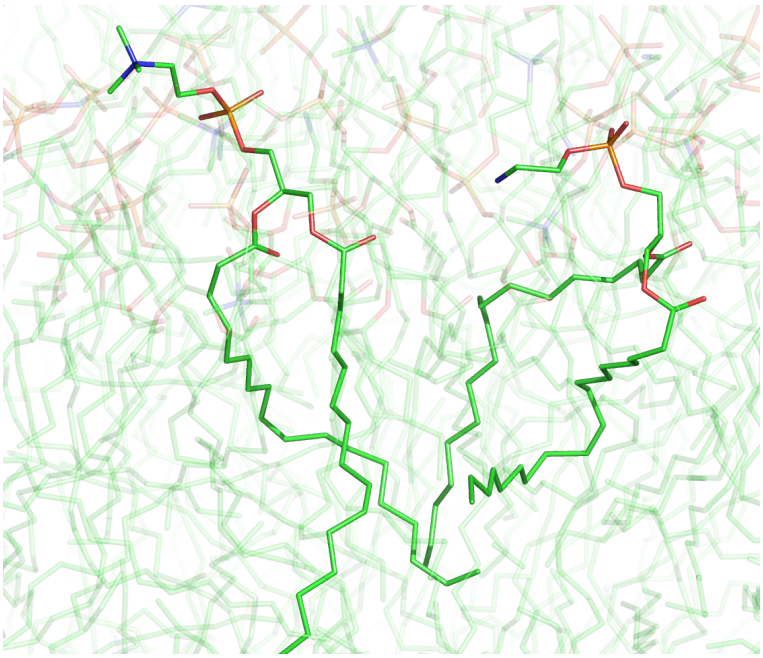

Figure 1.6 – Lipide DLPC à gauche d'un lipide DLPE dans une membrane équilibrée.

Le modèle utilisé pour les molécules d'eau est le TIP3P [116] qui est un modèle tout-atomes à 3 particules réelles et aucune virtuelle. La concentration ionique physiologique du cytosol de 140 mM KCl a été atteinte avec l'ajout de 148 K^+ et de 138 Cl^- dans 55,100 molécules d'eau pour obtenir une charge nette neutre du système.

1.4 Conclusion

Plusieurs questions restent sans réponse sur le fonctionnement des protéines de la famille des ABC Transporteurs. Bien que plusieurs structures tridimensionnelles soient disponibles, le processus de fermeture de ces protéines lors de la liaison d'allocrites ou de nucléotides et celui d'ouverture et d'expulsion des allocrites et des produits de l'hydrolyse de l'ATP sont encore inconnus. Nous proposons dans l'article présenté au chapitre 3 d'élucider un de ces mystères en soumettant la forme fermée de SAV1866 à des simulations de dynamique moléculaire en présence et en absence de ligand ADP

ainsi que des essais de dynamique moléculaire dirigée tentant de séparer la protéine pour obtenir la forme ouverte de cette dernière.

CHAPITRE 2

CONTRIBUTIONS DES AUTEURS À L'ARTICLE SUR LA PROTÉINE SAV1866

- Jean-François St-Pierre a écrit la première version du manuscrit et a effectué les simulations et les analyses présentées, incluant :
  - Équilibration de la membrane,
  - Insertion de la protéine SAV1866 dans la membrane et rééquilibration,
  - Exécution et analyse des dynamiques moléculaires (DM) avec et sans ATP,
  - Établissement du protocole de dynamique moléculaire dirigée (DMD),
  - Exécution et analyse des DMD ouvrant SAV1866, et des CMD et MD sur SAV1866 ouvert,
  - Présentation de l'analyse en discussion et des conclusions
- Alex Bunker a supervisé le travail de DM et Normand Mousseau le travail de DMD.
- Tomasz Róg a contribué à l'élaboration du potentiel de la membrane lipidique.
- Tous les auteurs ont contribué aux révisions et aux corrections du manuscrit.

CHAPITRE 3

ARTICLE : MOLECULAR DYNAMICS SIMULATIONS OF THE BACTERIAL ABC TRANSPORTER SAV1866 IN THE CLOSED FORM.

Jean-François St-Pierre
Département de Physique and Regroupement Québécois sur les Matériaux de Pointe, Université de Montréal, C.P. 6128, succursale centre-ville, Montréal (Québec) Canada H3C 3J7
Alex Bunker
Centre for Drug Research, Faculty of Pharmacy, University of Helsinki, PO Box 56, FI-00014, University of Helsinki, Finland
Tomasz Róg
Department of Physics, Tampere University of Technology, PO Box 692, FI-33101 Tampere, Finland.
Mikko Karttunen
Department of Chemistry, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1
Normand Mousseau
Département de Physique and Regroupement Québécois sur les Matériaux de Pointe, Université de Montréal, C.P. 6128, succursale centre-ville, Montréal (Québec) Canada H3C 3J7

3.1 Abstract

The ATP Binding Cassette (ABC) transporter family of proteins contains members involved in ATP-mediated import or export of ligands at the cell membrane. For the case of exporters, the translocation mechanism involves a large scale conformational change that involves a clothespin-like motion from an inward-facing open state, able to bind ligands and adenosine triphosphate (ATP), to an outward-facing closed state. Our work focuses on SAV1866, a bacterial member of the ABC transporter family for which the structure is known for the closed state. To evaluate the ability of this protein to undergo conformational changes at physiological temperature, we first performed conventional molecular dynamics (MD) on the co-crystalized adenosine diphosphate (ADP)-bound structure and on a nucleotide-free structure. With this assessment of SAV1866's stability, conformational changes were induced by steered molecular dynamics (SMD), in which the nucleotide binding domains (NBD) were pushed apart, simulating the ATP hydrolysis energy expenditure. We found that the trans-membrane domain is not easily perturbated by large scale motions of the NBDs.

3.2 Introduction

ABC transporters are a family of over 1,000 proteins involved in active, i.e., non-diffusive, ATP-hydrolysis dependent transport of ligands across cell membranes [24, 93, 155, 208]. They transport vital molecules such as lipids, steroids and vitamins [93] and, as a result, are involved in drug transport [219]. Mutations that affect the expression of ABC transporters can have severe consequences. Over-expression of ABC transporters is the leading cause of chemotherapy resistance in cancer treatment [81].

ABC transporters belong to the class of membrane-spanning proteins. It is well-known that the structural characterization of membrane proteins is particularly difficult, see, for example Ref.[194], and hence it has not been easy to obtain full crystal structures. The structures of individual domains have, however, enabled computational modeling studies, which have provided significant insight into the structure-function mechanisms of ABC transporters [1, 13, 102, 122, 186–188, 228, 269].

Structurally, all ABC transporters, whether eukaryotic or not, have two pairs of domains : The transmembrane domains (TMD) and the nucleotide binding domains (NBD). The former may have considerable sequence variations, but the latter are highly conserved [93]. Since full ABC transporters display a structural radial symmetry, they have two TMD and two NBD domains which can be encoded by an individual gene for each domain, by a gene containing one TMD and one NBD, or by a gene containing the full TMD–NBD–TMD–NBD sequence.

The NBDs hydrolyze two ATP molecules to provide the energy necessary for the efflux cycle. Both domain pairs are believed to move towards each other during the transport cycle, as the protein undergoes a transition from the open, ligand binding, to the closed, ligand expelling, conformation. Figure 3.1 shows crystal structures of two ABC transporters, mouse P-glycoprotein (permeability glycoprotein) [4] in the open conformation and its close bacterial homologue SAV1866 [44] in the closed conformation. The figure shows that the conformational change the protein undergoes during the efflux cycle involves both a large ($\sim$2 nm) displacement of the NBDs and the transformation from an open conformation with an inward facing cavity exposing the allocrite binding region in the TMDs to the cytosol and the membranw's inner leafleat to a closed conformation with an outward facing cavity in the TMDs when effluxing the allocrite.

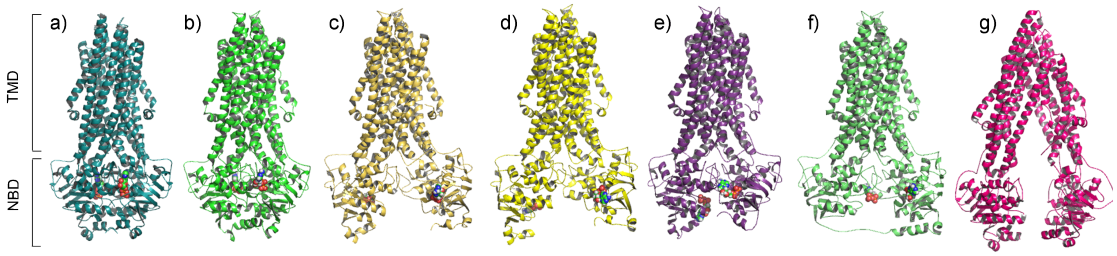

Figure 3.1 – Reference crystal structures : a) SAV1866 in closed conformation (pdb :2HYD) and g) mouse P-glycoprotein (pdb :3G5U) in the open conformation. Final results for a range of ADP-bound simulations : b) MD-ADP, c) SMD-ADP1, d) CMD-ADP1, e) MD-ADP-RF and f) SMD-ADP-H204A2 (see Table 3.I for details of the notation). ADP is shown in sphere representation.

Tableau 3.I – Reference table of all simulations performed in this work, the method used and their starting structures where MD, SMD and CMD stand for molecular dynamics, steered MD and constant-restrain MD, ML stands for membrane-less and RF for refolding.

Simulation name	Initial structure	Method	Length
MD-ADP	Crystal structure with ADP bound, inserted into membrane	MD	100 ns
MD-APO	Crystal structure inserted into membrane, ADP removed	MD	100 ns
SMD-ADP1	MD-ADP 100 ns structure	SMD	20 ns
SMD-ADP2	MD-ADP 100 ns structure	SMD	20 ns
SMD-ADP-ML	MD-ADP 100 ns structure without membrane	SMD	20 ns
SMD-ADP-H204A1	MD-ADP 100 ns structure with residues His204 mutated to Ala	SMD	20 ns
SMD-ADP-H204A2	MD-ADP 100 ns structure with residues His204 mutated to Ala	SMD	20 ns
SMD-APO1	MD-APO 100 ns structure	SMD	20 ns
SMD-APO2	MD-APO 100 ns structure	SMD	20 ns
SMD-APO-ML	MD-APO 100 ns structure without membrane	SMD	20 ns
CMD-ADP1	SMD-ADP1 20 ns structure	CMD	14 ns
CMD-ADP2	SMD-ADP2 20 ns structure	CMD	14 ns
CMD-APO1	SMD-APO1 20 ns structure	CMD	14 ns
MD-ADP-RF	CMD-ADP1 14 ns structure	MD	60 ns
MD-APO-RF	CMD-APO1 14 ns structure	MD	60 ns

Another protein of interest is MsbA for which three structures of lower resolution were obtained : An inward-facing open conformation and an outward facing closed conformation that are respectively similar to the P-glycoprotein and SAV1866 protein conformations mentioned above, but also a third conformation with closed NBD and inward-facing cavity [264].

In the current report, we are specifically interested in the *Staphylococcus aureus* multidrug transporter SAV1866. The ability of SAV1866 to efflux a large number of drugs makes it a good general model for multidrug ABC transporters, including human P-glycoprotein [256], which is involved in the transport of, for example, peptides, lipids and various xenobiotics. SAV1866 has a homodimer structure : The sequence of each component of the dimer contains one TMD and one NBD, with the two components of the dimer symmetrically oriented along a rotation axis perpendicular to the lipid bilayer. Aittoniemi *et al.* [1] performed simulations on SAV1866 to evaluate the effect of replacing the co-crystallized ADP by the active ligand ATP and Mg^{2+} in the NBDs. They found an asymmetric reorientation of the NBD interface regions towards the TMD interface. Becker *et al.* [13] saw increased constriction of the TMD helices in the membrane in the apo structure relative to the structure with ATP bound. Oliveira *et al.* found [185], by inserting the products of ATP hydrolysis, ADP + Mg^{2+} and inorganic phosphate (IP),

evidence of TMD separation close to the NBD interface in both the substrate-bound and product-bound structures. In the most recent simulation of truncated SAV1866 with single ATP + Mg^{2+}, Jones and George [113] observed a rotation in the NBDs of the SAV1866 homodimer in agreement with experimental results on heterodimer NBD domains of other ABC transporters, notably of P-glycoprotein [215, 282, 287]. This indicates that the otherwise symmetric NBDs of SAV1866 may also undergo asymmetric transformations during the efflux cycle.

We present atomistic simulations performed on the closed form of *S. aureus* ABC transporter SAV1866 in which we investigated the large scale structural motions that this protein may sample. We have completed two 100 ns MD simulations of the SAV1866 closed complex inserted in a dilinoleoylphosphatidylcholine (DLPC) and dilinoleoylphosphatidylethanolamine (DLPE) lipid bilayer, one in the presence of the co-crystallized ADP molecule and one with the ligand removed based on the hypothesis that these structures would favor conformation changes. We also investigated conformational changes when separating the two NBD using 20 ns SMD simulation runs. In the targeted MD performed by Weng *et al.* [269] on MsbA, all α-carbons (Cα) were forced to change conformation. We chose a different approach in our SMD simulations : A single harmonic potential between the centers of mass of the two NBDs is introduced in order to induce a transformation from the closed to the open conformation. Finally, we consider mutagenesis to test mechanisms for enhancing flexibility.

3.3 Methods

MD simulations were performed using the Gromacs 4 software package [88] with the Optimized Potentials for Liquid Simulations (OPLS-AA) all-atom force-field [115] and periodic boundary conditions. Electrostatic interactions were computed using the Particle-Mesh-Ewald method (PME) [43, 125] with a real space cut-off of 1.0 nm ; It was also ensured that charge groups were small, in order to avoid possible physical artifacts that have been reported in similar simulation geometries of nanotubes [271]. It has been shown that this choice is devoid of artifacts and leads to physically correct

behavior, both static and dynamic [197, 199] . The Lennard-Jones interactions were also cut off at 1.0 nm. SAV1866 crystal (pdb : 2HYD [44]) was inserted in a pre-equilibrated DLPC [210, 235]/DLPE lipid membrane of size 16.3×16.3 nm. A hole was created in the lipid membrane through the removal of all lipids that came into contact with the membrane-aligned protein [55, 102]. The remaining lipid membrane was composed of 217 DLPC and 175 DLPE lipids. The protein's termini were in their zwiterrionic form : The histidines 103, 457 and 559 of each dimer were protonated on the Nδ atom and all other histidines were protonated on the Nε atom. All aspartic acid, arginine, glutamic acid, and lysine were used in their default protonation states at physiological pH. Protein and membrane were solvated in a water box of 17 nm in height using the TIP3P water model [116]. K^+ and Cl^- ions were added to obtain a 140 mM concentration of KCl in order to model the cytosol while keeping the net charge of the system neutral. The simulation contained a total of $\sim$235,000 atoms. The protein's backbone coordinates were harmonically restrained for a short MD simulation of 425 ps at 310 K in which the gap between the membrane and protein disappeared. Any water molecule still present at the membrane/protein interface was manually removed.

Following a 1000-step steepest descent system minimization, a 2 ns MD relaxation with a smaller integration time step of 1 fs was executed in which the temperature was increased from 10 K to 310 K leading to a thermally equilibrated structure with a 0.28 nm backbone root-mean square deviation (RMSD) from the crystal structure. The size of the final periodic box after releasing the restraints was 11.6 nm $\times$ 11.6 nm $\times$ 17.7 nm. This ensures that the protein is always at least 4 nm away from its periodic images while it is aligned perpendicularly to the membrane plane. Production runs were executed using a 2 fs time step with covalent bond length constrained by LINCS [86]. A snapshot of the system's conformation was saved every 20 ps for analysis purposes.

All simulations were performed twice : Once with the co-crystalized ADP molecules and once on a nucleotide-free apoprotein (APO) structure. The APO conformation was generated by removing ADP from the ADP-bound system after 20 ns of simulation time followed by a 5 ps MD simulation with restrained protein backbone atoms and 60 ps of unrestrained MD at temperatures increasing from 10 K to 310 K to allow for equilibra-

tion of the water molecules in the ATP binding cavities. Initial conventional 100 ns MD simulations were completed using Berendsen weak coupling thermal and pressure bath fixed at 310 K and 1 bar with coupling time constants of 0.1 and 1.0 ps respectively [14]. System conformations were saved every 20 ps for analysis.

SMD simulations were completed by pushing the centers of mass of the two NBDs apart using a harmonic potential between the centers of mass (COM) of the two NBDs at rate of 0.1 nm/ns. The force constant was set to 2.5 MJ/(mol·nm^2). External forces were applied to 45 backbone atoms of residues ILE417-GLN421, ILE498-ALA504 and THR528-ALA533 from three strands of the parallel β-sheet found at the centers of each of the NBDs. The reasons for restraining these atoms are multiple. By applying the steering forces to a limited number of atoms with a COM located near the NBD's COM, we limit the possibility of generating unlikely structural deformations of the NBDs that artificially increasing the NBD's COM distances. The β-sheet structure of the selected atoms and its position at the center of the two halves of each NBD add to its structural stability. Also, the central position of the β-sheet's COM is less likely to induce bias in the separation angles observed when the NBDs break contact. Temperature and pressure were maintained through the Nosé-Hoover thermostat [95, 180] and a Parrinello-Rahman barostat [196] respectively, with the same coupling constants as used in the MD simulations. The work exerted by the harmonic force can be extracted through numerical integration of the force applied by the potential over time, multiplied by the total displacement.

For constant-restraint molecular dynamics (CMD), we used the same parameters as SMD, but with a null pull rate, i.e., we maintained a fixed harmonic bond length attached to the centers of mass of the two NBDs. Details concerning each method and starting conformation for all simulations are listed in Table 3.I. PyMol (http ://www.pymol.org/) was used for visualization.

3.4 Results and discussion

3.4.1 MD simulations

The RMSD as measured from the crystalline structure is shown for both the ADP-bound and the APO proteins in Figure 3.2 a). After a rapid growth in the first 2 ns of equilibration prior to the production runs to a RMSD of 0.28 ± 0.01 nm, the RMSD slowly increases to an average of 0.38 ± 0.01 nm and 0.36 ± 0.01 nm, respectively, during the last 20 ns of each 100 ns simulation. This is in line with the average of 0.34 nm, observed by Aittoniemi *et al.* [1] with ATP and Mg^{2+} present. While RMSD to the crystal structure of SAV1866 increases by only 0.8-1.0 nm during the production runs after equilibration, the protein is seen to explore a conformation subspace leading to RMSD values of 0.20 and 0.18 nm from the start of the production runs for MD-ADP and MD-APO respectively. RMSD on the pairs of TMDs and NBDs (Table 3.II) were also comparable to the previously published results, and slightly lower than the apoprotein simulations of Becker *et al.* [13] (Figure 2 in Ref. [13]) and the ATP-bound and ADP-IP bound simulations of Oliveira *et al.* (Figure 2 in Ref. [185]).

Tableau 3.II – Average Cα-RMSD measured over different domains between the reference initial conformation and the structures from last 20 ns of MD simulation for the ADP-bound (MD-ADP) and APO (MD-APO) SAV1866. Standard deviation for all points is 0.01 nm.

Domain	ADP-bound (nm)	APO (nm)
TMD1	0.40	0.33
NBD1	0.17	0.20
TMD2	0.42	0.39
NBD2	0.19	0.19
Both TMD	0.42	0.38
Both NBD	0.22	0.24

Looking at the RMS fluctuations for each of the atoms that correlate with the diffraction B factors [44], we find as expected, that the side chains in the ATP binding cavity of the APO structure show higher flexibility than the ADP-bound protein. However, there were no signs of large instabilities (data not shown). The most notable conformational changes occurred in the TMD region. In both the ADP bound and APO simulations, the

TMD becomes constricted (see Figure 3.3). This brings helix H1 into contact with facing helices H9 and H12 and also helices H3 and H6 into contact with H7. Opposites H6 and H12 (in grey) also come into contact. When examining the the number of water molecule within 0.7 nm of helices H6 and H12 throughout the MD simulations, we see a diminution of 50 water molecules for MD-ADP and 20 for MD-APO between the beginning of the simulation and at time 100 ns. This reduction is not correlated by a reduction of the solvent accessible surface area of the protein or of the cavity (Figure 3.2 b). Similar constriction was also observed by Becker *et al.* [13] in their APO structure simulation of SAV1866 over a 80 ns MD, but not in the ATP/Mg^{2+} bound structure.

The lipid density plots were obtained by counting the number of lipid head groups in a region 4.5 nm away from the center of mass of the membrane inserted protein's a.a., divided by the corresponding periodic box's surface area minus the occlusion disk of 4.5 nm radius that contains the inserted protein's TMDs. Figure 3.4 show that the TMD constriction is not caused by a pressure imbalance between the outer and inner lipid bilayer leaflets : We see no correlation between the closing motion of the TMD over the 100 ns simulation and the ratio of inner and outer leaflet lipid densities. On average, the outer leaflet lipid densities were 1.96 lipids/nm^2 and 1.94 lipids/nm^2 for the ADP-bound and APO simulations, respectively, while they are 1.87 lipids/nm^2 and 1.93 lipids/nm^2, respectively, in the inner leaflet. We also noticed the standard deviation in lipid density to be 22 % higher in the ADP-bound simulations than in the APO simulations.

The core helices H6 and H12 are the most deformed ones through the constriction in the bilayer zone, see Figure 3.5. In the ADP bound simulation, this deformation also involves a shortening of 0.36 nm of H12 and elongation of 0.17 nm of H6 calculated as the distance between residues GLY276 and ASP319 of each helix.

Helices H3-H4 and H9-H10, shown in Figure 3.6 and forming the core region of the lower TMD, are unaffected in the lower half but bend sharply at residue GLY183 of H4 to accommodate the constriction in a hinge-like motion at the stated residue.

The amino acid interactions at the NBDs interfaces differ from those observed by Aittoniemi *et al.* [1]. To compare our NBD interactions with theirs, we defined a contact ratio metric between two amino acids as the number of conformations where any atoms

of an amino acid of the first dimer's NBD (NBD1) is found at a distance of less then 0.3 nm from any atoms from an amino acid of second dimer's (NBD2), divided by the total number of saved conformation snapshot.

In the following NBD analysis, when a contact between two amino acids is stated, the first amino acid is in NBD1 and the second in NBD2. Therefore, if amino acid ARG474 from NBD1 is at a distance lower than 0.3 nm from ARG474 of NBD2 in 50 conformations out of 100, then we would say that the ARG474-ARG474 contact ratio is 0.50. In both our MD simulations, we see a reorientation of both ARG474 in the membrane plane (both amino acids remain parallel to the membrane plane) but unlike Aittoniemi *et al.* [1], we observed no stacking. The x-ray structure contact ARG474-ARG474 is also weak in our simulations with a contact ratio of no more than 0.11 in both MD simulations. The reorientation of the ARG474-ARG474 pair also involves a break in contact between GLN208 of TMD1 and ARG474 of NBD1 in both MD-ADP and MD-APO simulations, while the symmetrical contact is maintained between TMD2 and NBD2. Another contact of the NBDs with the TMDs which is observed to break is between TYR112 of TMD2 and GLY472 of NBD1 with a conserved symmetrical contact between TMD1 and NBD2. Aittoniemi *et al.* reported that ASP423-ARG474 and ARG474-ASP423 broke contact in their simulations prior to the formation of contacts ASP423-LYS483 and LYS483-ASP423, which freed ARG474 to adopt a stacked conformation along the axis perpendicular to the membrane. In our case, ASP423-ARG474 has a conserved contact ratio of 0.99 in MD-ADP and 0.20 in MD-APO, while the symmetrical ARG474-ASP423 has a contact ratio of 0.98 and 0.99, respectively. This does not prevent the formation of contact ASP423-LYS483 with contact ratio of ratio of 0.99 in MD-ADP and 0.89 in MD-APO, but it does prevent the formation of the symmetrical LYS483-ASP423 which was not found in either simulation. The later was replaced by a contact between LYS483-GLN422 with contact ratio of 0.83 in MD-ADP and 0.56 in MD-APO. This was made possible by the absence of Mg^{2+} in our simulations. Mg^{2+} interacts typically with GLN422.

In spite of these observations, we noticed only very small differences in the ADP-bound and APO NBD stability after simulations of 100 ns. To identify the opening modes

on computationally accessible timescales, it is necessary to use more forceful methods such as SMD.

3.4.2 Steered molecular dynamics

We can simulate the injection of energy into the NBDs by exerting a pressure on a group of atoms located at the center of each NBD. This is done through a harmonic potential of increasing equilibrium length connecting the various domains. For both ADP-bound and APO structures, we performed three 20 ns SMD simulations which is the amount of time needed to separate the NBDs by 2 nm. Seperations of more than 2 nm lead to unfolding of the NBDs and break of contacts with the TMDs which have no known biological relevence. In two of the cases, we replaced the lipid bilayer by water and ions to test the impact of the presence/absence of a membrane (see Table 3.I).

We observed two modes of NBD separation, Figure 3.7. The selection of a mode depended on the time it took to break the contacts between the two domains. As shown in Figure 3.8, the angle between the NBDs increases significantly in the plane parallel to the membrane before the 12th nanosecond in all simulations. This indicates a peeling separation where one side of the contact interface between the two NBDs breaks before the other. In most cases, the angle between the domains converges back to parallel when the external force is released. In two of the cases, SMD-ADP2 and SMD-ADP-ML, the contact was maintained throughout the simulations and the angle remained at 28 and 35 ± 2 degrees, respectively.

In the six SMD simulations, we computed the work of opening SAV1866 by taking the integral of the force distribution on the harmonic forcing separation. To do so, we integrate numericaly over time the force component of the pulling potential :

$$W = \int F v dt \tag{3.1}$$

where F is the instantenuous force on the harmonic potential and v is the pulling rate. Generally, we see that the presence of ADP and the presence of a lipid membrane have strong cumulative stabilizing effects (Table 3.III).

Tableau 3.III – Work of separating the NBDs for the six simulations and the resulting angle between the separated NBDs calculated in the protein membrane plane. Angle incertitude is the standard deviation of the last 2 ns of SMD simulation. See Table 3.I for a description of each system.

System	Work (kJ/mol)	Final inter-NBD angle (degrees)
SMD-ADP1	1050.34	2 ± 2
SMD-ADP2	1041.60	28 ± 2
SMD-ADP-ML	685.75	35 ± 2
SMD-APO1	629.35	7 ± 2
SMD-APO2	780.86	10 ± 1
SMD-APO-ML	572.60	5 ± 2

Although we used a slow pulling rate of 0.1 nm/ns, we were unable to sample the opening of the TMD domain as observed in P-glycoprotein structures (pdb : 3G5U [5]) or in both the open and closed inward-facing structure of MsbA [264] for any of the six simulations. In the lower extremities of the TMD a few contacts were broken during SMD, resulting in a semi-open state. In the crystal structure, TMDs make contacts with the NBDs through coupling with helices : The S108-N115 α-helix CH1 (CH3 on the second TMD) linking TMD helices H2 and H3 (H8-H9 on the second TMD) makes contacts with both NBDs and one ADP at the NBD's interface. The G209-F216 α-helix CH2 (CH4 on the second TMD) linking TMD helices H4 and H5 (H10-H11 on the second TMD) makes contact with the opposite NBD (see Figure 3.6). The open P-glycoprotein structure [5] features contacts between the CH1 and CH3 helices and their sequentially closest NBD domain which differs from our SMD results by a tilt of the NBDs bringing the C-terminal extremities closer while keeping the centers of mass of the two NBDs in place and opening the TMD core in a clothespin like motion. In all SMD simulations, we observe a break of contact between the NBDs and the CH1 and CH3 helices, but not with the CH2 and CH4 helices. We also see that in all cases, the ADP's phosphate groups stay in contact with the conserved Walker A motif G374-S381 [44, 154, 221].

Since the SMD simulations on wild-type SAV1866 structure did not generate separation events of the TMDs, we opted to extend the simulation time in which the protein's NBDs are separated.

3.4.3 Constant restrained molecular dynamics

Using constant restrained molecular dynamics (CMD) and the result of three wild-type SMD simulations, we extended the simulation time of the semi-opened states by 14 ns. We maintained the distance between the NBD centers of masses constant at 2.0 nm in order to let the protein to react to the external force. On short time scales, the results showed little deformation of the protein, as shown in Figure 3.9.

To ensure that the time scale was sufficient for large-scale motion, we tested the reversibility of the conformational change imposed by SMD. From the CMD conformation obtained, the refolding MD simulations were launched by removing the restraints. Unrestrained MD simulations for the final structure of SMD-ADP1 (parallel separation) and SMD-APO1 (skewed separation of 7.5 degrees) show a rapid refolding (25 ns) followed by little structure improvement in RMSD as shown in Figure 3.10.

In both cases, refolding was incomplete with a minimum RMSD of 0.40 nm in the ADP-bound simulation after 18 ns (MD-ADP-RF) and with a minimum RMSD of 0.51 nm after 23 ns for the APO simulation (MD-APO-RF) with initial unfolded RMSD were of 0.75 nm and 0.80 nm respectively. While the presence of ADP may be driving a faster refolding process in comparison to the APO simulation, it may also be hindering it at the end when the ADPs' phosphates are still in contact with the Walker A motif [221] and the other atoms are no longer in the crystal structure orientation. Irrespective of the impact of the ADP, these observations show that large-scale motion is possible within our simulation time scale.

3.4.4 Mutation assay

With the TMD stability being the limiting factor in observing the full opening of an ABC transporter, methods that target or destabilize the TMD may be needed. We attempted to destabilize the core interface of the TMD by the double H204A mutation. In Figure 3.7, His204 of helices H4 and H10 were highlighted in cyan and magenta respectively to illustrate the stability of these contacts after 2.0 nm of NBD separation. In the conventional MD simulations, His204 of H4 and H10 is in contact with its mirror image

96 % of the simulation time, with Gln116 of H3 and H9 91% of the time, Gln208 of the opposite H10 and H4 52 % of the time, as well as hydrophobic Val117 and Gly118 of H3 and H9 52 % and 59 % of the time respectively. Its position at the core of the cytosolic extremity of the TMD indicates that it must be one of the first residues to break contact when the TMDs separate in the opening process. We designed the H204A mutant to evaluate the possibility of separating the TMD through SMD. Starting from the 100 ns conformation of the ADP-bound MD-ADP, His204 side chain atoms were replaced by an Alanine methyl group. The system was relaxed and two SMD following the same protocol were performed. The first, SMD-ADP-H204A1, is similar to the previous SMD-ADP1 simulation with respect to the TMD stability and NBD inter-domain angle, converging to an angle of 7 ± 3 degrees. The second, SMD-ADP-H204A2, displays two new characteristics : 1) the c-terminal of one NBD stays in contact with its counterpart NBD throughout the simulation, effectively unfolding its secondary structure (see Figure 3.1 f). 2) the gap left by the Alanine mutation is not filled by neighboring amino acids like in the SMD-ADP-H204A1 run. Instead, a tunnel is left open, filled by roughly 8 water molecules.

3.5 Conclusions

By studying a bacterial homologue of P-glycoprotein, a protein of vast interest in drug research, our goal was to elucidate some elements of the conformational intermediates leading to the open and active form. SAV1866 is a very stable protein in MD simulation timescales. The conformational changes we observed in the membrane-inserted TMD region in both our ADP-bound and APO simulations were also found in APO simulation of Becker *et al.*, but not in the their ATP + Mg^{2+} simulation [13]. SAV1866 was also crystallized with a homologue of ATP, namely AMP-PNP + Na^{+} bound in its NBDs (pdb : 2ONJ [45]). Although more negative than ATP + Mg^{2+} (by one charge), it still holds the same structure as the ADP-bound case [44]. These conformational changes have not been previously reported [1, 185], and hence independent simulations and experiments are needed to examine if they have been overlooked in previous simulations

or are a simulation artifact.

Stability of the NBD during steered deformation is affected by the bilayer and the ADP, as suggested by the work exerted by the harmonic potential to steer the protein to a semi-open state. The stability of the TMD remains unaffected in all cases. The extra work needed to separate the ADP-bound NBDs is in agreement with the targeted MD results performed on MsbA [269] which reported increased stability for their ATP + Mg^{2+} simulations compared to the APO simulations. However, the separation pathways and the final states of the TMD were not influenced by the lipid bilayer or the ADP. Due to the small number of SMD simulations that were performed for each system we can't evaluate the error on the observed work exerted by the pulling potential.

The fact that in all our ADP-bound simulations the ADP stayed in contact with the Walker A motif [221] throughout the 20 ns of SMD simulation, is a clear indication that this motif may be the recruitment factor for ATP in an open conformation. The observed minimum angle peak of 10 degrees between the NBDs during separation may indicate that a peeling separation may be energetically preferable, at least initially. A similar NBD angle was observed in the MD simulations of Jones *et al.* [113] when the NBDs were populated by only one ATP and where the distance in the APO half of the NBD pair grew more distant. This suggest a coordinated sequential hydrolysis of ATP.

Deformations of the NBD are reversible on a short time scale by removing the restraints. The fast stabilization of these refolding simulations points to an unstable conformation forced by SMD. The observed stability suggests that the ATP hydrolysis energy exchange timescale between the ABC and TMD is beyond the scope of our MD simulations. Contrary to the simulations of Oliveira *et al.* [185] where a wide range of deformations including spontaneous separation of the TMDs in ATP-bound and ADP+IP bound simulations were observed, the cytosolic side of our MD simulations was uneventful. When comparing our simulation protocol to Oliveira *et al.* [185], none of their doubly protonated histidines that are singly protonated in our work are close to the cytosolic TMD regions shown to open and only the Hish534 might contribute to the stability of the NBDs. Our SMD assay on the H204A mutants demonstrated that a single point mutation was not sufficient to destabilize the TMD interface and provoke an opening of the

TMD.

Our present work was conducted using the hypothesis that the energy expenditure of ATP hydrolysis was the driving factor of the ABC and TMD separation, so our SMD protocol is akin to injecting energy close to the hydrolysis site. In light of our results, we must consider that either the opening is energy-driven on a much slower time scale that can be sampled through SMD, or the process is initiated by the allosteric effect of the presence of the products of ATP hydrolysis in the binding cavity, as was suggested in a recent numerical simulation [185].

3.6 Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (MK, NM), the Academy of Finland (TR,MK), Canada Research Foundation (NM) and the GALENOS program (JFSP, AB). We are grateful to the Réseau québécois de calcul de haute performance (RQCHP), SharcNet [www.sharcnet.ca] and the Finnish IT Centre for Science for their generous allocations of computer time.

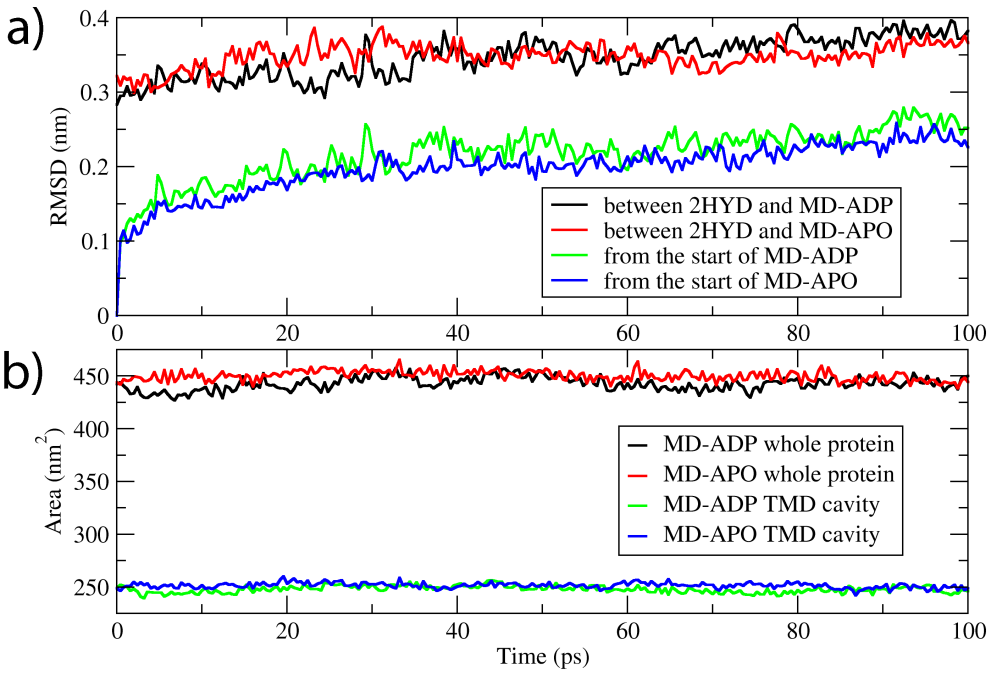

Figure 3.2 – a) Evolution of the backbone RMSD as measured from the SAV1866 crystal structure over the whole 100 ns MD trajectory for the ADP-bound (MD-ADP) and APO (MD-APO) SAV1866. Also presented is the evolution of the RMSD as measured from the start of the production simulations for MD-ADP and MD-APO. b) Solvent accessible surface area (SASA) of the whole protein excluding membrane contacts for MD-ADP and MD-APO and SASA of the TMD inner cavity helixes calculated on residues PHE17-SER89, ASN126-GLN200 and ALA250-SER307 of each TMD.

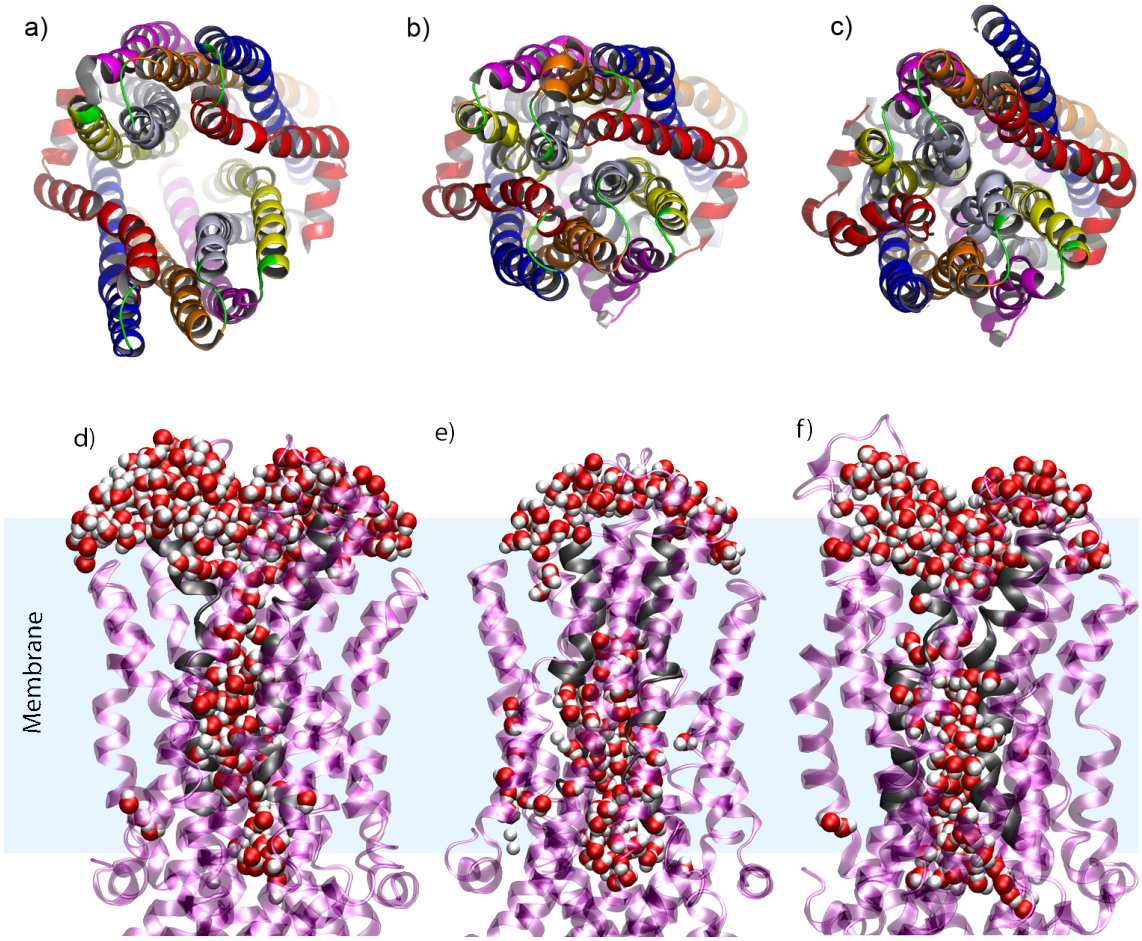

Figure 3.3 – Trans-membrane domain (TMD) helices viewed from the external side of the lipid membrane where (a) is the initial SAV1866 structure (PDB : 2HYD), (b) is the result of 100 ns of MD for the ADP-bound structure (MD-ADP) and (c) is the result of 100 ns for the APO structure (MD-APO). Color code for the first domain helices is H1(red), H2(blue), H3(yellow), H4(magenta), H5(orange), H6(grey). Helices H1 through H6 of the second homodimer TMD are labeled by the same color code respectively and are referenced in the text by the names H7 to H12 for clarity. Also presented, cross-membrane view (d) of the starting conformation of MD-ADP, and the 100 ns conformation (e) of MD-ADP and (f) MD-APO with the residues VAL277 to PHE303 of the H6 and H12 helices in grey and all the water molecules within 0.7 nm of these residues in van der Waals representation.

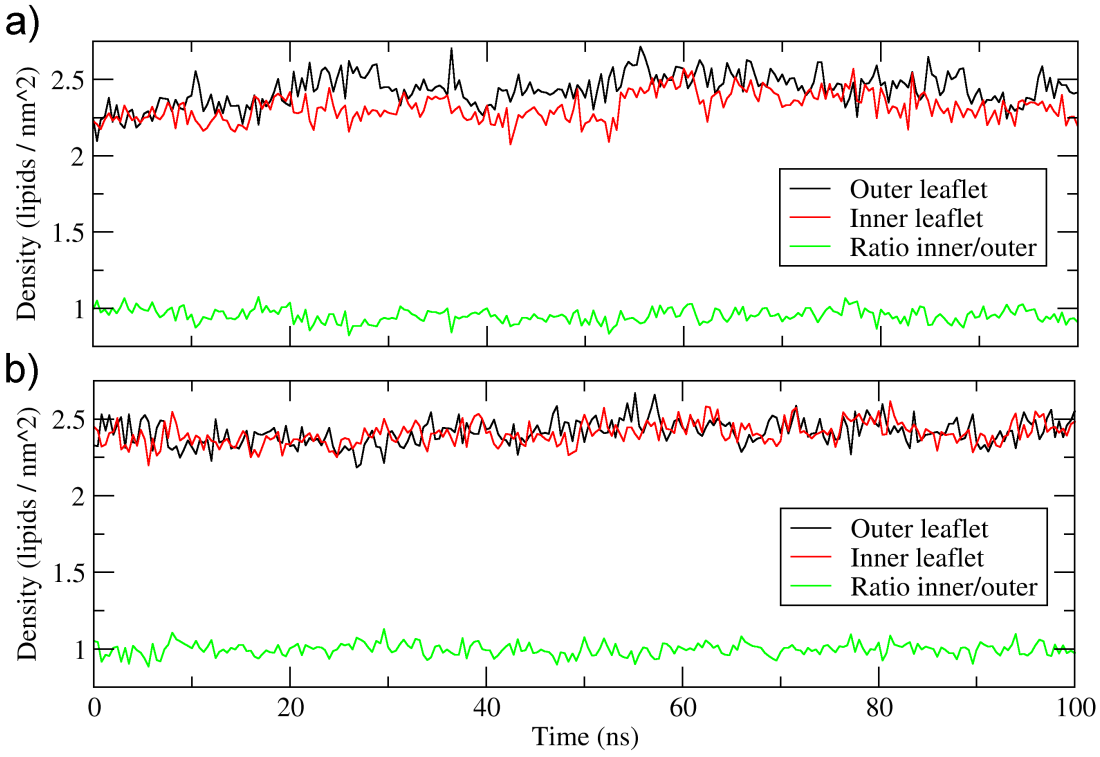

Figure 3.4 – Lipid density of the bilayer leaflets of a) ADP-bound MD simulation (MD-ADP), and b) the APO MD simulation (MD-APO).

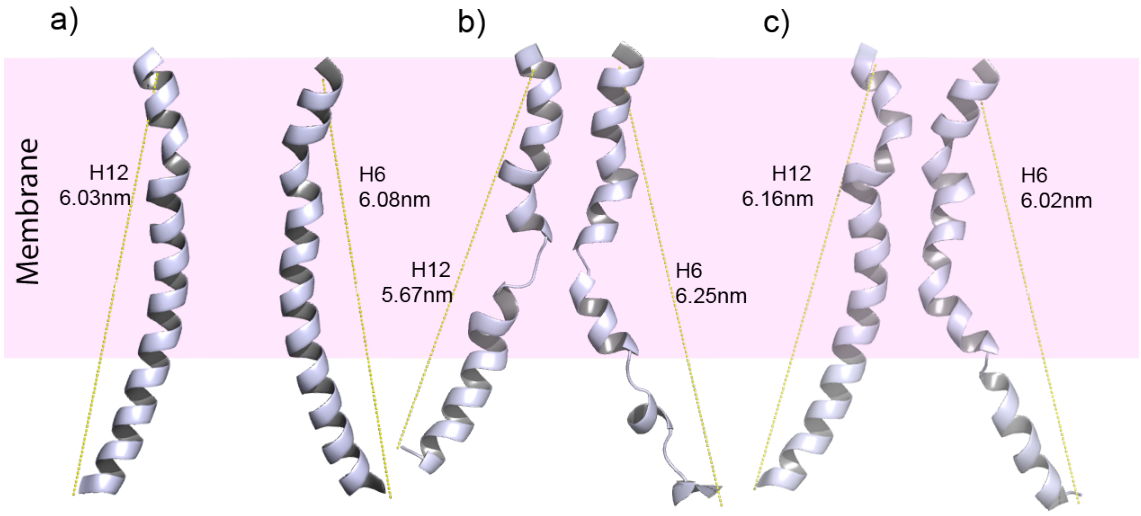

Figure 3.5 – TMD helices H6 and H12 where a) is the initial SAV1866 structure, b) is the result of 100 ns of MD for the ADP-bound structure and c) is the same result for the APO structure.

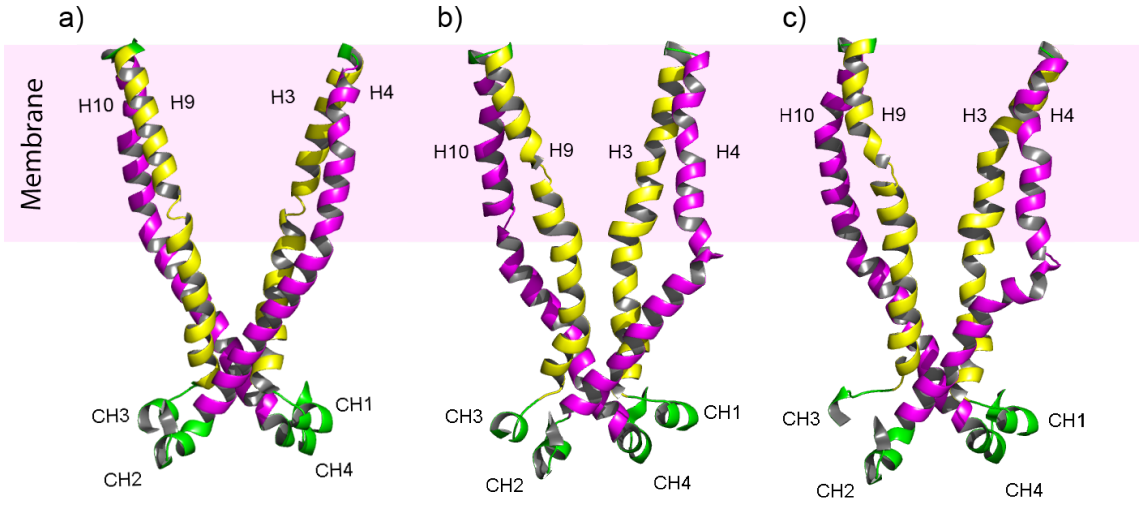

Figure 3.6 – TMD helices H3-H4 and H9-H10 where a) is the initial SAV1866 structure, b) is the result of 100 ns of MD for the ADP-bound structure and c) is the same result for the APO structure.

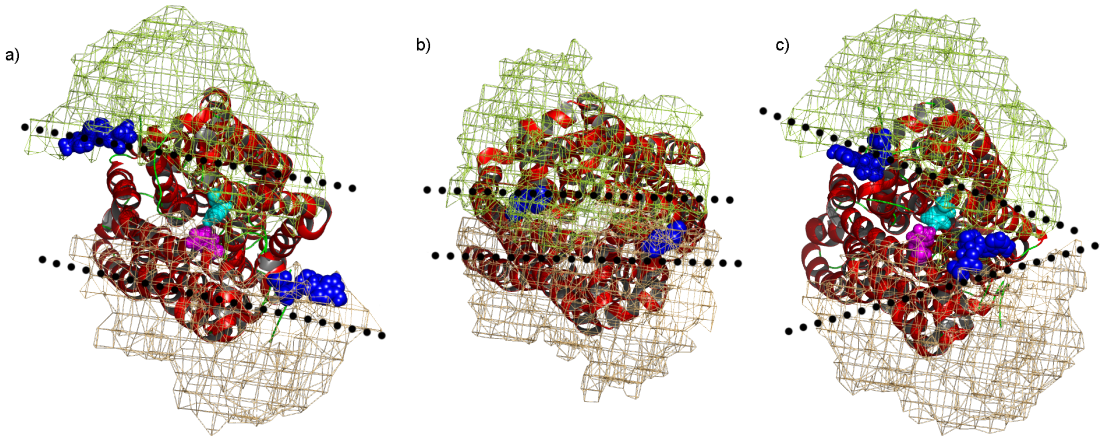

Figure 3.7 – View of the nucleotide binding domains (orange and green mesh) from a position perpendicular to the cytosolic side of the membrane after 20 ns of SMD for simulations for a parallel and a skewed NBD conformation, respectively : simulations a) SMD-ADP1 and c) SMD-ADP2 after 20 ns with ADP in dark blue and His204 of helix H4 (cyan) and H10 (magenta). Initial structure is presented in b). Dashed lines represent the approximate NBD-NBD surface interfaces.

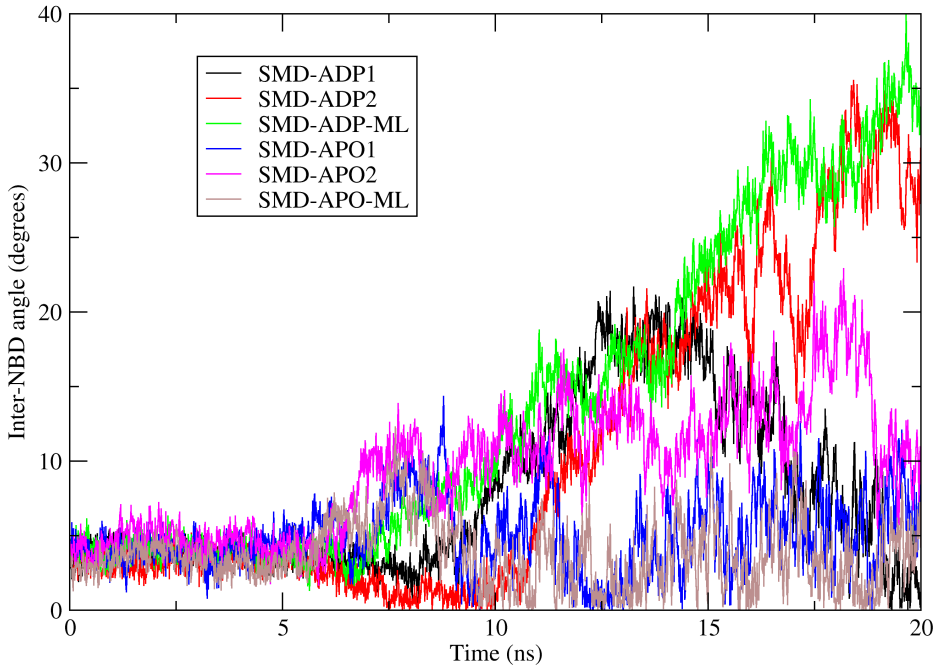

Figure 3.8 – Evolution of the angle between the contact planes of the two nucleotide binding domains as a function of the SMD simulation time.

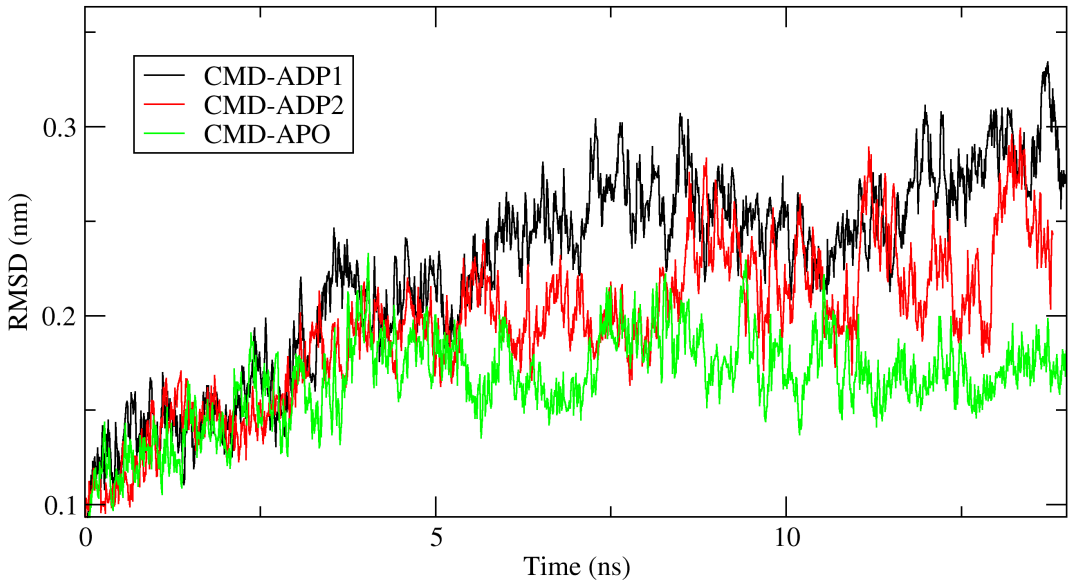

Figure 3.9 – Evolution of the RMSD during constant restrained MD simulation starting from the 20 ns time structure of SMD-ADP1, SMD-ADP2 and SMD-APO1 simulations.

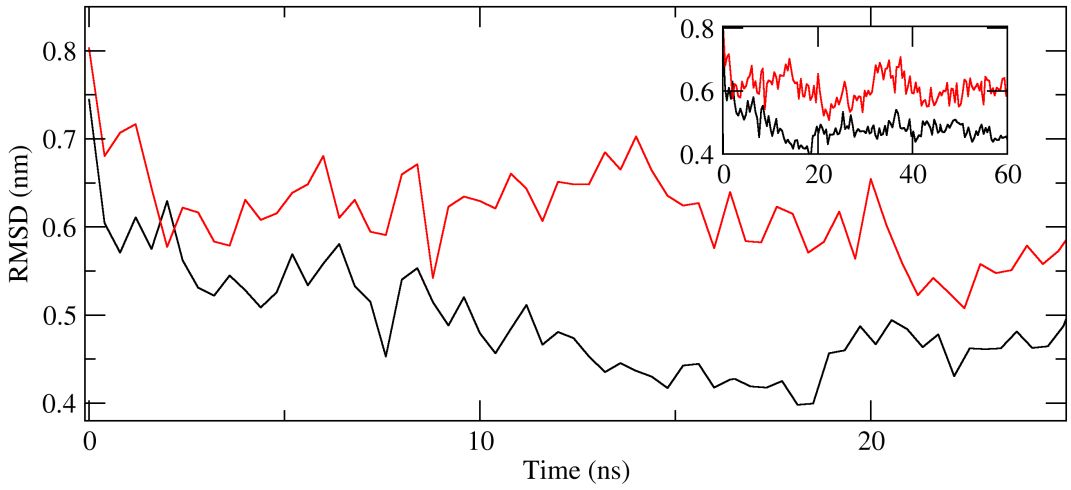

Figure 3.10 – RMSD evolution of the refolding simulation MD-ADP-RF (black) and MD-APO-RF (red) to the initial SMD conformations from SMD-ADP1 and SMD-APO1 respectively. Inset represents the whole 60 ns length of the refolding simulations.

CHAPITRE 4

APPROFONDISSENT DE L'ARTICLE SUR LA PROTÉINE SAV1866

L'approche utilisée pour échantillonner l'ouverture de la protéine SAV1866 avec la méthode de dynamique moléculaire dirigée (DMD) avait pour but de favoriser un échantillonnage le moins contraint possible de la trajectoire d'ouverture, limitant le biais introduit par des méthodes comme la dynamique moléculaire visée (DMV) (voir 1.2.2) dans laquelle des trajectoires minimisant plus rapidement le RMSD avec la structure ciblée sont favorisées. Dans le cas des structures de transporteurs ABC, les segments qui affichent le plus grand RMSD entre la structure ouverte et celle fermée sont les domaines liant les nucléotides (DLNs) dû à leur position à l'extrémité du levier créé par le mouvement de cisaillement des deux moitiés de la protéine. Ceci peut expliquer pourquoi dans les travaux de DMV examinant l'ouverture de MsbA, la perturbation des DLNs se fait ressentir très tôt alors que le réarrangement des hélices des domaines transmembranaires (DTMs) au niveau du feuillet externe de la membrane est observé en fin de simulation [269]. Des travaux récents non publiés du groupe d'Emad Tajkhorshid[1] sur l'ouverture de MsbA offrent une image à l'opposé de ceux de Weng *et al.* : Dans leurs essais comparant la DMV simple à d'autres combinaisons de modes de mouvement, la DMV nécessite la plus grande injection d'énergie pour échantillonner une ouverture de MsbA alors que les trajectoires dans lesquelles la fermeture de l'accès de l'extérieur de la membrane au DTM est observée en premier demandent moins d'énergie. Ceci laisse supposer que le fait que nous n'avons pu observer une séparation des DTMs lors des simulations de DMD dans lesquelles l'énergie de séparation est injectée dans les DLNs n'est pas en soit une invalidation de la méthode de DMD utilisée. Bien qu'il soit possible que les temps de simulations nécessaires à la DMD à un seul ressort pour échantillonner des trajectoires d'ouverture soient plus longs que les temps de calcul accordés à nos simulations, il est aussi possible que l'hypothèse de départ selon laquelle l'hydrolyse de l'adénosine triphosphate (ATP) dans les DLNs provoque tout d'abord une séparation de

[1] communication orale du 29 novembre 2011

ces domaines soit à réfuter. Aussi, ces résultats non publiés confèrent une plus grande importance à nos résultats de simulations de DM traditionnelles dans lesquelles nous voyons une fermeture de l'accès de l'extérieur de la membrane aux DTMs. Ces résultats ne semblent pas être corroborés par les résultats de simulations avec des nucléotides liés dans les DLNs des autres équipes de recherche.

Dans cette optique, il serait avantageux de pousser l'étude de l'ouverture de SAV1866 sur les systèmes résultants des nos simulations de DM à l'aide de simulations de DMD avec des points d'ancrage situés dans les DTMs au lieu des DLNs. Puisque nous avons déjà observée la fermeture de l'accès de l'extérieur de la membrane dans les DM, ce mode de mouvement n'a pas besoin d'être simulé par l'ajout d'une dimension et l'utilisation de méthodes d'échantillonnage de trajectoires multidimensionnelles. En utilisant différentes structures observées dans nos DM comme point de départ de DMD, il sera possible de déterminer si la fermeture de l'accès de l'extérieur de la membrane des DTMs n'est qu'un artefact de simulation ou si celle-ci facilite la séparation des DTMs. Notons toutefois que la protéine SAV1866 démontre une grande stabilité en simulation et une résilience aux changements conformationnels forcés. Tout projet de recherche tenant d'évaluer quantitativement les modes d'ouverture de cette protéine nécessitera probablement des temps de calcul imposants.

CHAPITRE 5

CALCUL D'ÉNERGIE LIBRE DE LIAISON DE ZPP À POP

Comme le nom l'indique, les méthodes de calcul de différence d'énergie libre permettent d'obtenir de façon précise la différence d'énergie libre entre deux états d'un système. Dans le cas de la méthode du potentiel de force moyenne (PFM), il est aussi possible de déterminer la nature du profil énergétique d'une réaction donnée, incluant l'énergie libre au point de transition. Bien qu'en théorie ces méthodes soient applicables à une transformation comme celle qu'entreprennent les transporteurs ABC pour passer d'une forme fermée à une forme ouverte, à notre connaissance, l'expérience n'a jamais été tentée sur un système aussi complexe. Les systèmes sur lesquelles le PFM est appliqué sont habituellement simples, tel que le transport d'un ion au travers d'un canal ionique ou la dissociation d'un ligand lié à la surface d'une protéine.

Les méthodes de dynamique moléculaire dirigée présentées au chapitre 1.2.1 ainsi que l'échantillonnage parapluie (EP) présenté en 5.2.2 peuvent toutes deux être utilisées pour calculer un PFM, mais d'abord nous désirons les tester sur un système de taille intermédiaire pour déterminer leur tractabilité. Notre choix s'est penché sur la protéine Prolyl oligopeptidase (POP) sur laquelle le groupe d'Alex Bunker en Finlande a acquis une certaine expérience de simulation [128, 129] et qui possède la caractéristique d'avoir un site de liaison de ses ligands entièrement isolé du solvant. Pour accommoder un ligand, la protéine doit obligatoirement passer par un changement de conformation. Or, le chemin que prennent les ligands pour accéder au site actif est à ce jour inconnu. Nous avons donc entrepris une étude de POP par le calcul du PFM à l'aide de la DMD et de l'EP pour déterminer la nature du chemin d'entrée des ligands et leur probabilité respective. Ce chapitre sert de brève introduction à la protéine POP, suivi d'une présentation des diverses méthodes de calcul de la différence d'énergie libre.

5.1 Prolyl oligoptidase

La famille des Prolyl oligopeptidases est un groupe d'endopeptidases capables de cliver des peptides après une proline interne. Chez les mammifères, POP est exprimée principalement au cerveau [101] et a pour substrat de nombreux neurpoeptides de moins de 30 a.a. résistants à la dégradation due à la présence de cette proline interne.

L'intérêt pharmacologique pour cette protéine est principalement dû au fait qu'il a été démontré que son inhibition peut réduire les déficits cognitifs dans des modèles animaux souffrant de symptômes de la maladie de Parkinson induit chimiquement [222]. Une liste des substrats potentiels de POP est disponible dans la revue de García-Horsman *et al.* [77]. Bien que plusieurs inhibiteurs de POP ont démontré leur potentiel *in vitro* et que certains se soient rendus aux tests chez les humains [175, 251], la localisation de POP dans le cerveau rend le développement de nouveaux inhibiteurs difficile puisque ceux-ci doivent pouvoir traverser la barrière hémato-encéphalique.

5.1.1 Structure et fonction

La raison principale de notre choix d'utiliser POP pour tester les méthodes de calcul d'énergie libre provient de sa structure. POP est composée d'un domaine en propulseur-β (DPβ) composé de 7 feuillets-β à 4 brins chacun et d'un domaine catalytique (DC) en repliement de type hydrolase α/β (figure 5.1 a) contenant la triade d'a.a. S554, D641 et H680 responsable du clivage des peptides. Plusieurs structures cristallographiques de POP ont été réalisées dont des mutants non-fonctionnels accompagnés de ligands [239–241] et des protéines actives et leurs inhibiteurs [71, 240] et leur traces est présentées en figure 5.4 (b). On y voit que la seule structure à haut RMSD avec les autres structures est POP de la bactérie *Myxococcus xanthus* [225](PDB : 2BKL) et la raison de cette déformation semble être due à l'emboîtement des copies de la protéine dans le réseau cristallin.

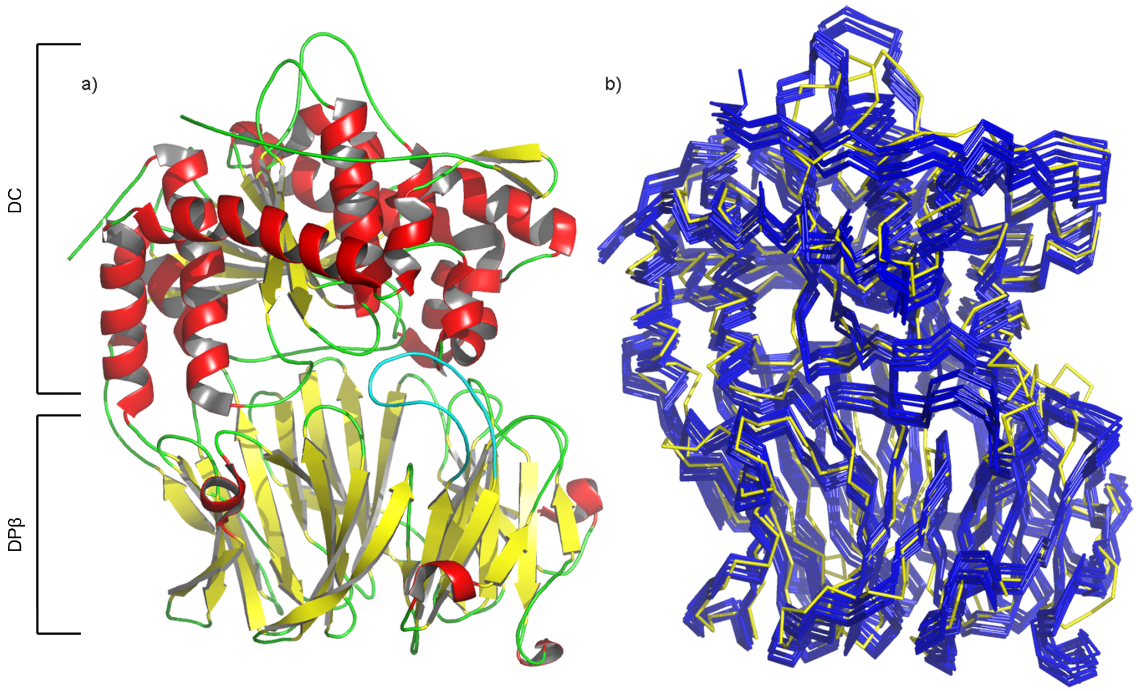

Figure 5.1 – Structure cristalline de POP avec a) PDB : 1QFS colorée par type de structure secondaire et b) la trace C-α superposée de plusieurs structures cristallographiques disponibles (PDB : 1H2W, 1H27, 1H2Z, 106Q, 1UOO, 1UOQ, 2EQ9 en bleu). En jaune, PDB : 2BKL

5.1.1.1 Site de liaison

La triade d'acides aminés S554, D641 et H680 localisée sur le DC se trouve à l'intérieur de la protéine et est entièrement coupée du solvant (figure 5.2 a). Plusieurs hypothèses ont donc été formulées concernant le chemin d'accès au site actif. Une hypothèse nous vient de la structure cristalline de *Sphingomonas capsulata* [225] dans laquelle les deux domaines de POP sont entièrement séparés. Bien que brièvement mentionné par Shan *et al.* , il repose un doute sur la raison de l'ouverture des deux domaines puisqu'il est mentionné que l'hélice-α en N-terminale est insérée dans l'interface inter-domaines des structures voisines du cristal. Cette insertion n'est pas visible dans la structure publiée, mais a été recréée en figure 5.2 (b). Cette même figure présente aussi la taille de la cavité intérieure du DPβ. Depuis, une autre structure d'un homologue bactérien de POP a été cristallisée sous sa forme ouverte (PDB : 3IUL) chez *Aeromonas punctata* [149].

Li *et al.* ont aussi réussi à cristalliser une forme fermée de la protéine (PDB : 3IVM) en ajoutant à la forme ouverte un inhibiteur, le Z-pro-prolinal (ZPP). De plus, ils ont démontré qu'en maintenant la structure dans sa forme ouverte, l'activité de la protéine était inhibée.

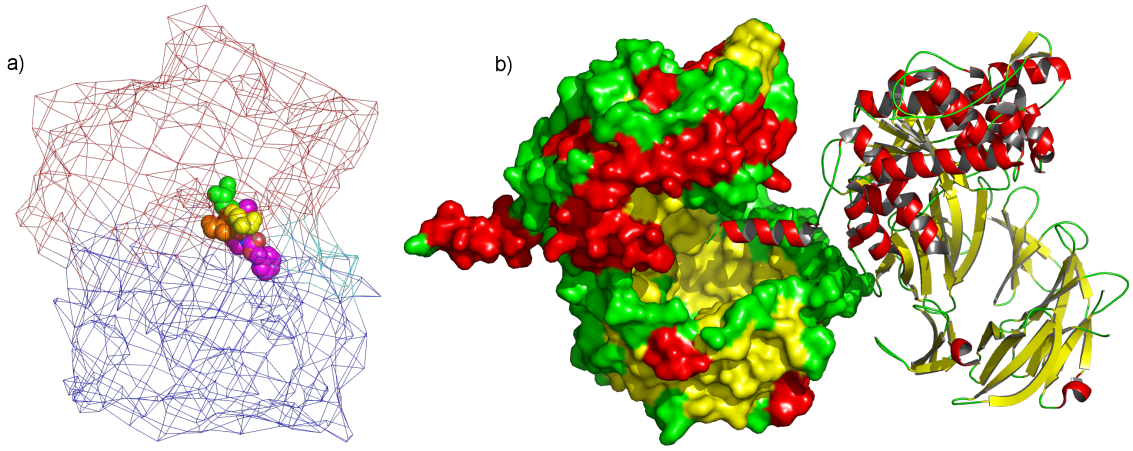

Figure 5.2 – En a), représentation en grillage de la structure externe des POP porcine (PDB : 1QFS) avec S554 (vert) lié de façon covalente à l'inhibiteur ZPP (magenta), D641 en jaune et H680 en orange. Le DC est peint en rouge, le DPβ en bleu, sauf la boucle flexible T190-N208 en cyan. À droite en b), exemple d'une structure de POP avec les domaines séparés provenant de *S. capsulata* (PDB : 1YR2).

L'hypothèse selon laquelle l'accès au site actif passe par une ouverture des deux domaines est renforcée par une étude de mutagénèse dirigée démontrant qu'il y a une forte réduction de l'activité lorsque les deux domaines sont liés de façon covalente entre l'a.a. muté T597C et l'a.a C255 [242]. La séparation complète des deux domaines n'est pas forcément la seule option puisque des cavités dans la zone inter-domaines ont été observées en dynamique moléculaire [73] et par microscopie électronique [244]. Les études de DM de POP de notre groupe ont échantillonné un événement susceptible d'être un chemin d'accès au site actif à l'interface des deux domaines [129]. On y voit une boucle flexible de 19 a.a. (en cyan dans la figure 5.2 a) située dans le DPβ et couvrant une partie de l'interface inter-domaines briser contact avec le DC pour s'étirer dans le solvant, puis se replier partiellement.

5.1.2 Inhibiteurs

Deux catégories d'inhibiteurs interagissent avec POP : les inhibiteurs ne formant pas de liaison covalente avec POP et ceux formant un lien de type hémiacétal entre le groupe aldéhyde (R-C(=O)H) du ligand et le groupe alcool (R-OH) de la sérine catalytique S554. Ces deux types d'inhibiteurs sont habituellement caractérisés par la présence de deux cycles semblables à deux prolines consécutives et leur mode d'inhibition est par compétition au site actif [142]. Le Z-pro-prolinal est un exemple typique des inhibiteurs formant un lien hémiacétal et est composé d'un cycle phényl en plus des deux prolines (voir Figure 7.1). Puisque les sites actifs sont hautement conservés entre les membres de la famille de POP, il serait avantageux de développer des inhibiteurs non-compétitifs ciblant de façon plus spécifique POP en inhibant le recrutement de ligand.

5.2 Calcul d'énergie libre

Plusieurs méthodes permettent de calculer la différence d'énergie libre entre deux états. L'approche directe consiste à échantillonner l'ensemble thermodynamique d'un système donné suffisamment pour que chaque état du système soit représenté proportionnellement à sa probabilité. Puisque la méthode n'est applicable qu'avec de très petits systèmes [99], nous ne la décrivons qu'en quelques mots : les états échantillonnés sont regroupés dans deux groupes d'état (par exemple, l'existence ou non d'un pont hydrogène spécifique entre deux molécules) et par comptage de la prévalence de chaque état on obtient deux concentrations de population ($[a_{pont-h}]$ et $[a_{libre}]$ respectivement). La constante d'association $K_{association}$ est obtenue à l'aide de ces concentrations :

$$K_{association} = \frac{1}{K_{dissociation}} = \frac{[a_{pont-h}]}{[a_{libre}]^2}, \tag{5.1}$$

où $K_{dissociation}$ est la constante de dissociation d'unité $1 mol/L$. On peut alors obtenir la différence d'énergie libre ΔG simplement par :

$$\Delta G = -k_{\mathrm{B}} T ln \frac{K_{dissociation}}{[1M]}, \tag{5.2}$$

où k_B est la constante de Boltzmann et T est la température absolue.

Il existe deux autres méthodes de calcul de différence d'énergie libre qui ne tiennent pas compte de la trajectoire empruntée par le ligand. La première, l'intégration thermodynamique (IT), consiste à muter un ligand hors d'un état donné et à le faire réapparaître dans un autre état. Dans sa forme la plus grossière, la différence d'énergie libre de chacune de ces deux étapes est donnée par :

$$\Delta G = \int_0^1 \langle \frac{dH(\lambda,x)}{d\lambda} \rangle_\lambda, \tag{5.3}$$

où $H(\lambda,x)$ est le Hamiltonien dont les termes sont atténués proportionnellement à la variable de couplage λ. Pour faciliter la convergence des résultats dans les cas complexes comme l'étude de la liaison d'un ligand à une protéine, les opérations de découplage et de recouplage sont divisées en plusieurs étapes consécutives. Deng et Roux définissent 5 étapes alchimiques menant à l'insertion du ligand dans la protéine et 3 étapes pour l'insertion du ligand dans le solvant [50]. Puisque le sens de l'intégration de l'équation 5.2 est sans importance, ces derniers préfèrent insérer le ligand plutôt que de le retirer. Ainsi, pour l'insertion du ligand, les contraintes de rotation et de translation sont tout d'abord ajoutées, puis les termes de répulsion des interactions de Lennard-Jones d'expression Ar^{-12}, suivis des mêmes termes dispersifs de Lennard-Jones d'expression $-Br^{-6}$, puis des termes électrostatiques et finalement le retrait des contraintes de translation et de rotation. Pour l'insertion dans le solvant, les étapes concernant les contraintes de translation et de rotation ne sont pas effectuées. La différence d'énergie libre totale est obtenue par la sommation des différences d'énergie libre de chacune de ces étapes.

Similairement, la méthode de perturbation de l'énergie libre (PEL) procède par mutation d'un système d'un état à un autre. Ici par contre, la mutation est faite vers un état similaire. Par exemple, on peut passer de l'état d'un ligand A lié à une protéine à l'état d'un autre ligand B, différent par un seul atome, lié à la même protéine. La différence d'énergie libre entre ces deux états est donnée par :

$$exp(-\frac{\Delta G(A_{lie} \rightarrow B_{lie})}{k_B T}) = \langle exp(-\frac{E_B - E_A}{k_B T}) \rangle_A, \tag{5.4}$$

où E_B et E_A sont les énergies des états B et A et les parenthèses triangulaires dénotent la moyenne d'ensemble provenant de la simulation A sur laquelle l'état B a été calqué. Pour obtenir une information d'un quelconque intérêt, on répète l'expérience pour le cas des ligands dans le solvant pour obtenir $\Delta G(A_{libre} \rightarrow B_{libre})$ et on soustrait les deux résultats :

$$\Delta\Delta G = \Delta G(A_{lie} \rightarrow B_{lie}) - \Delta G(A_{libre} \rightarrow B_{libre}) \tag{5.5}$$

ce qui permet d'obtenir la différence d'affinité entre deux ligands pour une protéine cible.

Ces deux méthodes ont l'avantage d'être relativement rapide et de pouvoir obtenir des résultats dont la précision est comparable à celle des méthodes expérimentales voir [32, 50, 261, 263]. Leur coût d'exécution est par contre encore trop élevé pour pouvoir être utilisé comme alternative aux méthodes d'amarrage de ligand pouvant cribler des bases de données de dizaines de milliers de ligands en peu de temps. Aussi, ces méthodes ne peuvent donner aucune information sur le chemin empruntés par le ligand pour atteindre son site de liaison. Leur utilité réside dans la capacité d'évaluer précisément la différence d'énergie libre entre deux systèmes dans le cas où la trajectoire empruntée est sans intérêt. Nous présentons donc maintenant deux méthodes permettant d'échantillonner les chemins possiblement emprunté par un ligand, de quantifier leur énergie libre de transition et de comparer leur probabilité respective.

5.2.1 DMD et équation de Jarzynski

La DMD a été introduite au chapitre 1.2.1 et nous élaborons ici son utilité pour caractériser quantitativement les trajectoires échantillonnées en calculant un PFM. Tout d'abord, il faut noter que les temps de simulations de trajectoires de DMD sont dans l'ordre de la dizaine de nanosecondes, les processus étudiés par DMD ne sont pas réversibles, i.e. l'entropie du système augmente au cours de la réaction. On peut ainsi obtenir le travail non-réversible en intégrant la force ressentie par le potentiel harmonique atta-

ché au ligand en déplacement du ligand. De l'équation 1.9, on dérive la force :

$$F_{dirigee}(X) = k(X - X_{dirigee}(t)), \tag{5.6}$$

où la position d'équilibre $X_{dirigee}$ évolue dans temps à vitesse v constante dans la direction $\overrightarrow{V_{dir}}$:

$$X_{dirigee}(t) = \overrightarrow{V_{dir}} \times vt + X_0, \tag{5.7}$$

et le travail W entre un état initial X_i et final X_f :

$$W(X_f - X_i) = W(\Delta X) = \int_{X_i}^{X_f} F(X)dX \tag{5.8}$$

Toutefois, Jarzynski a démontré qu'il était possible d'obtenir la différence d'énergie libre d'une transformation à partir d'un ensemble d'échantillons de la trajectoire de la transformation [107, 108]. L'équation de Jarzynski est définie comme :

$$exp(\frac{-\Delta G(\Delta X)}{k_B T}) = \langle exp(\frac{-W(\Delta X)}{k_B T}) \rangle \tag{5.9}$$

La moyenne de l'exponentielle des énergies de travail W fait que de rares événements de basse énergie dominent l'estimé de ΔG. Ces événements situés dans la queue de la distribution gaussienne des énergies de travail observées peuvent être difficiles à échantillonner. Un estimateur basé sur le premier et deuxième terme de l'expansion du cumulant de 5.9 a été proposé [195] :

$$\Delta G = \langle W \rangle - \frac{1}{2k_B T}(\langle W^2 \rangle - \langle W \rangle^2) + ... \tag{5.10}$$

Les deux premiers termes de cette équation peuvent être réécrits :

$$\Delta G = \langle W \rangle - \frac{Var(W)}{2k_B T} \tag{5.11}$$

L'expansion du cumulant permet donc d'estimer ΔG à partir de simulations échantillonnant la valeur moyenne de W et non les valeurs extrêmes qui domine l'équation

5.9. Cependant, l'équation 5.11 n'est un bon estimateur de ΔG que lorsque les énergies de travail échantillonnées suivent une distribution gaussienne, ce qui est le cas pour des constantes de ressort k suffisamment élevées [195].

5.2.1.1 Paramétrisation dans la littérature

La DMD au calcul de PFM a été appliquée à plusieurs reprises sur des systèmes simples. Kosztin *et al.* ont étudié trois trajectoires de sortie de l'acide rétionique (50 atomes) de son récepteur (230 a.a). Pour extirper le ligand caché à l'intérieur de la protéine, une constante de ressort k=1.6 MJ/(mol·nm^2) et une vitesse de tir de $3.2 nm/ns$ ont été choisies. Leurs simulations ont démontré peu de sensibilité à la vitesse de tir alors qu'un essai à une vitesse de $8.0 nm/ns$ induit le même type de déformation qu'observé à une vitesse de tir plus basse.

La méthode a aussi été employée pour étudier le dépliement de molécules d'ARN [157] et de protéines [274]. Dans ce dernier, Xiong *et al.* ont étudié l'effet de la vitesse de tir et du nombre de simulations pour obtenir un résultat à faible erreur comparativement à une simulation dans laquelle un peptide de poly-alanine est déplié à un taux très lent de $0.01 nm/ns$. Chaque groupe de simulation à une vitesse de tir donné a été conçu pour consommer la même quantité de temps CPU, i.e. le temps d'exécution d'une simulation de tir multiplié par le nombre de répétitions est identique pour les 3 groupes de vitesse de tir. Ils ont ainsi démontré que l'efficacité maximale était atteinte par la combinaison de vitesses de tir lentes ($0.1 nm/ns$ dans leur cas) combiné à un faible nombre de simulations (20).

Plus récemment, Baştuğ *et al.* ont comparé le transport d'un ion de K^+ au travers d'un nanotube et de la protéine Gramicidine A [9, 10]. En utilisant une constante de potentiel harmonique de 8.4 MJ/(mol·nm^2) et des vitesses de tir variant entre 1.00 nm/ns et 0.25 nm/ns, ils ont obtenu une erreur faible pour les essais sur nanotube comparativement à leurs résultats obtenus par échantillonnage parapluie. Cependant, les chemins au travers de la Gramicidine A, une protéine de 16 a.a., se sont avérés plus sensibles aux choix de paramètres et sont incohérents avec les simulations par EP deux fois moins coûteuses en temps de processeur, ce qui démontre les difficultés que peuvent rencontrer

les simulations sur des systèmes complexes.

L'expérience sur la Gramicidine A a tout de même été réessayée dernièrement par le groupe de Giorgino *et al.* et a démontré que par l'usage d'un grand nombre de simulations (1000) à des vitesses relativement rapides de 1.00 nm/ns, il était possible d'obtenir des résultats cohérents avec une erreur d'au plus 4.18 kJ/mol.

5.2.1.2 Méthode de tir

La méthode de tir présentée en section 1.2.2.1 basée sur la distance Euclidienne entre les deux points d'ancrage dans les DLNs était bien adaptée à la séparation de deux parties d'une protéine retenues entre elles par les DTMs et la membrane lipidique. Cependant, dans le cas d'un petit ligand se séparant complètement d'une protéine, il faut utiliser une méthode de tir assurant que le ligand quitte par un chemin que l'on désire échantillonner. Pour se faire, l'orientation de la protéine est fixée à l'aide d'un potentiel harmonique sur des C-α de l'ossature. Aussi, le potentiel harmonique imposé sur le ligand pour le tirer du site de liaison est défini par un vecteur de direction superposable à la trajectoire échantillonnée. Pour s'assurer que le ligand puisse s'adapter à toute courbure de la trajectoire de tir, le potentiel harmonique n'est appliqué que dans la direction du vecteur de tir tout en laissant libre tout mouvement dans le plan perpendiculaire à ce vecteur.

5.2.1.3 Direction du tir

Les 3 directions de tir tentées dans l'article du chapitre 7 sont ici présentées en figure 5.3. Le choix de ces trois directions est basé sur les hypothèses expérimentales présentées en section 5.1.1.1

5.2.2 Échantillonnage parapluie

L'échantillonnage parapluie (EP) est une méthode qui permet de briser une trajectoire comme celle examinée par DMD en plusieurs fenêtres étudiées séparément [247]. La différence principale ente l'EP et la DMD est que dans le premier, les positions d'équilibre

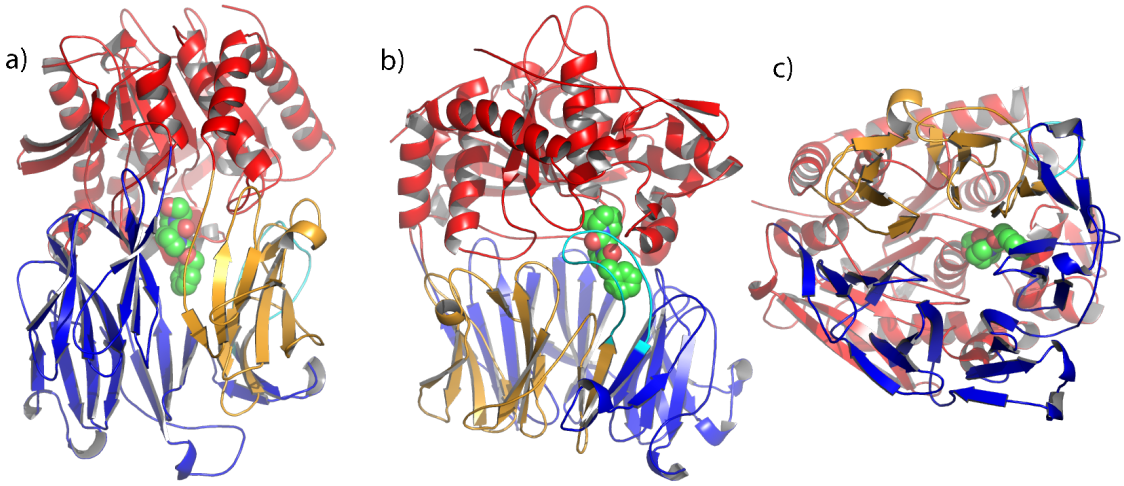

Figure 5.3 – Structure cristalline de POP (PDB : 1QFS) présentée sous les orientations utilisées dans les DMD avec vue de face de la direction par laquelle le ZPP est tiré hors de la protéine : a) entre le premier et septième feuillet-β du domaine en propulseur-β, b) par l'espace inter-domaines libéré lors des mouvements de la boucle flexible T190-N208, et c) par le centre du domaine en propulseur-β. Pour faciliter la compréhension, le domaine catalytique est peint en rouge (M1-D72 et K428-P710), la partie du propulseur-β avant la boucle flexible en orange (T73-A189), la boucle flexible en cyan et le reste du propulseur-β en bleu (K209-V427).

du potentiel harmonique sont fixées pour chaque fenêtre alors qu'en DMD elles évoluent suivant l'équation 5.7. De ce fait, le ligand n'est jamais tiré dans une fenêtre et la structure du système a le temps de converger vers un état à l'équilibre.

Chaque simulation évolue donc sous un potentiel biaisé forçant le système à échantillonner une partie de la coordonnée de réaction. En supposant qu'on utilise un potentiel harmonique biaisant une simulation autour d'une position d'équilibre ε_i dans une fenêtre i donnée, on obtient une équation du potentiel biaisé de la fenêtre :

$$E^i_{biais}(X) = E(X) + \frac{k}{2}(\varepsilon(X) - \varepsilon_i)^2 = E(X) + w_i(\varepsilon), \tag{5.12}$$

où l'opérateur $\varepsilon(X)$ est la projection du système sur la coordonnée de réaction. En se référant à Kästner [126], la distribution biaisée est alors décrite par :

$$P^i_{biais}(\varepsilon) = exp(\frac{-w_i(\varepsilon)}{k_B T}) \frac{\int exp(-\frac{E(X)\delta(\varepsilon(X)-\varepsilon_i)}{k_B T})dX}{\int exp(-\frac{E(X)+w_i(\varepsilon)}{k_B T})dX}, \tag{5.13}$$

où δ est une fonction de Dirac. Le potentiel de force moyenne pour chaque fenêtre est obtenu par :

$$G^i_{biais}(\varepsilon) = \frac{-ln(P^i_{biais}(X_\varepsilon))}{k_B T} - w_i(\varepsilon + K_i), \tag{5.14}$$

et la distribution sans biais de chaque fenêtre P^i :

$$P^i(\varepsilon) = P^i_{biais} exp(\frac{-w_i(\varepsilon)}{k_B T}) \langle exp(\frac{-w_i(\varepsilon)}{k_B T}) \rangle \tag{5.15}$$

Le choix de l'utilisation d'un potentiel harmonique pour biaiser l'échantillonnage de chaque fenêtre est arbitraire. Un potentiel de biais idéal serait $-G$, la valeur à estimer, puisque ceci forcerait un échantillonnage uniforme de la coordonnée de réaction.

5.2.2.1 Weighted histogram analysis method

Deux méthodes existent pour obtenir un potentiel de force moyenne global à partir des potentiels de force de chaque fenêtre $G^i_{biais}(\varepsilon)$. La méthode d'analyse des histogrammes par assignation de poids ou *Weighted histogram analysis method* (WHAM) [136] tente par une méthode itérative de déterminer les constantes K_i de l'équation 5.14. Pour ce faire, des poids p_i sont assignés aux distributions de chaque fenêtre pour trouver la distribution globale :

$$P(\varepsilon) = \sum_i p_i \times P^i(\varepsilon) \tag{5.16}$$

La valeur des poids p_i et des constantes K_i est alors résolue itérativement afin de diminuer l'erreur de la distribution $P(\varepsilon)$ à l'aide des fonctions :

$$p_i = N_i \frac{exp(\frac{-w_i(\varepsilon)+K_i}{k_B T})}{\sum_j (N_j \times exp(\frac{-w_j(\varepsilon)+K_j}{k_B T}))}, \tag{5.17}$$

et :

$$exp(\frac{-F_i}{k_B T}) = \int P(\varepsilon) exp(\frac{-w_i(\varepsilon)}{k_B T}) d\varepsilon, \tag{5.18}$$

où N_i et N_j sont les nombres d'échantillons pour les fenêtre i et j.

Une autre méthode permet d'obtenir un PFM à l'aide de simulations est l'intégration parapluie [127], mais n'est pas couverte dans ce travail puisqu'elle donne des résultats similaires à WHAM [126]. Notre choix d'utiliser la méthode WHAM repose encore une fois sur le fait que l'outil était disponible pour nos données au moment de notre analyse.

L'estimation de l'incertitude sur le profile d'énergie libre obtenue à l'aide de WHAM peut être calculée de deux façons à l'aide du rééchantillonnage Bootstrap [96] avec lequel des sous-ensembles des données sont utilisés pour générer des profile d'énergie libre, ce qui permet d'établir la robustesse du jeu de données. Si le nombre de fenêtres d'échantillonnage permet une couverture suffisante de chaque point de la trajectoire (au moins 10 fenêtres recouvrant chaque point), chaque fenêtre peut être considérée comme un point indépendant qui peut être échantillonné par l'algorithme de Bootstrap lors de la génération des nouveaux profils d'énergie libre. Si par contre le nombre de fenêtres est insuffisant pour assurer un recouvrement multiple de la coordonnée de réaction, l'algorithme de Bootstrap peut être utilisé peut être configuré pour rééchantillonner une partie des données de chaque fenêtre.

5.2.2.2 Couverture des fenêtres et constante de force

La méthode WHAM nécessite un recouvrement partiel de l'espace échantillonné entre chaque fenêtre et ses fenêtres voisines. On assure ce recouvrement en définissant des distances inter-fenêtres suffisamment courtes et des constantes de force du potentiel harmonique k suffisamment faibles. L'un des avantages de la méthode d'EP comparativement à la DMD est qu'il est possible de juger facilement quels endroits de l'espace de la coordonnée de réaction n'a pas bien été échantillonné et d'ajuster l'échantillonnage en ajoutant de nouvelles fenêtres avec des constantes de ressort plus élevées. Le choix des constantes de force est aussi dépendant de la taille et motilité du ligand déplacé par

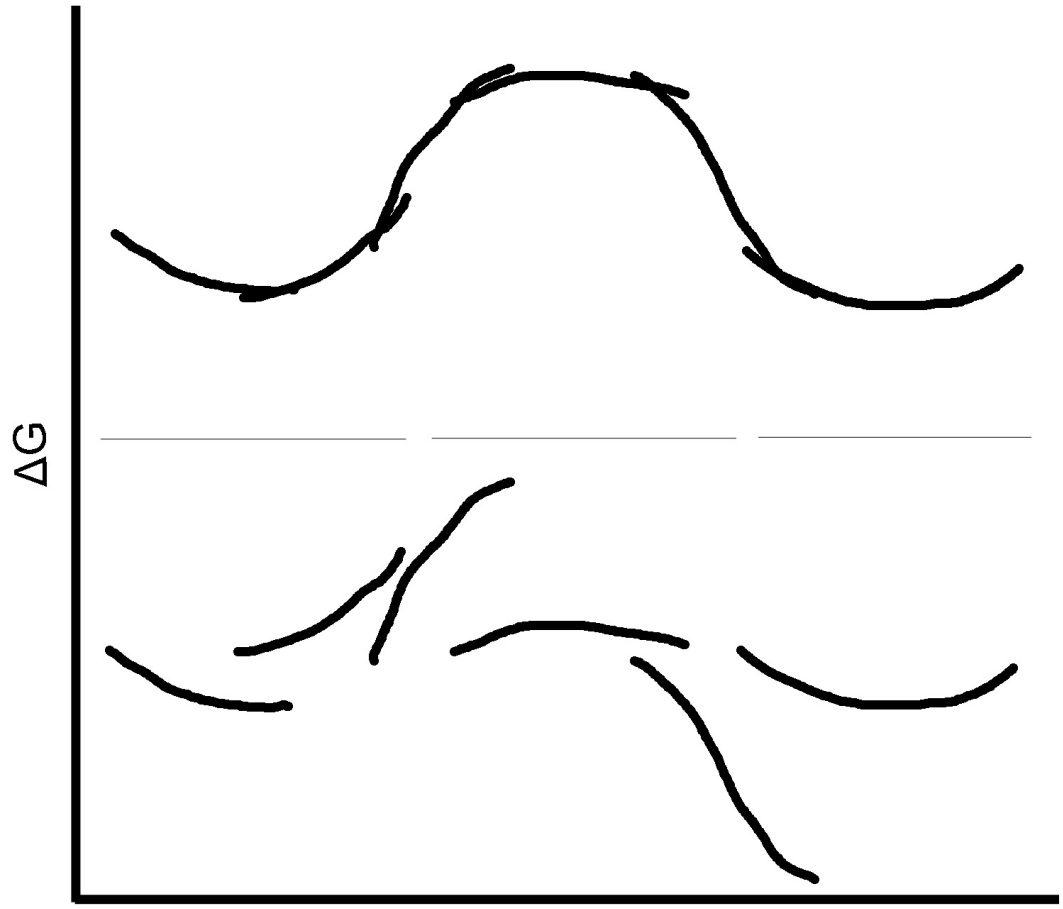

Figure 5.4 – Exemple d'assemblage des PFM au biais retiré obtenus par un EP à 6 fenêtres (bas) en un PFM global de la coordonnées de réaction.

EP. Dans les travaux de transport d'ions, des constantes de force k élevées ont été utilisées avec succès, variant entre 8.36 MJ/(mol·nm^2) et 16.72 MJ/(mol·nm^2) [30, 178] avec des espacements inter-fenêtre plus courts entre 0.02 nm et 0.05 nm. Pour les systèmes plus complexes tels que ceux impliquant l'insertion d'un lipide [273] ou d'un peptide [276] dans une membrane, on choisira des valeurs de k plus faibles de l'ordre de 1.00 MJ/(mol·nm^2) permettant un échantillonnage plus large de la coordonnée de réaction dans chaque fenêtre espacée de 0.05 nm.

5.3 Conclusion

Dans le chapitre 7 qui suit, nous proposons d'étudier les chemins de sortie du ligand ZPP de la protéine POP porcine (PDB : 1QFS). Une première évaluation de trois chemins de sortie probables sont tout d'abord échantillonnés à l'aide de simulations de DMD, puis les deux chemins les plus probables sont étudiés par EP afin de déterminer la différence d'énergie libre de la liaison du ZPP à POP et de l'état de transition de ce chemin.

CHAPITRE 6

CONTRIBUTION DES AUTEURS À L'ARTICLE SUR PROLYL OLIGOPEPTIDASE

- Jean-François St-Pierre a écrit la première version du manuscrit et a effectué les simulations et les analyses présentées, incluant :
  - Établissement du protocole de dynamique moléculaire dirigée (DMD),
  - Exécution des simulations et analyse des DMD pour les 3 trajectoires d'intérêt,
  - Paramétrisation du protocole d'échantillonnage parapluie (EP),
  - Établissement du protocole de dynamique moléculaire dirigée (DMD),
  - Exécution des simulations et analyse de simulations d'EP,
  - Présentation de l'analyse en discussion et des conclusions
- Alex Bunker et Normand Mousseau ont supervisé le travail de DMD et d'EP.
- Tous les auteurs ont contribué aux révisions et aux corrections du manuscrit.

CHAPITRE 7

ARTICLE : USE OF UMBRELLA SAMPLING TO CALCULATE THE ENTRANCE/EXIT PATHWAY FOR Z-PRO-PROLINAL INHIBITOR IN PROLYL OLIGOPEPTIDASE

Jean-François St-Pierre
Département de Physique and Regroupement Québécois sur les Matériaux de Pointe, Université de Montréal, C.P. 6128, succursale centre-ville, Montréal (Québec) Canada H3C 3J7
Mikko Karttunen
Department of Applied Mathematics, The University of Western Ontario, 1151 Richmond Street North, London (Ontario), Canada N6A 5B7
Normand Mousseau
Département de Physique and Regroupement Québécois sur les Matériaux de Pointe, Université de Montréal, C.P. 6128, succursale centre-ville, Montréal (Québec) Canada H3C 3J7
Tomasz Róg
Department of Physics, Tampere University of Technology, PO Box 692, FI-33101 Tampere, Finland.
Alex Bunker
Centre for Drug Research, Faculty of Pharmacy, University of Helsinki, PO Box 56, FI-00014, University of Helsinki, Finland

7.1 Abstract

Prolyl Oligopeptidase (POP), a member of the Prolyl Endopeptidase family, is known to play a role in several neurological disorders. Its primary function is to cleave a wide range of small oligopeptides, including neuroactive peptides. We have used force biased molecular dynamics simulation to study the binding mechanism of POP. We examined three possible binding pathways using Steered Molecular Dynamics (SMD) and Umbrella Sampling (US) on a crystal structure of porcine POP with bound Z-pro-prolinal (ZPP). Using SMD, an exit pathway between the first and seventh blade of the β-propeller domain of POP was found to be a non-viable route. US on binding pathways through the β-propeller tunnel and the TYR190-GLN208 flexible loop at the interface between both POP domains allowed us to isolate the flexible loop pathway as the most probable. Further analysis of that pathway suggests a long-range covariation of the inter-domain H-bond network which indicates the possibility of large-scale domain reorientation observed in bacterial homologues and hypothesized to also occur in human POP.

7.2 Introduction

Endopeptidases are a class of proteases that hydrolyse internal, i.e., non-terminal, peptide bonds. Prolyl Oligopeptidase (POP) is a proline-specific endopeptidase that cleaves oligopeptides ($<$ 30-mer) at the C-side of an internal proline. *In vitro* analysis found that a wide variety of neuroactive peptides substrates can be cleaved by POP [18, 77, 202, 207, 237] and *in vivo* analysis indicates that these are its actual metabolic substrates [77]. It has been found, although somewhat inconsistently, that certain POP inhibitors can reverse memory loss caused by amnesic agents, neurological disorders, and aging, making POP an important drug target [165, 278]. In addition, an alteration in POP enzyme activity has been measured in serum samples taken from patients suffering from Parkinson's and Alzheimer's disease [166], and multiple sclerosis [245]. Experimental evidence exists that POP might have a role beyond its peptidase function. Examples of this include protein-protein interactions [53, 77], intracellular transport [223], inflammation [75, 192], angiogenesis [158] and cancer development [158, 253]. The biochemical role of POP and its inhibitors have been reviewed in Refs. [18, 77, 142, 165, 202, 207, 237].

Several fundamental questions regarding access to the active site, gating, selection and the detailed inhibition mechanisms remain unanswered. Recent studies have combined a number of experimental and simulation techniques to address the details of inhibition mechanisms [141] as well as binding and gating mechanisms [244] and progress has been made in developing combinatorial libraries for POP. [37].

In this paper, our focus is on the active site access (exit) pathways. This is a fundamental question in understanding the peptidase function. What makes this question particularly challenging is its dynamic nature : entry and exit involve dynamic response. Hence, docking studies and short time-scale density functional calculations are not able to address this question. We study porcine POP, crystallized with bound ZPP inhibitor by Fülöp *et al.* [71] (PDB database ID 1QFS). The structure is composed of two domains : the protease catalytic domain with an α-β hydrolase fold composed of amino acids 1-71 and 436-710, and the seven-bladed β-propeller, comprised of amino acids 72-435. The

active site is on the surface of an internal cavity between the two protein domains. The ZPP inhibitor has a hydrophobic head that sterically blocks the active site and an aldehyde tail that forms a reversible covalent hemiacetal bond with the alcohol moiety of SER554 residue of the catalytic domain. The β-propeller domain has an unusual, mostly hydrophobic interaction between the first and seventh blade. This is called the "velcro rip" and has been proposed to act as a filter to the active site [70–72, 243]. Later studies have demonstrated that the domain by itself is more stable than in conjunction with the catalytic domain [118]. This finding is in agreement with a small scale computational study carried out by Fuxreiter *et al.* [73] and our previous computational work on POP with unbound inhibitor in the binding cavity [128] : both indicate that the β-propeller is a highly stable structure. It was suggested that the entry point is most likely through the H-bonded network of loosely structured loops that connect the two protein domains, in particular the location of the TYR190-GLN208 flexible loop.

Recent electron microscopy experiments suggest that the inter-domain region offers a wider entry-point than the β-propeller tunnel [244]. For example, Shan and collaborators find that the crystal structure of a distant POP relative, *Sphingomonas capsulata* (SC) (PDB database ID 1YR2), displays an open conformation in absence of the inhibitor where the two domains are separated, exposing the catalytic triad to the solvent [225]. Experimental evidence of an open form of a homologous protein have also been captured by x-ray diffraction crystallography of *Aeromonus punctata* prolyl endopeptidase [149]. In that study, Li *et al.* captured the closed form of the protein after soaking the open form of the crystalized protein in a bath of inhibitor. By maintaining the protein in this open form using glutaraldehyde driven lysine cross-linking, they obtained a complete absence of activity. This inter-domain large-scale motion has not yet been observed in experiments performed on mammalian POP. Cysteine cross-linking experiments binding the two domains together have, however, been shown to lead to strongly reduced protein activity [242].

Elucidating the mechanism through which substrates gain access to the active site, i.e. identifying the ligand entry pathway, would be beneficial for the development of new classes of inhibitors for mammalian POP. It is important to perform full MD simulations

with explicit solvent since water can access the active site and in some cases even have a decisive role [129]. To study the entry pathways, it is necessary to test the various possible trajectories directly. In this paper, we use SMD and US to measure the free energies of the three postulated exit/entry pathways of the ZPP inhibitor. Once we have determined the correct entry/exit pathway, we then study the interaction between inhibitor and the elements of the protein structure that comes in contact with it. Although SMD is first used to generate rough pathways, the bulk of our results are obtained using US. Using these methods we determine that the most probable exit pathway is through the loosely structured loops between the two domains opposite the inter-domain hinge.

7.3 Methods

7.3.1 Software, model, and simulation parameters

All simulations were performed using the GROMACS 4.0 simulation package [87] at constant pressure (1 bar) and temperature (310 K) (NPT). Temperature was maintained using the Nosé-Hoover thermostat [95, 180] and pressure using the Parrinello-Rahman barostat [196]. The coupling time constants were set to 0.1 and 1.0 ps for thermostat and barostat, respectively, and the protein and solvent were thermalized separately. Electrostatic interactions were computed using the Particle-Mesh-Ewald method (PME) [43, 125]. The Lennard-Jones interactions were cut off at 1.0 nm. The same cutoff was used for the real-space part of PME. Charge groups were chosen to be small to avoid artifacts that may arise if the charge groups are spatially too large [270].

To parameterize the POP and ZPP molecules and the solution ions, we used the OPLS-AA (Optimized Parameters for Liquid Simulations, AA stands for all-atom) potential set [115]. Partial charges on ZPP were taken from our previous work [128]. For water, the TIP3P model was used [116].

The initial structure was taken from our previous 100 ns MD study of POP with the ZPP inhibitor unbound in the active site [128]. The default physiological pH was also used to determine the protonation state of all amino acids. POP was solvated in a box of water of size $10 \times 10 \times 10$ nm. Potassium and chlorine ions were added to neutralize

the system to model physiological conditions (140 mM salt concentration). The solvated simulation box contained a total of 100,468 atoms.

Analysis and visualization were performed using the VMD (Visual Molecular Dynamics) package [98] and GROMACS [87, 153] analysis tools. Pathways were generated and evaluated using SMD and US, as described in the following sections.

7.3.2 SMD and US

In SMD, an external force, called force bias, $\vec{F}_{\mathrm{fb}}$ is applied to a single atom or a group of chosen atoms, through their center of mass. SMD is an irreversible approach and its use is based on the Jarzynski equality :

$$\langle \exp[-W/k_{\mathrm{B}}T] \rangle = \exp[-\Delta G/k_{\mathrm{B}}T], \tag{7.1}$$

where k_{B} is the Boltzmann constant, T is temperature and W is the total non-reversible work done on the system by $\vec{F}_{\mathrm{fb}}$ during a non-equilibrium transition between two states connected by a reaction coordinate λ [107, 108, 195]. The free energy difference between these two states is given by $\Delta G \equiv \Delta G(\lambda)$. The angular brackets stand for an ensemble average taken by repeating the simulation many times along the path connecting the initial and final states, λ_1 and λ_2. The essence of the Jarzynski equality is that it links rigorously the work done in a non-equilibrium process to the change in equilibrium free energy difference. To do so, a large number of simulations are performed in which a biasing potential is used to force a conformationnal change. The speed at which the transformation is done does not need be so small as to generate reversible work. It was shown that the exponential of the free energy difference of the tranformation is given by the average of the exponential of the work sampled in each individual trajectory using Equation 7.1 This method has been used in small systems, e.g., single molecule conformational changes [236, 274, 280] and, more rarely, in larger systems involving, e.g, protein-protein interactions [41].

We define the direction of $\vec{F}_{\mathrm{fb}}$ as $\hat{r}_{\mathrm{fb}}$. All other degrees of freedom are allowed to

react freely to this force. To drive the system, we apply a harmonic force

$$\vec{F}_{\mathrm{fb}}(\lambda) = k_{\mathrm{fb}}(\lambda \vec{r}_{\mathrm{fb}} - \vec{r}_{\mathrm{cm}}), \tag{7.2}$$

where k_{fb} is the force constant and $\vec{r}_{\mathrm{cm}}$ is the centre of mass. A force bias with a fixed value of k_{fb} is introduced with λ increasing from zero to one at a continuous rate as the simulation proceeds. This rate is known as the *pulling rate*.

Error estimates for Jarzinski equality are only well defined for pulling trajectories at near-equilibrium regime [80] and at this regime, it has been shown by Park and Schulten [195] that the Jarzynski equality is equivalent to calculating the free energy difference from the first and the second-order cumulants of the work done by the biasing force :

$$\Delta G = \langle W \rangle - \frac{\langle W^2 \rangle - \langle W \rangle^2}{2k_{\mathrm{B}}T} = \langle W \rangle - \overline{W_{dis}}, \tag{7.3}$$

where the second term is the dissipative work $\overline{W_{dis}}$ which is a function of the variance of the work. While both the Jarzynski equality and Equation 7.3 are formally correct in the thermodynamic limit, finite sampling leads to potential problems since ΔG depends exponentially on W, the result is easily dominated by the extremal values of the distribution [9, 182]. The impact of this sensitivity to rare pathways can be evaluated by comparing results obtained from these two methods.

US [212, 248] obtains the free-energy difference between two states from a set of equilibrated simulations. Like in SMD, a force bias is applied. However, now the configuration is equilibrated at each step, which we refer to as window. The value of the harmonic force constant used for each window is independent and can be set to a value that optimizes the efficiency with which the phase space is sampled. Parameters k_i and λ must be selected in such a way as to ensure that the phase space sampled by adjacent windows overlaps sufficiently, forming a continuous pathway between the initial and the final state. Results from all windows can then be combined using the weighted histogram analysis method (WHAM) [136] to provide the full thermodynamical evolution along the reaction coordinate.

Examples of systems where US has been successfully applied to include ion channels [30, 178], unfolding of the I27 titin domain [173] and the evaluation lipid transfer and peptide penetration in cellular membranes [273, 276]. SMD and US have been compared for ligand binding in the gramicidin A channel and Kv11.1 (also known as hERG) potassium channel [10]. Several comprehensive reviews are available [209, 281].

7.3.3 Application of SMD and US to the study of our system

SMD employing the Jarzynski equality was performed on the system with the applied bias force dislocating the ZPP molecule from the binding pocket along the three proposed pathways shown in Figure 7.1a. We generated one pull vector each for the pathways through the β-propeller tunnel and the flexible loop. For the third pathway, the velcro-rip junction between the first and seventh blade of the β-propeller, three different pull vectors (see Figure 7.1a) were attempted to test this suggested exit pathway [70–72, 243].

Since the center of mass and the orientation of the protein are not inherently conserved properties for the simulated system (protein + ZPP + solvent), it was necessary to add restraints to conserve the position and orientation of the protein. This was achieved by restraining the positions of a small number of α-carbon atoms positioned far from the sampled ZPP exit pathway with harmonic restoring forces with a force constant of 10 kJ/(mol $\times$ nm^2). Two sets of weak restraints were selected, for the flexible loop exit and β-propeller tunnel pathways, and in both cases the harmonic force constant on these restraints was two to three orders of magnitude smaller than the restraining force bias applied to the ZPP ligand. For the β-propeller tunnel pathway the restrained residues PRO7-ASP35, ASP431-GLY464, TYR510-LYS546 and ILE610-GLN629 were all located on the catalytic domain and did not form part of the inner cavity. For the flexible loop exit, the atoms selected for restraint span both domains : on the catalytic domain, the alpha-carbon atoms of the residues VAL427-LYS458 were restrained and on the beta-propeller, they were placed on amino acids TYR73-GLY108 and GLY288-LYS458. Although, formally, this position on both domains affect allosteric communication between the two domains, the restraints are weak enough and dispersed enough to minimize this

possibility. With the orientation of the protein maintained, the force bias is relative to the center of mass of the protein and the expression for the force bias vector becomes

$$\vec{F}_{\mathrm{fb}} = k_{\mathrm{fb}}(\lambda \vec{r}_{\mathrm{fb}} - (\vec{r}_{\mathrm{cm-ZPP}} - \vec{r}_{\mathrm{cm-protein}})). \tag{7.4}$$

Following the work of Tskhovrebova *et al.* [250], we first computed the free-energy barriers for the different exit pathways using SMD. This was performed with a harmonic force constant $k_{\mathrm{fb}} = 5$ MJ/(mol $\times$ nm^2) and pulling rates of 0.5 nm/ns as was used in previous work [159], and also of 0.1 nm/ns. To make it possible for the ZPP to exit POP, the simulation box was extended by 2.0 nm in the direction of the force bias for the β-propeller tunnel pulling vector which brings the total number of atoms in the periodic box to 130,004.

For our US, we used states obtained from the SMD as starting configurations. For each window, we selected a state where the value of the component of the displacement vector (the vector connecting the centers of mass of the POP and ZPP along the direction of the force bias) matched the value of λ for the window.

Only the flexible loop and the β-propeller exit pathways were studied with US. The pathways were divided into 46 and 48 windows respectively with the reaction coordinate values separated by 0.1 nm. In each window $k_{\mathrm{fb}} = 2.5$ MJ/(mol $\times$ nm^2) and force biased MD was performed for 10 ns and $8-9$ ns respectively. These initial simulations were started using conformations obtained from the SMD as shown in Tables 7.I and 7.II. To ensure sufficient overlap between distributions sampled in adjacent windows, undersampled regions were identified and simulations were launched in these windows.

For the loop opening, 14 extension windows were added in regions where the resistance to sampling was particularly high (energy barrier regions), the sampling was extended by 5-8 ns performed with a higher force constant of $5-10$ MJ/(mol $\times$ nm^2) to obtain a minimum sampling of $50,000$ conformations per bin of size 0.01 nm along the reaction coordinate. For the β-propeller tunnel exit, 20 extension simulations with a force constant of 5 MJ/(mol $\times$ nm^2) of length 5 ns each were added to reach a minimum sampling of $100,000$ conformations per bin of size 0.0095 nm. From this data, the free

energy difference profile of ZPP-POP unbinding was computed using WHAM [136] as implemented in GROMACS 4.5.3 [97]. These extension windows were initiated with the last conformation of a nearby window as indicated by the "source" columns of Tables 7.I and 7.II.

Tableau 7.I – Umbrella sampling windows parameters for the flexible loop exit. z is the reaction coordinate, the equilibrium distance between the ZPP and protein's center of mass for each window, k_{fb} the force constant of the spring restraining the ZPP at distance z, T is the length of time the window's MD simulation. The "Source" column indicates what was the source of the initial conformation of the window where SMD means it was extracted from the close position in the steered molecular dynamics and where a number points to the US window for which the last conformation was extracted.

Window	z(nm)	T(ns)	k_{fb}(MJ/(mol× nm^2))	Source	Window	z(nm)	T(ns)	k_{fb}(MJ/(mol× nm^2))	Source
1	0.3	10	2.5	SMD	18b	1.95	7.2	10	18
2	0.4	10.1	2.5	SMD	19	2.1	10.1	2.5	SMD
3	0.5	10	2.5	SMD	20	2.2	10.1	2.5	SMD
4	0.6	10.1	2.5	SMD	20b	21.4	10.1	10	20
5	0.7	10	2.5	SMD	21	2.3	10.1	2.5	SMD
6	0.8	10	2.5	SMD	22	2.4	5	2.5	SMD
7	0.9	10	2.5	SMD	23	2.5	10	2.5	SMD
8	1	10	2.5	SMD	24	2.6	10	2.5	SMD
8b	1	8.4	10	8	25	2.7	10.1	2.5	SMD
9	1.1	10	2.5	SMD	26	2.8	10.1	2.5	SMD
9b	1.05	8.4	10	9	27	2.9	7	2.5	SMD
9c	1.1	7.1	10	9	28	3	10	2.5	SMD
10	1.2	10.1	2.5	SMD	29	3.1	10	2.5	SMD
10b	1.15	8.4	10	10	30	3.2	10	2.5	SMD
10c	1.2	8.4	10	10	31	3.3	10	2.5	SMD
11	1.3	10	2.5	SMD	32	3.4	10	2.5	SMD
12	1.4	10	2.5	SMD	33	3.5	10	2.5	SMD
13	1.5	10	2.5	SMD	34	3.6	10	2.5	SMD
13b	1.47	10	10	13	35	3.7	10	2.5	SMD
14	1.6	10.1	2.5	SMD	36	3.8	10	2.5	SMD
15	1.7	10.1	2.5	SMD	37	3.9	10	2.5	SMD
15b	1.65	8.4	10	15	38	4	10	2.5	SMD
15c	1.7	8.5	10	15	39	4.1	10	2.5	SMD
16	1.8	10	2.5	SMD	40	4.2	10	2.5	SMD
16b	1.75	7.1	10	16	41	4.3	11	2.5	SMD
16c	1.8	8.5	10	16	42	4.4	10	2.5	SMD
17	1.9	10	2.5	SMD	43	4.5	10	2.5	SMD
17b	1.85	7.2	10	17	44	4.6	10	2.5	SMD
17c	1.9	7.1	10	17	45	4.7	10	2.5	SMD
18	2	10	2.5	SMD	46	4.8	10	2.5	SMD

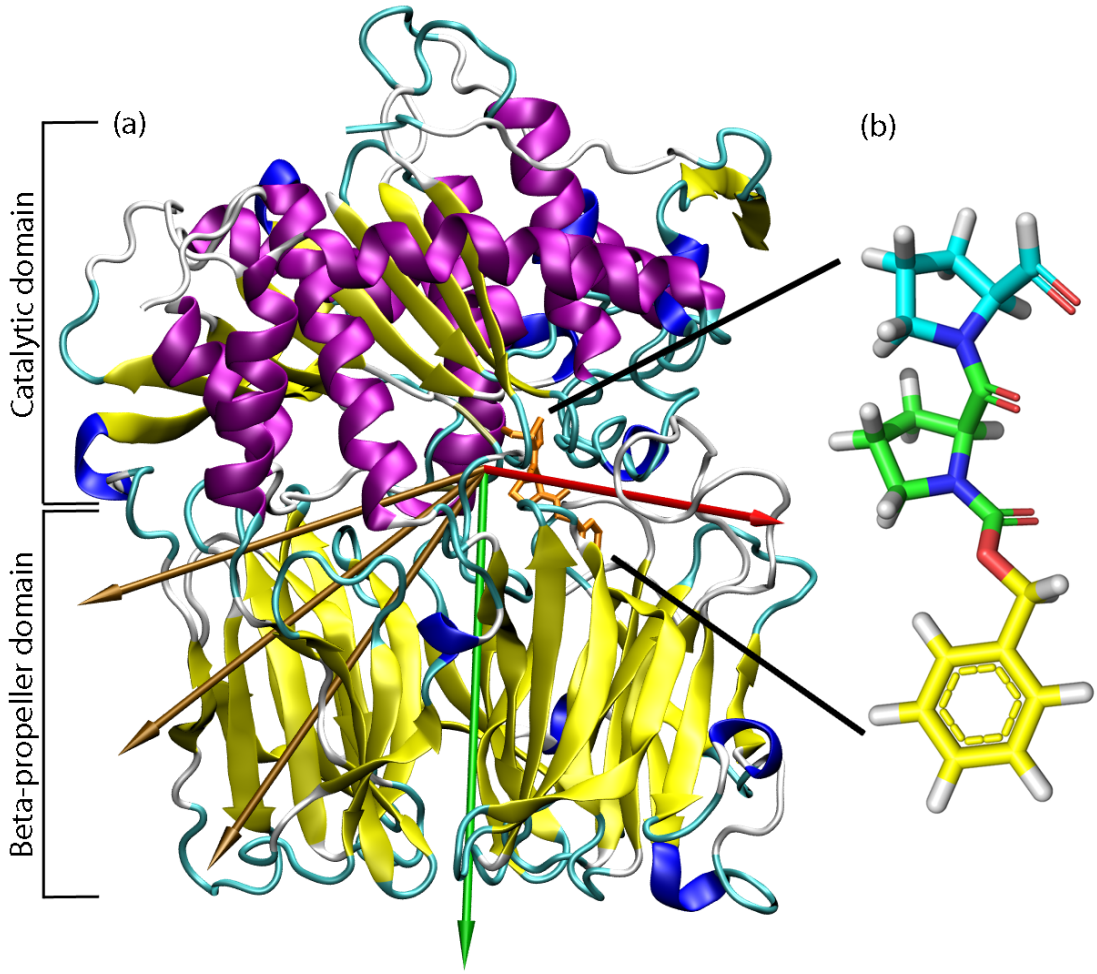

Figure 7.1 – (a) SMD ZPP-pulling vectors for the flexible loop exit (red), the β-propeller tunnel exit (green) and three possible exits through the velcro-rip of the β-propellor (golden arrows). The ZPP inhibitor inside is in orange. (b) Zoom on the ZPP with carbons colored by our definition of its 3 regions : PHE (yellow), PRO1 (green) and PRO2 (cyan) groups. In its inhibition mode, the PRO2 aldehyde group is involved in a covalent bond to the SER554 of the protein. Oxygen and Nitrogen atoms are left in red and dark blue respectively.

Tableau 7.II – Umbrella sampling windows parameters for the β-propeller exit. z is the reaction coordinate, the equilibrium distance between the ZPP and protein's center of mass for each window, k_{fb} the force constant of the spring restraining the ZPP at distance z, T is the length of time the window's MD simulation. The "Source" column indicates what was the source of the initial conformation of the window where SMD means it was extracted from the close position in the steered molecular dynamics and where a number points to the US window for which the last conformation was extracted

Window	z(nm)	T(ns)	k_{fb}(MJ/(mol$\times$ nm^2))	Source	Window	z(nm)	T(ns)	k_{fb}(MJ/(mol$\times$ nm^2))	Source
1	-0.15	9.1	2.5	SMD	27b	2.19	9.5	5	27
2	-0.05	9.7	2.5	SMD	28	2.35	5	2.5	SMD
3	0.05	8.6	2.5	SMD	29	2.45	7.3	2.5	SMD
4	0.15	9.4	2.5	SMD	29b	2.36	8.3	5	29
5	0.25	9.3	2.5	SMD	30	2.55	8.3	2.5	SMD
5b	0.2	5	5	5	30b	2.48	5	5	30
6	0.35	9.3	2.5	SMD	31	2.65	7.9	2.5	SMD
7	0.45	9.3	2.5	SMD	31b	2.59	5	5	31
8	0.55	9	2.5	SMD	32	2.75	6.3	2.5	SMD
9	0.65	9.4	2.5	SMD	33	2.85	7.8	2.5	SMD
10	0.59	5	5	SMD	34	2.95	8.7	2.5	SMD
11	0.75	8.5	2.5	SMD	34b	2.88	5	5	34
12	0.85	8.1	2.5	SMD	35	3.05	8.6	2.5	SMD
12b	0.8	5	5	12	35b	3	5	5	35
13	0.95	8.2	2.5	SMD	36	3.15	8.6	2.5	SMD
14	1.05	7.8	2.5	SMD	37	3.25	8.3	2.5	SMD
15b	1	5	5	15	38	3.35	8.5	2.5	SMD
16	1.15	9.2	2.5	SMD	38b	3.31	5	5	38
17	1.25	7.9	2.5	SMD	39	3.45	8.2	2.5	SMD
18	1.35	8.1	2.5	SMD	39b	3.5	5	5	39
18b	1.3	5	5	18	40	3.55	8.6	2.5	SMD
19	1.45	8.2	2.5	SMD	40b	3.6	5	5	40
20	1.55	7.7	2.5	SMD	41	3.65	8.4	2.5	SMD
21	1.65	5	2.5	SMD	41b	3.7	5	5	41
21b	1.6	8.2	5	21	42	3.75	8.5	2.5	SMD
22	1.75	5	2.5	SMD	42b	3.8	5	5	42
22b	1.72	9.2	5	22	43	3.85	8.7	2.5	SMD
23	1.85	8.9	2.5	SMD	44	3.95	8.8	2.5	SMD
24	1.95	5	2.5	SMD	45	4.05	8.5	2.5	SMD
24b	1.95	9.3	5	24	45b	4.05	5	5	45
25	2.05	5	2.5	SMD	46	4.15	8.5	2.5	SMD
25b	2.05	8.4	5	25	47	4.25	8.8	2.5	SMD
26	2.15	9.3	2.5	SMD	48	4.35	8.6	2.5	SMD
27	2.25	5	2.5	SMD					

7.4 Results

In order to describe the ZPP molecule and its interactions with its environment, we use the same formalism as our previous publication [128] to define the structure of the ZPP molecule in terms of 3 atomic groups. As shown in Figure 7.1b, PHE represents the aromatic phenyl head, PRO1 the middle proline, and PRO2 the terminal proline contai-

ning the aldehyde group (involved in the hemiacetal bond with SER554). In the following section we describe our SMD results followed by the US results, and an analysis of the exit pathway.

7.4.1 Exploration of the exit pathways using SMD

As the first attempt to estimate the free energy barriers associated with the exit pathways, we conducted a set of 3 SMD simulations with a pulling rate of 0.1 nm/ns and 35 SMD simulations at a rate of 0.5nm/ns for both the flexible loop and the β-propeller exit pathways with a total simulation time near of 250 ns per pulling direction. The work (W) required to move the ZPP over the entire path was calculated for each run. We computed the free energy difference from the distribution of W values using the two separate methods discussed previously, and found a significant discrepancy between their results. The pull rate was found to strongly influence the sampled pathways : in all 35 SMD trajectories with the higher pulling rate along the loop exit pathway, the ligand exited either around or through the TYR190-GLN208 flexible loop, while 2 of the 3 slower SMD pathways favored a concerted opening of the same loop. Moreover, the orientation of the ZPP at the exit of these two trajectories was reversed. The trajectory requiring the lowest work had ZPP oriented such that the phenyl group PHE was directed toward the catalytic domain and the PRO2 proline group toward the β-propeller.

The second trajectory showing a concerted opening of the flexible loop had a value of W that was greater by 40 kJ/mol than the previous pathway. In this case, ZPP exited the protein in a reversed orientation with the PHE and PRO2 groups directed toward the β-propeller and catalytic domain, respectively. The lack of convergence in the sampled work obtained at the slow pulling rate of 0.1 nm/ns indicates that a slower pulling rate would be required for the ZPP to adopt the preferred conformation in most pulling trials. Given the large number of trials necessary for sufficient statistics, and the high computational cost of each pulling trial, this option was not retained.

We also used SMD to investigate a third possible exit pathway proposed by a number of groups, which involves ZPP moving through the velcro-rip between the first and seventh blade of the β-propeller domain (see Figure 7.1) [70–72, 243]. We generated

3 SMD simulations using 3 different pulling vectors (gold vectors in Figure 7.1) and a pulling rate of 0.5 nm/ns. In all three cases, the ZPP molecule failed to pass through the velcro rip, either exiting by the β-propeller exit or through the hinge region linking the catalytic and the β-propeller domains. Clearly, the velcro-rip is extremely stable. This result is consistent with previous simulation results that indicated that the β-propeller domain is highly stable [73, 128]. Thus, this exit pathway was ruled out and the rest of our analysis focused only on the two remaining candidate pathways.

While SMD can serve to rule out a proposed exit pathway, this method requires too many simulations to converge. We did, however, use the configurations from generated pathways as starting points for US trajectories.

7.4.2 Free energy difference calculations with US

The initial configurations for launching MD in all the US windows was obtained from the SMD pathways. US for both the flexible loop and the β-propeller exit pathways was performed as described previously. The sum of all simulation times (i.e. including all windows) is equivalent to 539 ns of MD for the loop-exit and 499 ns for the β-propeller exit. In all individual simulation windows, statistics were accumulated after an initial 2 ns equilibration with positions, simulation times and force constants as listed in Tables 7.I and 7.II. To evaluate convergence toward equilibrium, we computed the root mean square deviation (RMSD) of the protein for each window. After 4 ns, the RMSD values converged, on average, to 0.20 nm $\pm$0.02 nm for the loop-exit and 0.20 nm $\pm$0.01 nm for the β-propeller pathway Figure 7.2.

Figure 7.3 shows the number of configurations sampled along the reaction coordinate for the loop and the β-propeller exit pathways, respectively, with bins of size 0.01 and 0.0095 nm. The smallest count for the loop exit is of 50,585 conformations at $z = 1.45$ nm and of 122,193 conformations at $z = 1.27$ nm for the β-propeller. They correspond to visiting times of 101 ps and 245 ps, respectively.

The resulting potential of mean force obtained from WHAM for the loop-exit and β-propeller exit is presented in Figure 7.4. The error bars give the standard deviation as calculated using 100 iterations of statistical bootstrapping using the histogram Baye-

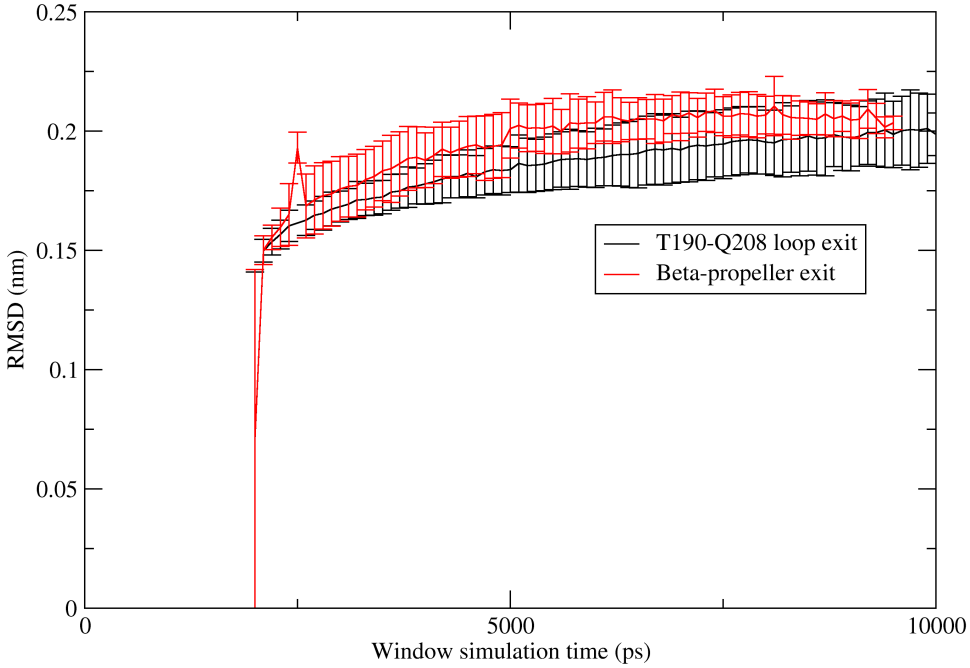

Figure 7.2 – Average root mean square deviation evolution in the loop-exit (black) and the β-propeller (red) as a function of the initial conformation of the protein in each window after 2 ns equilibration time. Error bars express the standard deviation over all windows.

sian bootstrap available in Gromacs version 4.5.3 [97]. For the loop-exit, we find the lowest free energy at position 0.8 nm in the cavity, the transition peak at 1.8 nm and the solvated free energy at 4.15 nm in the reaction coordinate. This reveals a free energy difference between the bound and free ZPP conformations of -18.5 ± 8.2 kJ/mol with a transition energy barrier of 25.1 ± 8.1 kJ/mol in the exit direction. For this interaction we can calculate a constant of inhibition $K_i = 0.8$ mM using the formula :

$$K_i = [1M]e^{\frac{-\Delta G}{RT}}, \tag{7.5}$$

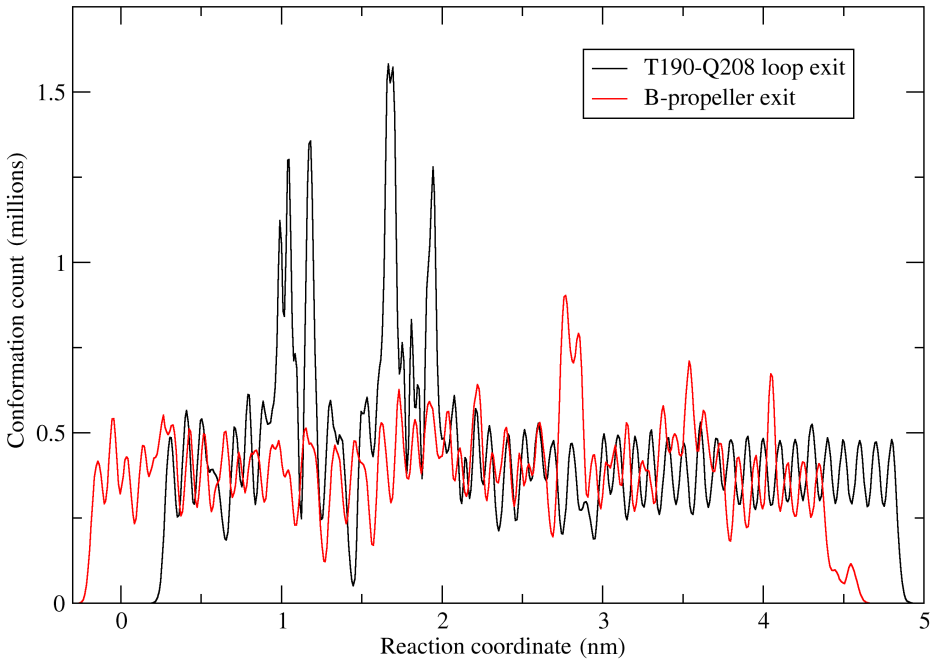

Figure 7.3 – Histogram of the reaction coordinate along the loop-exit as a function of displacement from the binding site using bin size of 0.01nm (black) and along the β-propeller tunnel exit using bin size of 0.0095 nm (red).

where R is the perfect gas constant and T is the absolute temperature. The accuracy of ΔG was calculated to be ± 8.2 kJ/mol , based on one standard deviation on the binding free energy difference, thus the confidence interval for K_i is $[32\mu\text{M}, 18\text{mM}]$. This inhibition constant is much weaker than the empirical value of $K_i = 0.35$ nM [11] since it does not include the formation of the favorable hemiacetal bond, an event not simulated in our study. However, the K_i found in our study is of the same order as other inhibitors who do not form an hemiacetal bond like suc-Gly-Pro-Nan with a $K_i = 0.278 \pm 0.35$ mM at pH 5.6 and Z-Gly-Pro-OH with a $K_i = 0.253 \pm 0.18$ mM at pH 8 or $K_i = 21.2 \pm 0.5\,\mu$M at pH 7.35 [238].

For the β-propeller tunnel exit, WHAM analysis performed on the US data (Fi-

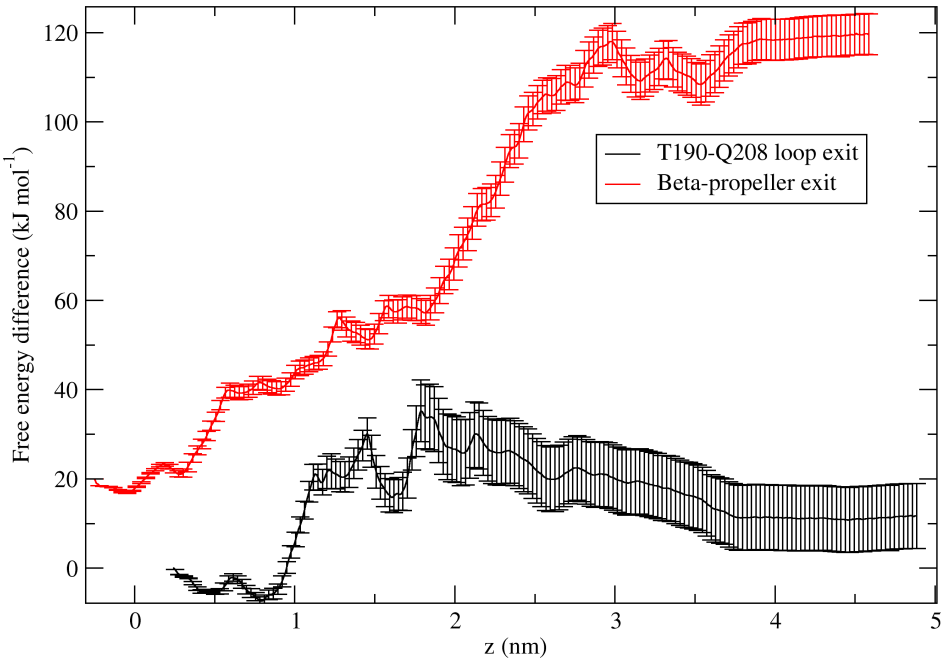

Figure 7.4 – Potential of mean force for the loop-exit (black) and β-propeller tunnel exit (red) as a function of displacement from the binding site. The red curve was shifter vertically for better legibility.Error bars express the standard deviation.

gure 7.4 in red) yielded a free energy minimum at -0.05 nm and a transition free energy of -101.6 ± 4.7 kJ/mol, corresponding to a plateau starting at $z = 4.0$ nm at which point the ZPP is free in the solvent. Contrary to the loop-exit, this pathway shows no global minimum, rising systematically as the ZPP moves out of the protein. Furthermore, the free energy barrier corresponds to a K_i of 7.6×10^{-18}M with confidence interval $[1.2\text{aM}, 50\text{aM}]$ based on one standard deviation of the free energy difference, nine orders of magnitude smaller than the experimental value, hence very strong binding.

These two observations suggest that the β-propeller exit is not sampled sufficiently, in spite of a total of 499 ns of MD simulation dedicated to this pathway. We hypothesize that the cause of this undersampling is the pathway's extreme constriction that

leads to very low mobility of the ZPP in the windows with ZPP position ranging from $z = 1.95$ nm to $z = 2.95$ nm presented in Figure 7.5, making it almost impossible to fully sample the accessible conformations along the trajectory. Thus even though the displacement may vary and overlap with neighboring windows, the phase space along the exit path is not properly sampled due to the strong dependence on initial conditions for the MD run in each window. To verify this hypothesis, we compared the average radius of gyration of ZPP for both the loop and the β-propeller pathways, the number of conformation clusters adopted by ZPP based on a RMSD clustering and the standard deviation of the angular distribution of ZPP as a function of the position along the reaction coordinate (Figure 7.6).

The above quantities are associated with the mobility and the conformational entropy for ZPP as each new window is explored. For all three properties, the window-to-window fluctuations are 20 to 30 % larger for the β-propeller exit than the flexible-loop exit. Since there is sufficient overlap between neighboring windows, we would expect that these three properties evolve smoothly from window to window, along the reaction coordinate. The large fluctuations observed for the β-propeller exit indicate rather that there is an imperfect overlap in the configuration samples within adjacent windows leading to sharp conformational transitions between windows. The overlap of the position of the center of mass of ZPP between neighbor windows does not translate to a POP-ZPP complex conformation overlap of the same windows. This suggests that the visited states are strongly influenced by the initial configuration selected for each window. While we cannot, in these conditions, extract an accurate free-energy barrier for the β-propeller exit pathway, the sampling difficulties observed can be directly associated with the presence of a significant free-energy barrier. This suggests that the pathway is significantly less favorable than the loop-exit. It is thus not unreasonable to conclude that this will not be the access pathway to the active site, leaving the loop-exit as the only feasible pathway.

7.4.3 Interaction between ZPP inhibitor and POP in loop-exit pathway

Now that the loop-exit, associated with three loosely-structured loops between the two protein domains, had been identified as the most probable entrance/exit pathway, we

investigate this pathway in further detail.

Tableau 7.III – Average probability of existence of the most persistent contacts in three regions of the reaction coordinate z for the PRO2 - body contacts and PRO2 - loop contacts. Regions units are in nm.

PRO2 - body						PRO2 - loop					
Region :	0.3-1.3	Region :	1.4-2.0	Region :	2.1-3.7	Region :	0.3-1.3	Region :	1.4-2.0	Region :	2.1-3.7
MET235	0.62	PHE173	0.85	PHE173	0.42	TYR190	0.41	TYR190	0.84	GLN192	0.42
PHE173	0.44	MET235	0.79	MET235	0.34	GLN208	0.32	ALA189	0.45	SER197	0.34
ILE591	0.35	ILE591	0.61	LYSH172	0.27	ASN205	0.15	ASN205	0.43	GLY195	0.30
ASN188	0.32	LYSH172	0.45	ILE591	0.26	SER203	0.06	GLN192	0.33	LYSH196	0.23
ARG252	0.31	ASN188	0.28	TRP150	0.19	LEU206	0.03	GLN208	0.31	ASP194	0.23
GLY236	0.26	SER174	0.20	ARG170	0.16	ALA189	0.01	SER203	0.18	GLN193	0.23
CYSH175	0.20	TRP234	0.19	ASN188	0.12			GLY195	0.04	ASP198	0.17
SER174	0.19	TRP150	0.15	SER174	0.09						
TRP234	0.19	ALA594	0.14	TRP234	0.06						
ALA594	0.18	TRP595	0.05	VAL171	0.04						
GLY237	0.17	LYSH233	0.02	ALA594	0.04						
LYSH172	0.15			GLU169	0.03						

We investigated the relation between the free-energy as a function of displacement z along the pathway with various structural properties to understand the nature of the energy barrier. As seen in Figure 7.4, two free energy peaks exist, at $z = 1.46$ and 1.8 nm, of approximately the same height, separated by a relatively deep local free-energy minimum at $z = 1.65$ nm. While it is difficult to determine the exact nature of these free-energy features, they are well correlated with specific alterations in the contacts and the H-bond network. Variation in the amino acids making contact with the different parts of ZPP can be linked to these transitions and are listed for the PRO2 (Table 7.III), PRO1 (Table 7.IV) and PHE (Table 7.V) groups. As the ZPP moves out of the protein, and z approaches 1.3 nm, it comes in contact with the cavity wall, in particular a group of hydrophobic side chains : MET235, ILE591, PHE173, TRP595. ZPP makes contact predominantly with positively charged ARG252 and ARG643, positioned on each side of the ligand (see Figure 7.7(a)). As it moves through this constrained region, however, the ligand also forms contact with hydrophobic amino acids PHE173, ILE591 and MET235 Figure 7.7(b), as well as with hydrophilic amino acids of the TYR190-GLN208 flexible loop Figure 7.7(c), strongly reducing ZPP's access to the solvent.

The ligand's outward displacement along the reaction coordinate is sterically constricted by a group of amino acids that block ZPP's direct access to the flexible loop. Looking more specifically at the configurations sampled around the free energy peak at

Tableau 7.IV – Average window probability of existence of the most persistent contacts in three regions of the reaction coordinate z for the PRO1 - body contacts and PRO1 - loop contacts. Regions units are in nm.

PRO1 - body						PRO1 - loop					
Region :	0.3-1.3	Region :	1.4-2.0	Region :	2.1-3.7	Region :	0.3-1.3	Region :	1.4-2.0	Region :	2.1-3.7
ILE591	0.67	ASN205	0.77	TRP150	0.25	TYR190	0.092	SER203	0.75	GLN192	0.44
MET235	0.56	TRP150	0.70	ASN205	0.14	LEU206	0.048	TYR190	0.57	LYSH196	0.40
TRP595	0.48	LYSH172	0.66	PHE173	0.10	SER203	0.031	GLN192	0.51	GLY195	0.36
PHE173	0.46	THR590	0.51	LYSH172	0.07			SER197	0.50	SER197	0.34
ALA594	0.36	ILE591	0.45	ILE591	0.05			GLU201	0.32	ASP198	0.18
ILE478	0.26	PHE173	0.29	VAL171	0.04			THR204	0.13	ARG170	0.18
ARG252	0.21	MET235	0.11	MET235	0.04			ASP194	0.13	ASP194	0.18
PHE476	0.16	TYR589	0.03	THR590	0.01			LYSH196	0.10	TYR190	0.16

Tableau 7.V – Average window probability of existence of the most persistent contacts in three regions of the reaction coordinate z for the PHE - body contacts and PHE - loop contacts. Regions units are in nm.

PHE - body						PHE - loop			
Region :	0.3-1.3	Region :	1.4-2.0	Region :	2.1-3.7	Region :	1.4-2.0	Region :	2.1-3.7
ILE591	0.84	THR590	0.92	TRP150	0.17	SER197	0.68	LYSH196	0.38
ARG643	0.87	LYSH172	0.81	ARG170	0.12	GLY199	0.61	SER197	0.38
PHE173	0.57	VAL645	0.56	PHE173	0.08	SER203	0.30	GLN192	0.32
TRP595	0.46	PHE173	0.53	THR590	0.06	ASP198	0.28	ASP198	0.31
THR590	0.36	ILE591	0.47	LYSH172	0.05	GLU201	0.23	GLY195	0.29
ASN555	0.24	ASP642	0.40	TYR589	0.03	THR202	0.12	SER203	0.24
HISB680	0.24	ARG643	0.36	LYSH23	0.03	THR200	0.08	GLU201	0.22
ASP149	0.23	TYR589	0.33			LYSH196	0.06	ASP194	0.18
LYSH172	0.22	VAL644	0.29			GLN192	0.05	ASN205	0.17
PHE476	0.21	VAL580	0.21						
ILE478	0.11	TRP150	0.19						
SER554	0.11	TRP595	0.06						

$z = 1.46$ nm, we see that ZPP is sterically constrained between THR590 on one side of the PHE group and TRP150, LYS172 and ARG643 on the other side. This reduces significantly the PRO1 and PRO2 groups' conformation flexibility, thus lowering the entropy available to the ZPP and raising the free energy difference.

The drop in free energy, as ZPP moves past $z = 1.46$, is associated with a decrease in the average number of H-bonds of 1.6 ± 0.4 (relative to the average number of H-bonds for the region $z = 0.3$ to $z = 1.25$ nm) between the whole protein and its TYR190-GLN208 loop, easing ZPP's diffusion pathway (Figure 7.8). Table 7.VI gives a list of H-bonds between the above loop and the rest of the protein that are modulated by the passage of ZPP through position $z \geq 1.46$ nm. For $z < 1.46$ the TYR190-GLN208 loop maintains an average of 11.75 ± 0.12 H-bonds with the rest of the protein. There are only 17 pairs of atoms forming H-bonds between the above loop and the protein that exists

for more than 25% of the simulation time and another 28 pairs of atoms present between 5% and 25% of that simulation time.

Tableau 7.VI – Average window probability of existence of the most persistent h-bonds between the TYR190-GLN208 loop and protein body in the 0.3nm to 1.8nm section of the loop exit pathway.

Donor	Acceptor	presence time ratio
SER203N	LYS588O	0.42
HIS593N	THR204OG	0.37
TYR190O	TRP234O	0.36
GLN208NE	TRP234O	0.33
LYS196NZ	GLU169O	0.31
ASN188ND2	GLN208OE1	0.30
SER203OG	THR590O	0.30
TRP234N	TYR190O	0.28
LYS196NZ	TRP150O	0.28

The second free energy barrier ($z = 1.8$ nm) can be linked with the breaking of H-bonds that connect the β-propeller and catalytic domains. The average number of H-bonds between these two domains increases from 15.5$\pm$0.22 at $z = 1.6$ nm to 19.4$\pm$0.15 at $z = 1.8$ nm falling back to 15.1$\pm$0.26 at $z = 1.95$ nm. Figure 7.8, which correlates well with the observed energy peak at $z = 1.8$ nm. The bulk of the variation in the H-bonds network between the two domains is not directly linked to the opening of the flexible loop into the solvent since we can see that the number of H-bonds between the TYR190-GLN208 loop and the catalytic domain varies only slightly, going from 1.5$\pm$0.1 H-bonds at $z = 1.6$ nm to 1.8$\pm$0.1 at $z = 1.8$ nm and decreases to 1.3$\pm$0.1 at $z = 1.95$ nm (data not presented).

To identify the H-bonds modulated by the position of ZPP, we calculated the correlation coefficient between the average number of H-bonds per US window and the probability of existence of each individual H-bond for that window. We define this probability of existence as the percentage of time over the length of a window for which the H-bond exists. Table 7.VII presents those H-bonds with (absolute value) minimum correlation coefficients of 0.4 for windows around the energy peaks. When combined, the fluctuations of this group of H-bonds (including those with negative correlation co-

Tableau 7.VII – Hydrogen bonds located at the inter-domain interface with activity modulated by the position of ZPP on the reaction coordinate as identified by the Pearson's correlation coefficient against the average number of h-bonds for two regions, $z = [1.3, 2.0]$ and $z = [1.05, 2.1]$. In both cases, the average probability of existence and standard deviation of each h-bond in their respective subset of windows is also given.

	Region $z = [1.3, 2.0]$			Region $z = [1.05, 2.1]$		
Hydrogen bond	Correl.(p-val.)	Prob.	Std-dev	Correl.(p-val.)	Prob.	Std-dev
GLN439NE2-GLN439HE22-ASP356O	0.51(0.044)	0.30	0.08	0.39(0.060)	0.32	0.09
THR597OG1-THR597HG1-GLY254O	0.50(0.049)	0.16	0.34	0.33(0.115)	0.11	0.28
THR686OG1-THR686HG1-ASN96OD1	0.45(0.080)	0.34	0.42	0.52(0.009)	0.32	0.42
SER148OG-SER148HG-ASP642OD2	0.44(0.088)	0.29	0.31	0.31(0.140)	0.27	0.35
LYSH75NZ-LYSH75HZ3-TYR71OH	0.43(0.096)	0.28	0.42	0.39(0.060)	0.27	0.42
SER148OG-SER148HG-ASP642OD1	0.41(0.114)	0.26	0.35	0.20(0.349)	0.26	0.32
LYSH677NZ-LYSH677HZ3-ASP122OD1	0.40(0.125)	0.33	0.37	0.26(0.220)	0.36	0.38
ARG128NH1-ARG128HH12-ASP641OD1	-0.40(0.125)	0.16	0.33	-0.40(0.053)	0.16	0.30
LYSH172NZ-LYSH172HZ3-ASP642OD2	-0.43(0.096)	0.16	0.27	0.07(0.745)	0.22	0.32
LYSH428NZ-LYSH428HZ3-GLU69OE1	-0.54(0.031)	0.77	0.31	-0.31(0.140)	0.81	0.26
LEU94N-LEU94H-ASP72OD1	-0.54(0.031)	0.19	0.35	-0.06(0.780)	0.23	0.38

efficients) explains 75% of all the inter-domain H-bonds fluctuations. Although a few H-bonds are formed and broken in the vicinity of the ZPP (SER148-ASP642, ARG128-ASP641, LYSH172-ASP642), many are located opposite the loop opening in the interface region near the velcro rip (THR686-ASN96, LYSH75-TYR71, LEU94-ASP72, LYSH428-GLU69 in Figure 7.9).

Once the ZPP has moved past the second free energy maximum, it gradually loses contact with the body of the protein, only keeping contact with the TYR190-GLN208 loop Figure 7.7(d) . Due to the higher mobility of the solvent-exposed ZPP, no strong contact dominates in this region. As the ZPP is pulled outwards, the flexible loop adopts an extended conformation to maintain contacts with it. The amino acids having the most frequent contacts with the ligand are those situated on the 192-198 segment of the loop which can conformationally extend the furthest into the solvent.

When examining the evolution of the H-bond network between the 2 domains of POP, we can see a drop in the average number of H-bonds as the ZPP moves outwards from the the second free energy peak ($z = 1.8$ nm), starting from 19.4 ± 0.15 and decreasing to 13.9 ± 0.17 at $z = 2.1$ nm. This number then increases as the ZPP moves further

along the trajectory up to a value of 18.4±0.22 at $z = 3.0$ nm. This increase of 4.5±0.4 H-bonds from $z = 2.1$ nm to $z = 3.0$ nm is mainly due to an increase of 1.9±0.3 H-bonds between the flexible loop and the catalytic domain indicating that the TYR190-GLN208 loop is folding back onto the protein. Correlation analysis between the presence of each individual H-bond and the average number of H-bonds (Table 7.VIII) shows fewer H-bonds forming atom pairs with high correlations : Specifically, all the H-bond forming pairs with an absolute value minimum correlation of 0.3 had to be selected to explain 75% of the variation of the average number of H-bonds.

Tableau 7.VIII – Hydrogen bonds located at the inter-domain interface with activity modulated by the position of ZPP on the reaction coordinate as identified by the Pearson's correlation coefficient against the average number of h-bonds for the hydrogen bonds with the highest correlation for the region $z = [1.75, 3.7]$. Also included are the average probability of existence and standard deviation of each h-bond in the selected subset of windows.

Hydrogen bond	Correl.(p-val.)	Prob.	Std-dev
THR597OG1-THR597HG1-GLY254O	0.50(0.008)	0.14	0.29
TRP150N-TRP150H-ASP642OD2	0.48(0.011)	0.15	0.29
THR686OG1-THR686HG1-ASN96OD1	0.48(0.012)	0.34	0.40
LYSH516NZ-LYSH516HZ3-ASP256OD1	0.43(0.025)	0.13	0.23
THR686OG1-THR686HG1-GLN95O	0.41(0.031)	0.11	0.27
ASN477ND2-ASN477HD22-TYR311OH	0.38(0.051)	0.21	0.18
THR686N-THR686H-ASN96OD1	0.38(0.051)	0.26	0.35
LYSH75NZ-LYSH75HZ3-TYR71O	0.37(0.058)	0.31	0.40
SER148N-SER148H-HISA640O	0.36(0.066)	0.19	0.22
ARG128NH2-ARG128HH22-ASP641OD1	0.32(0.099)	0.19	0.31
LYSH172NZ-LYSH172HZ3-ASP642OD1	0.32(0.106)	0.14	0.24
ARG128NH2-ARG128HH22-ALA682O	-0.31(0.113)	0.08	0.19
SER381N-SER381H-PRO482O	-0.34(0.086)	0.18	0.34

Interestingly, we observed a prolonged interaction between the TYR190-GLN208 loop and ZPP in the z-window where the inhibitor is completely solvated. Starting from a ZPP-POP distance of $z = 2.7$ nm, and moving further outwards from the protein, the large majority of the ligand's contacts are with amino acids on the flexible loop, even though ZPP's motion is only constrained along the reaction coordinates and the ligand can move freely in the perpendicular hyperplane. For example, in the $3.2 < z < 3.7$ nm

region, ZPP makes contacts with 5.4 amino acids compared to 12.6 contacts on average for the $0.3 < z < 2.0$ nm region of the trajectory(data not presented). Small in number, these interactions are nevertheless sufficient to stabilize ZPP's position and keep it in contact with the loop for at least 80% of the simulation time. This suggests that the role of the flexible loop is not simply to open up and leave a pathway open for the ZPP entrance. The TYR190-GLN208 loop could play an active role in recruiting the ZPP ligand by binding to it in the solvent and directing it to the entry pathway, helping the ZPP to go through a first free energy barrier at $z = 1.8$ nm.

7.5 Discussion

In this work, we identified the most probable pathway for binding ZPP to POP. Using SMD and US simulations, we eliminated two proposed pathways : first, through the velcro-rip and second through the β-propeller tunnel. The first pathway is ruled out because we were unable to even generate a pathway using SMD demonstrating the extreme resistance to this path. The resulting β-propeller exit trajectory free energy difference profile was unphysical. Detailed analysis suggests that this unphysical profile is associated with very constrained regions of the pathways where sampling is particularly difficult. Based on this evidence, we were also able to eliminate this pathway. The appropriate behavior of the ZPP when pulled through the flexible loop region and the physical nature of the free energy difference profile indicate that this is the correct access pathway.

Whether or not the access to the binding cavity involves a large domain reconfiguration could not be definitively resolved by our study. The natural substrates of POP, peptides of length < 30 residues, are much larger than the ZPP inhibitor. Thus the access mechanism of ZPP may not be the general pathway involved in its catalytic activity. The long-range destabilization of the H-bond network that was seen to occur as the ZPP left the binding cavity could possibly be interpreted as evidence of the large scale interdomain motion [225] that has been hypothesized to play a role in the access pathway. In addition, the prolonged association of the TYR190-GLN208 loosely structured loop

with the ZPP as it left the protein, provides evidence that this region of the protein could possibly have a role in ligand recognition and recruitment.

7.6 Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (MK, NM), the Academy of Finland (TR,MK), Canada Research Foundation (NM) and the GALENOS program (JFSP). We are grateful to the Réseau québécois de calcul de haute performance (RQCHP), SharcNet and the Finnish IT Centre for Science for their generous allocation of computer time.

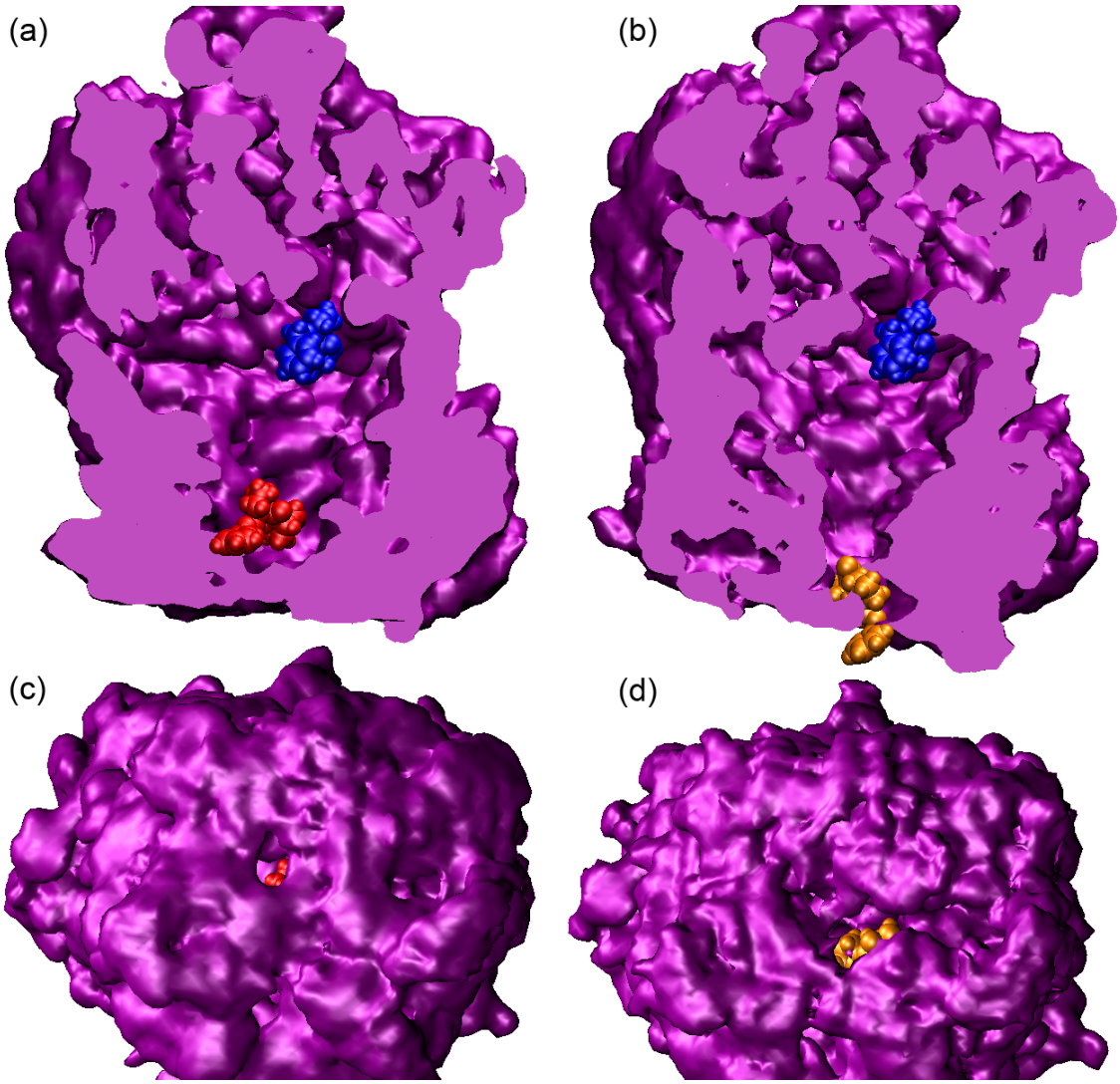

Figure 7.5 – Constriction of the β-propeller exit pathway. Position of the ZPP is presented (a) in red for window $z = 1.95$ nm and (b) in orange for $z = 2.95$ nm with ZPP from window $z = -0.15$ nm in blue as a reference to the starting position. Bottom view of the β-propeller tunnel is presented in (c) and (d) respectively. The protein's surface was computed from a volumetric density map averaged over the trajectory of the respective window.

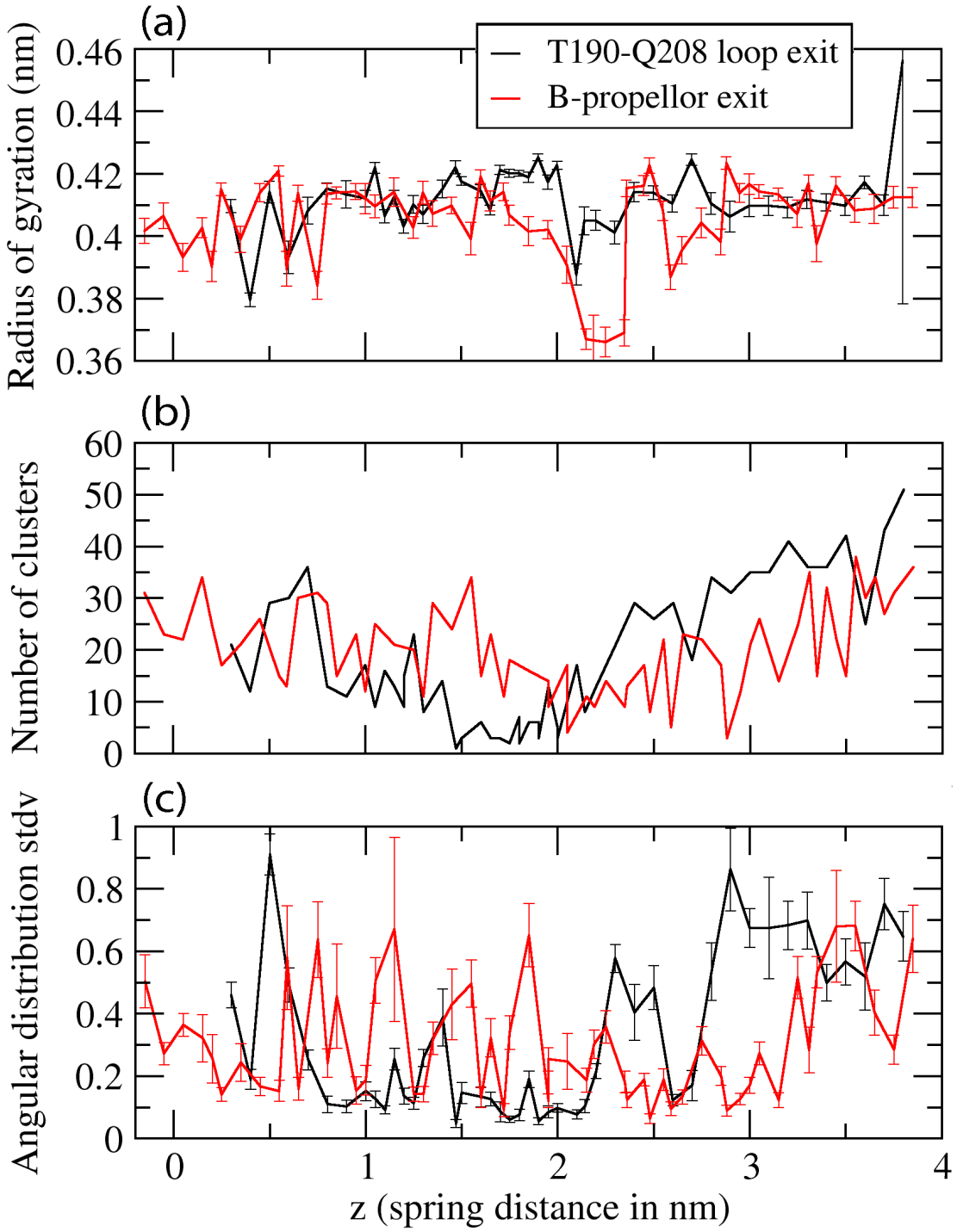

Figure 7.6 – (a) Average radius of gyration of ZPP as a function of the displacement from the binding site, (b) average number of conformation cluster using a RMSD clustering algorithm of cluster size 0.07 nm and (c) standard deviation of the angular distribution of ZPP as a function of displacement. Error bars in (a) and (c) are obtained through 5000 bootstrap evaluation of 10% of the available data and a confidence probability of 95%

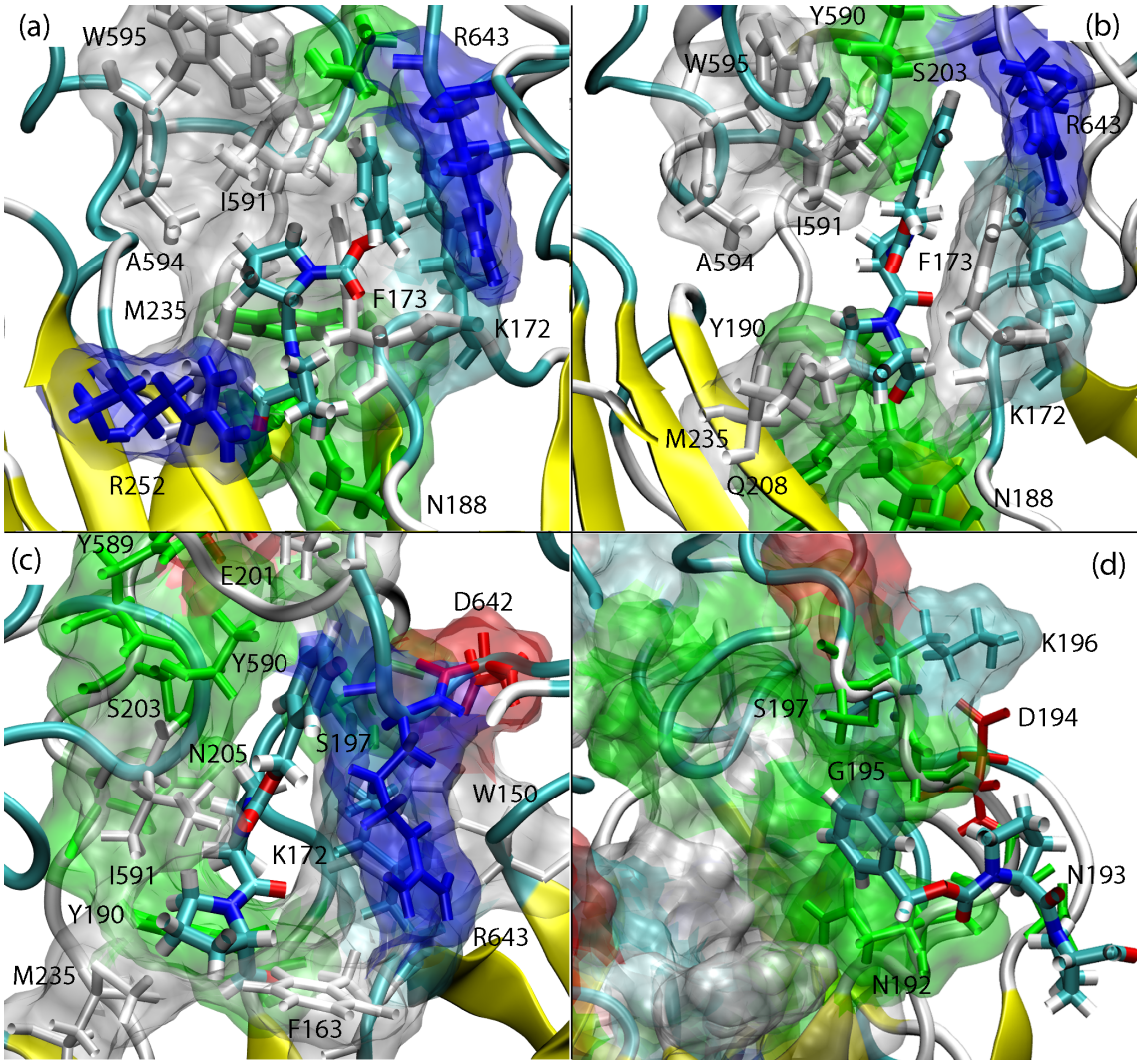

Figure 7.7 – Main amino acids (colored by type) making contact with ZPP from the windows $z = 1.0$ nm (a), $z = 1.3$ nm (b), $z = 1.6$ nm (c) and $z = 3.0$ nm (d).

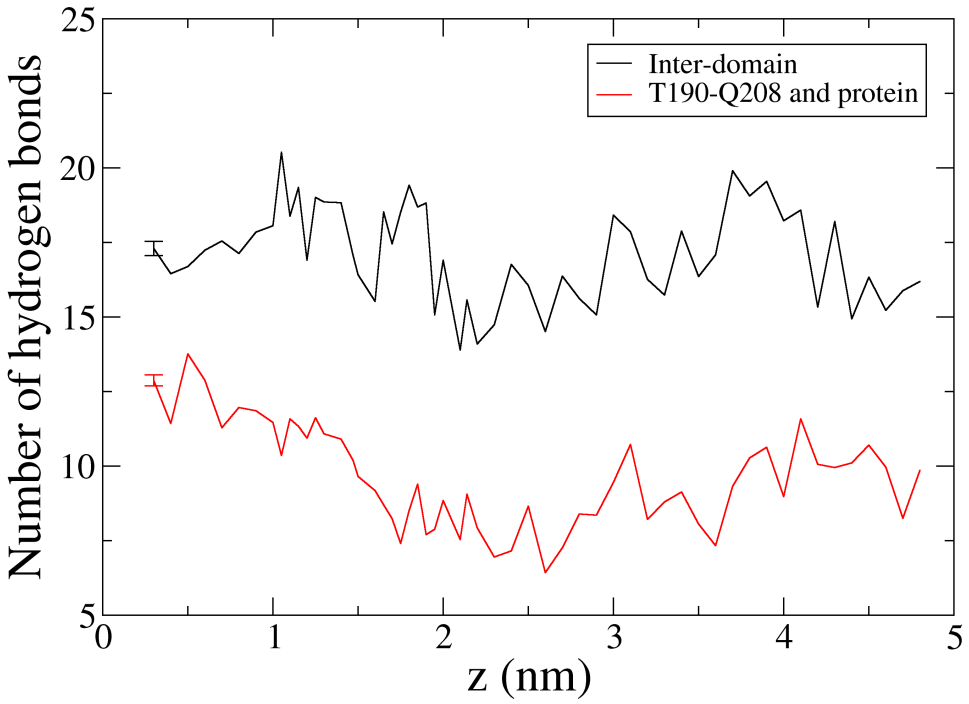

Figure 7.8 – Number of h-bonds formed between the two domains of POP (black) and between the TYR190-GLN208 flexible loop and the protein body (red) as a function of the spring equilibrium length.Maximum error evaluated to ± 0.26 and ± 0.18 respectively are obtained through 5000 bootstrap evaluation of 10% of the available data and a confidence probability of 95%

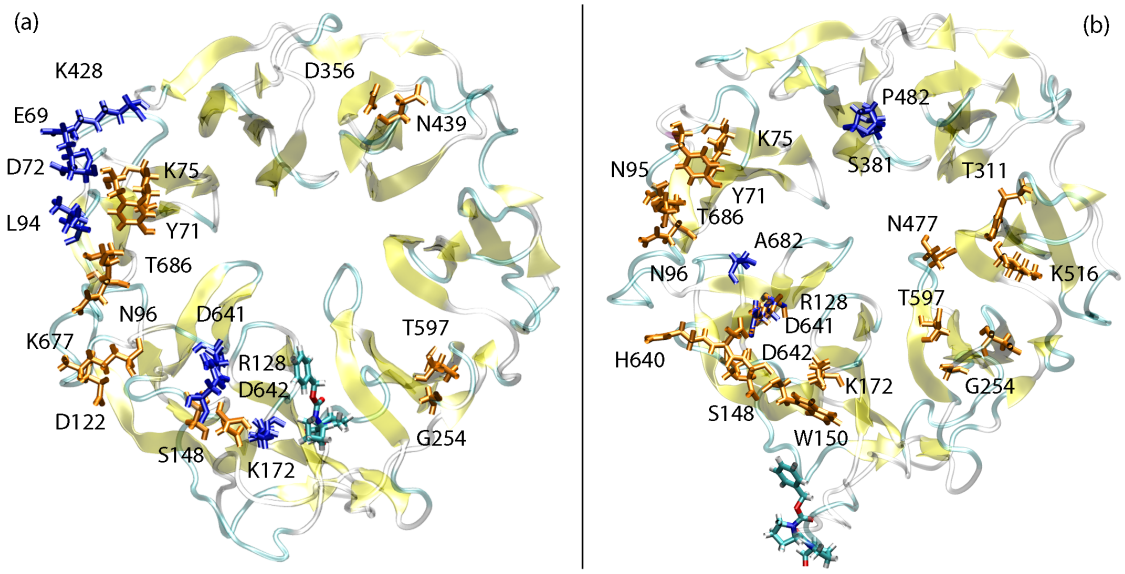

Figure 7.9 – View of the amino acids forming inter-domain H-bonds modulated by the position of ZPP on the reaction coordinate as identified by Pearson's correlation coefficient of the involved H-bond against the average number of H-bonds. Amino acids in orange are positively correlated with the average number of H-bonds while those in blue are negatively correlated. The catalytic domain is hidden to facilitate the view. ZPP crossing $z = 1.3$ nm (a) and $z = 3.0$ nm (b).

CHAPITRE 8

APPROFONDISSENT DE L'ARTICLE SUR LA PROLYL OLIGOPEPTIDASE

La première conclusion à tirer des travaux de dynamique moléculaire dirigée (DMD) et d'échantillonnage parapluie (EP) sur la protéine Prolyl-oligopeptidase (POP) est que pour des systèmes complexes caractérisés par de hautes énergies de transitions, la DMD ne peut fournir que des résultats qualitatifs. Dans le système relativement simple de POP avec le Z-Pro-Prolinale (ZPP) où seulement des changements de conformation de la protéine de faible amplitude étaient nécessaires à la sortie du ligand, la DMD n'a pu être utilisée pour obtenir des valeurs d'énergie libre de liaison plausibles. Pour pouvoir discriminer entre le chemin de sortie par la boucle flexible T190-N208 et le chemin par le trou central du propulseur-β, l'utilisation de la méthode de calcul d'énergie libre par EP fut nécessaire. Toutefois, les trajectoires de DMD de sortie par la boucle T190-N208 ont échantillonné un changement de conformation de cette boucle flexible qui avait déjà été observé dans des simulations de DM précédentes, soit le dépliement de la boucle et sa solvatation [129]. L'utilisation de cette trajectoire pour ensemencer les fenêtres d'EP a par la suite permis de constater que la boucle T190-N208 dépliée interagissait de façon significative avec le ligand ZPP dans les fenêtres d'EP où le ZPP se trouve à l'extérieur de la protéine. Une conformation affichant ce genre de contacte entre le ZPP et la boucle flexible est présenté en figure 8.1. Il est peu probable que cette observation eut été faite si les simulations d'EP avaient été ensemencées en forçant le positionnement du ZPP au même endroit, mais avec une protéine POP dont la boucle T190-N208 était à l'état repliée ; le dépliement de cette boucle dans les simulations de DM n'a été observé qu'à une seule reprise, et cela après 70 ns de simulation alors que les fenêtres d'EP ne sont simulées que 10 ns. Cette observation dans les simulations d'EP laisse présager que la boucle T190-N208 joue un rôle dans le recrutement de ligands.

La méthode de calcul de différences d'énergie libre par EP n'est pas sans lacunes. On observe dans la trajectoire de sortie du ZPP par le domaine en propulseur β (DPβ) qu'il peut être difficile pour la méthode d'EP de récupérer après avoir traverser une

barrière énergétique élevée. Bien que la méthode puisse donner l'impression d'avoir convergé vers un profil d'énergie libre de la trajectoire par le DPβ, on note une erreur de plusieurs ordres de grandeurs dans le calcul de la constante de liaison basée sur la différence d'énergie libre entre l'état du ZPP lié à POP et l'état du ZPP en solution comparativement aux valeurs expérimentales. Cependant, les résultats d'évaluation de la constante de liaison pour la trajectoire par la boucle flexible sont en accord avec les valeurs expérimentales et la plus faible énergie libre de transition observés pour cette trajectoire démontre que l'EP peut générer des profils énergétiques intéressants pour des trajectoires probables et discriminer les trajectoires improbables.

Bien que le système POP-ZPP puisse être classifié de difficulté intermédiaire, certaines trajectoires plus complexes n'ont pas été étudiées dans notre projet. Ainsi, l'interface inter-domaines et la boucle T190-N208 peuvent jouer un rôle dans le recrutement d'un ligand de petite taille comme le ZPP, mais il est peu probable que des ligands de 30 a.a. empruntent le même chemin étroit. Ici, l'hypothèse d'une séparation partielle ou complète entre les deux domaines POP semble plus adéquate, mais l'échantillonnage de la trajectoire d'ouverture de POP est complexifié par l'asymétrie du contact entre les deux domaines et des différents modes d'ouverture asymétrique qui en découlent. Une DMD à une dimension utilisant un potentiel reliant le centre de masse de chacun des deux domaines de POP permettrait d'identifier la zone de l'interface inter-domaines dont la cohésion est la plus faible, mais ne permettrait pas de déterminer s'il est énergétiquement avantageux pour la boucle T190-N208 de se déplier avant ou pendant la séparation des domaines. Il faudra avoir recours à des méthodes d'échantillonnage de trajectoires multidimensionnelles comme l'EP à plusieurs coordonnées de réactions [140]. Toutefois, l'utilisation de ces méthodes demandent des temps de calcul d'ordre $O(N \times M)$ au lieu de $O(N)$ où N et M sont le nombre de fenêtres divisant chacune des coordonnées de réactions. Pour être appliquées au système POP-ZPP, des infrastructures plus performantes seraient nécessaires pour le calcul de la DM, tel que le super-ordinateur Anton [227].

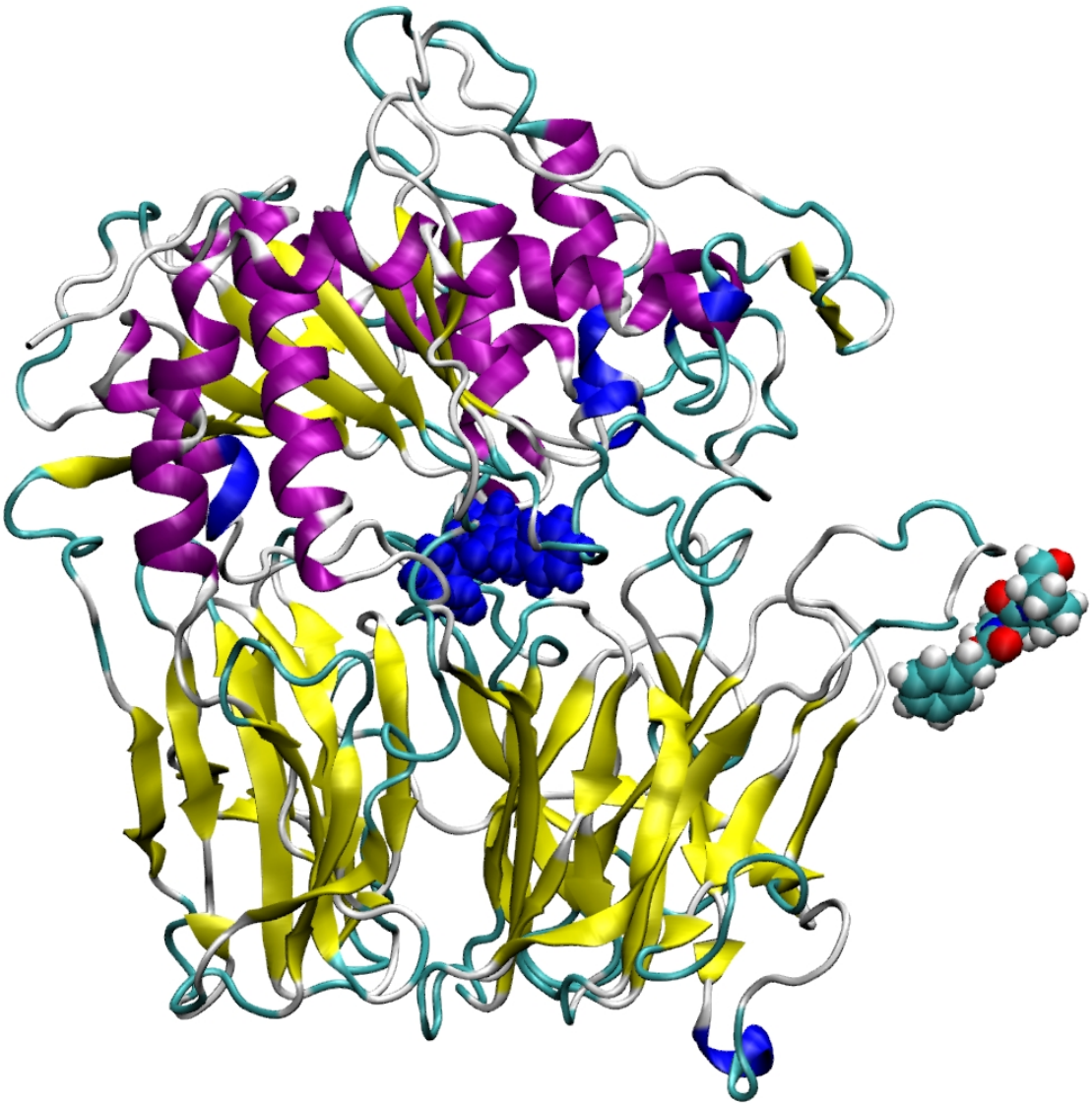

Figure 8.1 – Interaction entre le ZPP et la boucle flexible T190-N208 après 8 ns de simulation par échantillonnage parapluie avec une distance inter centre de masse $z =$ 3.5 nm sur la trajectoire de sortie par l'interface inter-domaines. La structure de ZPP à $z = 0.3$ nm est présenté en bleu pour fin de comparaison

CHAPITRE 9

PRÉDICTION DES STRUCTURES BOUCLES

La prédiction de la structure des protéines est un domaine qui a grandement évolué depuis les vingt dernières années, comme en témoigne la compétition bisannuelle du *Critical Assessment of Techniques for Protein Structure Prediction* (CASP) qui en est à sa neuvième édition [135]. Grâce à la modélisation par homologie, il est possible d'obtenir des prédictions de structure de haute qualité lorsque la structure de haute résolution d'une protéine homologue est disponible et lorsque leurs séquences sont hautement similaires [35]. Par contre, plusieurs facteurs peuvent affecter en partie la qualité de la structure prédite : certaines régions peuvent présenter de faibles similarités de séquence ou peuvent être entièrement manquantes de la séquence de la protéine patron, ou encore elles peuvent être mal résolue dans la structure du patron.

Les boucles sont des courtes séquences d'acides aminés qui relient des éléments de structure secondaire et qui affichent souvent de faibles taux d'identité de séquence comparativement au reste de la protéine. Bien que certaines boucles jouent un rôle important au site actif des protéines, la plupart se retrouvent à la surface des protéines, exposées au solvant, et sont caractérisées par une plus grande flexibilité. Due à cette flexibilité et au fait que leur séquence et leur structure sont faiblement conservées [27, 34], la prédiction de la structure des boucles est difficilement atteignable par les méthodes de prédiction par homologie traditionnelle.

Ce chapitre présente une revue de la littérature concernant les méthodes de prédiction des boucles, suivie d'une présentation de notre méthode ART-nouveau adaptée au problème de prédiction des boucles.

9.1 Définition de la métrique RMSDg

Avant de discuter de méthode de prédiction de structures boucles, il est nécessaire de définir une métrique universelle permettant de comparer la qualité des prédictions.

Traditionnellement, la qualité des prédictions par homologie est donnée par la racine carrée de la moyenne des distances inter-atomes au carré, ou *root mean square deviation* (RMSD) définie comme étant :

$$\sqrt{RMSD = \frac{\sum_{i=1}^{N}(r_i - r_i')^2}{N}}, \tag{9.1}$$

où r et r' sont les vecteurs de position de la protéine modélisée et de la protéine patron, N est le nombre de coordonnées équivalent à trois fois le nombre d'atomes. On calcule le RMSD après avoir superposé les structures modélisées sur leur patron afin de minimiser la contribution rotationnelle et translationnelle de l'orientation spatiale des protéines.

Lors de la prédiction de structure des boucles, le corps de la protéine est souvent fixé ou inexistant rendant l'utilisation du RMSD sur l'entièreté de la protéine inapplicable. Certaines études présentent leurs résultats en utilisant le RMSD calculé seulement sur la partie en boucle de la protéine. Cette approche présente l'inconvénient qu'elle ne tient pas compte de l'orientation de la boucle vis à vis du corps de la protéine et peut ainsi surestimer la qualité d'une prédiction. La métrique du RMSD globale (RMSDg) a donc été définie pour tenir compte du corps de la protéine : le RMSDg est toujours calculé sur les parties boucle de la protéine, mais seulement après superposition des corps fixes. Dans le cas où le corps de la protéine n'est pas utilisé, la superposition est faite sur les a.a. d'ancrage aux l'extrémités fixes N-terminale et C-terminale de la boucle.

9.2 Méthodes de prédiction de structures boucles

La prédiction de structure pour des boucles courtes peut aisément être obtenue par échantillonnage exhaustif de l'espace conformationnel de la boucle. Il y a huit ans de cela, par contre, aucune méthode ne pouvait systématiquement enrichir la qualité d'une structure boucle prédite comparativement à la simple copie d'une structure patron pour des boucles de plus de cinq a.a. [257]. Depuis, plusieurs méthodes ont vu le jour et permettent maintenant d'obtenir des résultats adéquats pour des boucles ayant jusqu'à douze a.a. Or, la question des longues boucles reste un sujet d'actualité.

9.2.1 Méthodes basées sur la connaissance

On dénombre deux familles de méthodes d'évaluation du repliement des boucles. La première et la plus rapide est basée sur l'utilisation de bases de données de boucles qui ont une certaine similarité de séquence ou de structure avec la protéine cible. Fernandez-Fuentes *et al.* définit une base de données de boucles, ArchDB, caractérisée par la structure secondaire des deux a.a. d'ancrage flanquant chaque boucle connue, leur orientation et leur distance ainsi que le nombre d'acides aminés dans la boucle et les angles dièdres ϕ et ψ de celle-ci [66]. Les boucles ainsi classées sont utilisées pour créer un profil de modèle de Markov caché caractérisant la structure de la classe boucle. La structure de la boucle cible est obtenue par alignement avec le profil correspondant.

L'idée d'utiliser la structure de a.a. d'ancrage de chaque côté de la boucle est reprise par d'autre groupes. Peng et Yang utilisent la base de données LSBSP1 [275] permetant d'obtenir une première prédiction des angles dièdres de segments de 9 a.a. couvrant la boucle et les 3.a.a d'ancrage de chaque côté à l'aide d'un réseau neuronal [200]. Les fragments de 9 a.a. sont ensuite alignés à l'aide d'un algorithme dynamique au profil de la boucle défini par l'orientation des a.a. d'ancrage. Finalement, les modèles ainsi générés sont filtrés pour éliminer ceux qui ont un mésappariement des a.a. d'ancrage, puis triés par qualité d'alignement de séquence, d'angles dièdres et des a.a. d'ancrage.

Hildebrand *et al.* utilisent la base de données LIP [172] pour laquelle seuls deux atomes d'ancrage sont utilisés de chaque côté des boucles [90]. Une cinquantaine de conformations de boucles échantillonnant le plus grand espace possible sont ensuite obtenues à l'aide d'un pointage simplement basé sur la similarité de séquence et l'alignement des quatre atomes d'ancrage. Similairement, la méthode FREAD [48] aligne des boucles entières aux a.a. d'ancrage de la protéines basée sur un facteur de similarité. Là où FREAD diffère est dans la définition de la similarité : Au lieu d'utiliser simplement une matrice de substitution évolutive tel que BLOSSUM62 pour aligner la séquence de la boucle avec les fragments de protéines de sa base de données, la méthode utilise une matrice de substitution basée sur la similarité structurelle des fragments et la probabilité que deux a.a. différents occupent la même position dans un fragment de structure don-

née. Avec l'augmentation du nombre de structures connues, les performances de FREAD semblent s'améliorer [33].

9.2.1.1 Limitations

Les méthodes entièrement basées sur l'utilisation de base de données sont limitées par l'étendue des bases de données. Le nombre de séquences possibles d'a.a. augmentant exponentiellement en fonction de la longueur de la séquence, les probabilités de trouver une séquence affichant une haute similarité pour de longues boucles diminue d'autant. De plus, le nombre de conformations possibles augmentant aussi exponentiellement avec la longueur de la séquence, la probabilité de trouver dans la base de données des patrons de boucles de haute homologie de séquence et de structure similaire est très faible pour les longues boucles. Elle sont aussi affectées par la qualité des structures composant la base de données, ces dernières étant obtenues par des méthodes expérimentales dont la résolution peut être aussi basse que 3.5 Å Finalement, il est difficile d'inclure dans les bases de données l'information concernant le corps de la protéine interagissant avec la boucle et affectant partiellement sa structure. Ces méthodes sont par-contre bien adaptées à l'évaluation rapide de courtes séquences et les trois présentées précédemment sont disponibles au travers de serveurs sur le Web.

9.2.2 Méthodes *ab initio*

Comme leur nom l'indique, les méthodes *ab initio* ne formulent pas d'hypothèses de départ sur les structures probables des séquences composant les boucles. L'effort est donc mis à l'échantillonnage de structures et à l'évaluation de leur probabilité par des méthodes de pointage habituellement basées sur des potentiels énergétiques physiques ou statistiques.

Une première classe de méthodes de *ab initio* consiste à échantillonner systématiquement l'espace des conformations des boucles. Fiser *et al.* décrivent une méthode rudimentaire d'échantillonnage où les atomes formant une boucle initiale étendue sont aléatoirement déplacés d'au plus 5 Å, puis sont passés par une minimisation par gra-

dient conjugué et deux simulations de recuits simulés avec et sans le corps de la protéine [67]. Après une dernière minimisation de l'énergie, la structure de plus basse énergie est sélectionnée. Ici, les étapes suivant le premier déplacement aléatoire ne servent qu'à minimiser la structure, les dynamiques moléculaires n'étant exécutées que pendant 40 ps. L'échantillonnage de l'espace des conformations repose donc entièrement sur les pas aléatoires dans l'espace cartésien. Plus récemment, l'idée de mouvements aléatoires dans l'espace cartésien a été reprise dans une méthode où la boucle est subdivisée en fragments solides séparés par des angles dièdres [160]. Ces fragments sont aléatoirement disposés dans l'espace, puis à l'aide d'un tableau de distances inter-atomiques minimales et maximales acceptables, les fragments sont itérativement déplacés jusqu'à ce qu'ils respectent les contraintes stériques et de liaisons covalentes de la boucle.

Une méthode plus efficace d'échantillonnage du nom de *Protein Local Optimization Program* (PLOP) a été développée par le groupe de Friesner en se basant sur l'échantillonnage dans l'espace des angles dièdres de la boucle [105]. Une base de données des angles ϕ et ψ de l'ossature que chaque a.a. peut occuper a préalablement été construite à partir de structures empiriques de haute résolutions. L'échantillonnage large de l'espace des angles ϕ et ψ de la boucle est assuré par le choix de conformations d'a.a. dissimilaires à partir de la base de données. Les boucles sont construites en deux fragments attachés aux a.a. d'ancrage les flanquant. Les conformations de fragments de boucles ainsi générées sont sommairement évaluées et celles entrant en contact stérique avec la protéine sont éliminés. Pour fermer les boucles, la distance entre chaque a.a. de fermeture des fragments en N-terminale et en C-terminale est évaluée et celles dont la distance est plus petite que 0.5 Å sont considérées comme étant des boucles closes. Le nombre de boucles conservées pour l'évaluation énergétique est tout d'abord diminué à l'aide d'un algorithme de clustering, puis les chaînes latérales sont ajoutées [104], puis l'énergie de la conformation est minimisée. La méthode est dite hiérarchique parce qu'elle est répétée à plusieurs reprises en imposant des contraintes sur l'espace des conformations de boucles échantillonnées basées sur les boucles de basse énergie de l'itération précédente. Une version ultérieure de la méthode PLOP [285] rajoute une étape hiérarchique supplémentaire avec des exécutions sur des sous-boucles, c'est à dire des boucles dont des

a.a. en N-terminale et/ou en C-terminale sont fixé en se basant sur des conformations de basse énergie trouvées à l'étape précédente. Dans [286] un terme diélectrique variable est ajouté puis le modèle de solvatation AGBNP a été implémenté [65]. Finalement, la méthode a été adaptée pour le cas où la conformation des chaînes latérales de tous les a.a. est inconnue [224] .

Spassov *et al.* proposent une autre méthode de recherche de conformations systématiques dans l'espace des angles dièdres [229]. Comme dans PLOP, les boucles sont construites en deux demi-fragments attachés au a.a. d'ancrages en N-terminale et C-terminale [105]. Cependant, les fragments générés utilisent une base de données des angles ϕ d'un a.a. et $\psi - 1$ de l'a.a. précédent. Cette base de données couvre donc l'espace des angles dièdres de chaque côté de la liaison entre le groupe carboxyle et l'azote de l'ossature des a.a. au lieu du traditionnel diagramme de Ramachandran [206] caractérisant les angles dièdres de chaque côté du carbone-α. Aussi, au lieu de considérer toutes les paires de demi-fragments formant naturellement une boucle close, toutes les paires de fragments dont l'énergie interne est suffisamment basse sont passées à une étape de minimisation permettant d'obtenir une boucle fermée.

La fermeture d'une boucle à l'aide de technique de dynamique moléculaire peut être coûteuse en temps de calcul. Une méthode a donc été proposée pour accélérer la fermeture de boucles par cinématique inversée [22]. Utilisée en robotique pour déterminer les mouvements angulaires d'un bras articulé nécessaires pour obtenir une translation d'une extrémité du bras, la méthode du nom de *Cyclic coordinate descent* (CCD) a été adaptée au mouvement de l'extrémité d'une boucle dans l'espace des angles dièdres de la boucle afin d'obtenir la fermeture. Dans leur état initial, les boucles sont générées aléatoirement avec leur extrémité N-terminale attachée au corps de la protéine et l'extrémité C-terminale est numériquement rattachée à ses a.a. d'ancrage.

Toujours basée sur la théorie de cinématique, une méthode analytique d'échantillonnage a été développée pour les boucles fermées en résolvant le problème de la rotation de 6 angles situés dans une boucle où les extrémités sont fixées [40, 164]. La méthode sélectionne aléatoirement 3 carbone-α dans la boucles et effectue une rotation des deux segments de boucles ainsi délimités. Les angles ϕ et ψ des 3 carbone-α correspondant

aux rotations de segments sont obtenus analytiquement par une extension du polynôme de 16^{ieme} degré du problème d'une boucle de 3 a.a. [265].

Il est aussi possible de traiter les problèmes de fermeture de la boucle dès la génération d'une conformation de boucle. Dans RAPPER [46, 51], les boucles sont construites par l'addition d'a.a. en N-terminale à l'aide d'une base de donnée d'angles ϕ et ψ. Après avoir été ajoutées au fragment de boucle en N-terminale,les conformations d'a.a. qui poussent la boucle à une distance trop éloignée du point d'ancrage en C-terminale sont éliminées.

Afin de pouvoir échantillonner la surface énergétique des boucles dans le référentiel des 6 angles dièdres de la solution analytique de Wedemeyer *et al.* [265], un algorithme de Monte-Carlo a été développé pour respecter la réversibilité microscopique et l'ergodicité [42, 171]. Le *Local Move Monte-Carlo* (LMMC) est tout d'abord appliqué à très haute température suivant un critère de Métropolis pour échantillonner largement l'espace des conformations, puis est soumis à un recuit simulé sur un nombre réduit de conformations représentant l'ensemble de l'espace. La méthode est par contre très coûteuse, demandant plus de 5 millions de pas de Monte-Carlo en recuite simulé par conformation.

Finalement, Olson *et al.* ont utilisé deux méthodes et comparé leur efficacité [189]. La première est l'algorithme de Monte-Carlo MONSSTER [130] définissant une grille spatiale de points distancés par 1.45 Å et où chaque acide aminé est représenté par une bille. Les pas de Monte-Carlo sont des déplacements cartésiens dans un espace discrétisé et l'échantillonnage est élargie par l'usage d'échanges de répliques à différentes températures. La deuxième méthode employée est une dynamique moléculaire avec échange de réplique (DMER) sur la section boucle des protéines.

9.2.2.1 Limitations

À l'exception de l'utilisation de DMER d'Olson *et al.* , toutes les méthodes *ab initio* énumérées précédemment on en commun le même défaut : elles essaient toutes d'échantillonner l'entièreté de l'espace de configurations des boucles. Puisque cet espace croît exponentiellement avec la longueur des séquences des boucles, ces méthodes font face au

choix d'échantillonner moins de conformations ou d'exécuter leur méthode des temps exponentiellement plus longs [7, 286]. Avec leur pas d'intégration d'au plus quelques femtosecondes, les méthodes de dynamique moléculaires peuvent traverser la surface énergétique d'une boucle pour ainsi dire sans discontinuité et sont donc mieux adaptées à découvrir et parcourir des entonnoirs de repliement [147] si ceux-ci existent. Cependant, les méthodes basées sur la dynamique moléculaire sont les plus lentes.

9.2.3 Potentiels énergétiques

L'échantillonnage de l'espace des conformations ne représente que la moitié du problème de la prédiction de la structure des boucles, l'autre moitié étant remplie par la classification des boucles échantillonnées. De nombreux potentiels énergétiques et statistiques ont été utilisés dans les articles précédemment cités et méritent d'être énumérés.

Les trois grands potentiels énergétiques, AMBER [38, 203, 262, 268], CHARMM [19] et OPLS [116, 117, 121], ont tous été utilisés ou adaptés par un groupe de recherche ou un autre.

AMBER a été adopté par le groupe de Blundell [46, 51] couplé à un modèle de solvattaion généralisé de Born [204]. Le groupe a aussi testé le potentiel RAPDF [214], plus rapide, mais moins discriminant.

Le potentiel CHARMM a été utilisé par plusieurs groupes, en totalité ou en partie. Fiser *et al.* n'utiliseront que les termes de liaisons et dièdres, optant pour un potentiel statistique pour les interactions de longue portée [67]. Xiange *et al.* [272] n'utilisent de CHARMM22 que les termes des angles dièdres et des rayons de van der Waals dans leur méthode tentant d'intégrer l'aspect entropique des boucles dans le terme énergétique. CHARM a aussi été utilisé par d'autres groupes [189, 229].

Pour leur part, le groupe de Friesner a adopté le potentiel OPLS qu'ils ont adapté [121] dans leurs travaux sur PLOP [105, 286].

Quelques potentiels ont été développés spécifiquement pour discriminer les formes natives des autres conformations de boucles. DFIRE [283] est un potentiel statistique qui lorsqu'appliqué sur un jeu de leurres est aussi précis qu'AMBER [268] avec un modèle de solvatation de Born généralisée [204] et que le potentiel OPLS [116] avec un modèle

de solvatation similaire [76], mais surpasse CHARMM [19] avec plusieurs modèles de solvatation [144, 218].

Un autre potentiel développé pour les boucles est le *Hydrophobic potential of mean force* (HPMF) de Lin *et al.* [152]. Cette méthode, couplée au potentiel AMBER [262] et à modèle de solvatation généralisé de Born [191] obtient des résultats supérieurs à DFIRE et OPLS sur un jeu de leurres [151].

Finalement, il a été proposé que l'information provenant de plusieurs potentiels énergétique peut être combinés à des fin de raffinement de prédictions. Sur un jeu de leurres, la méthode du consensus Pareto-optimale [150] a démontré qu'elle pouvait invariablement attribuer un score supérieur aux potentiels sur lesquels le consensus est construit.

9.3 ART-boucle

Nous nous sommes inspirés de notre expérience dans la simulation du repliement des protéines pour adapter la Technique d'Activation-Relaxation ART-nouveau [163, 176] au problème du repliement des boucles. La méthode permet d'échantillonner la surface énergétique d'un système de façon continue, traversant minima énergétiques et points de selle et permettant entre autre d'explorer les chemins de repliement de protéines [232, 267]. Nous posons l'hypothèse de départ que la méthode ART-boucle peut échantillonner efficacement la surface énergétique d'une boucle de grande taille et trouver les minima énergétiques d'intérêt sans avoir à explorer exhaustivement la surface puisque la méthode rejette systématiquement les minimums locaux non-physiques.

9.3.1 Méthode ART

La méthode ART est une méthode générique de parcours des surfaces énergétiques qui a trouvé application dans plusieurs systèmes de matière condensée. Sous la forme d'ART-nouveau, la méthode a démontré son efficacité dans l'étude de l'agrégation de peptides de courtes tailles [17, 54, 177, 216], mais aussi au repliement des protéines [266, 267] jusqu'à une taille de 60 a.a. [232]. Certaines variations de la méthode ont trouvé leur utilité dans les matériaux amorphes [59, 60, 110, 124, 258] ou encore de la dynamique

des mouvements d'éléments de structure secondaire dans les protéines [56, 57]. Au coeur de la méthode ART se trouvent quatre étapes permettant de passer d'un minimum local de la surface énergétique à un autre en passant par un point de selle de premier degré : la sortie du bassin harmonique, la convergence vers un point de selle, le saut du point de selle suivit de la convergence vers le nouveau minimum local, et finalement l'acceptation ou refus de la nouvelle structure.

9.3.1.1 Sortie du bassin harmonique

Une itération de la méthode ART commence avec une structure de notre système située dans un minimum local de la surface énergétique. À ce point, le gradient de force à une valeur nulle et la courbure de la surface énergétique est positive dans toutes les directions, d'où la désignation de bassin harmonique. Afin de détecter un point de selle, la position des atomes du système est perturbée dans une direction aléatoire. À chaque pas pris dans la direction aléatoire, quelques pas de minimisation sont pris pour atténuer la composante dure des angles et des longueurs des liaisons covalentes. La courbure de la surface énergétique est examinée par diagonalisation de la matrice Hessienne à l'aide de l'algorithme de Lanczos [131] qui ne nécessite qu'au plus une vingtaine d'appels au potentiel énergétique. La méthode d'ordre O(N) permet d'extraire les vecteurs propres de faible valeur propre de la matrice Hessienne. Lorsqu'un vecteur propre à valeur propre négative est détecté, l'algorithme passe en phase convergence.

9.3.1.2 Convergence vers un point de selle

Le vecteur propre à valeur propre négative pointe dans la direction de courbure d'un point de selle de premier degré présenté à la figure 9.1. Pour l'atteindre, les atomes du système sont poussés dans la direction du vecteur propre. À chaque pas, l'énergie du système est évaluée et des pas de minimisation de l'énergie sont pris dans la direction de la composante perpendiculaire au vecteur propre.Cette minimisation de l'énergie dans la direction perpendiculaire a pour effet de rapprocher la structure du sillon menant au point de selle et généralement diminue l'énergie de la conformation comparativement à

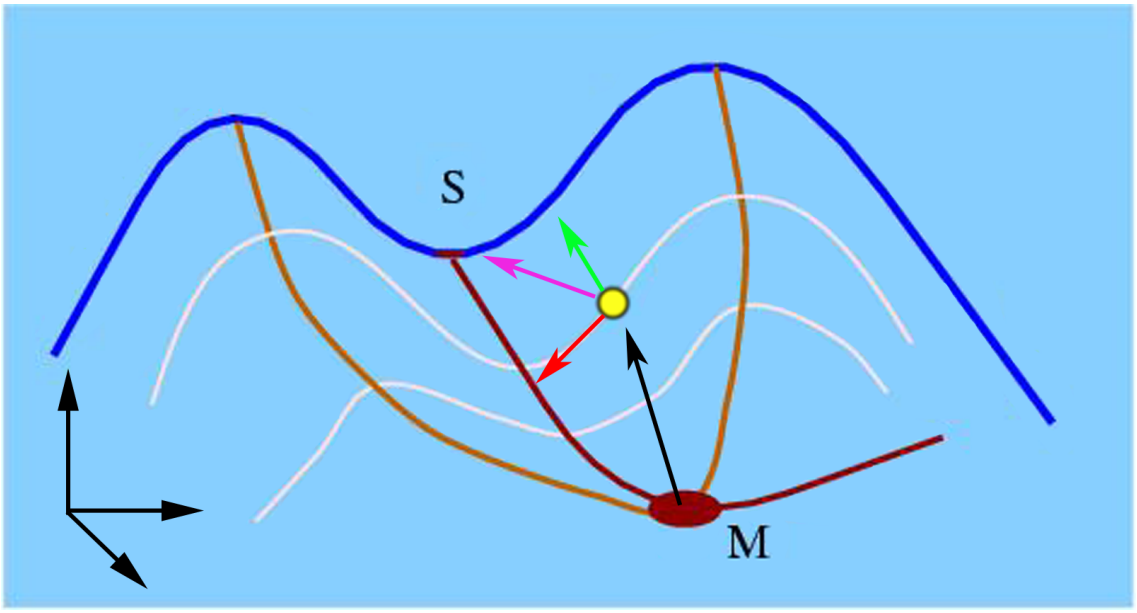

Figure 9.1 – Exemple de trajet emprunté par ART-nouveau pour atteindre un point de selle à partir d'un minimum. Suivant une direction aléatoire, la configuration est poussée hors du bassin harmonique (flèche noire) jusqu'au point où un vecteur propre de valeur propre négative est détecté (jaune). La configuration est alors poussée dans le sens de ce vecteur propre (flèche verte) tout en minimisant sont énergie dans l'hyperplan perpendiculaire (flèche rouge) résultant en un mouvement menant au point de selle (flèche rose). Image tirée de [230]

la structure générée à l'étape 9.3.1.1. Tout comme les minima des bassins harmoniques, les points de selle de premier ordre sont identifiés par un vecteur de force dont la norme est nulle. Puisque la méthode n'a pas pour but de caractériser aussi finement les points de selles que les minima énergétiques, la convergence vers un point de selle est considérée atteinte lorsque la composante parallèle au vecteur propre du vecteur de force est plus petite que 1.0 kcal/(mol$\times$Å) et que la composante perpendiculaire est plus petite que 2.0 kcal/(mol$\times$Å).

9.3.1.3 Saut du point de selle et convergence vers le nouveau minimum local

L'étape suivante consiste à minimiser l'énergie de la structure pour découvrir un nouveau minimum local. Pour s'assurer que la minimisation ne retourne pas la conformation

du système au minimum précédent, la conformation est légèrement poussée dans la direction s'éloignant de ce dernier. L'algorithme de minimisation présentement utilisé est la dynamique moléculaire amortie.

9.3.1.4 Acceptation ou refus de la nouvelle structure

La structure du nouveau minimum énergétique est finalement acceptée ou rejetée selon un critère de Metropolis [170] :

$$p_{accepter} = min(1, \exp \frac{-\Delta E}{k_{\mathrm{B}} \cdot T}), \tag{9.2}$$

où $P_{accepter}$ est la probabilité d'accepter la nouvelle structure, ΔE est la différence entre l'énergie de l'état final et celui de l'état initial, k_{B} est la constante de Boltzmann et T est la température simulée. Une différence d'énergie négative, donc un nouveau minimum plus stable que le précédent, mène à une acceptation automatique de la nouvelle structure. Puisque la méthode n'échantillonne pas la thermodynamique du système, le facteur de température T n'a pas d'effet sur les points de selles traversés et les nouveaux minima découverts, mais seulement sur la probabilité d'accepter un nouveau minimum.

9.3.2 Modifications apportées à ART-nouveau

Des essais préliminaires d'application de la méthode ART-nouveau sur des boucles ont démontré que, comparativement au problème du repliement des protéines entières, il était plus difficile d'échantillonner des minima distants les uns des autres dans un système où les deux extrémités sont attachées à des points d'ancrage. Les paramètres des deux premières étapes de la méthode ont donc été réajustés en fonction de la longueur des boucles simulées ce qui rend les comparaison de résultats parfois difficile.

Aussi, dans le souci d'échantillonner plus largement l'espace de conformations et d'assurer un recouvrement entre les simulations lancées à partir de structures initiales différentes, nous avons muni ART-boucle d'un algorithme de température variable inspiré de l'algorithme de Berendsen [14]. Notre algorithme ajuste la température en fonction de la différence entre la probabilité moyenne récente d'accepter des événements et la

probabilité désirée d'accepter un nouvel événement. Ainsi, lorsque le système se trouve pris dans un minimum local profond avec peu de probabilité de s'en sortir, la température est lentement augmentée jusqu'à atteindre un point où la probabilité d'accepter un nouveau minimum est égale à la probabilité désirée. Pour éviter l'effet contraire, c'est-à-dire de rester pris dans un bassin de minima où la différence d'énergie entre chaque minimum est faible et la probabilité d'accepter un nouveau minimum est élevée, on spécifie une température de Metropolis minimale.

La méthode est donc similaire au recuit simulé où la température oscille à une fréquence régulière, la différence étant que la fréquence d'oscillation de notre bain de température est hautement corrélée avec la différence d'énergie des minima locaux échantillonnés.

9.3.3 Potentiel OPEP

Le potentiel énergétique qui a été choisi pour nos simulations est le *Optimized Potential for Efficient peptide-structure Prediction* (OPEP) [52] que nous avons utilisé précédemment dans les simulations d'ART-nouveau sur des protéines ainsi que dans plusieurs simulations de dynamique moléculaire [26, 39, 137, 138, 179]. Dans OPEP, les a.a. ont une représentation réduite où toutes les chaînes latérales à l'exception des prolines sont représentées par une seule bille. Les atomes d'hydrogènes sur les C-α et sur l'anneau des prolines sont traités implicitement, tout comme le solvant.

Quelques modifications ont été apportées au potentiel pour l'accélérer. Tout d'abord, puisque seuls les atomes des boucles simulées peuvent bouger, le corps des protéines est fixé et les interactions de courte ou longue portée entre les atomes du corps ne sont pas calculése. De plus, le volume ellipsoïde d'espace que la boucle peut théoriquement occuper est pré-calculé et les atomes du corps de la protéine situés à une distance de plus de 16 Å de ce volume sont entièrement retirés de la fonction de potentiel. Ces modifications n'ont aucun effet sur la précision du calcul de force et d'énergie.

9.4 Conclusion

Dans l'article présenté au chapitre 11, nous présentons les résultats de l'échantillonnage et de la prédiction de structures à l'aide notre méthode ART modifiée de 3 jeux de boucles dont la taille varie entre 8 a.a. et 20 a.a.

CHAPITRE 10

CONTRIBUTION DES AUTEURS À L'ARTICLE SUR PRÉDICTION DE BOUCLES

- Jean-François St-Pierre a écrit la première version du manuscrit et a fait les manipulations suivantes :
  - Effectué les modifications au potentiel OPEP pour traiter le corps des protéines comme un potentiel d'arrière-plan pour les boucles flexibles.
  - Calibré les paramètres d'ART-nouveau pour obtenir des pas d'intégration de taille raisonnable pour chaque longueur de boucle simulée,
  - Implémenté l'algorithme de bain de température permettant d'obtenir une probabilité constante d'acceptation du critère de Métropolis,
  - Défini la métrique du rang d'agrégat (Cluster rank) permettant d'évaluer la taille de l'espace conformationnel échantillonné,
  - Exécuté les simulations et analysé les résultats.

  ,

- Normand Mousseau a supervisé le travail et contribué aux révisions et corrections du manuscrit

CHAPITRE 11

ARTICLE : LARGE LOOP CONFORMATION SAMPLING USING THE ACTIVATION RELAXATION TECHNIQUE ART-NOUVEAU METHOD.

Jean-François St-Pierre
Département de Physique and Regroupement Québécois sur les Matériaux de Pointe, Université de Montréal, C.P. 6128, succursale centre-ville, Montréal (Québec) Canada H3C 3J7
Normand Mousseau
Département de Physique and Regroupement Québécois sur les Matériaux de Pointe, Université de Montréal, C.P. 6128, succursale centre-ville, Montréal (Québec) Canada H3C 3J7

11.1 Abstract

We present an adaptation of the ART-nouveau energy surface sampling method to the problem of loop structure prediction. This method, previously used to study protein folding pathways and peptide aggregation, is well suited to the problem of sampling the conformation space of large loops by targeting probable folding pathways instead of sampling exhaustively that space. The number of sampled conformations needed by ART nouveau to find the global energy minimum for a loop was found to scale linearly with the sequence length of the loop for loops between 8 and about 20 amino acids. Considering the linear scaling dependance of the computation cost on the loop sequence length for sampling new conformations, we estimate the total computational cost of sampling larger loops to scale quadratically compared to the exponential scaling of exhaustive search methods.

11.2 Introduction

Protein structure prediction has had much success in predicting the ordered alpha and beta secondary structure components as, often, sequence alone can determine their conformation, especially with the help of previously known homologous protein structures [135]. Loop regions, however, adopt conformations that are not as easily predicted since they lack strict arrangement rules. Prediction is also complicated by the fact that these regions often show intrinsic mobility and, therefore, are not as well resolved by x-ray diffraction crystallography and nuclear magnetic resonance. Over the last ten years, considerable efforts have gone into developing and refining the prediction ability of three classes of loop-sampling algorithms. The fastest of these are knowledge-based methods that rely on structure databases and sequence homology to generate conformations [66, 90, 148, 200] and sport an accuracy as low as 2 Å for sequences length of up to 20 amino acids (a.a.) [33]. *Ab initio* methods, which build loop fragments from scratch and sample the conformation space in search of the lowest energy or best scoring conformations [22, 40, 46, 51, 67, 189, 211, 224, 272, 286], are more demanding computationally but they tend to lead to better results independent of the loop sequence. The last class of loop-sampling algorithms are hybrid methods that combine both algorithms for specific sequences [48, 249]. Although many of the previous methods were tested on independent datasets, a quantitative comparison of these methods is presented in Table 1 of Arnautova *et al.* [7].

Here, we focus on *ab initio* approaches. These all share a limitation on the maximum loop size they can effectively sample due to the exponential increase in conformational space with loop length. Most studies, until now, have been done on loop datasets of 13 a.a. or less. Their cost, in terms of computational effort, tend to increase exponentially with loop length, following the growth in conformational space [7, 286]. Sampling, moreover, is made more difficult by the constraints imposed by the fixed loop endings and is akin to protein folding in a confined environment, a problem which remains challenging.

In this work, we investigate the loop structure prediction problem by using the ART nouveau [163] energy landscape sampling method, which has been used to study protein

folding pathways and peptide aggregation of systems of up to 60 a.a [54, 232]. We show that the method, although somewhat heavy for short loops of 8 and 12 a.a., can handle large loops of 20 a.a. or more in a very competitive manner, providing both extensive configurational sampling as well as low-energy structures.

11.3 Methods

11.3.1 ART nouveau potential energy landscape exploration method

We adapted the Activation Relaxation Technique, ART nouveau [163, 232, 267], to the exploration of the constrained potential energy landscape of loop segments covalently bound at both extremities to a fixed protein body. ART nouveau is an iterative process consisting of four steps through which the conformation of an atomic system is moved from one local minimum on the potential energy surface to another nearby minimum passing through an adjacent first-order saddle point. (1) Starting from a local-energy minimum, the conformation is first deformed in a random direction taken in the 3N-dimensional loop space. It is pushed along this direction until it leaves the harmonic basin and the lowest eigenvector of the Hessian matrix become negative. (2) The conformation is pushed along the direction of negative curvature while its energy is minised in the perpendicular hyperplane until the force falls below a small threshold, indicating that the system has converged onto a first-order saddle point. (3) The conformation is then pushed slightly over this saddle point and relaxed, using a damped molecular dynamics, into a new energy minimum. (4) The move from the initial ot the final minimum is then accepted or rejected using a Metropolis criterion [170].

To avoid N^3 operations, the lowest eigenvalue and corresponding eigenvector are computed using the Lanczós algorithm with typically less than 16 force evaluations per step. Implementation details of this very competitive algorithm [162] can be found in Refs. [168, 232].

ART nouveau has been used to characterize the energy landscape of complex systems [120] as well as generate folding trajectories [232]. Here, we are interested in sampling the landscape of large loops and identify low-energy structures. To do so, we

elected to use a Metropolis algorithm [170] with adaptative temperature. In the original algorithm [162, 163], the acceptance probability is given by

$$P_a = min(1, exp(\frac{-\Delta E_c}{k_B T})), \tag{11.1}$$

where ΔE is the energy difference between consecutive minima $c-1$ and c, T is the Metropolis temperature and k_B is the Boltzmann constant. To keep the probability P_a constant and avoid getting trapped into deep basins, the Metropolis temperature here is adjusted on the fly by applying a Berendsen bath [14] on the acceptance probability of conformation c :

$$P_{avg}(c) = P_{avg}(c-1) + \frac{P_a - P_{avg}(c-1)}{\tau}, \tag{11.2}$$

where τ is the coupling parameter and P_{avg} is average effective acceptance rate over the previous w conformations defined as :

$$P_{avg}(c) = \frac{1}{w} \sum_{i=c-w}^{c} min(1, exp(\frac{-\Delta E_i}{k_B T})), \tag{11.3}$$

The Metropolis temperature for a given $P_{avg}(c)$ is solved iteratively. For our simulations, we selected a window size w of 15 conformations and a coupling τ of 20 with a target acceptance probability P_a of 50%. We also set a minimum Metropolis temperature of 300 K to prevent the system from freezing in shallow basins where the difference in energy between neighboring conformations is small.

For ART nouveau's exploration, the protein is divided into a fixed protein body and the flexible loop regions. Atoms in the fixed region were assigned based on the experimentally-derived native conformations. This procedure is similar to that used in previous studies of loop flexibility such as Refs. [67, 105, 145].

11.3.2 Dataset

In this study, we used two previously published datasets for the 8 and 12 a.a. loops, respectively. The first set, from Olson *et al.* [189], is a subset of a large database [23] and is composed of 25 8 amino acid loops from 22 proteins. The second set is a subset

of 38 loops of length 12 a.a. from the Fiser *et al.* dataset [67]. This later subset was used in a number of publications, either in part or as a whole [46, 146, 164, 211, 229].

Initial loop structures for the ART nouveau were generated by stretching the loop into an arc of length 3.25 Å times the number of loop amino acids using a harmonic potential applied onto the a.a. center of mass. Five stretched loop conformations were generated per protein with an angle between arc supporting planes of 30 degrees. Between 1 and 5 initial conformations were selected for the loop regions of size 8 a.a. and between 2 and 3 for loops of size 12 a.a., based on the potential energy and rejecting loops segments clashing with the protein body. By using stretched structures as initial conformation, we ensure that simulations are starting far away from the global energy minimum, decreasing possible biases of the initial state.

For the 8 a.a. loops, the standard Metropolis criterion, with a Metropolis temperature of 700 K, was used to accept or reject new local minima. For the 12 a.a. loops, we found that using the constant probability rate of Equation 11.3 yielded a wider sampling of the conformation space and therefore was used for these loops.

For both loop sets, the selected conformations were given 5 to 10 days of simulation time on single-core Intel Xeon 2.8 Ghz microprocessors. Simulation details are presented in Table 11.I and Table 11.II. In addition, 5 batches of preliminary simulations were executed to optimise the ART parameters on the 12 a.a. loop set. Although the results of these preliminary simulations are not included in the analysis presented in the next section, the search for a global energy minimum was done on all generated conformations including preliminary and test simulations.

For longer loop evaluation, we constructed a dataset of 10 proteins using the PISCES server [260] among all proteins with an X-ray structure of resolution lower than 2.0 Å, a maximum R-factor of 0.3, a sequence identity lower then 25% and a sequence length between 140 and 600 a.a. Regions with no defined secondary structure elements were identified using DSSP [119]. When the loop was found in a multimeric protein, the first chain containing the loop was used and it was verified that the loops did not interact with the removed meres. Due to the difficulty in finding long loop regions completely devoid of secondary structure, the 19 and 20 a.a. loops presented in Table 11.III have up to 3 a.a.

in bend or hydrogen bounded turn conformation, with the exception of 1ofl which also has 2 a.a. in α-helix conformation. Simulations were executed for 20 days on the same machines as above.

For analysis, we use the global definition of root mean square deviation (RMSD) in which the fixed portions of the proteins are superimposed before calculating the RMSD of the flexible loop region alone without further translations or rotations of the protein. Only the backbone atoms of the loop are included in the RMSD calculations.

11.3.3 OPEP force field

We have modified the Optimized Potential for Efficient peptide-structure Prediction (OPEP) [52] coupled to ART nouveau to allow faster sampling of loop regions. OPEP is a coarse grained potential for which all amino acids side chains are represented by a unified bead except for glycine and proline. All backbone heavy atoms and the hydrogen atom bound to the backbone nitrogen are also represented. To increase the efficiency of the energy computation, all interactions involving two atoms outside of the loop region were removed from the force field. These fixed protein body atoms formed a constant background potential for the docking of the free loop atoms. The forces and energy between the loop's atoms and the rest of the protein are calculated as usual, but the protein's body atoms are not allowed to change conformation. This potential was successfully used to study protein folding [39, 232, 267] and peptide aggregation [31, 54, 179, 216]. OPEP was recently compared to the AMBER99SB and OPLS-AA all-atom force field on two small peptides by parallel tempering metadynamics and was found to be in agreement with the two detailed potentials and could reproduce the features of the free-energy landscape at a much lower computational cost [12].

11.4 Results

The following analysis of the ART-nouveau method is divided into 3 sections. First, we evaluate the ability of the method and the OPEP potential to sample conformation of low energy in the vicinity of the native structure providing a proper score. Then, the

ability of the method to sample the conformational space and to find the global energy minima regardless of the native structure is presented for the short loops of 8 and 12 a.a. and the long loops of 19 and 20 a.a. Finally, we evaluate the scaling performance of the method as a function of loop-length.

11.4.1 Conformation scoring

The ability of the OPEP potential to find low energy loop conformations compatible with the crystallographic structure is presented in Table 11.I and Table 11.II. We find that for the 8 a.a. dataset, our results show a comparable accuracy to low-energy structures to that of Olson *et al.* [189]. More precisely, the lowest-energy conformations for our simulations are, on average, 3.50 Å (st. dev. 1.17 Å) away from the native structure, as compared with 3.89 Å for their latice-based work and 3.14 Å for their all-atom MD simulations. [189]

Even though the trajectories sample conformations within 1.75 Å of the native state for the longer 12 a.a loops, the lowest-energy structures show an average RMSD with respect to the native structure of 5.60 Å (st. dev. 2.53 Å). This discrepancy is due to the coarse-grained nature of the OPEP potential, which does not discriminate sufficiently between various steric packing, as well as to the rigid spatial representation of the non-loop protein regions which prevents structures from adopting the optimal conformations.

To test the impact of these two limitations, we induce flexibility by reconstructing the coarse-grained side-chains of the whole proteins using the SCWRL4 automated tool [133], then rescored the all-atom representations using dFIRE [284]. This analysis show that lower RMSD low energy structures were sampled with ART nouveau but improperly scored by the modified OPEP potential. The resulting average RMSD of the best scored conformations to the native structure is improved to 4.27 Å (st. dev. 1.87 Å), essentially identical to the average 4.32 Å obtained by Zhang *et al.* with a similar protocol [283] and slightly higher than other ab initio methods including FALCm4 that scores using dFIRE potential with 3.84 Å RMSD [146] and LOOPER with 4.08 Å RMSD [229]. Methods making use of predefined structures, such as ROSETTA do, of course, better : 3.62 Å RMSD for ROSETTA [211] and 2.3 Å RMSD for ROSETTA

with a kinetic closure algorithm [164]

The efficiency of the reconstruction and rescoring of the ART nouveau-generated datasets suggests that even though OPEP could not fully discriminate between the various energy minima on the energy landscape, ART nouveau samples the configurational space rather efficiently. This is confirmed by the fact that the smallest RMSD between trajectories and the experimentally-derived native structure is only, on average,1.23 Å and 1.75 Å, for the 8 a.a. (Table 11.I) and 12 a.a. loops (Table 11.II), respectively.

We observe similar results for the dataset composed of loops of 19 and 20 a.a, the RMSD of the conformations of lowest energy to the native conformations averages to 7.17 Å (st. dev. 3.21 Å) (Table 11.III), which is comparable to, or better than previously published results on loops of the same size, i.e., $\sim$7 Å for CABS, $\sim$9 Å for Rosetta and $\sim$12 Å for MODELLER on the Jamroz and Kolinski dataset [106], and 10.49 Å for MODELLER, 10.64 Å for RAPPER, 11.14 Å for PLOP and 7.64 Å for Original FREAD on the Choi and Deane dataset [33]. However, the lowest RMSD to the native structure observed in these 20 a.a. loop studies is of 2.98 Å (st. dev. 2.37 Å), consistently lower then the lowest RMSD for the above studies, a value that ranges between $\sim$ 4-5 Å [106] and 5.20-8.43 Å average RMSD [33]. This suggests that our prediction precision for large loops is competitive with previously published methods and that the ART method can sample conformations closer to the native structure even when this structure is not the global energy minimum of the employed potential.

The focus of this project is to evaluate the ability of the ART method to sample a wide range of loop structures and identify low energy conformations on a potential energy surface. We therefore leave aside the issue of proper scoring and prediction capacities, which are entirely dependent on the chosen energy potential, to analyze sampling capacities of ART-nouveau which is potential-independent. In the following analysis, RMSD values are calculated with respect to the global low energy conformations of the energy potential instead of the native conformation. For this purpose, the low computational cost of the OPEP potentiel is well suited since it allows for longer simulation times and wider sampling of the conformational space to identify the global low energy structures.

11.4.2 Exploration of the conformation space for the 8 a.a. and 12 a.a. loop dataset

ART nouveau's sampling ability can be evaluated by characterizing the volume of the conformation space sampled by the method and its ability to find the conformations of global lowest energy on the OPEP potential energy surface. In particular, proper care must be taken to insure that conformations are not trapped in local energy minima, away from the native state.

For the 8 a.a. loops, we see that in the 23 loops for which more than one simulation was executed, it was possible to recover he same lowest energy minimum, as defined with OPEP, at least two times or more in 18 of them (Table 11.I), suggesting that the exploration of the conformational space is sufficiently thorough to reach regularly the global-energy minimum. For the 12 a.a. loops, analysis of our preliminary simulations shows that in 6 cases, conformations of lowest energy were also sampled in the preliminary simulations, but not in the production simulations. The lowest energy conformations of these 6 preliminary simulations were used for further analysis. When considering all calibration and production runs, the global energy minimum was found two times or more for 31 of the 38 loops, 19 of which were found twice or more in the production runs of Table 11.II.

To understand how sampling occurs, we plot the evolution of the RMSD, averaged over three different subsets as a function of event number for both sets of loops in Figure 11.1. The black curve shows the averaged RMSD of the current ART event conformation, computed with respect to its respective global energy configuration. We see that this measure reaches a plateau after roughly 500 steps and remains around 2.4 Å for the 8 a.a. loops and 4.4 Å for the 12 a.a. loops. These distances are relatively near the maximum deformation distance achieved by stretching the loop, 4.63 and 7.49 Å respectively, which indicates that each trajectory samples widely the energy landscape.

The red curve shows the evolution of the RMSD of the lowest energy conformation identified, or TOP RMSD, for each trajectory launched and averaged over all simulations. This quantity shows how a group of trajectory can be used to identify the lowest-energy basin. When simulations are examined individually, we first see that not all runs

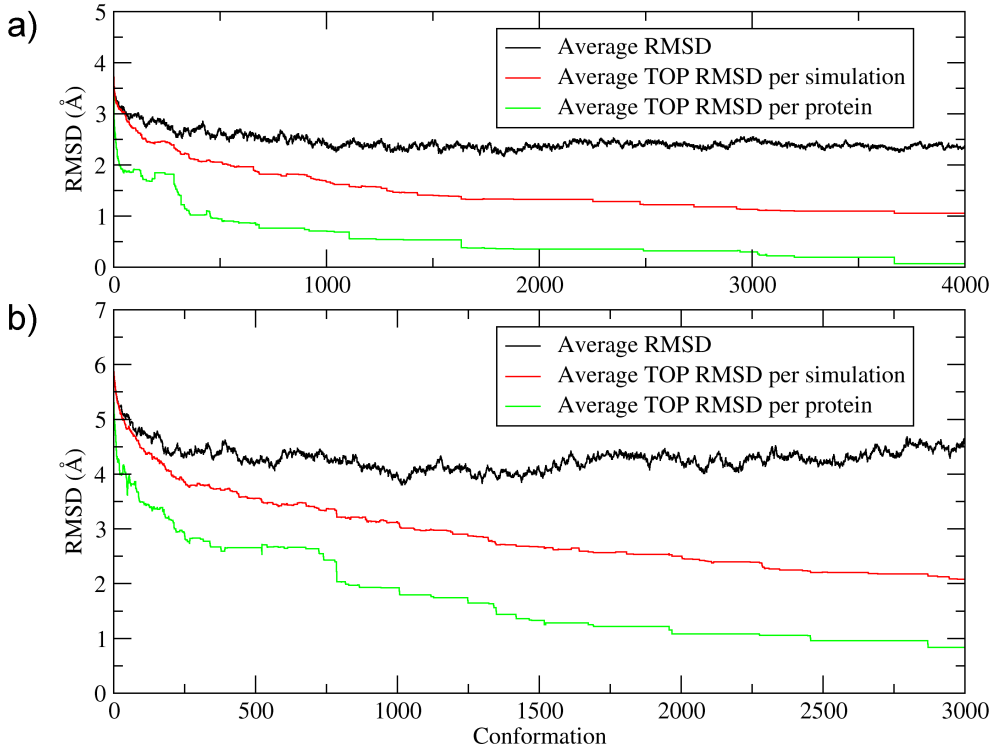

Figure 11.1 – RMSD evolution for the (a) 8 a.a. and (b) 12 a.a. loops. In black, the current conformation's RMSD of each simulation is calculated to the global energy minimum conformation of the sampled protein. The average TOP RMSD is calculated between the global energy minimum of a system and the lowest energy conformation found so far per simulation (red) or per protein (green).

for a given loop sequence sample the lowest energy conformation but that, overall, the probability of passing nearby this conformation increases with the number of steps, albeit at a constantly slower pace. After 4000 and 3000 steps, respectively, the 8 a.a. and 12 a.a. loop simulation sets reach an average *per simulation* RMSD value of 1.05 Å and 2.05 Å .

It is useful to follow convergence of the full set of simulations. In the same figure, we combine all runs for each loop and plot the overall average TOP RMSD for each sequence (green curve). As expected, we see a faster convergence in the first ART steps, leading to an average RMSD to the global energy structure of 0.1 Å after 3700-4000 conformations for the 8 a.a. loops and 1.25 Å after 1600 conformations, and 0.84 Å after 3000 conformations, for the 12 a.a. loops.

For the 12 a.a. loops, we see a lowest average RMSD of 1.0 Å because not all sequences manange to find their global energy-minimum structure in the production runs here presented. In some cases, these structures were only identified in the preliminary simulations.

While not all runs for one protein converge to the global energy minimum, all individual runs do overlap, suggesting that longer runs would allow all trajectories to find the global energy minimum. To see this, conformations were divided into clusters of maximum RMSD of 0.6 Å between each member using a hierarchical clustering algorithm and an average linkage clustering criteria [82]. The center of each cluster for a given simulation was compared to that of all other runs for the same loop and a new clustering is performed on this dataset. Figure 11.2 presents the average number of clusters with a maximum RMSD of 1.0 Å that are sampled in at least two runs for the three datasets studied here. For 8 a.a and 12 a.a. loops, we observe a linear increase as a function of increasing number of visited conformations, which indicates that, on average, trajectories continue to sample the configurational space without being trapped as simulations progress.

The size of the sampled conformational space can also be estimated by measuring the number of clusters within a fixed minimum RMSD between each other. The evolution of this RMSD rank is presented in Figure 11.3 for the 8 a.a. and 12 a.a. loops. In both cases,

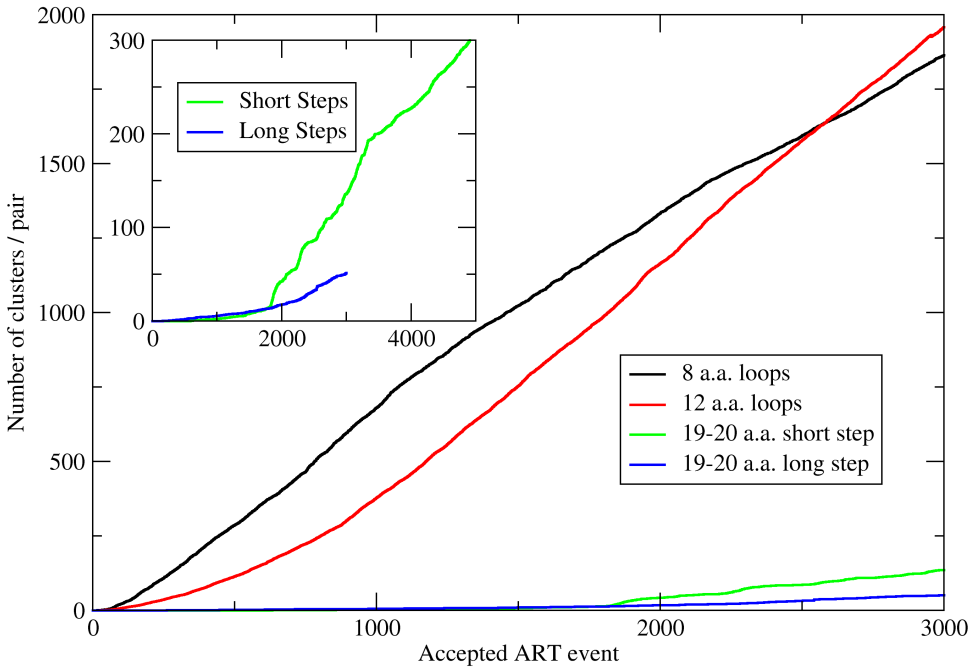

Figure 11.2 – Average number of clusters common between two simulations run for each protein defined by a RMSD distance of less than 1.0 Å between clusters central conformation. Inset is for short steps and long steps parametrization of the 19-20 a.a. loops simulations. Since the probability of two simulations overlapping is proportional to the square of the number of simulations, the plots are normalized by the number of pairs of simulations per protein.

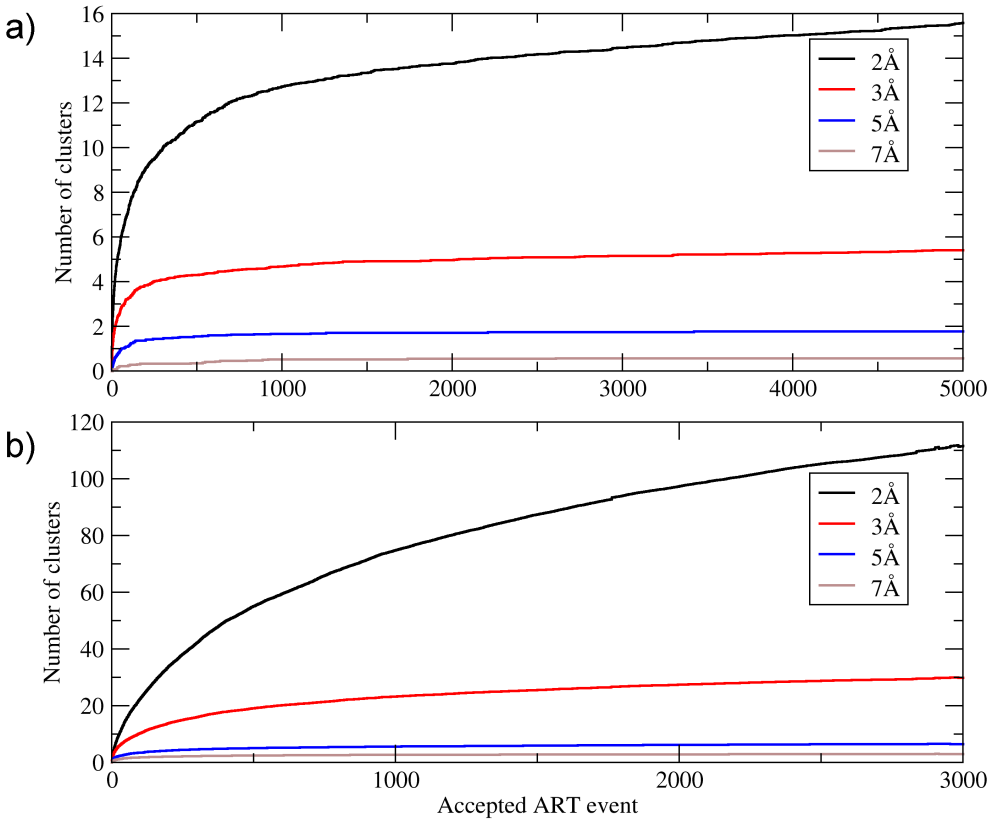

Figure 11.3 – Size of the largest group of clusters per simulation for the (a) 8 a.a. and (b) 12 a.a. loops with minimum RMSD between each member of the group greater then 2 Å (black) to 7 Å (brown).

we see a rapid increase in the number of clusters meeting the minimum RMSD cut-off for the 400 first conformations sampled followed by a slower linear stage to persists until the end of the runs.

Two different behaviors can be identified depending on the size of the RMSD cut-off. With higher minimum RMSD cut-offs, we measure the diameter of the hypervolume accessible to the loops. The rapid convergence of this quantity indicates that the initial configurations are chosen properly as they rapidly bring the various simulations in very different parts of configurational space.

With a minimum RMSD of 2 or 3 Å, the average number of clusters with minimum RMSD between each other gives us a sense of the finer sampling the configurational space. The continuous growth of the curves even after 300 to 5000 steps indicates that the various trajectories are still sampling the conformational space at a finer level.

11.4.3 Exploration of the conformation space for novel 19-20 a.a. loop dataset

The 19-20 a.a. loop dataset was constructed to test the efficiency and scaling of our method on larger model loops. As described below, because of the increase in configurational space, the parameters used in ART nouveau for sampling smaller loops is not optimal for the this dataset. Therefore, two different parameterizations are used on the 19-20 a.a. loops. The first one, dubbed "short step", produces conformations with an average inter-minimum RMSD of about 0.5 Å. It is the set used in our study of both 8 a.a. and 12 a.a. loop datasets. The second one, which we call "long step", generates conformations with an average of 1.1 Å RMSD displacement between adjacent minima. To obtain this increased travel distance between minima, we modified two parameters in the ART method. The first one is the number of iteration of the Lanczos routine used to find the eigenvector of lowest eigenvalue of the Hessian matrix [168]. By reducing from 12 to 4 iterations, the weight of the previous eigenvector, which is used as the seed direction in Lanczos, is more important, stabilizing the trajectory and reducing the impact of local fluctuations in Hessian curvature. The second modified parameter is the force threshold used in relaxing the forces in the perpendicular hyperplane to the activation relaxation. By increasing this threshold from 1.9 kcal/(mol $\cdot$ Å) to 2.5 kcal/(mol $\cdot$Å), the

probability of loosing a negative eigenvalue is further reduced. These modifications decrease the reliability of the saddle point and the physical basis for the initial minimum – saddle – final minimum pathway. However, here we are interested in moving through the landscape and mostly making sure that the visited minima are acceptable thermodynamically. Although fairly aggressive, this set of parameters allows us to keep a reasonable acceptance rate.

The details of both simulation sets are presented in Table 11.III. Comparing the average potential energy of the sampled minima for the two different parametrization sets, we see that the use of short steps leads to an average potential energy $\sim$7 kcal/mol lower and a continued convergence toward low energy structures compared to the long steps (data not shown). Also, 9 out of the 10 sequence-dependent global energy minima were found in the short step simulations.

We first characterize the sampling of the configurational space for the 19-20 a.a. loops by following the evolution of the conformation RMSD as a function of the number of generated conformations. As with the loops of 8 and 12 a.a., we observe that the average RMSD, measured from the native state, remains very high at 6.3 Å , close to the value of the initial conformation, 8 Å (Figure 11.4 (a) black).

However, the reduction of the average RMSD of the lowest energy conformation found so far in each run to the global minimum is much slower than with the smaller loops, reaching 5.0 Å in the first 5000 steps. This can be explained by the fact that out of 10 loop models, only the trajectories of protein 2i9i sampled the global minimum more than once (see Table 11.III). The larger conformation space of the longer proteins means that there is a smaller probability that two independent trajectories overlap, finding the same folding energy funnel to the global minima. Indeed, the simulation runs that do find the global energy minima evolve at at rate roughly two times slower than the 12 a.a. loops . Not surprisingly, the number of clusters shared between runs of the same 19-20 a.a. loop also grows at a slower rate that for shorter sequences. (Figure 11.2).

The evolution of the cluster rank metric for the short step simulations presented in Figure 11.5(a) also points to difficulties in sampling the conformation space in each simulation run using the short step parametrization. With a larger conformation space and

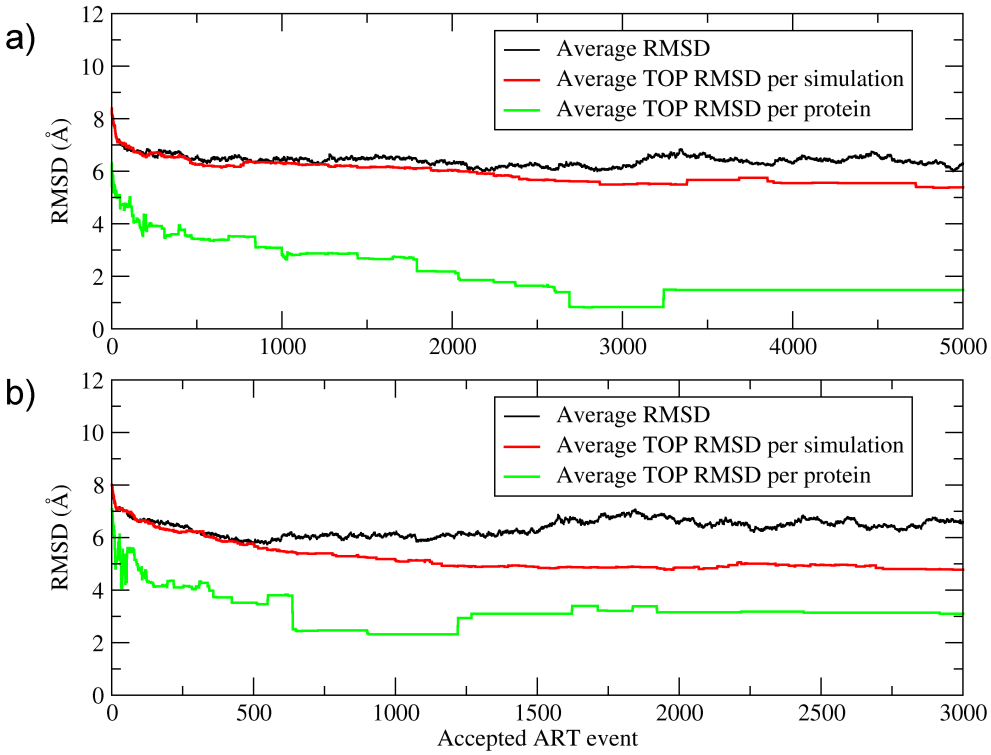

Figure 11.4 – RMSD evolution for the 19-20 a.a. loops using (a) short steps and (b) long steps parametrization. In black, the current conformation's RMSD of each simulation is calculated to the global energy minimum conformation of the sampled protein. The average TOP RMSD is calculated between the global energy minimum of a system and the lowest energy conformation found so far per simulation (red) or per protein (green).

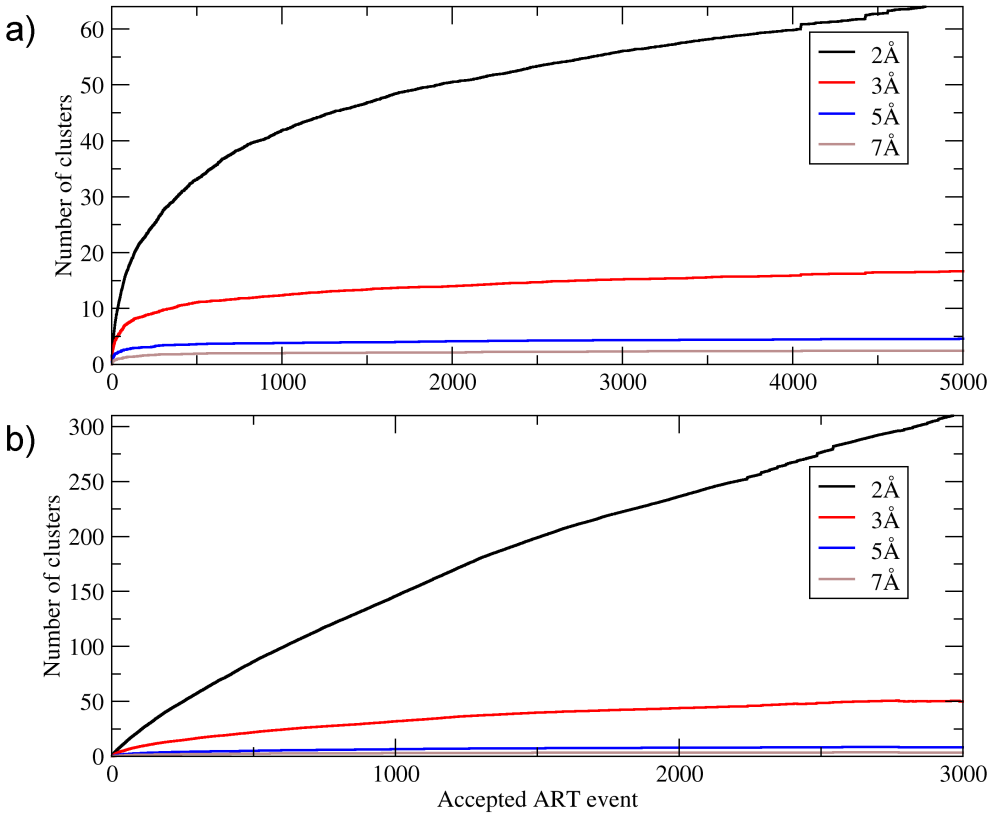

Figure 11.5 – Size of the largest group of clusters per simulation with minimum RMSD between each member of the group greater then 2 Å (black) to 7 Å (brown).

the possibility of larger RMSD between two loop conformations, the cluster rank metric or the 19-20 a.a. loops should increase at a higher rate than the 12 a.a. loops. However, the shorter average inter-minimum RMSD for the short step simulations (0.52 Å) compared to the 12 a.a. loops simulations (0.72 Å) may explain the lower cluster rank observed.

As mentioned above, to correct for the limited sampling of the short-step ART nouveau 19-20 a.a simulations, we also launched a number of runs with longer ART nouveau step. These newly generated trajectories sample saddle points of higher energy, and yet, sampling speed is greatly enhanced. Indeed, with the new parameters, we see that the number of clusters with minimum RMSD of 2 Å increases four times more rapidly in the first 1000 events (black in Figure 11.5 (b)) as compared to the short steps simulations (Figure 11.5 (a)). Moreover, the rate of increase does not slow down after the first 1000 events. On the other side, the median lowest RMSD structure to the global minimum is found at a distance of 1.92 Å (average 2.44 Å) compared to 3.71 Å (average 4.14 Å) for the short steps simulations even though the long step simulations are much less likely to find the global energy minimun of the various loops. This means that the long steps parametrization is better suited for wide sampling of the energy surface while the small steps parametrization are better for in-depth sampling and structure refinement. This is also demonstrated by the speed at which the long step simulations will find a low energy minimum of low RMSD (green line in Figure 11.4 (b)) compared to previous parametrization (Figure 11.4 (a)).

11.4.4 Scaling

ART manages to avoid the exponential increase in complexity of the conformational space as a function of the number of amino acides by not attempting to sample the whole configurational space, but rather sampling low-energy structures only through the generation of connected physical trajectories (at least, when using small steps). The time needed for the ART method to pass from one local minimum to neighboring minimum is proportional to the number of integration steps required to generate a new conformation, i.e. activate to a nearby saddle point and relax into a new minimum. The cost of each of

these step is, of course, dependant on computational efforts required by the force field. The modifications to the OPEP potential treating the protein's body as a background potential lead to a theoretical scaling of the force field computation time that is linear with the size of the loop (n) and the size of the protein (N), leading to an order of $O(n \times N)$. Experimental scaling results are presented in Table 11.IV where we see that, as expected, force field evalution times scale linearly with the protein's size with an average correlation coefficient of 0.98 ± 0.01 and scales linearly between loop size 8 a.a. and 12 a.a and sub-linearly between loops of 12 a.a. and 20 a.a (which can be explained by the presence of cut-offs for some parts of the potential). The average number of force field evaluation per even is not influenced greatly by the size of the loop with an average of 30000 ± 2000 evaluations (see Table 11.IV and Ref. [162]), and the total empirical scaling factor for the sampling of a new conformation is linear with the loop size.

Scaling is also measured by the number of sampled conformations needed to reach a given conformation of interest as a function of loop length. In both cases presented in Figure 11.1 (a) and (b), the RMSD measured with respected to the global energy minimum shows a fast collapse within the first 1500 sampled events, followed by a slow optimization. For the 19-20 a.a. loops, this collapse is evident in the 3000 first conformations sampled. What differs is the minimum RMSD to which the sampling converges after the initial fast collapse. As shown in Figure 11.6, after 500 sampled conformation, 29 of the 79 8 a.a. loop simulations have visited the energy global minimum for their respective sequence (36.7%) with 27 of the 108 12 a.a. loop simulations doing the same (25%). After 1500 conformations have been sampled, these ratio are 54.4% and 40.7% respectively. Therefore, to maintain the same probability, i.e. requiring that at least one simulation per protein finds its global minimum conformation in 1500 sampled conformations, we need 34% to 47% more generated conformations per 12 a.a. loop then per 8 a.a. loop, which is in line with the 50% increase in loop length. Combining this linear increase in the number of conformations needed to the previous linear increase in simulation time required to generate a conformation, we estimate the total computational efforts are quadratic with loop-length.

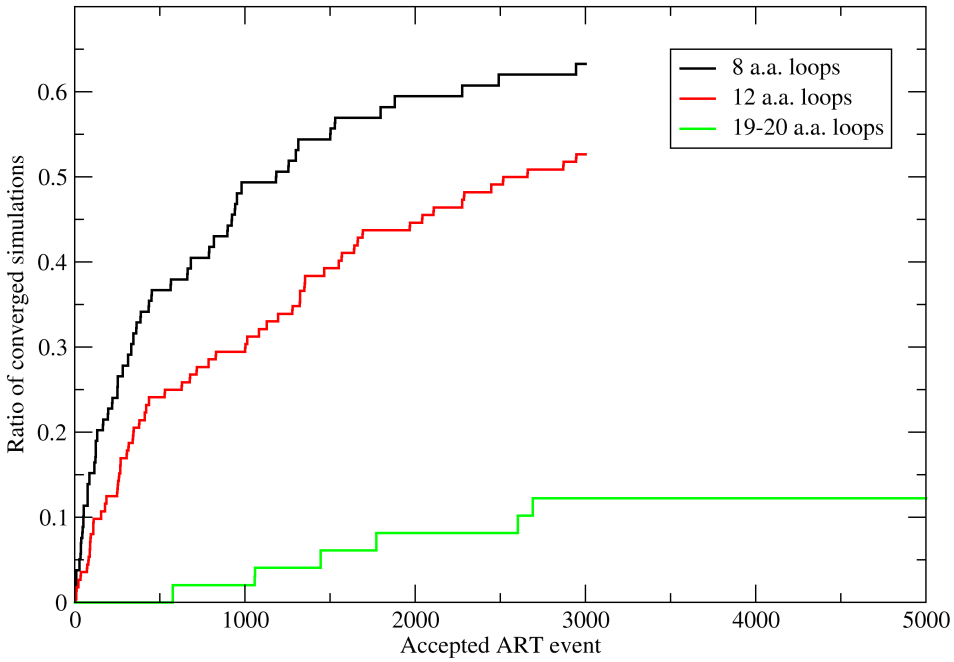

Figure 11.6 – Proportion of the number of simulation that have found their protein's global minimum loop structured as a function of the number of accepted conformations based on a 0.1 Å RMSD cut-off to the global minimum.

11.5 Discussion and conclusion

Small-loop structure prediction methods have seen significant improvements in terms of required efforts and achieved precision in the last decade. Loop predictions at the level of 1.25 Å RMSD are now available for 12 à. loops using methods that scale exponentially with system size [285]. Recent advances even boast lower than 1 Å RMSD precision on loops of up to 12 a.a. [164] or as low as 2 Å for loops of up to 20 a.a. that are identifiable by sequence homology and other similarity criterions [33].

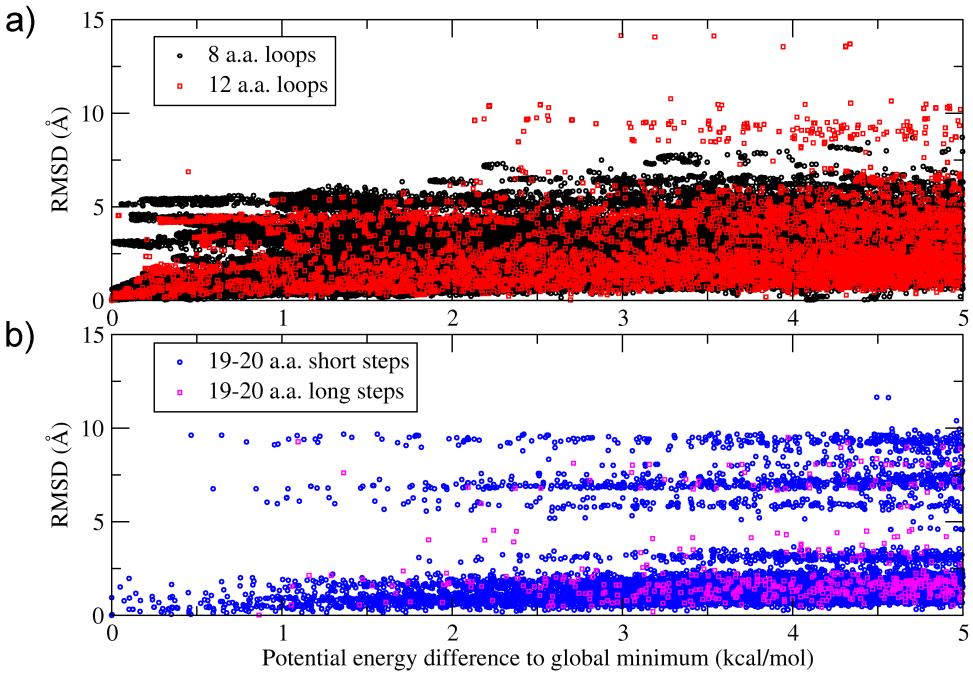

Figure 11.7 – Distribution of the RMSD to the global energy minimum structures for (a) small and (b) large loops for structures of potential energy 5 kcal/mol or less over the global minimum.

In this paper, we have shown that the ART nouveau method can be used to sample efficiently the conformation space of loops of 20 a.a. or more. In particular, ART nou-

veau is very competitive as compared with previously published methods on these large loops [33, 106], demonstrating an efficient sampling of a wide range of conformations and is also able to sample conformations of lower RMSD to the native structure. This advantage is likely due to the fact that events represent a physical trajectory with local minima connected through a common saddle point. Given that the conformation space increases exponentially with the loop length, large random moves are very likely to end up in unphysical parts of this space, something that is avoided with ART nouveau even with the relatively long steps used on the long loops. The trajectory we generate during the event, which attemps to follow a direction of negative curvature with all other 3N-1 directions near their minimum, ensures that.

By extensively sampling low-energy structures, ART nouveau can also provide useful information beyond the best score. While the proteins that were chosen for this study have well defined structures, it is interesting to note that our simulations sampled conformations of low energy and high RMSD to the global energy minimum for both the small and large loops as displayed in Figure 11.7 (a) and (b) respectively. For the 8 a.a. loops, multiple conformations with RMSD up to 6 Å to the global energy minimum structures were found within less then 1 kcal/mol above the global energy minimum. For the 12 a.a. and 19-20 a.a. loops, this value reaches 7 Å and 9 Å respectively on a few occasions. On more flexible loop targets, these distant conformations may well be populated at the equilibrium, playing biological role for these structures.

Since ART nouveau is a method that can be used with any underlying energy potential, its ability to find global energy minimum would not be altered by using a more faithful protein representation in which the global minima corresponds to the native structures. From the collected data on loops of size 8 a.a. to 20 a.a., we estimate the computational time requirements to scale roughly quadratically with the sequence length of the simulated loop.

The current adaptation of the ART-nouveau method is a promising tool to tackle the problem of long loop sampling. All the metrics presented show that the first 1500 sampled conformations are the most rewarding in there ability to minimize the RMSD to the global minimum and that it is preferable to launch multiple short simulation runs than

a few long ones. Results from a few test cases with the 19-20 a.a. loops demonstrate that, by alternating large and smaller moves, ART nouveau can avoid being trapped into the numerous basins associated with the longer-loops complex energy landscape, sampling the configurational space efficiently at rough and fine levels, leading to the identification of a number of competing states, slightly above the minimum-energy conformation, that could plan an important biological role.

11.6 Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada and Canada Research Chair Foundation. We are grateful to Calcul Québec for access to its computer infrastructure.

Tableau 11.I – Simulation details for the 8 a.a. loops of the Olson *et al.* dataset[189]. All RMSD are calculated with respect to the native loop structure and are presented in Å. RMSD initial is the distance between the initial stretched structures and the native conformation. Best RMSD corresponds to the structure of lowest RMSD and the energy rank is the percentage of conformation that have lower energy than this structure. "TOP RMSD OPEP" is the RMSD of the structure of lowest energy with the OPEP potential. The acceptance rate of a new conformation is averaged over all runs.

Protein	Nb runs	RMSD initial	Best RMSD	Energy rank (%)	TOP RMSD OPEP	Nb runs finding energy min.	Average accept. rate
1a62	3	3.21-5.59	0.72	96.6	2.73	3	0.40
1a62 2	4	4.31-6.76	0.15	55.0	3.18	2	0.31
1aac	3	3.42-5.06	2.01	71.8	2.89	2	0.35
1aba	4	3.75-6.67	0.36	85.1	3.05	1	0.37
1awd	2	4.02-6.06	0.11	32.0	3.94	2	0.34
1c52	3	2.68-4.44	2.65	92.1	3.75	3	0.21
1cbn	5	3.82-5.77	0.33	58.4	2.44	5	0.43
1hfc	3	2.84-6.18	0.43	97.4	2.89	3	0.33
1ig5	5	5.00-7.33	0.00	0.1	3.68	2	0.37
1lit	4	4.13-5.54	0.03	87.3	3.42	4	0.49
1msi	2	5.23-7.22	0.01	91.8	3.59	2	0.32
1nls	1	4.92	4.87	99.5	6.24	1	0.54
1nox	4	3.31-6.95	2.27	96.8	2.96	4	0.36
1opd	3	3.16-5.25	1.17	89.8	3.57	3	0.38
1plc	3	4.70-7.22	1.20	28.8	3.29	3	0.38
1plc 2	1	3.84	1.98	33.8	2.31	1	0.41
1ppn	3	2.15-5.79	0.52	32.8	1.20	3	0.29
1ppn 2	3	4.43-6.58	2.01	99.9	4.43	3	0.42
1ra9	4	3.86-5.42	2.29	70.9	4.36	1	0.32
1rat	3	3.71-5.46	2.99	89.9	4.02	1	0.26
1rro	3	3.18-7.64	0.01	57.1	4.87	3	0.23
1vwj	4	2.75-5.92	1.67	99.6	6.44	1	0.32
3nul	4	3.22-5.86	0.06	88.9	2.12	4	0.23
3seb	2	3.42-4.40	2.14	72.1	4.00	2	0.28
5pal	3	3.73-5.66	0.85	46.4	2.12	1	0.44
Average	3.2	4.71	1.23	71.0	3.50	2.4	0.35
Median			1.01	85.1	3.46		
St. Dev.			1.20	28.3	1.17		

Tableau 11.II – Simulation details for the 12 a.a. loops of the Fiser *et al.* dataset[67]. All RMSD are calculated with respect to the native loop structure and are presented in Å. RMSD initial is the distance between the initial stretched structures and the native conformation. Best RMSD corresponds to the structure of lowest RMSD and the energy rank is the percentage of conformation that have lower energy than this structure. Two scoring methods were compared to RMSD of the minimum energy conformation, first the OPEP simulation potential (TOP RMSD OPEP), then the dFIRE scoring method[284] (TOP RMSD dFIRE) after conversion of the coarse grained model to an all-atom representation using SCWRL4[133]. The acceptance rate of a new conformation is averaged over all runs.

Protein	Nb runs	RMSD initial	Best RMSD	Energy rank (%)	TOP RMSD OPEP	Nb runs finding energy minimum	TOP RMSD dFIRE	Average acceptance rate
154L	3	8.61-11.00	2.00	83.0	14.59	3	3.87	0.49
1ARP	3	4.93-7.71	2.40	99.0	5.77	1	5.61	0.49
1CTM	4	6.36-8.41	2.41	65.1	7.13	1	4.59	0.48
1DTS	3	5.11-7.33	2.70	85.9	5.44	1	3.68	0.50
1ECO	3	6.61-8.38	1.15	54.9	4.38	3	3.45	0.50
1EDE	3	5.73-6.65	2.35	76.3	6.32	0	3.20	0.49
1EZM	3	3.91-5.22	0.36	73.0	5.31	3	1.74	0.49
1HFC	3	8.11-9.09	3.01	66.2	11.31	3	7.90	0.50
1MSC	3	6.97-9.27	2.06	95.7	7.82	3	8.17	0.49
1ONC	4	7.20-8.43	1.51	77.1	4.80	1	3.54	0.49
1PBE	3	5.97-7.02	0.94	68.5	4.09	3	2.09	0.48
1PMY	3	4.74-5.85	2.21	71.5	4.78	0	3.43	0.48
1PRN	3	5.24-7.34	1.62	83.7	6.38	2	7.41	0.48
1RCF	3	6.24-9.59	2.22	83.0	4.09	3	4.06	0.48
1RRO	3	3.73-4.73	1.29	89.9	4.42	3	3.85	0.50
1SCS	2	5.80-10.09	0.34	49.5	3.32	2	2.9	0.49
1SRP	3	3.21-5.98	1.14	97.8	3.05	3	2.16	0.50
1TCA	2	6.20-8.42	3.04	6.9	5.11	0	5.21	0.48
1THG	2	5.70-6.41	1.73	24.4	2.58	1	2.92	0.49
1THW	2	5.76-8.14	3.63	99.9	9.61	0	9.45	0.49
1TML	3	7.98-8.70	1.18	17.2	3.85	3	2.93	0.49
1XIF	3	5.72-6.22	0.14	13.2	1.62	1	1.55	0.49
2CPL	4	7.16-9.14	2.84	72.4	6.59	1	5.34	0.49
2CYP	3	4.63-9.03	2.61	86.7	4.20	1	3.84	0.49
2EBN	3	5.68-9.97	2.52	88.1	7.98	1	4.70	0.50
2EXO	3	4.18-7.80	3.39	27.3	5.89	0	3.07	0.48
2PGD	3	5.74-7.88	1.39	95.4	7.36	1	3.09	0.48
2RN2	3	5.22-6.08	1.73	40.3	3.59	0	6.29	0.48
2SIL	3	7.59-9.10	0.00	30.9	3.61	2	1.87	0.49
2SNS	3	6.96-11.74	0.37	55.8	3.93	1	3.91	0.48
2TGI	3	6.75-7.32	1.70	76.1	3.23	3	3.17	0.50
3B5C	3	4.05-6.09	0.30	41.7	2.77	3	2.97	0.49
3CLA	3	4.53-8.99	2.84	30.2	5.46	1	5.80	0.48
3COX	3	5.04-5.67	2.02	89.8	5.61	0	4.84	0.48
3HSC	3	8.68-10.28	1.81	82.4	4.96	3	5.40	0.48
451C	2	7.41-7.65	2.92	80.2	5.93	2	6.11	0.48
4ENL	3	4.53-4.89	0.90	83.8	5.95	2	1.96	0.48
4I1B	3	7.42-8.23	0.01	75.3	10.2	2	6.25	0.49
Average	2.8	6.93	1.75	66.8	5.60	1.7	4.27	0.49
Median			1.77	75.7	5.21		3.85	
St. Dev.			0.98	26.4	2.53		1.87	

Tableau 11.III – Simulation details for the 19 to 20 a.a. loops dataset. Secondary structure a.a. is the number of a.a. in turn and bend conformation and, in the case of 1ofl, in α-helical conformation as annotated by DSSP [119]. RMSD initial is the distance between the initial stretched structures and the native conformation. SS and LS refer to the short step and long step parametrization respectively. Best RMSD corresponds to the structure of lowest RMSD while "TOP RMSD OPEP" is the RMSD of the structure of lowest energy with the OPEP potential. The acceptance rate of a new conformation is averaged over the number of runs.

Protein	Loop length	Loop	Secondary structure a.a.	RMSD init.	Nb runs SS	Nb runs LS	Nb runs finding energy minimum SS	Nb runs finding energy minimum LL	Best RMSD	TOP RMSD OPEP	Avg % acc. conf. SS	Avg % acc. conf. LL
1gwe	20	G406-D425	2	6.6-13.6	5	10	1	0	4.11	8.39	0.52	0.38
1ofl	20	Y434-N453	4	11.4-18.9	5	10	0	1	9.45	12.53	0.53	0.39
1q6z	20	V329-Q348	2	2.8-10.3	5	10	1	0	1.69	9.30	0.58	0.38
2ess	20	C139-P158	2	5.5-9.9	5	10	1	0	3.80	7.13	0.54	0.39
2gag	20	F445-P464	0	5.2-21.9	5	10	1	0	2.17	10.32	0.60	0.37
2i9i	20	H59-H78	0	2.2-12.7	5	10	2	0	0.83	1.10	0.56	0.41
2vk8	20	G343-S362	3	3.4-8.3	5	10	1	0	1.85	5.40	0.59	0.39
3cx5	20	N227-G246	3	3.9-8.9	5	10	1	0	1.80	6.98	0.57	0.39
3d3y	19	L218-I236	2	3.2-7.5	4	10	1	0	1.62	2.89	0.52	0.39
3igx	20	E261-I280	2	3.3-12.2	5	10	1	0	1.93	7.73	0.58	0.40
Average				8.26					2.98	7.18	0.56	0.39
Median									1.89	7.43		
St. Dev.									2.37	3.21		

Tableau 11.IV – Scaling parameters of the sampling of one new conformation through ART nouveau method. The protein size scaling factor represents the slope of the time needed for one force field evaluation in relation to the protein's size for three loop size obtained through linear regression. Also presented is the scaling factor correlation coefficient and the average total number of force field evaluations needed to sample one new local minimum. Abreviations "ss" and "ls" refer to the short step and long step parametrization of the 19-20 a.a. loop simulations.

Loop size	Protein size scaling factor (μs)	Protein size scaling correl.	Nb. force . per new conformation
8 a.a.	4.76	0.97	27123
12 a.a.	7.38	0.98	31572
20 a.a.	8.59	0.98	31596 (ss) 28030 (ls)

CHAPITRE 12

APPROFONDISSENT DE L'ARTICLE SUR LA PRÉDICTION DES STRUCTURES DE BOUCLES

Les travaux d'échantillonnage de conformations et de prédictions de structures de boucles à l'aide de notre méthode ART-nouveau modifiée ont démontré que la méthode est bien adaptée à l'échantillonnage de longues boucles de 19 à 20 a.a. tout en laissant présager une augmentation du temps de calcul d'ordre $O(N^2 \times M)$ où N la taille de la boucle et M est la taille de la protéine. Suivant cette performance, la méthode semble être une bonne candidate pour l'échantillonnage de boucles encore plus longues tel que la boucle à 60 a.a. liant les deux moitiés de P-glycoprotéine (Pgp). Avec la publication imminente d'une structure de Pgp humaine (PDB : 2YL4) et en comptant les structures ouvertes de Pgp de *Mus musculus* (PDB : 3G5U), la table est mise pour des projets de recherche visant l'étude du changement de conformation de la boucle manquante de Pgp.

Bien que notre méthode couplée au potentiel OPEP modifié affiche des prédictions de structures aussi précises que n'importe quelle autre méthode ayant été appliquée aux boucles de 19-20 a.a., les résultats de prédictions sur de plus petites boucles démontrent que notre modification du potentiel OPEP est sous-optimale. Des modifications supplémentaires telles que l'utilisation de chaînes latérales flexibles dans le corps de la protéine pourrait être envisagées au coût de perte de performances lors des calculs d'évaluation de l'énergie. Alternativement, le volume de ces chaînes latérales pourrait être diminué pour facilité l'empaquetage des boucles sur le corps des protéines, mais au coût d'une perte de précision sur le calcul de l'énergie. Une autre alternative intéressante serait de conserver l'usage du potentiel OPEP pour échantillonner les points de selles de la surface énergétique des boucles, puis de passer à une représentation tout-atome lors de la minimisation de l'énergie dans le but d'augmenter la précision lors de l'évaluation de l'énergie des minima énergétiques.

Finalement, nous avons démontré qu'en modifiant les paramètres d'ART, il est possible de passer d'un mode d'échantillonnage large de l'espace des configurations à un

échantillonnage précis d'un bassin de configurations. Il serait intéressant d'examiner si la combinaison de ces deux paramétrisations dans une méthode de recuit simulé augmenterait les performances de prédiction en conservant une échantillonnage suffisamment large pour éviter de rester pris dans des bassins de configurations locaux.

CONCLUSION

N.B. : La conclusion a entièrement été réécrite

Au cours de cette thèse sur la dynamique et la flexibilité des protéines, nous avons mis à l'épreuve plusieurs méthodes numériques de simulation basées sur la dynamique moléculaire (DM) dans le but d'échantillonner des trajectoires de liaison du ligand Z-pro-prolinal (ZPP) à la protéine Prolyl oligopeptidase (POP) (chapitre 7) ainsi qu'afin d'obtenir des trajectoires d'ouverture des domaines liant les nucléotides (DLN) de la protéine SAV1866. Nous avons aussi contribué à la résolution du problème de la prédiction des boucles en adaptant la méthode d'échantillonnage ART-nouveau à ce problème. De façon générale, nous avons vu que la qualité des résultats obtenus à l'aide des méthodes de DM, de dynamique moléculaire dirigée (DMD), d'échantillonnage parapluie (EP), de Monte Carlo (MC) employées dans les différents volets de cette thèse dépend grandement de la capacité d'échantillonner un espace de conformation d'intérêt. Nous avons évalué les avantages, mais aussi les limitations de ces méthodes lorsqu'elle sont appliquées à des systèmes larges et complexes.

Dû à son aspect dynamique, la DMD génère des trajectoires en perpétuelle convergence vers un état d'équilibre qui ne peut être atteint que lorsque cette convergence se produit plus rapidement que les déformations dues à la vitesse de tir. Nous avons observé dans le cas de POP et de son ligand que les résultats des simulations de tir générés par DMD ne peuvent être utilisés pour calculer des différences d'énergie libre de liaisons à l'aide de l'équation de Jarzynski pour un système de cette complexité avec les paramètres sélectionnés, même en utilisant une vitesse de tir lente de 0.1 nm/ns. L'utilisation de vitesses de tir plus lentes et la nécessité de répéter de nombreuses fois les simulations de tir rendent la méthode de DMD trop coûteuse en temps de calcul pour être utilisée dans le but d'obtenir des profils d'énergie libre de trajectoires. Toutefois, nous avons démontré que la méthode trouve son utilité lors de la génération de trajectoires initiales pour d'autres méthodes de calcul de profil d'énergie libre tel que l'échantillonnage parapluie (EP). En effet, certaines des structures de la trajectoire de DMD de sortie du ZPP par la boucle flexible T190-N208 dans lesquelles on voit une interaction

entre le ZPP et la boucle flexible étirée en solution et qui ont été utilisées pour ensemencer des fenêtres d'EP ont généré des simulations d'EP démontrant la stabilité de cette interaction. Dans les simulations avec la protéine SAV1866, la DMD n'a toutefois pas permis d'échantillonner des événements de séparation des domaines transmembranaires (DTM). Le vecteur de tir ayant été appliqué aux centres de masse des deux domaines liant le nucléotide (DLN), seuls ces domaines se sont séparés sans entraîner la séparation des DTM selon notre hypothèse de départ. Les résultats obtenus à l'aide de ces simulations de DTM sur SAV1866 permettent toutefois d'observer la stabilité de l'interaction entre l'adénosine diphosphate et le motif Walker A de chaque DLN ainsi que la fréquence des mode de séparation du DLN.

Dans le cas de la méthode d'échantillonnage parapluie (EP), nous avons démontré que le choix de la trajectoire échantillonnée pouvait avoir un grand impact sur la précision du profil d'énergie libre obtenue. Alors que l'EP permet d'obtenir une différence d'énergie libre de liaison comparable aux valeurs expérimentales dans le cas de la trajectoire passant par l'interface inter-domaines et la boucle T190-N208, la trajectoire passant par le tunnel du domaine en propulseur-β affiche une différence d'énergie libre plus de 4 fois plus large et une constante d'équilibre théorique irréaliste. Dans ce dernier cas, l'analyse de l'erreur par la méthode de Boostrap ne présente pas d'erreur évidente et l'espace de la coordonnée de réaction est échantillonné sans trou, ce qui semble indiquer que la trajectoire est contrainte et que les fenêtres n'échantillonnent pas les changements de conformation nécessaires à l'obtention d'états à l'équilibre de plus basse énergie libre.

Notre adaptation de la méthode ART-nouveau au problème de la prédiction des structures de boucles est une contribution importante à ce domaine. Dans le cas de la prédiction de longues boucles de 19 à 20 a.a., notre implémentation de la méthode avec le potentiel gros-grain OPEP modifié offre une capacité de prédiction équivalente aux autres méthodes tout en échantillonnant des conformations de boucles plus près de la structure native. De plus, la méthode affiche un temps d'exécution qui croît en $O(N^2 \times M)$ (où N et M sont le nombre d'acide aminés est la boucle et de la protéine respectivement) déterminé de façon empirique entre les jeux de boucles de 8 a.a. et de 12 a.a. alors que les autres méthodes d'échantillonnage dont les temps d'exécutions ont été publiés af-

fichent des taux de croissance du temps de calcul en $O(e^N)$ sur des séquences de tailles similaires [7, 286]. Nous croyons toutefois que les capacités de prédiction de la méthode doivent être améliorée en optimisant le potentiel énergétique ou en utilisant un potentiel plus précis afin d'augmenter l'utilité de la méthode.

En conclusion, les méthodes dirigées à une seule contrainte comme la DMD et l'EP ont leur utilité dans l'étude de systèmes complexes, mais aussi leur limites. L'utilisation d'une seule coordonnée de réaction, ou une seule contrainte, a pour effet de diminuer le biais imposé au système ainsi que le temps de calcul nécessaire à l'échantillonnage des coordonnées de réaction. Ces simulations permettent d'observer des événement ponctuels avec un minimum de biais tel que la formation ou le bris d'interactions entre des composantes du système tel que des ligands. Toutefois, pour échantillonner des événements de grande amplitude tel que la séparation complète des domaines de SAV1866, ou encore l'ouverture de l'interface entre le domaine en propulseur-β et le domaine catalytique de POP, il est fort probable que l'ajout de contraintes soit nécessaire.

BIBLIOGRAPHIE

[1] J. Aittoniemi, H. de Wet, F. M. Ashcroft, and M. S. P. Sansom. Asymmetric switching in a homodimeric abc transporter : A simulation study. *PLOS Computational Biology*, 6(4) :e1000762, 2010.

[2] M. H. Akabas, D. A. Stauffer, M. Xu, and A. Karlin. Acetylcholine-receptor channel structure probed in cysteine-substitution mutants. *Science*, 258(5080) : 307–310, 1992.

[3] S. G. Aller, J. Yu, A. Ward, Y. Weng, S. Chittaboina, R. Zhuo, P. M. Harrell, Y. T. Trinh, Q. Zhang, I. L. Urbatsch, and G. Chang. Structure of p-glycoprotein reveals a molecular basis for poly-specific drug binding. *Science*, 323(5922) :1718–22, 2009.

[4] S. G. Aller, J. Yu, A. Ward, Y. Weng, S. Chittaboina, R. Zhuo, P. M. Harrell, Y. T. Trinh, Q. Zhang, I. L. Urbatsch, and G. Chang. Structure of p-glycoprotein reveals a molecular basis for poly-specific drug binding. *Science*, 323(5922) :1718–1722, 2009. doi : 10.1126/science.1168750.

[5] S. G. Aller, J. Yu, A. Ward, Y. Weng, S. Chittaboina, R. Zhuo, P. M. Harrell, Y. T. Trinh, Q. Zheng, I. L. Urbatsch, and G. Chang. Structure of p-glycoprotein reveals a molecular basis for poly-specific drug binding. *Science*, 323 :1718–1722, 2009.

[6] C. Anezo, A. H. de Vries, H. D. Holtje, D. P. Tieleman, and S. J. Marrink. Methodological issues in lipid bilayer simulations. *Journal of Physical Chemistry B*, 107(35) :9424–9433, 2003.

[7] Y. A. Arnautova, R. A. Abagyan, and M. Totrov. Development of a new physics-based internal coordinate mechanics force field and its application to protein loop modeling. *Proteins : Structure, Function and Genetics*, 79(2) :477–498, 2011.

[8] A. D. Attie. Abca1 : at the nexus of cholesterol, hdl and atherosclerosis. *Trends in Biochemical Sciences*, 32(4) :172–179, 2007.

[9] T. Baştŭg and S. Kuyucak. Application of jarzynski's equality in simple versus complex systems. *Chemical Physics Letters*, 436 :383–387, 2007.

[10] T. Baştuğ, P-C. Chen, S. M. Patra, and S. Kuyucak. Potential of mean force calculations of ligand binding to ion channels from jarzynski's equality and umbrella sampling. *Journal of Chemical Physics*, 128 :155104, 2008.

[11] A. V. Bakker, S. Jung, R. W. Spencer, F. J. Vinick, and W. S. Faraci. Slow tight-binding inhibition of prolyl endopeptidase by benzyloxycarbonyl-prolyl-prolinal. *Biochemical Journal*, 271(2) :559–562, 1990.

[12] A. Barducci, M. Bonomi, and P. Derreumaux. Assessing the quality of the opep coarse-grained force field. *Journal of Chemical Theory and Computation*, 7(6) : 1928–1934, 2011.

[13] J.-P. Becker, F. Van Bambeke, P.l M. Tulkens, and M. Prevost. Dynamics and structural changes induced by atp binding in sav1866, a bacterial abc exporter. *Journal of Physical Chemistry B*, 114(48) :15948–15957, 2010.

[14] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics*, 81(8) :3684–3690, 1984.

[15] J. Berger and J. Gartner. X-linked adrenoleukodystrophy : Clinical, biochemical and pathogenetic aspects. *Biochimica Et Biophysica Acta-Molecular Cell Research*, 1763(12) :1721–1732, 2006.

[16] O. Berger, O. Edholm, and F. JÃd'hnig. Molecular dynamics simulations of a fluid bilayer of dipalmitoylphosphatidylcholine at full hydration, constant pressure, and constant temperature. *Biophysical Journal*, 72(5) :2002–2013, 1997.

[17] G. Boucher, N. Mousseau, and P. Derreumaux. Aggregating the amyloid a beta(11-25) peptide into a four-stranded beta-sheet structure. *Proteins : Structure, Function and Genetics*, 65(4) :877–888, 2006.

[18] I. Brandt, S. Scharpé, and A-M. Lampier. Suggested functions for prolyl oligopeptidase : A puzzling paradox. *Clinica Chimica Acta*, 377(1-2) :50 – 61, 2007.

[19] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. Charmm - a program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2) :187–217, 1983.

[20] A. H. Buchaklian and C. S. Klug. Characterization of the lsggq and h motifs from the escherichia coli lipid a transporter msba. *Biochemistry*, 45(41) :12539–46, 2006.

[21] A. Bunker, Pekka T. Männistö, JF St. Pierre, T Róg, P. Pomorski, and M. Karttunen. Molecular dynamics simulations of the enzyme catechol-o-methyltransferase : methodological issues. *SAR & QSAR in Environmental Research*, 19(1-2) :179–189, 2008.

[22] A. A. Canutescu and R. L. Dunbrack. Cyclic coordinate descent : A robotics algorithm for protein loop closure. *Protein Science*, 12(5) :963–972, 2003.

[23] A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Science*, 12(9) :2001–2014, 2003.

[24] L. M. S. Chan, S. Lowes, and B. H. Hirst. The abcs of drug transport in intestine and liver : efflux proteins limiting drug absorption and bioavailability. *European Journal of Pharmaceutical Sciences*, 21(1) :25–51, 2004.

[25] G. Chang. Structure of msba from e. coli : A homolog of the multidrug resistance atp binding cassette (abc) transporters. *Science*, 293 :1793–1800, 2001.

[26] Y. Chebaro, N. Mousseau, and P. Derreumaux. Structures and thermodynamics of alzheimer's amyloid-beta a beta(16-35) monomer and dimer by replica exchange molecular dynamics simulations : Implication for full-length a beta fibrillation. *Journal of Physical Chemistry B*, 113(21) :7668–7675, 2009.

[27] G. Chelvanayagam, G. Roy, and P. Argos. Easy adaptation of protein-structure to sequence. *Protein Engineering*, 7(2) :173–184, 1994.

[28] J. Chen, G. Lu, J. Lin, A. L. Davidson, and F. A. Quiocho. A tweezers-like motion of the atp-binding cassette dimer in an abc transport cycle. *Molecular Cell*, 12(3) : 651–661, 2003.

[29] L. F. Chen, H. Winkler, M. K. Reedy, M. C. Reedy, and K. A. Taylor. Molecular modeling of averaged rigor crossbridges from tomograms of insect flight muscle. *Journal of Structural Biology*, 138(1-2) :92–104, 2002.

[30] P. C. Chen and S. Kuyucak. Mechanism and energetics of charybdotoxin unbinding from a potassium channel from molecular dynamics simulations. *Biophysical Journal*, 96(7) :2577–88, 2009.

[31] W. Chen, N. Mousseau, and P. Derreumaux. The conformations of the amyloid-beta (21-30) fragment can be described by three families in solution. *Journal of Chemical Physics*, 125(8) :084911, 2006.

[32] C. Chipot, X. Rozanska, and S. Dixit. Can free energy calculations be fast and accurate at the same time ? binding of low-affinity, non-peptide inhibitors to the sh2 domain of the src protein. *Journal of Computer-Aided Molecular Design*, 19 (11) :765–770, 2005.

[33] Y. Choi and C. M. Deane. Fread revisited : Accurate loop structure prediction using a database search algorithm. *Proteins : Structure, Function and Genetics*, 78(6) :1431–1440, 2010.

[34] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *EMBO Journal*, 5(4) :823–826, 1986.

[35] S. Y. Chung and S. Subbiah. A structural explanation for the twilight zone of protein sequence homology. *Structure*, 4(10) :1123–1127, 1996.

[36] M. O. Collins, L. Yu, I. Campuzano, S. G. Grant, and J. S. Choudhary. Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder. *Molecular & Cellular Proteomics*, 7(7) :1331–48, 2008.

[37] G. Comellas, Z. Kaczmarska, T. Tarragø, M. Teixidó, and E. Giralt. Exploration of the one-bead one-compound methodology for the design of prolyl oligopeptidase substrates. *PLoS One*, 4 :e6222, 2009.

[38] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995). *Journal of the American Chemical Society*, 118(9) :2309–2309, 1996.

[39] S. Côté, P. Derreumaux, and N. Mousseau. Distinct morphologies for amyloid beta protein monomer : A beta(1-40), a beta(1-42), and a beta(1-40)(d23n). *Journal of Chemical Theory and Computation*, 7(8) :2584–2592, 2011.

[40] E. A. Coutsias, C. Seok, M. P. Jacobson, and K. A. Dill. A kinematic view of loop closure. *Journal of Computational Chemistry*, 25(4) :510–528, 2004.

[41] M. A. Cuendet and O. Michielin. Protein-protein interaction investigated by steered molecular dynamics : The tcr-pmhc complex. *Biophysical Journal*, 95 :3575–3590, 2008.

[42] M. Cui, M. Mezei, and R. Osman. Prediction of protein loop structures using a local move monte carlo approach and a grid-based force field. *Protein Engineering Design & Selection*, 21(12) :729–735, 2008.

[43] T. Darden, D. York, and L. Pedersen. Particle meshewald : An n log (n) method for ewald sums in large systems. *Journal of Chemical Physics*, 98 :10089 – 10092, 1993.

[44] R. J. Dawson and K. P. Locher. Structure of a bacterial multidrug abc transporter. *Nature*, 443(7108) :180–5, 2006.

[45] R. J. Dawson and K. P. Locher. Structure of the multidrug abc transporter sav1866 from staphylococcus aureus in complex with amp-pnp. *FEBS Letters*, 581(5) : 935–8, 2007.

[46] P. I. W. de Bakker, M. A. DePristo, D. F. Burke, and T. L. Blundell. Ab initio construction of polypeptide fragments : Accuracy of loop decoy discrimination by an all-atom statistical potential and the amber force field with the generalized born solvation model. *Proteins : Structure, Function and Genetics*, 51(1) :21–40, 2003.

[47] M. Dean, A. Rzhetsky, and R. Allikmets. The human atp-binding cassette (abc) transporter superfamily. *Genome Research*, 11(7) :1156–1166, 2001.

[48] C. M. Deane and T. L. Blundell. Coda : A combined algorithm for predicting the structurally variable regions of protein models. *Protein Science*, 10(3) :599–612, 2001.

[49] M. K. DeGorter, G. Conseil, R. G. Deeley, R. L. Campbell, and S. P. C. Cole. Molecular modeling of the human multidrug resistance protein 1 (mrp1/abcc1). *Biochemical and Biophysical Research Communications*, 365(1) :29–34, 2008.

[50] Y. Deng and B. Roux. Calculation of standard binding free energies : Aromatic molecules in the t4 lysozyme l99a mutant. *Journal of Chemical Theory and Computation*, 2(5) :1255–1273, 2006.

[51] M. A. DePristo, P. I. W. de Bakker, S. C. Lovell, and T. L. Blundell. Ab initio construction of polypeptide fragments : Efficient generation of accurate, representative ensembles. *Proteins : Structure, Function and Genetics*, 51(1) :41–55, 2003.

[52] P. Derreumaux. From polypeptide sequences to structures using monte carlo simulations and an optimized potential. *Journal of Chemical Physics*, 111(5) : 2301–2310, 1999.

[53] E. Di Daniel, C. P. Glover, E. Grot, M. K. Chan, T. H. Sanderson, J. H. White, C. L. Ellis, K. T. Gallager, J. Uney, J. Thomas, P. R. Maycox, and A. W. Mudge. Prolyl oligopeptidase binds to gap-43 and functions without its peptidase activity. *Molecular and Cellular Neuroscience*, 41 :373 – 382, 2009.

[54] X. Dong, W. Chen, N. Mousseau, and P. Derreumaux. Energy landscapes of the monomer and dimer of the alzheimer's peptide a beta(1-28). *Journal of Chemical Physics*, 128(12), 2008.

[55] R. O. Dror, D. H. Arlow, D. W. Borhani, M. ÃŸ. Jensen, S. Piana, and D. E. Shaw. Identification of two distinct inactive conformations of the Ã§2-adrenergic receptor reconciles structural and biochemical observations. *Proceedings of the National Academy of Sciences*, 106(12) :4689–4694, 2009.

[56] L. Dupuis and N Mousseau. Understanding the ef-hand folding pathway using non-biased interatomic potentials. *Journal of Chemical Physics*, (soumis), 2011.

[57] L. Dupuis and N Mousseau. Holographic multiscale method used with non-biased atomistic forcefields for simulation of large transformations in protein. *Journal of Physics : Conference Series*, (soumis), 2011.

[58] P. D. W. Eckford and F. J. Sharom. Abc efflux pump-based resistance to chemotherapy drugs. *Chemical Reviews*, 109(7) :2989–3011, 2009.

[59] F. El-Mellouhi and N. Mousseau. Ab initio characterization of arsenic vacancy diffusion pathways in gaas with siest-a-rt. *Applied Physics A-Materials Science & Processing*, 86(3) :309–312, 2007.

[60] F. El-Mellouhi, N. Mousseau, and L. J. Lewis. Kinetic activation-relaxation technique : An off-lattice self-learning kinetic monte carlo algorithm. *Physical Review B*, 78(15), 2008.

[61] S. W. Englander and L. Mayne. Protein folding studied using hydrogen-exchange labeling and 2-dimensional nmr. *Annual Review of Biophysics and Biomolecular Structure*, 21 :243–265, 1992.

[62] L. Federici, B. Woebking, S. Velamakanni, R. A. Shilling, B. Luisi, and H. W. van Veen. New structure model for the atp-binding cassette multidrug transporter lmra. *Biochemical Pharmacology*, 74(5) :672–678, 2007.

[63] S. E. Feller, D. Yin, R. W. Pastor, and A. D. MacKerell Jr. Molecular dynamics simulation of unsaturated lipid bilayers at low hydration : parameterization and comparison with diffraction studies. *Biophysical Journal*, 73(5) :2269–2279, 1997.

[64] Scott E. Feller, Klaus Gawrisch, and Alexander D. MacKerell. Polyunsaturated fatty acids in lipid bilayers : ? intrinsic and environmental contributions to their unique physical properties. *Journal of the American Chemical Society*, 124(2) : 318–326, 2001.

[65] A. K. Felts, E. Gallicchio, D. Chekmarev, K. A. Paris, R. A. Friesner, and R. M. Levy. Prediction of protein loop conformations using the agbnp implicit solvent model and torsion angle sampling. *Journal of Chemical Theory and Computation*, 4(5) :855–868, 2008.

[66] N. Fernandez-Fuentes, E. Querol, F. X. Aviles, M. J. E. Sternberg, and B. Oliva. Prediction of the conformation and geometry of loops in globular proteins : Testing archdb, a structural classification of loops. *Proteins : Structure, Function and Genetics*, 60(4) :746–757, 2005.

[67] A. Fiser, R. K. G. Do, and A. Sali. Modeling of loops in protein structures. *Protein Science*, 9(9) :1753–1773, 2000.

[68] E. L. Florin, V. T. Moy, and H. E. Gaub. Adhesion forces between individual ligand-receptor pairs. *Science*, 264(5157) :415–417, 1994.

[69] P. L. Freddolino, C. B. Harrison, Y. Liu, and K. Schulten. Challenges in protein-folding simulations. *Nature Physics*, 6(10) :751–758, 2010.

[70] V. Fülöp and D. T. Jones. β-propellers : structural rigidity and functional diversity. *Current Opinion in Structural Biology*, 9 :715 – 721, 1999.

[71] V. Fülöp, Z. Böcskei, and L. Polgár. Prolyl oligopeptidase an unusual β-propeller domain regulates proteolysis. *Cell*, 94(2) :161 – 170, 1998.

[72] V. Fülöp, Z. Szeltner, and L. Polgár. Catalysis of serine oligopeptidases is controlled by a gating filter mechanism. *EMBO Reports*, 1(3) :277 – 281, 2000.

[73] M. Fuxreiter, C. Magyar, T. Juhász, Z. Szeltner, L. Polgár, and I. Simon. Flexibility of prolyl oligopeptidase : Molecular dynamics and molecular framework analysis of the potential substrate pathways. *Proteins : Structure, Function and Genetics*, 60(3) :504 – 512, 2005.

[74] D. C. Gadsby, P. Vergani, and L. Csanady. The abc protein turned chloride channel whose failure causes cystic fibrosis. *Nature*, 440(7083) :477–483, 2006.

[75] A. Gaggar, P. L. Jackson, B. D. Noerager, P. J. O'Reilly, D. B. McQuaid, S. M. Rowe, J. P. Clancy, and J. E. Blalock. A novel proteolytic cascade generates an extracellular matrix-derived chemoattractant in chronic neutrophilic inflammation. *Journal of Immunology*, 180 :5662 – 5669, 2008.

[76] E. Gallicchio, L. Y. Zhang, and R. M. Levy. The sgb/np hydration free energy model based on the surface generalized born solvent reaction field and novel nonpolar hydration free energy estimators. *Journal of Computational Chemistry*, 23 (5) :517–529, 2002.

[77] J. A. García-Horsman, P. T. Männistö, and J. I. Venäläinen. On the role of prolyl oligopeptidase in health and disease. *Neuropeptides*, 41(1) :1 – 24, 2007.

[78] K. M. Giacomini, S. M. Huang, D. J. Tweedie, L. Z. Benet, K. L. Brouwer, X. Chu, A. Dahlin, R. Evers, V. Fischer, K. M. Hillgren, K. A. Hoffmaster, T. Ishi-

kawa, D. Keppler, R. B. Kim, C. A. Lee, M. Niemi, J. W. Polli, Y. Sugiyama, P. W. Swaan, J. A. Ware, S. H. Wright, S. W. Yee, M. J. Zamek-Gliszczynski, and L. Zhang. Membrane transporters in drug development. *Nature Reviews Drug Discovery*, 9(3) :215–36, 2010.

[79] C. Globisch, I. K. Pajeva, and M. Wiese. Identification of putative binding sites of p-glycoprotein based on its homology model. *ChemMedChem*, 3(2) :280–295, 2008.

[80] Jeff Gore, Felix Ritort, and Carlos Bustamante. Bias and error in estimates of equilibrium free-energy differences from nonequilibrium measurements. *Proceedings of the National Academy of Sciences*, 100(22) :12564–12569, 2003.

[81] M. M. Gottesman and V. Ling. The molecular basis of multidrug resistance in cancer : The early years of p-glycoprotein research. *FEBS Letters*, 580(4) :998–1009, 2006.

[82] Ilan Gronau and Shlomo Moran. Optimal implementations of UPGMA and other common clustering algorithms. *Information Processing Letters*, 104(6) :205–210, DEC 16 2007.

[83] H. Grubmuller, B. Heymann, and P. Tavan. Ligand binding : Molecular mechanics calculation of the streptavidin biotin rupture force. *Science*, 271(5251) :997–999, 1996.

[84] E. Hazai and Z. Bikadi. Homology modeling of breast cancer resistance protein (abcg2). *Journal of Structural Biology*, 162(1) :63–74, 2008.

[85] M. Hennessy and J. P. Spiers. A primer on the mechanics of p-glycoprotein the multidrug transporter. *Pharmacological Research*, 55(1) :1–15, 2007.

[86] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. Lincs : a linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12) :1463–1472, 1997.

[87] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. Gromacs 4 : Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3) :435–447, 2008.

[88] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. Gromacs 4 : Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of Chemical Theory and Computation*, 4(3) :435–447, 2008.

[89] C. F. Higgins and K. J. Linton. The atp switch model for abc transporters. *Nature Structural & Molecular Biology*, 11(10) :918–926, 2004.

[90] P. W. Hildebrand, A. Goede, R. A. Bauer, B. Gruening, J. Ismer, E. Michalsky, and R. Preissner. Superlooper-a prediction server for the modeling of loops in globular and membrane proteins. *Nucleic Acids Research*, 37 :W571–W574, 2009.

[91] A. Hillisch, M. Lorenz, and S. Diekmann. Recent advances in fret : distance determination in proteinâĂŞdna complexes. *Current Opinion in Structural Biology*, 11(2) :201–207, 2001.

[92] I. B. Holland and M. A. Blight. Abc-atpases, adaptable energy generators fuelling transmembrane movement of a variety of molecules organisms from bacteria to humans. *Journal of Molecular Biology*, 293(2) :381–399, 1999.

[93] K. Hollenstein, R. J. P. Dawson, and K. P. Locher. Structure and mechanism of abc transporter proteins. *Current Opinion in Structural Biology*, 17 :412–418, 2007.

[94] K. Hollenstein, D. C. Frei, and K. P. Locher. Structure of an abc transporter in complex with its binding protein. *Nature*, 446(7132) :213–216, 2007.

[95] W. G. Hoover. Canonical dynamics : Equilibrium phase-space distributions. *Physical Review A*, 31(3) :1695–1697, 1985.

[96] J. S. Hub, B. L. de Groot, and D. van der Spoel. g_wham-a free weighted histogram analysis implementation including robust error and autocorrelation estimates. *Journal of Chemical Theory and Computation*, 6(12) :3713–3720, 2010.

[97] J. S. Hub, B. L. de Groot, and D. van der Spoel. g_wham–a free weighted histogram analysis implementation including robust error and autocorrelation estimates. *Journal of Chemical Theory and Computation*, 6 :3713–3720, 2010.

[98] W. Humphrey, A. Dalke, and K. Schulten. Vmd : Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1) :33 – 38, 1996.

[99] P. H. Hünenberger, J. K. Granwehr, J.-N. Aebischer, N. Ghoneim, E. Haselbach, and W. F. van Gunsteren. Experimental and theoretical approach to hydrogen-bonded diastereomeric interactions in a model complex. *Journal of the American Chemical Society*, 119(32) :7533–7544, 1997.

[100] R. N. Hvorup, B. A. Goetz, M. Niederer, K. Hollenstein, E. Perozo, and K. P. Locher. Asymmetry in the structure of the abc transporter-binding protein complex btucd-btuf. *Science*, 317(5843) :1387–90, 2007.

[101] J. Irazusta, G. Larrinaga, J. González-Maeso, J. Gil, J. J. Meana, and L. Casis. Distribution of prolyl endopeptidase activities in rat and human brain. *Neurochemistry International*, 40(4) :337–345, 2002.

[102] A. Ivetac, J. D. Campbell, and M. S. P. Sansom. Dynamics and function in a bacterial abc transporter : Simulation studies of the btucdf system and its components. *Biochemistry*, 46(10) :2767–2778, 2007.

[103] S. Izrailev, S. Stepaniants, M. Balsera, Y. Oono, and K. Schulten. Molecular dynamics study of unbinding of the avidin-biotin complex. *Biophysical Journal*, 72(4) :1568–81, 1997.

[104] M. P. Jacobson, R. A. Friesner, Z. Xiang, and B. Honig. On the role of the crystal environment in determining protein side-chain conformations. *Journal of Molecular Biology*, 320(3) :597 – 608, 2002.

[105] M. P. Jacobson, D. L. Pincus, C. S. Rapp, T. J. F. Day, B. Honig, D. E. Shaw, and R. A. Friesner. A hierarchical approach to all-atom protein loop prediction. *Proteins : Structure, Function and Genetics*, 55(2) :351–367, 2004.

[106] M. Jamroz and A. Kolinski. Modeling of loops in proteins : a multi-method approach. *Bmc Structural Biology*, 10 :9, 2010.

[107] C. Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements : A master-equation approach. *Physical Review E*, 56(5) :5018–5035, 1997.

[108] C. Jarzynski. Hamiltonian derivation of a detailed fluctuation theorem. *Journal of Statistical Physics*, 98(1/2) :77–102, 2000.

[109] W. Curtis Johnson. Protein secondary structure and circular dichroism : A practical guide. *Proteins : Structure, Function and Genetics*, 7(3) :205–214, 1990.

[110] J.-F. Joly, L. Karim-Béland, F. El-Mellouhi, and N. Mousseau. Optimization of the kinetic activation-relaxation technique, an off-lattice and self-learning kinetic monte-carlo method. *Journal of Physics : Conference Series*, (soumis), 2011.

[111] P. M. Jones and A. M. George. Nucleotide-dependent allostery within the abc transporter atp-binding cassette - a computational study of the mj0796 dimer. *Journal of Biological Chemistry*, 282(31) :22793–22803, 2007.

[112] P. M. Jones and A. M. George. Opening of the adp-bound active site in the abc transporter atpase dimer : Evidence for a constant contact, alternating sites model for the catalytic cycle. *Proteins : Structure, Function and Genetics*, 75(2) :387–396, 2009.

[113] P. M. Jones and A. M. George. Molecular-dynamics simulations of the atp/apo state of a multidrug atp-binding cassette transporter provide a structural and mechanistic basis for the asymmetric occluded state. *Biophysical Journal*, 100(12) : 3025–3034, 2011.

[114] P. M. Jones, M. L. O'Mara, and A. M. George. Abc transporters : a riddle wrapped in a mystery inside an enigma. *Trends in Biochemical Sciences*, 34(10) :520–531, 2009.

[115] W. Jorgensen and J. Tirado-Rives. The opls potential functions for proteins. energy minimizations for crystals of cyclic peptides and crambin. *Journal of the American Chemical Society*, 110(6) :1657–1666, 1988.

[116] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics*, 79(2) :926–935, 1983.

[117] W. L. Jorgensen, D. S. Maxwell, and J. TiradoRives. Development and testing of the opls all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society*, 118(45) :11225–11236, 1996.

[118] T. Juhász, Z. Szeltner, V. Fulop, and L. Polgár. Unclosed β-propellers display stable structures : Implications for substrate access to the active site of prolyl oligopeptidase. *Journal of Molecular Biology*, 346(3) :907 – 917, 2005.

[119] W. Kabsch and C. Sander. Definition of secondary structure of protein given a set of 3d coordinates. *Biopolymers*, 22 :2577 – 2637, 1983.

[120] H. Kallel, N. Mousseau, and F. Schiettekatte. Evolution of the potential-energy surface of amorphous silicon. *Physical Review Letters*, 105(4), 2010.

[121] G. A. Kaminski, R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen. Evaluation and reparametrization of the opls-aa force field for proteins via comparison with accurate quantum chemical calculations on peptides. *Journal of Physical Chemistry B*, 105(28) :6474–6487, 2001.

[122] C. Kandt and D. P. Tieleman. Holo-btuf stabilizes the open conformation of the vitamin b12 abc transporter btucd. *Proteins : Structure, Function and Genetics*, 78(3) :738–753, 2010.

[123] C. Kandt, W. L. Ash, and D. P. Tieleman. Setting up and running molecular dynamics simulations of membrane proteins. *Methods*, 41(4) :475–88, 2007.

[124] L. Karim-Béland, P. Brommer, F. El-Mellouhi, J.-F. Joly, and N. Mousseau. The kinetic activation relaxation technique. *Physical Review E*, (sous presse), 2011.

[125] M. Karttunen, J. Rottler, I. Vattulainen, and C. Sagui. Electrostatics in biomolecular simulations : Where are we now and where are we heading ? *Current Topics in Membranes*, 60 :49 – 89, 2008.

[126] J. Kästner. Umbrella sampling. *Wiley Interdisciplinary Reviews : Computational Molecular Science*, 1(6) :932–942, 2011.

[127] J. Kästner and W. Thiel. Bridging the gap between thermodynamic integration and umbrella sampling provides a novel analysis method : "umbrella integration". *Journal of Chemical Physics*, 123(14), 2005.

[128] K. Kaszuba, T. Róg, J.-F. St-Pierre, P. T. Mannisto, M. Karttunen, and A. Bunker. Molecular dynamics study of prolyl oligopeptidase with inhibitor in binding cavity. *SAR & QSAR in Environmental Research*, 20(7-8) :595–609, 2009.

[129] K. Kaszuba, T. Róg, K. Bryl, I. Vattulainen, and M. Karttunen. Molecular dynamics simulations reveal fundamental role of water as factor determining affinity of binding of β-blocker nebivolol to β2-adrenergic receptor. *Journal of Physical Chemistry B*, 114 :8374–8386, 2010.

[130] A. Kolinski and J. Skolnick. Assembly of protein structure from sparse experimental data : An efficient monte carlo model. *Proteins : Structure, Function and Genetics*, 32(4) :475–494, 1998.

[131] L. Komzsik. *The Lanczos method : evolution and application*. Software, environments, tools. SIAM Society for Industrial and Applied Mathematics, Philadelphia, 2003. Louis Komzsik. ill. ; 26 cm.

[132] V. Krishnamani and J. K. Lanyi. Molecular dynamics simulation of the unfolding of individual bacteriorhodopsin helices in sodium dodecyl sulfate micelles. *Biochemistry*, 51(6) :1061–1069, 2012.

[133] G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack. Improved prediction of protein side-chain conformations with scwrl4. *Proteins : Structure, Function and Genetics*, 77(4) :778–95, 2009.

[134] P. Kruger, S. Verheyden, P. J. Declerck, and Y. Engelborghs. Extending the capabilities of targeted molecular dynamics : simulation of a large conformational transition in plasminogen activator inhibitor 1. *Protein Science*, 10(4) :798–808, 2001.

[135] A. Kryshtafovych and K. Fidelis. Protein structure prediction and model quality assessment. *Drug Discovery Today*, 14(7-8) :386–93, 2009.

[136] S. Kumar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of Computational Chemistry*, 13(8) :1011–1021, 1992.

[137] R. Laghaei, N. Mousseau, and G. Wei. Effect of the disulfide bond on the monomeric structure of human amylin studied by combined hamiltonian and temperature replica exchange molecular dynamics simulations. *Journal of Physical Chemistry B*, 114(20) :7071–7077, 2010.

[138] R. Laghaei, N. Mousseau, and G. Wei. Structure and thermodynamics of amylin dimer studied by hamiltonian-temperature replica exchange molecular dynamics simulations. *Journal of Physical Chemistry B*, 115(12) :3146–3154, 2011.

[139] G. Lamoureux, A. Javelle, S. Baday, S. Wang, and S. Berneche. Transport mechanisms in the ammonium transporter family. *Transfusion Clinique et Biologique*, 17(3) :168–75, 2010.

[140] A. Y. Lau and B. Roux. The hidden energetics of ligand binding and activation in a glutamate receptor. *Nature Structural & Molecular Biology*, 18(3) :283–U62, 2011.

[141] J. Lawandi, S. Toumieux, V. Seyer, P. Campbell, S. Thielges, L. Juillerat-Jeanneret, and N. Moitessier. Constrained peptidomimetics reveal detailed geometric requirements of covalent prolyl oligopeptidase inhibitors. *Journal of Medicinal Chemistry*, 52 :6672–6684, 2009.

[142] J. Lawandi, S. Gerber-Lemaire, L. Juillerat-Jeannere, and N. Moitessier. Inhibitors of prolyl oligopeptidases for the therapy of human diseases : Defining diseases and inhibitors. *Journal of Medicinal Chemistry*, 53 :3423–3438, 2010.

[143] J. Lawson, M. L. O'Mara, and I. D. Kerr. Structure-based interpretation of the mutagenesis database for the nucleotide binding domains of p-glycoprotein. *Biochimica Et Biophysica Acta-Biomembranes*, 1778(2) :376–391, 2008.

[144] T. Lazaridis and M. Karplus. Effective energy function for proteins in solution. *Proteins : Structure, Function and Genetics*, 35(2) :133–152, 1999.

[145] D. S. Lee, C. Seok, and J. Lee. Protein loop modeling using fragment assembly. *Journal of the Korean Physical Society*, 52(4) :1137–1142, 2008.

[146] J. Lee, D. Lee, H. Park, E. A. Coutsias, and C. Seok. Protein loop modeling by using fragment assembly and analytical loop closure. *Proteins : Structure, Function and Genetics*, 78(16) :3428–3436, 2010.

[147] P. E. Leopold, M. Montal, and J. N. Onuchic. Protein folding funnels - a kinetic approach to the sequence structure relationship. *Proceedings of the National Academy of Sciences of the United States of America*, 89(18) :8721–8725, 1992.

[148] C. Levefelt and D. Lundh. A fold-recognition approach to loop modeling. *Journal of Molecular Modeling*, 12(2) :125–139, 2006.

[149] M. Li, C. Chen, D. R Davies, and T. K Chiu. Induced-fit mechanism for prolyl endopeptidase. *Journal of Biological Chemistry*, 285(28) :21487–95, 2010.

[150] Y. H. Li, I. Rata, S. W. Chiu, and E. Jakobsson. Improving predicted protein loop structure ranking using a pareto-optimality consensus method. *Bmc Structural Biology*, 10 :14, 2010.

[151] M. S. Lin and T. Head-Gordon. Improved energy selection of nativelike protein loops from loop decoys. *Journal of Chemical Theory and Computation*, 4(3) : 515–521, 2008.

[152] M. S. Lin, N. L. Fawzi, and T. Head-Gordon. Hydrophobic potential of mean force as a solvation function for protein structure prediction. *Structure*, 15(6) : 727–40, 2007.

[153] E. Lindahl, B. Hess, and D. van der Spoel. Gromacs 3.0 : a package for molecular simulation and trajectory analysis. *Journal of Molecular Modeling*, 7(8) :306–317, 2001.

[154] K. Linton and C. Higgins. Structure and function of abc transporters : the atp switch provides flexible control. *Pflugers Archiv European Journal of Physiology*, 453 :555–567, 2007.

[155] K. J. Linton. Structure and function of abc transporters. *Physiology*, 22(2) :122–130, 2007.

[156] K. J. Linton and C. F. Higgins. Structure and function of abc transporters : the atp switch provides flexible control. *Pflugers Archiv European Journal of Physiology*, 453(5) :555–67, 2007.

[157] J. Liphardt, J. , S. Dumont, S. B. Smith, Jr. Tinoco, I., and C. Bustamante. Equilibrium information from nonequilibrium measurements in an experimental test of jarzynski's equality. *Science*, 296(5574) :1832–5, 2002.

[158] J.-M. Liu, M. Kusinski, V. Ilic, J. Bignon, N. Hajem, J. Komorowski, K. Kuzdaz, H. Stepien, and J. Wdzieczak-Bakala. Overexpression of the angiogenic tetrapeptide acsdkp in human malignant tumors. *Anticancer Res.*, 28(5A) :2813 – 2817, 2008.

[159] M. J. Liu, T. Sun, J. Hu, W. Chen, and C. Wang. Study on the mechanism of the btuf periplasmic-binding protein for vitamin b12. *Biophysical Chemistry*, 135 (1-3) :19 – 24, 2008.

[160] P. Liu, F. Q. Zhu, D. N. Rassokhin, and D. K. Agrafiotis. A self-organizing algorithm for modeling protein loops. *PLOS Computational Biology*, 5(8) :11, 2009.

[161] K. P. Locher, A. T. Lee, and D. C. Rees. The e. coli btucd structure : a framework for abc transporter architecture and mechanism. *Science*, 296(5570) :1091–8, 2002.

[162] E. Machado-Charry, L. Karim-Beland, D. Caliste, L. Genovese, T. Deutsch, N. Mousseau, and P. Pochet. Optimized energy landscape exploration using the ab initio based activation-relaxation technique. *Journal of Chemical Physics*, 135 (3) :034102, 2011.

[163] R. Malek and N. Mousseau. Dynamics of lennard-jones clusters : A characterization of the activation-relaxation technique. *Physical Review E*, 62(6 Pt A) : 7723–8, 2000.

[164] D. J. Mandell, E. A. Coutsias, and T. Kortemme. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature Methods*, 6(8) :551–552, 2009.

[165] P. T. Männistö, J. I. Venäläinen, A. Jalkanen, and J. A. García-Horsman. Prolyl oligopeptidase : a potential target for the treatment of cognitive disorders. *Drug News & Perspective*, 20(5) :293 – 305, 2007.

[166] D. Mantle, G. Falkous, S. Ishiura, P. J. Blanchard, and E. K. Perry. Comparison of proline endopeptidase activity in brain tissue from normal cases and cases with

alzheimer's disease, lewy body dementia, parkinson's disease and huntington's disease. *Clinica Chimica Acta*, 249 :129 – 139, 1996.

[167] M. Marchi and P. Ballone. Adiabatic bias molecular dynamics : A method to navigate the conformational space of complex molecular systems. *Journal of Chemical Physics*, 110(8) :3697–3702, 1999.

[168] M. C. Marinica, F. Willaime, and N. Mousseau. Energy landscape of small clusters of self-interstitial dumbbells in iron. *Physical Review B*, 83(9), 2011.

[169] R. Merkel, P. Nassoy, A. Leung, K. Ritchie, and E. Evans. Energy landscapes of receptor-ligand bonds explored with dynamic force spectroscopy. *Nature*, 397 (6714) :50–3, 1999.

[170] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6) :1087–1092, 1953.

[171] M. Mezei. Efficient monte carlo sampling for long molecular chains using local moves, tested on a solvated lipid bilayer. *Journal of Chemical Physics*, 118(8) : 3874–3879, 2003.

[172] E. Michalsky, A. Goede, and R. Preissner. Loops in proteins (lip)–a comprehensive loop database for homology modelling. *Protein Eng*, 16(12) :979–85, 2003.

[173] M. Mills and I. Andricioaei. An experimentally guided umbrella sampling protocol for biomolecules. *Journal of Chemical Physics*, 129(11) :114101, 2008.

[174] R. S. Molday, M. Zhong, and F. Quazi. The role of the photoreceptor abc transporter abca4 in lipid transport and stargardt macular degeneration. *Biochimica Et Biophysica Acta-Molecular and Cell Biology of Lipids*, 1791(7) :573–583, 2009.

[175] P. Morain, P. Lestage, G. De Nanteuil, R. Jochemsen, J.-L. Robin, D. Guez, and P.-A. Boyer. S 17092 : A prolyl endopeptidase inhibitor as a potential therapeutic

drug for memory impairment. preclinical and clinical studies. *CNS Drug Reviews*, 8(1) :31–52, 2002.

[176] N. Mousseau and G. T. Barkema. Traveling through potential energy landscapes of disordered materials : The activation-relaxation technique. *Physical Review E*, 57(2) :2419–2424, 1998.

[177] N. Mousseau, P. Derreumaux, and G. Gilbert. Navigation and analysis of the energy landscape of small proteins using the activation-relaxation technique. *Physical Biology*, 2(4) :S101–7, 2005.

[178] Morad Mustafa, Douglas J Henderson, and David D Busath. Free-energy profiles for ions in the influenza m2-tmd channel. *Proteins : Structure, Function and Genetics*, 76(4) :794–807, 2009.

[179] J. Nasica-Labouze, M. Meli, P. Derreumaux, G. Colombo, and N. Mousseau. A multiscale approach to characterize the early aggregation steps of the amyloid-forming peptide gnnqqny from the yeast prion sup-35. *PLOS Computational Biology*, 7(5), 2011.

[180] S. Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Molecular Physics*, 52 :255–268, 1984.

[181] A. F. Oberhauser, P. E. Marszalek, H. P. Erickson, and J. M. Fernandez. The molecular elasticity of the extracellular matrix protein tenascin. *Nature*, 393(6681) : 181–5, 1998.

[182] H. Oberhofer, C. Dellago, and P. L. Geissler. Biased sampling of nonequilibrium trajectories : can fast switching simulations outperform conventional free energy calculation methods ? *Journal of Physical Chemistry B*, 109(14) :6902–15, 2005.

[183] Z. Okten, L. S. Churchman, R. S. Rock, and J. A. Spudich. Myosin vi walks hand-over-hand along actin. *Nature Structural & Molecular Biology*, 11(9) :884–7, 2004.

[184] M. L. Oldham, D. Khare, F. A. Quiocho, A. L. Davidson, and J. Chen. Crystal structure of a catalytic intermediate of the maltose transporter. *Nature*, 450(7169) : 515–21, 2007.

[185] S. A. Oliveira, A. M. Baptista, and C. M. Soares. Conformational changes induced by atp-hydrolysis in an abc transporter : A molecular dynamics study of the sav1866 exporter. *Proteins : Structure, Function and Genetics*, 79(6) :1977–1990, 2011.

[186] E. O. Oloo. The dynamics of the mgatp-driven closure of malk, the energy-transducing subunit of the maltose abc transporter. *Journal of Biological Chemistry*, 281 :28397–28407, 2006.

[187] E. O. Oloo and D. P. Tieleman. Conformational transitions induced by the binding of mgatp to the vitamin b12 atp-binding cassette (abc) transporter btucd. *Journal of Biological Chemistry*, 279 :45013 –45019, 2004.

[188] E. O. Oloo, C. K.C Kandt, L. O. Megan, M. L. Mara, and D. P. Tieleman. Computer simulations of abc transporter components. *Biochemistry and cell biology*, 84(6) :900–911, 2006.

[189] M. A. Olson, M. Feig, and C. L. Brooks. Prediction of protein loop conformations using multiscale modeling methods with physical energy scoring functions. *Journal of Computational Chemistry*, 29(5) :820–831, 2008.

[190] M. L. O'Mara and D. P. Tieleman. P-glycoprotein models of the apo and atp-bound states based on homology with sav1866 and malk. *FEBS Letters*, 581(22) : 4217–22, 2007.

[191] A. Onufriev, D. Bashford, and D. A. Case. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins : Structure, Function and Genetics*, 55(2) :383–94, 2004.

[192] P. J. O'Reilly, M. T. Hardison, P. L. Jackson, X. Xu, R. J. Snelgrove, A. Gaggar, F. S. Galin, and J. E. Blalock. Neutrophils contain prolyl endopeptidase and generate the chemotactic peptide, pgp, from collagen. *Journal of Neuroimmunology*, 217(1-2) :51 – 54, 2009.

[193] E. Paci and M. Karplus. Forced unfolding of fibronectin type 3 modules : an analysis by biased molecular dynamics simulations. *Journal of Molecular Biology*, 288(3) :441–59, 1999.

[194] Y. Pan, B. B. Stocks, L. Brown, and L. Konermann. Structural characterization of an integral membrane protein in its natural lipid environment by oxidative methionine labeling and mass spectrometry. *Analytical Chemistry*, 81 :28–35, 2009.

[195] S. Park and K. Schulten. Calculating potentials of mean force from steered molecular dynamics simulations. *Journal of Chemical Physics*, 120(13) :5946–5961, 2004.

[196] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals : a new molecular dynamics method. *Journal of Applied Physics*, 52(12) :7182–7190, 1981.

[197] M. Patra, M. Karttunen, M.T. HyvÃűnen, E. Falck, P. Lindqvist, and I. Vattulainen. Molecular dynamics simulations of lipid bilayers : Major artifacts due to truncating electrostatic interactions. *Biophysical Journal*, 84(6) :3636 – 3645, 2003.

[198] M. Patra, M. Karttunen, M. T. Hyvönen, E. Falck, P. Lindqvist, and I. Vattulainen. Molecular dynamics simulations of lipid bilayers : Major artifacts due to truncating electrostatic interactions. *Biophysical Journal*, 84(6) :3636–3645, 2003.

[199] M. Patra, M. Karttunen, M. T. Hyvanen, E. Falck, and I. Vattulainen. Lipid bilayers driven to a wrong lane in molecular dynamics simulations by subtle changes in long-range electrostatic interactions. *The Journal of Physical Chemistry B*, 108(14) :4485–4494, 2004.

[200] H. P. Peng and A. S. Yang. Modeling protein loops with knowledge-based prediction of sequence-structure alignment. *Bioinformatics*, 23(21) :2836–2842, 2007.

[201] H. W. Pinkett, A. T. Lee, P. Lum, K. P. Locher, and D. C. Rees. An inward-facing conformation of a putative metal-chelate-type abc transporter. *Science*, 315(5810) :373–377, 2007.

[202] L. Polgár. The prolyl oligopeptidase family. *Cellular and Molecular Life Sciences*, 59(2) :349 – 362, 2002.

[203] J.W. Ponder and D.A Case. Force fields for protein simulatioin. *Advanced Protein Chemistry*, 66 :27–85, 2003.

[204] D. Qiu, P. S. Shenkin, F. P. Hollinger, and W. C. Still. The gb/sa continuum model for solvation. a fast analytical method for the calculation of approximate born radii. *Journal of Physical Chemistry A*, 101(16) :3005–3014, 1997.

[205] Q. Qu, P. L. Russell, and F. J. Sharom. Stoichiometry and affinity of nucleotide binding to p-glycoprotein during the catalytic cycle. *Biochemistry*, 42(4) :1170–7, 2003.

[206] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1) :95 – 99, 1963.

[207] D. Rea and V. Fülöp. Structure-function properties of prolyl oligopeptidase family enzymes. *Cell Biochemistry and Biophysics*, 44(3) :349 – 365, 2006.

[208] D. C. Rees, E. Johnson, and O. Lewinson. Abc transporters : the power to change. *Nature Reviews Molecular Cell Biology*, 10 :218–227, 2009.

[209] D. Rodriguez-Gomez and E. Darve. Assessing the efficiency of free energy calculation methods. *Journal of Chemical Physics*, 120(8) :3563–3578, 2004.

[210] T. Róg, H. Martinez-Seara, N. Munck, M. Oresic, K. Mikko, and I. Vattulainen. Role of cardiolipins in the inner mitochondrial membrane - insight gained through atom-scale simulations. *Journal of Physical Chemistry B*, 113 :3413–3422, 2009.

[211] C. A. Rohl, C. E. M. Strauss, D. Chivian, and D. Baker. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins : Structure, Function and Genetics*, 55(3) :656–677, 2004.

[212] B. Roux. The calculation of the potential of mean force using computer-simulations. *Comput. Phys. Commun.*, 91(1-3) :275–282, 1995.

[213] J.-P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints : molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3) :327–341, 1977.

[214] R. Samudrala and J. Moult. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology*, 275(5) :895–916, 1998.

[215] B. Sankaran, S. Bhagat, and A. E. Senior. Inhibition of p-glycoprotein atpase activity by beryllium fluoride. *Biochemistry*, 36(22) :6847–6853, 1997.

[216] S. Santini, G. Wei, N. Mousseau, and P. Derreumaux. Pathway complexity of alzheimer's beta-amyloid abeta16-22 peptide assembly. *Structure*, 12(7) :1245–55, 2004.

[217] T. Sato, A. Kodan, Y. Kimura, K. Ueda, T. Nakatsu, and H. Kato. Functional role of the linker region in purified human p-glycoprotein. *FEBS Journal*, 276(13) : 3504–16, 2009.

[218] M. Schaefer, C. Bartels, and M. Karplus. Solution conformations and thermodynamics of structured peptides : Molecular dynamics simulation with an implicit solvation model. *Journal of Molecular Biology*, 284(3) :835–848, 1998.

[219] A. H. Schinkel and J. W. Jonker. Mammalian drug efflux transporters of the atp binding cassette (abc) family : an overview. *Advanced drug delivery reviews*, 55 (1) :3–29, 2003.

[220] J. Schlitter, M. Engels, and P. Kruger. Targeted molecular dynamics : a new approach for searching pathways of conformational transitions. *Journal of Molecular Graphics*, 12(2) :84–9, 1994.

[221] E. Schneider and S. Hunke. Atp-binding-cassette (abc) transport systems : Functional and structural aspects of the atp-hydrolyzing subunits/domains. *FEMS Microbiology Reviews*, 22 :1–20, 1998.

[222] J. S. Schneider, M. Giardiniere, and P. Morain. Effects of the prolyl endopeptidase inhibitor s 17092 on cognitive deficits in chronic low dose mptp-treated monkeys. *Neuropsychopharmacology*, 26(2) :176–82, 2002.

[223] I. Schulz, U. Zeitschel, T. Rudolph, D. Ruiz-Carrillo, J.-U. Rahfeld, B. Gerhartz, V. Bigl, H.-U. Demuth, and S. Rossner. Subcellular localization suggests novel functions for prolyl endopeptidase in protein secretion. *Journal of Neurochemistry*, 94 :970 – 979, 2005.

[224] B. D. Sellers, K. Zhu, S. Zhao, R. A. Friesner, and M. P. Jacobson. Toward better refinement of comparative models : Predicting loops in inexact environments. *Proteins : Structure, Function and Genetics*, 72(3) :959–971, 2008.

[225] L. Shan, I. L. Mathews, and C. Khosla. Structural and mechanistic analysis of two prolyl endopeptidases : Role of interdomain dynamics in catalysis and specificity. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10) :3599 – 3604, 2005.

[226] F. J. Sharom. Abc multidrug transporters : structure, function and role in chemoresistance. *Pharmacogenomics*, 9(1) :105–27, 2008.

[227] D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo,

J. P. Grossman, C. Richard Ho, D. J. Ierardi, I. Kolossvary, J. L. Klepeis, T. Layman, C. McLeavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, and S. C. Wang. Anton, a special-purpose machine for molecular dynamics simulation. *Communications of the Acm*, 51(7) :91–97, 2008.

[228] J. Sonne, C. Kandt, G. H. Peters, F. Y. Hansen, M. O. Jensen, and D. P. Tieleman. Simulation of the coupling between nucleotide binding and transmembrane domains in the atp binding cassette transporter btucd. *Biophysical Journal*, 92(8) : 2727–34, 2007.

[229] V. Z. Spassov, P. K. Flook, and L. Yan. Looper : a molecular mechanics-based algorithm for protein loop prediction. *Protein Engineering Design & Selection*, 21(2) :91–100, 2008.

[230] J.-F. St-Pierre. *Méthodes de simulations moléculaires accélérées : application et développement*. 2006. Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de Maâtre ès sciences (M.Sc.) en Bio-informatique.

[231] J.-F. St-Pierre and N. Mousseau. Large loop conformation sampling using the activation relaxation technique art-nouveau method. *Proteins : Structure, Function, and Bioinformatics*, sous presse, 2012.

[232] J. F. St-Pierre, N. Mousseau, and P. Derreumaux. The complex folding pathways of protein a suggest a multiple-funnelled energy landscape. *Journal of Chemical Physics*, 128(4) :045101, 2008.

[233] J.-F. St-Pierre, M. Karttunen, N. Mousseau, T. Róg, and A. Bunker. Use of umbrella sampling to calculate the entrance/exit pathway for z-pro-prolinal inhibitor in prolyl oligopeptidase. *Journal of Chemical Theory and Computation*, 7(6) : 1583–1594, 2011.

[234] J.-F. St-Pierre, A. Bunker, T. Róg, M. Karttunen, and N. Mousseau. Molecular dynamics simulations of the bacterial abc transporter sav1866 in the closed form.

The Journal of Physical Chemistry B, 116(9) :2934–2942, 2012. doi : 10.1021/jp209126c.

[235] M. Stepniewski, A. Bunker, M. Pasenkiewicz-Gierula, M. Karttunen, and T. Rog. Effects of the lipid bilayer phase state on the water membrane interface. *Journal of Physical Chemistry B*, 114(36) :11784–11792, 2010.

[236] S. X. Sun. Equilibrium free energies from path sampling of nonequilibrium trajectories. *Journal of Chemical Physics*, 118(13) :5769–5775, 2003.

[237] Z. Szeltner and L. Polgár. Structure, function and biological relevance of prolyl oligopeptidase. *Current Protein & Peptide Science*, 9(1) :96 – 107, 2008.

[238] Z. Szeltner, V. Renner, and L. Polgár. Substrate- and ph-dependent contribution of oxyanion binding site to the catalysis of prolyl oligopeptidase, a paradigm of the serine oligopeptidase family. *Protein Science*, 9 :353–360, 2000.

[239] Z. Szeltner, D. Rea, T. Juhász, V. Renner, Z. Mucsi, G. Orosz, V. Fülöp, and L. Polgár. Substrate-dependent competency of the catalytic triad of prolyl oligopeptidase. *Journal of Biological Chemistry*, 277(47) :44597 – 44605, 2002.

[240] Z. Szeltner, D. Rea, V. Renner, V. Fülöp, and L. Polgár. Electrostatic effects and binding determinants in the catalysis of prolyl oligopeptidase. *Journal of Biological Chemistry*, 277(45) :42613 – 42622, 2002.

[241] Z. Szeltner, D. Rea, V. Renner, L. Juliano, V. Fülöp, and L. Polgár. Electrostatic environment at the active site of prolyl oligopeptidase is highly influential during substrate binding. *Journal of Biological Chemistry*, 278(49) :48786 – 48793, 2003.

[242] Z. Szeltner, D. Rea, T. Juhász, V. Renner, V. Fülöp, and L. Polgár. Concerted structural changes in the peptidase and the propeller domains of prolyl oligopeptidase are required for substrate binding. *Journal of Molecular Biology*, 340(3) : 627 – 637, 2004.

[243] Zoltan Szeltner, Veronika Renner, and Laszlo Polgár. The noncatalytic β-propeller domain of prolyl oligopeptidase enhances the catalytic capability of the peptidase domain. *Journal of Biological Chemistry*, 275(20) :15000 – 15005, 2000.

[244] T. Tarrago, J. Martin-Benito, E. Sabido, B. Claasen, S. Madurga, M. Gairi, J. M. Valpeusta, and E. Giralt. A new side opening on prolyl oligopeptidase revealed by electron microscopy. *FEBS Letters*, 583(20) :3344 – 3348, 2009.

[245] J. Tenorio-Laranga, F. Coret-Ferrer, B. Casanova-Estruch, M. Burgal, and J. A. García-Horsman. Prolyl oligopeptidase is inhibited in relapsing-remitting multiple sclerosis. *Journal of Neuroinflammation*, 7 :23, 2010.

[246] R. F. Tilton, J. C. Dewan, and G. A. Petsko. Effects of temperature on protein-structure and dynamics - x-ray crystallographic studies of the protein ribonuclease-a at 9 different temperatures from 98-k to 320-k. *Biochemistry*, 31 (9) :2469–2481, 1992.

[247] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation : Umbrella sampling. *Journal of Computational Physics*, 23(2) :187–199, 1977.

[248] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation : Umbrella sampling. *Journal of Computational Physics*, 23 :187–199, 1977.

[249] S. C. E. Tosatto, E. Bindewald, J. Hesser, and R. Manner. A divide and conquer approach to fast loop modeling. *Protein Engineering*, 15(4) :279–286, 2002.

[250] L. Tskhovrebova, J. Trinick, J. A. Sleep, and R. M. Simmons. Elasticity and unfolding of single molecules of the giant muscle protein titin. *Nature*, 387(6630) : 308–312, 1997.

[251] K. Umemura, K. Kondo, Y. Ikeda, T. Kobayashi, Y. Urata, and M. Nakashima. Pharmacokinetics and safety of jtp-4819, a novel specific orally active prolyl endopeptidase inhibitor, in healthy male volunteers. *British Journal of Clinical Pharmacology*, 43(6) :613–618, 1997.

[252] C. van der Does and R. Tampe. How do abc transporters drive transport ? *Biological Chemistry*, 385(10) :927–933, 2004.

[253] A. R. van Gool, R. Verkerk, D. Fekkes, S. Sleijfer, M. Bannink, W. H. Kruit, B. van der Holt, S. Scharpé, A. M. M. Eggermont, G. Stoter, and M. W. Hengeveld. Plasma activity of prolyl endopeptidase in relation to psychopathology during immunotherapy with ifn-α in patients with renal cell carcinoma. *Journal of Interferon & Cytokine Research*, 28(5) :283 – 286, 2008.

[254] S. Vanni, P. Campomanes, M. Marcia, and U. Rothlisberger. Ion binding and internal hydration in the multidrug resistance secondary active transporter norm investigated by molecular dynamics simulations. *Biochemistry*, 51(6) :1281–1287, 2012.

[255] V. Vasiliou, K. Vasiliou, and D. W. Nebert. Human atp-binding cassette (abc) transporter family. *Human Genomics*, 3(3) :281–90, 2009.

[256] S. Velamakanni, Y. Yao, D. A. Gutmann, and H. W. van Veen. Multidrug transport by the abc transporter sav1866 from staphylococcus aureus. *Biochemistry*, 47 (35) :9300–8, 2008.

[257] Č. Venclovas, A. Zemla, K. Fidelis, and J. Moult. Assessment of progress over the casp experiments. *Proteins : Structure, Function and Genetics*, 53(S6) :585–595, 2003.

[258] H. Vocks, M. V. Chubynsky, G. T. Barkema, and N. Mousseau. Activated sampling in complex materials at finite temperature : The properly obeying probability activation-relaxation technique. *Journal of Chemical Physics*, 123(24), 2005.

[259] J. E. Walker, M. Saraste, M. J. Runswick, and N. J. Gay. Distantly related sequences in the alpha-subunits and beta-subunits of atp synthase, myosin, kinases and other atp-requiring enzymes and a common nucleotide binding fold. *Embo Journal*, 1(8) :945–951, 1982.

[260] G. L. Wang and R. L. Dunbrack. Pisces : a protein sequence culling server. *Bioinformatics*, 19(12) :1589–1591, 2003.

[261] J. Wang, Y. Deng, and B. Roux. Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophysical Journal*, 91(8) :2798–2814, 2006.

[262] J. M. Wang, P. Cieplak, and P. A. Kollman. How well does a restrained electrostatic potential (resp) model perform in calculating conformational energies of organic and biological molecules ? *Journal of Computational Chemistry*, 21(12) : 1049–1074, 2000.

[263] L. Wang, B. J. Berne, and R. A. Friesner. On achieving high accuracy and reliability in the calculation of relative protein-ligand binding affinities. *Proceedings of the National Academy of Sciences of the United States of America*, 109(6) : 1937–1942, 2012.

[264] A. Ward, C. L. Reyes, J. Yu, C. B. Roth, and G. Chang. Flexibility in the abc transporter msba : Alternating access with a twist. *Proceedings of the National Academy of Sciences of the United States of America*, 104(48) :19005–10, 2007.

[265] W. J. Wedemeyer and H. A. Scheraga. Exact analytical loop closure in proteins using polynomial equations. *Journal of Computational Chemistry*, 20(8) :819–844, 1999.

[266] G. Wei, N. Mousseau, and P. Derreumaux. Complex folding pathways in a simple beta-hairpin. *Proteins : Structure, Function and Genetics*, 56(3) :464–474, 2004.

[267] G. H. Wei, N. Mousseau, and P. Derreumaux. Exploring the energy landscape of proteins : A characterization of the activation-relaxation technique. *Journal of Chemical Physics*, 117(24) :11379–11387, 2002.

[268] S. J. Weiner, P. A. Kollman, D. T. Nguyen, and D. A. Case. An all atom force-field for simulations of proteins and nucleic-acids. *Journal of Computational Chemistry*, 7(2) :230–252, 1986.

[269] J.-W. Weng, K.-N. Fan, and W.-N. Wang. The conformational transition pathway of atp binding cassette transporter msba revealed by atomistic simulations. *Journal of Biological Chemistry*, 285(5) :3053–3063, 2010.

[270] J. Wong-ekkabut, M. S. Miettinen, C. Dias, and M. Karttunen. Static charges cannot drive a continuous flow of water molecules through a carbon nanotube. *Nature Nanotechnology*, 5 :555 – 557, 2010.

[271] J. Wong-Ekkabut, M. S Miettinen, C. Dias, and M. Karttunen. Static charges cannot drive a continuous flow of water molecules through a carbon nanotube. *Nature Nanotechnology*, 5(8) :555–557, 2010.

[272] Z. X. Xiang, C. S. Soto, and B. Honig. Evaluating conformational free energies : The colony energy and its application to the problem of loop prediction. *Proceedings of the National Academy of Sciences of the United States of America*, 99 (11) :7432–7437, 2002.

[273] C. Xing and R. Faller. Density imbalances and free energy of lipid transfer in supported lipid bilayers. *Journal of Chemical Physics*, 131(17) :175104, 2009.

[274] H. Xiong, A. Crespo, M. Marti, D. Estrin, and A. E. Roitberg. Free energy calculationswith non-equilibrium methods : applications of the jarzynski relationship. *Theoretical Chemistry Accounts*, 116 :338–346, 2006.

[275] A. S. Yang and L. Y. Wang. Local structure prediction with local structure-based sequence profiles. *Bioinformatics*, 19(10) :1267–1274, 2003.

[276] S. Yesylevskyy, S.-J. Marrink, and A. E. Mark. Alternative mechanisms for the interaction of the cell-penetrating peptides penetratin and the tat peptide with lipid bilayers. *Biophysical Journal*, 97(1) :40–9, 2009.

[277] A. Yildiz, J. N. Forkey, S. A. McKinney, T. Ha, Y. E. Goldman, and P. R. Selvin. Myosin v walks hand-over-hand : single fluorophore imaging with 1.5-nm localization. *Science*, 300(5628) :2061–5, 2003.

[278] Y. Yoshimoto, K. Kado, F. Matsubara, N. Koriyama, H. Kaneto, and D. Tsura. Specific inhibitors for prolyl endopeptidase and their anti-amnesic effect. *Journal of pharmacobio-dynamics*, 10 :730 – 735, 1987.

[279] J. Young and I. B. Holland. Abc transporters : bacterial exporters-revisited five years on. *Biochim Biophys Acta*, 1461(2) :177–200, 1999.

[280] F. M. Ytreberg and D. M. Zuckerman. Single-ensemble nonequilibrium path-sampling estimates of free energy differences. *Journal of Chemical Physics*, 120 (23) :10876, 2004.

[281] F. M. Ytreberg, R. H. Swendsen, and D. M. Zuckerman. Comparison of free energy methods for molecular systems. *Journal of Chemical Physics*, 125 : 184114, 2006.

[282] J. Zaitseva, C. Oswald, T. Jumpertz, S. Jenewein, A. Wiedenmann, I. B. Holland, and L. Schmitt. A structural analysis of asymmetry required for catalytic activity of an abc-atpase domain dimer. *EMBO Journal*, 25(14) :3432–3443, 2006.

[283] C. Zhang, S. Liu, and Y. Q. Zhou. Accurate and efficient loop selections by the dfire-based all-atom statistical potential. *Protein Science*, 13(2) :391–399, 2004.

[284] H. Zhou and Y. Zhou. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science*, 11(11) :2714–26, 2002.

[285] K. Zhu, D. L. Pincus, S. W. Zhao, and R. A. Friesner. Long loop prediction using the protein local optimization program. *Proteins : Structure, Function and Genetics*, 65(2) :438–452, 2006.

[286] K. Zhu, M. R. Shirts, and R. A. Friesner. Improved methods for side chain and loop predictions via the protein local optimization program : Variable dielectric model for implicitly improving the treatment of polarization effects. *Journal of Chemical Theory and Computation*, 3(6) :2108–2119, 2007.

[287] J. K. Zolnerciks, C. Wooding, and K. J. Linton. Evidence for a sav1866-like architecture for the human multidrug transporter p-glycoprotein. *FASEB Journal*, 21(14) :3937–3948, 2007.