Université de Montréal

Développement et Optimisation des potentiels OPEP et simulations numériques de la protéine Huntingtine.

par
Vincent Binette

Département de physique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en physique

07 Avril, 2017

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Développement et Optimisation des potentiels OPEP et simulations numériques de la protéine Huntingtine.

présenté par:

Vincent Binette

a été évalué par un jury composé des personnes suivantes:

Rikard Blunck,	président-rapporteur
Normand Mousseau,	directeur de recherche
Radu Iftimie,	membre du jury

Mémoire accepté le: . . . . . .

RÉSUMÉ

Aux fondements de la vie, on retrouve les protéines qui agissent telles des nanomachines à l'échelle de la cellule. En effet, elles effectuent une immense diversité de tâches vitales à tout organisme vivant, allant de la transcription/traduction de l'ADN, au transport membranaire en passant par le métabolisme énergétique, etc. Les protéines sont des macromolécules composées de combinaisons des 20 acides aminés classiques. Sous l'effet des lois fondamentales de la physique, cette chaîne va se replier localement (structure secondaire) puis globalement (structure tertiaire). C'est cette dernière qui donne la fonction aux protéines et son étude se révèle donc primordiale. Les méthodes numériques offrent un complément indispensable à l'expérience pour la réalisation de cet immense défi. Ce mémoire s'articule autour de deux axes principaux. Le premier est l'étude du N-terminal de la protéine Huntingtine à l'aide de simulations numériques. Le second est l'optimisation et le développement de la famille de potentiel gros-grain OPEP.

La protéine Huntingtine est à l'origine du développement de la maladie d'Huntington et son N-terminal est crucial pour son agrégation et son interaction avec la membrane. Suite à la proposition d'un premier modèle expérimental du N-terminal de Huntingtine, nous avons déterminé son ensemble structurel, en solution et en membrane, à l'aide de simulations tout-atome et de techniques d'échantillonnage avancées. Les motifs structurels et interactions dominantes observés sont mis en relation avec les modèles détaillés de son agrégation et de son ancrage dans la membrane.

Finalement, les potentiels OPEP sont des potentiels gros-grain dont l'application s'est révélée un succès pour l'étude des protéines amyloïdes et pour la prédiction de structure tertiaire avec PEPFOLD. Le potentiel sOPEP fut optimisé sur de vastes ensembles de protéines avec comme résultat une amélioration de 25% au niveau des inégalités. La paramétrisation de sOPEP fut aussi testée à l'aide de simulation sur neuf protéines. En parallèle, nous développons le potentiel aaOPEP, qui portera la philosophie des potentiels OPEP en régime tout-atome. Ces potentiels sont intégrés à un code de dynamique moléculaire flexible qui permettra leur distribution.

Mots clés: Potentiel gros-grain, OPEP, protéine amyloïde, Huntingtine

ABSTRACT

Proteins are the vital machinery of the cell and therefore are fundamental to life. They play a wide variety of roles from the transcription/translation of DNA to membrane transport to energy metabolism and many more. In terms of structure, proteins are macromolecules composed by the combination of 20 basic amino acids. Under the fundamental laws of physic, this chain will fold itself locally (secondary structure) and then globally (tertiary). This three dimensional structure gives rise to the protein's function thus making the prediction of three dimensional structure a crucial aspect of protein study. Numerical techniques provide an essential tool to complement experiments in the study of proteins.

The following thesis is divided into two main research axes. The first one is the study, via numerical techniques, of the N-terminal of the Huntingtin protein. The second one is the optimization and the development of the coarse grained forcefield family OPEP.

First, we present our study of Huntingtin's N-terminal. The Huntingtin protein is associated with the Huntington disease and its N-terminal is crucial for the protein aggregation and anchoring inside the membrane. Following the proposition of a first experimental model of the N-terminal, we determined its structural ensemble in aqueous solution and membrane environment. To do so, we used powerful sampling techniques in the all-atom regime. Our discoveries are linked to Huntingtin's aggregation and membrane association models.

Finally, the last chapter discusses the optimization of the sOPEP and aaOPEP forcefield. The coarse grained OPEP forcefields were successfully applied to multiple studies, most notably for the simulation of amyloid peptides and the tertiary structure prediction using the PEPFOLD methodology. Vast protein ensemble were generated and used in the optimization of sOPEP resulting in 25% of improvement compared to the previous version on our decoys set. The parametrization is also validated by short molecular dynamic simulations on nine proteins. In parallel, we develop the new aaOPEP, a forcefield that will use the OPEP philosophy but in the all-atom regime. Those potential are included in a flexible molecular dynamic code that will be available to everyone when completed.

Key words: Coarse-Grained Potential, OPEP, Amyloid Protein, Huntingtin protein

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ANNEXES

REMERCIEMENTS

Je tiens premièrement à remercier le professeur Normand Mousseau pour m'avoir donné l'opportunité de travailler dans son groupe de recherche tout au long de mon baccalauréat et de ma maîtrise. Sous sa supervision, j'ai eu la chance de vivre des expériences enrichissantes. Ses conseils et son expérience ont grandement contribué à l'amérioration de mes aptidudes de recherche et je suis impatient de continuer à travailler sous sa supervision pour mon doctorat.

Je veux aussi remercier mon (ex-)collègue Sébastien Côté pour avoir accepté de jouer un rôle de mentor et pour la confiance qu'il m'a témoigné à mes débuts dans l'équipe. Il m'appris énormément sur la réalité des cycles supérieurs et sur la physique. J'ai beaucoup apprécié travailler avec lui. Sans cette expérience, mes études n'auraient pas été aussi enrichissantes !

Je remercie les collègues du groupe de recherche : Oscar Restrepo, Sami Mahmoud et Mickaël Trochet pour une belle ambiance au bureau.

Finalement, je dois beaucoup à ma famille pour leur appui indéfectible dans tout ce que j'entreprends, à mes amis pour les bons moments passés ensembles et à ma blonde pour faire ressortir le meilleur en moi !

CHAPITRE 1

INTRODUCTION

Les protéines sont des macromolécules biologiques aux fonctions et structures aussi diverses que complexes. Une analogie populaire dit que les protéines sont des machines agissant à l'échelle nanoscopique. En effet, certaines jouent le rôle d'enzyme qui permet de catalyser les réactions chimiques. Entre autres exemples, les protéases permettent de briser les protéines en petits fragments facilitant leur dégradation et les polymérases servent à assembler l'ADN ou l'ARN. Certaines protéines jouent aussi des rôles cruciaux au niveau du signalement, comme l'insuline qui permet l'absorption du glucose, et au niveau du transport comme l'hémoglobine qui permet le transport de l'oxygène. Certaines protéines forment des canaux qui font passer les ions à travers la membrane cellulaire, tels les canaux sodiques et potassiques. Ces exemples ne représentent qu'un petit échantillon de la panoplie de fonctions des protéines.

Dans la cellule, les informations nécessaires à la synthèse des protéines se retrouvent encodées dans le génome, soit l'ensemble du matériel génétique de celle-ci, qui est constitué d'ADN (parfois ARN chez les virus). Plus spécifiquement, les régions de l'ADN encodant la séquence des protéines se nomme région codante. Ces régions codantes de l'ADN sont ensuite transcrites en ARN par des enzymes nommées ARN polymerase puis l'ARN est traduit en protéine dans le ribosome. Le lien entre la séquence de l'ADN et les protéines est fondamental. Le Projet Génome Humain [1] permit d'obtenir la séquence du génome humain et les plus récentes estimations du nombre total de protéines uniques encodées dans celui-ci tournent autour de 19 000 [2].

Toutes les structures expérimentales des protéines présentement déterminées sont répertoriées dans la Protein Data Bank [3] qui fut inaugurée en 1971 et contenait à l'origine sept protéines. Aujourd'hui, elle est composée de 114 444 structures de près de 21 033 protéines uniques dont 4027 chez l'humain. Ces chiffres démontrent tous les efforts colossaux de la communauté scientifique dans ce grand projet et les améliorations importantes réalisées au cours des 45 dernières années. Malgré toutes ces avancées, beaucoup

de travail reste à accomplir afin d'obtenir les structures des milliers de protéines restantes (près de 15 000 juste chez l'humain). Face à un travail d'une telle ampleur, les méthodes numériques, modélisant les protéines au niveau atomique, deviennent de plus en plus incontournables. La modélisation des protéines se fait à l'aide d'un potentiel (ou champ de force) qui caractérise les paramètres des atomes ainsi que les interactions présentes entre eux. Les potentiels traditionnels tout-atome, modélisant individuellement tous les atomes du système (protéine, solvant etc.) demeurent très populaires et sont appliqués avec succès à une vaste gamme de problèmes. Par contre, ces derniers sont sévèrement limitées par les temps de calcul faramineux et les ressources informatiques limitées. C'est pourquoi les échelles de temps pertinentes à l'étude détaillée de la dynamique et thermodynamique des protéines restent trop souvent inaccessibles à l'aide d'un temps de simulation raisonnable. Les potentiels gros-grain forment une alternative aux potentiels traditionnels tout-atome en accélérant les simulations grâce à une diminution du nombre de degrés de liberté du système. La présente thèse s'articulera autour de deux thèmes principaux qui sont décrits individuellement dans ce qui suit. Une revue de littérature plus exhaustive est présentée en introduction pour chacun des thèmes aux chapitres 3 et 5.

1.1 La protéine Huntingtine

La protéine Huntingtine interagit avec plus de 350 partenaires et serait impliquée dans de nombreuses fonctions biologiques fondamentales comme la transcription, le métabolisme énergétique et bien plus. Phénomène intéressant, une expansion du segment polyglutamine situé dans son N-terminal est à l'origine du développement de la maladie neurodégénérative Huntington. Ce N-terminal, correspondant à l'exon1 du gène *Htt*, est suffisant pour reproduire le phénomène d'agrégation à l'origine de la maladie. C'est pourquoi une étude détaillée de la structure et des mécanismes d'agrégation du N-terminal s'impose. Celui-ci est composé de trois fragments distincts, une séquence de 17 acides aminés à son N-terminal, un domaine polyglutamine dont la taille détermine l'âge d'apparition et la sévérité de la maladie, et un segment riche en proline. Tous

ces fragments sont intrinsèquement désordonnés au niveau du monomère. Cette grande flexibilité de structure est complexe à étudier expérimentalement. Cela fait de ce peptide un candidat idéal pour les techniques numériques appliquées dans ce mémoire qui, dans cette situation, sont un complément essentiel aux expériences et permettent de dresser un portrait à haute résolution (au niveau atomique). Je présenterai aux chapitres 3 et 4 de ce mémoire les résultats de nos simulations sur le N-terminal de la protéine huntingtine. Les diverses méthodes numériques pertinentes à la compréhension de ce mémoire sont quant à elles décrites en détail au chapitre 2.

1.2 Développement et amélioration de OPEP

En biophysique numérique, les lois fondamentales de la physique, à l'origine du repliement des protéines, sont modélisées à l'aide d'un champ de force (potentiel). C'est le potentiel qui assure une représentation adéquate des protéines lors de la simulation. Parmi ces potentiels, on retrouve les potentiels gros-grain qui permettent une accélération substantielle des simulations en sacrifiant la complexité du modèle. OPEP est une famille de potentiels gros-grain qui fut employé avec succès pour la simulation de multiples protéines, dont plusieurs amyloïdes [4], et est l'outil de discrimination derrière la méthode PEPFOLD [5] faisant la prédiction de la structure secondaire et tertiaire des protéines uniquement à partir de la structure primaire.

Finalement, au chapitre 5 de ce mémoire, je présenterai les résultats de nos optimisations du potentiel sOPEP ainsi que les derniers avancements dans la paramétrisation du nouveau potentiel aaOPEP conçu et développé dans le groupe. Ce dernier est une extension au régime tout-atome des potentiels OPEP, une approche unique qui offrira un intermédiaire entre les potentiels tout-atome et gros-grain. Parallèlement, nous développons aussi un code de dynamique moléculaire qui permettra l'utilisation et la distribution des potentiels OPEP incluant sa version tout-atome aaOPEP.

CHAPITRE 2

MÉTHODOLOGIE

Au milieu des années 70, dans ce qui allait devenir la célèbre loi de Moore, Gordon Moore prédit que la puissance de calcul des ordinateurs doublerait à chaque deux ans, une tendance respectée jusqu'à tout récemment. À titre comparatif, si le milieu de l'automobile avait connu la même progression fulgurante que le domaine informatique jusqu'à nos jours, une Rolls Royce couterait moins de 100$ avec des performances de l'ordre de 1.18L/100km [6] (après vérification, ce n'est pas le cas) ! Le potentiel croissant des ordinateurs ne passa pas inaperçu des biologistes, chimistes et physiciens qui se lancèrent dans le développement de méthodes numériques dans les années 60s et 70s. Au début des années 60s, de multiples groupes travaillèrent de manière indépendante au développement des tous premiers champs de force basés sur des mesures expérimentales de spectroscopie, chaleur spécifique, calculs de mécanique quantique etc [12]. C'est notamment à l'institut Weizmann en Israël que le groupe du professeur Shneior Lifson reporta les premières minimisations énergétiques de protéines (myoglobine et lysozyme) [7]. En 1971, c'est Rhaman et Stillinger qui présentent les premières dynamiques moléculaires de molécules d'eau [8]. Les bases de la simulation numérique de protéines étaient posées et la crédibilité du domaine crût dans les années qui suivirent. Le domaine prit véritablement son envol avec l'apparition des premiers potentiels complets comme CHARMM, du groupe de Martin Karplus (prix Nobel de chimie 2013) et AMBER du groupe de Kollmann) [7]. Aujourd'hui, quelques 40 ans plus tard, l'amélioration colossale de l'informatique et le développement de nouveaux algorithmes numériques ont permis des avancées jugées jusqu'alors impossibles dans plusieurs domaines : simulation de protéines de l'ordre de la milliseconde [9], maîtriser le jeu de go [10], calculs de mécanique quantique à plusieurs corps [11] etc. Il n'y a aucun doute que les méthodes numériques sont devenues un partenaire essentiel à l'expérience et qu'elles recèlent un potentiel important pour l'avenir.

Dans ce chapitre, nous ferons un aperçu des diverses méthodes numériques perti-

nentes pour la compréhension de ce mémoire.

2.1 Les potentiels

Les protéines sont des macromolécules présentes à l'échelle nanoscopique. L'approche la plus rigoureuse et la plus juste pour les étudier viserait la résolution de la fonction d'onde en mécanique quantique. Malgré des avancées importantes au niveau de la performance des ordinateurs, cette approche demeure encore aujourd'hui impossible à réaliser en un temps de simulation raisonnable. De ce fait, l'approche la plus utilisée pour étudier numériquement les protéines est l'utilisation de la mécanique moléculaire classique. À la base de cette approche, on retrouve des fonctionnelles énergétiques simples composées de paramètres ajustables. Ces paramètres sont souvent déterminés empiriquement afin de reproduire les résultats expérimentaux ou les calculs de mécanique quantique. L'ensemble de ces fonctionnelles énergétiques et des paramètres forme ce que l'on appelle les potentiels (ou champs de force). Le développement des potentiels repose sur trois hypothèses physiques [12]. La première est l'hypothèse thermodynamique selon laquelle les molécules peuvent se replier naturellement vers leur état natif (de plus basse énergie). La seconde est l'hypothèse d'additivité à savoir que l'énergie associée à une structure peut s'écrire comme la somme de plusieurs fonctionnelles énergétiques plus simples (énergie des liens, des angles, van der Waals etc.). La troisième est l'hypothèse de transférabilité qui stipule que le potentiel dérivé à partir de structures expérimentales connues permet de faire la prédiction des structures encore inconnues. En bref, la qualité du potentiel dépend de sa capacité à reproduire (et prédire) numériquement les observations expérimentales. Avec la recherche, les potentiels deviennent de plus en plus sophistiqués et leur complexité varie grandement avec la qualité/quantité de détails désirés. Un contre coup de ces raffinements est que plus un potentiel est fin, plus les temps de calcul requis seront imposants. Il s'agit là du dilemme principal de quiconque travaille avec des méthodes numériques : trouver l'équilibre entre la vitesse et la précision du potentiel. Deux philosophies de potentiel naissent de ce dilemme : les potentiels tout-atome et les potentiels gros-grain. Une brève description de ces deux

familles et de certains de leurs représentants est proposée aux sections suivantes.

2.1.1 Potentiel tout-atome

Un potentiel tout-atome, comme son nom l'indique, est un potentiel représentant explicitement tous les atomes du système à l'étude. Ces potentiels permettent d'obtenir des résultats très fins avec en contrepartie des temps de simulation importants. Les potentiels AMBER [13] et CHARMM [14] sont quelques exemples de ce type de potentiel souvent utilisés.

Avec l'hypothèse de l'additivité, l'énergie totale est généralement divisée en plusieurs termes selon :

$$E = E_{\text{Lien}} + E_{\text{Angle}} + E_{\text{Torsion}} + E_{\text{Torsion Impropre}} + E_{\text{Non Liée}}$$

La forme et les paramètres impliqués dans chacun des termes varient avec le potentiel.

Par exemple, le potentiel AMBER est basé sur des fonctionnelles énergétiques simples composées d'interactions entre deux corps. Les spécificités de AMBER comprennent, entre autres, des charges partielles fixes et localisées sur chaque atome, l'utilisation explicite des atomes d'hydrogène, aucune fonctionnelle énergétique spécifique aux ponts-H et des paramètres d'angles diédraux déterminés à l'aide de calculs de mécanique quantique sur de petites molécules [13, 15]. Ce potentiel est encore continuellement amélioré, surtout au niveau des angles diédraux permettant de balancer les divers motifs de la structure secondaire. Tout particulièrement, des calculs de mécanique quantique réalisés sur des tétramères de glycine et d'alanine [15] et des dynamiques moléculaires de l'ordre de la microseconde [16] permirent d'optimiser les paramètres des angles diédraux (tout particulièrement des résidus Ile, Leu, Asn, Asp). Le potentiel AMBER fut utilisé pour nos simulations sur Htt17 en solution (au chapitre 4) et sur Htt17 en membrane (en annexe I).

Une autre famille de potentiels tout-atome largement utilisée est CHARMM [14]. Les potentiels CHARMM ont essentiellement la même forme que les potentiels AMBER, mais avec quelques différences clés [17]. En autre, CHARMM ajoute une fonction-

nelle énergétique pour les interactions 1-3 (Urey-Bradley), les charges dans CHARMM sont basées sur des calculs numériques de chimie quantique (HF/6-31G*) tandis que ceux d'AMBER sont basés sur un potentiel électrostatique (RESP), etc. Lors de simulations, le potentiel doit être choisi judicieusement afin qu'il soit adéquat pour le système à l'étude.

2.1.2 Potentiel gros-grain

En plus des potentiels tout-atome, on retrouve aussi les potentiels gros-grain. L'idée derrière ces potentiels est de diminuer le nombre de degrés de liberté du système afin d'augmenter l'échantillonnage et d'accélérer les simulations [18]. Le principe du gros-grain est de réunir un certain nombre d'atomes de la protéine en un seul centre d'interaction dont les paramètres physiques seront ensuite modélisés. Plusieurs modèles gros-grain furent développé allant d'un centre d'interaction (C_α), deux centres d'interaction (C_α, chaîne latérale), jusqu'à un modèle six centres d'interaction (chaîne principale tout-atome et un centre d'interaction par chaîne latérale) [19]. Les potentiels OPEP font partie de cette dernière catégorie. Comme OPEP est le sujet principal de cette thèse, une description détaillée de sa paramétrisation est présentée au chapitre 5.

Un autre modèle très populaire est celui du potentiel MARTINI [20]. Le modèle gros-grain de MARTINI est construit de telle sorte qu'en moyenne quatre atomes lourds (et leurs hydrogènes respectifs) sont unis dans un seul centre d'interaction. Cette unification offre une bonne balance entre la précision et la rapidité du potentiel. Chacune des billes est caractérisée par trois paramètres : un indiquant le type (polaire, non-polaire, apolaire ou chargé), un indiquant la capacité à former des ponts-H (donneur, accepteur, les deux ou aucun des deux) et un indiquant le degré de polarité (entre 1 et 5). On remarque plusieurs différences notables entre MARTINI et les potentiels OPEP. Dans MARTINI, l'eau est décrite explicitement, mais toujours en unissant quatre molécules en un centre d'interaction. De plus, billes chargées ont une charge explicite, localisées et leur interaction est décrite par une force de Coulomb. En comparaison, dans OPEP, ces deux interactions sont dissimulées à l'intérieur du potentiel entre chaînes latérales. La différence la plus fondamentale entre les deux potentiels est que la paramétrisation

de MARTINI est thermodynamique tandis que celle d'OPEP est structurelle [21]. En d'autres mots, les interactions non-liées de MARTINI sont paramétrisées pour chacune des billes à partir de données expérimentales thermodynamiques comme l'énergie libre de vaporisation, d'hydratation et de la séparation eau/huile. Plus précisément pour les protéines, la décomposition des différents acides aminés en bille fut testée en comparant les résultats de simulations et d'expériences pour les coefficients de partition eau-huile. Les simulations furent réalisées en ajoutant une faible concentration des acides aminés dans un mélange d'eau et de butane [22]. Dans le cas où deux conformations donnaient des résultats similaires, la meilleure fut déterminée à l'aide de simulation permettant de calculer le potentiel de force moyenne (Potential of mean force) en fonction de la distance avec une membrane lipidique [22]. Dans OPEP, la paramétrisation des interactions non-liées est déterminée à partir des structures expérimentales des protéines de la PDB. Plus de détails sur sa paramétrisation est donnée au chapitre 5

2.2 Méthodes d'échantillonnage de pointe

Une fois le potentiel conçu, il est maintenant possible d'étudier la dynamique et thermodynamique des protéines à l'aide de diverses méthodologies. Parmi ces méthodes, la dynamique moléculaire classique (MD) est un outil puissant et largement utilisé pour simuler l'évolution de systèmes moléculaires. Bien que les protéines sont à l'échelle du nanomètre, c'est la mécanique classique (newtonienne) qui est à la base de la MD et les effets quantiques sont ignorés. L'objectif de la MD est de faire évoluer le système en résolvant les équations du mouvement du système à partir d'une configuration initiale. On doit donc résoudre le système d'équations suivant :

$$\begin{aligned} \nabla V(\mathbf{x}(t)) &= F(\mathbf{x}) \\ \ddot{\mathbf{x}} &= \frac{F}{m} \end{aligned} \tag{2.1}$$

À partir de notre potentiel ($V(\mathbf{x}(t))$) à un temps t, on peut déterminer la force ($F(\mathbf{x})$) agissant sur chacun des atomes puis trouver leur accélération ($\ddot{\mathbf{x}}$) avec la seconde loi

de Newton et intégrer afin d'obtenir les nouvelles vitesses/positions de nos atomes à un temps $t + \delta t$.

Toutefois, l'échantillonnage de l'espace conformationnel pour des systèmes complexes demeure un obstacle important. En effet, les barrières énergétiques séparant les divers états métastables tendent à être beaucoup plus importantes que $k_b T$ avec comme résultat l'emprisonnement du système dans un état métastable sur de longues échelles de temps. Afin de contourner ce problème, de multiples techniques d'échantillonnage avancées furent développées et offrent une alternative à la MD. Dans ce chapitre, nous décrirons brièvement les diverses techniques pertinentes pour ce travail.

2.2.1 Échange de répliques hamiltonien

L'échange de répliques permet la simulation en parallèle de plusieurs versions du système, nommées répliques. L'idée de base derrière cette méthode est d'échantillonner une réplique dite "froide" pour extraire les statistiques pertinentes physiquement, une réplique "chaude" permettant l'exploration plus rapide de l'espace de phase en surmontant plus facilement les barrières énergétiques et des répliques intermédiaires servant uniquement à faire graduellement le pont entre la "froide" et la "chaude". Dans le cas de l'échange de répliques en température (PT), la réplique froide et chaude correspondent réellement à la plus basse et la plus haute température respectivement. Dans cette section, nous nous intéresserons plutôt à l'échange de répliques hamiltonien (HREX) dans sa version REST2 [23]. Pour HREX, les différentes répliques sont simulées à la même température. C'est plutôt l'énergie potentielle qui est échelonnée selon :

$$E_m^{REST2}(X) = \frac{\beta_m}{\beta_0} \cdot E_{pp}(X) + \sqrt{\frac{\beta_m}{\beta_0}} \cdot E_{pw}(X) + E_{ww}(X) \tag{2.2}$$

où E_{pp}, E_{pw} et E_{ww} est l'énergie associée respectivement aux interactions protéine/protéine, protéine/solvant et solvant/solvant. β_m et β_0 sont les bêtas thermodynamiques ($\beta = 1/k_b T$) respectivement à la température T_m et T_0 [23].

Plus en détail, l'énergie des angles diédraux, les paramètres de Lennard-Jones et les

charges du soluté sont modifiés respectivement par un facteur $\frac{\beta_m}{\beta_0}$, $\frac{\beta_m}{\beta_0}$ et $\sqrt{\frac{\beta_m}{\beta_0}}$. Le résultat d'un tel échelonnage est la réduction de la taille des barrières séparant les différents états métastables. Prenons la situation où une protéine se trouve dans un minimum du paysage énergétique. Si les tailles des barrières d'énergie entourant ce minimum sont trop grandes, il est extrêmement rare que la protéine puisse quitter ce minimum afin d'explorer la totalité de ses états accessibles et de potentiellement trouver le minimum global du système. Avec l'échelonnage proposé par HREX, la taille des barrières d'énergie se retrouve rapetissée pour les répliques "chaudes". Ainsi, une barrière pratiquement infransissable pour le système "froid" devient une barrière aisément franchissable pour le système "chaud". La vitesse et la qualité de l'échantillonnage se retrouvent donc grandement améliorées. La modification de l'énergie liée aux liens et aux angles n'est pas nécessaire pour une bonne accélération de l'échantillonnage.

Les répliques "chaudes", permettant l'accélération de l'échantillonnage en échelonnant la taille des barrières énergétiques, ne se retrouvent donc pas dans des conditions physiques pertinentes. Il est donc nécessaire de permettre aux diverses répliques de communiquer entre-elles afin de ramener les informations vers le milieu physiquement pertinent, soit en passant de la réplique "chaude" vers la réplique "froide". Ces transferts s'effectuent par l'échange des coordonnées spaciales entre les répliques à divers moments de la simulation. Durant la simulation, les répliques sont donc échangées entre les différentes "températures" et le taux d'acceptation des échanges entre deux répliques est calculé de manière générale avec :

$$\alpha = \min\left\{1, \exp\left[\frac{-U_i\left(r_j\right)+U_i\left(r_i\right)}{k_b T_i}+\frac{-U_j\left(r_i\right)+U_j\left(r_j\right)}{k_b T_j}\right]\right\} \tag{2.3}$$

où r_i/r_j représentent les positions de la réplique i et j et T_i/T_j représentent la température de la réplique i et j.

HREX a deux avantages comparativement à la version PT. HREX offre un meilleur taux d'échange entre les répliques et une meilleure transition entre les diverses "températures". Bref, il y a donc moins de répliques à simuler qu'en PT. De plus, échelonner le potentiel, une propriété extensive du système, permet de sélectionner des parties du sys-

tème à "chauffer", ce qui est impossible de faire pour la température qui est une propriété intensive. HREX est implémenté dans un logiciel complémentaire [24] au programme GROMACS [25, 26] et venant avec le programme PLUMED [27]. Nos simulations sur Htt17 en solution (chapitre 4) et en membrane (annexe I) sont réalisées avec cette méthode.

2.2.2 Métadynamique

Introduite en 2002, la métadynamique [28, 29] (MetaD) est une technique d'échantillonnage avancée permettant de reconstruire le paysage d'énergie libre (FES) sous-jacent au système en fonction d'un certain nombre de degrés de liberté pré-choisis et nommés variables collectives (CV). De plus, elle accélère l'échantillonnage en ajoutant, dans l'espace des CVs, un biais dépendant de l'historique du système qui "remplit" progressivement les minimums du FES et permet de guider le système vers de nouveaux états encore jamais visités. Ce biais peut s'écrire comme une somme de gaussiennes déposée le long de la trajectoire dans l'espace des CVs et décourage le système à revisiter les configurations déjà échantillonnées. Le biais est décrit par l'équation suivante :

$$V_G(\vec{s},t) = \int_0^t \mathrm{d}t' \omega \cdot \exp\left[-\sum_{i=1}^{d} \frac{\left(s_i(R) - s_i(R(t'))\right)^2}{2\sigma_i^2}\right] \tag{2.4}$$

où V_G est le potentiel de biais ajouté et prend la forme d'une somme de gaussienne, $\vec{s}$ est l'ensemble de d CVs dépendant des coordonnées microscopiques R, ω est une puissance et σ est la largeur des gaussiennes [29]. Une version schématisée du fonctionnement de la MetaD est présentée à la Figure 2.1.

Trois facteurs font de la MetaD une méthode particulièrement intéressante. (1) Elle permet d'échantillonner des événements rares en sortant le système des minimums locaux, (2) aucune information sur le FES n'est à connaître *a priori* et (3) le biais permet de déterminer un estimé du FES à une constante (C) près selon :

$$V_G(\vec{s}, t \to \infty) = -F(\vec{s}) + C \tag{2.5}$$

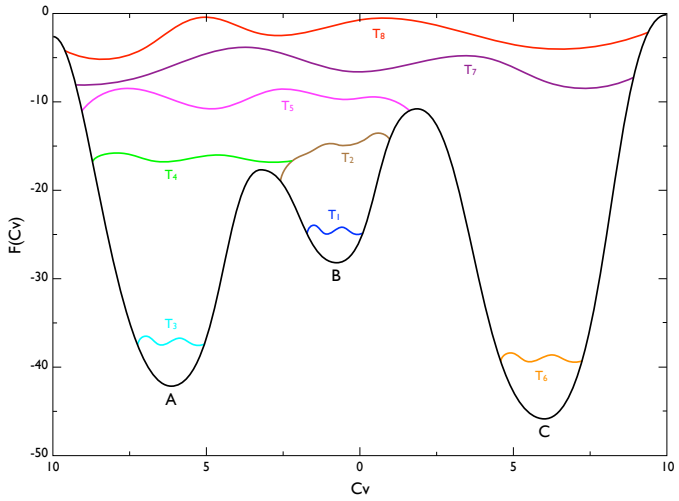

Figure 2.1 : Schéma représentant le fonctionnement de la MetaD avec en abscisse la valeur d'une CV quelconque et en ordonnée la valeur de l'énergie libre. Le trait noir correspond au FES en une dimension tandis que les traits de couleur représentent le potentiel de biais à divers moment. Dans ce cas, le système se trouve initialement dans le minimum B. Le biais remplit progressivement le FES de telle sorte que le minimum A est accessible au temps T_3 et le minimum C est accessible au temps T_6. Ultimement, une connaissance des biais permet de reconstruire le FES sous-jacent.

où F est l'énergie libre.

Par contre, la MetaD vient avec deux inconvénients majeurs [29]. (A) Le biais ne converge pas à la valeur d'énergie libre, mais fluctue autour de celle-ci causant un risque de remplissage excessif qui pousse le système à l'extérieur des régions physiquement pertinentes et rend l'atteinte de la convergence complexe à déterminer et (B) l'identification des CVs utiles à la description du système est ardue.

Le problème (A) fut adressé lors du développement de la métadynamique à recuit simulé [30] (WT-MetaD). Cette méthode introduit un biais progressivement amorti garantissant la convergence au résultat exact. Ce dernier prend maintenant une forme différente donnée par :

$$V_G(\vec{s},t) = k_b \cdot \Delta T \cdot \ln\left[1 + \frac{\omega \cdot N(\vec{s},t)}{k_b \cdot \Delta T}\right] \tag{2.6}$$

où k_B est la constante de Boltzmann, ΔT est un paramètre d'entrée avec des unités de

température et $N(\vec{s},t)$ est l'histogramme des variables collectives $\vec{s}$ calculé pour la simulation. En pratique, l'implémentation se fait en redimensionnant la hauteur des Gaussiennes (W) avec :

$$W = \omega \cdot \tau_G \cdot \exp\left[\frac{-V_G(\vec{s},t)}{k_B \Delta T}\right] \tag{2.7}$$

où τ_G est le taux de déposition des Gaussiennes.

Le problème (A) est maintenant réglé puisque le potentiel de biais converge sur de longues échelles de temps selon :

$$V_G(\vec{s}, t \to \infty) = -\frac{\Delta T}{T + \Delta T} F(\vec{s}) + C \tag{2.8}$$

et permet de concentrer les calculs uniquement sur les régions pertinentes physiquement.

L'identification des CVs est fondamentale au fonctionnement de la MetaD. Les CVs devraient correspondre aux coordonnées de réaction lentes du système qui limite l'exploration du FES. Pour des systèmes complexes, il est probable que les CVs soient difficilement identifiables ou qu'elles soient trop nombreuses pour une utilisation réaliste de la MetaD. Ces considérations sont à l'origine du problème (B). Pour passer outre cette problématique, il est pratique de combiner la MetaD avec une approche d'échange de répliques (en température ou hamiltonien) afin d'accélérer la relaxation du système le long des degrés de liberté oubliés par les CVs [31].

De plus, la WT-MetaD est particulièrement intéressante puisqu'elle permet de retrouver les distributions non-biaisées de n'importe quelles quantités n'étant pas les CVs grâce à diverses méthodologies de repondération [32, 33] et permet donc une comparaison quantitative avec l'expérience. Tous les outils de MetaD sont disponibles grâce au logiciel PLUMED [27] compatible avec les plus récentes versions du programme GROMACS [25, 26]. La WT-MetaD fut combinée à HREX pour nos simulations de Htt17 en solution(chapitre 4).

2.3 Prédiction structurelle

En plus des méthodes décrites à la section précédente, permettant d'étudier les protéines en les faisant avancer dans le temps, les potentiels sont aussi à la base de diverses méthodes de prédictions structurelles. PEPFOLD est une de ces méthodes et permet de prédire la structure native (expérimentale) de petites protéines à partir uniquement de leur séquence en acides aminés. La méthodologie derrière PEPFOLD, dans laquelle le potentiel gros-grain sOPEP (pour simplified OPEP) joue un rôle crucial, est présenté dans ce qui suit.

2.3.1 PEPFOLD

La structure des protéines peut être large et complexe. Elle peut par contre être décomposée en fragments (motifs, sous-structures) plus petits et communs/récurrents chez toutes les protéines [34]. L'ensemble de ces fragments forment ce que l'on appelle un alphabet structurel (SA) et les règles pour assembler les fragments les uns avec les autres se nomme une grammaire. La méthode PEPFOLD utilise ces deux concepts afin de reconstruire la structure tridimensionnelle des protéines à partir uniquement de leur séquence en acides aminés.

2.3.1.1 Alphabet Structurel et Grammaire

À la base de la méthode PEPFOLD, on retrouve un SA, un ensemble fini de lettres (fragments) permettant de reconstruire la structure de n'importe quelle protéine avec un certain niveau de précision [35]. Le SA utilisé par PEPFOLD est composé de 27 lettres correspondant chacune à un petit fragment de la chaîne principale composé de quatre carbones α, donc quatre acides aminés. Ces lettres peuvent être décrites de façon unique à l'aide de quatre descriptifs correspondant aux vecteurs entre les trois carbones α du fragment : $d_1 = d\left(C_\alpha^1 - C_\alpha^3\right)$, $d_2 = d\left(C_\alpha^1 - C_\alpha^4\right)$, $d_3 = d\left(C_\alpha^2 - C_\alpha^4\right)$ et de la projection C_α^4 sur le plan formé par les trois autres C_α. Ces quatre paramètres sont présentées graphiquement à la Figure 2.2 Les différents types de structure secondaire sont tous représentés à travers chacune de ces lettres qui sont présentées à la Figure 2.3. Par exemple, l'hélice

α est associée particulièrement aux lettres dénommées A et a. Nous avons maintenant un ensemble fini de fragments avec lequel on peut décomposer la structure de n'importe quelle protéine. Par contre, il manque toujours la grammaire, c'est à dire, les règles utiles afin de combiner les diverses lettres entre elles pour reconstruire la structure. La grammaire de PEPFOLD prend la forme de probabilités de transition entre chacune des lettres : si j'ai un A à la position n, qu'elles sont les probabilités de chacune des autres lettres de se retrouver à la position n+1. Tant le SA que la grammaire sont dérivés d'un modèle de Markov à variables cachées qui permet d'identifier la variabilité et les probabilités de transition entre chacune des lettres. Le modèle fut derivé indépendamment de deux ensembles de 250 protéines d'au moins 30 acides aminés et de moins de 30% de similarité de séquence de telle sorte que le SA soit parfaitement général.

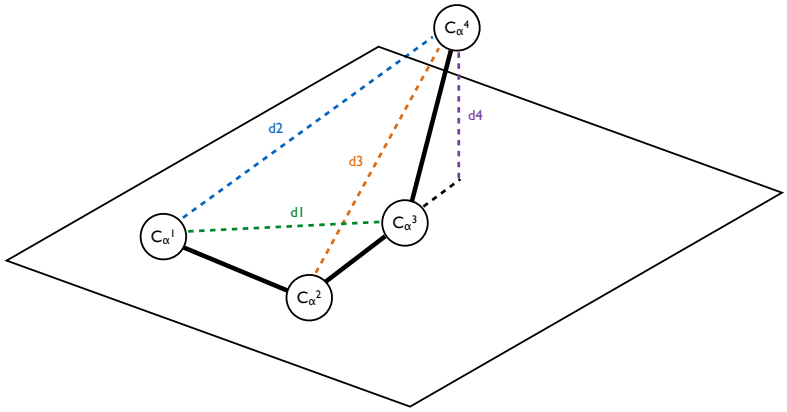

Figure 2.2 : Représentation graphique des quatres descriptifs du SA utilisé par PEPFOLD. Chacune des lettres est composé de quatre carbones α (quatre résidus). La distance d1 sépare le carbone C^1_α - C^3_α, d2 sépare C^1_α - C^4_α, d3 sépare C^2_α - C^4_α et d4 est la distance séparant le plan formé par C^1_α - C^2_α - C^3_α et C^4_α.

2.3.1.2 Algorithme Forward-Backward

L'algorithme Forward-Backward [36] est utilisé afin de déterminer les probabilités de chacune des lettres à encoder les fragments d'acide aminé. L'algorithme fonctionne en trois étapes. Premièrement, suivant l'initialisation du fragment i, les probabilités des fragments $i+1$, $i+2$, (...), $i+N$ sont déterminées par induction en prenant en considé-

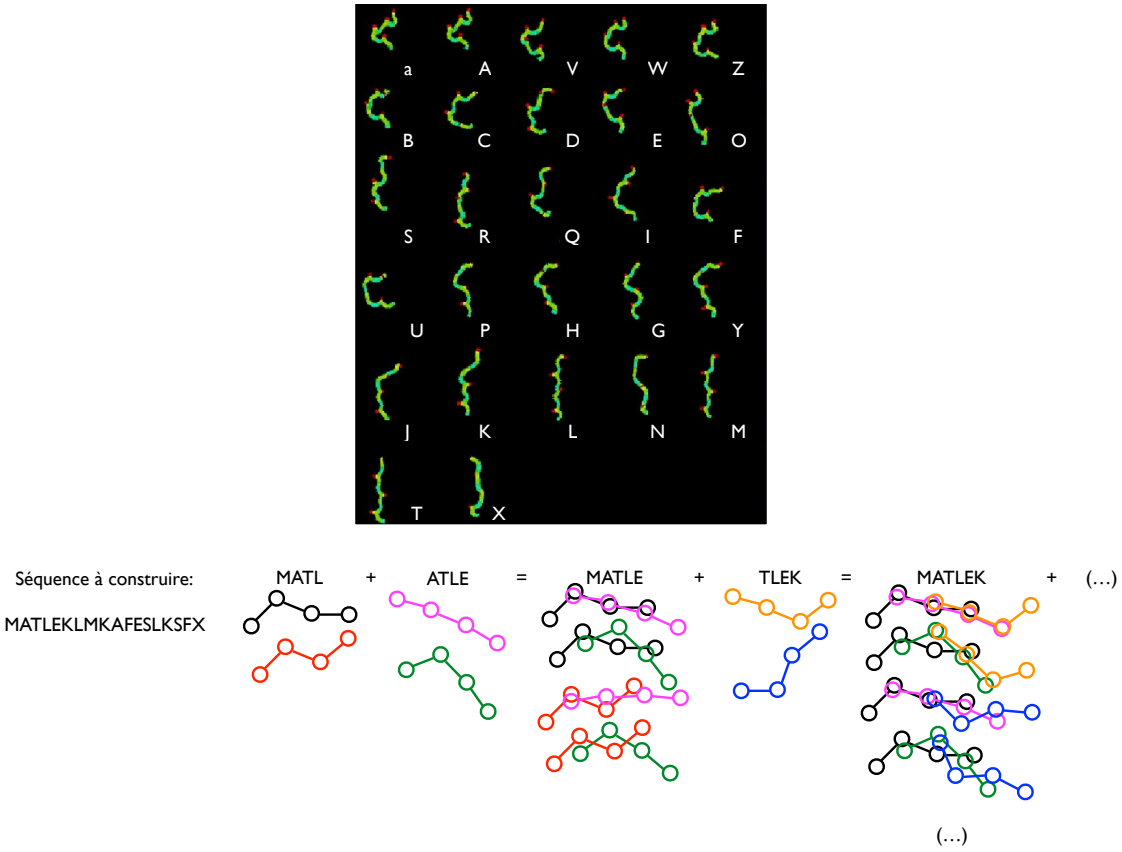

Figure 2.3 : (HAUT) : Structures des diverses lettres formant l'alphabet structurel utilisé dans PEPFOLD. L'image est tirée de [35]. (BAS) : Schéma de la reconstruction de la structure tridimensionnelle à partir de la séquence en acides aminés en superposant les trois derniers carbones α de la première lettre avec les trois derniers de la seconde. Les probabilités de chacune des lettres à représenter les quadruplets d'acide aminé ainsi que les probabilités de transition entre chacune des lettres sont dérivées d'un modèle de Markov à variables cachées.

ration toutes les lettres possibles pouvant résulter en l'acide aminé sélectionné. C'est la partie "Forward" de l'algorithme. De la même façon, on repart du fragment $i+N$ et on retourne vers le fragment i. C'est la partie "Backward" de l'algorithme. La dernière étape est le lissage des probabilités en combinant les résultats de la partie "Forward" et "Backward". Le résultat est maintenant une connaissance de la probabilité que chacune des lettres représente chacun des fragments d'acides aminés. Les lettres sont ensuite filtrées à chacune des positions pour conserver uniquement les plus probables.

2.3.1.3 Algorithme Glouton

L'algorithme Glouton (Greedy algorithm), est un algorithme qui tente de déterminer l'optimum global d'un problème en sélectionnant l'optimum local à chacune des étapes. Dans le cas qui nous intéresse, l'algorithme glouton tente d'identifier l'état fondamental et l'énergie fondamentale de la protéine en sélectionnant à chacune des étapes, les reconstructions de plus basse énergie [37]. Plus en détail, les lettres de construction sont assemblées de manière itérative. À partir d'une lettre initiale, la seconde est ajoutée en superposant ses trois premiers C_α avec les trois derniers de celle initiale (voir Figure 2.3). Ainsi, chaque lettre ajoutée agrandit la protéine d'un acide aminé [38]. L'idée derrière l'algorithme glouton est qu'après avoir généré la totalité des extensions possibles de la structure, seul un nombre maximal, H_{Best}, des meilleures structures sont conservées pour l'itération suivante. Il est important de noter que l'assemblage est local, mais que notre critère d'évaluation, l'énergie déterminée à l'aide du potentiel gros-grain sOPEP, est global ; une lettre moins appropriée localement pourrait permettre un meilleur assemblage des lettres subséquentes et améliorer la qualité globale de la reconstruction. L'algorithme glouton ne permet pas d'explorer cette possibilité, d'où la pertinence d'ajouter un élément stochastique. Ce dernier prend la forme suivante : aux H_{Best} meilleures structures sélectionnées à partir de toutes les structures générées C_i sont ajoutées des structures supplémentaires de l'ensemble restant (C_i-H_{Best}) [37]. En gardant certaines structures non-optimales pour la prochaine itération, on évite, en partie, les limitations énoncées précédemment.

2.3.2 BC-Score

Caractériser les similarités structurelles entre les protéines est essentiel pour l'identification et la classification de ces dernières. Le BC-Score (Binet-Cauchy score) est un outil développé récemment qui permet justement d'évaluer ces similarités de structure à une échelle locale. Ce score est normalisé entre 1 (structure parfaitement identique) et -1 (image miroir) où 0 est le BC-Score de deux conformations non-corrélées. Il est défini comme la somme normalisée des volumes de tous les tétraèdres dont les quatre

sommets sont formés par tous les triplets possibles de C_α et du centre géométrique de la protéine [39]. L'un de ces tétraèdres est présenté à la Figure 2.4. Il présente plusieurs avantages comparativement au calcul traditionnel de la déviation quadratique moyenne (RMSD). En effet, il (i) est indépendant de la taille de la protéine étudiée, (ii) permet une meilleure discrimination à moyenne portée, i.e. qu'un faible RMSD correspond à des structures semblables, mais qu'un large RMSD peut difficilement être relié à la présence/absence de structure et (iii) est plus simple et rapide à calculer que le RMSD [39].

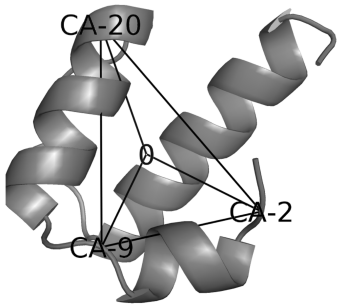

Figure 2.4 : Représentation graphique du tétraèdre formé par un triplets de carbone α (dans cette situation les carbones α des résidus 2, 9 et 20) avec le centre de la protéine O. La figure présentée ci-dessus est tirée de [39].

CHAPITRE 3

PROTÉINE HUNTINGTINE : UNE INTRODUCTION

3.1 La maladie d'Huntington

La maladie d'Huntington (HD) est une maladie neurodégénérative héréditaire et orpheline ; elle affecte entre 5 et 10 individus par tranche de 100 000. Elle est caractérisée par une vaste gamme de symptômes, variant considérablement selon le patient, mais que l'on peut regrouper en trois catégories [40]. Il y a premièrement les troubles moteurs comme les mouvements involontaires de type chorée, souvent le premier symptôme à se manifester, avec la perte de coordination et d'équilibre. On retrouve ensuite les symptômes psychiatriques comme la dépression, la démence et l'anxiété. Finalement, il y a les troubles cognitifs comme la perte de conscience de soi (self loss), de conscience spatiale et les pertes de mémoire. Ces exemples ne représentent qu'un petit échantillon de tout le spectre de symptômes que peut prendre la maladie. Au décès, on constate des pertes importantes (pouvant atteindre 30%) au niveau de la masse du cerveau selon la gravité de la maladie [41]. Il n'y a encore à ce jour aucun traitement menant à la guérison de HD.

Au niveau moléculaire, HD est causée par l'expansion anormale de la répétition du trinucléotide CAG dans le premier exon (Exon1) du gène *HTT* (voir Figure3.1). Cette caractéristique est partagée par huit autres maladies neurodégénératives dont six types d'ataxie spinocérébelleuse, la maladie de Kennedy (ou amyotrophie bulbo-spinale) et la maladie de Naito-Oyanagi. Le gène *HTT* contient l'information permettant de synthétiser la protéine Huntingtine (Htt). De son côté, la répétition du trinucléotide CAG encode un segment poly-glutamines (Q_N) qui se retrouve dans la région N-terminale de Htt. En circonstances normales, Q_N contient entre 16 et 20 répétitions de la glutamine. Par contre, l'expansion de Q_N au-dessus du seuil critique de 36 glutamines est associée à un repliement aberrant/mutant de Htt (mHtt). Ce mauvais repliement cause une transition pathologique associée à un gain/perte de fonction(s) toxique(s) et mène à son

agrégation en fibres amyloïdes. Plus Q_N est long, plus l'apparition de la maladie se fera précocement et plus les symptômes seront sévères [40].

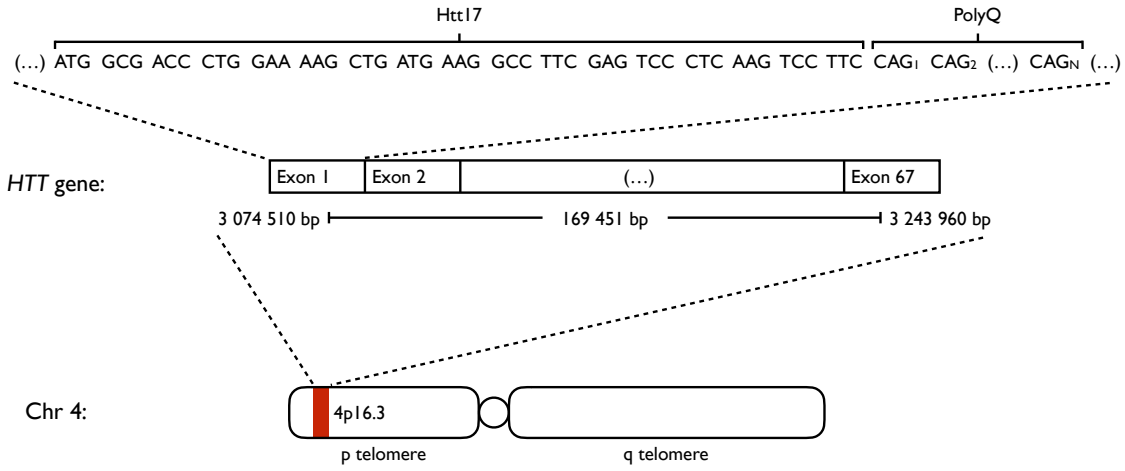

Figure 3.1 : Représentation schématique du gène *HTT* et de son positionnement sur le chromosome quatre.

Penchons-nous maintenant sur les conséquences de HD au niveau moléculaire [42]. Plusieurs observations indiquent qu'une mauvaise régulation des mitochondries est associée à HD [43]. En effet, mHtt se lie à la membrane externe de ces dernières et mène à la détérioration de la chaîne de transport d'électrons. De plus, mHtt fait obstacle au transport rétrograde et antérograde des mitochondries dans les axones empêchant la déposition de celles-ci aux sites à grande demande énergétique. Ces mécanismes pourraient causer le dysfonctionnement des neurones par des bouleversements du métabolisme énergétique et la promotion de dommage oxydatif. mHtt est aussi associée à une mauvaise régulation de la transcription [44] (mécanisme permettant de traduire le code de l'ADN en ARNm) en entravant le travail de plusieurs promoteurs et initiateurs de ce processus. Les méfaits de mHTT ne s'arrêtent pas là puisque de nombreux autres phénomènes sont aussi affectés par HD. Ainsi, on note la perte d'efficacité et une surutilisation de la protéostasie (pour la dégradation) mises en évidence par l'augmentation de l'agrégation et de la mauvaise localisation de protéines métastables, l'augmentation de protéines mal repliées et la perte de réponse suite à un choc thermique [42].

3.2 La protéine Huntingtine : structures et fonctions

Nous nous sommes intéressés dans la précédente section aux conséquences physiologiques, cellulaires et moléculaires de HD. Dans celle-ci, nous étudierons en détail la structure et les fonctions de Htt, qui, rappelons-le, est la protéine à l'origine de HD suite à l'expansion anormale de Q_N.

Au niveau de sa structure, Htt est une grande protéine comptant 3 144 acides aminés. Une version schématique de Htt^{NT} est présentée à la Figure 3.2. Son expression commence tôt, durant le développement embryonnaire, et se continue jusqu'à l'âge adulte. Htt est présente dans une grande diversité de tissus, mais plus particulièrement dans le système nerveux. Elle se retrouve principalement hors noyau, contrairement à la version mutante, mHtt, qui elle, s'accumule progressivement dans le noyau [45].

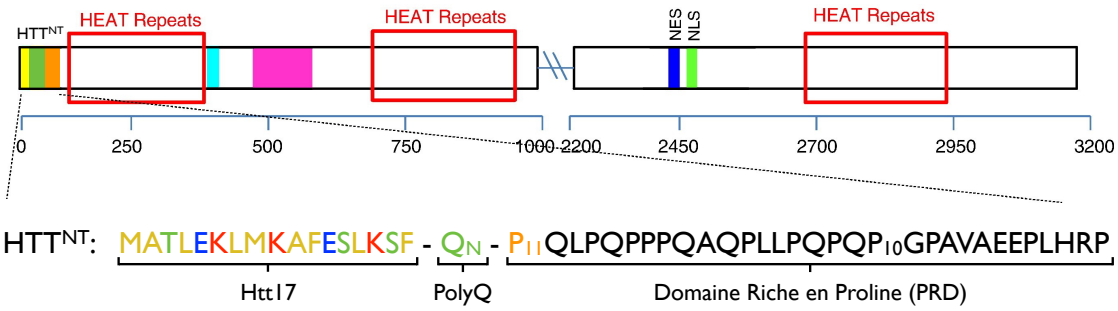

Figure 3.2 : Séquence d'acides aminés de Htt et plus spécifiquement de Htt^{NT}. Figure adaptée de [46].

La (très) vaste majorité des efforts de caractérisation de Htt vise la région du N-terminal (Htt^{NT}) (encodée par l'Exon1). Cela s'explique par le fait que cette région, englobant le segment Q_N, est suffisante pour reproduire le phénotype de HD *in vitro* [47] et *in vivo* [48]. En effet, l'analyse *post mortem* d'agrégats extraits de tissus cérébraux de patients atteints de HD montre qu'ils sont formés de Htt^{NT} plutôt que de la protéine entière [45] Comme l'Exon1 revêt un intérêt particulier, la prochaine section lui sera dédiée entièrement.

Pour ce qui est de la structure des 66 autres exons correspondants à 98% de la protéine, des analyses bio-informatiques ont identifié entre 16 et 36 répétitions HEAT for-

mant des structures hélicoïdales et permettant l'interaction inter/intra moléculaire [49]. Htt adopterait une panoplie de structures distinctes selon ses multiples partenaires d'interaction [50]. Globalement, cette partie de la protéine est beaucoup moins connue que Htt^{NT}. Par contre, de récents résultats *in vivo* et *in vitro* montrent que le segment C-terminal, obtenu lors de la protéolyse de Htt, inactive la dynamine-1 de la membrane du réticulum endoplasmique (ER) et mène à la vacuolisation et à la mort de celui-ci [51]. Ainsi, le fragment C-terminal serait toxique au même titre que le fragment N-terminal et l'étude de ce premier pourrait rapidement devenir incontournable.

Au niveau cellulaire, l'intégralité de la structure de Htt peut être altérée de façon plus ou moins importante. En effet, Htt peut être sujet à la protéolyse (le clivage de la protéine par des enzymes appelées peptidases) en de multiples sites et par une multitude de peptidases différentes. Ce phénomène est fondamental, puisque la protéolyse peut mener à la génération de fragments Htt^{NT} (contenant le segment Q_N) dont la migration vers le noyau résulte en l'apparition d'effets toxiques [52]. En plus de la génération de Htt^{NT}, la protéolyse peut aussi modifier les fonctions de Htt et activer l'apoptose (la mort cellulaire programmée) [51]. Outre la protéolyse, il fut démontré qu'un clivage aberrant de l'Exon1 de *HTT* mène à la transcription d'un petit ARNm polyadénylé qui est ensuite traduit en un peptide correspondant à Htt^{NT} [53].

En plus de la protéolyse, Htt peut aussi subir des transformations post-traductionnelles : phosphorylation (ajout d'un PO_3^{2-}), acétylation (ajout d'un CO-CH_3), palmitoylation (ajout d'un acide gras généralement sur une cystéine), ubiquitylation (ajout d'une Ubiquitine) et sumoylation (ajout d'une protéine SUMO sur une lysine) pour n'en nommer que quelques-unes. Ces modifications post-traductionnelles jouent un rôle important en modifiant la(les) fonctions(s) de Htt. Par exemple, l'acétylation de mHtt permet d'en régir la clairance (son élimination par un tissu) [54] et sa phosphorylation en S434/S536 atténue la protéolyse et la toxicité du fragment Q_N [55].

Biologiquement, Htt joue un rôle complexe et encore majoritairement incompris [56]. En effet, on lui a identifié plus de 350 partenaires d'interaction impliqués dans une foule de processus comme la transcription et la maintenance de l'ADN, l'épissage de l'ARN, l'endocytose, le transport, le métabolisme énergétique, l'homéostasie et la signalisation

cellulaire pour n'en nommer que quelques-uns.

Htt (ou certains de ses fragments) est aussi impliquée dans de nombreux processus associés à la membrane, comme le transport intra-membranaire. Htt se retrouve tant au niveau du cytoplasme que du noyau et il fut démontré qu'elle se lie à la membrane de multiples organites tel le ER, l'appareil de Golgi, les endosomes et les mitochondries [57, 58].

3.3 N-terminal

Dans cette section, nous nous pencherons sur le segment Htt^{NT} correspondant au peptide encodé par l'Exon1 du gène *HTT* et situé dans la région N-terminale de Htt. Il s'agit d'un segment incontournable puisqu'il est à l'origine du phénomène d'agrégation et de toxicité présent dans HD.

Htt^{NT} est composé de trois segments intrinsèquement désordonnés lorsqu'ils sont pris sous la forme de monomère. Le premier est une séquence de 17 acides aminés (Htt17) au caractère amphipathique et dont la séquence exacte est donnée par MATLEKLMKAFESLKSF. Le second est le fameux segment Q_N dont la taille (N) peut varier (N > 36 mène à la forme pathogène). Le dernier est une séquence de 38 acides aminés riche en proline (C38) et commençant par 11 prolines (P_{11}) (voir Figure 3.2). La suite de cette section passera en revue les différentes fonctions, structures et rôles joués dans l'agrégation de chacun de ces segments pris individuellement puis collectivement.

Malgré le consensus que la toxicité de Htt provient essentiellement de Htt^{NT}, le processus par lequel l'expansion du segment Q_N au-dessus du seuil critique l'affecte/le modifie est encore aujourd'hui inconnu. Deux hypothèses pourraient expliquer l'apparition de la toxicité [46]. La première est que l'expansion de Q_N introduit un changement de conformation au niveau du monomère de Htt^{NT} et que ce serait ce monomère qui, via un gain/perte de fonction, serait responsable de la toxicité. La seconde est que la toxicité viendrait d'un des différents types d'agrégats dont la présence dépendrait/serait favorisée dans le cas mutant. Avant de s'attaquer à ces deux hypothèses, regardons pour commencer la structure individuelle de chacun des segments de Htt^{NT}.

3.3.1 Q_N

Des observations d'agrégats de peptides contenant le fragment Q_N dans le noyau de neurones d'humains [45] et d'animaux [48] atteints de HD combinées au fait que son expansion est à l'origine de HD ont mené à l'hypothèse que, dans le cas mutant, la protéine Htt subirait un mauvais repliement associé à un gain/perte de fonction toxique. C'est pourquoi de multiples expériences se sont concentrées sur la structure de Q_N seul afin d'en déceler la structure et l'hypothétique mauvais repliement.

Des expériences RMN [59, 60] montrèrent qu'en solution, Q_N est flexible, n'adopte aucune structure stable et est exposé au solvant. De plus, ces dernières démontrèrent que peu importe la taille de N, Q_N adopte sensiblement la même structure indiquant que Q_N serait intrinsèquement toxique et que celle-ci serait dépendante de la taille de Q_N. Pour complémenter les résultats expérimentaux, de multiples simulations montrèrent que Q_N adopte une grande variété de structures allant de complètement désordonnées [61] à désordonnées mais avec la formation de feuillet β [62] ou d'hélice α [63].

3.3.2 Htt17 et C38

Si le domaine Q_N est crucial pour l'apparition de HD, les deux segments l'encadrant, Htt17 et C38, sont aussi fondamentaux puisqu'ils influencent la structure et la localisation de Htt^{NT}.

3.3.2.1 C38

Le fragment C38 situé au C-terminal du fragment Q_N est un segment riche en proline et commençant par 11 prolines consécutives (P_{11}). Des études *in vitro* de Htt^{NT} montrèrent que l'agrégation est semblable avec C38 ou uniquement $P_{10/11}$ tant au niveau de la morphologie que de la cinétique [64]. L'introduction de P_{10} au segment Q_N diminue le taux de formation et la stabilité des agrégats amyloïdes en stabilisant, hypothétiquement, une conformation du monomère non propice à l'agrégation [65]. Ces effets ne sont observés que lorsque P_{10} est connecté à Q_N via le C-terminal. Au niveau de sa structure, $P_{10/11}$ aurait tendance à former une hélice Poly-Proline de type-II (PPII),

mais aucune conformation ne serait particulièrement favorisée [46].

3.3.2.2 Htt17 en solution

Le segment Htt17 revêt un intérêt particulier puisque son ajout au segment Q_N change drastiquement le taux (augmentation marquée) et le mécanisme d'agrégation comparativement à Q_N seul [66] et sa séquence est fortement conservée chez les vertébrés [58]. Une étude détaillée de sa structure et de ses fonctions s'impose.

Au niveau du monomère de Htt17, de multiples expériences de dichroïsme circulaire (CD) indiquent la présence de structure α avec des proportions de 10% [67], 34% [68], 45% [58] et 55% [66] selon le protocole expérimental utilisé. Des analyses de résonnance magnétique nucléaire (RMN) suggèrent quant à elles que Htt17 ne possède pas de structure secondaire stable et que sa conformation serait plutôt désordonnée et aléatoire. On retrouve malgré tout un certain niveau de structure α passagère et localisée tout particulièrement entre les résidus Thr3 et Glu5 [66]. De plus, Htt17 adopterait une structure compacte et résistante à l'agrégation [66]. À cause de sa grande flexibilité et de son caractère désordonné, il n'y a toujours aucune structure expérimentale de Htt17 en solution.

Les résultats expérimentaux sont complémentés par de multiples simulations numériques. Ces dernières permettent d'obtenir un portait de Htt17 au niveau atomique, mais les résultats sont variés et dépendent grandement de la méthodologie, du traitement du solvant et du potentiel utilisé.

Des simulations de recuit simulé réalisées avec le potentiel AMBER2003 montrent que Htt17 a une forte tendance à former des hélices α, tout particulièrement entre les résidus L4 et K9 (voir Figure 3.3). Les états les plus représentés sont, dans l'ordre, deux segments hélice α séparés par un coude autour d'A10 et une hélice α complète. Dans les deux cas, le peptide a un caractère amphipathique [69]. D'un autre côté, des simulations de métadynamique aux échanges biaisés sur le monomère de Htt17 en solution montrent que Htt17 peuple quatre bassins cinétiques distincts. Dans le premier, représentant près de 75% de la population totale, on retrouve Htt17 en conformation désordonnée dont les résidus hydrophobes sont accessibles au solvant. Dans le second et le troisième,

réunissant près de 21% de la population totale, Htt17 adopte une structure d'hélice α entre les résidus 1-11 et 1-7 respectivement. Finalement, dans le dernier bassin contenant seulement 4% de la population totale, Htt17 adopte une conformation globulaire. Globalement, les auteurs évaluent à 29% le contenu en hélice α [70].

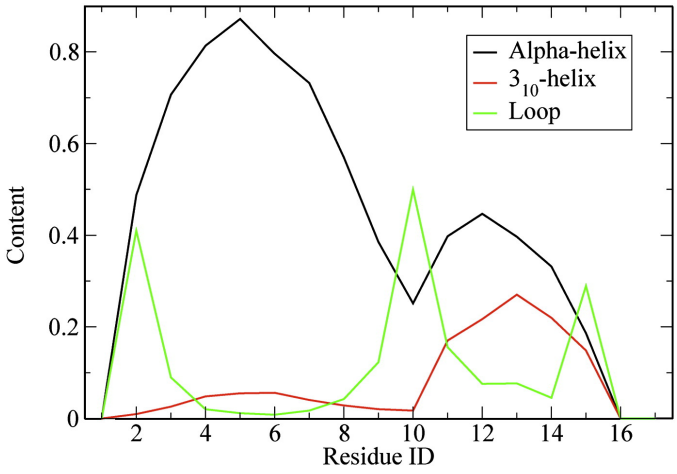

Figure 3.3 : Structure secondaire de Htt17 en fonction par résidus déterminée par simulation de recuit simulé. L'hélice α, l'hélice 3-10 et les boucles sont présentés respectivement en noir, rouge et vert. La figure est tirée de [69].

Bref, tant les résultats expérimentaux que numériques montrent que Htt17 en solution est plutôt désordonné avec une tendance plus ou moins forte à la formation de structure α.

3.3.3 Htt17 en membrane

En plus de sa présence en solution, Htt17 est indispensable pour l'interaction avec les membranes cellulaires et avec de nombreux organites. En effet, Htt17 agirait tel un signal de rétention cytosolique. Il est aussi requis pour l'association de Htt avec les mitochondries, le ER [57, 58] et l'appareil de Golgi [57].

Des expériences RMN en deux dimensions (2D-RMN) ont permis de déterminer la structure de Htt17 dans une solution de micelles composées de dodecylphosphocholine (DPC) et de déterminer son orientation dans une bicouche de phosphatidylcholine (PC).

Comme vu précédemment, en solution, Htt17 est principalement désordonné. Par contre, l'ajout de micelles induit un changement de structure de Htt17 qui adopte maintenant une conformation d'hélice α (>70%) au caractère amphipathique entre les résidus 6 et 17 [67]. Les résultats suggèrent qu'une structure similaire est présente dans la bicouche. Au niveau de l'orientation, Htt17 est orienté presque parallèlement à la membrane de la micelle. Les résidus hydrophobes pointent vers le centre tandis que les résidus hydrophiles s'accumulent sur la face opposée, vers le solvant. En plus des expériences en micelles, l'association de Htt17 avec la membrane d'une bicouche fut quantifiée en fonction de la composition lipidique de celle-ci. Ces expériences démontrent que Htt17 mène à une déstabilisation membranaire qui est plus prononcée pour des vésicules de POPC ou POPC/POPS tandis que la déstabilisation est réduite avec l'ajout de cholestérol. Peu importe la composition de la membrane, l'insertion de Htt17 diminue le paramètre d'ordre membranaire, une mesure de la flexibilité des lipides. De plus, Htt17 possède certaines caractéristiques clés, partagées par plusieurs autres peptides causant le rétrécissement et la courbure de la membrane [71]. Au niveau de son orientation, Htt17 est légèrement décalée par rapport à la surface de la bicouche, indiquant une possible interaction intermoléculaire médiée par la formation de ponts salins avec d'autres segments Htt17.

3.4 Apparition de la toxicité

Après avoir étudié en détail la structure du monomère de Q_N, C38 et Htt17, il faut maintenant replacer ces derniers dans le contexte plus global du peptide Htt^{NT} et regarder comment les résultats s'accordent avec les deux hypothèses concernant l'apparition de la toxicité.

3.4.1 Hypothèse 1 : Changement de structure

Une des hypothèses qui pourrait expliquer l'apparition de la pathologie est que le prolongement du segment Q_N au-dessus du seuil critique causerait un changement de conformation toxique au niveau du monomère de Htt^{NT}. Cette section offre un aperçu des différents changements structurels de Htt^{NT} selon la taille de Q_N et les différents

domaines présents.

Au niveau du monomère, tant Htt17 que Q_N sont majoritairement désordonnés en solution. Cette flexibilité structurelle fait en sorte qu'il est extrêmement complexe de les étudier expérimentalement au niveau atomique. Sur le sujet, des expériences de diffraction par rayon-X proposèrent une structure cristalline de Htt^{NT} attaché à une MALTOSE-BINDING-PROTEIN via le N-terminal avec $Q_{N=17}$ [72]. Dans ce contexte où Htt^{NT} est complet, Htt17 adopte une hélice α, le segment Q_{17} forme une hélice α dans sa région N-terminale mais demeure plus flexible par la suite et le domaine P_{11} forme une hélice PPII. Il faut par contre noter que la cristallisation de cette molécule résulta en la formation de trimères qui sont fort probablement un artefact de la méthode.

L'addition du segment Q_N à Htt17 changerait la structure de ce dernier pour une conformation plus étendue et propice à l'agrégation. Une expansion du segment Q_N favoriserait l'agrégation de deux façons : (i) en modifiant la structure compacte de Htt17 pour une nouvelle plus favorable et (ii) en augmentant l'efficacité de la nucléation [66].

Contrairement à l'expérience, les simulations numériques permettent aisément de réaliser des études au niveau atomique et plusieurs vinrent compléter les résultats expérimentaux.

Des simulations de Monte-Carlo/Metropolis sur Htt17-Q_N avec le solvant implicite ABSINTH [68] montrent que tant le segment Htt17, qui au départ a tendance à former des structures α, que Q_N deviennent de plus en plus désordonnés avec l'allongement de Q_N. Les résidus hydrophobes de Htt17 sont séquestrés dans le domaine Q_N causant une séparation claire des résidus hydrophiles. Cette séparation s'accentue avec l'agrandissement de Q_N. D'autres simulations tout-atome de MD discrète sur Htt17-Q_N-P_{11}-P_{10} en solvant implicite [62] étudièrent les effets des divers fragments de Htt^{NT} sur le segment Q_N. Lorsque Q_N est seul en solution, l'augmentation de sa taille augmente sa propension à former des structures riches en feuillets β tandis que l'ajout du segment P_{11}, formant une hélice PPII, réduit la tendance de Q_N à former des feuillets β. Finalement, le segment Htt17 adopte une hélice α dans la protéine native, mais l'agrandissement du segment Q_N cause un mauvais repliement modifiant Htt17 en un feuillet β. Finalement, des simulations tout-atome d'échange de répliques (125 μs x 8 répliques) réalisées sur

Htt17-Q_N-P_{11} avec le potentiel AMBER03 et un traitement implicite du solvant (Generalized Born) montrèrent que tant pour un segment Q_N ordinaire (N=17) que mutant (N=55), Htt17 adopte une hélice α du résidu 4 à 17, la région Q_N forme aussi une hélice α et la région P_{11} adopte une hélice PPII [73]. La différence de structure induite par une expansion du segment Q_N est une augmentation de l'accessibilité au solvant de Htt17 dans le cas mutant causée par une perte d'interaction avec le segment P_{11}.

En bref, nous sommes encore bien loin d'un consensus quant à la structure qu'adopte Htt^{NT} en solution et des changements structurels causés par une élongation de Q_N. En effet, les résultats expérimentaux au niveau atomique sont peu nombreux et les résultats numériques varient grandement selon la méthodologie, le potentiel et le traitement du solvant utilisé.

3.4.2 Hypothèse 2 : Agrégation

La seconde hypothèse serait que la toxicité ne viendrait pas du monomère mutant Htt^{NT} à proprement parler, mais plutôt que le phénomène d'agrégation et de fibrillation serait modifié de telle sorte qu'une (ou plusieurs) espèce serait maintenant toxique. Pour vérifier cette hypothèse, nous réviserons dans cette section les diverses connaissances entourant l'agrégation de Htt.

Des expériences de centrifugation analytique montrèrent que Htt17 seul ou suivi par un segment Q_{10} peut former des oligomères dont la distribution comprend le monomère, le tétramère, l'octomère et ainsi de suite pour les plus hauts ordres [74]. Le segment Q_N n'est pas requis pour la formation de ces oligomères. Semblablement, des analyses de spectrométrie de mobilité ionique (MS) révèlent que Htt17 se trouve sous la forme d'oligomères allant du monomère au tétramère [75]. Ces oligomères ont un large contenu d'hélice α et seraient cruciaux pour le développement des fibres.

Seuls les peptides dont le segment Q_N surpasse le seuil de huit glutamines maturent en agrégats amyloïdes par un lent processus d'apparition de feuillet β [74]. Un certain nombre d'éléments possédant les motifs d'hélice α est tout de même conservé dans l'agrégat amyloïde final.

Afin d'étudier l'effet de l'expansion du segment Q_N sur l'ensemble conformationnel

d'agrégation de Htt^{NT}, des expériences de spectroscopie de corrélation de fluorescence (FCS) furent réalisées. *In vitro*, ils montrèrent que la formation de dimères et de tétramères par le Htt^{NT} mutant dépend de la grandeur du segment Q_N et est suivie de la formation d'oligomères fibrillaires et sphériques puis par l'apparition de fibres amyloïdes. Une expérience similaire réalisée *in vivo* montre que le monomère de Htt^{NT} mutant est absent. Htt^{NT} mutant forme plutôt des oligomères dont la plus petite entité détectée est le tétramère. En fonction du temps d'incubation, on retrouve ensuite de petits agrégats puis de larges agrégats sédimentables. Des dégâts significatifs à l'ADN apparaissent dans ces cellules après six heures d'incubation [76].

Dans le modèle de nucléation classique, l'agrégation est décrite comme une série d'équilibres thermodynamiques pendant lesquels un monomère se joint à une espèce A_N pour former l'espèce A_{N+1}. L'addition d'un monomère à une espèce préexistante n'est pas favorable au départ, jusqu'à ce que l'espèce en question atteigne la taille du noyau critique (A_{N*}) à partir duquel l'ajout de monomère est plus favorable que son détachement. Le modèle de l'agrégation de Htt^{NT} développé à partir des éléments mentionnés ci-dessus est présenté de façon schématique à la Figure 3.4-(A). Le processus proposé est relativement similaire au modèle classique. En effet, Htt^{NT} est à l'équilibre entre le monomère et les divers oligomères dont la plus petite entité est le tétramère. Ces oligomères permettent un rapprochement entre les divers domaines Q_N qui font alors la transition vers des structures β et formeront des fibres amyloïdes. Ce modèle propose deux rôles principaux aux oligomères de Htt^{NT} : (i) certains formeront des noyaux à partir duquel se développera la croissance de la fibre et (ii) ceux ne participant pas à la nucléation serviraient de réservoir à monomère et supporteraient la croissance de la fibre [74].

De plus, l'agrégation pourrait être affectée de façon majeure par la présence de la membrane cellulaire/organite. En effet, le caractère amphipathique de Htt17 permet l'interaction réversible avec la membrane. De plus, son orientation, laissant le résidu PHE17 et le domaine Q_N exposés à l'eau et libres d'interagir avec d'autres domaines similaires, permettrait d'augmenter la concentration locale de Q_N à surface de la membrane et favoriserait le développement de structure β, d'agrégats toxiques et de fibres amyloïdes. Ce

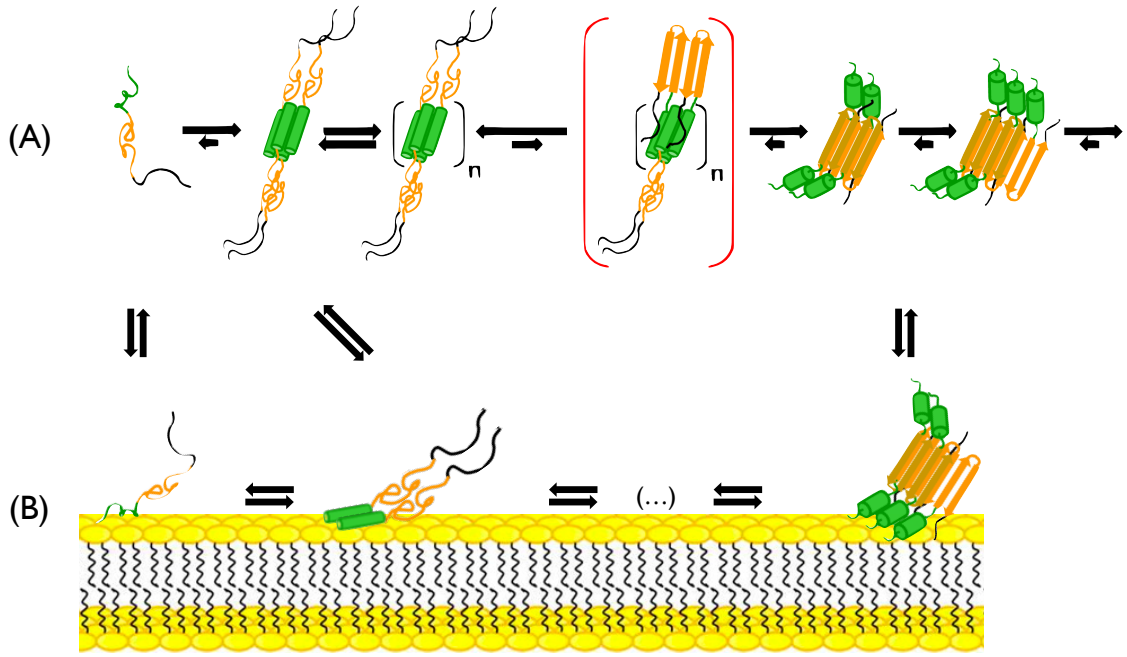

Figure 3.4 : (A) Modèle d'agrégation et de formation de fibres amyloïdes par Htt^{NT} en solution. Le monomère est initialement en équilibre avec le tétramètre et d'autres plus gros oligomères multiples de quatre. Suite à la formation du noyau critique, il y a apparition de structures β caractéristiques des fibres amyloïdes. (B) Modèle proposé lors de la présence de la membrane. Les segments Htt17, Q_N et C38 sont présentés respectivement en vert, orange et noir. La figure est adaptée de celle présentée dans [76].

phénomène pourrait être amplifié par une possible oligomérisation de Htt17 sur la membrane, similaire à celle proposée en environnement aqueux [67, 71]. Bref, l'ancrage de Htt17 dans la membrane pourrait aussi jouer un rôle clé dans l'agrégation en favorisant le processus de nucléation et l'agrégation.

Les résultats de simulation au niveau de la dimérisation semblent quant à eux montrer que la probabilité d'association observée est toujours plus grande pour Q_N seul que Htt17-Q_N et donc que Htt17 diminuerait la tendance naturelle de Q_N à former des oligomères. Dans ces simulations, la dimérisation est principalement le résultat de contacts entre les résidus glutamines et non entre les segments Htt17 [68].

Une nouvelle analyse de ces mêmes simulations à l'aide d'un modèle ultra-gros-grain et de dynamique brownienne [77] montra que Htt17 réduit l'enchevêtrement inter-

peptide des chaînes Q_N. En l'absence de Htt17, les petits agrégats sphériques et les larges agrégats se forment à différentes échelles de temps. En présence de Htt17, la réduction de l'enchevêtrement et l'ajout d'une orientation à l'association inter-peptide (causée par l'addition de Htt17) agissent de telle sorte que les petits et gros agrégats se forment sur la même échelle de temps.

Similairement, une combinaison de microscopie électronique et d'expériences de fluorescence [78] montra que le segment Htt17 augmente la tendance de Htt^{NT} à former des agrégats insolubles, accélère la formation de fibres et déstabilise les intermédiaires non-fibrillaires. De son côté, le segment C38 diminue la tendance de Htt^{NT} à former des agrégats insolubles, soit l'effet inverse de Htt17. Ensemble, il y a déstabilisation des intermédiaires non-fibrillaires due à Htt17 et diminution de la tendance à former des agrégats insolubles due à C38, le tout favorisant la formation de fibres ordonnées.

3.5 Motivation

Le fragment Htt^{NT} de la protéine Htt joue un rôle crucial dans le processus d'agrégation et la localisation de la protéine Htt. De plus, ce segment (sans la section C-terminale) est suffisant pour reproduire la toxicité présente dans HD. Il est composé de trois domaines : le domaine Htt17, le domaine Q_N et le domaine C38 commençant par P_{11}. Une connaissance poussée et au niveau atomique de la structure de ces trois fragments, en membrane et en solution, est requise et nécéssaire à l'obtention d'une meilleure compréhension de la maladie et au futur développement d'un traitement.

La flexibilité intrinsèque de Htt^{NT} est à l'origine de bien des casse-tête pour les expérimentalistes et il n'y a à ce jour aucune proposition de structure expérimentale. Les simulations numériques forment une approche complémentaire à l'expérience et permettent une description au niveau atomique. Si plusieurs furent réalisées sur une combinaison ou la totalité des trois domaines de Htt^{NT}, les résultats dépendent grandement de la méthodologie appliquée, du champ de force et du traitement du solvant [62, 68, 73, 77]. Or, il n'y a présentement aucune simulation réalisée sur Htt^{NT} en solvant explicite. Il devient donc nécessaire de réaliser des simulations à l'aide des méthodes de simula-

tion les plus puissantes et les plus avancées afin d'offrir un portrait robuste de l'ensemble structurel de Htt^{NT} en solution et de minimiser les effets de la méthode sur les observations. Pour combler ces lacunes, nous avons réalisé des simulations tout-atome en solvant explicite à l'aide du potentiel AMBER99sb*-ILDN, un potentiel reconnu comme particulièrement adéquat pour l'étude de Htt17 [79], et à l'aide des méthodes de simulation les plus avancées alliant WT-MetaD et HREX. Une description détaillée de ces deux méthodes se trouve à la section 2.2.2 et 2.2.1.

Nos études de Htt^{NT} sont réalisées le long de deux axes : en membrane et en solvant. Dans un premier temps, nous avons réalisé des simulations du monomère de Htt17 dans une bicouche lipidique. Ces simulations avaient comme objectifs d'étudier le modèle expérimental de Htt17 en environnement de micelles, de décrire les possibles modifications induites par le remplacement de la micelle par une bicouche, d'identifier des motifs structuraux à l'origine de son interaction avec la membrane et finalement de quantifier les répercussions de son insertion sur la membrane. La validité de nos simulations fut confirmée grâce à la comparaison avec plusieurs mesures expérimentales tels les déplacements chimiques et la profondeur d'insertion du peptide. Un résumé des principaux résultats est présenté à l'annexe I. Dans un second temps, nous avons étudié l'ensemble structurel des monomères de Htt17, Htt17Q_{17} et Htt17$Q_{17}P_{11}$ en solution avec comme objectifs d'identifier les motifs structuraux qui pourraient être à l'origine du phénomène d'agrégation ou de l'interaction avec la membrane ainsi que d'obtenir un portrait quantitatif de l'impact qu'ont les différents fragments les uns avec les autres. Il est essentiel, avant de se lancer dans l'étude des systèmes plus complexes (Htt^{NT} avec un segment Q_N mutant ou l'oligomérisation), d'obtenir des résultats robustes sur l'ensemble structurel du Htt^{NT} sain. Nous avons déterminé qu'un segment de 17 glutamines (Q_{17}) permettait de garder les temps de simulation dans la limite du raisonnable tout en permettant une comparaison directe avec la structure cristalline [72] et en permettant un aperçu pertinent des phénomènes biophysiques. De cette façon, les modifications structurelles causées par une élongation du segment Q_N pourront être comparés avec ces simulations de référence. Nous avons néanmoins tenté d'identifier, à partir de nos informations sur le monomère, certains motifs qui pourraient jouer un rôle important dans l'oligomérisa-

tion et l'ancrage à la membrane. Des études plus détaillées devront être réalisées afin de confirmer/infirmer nos hypothèses. La qualité de nos simulations fut établie en comparant avec les mesures expérimentales de RMN et de CD. Ces résultats sont présentés au chapitre 4.

CHAPITRE 4

ÉTUDE DE LA PROTÉINE HUNTINGTINE EN SOLUTION

Ma Contribution : Pour cet article, je mis en place et réalisai toutes les simulations de HREX-MetaD pour chacun de nos trois systèmes. J'analysai en totalité les résultats pour les simulations de HREX-MetaD et en partie pour ceux de PT-MetaD, placés dans les matériels supplémentaires. Au niveau de l'écriture de l'article, ma contribution se retrouve principalement à la section résultat, analyse et dans les matériels supplémentaires. Je réalisai aussi les figures présentées.

Free energy Landscape of the Amino-terminal Fragment of Huntingtin in Aqueous Solution

Vincent Binette$^{\triangle}$, Sébastien Côté$^{\triangle}$, and Normand Mousseau*

$\triangle$ Vincent Binette and Sébastien Côté contributed equally to this work.

*Correspondence : normand.mousseau@umontreal.ca.

Editor : Michael Feig.

Article publié dans le Biophysical Journal : Binette, V., Côté, S., & Mousseau, N. (2016). Free-Energy Landscape of the Amino-Terminal Fragment of Huntingtin in Aqueous Solution. Biophysical journal, 110(5), 1075-1088.

4.1 Abstract

The first exon of Huntingtin – a protein with multiple biological functions whose misfolding is related to Huntington's disease – modulates its localization, aggregation and function within the cell. It is composed of a 17-amino-acid amphipathic segment (Htt17), an amyloidogenic segment of consecutive glutamines (Q_N), and a proline-rich segment. Htt17 is of fundamental importance : it serves as a membrane anchor to control the loca-

lization of huntingtin, it modulates huntingtin's function through post-translational modifications, and it controls the self-assembly of the amyloidogenic Q_N segment into oligomers and fibrils. Experimentally, the conformational ensemble of the Htt17 monomer as well as the impact of the polyglutamine and proline-rich segments remain, however, mostly uncharacterized at the atomic level due to their intrinsic flexibility. Here, we unveil the free energy landscape of Htt17, Htt17Q_{17} and Htt17$Q_{17}P_{11}$ using Hamiltonian replica exchange combined to well-tempered metadynamics. We characterize the free energy landscape of these three fragments in terms of a few selected collective variables. Extensive simulations reveal that the free energy of Htt17 is dominated by a broad ensemble of configurations that agree with solution NMR chemical shifts. Addition of Q_{17} at its carboxy-terminus reduces the extent of the main basin to more extended configurations of Htt17 with lower helix propensity. Also, the aliphatic carbons of Q_{17} partially sequester the non-polar amino acids of Htt17. For its part, addition of $Q_{17}P_{11}$ shifts the overall landscape to a more extended and helical Htt17 stabilized by interactions with Q_{17} and P_{11}, which are almost exclusively forming a PPII-helix, as well as by intramolecular H-bonds and salt-bridges. Our characterization of Huntingtin's amino-terminus provides insights on the structural origin of its ability to oligomerize and interact with phospholipid bilayers, processes closely linked to the biological functions of this protein.

4.2 INTRODUCTION

Huntingtin is a large ubiquitous protein of more than three thousand amino acids [80, 81]. It is essential to embryonic development [82], it interacts with many proteins through its 36 HEAT repeats [83, 84], it is involved in intracellular organelles and vesicular trafficking [85] as well as transcription and axonal transport [86]. The exon 1 of huntingtin – consisting of an amphipathic sequence of 17 amino acids (Htt17), an amyloidogenic polyglutamine region (Q_N), and a segment of 36 amino acids rich in prolines – is closely linked to Huntingtin's functions. This segment contains a nuclear export sequence that controls the localization of huntingtin between the cytoplasm and the nucleus [87, 88], it can undergo post-translational modifications affecting the localization

and function of huntingtin [89–93], and it is responsible for the localization of huntingtin to the mitochondria and the Golgi [57, 58].

Over the past years, Huntingtin attracted considerable attention as it is an amyloid protein associated to Huntington's disease, an autosomal dominant genetic disorder [94]. Its assembly into amyloid fibrils is triggered *in vivo* by the expansion of the consecutive segment of glutamines at its first exon above a specific threshold. This characteristic behaviour, which is shared by at least 10 other proteins, is termed the polyglutamine/CAG repeat disorder and is associated to several disorders [40, 95, 96]. More specifically to Huntington's disease, the huntingtin protein misfolds, self-assembles, and mislocalizes in the cell when the Q_N region has more than 36 repeats causing deleterious effects by gain- and loss-of-function through various nuclear and extra-nuclear pathways [97–99]. Huntingtin amino-terminus fragments, that can be generated *in vivo* by proteolytic cleavage, are found in post-mortem brain tissue [45] and are involved in the pathogenesis of Huntingtin's disease [100, 101]. The first exon, more precisely, is closely linked to the cytotoxicity as it is sufficient to cause Huntington's phenotype both *in vivo* [48, 102] and *in vitro* [47, 103, 104]. This segment also controls the toxicity, localization and clearance of mutant huntingtin through posttranslational modifications [89–93]. It furthermore interacts with the TRiC chaperonin mainly through its Htt17 segment suppressing the misfolding and aggregation of huntingtin [105].

The neighboring regions of Q_N in the first exon are crucial to control its misfolding and amyloidogenesis [46]. In fact, the aggregation of full-length huntingtin exon 1 is very similar to that of $Htt17Q_NP_{10}$ showing the importance of the amino acids right next to Q_N [64]. For instance, the presence of the Htt17 segment accelerates the fibrillation kinetics of Q_N [66, 106], while the P_N segment decelerates it [65, 107]. The nucleus size as well as the overall aggregation pathways of Q_N are also strongly affected by the presence of Htt17 [74, 78, 105, 108]. Some experimental results indicate that it causes the aggregation to split into two main pathways in direct kinetic competition : it proceeds either (i) through the formation of α-helical tetrameric bundles of Htt17 that increase the local concentration of Q_N favoring the nucleation of beta-sheeted structures in it, or (ii) through an unfavorable nucleation in the monomeric Q_N [108]. The Htt17 segment

could also facilitate the formation and stability of β-sheeted structures in Q_N by interacting directly with it [105, 109]. Others suggest a less direct role for Htt17 where it could destabilize non-fibrillar aggregates by reducing the entanglement of Q_N [77], thus accelerating the formation of fibrils [78].

Given the importance of Htt17, a characterization of its conformational ensemble at the monomer level could shed light on the atomistic features responsible for its aggregation. Experiments suggest that Htt17 samples transient helical configurations in aqueous solution as circular dichroism data indicates the presence of helical structures [58, 66, 68, 71] and as solution NMR suggests no stable secondary structure motif [66]. Due to its intrinsic flexibility and the absence of stable secondary structure motif, the Htt17 monomer yields too few NMR constraints in aqueous solution to extract any viable three-dimensional structural model [66]. A X-rays model of the chimeric maltose-binding protein – huntingtin exon 1 (MPB-Htt17Q_{17}-Ex1) protein also suggests that Htt17 can adopt helical structures, while the Q_N region is mostly disordered and the P_{11} is a polyproline type-II helix [72]. Few notable simulations of Htt17 complemented these experimental results by describing, to some extend, its conformational ensemble either using (i) all-atom, explicit solvent simulated tempering molecular dynamics [69], (ii) all-atom, implicit solvent Monte Carlo [68], or (iii) all-atom, explicit solvent bias-exchanged metadynamics [70] simulations. All agree that Htt17 samples a broad ensemble of helix/coil structures. Other simulations were aimed at characterizing the overall effect of increasing the Q_N length in the context of huntingtin exon 1 [62, 68, 73].

The Htt17 is also crucial for the localization of huntingtin in the cell, in part, through direct membrane interactions [57, 58]. More recently, the structure of Htt17 in the presence of DPC micelles has been resolved using solution NMR : it is an alpha-helix from residues 6 to 17, while the rest of the sequence is disordered and highly flexible [67, 71]. Results from solid-state NMR [67] and Hamiltonian replica-exchange all-atom simulations [110] indicate that Htt17 is also an alpha-helix in the context of a membrane bilayer. As the formation of alpha-helical structures in Htt17 prior to its binding seems to favor its membrane partitioning [79], understanding the conformational ensemble of Htt17 in aqueous solution could then unveil motifs beneficial to membrane-binding. In

the context of exon 1, the effect of Q_N and P_{11} on the occurrence of such motifs could explain their modulation of Htt17 binding affinity as observed experimentally [111].

Focusing on the identification of Htt17's structural features at the origin of its oligomerization and membrane partitioning, we investigate the free energy landscape of the monomeric Htt17 using all-atom, explicit solvent Hamiltonian replica-exchange metadynamics. Such simulation protocol favors the correct sampling of the entire conformational space physically available to the protein. We also quantify the effect on Htt17's global free energy landscape of adding the amyloidogenic Q_N region as well as the P_{11} segment. Overall, such detailed information is necessary to rationalize the importance of Htt17 in addition to paving the way for the investigation of the oligomerization and membrane binding processes per se using such a similarly stringent simulation protocol.

4.3 MATERIALS AND METHODS

In this study, we use Hamiltonian replica exchange (HREX) and parallel tempering (PT) combined with well-tempered metadynamics (MetaD) [24, 28, 30, 32, 112] simulations to investigate the free energy landscape of the 17-amino-acid amino-terminus segment (Htt17) of the huntingtin protein in aqueous solution. We also quantify the impact of adding the amyloidogenic polyglutamine (Q_{17}) and the polyproline (P_{11}) segments on Htt17's free energy landscape. The amino acid sequence of Htt17 is MATLEKLMKAFESLKSF and an amidated carboxy-terminus is used for all peptide constructs. All simulations are summarized in Table 4.I. We focus on the HREXMetaD simulations in the main text, while the PTMetaD simulations are presented in the Supporting Material.

Simulations protocols. Our molecular dynamics simulations are done with the Gromacs package version 4.6.5 [25, 113–115] combined with the PLUMED plug-in version 2.0.2 [27] to perform the well-tempered MetaD [30] and HREX [24] parts of our simulations, as described below. We use the all-atom forcefield AMBER99sb*-ILDN [15, 116, 117] as it offers helix/coil-balanced sampling for the conformational ensemble of small and mostly disordered peptides with transient α-helical structures [116, 118], which is similar to Htt17 in aqueous solution [58, 66, 68, 71]. It is also recognized as one of

Tableau 4.I : Summary of the performed simulations.

Simulations	Type	Initial conf.	Time per replica μs	Time μs
Htt17_nmr	HREXMetaD	NMR	0.9×16	14.4
Htt17_coil	HREXMetaD	coil	0.9×16	14.4
$Htt17Q_{17}$	HREXMetaD	NMR/coil	0.9×24	21.6
$Htt17Q_{17}P_{11}$	HREXMetaD	NMR/coil/coil	0.9×24	21.6
Htt17_nmr_remd	PTMetaD	NMR	0.9×64	57.6
Htt17_coil_remd	PTMetaD	coil	0.9×64	57.6
Htt17_grf	HREXMetaD	NMR	0.25×16	4
$Htt17Q_{17}P_{11}$ _pro	HREXMetaD	NMR/coil/coil	0.2×24	4.8

All simulations are done in the NVT ensemble in a rhombic dodecahedron periodic cell ($\alpha = 60^\circ$; $\beta = 90^\circ$; $\gamma = 60^\circ$; $a = b = c = 5.35$ nm and 3500 water molecules for Htt17, $a = b = c = 6.80$ nm and 7000 water molecules for $Htt17Q_{17}$, and $a = b = c = 8.12$ nm 10000 water molecules for $Htt17Q_{17}P_{11}$). This set-up is sufficient to ensure that our system, which is monomeric, interact very little with its periodic image. In spite of a formal dilution of 3.5 mM, it is therefore largely equivalent to a highly diluted system allowing comparison to NMR studies done at 40 μM concentrations. We combined well-tempered metadynamics (MetaD) to two other sampling enhancing simulation types : Hamiltonian replica-exchange (HREX) and parallel-tempering (PT). The simulations on Htt17 are started from two different initial configurations : a fully random coil structure and its NMR model in the presence of DPC detergent micelles (PDB 2LD2). The latter configuration is disordered from residues 1 to 5 and an α-helix for the rest of the sequence [67, 71]. In the initial state of the simulations on $Htt17Q_{17}$ and $Htt17Q_{17}P_{11}$, Htt17 is taken as the NMR model, while Q_{17} and P_{11} are completely disordered. We focus on the HREX simulations in the main text, and we present the PT simulations in the Supporting Material. The last two simulations Htt17_grf and $Htt17Q_{17}P_{11}$ _pro are tests on the validity of the potentials and are presented and discussed in the Supporting Material.

the best forcefield to study protein folding [119–122]. Our simulations are performed in the NVT ensemble and the temperature is maintained by the Bussi–Donadio–Parrinello thermostat with a coupling constant of 0.1 ps [123]. Van der Waals and short range electrostatic interactions are cutoff at 1.0 nm. Long range electrostatics are computed using smooth Particle-Mesh Ewald [124, 125]. Bond lengths and TIP3P water geometry are respectively constrained using LINCS [126] and SETTLE [127] allowing an integration time step of 2 fs. The center-of-mass motion is removed every 20 fs. Configurations are

saved every 4 ps for analysis.

We use Hamiltonian replica-exchange metadynamics (HREXMetaD) to efficiently sample the conformational ensemble and unveil the free energy landscape of Htt17 in aqueous solution. This method combines two sampling enhancing techniques : Hamiltonian replica-exchange (HREX) [23, 63] and metadynamics (MetaD) [28, 30]. MetaD introduces a history-dependent bias constructed by adding gaussians in the energy space to previously visited states along a set of specified collective variables (CVs). This increases the overall sampling at the same time as reconstructing the free energy landscape along those CVs ($\vec{S}$) as the introduced history-dependent biased potential converges to

$$V(\vec{S},t \to \infty) = -\frac{\Delta T}{T+\Delta T} F(\vec{S}) + C \tag{4.1}$$

where V($\vec{S}$,t) is the biased potential, F($\vec{S}$) is the free energy, T is the temperature, C is an irrelevant constant, and ΔT is a parameter that controls the extend of barrier heights sampled in the well-tempered flavour of MetaD [30]. It is also possible, as we do in this study, to reconstruct the free energy landscape along any omitted CV given sufficient sampling [32, 33]. The free energy landscape is normalized so that all free energies are measured with respect to the most stable structure for each simulation, which is set at 0 kJ/mol.

For correct free energy landscape reconstruction, all slow CVs need to be considered, however the maximum number of CVs computationally accessible is about 2-3 for MetaD which is clearly not enough to model protein folding [29, 31, 128]. One of the most efficient way to avoid this limitation is to couple MetaD with a replica exchange scheme such as Hamiltonian replica exchange (HREX) as this technique, which is widely use to simulate protein folding by its own, increases the probability of escaping free energy minima by allowing exchanges between simultaneous MD simulations at different Hamiltonians [23, 24]. Using replica exchange schemes such as HREX and parallel tempering (PT) together with MetaD allows to correctly sample other CVs not explicitly taken into account by the time-dependent biased potential as demonstrated for proteins with similar conformational ensemble to Htt17 [129–131].

For the MetaD part of our hybrid simulations, we use two CVs to bias the α-helical character (S_α) and the radius of gyration (S_{gyr}) of the peptide :

$$S_\alpha = \sum_{i=0}^{13} \frac{1-\left(\frac{d_i}{d_0}\right)^6}{1-\left(\frac{d_i}{d_0}\right)^{12}} \tag{4.2}$$

$$S_{gyr} = \left(\frac{\sum_{i=0}^{17} |r_i - r_{\mathrm{COM}}|^2}{\sum_{i=0}^{17} m_i} \right) \tag{4.3}$$

where the sum in S_α is over the 13 possible α-helix hydrogen bond distances d_i between main chain H_N – O couples separated by 4 residues, d_0 is 0.3 nm, the sum in S_{gyr} is over all C_α, r_i and m_i are the current C_α coordinate and mass respectively, and r_{COM} is the center-of-mass coordinate. Note that by construction $\max(S_\alpha) = 13$, but the single α-helix has $S_\alpha \sim 12.0$ as d_i is about 0.20–0.25 nm for a hydrogen bond. This choice of CVs is motivated by the fact that the Htt17 peptide in aqueous solution has an average α-helix probability of 10 to 55% according to circular dichroism (CD) spectroscopy measurements [58, 66, 68, 71], but no stable α-helix as determined by solution NMR experiments [66]. The free energy landscape of peptides with similar conformational ensemble to Htt17 were also characterized using this set of CVs [129–131]. During our simulations, a new gaussian is added to the biased potential every 4 ps with standard deviations of 0.1 and 0.01 nm along S_α and S_{gyr} respectively, and their initial height is 0.5 kJ/mol. The bias factor of the well-tempered scheme is set to 15.

The HREXMetaD simulations are performed at 303K using 16 scales for Htt17 and 24 scales for Htt17Q_{17} and Htt17$Q_{17}P_{11}$ spanning 1.0 to 0.3 with intermediate scales specified by a geometric distribution as previously done [24]. Exchanges between neighboring scales are attempted every 4 ps resulting in an exchange rate of about 20–40%.

Simulated systems. All performed simulations are summarized in Table 4.I. The three investigated fragments of Huntingtin amino-terminus – Htt17, Htt17Q_{17} and Htt17$Q_{17}P_{11}$ – are simulated using HREXMetaD at 303K. The P_{11} segment corresponds to the first complete proline repeat sequence of the 36-amino-acid proline-rich segment connected to Q_N in huntingtin. Addition of both Htt17 and P_{11} are sufficient to reproduce the main

characteristics of the aggregation of Huntingtin's first exon [64]. The two initial states for Htt17 are a random coil structure and the solution NMR model (PDB 2LD2) determined in the presence of DPC micelles [67, 71]. In this latter state, Htt17 is an α-helix from residues 6 to 17 and disordered for the first five residues. For Htt17Q_{17} and Htt17$Q_{17}P_{11}$, Htt17 is the NMR model, while Q_{17} and P_{11} are disordered. Random coil configurations are generated with 100 ns high temperature (600K) simulations starting from initially totally extended structures.

Peptides are solvated in a rhombic dodecahedron periodic cell and neutralized by the addition of two chloride ions. All systems are energy minimized using the conjugate gradient algorithm, and are equilibrated in the NPT ensemble at 303K for 5 ns restraining the protein backbone atoms to their initial position using a harmonic potential. All replicas are further independently equilibrated at their respective Hamiltonian scale in the NVT ensemble for 10 ns.

Our analysis is performed using in-house, GROMACS and PLUMED utilities. The secondary structure is computed using STRIDE [132] and chemical shifts using SPARTA+ [133] and Camshift [134]. All computed quantities are re-weighted to remove the bias introduced during the MetaD simulations as previously described [33] using a python implementation by Ludovico Sutto (University College London) that is available to the PLUMED community. The free energies are re-weighted using a recently developed time-independent free energy estimator [33]. GROMACS utilities are used to compute the structural clusters using the gromos algorithm with a RMSD cutoff of 0.15 nm on the backbone atoms (g_cluster) [135], the solvent accessible surface area of the non-polar residues (g_sas) [136], and the occurrence of H-bonds using a cutoff of 0.35 nm on the donor-acceptor distance and of 30° on the hydrogen-donor-acceptor angle (g_hbond). Salt-bridges are considered when the distance between two oppositely charged moieties is less than 0.4 nm [137].

Errors correspond to one statistical standard deviation computed on the converge interval with 20-ns subsets.

Convergence. We assess the convergence of our simulations using three quantitative criteria as shown in Figure S1 for Htt17, Figure S2 for Htt17Q_{17} and Figure S3 for

Htt17$Q_{17}P_{11}$. First, we track the evolution of the global uncertainty on the free energy to identify the time at which it becomes small enough. Second, we monitor the sum of the free energy bias added each 10 ns as a function of time to confirm that the added biases become small enough at some time. Third, we compute the two-dimensional free energy uncertainty landscape as a function of S_α and S_{gyr} on the time-interval of convergence determined from the two previous criteria to confirm that the errors are located in unimportant regions of the landscape. All times presented in this article correspond to real simulation time but are not associated to a real physical trajectory due to the exchanges between replicas.

We confirm that the choice of initial state does not impact our results on Htt17 as described in details in the Supporting Material. To do so, we confirm that the convergence analysis and free energy landscape generated starting from the NMR model obtained in the presence of DPC micelles and a random coil state are very similar (compare Figure S1 and Figure S4). Moreover, as described in details in the Supporting Material, we ensure that our results on Htt17 are mostly independent of the sampling method by comparing HREXMetaD to PTMetaD since the latter one is most often used. Additional simulations probing the effect of a Generalized-reaction field on Htt17 (Htt17_grf) and the effect of the AMBER99sb*-ILDNP forcefield, with revised proline parameters, on Htt17$Q_{17}P_{11}$ (Htt17$Q_{17}P_{11}$_pro) were done using slightly modified parameters to accelerate the simulation ; new gaussian is added to the potentiel every 1 ps (instead of 4 ps) and the standard deviation of S_{rg} was increase to 0.03 nm (instead of 0.01 nm). An in-depth discussion on these simulations can be found in the Supporting Material.

In complement, we probe the quality of the sampling of our HREX simulations by monitoring the replicas visiting the first scale that is used in our analysis as well as the secondary structure content as a function of the scaling (Figure S5). We find great diffusion in the replica space and a more disordered peptide at larger scales.

Overall, our convergence evaluation indicates that the following analysis time intervals are suitable : 400–900 ns for Htt17_nmr (Figure S1), 500–900 ns for Htt17Q_{17} (Figure S2), and 500–900 ns for Htt17$Q_{17}P_{11}$ (Figure S3). All analysis presented are thus performed on these intervals.

4.4 RESULTS

4.4.1 Htt17

The two-dimensional free energy surface (FES) of the Htt17 sequence in terms of the two biased CVs (S_α, number of helical H-bonds ; S_{gyr}, gyration radius) for the Htt17_nmr simulation is shown in Figure 4.1. It is characterized by a single large basin with the configurations below the 5 kJ/mol isoline being bound by 2 to 6 helical H-bonds and a gyration radius between 0.6 and 0.8 nm. In this region, the free energy average is 4.8 kJ/mol and gradually rises to an average of 9.1 kJ/mol as the number of helical H-bonds increases. These latter conformations are less collapsed than those below the 5 kJ/mol isoline as indicated by their larger gyration radius.

The per residue secondary structure of Htt17 is shown in Figure 4.2. We find that the first half of the peptide forms an α-helix (residue 3 to 7, $\sim$40–55%) more often than the second half ($<$35%). We also observe that residues 10 to 13 are very likely to form a turn indicating a population of two-helix bundle conformations. Overall, the peptide is mostly unstructured with only $29.3 \pm 0.7\%$ of α-helix probability, in agreement with measurements from several circular dichroism (CD) experiments [58, 66, 68, 71], and a negligible amount of β-sheet and β-bridge totalizing $1.2 \pm 0.2\%$.

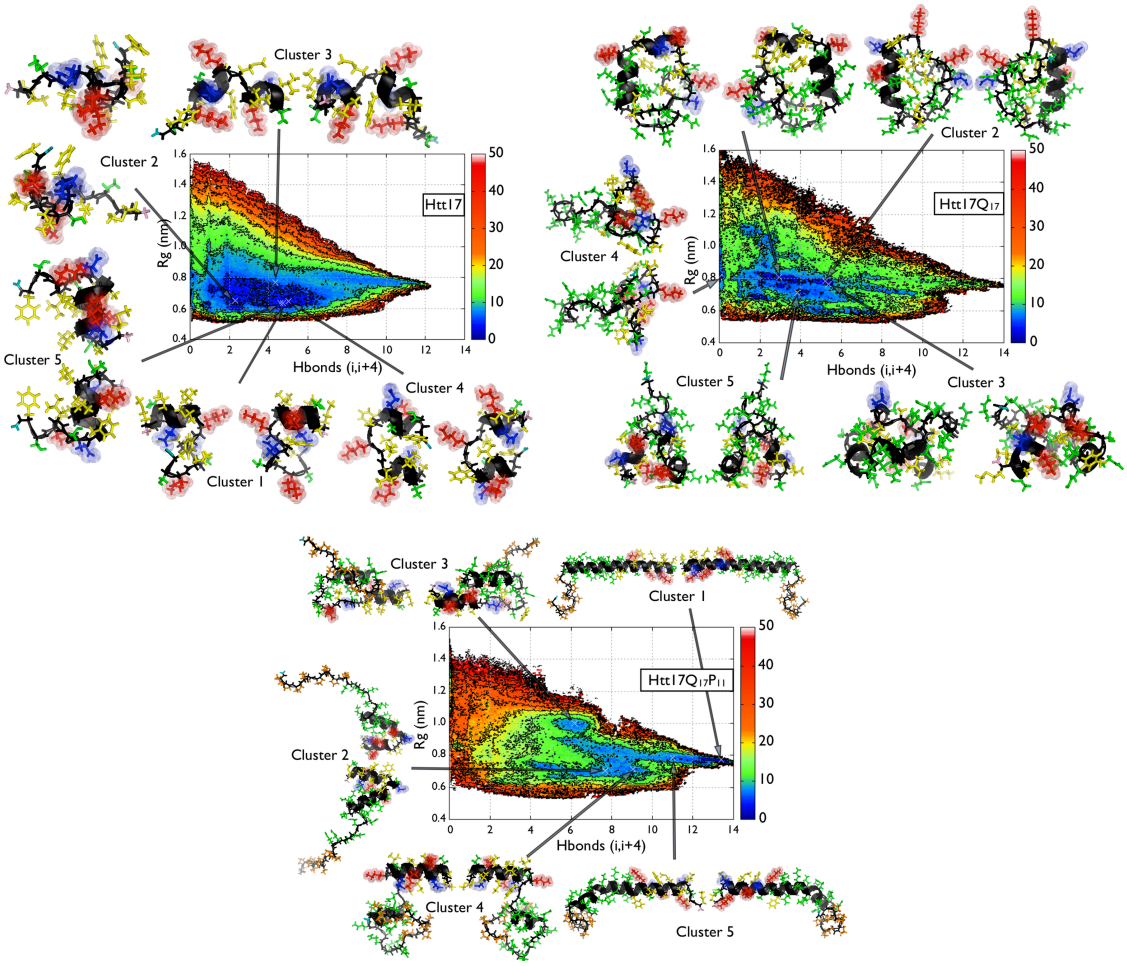

Figure 4.1 : The free energy landscapes of Htt17 (Htt17_nmr) (shown in the top left panel), $Htt17Q_{17}$ (shown in the top right panel) and $Htt17Q_{17}P_{11}$ (shown in the bottom panel). The horizontal and vertical axes respectively represent the number of helical H-bonds (S_{α}) and the gyration radius (S_{gyr}). The number of helical H-bonds is computed on the first 13 residues for Htt17 and the first 17 residues for $Htt17Q_{17}$ and $Htt17Q_{17}P_{11}$ to include possible H-bond formation with the Q_{17} domain. The energy isolines are drawn every 5 kJ/mol. The uncertainty on the free energy landscapes of Htt17, $Htt17Q_{17}$ and $Htt17Q_{17}P_{11}$ are respectively shown in Figures S1, S2 and S3. The uncertainty is always smaller than 1 kJ/mol on the relevant parts of the landscapes. In addition, the cluster analysis of the most representative conformations populating the FES below 5 kJ/mol and below 8 kJ/mol for $Htt17Q_{17}P_{11}$. The negatively charged, positively charged, non polar and polar residues of Htt17 are shown in blue, red, yellow and green. The Q_{17} and the P_{11} segments are respectively coloured in green and orange. The backbone is coloured in black, the amino-terminus in pink and the carboxy-terminus in teal.

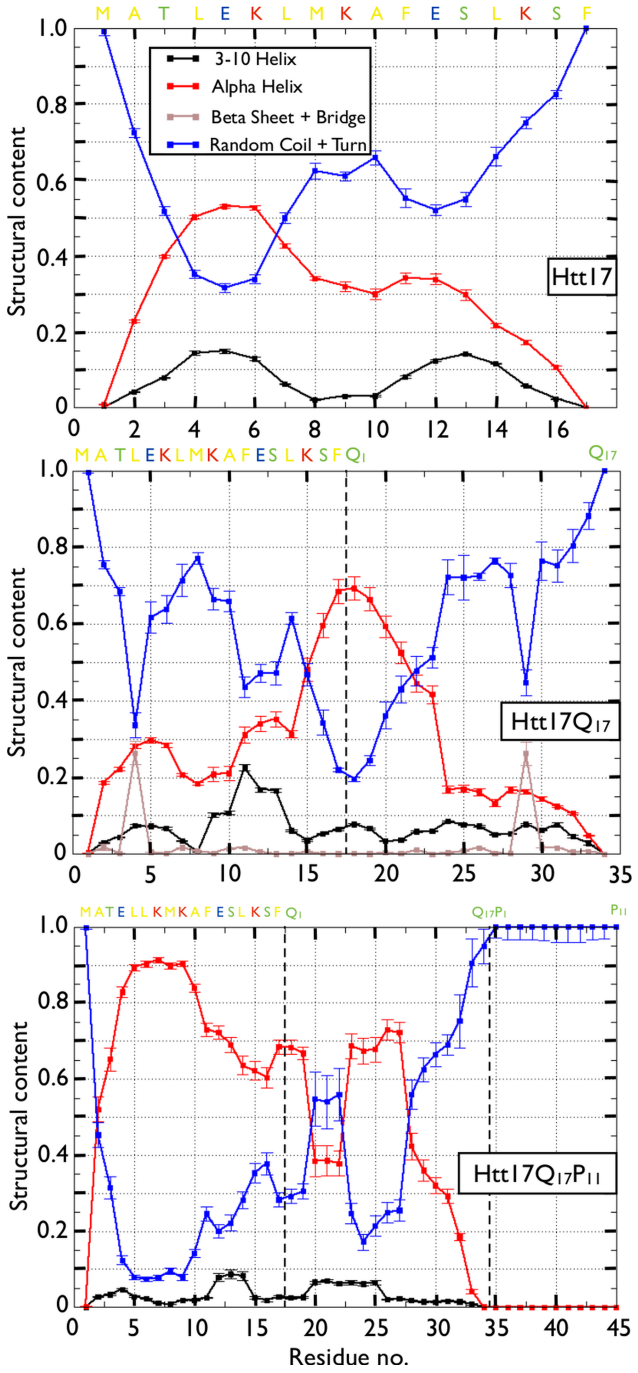

Figure 4.2 : From top to bottom, the per residue secondary structures of Htt17 (Htt17_nmr), $Htt17Q_{17}$ and $Htt17Q_{17}P_{11}$. The probability of α-helix, 3-10 helix, β-bridge and β-strand, and turn/coil are respectively shown in red, black, brown and blue. The vertical black dotted line for $Htt17Q_{17}$ and $Htt17Q_{17}P_{11}$ indicates the end of the Htt17 segment and of the Q_{17} segment.

We also evaluate the presence of structural elements using the secondary H^{α} chemical shift and NOE signal sampled in our simulations (Figure 4.3) to compare with the solution NMR experiment of Wetzel and co-workers [66]. The secondary chemical shifts are in very good agreement with the NMR measurements ; except for residues Lys9, Ala10 and Ser16 that are more extended in our simulations. For the most part, the secondary chemical shifts are small and negative indicating weak helical features. The high H^{N}(i)-H^{N}(i+1) and low H^{α}(i)-H^{N}(i+2) NOE signals combined with small H^{α}(i)-H^{N}(i+3) NOE intensities indicate a global α-helix average of about 30% without any individual residue showing more than 55% probability, which slightly overestimates the α-helical propensity but remains compatible with the NMR data (see Supporting Material for an in-depth discussion).

The main configurations sampled by Htt17 in basins below the 5 kJ/mol isoline are depicted by their cluster's center in Figure 4.1. In line with our previous analysis, the first residues of the Htt17 peptide have a greater tendency to be helical, while the last residues are mostly unstructured and the central part of Htt17 (residues 10 to 13) forms a turn bringing the amino- and the carboxy-terminus close to each other (see clusters 1, 3 and 4). The non-polar residues are mainly accessible to the solvent This observation is confirmed by the analysis of the two-dimensional FES as a function of the solvent accessible surface area (SASA) of Htt17's non-polar residues and the number of helical H-bonds (Figure S6).

The contact map between each residue is indicative of the mostly flexible and disordered tertiary structure of Htt17 as most contacts are between neighboring residues (Figure S7). We observe nevertheless the presence of three electrostatic contacts : Glu5–Lys9, Glu12–Lys9 and Glu12–Lys15 with a probability of 50.3%, 42.3% and 64.0%, respectively. The formation of a stable salt-bridge occurs less often however with probability of 4.2 $\pm$ 0.1%, 11.0 $\pm$ 1.0% and 12.3 $\pm$ 0.3%, respectively. We also note a long range non-polar contact between Met8 and Phe17 in 24.1% of the sampled conformations that seems to be involved in the formation of the turn between residues 10 and 13 as well as in the destabilization of the secondary structure in the second half of the peptide (see clusters 1 and 4 of Figure 4.1).

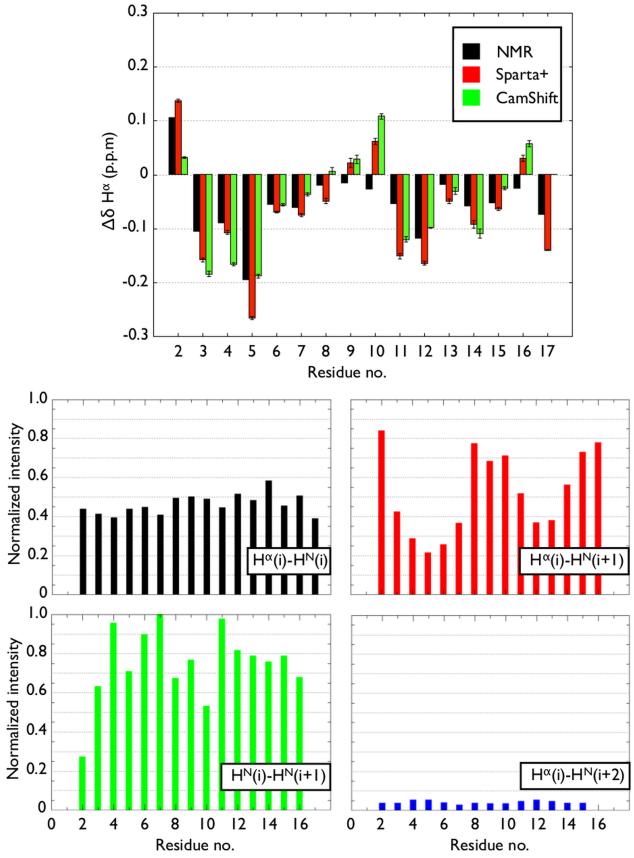

Figure 4.3 : **(A)** The H^{α} secondary chemical shift per residue computed with SPARTA+ (red) and CamShift (green) are compared to the NMR measurements (black) on Htt17 in aqueous solution [66]. The secondary shifts are obtained by subtracting the corrected coil value specific to each amino acid type to their H^{α} chemical shift [138]. **(B)** The computed intensities of the interproton NOEs for all residues between the H^{α} of residue i and the H^{N} of residues i, i+1 and i+2, as well as between the H^{N} of residues i and i+1 are compared to the equivalent NMR measurements on Htt17 in aqueous solution [66].

4.4.2 Htt17Q_{17}

We now investigate the changes induced on the FES of Htt17 due to the addition of the Q_{17} segment at its carboxy-terminus. The two-dimensional FES of Htt17Q_{17} in terms of S_{α} and S_{gyr} – only on the Htt17 region – is shown in Figure 4.1. In the largest basin, we observe three distinct minima characterized by a similar gyration radius around 0.8 nm, but with a different number of helical H-bonds (either 0.0 or between 2.0 and 6.0).

Addition of the Q_{17} region modifies key features of the FES of Htt17 as it becomes more extended (more configurations having a larger gyration radius) and less structured (fewer configurations having a large number of helical H-bonds) as shown in Figure 4.1. More precisely, the free energy of the conformations with a large number of helical H-bonds significantly increases from an average of 9.1 kJ/mol for Htt17 alone to 14.6 kJ/mol for Htt17Q_{17}.

In terms of secondary structure, we observe a significant loss of helical propensity for residues 2 to 10, while that of the remaining residues in Htt17 greatly increases upon addition of the Q_{17} segment (Figure 4.2). Even though the Htt17 segment in Htt17Q_{17} has an overall α-helix probability that is unchanged with respect to Htt17 alone (30.4 $\pm$ 1.4 % vs. 29.3 $\pm$ 0.7%, respectively), the per residue probability is very different : there is a significant amino-to-carboxy-terminus shift of the helical probability that is directly due to the presence of the Q_{17} as the α-helix in Htt17 continues up to the first six glutamines. The remaining part of the Q_{17} is however mostly disordered. We also note the presence of a turn between Glu5 and Leu7 ($\sim$40-45%). The amount of β-structure is still negligible (except for Leu4).

A cluster analysis of the structures characterized by a free energy below 5 kJ/mol deepens the atomistic insights (Figure 4.1). The first five clusters can be classified in three main categories : (1) no helical H-bond in Htt17 and a fully disordered Q_{17} (clusters 2), (2) two small α-helix fragment at both end of Htt17 and a disordered Q_{17} (cluster 4 and 5) and (3) an α-helix spanning the last residues of Htt17 and the first glutamines of Q_{17} (clusters 1 and 3). In these clusters, the Htt17 segment adopts a u-shaped topology with a turn around Leu7 bringing its non-polar residues close together as shown by Htt17Q_{17}'s contacts map (Figure S7). More precisely, we identify long-range non-polar contacts between Leu4–Phe17 (20.8%), Met8–Phe17 (20.4%) and Phe11–Phe17 (29.1%) that form a non-polar cluster, which is further isolated from the solvent by the aliphatic carbons of the glutamines (for example, see clusters 1 and 3 of Figure 4.1). As in Htt17 alone, we observe electrostatic contacts/salt-bridge between Glu5–Lys9 (44.5% / 6.9 $\pm$ 0.2%), Glu12–Lys9 (51.1% / 23 $\pm$ 2%) and Glu12–Lys15 (48.9% / 12.9$\pm$ 0.2%). The charged residues Glu5/Lys6/Lys9/Glu12/Lys15 of the Htt17 domain also interact with the Q_{17}

domain with a probability of 54.6%, 50.8%, 48.2%, 40.7%, 83.9% respectively. For its part, the glutamines aliphatic carbons of the Q_{17} segment interacts a lot with the non-polar residues of the Htt17 segment. As a consequence, the resulting non-polar SASA decreases in contrast to Htt17 alone (Figure S6). We also note the presence of several main-chain/side-chain and main-chain/main-chain interactions between Htt17 and Q_{17} ($\sim$ 30%, Figure S7).

The FES of the Q_{17} region shows that it is mostly disordered and collapsed (Figure S8). A cluster analysis on the configurations below 4 kJ/mol shows that the first glutamines are α-helical, while the remaining region is disordered independently of Htt17's structure, which is either an α-helix from residues Ser13 to Phe17 (clusters 1, 4 and 5) or from residues Ala2 to Met8 (clusters 2). At high helical H-bonds, we note the presence of a very narrow minimum characterized by an almost fully α-helical Q_{17} with Htt17 adopting the same structure up to residue Glu12.

4.4.3 Htt17$Q_{17}P_{11}$

We finally probe the effects of the addition of the Q_{17} and P_{11} segments on the FES of Htt17. The two-dimensional FES of Htt17$Q_{17}P_{11}$ in terms of S_α and S_{gyr} – only on the Htt17 region – is shown in Figure 4.1. The resulting FES unveils a striking shift toward the α-helix as the FES is characterized by a single minimum with a number of helical H-bonds between 12 and 14 and a gyration radius between 0.75 and 0.85 nm. Most of the structures sampled by Htt17 alone or Htt17Q_{17} are therefore less stable upon the addition of P_{11}.

Analysis of the secondary structure indicates that the Htt17 region adopts an α-helical conformations in 70.9 $\pm$ 1.6% of the time, a drastic increase compared to both Htt17 (29.3 $\pm$ 0.7%) and Htt17Q_{17} (30.4 $\pm$ 1.4 %) (Figure 4.2). The probability is notably high for residues 5 to 9 with more than 90%. The Q_{17} domain has 44.8 $\pm$ 2.5% of α-helical content with the residues near Htt17 (first, second and sixth to tenth glutamines) having the largest probability, while glutamines 3 to 5 and those near P_{11} mostly form turn/coil structures. Finally, the P_{11} domain almost exclusively forms a PPII-helix characterized by Φ and Ψ dihedral angles respectively near of -75° and 150° according

to an analysis of the Ramachandran plot for every proline (data not shown).

The addition of P_{11} causes drastic changes in structure as it doubles the helical content of both Htt17 and the Q_{17} domain. We further quantify its effect from a cluster analysis on the structures found below 8 kJ/mol. In contrast to the important structural diversity of Htt17 and Htt17Q_{17} in solution, Htt17 within Htt17$Q_{17}P_{11}$ has a strong tendency of forming an α-helix as the first cluster is composed of more than 25% of the sampled structures. The depicted conformation for the first cluster shows the Q_{17} as a fully formed α-helix and the P_{11} region extends away from Htt17 and Q_{17}. We note a clear separation between the polar and non-polar residues of Htt17. The formers are interacting mostly with Q_{17} as shown on the Htt17$Q_{17}P_{11}$ contacts map (Figure S7). More specifically, there are contacts between Glu5/Lys6/Lys9/Glu12/Lys15 of the Htt17 domain and the glutamines with a probability of 25.0%, 21.0%, 61.1%, 57.6% and 85.7%, respectively. Salt-bridges are also present between Lys9–Glu12 ($45 \pm 4\%$) and Lys15–Glu12 ($14 \pm 2\%$) stabilizing Htt17 conformation. For its part, the P_{11} domain interact mostly with residues surrounding Lys9 (33.5%), Ser13 (50.0%) and Leu14 (24.0%) and Phe17 (48.8%) via mostly their side-chain. As for the non-polar residues in Htt17, they are all located on the same side of the peptide and fully accessible to the solvent as shown by a drastic increase of the non-polar SASA when compare to the Htt17 and Htt17Q_{17} peptides (Figure S6).

The FES of Q_{17} unveils three minima with distinct number of H-bonds (either 0, 2.5 and 7.5) as shown in Figure S8. A cluster analysis on the structures found inside those region (below 4 kJ/mol) shows that the most important cluster of the Q_{17} domain has an important helical propensity up to the tenth glutamine and that the remaining glutamines are mostly unstructured. The other clusters depict the Q_{17} domain as fully unstructured independently of Htt17's structure that is either a fully formed α-helix (clusters 2, 4 and 5) or mostly unstructured (cluster 3).

4.5 DISCUSSION

Numerous experiments indicate that the Huntingtin amino-terminus is crucial for its biological functions. More specifically, the first 17-amino-acid segment (Htt17), which is right before the amyloidogenic polyglutamine segment (Q_N), is directly involved in the membrane interactions and aggregation of Huntingtin. In this study, we quantify the conformational ensemble of three fragments of Huntingtin amino-terminus – Htt17, Htt17Q_{17} and Htt17$Q_{17}P_{11}$ – in terms of free energy surfaces, secondary structures, contact maps and clusters. Our results demonstrate the effects of Q_{17} and P_{11} on the conformational ensemble of Htt17 and, taken together with other studies, they provide insights on motifs at the origin of Htt17's membrane-binding and oligomerization.

4.5.1 Htt17 samples a wide variety of coil/helix structures

Experiments indicate that Htt17 has a helical population of about 10–55% in aqueous solution using circular dichroism (CD) [58, 66, 68, 71], but no stable structural motif according to solution NMR measurements [66]. Recently results from ion mobility spectrometry-mass-spectrometry (IMS-MS) coupled to molecular dynamics simulations suggest that Htt17 populates two kinds of helical monomer with an α-helix from the amino-terminus to residue Lys9–Ala10 : (i) a compact structure characterized by an unstructured region between residues Phe11 and Phe17 that turn back on itself and brings the amino- and carboxy-terminus closer to each other and (ii) an extended structure characterized by a 3-10 helix spanning residues Ala10 to Glu12 and where the carboxy-terminus region is extended away from Htt17 [75]. Taken together, these observations suggest that the structural ensemble of Htt17 consists of a wide variety of flexible helix/coil conformations.

Previous simulations on Htt17 suggest such a conformational ensemble [68–70]. More precisely, simulated tempering simulations with the AMBER03 forcefield and explicit solvent (TIP3P) show that the conformational ensemble of Htt17 contains about 70% of diverse two-helix bundles with a loop around Ala10, while the rest of the ensemble populates a single straight helix or disordered configurations [69]. In this work,

residues 3 to 6 have the highest α-helix propensity and that the sampled configurations are mostly stabilized either by charged interactions or sequestration of the non-polar residues. Other simulations were performed using Monte Carlo with the ABSINTH implicit solvent forcefield showing that Htt17 has an α-helix probability of 34% and that it is mostly collapsed upon itself to sequester its non-polar residues [68]. Bias-exchange metadynamics, for its part, unveiled the free energy landscape of Htt17 using explicit solvent, all-atom simulations (AMBER99/TIP3P) [70]. In this work, Carloni *et al.* observe that the free energy landscape using 6 collective variables (CVs) is mainly made of four basins and that the transitions from one basin to the others occur on the microsecond timescale. The resulting conformational ensemble of Htt17 is largely disordered (75%) and helical (25%) with a global α-helix probability of 29% notably for residues 1 to 7. They also note that the disordered configurations of the largest basin have their non-polar residues largely accessible to the solvent.

This is in line with the trend depicted by our simulations in terms of the free energy landscape (Figure 4.1) and the secondary structure propensity (Figure 4.2). Our predicted secondary structure is characterized by a global α-helix probability of $29.3 \pm 0.7\%$ that is similar to the values obtained in the aforementioned simulations – 43% [69], 34% [68] and 29% [70] – and CD experiments – 10% [71], 34% [68], 45% [58] and 55% [66] – on Htt17 in aqueous solution. Our results also indicate that residues 3 to 5 have the highest helix propensity (Figure 4.2, about 40–50%) in agreement with other simulation protocols [69, 70]. Finally, we find that Htt17 forms various two-helix, single helix, helix/coil and coil conformations as previously observed [69, 70, 75]. The probability of structured conformations is however lower in our simulations than in Ref [69], which might be due to AMBER03 slightly overstabilizing helical structures in helix/coil peptides when compared to the AMBER99sb*-ILDN force field as indicated by other studies [116, 118].

In terms of tertiary structure, our simulations indicate that Htt17's non-polar residues are mostly accessible to the solvent (Figure S6) in agreement with previous bias-exchange metadynamics simulations [70]. In addition, mainly short-range contacts between neighbouring residues are populated in Htt17. Still, a non-polar contact between

Met8 and Phe17 occurs in 24.1% of the sampled structures. It could be crucial in the formation of the turn between residue Ala10 and Ser13 therefore leading to the destabilization of the second half of Htt17.

Finally, we provide a detailed analysis indicating that the structural ensemble sampled in our simulations is consistent with the only solution NMR experiment done on Htt17 in an aqueous environment [66] in terms of secondary H^{α} chemical shifts and interproton NOE distances (Figure 4.3 and Supporting Material). We are thus confident that our results yields relevant insights on the structural ensemble of Htt17.

4.5.2 Addition of Q_{17} reduces Htt17's non-polar residues accessibility to the solvent

Fluorescence-based resonance energy transfer (FRET) experiments indicate that Htt17 is in a collapsed state and that it becomes more extended upon addition of the polyglutamine segment [66]. CD spectra suggest an increase of helix propensity with the polyglutamine length, but it is unknown if it is due to the Q_N or Htt17 segments [74]. Data from X-ray crystallography on a chimeric protein containing Huntingtin exon 1 supplement this by indicating that an α-helix in Htt17 can extend to the Q_N region [72]. In both studies, absence of β-sheet is observed. For its part, Monte-Carlo simulations using the implicit solvent ABSINTH potential show that addition of the Q_N domain disorders Htt17 in a length-dependent manner, while the Q_N segment itself remains disordered [68]. Pappu *et al.* also find that the non-polar residues of Htt17 lie in interdomain interface between Htt17 and Q_N.

Our results complement these experiments by showing that the Q_N region in $Htt17Q_{17}$ is mostly disordered, but that it can sample α-helices with a probability of $27.8 \pm 1.3\%$, particularly for the first glutamines (Figure 4.2). The Q_N region also induces an amino-to-carboxy-terminus shift of the helical probability in the Htt17 region but leaves unchanged its global α-helical probability from $29.3 \pm 0.7\%$ to $30.4 \pm 1.4\%$. In line with experimental results, we find negligible amount of β structures for both the Htt17 and Q_N regions at the monomer level. We also find that the aliphatic carbons of the Q_N domain interact directly with the non-polar residues of Htt17 dramatically reducing their solvent accessibility (Figures S6 and S7), in agreement with previous simulations [68].

Overall, we observe that the Q_N region modifies the conformational ensemble of Htt17 already at the monomer level, which could have a direct impact on its aggregation and membrane-binding affinities as discussed next. This indicates that not only does Htt17 influence Q_N as previously determined experimentally, but that the opposite also occurs and that the interplay between Htt17 and Q_N might be more complex than previously thought. As demonstrated in the Supporting Material, these results are replicated qualitatively using a different electrostatic scheme as well as a newly proposed set of proline parameters, they are not, therefore, an artefact of the simulation conditions.

4.5.3 Htt17 is more structured upon addition of $\mathbf{Q_{17}P_{11}}$

Circular dichroism (CD) measurements show that the addition of a P_{10} domain to Htt17Q_{37} reduces the α-helix probability from more than 50% to around 30% [74]. However, CD is unable to tell the localization of these structural changes. Using HPLC sedimentation assay, Wetzel et al. also reported that the aggregation of Htt17Q_{35} is quicker than for Htt17$Q_{37}P_{10}$, although the latter is still much faster than a Q_N domain of similar length alone. ThT fluorescence kinetic profile monitoring the growth of the fibril showed mostly no difference between Htt17Q_{30} and Htt17Q_{30}C38, where C38 is the full-length proline-rich region starting with P_{11}, indicating that the aggregation mechanism is dominated by Htt17 [68]. It is also found that C38 acts as a solubilizing module that weakens the driving force toward the formation of insoluble aggregates. X-ray crystallography on a chimeric protein containing Huntingtin exon 1 suggests that the Htt17 can populate α-helix configurations, while the Q_N region is mostly unstructured except for the first glutamines that can populate an α-helix [72]. The first prolines in the proline-rich region, for their part, are characterized by a PPII-helix.

Only one set of simulations has been performed on Htt17Q_NP_{11} to our knowledge [73]. These all-atom replica exchange molecular dynamics simulations with the FF03 force field and implicit solvent suggested that both Htt17 and Q_N adopt mostly α-helical conformations, while the P_{11} forms a PPII-helix. In these simulations, the α-helical content is especially large between residues 4 and 17 of Htt17, and the P_{11} region lies anti-parallel to the Htt17 region when there are 17 glutamines in Q_N, but not when there

are 55 glutamines (above than the pathological threshold of 36 repeats).

In our simulation, the P_{11} domain stabilize Htt17 as an almost fully formed α-helix with more than 70% of helical content. The Q_{17} domain adopts an α-helix conformation $44.8 \pm 2.5\%$ of the time. Our results differs from the secondary structure signal from CD [74] and are surprising overall. Indeed, the increase of the non-polar SASA and the decreased number of contacts between Htt17 and Q_{17} are two strong destabilizing factors present in our simulations. The difference with experiment is perhaps due to the length of the Q_N domain used ; longer Q_N (as in the CD experiment) might mitigate the stabilizing effects of P_{11}. An other simulation protocol show a similar α-helical population in Htt17 in the context of $Htt17Q_{17}P_{11}$, but significantly more α-helix in Q_{17} [73]. Ultimately, other simulation and experimental protocol will be needed to unveil the origin of this dissimilarity.

4.5.4 Motifs relevant to membrane-binding and oligomerization

Htt17 is crucial to the localization of Huntingtin in the cell [57, 58, 87–93] and adopts an α-helical conformation in the presence of micelles, vesicles and phospholipid membranes as shown by CD spectroscopy [58, 71], solution NMR [71], solid-state NMR [67] and Hamiltonian replica-exchange simulations [110]. Solid-state NMR and Hamiltonian replica-exchange simulations also indicate that the amphipatic plane of Htt17 is aligned parallel to the membrane surface with its non-polar residues facing the hydrophobic core of the membrane. The presence of α-helical structures in Htt17 prior to membrane-binding has been shown to ease its insertion in the membrane [79].

Htt17 also drastically modifies the aggregation of the amyloidogenic Q_N segment. Three main models have been proposed to describe Htt17's role in the aggregation mechanism of Huntingtin : (i) the formation of tetrameric α-helical bundles of Htt17 that increases the local concentration of Q_N favoring the nucleation of β-sheeted structures in the latter region [74], (ii) the reduction in the entanglement of Q_N destabilizing the non-fibrillar aggregates [78], and (iii) the direct interaction between Htt17 and Q_N favoring formation of extended structures in the latter region [105]. Solid-state NMR indicates, for its part, that the core of a $Htt17Q_{30}P_{10}K_2$ amyloid fibril is formed by Q_N, while Htt17

and P_{10} respectively form an α-helix and a PPII-helix [139, 140].

The first aggregation model, more specifically, is based on sedimentation velocity experiments that indicate that Htt17 and Htt17$Q_{10}K_2$ are mostly monomeric in solution with low level of compact oligomers that correspond to, in decreasing population, tetramer, octomer, dodecamer and so on [74]. The aggregation-enhancing property of Htt17 with respect to Q_N alone can be then explained by the formation of reversible α-helical tetrameric bundles via Htt17. Namely, these tetramers assemble into higher order oligomers that increase the local concentration of the amyloidogenic Q_N segment easing the nucleation of the β-sheeted structures necessary to the formation of amyloid fibrils.

These previous investigations indicate common motifs in Htt17 – the formation of helical structures and the sequestration of its non-polar residues – that are of fundamental to both its aggregation and membrane-binding enhancing properties. We now discuss how our observations on the Htt17, Htt17Q_{17} and Htt17$Q_{17}P_{11}$ monomers are related to these models.

We observe in our simulations the presence of α-helical conformations ($29.3 \pm 0.7\%$) in Htt17. We also quantify, more specifically, the presence of highly α-helical structures in Htt17 by comparing the sampled ensemble in aqueous solution to its membrane-bound state. We compute the RMSD with respect to the α-helical structure sampled in a POPC bilayer [110] and reconstruct the two-dimensional FES of Htt17 in terms of this RMSD and the number of helical H-bonds (Figure 4.4). We observe a broad basin between 1.0 and 7.5 helical H-bonds and 0.3 and 0.7 nm RMSD, corresponding to structures that are different from their membrane counterpart. The free energy gradually rises as the structural similarity to the membrane state increases indicating that a highly α-helical structure similar to the membrane-bound state is not stable in aqueous solution.

We note, nonetheless, that some configurations in the basin below the 4 kJ/mol isoline possess a motif that could initiate the membrane-binding and the formation of the tetrameric α-helical bundle : the first residues of Htt17, particularly between residues Thr3 to Lys6, can form an α-helix ($\sim$50%, Figure 4.2). The presence of such motif has also been observed in other simulations [70]. Moreover, our results show that the non-polar residues of Htt17 are mostly accessible to the solvent (Figure S6), particularly

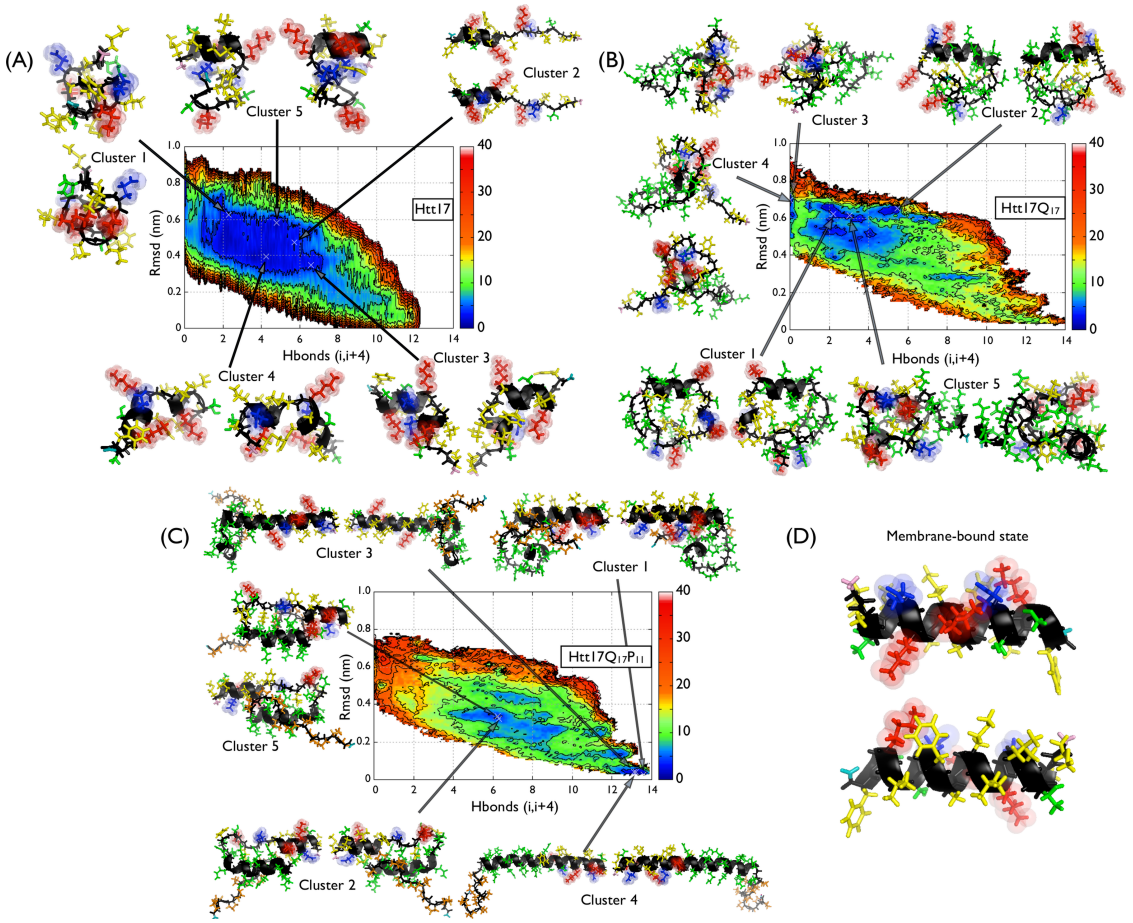

Figure 4.4 : Comparison of the conformational ensemble of Htt17 to the membrane-bound state [110]. The FES of the Htt17 segment as a function of the backbone RMSD measured between Htt17 in solution and the membrane-bound state (vertical axis) and the number of helical H-bonds (S_α, horizontal axis) is shown along with the representative principal conformation clusters. The membrane-bound state is depicted in **(D)** The negatively charged, positively charged, non polar and polar residues of Htt17 are shown in blue, red, yellow and green. The Q_{17} and the P_{11} segments are respectively coloured in green and orange. The backbone is coloured in black, the amino-terminus in pink and the carboxy-terminus in teal. Energy isolines are drawn every 4 kJ/mol.

Met8, Phe11 and Phe17, which could promote the association of Htt17 with other Htt17 segments and its anchoring to a phospholipid membrane.

Upon addition of Q_{17}, the population of a highly helical Htt17 segment is significantly reduced (Figure 4.2), moving away from the membrane-bound state (Figure 4.4).

This is explained by the α-helix motif between residues 3 and 7 being less stable than for Htt17 alone even though a new α-helical motif starting at residue 15 in Htt17 and extending in the first glutamines of Q_{17} is formed. This shift of the position of the α-helical motif could result in a shift of Htt17Q_N's primary interaction site with phospholipid membranes and nucleation site for the tetrameric bundles formation during oligomerization.

We also observe that the non-polar residues of Htt17 are globally less exposed due to interactions with the aliphatic carbons of the glutamines (Figures S6 and S7). Together with the α-helical shift, this could affect the aggregation and membrane-binding pathways as the Q_N region needs to move away from the Htt17 region to free the non-polar residues for these events to proceed. This might be one of the rate limiting steps for the tetrameric bundle formation as intrapeptide Htt17–Q_N interactions needs to be dominated by interpeptide interactions between Htt17–Q_N or Q_N–Q_N during dimerization. Previous simulations indeed indicate that the Q_N–Q_N interaction dominate in the dimer [68].

Our results also suggest that the stability of Htt17 in a fully formed α-helix state drastically increases with the addition of the $Q_{17}P_{11}$ domain (Figures 4.1 and 4.2). This results in an important population of membrane-bound like states characterized by a RMSD below 0.1 nm (Figure 4.4) and a high solvent accessibility for the non-polar residues of Htt17 (Figure S6). Both the α-helical character and the non-polar residues accessibility of Htt17 due to the combined addition of Q_N and P_{11} could promote membrane-binding, as observed experimentally [111] and numerically [141], as well as the formation of Htt17 tetrameric bundle.

A priori, this observation from our simulation on the P_{11} role seems to contradict previous experiments. In some studies, the addition of a P_{10} domain decreases the rate of formation and the stability of amyloid-like aggregates leaving the nucleation mechanism unchanged compared to Q_N domain alone [65]. P_{10} would therefore stabilize conformations incompatible with aggregation. Other studies show, however, that the proline-rich segment C38 – starting with P_{11} and located at the carboxy-terminus of Q_N – increases the overall solubility of Htt17Q_NC38, weakening the driving force toward the forma-

tion of insoluble aggregates, but preserving a rate of fibril formation similar to that of Htt17Q_N [78]. Consequently, the formation of structural features that could favor the aggregation of Htt17 in our simulations might not be sufficient to enhance the overall oligomerization. The slowing effect of P_{11} could then occur latter in the aggregation process or be caused by another phenomena such as induced structural changes in the Q_N domain. In that regards, our simulations on the monomer indicate that the structural flexibility of Q_N is reduced by the incorporation of P_{11} (Figure S8)

Other models have been proposed for huntingtin aggregation. In the first one, both Htt17 and C38 – the proline-rich segment starting with P_{11} and located at the carboxy-terminus of Q_N – modulate the aggregation of Q_N by controlling the intrinsic structural heterogeneity of this amyloidogenic segment [78]. Such model is corroborated by dynamic light scattering experiments done on a polyglutamine domain fused to a heterotetrameric coiled-coil system that show that even a disordered N-terminal can facilited the structuring of the polyQ domain [142]. Fibrillation is promoted by Htt17 destabilizing the intermediate non-fibrillar structures and P_{11} destabilizing the intermediate insoluble aggregates. The role for Htt17 was unveiled using Monte-Carlo [68] and mesoscopic [77] simulations that investigated respectively the dimerization and large-scale aggregation of Htt17Q_N. The Htt17 segment would then reduce the entanglement within the Q_N segment and introduces a barrier to intermolecular associations that brings the formation of small spherical structures (soluble oligomers) and large linear aggregates (insoluble fibrils) on similar timescale. In the second model, the amyloidogenic Q_N segment would interact directly with Htt17 to promote fibrillation through the formation of extended motifs in Q_N [105].

The relations between our simulations and these models remain, nevertheless, more limited as we have focused our investigation on the Htt17 segment and do not have a simulation on the Q_N segment alone. We observe, nonetheless, that Q_{17} in the presence of Htt17 adopts a variety of structures that are mostly disordered (Figure S8). It also interacts directly with Htt17 (Figure S7) as previously suggested experimentally [105]. Addition of P_{11} leads to a more compact Q_{17} region that interacts much more with itself and that this could reduce entanglement during oligomerization.

4.6 CONCLUSION

We studied three fragments amino-terminus of Huntingtin – Htt17, Htt17Q_{17} and Htt17$Q_{17}P_{11}$ – with special consideration to the first 17-amino-acids segment (Htt17) that is crucial for its oligomerization and membrane binding. We applied a novel combination of two sampling enhancing techniques – Hamiltonian replica-exchange and well-tempered metadynamics (HREXMetaD) – to have a thorough understanding of the modifications on Htt17's structural ensemble due to the addition of the amyloidogenic Q_N segment and the polyproline segment (P_{11}).

We find that the structural ensemble of Htt17 is characterized by a wide variety of helix/coil conformations. The addition of the Q_{17} domain results in an amino-to-carboxy-terminus shift of the helical content and it decreases the solvent accessibility of Htt17's non-polar residues by interacting directly with it. The addition of both Q_{17} and P_{11} drastically changes the structural ensemble of Htt17 towards more structured conformations with more exposed non-polar surfaces.

Careful comparison with experimental aggregation and membrane-binding models reveals that Htt17 possesses crucial features essential to these processes whether it is combined with Q_{17}, $Q_{17}P_{11}$ or alone. We find that the position and the type of motifs are very different depending on the adjacent sequences to Htt17 showing that all these neighboring regions strongly impact each other already at the monomer level.

Our results also provide a strong basis for further study of more complex situations such as Htt17$Q_N P_{11}$ oligomerization and membrane-binding using a similar simulation protocol (HREXMetaD). We find this novel protocol to offer good sampling at a moderate computational cost and to scale very well with the number of particles as it is essentially size independent.

4.7 SUPPLEMENTARY MATERIAL

See Annex II

4.8 AUTHOR CONTRIBUTIONS

Designed research : VB SC. Performed research : VB SC. Contributed analytic tools : VB SC. Analyzed data : VB SC NM. Wrote the paper : VB SC NM.

4.9 ACKNOWLEDGMENTS

The authors thank the PLUMED community – particularly Giovanni Bussi, Pratyush Tiwary and Ludovico Sutto – for helpful discussions and advices. They also thank Ronald Wetzel and In-Ja Byeon for kindly providing their NMR data and for their insight on our results. Computations were made on the supercomputers Briarée from Université de Montréal and Mammouth from Université de Sherbrooke, managed by Calcul Québec and Compute Canada. The operation of these supercomputers is funded by the Canada Foundation for Innovation, Nano Québec, RMGA, and the Fonds de Recherche Québécois sur la Nature et les Technologies. This work was funded by the Canada Research Chairs program, the Natural Sciences and Engineering Research Council of Canada, the Fonds de Recherche Québécois sur la Nature et les Technologies, and the Fonds de Recherche en Santé du Québec.

CHAPITRE 5

DÉVELOPPEMENT ET OPTIMISATION DES POTENTIELS OPEP.

La structure tridimensionnelle des protéines est liée à leurs rôles biologiques. C'est pourquoi la prédiction de cette structure uniquement à partir de la séquence d'acides aminés demeure un des principaux objectifs de la biologie structurelle. Au chapitre 4 et I, nous avons présenté les résultats sur l'ensemble structurel de Htt17 en solution et en membrane. L'étude de ce genre de peptide représente un défi de taille à cause de son caractère désordonné au niveau du monomère et de la formation de larges oligomères et fibres amyloïdes. Pour étudier l'oligomérisation de ce genre de peptide, il est particulièrement intéressant de se tourner vers des approches moins couteuses au niveau informatique pour faciliter l'échantillonnage et avoir accès à des systèmes plus gros. Dans ce chapitre, nous décrirons nos optimisations du potentiel sOPEP (section 5.3) ainsi que le développement du nouveau potentiel aaOPEP (section 5.4), menant la philosophie de OPEP en régime tout-atome.

5.1 Introduction aux potentiels OPEP

À la fin des années 1990, Philippe Derreumaux introduit le potentiel gros-grain avec solvant implicite OPEP, pour Optimized Potential for Efficient peptide-structure Prediction [143]. Dès son introduction, OPEP fut utilisé avec succès pour la description du repliement d'un modèle d'épingle à cheveux β [144], l'étude de protéines amyloïdes [4] et mena à de relativement bonnes prédictions de la structure tertiaire de huit petits peptides [145].

La représentation gros-grain de OPEP, présentée à la Figure 5.1, est basée sur un modèle à six billes : une bille par atome de la chaîne principale (N, C_α, C, O, H_N) et une bille pour chacune des chaînes latérales, sauf pour la proline qui est représentée en tout-atome (sauf l'hydrogène). Les potentiels gros-grains sont pratiques puisqu'ils permettent de diminuer significativement le temps de simulation en réduisant le nombre

d'atomes à simuler et en simplifiant le paysage énergétique sous-jacent au prix d'une perte de raffinement uniquement accessible au niveau tout-atome.

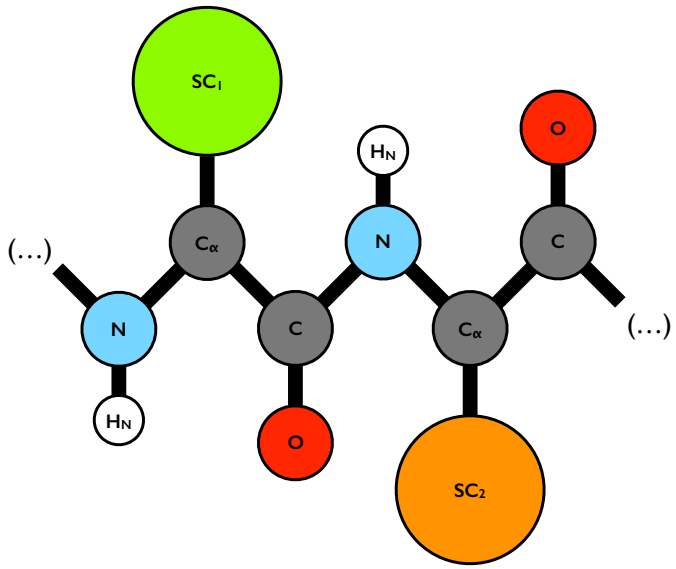

Figure 5.1 : Schéma du modèle gros grains six billes de OPEP.

Des modifications significatives de OPEP survinrent avec le développement de OPEPv3.0 [146] et la description des termes énergétiques est détaillée dans ce qui suit. L'énergie de OPEP peut donc s'écrire en trois termes :

$$E = E_{Locale} + E_{Non-Liée} + E_{Pont-H} \tag{5.1}$$

Le terme d'énergie locale inclus l'énergie des liens, des angles et des angles diédraux de la chaîne principale et latérale selon :

$$E_{local} = w_b \sum_{liens} K_b (r - r_0)^2 + w_\alpha \sum_{angles} K_\alpha (\alpha - \alpha_0)^2 + w_\Omega \sum_{dièdraux} K_n (1 + \cos(n\Omega + \delta_n))$$
$$+ w_{\phi,\psi} \left(\sum_{\phi} k_{\phi\psi} (\phi - \phi_0)^2 + \sum_{\psi} k_{\phi\psi} (\psi - \psi_0)^2 \right)$$

Les constantes de force ainsi que les valeurs à l'équilibre sont tirées du potentiel

AMBER.

Le terme d'énergie non-liée, quant à lui, contient les interactions de van der Waals divisées entre les courtes portées, longues portées (j>i+4), les C_α-C_α et les chaînes latérales/chaînes latérales selon :

$$E_{Non-Liée} = w_{1,4}\sum_{1,4} E_{VdW} + w_{C_\alpha,C_\alpha}\sum_{c_\alpha,c_\alpha} E_{VdW} + w_{1>4}\sum_{M',M'} E_{VdW} + w_{1>4}\sum_{M',C_\alpha} E_{VdW} + w_{1>4}\sum_{M,S_c} E_{VdW} + \sum_{S_c,S_c} w_{S_c,S_c} E_{VdW}$$

M représente tous les atomes de la chaîne principale tandis que M' représente uniquement les atomes N, C, O et H. E_{VdW} prend la forme suivante :

$$E_{VdW} = \varepsilon_{ij}\left(\left(\frac{r_{ij}^0}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^0}{r_{ij}}\right)^{6}\right) H\left(\varepsilon_{ij}\right) - \varepsilon_{ij}\left(\frac{r_{ij}^0}{r_{ij}}\right)^{6} H\left(-\varepsilon_{ij}\right)$$

Où $H(x)$ est la fonction Heaviside de telle sorte que $H(x) = 1$ pour $x \geq 0$ et $H(x) = 0$ pour $x < 0$. Le potentiel de type 12-6 ($H(x) = 1$) est utilisé pour toutes les interactions sauf entre les chaînes latérales. Dans ce second cas, $H(x) = 1$ pour les interactions hydrophobes ou des charges inverses sinon $H(x) = 0$ et le potentiel répulsif est utilisé. Les paramètres initiaux de ε_{ij} entre les chaînes latérales sont tirés d'un potentiel statistique raffiné [147] dérivé de la matrice de contacts entre chaînes latérales obtenue à partir d'un ensemble d'apprentissage de 224 protéines [148].

OPEPv3.0 introduit un terme de coopérativité (E_{HB2}) pour reproduire la coopérativité des ponts-H amides telle que dérivée par des calculs de mécanique quantique et le terme des ponts-H peut être divisé en deux parties, une à deux corps (E_{HB1}) et une à quatre corps (E_{HB2}), selon :

$$E_{Pont-H} = E_{HB1} + E_{HB2}$$

$$E_{HB1} = w_{HB1} \sum_{ij, j \geq i+4} \varepsilon_{HB1} \left(5 \left(\frac{\sigma}{r_{ij}} \right)^{12} - 6 \left(\frac{\sigma}{r_{ij}} \right)^{10} \right) \cos^2 \left(\alpha_{ij} \right) H \left(\alpha_{ij} - 90^\circ \right)$$

$$E_{HB2} = \sum_{\substack{ij,kl \\ (k,l)=(i+1,j+1) \\ \cap (j,l)=(i+4,k+4)}} w_\alpha^{coop} \varepsilon_\alpha^{coop} \exp \left(\frac{-\left(r_{ij} - \sigma\right)^2}{2} \right) \exp \left(\frac{-\left(r_{kl} - \sigma\right)^2}{2} \right)$$

$$+ \sum_{\substack{ij,kl \\ (k,l)=(i+2,j-2) \\ \cup (k,l)=(i+2,j+2)}} w_\beta^{coop} \varepsilon_\beta^{coop} \exp \left(\frac{-\left(r_{ij} - \sigma\right)^2}{2} \right) \exp \left(\frac{-\left(r_{kl} - \sigma\right)^2}{2} \right)$$

La distance r_{ij} est celle séparant l'oxygène du groupement carbonyle de l'hydrogène du groupement amide. L'énergie des ponts-H E_{HB1} est divisée entre courte portée (j=i+4) et longue portée (j>i+4) pour différencier entre les ponts-H d'hélice α et de feuillet β.

Chacun des termes énergétiques est affecté d'un poids (w), qui permet une calibration flexible du potentiel. Au total, on retrouve 261 poids : 210 pour les interactions entre chaînes latérales ($w_{Sc,Sc}$), 1 pour les liens (w_b), 1 pour les angles de liaisons (w_α), 1 pour les angles diédraux (w_Ω), 1 pour les angles diédraux ϕ/ψ ($w_{\phi,\psi}$), 1 pour les interactions C_α-C_α (w_{C_α,C_α}), 1 pour les interactions Lennard-Jones et Lennard-Jones 1-4 ($w_{1>4}$ et $w_{1,4}$), 1 pour les ponts-H et ponts-H 1-4 ($w_{1>4}^{HB1}$ et $w_{1,4}^{HB1}$), 1 pour la coopérativité des ponts-H α et β (w_α^{coop} et w_β^{coop}) et 40 poids associés à la tendance naturelle de chaque résidu à former des structures α et β (w_α^R et w_β^R). Ces poids sont optimisés sur une banque de protéines de telle sorte que la structure native de chacune soit plus stable que les autres, sous-divisés en deux catégories : les Quasi-Natives et les Non-Natives. Ces contraintes correspondent au système d'inégalités suivant :

$$
\begin{aligned}
E_{Native} &< E_{Quasi-Native} \\
E_{Native} &< E_{Non-Native} \\
E_{Quasi-Native} &< E_{Non-Native}
\end{aligned}
\tag{5.2}
$$

La classification des leurres dans les catégories de structure Native, Quasi-Native et Non-Native est réalisée à l'aide du BCscore, un indice décrivant les similarités structurelles (décrit en détails à la section 2.3.2) . Les structures Natives sont très similaires à la structure expérimentale (BCscore $\geq$ 0.9), les Quasi-Natives sont similaires à la structure expérimentale (0.7 $\leq$ BCscore < 0.9) et les Non-Natives sont très différentes de la structure expérimentale (-1 $\leq$ BCscore < 0.7).

L'optimisation des poids est réalisée en maximisant le nombre d'inégalités satisfaites à l'aide d'un algorithme génétique. La première optimisation des poids de OPEPv3.0 se fit sur un ensemble de paramétrisation composé de 13 protéines : 1ABZ, 1DV0, 1E0M, 1ORC, 1PGB, 1QHK, 1SHG, 1SS1, 1VII, 2CI2, Betanova, 1PGBF et 2CRO-fisa. Des structures Quasi-Natives et Non-Natives furent générées couvrant un vaste ensemble de conformations et de structures secondaires pour un total de 7627 leurres. On note au total, une amélioration significative de 42% des inégalités non-satisfaites comparativement aux poids initiaux. La nouvelle paramétrisation fut testée sur un ensemble de validation comprenant 16 nouvelles protéines pour un total de 20 926 leurres. Au total, OPEPv3.0 fut en mesure d'identifier les structures Natives ou Quasi-Natives pour 24 des 29 protéines utilisées.

Cette nouvelle paramétrisation de OPEP fut appliquée avec succès pour des simulations de MD de la protéine Aβ et β2m [149] et de la protéine amyline [150] ainsi que pour des simulations de PT sur Aβ [151] et six petits peptides [152]. De plus, sa validité fut testée à l'aide de la MetaD [129].

L'amélioration et le peaufinement de OPEP s'est poursuivi dans les années suivantes. Dans la version OPEPv4.0 [153], de nouvelles interactions chaînes latérales/chaînes latérales pour les hélices α sont ajoutées et la formulation des interactions non-liées est

modifiée pour :

$$E_{VdW} = w_{S_c,S_c}\varepsilon_{i,j}\left(\left(\frac{G(r_{ij}^0)}{r_{ij}}\right)^6 \exp\left(-2r_{ij}\right) + 0.6563701\left(\tanh\left[2\left(r_{ij}-r_{ij}^0-0.5\right)\right]-1\right)\right) H\left(\varepsilon_{i,j}\right) - \varepsilon_{i,j}\left(\frac{r_{ij}^0}{r_{ij}}\right) H\left(-\varepsilon_{i,j}\right)$$

La forme de cette interaction est inspirée d'un modèle gros-grain de l'ARN [154]. Pour ce qui est des nouvelles interactions entre chaînes latérales, une analyse statistique de 11 structures de la PDB permit d'identifier des contacts (i,i+3) et (i,i+4) dans les hélices α qui sont traités distinctement dans OPEPv4.0. Les nouvelles interactions (i,i+3) et (i,i+4) sont entre les résidus Lys-Glu, Lys-Asp, Glu-Arg et Asp-Arg tandis que les nouvelles (i,i+4) sont entre Lys-Gln, Lys-Leu, Ala-Arg, Ala-Gln, Ala-Glu, Leu-Glu et Ile-Lys. À chacune de ces nouvelles interactions est associée un nouveau poids (w_{S_c,S_c}) qui s'ajoute aux 261 poids de OPEPv3.0. Les 11 nouveaux poids furent optimisés à l'aide d'un algorithme génétique avec les 4 poids associés aux ponts-H en conservant les 257 autres constants. OPEPv4.0 fut testé en MD, où il permit de préserver la structure expérimentale de 17 protéines, et en PT où il retrouva la structure RMN de 3 peptides.

Finalement, une nouvelle paramétrisation améliorant la description des interactions entre chaînes latérales chargées mais de signe opposé (Arg-Asp, Arg-Glu, Lys-Asp, Lys-Glu) est ajoutée dans OPEPv5.0 sans ajouter explicitement des termes électrostatiques [21]. Le nouveau potentiel de ces chaînes latérales fut déterminé à l'aide d'une méthode d'inversion de Boltzmann itérative sur la distribution des distances entre les centres de masse de ces chaînes latérales. Les distributions furent extraites de simulation de MD tout-atome avec solvant explicite. Finalement, le poid du potentiel OPEP (w) pour chacune de ces interactions fut déterminé à l'aide de simulation de PT sur quelques protéines tests en calculant certains paramètres comme le RMSD entre les structures expérimentales et simulées, les courbes de chaleur spécifique, les structures secondaires présentes etc. Il fut déterminé que les meilleurs résultats étaient obtenus pour un poid variant entre

0.9 et 1.0. C'est dans cet intervalle que la stabilité des petites protéines testées était la meilleure et que la température de fusion de l'épingle à cheveux β et la protéine HPr correspond le mieux aux valeurs expérimentales. Cette nouvelle version de OPEP et ces nouveaux paramètres ont un impact positif dans la stabilisation des structures secondaires et améliorent les tendances d'agrégation. Cette version de OPEP est particulièrement pratique pour des systèmes dans lesquels les ponts salins jouent un rôle important.

Il est à noter qu'il n'y a pas de solvant explicite dans les potentiels OPEP. L'ojectif est d'intégrer directement les interactions avec le solvant à l'interieur des divers poids (w) de OPEP lors de la paramétrisation afin de reproduire ces phénomènes implicitement.

Parallèlement au développement des potentiels OPEP, le nouveau potentiel sOPEP, pour OPEP simplifié (simplified OPEP), est introduit lors du développement de l'approche PEPFOLD prédisant la structure de peptides et mini-protéines [5]. Une description détaillée de PEPFOLD est présentée à la section 2.3.1. sOPEP fut développé à partir de la version optimisée de OPEPv3.1 [146]. Comparativement à OPEP qui est conçu pour faire la simulation de protéines dans l'espace des coordonnées cartésiennes, les prédictions de PEPFOLD sont réalisées en reconstruisant la protéine dans un espace discret en superposant des fragments de quatre résidus à l'aide d'un algorithme glouton. Pour que le nouveau potentiel soit adapté à la reconstruction glouton, OPEPv3.1 fut modifié pour donner naissance à sOPEP. Comme la reconstruction se fait à partir de fragments prédéterminés, le potentiel chaîne latérale/chaîne latérale fut modifié pour éviter l'apparition de collisions stériques en contrôlant le paramètre r^0_{ij}, la distance optimale d'interaction, ε_{ij} l'énergie au minimum et R^0_{ij}, la distance à partir de laquelle l'énergie devient répulsive. Le nouveau terme énergétique prend la forme :

$$E_{S_c,S_c}\left(r_{ij}\right) = \begin{cases} -\varepsilon_{ij} \times \frac{2\times R^0_{ij} - r^0_{ij}}{r_{ij}} & \text{pour } \varepsilon_{ij} < 0 \\ \varepsilon_{ij}\left(\left(\frac{r^0_{ij}-p_{ij}}{r_{ij}-p_{ij}}\right)^1 2 - 2\left(\frac{r^0_{ij}-p_{ij}}{r_{ij}-p_{ij}}\right)^6\right) & \text{autre} \end{cases} \tag{5.3}$$

Une comparaison graphique des différents termes d'énergie non-liée est présentée à la figure 5.2. Les paramètres r^0_{ij} et R^0_{ij} furent déterminés par une analyse d'une banque

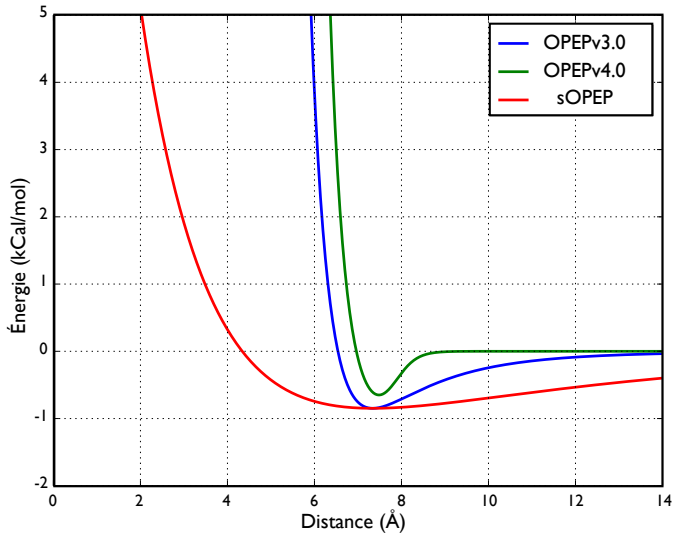

Figure 5.2 : Comparaison de la fonction énergétique pour les interactions de van der Waals entre les chaînes latérales pour OPEPv3.0 (bleu), OPEPv4.0 (vert) et sOPEP(rouge). Les courbes sont tracées spécifiquement pour l'interaction TRP-TRP.

de protéines et le paramètre p_{ij} est obtenu avec la solution de $E(R^0_{ij}) = 0$. Les nouveaux paramètres d'interaction font que l'énergie chaîne latérale/chaîne latérale est plus lisse à courte distance. Dans un premier temps, seuls les 210 poids associés aux interactions chaîne latérale/chaîne latérale furent optimisés à l'aide d'un ensemble de paramétrisation similaire à celui d'OPEPv3.0 [146]. L'optimisation des poids fut ensuite raffinée sur un ensemble de 56 protéines allant jusqu'à 52 acides aminés pour la nouvelle version améliorée PEPFOLDv2.0 [155].

5.2 Protocole d'optimisation

PEPFOLDv2.0 fut en mesure de générer des structures Native et Quasi-Native pour 80% des 56 protéines entre 25 et 52 acides aminés testées [155] et se compare donc avantageusement à PEPFOLDv1.0 et à la très populaire méthode ROSETTA [156]. Par contre, il y a toujours place à amélioration et surtout à l'expansion de PEPFOLD à un régime de protéines plus grandes. Dans cette section, nous détaillerons notre procédure d'optimisation employée pour perfectionner les paramètres de sOPEP et de aaOPEP.

Les potentiels OPEP sont développés en optimisant les fonctions énergétiques et leurs paramètres sur de vastes ensembles d'apprentissage de protéines et petits peptides. Dans la continuité de cette approche, nous avons développé un protocole d'optimisation en trois phases. La première est la construction d'une banque de protéines formant l'ensemble de paramétrisation et de validation. La seconde est la génération des leurres. La dernière est l'optimisation des paramètres à l'aide d'un algorithme génétique. Nous décrirons plus en détail chacune de ces étapes dans ce qui suit.

5.2.1 Identification des protéines cibles

La première étape du protocole est l'identification des protéines cibles qui composeront nos ensembles de paramétrisation et de validation. Pour ce faire, toutes les protéines de moins de 70 acides aminés sont extraites de la PDB. Suivant cette extraction, une série de filtrations en neuf critères, d'abord automatique puis manuelle, permet de conserver (i) les protéines monomériques (ii) avec plusieurs modèles expérimentaux (iii) aux acides aminés standards dont (iv) le pH est entre 5.5 et 8.5 et de rejeter (v) les protéines membranaires, (vi) les protéines amyloïdes, (vii) celles se liant à des ions (viii) ou un ligand et (ix) avec plus de 30% de similarité de séquence. La carte de contacts entre chaînes latérales est ensuite tracée pour déterminer quelles protéines feront partie de quel ensemble. L'objectif est d'obtenir un jeu de paramétrisation représentant optimalement les 210 interactions possibles entre les chaînes latérales, central au modèle d'OPEP, et une bonne diversité de structures secondaires. Finalement, le jeu de paramétrisation est déterminé à l'aide d'une procédure Monte-Carlo visant à déterminer un sous-ensemble de protéines maximisant le nombre de contacts tout en conservant la distribution initiale des topologies pour échantillonner toutes les structures secondaires présentes. Les protéines non retenues sont placées dans l'ensemble de validation. Pour toutes les protéines retenues, le coeur rigide (RC) est calculé similairement à la procédure décrite dans [5]. Tout d'abord, tous les modèles expérimentaux furent superposés avec iSuperpose, disponible sur le serveur MOBYLE [157], puis la déviation quadratique moyenne sur les C_α (C_αRMSF) fut calculée. Les résidus pour lesquels C_αRMSF > 1.5 Å sont considérés comme flexibles et exclus du RC. Le RC établit la section de la protéine sans fluctua-

tion (rigide) et c'est donc sur cette zone uniquement qu'est déterminé si une structure est semblable à celle expérimentale, le reste pouvant grandement varier même selon les modèles expérimentaux.

5.2.2 Génération et classification des leurres

Suite à l'identification des protéines cibles et la construction de notre ensemble de paramétrisation et de validation, il faut maintenant générer les leurres sur lesquels l'optimisation sera lancée. La génération des leurres doit se faire selon deux critères essentiels : (1) leur distribution doit couvrir uniformément une vaste gamme de structures différentes, telle que déterminée à l'aide du BC-Score, une mesure caractérisant les similarités structurelles entre protéines et (2) ils doivent représenter une vaste gamme de structures secondaires. En bref, on veut que la paramétrisation du potentiel soit en mesure de discriminer les structures natives parmi une multitude d'autres structures. Les leurres doivent être générés de telle sorte qu'ils compétitionnent avec la structure native afin que la paramétrisation soit robuste et réponde à cet objectif.

5.2.3 Optimisation via l'algorithme génétique et validation

Une fois les protéines cibles identifiées et les leurres générés, la dernière (et principale) étape est l'optimisation des paramètres à l'aide d'un algorithme génétique. Les différents poids sont optimisés en trois étapes, à l'aide d'un algorithme génétique, de telle sorte que le maximum d'inégalités, selon l'équation 5.2, soit satisfaits. La première étape vise l'optimisation des 210 poids des interactions chaînes latérales/chaînes latérales en gardant les autres nuls. La seconde vise l'optimisation des poids liés aux interactions des ponts-H, en gardant les 210 premiers constants. Finalement, la dernière étape optimise les autres poids en conservant tous les premiers constants. Les poids initiaux de l'optimisation sont tous de un, c'est-à-dire que les paramètres du potentiel sont identiques à ceux déterminés précédemment à l'aide d'un potentiel statistique [147].

5.3 Optimisation du potentiel sOPEP

Le protocole d'optimisation décrit précédemment fut appliqué pour l'optimisation du potentiel sOPEP. Suivant l'extraction et la filtration des protéines cibles, 157 protéines furent retenues afin de former l'ensemble de paramétrisation et de validation. De ces 157 protéines, 62 ont moins de 50 acides aminés et 95 ont entre 51 et 70 acides aminés et sont réunies sous l'étiquette de Prot_0to50aa et Prot_51to70aa respectivement. Le jeu de paramétrisation est déterminé de manière indépendante pour les deux banques de protéines (Prot_0to50aa et Prot_51to70aa) qui sont alors combinées pour former le jeu de paramétrisation total qui servira à l'optimisation des différents poids de sOPEP avec l'algorithme génétique (voir section 5.2.3). Les protéines restantes composeront le jeu de validation qui permettra de certifier l'amélioration de la nouvelle paramétrisation sur les prédictions de PEPFOLD. Pour chacune des protéines, les cartes de contacts furent générées et une procédure Monte-Carlo fut appliquée pour trouver le sous-ensemble maximisant le nombre de contacts. Cette procédure est répétée de façon indépendante 25 fois sur nos deux banques. Des 25 sous-ensembles, celui maximisant le nombre de contacts total et minimisant le nombre de types de contact sans population est utilisé. Les protéines n'ayant pas été retenues dans le processus sont placées dans l'ensemble de validation. Au total 30 protéines de Prot_0to50aa et 50 de Prot_51to70aa, identifiés dans le tableau 5.I, composent l'ensemble de paramétrisation ce qui laisse 32 protéines de Prot_0to50aa et 45 de Prot_51to70aa pour l'ensemble de validation. La carte de contacts entre les diverses chaînes latérales est placée à la figure 5.3 tant pour l'ensemble de paramétrisation que de validation.

En résumé, l'ensemble de paramétrisation comprend 30 protéines de 50 acides aminés et moins et 50 protéines de 51 à 70 acides aminés pour un total de 80 protéines. Les cartes de contacts, présentées à la figure 5.3, montrent une représentation adéquate des 210 interactions possibles entre chaînes latérales. Notons l'exception de la glycine, le plus petit des résidus, et la cystéine, pouvant faire des ponts disulfures non-désirés, dont le nombre de contacts reste assez bas, mais toujours non-nul. L'ensemble de validation est quant à lui composé de 32 protéines de 50 acides aminés et moins et de 45 protéines

de 51 à 70 acides aminés pour un total de 77 protéines.

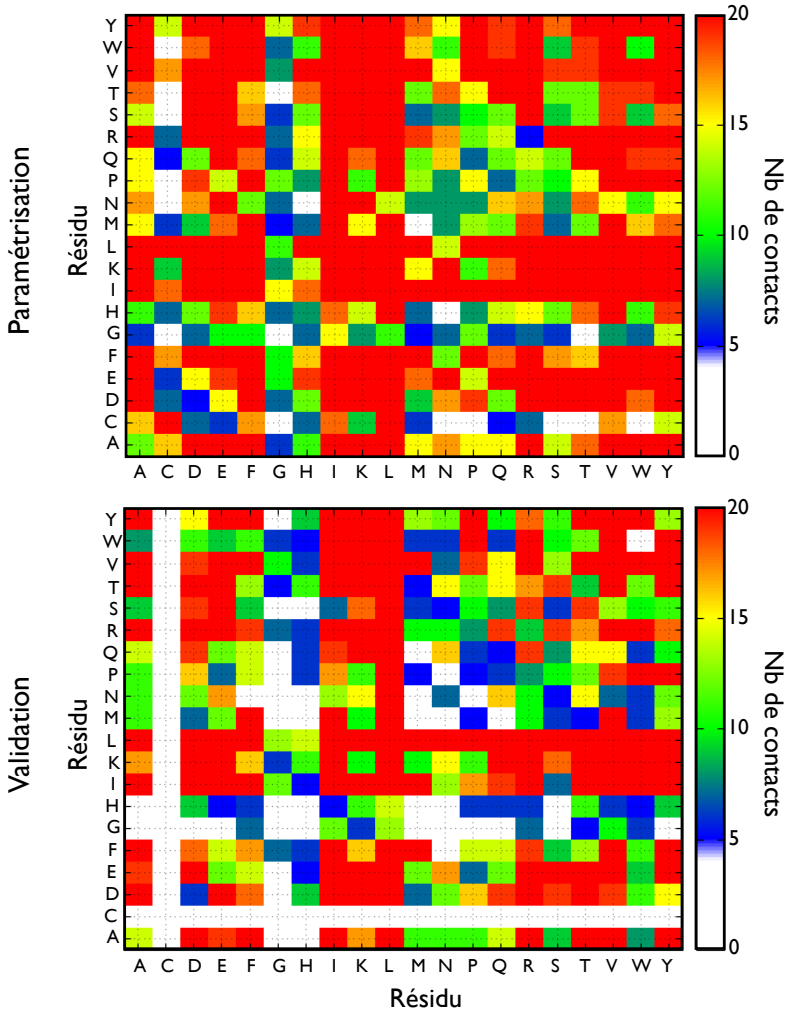

Figure 5.3 : Carte de contact entre les diverses chaînes latérales pour l'ensemble de paramétrisation (haut) et l'ensemble de validation (bas).

La deuxième étape du protocole d'optimisation est la génération des leurres pour chacune des protéines. Les leurres furent générés à l'aide de diverses MD avec comme structure initiale ou la structure expérimentale ou les prédictions de PEPFOLDv2 (dans le cas de Prot_0to50aa). Les quatre types de MD appliquées sont (i) une MD sans contrainte à 300K, (ii) une MD sans contrainte à 500K, (iii) une MD avec contrainte sur les angles diédraux du RC et (iv) une MD avec contrainte sur les structures secondaires. Addition-

Tableau 5.I : Tableau des protéines utilisées dans l'ensemble de paramétrisation et de validation.

Ensemble	Classe	Nb protéines	Liste
Paramétrisation	Prot0to50	30	1b03 1bhi 1fvy 1i6c 1ify 1jjs 1nd9 1pgy 1pv0 1spw 1vpu 1w4g 1wr7 1x32 1yiu 1zrj 2bbp 2ekk 2j8p 2k76 2k9d 2l92 2l9v 2m6o 2m8j 2msu 2oru 2ysb 2ysi 4uzw
	Prot51to70	50	1cok 1cpz 1du6 1f0z 1fex 1go5 1hyw 1ity 1ne3 1q1v 1qpm 1qxf 1rq6 1u97 1w4i 1wcn 1wgn 1wji 1y2y 1zv6 2b7e 2bby 2bzt 2c0s 2cob 2cp9 2cw1 2d9a 2dak 2dt6 2e45 2hbp 2jrr 2jtv 2k3b 2k85 2kac 2kaf 2l54 2lhc 2loj 2lss 2m4h 2m4y 2m7o 2mdj 2mck 2rnq 2xk0 4c26
Validation	Prot0to50	32	1by0 1dv0 1e0n 1ed7 1hu7 1k1v 1k91 1oeg 1rij 1s4j 1ywj 1yyb 1zn5 2bn6 2dmv 2e4e 2evq 2jnh 2jof 2jp5 2kbl 2ki0 2l0g 2lma 2lrx 2mdu 2mih 2mlj 2mwf 2p81 2wxc 2ysc
	Prot51to70	45	1g2h 1gyf 1gyz 1irz 1k8b 1n87 1r4g 1r73 1rzs 1uxd 1v65 1w3d 1wgl 1wmv 1yez 1z4h 1zwv 1zxg 2a63 2coo 2dii 2do3 2dy8 2eqi 2fce 2fmr 2jtm 2jvd 2k2a 2k57 2k5k 2kt2 2kz9 2l4m 2l8d 2l8n 2l93 2lsm 2m2l 2mi6 2mkx 2v0e 2wqg 2ymj 3gcc

nellement, nous avons construit des leurres plus structurés. Les leurres en hélice α sont générés à l'aide du programme PYMOL tandis que ceux en feuillet β le sont à l'aide de simulation de MD biaisée pour la formation des ponts-H. Au total, plusieurs dizaines de milliers de structures furent générées. Pour éviter un excès de structures similaires et un temps d'optimisation excessif, les leurres furent triés de telle sorte que seuls ceux ayant un BC-Score unique sont conservés. Dans le cas où deux structures ont le même BC-Score, celle ayant l'énergie minimale est conservée. Finalement, les leurres furent classés en trois catégories selon leur BC-Score (description détaillée de cette mesure à la section 2.3.2). Les structures Natives, Quasi-Natives et Non-Natives ont respectivement

un BC-Score plus grand que 0.9, entre 0.7 et 0.9 et plus petit que 0.7. On compte un total de 26 773 806 inégalités à résoudre : 6 444 046 inégalités entre la classe Native et Quasi-Native, 1 808 430 inégalités entre la classe Native et Non-Native et 18 521 330 inégalités entre la classe Quasi-Native et Non-Native.

La dernière étape du protocole est l'optimisation à l'aide de l'algorithme génétique. Au total, pour sOPEP, on veut optimiser 218 poids : 210 pour les interactions chaînes latérales/chaînes latérales, deux pour les ponts-H (courte et longue portée), deux pour la coopérativité α/β, un pour les interactions C_α/C_α, un pour les interactions Lennard-Jones, un pour les angles PHI-PSI et un pour les interactions Lennard-Jones 1-4. Afin de ne pas trop s'éloigner des paramètres originaux, nous avons limité les valeurs possibles que peuvent prendre les poids. Une limite souple (soft wall) est placée à 0.75/1.25, et peut être franchie avec compensation d'une pénalité, tandis qu'une limite maximale (hard wall), ne pouvant être dépassée, est placée à 0.45/1.55. Nous avons choisi ces limites en comparant les énergies initiales des différents leurres. La limite souple et maximale permettent respectivement de modifier la tendance et les fluctuations énergétiques pour discriminer entre une structure Native et Non-Native au niveau de l'énergie.

Les statistiques concernant la satisfaction des inégalités sont présentées au tableau 5.II. Suivant l'optimisation des paramètres, 80.9% des inégalités sont résolues pour notre ensemble de paramétrisation. Nous avons de plus calculé le taux de corrélation entre l'énergie totale de chacun des leurres et le BCscore. La corrélation doit être négative puisque l'on désire que l'énergie soit de plus en plus faible plus on s'approche de la structure expérimentale. Cette tendance est respectée pour toutes les protéines de l'ensemble de paramétrisation sauf quatre (1pv0, 2cp9, 2j8p et 4c26). Le taux de corrélation moyen entre toutes les protéines est de -0.33. Comparons maintenant ces résultats avec la précédente paramétrisation de sOPEP. Sur notre ensemble de paramétrisation, sOPEP V1.0 permet de résoudre 74.1% des inégalités totales. On note une amélioration globale de +6.8% du nombre d'inégalités satisfaites et donc une amélioration de 26.3% des inégalités initiales non-résolues avec notre nouvelle optimisation. Cette amélioration au niveau des différentes inégalités ne se traduit par contre pas par une amélioration du taux de corrélation moyen qui reste inchangé à -0.33. Sept protéines sont problématiques avec

sOPEP V1.0 soit 1pv0, 2c0s, 2cp9, 2j8p, 2m4h, 2mck, 2oru. Notre nouvelle paramétrisation améliore les résultats pour les protéines 2c0s, 2m4h, 2mck et 2oru mais les détériore pour 4c26.

Tableau 5.II : Comparaison du taux de satisfactions des inégalités pour sOPEP v1.0 et notre nouvelle optimisation. Les données sont tirées de nos simulations de 15 ns à 50 K.

Optimisation	Inégalités satisfaites	Taux de satisfaction	Corrélation
sOPEP v1.0	19 835 622 / 26 773 806	74.1%	-0.33
sOPEP nouveau	21 660 703 / 26 773 806	80.9%	-0.33

Afin de vérifier la validité de notre optimisation, nous avons aussi lancé une série de simulations de MD à basse température (25K, 50K et 100K) sur neuf protéines de 43 à 70 acides aminés ; 15 ns pour chaque protéine et chaque température. Toutes les simulations commencent du modèle expérimental extrait de la PDB. Pour ces neufs protéines, la structure fut déterminée à l'aide de RMN en solution. De ces neuf protéines, cinq ont un excellent taux de corrélation entre l'énergie et le BCscore (< -0.65) soit 1f0z, 2bzt, 2jrr, 2jtv et 2lss. Les quatre restantes, 1wgn, 2k3b, 2m8j et 2mck, sont celles montrant le plus grand taux d'amélioration comparativement à sOPEP V1.0. Notons que pour la protéine 2mck, le taux de corrélation passe de positif (+0.49) dans sOPEP v1.0 à négatif (-0.09) dans notre nouvelle version, une amélioration considérable. Ces neuf protéines offrent un bon bilan des points forts et faibles de notre nouvelle optimisation. L'analyse présentée dans les prochains paragraphes est tirée des simulations à 50 K. Une analyse visuelle du BCscore et de l'énergie potentielle en fonction du temps montre que les simulations convergent pour toutes les protéines après sept nanosecondes. La simulation fut poursuivie jusqu'à 45 ns pour 2jtv sans changement majeur. Les moyennes sont donc calculées sur les huit dernières nanosecondes. Nous terminerons en décrivant les étapes futures de l'optimisation.

Commençons par une brève discussion sur l'effet de la température en considérant les simulations à 25 K et 100 K. Il est essentiel de noter au départ que la paramétrisation de sOPEP se fait à température nulle. Ainsi les températures de transitions obtenues ne

sont pas rigoureuses et ne sont pas physiques. Les températures sont choisies arbitrairement. À basse température (25 K), la structure native est mieux conservée pour toutes les protéines sauf pour 2jrr et 2jtv qui elles sont mieux conservées à 50 K. À plus haute température (100 K), on note un dépliement important de 1f0z et 2jtv tandis que les autres protéines restent relativement stables. Dans le cas de 1f0z, les structures obtenues restent tout de même dans la famille Quasi-Native des leurres avec un BCscore de 0.819.

Commençons par une analyse des cinq protéines ayant un excellent taux de corrélation (< -0.65). À 50 K, les protéines 1f0z et 2jtv demeurent très stables avec un BCscore moyen à convergence de 0.935 et 0.927 respectivement, ce qui les garde dans la catégorie Native des leurres. Les structures natives des trois autres protéines sont en comparaison moins stables avec un BCscore moyen de 0.834, 0.898 et 0.892 pour 2bzt, 2jrr et 2lss respectivement. Elles demeurent tout de même dans la catégorie Quasi-Native et plus près de la Native que de la Non-Native. En terme de RMSD, calculé sur les carbones α et uniquement sur le RC (C_α RMSD), 1f0z, 2bzt, 2jrr, 2jtv et 2lss convergent respectivement à 2.479 Å, 3.663 Å, 2.271 Å, 2.401 Å et 2.951 Å. La comparaison entre la structure expérimentale (native) et la structure obtenue après 15 ns de MD à 50 K est présentée à la figure 5.4 pour toutes les protéines. Les résidus colorés (rouge/bleu) correspondent à ceux dont la distance C_α-C_α dévie de la moyenne de 50% ou plus. Pour la protéine 1f0z, ces résidus se retrouvent dans la boucle connectant le premier segment hélice α avec le prochain brin β. Pour la protéine 2bzt, ces résidus composent principalement la région désordonnée du N-terminal ainsi que la boucle reliant la première à la seconde hélice α. Pour 2jrr, les résidus problématiques forment le premier brin β de la protéine. Pour 2jtv, ces résidus se situent principalement dans les régions connectrices entre les différents brins β ainsi que dans la dernière hélice α. Finalement et similairement, les régions problématiques pour 2lss se situent dans les boucles reliant les divers éléments de structure secondaires. Globalement, on remarque que les plus grandes déviations entre la structure expérimentale et la structure stable donnée par notre paramétrisation de sOPEP sont dans les régions flexibles (coudes/boucles) reliant les divers brins β et hélice α. Au niveau des structures secondaires, on note une légère surstabilisation de l'hélice α pour 1f0z (21% vs 12%), 2bzt (63% vs 54%) et 2jrr (6% vs 0%) comparativement à la

structure expérimentale. Le taux de feuillet β est quant à lui bien conservé pour toutes les protéines testées. Ces simulations de MD montrent que les structures expérimentales restent relativement stables avec la nouvelle version de sOPEP. En effet, les protéines restent dans une région de BCscore associée à des leurres natifs ou bien quasi-natifs.

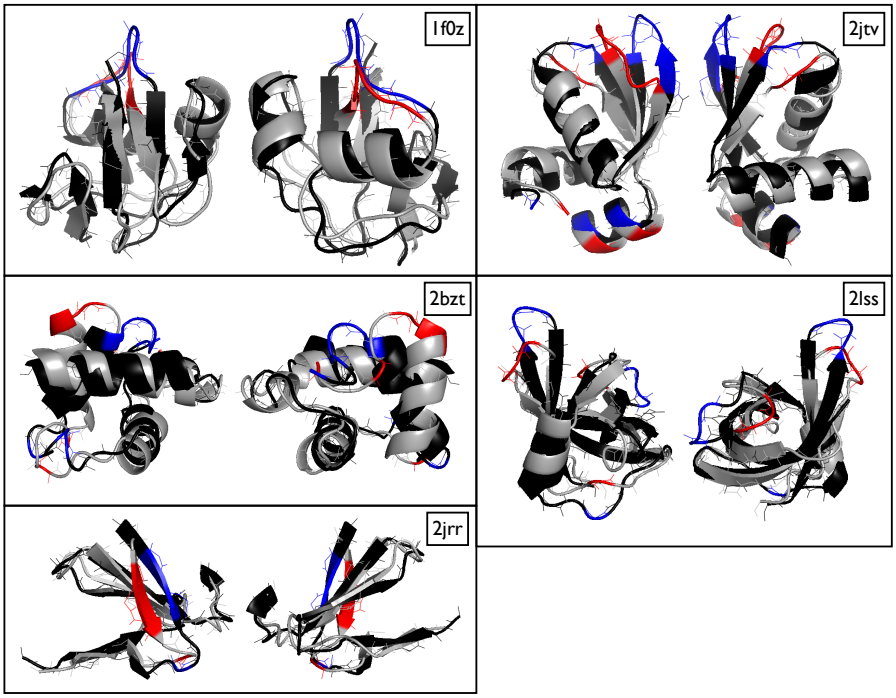

Figure 5.4 : Stabilité de la structure native en MD à 50 K à l'aide du potentiel sOPEP. La structure native (en noir) est comparée à la structure suivant 15 ns de MD à 50K (en gris). Les résidus dont la distance C_α-C_α est supérieure à la moyenne de 50% sont colorés en bleu pour la structure native et en rouge pour la structure finale. Seuls les résidus composant le RC sont présentés tandis que les autres sont cachés.

Nous explorerons dans ce qui suit les différences entre notre nouvelle optimisation et la version 1.0 de sOPEP. Pour ce faire, nous avons réalisé de nouvelles simulations de MD à 50 K avec sOPEP v1.0 et sOPEP nouveau sur quatre protéines 1wgn, 2k3b, 2m8j et 2mck. Les simulations partent de la structure expérimentale. Ces quatre protéines furent sélectionnées puisqu'elles voient leur taux de corrélation entre l'énergie et le BCscore grandement s'améliorer d'une version à l'autre. Le bilan des simulations pour chacune des protéines est présenté au tableau 5.III. Pour deux protéines sur quatre,

1wgn et 2k3b, on voit que les résultats s'améliorent considérablement tant au niveau du BCscore que du C_α RMSD. Les structures à l'équilibre demeurent dans la famille Native avec sOPEP nouveau tandis qu'elles tombent dans la famille Quasi-Native avec sOPEP v1.0. Pour la protéine 2mck, la structure expérimentale demeure plus stable au niveau du BCscore avec sOPEP v1.0 qu'avec OPEP nouveau. Par contre, on note une amélioration significative (20%) au niveau du C_α RMSD. Finalement, la protéine 2m8j demeure plus stable avec sOPEP v1.0 qu'avec OPEP nouveau, tant au niveau du BCscore que du C_α RMSD. L'amélioration du taux de corrélation ne semble donc pas avoir d'effet sur la structure expérimentale. Au niveau de la structure secondaire, tant sOPEP v1.0 que sOPEP nouveau conservent bien les divers motifs, tant d'hélice α que de feuillet β pour trois protéines sur quatre. Pour 2k3b, on observe que sOPEP v1.0 sous-stabilise les structures β de la protéine par 30.508% tandis que OPEP nouveau les surestime de 20.320%. La génération de leurres aux structures β plus variées devrait permettre de résoudre cette problématique. Attardons-nous maintenant aux résidus problématiques. Similairement aux observations faites sur les cinq protéines à excellent taux de corrélation, les résidus contribuant le plus au C_α RMSD se situent dans les régions connectrices flexibles (coudes et boucles) reliant les divers motifs de la structure secondaire. On note la récurrence de plusieurs acides aminés dans les régions problématiques. Les résidus chargés, tout particulièrement l'Asp, la Lys et l'Arg, comptent pour près de 25% des résidus problématiques identifiés. Les plus petits résidus, l'Ala, la Gly et la Ser, composent quant à eux 33% des résidus problématiques. On note que tous les contacts Asp/Lys/Arg avec Ala/Gly/Ser sont considérés comme répulsifs dans le potentiel statistique utilisé par sOPEP. Reconsidérer le caractère répulsif de ces interactions serait une solution potentielle afin d'améliorer les prédictions au niveau des coudes et boucles.

En conclusion, notre nouvelle optimisation de sOPEP permet une amélioration de 26.3% du taux de résolution des inégalités mais n'a aucun effet sur le taux de corrélation moyen qui reste inchangé à -0.33. Des MD à faible température sur cinq protéines ayant un excellent taux de corrélation montrent que la structure expérimentale demeure dans la famille Native ou Quasi-Native dans tous les cas. Quatre MD additionnelles avec sOPEP v1.0 et sOPEP nouveau montrent que notre nouvelle optimisation amé-

Tableau 5.III : Comparaison des simulations de MD entre le potentiel sOPEP v1.0 et la nouvelle paramétrisation. Les données sont tirées de nos simulations de 15 ns à 50 K. Les moyennes présentées pour le BCscore et le C_α RMSD sont calculées sur l'intervalle de convergence. Les lignes Δ-α et Δ-α présentent la différence de structure secondaire (hélice α et feuillet β) avec le modèle expérimental.

Protéine		sOPEP v1.0	sOPEP nouveau
	Taux de Corrélation	-0.15	-0.56
	BCscore	0.852	0.979
1wgn	C_α RMSD (Å)	3.279	1.367
	Δ-α	+3.175%	+3.175%
	Δ-β	0.000%	0.000%
	Taux de Corrélation	-0.27	-0.78
2k3b	BCscore	0.773	0.926
	C_α RMSD (Å)	4.263	2,263
	Δ-α	0.000%	0.000%
	Δ-β	-30.508%	+20.320%
	Taux de Corrélation	-0.05	-0.42
	BCscore	0.625	0.615
2m8j	C_α RMSD (Å)	4.360	5.131
	Δ-α	0.000%	0.000%
	Δ-β	0.000%	0.000%
	Taux de Corrélation	0.49	-0.05
	BCscore	0.700	0.622
2mck	C_α RMSD (Å)	7.217	5.768
	Δ-α	0.000%	0.000%
	Δ-β	+2.899%	+0.226%

liore considérablement la stabilité de la structure expérimentale pour deux protéines au niveau du BCscore et trois au niveau du C_α RMSD. On note que les résidus problématiques se situent presque exclusivement dans les régions flexibles reliant les divers motifs de la structure secondaire. Une grande proportion de ces résidus sont chargés (Asp,Lys,Arg) ou possèdent une petite chaîne latérale (Ala,Gly,Ser). Il devrait être possible d'appréhender cette problématique en modifiant le caractère répulsif de certaines

interactions chaînes-latérales/chaîne-latérales. Je suis confiant que les résultats pourront être raffinés. En effet, l'optimisation et la validation de sOPEP se poursuivront dans les mois à venir via trois avenues principales. La première est la génération d'un plus grand nombre de leurres aux structures secondaires variées. Une nouvelle version du serveur MOBYLE [157] devrait être disponible sous peu et permettre de les générer avec plus de facilité et d'efficacité. La seconde est de tester différentes modifications au niveau des termes énergétiques entre chaînes latérales. Sans changer les poids initiaux du potentiel statistique [147], certains termes pourraient passer d'attractifs à répulsifs et vice-versa afin de mieux représenter l'ensemble des contacts présents dans les structures expérimentales. La dernière est la vérification de l'optimisation sur les prédictions de PEPFOLD, une application prometteuse pour nos nouveaux résultats.

5.4 Optimisation et développement de aaOPEP : Un aperçu

Nous développons présentement dans le groupe un nouveau potentiel, aaOPEP, qui étendra la philosophie des potentiels OPEP dans le régime tout-atome. En effet, aaOPEP conservera les interactions spécifiques entre chaînes latérales à l'aide d'une paramétrisation tirée d'un potentiel statistique et les poids permettant d'imiter implicitement les effets du solvant, les interactions électrostatiques, etc. De plus, nous conservons la coopérativité entre les ponts-H associés à la formation de structure secondaire α et β. Le tout-atome permettra un meilleur empaquetage des chaînes latérales et permettra de corriger certains problèmes observés à ce niveau dans les boucles [158] et les protéines globulaires [159]. Les ponts-H entre chaîne-latérale et chaîne-principale seront aussi permis. Ce nouveau potentiel offre une approche intermédiaire entre les potentiels gros-grain et tout-atome. En effet, sa représentation tout-atome permettra un meilleur niveau de précision que le gros-grain, tout en demeurant moins coûteux au niveau informatique grâce à sa représentation implicite du solvant.

Similairement à sOPEP, les interactions de aaOPEP sont affectées d'un poids permettant le peaufinement de la paramétrisation du potentiel. Dans le cas de aaOPEP, on compte 263 poids : 210 associés aux interactions chaînes latérales/chaînes latérales,

quatre pour les ponts-H (i,j=i+3), (i,j=i+4), (i,j=i+5) et (i,j>i+5), deux pour la coopérativité α et β, 40 pour la propension de chaque résidu à former des structures α et β, un pour les interactions C_α-C_α, deux pour les interactions Lennard-Jones et Lennard-Jones 1-4, un pour les liens atomiques, un pour les angles, un pour les angles dihédraux et un pour les angles diédraux ϕ/ψ. Comme les chaînes latérales sont maintenant tout-atome, les 210 poids associés à chacune des interactions possibles entre chaînes latérales affectent maintenant certains atomes lourds de celles-ci selon la situation. Dans le cas de deux chaînes latérales non-polaires ou une polaire et une non-polaire, les carbones aliphatiques sont désignés, sauf si les deux sont aromatiques, alors les carbones du cycle sont désignés. Finalement, dans le cas de deux résidus polaires, les résidus impliqués dans l'interaction adéquate (pont-H ou pont salin) sont désignés.

Au cours des prochains mois, les leurres générés pour sOPEP seront convertis en tout-atome à l'aide de SCWRL4 [160] puis minimisés afin d'éviter les collisions stériques tout en conservant les positions de la chaîne principale. Par la suite, un protocole d'optimisation similaire à celui décrit précédemment sera appliqué aux paramètres de aaOPEP. Cette paramétrisation à température nulle sera ensuite étendue à température non nulle. Ce nouveau potentiel offrira une vaste gamme d'application, entre autres dans le peaufinement tout-atome des résultats de PEPFOLD et dans l'étude des protéines amyloïdes. À suivre !

CHAPITRE 6

CONCLUSION

Les protéines sont des nanomachines sophistiquées aux fonctions essentielles à tout organisme. On les retrouve dans presque tous les processus biologiques allant de la réplication et la traduction de l'ADN, au transport membranaire en passant par la catalyse de réactions chimiques et la dégradation. Ces quelques exemples ne sont qu'un mince échantillon de l'immense diversité des tâches qu'accomplissent ces molécules. C'est la structure tridimensionnelle (tertiaire) des protéines qui leur permettent d'accomplir leurs tâches. Au niveau primaire, les protéines sont des chaînes composées des 20 acides aminés de base avec chacun leurs caractéristiques propres. Sous l'action des lois fondamentales de la physique, cette chaîne se replie sur elle même localement pour former la structure secondaire puis globabement pour former la structure tertiaire, à l'origine de la fonction.

6.1 La protéine Huntingtine

Au chapitre 4, nous avons étudié le monomère de Htt^{NT} en solution afin de comprendre les impacts structurels de chacun des trois fragments qui le composent et d'identifier des motifs potentiellement liés à son agrégation et à son interaction avec la membrane. Nos résultats indiquent que Htt17 seul adopte une grande variété de conformations désordonnées, mais avec une certaine tendance à la formation d'hélice α, tout particulièrement dans pour sa première moitié. Les résidus apolaires restent accessibles au solvant. Nos résultats concordent avec les mesures expérimentales de déplacements chimiques et de dichroïsme circulaire. Ces deux caractéristiques pourraient jouer un rôle crucial dans la formation de tétramètres par Htt17 ou pour son interaction avec la membrane. L'ajout du segment Q_N modifie grandement la structure de Htt17. Si la quantité d'hélice α reste globalement inchangée, c'est maintenant plutôt sa seconde moitié qui est structurée. Cette conformation se propage dans les premiers résidus de Q_N. Globablement,

Htt17 et Q_N forment une sorte d'anneau permettant de mieux isoler les résidus apolaires au coeur de celui-ci. Finalement, l'ajout de P_{11} a comme effet de stabiliser Htt17 dans une structure d'hélice α presque complète. En bref, nos résultats suggèrent que l'initialisation du mécanisme d'agrégation et d'ancrage à la membrane prendra une différente forme selon les fragments de Htt^{NT} présents. En effet, la structure de Htt17 tant dans la membrane [67] que dans les fibres amyloïdes [74], adopte une structure d'hélice α. Une transition vers l'hélice α devient nécessaire tant pour Htt17 que Htt17-Q_{17}, tandis que nos résultats montrent que Htt17 dans Htt17-Q_{17}-P_{11} se retrouve en solution déjà sous la forme d'hélice α. De plus, l'accord entre nos résultats et l'expérience montre que notre protocole de simulation, alliant MetaD et HREX, est robuste. La continuation du projet sera d'utiliser notre méthodologie afin d'étudier l'initialisation de l'agrégation de Htt^{NT} via la formation de tétramère par Htt17 en solution. Une autre avenue serait de continuer l'exploration de Htt^{NT} en membrane afin de complémenter nos précédents résultats, présentés dans l'annexe I, par une étude de la potentielle dimérisation de Htt17 en membrane. Finalement, une étude de la transition solution/membrane de Htt17 permettrait de faire le pont entre nos résultats obtenus dans ces deux environnements.

6.2 Développement et amélioration de OPEP

La dernière portion de cette thèse discutait du protocole d'optimisation appliqué pour le peaufinement des paramètres de sOPEP et aaOPEP. Nous avons premièrement identifié un ensemble de protéines entre 0 et 70 acides aminés que nous avons distribué dans un ensemble de paramétrisation et un ensemble de validation. Pour chacune des protéines, nous avons généré de multiples structures afin de couvrir une plage variée de structures secondaires et de conformations différentes. Enfin, le jeu d'apprentissage fut utilisé afin d'optimiser, à l'aide d'un algorithme génétique, les paramètres de sOPEP de telle sorte qu'il soit en mesure de mieux discriminer les structures natives (expérimentales) des autres structures. Suite à l'optimisation, nous avons observé une amélioration de 25% du taux de résolution des inégalités comparativement à la première version de sOPEP. Les améliorations furent confirmées à l'aide de simulations de dynamique moléculaire à

basse température sur les structures expérimentales de neuf protéines.

Au cours des prochains mois, l'optimisation de sOPEP se poursuivra le long de trois axes principaux. Le premier est le peaufinement des structures de l'ensemble de paramétrisation et de validation à l'aide d'une version de PEPFOLD mise à jour sur le serveur MOBYLE afin d'ajouter des leurres aux structures secondaires variées. La seconde est la révision des interactions entre chaînes latérales. Les dynamiques moléculaires effectuées pour ce mémoire permirent d'identifier certains problèmes avec les chaînes latérales. Les paramètres de cette fonctionnelle d'énergie seront peaufinés de même que le caractère attractif/répulsif de certaines chaînes-latérales, notamment les chargés (Asp, Lys, Arg) et ceux ayant une petite chaîne latérale (Ala, Gly, Ser) seront révisés. Ces modifications devraient permettre de corriger la situation pour les protéines problématiques. Finalement, nous tenterons d'intégrer les 40 poids associés à la tendance naturelle des résidus à former des hélices α/feuillets β de OPEPv3.0 dans sOPEP. Cette dernière modification devrait permettre de briser les structures secondaires pour mieux représenter les régions plus désordonnées.

Similairement, nous développons présentement le potentiel aaOPEP qui portera la philosophie de OPEP au régime tout-atome. Du côté théorique, les fonctionnelles de aaOPEP et les termes énergétiques à considérer seront déterminés. Suivant cela, nous optimiserons aaOPEP similairement à sOPEP. Ce nouveau potentiel servira à ajouter une étape de raffinement tout-atome des prédictions de PEPFOLD afin d'en améliorer les capacités de prédiction. Cette première paramétrisation à température nulle sera ensuite étendue à un régime de température non nulle. Une fois complété, le potentiel aaOPEP permettra de simuler des systèmes plus gros sur des échelles de temps plus longues que les potentiels traditionnels tout-atome. Son application à l'étude des protéines amyloïdes, formant des oligomères et de larges fibres, est une avenue prometteuse qui sera explorée. Finalement, les potentiels OPEP seront inclus dans une code de dynamique moléculaire (opep_sim) qui sera offerte à la communauté scientifique. À suivre !

BIBLIOGRAPHIE

[1] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507) :1304–1351, 2001.

[2] Iakes Ezkurdia, David Juan, Jose Manuel Rodriguez, Adam Frankish, Mark Diekhans, Jennifer Harrow, Jesus Vazquez, Alfonso Valencia, and Michael L Tress. Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human molecular genetics*, 23(22) :5866–5878, 2014.

[3] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1) :235–242, 2000.

[4] Normand Mousseau and Philippe Derreumaux. Exploring the early steps of amyloid peptide aggregation by computers. *Accounts of chemical research*, 38(11) :885–891, 2005.

[5] Julien Maupetit, Philippe Derreumaux, and Pierre Tufféry. A fast method for large-scale de novo peptide and miniprotein structure prediction. *Journal of computational chemistry*, 31(4) :726–738, 2010.

[6] David J Lilja. *Measuring computer performance : a practitioner's guide*. Cambridge university press, 2005.

[7] Michael Levitt. The birth of computational structural biology. *Nature Structural & Molecular Biology*, 8(5) :392–393, 2001.

[8] Aneesur Rahman and Frank H Stillinger. Molecular dynamics study of liquid water. *The Journal of Chemical Physics*, 55(7) :3336–3359, 1971.

[9] David E Shaw, Ron O Dror, John K Salmon, JP Grossman, Kenneth M Mackenzie, Joseph A Bank, Cliff Young, Martin M Deneroff, Brannon Batson, Kevin J

Bowers, et al. Millisecond-scale molecular dynamics simulations on anton. In *High performance computing networking, storage and analysis, proceedings of the conference on*, pages 1–11, 2009.

[10] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587) :484–489, 2016.

[11] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325) :602–606, 2017.

[12] Tamar Schlick. *Molecular modeling and simulation : an interdisciplinary guide : an interdisciplinary guide*, volume 21. Springer Science & Business Media, 2010.

[13] Wendy D Cornell, Piotr Cieplak, Christopher I Bayly, Ian R Gould, Kenneth M Merz, David M Ferguson, David C Spellmeyer, and Thomas and Fox. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19) :5179–5197, 1995.

[14] Alex D MacKerell Jr, Donald Bashford, MLDR Bellott, Roland Leslie Dunbrack Jr, Jeffrey D Evanseck, Martin J Field, Stefan Fischer, Jiali Gao, H Guo, Sookhee Ha, et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins†. *The journal of physical chemistry B*, 102(18) :3586–3616, 1998.

[15] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins : Structure, Function, and Bioinformatics*, 65(3) :712–725, 2006.

[16] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L Klepeis, Ron O Dror, and David E Shaw. Improved side-chain torsion potentials for

the amber ff99sb protein force field. *Proteins : Structure, Function, and Bioinformatics*, 78(8) :1950–1958, 2010.

[17] Alexander D Mackerell. Empirical force fields for biological macromolecules : overview and issues. *Journal of computational chemistry*, 25(13) :1584–1604, 2004.

[18] Steve O Nielsen, Carlos F Lopez, Goundla Srinivas, and Michael L Klein. Coarse grain models and the computer simulation of soft materials. *Journal of Physics : Condensed Matter*, 16(15) :R481, 2004.

[19] Valentina Tozzini. Coarse-grained models for proteins. *Current opinion in structural biology*, 15(2) :144–150, 2005.

[20] Siewert J Marrink and D Peter Tieleman. Perspective on the martini model. *Chemical Society Reviews*, 42(16) :6801–6822, 2013.

[21] Fabio Sterpone, Phuong H Nguyen, Maria Kalimeri, and Philippe Derreumaux. Importance of the ion-pair interactions in the opep coarse-grained force field : parametrization and validation. *Journal of chemical theory and computation*, 9(10) :4574–4584, 2013.

[22] Luca Monticelli, Senthil K Kandasamy, Xavier Periole, Ronald G Larson, D Peter Tieleman, and Siewert-Jan Marrink. The martini coarse-grained force field : extension to proteins. *Journal of chemical theory and computation*, 4(5) :819–834, 2008.

[23] Lingle Wang, Richard A Friesner, and BJ Berne. Replica exchange with solute scaling : A more efficient version of replica exchange with solute tempering (rest2). *The Journal of Physical Chemistry B*, 115(30) :9431–9438, 2011.

[24] Giovanni Bussi. Hamiltonian replica exchange in gromacs : a flexible implementation. *Molecular Physics*, 112(3-4) :379–384, 2014.

[25] Sander Pronk, Szilárd Páll, Roland Schulz, Per Larsson, Pär Bjelkmar, Rossen Apostolov, Michael R Shirts, Jeremy C Smith, Peter M Kasson, David van der Spoel, et al. Gromacs 4.5 : a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, page btt055, 2013.

[26] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C Smith, Berk Hess, and Erik Lindahl. Gromacs : High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1 :19–25, 2015.

[27] Gareth A Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. Plumed 2 : New feathers for an old bird. *Computer Physics Communications*, 185(2) :604–613, 2014.

[28] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20) :12562–12566, 2002.

[29] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. Metadynamics. *Wiley Interdisciplinary Reviews : Computational Molecular Science*, 1(5) :826–843, 2011.

[30] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics : A smoothly converging and tunable free-energy method. *Physical review letters*, 100(2) :020603, 2008.

[31] Ludovico Sutto, Simone Marsili, and Francesco Luigi Gervasio. New advances in metadynamics. *Wiley Interdisciplinary Reviews : Computational Molecular Science*, 2(5) :771–779, 2012.

[32] Massimiliano Bonomi, Alessandro Barducci, and Michele Parrinello. Reconstructing the equilibrium boltzmann distribution from well-tempered metadynamics. *Journal of computational chemistry*, 30(11) :1615–1621, 2009.

[33] Pratyush Tiwary and Michele Parrinello. A time-independent free energy estimator for metadynamics. *The Journal of Physical Chemistry B*, 119(3) :736–742, 2014.

[34] T Alwyn Jones and Soren Thirup. Using known substructures in protein model building and crystallography. *The EMBO Journal*, 5(4) :819, 1986.

[35] Anne-Cloude Camproux, R Gautier, and P Tuffery. A hidden markov model derived structural alphabet for proteins. *Journal of molecular biology*, 339(3) :591–605, 2004.

[36] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2) :257–286, 1989.

[37] Pierre Tuffery, Frédéric Guyon, and Philippe Derreumaux. Improved greedy algorithm for protein structure reconstruction. *Journal of computational chemistry*, 26(5) :506–513, 2005.

[38] Rachel Kolodny, Patrice Koehl, Leonidas Guibas, and Michael Levitt. Small libraries of protein fragments model native protein structures accurately. *Journal of molecular biology*, 323(2) :297–307, 2002.

[39] Frédéric Guyon and Pierre Tufféry. Fast protein fragment similarity scoring using a binet–cauchy kernel. *Bioinformatics*, 30(6) :784–791, 2014.

[40] Harry T Orr and Huda Y Zoghbi. Trinucleotide repeat disorders. *Annu. Rev. Neurosci.*, 30 :575–621, 2007.

[41] Jean-Paul Vonsattel, Richard H Myers, Thomas J Stevens, Robert J Ferrante, Edward D Bird, and Edward P Richardson Jr. Neuropathological classification of huntington's disease. *Journal of Neuropathology & Experimental Neurology*, 44(6) :559–577, 1985.

[42] John Labbadia and Richard I Morimoto. Huntington's disease : underlying molecular mechanisms and emerging concepts. *Trends in biochemical sciences*, 38(8) :378–385, 2013.

[43] Fanny Mochel and Ronald G Haller. Energy deficit in huntington disease : why it matters. *The Journal of clinical investigation*, 121(2) :493–499, 2011.

[44] Tamara Seredenina and Ruth Luthi-Carter. What have we learned from gene expression profiles in huntington's disease ? *Neurobiology of disease*, 45(1) :83–98, 2012.

[45] Marian DiFiglia, Ellen Sapp, Kathryn O Chase, Stephen W Davies, Gillian P Bates, JP Vonsattel, and Neil Aronin. Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain. *Science*, 277(5334) :1990–1993, 1997.

[46] Ronald Wetzel. Physical chemistry of polyglutamine : intriguing tales of a monotonous sequence. *Journal of molecular biology*, 421(4) :466–490, 2012.

[47] G Cisbani and F Cicchetti. An in vitro perspective on the molecular mechanisms underlying mutant huntingtin protein toxicity. *Cell death & disease*, 3(8) :e382, 2012.

[48] Laura Mangiarini, Kirupa Sathasivam, Mary Seller, Barbara Cozens, Alex Harper, Colin Hetherington, Martin Lawton, Yvon Trottier, Hans Lehrach, Stephen W Davies, et al. Exon 1 of the hd gene with an expanded cag repeat is sufficient to cause a progressive neurological phenotype in transgenic mice. *Cell*, 87(3) :493–506, 1996.

[49] Gareth A Palidwor, Sergey Shcherbinin, Matthew R Huska, Tamas Rasko, Ulrich Stelzl, Anup Arumughan, Raphaele Foulle, Pablo Porras, Luis Sanchez-Pulido, Erich E Wanker, et al. Detection of alpha-rod protein repeats using a neural network and application to huntingtin. *PLoS Comput Biol*, 5(3) :e1000304, 2009.

[50] Ihn Sik Seong, Juliana M Woda, Ji-Joon Song, Alejandro Lloret, Priyanka D Abeyrathne, Caroline J Woo, Gillian Gregory, Jong-Min Lee, Vanessa C Wheeler, Thomas Walz, et al. Huntingtin facilitates polycomb repressive complex 2. *Human molecular genetics*, 19(4) :573–583, 2010.

[51] Marie-Thérèse El-Daher, Emilie Hangen, Julie Bruyère, Ghislaine Poizat, Ismael Al-Ramahi, Raul Pardo, Nicolas Bourg, Sylvie Souquere, Céline Mayet, Gérard Pierron, et al. Huntingtin proteolysis releases non-polyq fragments that cause toxicity through dynamin 1 dysregulation. *The EMBO journal*, 34(17) :2255–2271, 2015.

[52] Rona K Graham, Yu Deng, Elizabeth J Slow, Brendan Haigh, Nagat Bissada, Ge Lu, Jacqueline Pearson, Jacqueline Shehadeh, Lisa Bertram, Zoe Murphy, et al. Cleavage at the caspase-6 site is required for neuronal dysfunction and degeneration due to mutant huntingtin. *Cell*, 125(6) :1179–1191, 2006.

[53] Kirupa Sathasivam, Andreas Neueder, Theresa A Gipson, Christian Landles, Agnesska C Benjamin, Marie K Bondulich, Donna L Smith, Richard LM Faull, Raymund AC Roos, David Howland, et al. Aberrant splicing of htt generates the pathogenic exon 1 protein in huntington disease. *Proceedings of the National Academy of Sciences*, 110(6) :2366–2370, 2013.

[54] Hyunkyung Jeong, Florian Then, Thomas J Melia, Joseph R Mazzulli, Libin Cui, Jeffrey N Savas, Cindy Voisine, Paolo Paganetti, Naoko Tanese, Anne C Hart, et al. Acetylation targets mutant huntingtin to autophagosomes for degradation. *Cell*, 137(1) :60–72, 2009.

[55] Birgit Schilling, Juliette Gafni, Cameron Torcassi, Xin Cong, Richard H Row, Michelle A LaFevre-Bernt, Michael P Cusack, Tamara Ratovitski, Ricky Hirschhorn, Christopher A Ross, et al. Huntingtin phosphorylation sites mapped by mass spectrometry modulation of cleavage and toxicity. *Journal of Biological Chemistry*, 281(33) :23686–23697, 2006.

[56] Frédéric Saudou and Sandrine Humbert. The biology of huntingtin. *Neuron*, 89(5) :910–926, 2016.

[57] Erica Rockabrand, Natalia Slepko, Antonello Pantalone, Vidya N Nukala, Aleksey Kazantsev, J Lawrence Marsh, Patrick G Sullivan, Joan S Steffan, Stefano L Sensi, and Leslie Michels Thompson. The first 17 amino acids of huntingtin modulate its sub-cellular localization, aggregation and effects on calcium homeostasis. *Human molecular genetics*, 16(1) :61–77, 2007.

[58] Randy Singh Atwal, Jianrun Xia, Deborah Pinchev, Jillian Taylor, Richard M Epand, and Ray Truant. Huntingtin has a membrane association signal that can modulate huntingtin aggregation, nuclear entry and toxicity. *Human molecular genetics*, 16(21) :2600–2615, 2007.

[59] Laura Masino, Geoff Kelly, Kevin Leonard, Yvon Trottier, and Annalisa Pastore. Solution structure of polyglutamine tracts in gst-polyglutamine fusion proteins. *FEBS letters*, 513(2-3) :267–272, 2002.

[60] Fabrice AC Klein, Annalisa Pastore, Laura Masino, Gabrielle Zeder-Lutz, Hélène Nierengarten, Mustapha Oulad-Abdelghani, Danièle Altschuh, Jean-Louis Mandel, and Yvon Trottier. Pathogenic and non-pathogenic polyglutamine tracts have similar structural properties : towards a length-dependent toxicity gradient. *Journal of molecular biology*, 371(1) :235–244, 2007.

[61] Xiaoling Wang, Andreas Vitalis, Matthew A Wyczalkowski, and Rohit V Pappu. Characterizing the conformational ensemble of monomeric polyglutamine. *Proteins : Structure, Function, and Bioinformatics*, 63(2) :297–311, 2006.

[62] Vinal V Lakhani, Feng Ding, and Nikolay V Dokholyan. Polyglutamine induced misfolding of huntingtin exon1 is modulated by the flanking sequences. *PLoS Comput Biol*, 6(4) :e1000772, 2010.

[63] Rozita Laghaei and Normand Mousseau. Spontaneous formation of polygluta-

mine nanotubes with molecular dynamics simulations. *The Journal of chemical physics*, 132(16) :165102, 2010.

[64] Bankanidhi Sahoo, David Singer, Ravindra Kodali, Thole Zuchner, and Ronald Wetzel. Aggregation behavior of chemically synthesized, full-length huntingtin exon1. *Biochemistry*, 53(24) :3897–3907, 2014.

[65] Anusri Bhattacharyya, Ashwani K Thakur, Veronique M Chellgren, Geetha Thiagarajan, Angela D Williams, Brian W Chellgren, Trevor P Creamer, and Ronald Wetzel. Oligoproline effects on polyglutamine conformation and aggregation. *Journal of molecular biology*, 355(3) :524–535, 2006.

[66] Ashwani K Thakur, Murali Jayaraman, Rakesh Mishra, Monika Thakur, Veronique M Chellgren, In-Ja L Byeon, Dalaver H Anjum, Ravindra Kodali, Trevor P Creamer, James F Conway, et al. Polyglutamine disruption of the huntingtin exon 1 n terminus triggers a complex aggregation mechanism. *Nature structural & molecular biology*, 16(4) :380–389, 2009.

[67] Matthias Michalek, Evgeniy S Salnikov, and Burkhard Bechinger. Structure and topology of the huntingtin 1–17 membrane anchor by a combined solution and solid-state nmr approach. *Biophysical journal*, 105(3) :699–710, 2013.

[68] Tim E Williamson, Andreas Vitalis, Scott L Crick, and Rohit V Pappu. Modulation of polyglutamine conformations and dimer formation by the n-terminus of huntingtin. *Journal of molecular biology*, 396(5) :1295–1309, 2010.

[69] Nicholas W Kelley, Xuhui Huang, Stephen Tam, Christoph Spiess, Judith Frydman, and Vijay S Pande. The predicted structure of the headpiece of the huntingtin protein and its implications on huntingtin aggregation. *Journal of molecular biology*, 388(5) :919–927, 2009.

[70] Giulia Rossetti, Pilar Cossio, Alessandro Laio, and Paolo Carloni. Conformations of the huntingtin n-term in aqueous solution from atomistic simulations. *FEBS letters*, 585(19) :3086–3089, 2011.

[71] Matthias Michalek, Evgeniy S Salnikov, Sebastiaan Werten, and Burkhard Bechinger. Membrane interactions of the amphipathic amino terminus of huntingtin. *Biochemistry*, 52(5) :847–858, 2013.

[72] Mee Whi Kim, Yogarany Chelliah, Sang Woo Kim, Zbyszek Otwinowski, and Ilya Bezprozvanny. Secondary structure of huntingtin amino-terminal region. *Structure*, 17(9) :1205–1212, 2009.

[73] Maciej Długosz and Joanna Trylska. Secondary structures of native and pathogenic huntingtin n-terminal fragments. *The Journal of Physical Chemistry B*, 115(40) :11597–11608, 2011.

[74] Murali Jayaraman, Ravindra Kodali, Bankanidhi Sahoo, Ashwani K Thakur, Anand Mayasundari, Rakesh Mishra, Cynthia B Peterson, and Ronald Wetzel. Slow amyloid nucleation via α-helix-rich oligomeric intermediates in short polyglutamine-containing huntingtin fragments. *Journal of molecular biology*, 415(5) :881–899, 2012.

[75] James R Arndt, Samaneh Ghassabi Kondalaji, Megan M Maurer, Arlo Parker, Justin Legleiter, and Stephen J Valentine. Huntingtin n-terminal monomeric and multimeric structures destabilized by covalent modification of heteroatomic residues. *Biochemistry*, 54(28) :4285–4296, 2015.

[76] Bankanidhi Sahoo, Irene Arduini, Kenneth W Drombosky, Ravindra Kodali, Laurie H Sanders, J Timothy Greenamyre, and Ronald Wetzel. Folding landscape of mutant huntingtin exon1 : Diffusible multimers, oligomers and fibrils, and no detectable monomer. *PloS one*, 11(6) :e0155747, 2016.

[77] Kiersten M Ruff, Siddique J Khan, and Rohit V Pappu. A coarse-grained model for polyglutamine aggregation modulated by amphipathic flanking sequences. *Biophysical journal*, 107(5) :1226–1235, 2014.

[78] Scott L Crick, Kiersten M Ruff, Kanchan Garai, Carl Frieden, and Rohit V Pappu. Unmasking the roles of n-and c-terminal flanking sequences from exon 1 of hun-

tingtin as modulators of polyglutamine aggregation. *Proceedings of the National Academy of Sciences*, 110(50) :20075–20080, 2013.

[79] Sébastien Côté, Guanghong Wei, and Normand Mousseau. Atomistic mechanisms of huntingtin n-terminal fragment insertion on a phospholipid bilayer revealed by molecular dynamics simulations. *Proteins : Structure, Function, and Bioinformatics*, 82(7) :1409–1427, 2014.

[80] Alan H Sharp, Scott J Loev, Gabriele Schilling, Shi-Hua Li, Xiao-Jiang Li, Jun Bao, Molly V Wagster, Joyce A Kotzuk, Joseph P Steiner, Amy Lo, et al. Widespread expression of huntington's disease gene (it15) protein product. *Neuron*, 14(5) :1065–1074, 1995.

[81] Marian DiFiglia, Ellen Sapp, Kathryn Chase, Cordula Schwarz, Alison Meloni, Christine Young, Eileen Martin, Jean-Paul Vonsattel, Robert Carraway, Steven A Reeves, et al. Huntingtin is a cytoplasmic protein associated with vesicles in human and rat brain neurons. *Neuron*, 14(5) :1075–1081, 1995.

[82] Mabel P Duyao, Anna B Auerbach, Angela Ryan, Francesca Persichetti, et al. Inactivation of the mouse huntington's disease gene homolog hdh. *Science*, 269(5222) :407, 1995.

[83] Miguel A Andrade and Peer Bork. Heat repeats in the huntington's disease protein. *Nature genetics*, 11(2) :115–116, 1995.

[84] Marcy E MacDonald. Huntingtin : alive and well and working in middle management. *Sci. STKE*, 2003(207) :pe48–pe48, 2003.

[85] Juliane P Caviston and Erika LF Holzbaur. Huntingtin as an essential integrator of intracellular vesicular trafficking. *Trends in cell biology*, 19(4) :147–155, 2009.

[86] Zhiqiang Zheng and Marc I Diamond. Huntington disease and the huntingtin protein. *Progress in molecular biology and translational science*, 107 :189–214, 2011.

[87] T Maiuri, T Woloshansky, J Xia, and R Truant. The huntingtin n17 domain is a multifunctional crm1 and ran-dependent nuclear and cilial export signal. *Human molecular genetics*, page dds554, 2013.

[88] Zhiqiang Zheng, Aimin Li, Brandon B Holmes, Jayne C Marasa, and Marc I Diamond. An n-terminal nuclear export signal regulates trafficking and aggregation of huntingtin (htt) protein exon 1. *Journal of Biological Chemistry*, 288(9) :6063–6071, 2013.

[89] Joan S Steffan, Namita Agrawal, Judit Pallos, Erica Rockabrand, Lloyd C Trotman, Natalia Slepko, Katalin Illes, Tamas Lukacsovich, Ya-Zhen Zhu, Elena Cattaneo, et al. Sumo modification of huntingtin and huntington's disease pathology. *Science*, 304(5667) :100–104, 2004.

[90] Xiaofeng Gu, Erin R Greiner, Rakesh Mishra, Ravindra Kodali, Alex Osmand, Steven Finkbeiner, Joan S Steffan, Leslie Michels Thompson, Ronald Wetzel, and X William Yang. Serines 13 and 16 are critical determinants of full-length human mutant huntingtin induced disease pathogenesis in hd mice. *Neuron*, 64(6) :828–840, 2009.

[91] Charity T Aiken, Joan S Steffan, Cortnie M Guerrero, Hasan Khashwji, Tamas Lukacsovich, Danielle Simmons, Judy M Purcell, Kimia Menhaji, Ya-Zhen Zhu, Kim Green, et al. Phosphorylation of threonine 3 implications for huntingtin aggregation and neurotoxicity. *Journal of Biological Chemistry*, 284(43) :29427–29436, 2009.

[92] Leslie Michels Thompson, Charity T Aiken, Linda S Kaltenbach, Namita Agrawal, Katalin Illes, Ali Khoshnan, Marta Martinez-Vincente, Montserrat Arrasate, Jacqueline Gire O'Rourke, Hasan Khashwji, et al. Ikk phosphorylates huntingtin and targets it for degradation by the proteasome and lysosome. *The Journal of cell biology*, 187(7) :1083–1099, 2009.

[93] Randy Singh Atwal, Carly R Desmond, Nicholas Caron, Tamara Maiuri, Jianrun

Xia, Simonetta Sipione, and Ray Truant. Kinase inhibitors modulate huntingtin cell localization and toxicity. *Nature chemical biology*, 7(7) :453–460, 2011.

[94] Gillian P Bates. History of genetic disease : the molecular genetics of huntington disease—a history. *Nature Reviews Genetics*, 6(10) :766–773, 2005.

[95] Huda Y Zoghbi and Harry T Orr. Glutamine repeats and neurodegeneration. *Annual review of neuroscience*, 23(1) :217–247, 2000.

[96] Jennifer R Gatchel and Huda Y Zoghbi. Diseases of unstable repeat expansion : mechanisms and common principles. *Nature Reviews Genetics*, 6(10) :743–755, 2005.

[97] Caroline L Benn, Christian Landles, He Li, Andrew D Strand, Ben Woodman, Kirupa Sathasivam, Shi-Hua Li, Shabnam Ghazi-Noori, Emma Hockly, Syed MNN Faruque, et al. Contribution of nuclear and extranuclear polyq to neurological phenotypes in mouse models of huntington's disease. *Human molecular genetics*, 14(20) :3065–3078, 2005.

[98] James J Ritch, Antonio Valencia, Jonathan Alexander, Ellen Sapp, Leah Gatune, Gavin R Sangrey, Saurabh Sinha, Cally M Scherber, Scott Zeitlin, Ghazaleh Sadri-Vakili, et al. Multiple phenotypes in huntington disease mouse neural stem cells. *Molecular and Cellular Neuroscience*, 50(1) :70–81, 2012.

[99] Chiara Zuccato, Marta Valenza, and Elena Cattaneo. Molecular mechanisms and potential therapeutical targets in huntington's disease. *Physiological reviews*, 90(3) :905–981, 2010.

[100] Yun J Kim, Yong Yi, Ellen Sapp, Yumei Wang, Ben Cuiffo, Kimberly B Kegel, Zheng-Hong Qin, Neil Aronin, and Marian DiFiglia. Caspase 3-cleaved n-terminal fragments of wild-type and mutant huntingtin are present in normal and huntington's disease brains, associate with membranes, and undergo calpain-dependent proteolysis. *Proceedings of the National Academy of Sciences*, 98(22) :12784–12789, 2001.

[101] Tamara Ratovitski, Masayuki Nakamura, James D'Ambola, Ekaterine Chighladze, Yideng Liang, Wenfei Wang, Rona Graham, Michael R Hayden, David R Borchelt, Ricky R Hirschhorn, et al. N-terminal proteolysis of full-length mutant huntingtin in an inducible pc12 cell model of huntington's disease. *Cell Cycle*, 6(23) :2970–2981, 2007.

[102] Stephen W Davies, Mark Turmaine, Barbara A Cozens, Marian DiFiglia, Alan H Sharp, Christopher A Ross, Eberhard Scherzinger, Erich E Wanker, Laura Mangiarini, and Gillian P Bates. Formation of neuronal intranuclear inclusions underlies the neurological dysfunction in mice transgenic for the hd mutation. *Cell*, 90(3) :537–548, 1997.

[103] Qi Charles Zhang, Tzu-lan Yeh, Alfonso Leyva, Leslie G Frank, Jason Miller, Yujin E Kim, Ralf Langen, Steven Finkbeiner, Mario L Amzel, Christopher A Ross, et al. A compact β model of huntingtin toxicity. *Journal of Biological Chemistry*, 286(10) :8188–8196, 2011.

[104] Leslie G Nucifora, Kathleen A Burke, Xia Feng, Nicolas Arbez, Shanshan Zhu, Jason Miller, Guocheng Yang, Tamara Ratovitski, Michael Delannoy, Paul J Muchowski, et al. Identification of novel potentially toxic oligomers formed in vitro from mammalian-derived expanded huntingtin exon-1 protein. *Journal of Biological Chemistry*, 287(19) :16017–16028, 2012.

[105] Stephen Tam, Christoph Spiess, William Auyeung, Lukasz Joachimiak, Bryan Chen, Michelle A Poirier, and Judith Frydman. The chaperonin tric blocks a huntingtin sequence element that promotes the conformational switch to aggregation. *Nature structural & molecular biology*, 16(12) :1279–1285, 2009.

[106] Susan W Liebman and Stephen C Meredith. Protein folding : sticky n17 speeds huntingtin pile-up. *Nature chemical biology*, 6(1) :7–8, 2010.

[107] Gregory Darnell, Joseph PRO Orgel, Reinhard Pahl, and Stephen C Meredith. Flanking polyproline sequences inhibit β-sheet structure in polyglutamine seg-

ments by inducing ppii-like helix structure. *Journal of molecular biology*, 374(3) :688–704, 2007.

[108] Murali Jayaraman, Rakesh Mishra, Ravindra Kodali, Ashwani K Thakur, Leonardus MI Koharudin, Angela M Gronenborn, and Ronald Wetzel. Kinetically competing huntingtin aggregation pathways control amyloid polymorphism and properties. *Biochemistry*, 51(13) :2706–2716, 2012.

[109] Sébastien Côté, Guanghong Wei, and Normand Mousseau. All-atom stability and oligomerization simulations of polyglutamine nanotubes with and without the 17-amino-acid n-terminal fragment of the huntingtin protein. *The journal of physical chemistry B*, 116(40) :12168–12179, 2012.

[110] Sébastien Côté, Vincent Binette, Evgeniy S Salnikov, Burkhard Bechinger, and Normand Mousseau. Probing the huntingtin 1-17 membrane anchor on a phospholipid bilayer by using all-atom simulations. *Biophysical journal*, 108(5) :1187–1198, 2015.

[111] Kathleen A Burke, Karlina J Kauffman, C Samuel Umbaugh, Shelli L Frey, and Justin Legleiter. The interaction of polyglutamine peptides with lipid membranes is regulated by flanking sequences associated with huntingtin. *Journal of Biological Chemistry*, 288(21) :14993–15005, 2013.

[112] Yanier Crespo, Fabrizio Marinelli, Fabio Pietrucci, and Alessandro Laio. Metadynamics convergence law in a multidimensional system. *Physical Review E*, 81(5) :055701, 2010.

[113] Berk Hess, Carsten Kutzner, David Van Der Spoel, and Erik Lindahl. Gromacs 4 : algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation*, 4(3) :435–447, 2008.

[114] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E Mark, and Herman JC Berendsen. Gromacs : fast, flexible, and free. *Journal of computational chemistry*, 26(16) :1701–1718, 2005.

[115] Herman JC Berendsen, David van der Spoel, and Rudi van Drunen. Gromacs : a message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1) :43–56, 1995.

[116] Robert B Best and Gerhard Hummer. Optimized molecular dynamics force fields applied to the helix- coil transition of polypeptides. *The Journal of Physical Chemistry B*, 113(26) :9004–9015, 2009.

[117] Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. How robust are protein folding simulations with respect to force field parameterization ? *Biophysical journal*, 100(9) :L47–L49, 2011.

[118] Stefano Piana, John L Klepeis, and David E Shaw. Assessing the accuracy of physical models used in protein-folding simulations : quantitative evidence from long molecular dynamics simulations. *Current opinion in structural biology*, 24 :98–105, 2014.

[119] Kresten Lindorff-Larsen, Paul Maragakis, Stefano Piana, Michael P Eastwood, Ron O Dror, and David E Shaw. Systematic validation of protein force fields against experimental data. *PloS one*, 7(2) :e32131, 2012.

[120] Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. Protein folding kinetics and thermodynamics from atomistic simulation. *Proceedings of the National Academy of Sciences*, 109(44) :17845–17850, 2012.

[121] Kyle A Beauchamp, Yu-Shan Lin, Rhiju Das, and Vijay S Pande. Are protein force fields getting better ? a systematic benchmark on 524 diverse nmr measurements. *Journal of chemical theory and computation*, 8(4) :1409–1414, 2012.

[122] Elio A Cino, Wing-Yiu Choy, and Mikko Karttunen. Comparison of secondary structure formation using 10 different force fields in microsecond molecular dynamics simulations. *Journal of chemical theory and computation*, 8(8) :2725–2740, 2012.

[123] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of chemical physics*, 126(1) :014101, 2007.

[124] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald : An n·log (n) method for ewald sums in large systems. *The Journal of chemical physics*, 98(12) :10089–10092, 1993.

[125] Ulrich Essmann, Lalith Perera, Max L Berkowitz, Tom Darden, Hsing Lee, and Lee G Pedersen. A smooth particle mesh ewald method. *The Journal of chemical physics*, 103(19) :8577–8593, 1995.

[126] Berk Hess, Henk Bekker, Herman JC Berendsen, Johannes GEM Fraaije, et al. Lincs : a linear constraint solver for molecular simulations. *Journal of computational chemistry*, 18(12) :1463–1472, 1997.

[127] Shuichi Miyamoto and Peter A Kollman. Settle : an analytical version of the shake and rattle algorithm for rigid water models. *Journal of computational chemistry*, 13(8) :952–962, 1992.

[128] Alessandro Laio and Francesco L Gervasio. Metadynamics : a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Reports on Progress in Physics*, 71(12) :126601, 2008.

[129] Alessandro Barducci, Massimiliano Bonomi, and Philippe Derreumaux. Assessing the quality of the opep coarse-grained force field. *Journal of chemical theory and computation*, 7(6) :1928–1934, 2011.

[130] Carlo Camilloni, Daniel Schaal, Kristian Schweimer, Stephan Schwarzinger, and Alfonso De Simone. Energy landscape of the prion protein helix 1 probed by metadynamics and nmr. *Biophysical journal*, 102(1) :158–167, 2012.

[131] Carlo Camilloni, Davide Provasi, Guido Tiana, and Ricardo A Broglia. Exploring the protein g helix free-energy surface by solute tempering metadynamics. *Proteins : Structure, Function, and Bioinformatics*, 71(4) :1647–1654, 2008.

[132] Dmitrij Frishman and Patrick Argos. Knowledge-based protein secondary structure assignment. *Proteins : Structure, Function, and Bioinformatics*, 23(4) :566–579, 1995.

[133] Yang Shen and Ad Bax. Sparta+ : a modest improvement in empirical nmr chemical shift prediction by means of an artificial neural network. *Journal of biomolecular NMR*, 48(1) :13–22, 2010.

[134] Kai J Kohlhoff, Paul Robustelli, Andrea Cavalli, Xavier Salvatella, and Michele Vendruscolo. Fast and accurate predictions of protein nmr chemical shifts from interatomic distances. *Journal of the American Chemical Society*, 131(39) :13894–13895, 2009.

[135] Xavier Daura, Raymond Suter, and Wilfred F van Gunsteren. Validation of molecular simulation by comparison with experiment : rotational reorientation of tryptophan in water. *The Journal of chemical physics*, 110(6) :3049–3055, 1999.

[136] Frank Eisenhaber, Philip Lijnzaad, Patrick Argos, Chris Sander, and Michael Scharf. The double cubic lattice method : efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *Journal of Computational Chemistry*, 16(3) :273–284, 1995.

[137] David J Barlow and JM Thornton. Ion-pairs in proteins. *Journal of molecular biology*, 168(4) :867–885, 1983.

[138] Stephan Schwarzinger, Gerard JA Kroon, Ted R Foss, John Chung, Peter E Wright, and H Jane Dyson. Sequence-dependent correction of random coil nmr chemical shifts. *Journal of the American Chemical Society*, 123(13) :2970–2978, 2001.

[139] VN Sivanandam, Murali Jayaraman, Cody L Hoop, Ravindra Kodali, Ronald Wetzel, and Patrick CA van der Wel. The aggregation-enhancing huntingtin n-terminus is helical in amyloid fibrils. *Journal of the American Chemical Society*, 133(12) :4558–4566, 2011.

[140] Cody L Hoop, Hsiang-Kai Lin, Karunakar Kar, Zhipeng Hou, Michelle A Poirier, Ronald Wetzel, and Patrick CA van der Wel. Polyglutamine amyloid core boundaries and flanking domain dynamics in huntingtin fragment fibrils determined by solid-state nuclear magnetic resonance. *Biochemistry*, 53(42) :6653–6666, 2014.

[141] Anu Nagarajan, Sudi Jawahery, and Silvina Matysiak. The effects of flanking sequences in the interaction of polyglutamine peptides with a membrane bilayer. *The Journal of Physical Chemistry B*, 118(24) :6368–6379, 2014.

[142] Bashkim Kokona, Zachary P Rosenthal, and Robert Fairman. Role of the coiled-coil structural motif in polyglutamine aggregation. *Biochemistry*, 53(43) :6738–6746, 2014.

[143] Philippe Derreumaux. From polypeptide sequences to structures using monte carlo simulations and an optimized potential. *The Journal of chemical physics*, 111(5) :2301–2310, 1999.

[144] Guanghong Wei, Normand Mousseau, and Philippe Derreumaux. Complex folding pathways in a simple β-hairpin. *Proteins : Structure, Function, and Bioinformatics*, 56(3) :464–474, 2004.

[145] François Forcellino and Philippe Derreumaux. Computer simulations aimed at structure prediction of supersecondary motifs in proteins. *Proteins : Structure, Function, and Bioinformatics*, 45(2) :159–166, 2001.

[146] Julien Maupetit, P Tuffery, and Philippe Derreumaux. A coarse-grained protein force field for folding and structure prediction. *Proteins : Structure, Function, and Bioinformatics*, 69(2) :394–408, 2007.

[147] Marcos R Betancourt and D Thirumalai. Pair potentials for protein folding : choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Science*, 8(02) :361–369, 1999.

[148] Jeffrey Skolnick, Adam Godzik, Lukasz Jaroszewski, et al. Derivation and testing of pair potentials for protein folding. when is the quasichemical approximation correct ? *Protein science*, 6(3) :676–688, 1997.

[149] Yan Lu, Philippe Derreumaux, Zhi Guo, Normand Mousseau, and Guanghong Wei. Thermodynamics and dynamics of amyloid peptide oligomerization are sequence dependent. *Proteins : Structure, Function, and Bioinformatics*, 75(4) :954–963, 2009.

[150] Yuxiang Mo, Yan Lu, Guanghong Wei, and Philippe Derreumaux. Structural diversity of the soluble trimers of the human amylin (20–29) peptide revealed by molecular dynamics simulations. *The Journal of chemical physics*, 130(12) :125101, 2009.

[151] Yassmine Chebaro, Normand Mousseau, and Philippe Derreumaux. Structures and thermodynamics of alzheimer's amyloid-β aβ (16- 35) monomer and dimer by replica exchange molecular dynamics simulations : Implication for full-length aβ fibrillation. *The Journal of Physical Chemistry B*, 113(21) :7668–7675, 2009.

[152] Yassmine Chebaro, Xiao Dong, Rozita Laghaei, Philippe Derreumaux, and Normand Mousseau. Replica exchange molecular dynamics simulations of coarse-grained proteins in implicit solvent. *The journal of physical chemistry B*, 113(1) :267–274, 2008.

[153] Yassmine Chebaro, Samuela Pasquali, and Philippe Derreumaux. The coarse-grained opep force field for non-amyloid and amyloid proteins. *The Journal of Physical Chemistry B*, 116(30) :8741–8752, 2012.

[154] Samuela Pasquali and Philippe Derreumaux. Hire-rna : a high resolution coarse-grained energy model for rna. *The Journal of Physical Chemistry B*, 114(37) :11957–11966, 2010.

[155] Yimin Shen, Julien Maupetit, Philippe Derreumaux, and Pierre Tufféry. Improved

pep-fold approach for peptide and miniprotein structure prediction. *Journal of Chemical Theory and Computation*, 10(10) :4745–4758, 2014.

[156] David E Kim, Dylan Chivian, and David Baker. Protein structure prediction and analysis using the robetta server. *Nucleic acids research*, 32(suppl 2) :W526–W531, 2004.

[157] Bertrand Néron, Hervé Ménager, Corinne Maufrais, Nicolas Joly, Julien Maupetit, Sébastien Letort, Sébastien Carrere, Pierre Tuffery, and Catherine Letondal. Mobyle : a new full web bioinformatics framework. *Bioinformatics*, 25(22) :3005–3011, 2009.

[158] Jean-François St-Pierre and Normand Mousseau. Large loop conformation sampling using the activation relaxation technique, art-nouveau method. *Proteins : Structure, Function, and Bioinformatics*, 80(7) :1883–1894, 2012.

[159] L Dupuis and Normand Mousseau. Understanding the ef-hand closing pathway using non-biased interatomic potentials. *The Journal of chemical physics*, 136(3) :035101, 2012.

[160] Georgii G Krivov, Maxim V Shapovalov, and Roland L Dunbrack. Improved prediction of protein side-chain conformations with scwrl4. *Proteins : Structure, Function, and Bioinformatics*, 77(4) :778–795, 2009.

[161] Burkhard Bechinger and Christina Sizun. Alignment and structural analysis of membrane polypeptides by 15n and 31p solid-state nmr spectroscopy. *Concepts in Magnetic Resonance Part A*, 18(2) :130–145, 2003.

[162] Evgeniy Salnikov, Philippe Bertani, Jan Raap, and Burkhard Bechinger. Analysis of the amide 15n chemical shift tensor of the cα tetrasubstituted constituent of membrane-active peptaibols, the α-aminoisobutyric acid residue, compared to those of di-and tri-substituted proteinogenic amino acid residues. *Journal of biomolecular NMR*, 45(4) :373–387, 2009.

[163] Christopher Aisenbrey and Burkhard Bechinger. Tilt and rotational pitch angle of membrane-inserted polypeptides from combined 15n and 2h solid-state nmr spectroscopy. *Biochemistry*, 43(32) :10502–10512, 2004.

[164] Takaharu Mori, Fumiko Ogushi, and Yuji Sugita. Analysis of lipid surface area in protein–membrane systems combining voronoi tessellation and monte carlo integration methods. *Journal of computational chemistry*, 33(3) :286–293, 2012.

[165] William J Allen, Justin A Lemkul, and David R Bevan. Gridmat-md : A grid-based membrane analysis tool for use with molecular dynamics. *Journal of computational chemistry*, 30(12) :1952–1958, 2009.

[166] Ulrich HE Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chemical Physics Letters*, 281(1) :140–150, 1997.

[167] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters*, 314(1) :141–151, 1999.

[168] Meher K Prakash, Alessandro Barducci, and Michele Parrinello. Replica temperatures for uniform exchange and efficient roundtrip times in explicit solvent parallel tempering simulations. *Journal of chemical theory and computation*, 7(7) :2025–2027, 2011.

[169] Wolfgang Weber, Philippe H Hünenberger, and J Andrew McCammon. Molecular dynamics simulations of a polyalanine octapeptide under ewald boundary conditions : influence of artificial periodicity on peptide conformation. *The Journal of Physical Chemistry B*, 104(15) :3668–3675, 2000.

[170] Abil E Aliev, Martin Kulke, Harmeet S Khaneja, Vijay Chudasama, Tom D Sheppard, and Rachel M Lanigan. Motional timescale predictions by molecular dynamics simulations : Case study using proline and hydroxyproline sidechain dynamics. *Proteins : Structure, Function, and Bioinformatics*, 82(2) :195–215, 2014.

Annexe I

Annexe 1 : Étude de la protéine Huntingtine en membrane

Suite à la proposition d'un premier modèle expérimental de Htt17 en environnement de micelles par Michalek *et al.* [67, 71], nous avons décidé, en collaboration avec ces derniers, de raffiner le modèle proposé à l'aide de simulations numériques tout-atome de Htt17 dans une bicouche de phospholipide. Les fruits de ce travail sont publiés dans le Biophysical Journal [110]. Ce chapitre se contentera d'une description détaillée de ma contribution à l'article mentionné et présentera un résumé des principaux résultats.

Ma contribution à ce projet peut être divisée en deux parties. Premièrement, je fis la mise en place du protocole de simulation HREX (décrit en détail à la section 2.2.1). Je déterminai, entre autres, le nombre et la distribution de répliques nécessaires pour un bon échantillonnage et je m'assurai du bon fonctionnement de la méthode avec notre système.

De plus, je développai certains outils d'analyse permettant une comparaison raffinée avec les résultats expérimentaux. Tout particulièrement, je développai l'outil d'analyse opep_nmr, qui à partir de l'orientation de la protéine par rapport à la membrane, permet de calculer les déplacements chimiques N^{15} et les déplacements quadrupolaires (quadrupolar splitting) H^2, deux mesures déterminées expérimentalement pour Htt17 à l'aide de RMN en solide.

I.1 Déplacement chimique et orientation

À cause des variations spatiales de la densité d'électron, les déplacements chimiques ont un caractère anisotrope, c'est-à-dire qu'ils dépendent de l'orientation avec laquelle ils sont mesurés. En solution, les déplacements chimiques sont moyennés sur un ensemble de structures libres de tourner ce qui a pour effet de faire disparaître l'anisotropie. Ce n'est pas le cas en environnement solide dans lequel les déplacements chimiques dépendent de l'orientation des molécules avec le champ magnétique. Ces propriétés furent

exploitées pour, entre autres, déterminer des structures de peptides en bicouche lipidique. L'essentiel de la théorie est présenté dans [161] et [162] et nous résumerons dans ce qui suit. L'interaction anisotropique peut être décrite par un tenseur d'ordre deux :

$$\sigma = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix}$$

Où σ_{zz} correspond à la composante parallèle au champ magnétique et est la valeur mesurée en RMN. On peut aussi l'écrire dans le système des axes principaux dans lequel cette matrice est diagonale :

$$\sigma = \begin{bmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{11} & 0 \\ 0 & 0 & \sigma_{22} \end{bmatrix}$$

On peut obtenir n'importe quel système de coordonnées à partir de celui-ci à l'aide de rotations autour des angles d'Euler (Φ, Θ, Ψ). Φ est une rotation autour de l'axe z initial et mène vers le système de coordonnées (x', y', z'), Θ est une rotation autour de y' vers le système de coordonnées (x'', y'', z'') et finalement Ψ est une rotation autour de z''.

En fonction des angles d'Euler, on peut trouver la valeur du déplacement chimique mesuré σ_{zz} avec les matrices de rotation. Le résultat est donné par :

$$\sigma_{zz} = \sigma_{11}\sin^2(\Theta)\cos^2(\Phi) + \sigma_{22}\sin^2(\Theta)\sin^2(\Phi) + \sigma_{33}\cos^2(\Theta) \qquad \text{(I.1)}$$

L'orientation de σ_{11}, σ_{22} et σ_{33} dans le plan moléculaire est présenté à la figure I.1. σ_{22} est perpendiculaire au plan peptidique contrairement à σ_{11} et σ_{33}. σ_{11} fait un angle $\beta = 16^\circ$ avec le lien N-H. Les valeurs de σ_{11}, σ_{22} et σ_{33} sont respectivement 56, 81 et 223 ppm, similairement à celles utilisées expérimentalement [71].

Pour ce qui est du déplacement quadrupolaire H^2 ($\Delta\mu_Q$) [163], sa mesure est liée directement à l'orientation du lien C_α-C_β selon :

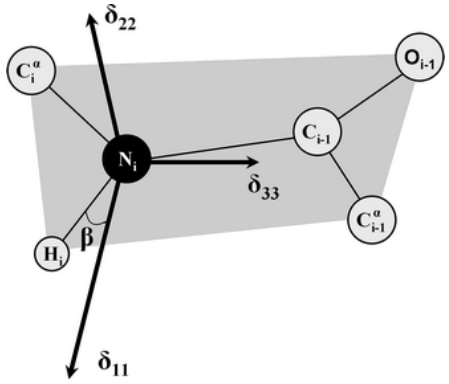

Figure I.1 : Localisation des tenseurs de déplacement chimique dans le plan moléculaire. Les valeurs de σ_{11}, σ_{22} et σ_{33} sont respectivement de 56, 81 et 223 ppm et $\beta = 16^\circ$ selon les données RMN [71]. Figure tirée de [162].

$$\Delta\mu_Q = \frac{3}{2}\left(\frac{e^2qQ}{h}\right)\left(\frac{3\cos^2\theta - 1}{2}\right) \tag{I.2}$$

Où θ est l'angle entre le champ magnétique et le lien C_α-C_β et e^2qQ/h est la constante quadrupolaire de couplage (static quadrupolar coupling constant). Dans nos analyses, cette constante vaut 74 kHz similairement aux expériences RMN [67].

Ces deux mesures sont complémentaires et, ensemble, permettent de déterminer l'alignement du peptide dans la membrane en fonction de l'angle d'inclinaison (tilt) et de tangage (pitch) [163].

I.2 Principaux résultats

Comme nous l'avons vu précédemment, Htt17 est crucial pour l'interaction avec la membrane de plusieurs organites comme le ER et l'appareil de Golgi [57, 93]. Suite au dévoilement d'une première structure expérimentale pour Htt17 en environnement de micelles [71], nous avons décidé, en collaboration avec les auteurs, de raffiner le modèle au niveau atomique à l'aide de simulations numériques tout-atome. Ces simulations furent effectuées dans un environnement membranaire afin de déterminer : (1) Est-ce que la structure obtenue en micelle est stable dans une bicouche lipidique ? (2) Quelles

sont les interactions dominantes entre Htt17 et la membrane ? (3) Quelle est l'orientation du peptide ? (4) Quelles sont les conséquences sur la membrane de l'insertion de Htt17 ? Pour ce faire, nous avons réalisé des simulations de MD et de HREX en les commençant à partir de deux structures initiales différentes : le modèle expérimental déterminé par RMN et un modèle structuré en hélice α. Au total, 11 simulations indépendantes de MD de 1000 ns chacune furent lancées pour chacune de nos structures initiales. Pour HREX, trois simulations indépendantes de 16 répliques furent lancées à partir de la structure α pour un total de 4000 ns (250 ns x 16 répliques) et une simulation de 16 répliques fut lancée à partir du modèle RMN pour un total de 8000 ns (500 ns x 16 répliques).

I.2.1 Structure Secondaire

Le modèle déterminé expérimentalement en environnement de micelles adopte une structure d'hélice α amphipatique entre les résidus 6-17 [67]. Avec nos simulations partant de ce modèle, nous avons observé de larges fluctuations structurelles entre les résidus 6 et 9 (voir Fig I.2). Ces dernières sont causées par le dépliement de l'hélice dans cette région, indiquant que la structure expérimentale déterminée en micelles n'est pas stable en membrane. À titre comparatif, nos simulations de MD et de HREX partant du modèle d'hélice α montrent que ce deuxième modèle reste très stable en membrane avec un taux d'hélice α dépassant le 80% et de faibles fluctuations structurelles (Fig I.2). Pour s'assurer que ce phénomène n'est pas simplement causé par un piètre échantillonnage, nous avons réalisé une nouvelle simulation HREX mais cette fois en partant du modèle expérimental. En fonction du temps, on observe une augmentation marquée du taux d'hélice α indiquant que le repliement du modèle vers une conformation beaucoup plus structurée (Fig I.2). Nos résultats indiquent que Htt17 adopterait une hélice α plus stable et plus structurée en membrane qu'en environnement micellaire.

I.2.2 Orientation

Le modèle expérimental déterminé en environnement de micelles fut aussi étudié par RMN en membrane de phospholipide [71]. Dans cet environnement, Htt17 s'oriente

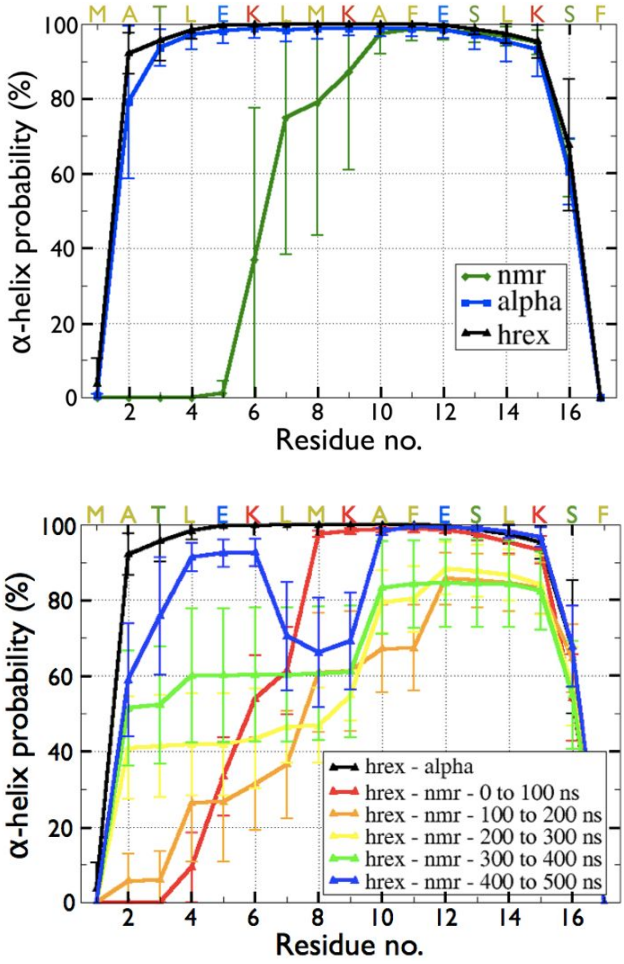

Figure I.2 : (HAUT) : Structure secondaire par résidu de Htt17 en environnement membranaire telle que calculé par STRIDE [132]. Les résultats présentés en vert, bleu et noir correspondent respectivement aux MDs partant de la structure expérimentale, aux MDs partant d'une structure α et les simulations de HREX partant de la structure α. Les résultats et les barres d'erreur correspondent à la moyenne et la déviation standard mesurée sur l'intervalle de convergence (250-1000 ns) des 11 simulations indépendantes pour la MD. Pour HREX, la moyenne et la déviation standard sont calculées sur l'intervalle de convergence (50-250 ns) par tranche de 20 ns. (BAS) : Structure secondaire par résidu de Htt17 en fonction du temps pour une simulation de HREX partant du modèle expérimental.

presque parallèlement à la membrane avec un angle d'inclinaison (tilt) de $103 \pm 5^{\circ}$ de telle sorte que la majorité des résidus apolaires est cachée dans la membrane tandis que la majorité des résidus hydrophiles pointe vers le solvant, un résultat peu surprenant à

cause du caractère amphipatique de Htt17. Par contre, l'angle de tangage (pitch) mesuré est plus grand ($137 \pm 5°$) que ne laisse entendre une simple superposition du plan hydrophobe de Htt17 avec la surface de la membrane. Cette orientation place le résidu Glu12 et Lys15 au niveau de l'interface de la membrane. Ce résultat laisse présager une possible dimérisation, médiée par la formation de ponts salins, de Htt17 dans la membrane. Du côté de nos simulations, nous avons observé que Htt17 est presque parfaitement parallèle à la membrane (voir Fig I.3). Les angles d'inclinaison et de tangage, déterminés à l'aide d'opep_nmr, sont de $94\pm5°/76\pm5°$, $87\pm5°/85\pm5°$, $91\pm5°/95\pm5°$ pour nos simulations de MD partant du modèle expérimental, nos simulations de MD partant du modèle α et nos simulations de HREX partant du modèle α respectivement. Nous n'avons pas observé un décalage au niveau de l'angle de tangage indiquant qu'une telle orientation ne serait pas stable au niveau du monomère. Des simulations en membrane sur une possible dimérisation de Htt17 seraient nécessaires. Nous reviendrons sur l'oligomérisation à la prochaine section. Finalement, au niveau des interactions avec la membrane, on observe la séquestration des résidus apolaires (Leu7, Phe11, Leu14 et Phe17) de Htt17 dans la membrane (voir Fig I.3). L'insertion est moins importante pour les résidus du N-terminal de la MD partant du modèle RMN et ces résidus restent plutôt désordonnés.

I.2.3 Impact sur la membrane

Finalement, nous avons caractérisé les effets de Htt17 sur les propriétés de la membrane. Nous avons observé un rétrécissement de la membrane avec une diminution de la surface par lipides proche de Htt17 (voir Fig I.4). Ces perturbations sont causées par l'étirement des chaînes lipidiques près de Htt17 afin d'en couvrir la surface non polaire. Des prédictions théoriques montrent que de telles déformations de la membrane favorisent le processus de dimérisation du peptide inséré. Bien que nos mesures de l'orientation de Htt17 ne semblent pas indiquer une orientation propice à la dimérisation, les effets sur la membrane observés sont compatibles avec cette proposition expérimentale. De plus, les résidus polaires de Htt17 sont accessibles au solvant et restent disponibles pour la formation de ponts salins. Suivant cette dimérisation, Htt17 pourrait ensuite former des oligomères plus grands et mener à la formation de fibres amyloïdes.

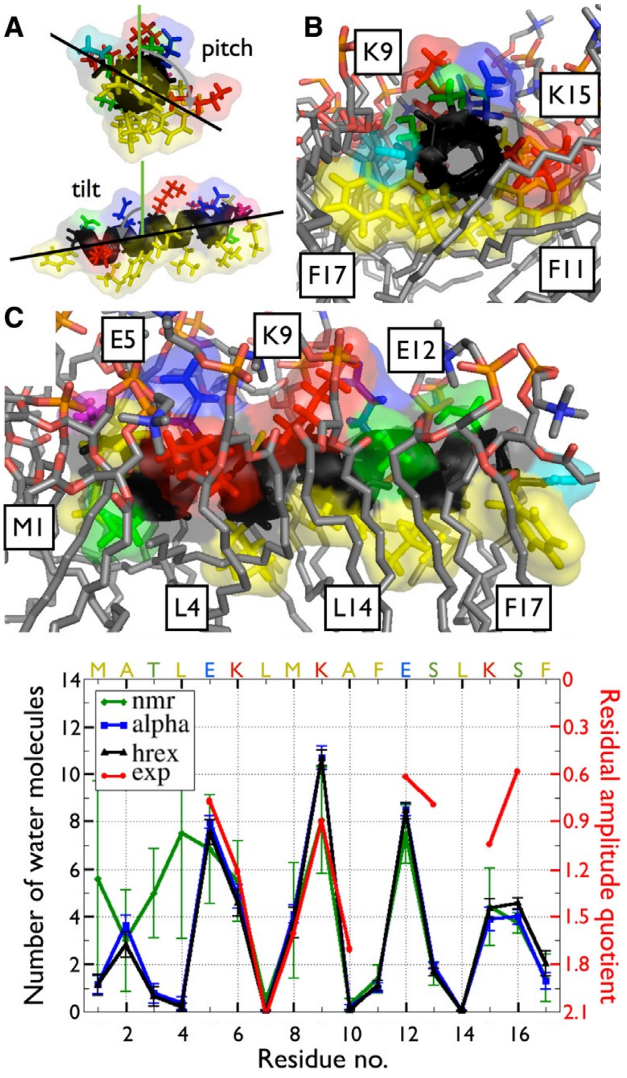

Figure I.3 : (HAUT) : (A) Définition de l'angle d'inclinaison (tilt) et de tangage (pitch). Orientation moyenne de Htt17 dans la membrane vue du C-terminal parallèlement à la membrane (B) et perpendiculairement à la membrane (C). Les résidus apolaires, polaires, chargés positivement et chargés négativement sont montrés respectivement en jaune, vert, rouge et bleu tandis que la chaîne principale est présentée en noir. (BAS) : Nombre de contacts entre molécules d'eau et les différents résidus. Un contact est défini lorsque la distance est moins de 0.35 nm. À nouveau, les résultats présentés en vert, bleu et noir correspondent respectivement aux MDs partant de la structure expérimentale, aux MDs partant d'une structure α et les simulations de HREX partant de la structure α. On retrouve en rouge les résultats expérimentaux de "Residual amplitude quotient" indiquant le degré d'insertion de Htt17 dans la membrane.

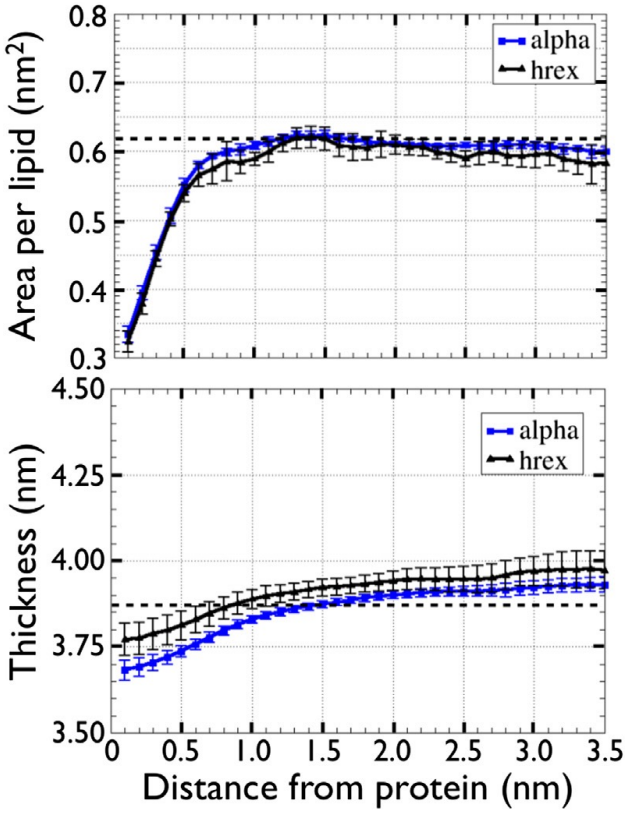

Figure I.4 : Perturbation de la membrane avec l'insertion de Htt17. Les résultats présentés sont pour les simulations de MD (en bleu) et HREX (en noir) partant de la structure α. (HAUT) : Surface moyenne par lipide calculée avec VTMC [164] et (BAS) : Épaisseur moyenne de la membrane calculée avec GRIDMAT [165] en fonction de la distance avec Htt17. Les résultats et les barres d'erreur correspondent à la moyenne et la déviation standard mesurée sur l'intervalle de convergence (250-1000 ns) des 11 simulations indépendantes pour la MD. Pour HREX, la moyenne et la déviation standard sont calculées sur l'intervalle de convergence (50-250 ns) par tranche de 20 ns.

La présence de la membrane pourrait jouer un rôle crucial dans l'apparition de la toxicité.

I.3 Conclusion

En conclusion, suite à la publication d'un premier modèle expérimental pour Htt17 en environnement de micelles [67], nous avons raffiné le modèle au niveau atomique à l'aide de simulations numériques dans une bicouche de phospholipide. Selon nos simu-

lations, Htt17 formerait une hélice α stable et orientée presque parallèlement à la membrane de telle sorte que les résidus apolaires pointent vers la membrane et les résidus polaires restent accessibles au solvant. L'insertion de Htt17 cause des perturbations de la membrane compatibles avec l'hypothèse d'une dimérisation. Ma contribution à ce projet, soit la mise en place de la méthode d'échantillonnage avancée HREX et le développement de l'outil d'analyse opep_nmr, permettant le calcul des déplacements chimiques, apporte une réelle bonification des résultats. En effet, HREX permit d'observer directement le repliement de Htt17 sur la membrane et opep_nmr permit une comparaison fine et directe avec les résultats expérimentaux.

Annexe II

Annexe 2 : Étude de la protéine Huntingtine en solution : Matériel supplémentaire

Supporting Material – Free-energy Landscape of the Amino-terminal Fragment of Huntingtin in Aqueous Solution

Vincent Binette[△], Sébastien Côté[△,*], and Normand Mousseau[*]

Département de Physique and Groupe de recherche sur les protéines membranaires (GEPROM), Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal (Québec), Canada

HREXMetaD on Htt17_coil

We perform a second HREXMetaD simulation on Htt17 starting from a random coil state (Htt17_coil) to assess the robustness of our simulations starting from the NMR model obtained in the presence of DPC micelles (Htt17_nmr). Both systems have the same size and number of water molecules. Both simulations are run at 303K using 16 scales spanning 1.0 to 0.3 with the same intermediate scales. Exchanges between neighboring scales are attempted every 4 ps resulting in an exchange rate of about 20–40%.

The free energy surfaces (FES) in terms of S_α and S_{gyr} for these two simulations indicate that the main features and the extend of the basin are very similar in both simulation sets (compare Figure S1 on Htt17_nmr to Figure S4 on Htt17_coil). The global α-helix probability is also similar : 29.3 $\pm$ 0.7% for Htt17_nmr and 26.9 $\pm$ 0.3% for Htt17_coil. While residues 7 to 12 are less structured in Htt17_coil, the main features of the per residue secondary structure are also preserved : moderate helical content for the first residues, presence of a turn between residues 10 and 13 and more disordered or the last residues (Figure S9). The FES in terms of the number of helical H-bonds (horizontal axis, S_α) and the gyration radius (vertical axis, S_{rg}) displays a similar conformational ensemble characterized by two main structures : two-helix bundle structures (clusters 1,2,4,5), with the first half of the peptide more structured than the second half (clus-

ter 1,4) and almost fully random structures (clusters 3) as shown in Figure S9. Both kinds of structures were also identified in the simulation Htt17_nmr. The FES in terms of the number of helical H-bonds (S_α) and the solvent accessible surface area (SASA) of Htt17's non-polar residues is also similar in both cases as most of the structural ensemble is located between 3 and 5 nm^2 (data not shown). Finally, the contact maps show that the key non-polar and electrostatic contacts are preserved : Met8–Phe17 (24.1% for Htt17_nmr vs. 26.8% for Htt17_coil), Glu5–Lys9 (50.3% for Htt17_nmr vs. 44.5% for Htt17_coil), Glu12–Lys9 (42.3% for Htt17_nmr vs. 38.2% for Htt17_coil), and Glu12–Lys15 (64.0% for Htt17_nmr vs. 58.5 for Htt17_coil) (data not shown).

Overall, we observe an excellent agreement between these two simulations that start from the two very different initial states indicating adequate convergence assessment and sampling of the conformational ensemble.

HREXMetaD vs. PTMetaD for Htt17

In addition to our HREXMetaD simulation on Htt17, we use a second methodology that is very popular – parallel tempering metadynamics (PTMetaD) – to compute the free energy surface of Htt17 in terms of S_α and S_{gyr}. Parallel tempering is often used on its own to simulate protein folding because it increases the probability of escaping free energy minima by allowing exchanges between simultaneous MD simulations at different temperatures [166, 167]. Similarly to HREXMetaD, the combination of MetaD and PT dumbed PTMetaD allows one to correctly sample other CVs not explicitly taken into account by the time-dependent biased potential as demonstrated from proteins with similar conformational ensemble to Htt17 [129–131]. The temperature distribution used for PT spans 278 to 646K and the intermediate temperatures are determined using a recent protocol and requiring an exchange rate of approximately 20% for a total of 64 replicas [168].

The free energy surface in terms of the number of helical H-bonds (S_α) and gyration radius (S_{gyr}) obtained using PTMetaD is shown in Figure S10. We observe that its extent is very similar to that obtained using HREXMetaD (Figure S9), while there are two

minor differences : (i) the FES minimum is now bounded between 3 and 6 helical H-bonds – instead of between 2 and 6 for the HREXMetaD simulations – and has narrower gyration radius bracket and (ii) the fully random structures are slightly less favoured when using PTMetaD.

Even if these changes in the FES lead to a slight increase of the α-helical propensity from 29.3 $\pm$ 0.7% (HREXMetaD) to 38$\pm$ 3% (PT-MetaD), the main features of the secondary structure per residue profile are unchanged with the first half of the peptide being more structured than the second half (compare Figures S9 and S10). From the same Figures, the cluster analysis of the structures sampled in the FES minima (below 5 kJ/mol) further indicates that our HREXMetaD and PTMetaD simulations sample a similar structural ensemble. Indeed, in agreement with our HREXMetaD simulations, we see that Htt17 mostly adopts a two-helix bundle structure (see clusters 1,2 and 4).

The good agreement between our PTMetaD and HREXMetaD simulations demonstrates the robustness of the sampling in both methodologies although we believe that HREX might escape local minima faster than PT because the configurations have significantly less replicas to diffuse in. The use of HREXMetaD might then reduce the probability that MetaD adds wrong biases to the free energy landscape when compared to PTMetaD.

Comparison to the solution NMR experiment on Htt17

With large intensities for H^{α}(i)-H^{N}(i+1) NOEs, medium intensities for H^{α}(i)-H^{N}(i) NOEs and very small ones for medium range NOEs, the structural ensemble sampled during our simulations is largely compatible with the NMR experiment on Htt17 in aqueous solution indicating that it is mostly unstructured in solution [66] (Figure S11).

We refine our analysis by comparing the interproton NOEs for our most scaled replica (replica 16) to our unscaled replica (replica 1) that populate a very different conformational ensemble : the helix propensity is only 3.3 $\pm$ 0.1% for the most scaled replica, while it is 36.9 $\pm$ 0.9% for the unscaled replica. We find an almost identical trend for H^{α}(i)-H^{N}(i) interproton distances, slightly weaker H^{N}(i)-H^{N}(i+1) and H^{α}(i)-H^{N}(i+2) in-

tensities, and stronger H^{α}(i)-H^{N}(i+1) intensities (Figure S11). For its part, the medium-range H^{α}(i)-H^{N}(i+3) NOEs are weaker (data not shown). Taken together, this indicates that the structural ensemble in terms of NOEs of the unscaled replica is dominated by mostly random conformations as the H^{α}(i)-H^{N}(i+1) intensities are very large with a small population of helical conformations as the H^{α}(i)-H^{N}(i+2) and H^{α}(i)-H^{N}(i+3) intensities are stronger compared to the most scaled replica.

We also compute the interproton NOE intensities for two extreme cases : a perfect α-helix and a completely extended conformation. The "perfect" conformations are build with PYMOL and then minimized with the conjugate gradient method to avoid structural clashes. The results are shown in Figure S11. Comparing with both sets, we conclude the intensities observed are consistent with mostly disordered structures.

We note, moreover that we find very low H^{N}(i)-H^{N}(i+1) intensities (below 0.05) except for two residues where the intensities drastically increase to 0.7 indicating that only small structural changes can lead to large fluctuations of this NOE.

Overall, our investigation of the interproton distances shows that a globally good agreement between our simulations and NMR experiments. Indeed, three out of four interproton distances are well conserved with large H^{α}(i)-H^{N}(i+1), medium H^{α}(i)-H^{N}(i) and very small H^{α}(i)-H^{N}(i+2) intensities. We find high intensities for the H^{N}(i)-H^{N}(i+1) NOEs, which seems conflicting with the NMR experiment showing very weak intensities. The lack of sequential H^{N}(i)-H^{N}(i+1) NOE in the NMR experiment is an indicator of a mostly disorder structural ensemble.A more thorough examination of our simulations leads us to believe that the difference is due only to very small and local structural changes as both the fully extended conformation and the structural ensemble sampled by the most scaled replica (having only $3.3 \pm 0.1\%$ of helical content) have H^{N}(i)-H^{N}(i+1) intensities of 0.7 and higher. The presence of H^{α}(i)-H^{N}(i+3) NOEs indicates a small population of helical structures, not found in NMR experiments. We thus conclude that structural ensemble of Htt17 is, at the exception of very local flexibility and small overestimation of the helical content, in agreement with this experiment.

Particle-Mesh Ewald vs. Generalized reaction field

Long-range electrostatic calculation computed using Particle-Mesh Ewald (PME) can, in some cases, artificially increase the α-helical content of peptides possessing an α-helical propensity [169]. To probe the magnitude of this effect on our system, we redid a simulation for the Htt17(2LD2) system and replaced the PME calculation by a Generalized Reaction Field (GRF).

The resulting two-dimensional FES shows a single large minimun bounded by between 2 and 6 helical H-bonds and a gyration radius between 0.6 and 0.8 nm S12. Comparison with the same simulation done with PME shows that the use of the GRF does not alter the position of the free-energy minima. However, the conformations with more than 6 helical H-bonds are slightly destabilized as the free-energy slowly increases with the number of helical H-bonds compared to its PME counterpart. A more detailed analysis of the secondary structure reveals that the GRF reduces the α-helical content passing from $29.7 \pm 0.7\%$ with PME to $17 \pm 4\%$. The per residue secondary structure profile is entirely conserved during the process with the first half of the peptide being more structured than the second one. In both simulation, we observe a mix of two-helix, single helix, helix/coil and coil conformations.

Overall, the results of this simulation indicate that GRF slightly destabilizes the α-helix in our system. The secondary structure per residue profiles of both simulations are nonetheless in very good agreement and we conclude that the GRF and PME scheme show coherent qualitative behavior for Htt17.

Probing the effects of revised proline parameters

In order to probe more meticulously the effect of the poly-proline domain, we simulated Htt17$Q_{17}P_{11}$ using the a recent version of the AMBER99sb*-ILDN forcefield that includes revised proline parameters (AMBER99sb*-ILDNP), derived from fitting experimental correlation times and NMR J couplings [170].

The newly obtained FES shows two main minima, the first one characterized by a set of states with 7 to 10 helical H-bonds and a gyration radius of 0.8 nm and a second

one characterized by around 6 helical H-bonds and a gyration radius slightly lower than 1.0 nm (see Figure S13). Both these regions are also very stable in the simulation using the AMBER99sb*-ILDN forcefield. The main discrepancy between both FES is the decrease of stability of the fully formed α-helix (more than 12 helical H-bonds) by about 6.5 kJ/mol.

In more details, the resulting secondary structure per residue profile is presented in Figure S13. The Htt17 and Q_{17} domain are characterized respectively by $66 \pm 12\%$ and $18 \pm 7\%$ of α-helical content. Those results indicate that revised proline parameters slightly reduce the α-helical content of the Htt17 domain (from $70 \pm 1.6\%$ to $69 \pm 12\%$) but decrease the α-helical content of the Q_{17} domain (from $45 \pm 3\%$ to $18 \pm 7\%$). Despite the decrease of the helical content, the qualitative behavior of secondary structure per residue profile is well preserved between the two forcefield. Indeed, with both forcefield, Htt17 adopts an helical conformation that is maintained throughout the first few glutamines and the helical propensity slowly decreases toward the P_{11} domain which forms a PPII-helix. In term of structure, a cluster analysis of the structure found inside the FES minima (below 8kJ/mol) reveals that the most populated cluster depicts Htt17 as an α-helix between residues 3 and 16 and the Q_17 is entirely unstructured. Htt17 non-polar residues are mostly accessible to the solvent, in agreement the simulations done with AMBER99sb*-ILDN.

In summary, the revised proline parameters do not cause drastic changes of structure for both the Htt17 and P_{11} domain. Indeed, Htt17 is still mostly helical with a very similar secondary structure profile, although the fully formed helix is less present. The P_{11} domain, for its part, maintain the PPII-helix. The main difference causes by the new forcefield is on the the Q_{17} domain which is now less structured.

Supporting Figures

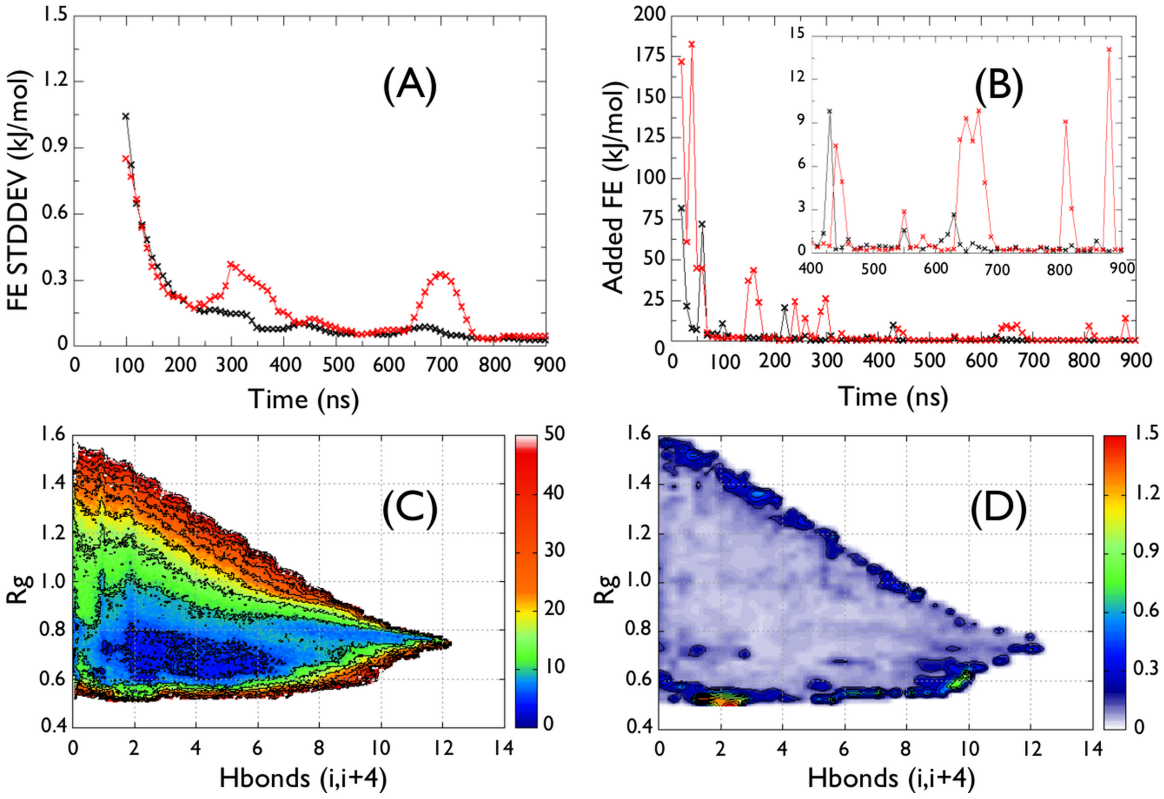

Figure S1 : Convergence assessment of the Htt17_nmr simulation. **(A)** Running average of the standard deviation of the 1D-FES (S_α in black and S_{gyr} in red) over 100 ns time-windows. **(B)** Total addition of free energy to the FES every 10 ns. **(C)** The 2D-FES (S_α ;S_{gyr}) and **(D)** its uncertainty computed on the convergence interval (400–900 ns), which is determined from the small modifications of the FES after 400 ns shown in (A) and (B). We observe that the uncertainty on the FES is mostly located to its border, while it is low ($<$ 0.5 kJ/mol) inside the basin. Energy isolines are drawn every 5 kJ/mol for **(C)** and 0.15 kJ/mol and **(D)**.

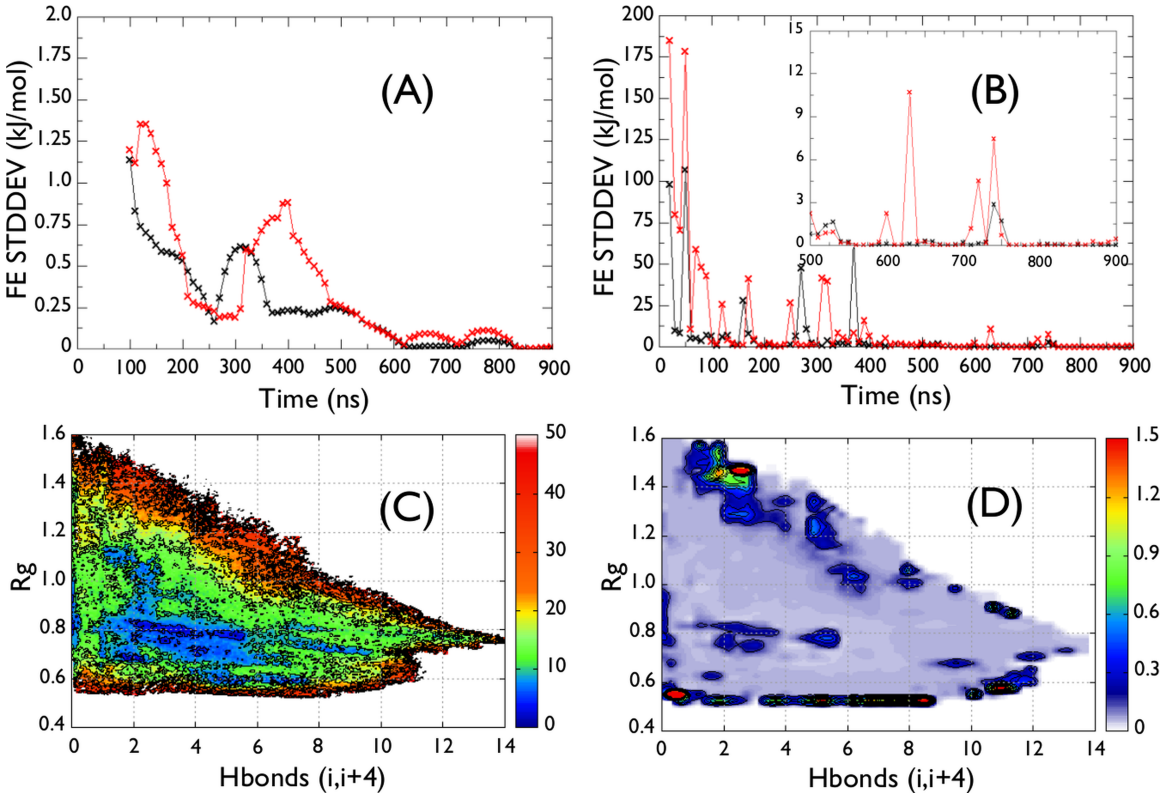

Figure S2 : Convergence assessment of the Htt17Q$_{17}$ simulation. **(A)** Running average of the standard deviation of the 1D-FES (S_α in black and S_{gyr} in red) over 100 ns time-windows. **(B)** Total addition of free energy to the FES every 10 ns. **(C)** The 2D-FES (S_α ;S_{gyr}) and **(D)** its uncertainty computed on the convergence interval (500–900 ns), which is determined from the small modifications of the FES after 500 ns shown in (A) and (B). We observe that the uncertainty on the FES is mostly located to its border, while it is low ($<$ 1.0 kJ/mol) inside the basin. Energy isolines are drawn every 5 kJ/mol for **(C)** and 0.15 kJ/mol and **(D)**.

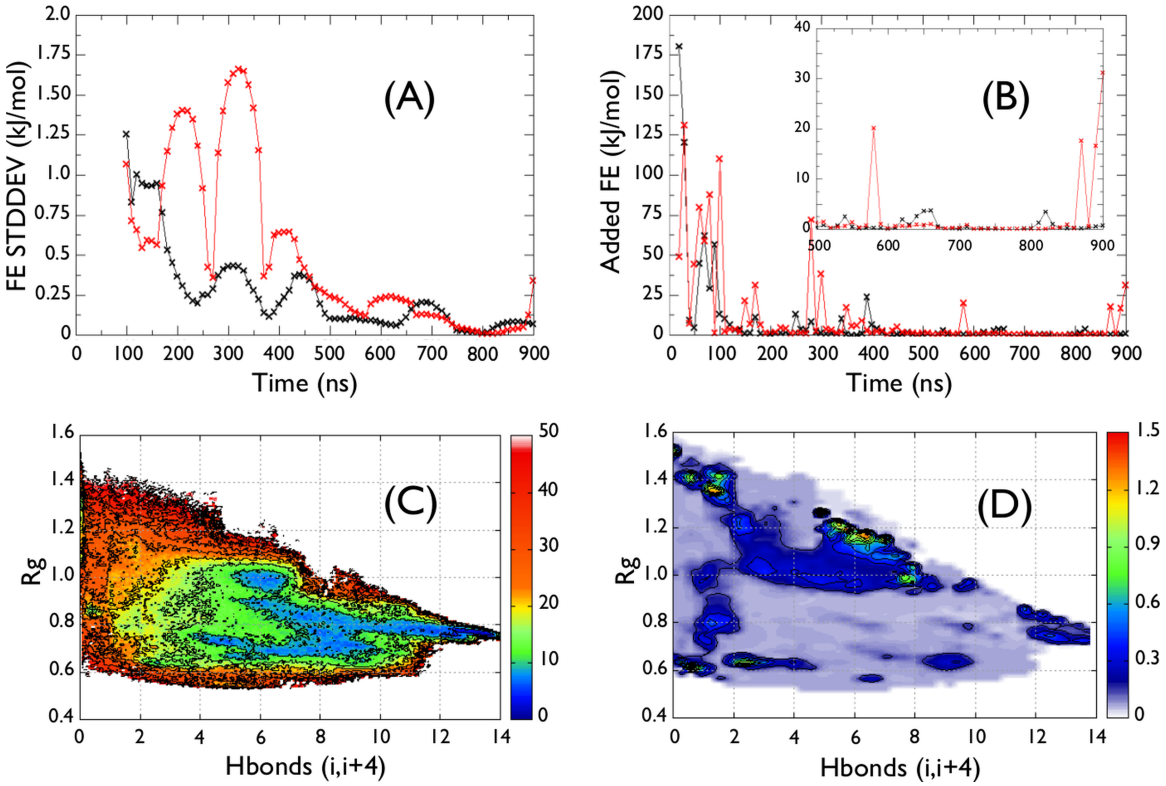

Figure S3 : Convergence assessment of the Htt17$Q_{17}P_{11}$ simulation. **(A)** Running average of the standard deviation of the 1D-FES (S_α in black and S_{gyr} in red) over 100 ns time-windows. **(B)** Total addition of free energy to the FES every 10 ns. **(C)** The 2D-FES (S_α ;S_{gyr}) and **(D)** its uncertainty computed on the convergence interval (500–900 ns), which is determined from the small modifications of the FES after 400 ns shown in (A) and (B). We observe that the uncertainty on the FES is mostly located to its border, while it is low ($<$ 1.0 kJ/mol) inside the basin. Energy isolines are drawn every 5 kJ/mol for **(C)** and 0.15 kJ/mol and **(D)**.

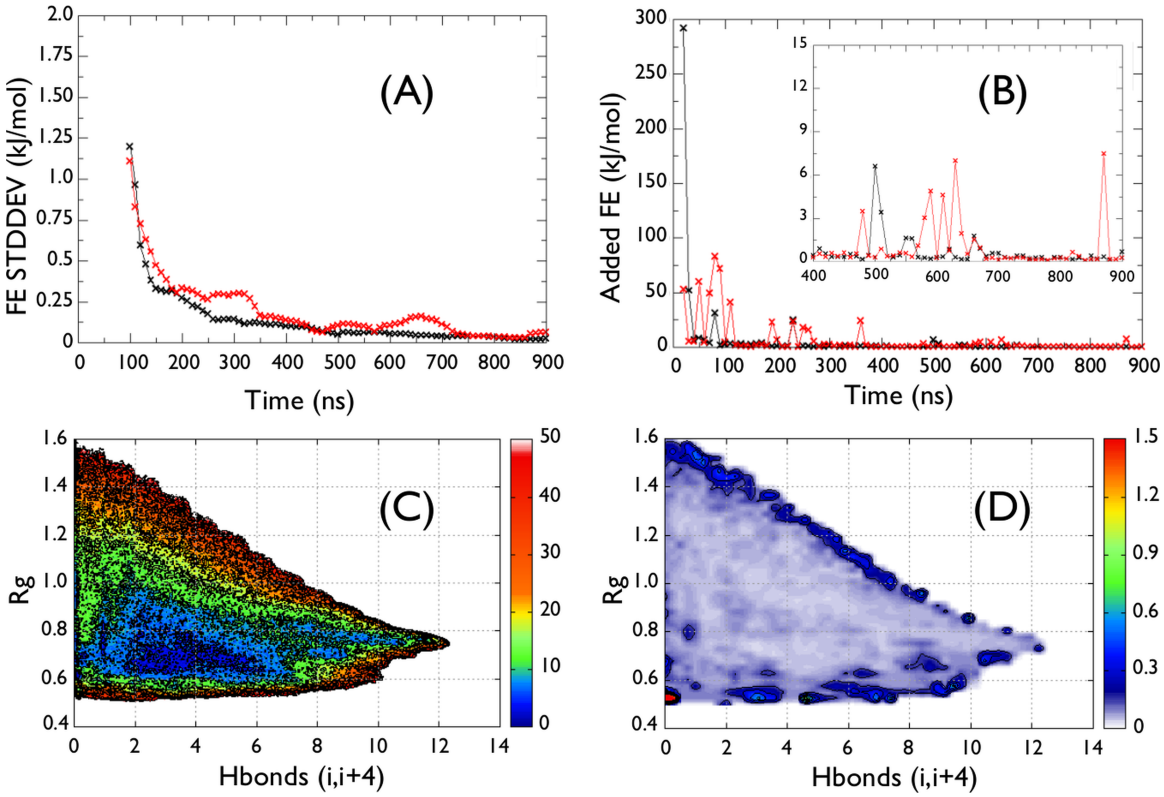

Figure S4 : Convergence assessment of the Htt17_coil simulation. **(A)** Running average of the standard deviation of the 1D-FES (S_α in black and S_{gyr} in red) over 100 ns time-windows. **(B)** Total addition of free energy to the FES every 10 ns. **(C)** The 2D-FES (S_α ;S_{gyr}) and **(D)** its uncertainty computed on the convergence interval (400–900 ns), which is determined from the small modifications of the FES after 400 ns shown in (A) and (B). We observe that the uncertainty on the FES is mostly located to its border, while it is low ($<$ 0.5 kJ/mol) inside the basin. Energy isolines are drawn every 5 kJ/mol for **(C)** and 0.15 kJ/mol and **(D)**.

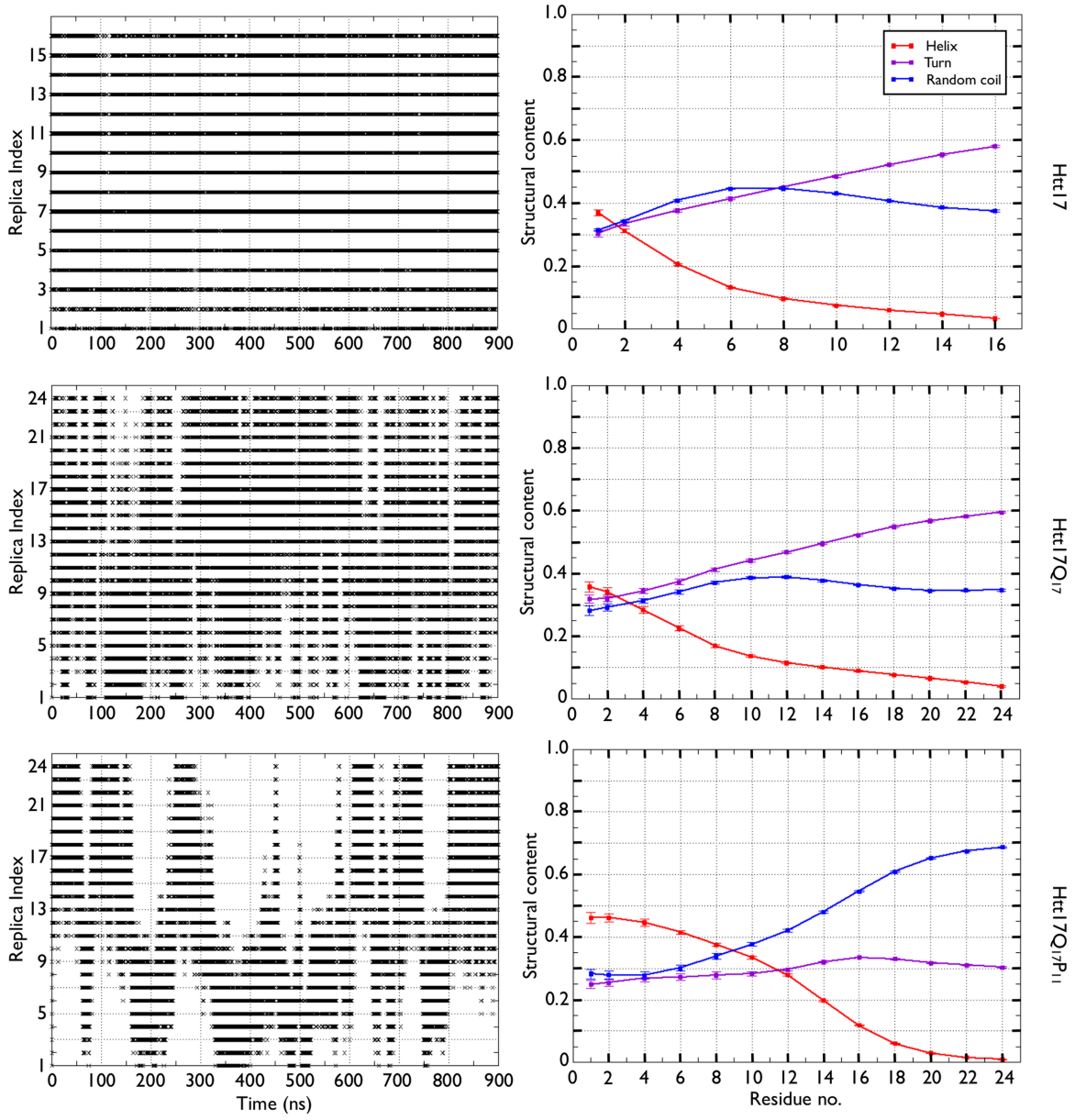

Figure S5 : Sampling assessment of the HREX simulations. The left panel shows the replica index visiting the first scale. The right panel shows the secondary structure as a function of the scaling. $Htt17Q_{17}$ and $Htt17Q_{17}P_{11}$ are respectively shown from top to bottom.

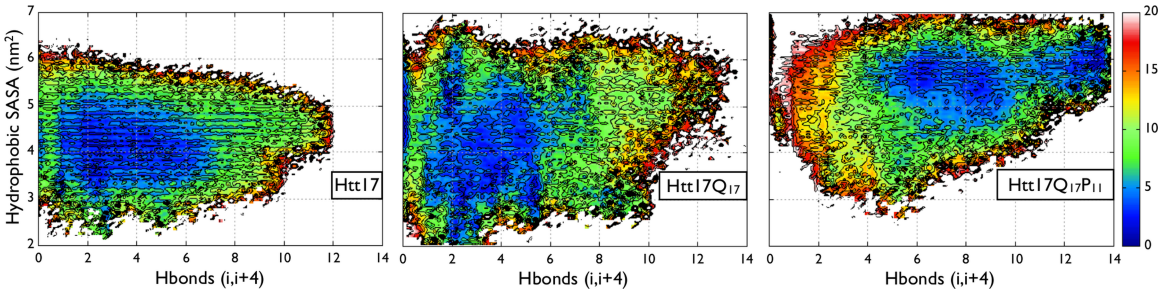

Figure S6 : The FES of the Htt17 segment as a function of the number of helical H-bonds (S_α, horizontal axis) and SASA of Htt17's non-polar residues (vertical axis) is shown for Htt17, $Htt17Q_{17}$ and $Htt17Q_{17}P_{11}$ from left to right. Energy isolines are drawn every 2 kJ/mol.

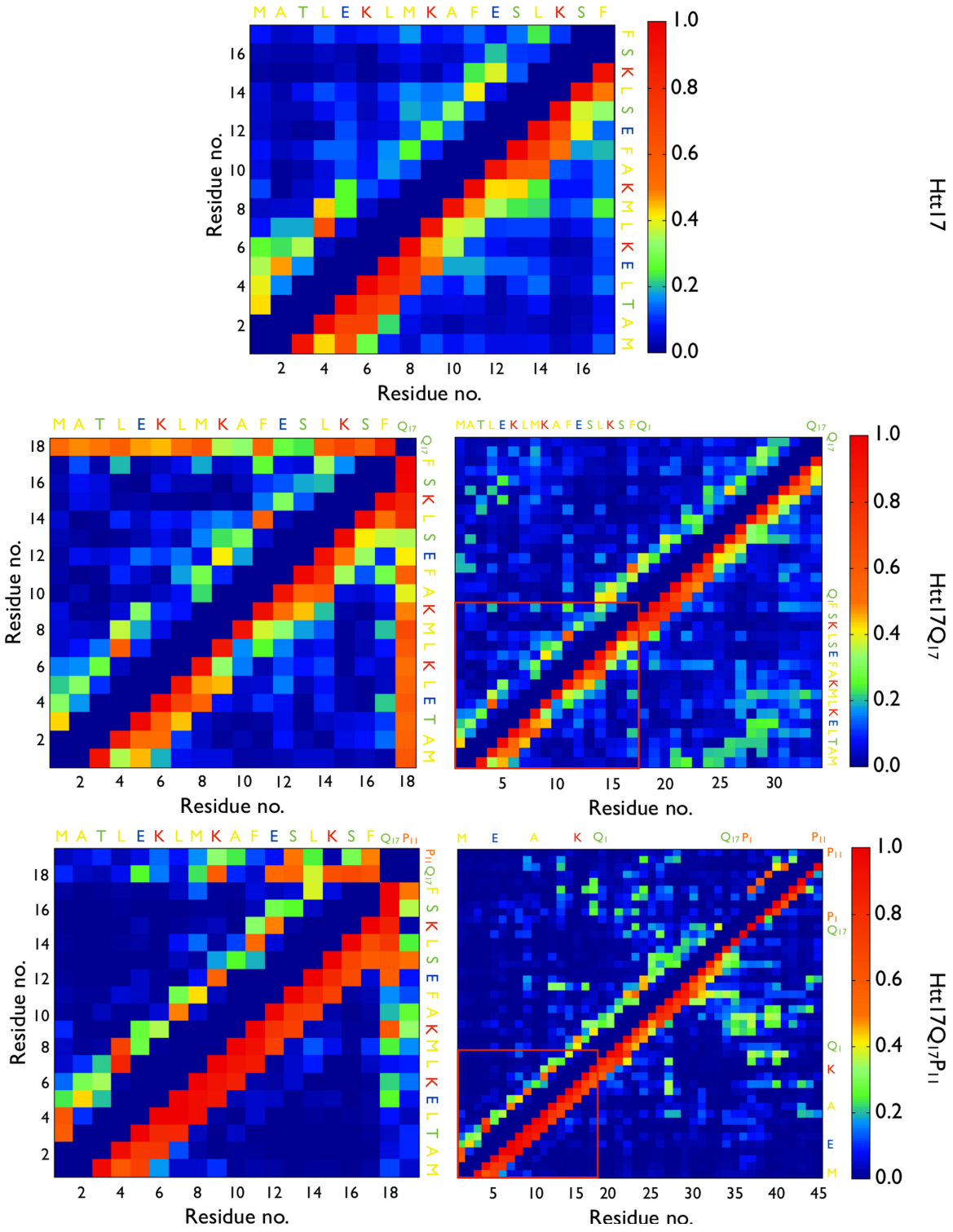

Figure S7 : Contact maps of Htt17_nmr, Htt17Q_{17} and Htt17$Q_{17}P_{11}$ are shown from top to bottom. The side-chain/side-chain and the total number of contacts are respectively displayed on the upper and lower halves of the contact maps. For Htt17Q_{17} and Htt17$Q_{17}P_{11}$, the global propensity of Q_{17}/Htt17 and P_{11}/Htt17 contacts (left column) and the per residue probability of each individual glutamines and prolines (right column) are shown. The red square indicates the Htt17/Htt17 contacts when appropriate.

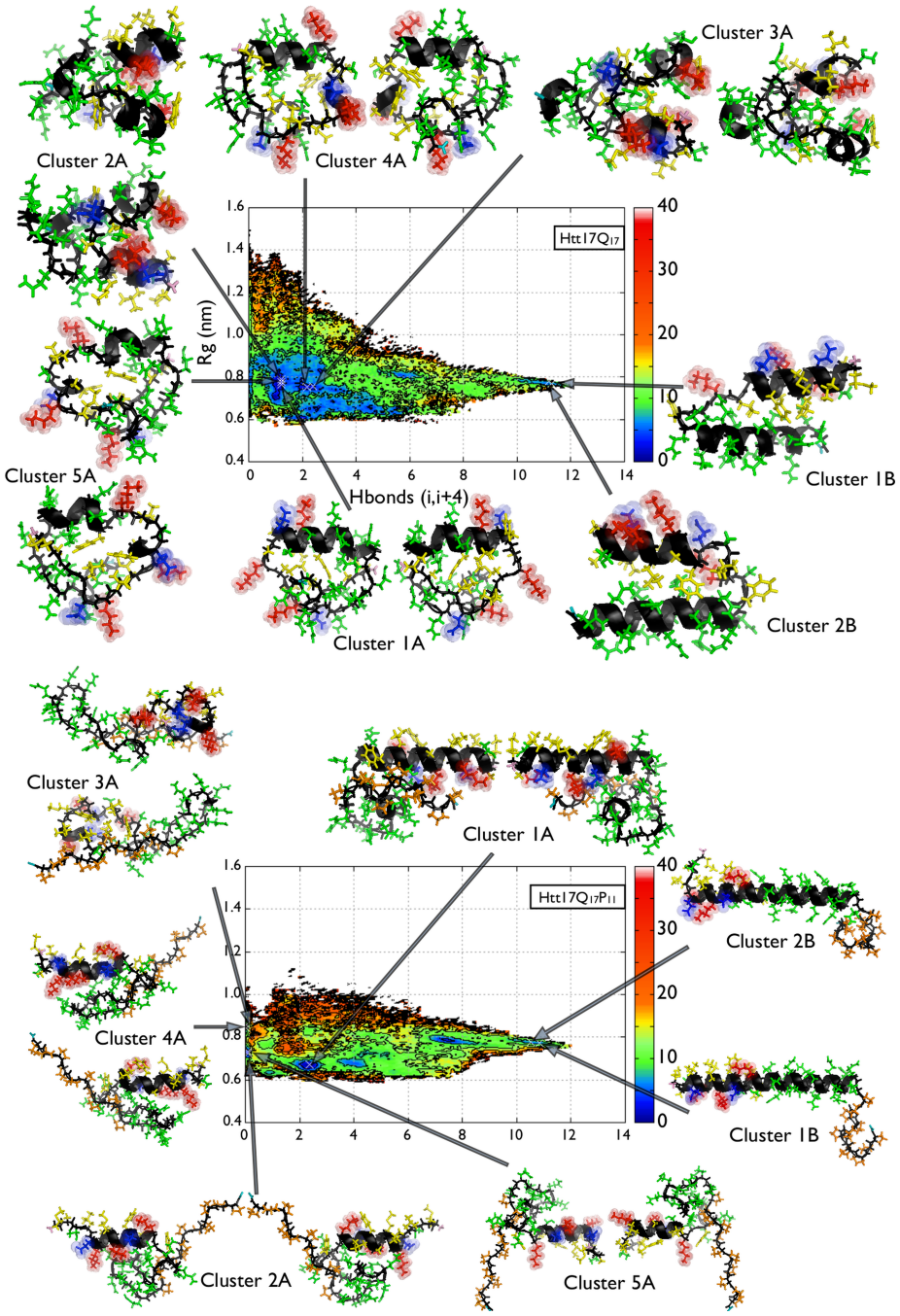

Figure S8 : The FES of the Q_{17} segment as a function of the number of helical H-bonds (S_α, horizontal axis) and gyration radius (S_{gyr}, vertical axis) is shown for Htt17Q_{17} (top) and Htt17$Q_{17}P_{11}$ (bottom). The main clusters of the conformations inside the main basin (below 4 kJ/mol) and those with more than 9.5 helical H-bonds (below 8 kJ/mol) are displayed around the FES. Energy isolines are drawn every 4 kJ/mol.

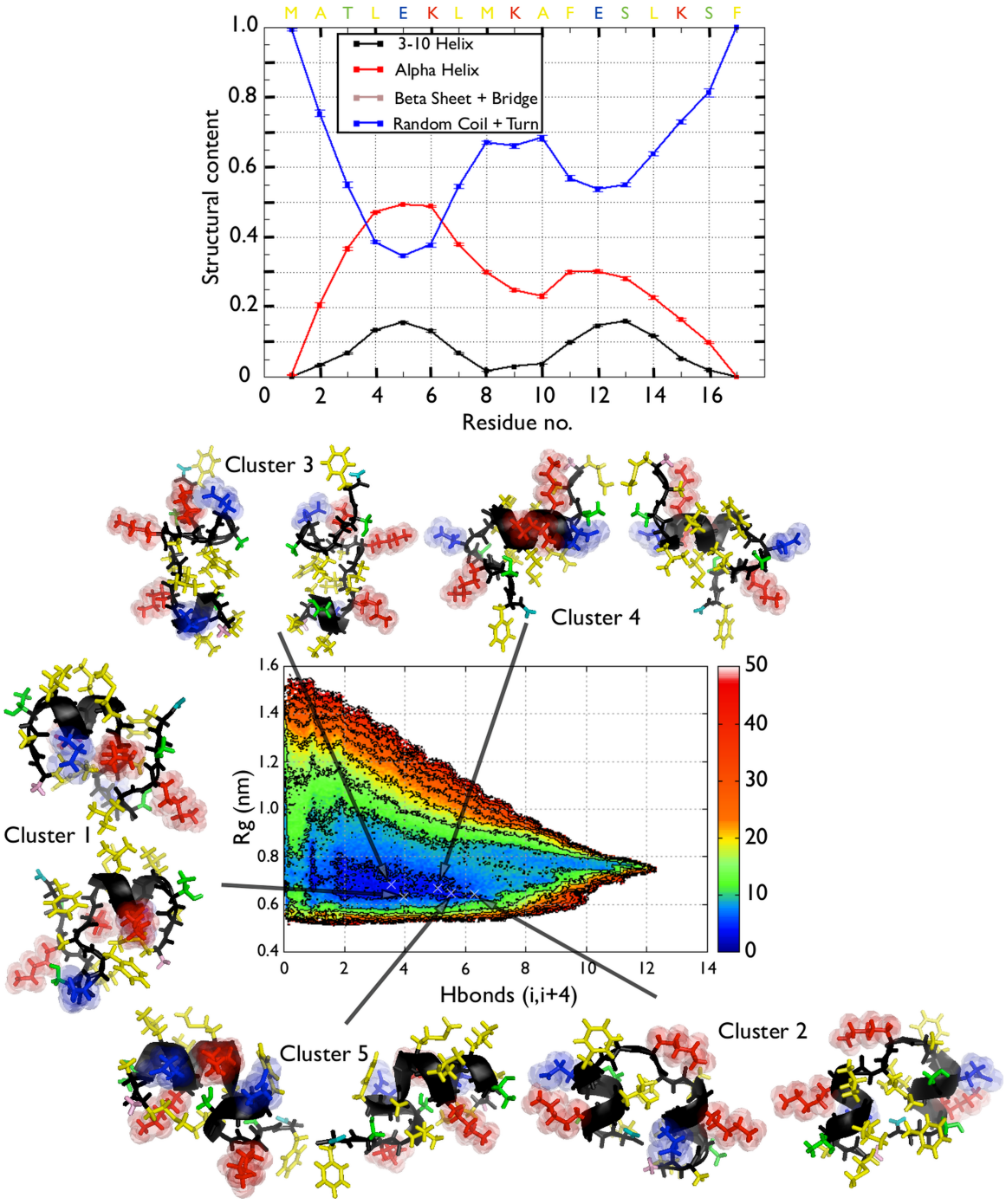

Figure S9 : The per residue secondary structure of Htt17 from the HREXMetaD simulation starting from the random structure is shown in the top panel. The probability of α-helix, 3-10 helix, β-bridge and β-strand, and all other motifs are respectively shown in red, black, brown and blue. The FES of the Htt17 segment as a function of the number of helical H-bonds (S_α, horizontal axis) and gyration radius (S_{gyr}, vertical axis) is shown in the bottom panel. Energy isolines are drawn every 5 kJ/mol. The FES is surrounded by the cluster center of the representative structures found below 5 kJ/mol.

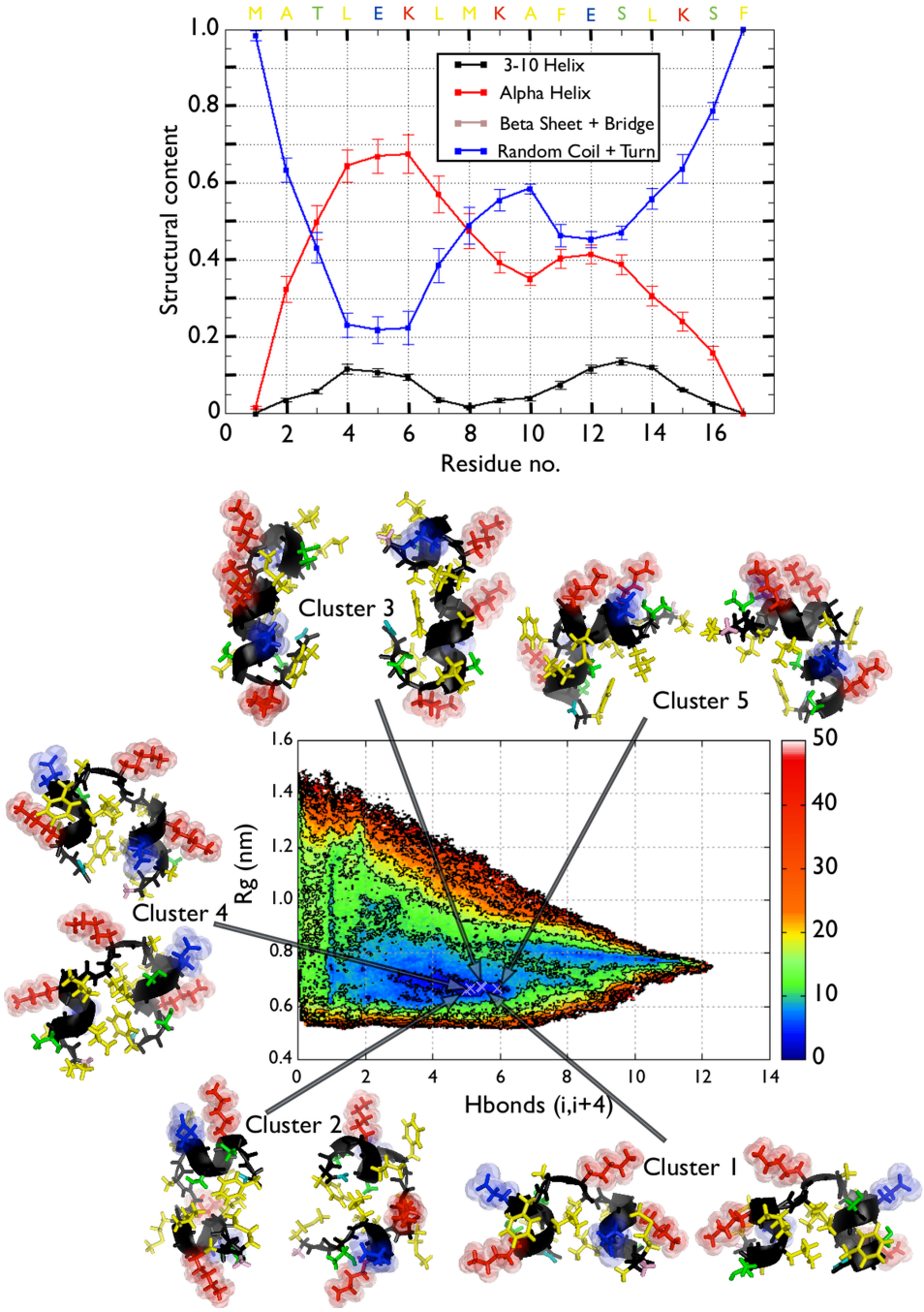

Figure S10 : The per residue secondary structure of Htt17 from the PTMetaD simulation starting from a random coil structure is shown in the top panel. The probability of α-helix, 3-10 helix, β-bridge and β-strand, and all other motifs are respectively shown in red, black, brown and blue. The FES of the Htt17 segment as a function of the number of helical H-bonds (S_α, horizontal axis) and gyration radius (S_{gyr}, vertical axis) is shown in the bottom panel. Energy isolines are drawn every 5 kJ/mol. The FES is surrounded by the cluster center of the representative structures found below 5 kJ/mol.

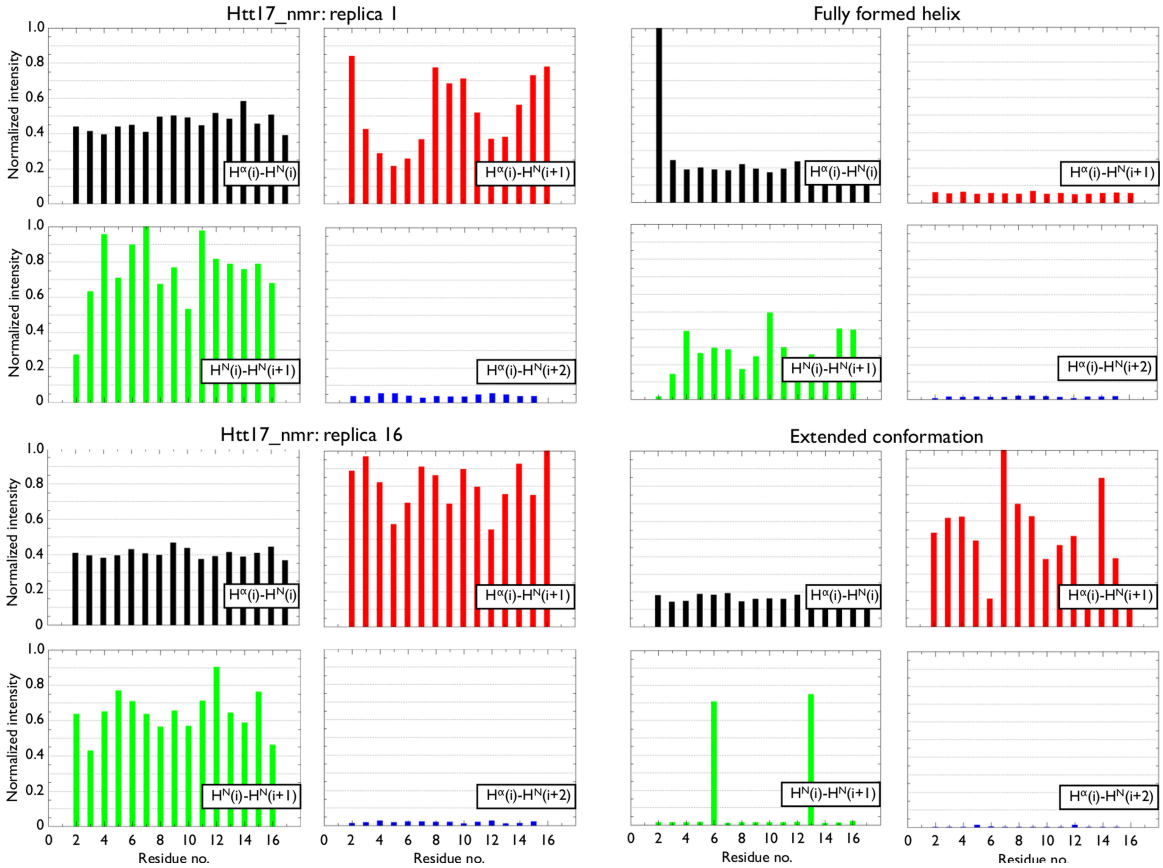

Figure S11 : The computed intensities of the interproton NOEs for all residues between the H^{α} of residue i and the H^{N} of residues i, i+1 and i+2, as well as between the H^{N} of residues i and i+1. The top left panel shows the NOEs for the analysis replica, the bottom left panel shows the NOEs for the most scaled replica, the top right panel shows the NOEs of a fully formed α-helix and the bottom right panel shows the NOEs of a fully extended conformation.

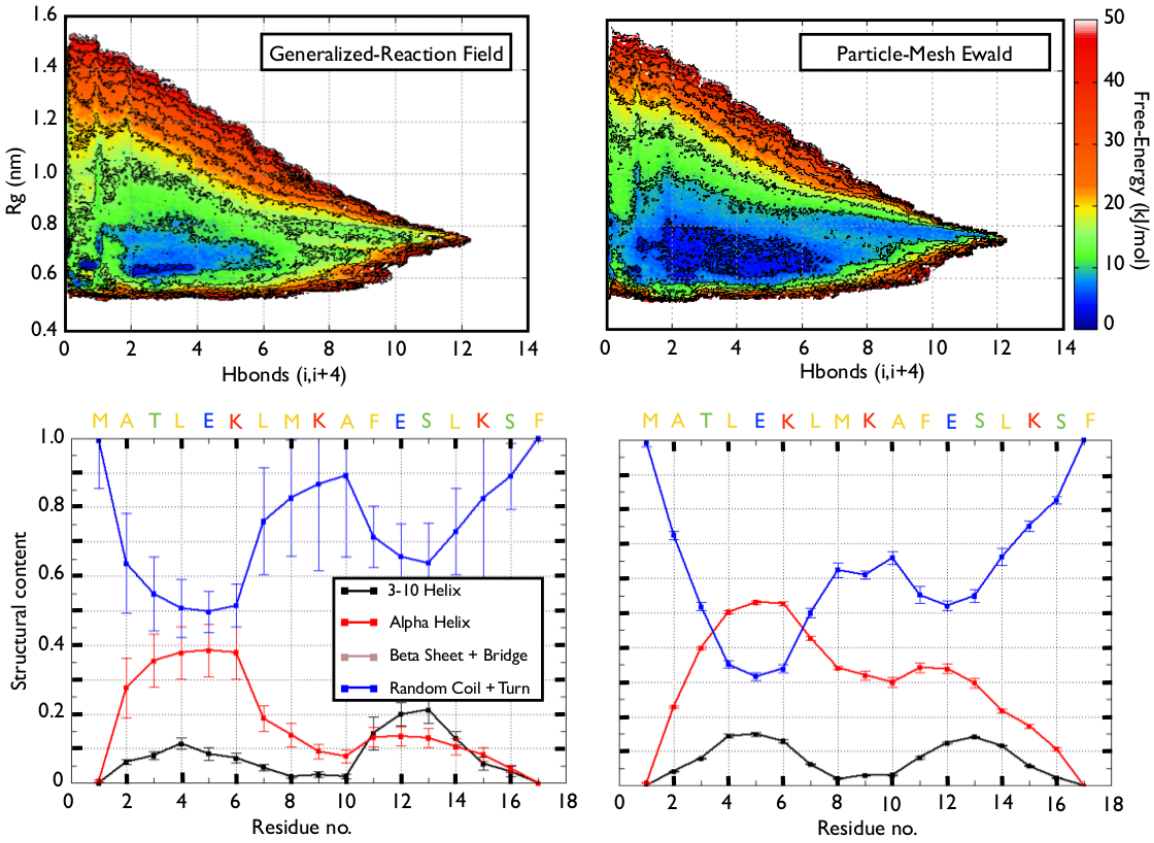

Figure S12 : The FES (shown in the top row) and secondary structure profile per residue (shown in the bottom row) of Htt17 using a Generalized-Reaction Field (shown in the left column) or the Particule-Mesh Ewald scheme (shown in the right column). The probability of α-helix, 3-10 helix, β-bridge and β-strand and turn/coil are respectively shown in red, black, brown, and blud. The vertical black dotted lines indicates respectively the end of the Htt17 segment and the end of the Q_{17} segment. The average and standard deviation are computed on the interval of convergence using 20-ns time windows.

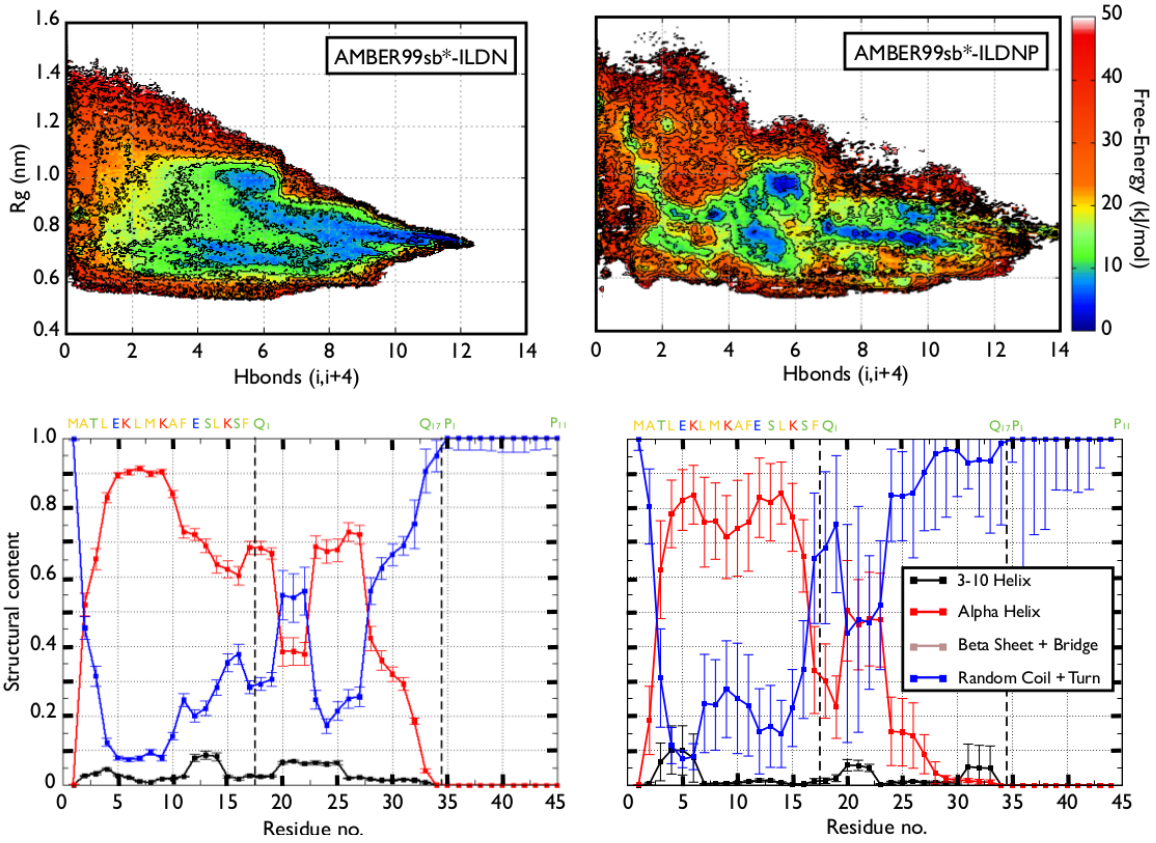

Figure S13 : The FES (shown in the top row) and secondary structure profile per residue (shown in the bottom row) of Htt17$Q_{17}P_{11}$ using the AMBER99sb*-ILDN forcefield (shown in the left column) or the AMBER99sb*-ILDNP forcefield, with the improved proline parameters (shown in the right column). The probability of α-helix, 3-10 helix, β-bridge and β-strand and turn/coil are respectively shown in red, black, brown, and blud. The vertical black dotted lines indicates respectively the end of the Htt17 segment and the end of the Q_{17} segment. The average and standard deviation are computed on the interval of convergence using 20-ns time windows.