Conception de peptide cyclique et étude de l'interaction protéine-protéine par des méthodes d'échantillonnage accélérée

Jaysen Sawmynaden

Sorbonne Université

École doctorale n°397 : Physique et Chimie des Matériaux

Institut de Minéralogie, de Physique des Matériaux et de CosmochimieUnité de Bioinformatique et Biophysique

Conception de peptide cyclique et étude de l'interaction protéine-protéine par des méthodes d'échantillonnage accélérée

Thèse de doctorat présenté par
Jaysen SAWMYNADEN

Dirigée par Fabio PIETRUCCI
Dirk STRATMANN
Jacques CHOMILIER

pour obtenir le grade de
DOCTEUR DE SORBONNE UNIVERSITÉ

Soutenance prévue le 16 octobre 2020 devant un jury composé de :

Mme. Anne-Claude Camproux	Professeur – Université Paris Diderot	Présidente du jury
M. Dirk Stratmann	MCU – Sorbonne Université	Co-encadrant
M. Fabio Pietrucci	MCU – Sorbonne Université	Co-encadrant
M. Jacques Chomilier	Directeur de recherche CNRS	Directeur de thèse
Mme. Jessica Andreani	CEA – Saclay	Examinatrice
M. Manuel Dauchez	Professeur - Université de Reims Champagne-Ardenne	Rapporteur
M. Matthieu Montes	Professeur - CNAM	Rapporteur
M. Patrick Fuchs	MCU Université de Paris, HDR	Examinateur

RÉSUMÉ

Les protéines sont des macromolécules biologiques présentes dans toutes les cellules vivantes et codées dans le code génétique. Ellels assurent la grande majorité des fonctions biologiques (enzyme, reconnaissance de molécules...). Dans le milieu cellulaire, l'activité (inactivation ou l'inhibition) d'une protéine est souvent conditionnée par l'interaction protéines protéines (IPP).

Pouvoir inhiber ces IPP a un intérêt pour comprendre des mécanismes moléculaires au sein de la cellule ou bien dans le développement de nouveaux médicaments. Par rapport à de petites molécules, les peptides cycliques (qui sont des protéines de moins de 50 résidus) sont particulièrement adaptés pour bloquer des IPP. En effet, les peptides cycliques possèdent une grande surface d'interaction tout en ayant une bonne stabilité conformationnelle ce qui, en théorie, leur permet d'avoir une bonne spécificité et affinité pour leur cible. A l'heure actuelle, peu de méthodes sont disponibles pour prédire les structures les plus probables de peptides cycliques, de même pour ce qui est de la prédiction des IPP et des mécanismes qui ont lieu durant le processus d'association.

Dans cette thèse nous présentons les travaux effectués dans l'élaboration d'un protocole qui vise à concevoir et échantillonner le paysage conformationnel de peptides cycliques. Cet échantillonnage des conformations est réalisé à l'aide de dynamique moléculaire avec échange de répliques. Nous montrons qu'avec cette méthode, les prédictions des structures les plus probables sont proches des conformations résolues expérimentalement. Enfin pour ce qui est de l'étude des interactions IPP et des mécanismes qui ont lieu durant le processus d'association, nous présentons nos travaux dans l'élaboration d'un protocole de métadynamique avec biais échangés. Cette méthode d'échantillonnage accéléré de dynamique moléculaire permet de prédire l'affinité de liaison d'un complexe protéine protéine ainsi que les états métastables. Pour cette partie nous présentons deux applications sur un système protéine peptide cyclique et protéine protéine. Nous montrons également que le choix des variables collectives est important dans l'élaboration d'une méthode fiable et que la prédiction des constantes cinétiques reste encore difficile d'accès.

TABLE DES MATIÈRES

LISTE DES ABRÉVIATIONS ET ANNOTATIONS

Interaction protéine protéine : IPP
van der Walls : VDW
Dynamique moléculaire : DM
Replica exchange molecular dynamics : REMD
VC : variable collective
Métadynamique avec biais échangés (metad-BE)
Root mean square deviation : RMSD
Profil de densité de probabilité : PDP
Kullback–Leibler : KL
JSD : *Jensen–Shannon divergence*
MCC : Monte Carlo cinétique

CHAPITRE I

INTRODUCTION

I.1 PROTÉINES

Les protéines sont des macromolécules biologiques présentes dans toutes les cellules vivantes et codées dans le code génétique. Elles sont formées d'une ou plusieurs chaînes polypeptidiques. Chacune de ces chaînes est constituée de briques élémentaires, appelées acides aminés et qui sont reliés entre eux par des liaisons peptidiques (Figure I.1.2).

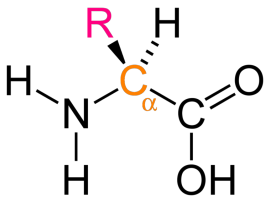


Fig. I.1.1 *Schéma d'un acide aminé*

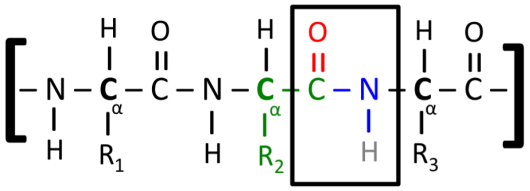


Fig. I.1.2 *Les liaisons peptidiques lient les acides aminés*

Les propriétés physico-chimiques des acides aminés qui composent les protéines dépendent de leur chaîne latérale. Dans le vivant, 22 acides aminés sont protéinogènes, dont 20 standards (la sélénocystéine et la pyrolysine sont, quant à elles, codés indirectement). Ces 20 acides aminés sont généralement regroupés en cinq catégories suivant leurs propriétés physico-chimiques (Tab I.1.1).

Liste des acides aminés				
Non polaire	Polaire	Chargé +	Chargé -	Aromatique
Glycine	Sérine	Lysine	Aspartate	Phénylalanine
Alanine	Thréonine	Arginine	Glutamate	Tyrosine
Valine	Cystéine	Histidine		Tryptophane
Leucine	Proline			
Isoleucine	Asparagine			
Méthionine	Glutamine			

TABLE I.1.1 – Liste des 20 acides aminés standards

A l'exception de la glycine, les acides aminés sont asymétriques. Il existe deux formes d'acides aminés suivant leur chiralité (figure I.1.3), une forme L et une forme D. La forme L est celle qui compose majoritairement les protéines. Tandis que les acides aminés sous forme D sont présents chez les orga-

nismes supérieurs[1,2] dans certains neuropeptides et peptides opioïdes ou bien sous forme libre (D-Asp et D-Ser) dans certains tissus[1,3,4].

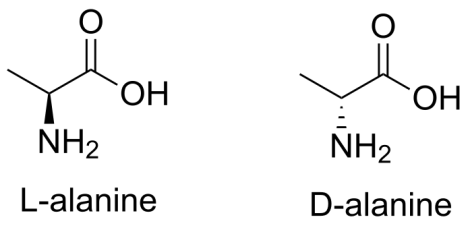


Fig. I.1.3 *Les acides aminés sous forme L composent majoritairement les protéines.*

I.2 STRUCTURES

L'analyse des génomes d'eucaryotes, d'eubactéries et d'archées[5] (effectuée en 2012) montre une disparité concernant la taille moyenne des protéines chez ces différents taxons. Les séquences protéiques ont en moyenne une longueur de 472 résidus chez les eucaryotes, tandis qu'elle est de 320 et 283 résidus, chez les procaryotes et archées (figure I.2.1). Cette longueur plus importante chez les eucaryotes s'explique par la duplication et la fusion de gènes.

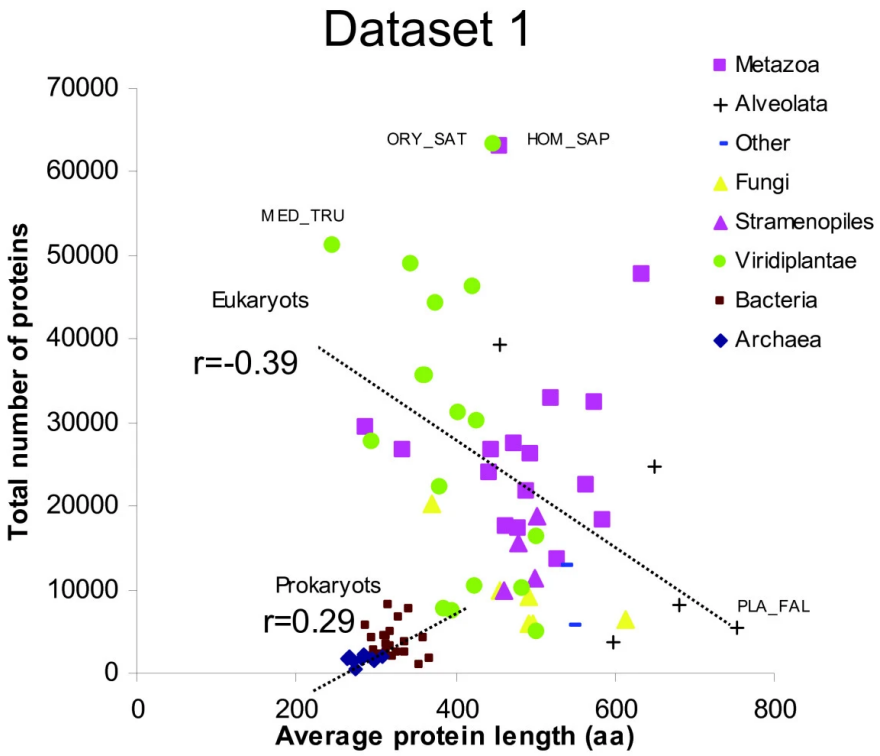


Fig. I.2.1 *Graphique issu de la ref[5]. Sur un jeu de données contenant 7.3 millions de séquences protéines (1.2 millions provenant de 1 302 procaryotes et 6.1 millions issus de 140 espèces d'eucaryotes), la taille moyenne des protéines est différente chez les eucaryotes, eubactéries et archées. Les eucaryotes ont en moyenne des protéines plus longue à cause de la duplication et la fusion de gènes qui ont lieu au cours de l'évolution.*

L'énergie contenue dans les liaisons hydrogènes, les ponts disulfures, l'attraction entre les charges (positives et négatives) et les radicaux hydrophobes ou hydrophiles, structurent la protéine. La structure d'une protéine peut être décomposée en quatre niveaux[6]. Le niveau de base de la structure des protéines, appelé structure primaire, est la séquence linéaire des acides aminés. Cependant, une protéine ne garde jamais une forme strictement linéaire.

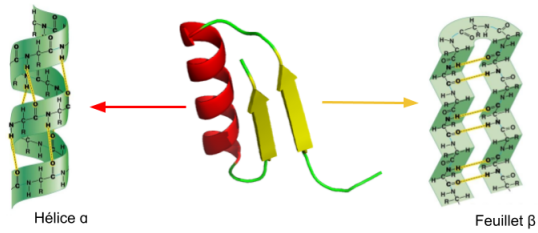


Fig. I.2.2 *Les interactions non covalentes structurent localement les protéines.*

Le premier niveau d'organisation structurelle est la structure secondaire (Figure I.2.2). Elle est due principalement aux liaisons hydrogènes formées par le squelette peptidique. Les structures qui la composent sont les feuillets β, les coudes, les hélices (α, π, poly-proline...).

Cette structuration de la protéine induit donc des changements de conformations au niveau du squelette peptidique. Ainsi les angles entre les acides aminés ne sont pas fixes. Trois angles dihèdres sont utilisés au niveau du squelette peptidique (figure I.2.3).

— ϕ : défini par les quatre atomes successifs $CO_{(1)} - NH_{(2)} - C_{\alpha(2)} - CO_{(2)}$
— psi : défini par les quatre atomes successifs $NH_{(2)} - C_{\alpha(2)} - CO_{(2)} - NH$
— ω : défini par les quatre atomes successifs $C_{\alpha(1)} - CO_{(1)} - NH_{(2)} - C_{\alpha(2)}$

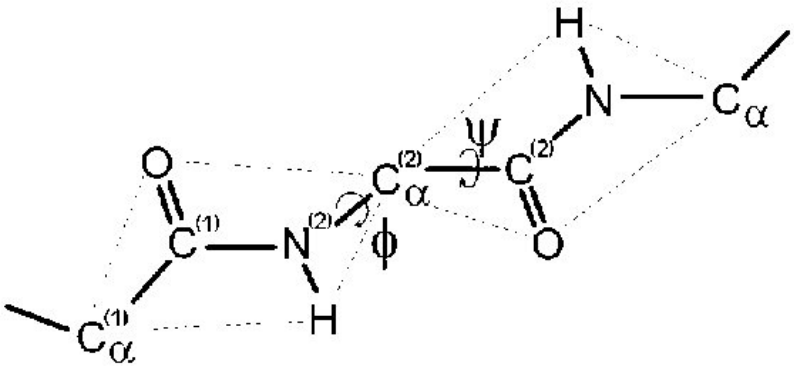


Fig. I.2.3 *Figure issue de la ref*[7] *qui représente les angles ϕ et ψ. Les protéines adoptent des structures secondaires, ainsi les angles entre les acides aminés ne sont pas fixes. Les valeurs des angles ϕ et ψ permettent de caractériser les structures secondaires.*

La liaison peptidique étant plane, l'angle ω vaut 180° (pour les protéines) et seuls les angles ϕ et ψ ont leurs valeurs qui varient. En étudiant les combinaisons des angles ϕ et ψ au sein des protéines et en les projetant sur un diagramme 2D[8] (figure I.2.4), Ramachandran a observé que la majorité des acides aminés prennent certaines valeurs ϕ et ψ. Cette spécificité s'explique par les contraintes stériques qui limitent les configurations possibles pour les angles dihèdres au niveau du squelette peptidique.

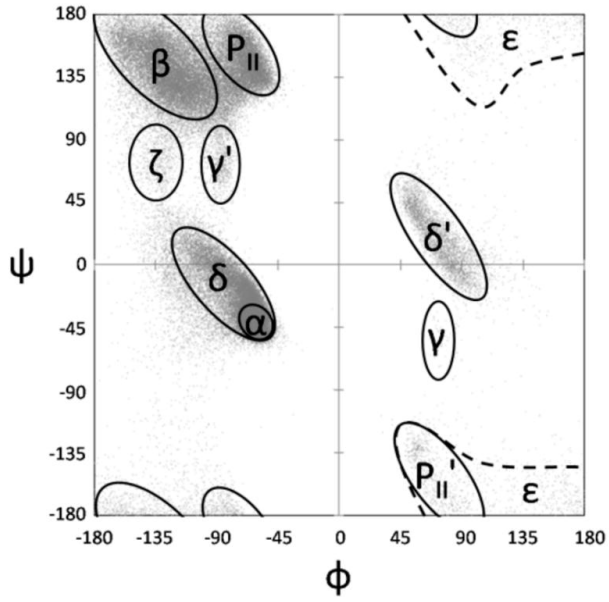


Fig. I.2.4 *Figure issue de la ref[9] qui représente le diagramme de Ramachandran avec les annotations indiquant des zones correspondant à un type de structure secondaire.*

La structure tertiaire est l'organisation globale de la protéine (figure I.2.5). Les interactions hydrophobes, les liaisons ioniques, les interactions de van der Walls (VDW) et les liaisons covalentes (comme les ponts disulfures) contribuent au repliement de la protéine.

Enfin la structure quaternaire est le dernier niveau d'organisation. Elle concerne les complexes constitués de plusieurs chaînes protéiques (appelées sous-unités) ayant chacune sa propre structure tertiaire. C'est cet assemblage des sous-unités qui confère à la protéine sa fonction (figure I.2.6).

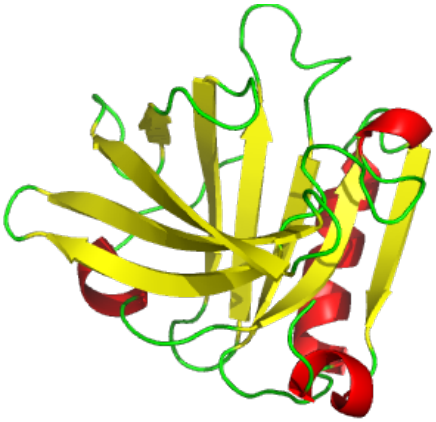


Fig. I.2.5 *Structure tertiaire d'une protéine.*

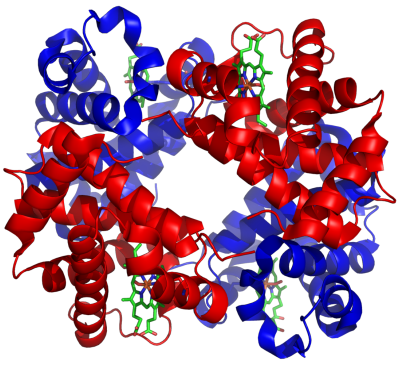


Fig. I.2.6 *L'hémoglobine est un tétramère.*

A l'instar des protéines, les peptides sont des macromolécules constituées d'acides aminés. Chimiquement semblable aux protéines, la dénomination peptide réfère (de manière générale) à des protéines de petite taille (2 à 50 acides aminés) adoptant des structures secondaires mais peu de structures tertiaires[10]. Les peptides sont présents aussi bien chez les eucaryotes que les procaryotes et ont des fonctions diverses et variées (neuropeptides, peptide endocrinien, peptide rénal...)[11]).

I.3 PROBLÉMATIQUE

La plupart des fonctions biologiques sont assurées par les protéines. En effet, l'enchaînement des acides aminés permet d'avoir une grande variété de structures et de fonctions :

— catalyse de réaction chimique (enzyme)
— transport de molécules (hémoglobine, canaux ioniques...)
— messagers (insuline, hormone de croissance...)
— reconnaissance de molécules (immunoglobulines, récepteurs cellulaire...)
— structure cellulaire (collagène, protéines du cytosquelette...)

Ces fonctions se font dans la grande majorité par le biais d'interactions protéine-protéine (IPP). On estime qu'il y a plus de 600 000 IPP chez l'homme[12]. Plus de 420 000 IPP uniques ont été répertoriées (d'après la base de données BioGRID[13]) et peuvent être aussi bien homomériques qu'hétéromériques. Les IPP peuvent être formés à la suite d'interaction forte avec longue durée d'association ou bien suite à de faibles interactions et avec une durée de vite courte (type transitoire)[14].

Concernant les interfaces des IPP, elles peuvent être constituées d'un coeur hydrophobe entouré de résidus polaires ou bien d'une mosaïque de résidus hydrophobes et polaires[15]. Enfin au niveau structural, les IPP peuvent être classées suivant les modifications structurales induites par l'association et les types de molécules qui interagissent[16] (figure I.3.1).

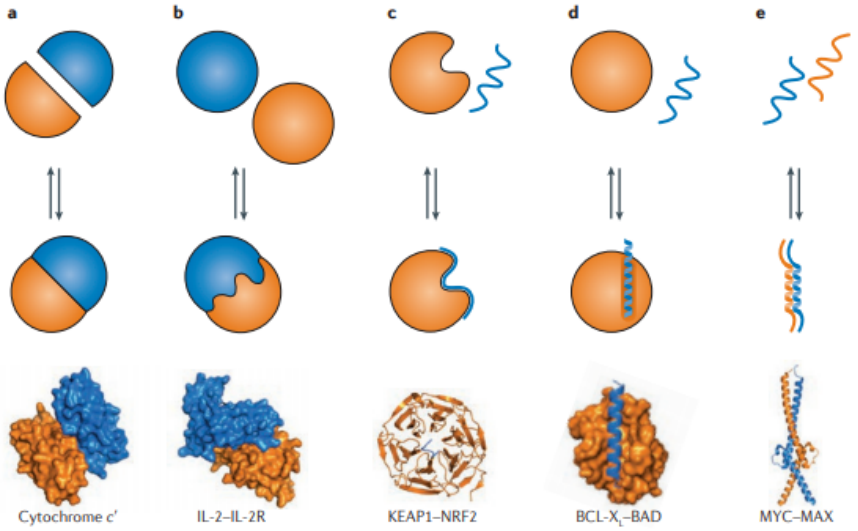


Fig. I.3.1 *Les différents types de structures présents dans les IPP (figure issu de la ref[14]). Les premières lignes sont une représentation schématique des protéines avec leurs partenaires (protéine ou peptide), tandis que la dernière ligne est un exemple de complexe adoptant ce type de conformation. a) complexe formé par deux protéines globulaires sans modification structurale (PDB 2CCY). b) complexe formé par deux protéines globulaires avec modification structurale (PDB 1Z92). c) complexe formé par une protéine globulaire rigide avec un peptide linéaire (PDB 2DYH). d) complexe formé par une protéine globulaire flexible avec un peptide linéaire (PDB 2XAO). e) complexe formé par deux peptides linéaires (PDB 1NKP)*

Bloquer les IPP permet d'étudier l'impact sur les interactions entre les différentes molécules biochimiques et les perturbations de l'activité au sein de la cellule (ou de l'organisme). A l'heure actuelle, environ 12 000 de ces IPPs ont été la cible d'un inhibiteur[17][18][19]. Pour ces raisons, l'étude de ces IPPs et leurs inhibitions a un grand intérêt scientifique, mais également un fort potentiel dans la découverte de

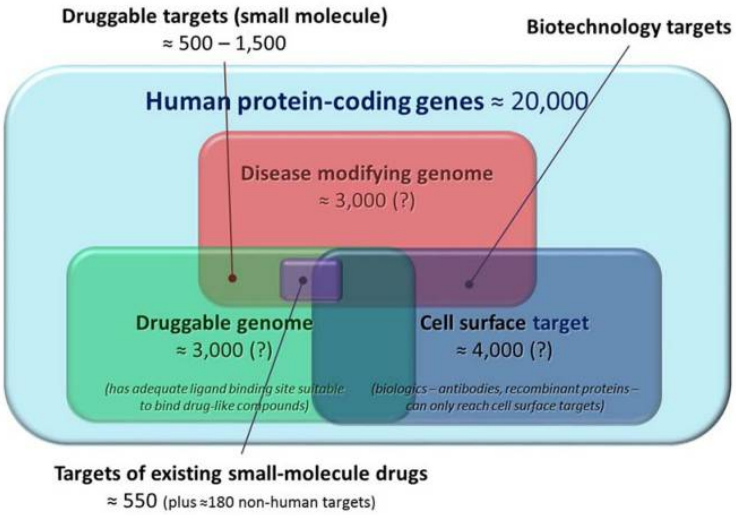


Fig. I.3.2 *Environ 3000 protéines sont impliquées dans des maladies, l'utilisation de petites molécules permet, potentiellement, de modifier l'activité d'une fraction de ces protéines tandis que pour les rest des protéines d'autres approches doivent être utilisées*[21].

nouvelles cibles thérapeutiques.

L'utilisation de petites molécules, traditionnellement avec une masse moléculaire inférieure à 500 Da, a l'avantage d'avoir une bonne biodisponibilité (mesure qui représente la part d'une substance active dans la circulation systémique après une administration autre qu'intraveineuse)[20]. En 2012, 364 médicaments non protéiques disponibles à la vente[21] ciblaient des protéines humaines. L'analyse des structures de protéines présentes dans la PDB révèle qu'environ 10% des protéines possèdent des poches pouvant servir de sites d'interactions pour de petites molécules[22 23 24]. Cependant seulement 10% de ces protéines sont impliquées dans des maladies (fig I.3.2). Ainsi l'inhibition des protéines par de petites molécules a un potentiel limité par rapport à l'ensemble des protéines exprimées chez l'humain (figure I.3.2).

Les interactions entre les protéines impliquent généralement plusieurs acides aminés avec de grandes surfaces d'interactions (entre 12 000 à 22 000 Å²)[14]. Du fait de leur taille, les petites molécules n'offrent pas une grande surface d'interaction, ce qui (potentiellement) ne garantit pas une bonne spécificité pour leur cible (figure I.3.3). Ce manque de spécificité peut engendrer des interactions non désirées avec d'autres molécules et donc induire des effets secondaires lors de tests *in vivo*.

A contrario, l'utilisation de grandes molécules (supérieur à 5000 Da) permet d'avoir une bonne spécificité. Par contre leur biodisponibilité par voie orale est très mauvaise et nécessite des injections intraveineuses. Pour surmonter cette limitation, l'utilisation de peptides peut être envisagée. Comme expliqué précédemment, les peptides ont une grande surface d'interaction et sont donc de parfaits candidats pour cibler des interactions protéine protéine. En outre, ils sont simples à synthétiser, ils offrent une grande variété de structures et peuvent avoir une bonne affinité et spécificité pour leurs cibles.

D'après la base de donnée StratPep[25], environ 3700 peptides avec une activité biologique ont été répertoriés. Bien que possédant un fort potentiel thérapeutique, les peptides dans leur forme linéaire ont l'inconvénient d'avoir une grande variabilité conformationnelle. En effet, du fait de leur taille plus réduite par rapport à des protéines, les peptides adoptent majoritairement des structures secondaires. Les degrés de libertés sont plus importants que des protéines stabilisées par des structures tertiaires. Pour ces raisons les peptides linéaires peuvent adopter différentes conformations, réduisant l'affinité de liaison pour leurs cibles. En outre, sans modification chimique, les peptides linéaires sont rapidement

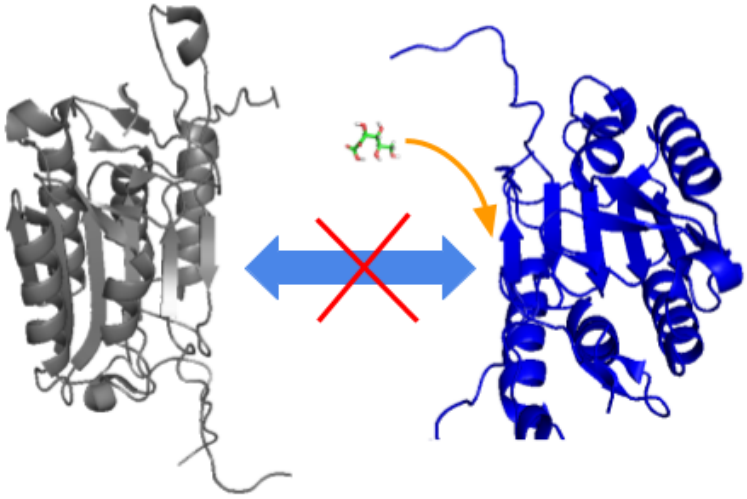


Fig. I.3.3 *Un petit ligand (en vert) n'a pas une une grande surface d'interaction pour inhiber l'association des deux protéines (gris et bleu)*

dégradés par les protéases de l'organisme lorsque qu'ils sont ingérés par voie orale (il y a donc une faible biodisponibilité). Une manière de stabiliser la conformation des peptides linéaires et d'augmenter leur résistance face aux protéases est de les cycliser[26] (figure I.3.4).

La cyclisation d'un peptide linéaire peut se faire de différentes manières. Il existe les cyclisassions simples avec le squelette peptidique en Nter-Cter, une cylisation impliquant deux chaînes latérales (pont disulfure, liaison amide), la cylisation entre chaîne latérale et squelette peptidique ou bien en utilisant plusieurs liaisons covalentes. Dans la nature, des peptides cycliques sont produits naturellement[27] chez des organismes avec des fonctions diverses comme par exemple d'antibiotique (daptomycine, polymyxine) ou bien d'immunosuppresseurs (cylcosporine A). Les peptides cycliques ont un potentiel pharmacochimique très important puisqu'ils peuvent avoir une grande variabilité de structure avec une stabilité conformationnelle, tout en étant résistants aux protéases de l'organisme (ce qui assure une meilleur biodisponibilité que des peptides linéaires). Pour ces raisons, l'utilisation de peptide cyclique comme inhibiteur d'IPP constitue un domaine de recherche très actif et une classe de biomédiacament à fort potentiel[28].

En 2017, on dénombrait 40 médicaments fondés sur des peptides cycliques et disponibles dans le commerce[26] et ce nombre est voué à augmenter dans les années à venir[28].Cependant plusieurs limitations sont présentes dans l'élaboration de peptide cyclique inhibiteur. La première est la perdition de la stabilité conformationnelle des peptides cycliques. En effet, contrairement aux protéines, dont un grand ensemble de structures 3D est disponible (plus de 150 000 dans la PDB[29]), peu de structures de peptides cycliques libre ont été résolues. En outre, les diagrammes de Ramachandran pour des peptides cycliques ont montré que leurs angles dièdres ϕ et ϕ peuvent avoir des valeurs différentes de celles observés pour des peptides linéaires ou bien des protéines globulaires[30].

A l'heure actuelle, il existe des outils de prédictions de structures de peptides cycliques. Nous pouvons citer le logiciel PepLook[31] qui a été le premier à prendre en compte les résidus sous forme D.Le projet Rosetta commons propose le stand alone Simple Cyclic Peptide Prediction[32] (qui prend en charge les résidus sous forme D). EGSCyP[33] est la seule méthode d'exploration exhaustive pour les pentapeptides cycliques qui prend en charge les résidus sous forme D et N-méthylés. Enfin d'autres méthodes utilisant une approche par modèle (fragment ou boucle) existent[34 35 36 37] mais ont souvent des limita-

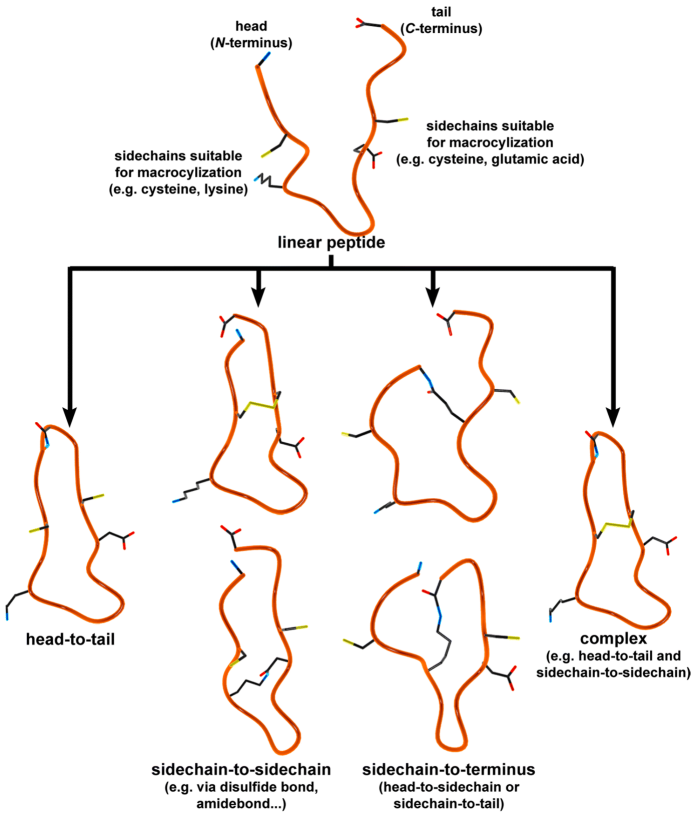


Fig. I.3.4 *Figure issue de la revue[26]. La cyclisation d'un peptide linéaire peut se faire de différentes manières. La cyclisation tête à queue Nter-Cter, via une une liaison covalente impliquant une chaîne latérale (pont disulfure, liaison amide), ou bien en utilisant plusieurs liaisons covalantes*

tions dans l'utilisation d'acide aminé (pas de forme D ou bien de résidus non naturels) ou dans le nombre de résidus qui composent les peptides.

Un point qui importe dans l'élaboration d'un inhibiteur est son affinité pour sa cible et son temps d'occupation avec celle-ci. Les protéines interagissent entre elles via des liaisons non covalentes au niveau de leurs chaînes latérales. Les forces régissant ces interactions peuvent être hydrophobes, polaires (liaisons hydrogènes), de type van der Waals ou bien électrostatiques (formation de ponts salins). Plusieurs méthodes expérimentales permettent d'étudier les interactions protéiques (*fluorescence resonance energy transfer* ,co-précipitation...). En plus de leur coût, ces méthodes nécessitent d'isoler les protéines d'intérêts. Suivant les systèmes étudiés, cette étape de purification et d'isolation peut s'avérer complexe, voire problématique (comme par exemple pour des protéines transmembranaires). Des méthodes *in silico* ont été développées pour prédire les structures de complexe ligand récepteur (LR), leur affinité de liaisons et les constantes cinétique. Ces dernières années, des améliorations ont été faites dans le domaine avec notamment des outils de prédiction de complexe LR[38 39] et des énergies libres relativement peu coûteux en calcul (exemple méthode de *docking*)[40 41 42].

A l'heure actuelle, la principale limitation est la prédiction des constantes d'association (k_{on}) et de dissociation (k_{off}). La première mesure l'affinité du ligand pour sa cible, lorsque les deux composés sont sous forme libre. Sa valeur dépend des concentrations des réactifs. Le k_{on} peut être estimé avec des outils peu coûteux en temps de calcul[43 44]. La constante de dissociation, quant à elle, peut être vue comme l'inverse du temps de résidence du ligand sur son récepteur (elle ne dépend pas de la concentration des réactifs). La dissociation d'un complexe protéique est un processus qui nécessite souvent des modifi-

cations des interfaces du ligand et du récepteur, avec des ruptures de liaisons hydrogènes, des ponts salins et des interactions hydrophobes ou bien polaires. Pour ces raisons, il est compliqué d'avoir des estimations de k_{off} avec des méthodes peu coûteuses en temps de calcul. Généralement, il est nécessaire d'utiliser des simulations de dynamique moléculaire biaisée pour avoir accès à cette constante cinétique d'un système.

Dans cette thèse nous présentons les travaux effectués dans l'élaboration d'un protocole qui vise à concevoir et échantillonner (de manière automatique) le paysage conformationnel de peptides, ainsi que dans la prédiction de l'affinité de liaison d'un complexe protéine protéine avec ses constantes cinétiques. Le premier chapitre de la thèse détaille les principes généraux sur lesquels nous nous appuyons pour nos travaux. Nous évoquons les principes de l'affinité de liaison, les constantes cinétiques, et également les méthodes d'échantillonnage accéléré utilisées. Le second concerne la mise en place d'un protocole d'échantillonnage conformationnel de peptides à l'aide de dynamique moléculaire avec répliques échangées. Nous présentons les résultats obtenus à partir d'un jeux de neuf peptides cycliques ayant fait l'objet d'études poussées dans la littérature, aussi bien numériquement qu'expérimentalement, ainsi que les limites de notre protocole. Enfin, le dernier chapitre concerne la mise en place d'un protocole généraliste pour prédire l'affinité de liaison et l'utilisation d'un modèle markovien pour prédire le k_{off}. Dans cette partie nous présentons l'application de notre protocole sur le complexe protéique barnase barstar.

CHAPITRE II

MÉTHODES

II.1 DYNAMIQUE MOLÉCULAIRE

La dynamique moléculaire (DM) est une méthode de simulation numérique qui reproduit les mouvements d'un ensemble de particules d'un système au cours du temps. La DM peut être classée en deux catégories suivant les forces appliquées au système :

— *ab initio* : applique des interactions quantiques sur des petits systèmes (quelques dizaines d'atomes)

— classique : les différentes interactions sont définies dans un champ de force empirique.

C'est cette dernière catégorie qui a été appliquée dans cette thèse, à l'aide du logiciel GROMACS[45]. Les mouvements des particules sont simulés en se fondant sur le second principe des lois de Newton (eq (II.1.1)). Dans un référentiel galiléen, la somme des forces qui s'exercent sur une particule donnée est égale au produit de sa masse par son vecteur d'accélération :

$$\sum \overrightarrow{f_i}(t) = m.\overrightarrow{a}(t) \tag{II.1.1}$$

— $\overrightarrow{f_i}$: différentes forces appliquées à la particule.

— $\overrightarrow{a}$: le vecteur d'accélération de la particule.

— $\overrightarrow{m}$: la masse de la particule.

Le vecteur d'accélération de la particule $\overrightarrow{a}$ correspond à la dérivée temporelle du vecteur vitesse $\overrightarrow{a} = \frac{d\overrightarrow{v}}{dt}$. Le vecteur vitesse ($\overrightarrow{v}$), quant à lui, correspond à la dérivé temporelle de la position $v = \frac{dx}{dt}$ (eq (II.1.2)) :

$$\frac{d\overrightarrow{v}}{dt} = \frac{d^2x}{dt^2} \tag{II.1.2}$$

— $\frac{\overrightarrow{dv}}{dt}$: dérivée temporelle du vecteur vitesse appliqué à la particule.

— x : coordonnées de la particule.

En reprenant l'équation (II.1.1), l'accélération (qui est la dérivée temporelle du vecteur vitesse) équivaut à $\frac{\sum \overrightarrow{f_i}(t)}{m} = \frac{d\overrightarrow{v}}{dt}$.

Ainsi pour un intervalle de temps très court δ_t, la position de la particule à l'instant $t+1$ peut être est donnée par l'équation (II.1.3) :

$$x_{t+1} = x_t + \frac{dx}{dt} \tag{II.1.3}$$

Or comme détaillé précédemment $\frac{dx}{dt}$ équivaut à $\overrightarrow{v}$. En remplaçant le terme, il est alors possible de calculer la nouvelle position de la particule en utilisant la vitesse ((II.1.4)) :

$$x_{t+1} = x_t + \delta_t \overrightarrow{v_t} \tag{II.1.4}$$

En appliquant le même raisonnement il est possible de déterminer la nouvelle vitesse de la particule (eq (II.1.5)) :

$$v_{t+1} = v_t + \frac{dv}{dt} \Leftrightarrow v_{t+1} = v_t + \delta_t \frac{\sum \vec{f_i}(t)}{m} \tag{II.1.5}$$

En pratique la résolution numérique de ces équations est approximée en utilisant l'algorithme de Verlet[46]. Ce dernier effectue un développement de Taylor au troisième ordre sur les positions x_t, un avec δ_t et un autre avec un $-\delta_t$ sur le temps (eq (II.1.6)).

$$\begin{aligned} x_{t-\delta_t} &= x_t - \delta_t \vec{v_t} + \delta_t^2 \frac{\sum \vec{f_i}(t)}{2m} - \delta_t^3 \frac{1}{3!} \frac{d^3x}{dt^3} + \mathcal{O}(\delta_t^4) \\ x_{t+\delta_t} &= x_t + \delta_t \vec{v_t} + \delta_t^2 \frac{\sum \vec{f_i}(t)}{2m} + \delta_t^3 \frac{1}{3!} \frac{d^3x}{dt^3} + \mathcal{O}(\delta_t^4) \end{aligned} \tag{II.1.6}$$

La somme de ces deux expressions donne :

$$x_{t+\delta_t} = 2x_t - x_{t-\delta_t} + \delta_t^2 \frac{\sum \vec{f_i}(t)}{m} + \mathcal{O}(\delta_t^4)$$

Soit

$$x_{t+\delta_t} = 2x_t - x_{t-\delta_t} + \delta_t^2 \vec{a}$$

Ainsi pour chaque intervalle de temps donnée (δ_t) la position est calculée. Cependant cet algorithme ne donne pas d'information sur la vitesse. Connaître les valeurs des vecteurs vitesses peut-être très utile, par exemple pour calculer l'énergie cinétique, ou d'autres grandeurs utilisant la vitesse. Il est possible de les obtenir en calculant le déplacement de la particule sur l'intervalle de temps $2\delta_t$ (eq (II.1.7)) :

$$v(t) = \frac{x_{t+\delta_t} - x_{t-\delta_t}}{2\delta_t} \tag{II.1.7}$$

Dans le cas des dynamiques moléculaires effectuées, une variante de l'algorithme de Verlet est utilisée, l'algorithme saute mouton (*leapfrog integration*) qui donne les vitesses à $+\delta_{\frac{t}{2}}$ et les positions à δ (figure II.1.1).

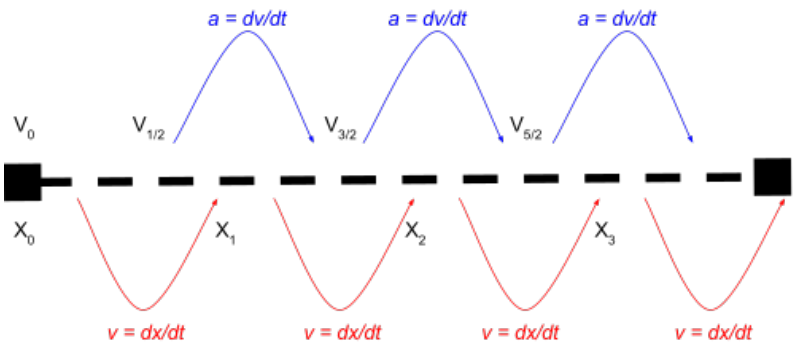


Fig. II.1.1 *Représentation de l'algorithme lleapfrog. La nouvelle vitesse calculée va servir à déterminer la nouvelle position lors du calcul suivant.*

.

Les caractéristiques des particules dépendent de leur nature, de même pour les forces qui peuvent s'appliquent sur eux. L'ensemble des paramètres qui régissent la structure et l'énergie d'un système sont définis dans ce que l'on appelle le champ de force. Les parties suivantes détaillent le principe d'un champ de force tout atomes et les différentes méthodes de modélisations du solvant utilisées.

II.2 CHAMP DE FORCE

Le champ de forces comprend différents paramètres utilisés pour la dynamique moléculaire, tels que les charges, les masses des atomes, les longueurs des liaisons atomiques... Dans le cas des champs de forces atomistiques, l'énergie que peut adopter un système est modélisée par une énergie potentielle (figure II.2.1 et équation (II.2.1)) somme de deux composantes : les énergies des termes liés (qui modélisent des interactions covalentes) et non liés (qui correspondent aux interactions à longue portée).

$$E_{\text{Potentiel}} = E_{\text{liée}} + E_{\text{non-liée}} \tag{II.2.1}$$

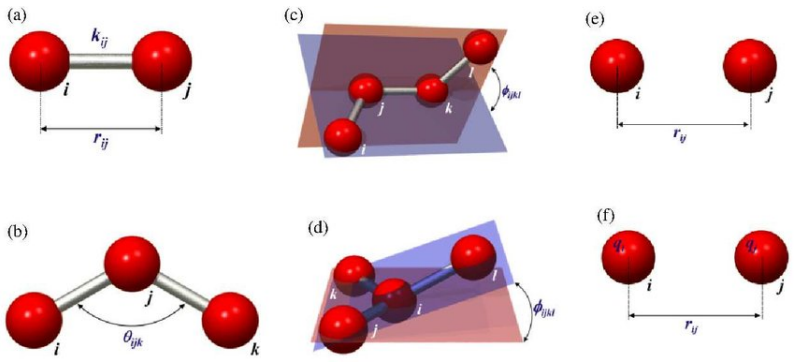


Fig. II.2.1 *Représentation des composantes de l'énergie potentielle. Figure issue de la Ref*[47]*. a) liaison covalente (k_{ij}) entre les atomes (i et j), distants de r_{ij}. b) angle θ_{ijk} formé par les atomes ijk. c) angle dièdre (ϕ_{ijkl}) entre les plans ijk et jkl. d) angle dièdre impropre (ϕ_{ijkl}) entre les plans ijk et jkl. e) interaction de van der Waals entre deux atomes (i et j) distants de r_{ij}. f) interaction électrostatique entre deux atomes (i et j) distants de r_{ij}.*

.

Les énergies des termes liés :
- énergie des liaisons covalentes
- énergie des angles de valence
- énergie des angles dièdres (rotation autour d'une liaison covalente)

Les termes non liés :
- énergie électrostatique
- énergie de van der Waals

Les paramètres des différents termes qui composent l'énergie potentielle dépendent du champ de forces utilisé. De nombreux champs de forces ont été développés pour différents domaines d'applications (lipides, protéines, acides aminés...) et différents modes de modélisation. Certains champs de forces modélisent tous les atomes (modèle atomistique), tandis que d'autres regroupent des atomes pour réaliser une modélisation en gros grain. Le couple solvant/champ de forces est un critère important en dynamique moléculaire. *Best et al.* ont réalisé un comparatif de 12 champs de forces avec des polyalanines[48]. Bien que les structures les plus représentatives soient cohérentes avec les données expérimentales, les auteurs ont montré que les champs de forces ont tendance à favoriser la présence d'hélices dans le cas de peptides linéaires. Sur les 12 champs de forces, seuls Amber03[49], CHARMM27/cmap[50 51], OPLS-aa/L[52] et Gromos43a1[53] ont une proportion de brinsβ en adéquation avec les données RMN.

SOLVANT EXPLICITE ET IMPLICITE

De par ses interactions polaires et ses liaisons hydrogènes, l'eau contribue aussi bien dans les interactions protéine protéine qu'à leurs repliements[54]. Le choix de la représentation du solvant est un critère important en dynamique moléculaire. On distingue deux catégories de solvatation : explicite et implicite. En solvant explicite, le modèle est atomistique et chaque molécule de solvant est rigide et est modélisée par plusieurs particules. Suivant le modèle choisi, le nombre de particules composant une molécule d'eau peut aller de 3 à 6 (trois atomes avec les doublets non liants et un site chargé). En plus du nombre de particules, la topologie (longueurs et angles) est également différente.

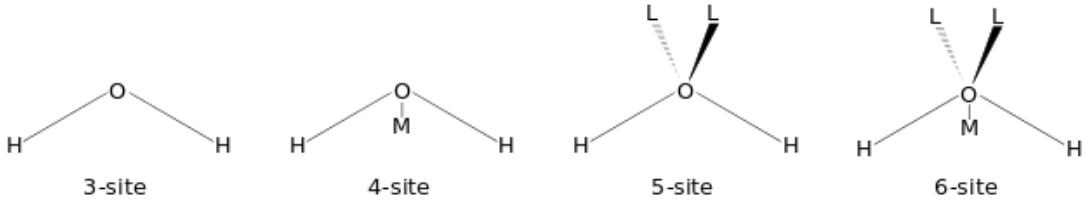


Fig. II.2.2 *Modèles de solvants explicites utilisés en dynamique moléculaire. A partir du modèle 4 sites, le doublet non liant est modélisé, soit au niveau de sa charge (M) ou bien de son occupation dans l'espace sites (L)*

Dans ces modèles de solvant explicite une molécule d'eau interagit avec le reste du solvant (ou les autres molécules) via des interactions non liantes (électrostatique et van der Waals). Cependant, simuler toutes les molécules d'eau est coûteux en temps de calcul. Il est parfois préférable d'utiliser un continuum qui va modéliser implicitement le comportement de l'eau aux voisinages des molécules. En plus d'être moins coûteux en temps de calcul, la viscosité est moindre comparée à un solvant explicite. Ce qui peut être considéré comme un inconvénient s'avère être un avantage dans l'étude de l'échantillonnage du paysage conformationnel puisqu'il est accéléré. En l'absence de modèle explicite de solvant, le solvant implicite approxime l'énergie libre de solvatation, en la décomposant en deux termes : la contribution non polaire et polaire[55]. La contribution non polaire peut être modélisée par l'équation (II.2.2) :

$$\Delta G_{\mathrm{np}} = \gamma A \tag{II.2.2}$$

Avec A la surface totale accessible au solvant et γ le coefficient de tension de surface.

Enfin le terme polaire est celui qui contribue le plus dans l'énergie de solvatation. Différents modèles existent, dont le generalized Born (GB)[56] (figure II.2.3 et équation (II.2.3)). Dans ce modèle chaque atome, dans une molécule, est représenté par une sphère de rayon (R) avec une charge q à son centre. Le solvant est modélisé par un continuum qui a va réduire les interactions électrostatiques ("effet d'écran des charges") du soluté par rapport au vide. Pour ce faire une constante diélectrique (ϵ_{out}) différente du voisinage de la molécule (ϵ_{in}) est appliquée.

$$\Delta G_{\mathrm{GB}} = -\frac{1}{2}\left(\frac{1}{\epsilon_{\mathrm{in}}} - \frac{1}{\epsilon_{\mathrm{out}}}\right)\sum_{i,j}\frac{q_{\mathrm{i}}q_{\mathrm{j}}}{f_{i,j}^{GB}(r_{i,j})} \tag{II.2.3}$$

— $f_{i,j}^{GB} = [r_{ij}^2 + R_i R_j exp(\frac{-r_{ij}^2}{4R_iR_j})]^{\frac{1}{2}}$
— r_{ij} est la distance entre les atomes i et j.
— q_i et q_j sont les charges partiels des atomes i et j.
— ϵ_{in} est la constante diélectrique interne de la molécule (fixée à 1)
— ϵ_{out} est la constante diélectrique de l'eau à 300K (80).

— R_i et R_j sont les rayons de Born effectifs des atomes i et j, peut être approximé comme la distance du centre de l'atome à la surface de la molécule.

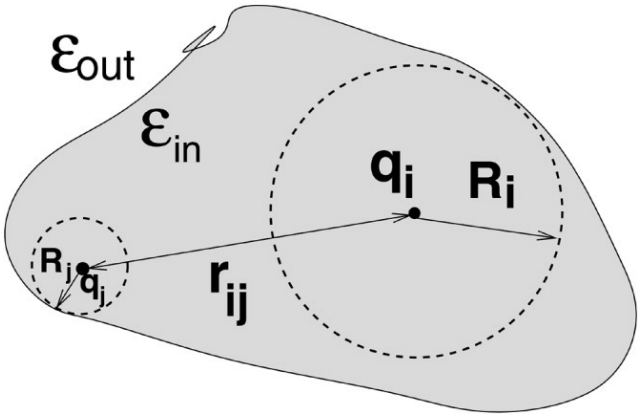


Fig. II.2.3 *Figure issue de la réf*[57]*. Représentation de la modélisation du solvant implicite generalized Born.* r_{ij} *est la distance entre les atomes* i *et* j*,* q_i *et* q_j *leurs charges partielles.* ϵ_{in} *et* ϵ_{out} *sont les constantes diélectriques interne de la molécule (1) et du milieu externe (80)*

Le volume de chaque atome dépend du rayon R. Dans le modèle BG, le rayon R correspond au rayon intrinsèque de l'atome (ρ) (qui peut être approximé comme le rayon de van der Waals) et de son environnement proche. Différents modèles de solvant implicite dérivés du modèle GB ont été proposés par la communauté (en calculant différemment R). Le modèle GB-HCT (Hawkins, Cramer et Truhlar)[58] estime le rayon R avec un coût de calcul réduit. Toutefois GB-HCT sous-estime le rayon R pour les atomes enfouis. Pour cette raison, d'autres modèles de solvant implicite ont été proposés. Le modèles GB-OBC (Onufriev, Bashford et Case)[59] utilise des paramètres empiriques pour corriger le terme R des atomes enfouis. Les solvants implicites Neck[60] et son dérivé Neck2[55] quant à eux améliorent la modélisation de R lorsque les distances inter-atomiques sont petites (inférieures à une molécule d'eau).

Dans le cas du solvant implicite, *Irene Maffucci et Alessandro Contini* ont réalisé un comparatif de différents couples champs de forces/ solvant implicite GB[61]. Le jeux de données utilisé comprend 8 peptides dont les structures ont été résolues. Deux peptides en hélices, trois adoptant une structure en épingle à cheveux β (β hairpin) et enfin trois peptides intrinsèquement désordonnée (PID). Les résultats du comparatif révèlent que le solvant implicite GB-Neck2 a de meilleurs résultats que les solvants GB-HCT et GB-OBC. En outre, il n'existe pas de couple champ de force/solvant implicite idéal dans la formation des structures secondaires en hélice, des β hairpins et des IDP. Les champs de forces ff99SB, ff99SBildn, et ff99SBildn-ϕ avec le solvant GB-Neck2 sont les plus précis[61]. Dans le cas de peptides adoptant une structure secondaire, le champ de forces Amber 96 couplé avec le solvant implicite GB-HCT ou GB-OBC(II) donne de meilleurs résultats. Afin de faire une comparaison future de nos résultats avec ceux obtenus par Maud jusot[33], les simulations de REMD ont été faites avec le champ de forces Amber 96 et le solvant implicit GB-OBC.

II.3 Dynamique moléculaire avec répliques échangées

L'étude du paysage conformationnel de peptides peut se faire à l'aide de simulations de dynamique moléculaire. Cependant l'obtention d'un échantillonnage de tout l'espace conformationnel qui soit ergodique, avec une distribution de Boltzmann, est pratiquement impossible avec une seule dynamique moléculaire. En effet la probabilité d'observer ($P(x)$) une réplique avec un état particulier, dépend de son énergie potentielle ($U(x)$) et de sa température (T) (équation (II.3.1)).

$$P(x) \propto \exp -\frac{U(x)}{k_B T} \tag{II.3.1}$$

Avec k_B, la constante de Boltzmann.

Ainsi pour s'échapper des minima énergétiques, le système doit franchir des barrières énergétiques. Plus ces barrières sont hautes et moins il est probable que la molécule d'intérêt explore de nouvelles conformations (Figure II.3.1).

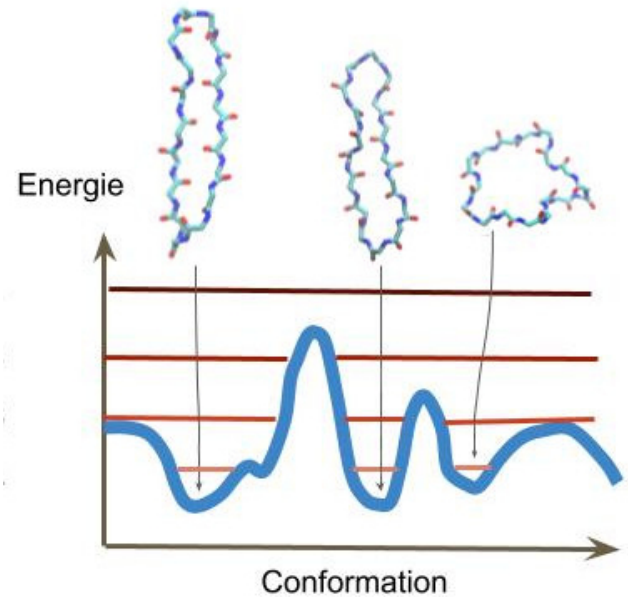


Fig. II.3.1 *Schéma représentant les conformations accessibles à des températures données. A plus haute température, toutes les conformations sont accessibles tandis qu'à plus basse température le système est bloqué dans des minima d'énergie*

Une des possibilités pour surmonter cette limitation est de réaliser des simulations de dynamiques moléculaires avec des répliques échangées (REMD)[62]. Cette méthode de dynamique moléculaire, permet à des systèmes moléculaires de franchir des barrières d'énergie libre infranchissable dans un temps raisonnable de simulation. En d'autre terme la REMD permet d'accélérer des processus, comme par exemple le changement de conformations de molécules[63], avec l'identification d'état de transitions et même d'association. Par exemple dans le cadre de l'étude du peptide amylin, *Ruxi Qi* et ses collaborateurs ont utilisé des simulations de REMD pour étudier les changements conformationnels du peptide (passant d'une hélice α à un brin β[64]) qui ont lieu avant l'association des protéines entre elles.

La REMD utilise plusieurs dynamiques moléculaires d'un même système avec des températures d'équilibre différentes (et donc des énergies différentes). Durant la simulation, les répliques vont tenter de s'échanger périodiquement leurs conformations (sans leur température) en suivant un critère de Metropolis (Equation II.3.2). De cette manière, les trajectoires à basse température peuvent franchir les

barrières énergétiques et échantillonner l'espace conformationnel de la molécule (Figure II.3.2). L'analyse quant à elle se fait sur la trajectoire à 300 K.

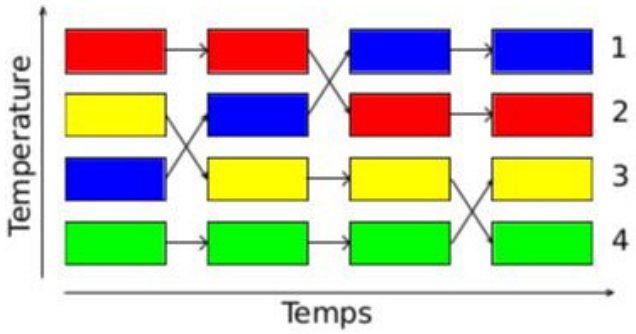


Fig. II.3.2 *Schéma représentant les tentatives d'échanges entre les réplica au cours de la simulation.*

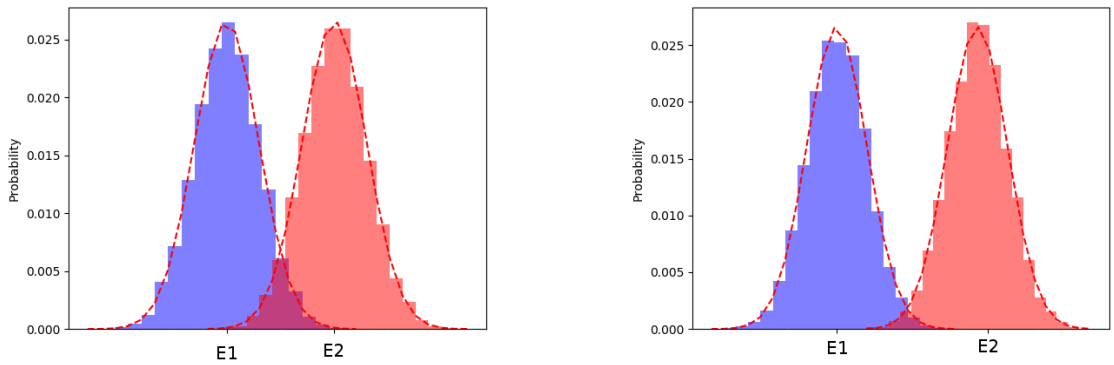
Le critère de Metropolis (équation (II.3.2)) donne la probabilité d'échange entre deux répliques à un instant donné. Il dépend de l'énergie des replica (E_1 et E_2), leurs températures (T_1 et T_2) et de la constante de Boltzmann (k_B) :

$$p = min(1, \frac{exp(-\frac{E_2}{k_B T_1} - \frac{E_1}{k_B T_2})}{exp(-\frac{E_1}{k_B T_1} - \frac{E_2}{k_B T_2})})$$

Ce qui donne :

$$p = min(1, exp((E_1 - E_2)(\frac{1}{k_B T_1} - \frac{1}{k_B T_2}))) \tag{II.3.2}$$

Les différentes énergies que peut prendre une molécule au cours d'une simulation peuvent être modélisées comme une loi normale (d'après le théorème central limite), centré sur une valeur moyenne E. Plus les moyennes des énergies de deux replica sont proches et plus la probabilité qu'ils aient la même valeur d'énergie augmente (figure II.3.3). Les replica ayant le même nombre d'atomes et étant tous à une pression constante de 1 bar, la modulation de l'énergie se fait uniquement en appliquant des températures différentes entre les replica.

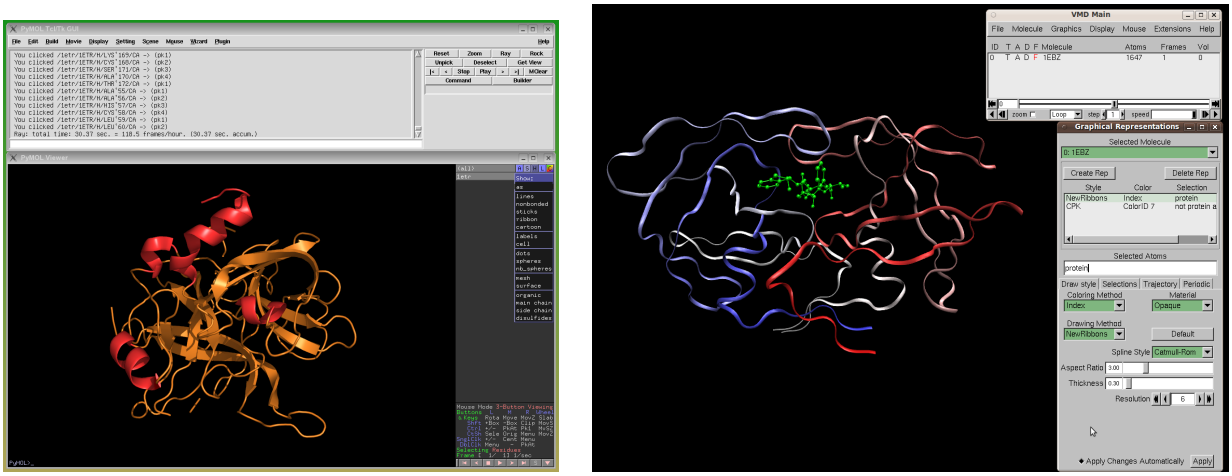


(a) *E1 et E2 proches* (b) *E1 et E2 éloignés*

Fig. II.3.3 *Schéma illustrant le recouvrement des énergies des replica dans le cas où leurs températures sont similaires ou non*

II.4 VISUALISATION DES STRUCTURES

Les dynamiques moléculaires réalisées avec le logiciel GROMACS[45] génèrent des fichiers binaires au format xtc. Ce format contient les coordonnées de tous les atomes, pour chaque *frame* générée au cours de la simulation. Afin de visualiser l'évolution du système au cours du temps, le logiciel VMD[65] (Visual Molecular Dynamics) est utilisé. Enfin pour les frames d'intérêt, les structures 3D peuvent être converties au format PDB. Ce fichier texte contient au minimum le nom des atomes et leur coordonnées. La visualisation des fichiers PDB s'est faite avec VMD mais également avec le logiciel Pymol[66] (fig II.4.1).



(a) *Le logiciel Pymol* (b) *Le logiciel VMD*

Fig. II.4.1 *Logiciels utilisés pour la visualisation de structures et de trajectoires de dynamiques moléculaires*

II.5 Energie libre de Gibbs et constante cinétique

A pression et température constantes (ce qui est le cas en biologie) une réaction chimique est possible uniquement si elle est favorable thermodynamiquement (figure II.5.1, équation (II.5.1)). Le sens d'évolution d'une réaction chimique peut être décrit par la fonction d'énergie libre de Gibbs (par la suite nommé énergie libre). La différence d'énergie ΔG entre l'état final (G_{final}) et l'état initial ($G_{t=0}$).

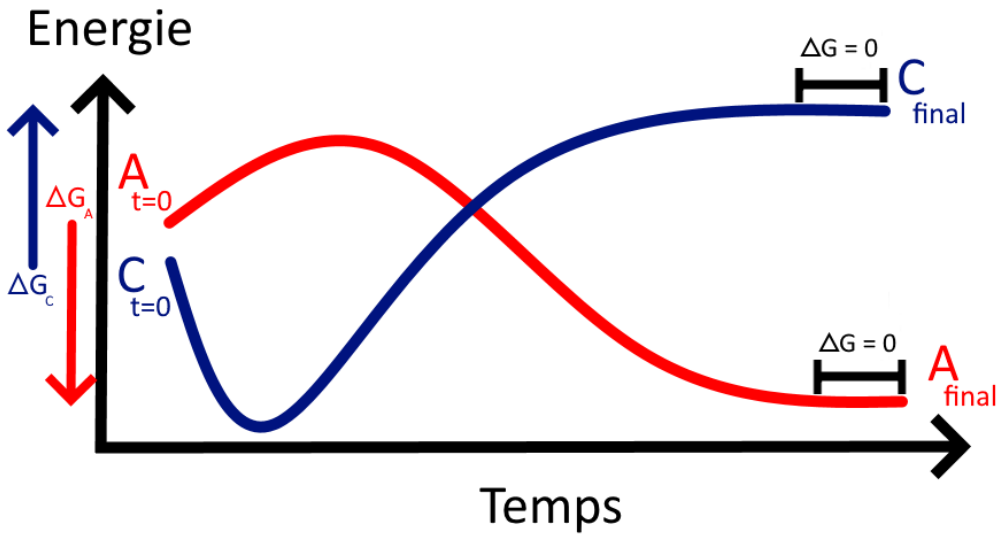


Fig. II.5.1 *Une réaction chimique est favorable lorsque la différence d'énergie entre l'état initial et final est négative (réaction A). Dans le cas contraire un apport d'énergie au système est nécessaire (réaction C).*

$$\Delta G = G_{\text{final}} - G_{\text{t=0}} \tag{II.5.1}$$

Dans le cas où ΔG < 0, le processus est exergonique. Dans le cas contraire, ΔG > 0, le processus est endergonique et un apport d'énergie du milieu extérieur est nécessaire pour que la réaction ait lieu. Enfin ΔG peut-être décomposé en deux paramètres : l'entropie S et l'enthalpie H pour une température T (équation (II.5.2)).

$$\Delta G = \Delta H - T\Delta S \tag{II.5.2}$$

L'enthalpie H correspond à l'énergie interne du système (U) tandis que l'entropie correspond au degré de désordre d'un système au niveau microscopique. Au cours d'une réaction chimique, l'entropie augmente jusqu'à un maximum qui correspond à l'état d'équilibre du système.

L'énergie libre se mesurant à pression et température constantes, un état standard a été défini. En biochimie il correspond à une réaction se produisant à une pression de 10^5 Pa, une température de 298 K, avec une concentration des solutés de 1 M et une concentration de l'eau à 55,5 M. La relation entre l'énergie libre standard ($\Delta G°$) et l'énergie libre est donnée par l'équation (II.5.3) :

$$\Delta G = \Delta G° + RTln(C) \tag{II.5.3}$$

Avec C le rapport de concentration entre les produits et les réactifs. Ainsi dans le cas d'une association entre un ligand (L) et un récepteur (R), le terme C devient : $C = \frac{[RL]}{[R].[L]}$. Lorsque qu'une réaction est

à l'équilibre, la variation de l'énergie libre est nulle ($\Delta G = 0$). Par exemple dans la figure II.5.1 pour un temps avancé l'énergie libre ne varie plus. En outre, les concentrations des ligands [L] et récepteurs [R] libres restent constantes, de même pour la concentration du complexe ligand récepteur [LR]. En réalité deux processus se compensent, la formation du complexe et sa dissociation. En définissant un taux d'association (k_{on}) par unité de temps et un taux de dissociation (k_{off}) par unité de temps, on peut relier les concentrations des récepteurs et ligands libres à la concentration de leur forme complexée (équation (II.5.4)) [67] :

$$[R][L]k_{on} = [LR]k_{off} \tag{II.5.4}$$

— k_{on} le taux (vitesse) d'association ($M^{-1}.s^{-1}$)
— k_{off} le taux de dissociation (s^{-1})

En regroupant les termes k_{off} et k_{on}, on retrouve le rapport de concentration entre les produits et les réactifs. On définit la constante de dissociation à l'équilibre (K_d) qui dépend uniquement de la réaction chimique considérée et reflète l'affinité du ligand pour le récepteur (équation (II.5.5)).

$$\frac{[R] * [L]}{[LR]} = \frac{k_{off}}{k_{on}} = K_d \tag{II.5.5}$$

Ainsi l'équation correspondant à l'énergie libre peut s'écrire sous la forme de l'équation (II.5.6) :

$$\Delta G = \Delta G° + RTln(K_d) \tag{II.5.6}$$

A l'équilibre, la variation de ΔG est nulle, ce qui donne l'équation (II.5.7) :

$$\Delta G° = -RTln(K_d) \tag{II.5.7}$$

Différentes techniques expérimentales permettent de déterminer les énergies libres d'un processus, comme par exemple la calorimétrie, la *fluorescence resonance energy transfer* ,la co-précipitation... Toutefois il peut être fastidieux et coûteux de tester expérimentalement le ΔG_b (l'énergie libre de liaison) de certains systèmes. Par exemple, évaluer expérimentalement l'affinité de liaison de deux protéines nécessite dans un premier temps d'exprimer et d'isoler les protéines d'intérêts. Cette phase d'extraction et purification peut être un processus complexe qui ajoute une difficulté supplémentaire avant un test *in vitro*. En outre, l'accès aux différents états métastables au cours du processus d'association n'est pas forcément accessible avec ces méthodes. Pour ces raisons la prédiction de l'affinité de liaison entre protéine-protéine (ou bien protéine-peptide) à partir de données structurales est un domaine de recherche très actif.

II.6 PRÉDICTION DE L'AFFINITÉ DE LIAISON

Le développement d'une méthode robuste et fiable de prédiction de l'affinité de liaison est un enjeu majeur, qui permet d'accélérer le développement de nouveaux médicaments ou d'inhibiteurs, tout en réduisant une partie du coût de la recherche et développement[68]. Les premières méthodes pour prédire l'affinité de liaisons (ΔG_b) consistaient à un emboîtement (technique de *docking*) de molécules rigides par identification géométrique[69]. Par la suite, ces méthodes se sont enrichies avec la prise en compte de la surface accessible au solvant[70 71] puis des acides aminés clés (*hotspots*) dans le processus de dimérisation. Soit en analysant les acides aminés conservés au sein d'une famille de protéines ou soit en utilisant des données expérimentales telles que celles produites par la *fluorescence resonance energy transfer* (FRET), *surface plasmon resonance*, *isothermal titration calorimetry*[72 13 73].

Dans le domaine du docking, le jeu de données utilisé pour entraîner les méthodes influe sur la précision des prédictions. Les résultats sont variables suivant le système étudié (complexe protéine-protéine, protéine-peptide[42]...). De nombreuses avancées ont été réalisées ces dernières années et les prédictions d'énergie libre ont gagné en fiabilité. PRODIGY (PROtein binDIng enerGY prediction)[41], l'un des outils de docking de référence actuellement, a montré qu'il est possible d'obtenir des valeurs d'énergie libre d'association (ΔG_b) de deux protéines en se fondant sur les contacts formés lorsqu'ils sont associés. Les auteurs ont utilisé un jeu de données composé de 81 complexes de protéines où les valeurs de δG ont été mesurées expérimentalement (les valeurs vont de -4.3 à -18.6 kcal/mol). Leur méthode de prédiction d'énergie libre de liaison de complexe protéine protéine a une corrélation de Pearson de -0.73 avec une une erreur quadratique moyenne de 1.89 kcal/mol (figure II.6.1).

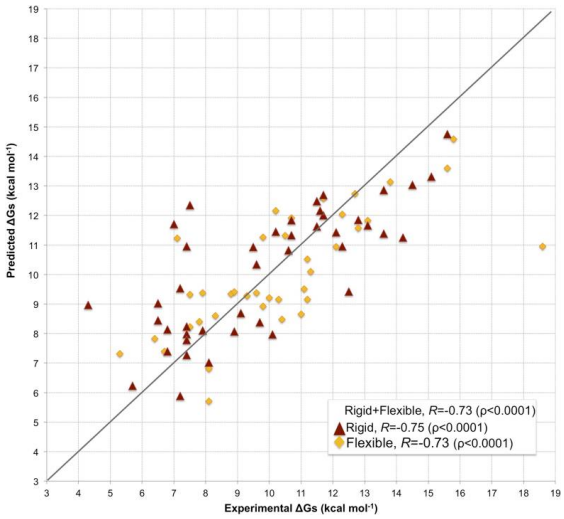


Fig. II.6.1 *Graphique issu de la réf[41] qui représente les prédictions d'énergie libre obtenues avec PRODIGY sur 81 protéines. La corrélation de Pearson est de -0.73 avec une erreur quadratique moyenne de 1.89 kcal/mol. En utilisant uniquement les protéines dont le RMSD entre la forme libre et complexée (au niveau des interfaces de dimérisation) est inférieur à 1 Å, la corrélation de Pearson est de -0.75 avec un erreur quadratique moyenne de 1.88 kcal/mol. Pour les protéines dont le RMSD entre la forme libre et complexée (au niveau des interfaces de dimérisation) est supérieur à 2 Å, la corrélation de Pearson est de -0.73 avec un erreur quadratique moyenne de 1.88 kcal/mol.*

Les méthodes de docking ont l'avantage d'être très rapides et peu coûteuses en temps de calcul. Cette vitesse d'exécution a comme contre-partie qu'il n'est pas possible de reconstruire le chemin de transition menant d'un état dissocié à un état complexé. Prédire si le système doit franchir de grandes

barrières d'énergie libre ou bien estimer les taux d'association k_{on} et de dissociation k_{off} (qui sont des paramètres importants dans l'élaboration d'inhibiteur) n'est pas possible avec ces méthodes. En effet, l'association d'un ligand à son récepteur est un processus complexe et les approximations faites ne permettent pas d'avoir une méthode suffisamment fiable et précise. L'énergie libre de liaison entre deux protéines met en jeu l'enthalpie et l'entropie du système. Omettre le solvant, ou bien représenter le récepteur et le ligand comme des corps rigides, ne permet pas de quantifier la contribution de l'entropie du système. Si l'on souhaite reconstruire le chemin de transition menant d'un état libre à un état complexé, il est nécessaire de prendre en compte certains mécanismes clés (comme par exemple le changement de conformation du ligand et du récepteur, la désolvatation de leurs interfaces d'interactions ou bien les interactions hydrophobes et polaires).

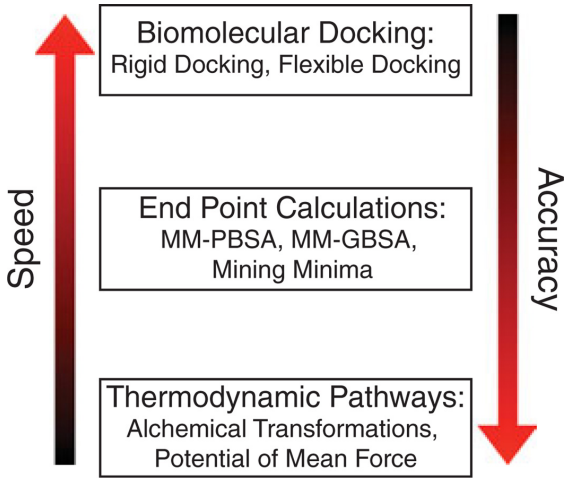


Fig. II.6.2 *Figure issue de la réf*[74]*. Différentes méthodes et approches in silico de prédiction de l'affinité de liaison existent. Les méthodes d'amarrage (docking) sont peu coûteux en temps de calcul mais font des approximations. Il n'est pas possible de déterminer si deux molécules doivent franchir des barrières d'énergie libre pour s'associer. Pour répondre à cette question, il est nécessaire de reconstruire le chemin de transition depuis l'état dissocié à l'état complexé. Des méthodes in silico peuvent échantillonner ce chemin de transition au prix d'un temps de calcul plus élevé.*

Comme expliqué précédemment, la MD simule le mouvement des atomes et peut donc être utilisée pour prédire le ΔG_b des systèmes moléculaires et le K_d (équation (II.5.7)). Le K_d peut être également interprété comme étant le rapport entre deux probabilités d'état du système (P_B et P_A). En réalisant une longue simulation, il est théoriquement possible de déterminer les différentes probabilités à l'équilibre et donc estimer ΔG_b (équation (II.6.1)) :

$$\Delta G = \Delta G° + RT\,ln(\frac{P_A}{P_B}) \tag{II.6.1}$$

Cependant l'utilisation de simulation classique de DM a des limitations pour ce genre d'étude comme expliqué dans la section suivante.

LE PROBLÈME DES ÉCHELLES DE TEMPS

Dans le cas d'un processus réversible dans lequel un système est supposé passer d'un état A à un état B, l'énergie libre peut être calculée comme le rapport des probabilités de ces états P_A/P_B. En réalisant une longue simulation de DM et en supposant qu'elle soit ergodique, la valeur moyenne de $\overline{P_A}$ ou de $\overline{P_B}$ est égale à la moyenne d'un grand nombre de mesures prises de P_A et P_B dans le temps (équation (II.6.2)).

$$\overline{P_A} = \frac{1}{M} \sum_{i=1}^{i=M} P_{A_i} \tag{II.6.2}$$

Avec M le nombre total de mesures effectuées.

Malgré la montée en puissance des processeurs, à l'heure actuelle les temps de simulations en DM pour des systèmes à plusieurs milliers d'atomes ne dépassent pas l'ordre de la μs (figure II.6.3). Avec cet ordre de grandeur, certain processus restent difficilement accessibles avec la DM classique. Par exemple le repliement des protéines ou la dissociation d'un ligand à son récepteur sont à des échelles de temps supérieures à la μs. En effet le système doit franchir des barrières d'énergie libre, plus ces barrières sont importantes, moins il est probable d'observer ces évènement dans un temps relativement court (cette probabilité correspond à $exp(\frac{-\Delta G}{kT})$).

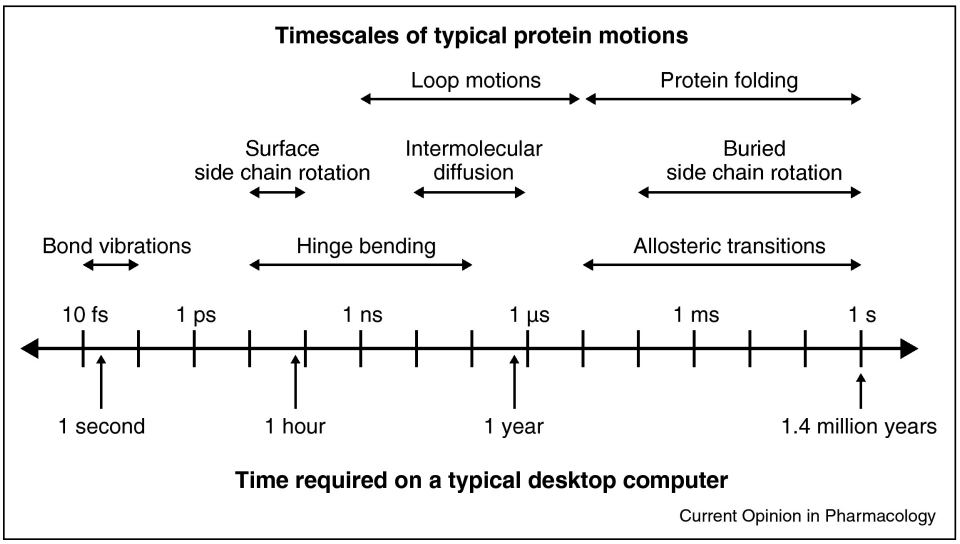


Fig. II.6.3 *Figure issue de la Ref[75]. Échelle de temps des différents évènements moléculaire et le temps de calcul estimé pour leur simulation sur un ordinateur de bureau (datant de 2010). Certains événements biologiques nécessitant un temps de simulation supérieur à la μs (repliement de protéine, mécanisme d'allostérie...) ne sont pas accessibles dans un temps raisonnable en dynamique moléculaire classique*

De ce constat découlent deux problèmes. Le premier concerne l'hypothèse d'ergodicité qui n'est plus valide et la seconde est que si le système n'atteint pas l'état B à cause d'un temps de simulation trop court, l'estimation de l'énergie libre sera alors erronée. En effet, la probabilité d'observer l'état B valant 0, l'énergie libre associé à cet évènement aura une valeur infinie :

$$\Delta G_{A-B} = -RTln(0/P_A) = \infty$$

Pour ces raisons, l'utilisation de la DM n'est pas suffisante et des méthodes d'échantillonnage accéléré ont été développées pour surmonter ces limitations. Leur principe est de perturber le système en apportant de l'énergie de manière à ce qu'il puisse s'échapper des minima d'énergie libre et explorer de manière réversible différentes configurations. Les chapitres III et V présentent des méthodes d'échantillonnage accéléré pour l'étude du paysage conformationnel et des interactions protéine protéine.

ÉCHANTILLONAGE ACCÉLÉRÉ

Comme expliqué précédemment, l'énergie libre de liaison (ΔG_b) entre un ligand et son récepteur est reliée au K_d par l'équation (II.5.7). Le K_d correspond au rapport des concentrations produits/réactifs ou bien au rapport $\frac{k_{off}}{k_{on}}$ (équation (II.5.5)).

Le taux de dissociation est très utile dans l'élaboration (ou l'optimisation) d'inhibiteur[76] d'interaction protéine protéine. En effet, le k_{off} correspond à l'inverse du temps de résidence (τ) du ligand à son récepteur et ne dépend pas de la concentration des réactifs. Ainsi dans la conception d'un inhibiteur, on va chercher à avoir un ligand qui se lie favorablement à son récepteur mais également qui occupe le plus longtemps possible le site de liaison.

Déterminer les valeurs du k_{off} et k_{on} par des méthodes *in silico* est un des enjeux majeurs de ces dernières années. Là où les méthodes de docking permettent d'obtenir des valeurs de ΔG_b fiables pour des systèmes de protéines-protéines[30 32], une estimation fiable du k_{off} n'est pas encore d'actualité. Il est ainsi nécessaire d'utiliser d'autres moyens pour estimer le taux de dissociation. Simuler le système avec de la DM classique n'est également pas possible, à l'heure actuelle. En effet, en biologie les temps de résidence de ligand à une protéine cible sont généralement compris entre la minute et l'heure. Les temps de simulation nécessaires (figure II.6.3) pour espérer observer au moins une dissociation ne sont pas possibles dans des simulations atomistiques (contrairement aux simulations en gros grains où des dissociations spontanées de protéines sont possibles).

Comme détaillé dans la revue en Ref.[77], différentes approches existent pour estimer les constantes de cinétique à l'aide de simulations de dynamique moléculaire accéléré[78 77] et de l'utilisation d'un modèle markovien. L'idée sur laquelle reposent les méthodes d'échantillonnage accéléré est de perturber le système moléculaire pour qu'il puisse sortir des minima d'énergie libre dans un temps limité. Par exemple la REMD[62] (voir chapitre III), procède à des échanges avec des répliques à plus hautes températures (et donc à plus hautes énergies). Cependant cette approche est indiquée dans le cas de systèmes qui dépendent d'un processus où les barrières d'énergie libre à franchir sont peu élevées. Dans le cas des IPP, l'ordre de grandeur des énergies pour observer une dissociation est beaucoup trop élevé pour l'utilisation de la REMD. A titre d'exemple, pour le complexe barnase barstar l'énergie d'association est de -90 kJ/mol, ce qui équivaut à une température supérieure à 10 000 K ($\frac{\Delta G_b}{k_B}$). En plus du fait d'utiliser un nombre important de répliques pour obtenir un bon recouvrement des températures (entre les replica), les champs de forces utilisées ne sont tout simplement pas adaptés à de telles températures de simulation. Le processus d'association/dissociation est en compétition avec la dénaturation des monomères, et la vaporisation de l'eau.

II.7 MÉTADYNAMIQUE

La REMD n'étant pas adaptée dans l'étude des IPP, l'approche la plus efficace consiste à perturber le système en appliquant une force artificielle (ou, de façon équivalente, un potentiel de biais), avec l'objectif d'accélérer certaines transformations sélectionnées : une méthode très répandue est la métadynamique[79]. Cette méthode, comme d'autres méthodes d'échantillonnage accéléré (umbrella sampling[80], steered molecular dynamics[81], adaptive biasing force...), applique des forces artificielles sur un petit nombre de degrés de liberté du système, appelés variables collectives (VC). Les VC sont en général des fonctions des coordonnées atomiques, telles que distances, angles, nombre de liaisons hydrogènes, contacts hydrophobes, etc.

En appliquant des forces sur des VCs, il est possible d'accélérer un processus de façon réversible (aller-retour) et de reconstruire le profil d'énergie G(q) en fonction de la VC choisie. Le choix des VC à utiliser est primordial et peut être complexe pour des systèmes de plusieurs milliers d'atomes. Une bonne VC dépendra essentiellement du système et de la problématique étudiée. La principale caractéristique est qu'il doit permettre de suivre l'évolution du processus d'intérêt (par exemple, le repliement d'une protéine ou, dans notre cas, l'interaction entre deux protéines) à partir de ses variations.

Dans le cas de la métadynamique, le biais introduit dans le système est donné via un potentiel historique-dépendant (eq (II.7.1), figure II.7.1). Ce dernier est construit comme la somme de gaussiennes déposées à un intervalle de temps donné t et centrées sur les coordonnées de la VC *(s(x))* du système.

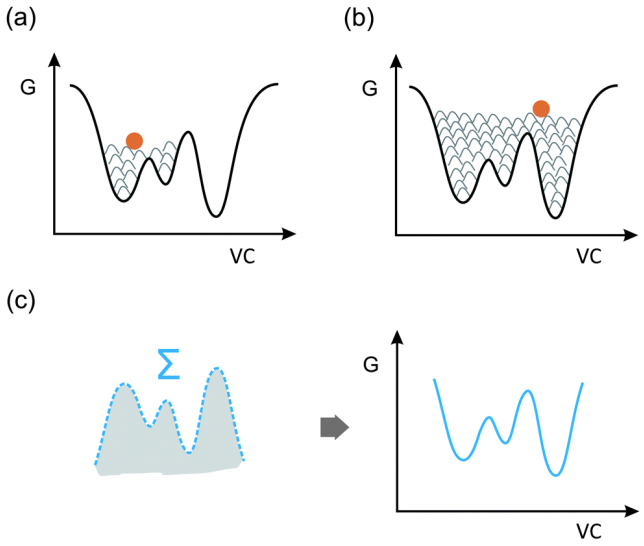


Fig. II.7.1 *Figure issue de la Ref[82] : la métadynamique est une technique d'échantillonage avancé dans laquelle le système moléculaire (représenté par le point orange) est perturbé par l'introduction d'énergie. De cette manière le système peut s'échapper des minima d'énergie libre (a) et explorer les différentes conformations disponibles dans l'espace de la variable collective (VC) utilisée (b). L'énergie introduite lors de la simulation dépend d'un potentiel de biais historique-dépendant. (c) Somme du potentiel des biais introduits en fonction de la VC. Cette dernière correspond à l'opposé de l'énergie libre (d) à une constante additive près.*

$$V_G(s(x),t) = w \sum_{t'=\tau_G,2\tau_G,\ldots} \exp\left(-\frac{(s(x)-s(t'))^2}{2\Delta s^2}\right) \tag{II.7.1}$$

— V_G : Potentiel de biais.
— w : Hauteur de la gaussienne (énergie donnée).

— Δs : Largeur de la gaussienne.
— t' : Temps de dépôt des gaussiennes.

La hauteur des gaussiennes correspond à l'énergie introduite. Plus cette énergie est grande et plus vite le système peut s'échapper des bassins d'énergie libre. La largeur des gaussiennes doit être plus petite que la largeur des bassins d'énergie libre pour induire un gradient significatif, et donc une accélération significative des processus de transition. Cependant, des gaussiennes très étroites permettent de reconstruire un paysage d'énergie libre très précis mais au prix d'un temps de simulation plus long, ce qui amène à chercher un compromis. De manière empirique, la valeur donnée à Δs correspond à environ 1/20 – 1/50 de l'intervalle de variation de la variable collective explorée.

PROFIL D'ÉNERGIE LIBRE

Lorsque la DM est suffisamment avancée, dans les situations favorables, le potentiel de biais compense tous les bassins d'énergie libre, donnant un profil approximativement plat (figure II.7.1). Le système peut explorer de manière uniforme toutes les conformations possibles dans l'espace de la variable collective dans lequel le biais est appliqué. Quand cette situation se produit (malheureusement pas toujours, cela dépend en premier du choix de la VC), on parle de convergence de la simulation : la diffusion de la VC représente un processus stationnaire et il est alors possible de reconstruire le profil d'énergie libre de la variable collective. Ce dernier correspond à l'opposé du potentiel de biais (figure II.7.1) (à moins de fluctuations de la taille des gaussiennes et à moins d'une constante additive près).

Il est donc essentiel de ne pas arrêter la simulation trop tôt sous peine d'obtenir une mauvaise estimation du profil d'énergie libre. Une comparaison du profil d'énergie dans le second et dernier tiers de la simulation est réalisée avec le module *sum_hills* de plumed[83] ou bien le module METAGUI[84] de VMD. Dans le cas où les profils sont similaires, la simulation a convergé. Dans le cas contraire, cela signifie que la somme du potentiel de biais et du profil d'énergie libre n'est pas "plate" et que certaines régions dans l'espace de la VC ne sont pas encore échantillonnées de façon réversible.

La construction du profil d'énergie libre se fait en utilisant la méthode d'analyse des histogrammes pondérés (*weigthed Weighted Histogram Analysis Method*[85]). Cet algorithme discrétise l'espace de la variable collective (q) en plusieurs sous ensemble (K). Le profil d'énergie libre est construit en calculant la différence d'énergie libre entre les différents sous ensembles et en minimisant l'erreur de manière itérative en réduisant l'erreur statistique. Cet algorithme est implémenté dans le logiciel METAGUI[84].

II.8 MÉTADYNAMIQUE AVEC BIAIS ÉCHANGÉS

Dans le cas de la dynamique de protéines, que cela soit pour le repliement ou les interactions PPI, il est difficile de choisir une ou deux VC qui capturent tous les degrés de libertés importants pendant le processus étudié. La raison est que l'évolution du système dépend de plusieurs variables réactionnelles : les angles dièdres, les interactions polaires et hydrophobes, l'eau proche de la protéine à l'interface (très importante), les modifications structurales (formation/destruction de structure secondaire), les ponts salins... De plus, le processus de repliement ou d'association peut passer à travers plusieurs étapes, avec un ordre variable : ABCDE, ACDBE, etc. Ainsi pour bien décrire le paysage d'énergie libre, il est souvent nécessaire d'utiliser plusieurs VC. Si le processus nécessite, typiquement quatre ou huit VC et que l'on en utilise seulement deux, on tombe dans l'hystérésis, c'est à dire que la simulation de la transformation à l'aller et au retour présente des profils de biais différents[79].

La métadynamique avec biais échangés (metad-BE) fait face à ce problème en utilisant plusieurs répliques du système, à la même température, chacune biaisée sur une VC différente, de façon à utiliser, dans l'ensemble des répliques, un nombre de VC > 2. Périodiquement, on échange les répliques selon le critère de Metropolis (equation II.3.2), comme pour la REMD :

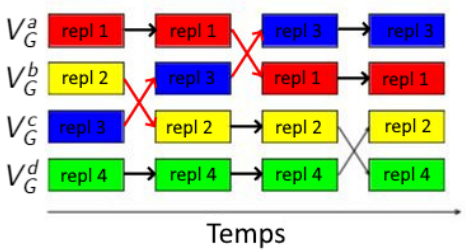


Fig. II.8.1 *Un potentiel de métadynamique est appliqué à chaque replica. Au cours de la simulation les replica vont tenter de s'échanger suivant un critère de Metropolis (flèche noire). Si le critère de Metropolis est accepté, il y a un échange de replica (flèche rouge).*

La metad-BE est une méthode d'échantillonnage accélérée qui a été utilisé pour différentes problématiques appliquées à de nombreux systèmes, tels que le repliement de protéines[86], les interactions protéines ligands[87], l'agrégation d'amyloide[88], ou les IPP[89]

II.8.1 Choix des variables collectives pour PPI

L'association d'un ligand à son récepteur est un processus complexe qui met en jeux de nombreux mécanismes et paramètres. Par exemple, les changements de conformations du ligand et du récepteur, leurs surfaces accessibles au solvant, la désolvatation de leurs interfaces d'interactions ou bien les interactions hydrophobes et polaires. Afin d'obtenir le protocole le plus généraliste possible, nous avons utilisé des variables collectives globales qui sont non système dépendant. Cette partie détaille les variables collectives utilisées dans l'établissement de ce protocole.

Contacts polaires et hydrophobes

Les liaisons hydrogènes, les ponts salins et les contacts hydrophobes contribuent à la bonne association d'un complexe ou non. L'échantillonnage des contacts permet de déterminer la configuration optimale du complexe avec l'énergie libre la plus basse.

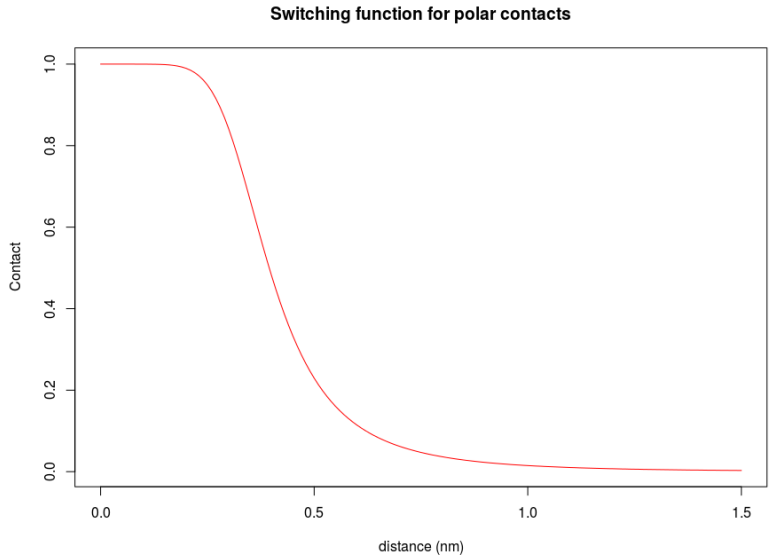


Fig. II.8.2 *Fonction de probabilité de contact, qui donne une valeur de contact (compris entre 0 et 1) en fonction de la distance entre 2 particules. Pour une distance inférieure à 0.25 nm, la valeur de contact est de 1, puis diminue à mesure que la valeur de la distance augmente (un demi contact correspond à une distance de 0.4 nm).*

La modélisation de ces interactions se fait à l'aide d'une fonction de probabilité de contact *COORDINATION* (équation (II.8.1)). Cette dernière additionne les contacts entre deux listes d'atomes définies, celle du récepteur (i) et du ligand (j).

$$\sum_i \sum_j s_{ij} \tag{II.8.1}$$

Avec S, une fonction de probabilité de contact.

Pour ne pas comptabiliser les contacts de façon binaire, une fonction de probabilité de contact est utilisée (équation (II.8.2)). Cette fonction donne une valeur de contact (compris entre 0 et 1) par rapport à la distance entre 2 particules. Pour une distance inférieure à 0.25 nm, la valeur de contact est de 1, puis diminue à mesure que la valeur de la distance augmente (un demi contact correspond à une distance de 0.4 nm) jusqu'à atteindre 0 pour une distance supérieur à 0.7 nm.

$$s_{ij} = \frac{1-(\frac{r_{ij}-d_0}{r})^n}{1-(\frac{r_{ij}-d_0}{r_0})^m} \tag{II.8.2}$$

— r_{ij} est la distance entre les atomes i et j
— d_0 paramètre de la fonction de probabilité de contact, par défaut sa valeur est 0.
— r_0 paramètre de la fonction de probabilité de contact, il correspond à *R_0*.
— n et m sont des paramètres de la fonction de contact, par défaut ils valent 6 et 0.

SOLVATATION

La contribution du solvant dans l'association de molécules est un paramètre à prendre en compte. Il a été montré que la dynamique de l'eau au voisinage des protéines était différente du reste du *bulk*. Les fluctuations géométriques (à la surface de la protéine) et les liaisons hydrogènes entre le solvant et la protéine contribuent à un ralentissement des molécules d'eau d'un facteur 3 à 5[90 91 92]. Ainsi la désolvatation des interfaces d'interactions des protéines nécessite de rompre les liaisons hydrogènes formées avec le solvant (ce qui constitue une barrière énergétique à franchir pour le système).

CHEMIN DE TRANSITION

Enfin la dernière variable collective correspond au chemin menant de l'état libre à l'état associé. Pour cette variable collective, biaiser la distance entre les deux protéines peut sembler judicieux. Toutefois, l'espace conformationnel à explorer augmente à mesure que la distance grandit. En outre, la distance ne donne pas d'information sur l'état dans lequel se trouvent les protéines (complexé ou libre), ne prévient pas les contacts non désirés et ne limite pas assez l'espace à explorer (Fig II.8.3).

Fig. II.8.3 *Pour une même distance entre deux centres de masses, les protéines peuvent être libres ou bien complexées si une des protéines effectue une rotation.*

Pour ces raisons le *path cv*[93] est utilisé. Cette variable collective, à deux dimensions, relie deux conformations à partir de structures de références (figure II.8.4) :

— La dimension S (valeur continue) indique la structure de référence la plus proche que notre système adopte.
— La dimension Z renseigne à quel point le système est distant par rapport à la structure de référence la plus proche.

Le *path cv* permet de connaître l'état dans lequel est le système grâce à la dimension S, mais également à quel point le chemin de transition optimal diverge de celui indiqué (par les structures de références) via la composante Z. De plus, en limitant l'exploration en Z il est possible de restreindre l'exploration conformationnelle. Cette variable collective a été utilisée dans l'étude d'association d'un ligand à un récepteur, tel que 3 récepteurs couplés aux protéines G et leurs ligands[94].

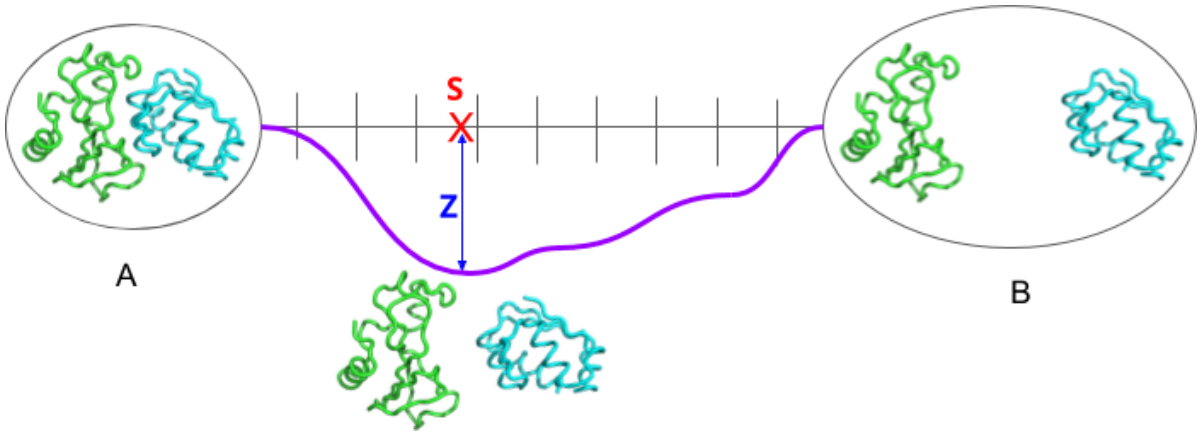


Fig. II.8.4 *Des structures de références (symbolisées par les graduations) relient l'état A et B. En autorisant le système à s'éloigner de l'axe S, le chemin de transition optimal (en violet sur le schéma) peut être exploré.*

Le calcul de S (équation (II.8.3)) et Z (équation (II.8.4)) se fait en fonction d'une métrique. Dans le cas de notre protocole, le *Mean Square Deviation* (MSD) a été choisi. Au cours d'une simulation de dynamique moléculaire le calcul du MSD de la structure d'intérêt X est réalisé par rapport aux N structures de références.

$$S(X) = \frac{\sum_{i=1}^{N} i exp(-\lambda|X - X_i|)}{\sum_{i=1}^{N} exp(-\lambda|X - X_i|)} \tag{II.8.3}$$

$$Z(X) = -\frac{1}{\lambda} log(\sum_{i=1}^{N} exp(-\lambda|X - X_i|)) \tag{II.8.4}$$

— i indice de la structure de référence
— λ paramètre
— $|X - X_i|$ distance absolue de la frame X avec la structure de référence X_i. Lorsque la distance est proche de 0, l'exponentielle vaut 1 et tendra vers 0 dans le cas contraire.

Enfin le paramètre λ est obtenu en calculant le MSD entre toutes les structures de références (eq (II.8.5)).

$$\lambda = \frac{2.3(N-1)}{\sum_{i=1}^{N-1} |X_i - X_{i+1}|} \tag{II.8.5}$$

NB : L'unité choisie dans le calcul du MSD et de lambda doit correspondre aux unités utilisées par le moteur de dynamique moléculaire (nm pour Gromacs).

II.8.2 Reconstruction d'une hyper-surface d'énergie libre à partir des répliques

En combinant les différents potentiels de métadynamique utilisés, il est possible de construire la surface (ou l'hyper surface lorsque le nombre de biais est supérieur à trois) d'énergie libre permettant ainsi de visualiser les chemins de transitions entre différents états. Chaque réplique étant biaisée sur différentes variables collectives, l'obtention de cette surface d'énergie libre se fait à l'aide de *weighted histogram analysis method*[85] implémenté dans l'extension VMD METAGUI[84]. Ce module de VMD permet de partitionner les structures en fonction des valeurs de variables collectives et de leur attribuer une valeur d'énergie libre. Ces résultats sont sauvegardés dans un fichier MICROSTATE qui est par la suite utilisé pour l'analyse. Le fichier MICROSTATE contient l'énergie libre en fonction des quatre variables collectives (a, b, c et d).

Par la suite des projections 3D, 2D et 1 D peuvent être réalisées en intégrant les dimensions non projetées. Par exemple la projection 2D en fonction de deux variables collectives se fait en sommant les énergies des deux autres variables qui ne sont pas projetées (équation (II.8.6)) :

$$F_{(a,b)} = -k_B T \log(\sum_{c,d} \exp(-\beta F_{(a,b,c,d)})) \tag{II.8.6}$$

Avec $k_B T = 2.49$ kJ/mol et $\beta = 1/k_B T$.

II.8.3 Prédiction des taux d'association et dissociation

Les processus menant aux différents états métastables sont accélérés au cours de la simulation de métadynamique avec biais échangés. De ce fait, les taux d'association (k_{on}) et de dissociation (k_{off}) ne peuvent pas être obtenus directement. Cependant, il est possible d'obtenir les valeurs de ces constantes en s'appuyant sur les résultats précédents. En effet, la métadynamique avec biais échangés fournit un paysage énergétique du système. En le partitionnant, on discrétise cet espace 4D dans lequel chaque groupe correspond à une conformation du système[95 77 96] contenue dans un hypercube 4D dans lequel chaque côté correspond à une dimension de VC. Pour cette étape, l'hypothèse importante est que le partitionnement des conformations suivant leur VC regroupe ensemble les structures qui sont proches et que le passage menant d'un groupe à un groupe adjacent se fait par des transitions rapides (figure II.8.5).

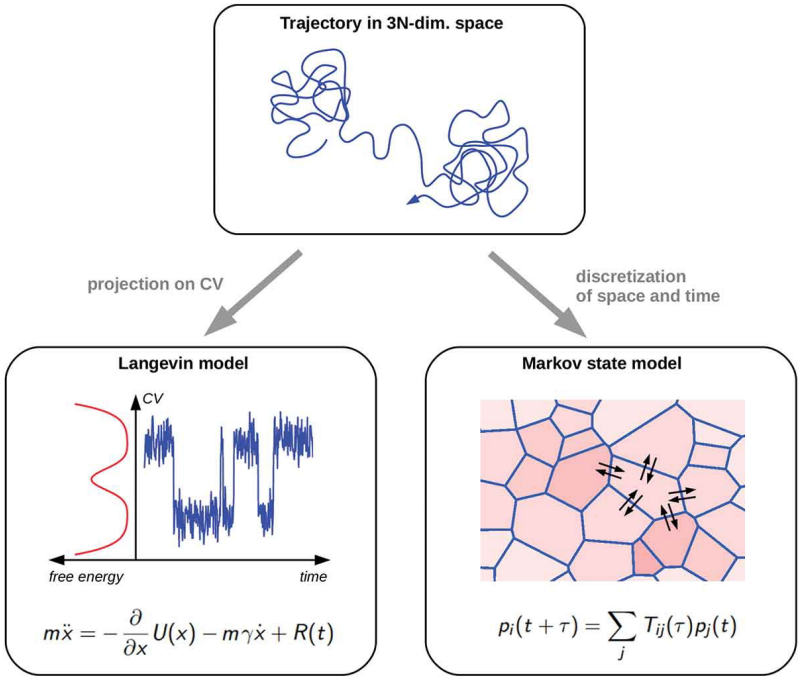


Fig. II.8.5 *Figure issue de la ref[77], où deux approches pour construire un modèle cinétique à partir d'une simulation sont présentées. Le premier aboutit a un modèle de Langevin et le second à un modèle Markovien (c'est cette dernière que nous appliquons).*

Puis à l'aide de plusieurs DM tout atome et non biaisées (avec des vitesses initiales différentes), les taux de transition entre les groupes adjacents sont calculés pour obtenir une matrice cinétique. Par exemple, le taux de transition pour passer d'un groupe i à j est estimé en calculant l'inverse du temps moyen nécessaire pour aller de i à j une première fois (nombre de fois où le système passe de i à j divisé par le temps passé dans le groupe i). A la fin de cette étape, une matrice cinétique est obtenue en appliquant l'équation de Fokker-Planck comme dans les références[97 98]. Dans le cas d'une diffusion dans un espace 1D, le passage d'un groupe i au groupe $i+1$ ou $i-1$ est défini par l'équation (II.8.7).

$$k_{(i)(i\pm1)} = \frac{D}{ds^2} \tag{II.8.7}$$

avec $k_{(i)(i\pm1)}$ le taux de transition, ds le côté d'un hypercube de VC et la matrice de diffusion D.

Dans cette équation, l'unique inconnue est la matrice de diffusion. Son estimation se fait par la méthode du maximum de vraisemblance[99], en utilisant les données issues de DM non biaisée précédem-

ment réalisées. Le principe est de sauvegarder les coordonnées des VC des DM non biaisées à des pas fixes (nommé *lag time* ou bien temps de latence). Puis des simulations de Monte Carlo cinétique (MCC) sont lancées. Contrairement à la dynamique moléculaire où tout le système est modélisé, ici seul les sauts entre les différents états de notre système sont simulés. Cette méthode peu gourmande en temps de calcul permet d'explorer des échelles de temps importantes ce qui est utile lorsque l'on souhaite étudier des phénomènes à probabilité faible en DM.

Le début de chaque simulation se fait en partant d'une coordonnée de variable collective qui a été visité par l'une des DM et en utilisant un D quelconque. À chaque étape de la simulations de MCC, le système dispose de plusieurs états accessibles à différentes probabilités. Le choix de D est fait de manière itérative afin de maximiser la vraisemblance des taux de transition observés avec les DM libres.

Le choix du *lag time* est important un intervalle trop court et le processus ne sera pas Markovien. Pour estimer le *lag time* minimum à utiliser, une manière de procéder est de calculer D à des *lag time* croissants. La matrice de diffusion D converge pour ne varier que faiblement. En analysant les temps moyens nécessaires pour passer de l'état dissocié à l'état associé, il est possible d'estimer le k_{on} qui correspond à l'inverse de ce temps. Le k_{off} quant à lui est déduit en reprenant les équations (II.5.7) et (II.5.5).

CHAPITRE III

ÉCHANTILLONAGE ACCÉLÉRÉ POUR L'ÉTUDE DU PAYSAGE CONFORMATIONNEL DES PEPTIDES CYCLIQUES

III.1 MÉTHODES D'ÉCHANTILLONNAGE DU PAYSAGE CONFORMATIONNEL DES PEPTIDES CYCLIQUES

Contrairement aux protéines, dont un grand ensemble de structure 3D est disponible (plus de 150 000 dans la PDB[29]), peu de structure de peptides cycliques libre ont été résolues. Pouvoir prédire les structures les plus probables a donc un grand intérêt. Cependant un peptide interagissant avec sa cible n'aura pas forcément la même conformation qu'à l'état libre. Dans le cas d'étude d'interaction protéine peptide cyclique, il est intéressant d'étudier le paysage conformationnel, c'est-à-dire l'ensemble des conformations que peut adopter la molécule, en plus des structures les plus probables.

A l'heure actuelle, il existe quelques outils de prédictions de structures de peptides cycliques. Nous pouvons citer le logiciel PepLook[31] qui a été le premier à prendre en compte les résidus sous forme D (toutefois leur serveur web ne fonctionne plus). PEPstrMOD[100] quant à lui, gère les acides aminés sous forme L et D accompagné de modification chimique. Cette méthode utilise une minimisation suivi d'une dynamique moléculaire courte (100 ps) pour prédire la conformation la plus probable (il y a donc une influence de la structure de départ sur la prédiction finale). Le projet Rosetta commons propose le logiciel *Simple Cyclic Peptide Prediction*[32] qui prend en charge les résidus sous forme D et N-méthylés en utilisant une approche robotique, mais a l'inconvénient d'être très coûteux en temps de calcul. EGSCyP[33] est la seule méthode d'exploration exhaustive pour les pentapeptides cycliques qui prend en charge les résidus sous forme D et N-méthylés. Enfin d'autres méthode utilisant une approche par modèle (fragment ou boucle) existent[34 35 36 37] mais ont souvent des limitations dans l'utilisation d'acide aminé (pas de forme D ou bien de résidus non naturels) ou dans le nombre de résidu qui composent les peptides.

Dans ce chapitre, nous présentons une méthode automatique pour échantillonner des peptides cycliques à l'aide de la REMD. Cette méthode d'échantillonnage accéléré a déjà été utilisée dans l'étude de peptides cycliques[101 33]. Dans l'optique de mettre en place un serveur web, nous avons évalué notre protocole. Pour ce faire, nous avons étudié le paysage conformationnel de 9 peptides cycliques (de 7 à 12 résidus) composés d'acides aminés naturels protéinogènes sous forme L et D (tableau III.1.1 et III.1.1). Ces peptides ont été créé dans le cadre d'une étude de conception de peptide cycliques stables, dans le groupe de David Baker[102].

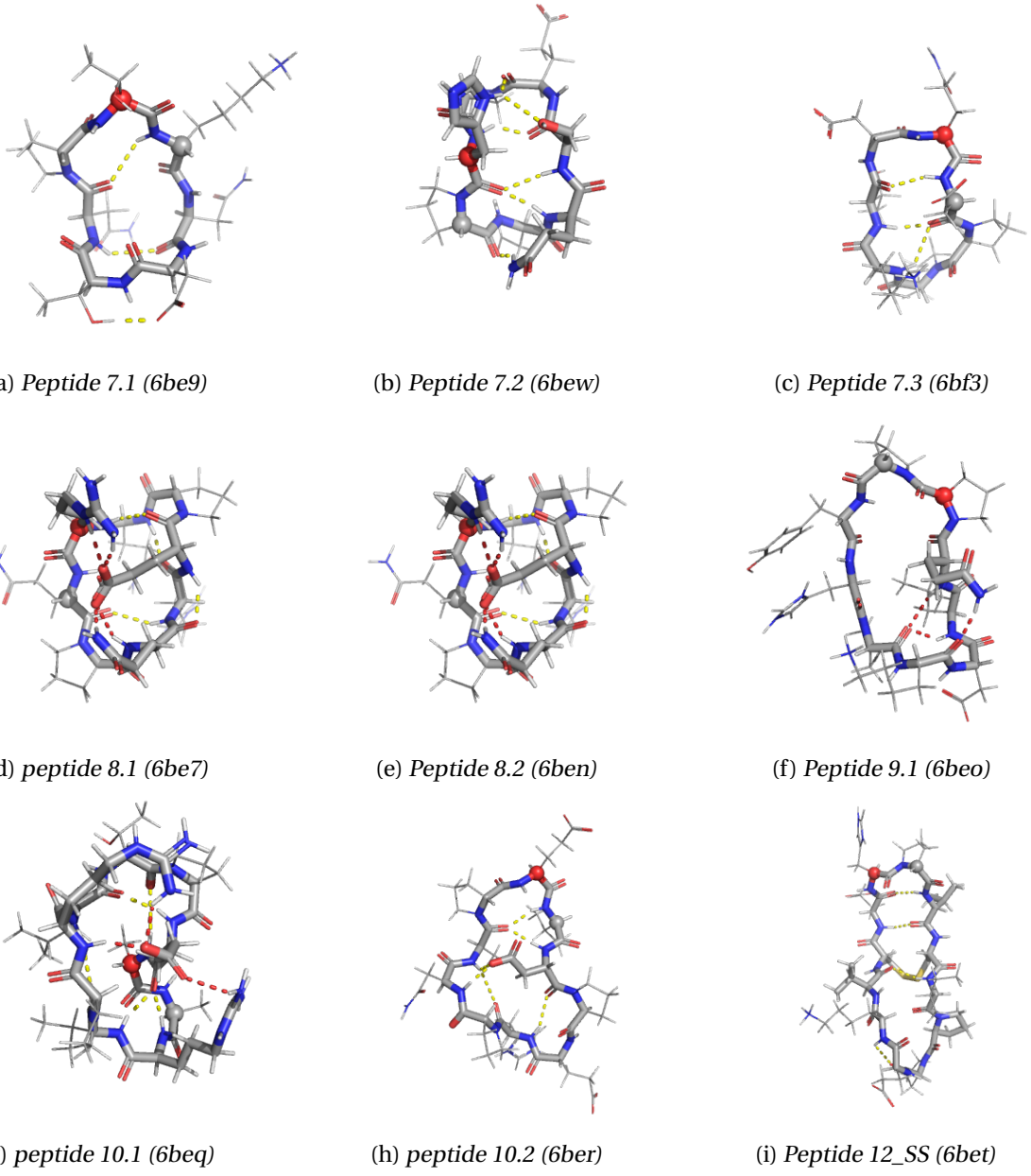
Dans cette étude, l'outil Rosetta (*simple_cycpep_prediction*)[32] est utilisé dans le cas de la prédic-

tion des structures de peptides cycliques. Cette méthode utilise une polyglicine linéaire dans lequel les chaînes latérales des résidus sont ajoutés et où la cyclisation en Nter et Cter s'effectue à l'aide d'un algorithme issu de la robotique, la cinématique inverse[35]. En plus de prédire les conformations les plus stables pour des peptides cycliques, les auteurs ont effectué une étude par mutagénèse (*in silico*) afin de déterminer les acides aminés qui stabilisent (ou déstabilisent) le plus une conformation. Enfin une résolution expérimentale (par RMN) pour 14 peptides cycliques prédits comme stables a été réalisée.

Code PDB	Peptide	séquence	atomes
6BE9	7.1	TkNDTnp	104
6BEW	7.2	hPdqsep	100
6BF3	7.3	QDPpKtd	105
6BE7	8.1	DDPTprQq	124
6BEN	8.2	rQpqRePQ	142
6BEO	9.1	pPYhPKDLq	150
6BEQ	10.1	AARvpRltPE	160
6BER	10.2	EvDPehpNap	141
6BET	12_SS	HpvCIPpEkVCe	184

TABLE III.1.1 – Code PDB, nom et séquences des peptides cycliques utilisés pour l'analyse du paysage conformationnel. Les structures sont issues de l'étude effectuée par le groupe de David Baker[102]

Pour l'ensemble de ces peptides, la cyclisation est de tête à queue au niveau du squelette peptidique (entre Nter et Cter). En plus de la cyclisation en Nter-Cter, le peptide 6BET est doublement cyclisé avec un pont disulfure formé par deux cystéines. La stabilité conformationnelle de ces peptides se fait par plusieurs mécanismes. L'alternance d'acides aminés sous forme D et L induit des contraintes stériques qui vont limiter les conformations accessibles. Par exemple une séquence L-Pro D-Pro va favoriser la présence d'un coude[103 102]. Enfin des liaisons hydrogènes au sein du squelette peptidique ou avec les chaînes latérales ainsi que les ponts salins entre résidus chargés stabilisent ces peptides. Bien que ces peptides soient très stables au niveau conformationnel, nous estimons qu'ils constituent un bon jeu de données de départ. En effet, l'ensemble des ces peptides cycliques ont été résolues sous forme libre. Ils utilisent des acides aminés sous forme L ou D et le nombre de résidus va au delà de huit acides aminés.



(a) *Peptide 7.1 (6be9)* (b) *Peptide 7.2 (6bew)* (c) *Peptide 7.3 (6bf3)*

(d) *peptide 8.1 (6be7)* (e) *Peptide 8.2 (6ben)* (f) *Peptide 9.1 (6beo)*

(g) *peptide 10.1 (6beq)* (h) *peptide 10.2 (6ber)* (i) *Peptide 12_SS (6bet)*

Fig. III.1.1 *Représentation des peptides cycliques utilisés. Les liaisons hydrogènes au niveau du squelette peptidique sont représentées en pointillés rouges. Pour des soucis de visibilité, seules les chaînes latérales formant des ponts salins ou des liaisons hydrogènes sont représentées. Leurs interactions sont symbolisées en pointillés jaunes.*

III.2 PROTOCOLE

Une partie du travail effectué durant la thèse a été d'automatiser la conception des peptides linéaires et cycliques. Afin d'avoir un outil simple d'installation, facile à mettre à jour et facilement transposable d'une machine à l'autre, nous avons choisi d'utiliser le logiciel Docker. Ce programme permet d'utiliser des applications et leurs dépendances, via des conteneurs isolés. Docker a l'avantage de fonctionner sur windows et linux et de pouvoir être exécuté sur plusieurs types de plateformes (serveur, ordinateur de bureau...) et d'être simple d'utilisation (les différentes étapes d'installation sont renseignées dans un fichier texte *Dockerfile*).

La génération des peptides est effectuée de la façon suivante : Une fois que l'utilisateur a spécifié les séquences en acides aminés souhaitées, un squelette peptidique est générée automatiquement à partir d'un modèle 3D constitué de glycines. Les chaînes latérales sont rajoutées avec SCWRL 3 [104]. La chiralité des résidus est ensuite modifiée à l'aide d'Ambertools [105] à l'exception des prolines, qui ne sont pas prises en charge par cet outil. Pour ces dernières, la structure d'une D-proline est utilisée comme patron et est superposée (via biopython [106 107]) sur le squelette peptidique des prolines dont la chiralité doit être modifié. Dans le cas de peptides linéaires, un groupement acétyle et amide sont utilisés pour neutraliser les terminaisons Nter et Cter. La génération des fichiers de topologie pour GROMACS 5.1 [108] est réalisée à l'aide d'Ambertools. Enfin en ce qui concerne les peptides possédant des résidus alanine N-méthylés. Leur conception a été faite manuellement et n'est pas encore implémentée et la paramétrisation du champ de forces des résidus non standard a été effectuée en utilisant le serveur R.E.D [109].

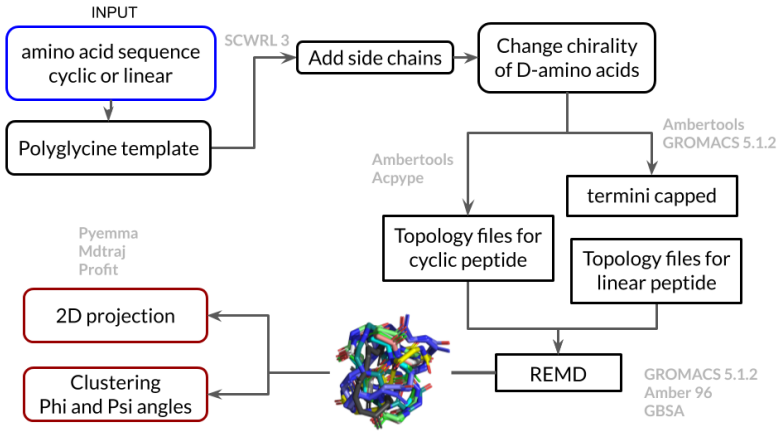


Fig. III.2.1 *REMD server flow chart*

Enfin après minimisation des structures, une vérification de la chiralité des acides aminés est effectuée afin de s'assurer qu'il n'y a pas d'erreur dans les structures avant de lancer nos simulations. A notre connaissance, il n'y a pas de bibliothèque python libre de droit qui permette de déterminer la chiralité d'un acide aminé. Nous avons donc implémenté un algorithme, ce dernier procède de la façon suivante :

- Translation du carbone α de l'acide aminé d'intérêt à l'origine d'un repère othonormé
- Rotation du carbone α de manière à aligner son hydrogène sur l'axe X et le groupement carbonyle sur l'axe Z.
- Utilisation de la règle "COORN" pour déterminer la chiralité de l'acide aminé (figure III.2.2) en comparant la composante Y des coordonnées de l'azote et du premier carbone de la chaîne latérale.

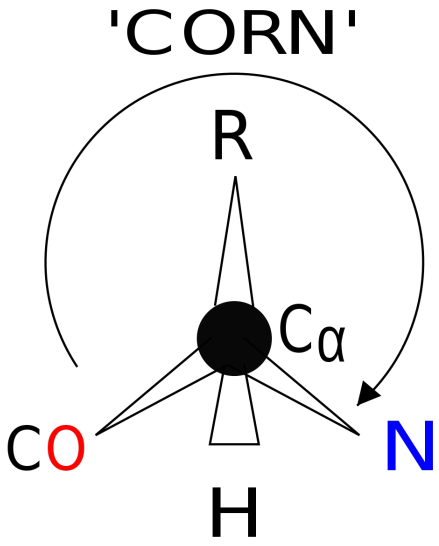


Fig. III.2.2 *Après avoir aligné l'atome d'hydrogène du carbone α sur un axe, si le passage de l'atome du carbone du groupement carboxyle ("COO"), du premier carbone de la chaîne latérale ("R") et de l'azote (N) se fait dans le sens horaire, alors l'acide aminé est sous forme L. Dans le cas contraire, il est sous forme D.*

Pour explorer le paysage conformationnel sans être trop coûteux en temps de calcul, les simulations sont faites en solvant implicite (GBSA) avec une tentative d'échange toutes les picosecondes [110]. Chaque simulation de REMD est précédée par une minimisation et équilibration du système. La minimisation se fait en solvant implicite (GBSA) et est réalisée par un algorithme utilisant le gradient conjugué [111], avec comme critère d'arrêt la convergence du système (moins de 10 kJ/mol/nm entre deux pas successifs). Si celui-ci n'est pas atteint, l'étape de minimisation s'arrête après 50 000 itérations. L'équilibration se fait en réalisant une dynamique moléculaire de 100 ps en condition NPT (conservation tout au long de la simulation du nombre d'atomes, de la pression et de la température), avec un pas d'intégration de 2 fs. Enfin une simulation de REMD est réalisée (les structures sont sauvegardées toutes les 10 ps).

Concernant l'analyse, seule la trajectoire à 300K est utilisée. Afin de ne pas être influencé par le choix des conditions initiales, les premiers 10% de la simulation sont retirés. Une fois cette étape réalisée, l'ensemble des cinq simulations sont concaténées pour créer une trajectoire de référence qui contient toutes les conformations explorées.

III.2.1 Caractérisation des structures

A la fin d'une simulation, plusieurs milliers de conformations sont obtenues. Dans le but de caractériser l'ensemble des structures, une approche possible est de les comparer par rapport à une structure de référence à l'aide du root mean square deviation (RMSD) (eq (III.2.1)).

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n}(Xi - Yi)^2}{n}} \tag{III.2.1}$$

Avec Xi, les coordonnées atomiques de l'atome i issu de la structure X et Yi, les coordonnées atomiques de l'atome i dans la structure Y.
Cette mesure permet de déterminer à quel point une conformation observée diverge par rapport à la référence. De faibles valeurs de RMSD correspondent à des conformations proches de la structure de référence, tandis que des valeurs importantes (supérieures à 1.5 Å, typiquement) indiquent des structures

divergentes.
L'utilisation du RMSD seul n'est pas suffisante pour caractériser un ensemble de conformations. En effet, l'une des limitations d'une comparaison par rapport à une structure de référence est que pour une même grande valeur de RMSD, les conformations présentes peuvent être très divergentes entre elles. Le rayon de giration est utilisé comme seconde métrique pour caractériser les structures. Cette mesure renseigne sur la compacité des molécules. Elle est calculé sur les atomes lourds du squelette peptidique (C, O et N) en utilisant la formule (III.2.2)).

$$R_{gyr} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(r_k - r_{CM})^2} \tag{III.2.2}$$

— N, le nombre d'atomes qui composent le squelette peptidique
— r_k les coordonnées de l'atome k
— r_{CM} les coordonnées moyennes des atomes du squelette peptidique. $r_{CM} = \frac{1}{N}\sum_{k=1}^{N}(r_k)$

Une projection en 2D du paysage conformationnel en fonction du RMSD (par rapport à la structure moyenne de la simulation ou bien par rapport à une structure de référence) et du rayon de giration est effectuée avec Pyemma[112]. Une énergie libre est attribuée à chaque conformation suivant sa probabilité d'apparition en appliquant l'équation (III.2.3) :

$$F = -k_B T * ln(P) \tag{III.2.3}$$

Avec k_B la constante de Boltzmann, P la probabilité d'occurence de la conformation

A partir des projections 2D, un partitionnement supervisé avec la méthode des *kmeans*[113] est effectué. Ce partitionnement étant supervisé, le nombre de groupes utilisés est répertorié dans le tableau III.2.1 :

Code PDB	Peptide	séquence	kmeans utilisés
6BE9	7.1	TkNDTnp	1
6BEW	7.2	hPdqsep	3
6BF3	7.3	QDPpKtd	2
6BE7	8.1	DDPTprQq	2
6BEN	8.2	rQpqRePQ	3
6BEO	9.1	pPYhPKDLq	4
6BEQ	10.1	AARvpRltPE	3
6BER	10.2	EvDPehpNap	3
6BET	12_SS	HpvCIPpEkVCe	2

TABLE III.2.1 – Code PDB, nom, séquence des peptides cycliques et nombre de groupes utilisés pour le partitionnement avec la méthode des *kmeans*

En plus de la méthode des *kmeans*, un partitionnement non supervisé des conformations est réalisé en utilisant le *regular space clustering*[114] sur les angles phi et psi. Cette méthode de regroupement partitionne les structures en M différents groupes dans lesquels chaque élément a une distance (par rapport au centroïde du groupe) inférieure à un seuil donnée (T). Si ce n'est pas le cas, la conformation constitue un nouveau groupe. Une fois les centres créés, une tesselation de Voronoï est réalisée.

Une question légitime que l'on pourrait se poser à propos de ces méthodes est de savoir si les conformations majoritaires trouvées sont similaires et si elles sont cohérentes d'une méthode à l'autre. Dit autrement, cela revient à se demander si les conformations appartenant à un groupe d'après le *regular*

space clustering correspondent à un même groupe d'après la méthode des *kmeans* (comme illustré dans la figure III.2.3).

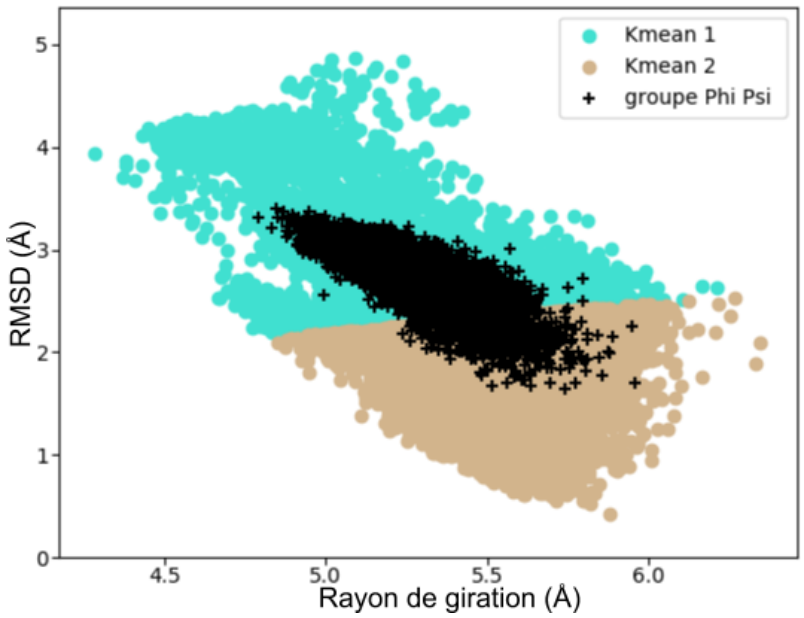


Fig. III.2.3 *Exemple de comparaison avec les méthodes de kmeans et de regular space clustering. Les conformations sont projetées suivant leur rayon de giration et de RMSD (calculé à partir du squelette peptidique 12_SS de la structure RMN). Pour des questions de lisibilité, la densité des points n'est pas représentée, mais seulement les régions obtenues avec les différentes méthodes de partitionnement. Avec la méthode des kmeans, le groupe 1 contient 71.3% des conformations tandis que le groupe 2 contient 28.7%. Le groupe trouvé avec la méthode regular space clustering contient 26.7% des conformations totales. L'analyse de sa constitution par rapport à la méthode des kmeans révèle que 92.2% et 7.8% des structures appartiennent respectivement aux groupe 1 et 2.*

III.2.2 Cohérence de la simulation

La REMD étant une exploration aléatoire, la proportion des conformations adoptées par le système évolue au cours de la simulation. Il est primordial de déterminer la convergence de nos simulations, c'est à dire que l'exploration de nouvelles conformations s'arrête (ou est très faible en prolongeant les simulations) et où les proportions des conformations majoritaires restent stable.

Pour chaque peptide cyclique, cinq simulations de REMD (d'une durée de 800 ns) sont effectuées. Ces simulations utilisent les mêmes nombres de replica et de températures (300, 313, 329, 347, 367, 391, 418 et 450 K)[115], mais diffèrent au niveau des vitesses initiales. Évaluer la cohérence des 5 simulations est intéressant sur deux aspects.

Le premier est qu'il indispensable pour pouvoir conclure sur les conformations les plus probables. Si les simulations ne sont pas cohérentes, cela signifie que les proportions des conformations sont susceptibles d'évoluer dans le temps. Ainsi il ne sera alors pas possible de classer les conformations les plus probables adoptées par nos peptides. Enfin le second aspect est qu'il permet d'obtenir un *benchmark* du temps de simulation de REMD minimum nécessaire pour espérer que les proportions des conformations majoritaires soient stables.

Pour répondre à la question de la cohérence des simulations de REMD, nous avons décidé de comparer (pour chaque système) les proportions des valeurs de RMSD dans chacune des cinq simulations de REMD avec les proportions correspondant à l'ensemble des valeurs de RMSD obtenues en combinant ces cinq simulations (afin d'obtenir une proportion de référence). Le RMSD est calculé en comparant les conformations échantillonnées au cours de la simulation et la première structure RMN du peptide correspondant (tab III.1.1).

Il est possible de quantifier les différentes conformations issues de la simulation en fonction des valeurs de RMSD, sous forme d'un histogramme de densité de probabilité du RMSD. Cette représentation graphique représente la densité observée en fonction des différents intervalles (appelés amplitude de la classe) de valeurs de RMSD. Dans notre cas chaque amplitude de classe vaut 0.1 Å). Dans cette représentation, la proportion (P) dans une amplitude de classe dx est obtenue en calculant l'aire du rectangle par l'équation (III.2.4) :

$$P = h * dx \tag{III.2.4}$$

— h : la hauteur de l'histogramme (densité de fréquence)
— dx : intervalle

L'inconvénient de l'histogramme est qu'il est une construction non-continue. Cependant avec un échantillon suffisamment grand et une amplitude de classe dx suffisamment petite, l'histogramme peut être vu comme la courbe de densité de probabilité des valeurs de RMSD pour chaque peptide (fig III.2.4).

Ce profil, dont l'aire totale vaut 1, représente la probabilité d'observer une conformation donnée avec une certaine valeur de RMSD. Pour obtenir la probabilité qu'un peptide adopte un intervalle de valeurs de RMSD, il est nécessaire d'intégrer sur la région dont on souhaite estimer la probabilité. Le passage de l'histogramme au profil de densité de probabilité (PDP) se fait par la méthode du noyau[116]. Cette méthode non paramétrique permet de construire le profil de densité de probabilité à partir des valeurs de RMSD obtenues. Dans le cas de notre étude nous avons utilisé la fonction gaussian_kde de la bibliothèque python scipy pour réaliser automatiquement les profils de densités.

Dans le cas de notre étude, ce sont les PDP qui sont utilisés pour évaluer la cohérence des 5 simulations de chaque système avec un PDP de référence (obtenu en combinant les valeurs de RMSD des 5 simulations). Cette courbe peut-être assimilée à la densité moyenne des cinq simulations de REMD (elle équivaut à une simulation de 5 * 800 ns). Par la suite cette distribution de référence est comparée aux PDP des valeurs de RMSD de chaque simulation, à différents intervalles de temps. Pour quantifier

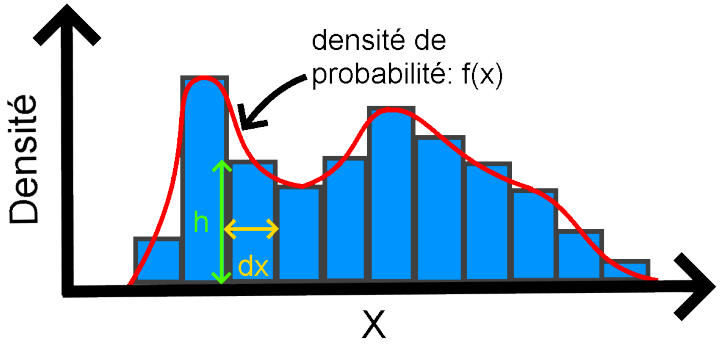


Fig. III.2.4 *Représentation d'un histogramme de densité pour une variable aléatoire continue X. Le calcul d'une proportion d'une classe se fait en intégrant l'aire du rectangle. Le profil de densité quant à lui est une fonction dont l'aire sous la courbe permet de calculer la probabilité que X soit compris dans un intervalle. Lorsque le jeu de données est important, le profil de densité peut être approximé en utilisant un histogramme de densité avec des amplitudes de classe petite.*

la différence (mais aussi la similitude) entre deux distributions de probabilités P et Q, la divergence de Kullback–Leibler de P par rapport à Q (*KL*) peut être utilisée[117] (eq (III.2.5)).

$$D_{\mathrm{KL}}(P||Q) = \int_{-\infty}^{\infty} p(x) log(\frac{p(x)}{q(x)}) dx \tag{III.2.5}$$

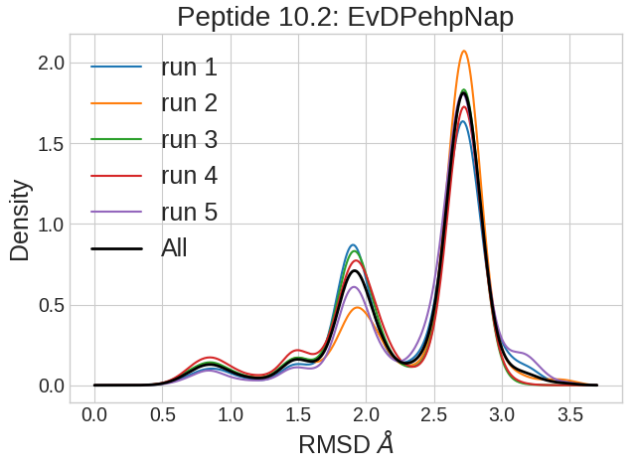
Où p et q sont les densités de probabilité de P et Q

Dans notre cas, P représente la densité de probabilité de référence (construite en utilisant les valeurs de RMSD des cinq simulations de REMD) et Q à la densité de probabilité d'une seule simulation. Pour deux distributions identiques le *KL* vaudra 0 et la valeur sera de plus en plus grande à mesure que les distributions divergent. Bien que très utile, la *KL* a deux principaux inconvénients. Sa valeur maximale n'est pas bornée et enfin la divergence de *KL* n'est pas symétrique. Ainsi, la valeur de la divergence entre P et Q diffère de celle entre Q et P. Pour ces raisons, la divergence de Jensen–Shannon (*JSD*) est utilisée[118] (eq (III.2.6)). Cette dernière dérive de la divergence de *KL* et a pour avantage d'être symétrique et d'être bornée à 1 en valeur maximale.

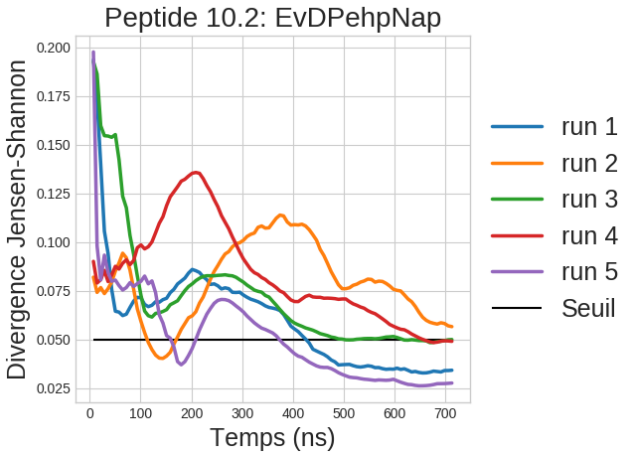
$$JSD(P||Q) = \frac{1}{2} D_{\mathrm{KL}}(P||M) + \frac{1}{2} D_{\mathrm{KL}}(Q||M) \tag{III.2.6}$$

Avec $M = \frac{1}{2}(P+Q)$

La valeur seuil de JSD, à partir de laquelle nous considérons que deux distributions diffèrent, est choisie à 0,05. Cette valeur a été prise pour être suffisamment restrictive pour discriminer des situations similaires à celle rencontrée avec le peptide 10.2. En effet dans ce cas, l'espace conformationnel exploré est similaire dans les cinq simulations, mais les proportions des conformations majoritaires diffèrent (figure III.2.5a et III.2.5b). Ainsi avec ce critère nous pouvons déterminer si les simulations explorent les mêmes conformations dans les mêmes proportions (avec des profils de densité similaires) ou bien sont bloquées dans un minimum d'énergie libre. Dans ce cas il est alors nécessaire de prolonger les simulations.



(a) *Représentation du profil de densité des cinq simulations et du profil de densité de référence (All) construit en utilisant les valeurs de RMSD des cinq simulations. Dans cet exemple, l'espace conformationnel exploré est similaire dans les cinq simulations, mais les proportions des valeurs de RMSD majoritaire diffèrent, notamment pour la simulation 2.*



(b) *La divergence de Jensen–Shannon permet de quantifier la divergence du profil de densité d'une simulation par rapport au profil de densité de référence (obtenu en utilisant les valeurs de RMSD des cinq simulations). Au delà de la valeur seuil de 0.05, les deux profils sont considérés comme divergentes, comme c'est le cas avec la simulation 2.*

Fig. III.2.5 *Évaluation de la convergence*

Après avoir retiré les premiers 10% de la simulation, la durée totale de celle-ci est de 720 ns. La divergence JSD est calculée tous les les 7.2 ns (pas correspondant à 1 centième du temps de simulation). De cette manière, nous pouvons déterminer le temps à partir duquel une simulation est cohérente (en terme d'espace conformationnel et de proportion des conformations explorées) avec la distribution de référence.

La divergence *JSD* quantifie la différence entre deux distributions. Calculer la divergence *JSD* de la distribution de référence à différents intervalles de temps revient à déterminer le temps nécessaire pour que la distribution de référence soit similaire avec elle-même. Or la similarité du profil de densité de RMSD de référence sera toujours similaire à elle-même à mesure que l'on utilise un de temps de simulation proche du temps total.

En outre, les simulations de REMD sont des explorations aléatoires, les effectifs des conformations peuvent changer d'une simulation à l'autre. Au final dans notre étude, la divergence *JSD* est utilisée pour déterminer le temps nécessaire pour qu'une simulation soit similaire à l'ensemble des cinq simulations. Pour estimer le temps moyen pour lequel une simulation converge, il suffit de moyenner ces temps obtenus pour chaque peptide. Nous avons fait le choix de calculer ce temps moyen uniquement pour les peptides cycliques pour lesquels les cinq simulations de REMD ont une valeur de divergence *JSD* inférieure à 0,05 à la fin de la simulation.

En effet si une ou plusieurs simulations ont une divergence *JSD* supérieure à 0.05, cela signifie que les proportions des conformations ou bien l'espace conformationnel exploré ne sont pas similaires d'une simulation à une autre. Dans ce cas, le temps moyen pour arriver à convergence sera considéré comme étant supérieur à 720 ns. En outre, il ne sera pas possible de classer les conformations majoritaires puisque que les proportions observées en combinant les cinq simulations de REMD sont susceptibles d'évoluer en prolongeant la simulation comme l'illustre la figure III.2.6.

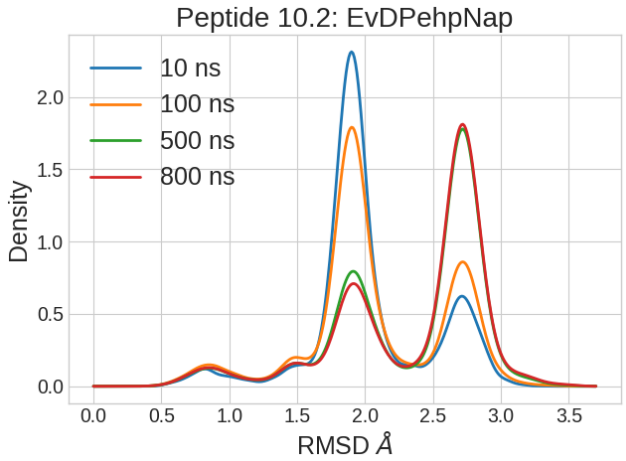


Fig. III.2.6 *Représentation du profil de densité du RMSD de référence (construit en utilisant les valeurs de RMSD des cinq simulations) pour différents temps de simulations. Dans le cas où les cinq simulations de REMD ne sont pas cohérentes entre elles, l'espace conformationnel et les proportions des conformations explorées ne sont pas similaires. Il est important de vérifier la bonne cohérence des simulations pour pouvoir classer les conformations les plus probables. Par exemple, le profil de densité de référence change à mesure que les simulations de REMD sont prolongées. Pour 10 et 100 ns les conformations majoritaires ont un RMSD moyen de 1.7 Å. Après 500 ns le RMSD des conformations majoritaire est de 2.7 Å.*

III.2.3 Diffusion des températures

L'échantillonnage des conformations en REMD dépend notamment du nombre de répliques utilisé mais aussi de l'intervalle de temps utilisé pour effectuer les tentatives d'échanges entre les replica [119 120]. Il n'y a pas de consensus concernant le choix du temps entre chaque tentative d'échange. L'intervalle utilisé peut aller de 0.01 à 100 ps [62 121] suivant le système mais surtout du processus étudié. Dans le cas de l'étude de repliement d'un peptide linéaire de 21 acides aminés fait par *Zhang et al.*, il a été déterminé qu'une tentative d'échange toutes les 1 ps donne des résultats en adéquation avec les résultats expérimentaux. Cet intervalle de temps entre les échanges est également utilisé dans l'étude de pentapeptides cycliques RGD [122]. Pour ces raisons le temps entre chaque tentative d'échange est fixé à 1 ps pour toutes les simulations de REMD.

Ce paramètre étant le même d'un système à l'autre, il est nécessaire d'évaluer la bonne diffusion des replica dans l'espace des températures. C'est-à-dire que le temps de séjour d'une réplique est à peu près équivalent pour n'importe quelle température. Le taux d'échange moyen observé au cours de la simulation renseigne uniquement sur les échanges de replica deux à deux, mais ne donne aucune indication sur la bonne diffusion dans l'espace des températures.

Un graphique illustrant la proportion du temps de séjour de chaque replique (à l'exception du peptide 7.1 dont les fichiers log ont été supprimés) pour les différentes températures est tracé pour chacune des simulations (figure III.2.7).

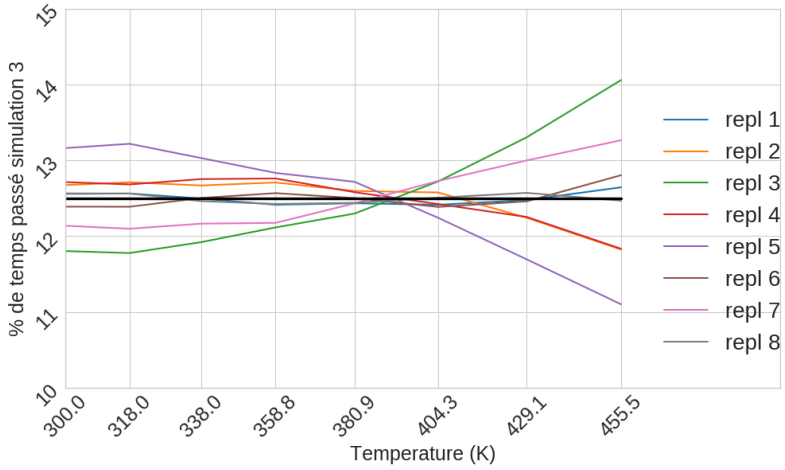
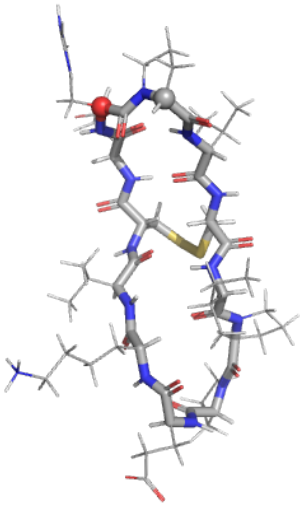


Fig. III.2.7 *Graphique qui représente le temps de séjour de chaque réplique pour les différentes température pour une simulation de REMD. Durant la simulation de REMD, les replica doivent avoir un temps de visite pour les différentes températures équivalentes. Dans le cas d'une mauvaise diffusion, la simulation à 300K aura du mal à s'échapper de minima d'énergie libre.*

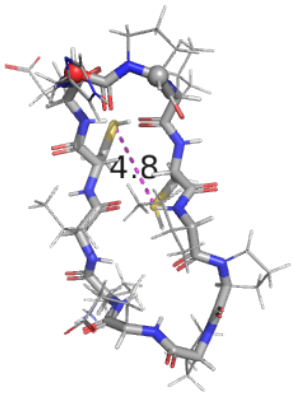
III.2.4 ÉCHANTILLONNAGE

Pour chaque peptide, cinq simulations de REMD de 800 ns ont été faites avec huit replica. L'analyse des structures s'est faite à partir de la trajectoire à 300K. Pour chaque peptide cyclique, les cartes d'énergie libre et le partitionnement avec la méthode *regular space* ont été réalisées en utilisant les cinq simulations de REMD à 300 K.

Enfin, concernant le peptide cyclique HpvCIPpEkVCe, ce dernier possède un pont disulfure (figure III.2.8a) Afin de ne pas être bloqué dans une configuration tout au long de la simulation, cette liaison n'a pas été modélisée (dans le fichier de topologie correspondant à notre molécule). Avant d'effectuer les projections 2D et le partitionnement des conformations sur les angles phi et psi, un premier filtrage des structures est réalisé pour ce système (figure III.2.8b). Seules les conformations pour lesquelles la distance entre les deux atomes de soufres des résidus cystéines est inférieure ou égale à 5 Å sont conservées.



(a) *Le peptide 12_SS est doublement cyclisé. En plus de la cyslisation Nter-Cter, le peptide possède deux résidus cystéines qui forment un pont disulfure.*



(b) *Dans notre simulation, le pont disulfure du peptide 12_SS n'est pas modélisé. Les cystéines ne forment pas de liaison covalente entre elles. Pour l'analyse des structures, seules les conformations où les atomes de soufres sont distants au maximum de 5 Å sont utilisées.*

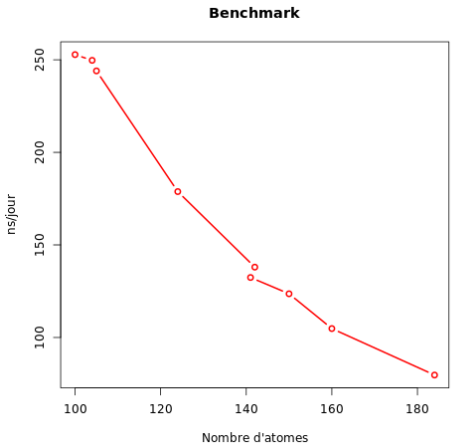
Fig. III.2.8 *Représentation du squelette peptidique et des résidus cystéines du peptide cyclique 12_SS HpvCIPpEkVCe*

Dans le cas de notre projet, les simulations de REMD sont utilisées pour générer diverses conformations en vue de réaliser un criblage virtuel par une méthode de docking entre un peptide et une protéine cible. Dans l'optique de mettre en place un serveur Web, nous souhaitons évaluer notre protocole à partir d'un ensemble de structures de peptides cycliques sous formes libres résolues expérimentalement [102] (tableau III.1.1). Nous souhaitons déterminer le temps nécessaire pour obtenir une conformation avec un RMSD inférieur ou égal à 1 Å pour différentes tailles de peptides cycliques. Enfin nous souhaitons évaluer si les structures les plus probables obtenues avec notre protocole de REMD correspondent aux structures résolues expérimentalement (au niveau du squelette peptidique). Ceci afin d'établir un benchmark en vue de déterminer les ressources minimales et les limites de notre protocole.

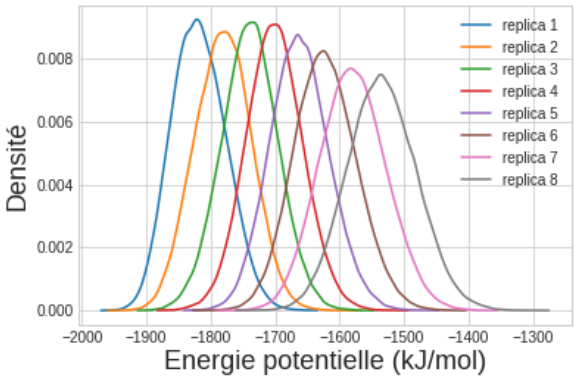
III.3 RÉSULTATS

III.3.1 ÉCHANTILLONNAGE ET CONVERGENCE

Les temps de production théorique, pour 24 heures de simulations, sont indiqués dans le tableau III.3.1 et sur la figure III.3.1a. Les simulations ont été effectuées sur le cluster Occigen, sur 8 cœurs d'un processeur Haswell cadencé à 2.6 GHz. Le temps de production diminue à mesure que le nombre d'atomes du système augmente. Il passe de 252 ns par jour pour 100 atomes à 80 ns par jour pour 184 atomes. Au delà de 160 atomes la production théorique passe sous les 100 ns quotidien. Enfin comme le montre le profil de densité des énergies potentielles (figure III.3.1b), entre répliques voisines, le recouvrement des profil d'énergie avec 8 répliques est important, ce qui permet d'avoir un bon taux d'échange. A titre d'exemple pour le peptide 10.1 (constitué de 160 atomes), à la fin de la simulation le taux d'échange entre les replica est de l'ordre de 0.5 (0.55, 0.51, 0.49, 0.51, 0.52, 0.54 et 0.56).



(a) *Temps de production théorique par jour pour une simulation de REMD avec 8 replica.*



(b) *Densité de probabilité des énergies pour chaque replica pour le peptide 10.1*

Fig. III.3.1 *Evaluation du protocole de REMD.*

N°	séquence	atomes	ns par jour	Temps (ns) pour RMSD <=1 Å					Temps moyen
				run 1	run 2	run 3	run 4	run 5	
7.1	TkNDTnp	104	249.7	0.04	0.03	0.03	0.07	0.09	0.05+/-0.02
7.2	hPdqsep	100	252.8	0.0	0.0	0.0	0.0	0.0	0.0+/-0.0
7.3	QDPpKtd	105	244.0	0.01	0.01	0.01	0.03	0.01	0.01+/-0.01
8.1	DDPTprQq	124	138.0	0.76	0.29	0.02	0.15	0.02	0.3+/-0.3
8.2	rQpqRePQ	142	138.0	15.9	5.0	3.9	11.4	0.73	7.4+/-6.0
9.1	pPYhPKDLq	150	123.6	20.1	2.0	2.9	1.1	78.3	20.88+/-20
10.1	AARvpRltPE	160	104.8	373.7	145.0	45.3	116.4	218.2	179.7+/-111
10.2	EvDPehpNap	141	139.0	74.9	10.5	23.3	2.8	8.8	30.0+/-30
12_SS	HpvCIPpEkVCe	184	80.7	220.8	162.9	178.1	166.1	248.1	195.20+/-34

TABLE III.3.1 – Durées de simulations de REMD produites quotidiennement pour chacun des peptides de cette étude et temps (ns) pour obtenir une conformation avec un RMSD (calculé par rapport au squelette peptidique de la structure de référence) inférieur ou égale à 1Å pour chaque simulation.

Pour chaque conformation générée, le RMSD du squelette peptidique a été calculé par rapport à la première structure RMN présente dans les fichiers PDB. Le RMSD minimum obtenu durant la simula-

tion a été extrait, afin de déterminer le temps nécessaire pour obtenir une conformation avec un RMSD inférieure où égal à 1 Å. L'analyse du RMSD minimum révèle que pour les peptides cycliques à 7 résidus, l'obtention d'une conformation similaire à la structure de référence est immédiate après l'équilibration (tab III.3.2 et figure III.3.7). En effet quelques étapes de dynamiques moléculaires suffisent pour obtenir une structure avec un RMSD inférieur à 1 Å. Pour les peptides cycliques 8.1 et 8.2, le temps moyen est respectivement 0.3 ± 0.3 et 7.4 ± 6.0 ns.

A mesure que la séquence des peptides augmente, le temps moyen pour échantillonner une conformation similaire à la structure de référence augmente, avec des disparités entre les peptides à 10 résidus. Ainsi pour les systèmes 9.1, 10.1 et 10.2, les temps respectifs sont 20.8±29.6, 179.7 ± 111 et 30.0±30 ns. Cette différence d'échantillonage entre les simulations pourrait être due à un mauvais échange entre les replica (figure III.3.2 et III.3.3). Si les replica à haute température ne s'échangent pas avec les basses températures, la replique à 300 K peut être bloquée dans un minimum d'énergie libre. Afin de vérifier cette hypothèse, l'évolution des valeurs des RMSD minimum des 8 replica pour chaque simulation a été tracée pour les systèmes 10.1 et 10.2. L'évolution des valeurs des RMSD minimum des 8 replica pour chaque simulation révèle que le RMSD minimum évolue de la même façon chez les 8 replica. Ainsi nous pouvons écarter l'hypothèse d'un problème d'échange entre les replica pour expliquer cette différence de temps pour obtenir une conformation avec un RMSD de 1 Å.

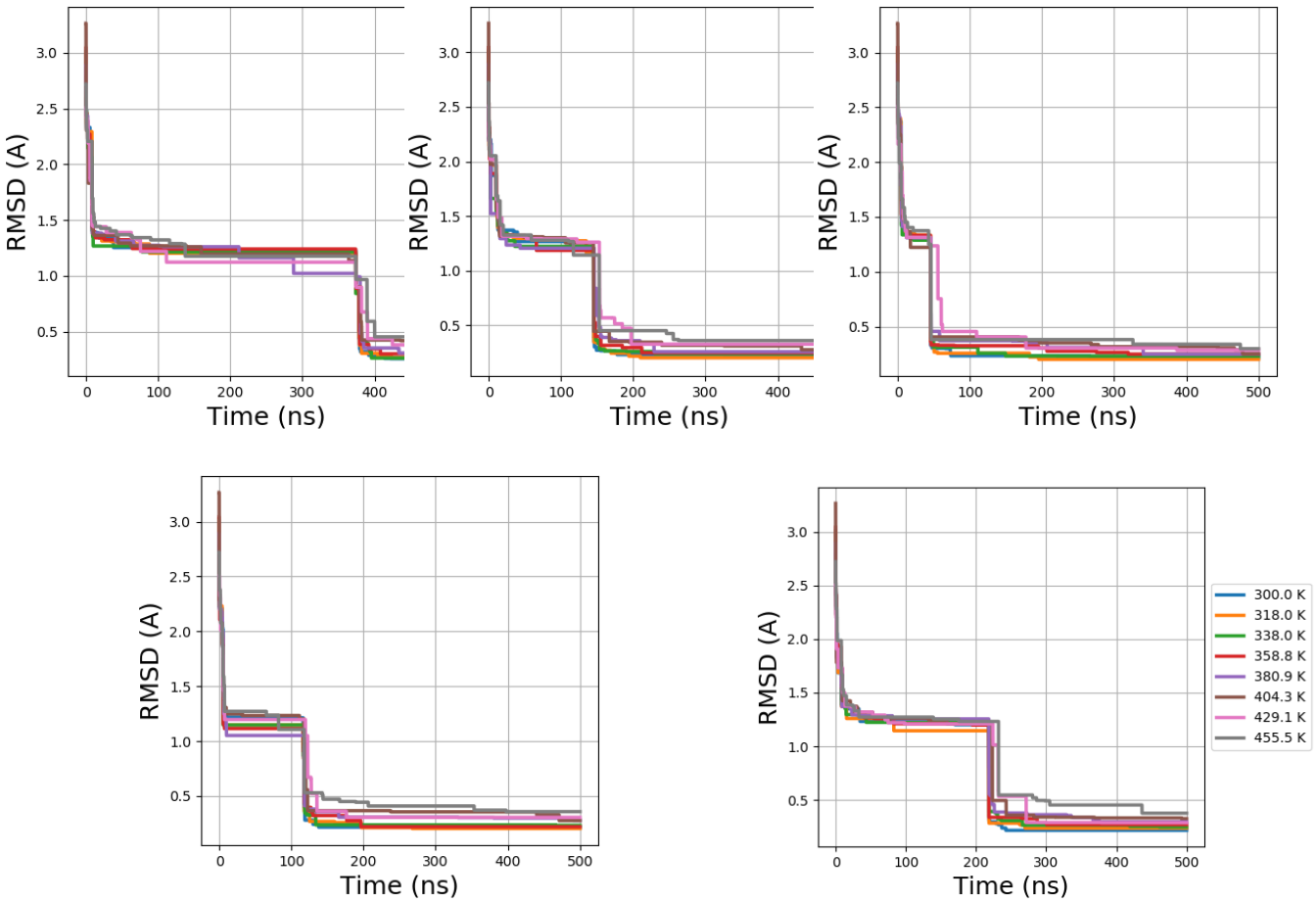


Fig. III.3.2 *Evolution du RMSD minimum des 8 replica pour chaque simulation pour le peptide 10.1. Le RMSD minimum évolue de la même façon pour les 8 replica pour atteindre les 1 Å, ce qui écarte l'hypothèse d'un échange trop faible pour expliquer la différence du temps pour obtenir un squelette peptidique avec un RMSD de 1 Å (calculé par rapport à la structure RMN expérimentale de code PDB 6beq)*

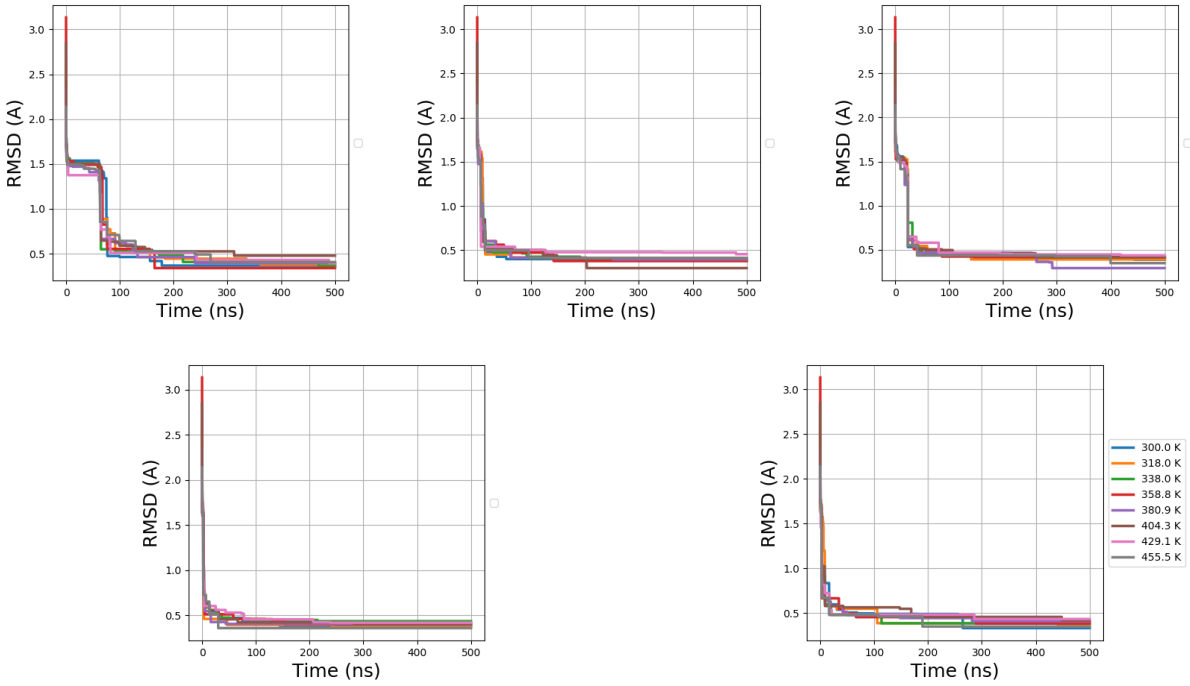


Fig. III.3.3 *Evolution du RMSD minimum des 8 replica pour chaque simulation pour le peptide 10.2. Le RMSD minimum évolue de la même façon pour les 8 replica pour atteindre les 1 Å, ce qui écarte l'hypothèse d'un échange trop faible pour expliquer la différence du temps pour obtenir un squelette peptidique avec un RMSD de 1 Å (calculé par rapport à la structure RMN 6ber)*

Nous pouvons supposer que cette augmentation du temps moyen s'explique par le fait que le squelette peptidique est moins contraint. Le nombre de conformations accessibles pour ce dernier est plus important. Les projections 2D des conformations (figure III.3.14 à III.3.30) révèlent que plusieurs conformations correspondent à des minima d'énergie. Ainsi il est probable que pour ces systèmes, les peptides soient bloqués dans des configurations correspondant à ces minima d'énergie libre.

Le peptide cyclique 12_SS est doublement cyclisé, avec un pont disulfure en plus de la liaison peptidique Nter-Cter. Pour rappel, cette liaison soufre soufre n'est pas prise en compte dans la topologie durant la simulation afin que le système puisse adopter des configurations qui nécessiteraient de casser cette liaison. Pour l'analyse du RMSD, nous avons pris l'ensemble de la trajectoire afin de déterminer le temps nécessaire pour obtenir un squellette peptidique proche de la structure RMN malgré l'abscence de cette liaison covalente supplémentaire. Le temps moyen pour obtenir une conformation à 1Å est de 195.2±33.6 ns. Par la suite, les structures pour lesquelles la distance entre les atomes de soufres des résidus cystéines est inférieure ou égale à 5 Å ont été gardées (ce qui représente 24.6±6.5% des conformations) pour réaliser les projections 2D et le partitionnement sur les angles ϕ et ψ.

Nous avons donc une estimation du temps moyen pour obtenir un squelette peptidique avec un RMSD inférieur ou égal à 1 Å de la structure de référence. Cependant, ce temps ne renseigne pas sur les caractéristiques des conformations les plus échantillonnées par nos simulations. La REMD étant une marche aléatoire, il est important de déterminer si la durée des simulations est suffisante pour échantillonner l'espace conformationnel et s'assurer que les proportions des différentes conformations ne vont pas évoluer par la suite.

Nous considérons un système comme "convergé" lorsque les cinq profils de densité de RMSD des cinq simulations de REMD ont leur divergence JSD inférieure à 0.05 par rapport au profil de référence (qui est lui calculé en prenant l'ensemble des valeurs de RMSD des cinq simulations). La valeur JSD a été calculée à un intervalle de temps correspondant à 1 centième de la simulation (c'est-à-dire tous les 7.2 ns). Comme indiqué dans le tableau III.3.2, les cinq peptides 7.1, 7.2, 7.3, 8.1 et 8.2 convergent.

Il faut moins de 50 ns pour obtenir un profil similaire au profil de référence (qui est le profil de densité construit à partir de l'ensemble des valeurs de RMSD des cinq simulations), pour les systèmes 7.1, 7.2, 7.3 et 8.1. Pour le peptide 8.2 les 5 simulations ont toutes convergées après 208.8 ns. Pour les peptides avec plus de 8 résidus, plusieurs simulations ont une divergence JSD supérieure à 0.05 à la fin de la dynamique moléculaire (figure III.3.10). Pour ces systèmes, il est donc nécessaire de simuler plus de 800 ns pour obtenir une bonne cohérence entre les cinq simulations (en terme de proportion des effectifs les plus peuplés et d'espace conformationnel exploré). Cependant, nous pouvons observer une disparité entre ces quatre peptides cycliques.

En effet pour le peptide 10.2 (séquence EvDPehpNap), une seule simulation possède une divergence JSD supérieur à 0.05. Comme l'illustre le graphique (figure III.3.10), cette dernière est proche du seuil de 0.05. Nous pouvons supposer que prolonger la simulation de 100 ns supplémentaire permettrait d'avoir l'ensemble des cinq simulation avec un seuil JSD infériur à 0.05.

Pour le peptide 9.1, les conformations les plus peuplées ne correspondent pas d'une simulation à l'autre. Par exemple la simulation 5 a un profil différent de la simulation 4 (figure III.3.4). Il est donc primordial de prolonger les simulations si l'on souhaite classer les conformations les plus probables.

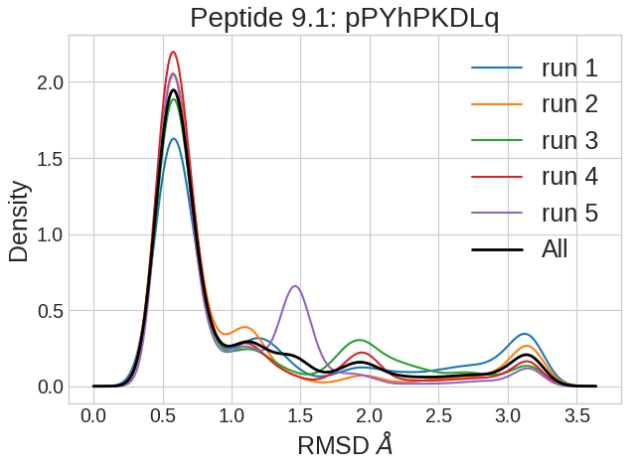


Fig. III.3.4 *Pour le peptide 9.1, les maxima des profils de RMSD ne sont pas situés dans les mêmes valeurs de RMSD pour les cinq simulations de REMD. Les proportions des conformations les plus peuplées ne sont pas équivalentes, il est donc nécessaire de prolonger les simulations*

Enfin pour les peptides 10.1 et 12_SS, les maxima des profils de densité de RMSD correspondent aux mêmes valeurs de RMSD (figure III.3.26, III.3.28 et III.3.30). Cependant les proportions d'une simula-

tion à l'autre ne sont pas équivalentes. Là encore il est nécessaire de prolonger ces simulations. Pour les peptides cycliques 9.1, 10.1, 10.2 et 12_SS, à défaut d'avoir des simulations cohérentes entre elles, nous pouvons supposer que les conformations les plus probables se situent dans les maxima du profil de densité du RMSD de référence (qui correspond à l'ensemble des valeurs de RMSD des 5 simulations)

N°	séquence	Temps de convergence (JSD <0.05)					Temps moyen	Nombre de simulations JSD <= 0.05
		run 1	run 2	run 3	run 4	run 5		
7.1	TkNDTnp	21.6	7.2	7.2	7.2	7.2	10.1+/-5.8	5
7.2	hPdqsep	28.8	7.2	7.2	50.4	21.6	23.0+/-16.0	5
7.3	QDPpKtd	7.2	7.2	7.2	14.4	7.2	8.6+/-2.9	5
8.1	DDPTprQq	7.2	7.2	7.2	7.2	7.2	7.2+/-0.0	5
8.2	rQpqRePQ	72.0	208.8	28.8	79.2	21.6	82.1+/-67.3	5
9.1	pPYhPKDLq	>720	>720	604.8	>720	>720	non convergé	1
10.1	AARvpRltPE	>720	>720	>720	>720	640.9	non convergé	1
10.2	EvDPehpNap	432.0	>720	619.3	662.5	374.4	non convergé	4
12_SS	HpvCIPpEkVCe	560.5	442.8	>720	>720	>720	non convergé	2

TABLE III.3.2 – Temps de convergence (en ns) pour obtenir un profil de densité de RMSD similaire au profil de référence (ce dernier correspond à la moyenne des profils de densité des cinq simulations de REMD à 300K). La valeur de temps indique pour une simulation le nombre de ns à partir duquel la divergence JSD est inférieure à 0.05. > 720 ns indique qu'après 720 ns, le PDP de la simulation est différente de celle de référence.

Concernant le temps de séjour des replica pour les différentes températures (figure III.3.11, III.3.12 et III.3.13). Nous observons que d'une simulation à l'autre (pour un même système) la proportion du temps de séjour peut varier. Pour plusieurs simulations, il y a une asymétrie dans le temps de séjour à basse et à haute température chez certains replica. Cette asymétrie traduit qu'une ou plusieurs répliques restent trop longtemps dans un intervalle de température, il y a donc moins de diffusions dans la dimension des températures ce qui peut impacter l'échantillonnage de nouvelles conformations.

Pour les peptides 9.1, 10.1 et 12_SS cette asymétrie s'observe dans au moins quatre simulations pour plusieurs replica. Concernant le taux d'échange entre les réplica ce dernier est supérieur à 0.6 pour des peptides à 7 résidus, à mesure que le système possède plus d'atome, ce taux d'échange moyen diminue passant à 0.52 pour les peptides cycliques 10.1 et 12_SS (tableau III.3.3).

Peptide	Sequence	Taux d'échange moyen entre replica
7.1	TkNDTnp	0.64+/-0.01
7.2	hPdqsep	0.66+/-0.01
7.3	QDPpKtd	0.64+/-0.00
8.1	DDPTprQq	0.59+/-0.01
8.2	rQpqRePQ	0.57+/-0.01
9.1	pPYhPKDLq	0.59+-/0.01
10.1	AARvpRltPE	0.52+/-0.02
10.2	EvDPehpNap	0.60+/-0.01
12_SS	HpvCIPpEkVCe	0.52+/-0.02

TABLE III.3.3 – Taux d'échange moyen entre les replica au cours des simulations de REMD. Le taux d'échange est supérieur à 0.6 pour des peptides à 7 résidus. Plus le système a d'atome et plus ce taux d'échange moyen diminue passant à 0.52 pour les peptides cycliques 10.1 et 12_SS

Avec notre jeux de données, les 5 simulations de REMD convergent uniquement pour des systèmes ne dépassant pas 8 acides aminés. Au delà de ce nombre, nous observons uniquement 1 ou 2 simulations qui ont une divergence JSD inférieure à 0.05. Cette relation peut s'expliquer par deux phénomènes. Le premier est les contraintes au niveau du squelette peptidique. En effet, ce dernier est de plus en plus contraint à mesure que le nombre d'acides aminés diminue, réduisant ainsi le nombre de conformations pouvant être adoptées par le système. Enfin la dernière raison concerne la diffusion des replica. Au delà de 8 résidus (figure III.3.12 et III.3.12), plusieurs diffusent moins dans les hautes et basses températures, nous pouvons supposer que ce problème de diffusion impacte l'échantillonnage (ce point sera vérifié plus tard).

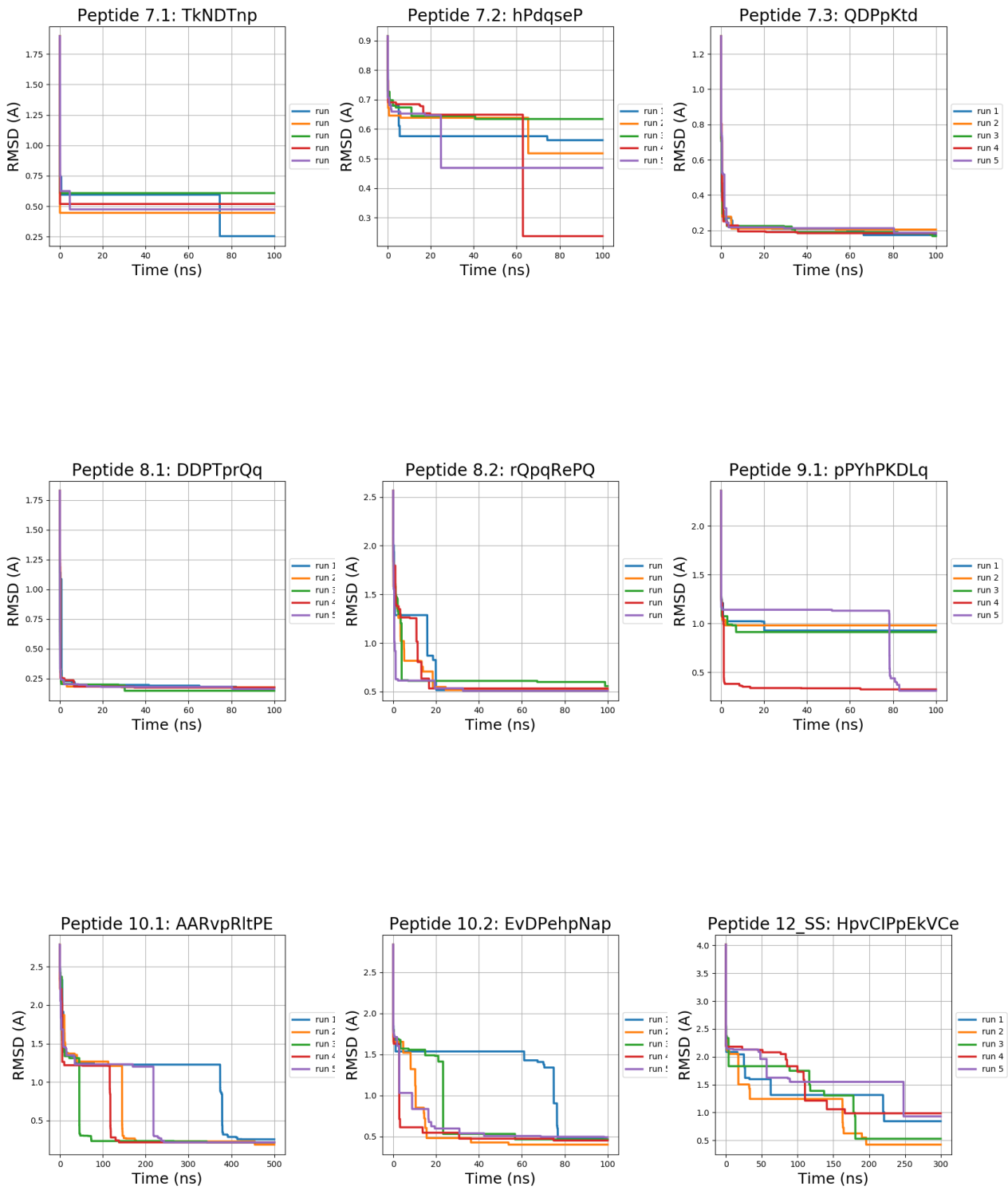


Fig. III.3.7 *Evolution du RMSD minimum au cours de la simulation pour chaque condition initiale de vitesses (runs).*

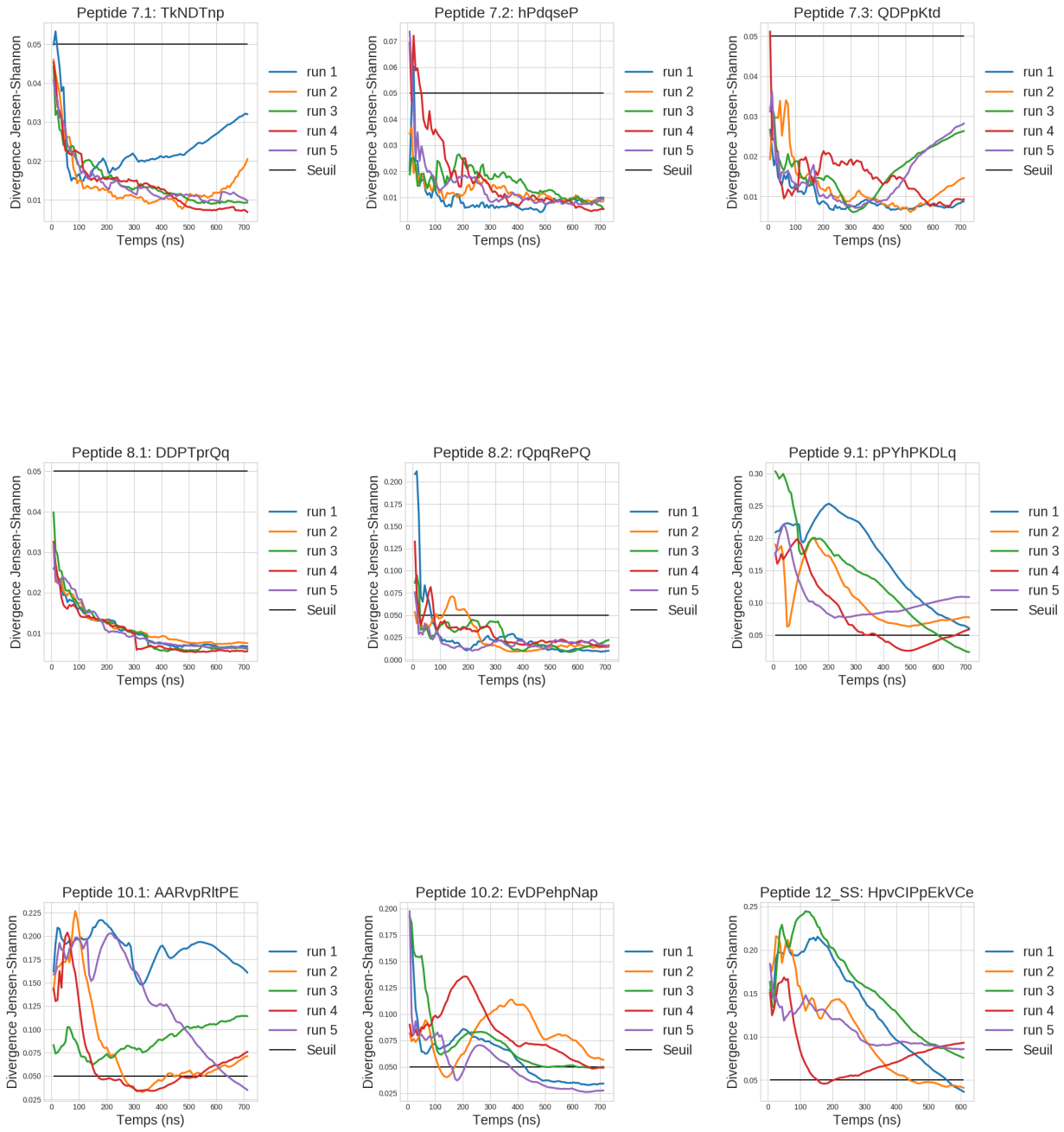


Fig. III.3.10 *Évolution de la divergence de Jenson-Shannon (JSD) de chaque simulation par rapport à une distribution de référence qui correspond au profil de RMSD moyen de chaque peptide (en utilisant les cinq simulations). Afin de ne pas être influencé par le début de la simulation, les profils de densité ont été calculés après avoir retiré les premiers 10% de la simulation (soit 80 ns). La divergence de JSD est calculée pour un pas correspondant à 1 centième de la simulation (toutes les 7,2 ns). Une simulation converge lorsque sa divergence JSD est inférieure à 0.05.*

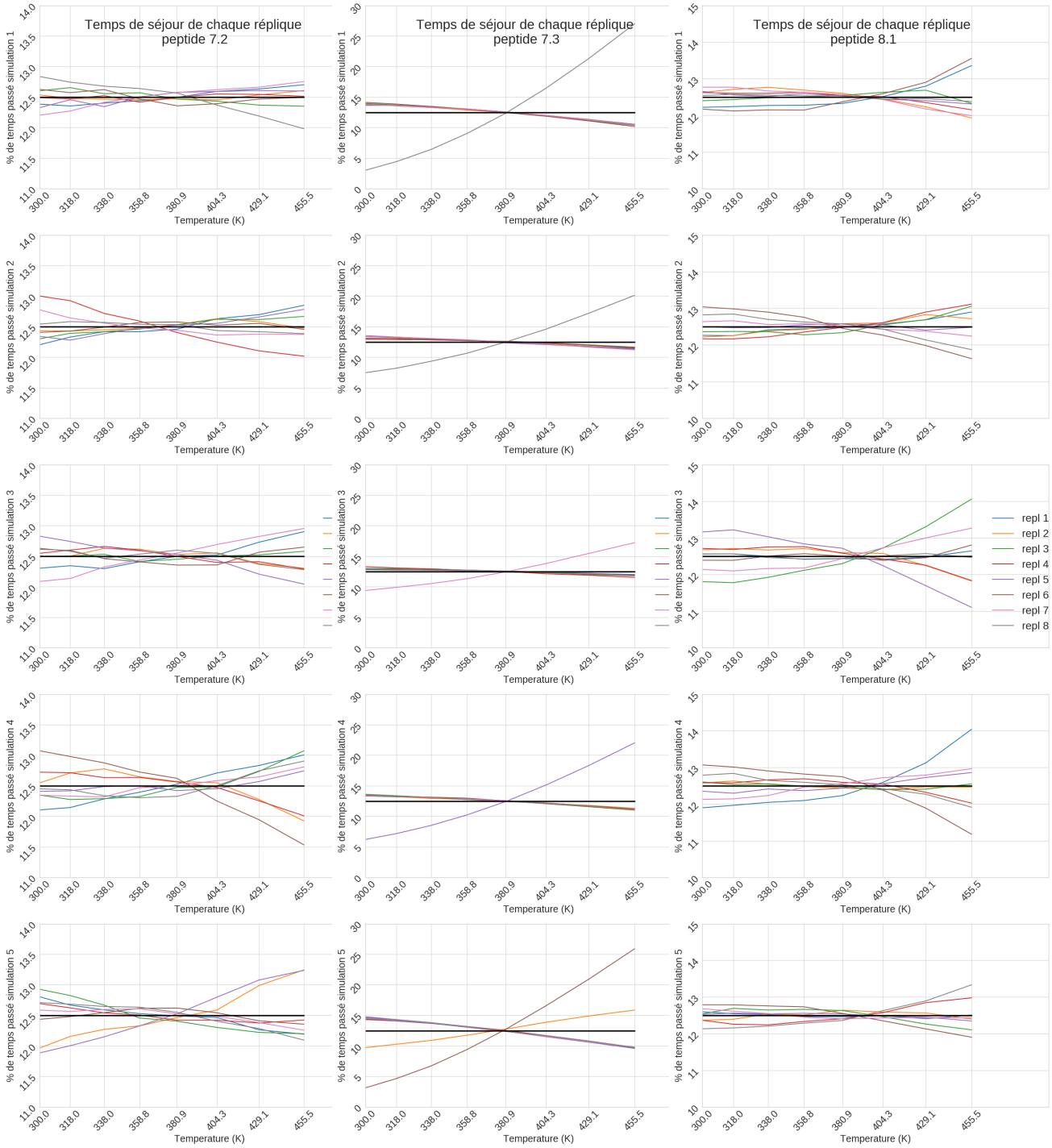


Fig. III.3.11 *Temps de séjour des replica aux différentes températures. Chaque ligne coressspond à une simulation de REMD. Le graphique permet de visualiser la bonne diffusion dans l'espace des températures des replica. Dans un modèle idéal, les replica ont un temps de séjour identique pour les différentes températures (ligne noire).*

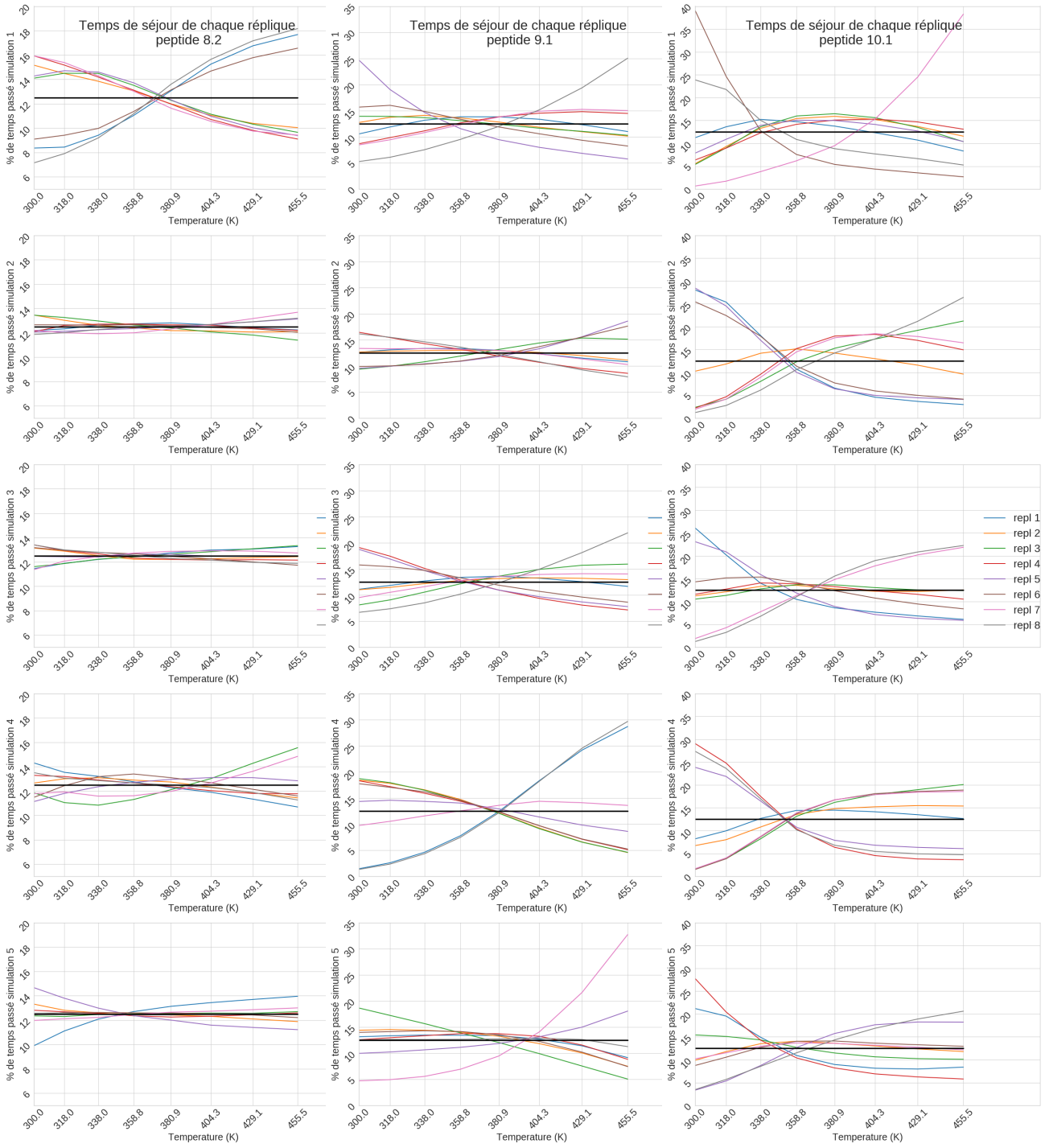


Fig. III.3.12 *Temps de séjour des replica aux différentes températures. Chaque ligne coressspond à une simulation de REMD. Le graphique permet de visualiser la bonne diffusion dans l'espace des températures des replica. Dans un modèle idéal, les replica ont un temps de séjour identique pour les différentes températures (ligne noire)*

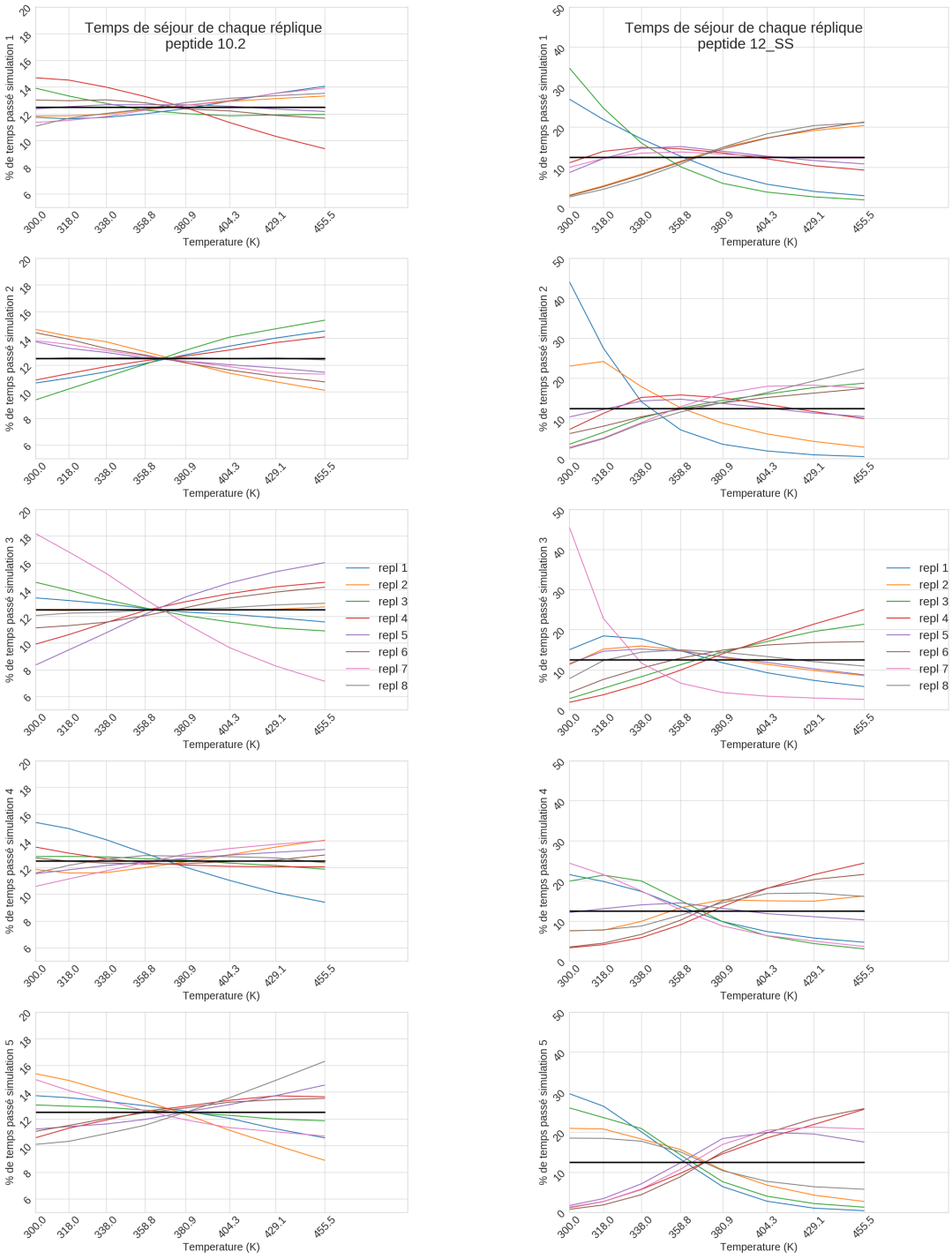


Fig. III.3.13 *Temps de séjour des replica aux différentes températures. Le graphique permet de visualiser la bonne diffusion dans l'espace des températures, les replica. Dans un modèle idéal, les replica ont un temps de séjour identique pour les différentes températures.*

III.3.2 Carte d'énergie libre

L'évolution du RMSD minimum donne une indication sur le temps nécessaire pour obtenir une conformation avec un RMSD (au niveau du squelette peptidique) inférieur ou égal à 1 Å par rapport à la structure résolue expérimentalement. Cependant cette analyse ne donne aucune indication sur les conformations en terme de proportion. Pour analyser l'ensemble des conformations produites, des projections de l'énergie libre sont réalisées suivant le RMSD et le rayon de giration (après avoir retiré les premiers 10% de chacune des 5 simulations). Les cartes présentées utilisent l'ensemble des cinq simulation et sont accompagnées des centroïdes des clusters obtenus avec la méthode des *kmeans*. Les conformations majoritaires obtenues sont renseignées dans le tableau III.3.4. Ces cartes sont accompagnées des profils de densité du RMSD des cinq simulations (figure III.3.14 à III.3.14). Ces dernières permettent de visualiser l'espace conformationnel exploré par les cinq simulations de REMD.

A l'exception du peptide cyclique 9.1, les profils de densité du RMSD montrent que l'espace conformationnel exploré est similaire dans les cinq simulations de REMD (c'est à dire que les maxima et minima des profils de densité des 5 simulations correspondent entre eux). Toutefois les proportions des conformations majoritaires diffèrent d'une simulation à l'autre, comme c'est le cas pour les systèmes 9.1, 10.1, 10.2 et 12_SS. Ce résultat n'est pas surprenant car seulement 5 peptides cycliques (7.1, 7.2, 7.3, 8.1 et 8.2) ont eu toutes leurs simulations convergées. Pour ces systèmes, nous pouvons classer les conformations les plus probables. Avec notre protocole de REMD, le cluster le plus peuplé pour les systèmes 7.2, 7.3 et 8.1 a un RMSD inférieur à 1 Å et représente plus de 70% des effectifs.

Concernant le système 7.1 (figure III.3.14), il n'y a qu'un seul cluster, ce peptide cyclique est donc très stable. En comparant le squelette peptidique de la conformation correspondant au centroïde à celle de la structure RMN, la valeur de RMSD est de 1.6 Å.

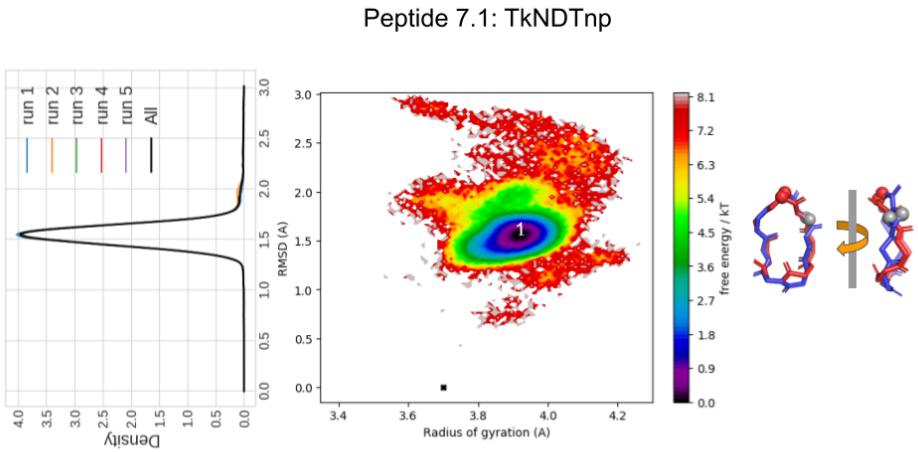


Fig. III.3.14 *De gauche à droite : profil de densité des valeurs de RMSD calculées sur le squelette peptidique par rapport à la structure RMN pour les cinq simulations. Carte d'énergie libre pour le peptide. Superposition de la conformation de référence (rouge) avec les centroïdes des groupes issus du partitionnement des kmeans et population des clusters (les carbones α des résidus 1 et 2 sont représenté en sphère rouge et grise). La croix noire représente la structure de référence.*

Afin de comprendre cette différence de structure, une carte de probabilité de contacts des chaînes latérales dans ce cluster a été réalisée. Cette carte donne la probabilité pour qu'une chaîne latérale (sans

prendre en compte les atomes d'hydrogènes) soit à une distance inférieure à 5 Å d'une autre chaîne latérale (figure III.3.15).

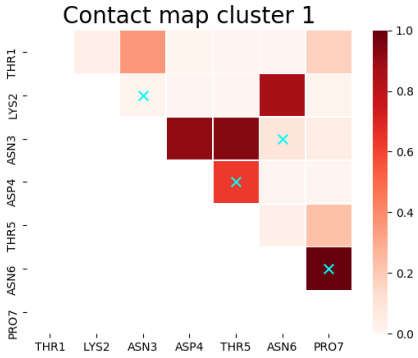


Fig. III.3.15 *Carte de contact des chaînes latérales observé pour le peptide 7.1 TkNDTnp. Cette carte donne la probabilité pour qu'une chaîne latérale soit à une distance inférieure à 5 Å d'une autre chaîne latérale. Les croix bleues correspondent aux contacts observés dans la structure résolue expérimentalement. Le centroïde de ce cluster a un RMSD (calculé par rapport au squelette peptidique de la première structure RMN) de 1.1Å*

La carte de probabilité de contacts révèle que plusieurs chaînes latérales interagissent au cours de la simulation pour former des interactions non natives. Ainsi nous pouvons supposer que cette différence de RMSD entre la structure résolue et le centroïde trouvé avec notre simulation de REMD est due à ces interaction non natives qui vont stabiliser le peptide cyclique dans une configuration différentes de celle observée en RMN.

Pour le peptide 7.2 (hPdqsep), Le groupe le plus peuplé contient 73.8% des effectifs (cluster 1) et le centroïde correspondant a squelette peptidique qui a un RMSD de 0.9 Å par rapport à la structure résolue expérimentalement. Les deux autres groupes quant à eux représentent 10.2% (cluster 2) et 16.0% (cluster 3) des effectifs et leurs centroïdes respectifs ont comme valeur de RMSD 2.2 et 1.8 Å.

Dans le premier *cluster*, 6 des 7 interactions natives sont présentes dans plus de 60% des conformations. Deux autres interactions non natives sont également présentes entre les résidus GLU6 et SER5 et GLN4 et ASP3. Pour les groupes 2 et 3, le nombre de contacts natifs observé dans plus de 50% des conformations diminue. Le groupe 2 possède 5 interactions non natives présentes dans plus de 50% des structures de ce groupe. Tandis que dans le groupe 3, ce nombre tombe à 1.

Enfin pour le dernier heptapeptide (7.3), l'échantillonnage du paysage conformationnel prédit la présence de deux configurations majoritaires pour le squelette peptidique (figure III.3.18). Les deux centroides ont des valeurs de RMSD de 0.5 et 1 Å, les conformations présentes dans ces *clusters* sont donc proches de la structure résolue expérimentalement. Le nombre de contacts natifs présents dans plus de 50% des conformations est de 8 (sur un total de 9) dans les deux groupes (8 sur 9). Au final seule la proportion de ces contacts de chaînes latérale diffère entre les *clusters*.

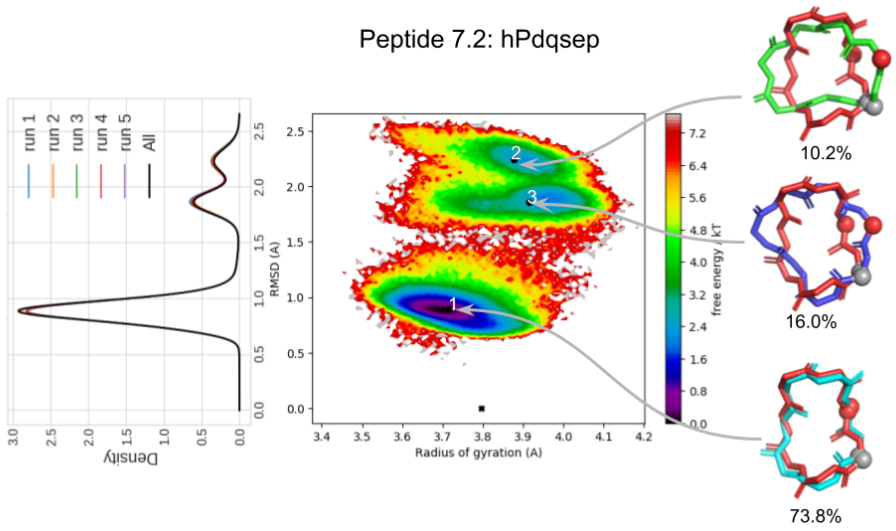


Fig. III.3.16 *De gauche à droite : profil de densité des valeurs de RMSD calculées sur le squelette peptidique par rapport à la structure RMN pour les cinq simulations. Carte d'énergie libre pour le peptide. Superposition de la conformation de référence (rouge) avec les centroïdes des groupes issus du partitionnement des kmeans et population des clusters (les carbones α des résidus 1 et 2 sont représenté en sphère rouge et grise). La croix noire représente la structure de référence.*

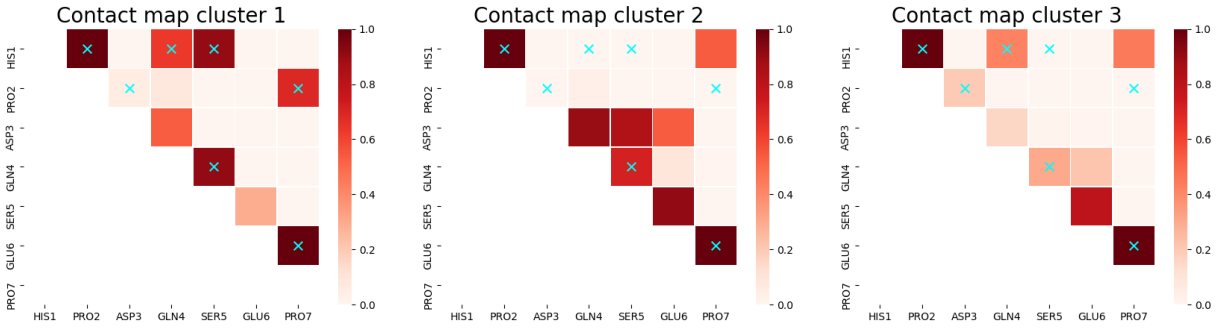


Fig. III.3.17 *Carte des interactions du peptide 7.2 (hPdqsep). Cette carte donne la proportion de contacts observés au sein des clusters trouvés avec la méthode des kmeans (figure III.3.16) et les contacts correspondant à la structure résolu (représentée par des croix). Un contact est défini lorsque la distance entre les atomes lourds des chaînes latérales de deux résidus est inférieure à 5 Å). Les effectifs des clusters par rapport à l'ensemble des conformations observées et le RMSD (calculé par rapport au squelette peptidique de la première structure RMN) de leur centroïde sont respectivement : 73.8% (0.9Å), 10.2% (2.2Å) et 16.0% (1.8Å).*

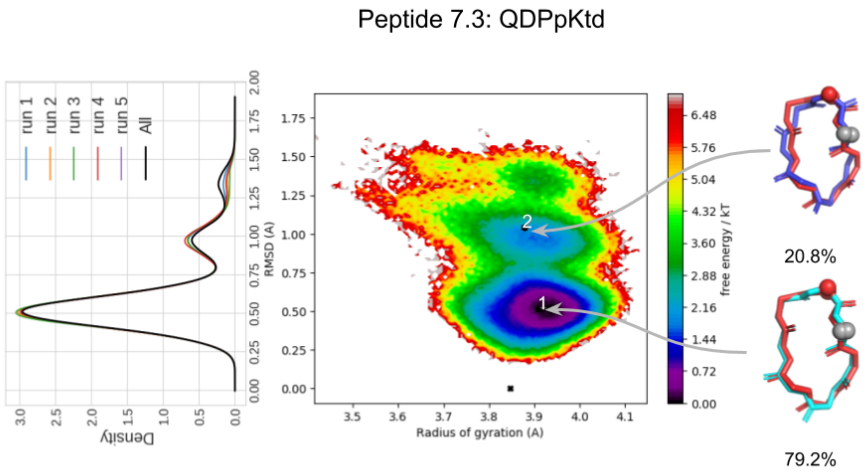


Fig. III.3.18 *De gauche à droite : profil de densité des valeurs de RMSD calculées sur le squelette peptidique par rapport à la structure RMN pour les cinq simulations. Carte d'énergie libre pour le peptide. Superposition de la conformation de référence (rouge) avec les centroïdes des groupes issus du partitionnement des kmeans et population des clusters (les carbones α des résidus 1 et 2 sont représentés en sphère rouge et grise). La croix noire représente la structure de référence.*

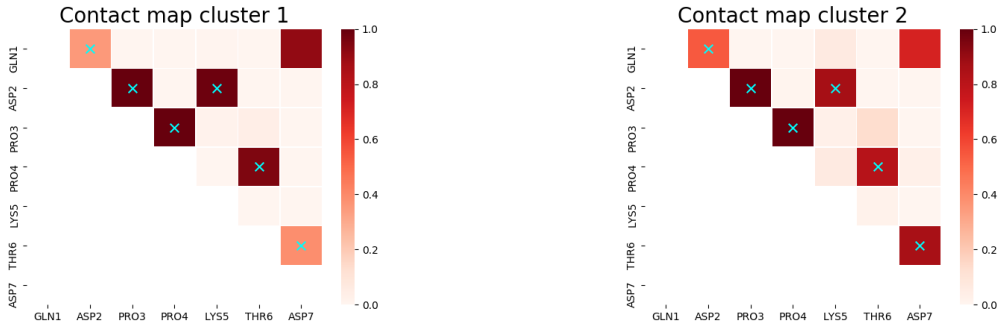


Fig. III.3.19 *Carte des interactions du peptide 7.3 (QDPpKtd). Cette carte donne la proportion de contacts observés au sein des clusters trouvés avec la méthode des kmeans (figure III.3.18) et les contacts correspondant à la structure résolue (représentée par des croix). Un contact est défini lorsque la distance entre les atomes lourds des chaînes latérales de deux résidus est inférieure à 5 Å). Les effectifs des clusters par rapport à l'ensemble des conformations observés et le RMSD (calculé par rapport au squelette peptidique de la première structure RMN) de leur centroïde sont respectivement : 79.2% (0.5Å) et 20.8% (1.0Å).*

L'exploration en REMD du peptide 8.1 (DDPTprQq) prédit deux conformations stables (figure III.3.20) dont les centroïdes ont un RMSD de 0.6 et 0.3 Å. Les contacts natifs sont présents dans les deux groupes et seule l'interaction entre les résidus ASP1 et GLN8 est présente dans moins de 20% structures (figure III.3.21).

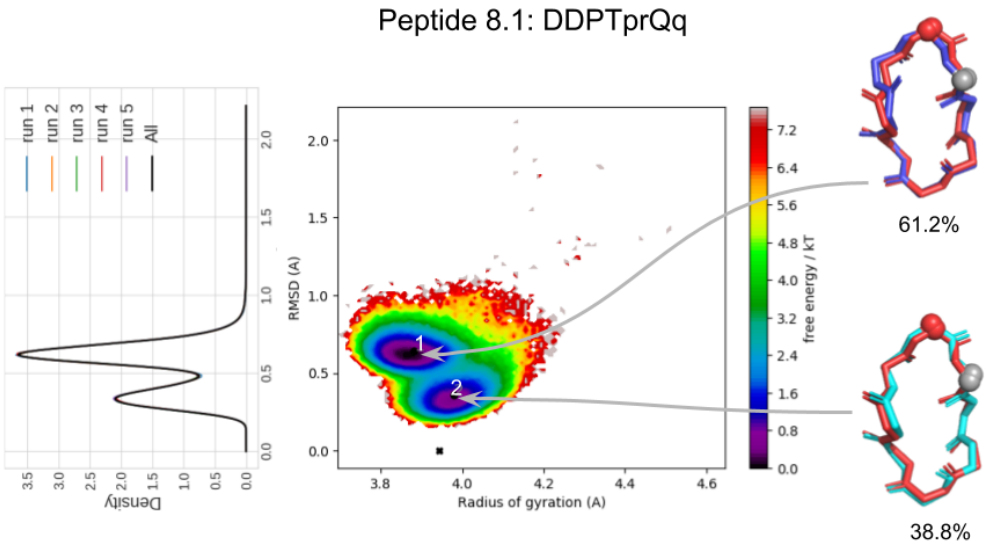


Fig. III.3.20 *De gauche à droite : profil de densité des valeurs de RMSD calculées sur le squelette peptidique par rapport à la structure RMN pour les cinq simulations. Carte d'énergie libre pour le peptide. Superposition de la conformation de référence (rouge) avec les centroïdes des groupes issus du partitionnement des kmeans et population des clusters (les carbones α des résidus 1 et 2 sont représentés en sphère rouge et grise). La croix noire représente la structure de référence.*

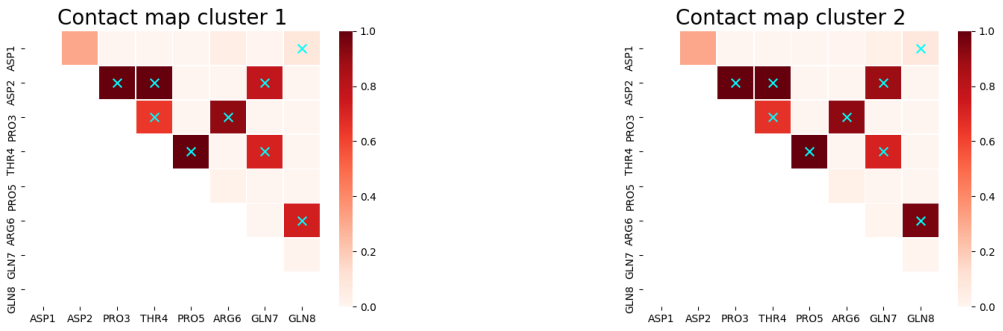


Fig. III.3.21 *Carte des interactions du peptide 8.1 (DDPTprQq). Cette carte donne la proportion de contacts observés au sein des clusters trouvés avec la méthode des kmeans (figure III.3.20) et les contacts correspondant à la structure résolue (représentée par des croix). Un contact est défini lorsque la distance entre les atomes lourds des chaînes latérales de deux résidus est inférieure à 5 Å). Les effectifs des clusters par rapport à l'ensemble des conformations observés et le RMSD (calculé par rapport au squelette peptidique de la première structure RMN) de leur centroïde sont respectivement : 61.2% (0.6Å) et 38.8% (0.3Å).*

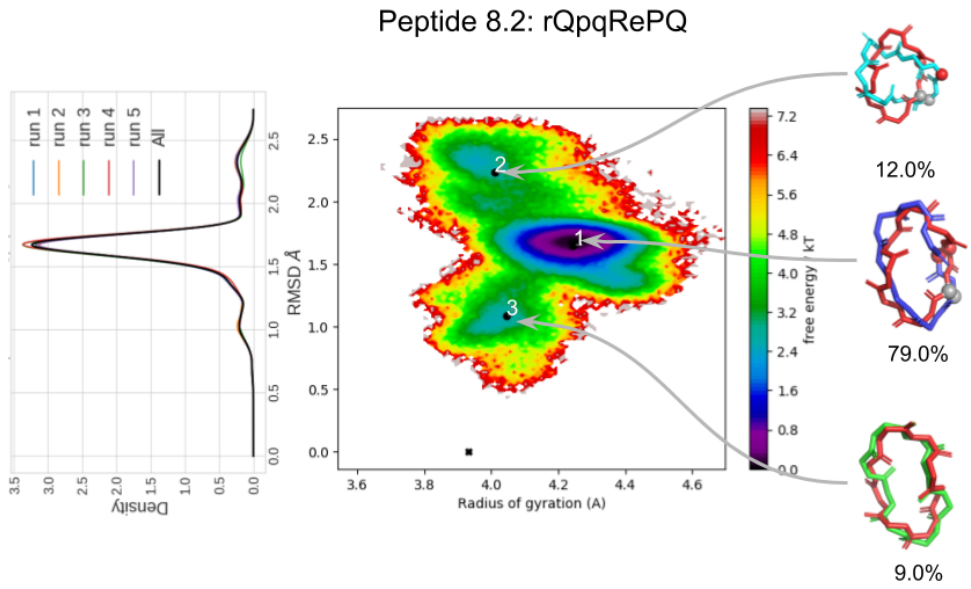


Fig. III.3.22 *De gauche à droite : profil de densité des valeurs de RMSD calculées sur le squelette peptidique par rapport à la structure RMN pour les cinq simulations. Carte d'énergie libre pour le peptide. Superposition de la conformation de référence (rouge) avec les centroïdes des groupes issus du partitionnement des kmeans et population des clusters (les carbones α des résidus 1 et 2 sont représentés en sphère rouge et grise). La croix noire représente la structure de référence.*

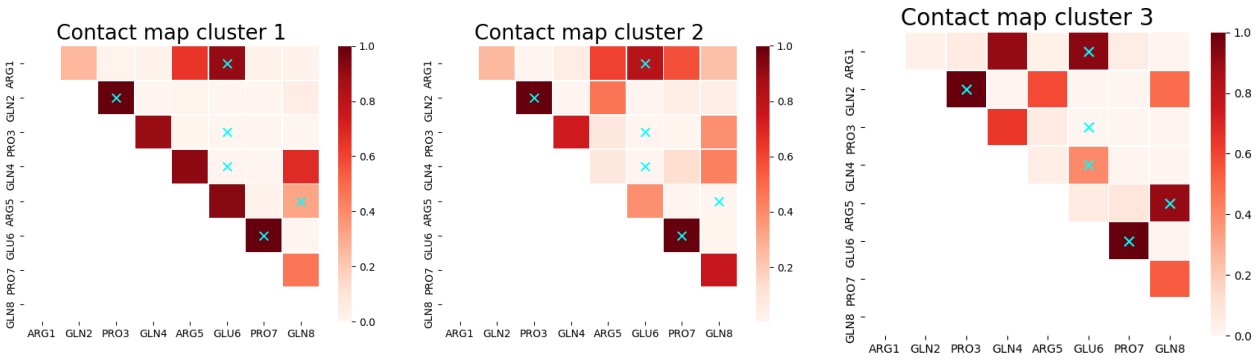


Fig. III.3.23 *Carte des interactions du peptide 8.2 (rQpqRePQ). Cette carte donne la proportion de contacts observés au sein des clusters trouvés avec la méthode des kmeans (figure III.3.22) et les contact correspondant à la structure résolue (représentée par des croix). Un contact est défini lorsque la distance entre les atomes lourds des chaînes latérales de deux résidus est inférieure à 5 Å). Les effectifs des cluster par rapport à l'ensemble des conformations observés et le RMSD de leur centroïde sont respectivement : 79% (1.6Å), 12% (2.2Å) et 9.0% (1.1Å)*

Pour le peptide 8.2 (séquence rQpqRePQ), le groupement le plus peuplé (cluster 1 dans la figure III.3.23) représente 79% des effectifs et est caractérisé par un RMSD de 1.6 Å. La carte de contacts révèle la présence d'interactions non natives par rapport à la structure RMN. De manière générale les chaînes latérales des résidus ont souvent une distance inférieure à 5 Å de leurs voisins directs. Concernant les

autres interactions, le résidu GLU6 interagit avec les résidus ARG1 et ARG5 pour former un pont salin alors qu'au sein de la structure RMN, le résidu GLU6 forme un contact uniquement avec le résidu ARG1. Cette interaction est également présente dans le groupe 2. Les résidus GLN4 et GLN8 ont leurs chaînes latérales distantes de moins de 5 Å. Pour ce dernier cas, l'acide aminé GLN8 forme une liaison hydrogène avec le squelette peptidique au niveau de l'azote de ARG5. Le deuxième groupe qui contient 12% des conformations contient des structures qui ont un RMSD supérieur à 2 Å par rapport à la structure RMN. Dans ce groupe seulement deux contacts natifs sont présents, le reste étant des contacts alternatifs. De précédentes études de REMD de peptide linéaire ont montré que l'utilisation de solvant implicite favorisaient la formation de ponts salins [123 124 125 126]. Nous pouvons donc supposer que l'interaction des résidus GLU6 et ARG1 est un artefact induit par l'utilisation de notre couple solvant implicite et champ de force.

Enfin les conformations similaires à la structure native se trouvent dans le troisième groupe qui représente 9.0% des effectifs (fig III.3.22. Cinq interactions natives sur six sont présentes avec des contacts alternatifs au niveau des chaîne latérale. Cependant contrairement aux groupes 1 et 2, les résidus ARG5 et GLU6 n'interagissent pas. Etant donné que le groupe 3 contient les structures qui ont un RMSD proche de 1 Å, nous pouvons supposer que les contacts alternatifs trouvés au sein de ce groupe ont peu d'impact sur la conformation du squelette peptidique (par rapport à la structure RMN). En outre, nous pouvons supposer que la formation du pont salin entre les résidus ARG5 et GLU6 est le principal contributeur à la stabilisation du squelette peptidique dans une conformation alternative (par rapport à la structure RMN).

Pour les autres systèmes (peptides 9.1, 10.1, 10.2 et 12_SS) les simulations de REMD n'ont pas toutes convergées, il n'est pas possible de classer les *clusters* en fonction de leurs effectifs. En effet, les proportions des conformations majoritaires sont amenées à évoluer si les dynamiques moléculaires sont prolongées. Cependant, étant donné la longueur de nos simulations (800 ns), nous pouvons supposer que les maxima observés dans les profils de densité de référence (construits en utilisant l'ensemble des valeurs de RMSD des 5 simulations de REMD) correspondent à des conformations prédites comme stables.

A mesure que le nombre d'acides aminés augmente, les peptides cycliques sont moins contraints au niveau du squelette peptidique. Au delà de huit résidus, nous pouvons observer des bassins d'énergie correspondant à des valeurs de RMSD de plus 3 Å, indiquant une grande flexibilité (tableau III.3.4).

Le peptide 9.1 possède 4 configurations probables qui ont comme valeur de RMSD 1.2, 3.0, 0.6 et 2.0 Å III.3.24).

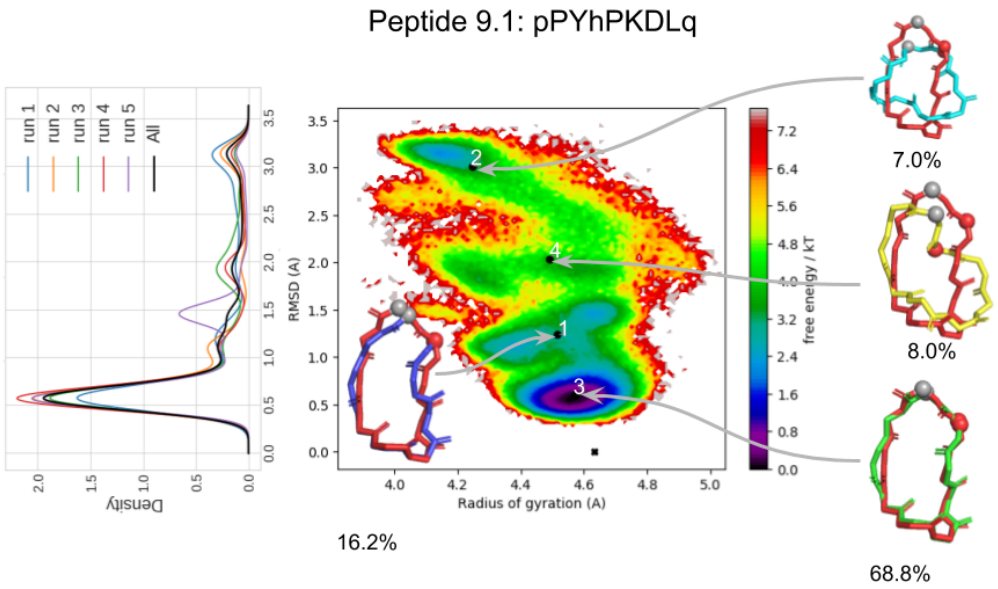


Fig. III.3.24 *De gauche à droite : profil de densité des valeurs de RMSD calculées sur le squelette peptidique par rapport à la structure RMN pour les cinq simulations. Carte d'énergie libre pour le peptide. Superposition de la conformation de référence (rouge) avec les centroïdes des groupes issus du partitionnement des kmeans et population des clusters (les carbones α des résidus 1 et 2 sont représentés en sphère rouge et grise). La croix noire représente la structure de référence.*

Les cartes de contacts indiquent l'absence de pont salin pour les conformations majoritaires (entre les résidus HIS4, LYS6 et ASP7) de chaque *cluster* (figure III.3.25). Le premier groupe a 7 et 9 interactions natives qui sont présentes dans plus de 50% des conformations. Bien que plusieurs contacts alternatifs entre les chaînes latérales soient présents dans se groupe, ils sont observés dans moins de 50% des structures.

Les conformations les plus proches de la structure RNM se trouvent dans le groupe 3. Comme avec le cluster 1, 7 des 9 interactions natives sont présentes à plus dans plus de 50% des conformations de ce groupe. 6 sont même présentes dans plus de 90% des conformations. Les résidu LYS6 et GLN9 ont leurs chaînes latérales distantes à moins de 5 Å dans plus de 50% des conformations. Étant donnée que le centroïde de ce groupe a une valeur de RMSD de 0.6 Å, nous pouvons supposer que ce contact ne déstabilisent pas la conformation du squelette peptidique.

Pour les groupes 2 et 4, les contacts natifs présents dans plus de 50% des conformations sont au nombre de 5. Le groupe 2 possède 4 contacts alternatifs par rapport à la structure RMN, ce qui induit une conformation différente de cette dernière (le centroïde a une valeur de RMSD de 3.0 Å). Le groupe 4 quant à lui présente des contacts alternatifs également, mais ces dernières sont dans des proportions inférieurs à 50% des effectifs. Ainsi dans ce cas, c'est l'absence de contacts natifs qui induit une conformation différentes par rapport à la structure RMN.

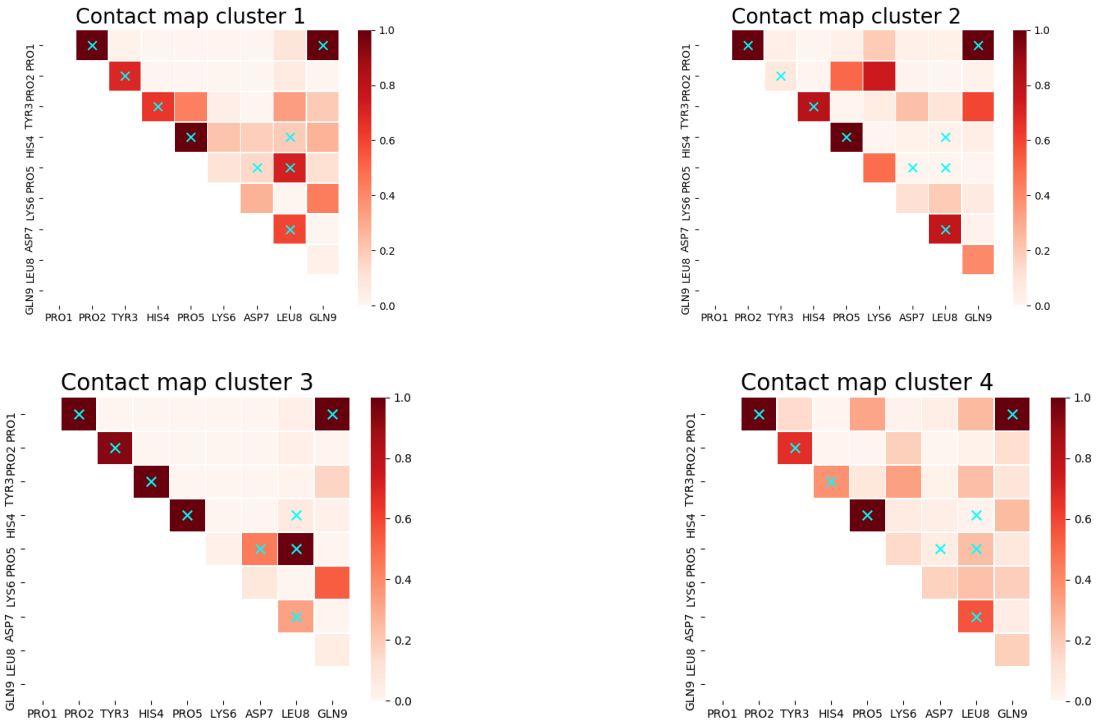


Fig. III.3.25 *Carte des interactions du peptide 9.1 (pPYhPKDLq). Cette carte donne la proportion de contacts observés au sein des clusters trouvés avec la méthode des kmeans (figure III.3.24) et les contacts correspondant à la structure résolue (représentée par des croix). Un contact est défini lorsque la distance entre les atomes lourds des chaînes latérales de deux résidus est inférieure à 5 Å). Les effectifs des clusters par rapport à l'ensemble des conformations observés et le RMSD de leur centroïde sont respectivement : 16.2% (1.2Å), 7.0% (3.0Å), 68.8% (0.6Å) et 8.0% (2.0Å).*

Pour les systèmes 10.1 (AARvpRltPE) et 10.2 (EvDPehpNap), trois configurations probables sont présentes (figure III.3.26 et III.3.28). Les valeurs de RMSD des centroïdes valent 0.5, 1.7 et 2.8 Å pour le peptide AARvpRltPE, tandis que pour le peptide EvDPehpNap leurs valeurs sont 1.0, 1.9 et 2.7 Å.

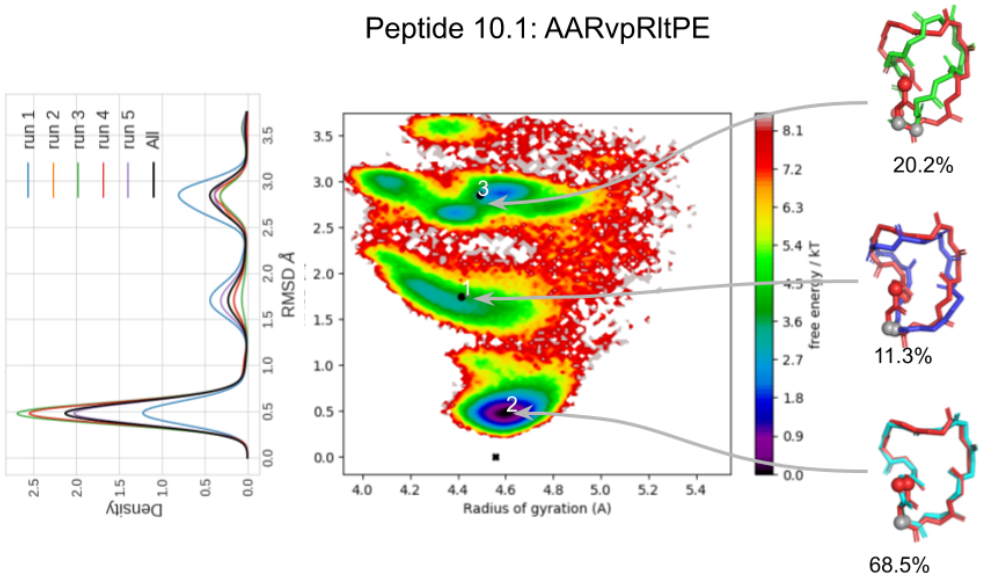


Fig. III.3.26 *De gauche à droite : profil de densité des valeurs de RMSD calculées sur le squelette peptidique par rapport à la structure RMN pour les cinq simulations. Carte d'énergie libre pour le peptide. Superposition de la conformation de référence (rouge) avec les centroïdes des groupes issus du partitionnement des kmeans et population des clusters (les carbones α des résidus 1 et 2 sont représentés en sphère rouge et grise). La croix noire représente la structure de référence.*

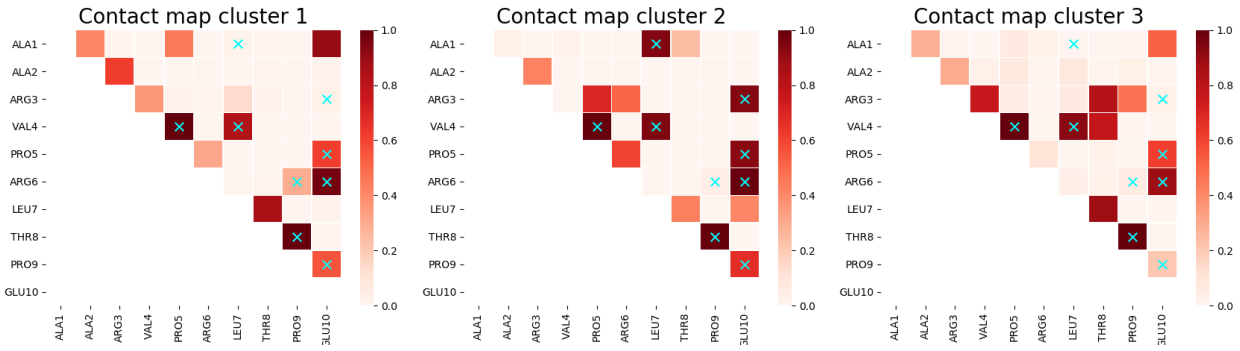


Fig. III.3.27 *Carte des interactions du peptide 10.1 (AARvpRltPE). Cette carte donne la proportion de contacts observés au sein des clusters. Un contact est défini lorsque la distance entre les atomes lourds des chaînes latérales de deux résidus est inférieure à 5 Å). Dans l'ordre d'apparition des cartes, les valeurs de RMSD des centroïdes sont respectivement 0.5, 1.7 et 2.9 Å*

Le *cluster* 2 (figure III.3.27) du peptide 10.1 contient les conformations dont le squelette peptidique présente des valeurs de RMSD proches de la structure de RMN. En effet le centroïde de ce groupe a un RMSD de 0.5 Å. Sur les 9 contacts natifs, 8 sont présents dans plus de 90% des conformations, dont no-

tamment un pont salin entre ARG3-GLU10 et ARG6-GLU10. Cette interaction des chaînes latérales entre les résidus ARG3 et GLU10 n'est pas présente dans les groupes 1 et 3. Enfin dans les interractions alternatives entre les chaînes latérales, il y a les résidus ARG3 et ARG6 qui sont à moins de 5 Å entre eux, mais également du résidu PRO5. Ces contacts sont dus à l'utilisation du solvant implicite et à l'interaction avec le résidu GLU10. En interagissant avec ce dernier, les chaînes latérales des arginines se rapprochent diminuant ainsi leur distance entre eux et avec le résidu PRO5.

Dans le groupe 1, 6 contacts natifs sont présents dans plus de 50% des conformations tandis que pour les contacts alternatifs il est au nombre de 3. Ces derniers concernent des résidus avec leurs voisins directs. Enfin dans le groupe 3, les contacts natifs présents dans plus de 50% des conformations sont au nombre de 5. Dans ce groupe, 6 contacts alternatifs (observés chez plus de 50% des structures de ce groupe) sont présents. Ces contacts impliquent le résidu ARG3 avec VAL4, THR8 et PRO9, ainsi que le THR8-VAL4 et ALA1-GLU10.

Pour le peptide 10.2 (EvDPehpNap), 3 bassins d'énergie libre sont présents. Pour les contacts natifs (figure III.3.27), plusieurs résidus ont leur chaîne latérale à une distance inférieure à 5 Å des résidus adjacents (en terme de numérotation). Pour ce qui est des contacts de chaînes latérales avec d'autres résidus non adjacents, l'ASP3 est en contact avec les résidus PRO7, ASN8, ALA9 et PRO10, tandis que HIS6 et ALA9 sont distant de moins de 5 Å.

La distance entre ces chaînes latérales ne traduit pas une interaction entre les résidus (via des liaisons hydrogènes ou ponts salins), mais plutôt une configuration du squelette peptidique qui va rapprocher les chaînes latérales de ces résidus à une distances inférieure à 5 Å.

La carte de contacts du groupe 1 (III.3.29) révèle que les conformations présentes dans ce groupe ont 6 contacts (sur 11) similaires à la structure native qui sont présents dans plus de 50% des conformations. Pour les groupes 2 et 3, ce nombre est respectivement de 6 et 9.

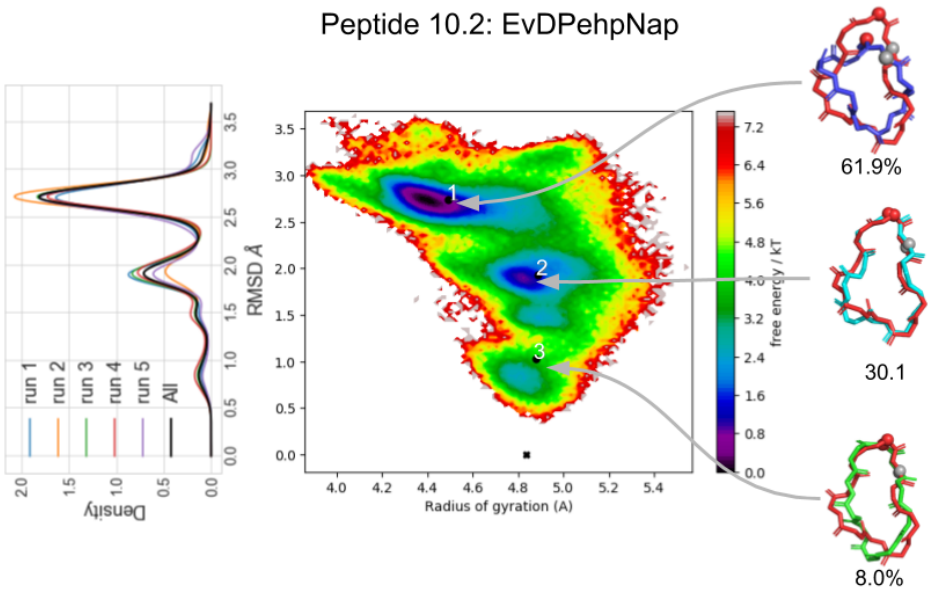


Fig. III.3.28 *De gauche à droite : profil de densité des valeurs de RMSD calculées sur le squelette peptidique par rapport à la structure RMN pour les cinq simulations. Carte d'énergie libre pour le peptide. Superposition de la conformation de référence (rouge) avec les centroïdes des groupes issus du partitionnement des kmeans et population des clusters (les carbones α des résidus 1 et 2 sont représentés en sphère rouge et grise). La croix noire représente la structure de référence.*

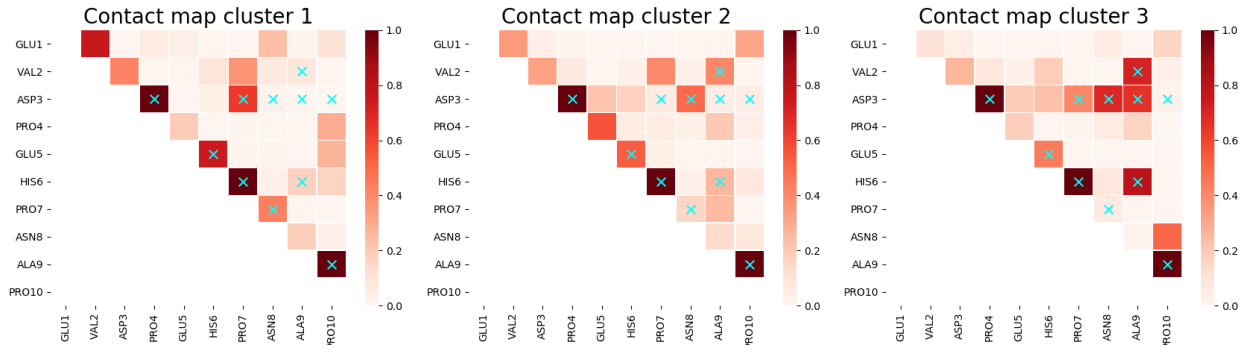


Fig. III.3.29 *Carte des interactions du peptide 10.2 (EvDPehpNap). Cette carte donne la proportion de contacts observés au sein des clusters. Un contact est défini lorsque la distance entre les atomes lourds des chaînes latérales de deux résidu est inférieure à 5 Å). Dans l'ordre d'apparition des cartes, les valeurs de RMSD des centroïdes sont respectivement 2.7, 1.9 et 1.0 Å*

Enfin le peptide 12_SS est doublement cyclisé avec un pont disulfure formé par deux résidus cystéines. Cette liaison covalente n'est pas modélisée dans notre simulation afin que le système ne soit pas bloqué dans une configuration particulière. Seules les structures pour lesquelles les atomes de soufres des deux cysteines sont à moins de 5 Å sont conservées (ce qui représente 24.6±6.5% des structures obtenus au cours de la simulation). Les deux centroïdes des configurations majoritaires possèdent un RMSD de 1.8 et 2.7 Å (figure III.3.30). Ce système est particulier car bien que les deux résidus cystéines soient proches, les contacts natifs n'impliquant pas des résidus adjacents ne sont pas présents (ou sont observés dans moins de 50% des structures) dans les deux *clusters*.

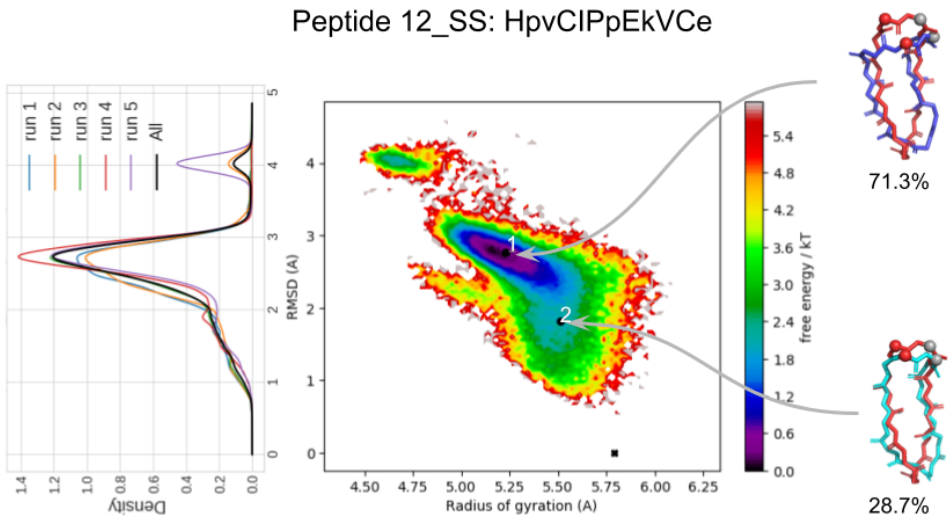


Fig. III.3.30 *De gauche à droite : profil de densité des valeurs de RMSD calculées sur le squelette peptidique par rapport à la structure RMN pour les cinq simulations. Carte d'énergie libre pour le peptide. Superposition de la conformation de référence (rouge) avec les centroïdes des groupes issus du partitionnement des kmeans et population des clusters (les carbones α des résidus 1 et 2 sont représentés en sphère rouge et grise). La croix noire représente la structure de référence.*

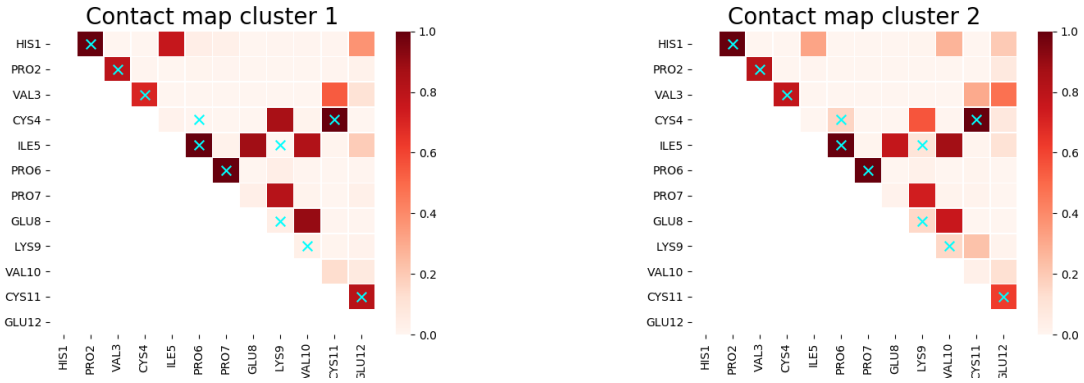


Fig. III.3.31 *Carte des interactions du peptide 12_SS (HpvCIPpEkVCe). Cette carte donne la proportion de contacts observés au sein des clusters. Un contact est défini lorsque la distance entre les atomes lourds des chaînes latérales de deux résidu est inférieure à 5 Å). Dans l'ordre d'apparition des cartes, les valeurs de RMSD des centroïds sont respectivement 2.7 et 1.8 Å*

N°	séquence	atomes	RMSD (Å) des clusters les plus peuplés
7.1	TkNDTnp	104	1.5 (100%)
7.2	hPdqsep	100	0.9 (73.8%), 1.8 (16.0%), 2.2 (10.2%)
7.3	QDPpKtd	105	0.5 (79.2%), 1.0 (20.8%)
8.1	DDPTprQq	124	0.6 (61.2%), 0.3 (38.8%)
8.2	rQpqRePQ	142	1.6 (79.0%), 2.2 (12%), 1.1 (9.0%)
9.1	pPYhPKDLq	150	0.6 (68.8%), 2.0 (8.0%), 3.0 (7.0%), 1.2 (16.2)
10.1	AARvpRltPE	160	0.5 (68.5%), 1.7 (11.3%), 2.9 (20.2%)
10.2	EvDPehpNap	141	2.7 (61.9%), 1.9 (30.1%), 1.0 (8.0%)
12_SS	HpvCIPpEkVCe	184	2.7(71.3%), 1.8 (28.7%)

TABLE III.3.4 – Récapitulatif du partitionnement supervisé (*kmeans*) sur le RMSD et le rayon de giration. La proportion des structures présentes dans chaque groupe est renseignée, ainsi que le RMSD de leur centroïde (calculé par rapport au squelette peptidique de la structure RMN)

La projection des conformations en fonction du rayon de giration et du RMSD permet de discrétiser les différentes configurations échantillonnées avec la REMD. Avec le partitionnement supervisé *kmeans*, nous pouvons constater que pour les peptides cycliques 7.2 et 7.3, plus de 70% des conformations appartiennent à un groupe dans lequel le centroïde a un RMSD inférieur où égale à 1Å (calculé par rapport au squelette peptidique de la structure RMN). Tandis que pour les autres systèmes, les structures similaires à la configuration de référence (RMSD < 1 Å) se trouvent dans l'un des trois premiers groupes, à l'exception des peptides 7.1 (qui n'a qu'un seul groupe) et 12_SS (dont les centroïdes des groupes majoritaires ont un RMSD supérieur à 1 Å).

A titre de comparaison, les valeurs de RMSD (calculés sur le squelette peptidique) des structures prédites par ROSETTA[102] par rapport à la première structure RMN dans les fichiers PDB sont indiquées dans le tableau III.3.5. Dans l'ensemble les prédictions réalisées par l'équipe de David Baker ont des valeurs de RMSD plus petites qu'avec notre méthode.

N°	séquence	RMSD (Å)
7.1	TkNDTnp	1.5
7.2	hPdqsep	0.7
7.3	QDPpKtd	0.4
8.1	DDPTprQq	0.6
8.2	rQpqRePQ	0.9
9.1	pPYhPKDLq	0.24
10.1	AARvpRltPE	0.3
10.2	EvDPehpNap	0.7
12_SS	HpvCIPpEkVCe	1.9

TABLE III.3.5 – Valeur de RMSD des conformations prédites[102] par rapport à la première structure RMN présente dans le fichier PDB. Le RMSD est calculé sur le squelette peptidique.

Ainsi l'outil de Rosetta[32] est plus précis dans la prédiction des structures que l'échantillonnage par REMD. Dans leur article, l'équipe de Baker a réalisé une étude de la stabilité par mutagénèse *in silico* dans lequel chaque acide aminé a été muté par chacun des 18 acides aminés restants (la glycine n'a pas été utilisée) avec la même chiralité, ainsi qu'un changement de chiralité en utilisant une alanine.

Le coût en calcul pour chacun des peptides cycliques synthétisés n'est pas précisé. Les auteurs de l'article ont donné l'exemple d'une commande permettant d'échantillonner un peptide cyclique de 7

glycines en utilisant 131 072 CPU pendant une heure (A titre d'exemple, la simulation de REMD du peptide 7.3 a un coût de 768 CPU heures). Les auteurs de l'article ont également évalué la stabilité des peptides cycliques en réalisant 10 dynamique moléculaire (en solvant explicite TIP3P[127] en utilisant le champ de force AMBER99-ILDN[128] de 100 ns. Pour chaque système, la structure RMN a été comparée à la structure de la conformation prédite comme la plus stable d'après Rosetta. Pour 8 systèmes, la valeur du RMSD (calculé sur les carbones α) est au alentour de 1 Å, 75% du temps.

Pour le peptide 7.3, deux simulations d'environ 1 μs ont été effectuées. Pour une simulation, le système semble bloqué dans un minimum d'énergie libre au cours duquel il adopte des conformations avec un RMSD supérieur à 1.5 Å. Bien que la structure de référence utilisée pour calculer le RMSD ne soit pas la même que la nôtre (ici c'est la structure prédite), cette simulation est intéressante puisque qu'elle montre qu'un heptapeptide peut adopter différentes conformations. Le temps nécessaire pour que le système franchisse les barrières d'énergies libres pour adopter ces conformations alternatives est supérieur à 600 ns en dynamique moléculaire classique.

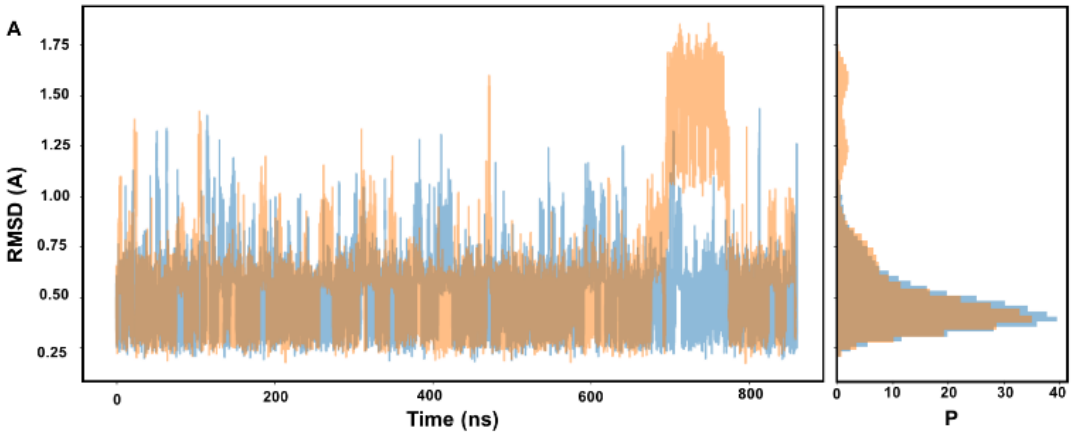


Fig. III.3.32 *Évolution du RMSD au cours d'une simulation de dynamique moléculaire en solvant explicite pour le système 7.3, d'après l'étude de David Baker[102]. Deux simulations ont été lancées (bleu et orange) en utilisant la structure RMN comme point de départ. Le RMSD a été calculé en comparant les carbones α de la structure prédite. Le peptide cyclique passe plus de 75% du temps dans une conformation proche à la structure prédite.*

III.3.3 PARTITIONNEMENT NON SUPERVISÉ DES STRUCTURES

Les cartes d'énergie libre projettent les conformations en fonction du rayon de giration et du RMSD. Toutefois, le partitionnement qui s'en suit a deux limitations. La première concerne le RMSD. Bien que très utile pour déterminer si des structures sont similaires, cette métrique a comme inconvénient de ne pas être informative pour de grandes valeurs de RMSD. En effet au sein d'un groupe composé de structures ayant toutes une même valeur élevée de RMSD, nous pouvons retrouver des structures différentes les unes des autres. En outre, dans le cas de la mise en place d'un serveur web, la structure de référence utilisée pour calculer le RMSD sera la structure moyenne de la simulation. Enfin, la seconde limitation concerne la méthode de partitionnement utilisée qui est supervisée. Il est nécessaire de spécifier à l'avance le nombre de groupes que l'on souhaite obtenir. Suivant les cas, la détermination de ce nombre peut être ambiguë.

Afin de s'affranchir de ces contraintes et d'automatiser les analyses dans le cas de la mise en place d'un serveur, un partitionnement non supervisé (et sans projection) sur les angles ϕ et ψ du squelette peptidique (avec la méthode *regular space clustering*[114]) a été effectué. Les valeurs de RMSD des centroïdes des clusters sont indiquées dans le tableau III.3.6.

N°	séquence	atomes	RMSD (Å) des clusters les plus peuplés
7.1	TkNDTnp	104	1.4 (76.8%), 1.5 (13.5%)
7.2	hPdqsep	100	0.8 (47.2%), 0.8 (13.3%), 2.0 (10.3%)
7.3	QDPpKtd	105	0.5 (53.9%), 0.4 (10.0%), 0.4 (9.3%)
8.1	DDPTprQq	124	0.7 (58.7%), 0.4 (35.2%)
8.2	rQpqRePQ	142	1.7 (68.4%), 2.4 (6.9%)
9.1	pPYhPKDLq	150	0.7 (62.5%)
10.1	AARvpRltPE	160	0.6 (36.7%), 0.7 (21.6%), 2.8 (5.3%)
10.2	EvDPehpNap	141	2.5 (23.0%), 2.8 (19.3%), 1.9 (9.0%)
12_SS	HpvCIPpEkVCe	184	2.9 (26.9%), 2.8 (15.9%), 2.5 (12.7%)

TABLE III.3.6 – Récapitulatif du partitionnement non supervisé sur les angles phi et psi. La proportion des structures présentes dans chaque groupe est renseignée, ainsi que le RMSD de leur centroïde (calculé par rapport au squelette peptidique de la structure RMN). La distance minimale entre les centroïdes est de 2.0° pour les peptides ayant moins de 10 résidus et 2.5° pour les peptides avec 10 et 12 résidus.

Au niveau des résultats, les groupes les plus peuplés avec le partitionnement sur les angles ϕ et ψ ont des valeurs de RMSD similaires à celles trouvées avec le partitionnement des *kmeans*[113] (tableau III.3.4 et III.3.6).

Pour chacun des groupes obtenus avec le partitionnement non supervisé, nous avons indiqué à quel groupe il correspondait avec la méthode des *kmeans* (tableau III.3.7). Dans le cas où un groupe obtenu avec le partitionnement *regular space clustering* est constitué de plusieurs *kmeans*,les proportions de ces derniers sont également renseignées.

Par exemple pour le peptide 7.2 (hPdqseP), les trois groupes les plus peuplés avec la méthode non supervisée ont comme effectif 47.2%, 13.3% et 10.3%. Les conformations des deux premiers groupes appartiennent au même groupe créé avec la méthode des *kmeans* (*cluster* 1 qui regroupe 73.8% des conformations totales). Le 3ème groupe contient 10.32% des structures totales. Pour ce dernier, les conformations appartiennent soit au *clusters* 3 (10.26%) et 2 (0.06%).

Comme le montre le tableau III.3.7, plusieurs *clusters* obtenus avec le partitionnement non supervisé sont constitués de plusieurs *kmeans*. Lorsque ce cas se produit, il y a un seul kmeans majoritaire qui représente plus de 90% des structures de ce groupe. La seule exception concerne le peptide 12_SS (HpvCIPpEkVCe) pour lequel le 3 groupe obtenu avec le *regular space clustering* est constitué de 2 kmeans

pour lequel lequel le kmeans minoritaire représente 1/3 des structures de ce groupe.

Ce système illustre bien les limites du partitionnement avec les *kmeans* tel que nous l'utilisons. En effet comme l'illustre la figure III.2.3, le partitionnement des conformations avec cette méthode n'est pas optimal puisqu'il est fait suivant une valeur seuil de RMSD (environ 2Å). Or pour de grandes valeurs de RMSD (> 2 Å), les conformations peuvent être différentes les unes des autres.

Enfin l'utilisation du *regular space clustering* permet de mieux discrétiser des conformations lorsque les valeurs de RMSD sont élevées. En effet pour le peptide 10.2, en partitionnant la carte 2D avec la méthode des kmeans. l'algorithme constitue un groupe qui inclut les conformations à plus de 2.5 Å et qui représente plus de 65% des conformations totales. Or les structures de ce groupe sont hétérogènes (les valeurs de rayons de gyrations vont de 4.0 à 5.4 Å). En s'affranchissant de la métrique RMSD, le partionnement sur les angles ϕ et ψ permet de mieux discrétiser des conformations. Ainsi au delà de 2.5 Å, les structures ne partagent pas toutes les mêmes caractéristiques et l'on peut trouver deux *clusters* représentant plus de 10 % des conformations totales.

Au final, le partitionnement non supervisé rassemble des conformation de manière similaire avec la méthode des *kmeans* pour 8 des 9 peptides cycliques (c'est-à-dire que chaque groupe obtenus avec le *regular space clustering* correspond à plus de 90% à un groupe de *kmeans*). Toutefois cette méthode s'avère plus intéressante que le partitionnement sur les cartes d'énergie libre puisqu'il permet de s'affranchir de la métrique RMSD. Ainsi les deux méthodes sont complémentaire, la projection 2D est un moyen simple de visualiser l'espace conformationnel tandis que le partitionnement non supervisé sur les angles ϕ et ψ discrétise mieux les conformations lorsqu'elles sont différentes de la structure de référence (RMSD > 2Å).

Peptide	Sequence	% des effectif présent dans le cluster Phi Psi	Correspondance avec le partionnement Kmeans	RMSD (A)
7.1	TkNDTnp	76.79	1	1.5+/-0.1
		13.48	1	1.5+/-0.1
7.2	hPdqseP	47.22	1	0.9+/-0.1
		13.3	1	0.8+/-1
		10.26	3	1.9+/-0.1
		0.06	2	2.0+/-0.02
7.3	QDPpKtd	53.16	2	0.53+/-0.1
		0.73	1	0.8+/-0.1
		9.87	2	0.8+/-0.1
		0.08	1	0.8+/-0.1
		9.32	2	0.4+/-0.1
		0.02	1	0.8+/-0.0
8.1	DDPTprQq	58.15	1	0.6+/-0.1
		0.51	2	0.5+/-0.0
		33.66	2	0.4+/-0.1
		1.51	1	0.6+/-0.1
8.2	rQpqRePQ	68.32	1	1.7+/-0.1
		0.04	3	1.3+/-0.0
		0.02	2	1.9+/-0.0
		6.94	2	2.3+/-0.1
		0.03	1	1.9+/-0.1
9.1	pPYhPKDLq	60.76	3	0.6+/-0.1
		1.76	1	1.0+/-0.1
10.1	AARvpRltPE	36.65	2	0.5+/-0.1
		0.03	1	1.32+/-0.1
		21.6	2	0.5+/-0.1
		5.28	3	2.9+/-0.1
		0.1	1	2.2+/-0.0
10.2	EvDPehpNap	22.98	1	2.7+/-0.1
		0.02	2	2.3+/-0.1
		19.3	1	2.8+/-0.1
		9.05	2	1.9+/-0.1
12_SS	HpvCIPpEkVCe	24.66	1	2.6+/-0.2
		2.09	2	2.1+/-0.2
		15.14	1	2.8+/-0.2
		0.8	2	2.15+/-0.1
		8.46	1	2.6+/-0.2
		4.23	2	2.1+/-0.1

TABLE III.3.7 – Comparaison du partitionnement non supervisé sur les angles phi et psi avec le partitionnement supervisé (méthode *kmeans*). La proportion des structures présentes dans chaque groupe est renseignée, ainsi que le RMSD de leur centroïde (calculé par rapport au squelette peptidique de la structure cristallographique). Lorsque qu'un groupe obtenu avec la méthode *regular space clustering* se retrouvent dans plusieurs groupes de *kmeans* il occupent plusieurs ligne sans séparation. La distance minimale entre les centroïdes est de 2.0°pour les peptides ayant moins de 10 résidus et 2.5°pour les peptides avec 10 et 12 résidus.

III.3.4 COMPARAISON 24 VS 8 REPLICA

Les observations des simulations de REMD avec différentes vitesses initiales révèlent une différence du temps nécessaire pour échantillonner le paysage conformationnel. Cette différence peut s'expliquer par le nombre de replica utilisé. Comme détaillé dans les paragraphes précédents, au delà de 8 acides aminés, une ou plusieurs replica n'ont pas un temps de séjour équivalent pour toutes les températures.

Pour rappel, la REMD est une méthode d'échantillonnage aléatoire où les conformations sont échangées suivant un critère de Metropolis (figure II.3.3). Ce critère dépend de la différence d'énergie entre les replica. Différentes études ont discuté sur l'influence du taux des fréquences des échanges, des températures utilisées et l'influence sur la convergence d'une simulation de REMD[129 119]. La fréquence des tentavies d'échange étant fixe dans notre protocole, ene façon d'améliorer cet échantillonnage est de multiplier le nombre de replica.

Dans le but de mettre en place un serveur web, où le temps des jobs est limité, nous avons voulu déterminer si l'augmentation du nombre de repliques avait un impact sur l'échantillonnage des conformations[119]. C'est-à-dire si en multipliant leur nombre, nous pouvons obtenir une conformation proche de la structure native pour un temps de simulation plus court.

Nous avons effectué 5 simulations de REMD avec 24 replica pour les peptides 10.1 et 10.2. L'objectif est de déterminer si l'augmentation du nombre de réplique améliore le taux de séjour des replica, permet de diminuer le temps moyen pour échantillonner la conformation de référence ainsi que le temps de convergence.

Le peptide 10.1 (AARvpRltPE) a été choisi car le temps nécessaire pour obtenir des simulations cohérentes est supérieur à 800 ns. Le peptide 10.2 (EvDPehpNap) quant à lui semble converger aux alentours de 800 ns avec 8 replica (à la fin des simulations, seulement 1 run a eu une divergence JSD supérieur à 5).

Le graphique III.3.33 représente le temps de diffusion des différentes simulations pour 24 replica. Pour le peptide 10.1, il ne semble pas avoir d'amélioration, plusieurs replica ont un temps de séjour différent entre les basses et hautes températures. Par contre pour le peptide 10.2, le temps de séjour est plus homogène dans les 5 simulations. Concernant le taux moyen d'échange entre répliques, il est de 0.52 ± 0.02 pour le peptide 10.1 et 0.60 ± 0.01 pour le peptide 10.2, tandis qu'avec 24 replica le taux passe à 0.85 ± 0.02 et 0.87 ± 0.01 respectivement.

Le temps moyen pour obtenir une conformation avec un RMSD inférieur à 1 Å avec 8 réplica est respectivement de 179.7 ± 111 ns et 30.0 ± 30 pour les peptides 10.1 et 10.2. Avec 24 replica le temps d'échantillonnage passe à 53.1 ± 46 ns et 10.5 ± 8. Ainsi il y a une amélioration d'un facteur 3 pour le temps moyen d'échantillonnage d'une structure avec un squelette peptidique proche de la structure RMN (RMSD inférieur ou égal à 1Å).

Concernant la convergence des simulations, pour le peptide 10.1, quatre simulations ont leur divergence *JSD* supérieure à 0.05. L'unique simulation qui a une divergence *JSD* inférieure à 0.05 franchit le seuil à 500 ns. Enfin pour le peptide 10.2, les 5 simulations ont leur divergence JSD inférieure à 0.05. Le temps moyen pour franchir ce seuil est de 432.0 ± 103 ns.

Au final l'utilisation 24 replica permet d'améliorer le taux d'échange entre les replica. Ce taux augmentant, les conformations issues des replica à haute température peuvent atteindre plus facilement la replique à 300K. Enfin le temps moyen d'échantillonnage de la structure de référence diminue.

Concernant la convergence, le peptide cyclique 10.1 ne converge toujours pas et seul le peptide 10.2 a ses 5 simulations qui ont une divergence *JSD* inférieure à 0.05 (tableau III.3.8). L'analyse du temps de présence des replica pour les différentes température révèle que le peptide 10.1 a plusieurs replica qui ne se diffusent pas assez dans l'espace des températures, tandis que pour le peptide 10.2 le temps de séjour des replica (figure III.3.33) (pour le peptide 10.2) est amélioré.

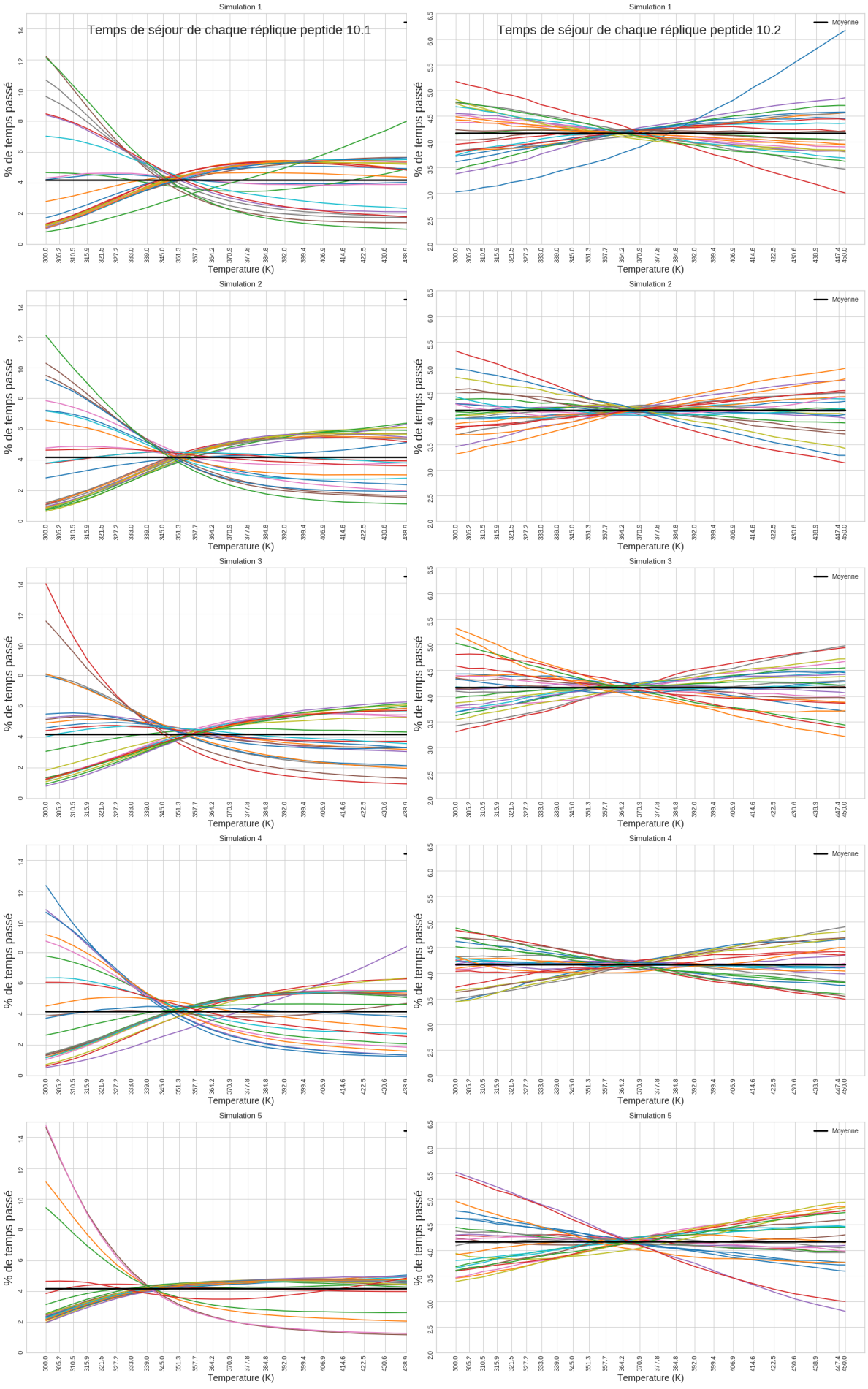


Fig. III.3.33 *Représentation du temps de séjour de chaque réplique pour différentes températures, pour les peptides 10.1 (AARvpRltPE) et 10.2 (EvDPehpNap)*

Peptide	Nombre de réplica	Temps (ns) pour RMSD <= 1 Å	divergence JS < 0.05
10.1	8	373.7, 145.0, 45.3, 116.4, 218.2	1
10.1	24	57.7, 6.58, 25.0, 36.8, 139.3	1
10.2	8	74.9, 10.5, 23.3, 2.8, 8.8	4
10.2	24	26.0, 8.5, 3.6, 13.7, 5.8	5

TABLE III.3.8 – Tableau récapitulatif du temps pour échantillonner une conformation avec un squelette peptidique inférieur ou égale à 1 Å par rapport à la structure RMN et le nombre de simulations qui ont un PDF similaire au PDF de référence (construit en prenant l'ensemble des valeurs de l'ensemble des 5 simulations)

III.3.5 Conclusion

Nous avons mis en place une méthode utilisant le REMD pour créer et échantillonner automatiquement les conformations de peptides cycliques avec des acides aminés sous forme D. Au vu des résultats, nous pouvons dire que ce protocole de REMD avec 8 réplica permet d'échantillonner en moins de 100 ns les structures RMN pour des peptides cycliques allant jusqu'à 9 résidus. Au delà, notre protocole de REMD nécessite des simulations dépassant 100 ns pour obtenir au moins une fois une conformation proche de la structure RMN ou bien d'augmenter le nombre de réplica.

Le temps de convergence varie également d'un système à l'autre, mais également d'une simulation à l'autre. Cette différence s'explique par le nombre d'acides aminés qui composent le peptide cyclique. Pour des séquences inférieures à 7 acides aminés, le squelette peptidique est contraint (cas des peptides 7 à résidus). Les degrés de libertés sont moindres et par conséquent les conformations accessibles sont plus limitées. Un autre facteur qui a une influence est la séquence en acides aminés. L'enchaînement d'acides aminés de formes D et L peut engendrer des contraintes stériques qui limitent les configurations possibles et favorisent la formation de structures secondaires. Par exemple une D-Pro suivi d'un acide aminé sous forme L favorise la formation d'une structure secondaire en épingle (β hairpin)[103]. L'utilisation d'une chaîne latérale chargée peut également avoir un impact. En effet du fait de l'absence de solvant explicite, ces résidus forment plus facilement des ponts salins, limitant ainsi l'exploration d'autres configurations.

Le dernier point concerne le taux d'échange et les températures utilisés. A mesure que le nombre d'atomes du système augmente, moins il y a d'échange entre les replica et par conséquent moins le système a la possibilité de s'échapper de minima d'énergie libre dans la simulation à 300K. Concernant les températures, ces derniers influent également sur les niveau d'énergie accessible aux systèmes. Le peptide 10.1 a montré que l'augmentation des replica ne résout pas les soucis de convergence et que de grandes barrières d'énergies libre doivent être franchi pour passer d'un bassin d'énergie libre à l'autre. L'utilisation de température au delà de 450K permettrait d'améliorer l'échantillonnage et donc la convergence des simulations (ce point sera vérifié plus tard).

Concernant la prédiction des structures RMN, les résultats varient d'un système à l'autre. Pour 7 peptides cycliques les conformations prédites ont des RMSD inférieurs à 1.6 Å. Pour les autres systèmes, les simulations n'ayant pas convergé, nous ne pouvons pas classer les conformations majoritaires en terme d'effectif. Toutefois nous pouvons supputer que les bassins d'énergies libres trouvés correspondent aux structures les plus probables. Pour le système 12_SS, les conformations présentes dans le bassin d'énergie libre ne correspondent pas à la structure RMN. Toutefois les simulations de REMD arrivent à échantillonner des structures avec un RMSD inférieur à 1 Å. Afin d'améliorer cet échantillonnage une approche possible (en perspective) est de modifier le champs de force de façon à ce que les atomes de soufre des résidus cystéines soient attractifs entre eux.

Enfin ces résultats ont montré la complémentarité des deux méthodes de partitionnement. L'utilisation d'une projection 2D en fonction du RMSD et du rayon de giration permet d'avoir une idée intuitive de l'espace conformationnel des peptides. Il s'avère efficace lorsque les valeurs de RMSD ne sont pas élevées. Par contre l'inconvénient majeur est de devoir spécifier le nombre de groupe désiré (ce qui complique la mise en place d'une analyse en ligne automatisé avec cette méthode) et de reposer sur une projection 2D utilisant comme métrique le RMSD. Le partitionnement avec la méthode du regular space clustering, quant à lui, permet de s'affranchir de la métrique RMSD et de partitionner de manière non supervisée les conformations. En outre il s'avère particulièrement utile lorsque qu'aucune structure de référence n'est disponible.

CHAPITRE IV

ÉCHANTILLONNAGE ACCÉLÉRÉ POUR L'ÉTUDE DES INTERACTIONS PROTÉINE-PROTÉINE

IV.1 INTRODUCTION

Dans le milieu cellulaire, l'activité (inactivation ou l'inhibition) d'une protéine est souvent conditionnée par l'interaction avec d'autres molécules, notamment des peptides. Pour comprendre le fonctionnement d'une protéine, il est nécessaire de prendre en compte les interactions entre récepteur et ligand. Dans ce chapitre nous présentons l'application de notre protocole de metad-BE sur le système "test" barnase (une ribonuclease) et son inhibiteur naturel la barstar. Ce complexe protéique est très étudié aussi bien expérimentalement que théoriquement. En effet le complexe contient un nombre raisonnable d'acides aminés (moins de 200) et les valeurs d'énergie libre d'association (ΔG), de k_{on} et k_{off} ont été mesurées expérimentalement [130 131]. En outre différentes études ont permis de connaître les acides aminés clés intervenant dans le processus d'association [132 133 131].

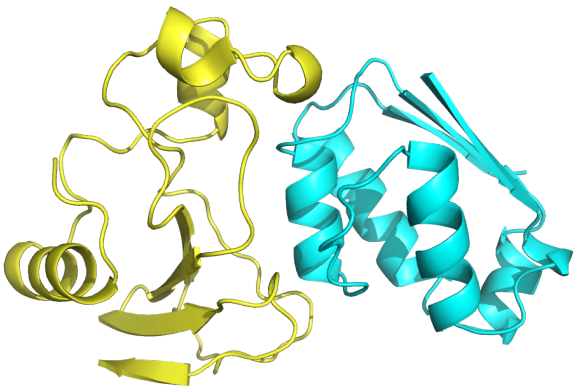


Fig. IV.1.1 *Complexe (1brs) de la ribonuclease barnase (jaune) et son inhibiteur barstar (cyan)*

Nous souhaitons évaluer si les paramètres choisis dans notre protocole permettent de retrouver les interactions protéine protéine (PPI) avec la ribonuclease barnase et son inhibiteur naturel la barstar, ainsi que les valeurs d'énergie libre de liaison et de k_{on} et k_{off}. L'objectif final est de pouvoir transposer ce protocole dans d'autres systèmes tel que la caspase 3 avec un peptide cyclique.

IV.1.1 Le problème de l'interaction entre protéine et ligand

Généralement les protéines interagissent entre elles via des liaisons non covalentes au niveau de leurs chaînes latérales. Les forces régissant ces interactions peuvent être hydrophobes, polaires (liaisons hydrogènes), forces de van der Waals ou bien électrostatiques (formation de ponts salins). Connaître ces interactions permet de comprendre les mécanismes régulant l'activité d'une protéine et aide dans le développement de potentiels inhibiteurs.

Pour rappel, à l'équilibre deux processus ont cours, l'association et la dissociation du ligand (L) à son récepteur (R) :

$$R + L \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} RL$$

Le k_{on} et k_{off} correspondent réciproquement aux constantes d'association (unité $M^{-1}.s^{-1}$) et de dissociation (unité s^{-1}). La première mesure la cinétique d'affinité du ligand pour sa cible, lorsque les deux composés sont sous forme libre. Sa valeur dépend des concentrations des réactifs. La constante de dissociation, quant à elle, peut être vue comme l'inverse du temps de résidence du ligand sur son récepteur (elle ne dépend pas de la concentration des réactifs). Enfin le rapport k_{off} et k_{oon} correspond à la constante de dissociation à l'équilibre [134].

$$K_D = \frac{k_{off}}{k_{on}}$$

Cette valeur mesure la force de l'interaction de liaison entre une biomolécule et son ligand de liaison. Ces mesures sont utiles dans l'élaboration d'inhibiteur d'IPP ou de médicament. Ainsi dans la conception d'un inhibiteur, on va chercher à avoir un ligand qui se lie favorablement à son récepteur mais également qui occupe le plus longtemps possible le site de liaison ($1/k_{off}$ le plus grand possible) [76].

Comme évoqué dans le chapitre II.5, il existe différentes méthodes expérimentales pour étudier les interactions protéiques : FRET (fluorescence resonance energy transfer),co-précipitation [135 136 137]... En plus des coûts et de la difficulté de ces expérimentations, des complications peuvent survenir en amont des mesures expérimentales, comme par exemple l'isolation des protéines d'intérêt (notamment pour des protéines transmembranaires). Enfin l'accès aux différents états métastables au cours du processus d'association n'est pas forcément accessible avec. Pour ces raisons la prédiction de l'affinité de liaison entre protéine-protéine (ou bien protéine-peptide) à partir de données structurales est un domaine de recherche très actif.

Pouvoir prédire théoriquement les PPI (avec leur affinité de liaison) est un enjeu majeur. Différentes méthodes permettent de prédire la formation de complexes protéine/protéine avec un coût en calcul relativement faible [74 42 30 32] et des résultats variables suivant le jeu de données employé. Par exemple PRODIGY[41], l'un des outils de docking de référence actuellement, a sur un jeu de 81 complexes de protéine-protéine une erreur quadratique moyenne de 1.89 kcal/mol). Par contre pour ce qui est de la prédiction du k_{off} et des différents processus entre l'état dissocié et associé, il est nécessaire d'utiliser des méthodes d'échantillonnages accélérées. Ces méthodes sont généralement coûteuses en temps de calcul [95 78 77 138] mais elles permettent de reconstruire le chemin de transition à l'aide de dynamiques moléculaires avancées. Cet échantillonnage accéléré permet de vérifier les points suivants :

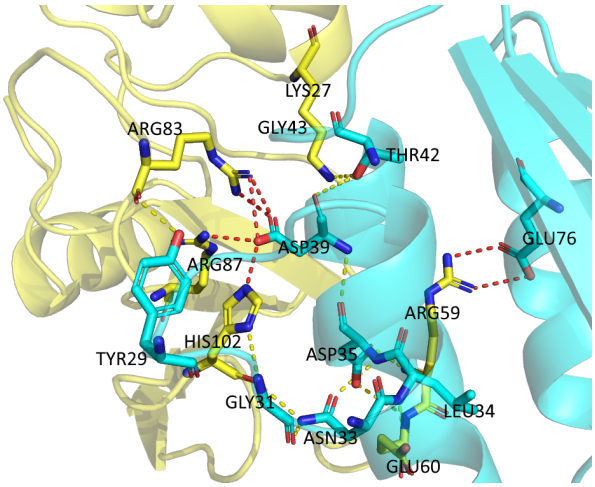
1. si A et B forment un complexe.
2. la structure du complexe.

3. l'affinité de liaison (ΔG_b), c'est à dire la stabilité du complexe par rapport à l'état dissocié.
4. k_{on} et k_{off}.
5. le mécanisme détaillé d'association (utile pour dessiner une stratégie pour modifier l'interaction).

Les trois premiers points peuvent être réalisés avec les méthodes de docking, mais l'association d'un ligand à son récepteur est un processus complexe et les approximations faites ne permettent pas toujours d'avoir une méthode suffisamment fiable et précise. En effet, l'énergie libre de liaison entre deux protéines met en jeu l'enthalpie et l'entropie du système. Omettre le solvant, ou bien représenter le récepteur et le ligand comme des corps rigides, ne permet pas de quantifier la contribution de l'entropie du système. De plus si l'on souhaite reconstruire le chemin de transition menant d'un état libre à un état complexé, il est nécessaire de prendre en compte certains mécanismes clés (comme par exemple le changement de conformation du ligand et du récepteur, la désolvatation de leurs interfaces d'interactions ou bien les interactions hydrophobes et polaires).

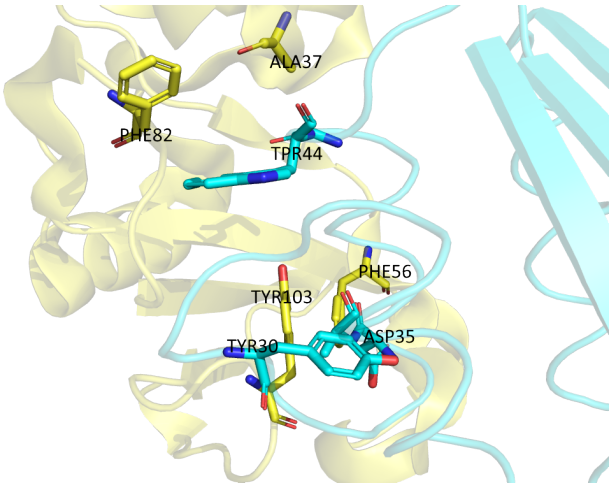
IV.1.2 LE COMPLEXE BARNASE-BARSTAR

Nous avons choisi comme système pour étudier les PPI la barnase (une ribonuclease) et son inhibiteur la barstar. Les deux protéines ont une taille raisonnable (environ une centaine d'acides aminés). Les RMSD du squelette peptidique des protéines varient peu entre l'état libre et associé des fichiers PDB. Il vaut 1.5 Å pour la barnase (comparaison avec la chaîne A du complexe barnase barnase dans le fichier PDB 1BRS et la barnase sous forme libre dans le fichier PDB 1BNR) et 1.1 Å (comparaison avec la chaîne D du du complexe barnase barstar dans le fichier PDB 1BRS et la chaine A barstar sous forme libre dans le fichier PDB 1BNR), ce qui simplifie l'étude de l'association et dissociation avec des simulations.



(a) *Détail des interactions polaires observées dans la structure résolue par cristalographie (1BRS). Les liaisons hydrogènes (pointillés verts) et les ponts salins (pointillé rouges) assurent une grand stabilité au complexe barnase (jaune) barstar (cyan)*

IV.1.2b



(b) *Détail des interactions hydrophobes observées dans la structure résolue par cristallographie (1BRS). La barnase (jaune) et la barstar (cyan).*

Fig. IV.1.2 *Représentation des contacts polaires et hydrophobes de la barnase et la barstar.*

Au niveau structural, les protéines interagissent essentiellement par des liaisons hydrogènes et des ponts salins (figure IV.1.3). Une fois les protéines associées, ces interactions assurent au complexe une grande stabilité. En effet, la constante de dissociation à l'équilibre (K_d) mesurée expérimentalement[139] par fluorimétrie est comprise entre 10^{-14} et $10^{-13} M^{-1}$. Enfin l'énergie libre de liaison et les taux d'association et de dissociation sont respectivement -18.9 kcal/mol, $6 \cdot 10^8 s^{-1} M^{-1}$ et $8.0 \cdot 10^{-6}$ s^{-1}[130]

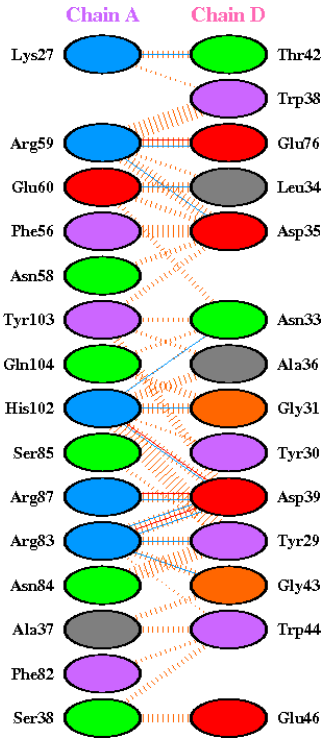


Fig. IV.1.3 *Carte des interactions du complexe barnase (chaine A) et bastar (chaine D) obtenues avec PDBsum pour la structure 1BRS. Les ponts salins (ligne rouge) et les liaisons hydrogènes (ligne bleu) contribuent à la stabilité du complexe.*

Plusieurs travaux utilisant de la dynamique moléculaire (DM) pour étudier les mécanismes de liaisons avec des prédictions de l'affinité et des constantes cinétiques ont été publiés [140 141 142 143 144].

Ces travaux ont montré la difficulté d'utiliser la DM pour ce type de problème. En effet la dissociation spontanée dans ces systèmes est impossible à observer en DM classique dans l'échelle de temps accessible à la simulation. Pour rappel, les ressources en calcul actuelles permettent d'avoir des durées de simulations de l'ordre de la μs pour des systèmes de plusieurs milliers d'atomes. [142 144 138]. Cette échelle de temps reste bien inférieure à $1/k_{off}$ (figure II.6.3). Tandis que l'association est très rapide ($k_{on} = 6 \cdot 10^8 s^{-1}.M^{-1}$), donc accessible en DM. En outre, l'absence de méthodes consensus pour estimer les affinités de liaison et les constantes cinétiques [77] est un frein supplémentaire dans l'étude des PPI avec la dynamique moléculaire.

Pour ces raisons, le système barnase barstar représente un système test idéal. Notre étude vise à montrer la possibilité de reconstruire ΔG_b et les constantes cinétiques par dynamique moléculaire avec des temps de simulation inférieurs à la μs. A la différence de *T. Chong* [138] et de *F. Noé* [142], qui ont utilisé des simulations non biaisées, nous souhaitons faire usage de méthodes d'échantillonage accéléré, comme d'ailleurs dans l'étude récente du groupe de David E. Shaw [144] qui a permis, pour la première fois, de reconstruire affinité et cinétique pour cinq complexes différents. Dans notre cas on utilise la métadynamique avec biais échangé sur des variables réactionnelles généralistes (voir section II.8.1). Ceci afin de proposer à plus long terme un protocole transposable à des complexes différents.

IV.2 PROTOCOLE DE SIMULATION

L'ensemble des simulations de DM ont été faites avec le champ de force amber99-ILDN*[128], le solvant explicite TIP3P[145] et le logiciel GROMACS[45]. Les structures utilisées pour la barnase et la barstar proviennent du complexe protéique présent dans le fichier PDB 1BRS[146] (les chaînes A et D ont été utilisées). 27069 molécules d'eau ont été utilisées accompagnées de 3 ions NA+ pour neutraliser les charges du système.

Avant de réaliser la production des dynamiques moléculaires, une minimisation et une équilibration ont été effectués. La minimisation du système s'est faite avec l'algorithme du gradient conjugué. L'équilibration s'est faite en 2 parties. Une équilibration de 100 ps à 300K en condition NVT (en utilisant le thermostat V-rescale), suivie d'une équilibration de 10 ns à 300 K en condition NPT (avec le barostat Berendsen)[147]. Enfin les dynamiques moléculaires de production (simulation de metad-BE et DM classique) ont été effectuées en condition NPT avec le barostat Parrinello-Rahman[148].

Comme expliqué dans le chapitre II.5 nos simulations de métadynamique avec biais échangés (metad-BE) utilisent 4 variables collectives (VC) : le *path CV* (ou variable de chemin), les interactions hydrophobes, les interactions polaires et enfin l'eau au voisinage de l'interface de la barnase. Les résidus se trouvant à moins de 5 Å des deux protéines ont été utilisés pour définir l'interface d'interaction. C'est sur cette interface qu'ont été appliqués les biais, la sélection des atomes a été faite à l'aide d'un script Python.

La variable de chemin[93] est définie à partir de l'interface d'interaction. Cette VC est à 2 dimensions (S et Z) permet de définir un espace conformationnel dans lequel le système peut évoluer autour d'un chemin de référence (S) avec la possibilité de s'en éloigner plus ou moins (Z) (section II.7 et figure IV.2.1).

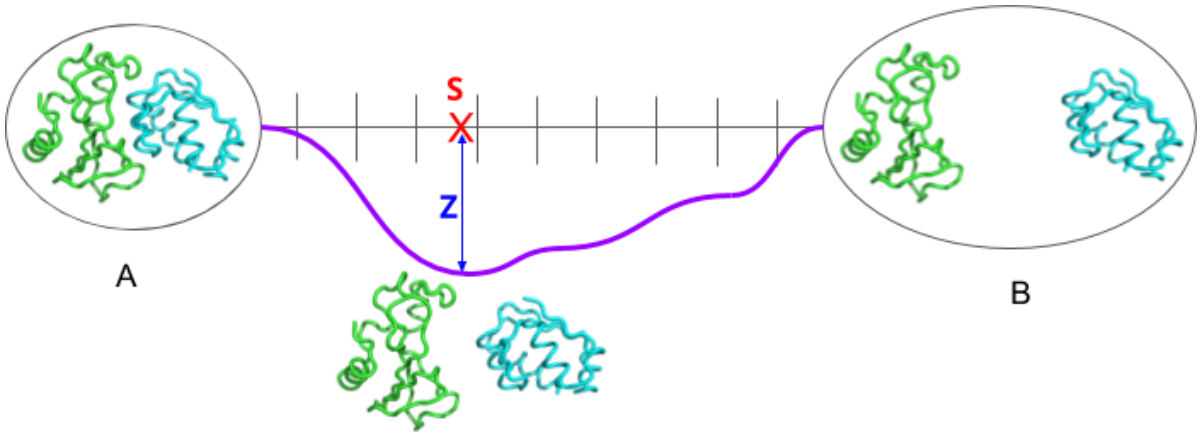


Fig. IV.2.1 *Des structures de références (symbolisées par les graduations) relient l'état A et B. En autorisant le système à s'éloigner de l'axe S, la variable de chemin optimal (en violet sur le schéma) peut être explorée.*

L'interface de la barnase a été translaté de manière à créer 7 structures de références qui servent à connecter l'état complexé à l'état dissocié. La translation de la barstar se fait uniquement sur l'axe X et les structures de références sont choisies de manière à ce que le RMSD entre une conformation et la suivante soit de 2Å. Notre variable de chemin n'étant pas le chemin le plus favorable thermodynamiquement, le complexe barnase et barstar est autorisé à s'en éloigner (figure IV.2.1). Pour limiter l'exploration et éviter des contacts non désirés entre la barnase et la barstar, nous avons restreint l'espace dans lequel les deux protéines peuvent évoluer grâce à la composante Z de la variable de chemin. Nous autorisons la barstar à diverger jusqu'à une distance de 1 nm du chemin défini avec des rotations jusqu'à 30° (sur les axes x,y,z). Nous avons calculé la valeur de Z (voir chapitre II.5) dans ces conditions limites pour déterminer

la valeur de Z maximum autorisée. Pour que le système explore l'espace dans la dimension S, les biais sont introduits sur les atomes lourds des résidus qui composent l'interface d'interaction (qui sont les résidus se trouvant à moins de 5 Å des deux protéines). L'espace de cette VC va de 1 à 7. L'énergie fournie est obtenue en sommant les gaussiennes de hauteur 0.5 kJ/mol et d'écart type de 0.15 introduite dans la composante *S* de cette variable. Pour cette VC, l'énergie est introduite toutes les 16 ps.

Pour les interactions hydrophobes, les biais sont appliqués sur les atomes des carbones des chaînes latérales. Cette VC comptabilise tous les contacts entre les atomes de carbones, entre les résidus hydrophobes de la barnase et de la barstar. Ainsi le nombre de contacts est une combinatoire. Les biais sont appliqués pour un nombre de contacts compris entre 1.5 et 110 contacts. L'énergie est introduite via le potentiel de biais toutes les 8 ps. L'énergie fournie est obtenue en sommant les gaussiennes de hauteur 0.5 kJ/mol et d'écart type de 1.5 contact qui ont été déjà déposé aux coordonnée de VC où se trouve le système (pour plus de détail voir la section II.7).

Les contacts polaires sont quant à eux comptabilisés comme des paires de contacts. Les biais sont introduits de manière à casser/former des interactions polaires préalablement définies (pont salin, liaisons hydrogènes). Les biais sont appliqués sur les atomes des résidus polaires qui participent à l'interaction (O, N). L'espace de cette VC va de 0.3 à 25 contacts. Les biais sont introduits toutes les 8 ps. Ils sont modélisés par une gaussienne de hauteur 0.5 kJ/mol et d'écart type de 0.3 contact.

Enfin la dernière variable collective est l'eau au voisinage de la barnase. Plusieurs études ont montré que la dynamique de l'eau au voisinage des protéines était différente du reste du *bulk*. Les fluctuations géométriques (à la surface de la protéine) et les liaisons hydrogènes entre le solvant et la protéine contribuent à un ralentissement des molécules d'eau d'un facteur 3 à 5[90 91 92]. Ce qui est plus important, les degrés de liberté correspondant à l'insertion ou à l'élimination de molécules d'eau à l'interface peuvent représenter une variable très lente sur l'échelle des temps des changements de conformation des protéines, ce qui demande leur inclusion parmi les VC à accélérer à travers le biais, comme démontré par plusieurs études[87 149]. Ainsi la désolvatation des interfaces d'interactions des protéines constitue une barrière d'énergie libre à franchir (les protéines doivent rompre les liaisons hydrogènes formées avec le solvant). A l'instar des contacts hydrophobes, les interactions avec l'eau sont combinatoires. Les contacts sont comptés entre les atomes O et N des résidus polaires de la barnase et les atomes d'oxygènes de l'eau. L'espace de cette VC va de 30 à 300 contacts. Les biais sont introduits toutes les 8 ps et sont modélisés par une gaussienne de hauteur 0.5 kJ/mol et d'écart type 5.0 contacts.

IV.3 RÉSULTATS

IV.3.1 PAYSAGE D'ÉNERGIE LIBRE ET STRUCTURE DU COMPLEXE PROTÉIQUE

Notre simulation utilisant 4 variables collectives, le paysage d'énergie libre résultant est en 4 dimensions. Pour faciliter la visualisation, les projections 2D de cette hypersurface sont présentées dans la figure IV.3.1. Les projections 2D montrent, qu'avec les paramètres appliqués, le système barnase barstar évolue dans une région limitée de l'espace 4D et que le paysage énergétique a une forme d'entonnoir. C'est-à-dire que durant l'association, le système n'a pas de grande barrière d'énergie libre à franchir. Ce résultat est en adéquation avec de précédentes études [140 138 144], alors que l'exploration est limitée dans la composante Z de la variable de chemin.

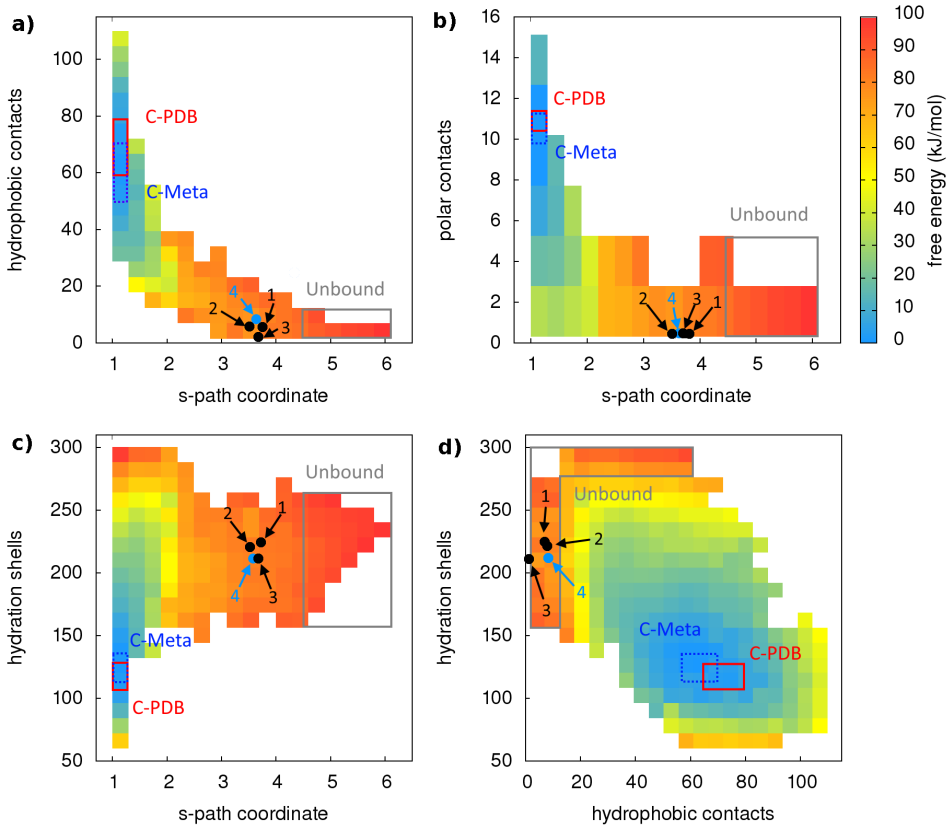


Fig. IV.3.1 *Projection 2D du paysage énergétique en fonction de deux variables collectives : a) contacts hydrophobes et composante S de la variable de chemin ; b) contacts polaires et composante S ; c) contacts polaires et solvatation de la barnase (hydratation shell) ; d) nombre de contacts entre barnase et eau, et contacts hydrophobes. Les valeurs de variables collectives correspondant à la structure résolue par cristallographie sont représentés par la région C-PDB (complexe-PDB), tandis que les valeurs correspondant au complexe trouvé avec la metad-BE sont délimitées par la région C-Meta (complexe métadynamique avec biais échangés). Les coordonnées des structures utilisées pour les trajectoires non biaisées sont numérotées de 1 à 4. Les trajectoires issues des points de départ 1 à 3 ne se dissocient pas, contrairement au point de départ 4. La région dans l'espace de variables collectives dans laquelle les deux protéines sont dissociées est indiquée avec unbound.*

Dans notre simulation l'état dissocié (qui correspond à toutes les structures pour lesquelles les atomes lourds de la barnase et la bastar sont distants de plus de 1 nm) est atteint lorsque $S > 4.5$ (figure IV.3.1).

L'ensemble des conformations ayant une valeur d'énergie libre de 5 kJ/mol au maximum identifie le complexe. Ces conformations ont un I-RMSD (RMSD calculé sur les atomes du squelette peptidique des résidus qui se trouvent à moins de 10 Å de l'autre protéine dans la structure 1BRS) de 0.9 ± 0.2 Å, L-RMSD $=1.9 \pm 0.5$ (RMSD calculé sur le squelette peptidique de la barstar après avoir aligné la barnase à la barnase de la conformation de référence) et une fraction de contacts natifs de 0.83 ± 0.04. Au vu des valeurs, la qualité de notre prédiction est considérée comme moyenne d'après le score CAPRI.

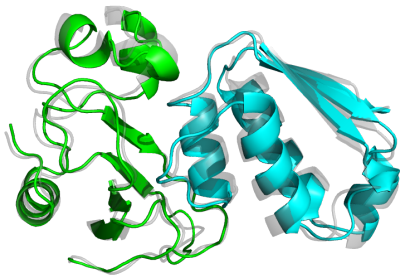


Fig. IV.3.2 *Structure issue du cluster avec la plus basse énergie. La barnase et la barstar sont représentées en vert et cyan. La structure de référence (résolue par cristallographie) est en gris. Le I-RMSD à la région de l'interface vaut* 0.9 ± 0.2 *Å, L-RMSD* $= 1.9 \pm 0.5$ *et la fraction de contacts natifs* $= 0.83 \pm 0.04$.

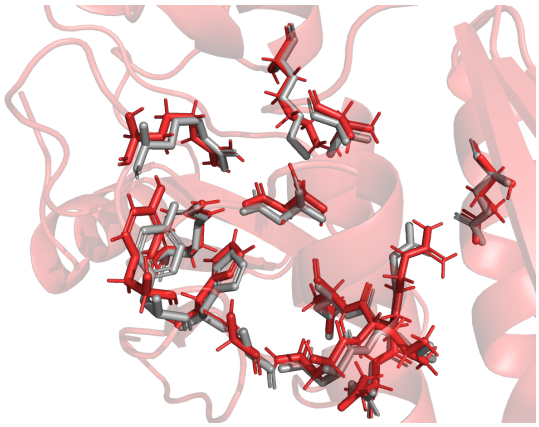


Fig. IV.3.3 *Superposition des interactions barnase et barstar prédites par la metad-BE (rouge) et celle de la structure 1BRS (gris)*

Le résultat montre que notre protocole produit des complexes proches de la structure expérimentale. De même, les valeurs des variables collectives sont similaires (figure IV.3.1, tableau IV.3.1) entre le complexe prédit et celle correspondant à la structure de référence [146] (après une dynamique moléculaire classique de 1 ns).

	Path CV	Hydrophobic contacts	Polar contacts	hydration shell
Structure de référence	1.09 ± 0.01	72 ± 7	11 ± 0.6	120 ± 10
Complexe prédit	1.15 ± 0.02	62.23 ±7	10 ± 1.2	128 ± 12

TABLE IV.3.1 – Comparaison des valeurs des variables collectives entre la structure issue du cluster avec la plus basse énergie et la structure de référence (résolue par cristallographie)

L'analyse de l'état complexé obtenu révèle que notre simulation retrouve les interactions clés, telles que la formation de ponts salins qui stabilisent le complexe formé par les deux protéines[132] (figure IV.3.3). Ces résultats montrent que le métadynamique avec biais échangé arrive à trouver le bon complexe mais également les bonnes interactions.

Le temps total de notre simulation est 556 ns, l'estimation de l'énergie libre de liaison (ΔG_b) obtenue est de -21.7 kcal/mol. Après application de la correction pour être en condition standard (c'est-à-dire une concentration de 1M), l'énergie libre standard de liaison (ΔG_b^o) vaut -23.4 ±2.1 kcal/mol. L'estimation de l'écart type n'est pas fournie car dans notre première simulation de metad-BE, la division en trois blocs temporels de notre simulation ne nous permet pas de le calculer. En effet, l'échantillonnage des quatre VC dans le le second et dernier tiers de la simulation n'est pas la même (figure IV.3.4). Une meilleure estimation de la barre d'erreur sera fournie après avoir analysé les nouvelles simulations qui ont été faites (mais pas encore analysées durant la rédaction du manuscrit de thèse).

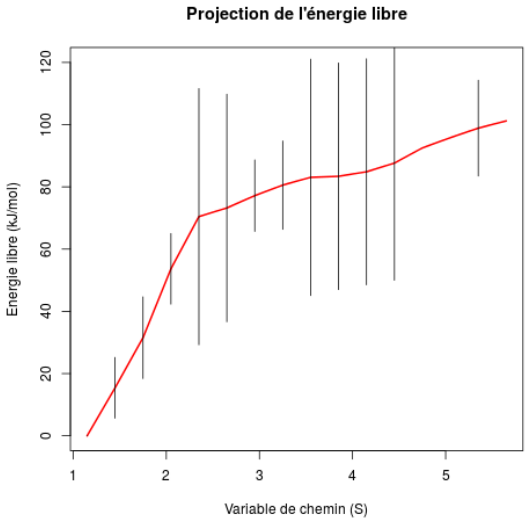


Fig. IV.3.4 *Projection du paysage d'énergie libre en fonction de la VC variable de chemin. Les écart-types sont représentés par les barres d'erreur. Dans le dernier tiers de la simulation, certaines régions ne sont pas explorées. Il n'est donc pas possible d'assigner les valeurs d'écart-types correspondantes.*

En comparaison, la valeur expérimentale obtenue par mesure de fluorescence à T=25 C°et pH=8 est -18.9 kcal/mol[130]. L'estimation obtenue par D.E.Shaw et collègues est de -19.2 kcal/mol[144]. Ce dernier a effectué une simulation de 400 µs en utilisant le champ de force AMBER99-ILDN*[128] (mais avec des corrections d'angle de torsion), le solvant TIP3P et la méthode d'échantillonage accéléré *tempered binding*. Cette méthode d'échantillonnage accéléré biaise les interactions électrostatiques entre les protéines en utilisant un facteur multiplicateur λ défini par les auteurs de l'étude (et qui a un nombre fini de valeurs possibles). A l'instar d'une simulation de REMD, à intervalle régulier, des échanges entre des valeurs adjacentes de λ sont tentés. En modifiant les interactions électrostatiques entre les protéines, le système

peut se dissocier et s'associer plusieurs fois au cours de la simulation.

L'une des critiques possibles concernant cette simulation concerne le fait que l'exploration conformationnelle de notre système est restreinte, via un mur sur la composante Z des variables de chemin. Pour rappel, l'utilisation de ce mur permet de rationaliser l'exploration dans l'espace des variables collectives, d'éviter des interactions non désirées et la formation de complexes différents de ceux que l'on souhaite évaluer. Ainsi les résultats présentés ne concernent qu'un sous ensemble des configurations possibles. Nous avons testé l'effet de varier la position du mur. Finalement, afin de valider nos résultats obtenus avec la metad-BE, une comparaison avec une série de simulations de dynamique moléculaire non biaisée a été effectuée, comme décrit dans la section suivante.

Variable collective					fnat	I RMSD
S	Z	hydrophobe	polaire	H shell		
3.76	0.17	5.4	0.06	224	0.06	6.7
3.5	0.16	6.9	0.07	219	0.06	6.2
3.7	0.11	0.45	0.04	210	0.12	6.3
6.7	0.22	8.22	0.12	211	0.12	6.9

TABLE IV.3.2 – Caractéristiques des 4 conformations qui ont été utilisées pour réaliser les dynamiques moléculaires libres. Les valeurs des variables collectives sont indiquées dans les 5 premières colonnes. S et Z sont les deux composantes des variables de chemin. Hydrophobe et polaire comptabilisent les contacts et H shell est la couche d'eau présente au niveau de l'interface de la barnase. Les contacts natifs de ces structures sont comparés entre la structure résolue (Référence) et le complexe le plus stable trouvé par notre simulation de metad-BE

IV.3.2 MÉCANISMES D'ASSOCIATION

Basé sur l'étude en ref. [144] (figure IV.3.5), l'état de transition pour la barnase barstar a comme caractéristique d'être hydraté au niveau de l'interface de dimérisation et d'avoir moins de 20% de contacts natifs formés. Ces contacts natifs sont définis comme étant tous les atomes lourds des deux monomères se trouvant à moins de 4.5 Å dans la structure 1BRS. Le comptage s'effectue en utilisant la formule décrite dans la référence **??**.

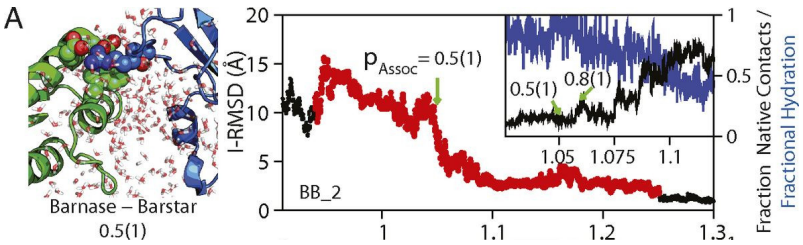


Fig. IV.3.5 *Caractéristiques d'un des états de transitions de la barnase barstar*[144]*. Moins de 20% de contacts natifs sont formés et le I-RMSD est de 10 Å.*

Quatre conformations avec moins de 20% de contacts natifs et une interface hydratée (IV.3.2) ont été choisies comme point de départ pour lancer, à partir de chacune, 50 dynamiques moléculaires libres de 100 ns initialisées à partir de vitesses atomiques tirées de la distribution de Maxwell-Boltzmann. Ces trajectoires ont pour objectif de localiser l'état de transition pour le processus d'association, et aussi de valider les mécanismes d'association et la structure du paysage d'énergie libre trouvés avec les simulations biaisée. De plus, les trajectoires libres nous ont permis de construire un modèle markovien qui donne accès à la cinétique du système.

Les différentes conformations explorées par les trajectoires libres peuvent être caractérisées comme dissociées (état D en Figure 8.1), formant le complexe natif (état C en figure 8.1), ou bien formant des contacts mais pas identiques à ceux du complexe natif. Ce dernier cas, d'ailleurs, peut-être séparé en deux groupes : les conformations avec un nombre important de contacts (natifs ou pas), donnant lieu à des complexes avec une certaine stabilité, indiqués avec CA, ou bien avec un nombre très faible de contacts (donc avec une stabilité prévue comme faible), indiquées avec CF.

Afin d'identifier les structures comme appartenant aux quatre groupes ci-dessus, nous avons analysé plusieurs indicateurs. Parmi eux, la surface enfouie (*buried surface area*, notée SE) est utilisée pour comptabiliser le nombre de simulations dans lesquelles la barnase et la barstar forment un complexe. Cette métrique permet d'identifier tous les états complexés (même ceux alternatifs de la structure réso-

Structure	% état C	% état CA	% état D	% état CF
1	14	6	0	80
2	14	0	0	86
3	22	8	0	70
4	24	6	4	70

TABLE IV.3.3 – Résultat des trajectoires libres. Les structures forment un complexe pour une surface enfouie > 1000 Å². Lorsque la composante S de la variable de chemin est inférieure à 1.5, le complexe est considéré natif (C), dans le cas contraire c'est un complexe alternatif (CA). Les protéines sont considérées comme dissociées (D) lorsque la distance entre leurs atomes lourds est supérieure à 1 nm. Le groupe CF identifie les groupes pour lequel les protéines interagissent faiblement (0<SE<1000 A²)

Variable collective					fnat	I RMSD
S	Z	hydrophobe	polaire	H shell		
1.13+/-0.07	0.03+/-0.02	71+/-9	8+/-2	114+/-15	0.75+/-0.08	1.2+/-0.5
1.15+/-0.09	0.03+/-0.02	65+/-7	9+/-2	139+/-20	0.78+/-0.06	1.2+/-0.4
1.14+/-0.06	0.04+/-0.01	68+/-11	8+/-2	145+/-15	0.76+/-0.07	1.2+/-0.5
1.14+/-0.05	0.04+/-0.05	68+/-8	9+/-3	141+/-21	0.75+/-0.11	1.4+/-1.0

TABLE IV.3.4 – Caractéristiques des structures qui forment le complexe natif (état C). Le I-RMSD est calculé sur les carbones alpha des résidus de l'interface. Les valeurs des variables collectives sont indiquées dans les 5 premières colonnes. S et Z sont les deux composantes des variables de chemin. Hydrophobe et polaire comptabilisent les contacts et H shell est la couche d'eau présente au niveau de l'interface de la barnase. Les contacts natifs de ces structures sont comparés par rapport à la structure résolue [146] et le complexe le plus stable trouvé par notre simulation de métadynamique avec biais échangé. Le comptage s'effectue en utilisant la formule décrite dans la référence [150]

lue), donc C, CA et CF. Nous considérons que la barnase et la barstar sont associées fortement (états C, CA) lorsque la surface enfouie (SE) est supérieure ou égale à 1000 Å², tandis qu'une valeur 0<SE<1000 identifie le groupe CF. En plus de la SE, la variable S (variable de chemin) est utilisée pour identifier les conformations qui forment un complexe similaire au complexe natif (groupe C) pour $1 < S < 1.5$.

L'analyse des trajectoires (tableau IV.3.3) montre que le taux d'association des protéines est entre 14% à 30 % suivant le point de départ. En outre, à l'exception du point de départ n°2, les protéines forment des complexes alternatifs minoritaires par rapport à la structure de référence. Enfin, concernant la dissociation des protéines, cet évènement s'est produit uniquement dans la trajectoire issue du point de départ n°4. Dans le reste des cas, les protéines interagissent via quelques résidus. Donc, nous pouvons conclure que la structure 4 peut-être identifiée comme la plus proche de l'ensemble des états de transition : ces derniers sont les structures qui relaxent spontanément vers les états C et D avec une probabilité similaire. Nous pourrions mieux caractériser l'ensemble de ces états en considérant davantage de structures initiales pour les dynamiques libres et en rallongeant la durée de ces dernières (afin de réduire la fraction de complexes non-natifs CA et CF). A titre d'exemple D.E.Shaw et collaborateurs ont réalisé 61 simulations de 5.5 μs pour leur DM libre.

Concernant les bonnes associations (état C), leurs caractéristiques sont renseignées dans le tableau IV.3.4. Les valeurs des variables collectives sont dans l'intervalle des valeurs correspondant à la structure de référence et prédite.

Afin de saisir les interactions types qui ont lieu lors de l'association de la barnase et de la barstar, des cartes de contacts sont réalisées où sont répertoriées les résidus dont les carbones α ont entre eux une distance inférieure à 1 nm (figures IV.3.7 et IV.3.8).

Variable collective					fnat	I RMSD
S	Z	hydrophobe	polaire	H shell		
2.2+/-0.7	0.4+/-0.3	34+/-14	1.8+/-0.5	196+/-14	0.29+/-0.09	6.2+/-3.1
X	X	X	X	X	X	X
2.2+/-0.9	0.4+/-0.1	32+/-17	3.2+/-1.0	179+/-13	0.30+/-0.10	7.1+/-1.5
1.82+/-0.3	0.5+/-0.1	25+/-4	2.2+/-0.2	169+/-8	0.35+/-0.06	7.2+/-0.0

TABLE IV.3.5 – Caractéristiques des structures qui forment un complexe alternatif. Le I-RMSD est calculé sur les carbones alpha des résidus de l'interface. Les valeurs des variables collectives sont indiquées dans les 5 premières colonnes. S et Z sont les deux composantes des variables de chemin. Hydrophobe et polaire comptabilisent les contacts et H shell est la couche d'eau présente au niveau de l'interface de la barnase. Les contacts natifs de ces structures sont comparés par rapport à la structure résolue (la référence) et le complexe le plus stable trouvé par notre simulation de métadynamique avec biais échangé (cluster)

Les mécanismes menant à l'association trouvés avec les dynamiques libres sont similaires à ceux décrits dans d'autres études [151] [142]. Pour des valeurs de variable S (variable de chemin) comprises entre 4.1 et 2.95, la barstar interagit avec une boucle de la barnase. Du fait de la flexibilité de ces boucles, la barstar peut se réorienter autour de ce point de contact. Puis aux alentours de S valant 2, le système atteint un état de pré-association. Les résidus bn-Arg59 et bs-Asp35 (figure IV.3.7) interagissent et forment un pont salin.

Il a été montré dans de précédentes études que le résidu bn-Arg59 et bs-Asp35 sont des résidus clés dans l'association et que la mutation de la bn-59 en alanine réduit le taux d'association d'un facteur 10 [152].

L'étape suivante consiste à un "ancrage" de la barnase avec la barstar via l'interaction d'une seconde boucle de la barnase (résidus 35-41) avec la barstar (notamment le résidu GLU46).

Une fois cette interaction effectuée, les protéines vont augmenter leurs contacts hydrophobes et polaires, jusqu'à la formation d'un complexe similaire à celui natif (à partir d'un S de 1.45). La combinaison optimale de contact hydrophobe, polaire et d'eau à l'interface est atteinte lorsque S vaut 1.1.

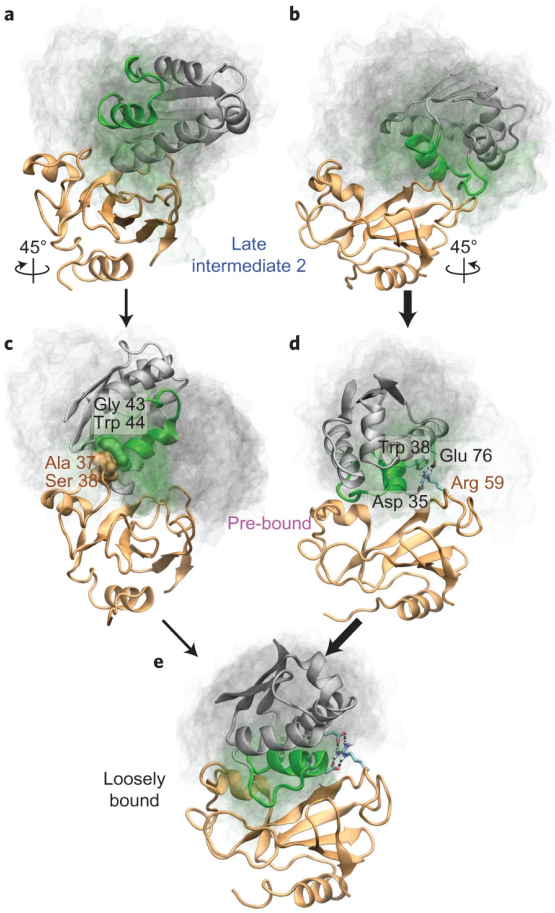


Fig. IV.3.6 *Noé et ses collaborateurs*[142] *décrivent deux mécanismes menant à l'association. Le chemin a, c, e ou bien b, d, e. Ces interactions clés sont trouvées dans la carte pour un S compris entre 1.75 et 2.*

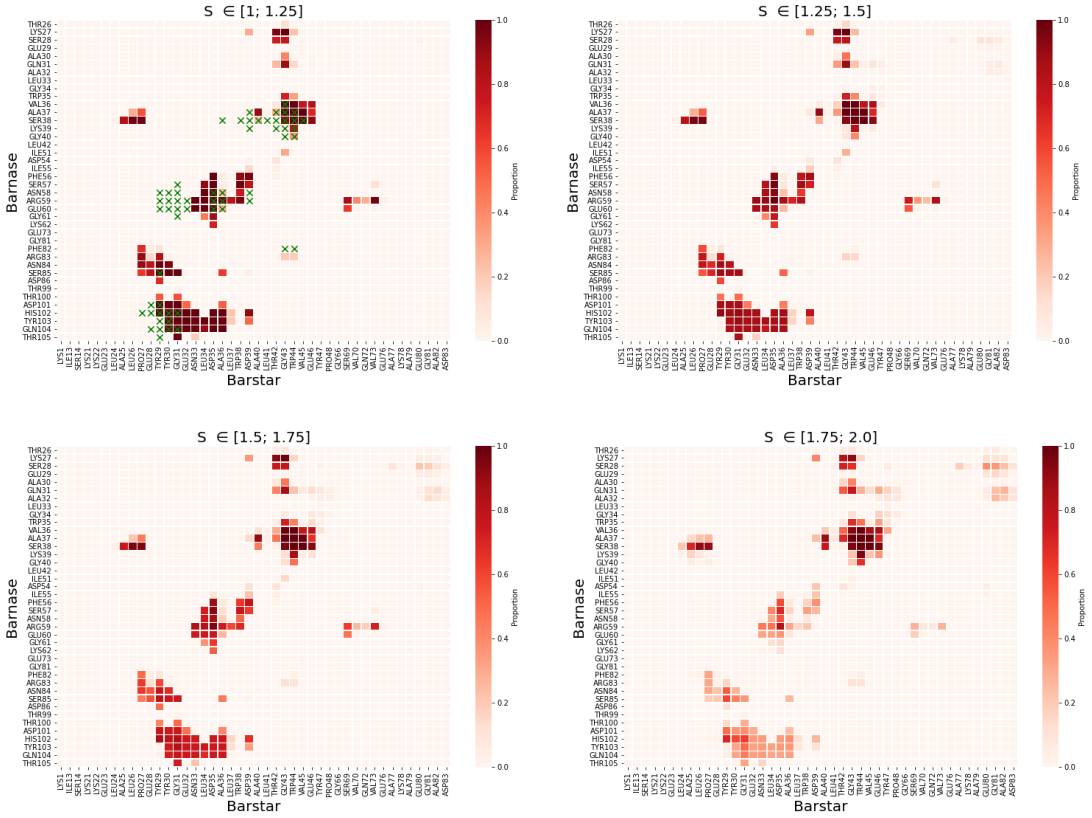


Fig. IV.3.7 *Détail des interactions rencontrées lors de l'association de la barnase (en ordonnée) et de la barstar (abscisse) pour différents intervalles de S. Les contacts observés dans la structure 1BRS sont représentés par les croix.*

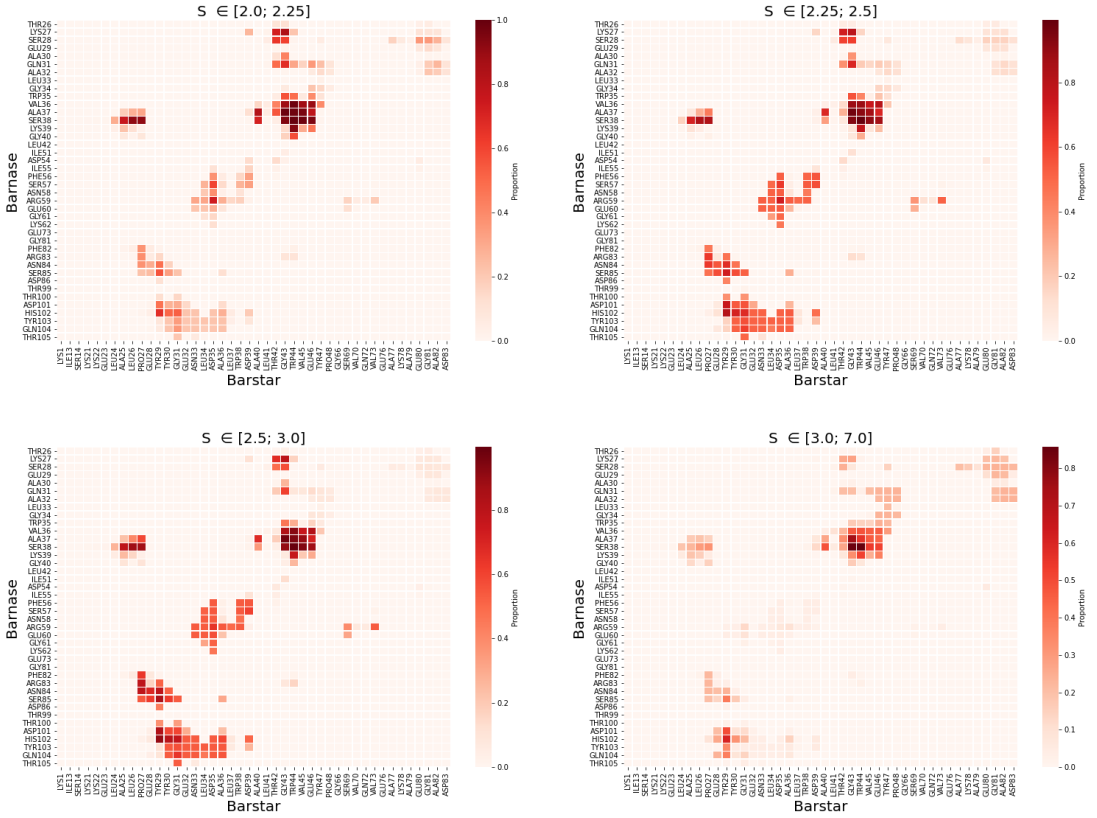


Fig. IV.3.8 *Détail des interactions rencontrées lors de l'association de la barnase (en ordonnée) et de la barstar (abscisse) pour différents intervalles de S.*

IV.3.3 CINÉTIQUE

Les constantes d'association (k_{on}) et de dissociation (k_{off}) ne peuvent pas être obtenues directement depuis la simulation de metad-BE, qui accélère les processus de transition entre les états métastables. Toutefois, à partir du paysage énergétique obtenu, il est possible d'associer une probabilité à chaque conformation assignée à un *cluster* par la méthode du WHAM[85]. Dès lors, nous obtenons une matrice de probabilités en 4D et à partir de cette dernière il est possible de construire un modèle Markovien[95], qui donne le taux de transition entre les états métastables voisins (comme détaillé dans la partie méthode).

L'un des paramètres importants à connaître est le temps de latence (*lag time*) minimum. Pour rappel ce temps correspond à la résolution temporelle du modèle markovien, donc à l'échelle de temps minimale accessible au modèle. Ce paramètre est important car il influe sur l'estimation du coefficient de diffusion. Pour des temps de latences suffisamment grand le coefficient de diffusion D devrait converger vers une valeur précise. La figure IV.3.9 représente la valeur de D pour des *lag time* compris entre 30 et 500 ps. C'est à partir de 300 ps que les variations sont minimes pour toutes les VC (figure IV.3.9).

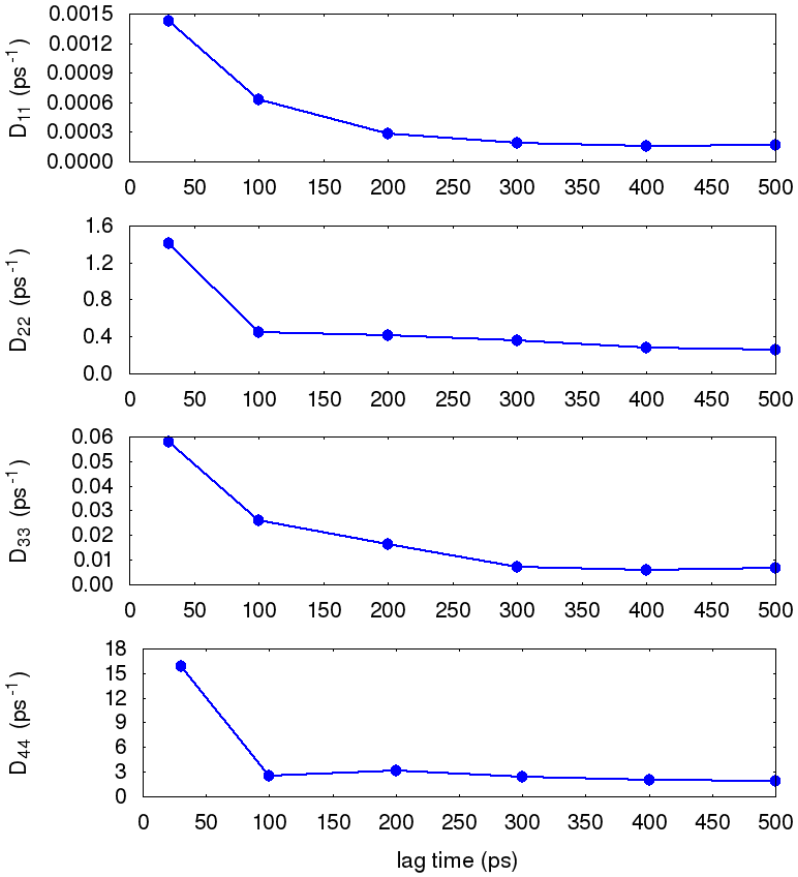


Fig. IV.3.9 *Coefficients de diffusion pour les 4 VC estimées par la méthode de maximum de vraisemblance*

En utilisant un temps de latence de 500 ps, 100 simulations de Monte-Carlo cinétique ont été réalisées. Ces simulations permettent de modéliser la diffusion "par des sauts" entre cases voisines dans une grille qui discrétise l'espace de VC, et d'observer les processus lents (qui correspondent à des évènements rares en DM). Pour chaque simulation, les trajectoires débutent à l'état dissocié (les valeurs de VC sont : 5.35, 4.21, 1.53, 234) et finissent à l'état associé (valeurs de VC 1.15, 74.7, 8.94, 126). Le temps

moyen nécessaire pour que la barnase et la barstar s'associent est de 26 ± 2 ns, ce qui correspond à un $k_{on} = 3.8 \pm 0.3 \cdot 10^7$ $M^{-1}.s^{-1}$. Du fait de la grande stabilité du complexe, le taux de dissociation (k_{off}) ne peut pas être obtenu directement. Toutefois, connaissant le ΔG_b^o et le k_{on}, il est possible de le calculer. En reprenant l'équation (II.5.4) et (II.5.5), nous obtenons l'équation (IV.3.1) :

$$\frac{k_{on}}{k_{off}} = \frac{[RL]}{[R][L]} = \frac{1}{K_d} = \frac{1}{c_o} e^{-\Delta G_b^o / kT} \tag{IV.3.1}$$

avec

- k_{on} : taux d'association
- k_{off} : taux de dissociation
- $[R][L]$: proportion des conformations sous forme libre
- $[RL]$: proportion des conformations sous forme complexée
- K_d : constante de dissociation
- c_o : concentration standard (1 M)

Notre estimation du k_{on} est de $2.3 \cdot 10^7 M^{-1} s^{-1}$. A titre de comparaison la valeur expérimentale est de $6 \cdot 10^8 s^{-1} M^{-1}$ [130]. Dans leur simulation D.E.Shaw et ses collaborateurs [144] obtiennent une prédiction plus faible du k_{on} ($4.4 \cdot 10^6 M^{-1} s^{-1}$), avec l'apport des corrections, leur estimation du k_{on} passe à $2.3 \cdot 10^7 M^{-1} s^{-1}$ (des angles de torsion et des biais introduits). Enfin dans le modèle Markovien de Noé [142] et ses collaborateurs, les prédictions de k_{on} vont de $2.8 \cdot 10^{-8}$ à $2.65 \cdot 10^{-8} M^{-1} s^{-1}$.

Concernant le k_{off} notre estimation est de $k_{off} = 4.4 \cdot 10^{-10}$ s^{-1} contre $8.0 \cdot 10^{-6}$ s^{-1} [130] pour la mesure expérimentale. On peut noter que le désaccord s'explique quantitativement avec notre surestimation de ≈ 4 kcal/mol de l'énergie libre d'association. L'estimation faite par D.E.Shaw et ses collaborateurs, quant à elle, est indiquée uniquement pour la simulation avec les corrections, sa valeur est $2.3 \cdot 10^{-7} s^{-1}$. Enfin chez Noé [142] et ses collaborateurs, les prédictions vont de $3.0 \cdot 10^{-6}$ à $1.0 \cdot 10^{-1} s^{-1}$.

Il faut noter que l'équation qui permet de calculer le coefficient k_{on} dépend du temps moyen pour que les protéines s'associent une première fois (τ), avec $n = 100$ simulations de MCC, nous obtenons une moyenne et un écart type de $\langle \tau \rangle = 26$ ns, $\sigma_\tau = 22$ ns. Un écart type similaire à la moyenne est propre à une loi de Poisson qui correspond aux évènements rares (ce qui correspond à notre cas). L'erreur type de la moyenne vaut $\epsilon_\tau = \sigma_\tau / \sqrt{n} = 2$ ns. La propagation de cette incertitude sur le taux d'association $k_{on} = 1/\tau$ donne comme erreur $\epsilon_k = |dk/d\tau| \epsilon_\tau = \epsilon_\tau / \langle \tau \rangle^2$. L'incertitude indiquée pour le k_{on} est faite sur le résultat des simulations de Monte Carlo cinétique, mais il ne faut pas oublier que l'énergie libre contribue dans l'estimation du coefficient de diffusion [95] (voir chapitre II.5), ainsi que dans l'estimation du k_{on} (équation (IV.3.1)). Les erreurs sur le calcul du paysage d'énergie libre ont un effet exponentiel sur l'estimation de k_{on}. La durée limitée de notre simulation de metad-BE conduit à une erreur importante sur le paysage d'énergie libre (une meilleure estimation sera faite après analyse des nouvelles simulations de metad-BE réalisées).

IV.3.4 Conclusion

Notre première simulation de metad-BE montre que cette méthode d'échantillonnage accéléré permet de retrouver les mêmes IPP que celles présentes dans la structure résolue par cristallographie, mais également les processus menant à l'association. En effet les résidus impliqués qui forment l'état de pré-association sont similaires à ceux décrits dans de précédentes études théoriques et également expérimentales. Ces résultats nous confortent dans l'utilisation de cette méthode pour prédire la structure des complexes protéine-protéine et pour reconstruire le chemin de transition entre la forme libre et complexé.

Concernant la valeur d'énergie libre d'association. La simulation de metad-BE donne une prédiction de ΔG_b^o de -23.4 kcal/mol après correction pour extrapoler à la concentration standard, à comparer avec la valeur expérimentale de -18.9 kcal/mol. Les données analysées avec ce premier run montrent cependant une barre d'erreur importante sur notre valeur, indiquant la nécessité d'augmenter la durée des simulations pour améliorer l'échantillonnage. Cette erreur sera précisément quantifiée et réduite en utilisant de nouvelles simulations de metad-BE (déjà effectuées). L'étude de D.E.Shaw donne une meilleure estimation de ΔG_b^o prédit[144], mais en utilisant une correction des angles de torsions pour le squelette peptidique pour mieux correspondre à la structure 1BRS, ainsi qu'au prix de ressources de calcul beaucoup plus importantes que les nôtres. La durée cumulée de nos simulations de metad-BE est mille fois plus courte (530 ns) que celle de D.E.Shaw (400 µs).

La reconstruction du paysage énergétique à partir de notre simulation de métadynamique avec biais échangés indique que la forme de l'hypersurface d'énergie libre correspond à un entonnoir (figure IV.3.1) dans lequel il n'y a pas de grande barrière d'énergie libre à franchir entre l'état dissocié et associé. Ce résultat est en accord avec d'autres reconstructions du profil d'énergie libre du système barnase barstar réalisées dans le passé[140 153 154].

Le complexe prédit comme étant le plus stable (avec notre simulation de metad-BE) est proche de la structure résolue par cristallographie (le I-RMSD vaut 0.9 Å et la fraction de contacts natifs est de 0.83). Pour ce qui est du processus d'association, celui-ci est mené dans un premier temps par des interactions électrostatiques qui engendrent la formation de contacts clés entre les deux protéines, puis à la formation du complexe. La simulation de metad-BE retrouve bien les mêmes mécanismes clés dans l'association des protéines, notamment la formation d'états de pré-association (obtenus avec l'interaction des résidus bn Arg59 et bs Asp35 ou bien bn Ser38 et bs Tpr44)[132 133 131].

Enfin concernant les estimations des constantes de cinétique, la métadynamque ne permet pas d'accéder directement à ces grandeurs. Toutefois, il est possible d'estimer les taux d'association et de dissociation en utilisant un modèle markovien paramétré à partir du paysage énergétique et de la matrice de diffusion estimée à partir de trajectoires libres. Nous retrouvons comme valeur pour la constante d'association (k_{on}) $3.8 \pm 0.3 \cdot 10^7$ $M^1 s^{-1}$ et pour la constante de dissociation (K_{off}) $4.4 \cdot 10^{-10}$ s^{-1}. Les valeurs mesurées sont de $6 \cdot 10^8 s^{-1} M^{-1}$ pour le k_{on} et $8 \cdot 10^{-6} s^{-1}$ pour le K_{off}. Notre prédiction du K_{on} est 10 fois plus grande que la valeur théorique, tandis que l'estimation du k_{off} est différente de 4 ordre de grandeurs. Pour rappel, les valeurs de k_{on} et K_{off} obtenues dépendent exponentiellement de l'estimation du paysage énergétique. Notre surestimation du ΔG_b de $\approx$ 4 kcal/mol explique la différence entres les constantes cinétiques théoriques et mesurées.

A titre de comparaison, D.E.Shaw et ses collaborateurs[144] obtiennent une prédiction plus faible du k_{on} ($4.4 \cdot 10^6 M^{-1} s^{-1}$) lorsque l'on prend en compte uniquement la simulation avec le même champ de force/solvant et sans correction sur les torsions du squelette peptidique. Avec l'apport des corrections, l'estimation du k_{on} passe à $2.3 \cdot 10^7 M^{-1} s^{-1}$. La valeur du k_{off}, quant à elle, est indiquée uniquement pour la simulation avec les corrections, sa valeur est $2.3 \cdot 10^{-7} s^{-1}$. Enfin dans le modèle markovien de Noé[142] et ses collaborateurs, les prédictions de k_{on} vont de $2.8 \cdot 10^{-8}$à$2.65 \cdot 10{-}8 M^{-1} s^{-1}$ et le k_{off} $3.0 \cdot 10^{-6}$à$1.0 \cdot 10{-}1 s^{-1}$.

Ces variations dans l'estimation du k_{on} et k_{off} dans ces différentes études montrent à quel point leur estimation est difficile à l'heure actuelle. Le choix du solvant/champ de force et l'apport de corrections sur les angles influent sur la précision des prédictions. En outre, la méthode utilisée impacte également cette prédiction. Dans notre cas, la source principale d'erreur correspond à l'estimation du paysage d'énergie libre. Pour cette raison, nous avons effectué trois nouvelles simulations de metad-BE dont les analyses seront faites prochainement et qui permettront de réduire les erreurs statistiques dans le calcul de l'énergie libre d'association ainsi que des constantes cinétiques.

CHAPITRE V

CONCEPTION DE PEPTIDES CYCLIQUES POUR L'INHIBITION DE LA DIMÉRISATION DES CASPASES

Comme expliqué dans les chapitres précédents, les peptides cycliques sont une catégorie de molécule avec un fort potentiel pharmaceutique, notamment pour perturber des IPPs[20]. Dans le cadre du projet Emergence de Sorbonne Université (effectué au début de la thèse), nous avons tenté d'utiliser les peptides cycliques comme inhibiteur pour la caspase 3. Les caspases (cysteine-aspartic protease) sont une famille de protéases participant à la régulation cellulaire en contrôlant la voie inflammatoire et apoptotique[155 156] (figure V.0.1).

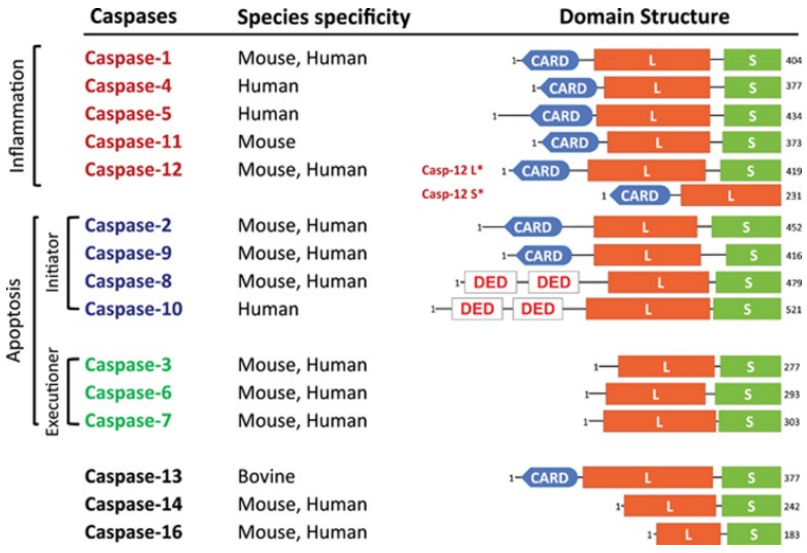
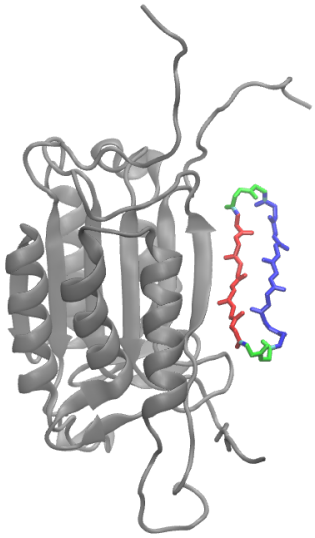
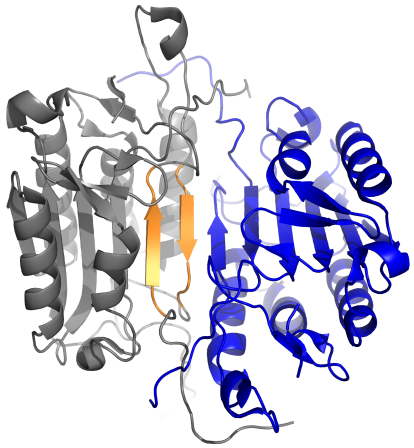


Fig. V.0.1 *Figure issu de la ref[156] Classifications des caspases. Cette famille de protéines régule l'activité cellulaire en enclenchant la mort cellulaire*



(a) *Pour bloquer l'activité catalytique de la caspase 3, nous souhaitons empêcher sa dimérisation à l'aide d'un peptide cyclique basé sur le feuillet β présent au niveau de l'interface de dimérisation de la caspase 3*



(b) *Représentation de la caspase 3. La protéine a une activité catalytique lorsque les monomères (coloriés en gris et en bleu) s'associent pour former un complexe. Le feuillet β (jaune) est la partie de la protéine qui a servi à la conception d'un peptide cyclique*

L'activité des caspases est régulée au sein de la cellule. Après traduction ces protéines sont inactives (zymogen). L'activation des caspases est initiée par un signal intracellulaire ou par l'activation des récepteurs de mort cellulaire. Suite à ce signal, les procaspases initiatrices (caspases 2,8,9 et 10) vont avoir des modifications post traductionnelles. Elles vont se cliver et s'associer en dimères, passant d'une forme inactive à une formes active qui va ensuite recruter les caspases effectrices (caspase 3, 6 et 7). Ces dernières vont à leur tour se cliver et s'associer sous forme d'hétérotetramère. Les changements conformationnels engendrés par cette association permettent ensuite aux caspsases effectives d'avoir une activité catalytique [157]. L'activité catalytique de ces protéines se fait au niveau de leur site actif composé de la séquence GLN–ALA–CYS–XXX–GLY (où XXX est Arg, Gln ou Gly). L'hydrolise de protéines par les caspases se fait uniquement au niveau des résidus aspartates et se déroule de la manière suivante.

La chaîne latérale d'une histidine va déprotoner la cystéine du site actif au niveau du groupement thiol. Puis via une attaque nucléophile, la cystéine va cliver son substrat au niveau du groupement carbonyle, libérant ainsi la partie N-terminale. Le résidu histidine revient à son état déprotoné et une liaison thioester entre la cysteine et le reste du substrat est formée. Enfin le fragment C-terminal du substrat est libéré grâce à l'hydrolyse de la liaison thioester, régénérant l'enzyme dans sa forme active.

De par l'implication des caspases dans la voie apoptotique, les maladies neurodégénératives et certains cancers, le développement d'inhibiteurs a un intérêt scientifique et médical. A l'heure actuelle, la plupart des inhibiteurs développés ciblent directement le site actif (deux inhibiteurs de ce type sont en phase de test clinique). Cependant la conservation de ce site au sein des caspases ne permet pas d'avoir une spécificité pour un seul membre de la famille. Ainsi, c'est l'activité catalytique de l'ensemble des caspases qui se trouve modifiée. Une autre approche possible pour inhiber l'activité catalytique consiste à cibler directement le site de dimérisation au niveau de l'interface, comme cela a été fait sur la caspsase-9 à l'aide de peptides stabilisés [158]. De par sa taille et sa variabilité en acides aminés, ce site de dimérisation

est spécifique entre les différentes caspases et est également un site allostérique. Ainsi en maintenant les caspases sous forme monomérique, il est possible d'inhiber leur activité catalytique sans avoir à cibler le site actif.

Nous avons utilisé une approche *in silico* s'appuyant sur la conception et l'étude de peptides cycliques comme inhibiteurs compétitifs de la caspase-3. Ces peptides ont été extraits de l'interface dans le but d'obtenir une bonne affinité et spécificité pour la caspsase-3 (figure V.0.2b). Ce feuillet β a été cyclisé en Nter Cter via l'ajout de deux acides aminés de part et d'autre du feuillet. L'objectif étant de former un coude qui stabilise le feuillet β (et donc la conformation du peptide).

Différents couples d'acides aminés permettent la formation d'un coude, soit en utilisant des acides aminés sous forme L (Asn-Gly), soit en alternant des acides aminés sous forme D et L (D-Pro suivi d'un acide aminé L) ou bien par l'ajout de groupement méthyle sur les azotes du squelette peptidique (N-méthyle) au niveau des coudes[103]. Outre la baisse de l'entropie des peptides cycliques (du fait de l'encombrement stérique), il a été montré que l'utilisation de résidus N-methylés permet d'augmenter la spécificité des peptides cycliques pour leur substrat[159] et également d'accroître la lipophilie, rendant potentiellement plus facile le franchissement des barrières lipidiques telles que la membrane plasmique[160].

V.1 PROTOCOLE

ÉCHANTILLONNAGE DE PEPTIDE CYCLIQUE

Les peptides cycliques ont été conçus à partir d'un fragment de feuillet β présent au niveau de l'interface de la dimérisation de la caspase 3. Ce feuillet a été cyclisé en Nter-Cter par l'ajout de 2 acides aminés de part et d'autre du feuillet. Le choix des acides aminés qui composent les coudes est basé sur l'étude de D. Ghosh *et al.*[103]. Dans leur travaux, les auteurs ont évalué la stabilité de feuillets β connectés par différents coudes (V.1.1). Leur principale conclusion est qu'un coude x'X' (acide aminé sous forme D N-méthylé suivit d'un acide aminé sous forme L N-méthylé) favorise le repliement en feuillet β du peptide.

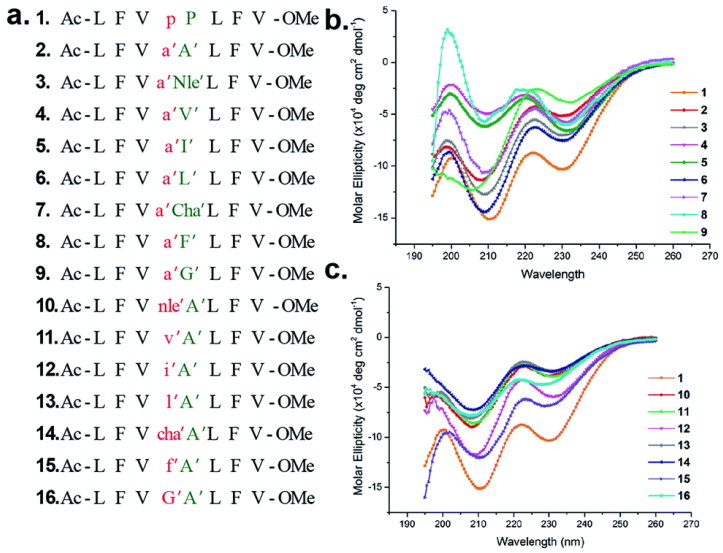


Fig. V.1.1 *Figure issue de la réf*[103] *dans laquelle la stabilité en tant que coude des différentes paires d'acides aminés est testée. a) les séquences en acides aminés utilisées. Les acides aminés sous forme D sont en minuscule et les acides aminés N-methylés sont indiqués par " ' ". b et c) spectres de dichroïsme circulaire correspondant aux différents peptides.*

Cinq peptides cycliques ont été conçus (tableau V.1.1), deux avec 17 acides aminés et trois avec 16 résidus. Afin d'évaluer la stabilité de ces peptides cycliques, nous avons échantillonné leur paysage conformationnel à l'aide de simulations de REMD de 240 ns minimum. Le protocole utilisé pour les simulations est similaire à celui présenté dans le chapitre III. Les peptides 2 et 5 ont été prolongés par la suite pour atteindre respectivement 600 et 500 ns de temps de simulation. Chaque simulation de REMD en solvant implicite GBSA est précédée par une minimisation et équilibration du système. La minimisation est réalisée par un algorithme utilisant le gradient conjugué, avec comme critère d'arrêt la convergence du système (moins de 10 kJ/mol/nm entre deux pas successifs). Si ce critère n'est pas satisfait, l'étape de minimisation s'arrête après 50 000 itérations. L'équilibration se fait en réalisant une dynamique moléculaire de 100 ps en condition NPT (conservation tout au long de la simulation du nombre d'atomes, de la pression et de la température), avec un pas d'intégration de 2 fs. Enfin une simulation de REMD est réalisée avec 8 replica.

L'analyse des conformations est faite à partir de la trajectoire à 300 K après avoir retiré les premiers 10% de la trajectoire (afin de ne pas être influencé par le point de départ de la simulation). Un partitionnement non supervisé est réalisé sur les angles ϕ et ψ avec la méthode *regular space clustering* afin de déterminer les conformations les plus probables. La durée des simulation est de 240 ns à l'exception des

peptides 2 et 4 (coude a'A')

Peptide	Sequence
1	pPCIVSMLpPDFLYAYS
2	a'A'CIVSMLa'A'DFLYAYS
3	a'L'CIVSMLa'L'DFLYAY
4	a'A'CIVSMLa'A'DFLYAY
5	NGCIVSMLNGDFLYAY

TABLE V.1.1 – Séquence des peptides cycliques conçus. La séquence est basée sur le feuillet β présent à l'interface du dimère de la caspase 3. Différents coudes sont testés afin d'évaluer leur impact sur la stabilité conformationnelle.

METADYNAMIQUE AVEC BIAS ÉCHANGÉ

Avant de réaliser un test *in vitro*, pour évaluer l'affinité de liaison de nos peptides cycliques pour la caspase 3, nous avons voulus classer nos peptides cycliques par rapport à la prédiction d'affinité de liaison envers la caspase 3. Différentes méthodes permettent de prédire la formation de complexes protéine/protéine avec un coût en calcul relativement faible[74 42 30 32]. Toutefois à notre connaissance, il n'existe pas de méthode fiable et robuste pour prédire les interactions protéine et peptide cyclique N-méthylé.

L'association d'un ligand à son récepteur est un processus complexe, l'énergie libre de liaison entre deux protéines met en jeu l'enthalpie et l'entropie du système. Omettre le solvant, ou bien représenter le récepteur et le ligand comme des corps rigides, ne permet pas de quantifier la contribution de l'entropie du système. Il est nécessaire de prendre en compte certains mécanismes clés (comme par exemple le changement de conformation du ligand et du récepteur, la désolvatation de leurs interfaces d'interactions ou bien les interactions hydrophobes et polaires). Une première simulation de metad-BE a été réalisée (voire section II.7) sur le peptide 2 avec quatre variables collectives (VC) (figure V.1.2), mais avec un protocole différent de celui présenté dans le chapitre II.5 (la simulation utilise un protocole non optimisé).

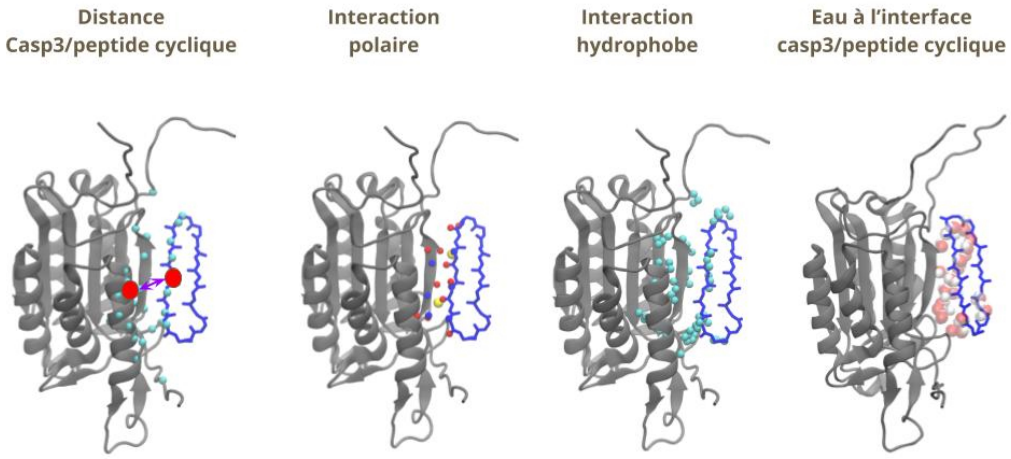


Fig. V.1.2 *Les variables collectives utilisées pour la simulation de metaBE. De gauche à droite, distance entre les centres de masse (en utilisant les carbones α) entre l'interface de dimérisation de la caspase 3 et du peptide cyclique. Contact polaire entre les atomes des résidus polaires de la caspase 3. Contacts hydrophobes entre la caspase 3 et le peptide cyclique. L'eau entre la casapse 3 et le peptide cyclique.*

La première VC est la distance entre deux centres de masse (c'est-à-dire la position moyenne calculée sur sur un groupe de carbones α), le centre de masse de l'interface de dimérisation de la caspase 3 et le centre de masse du brin β du peptide cyclique (uniquement la partie qui est sensée interagir avec la caspase 3). Le domaine de définition de cette VC va de 0.3 (état associé) à 3 nm (état dissocié). L'énergie fournie est obtenue en sommant les gaussiennes de hauteur 1 kJ/mol et d'écart type de 0.03 nm qui ont été déjà déposées aux coordonnées (de la VC) où du système. Pour cette VC, l'énergie est introduite toutes les 4 ps (pour plus de détail voir la section II.7).

Pour les interactions hydrophobes, les biais sont appliqués sur les chaînes latérales, mais uniquement sur les atomes de carbone. Cette VC comptabilise tous les contacts entre les atomes de carbones, entre les résidus hydrophobes présents sur les interface de la caspase 3 et du peptide cyclique. Ainsi le nombre de contacts est une combinatoire. Les biais sont appliqués pour un nombre de contacts compris entre 4 et 150 contacts. Les biais sont introduits toutes les 4 ps et sont modélisés par une gaussienne de hauteur 1 kJ/mol et d'écart type de 2.0 contacts qui ont été déjà été déposés aux coordonnées de VC où se trouve le système (pour plus de détail voir la section II.7).

Les contacts polaires sont comptabilisés sur les atomes d'oxygènes et d'azotes des chaînes latérales des résidus polaires. Les biais sont introduits de manière à casser/former des ponts salin et des liaisons hydrogènes. L'espace de cette VC va de 2.0 à 25 contacts. Les biais sont introduits toutes les 4 ps et sont modélisés par une gaussienne de hauteur 1 kJ/mol et d'écart type de 0.25 contacts.

Enfin la dernière variable collective, l'eau au voisinage de la caspace 3 et du peptide cyclique. Plusieurs études ont montré un ralentissement des molécules d'eau d'un facteur 3 à 5[90 91 92] au voisinage des protéines. Ainsi, la désolvatation des interfaces d'interactions des protéines nécessite de rompre les liaisons hydrogènes formées avec le solvant. Ce phénomène constitue une barrière d'énergie libre à franchir (les protéines doivent rompre les liaisons hydrogènes formées avec le solvant). Pour cette VC l'eau à l'interface correspond aux molécules d'eau distantes de moins de 1 nm des atomes utilisés par la VC contact polaire. A l'instar des contacts hydrophobes et polaires, le nombre donné est combinatoire (une même molécule d'eau peut être comptée plusieurs fois si elle est à moins de 1 nm de distance de plusieurs atomes O et N de résidus polaires). L'espace de cette VC va de 2 à 60 contacts. Les biais sont introduits toutes les 4 ps et sont modélisés par une gaussienne de hauteur 1 kJ/mol et d'écart type 0.25

contacts.

Les simulations de metad-BE ont été réalisées sur le cluster OCCIGEN du CINES avec le logiciel GROMACS 5.2 et PLUMED 2.3, le champ de force amber99[161], le solvant explicite TIP4P[145] et en utilisant le monomère de caspase 3 présent dans le fichier PDB 2j30[162], ainsi que peptide cyclique 2 (tableau V.2.1). 33432 molécules d'eau ont été ajoutées afin de solvater le système. Une minimisation, puis une équilibration (100 ps en NVT et 500 en NPT avec le thermostat Berendsen) ont été réalisées avant la phase de production. Afin de quantifier le temps de calcul nécessaire, de courtes simulations ont été réalisées. Une fois cette étape réalisée, une simulation de metad-BE a été effectuée. La durée totale par réplique est de 160 ns (soit une durée totale de 640 ns).

V.2 RÉSULTAT

ÉCHANTILLONNAGE DES PEPTIDES CYCLIQUES

Le partitionnement non supervisé sur les angles ϕ et ψ révèle des disparités dans le nombre de groupes contenant plus de 10% des effectifs (tableau V.2.1). Il va de 0 à 2 suivant la séquence des peptides. Le premier peptide possède 4 *cluster* dont un qui contient plus de 46% des conformations. Toutefois l'analyse du centroïde de ce *cluster* révèle que le peptide cyclique (figure V.2.1) a perdu le feuillet β et adopte une conformation alternative (figure V.2.1). Le peptide cyclique 2 n'a aucun *cluster* qui contient plus de 10% des effectifs. Cette faible proportion indique que ce peptide adopte des conformations diverses au cours de la simulation.

Concernant les peptides cycliques symétriques (qui ont 16 résidus), le groupe le plus peuplé pour le peptide 3 contient seulement 6.0% des conformations. Là encore, cette faible proportion indique que ce peptide adopte plusieurs conformations minoritaires au cours de la simulation. Enfin les peptides 4 et 5 possèdent 2 *clusters* qui contiennent plus de 10% des conformations (tableau V.2.1). Pour ce qui est des conformations, seul le peptide 4 conserve un feuillet β (figure V.2.2a). En outre, nous observons un décalage des résidus a'A', ces derniers font partie des brins β au lieu de former le coude.

Les résultats de ces simulations de REMD révèlent plusieurs choses. La première est que parmi les peptides à 16 résidus, le coude a'A' semble celui qui stabilise le mieux le feuillet β. En effet, le peptide 4 adopte cette structure dans plus de 30% de ses conformations. Enfin l'autre résultat est le décalage de la séquence a'A' qui se trouve au niveau des brins β. Nous pouvons supposer que l'utilisation des doubles coudes contraint le squelette peptique conduisant à ce décalage (ce qui pourrait expliquer l'absence de feuillet β dans les clusters les plus peuplés chez les autres peptides).

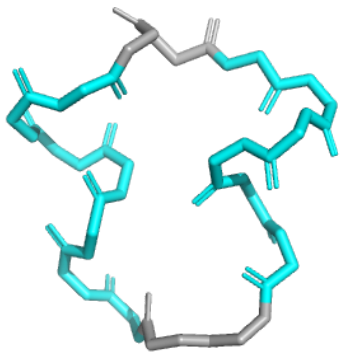
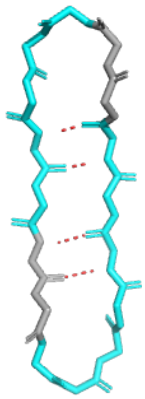
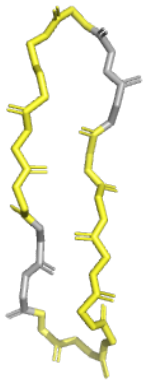


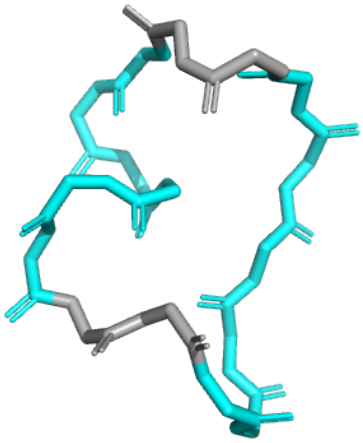
Fig. V.2.1 *Le peptide 1 possède un cluster qui contient 46% des conformations. La figure représente le squelette peptidique du centroïde de ce cluster. Les résidus qui forment les coudes sont colorés en gris.*



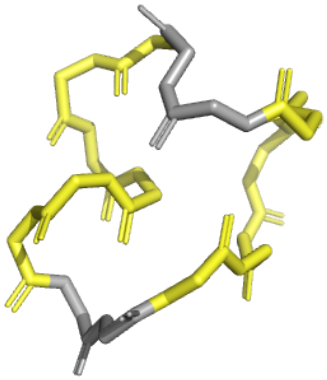
(a) *Représentation du centroïde du peptide 4 issu du cluster 1 (17.9%), les résidus qui forment le coude sont coloriés en gris.*



(b) *Représentation du centroïde du peptide 4 issu du cluster 2 (10.0%), les résidus qui forment le coude sont coloriés en gris.*



(c) *Peptide 5 cluster 1 (29.1%)*



(d) *Peptide 5 cluster 2 (12.9%)*

Fig. V.2.2 *Représentation des centroïdes issus des groupes contenant plus de 10% des conformations. Le peptide 4 conserve une structuration en feuillet β, tandis que le peptide 5 adopte une conformation différente dans ses deux clusters les plus peuplés.*

Peptide	Sequence	Temps de simulation (ns)	Clusters les plus peuplés (>5%)
1	pPCIVSMLpPDFLYAYS	240	46.0, 9.3, 9.5, 6.0
2	a'A'CIVSMLa'A'DFLYAYS	600	7.3, 5.3
3	a'L'CIVSMLa'L'DFLYAY	240	6.0
4	a'A'CIVSMLa'A'DFLYAY	500	17.9, 10.0, 5.8, 5.5
5	NGCIVSMLNGDFLYAY	240	29.1, 12.9, 6.8, 5.3

TABLE V.2.1 – Peptides cycliques échantillonnés. La séquence est basée sur le feuillet β présent à l'interface du dimère de la caspase 3. Différents coudes sont testés afin d'évaluer leur impact sur la stabilité conformationnelle. Les temps des simulations de REMD sont indiqués dans la colonne du milieu. Afin de discrétiser les conformations, un partionnement non supervisé est réalisé en utilisant l'algorithme *regular space clustering*(voir chapitre II.5). Les effectifs des groupes contenant plus de 5% des conformations sont indiqués dans la dernière colonne.

Évaluation des ressources nécessaires

Afin d'évaluer les ressources en calcul nécessaires et l'impacte des choix des VC sur le temps de production, un benchmark a été réalisé en faisant des simulations courtes. Les temps sont indiqués dans le tableau V.2.2.

Le premier résultat de ce benchmark est la différence de production entre GROMACS 4.5.5 avec PLUMED 1.3 et GROMACS 5.1.4 avec PLUMED 2.3. Les performances sont différentes pour un même nombre de CPU utilisés. GROMACS 4.5.5 avec PLUMED 1.3 (cette version a été installée par nous mêmes) produit 2.4 fois plus que GROMACS 5.1.4 avec PLUMED 2.3. Toutefois des problèmes d'instabilités du logiciel surviennent lors de la production des résultats. Pour la suite nous nous sommes focalisés sur les performances de GROMACS 5.1.4 avec PLUMED 2.3

Pour le reste, nous pouvons constater qu'à mesure que le nombre de processeurs augmente, plus le temps de production théorique augmente. Il passe de 1.76 ns/jour avec 48 coeurs à 3.78 ns/jour avec 192 CPU utilisés. En outre la variable collective de l'eau a un impact sur le temps de production. Son utilisation réduit le temps de production théorique d'un facteur 16. Cette VC comptabilise les molécules d'eau au contact de deux interfaces de protéines (ce qui revient à comparer trois listes d'atomes). Ainsi la complexité de calcul est plus élevée par rapport à une comparaison d'éléments entre deux listes d'atomes.

Au vu de ces résultats, nous avons décidé d'utiliser pour ce système la configuration GROMACS 5.1.4 avec PLUMED 2.3 et 96 CPU.

PLUMED	GROMACS	nb node	nb CPU	condition	ns/day	ns/CPU hour
2.3	5.1.4	2	48	all the CV	1.76	$1.53 \cdot 10^{-3}$
2.3	5.1.4	4	96	all the CV	2.49	$1.08 \cdot 10^{-3}$
2.3	5.1.4	4	96	without water bridge	35	$15.2 \cdot 10^{-3}$
2.3	5.1.4	8	192	all the CV	3.78	$0.82 \cdot 10^{-3}$
2.3	5.1.4	10	240	all the CV	4.29	$0.74 \cdot 10^{-3}$
1.3	4.5.5	2	48	all the CV	4.31	$3.74 \cdot 10^{-3}$

TABLE V.2.2 – Benchmark des temps de productions théoriques en fonction du nombre de coeurs et des VC utilisées. L'eau à l'interface ralentit la simulation de DM par rapport aux autres VC. De gauche à droite : version de PLUMED et de GROMACS, nombre de noeuds utilisés sur OCCIGEN, nombre de processeurs, utilisation ou non de toutes les VC, nombre de ns/jour théorique et nombre de ns/heure de CPU.

D'après le profil de l'énergie libre en fonction de la distance (entre les centres de masse des interfaces du peptide cyclique et de la caspase 3), l'association du peptide cyclique n'est pas favorable (ΔG est proche de 0 kJ/mol) (figure V.2.3). L'analyse des structures échantillonnées au cours de la simulation révèle que la VC "distance" n'est pas optimale dans la manière dont nous l'avons utilisée. Nous nous retrouvons dans une situation similaire à celle illustrée par la figure V.2.4 ; les molécules peuvent toujours interagir entre elles au delà de 1 nm.

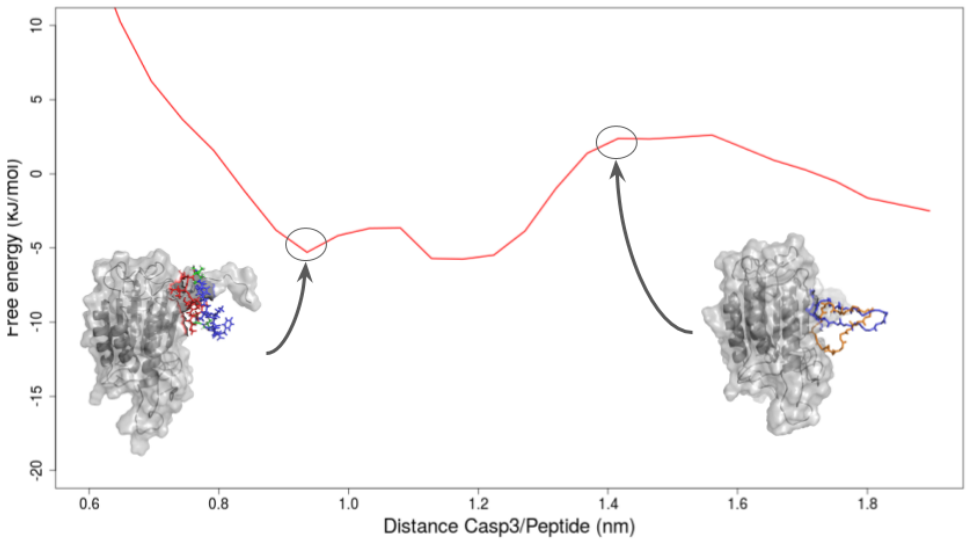


Fig. V.2.3 *Profil de l'énergie libre en fonction de la distance entre la caspase 3 et le peptide cyclique 2. Pour de grandes distances entre les centres de masses, les deux molécules peuvent interagir. Cette variable collective n'est donc pas informative sur l'état dissocié ou complexé du système*

Fig. V.2.4 *Pour une même distance entre deux centres de masses, les protéines peuvent être libres ou bien complexées si une des protéines effectue une rotation.*

V.3 CONCLUSION

Sur l'ensemble des peptides cycliques conçus, le peptide 5 semble être le plus stable, l'utilisation d'un coude a'A' stabilise le peptide qui adopte une structuration en feuillet β. Toutefois les acides aminés qui forment le coude ne sont pas ceux prévus. Ce décalage semble indiquer que l'utilisation de 2 coudes engendre d'importantes contraintes stériques qui déstabilisent le feuillet β. Afin de confirmer cette hypothèse, des simulations de REMD de peptide linéaire avec un seul coudes a'A', pP, NG et a'L' devront être faites pour étudier leur impact sur le repliement des peptides.

Concernant l'affinité de liaison de ce peptide cyclique avec la caspase 3, nous n'avons pas réalisé de mesure *in vitro*, l'incorporation d'un coude a'A' n'étant pas possible via la plateforme de chimie de l'université. Cette synthèse étant compliquée, nous avons voulu finaliser notre protocole de metadyn-BE pour classer les peptides cycliques en fonction de leur affinité de liaison vu d'un test *in vitro* avant de débuter la synthèse des meilleurs candidats.

Enfin notre simulation de metad-BE a montré que la VC "distance" n'est pas optimale. Cette dernière doit être remplacée par une autre VC, telle que la variable de chemin. Dans l'optique d'améliorer le temps de production, la VC correspondant à l'eau à l'interface a été modifiée également. Les biais ne sont plus appliqués pour désolvater les deux interfaces, mais uniquement l'interface d'une molécule. Afin de valider ce nouveau protocole, nous avons décidé de le tester sur le système barnase et barstar, qui fait référence en la matière. Les résultats de ce protocole ont été présentés dans le chapitre IV, sur ce complexe protéique bien connu qui a l'avantage d'avoir des valeurs expérimentales d'énergie libre, k_{on} et k_{off} et serq par la suite de nouveau appliqué sur les système caspase 3/peptides cycliques.

CHAPITRE VI

CONCLUSION

VI.1 DISCUSSION/CONCLUSION

L'étude des interactions protéine peptide est un domaine vaste qui a de nombreux embranchements. Au cours de la thèse nous avons essayé de développer un protocole de metad-BE pour les interactions protéine-peptide, ainsi que pour échantillonner la stabilité conformationnelle de peptides avec comme application la conception d'un peptide inhibiteur pour la caspase 3. De ces objectifs initiaux ont découlé plusieurs problématiques.

La première a été de concevoir *in silico* des peptides cycliques et d'évaluer leur stabilité conformationnelle. Plusieurs méthodes existent pour prédire les structures de peptides. Toutefois elles comportent des limitations soit en terme du nombre de résidus, dans l'utilisation d'acides aminés non standards (forme D ou bien acide aminé N-methylé) ou bien dans leur temps de calcul. Nous avons voulu mettre en place une procédure automatique pour échantillonner le paysage conformationnel de peptides avec des simulations de REMD (l'image Docker sera bientôt déployée sur les serveurs de RPBS). Les simulations de REMD convergent en moyenne en moins de 100 ns pour 5 peptides cycliques. Pour 7 des 9 peptides cycliques le centroïde du cluster le plus peuplé a un RMSD inférieur à 1.6 Å par rapport à la structure expérimentale. Ce résultat montre que les prédictions faites en utilisant la REMD sont proches des conformations de référence le tout pour un coût en calcul relativement faible. Enfin pour ce qui est de la conception de peptide inhibiteur, les résultats des REMD indiquent que le coude a'A' stabilise le mieux le feuillet β mais que cette stabilisation se fait au prix d'un décalage de la séquence a'A' qui se trouve au niveau des brins β.

Au cours de ces trois dernières années, nous avons voulu mettre au point un protocole qui vise à prédire l'affinité de liaison, les taux d'association/dissociation, mais également les mécanismes menant de l'état libre à l'état complexé. Différentes méthodes d'échantillonnage accélérée existent et à l'heure actuelle, il n'y a pas de techniques consensus pour répondre à cette problématique[77]. Nous avons voulu proposer un protocole de metad-BE qui utilise (à nos yeux) les variables collectives (VC) les plus généralistes possibles. Ceci dans le but d'être le moins système dépendant et le plus facilement adaptable d'un complexe protéique à l'autre. Le système caspase 3/peptide cyclique a montré que le choix des VC n'est pas trivial. De grandes valeurs de distance entre deux centres de masses ne garantissent pas l'absence de contact, tandis que l'eau à l'interface des deux protéines engendre des coûts de calculs élevés. Nous avons donc décidé de changer les VC utilisées et de calibrer les paramètres sur un système test connu, la barnase barstar. Notre première simulation de production a montré qu'avec la metad-BE et l'utilisation de quatre VC, il était possible de retrouver les même IPP que ceux présents dans la structure résolue par cristallographie, ainsi que les mécanismes clés menant à l'association des protéines. Pour ce qui est de l'énergie libre de liaison, notre prédiction surestime sa valeur. Cependant nos temps de simulations sont

beaucoup plus courts et nous pensons améliorer nos prédictions en prenant en compte les nouvelles simulations de metad-BE réalisées. Ces résultats seront prochainement analysés dans l'optique d'une publication. Concernant les taux d'association et de dissociation, leur estimation reste à l'heure actuelle compliquée. Il est nécessaire de construire des modèles à partir de méthodes d'échantillonnages accélérées et comme montré dans le chapitre IV la principale source d'erreur correspond à l'estimation du paysage d'énergie libre.

VI.2 PERSPECTIVE

Dans la perspective d'une seconde version de ce serveur, nous souhaiterions automatiser l'utilisation de résidus N-méthylés (pour le moment les modifications de la topologie sont faites manuellement) ainsi que modifier le potentiel de Lennard-Jones pour que les atomes de soufre des résidus cystéines (qui sont sensés former un liaison disulfure) s'attirent. D'autres axes d'améliorations sont également possibles. Nous pouvons imaginer la mise en place d'une *pipeline* qui génère des conformations à l'aide de simulations de REMD de 100 ns, puis classe les structures en fonction de leur probabilité de présence en vue de faire un criblage virtuel massif avec du *docking*. Cet échantillonage par la REMD a tout de même des limites : la première est l'utilisation du couple champ de force/solvant qui n'est pas le plus précis. L'eau n'étant pas présente physiquement, les chaînes latérales sont beaucoup plus mobiles et forment plus facilement des ponts salins/liaisons hydrogènes, ce qui peut bloquer le peptide dans des conformations non réalistes. Pour résoudre ce problème, une piste à envisager est d'effectuer une simulation de DM en solvant explicite avec les centroïdes des *clusters* les plus peuplés en utilisant le champ de force RSFF2[30].

Enfin, concernant la metad-BE, nous avons trouvé des VC généralistes qui à nos yeux permettent de capturer les évènements clés qui ont lieu lors d'une association dissociation. Ces VC n'étant pas spécifiques à un système donné, elles peuvent être théoriquement utilisées dans l'étude d'autres complexes protéine-protéine ou bien protéine-peptide. Afin de le confirmer, nous souhaitons appliquer ce protocole de metad-BE pour d'autres complexes protéiques. Ces nouvelles simulations nous permettrons d'avoir un recul sur la précision du couple champ de force/solvant sur les prédictions d'énergie libre de liaison, ainsi que d'éventuels paramètres de VC à modifier. Une fois cette phase de test réalisée, nous aimerions appliquer ce protocole dans le cadre du développement d'un peptide cyclique inhibiteur pour la caspase 3.

BIBLIOGRAPHIE

[1] N. Fujii, "D-amino acids in living higher organisms," *Origins of Life and Evolution of the Biosphere*, vol. 32, no. 2, pp. 103–127, 2002.

[2] G. Genchi, "An overview on d-amino acids," *Amino Acids*, vol. 49, no. 9, pp. 1521–1533, 2017.

[3] T. Nishikawa, "Metabolism and functional roles of endogenous d-serine in mammalian brains," *Biological and Pharmaceutical Bulletin*, vol. 28, no. 9, pp. 1561–1565, 2005.

[4] T. Furuchi and H. Homma, "Free d-aspartate in mammals," *Biological and Pharmaceutical Bulletin*, vol. 28, no. 9, pp. 1566–1570, 2005.

[5] A. Tiessen, P. Pérez-Rodríguez, and L. J. Delaye-Arredondo, "Mathematical modeling and comparison of protein size distribution in different plant, animal, fungal and microbial species reveals a negative correlation between protein size and protein number, thus providing insight into the evolution of proteomes," *BMC research notes*, vol. 5, no. 1, p. 85, 2012.

[6] K. Linderstrøm-Lang, *Proteins and enzymes*, vol. 6. Stanford university press, 1952.

[7] R. Tycko, D. P. Weliky, and A. E. Berger, "Investigation of molecular structure in solids by two-dimensional nmr exchange spectroscopy with magic angle spinning," *The Journal of chemical physics*, vol. 105, no. 18, pp. 7915–7930, 1996.

[8] G. N. Ramachandran, "Stereochemistry of polypeptide chain configurations," *J. Mol. Biol.*, vol. 7, pp. 95–99, 1963.

[9] S. A. Hollingsworth and P. A. Karplus, "A fresh look at the ramachandran plot and the occurrence of standard structures in proteins," *Biomolecular concepts*, vol. 1, no. 3-4, pp. 271–283, 2010.

[10] Y. A. Haggag, A. A. Donia, M. A. Osman, and S. A. El-Gizawy, "Peptides as drug candidates : Limitations and recent development perspectives," *Biomedical Journal*, vol. 1, p. 3, 2018.

[11] A. Kastin, *Handbook of biologically active peptides*. Academic press, 2013.

[12] M. P. Stumpf, T. Thorne, E. de Silva, R. Stewart, H. J. An, M. Lappe, and C. Wiuf, "Estimating the size of the human interactome," *Proceedings of the National Academy of Sciences*, vol. 105, no. 19, pp. 6959–6964, 2008.

[13] R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. O'Donnell, G. Leung, R. McAdam, *et al.*, "The biogrid interaction database : 2019 update," *Nucleic acids research*, vol. 47, no. D1, pp. D529–D541, 2019.

[14] D. E. Scott, A. R. Bayly, C. Abell, and J. Skidmore, "Small molecules, big targets : drug discovery faces the protein–protein interaction challenge," *Nature Reviews Drug Discovery*, vol. 15, no. 8, p. 533, 2016.

[15] T. A. Larsen, A. J. Olson, and D. S. Goodsell, "Morphology of protein–protein interfaces," *Structure*, vol. 6, no. 4, pp. 421–427, 1998.

[16] T. L. Blundell, B. L. Sibanda, R. W. Montalvão, S. Brewerton, V. Chelliah, C. L. Worth, N. J. Harmer, O. Davies, and D. Burke, "Structural biology and bioinformatics in drug design : opportunities and challenges for target identification and lead discovery," *Philosophical Transactions of the Royal Society B : Biological Sciences*, vol. 361, no. 1467, pp. 413–423, 2006.

[17] A. E. Modell, S. L. Blosser, and P. S. Arora, "Systematic targeting of protein–protein interactions," *Trends in pharmacological sciences*, vol. 37, no. 8, pp. 702–713, 2016.

[18] H. Ruffner, A. Bauer, and T. Bouwmeester, "Human protein–protein interaction networks and the value for drug discovery," *Drug discovery today*, vol. 12, no. 17-18, pp. 709–716, 2007.

[19] L. Mabonga and A. P. Kappo, "Protein-protein interaction modulators : advances, successes and remaining challenges," *Biophysical reviews*, pp. 1–23, 2019.

[20] D. J. Craik, D. P. Fairlie, S. Liras, and D. Price, "The future of peptide-based drugs," *Chem. Biol. Drug Des.*, vol. 81, no. 1, pp. 136–147, 2013.

[21] D. Bojadzic and P. Buchwald, "Toward small-molecule inhibition of protein–protein interactions : General aspects and recent progress in targeting costimulatory and coinhibitory (immune checkpoint) interactions," *Current topics in medicinal chemistry*, vol. 18, no. 8, pp. 674–699, 2018.

[22] G. L. Verdine and L. D. Walensky, "The challenge of drugging undruggable targets in cancer : lessons learned from targeting bcl-2 family members," *Clinical cancer research*, vol. 13, no. 24, pp. 7264–7270, 2007.

[23] A. P. Russ and S. Lampel, "The druggable genome : an update.," *Drug discovery today*, vol. 10, no. 23-24, p. 1607, 2005.

[24] A. L. Hopkins and C. R. Groom, "The druggable genome," *Nature reviews Drug discovery*, vol. 1, no. 9, pp. 727–730, 2002.

[25] J. Wang, T. Yin, X. Xiao, D. He, Z. Xue, X. Jiang, and Y. Wang, "Strapep : a structure database of bioactive peptides," *Database*, vol. 2018, 2018.

[26] J. Koehnke, J. Naismith, and W. A. Van der Donk, *Cyclic Peptides : From Bioorganic Synthesis to Applications*, vol. 6. Royal Society of Chemistry, 2017.

[27] M. A. Abdalla and L. J. McGaw, "Natural cyclic peptides as an attractive modality for therapeutics : a mini review," *Molecules*, vol. 23, no. 8, p. 2080, 2018.

[28] A. Zorzi, K. Deyle, and C. Heinis, "Cyclic peptide therapeutics : past, present and future," *Current opinion in chemical biology*, vol. 38, pp. 24–29, 2017.

[29] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000.

[30] H. Geng, F. Jiang, and Y.-D. Wu, "Accurate structure prediction and conformational analysis of cyclic peptides with residue-specific force fields," *The journal of physical chemistry letters*, vol. 7, no. 10, pp. 1805–1810, 2016.

[31] J. Beaufays, L. Lins, A. Thomas, and R. Brasseur, "In silico predictions of 3d structures of linear and cyclic peptides with natural and non-proteinogenic residues," *Journal of Peptide Science*, vol. 18, no. 1, pp. 17–24, 2012.

[32] G. Bhardwaj, V. K. Mulligan, C. D. Bahl, J. M. Gilmore, P. J. Harvey, O. Cheneval, G. W. Buchko, S. V. Pulavarti, Q. Kaas, A. Eletsky, *et al.*, "Accurate de novo design of hyperstable constrained peptides," *Nature*, vol. 538, no. 7625, pp. 329–335, 2016.

[33] M. Jusot, D. Stratmann, M. Vaisset, J. Chomilier, and J. Cortés, "Exhaustive exploration of the conformational landscape of small cyclic peptides using a robotics approach," *Journal of chemical information and modeling*, vol. 58, no. 11, pp. 2355–2368, 2018.

[34] I.-J. Chen and N. Foloppe, "Tackling the conformational sampling of larger flexible compounds and macrocycles in pharmacology and drug discovery," *Bioorganic & medicinal chemistry*, vol. 21, no. 24, pp. 7898–7920, 2013.

[35] D. J. Mandell, E. A. Coutsias, and T. Kortemme, "Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling," *Nature methods*, vol. 6, no. 8, pp. 551–552, 2009.

[36] Y. Zhang, "I-tasser server for protein 3d structure prediction," *BMC bioinformatics*, vol. 9, no. 1, p. 40, 2008.

[37] S. M. McHugh, J. R. Rogers, S. A. Solomon, H. Yu, and Y.-S. Lin, "Computational methods to design cyclic peptides," *Current opinion in chemical biology*, vol. 34, pp. 95–102, 2016.

[38] G. Weng, J. Gao, Z. Wang, E. Wang, X. Hu, X. Yao, D. Cao, and T. Hou, "Comprehensive evaluation of fourteen docking programs on protein–peptide complexes," *Journal of Chemical Theory and Computation*, vol. 16, no. 6, pp. 3959–3969, 2020.

[39] Y. Zhang and M. F. Sanner, "Autodock crankpep : combining folding and docking to predict protein–peptide complexes," *Bioinformatics*, vol. 35, no. 24, pp. 5121–5127, 2019.

[40] P. L. Kastritis, I. H. Moal, H. Hwang, Z. Weng, P. A. Bates, A. M. Bonvin, and J. Janin, "A structure-based benchmark for protein–protein binding affinity," *Protein Science*, vol. 20, no. 3, pp. 482–491, 2011.

[41] A. Vangone and A. M. Bonvin, "Contacts-based prediction of binding affinity in protein–protein complexes," *elife*, vol. 4, p. e07454, 2015.

[42] T. Vreven, I. H. Moal, A. Vangone, B. G. Pierce, P. L. Kastritis, M. Torchala, R. Chaleil, B. Jiménez-García, P. A. Bates, J. Fernandez-Recio, *et al.*, "Updates to the integrated protein–protein interaction benchmarks : docking benchmark version 5 and affinity benchmark version 2," *Journal of molecular biology*, vol. 427, no. 19, pp. 3031–3041, 2015.

[43] P. L. Kastritis and A. M. Bonvin, "On the binding affinity of macromolecular interactions : daring to ask why proteins interact," *Journal of The Royal Society Interface*, vol. 10, no. 79, p. 20120835, 2013.

[44] X. Du, Y. Li, Y.-L. Xia, S.-M. Ai, J. Liang, P. Sang, X.-L. Ji, and S.-Q. Liu, "Insights into protein–ligand interactions : mechanisms, models, and methods," *International journal of molecular sciences*, vol. 17, no. 2, p. 144, 2016.

[45] H. J. Berendsen, D. van der Spoel, and R. van Drunen, "Gromacs : a message-passing parallel molecular dynamics implementation," *Computer physics communications*, vol. 91, no. 1-3, pp. 43–56, 1995.

[46] L. Verlet, "Computer" experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules," *Physical review*, vol. 159, no. 1, p. 98, 1967.

[47] M. Kouza, "Numerical simulation of folding and unfolding of proteins," 08 2013.

[48] R. B. Best, N.-V. Buchete, and G. Hummer, "Are current molecular dynamics force fields too helical?," *Biophysical journal*, vol. 95, no. 1, pp. L07–L09, 2008.

[49] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, *et al.*, "A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations," *Journal of computational chemistry*, vol. 24, no. 16, pp. 1999–2012, 2003.

[50] A. D. MacKerell Jr, D. Bashford, M. Bellott, R. L. Dunbrack Jr, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, *et al.*, "All-atom empirical potential for molecular modeling and dynamics studies of proteins," *The journal of physical chemistry B*, vol. 102, no. 18, pp. 3586–3616, 1998.

[51] A. D. MacKerell Jr, M. Feig, and C. L. Brooks, "Improved treatment of the protein backbone in empirical force fields," *Journal of the American Chemical Society*, vol. 126, no. 3, pp. 698–699, 2004.

[52] 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," *The Journal of Physical Chemistry B*, vol. 105, no. 28, pp. 6474–6487, 2001.

[53] W. R. Scott, P. H. Hünenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Krüger, and W. F. van Gunsteren, "The gromos biomolecular simulation program package," *The Journal of Physical Chemistry A*, vol. 103, no. 19, pp. 3596–3607, 1999.

[54] M.-C. Bellissent-Funel, A. Hassanali, M. Havenith, R. Henchman, P. Pohl, F. Sterpone, D. van der Spoel, Y. Xu, and A. E. Garcia, "Water determines the structure and dynamics of proteins," *Chemical reviews*, vol. 116, no. 13, pp. 7673–7697, 2016.

[55] H. Nguyen, D. R. Roe, and C. Simmerling, "Improved generalized born solvent model parameters for protein simulations," *Journal of chemical theory and computation*, vol. 9, no. 4, pp. 2020–2034, 2013.

[56] W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson, "Semianalytical treatment of solvation for molecular mechanics and dynamics," *Journal of the American Chemical Society*, vol. 112, no. 16, pp. 6127–6129, 1990.

[57] D. A. Case, T. E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. M. Merz Jr, A. Onufriev, C. Simmerling, B. Wang, and R. J. Woods, "The amber biomolecular simulation programs," *Journal of computational chemistry*, vol. 26, no. 16, pp. 1668–1688, 2005.

[58] G. D. Hawkins, C. J. Cramer, and D. G. Truhlar, "Pairwise solute descreening of solute charges from a dielectric medium," *Chemical Physics Letters*, vol. 246, no. 1-2, pp. 122–129, 1995.

[59] 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 Bioinformatics*, vol. 55, no. 2, pp. 383–394, 2004.

[60] J. Mongan, C. Simmerling, J. A. McCammon, D. A. Case, and A. Onufriev, "Generalized born model with a simple, robust molecular volume correction," *Journal of chemical theory and computation*, vol. 3, no. 1, pp. 156–169, 2007.

[61] I. Maffucci and A. Contini, "An updated test of amber force fields and implicit solvent models in predicting the secondary structure of helical, β-hairpin, and intrinsically disordered peptides," *Journal of chemical theory and computation*, vol. 12, no. 2, pp. 714–727, 2016.

[62] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chemical physics letters*, vol. 314, no. 1-2, pp. 141–151, 1999.

[63] U. H. Hansmann and Y. Okamoto, "Prediction of peptide conformation by multicanonical algorithm : New approach to the multiple-minima problem," *Journal of computational chemistry*, vol. 14, no. 11, pp. 1333–1338, 1993.

[64] R. Qi, Y. Luo, B. Ma, R. Nussinov, and G. Wei, "Conformational distribution and α-helix to β-sheet transition of human amylin fragment dimer," *Biomacromolecules*, vol. 15, no. 1, pp. 122–131, 2013.

[65] W. Humphrey, A. Dalke, and K. Schulten, "VMD – Visual Molecular Dynamics," *Journal of Molecular Graphics*, vol. 14, pp. 33–38, 1996.

[66] Schrödinger, LLC, "The PyMOL molecular graphics system, version 1.8." November 2015.

[67] J. Corzo, "Time, the forgotten dimension of ligand binding teaching," *Biochemistry and Molecular Biology Education*, vol. 34, no. 6, pp. 413–416, 2006.

[68] R. C. Mohs and N. H. Greig, "Drug discovery and development : Role of basic biological research," *Alzheimer's & Dementia : Translational Research & Clinical Interventions*, vol. 3, no. 4, pp. 651–657, 2017.

[69] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin, "A geometric approach to macromolecule-ligand interactions," *Journal of molecular biology*, vol. 161, no. 2, pp. 269–288, 1982.

[70] A. A. Bogan and K. S. Thorn, "Anatomy of hot spots in protein interfaces," *Journal of molecular biology*, vol. 280, no. 1, pp. 1–9, 1998.

[71] J. Liang, C. Woodward, and H. Edelsbrunner, "Anatomy of protein pockets and cavities : measurement of binding site geometry and implications for ligand design," *Protein science*, vol. 7, no. 9, pp. 1884–1897, 1998.

[72] I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg, "Dip : the database of interacting proteins," *Nucleic acids research*, vol. 28, no. 1, pp. 289–291, 2000.

[73] K. Yugandhar and M. M. Gromiha, "Computational approaches for predicting binding partners, interface residues, and binding affinity of protein–protein complexes," in *Prediction of Protein Secondary Structure*, pp. 237–253, Springer, 2017.

[74] J. Wereszczynski and J. A. McCammon, "Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition," *Quarterly reviews of biophysics*, vol. 45, no. 1, p. 1, 2012.

[75] M. C. Zwier and L. T. Chong, "Reaching biological timescales with all-atom molecular dynamics simulations," *Current opinion in pharmacology*, vol. 10, no. 6, pp. 745–752, 2010.

[76] R. A. Copeland, "Drug–target interaction kinetics : underutilized in drug optimization?," 2016.

[77] C. Camilloni and F. Pietrucci, "Advanced simulation techniques for the thermodynamic and kinetic characterization of biological systems," *Advances in Physics : X*, vol. 3, no. 1, p. 1477531, 2018.

[78] D. B. Kokh, M. Amaral, J. Bomke, U. Grader, D. Musil, H.-P. Buchstaller, M. K. Dreyer, M. Frech, M. Lowinski, F. Vallee, *et al.*, "Estimation of drug-target residence times by tau-random acceleration molecular dynamics simulations," *Journal of Chemical Theory and Computation*, vol. 14, no. 7, pp. 3859–3869, 2018.

[79] A. Laio and M. Parrinello, "Escaping free-energy minima," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12562–12566, 2002.

[80] J. Kästner, "Umbrella sampling," *Wiley Interdisciplinary Reviews : Computational Molecular Science*, vol. 1, no. 6, pp. 932–942, 2011.

[81] S. Izrailev, S. Stepaniants, B. Isralewitz, D. Kosztin, H. Lu, F. Molnar, W. Wriggers, and K. Schulten, "Steered molecular dynamics," in *Computational molecular dynamics : challenges, methods, ideas*, pp. 39–65, Springer, 1999.

[82] V. Van Speybroeck, K. De Wispelaere, J. Van der Mynsbrugge, M. Vandichel, K. Hemelsoet, and M. Waroquier, "First principle chemical kinetics in zeolites : the methanol-to-olefin process as a case study," *Chemical Society Reviews*, vol. 43, no. 21, pp. 7326–7357, 2014.

[83] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, "Plumed 2 : New feathers for an old bird," *Computer Physics Communications*, vol. 185, no. 2, pp. 604–613, 2014.

[84] T. Giorgino, A. Laio, and A. Rodriguez, "Metagui 3 : A graphical user interface for choosing the collective variables in molecular dynamics simulations," *Computer Physics Communications*, vol. 217, pp. 204–209, 2017.

[85] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method," *Journal of computational chemistry*, vol. 13, no. 8, pp. 1011–1021, 1992.

[86] S. Piana and A. Laio, "A bias-exchange approach to protein folding," *The journal of physical chemistry B*, vol. 111, no. 17, pp. 4553–4559, 2007.

[87] F. Pietrucci, F. Marinelli, P. Carloni, and A. Laio, "Substrate binding mechanism of hiv-1 protease from explicit-solvent atomistic simulations," *Journal of the American Chemical Society*, vol. 131, no. 33, pp. 11811–11818, 2009.

[88] F. B. Baghal, X. Biarnes, F. Pietrucci, A. Laio, and F. Affinito, "Simulation of amyloid nucleation with bias-exchange metadynamics," *Biophysical Journal*, vol. 102, no. 3, p. 242a, 2012.

[89] T. Feng and K. Barakat, "Molecular dynamics simulation and prediction of druggable binding sites," in *Computational Drug Discovery and Design*, pp. 87–103, Springer, 2018.

[90] D. S. Grebenkov, Y. A. Goddard, G. Diakova, J.-P. Korb, and R. G. Bryant, "Dimensionality of diffusive exploration at the protein interface in solution," *The Journal of Physical Chemistry B*, vol. 113, no. 40, pp. 13347–13356, 2009.

[91] M. Marchi, F. Sterpone, and M. Ceccarelli, "Water rotational relaxation and diffusion in hydrated lysozyme," *Journal of the American Chemical Society*, vol. 124, no. 23, pp. 6787–6791, 2002.

[92] F. Sterpone, G. Stirnemann, and D. Laage, "Magnitude and molecular origin of water slowdown next to a protein," *Journal of the American Chemical Society*, vol. 134, no. 9, pp. 4116–4119, 2012.

[93] D. Branduardi, F. L. Gervasio, and M. Parrinello, "From a to b in free energy space," *The Journal of chemical physics*, vol. 126, no. 5, p. 054103, 2007.

[94] N. Saleh, P. Ibrahim, G. Saladino, F. L. Gervasio, and T. Clark, "An efficient metadynamics-based protocol to model the binding affinity and the transition state ensemble of g-protein-coupled receptor ligands," *Journal of chemical information and modeling*, vol. 57, no. 5, pp. 1210–1217, 2017.

[95] F. Marinelli, F. Pietrucci, A. Laio, and S. Piana, "A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations," *PLoS computational biology*, vol. 5, no. 8, 2009.

[96] P. Tiwary and M. Parrinello, "From metadynamics to dynamics," *Physical review letters*, vol. 111, no. 23, p. 230602, 2013.

[97] G. Hummer, "Position-dependent diffusion coefficients and free energies from bayesian analysis of equilibrium and replica molecular dynamics simulations," *New Journal of Physics*, vol. 7, no. 1, p. 34, 2005.

[98] D. Bicout and A. Szabo, "Electron transfer reaction dynamics in non-debye solvents," *The Journal of chemical physics*, vol. 109, no. 6, pp. 2325–2338, 1998.

[99] J. Aldrich *et al.*, "Ra fisher and the making of maximum likelihood 1912-1922," *Statistical science*, vol. 12, no. 3, pp. 162–176, 1997.

[100] S. Singh, H. Singh, A. Tuknait, K. Chaudhary, B. Singh, S. Kumaran, and G. P. Raghava, "Pepstrmod : structure prediction of peptides containing natural, non-natural and modified residues," *Biology direct*, vol. 10, no. 1, p. 73, 2015.

[101] V. A. Voelz, K. A. Dill, and I. Chorny, "Peptoid conformational free energy landscapes from implicit-solvent molecular simulations in amber," *Peptide Science*, vol. 96, no. 5, pp. 639–650, 2011.

[102] P. Hosseinzadeh, G. Bhardwaj, V. K. Mulligan, M. D. Shortridge, T. W. Craven, F. Pardo-Avila, S. A. Rettie, D. E. Kim, D.-A. Silva, Y. M. Ibrahim, *et al.*, "Comprehensive computational design of ordered peptide macrocycles," *Science*, vol. 358, no. 6369, pp. 1461–1466, 2017.

[103] D. Ghosh, P. Lahiri, H. Verma, S. Mukherjee, and J. Chatterjee, "Engineering β-sheets employing n-methylated heterochiral amino acids," *Chemical Science*, vol. 7, no. 8, pp. 5212–5218, 2016.

[104] Q. Wang, A. A. Canutescu, and R. L. Dunbrack Jr, "Scwrl and molide : computer programs for side-chain conformation prediction and homology modeling," *Nature protocols*, vol. 3, no. 12, p. 1832, 2008.

[105] R. Duke, T. Giese, H. Gohlke, A. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. Lee, *et al.*, "Ambertools 16," *University of California, San Francisco*, 2016.

[106] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, *et al.*, "Biopython : freely available python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.

[107] T. Hamelryck and B. Manderick, "Pdb file parser and structure class implemented in python," *Bioinformatics*, vol. 19, no. 17, pp. 2308–2310, 2003.

[108] M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "Gromacs : High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1, pp. 19–25, 2015.

[109] E. Vanquelef, S. Simon, G. Marquant, E. Garcia, G. Klimerak, J. C. Delepine, P. Cieplak, and F.-Y. Dupradeau, "Red server : a web service for deriving resp and esp charges and building force field libraries for new molecules and molecular fragments," *Nucleic acids research*, vol. 39, no. suppl_2, pp. W511–W517, 2011.

[110] D. J. Earl and M. W. Deem, "Parallel tempering : Theory, applications, and new perspectives," *Physical Chemistry Chemical Physics*, vol. 7, no. 23, pp. 3910–3916, 2005.

[111] H. B. Curry, "The method of steepest descent for non-linear minimization problems," *Quarterly of Applied Mathematics*, vol. 2, no. 3, pp. 258–261, 1944.

[112] M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, "PyEMMA 2 : A Software Package for Estimation, Validation, and Analysis of Markov Models," *Journal of Chemical Theory and Computation*, vol. 11, pp. 5525–5542, Oct. 2015.

[113] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA, 1967.

[114] J.-H. Prinz, H. Wu, M. Sarich, B. Keller, M. Senne, M. Held, J. D. Chodera, C. Schütte, and F. Noé, "Markov models of molecular kinetics : Generation and validation," *The Journal of chemical physics*, vol. 134, no. 17, p. 174105, 2011.

[115] A. Patriksson and D. van der Spoel, "A temperature predictor for parallel tempering simulations," *Physical Chemistry Chemical Physics*, vol. 10, no. 15, pp. 2073–2077, 2008.

[116] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[117] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society : Series B (Methodological)*, vol. 28, no. 1, pp. 131–142, 1966.

[118] I. Dagan, L. Lee, and F. Pereira, "Similarity-based methods for word sense disambiguation," *arXiv preprint cmp-lg/9708010*, 1997.

[119] D. Sindhikara, Y. Meng, and A. E. Roitberg, "Exchange frequency in replica exchange molecular dynamics," *The Journal of chemical physics*, vol. 128, no. 2, p. 01B609, 2008.

[120] R. Qi, G. Wei, B. Ma, and R. Nussinov, "Replica exchange molecular dynamics : A practical application protocol with solutions to common problems and a peptide aggregation and self-assembly example," in *Peptide Self-Assembly*, pp. 101–119, Springer, 2018.

[121] M. Cecchini, F. Rao, M. Seeber, and A. Caflisch, "Replica exchange molecular dynamics simulations of amyloid peptide aggregation," *The Journal of chemical physics*, vol. 121, no. 21, pp. 10748–10756, 2004.

[122] A. E. Wakefield, W. M. Wuest, and V. A. Voelz, "Molecular simulation of conformational preorganization in cyclic rgd peptides," *Journal of Chemical Information and Modeling*, vol. 55, no. 4, pp. 806–813, 2015.

[123] A. Okur, L. Wickstrom, and C. Simmerling, "Evaluation of salt bridge structure and energetics in peptides using explicit, implicit, and hybrid solvation models," *Journal of Chemical Theory and Computation*, vol. 4, no. 3, pp. 488–498, 2008.

[124] R. Zhou and B. J. Berne, "Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water?," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12777–12782, 2002.

[125] J. W. Pitera and W. Swope, "Understanding folding and design : Replica-exchange simulations of"trp-cage"miniproteins," *Proceedings of the National Academy of Sciences*, vol. 100, no. 13, pp. 7587–7592, 2003.

[126] R. Zhou, "Free energy landscape of protein folding in water : explicit vs. implicit solvent," *Proteins : Structure, Function, and Bioinformatics*, vol. 53, no. 2, pp. 148–161, 2003.

[127] J. Leszczynski and M. K. Shukla, *Practical aspects of computational chemistry*. Springer, 2012.

[128] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the amber ff99sb protein force field," *Proteins : Structure, Function, and Bioinformatics*, vol. 78, no. 8, pp. 1950–1958, 2010.

[129] F. Rao and A. Caflisch, "Replica exchange molecular dynamics simulations of reversible folding," *The Journal of chemical physics*, vol. 119, no. 7, pp. 4035–4042, 2003.

[130] G. Schreiber and A. R. Fersht, "Interaction of barnase with its polypeptide inhibitor barstar studied by protein engineering," *Biochemistry*, vol. 32, no. 19, pp. 5145–5150, 1993.

[131] G. Schreiber, C. Frisch, and A. R. Fersht, "The role of glu73 of barnase in catalysis and the binding of barstar," *Journal of molecular biology*, vol. 270, no. 1, pp. 111–122, 1997.

[132] Y. Urakubo, T. Ikura, and N. Ito, "Crystal structural analysis of protein–protein interactions drastically destabilized by a single mutation," *Protein Science*, vol. 17, no. 6, pp. 1055–1065, 2008.

[133] G. Schreiber and A. R. Fersht, "Energetics of protein-protein interactions : Analysis ofthe barnase-barstar interface by single mutations and double mutant cycles," *Journal of molecular biology*, vol. 248, no. 2, pp. 478–486, 1995.

[134] R. H. Folmer, "Drug target residence time : a misleading concept," *Drug Discovery Today*, vol. 23, no. 1, pp. 12–16, 2018.

[135] J. R. Lakowicz, *Principles of fluorescence spectroscopy*. Springer science & business media, 2013.

[136] J. A. Dean, *Analytical chemistry handbook*, vol. 1.

[137] C. R. Clark, "A stopped-flow kinetics experiment for advanced undergraduate laboratories : Formation of iron (iii) thiocyannate," *JChEd*, vol. 74, no. 10, p. 1214, 1997.

[138] A. S. Saglam and L. T. Chong, "Protein–protein binding pathways and calculations of rate constants using fully-continuous, explicit-solvent simulations," *Chemical science*, vol. 10, no. 8, pp. 2360–2372, 2019.

[139] R. W. Hartley, "Barnase-barstar interaction.," *Methods in enzymology*, vol. 341, p. 599, 2001.

[140] L. Wang, S. W. Siu, W. Gu, and V. Helms, "Downhill binding energy surface of the barnase–barstar complex," *Biopolymers*, vol. 93, no. 11, pp. 977–985, 2010.

[141] M. Hoefling and K. E. Gottschalk, "Barnase–barstar : From first encounter to final complex," *Journal of structural biology*, vol. 171, no. 1, pp. 52–63, 2010.

[142] N. Plattner, S. Doerr, G. De Fabritiis, and F. Noé, "Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and markov modelling," *Nature chemistry*, vol. 9, no. 10, p. 1005, 2017.

[143] D. Suh, S. Jo, W. Jiang, C. Chipot, and B. Roux, "String method for protein–protein binding free-energy calculations," *Journal of Chemical Theory and Computation*, vol. 15, no. 11, pp. 5829–5844, 2019.

[144] A. C. Pan, D. Jacobson, K. Yatsenko, D. Sritharan, T. M. Weinreich, and D. E. Shaw, "Atomic-level characterization of protein–protein association," *Proceedings of the National Academy of Sciences*, vol. 116, no. 10, pp. 4244–4249, 2019.

[145] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *The Journal of chemical physics*, vol. 79, no. 2, pp. 926–935, 1983.

[146] A. M. Buckle, G. Schreiber, and A. R. Fersht, "Protein-protein recognition : Crystal structural analysis of a barnase-barstar complex at 2.0-. ang. resolution," *Biochemistry*, vol. 33, no. 30, pp. 8878–8889, 1994.

[147] H. J. Berendsen, J. v. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *The Journal of chemical physics*, vol. 81, no. 8, pp. 3684–3690, 1984.

[148] M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals : A new molecular dynamics method," *Journal of Applied physics*, vol. 52, no. 12, pp. 7182–7190, 1981.

[149] P. Tiwary, J. Mondal, and B. J. Berne, "How and when does an anticancer drug leave its binding site?," *Science advances*, vol. 3, no. 5, p. e1700014, 2017.

[150] Z. Zhang and H. S. Chan, "Transition paths, diffusive processes, and preequilibria of protein folding," *Proceedings of the National Academy of Sciences*, vol. 109, no. 51, pp. 20919–20924, 2012.

[151] G. Schreiber and A. R. Fersht, "Rapid, electrostatically assisted association of proteins," *Nature structural biology*, vol. 3, no. 5, pp. 427–431, 1996.

[152] C. Frisch, A. R. Fersht, and G. Schreiber, "Experimental assignment of the structure of the transition state for the association of barnase and barstar," *Journal of molecular biology*, vol. 308, no. 1, pp. 69–77, 2001.

[153] K. Moritsugu, T. Terada, and A. Kidera, "Energy landscape of all-atom protein-protein interactions revealed by multiscale enhanced sampling," *PLOS Comput Biol*, vol. 10, no. 10, p. e1003901, 2014.

[154] A. Spaar, C. Dammer, R. R. Gabdoulline, R. C. Wade, and V. Helms, "Diffusional encounter of barnase and barstar," *Biophysical journal*, vol. 90, no. 6, pp. 1913–1924, 2006.

[155] D. R. McIlwain, T. Berger, and T. W. Mak, "Caspase functions in cell death and disease," *Cold Spring Harbor perspectives in biology*, vol. 5, no. 4, p. a008656, 2013.

[156] S. Shalini, L. Dorstyn, S. Dawar, and S. Kumar, "Old, new and emerging functions of caspases," *Cell Death & Differentiation*, vol. 22, no. 4, pp. 526–539, 2015.

[157] P. Tawa, K. Hell, A. Giroux, E. Grimm, Y. Han, D. Nicholson, and S. Xanthoudakis, "Catalytic activity of caspase-3 is required for its degradation : stabilization of the active complex by synthetic inhibitors," *Cell Death & Differentiation*, vol. 11, no. 4, pp. 439–447, 2004.

[158] K. L. Huber, S. Ghosh, and J. A. Hardy, "Inhibition of caspase-9 by stabilized peptides targeting the dimerization interface," *Peptide Science*, vol. 98, no. 5, pp. 451–465, 2012.

[159] C. Mas-Moruno, J. G. Beck, L. Doedens, A. O. Frank, L. Marinelli, S. Cosconati, E. Novellino, and H. Kessler, "Increasing $\alpha v\beta 3$ selectivity of the anti-angiogenic drug cilengitide by n-methylation," *Angewandte Chemie International Edition*, vol. 50, no. 40, pp. 9496–9500, 2011.

[160] J. Chatterjee, C. Gilon, A. Hoffman, and H. Kessler, "N-methylation of peptides : a new perspective in medicinal chemistry," *Accounts of chemical research*, vol. 41, no. 10, pp. 1331–1342, 2008.

[161] J. 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*, vol. 21, no. 12, pp. 1049–1074, 2000.

[162] B. Feeney, C. Pop, P. Swartz, C. Mattos, and A. C. Clark, "Role of loop bundle hydrogen bonds in the maturation and activity of (pro) caspase-3," *Biochemistry*, vol. 45, no. 44, pp. 13249–13263, 2006.

Remerciements

Je tiens à remercier mes encadrants Jacques CHOMILIER, Dirk STRATMANN et Fabio PIETRUCCI pour ces 3 années de thèse. Leur patience et leur pédagogie ainsi que leur soutien pour les derniers mois de la thèse m'a été d'une grande aide pour supporter l'année.

Un grand merci à Manuel DAUCHEZ et Matthieu MONTES pour avoir accepter de lire ma thèse ainsi qu'aux autres membres de mon jury de thèse (Anne-Claude CAMPROUX, Jessica ANDREANI et Patrick FUCHS). Je remercie également Guillaume Fiquet et l'équipe BIBIP pour m'avoir accueilli dans le labo. Je dédie également ma thèse à Maud JUSOT, mon ancienne maîtresse de stage, pour l'initiation à la vie de thésard, ainsi qu'à Guillaume POSTIC.

Une pensé pour mes amis du master de bioinformatique, dont notamment Florence et Julie pour les après-midis jeux de société et cuisine. Je remercie également Laura, Léon, Dania, Appoline, Julie(tte) et Billy pour les discussions non scientifiques durant les pauses. Avec le temps, ils sont devenus plus que de simples collègues.

Un grand merci pour mes camarades de la 412 avec qui on a contribué à l'animation du couloir 22-23. Nicolas qui m'a enseigné la théorie de la montre, Baptiste avec ses mimes de boxe thaï, Antoine pour le barbe rousse et Romain le dernier venu. Ce dernier m'a montré qu'il était possible d'imiter des dinosaures, fossiliser une pomme de terre dans de l'époxy et manger de la colle sans qu'aucune personne ne se posent des questions sur sa santé mentale.

Enfin une dernière pensée pour Marion et Aniela qui ont été là du début à la fin et ont gardé le moral malgré mes blagues douteuses.

SUMMARY

Proteins are molecules involved in biological function. Most of them interact with other proteins. Disturb protein protein interactions has high potential in the development of new drugs and in understanding biological mechanisms (pathway, molecular transportation...). Protein portein interaction involve relatively large protein surfaces lacking well-definied binding pockets for currently therapeutic drugs. One way to target and disturb those IPP it is to find molecules can mimic IPP interaction. Cyclic peptides could be good candidat with a good specificity and target for their targets. Indeed cyclization stabilize and increase their resistance again protease. However make experimental experience to check their binding affinity can be complicated. In this thesis we present method to sample Cyclic peptides's conformational landscape and predicted their binding affinity for their targets with enhanced sampling method