Solute Energy Based REMD: Developments and Applications to Prion Protein Misfold Predictions

THÈSE N^{O} 4103 (2008)

PRÉSENTÉE LE 18 JUILLET 2008
À LA FACULTE SCIENCES DE BASE
LABORATOIRE DE CHIMIE ET BIOCHIMIE COMPUTATIONNELLES
PROGRAMME DOCTORAL EN CHIMIE ET GÉNIE CHIMIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Pascal BAILLOD

biologiste diplômé de l'Université de Lausanne
de nationalité suisse et originaire de Gorgier (NE)

acceptée sur proposition du jury:

Prof. L. Helm, président du jury
Prof. U. Röthlisberger, directrice de thèse
Prof. M. Dal Peraro, rapporteur
Dr O. Michielin, rapporteur
Dr J. Requena, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2008

Contents

Summary **iii**

Résumé **vii**

1 Molecular dynamics and replica-exchange molecular dynamics **1**
- 1.1 Molecular dynamics . . . 2
  - 1.1.1 Empirical force fields . . . 3
- 1.2 Replica-Exchange Molecular Dynamics . . . 4
- 1.3 References . . . 9

2 Testing the validity of solute energy based replica-exchange molecular dynamics **11**
- 2.1 Abstract . . . 12
- 2.2 Introduction . . . 12
- 2.3 Methods . . . 15
- 2.4 Results and Discussion . . . 17
  - 2.4.1 Protein contributions to the total potential energy . . . 17
  - 2.4.2 Potential energy distributions . . . 17
  - 2.4.3 Energy histogram overlaps . . . 17
  - 2.4.4 Relative average energies of conformational clusters . . . 19
- 2.5 Conclusions . . . 22
- 2.6 References . . . 25

3 Exploring protein conformational space with replica exchange molecular dynamics: possibilities & limitations **31**
- 3.1 Introduction . . . 32
- 3.2 Methods . . . 34
- 3.3 Results and Discussion . . . 36
  - 3.3.1 Extent of conformational sampling in the reference MD simulation . 36
  - 3.3.2 Extent of conformational sampling in the *REMDpe* simulation . . 37
  - 3.3.3 High temperature non-native collapsed structures . . . 40
  - 3.3.4 Too short inter-exchange times cause strong sampling biases . . . 42
- 3.4 Conclusions . . . 47
- 3.5 Appendix . . . 48

3.5.1 Average protein and total potential energy per structural group monitored in 30 successive trajectory windows 48
3.5.2 Average protein and total potential energy in function of the location on the conformational landscape 48
3.5.3 Exchange attempt statistics as a function of structural groups involved 49
3.5.4 Correlation of protein energy and radius of gyration or fraction of native contacts 49
3.5.5 Doppel protein *REMDpe* simulation: Comparable results 50
3.6 References 59

4 A replica-exchange molecular dynamics study of prion misfolding and β-rich folds 63
4.1 Introduction 64
4.2 Methods 66
4.3 Results and Discussion 67
4.4 Conclusions 76
4.5 References 77

5 What makes doppel different? 81
5.1 Introduction 82
5.2 Methods 85
5.3 Results and Discussion 86
5.3.1 Reference MD simulations 86
5.3.2 *REMDpe* simulations 87
5.3.3 Secondary structure propensities 92
5.3.4 β-rich folds 92
5.4 Conclusions 99
5.5 Appendix: Sequence alignments 100
5.5.1 Structure based sequence alignment of PrP of different species . . . 100
5.5.2 Structure based sequence alignment of mouse PrP and Dpl 101
5.6 References 102

6 Outlook 109
6.1 Protein energy as a function of α and β content 110
6.2 Dissecting the mechanisms of protein unfolding 110
6.3 Electrostatic properties and aggregation propensities of β-rich folds 110

Acknowledgments 113

Curriculum Vitae 117

Summary

Molecular dynamics (MD) simulations have increasingly contributed to the understanding of biomolecular processes, allowing for predictions of thermodynamic and structural properties. Unfortunately, the holy grail of protein structure prediction was soon found to be severely hampered by the very rugged free energy surface of proteins, with small relative free energies separating native, folded protein conformations from unfolded states. These multiple minima frequently trap present-day protein MD simulations permanently. In order to allow the simulation to escape minima and explore wider portions of conformational space, enhanced sampling techniques were developed. One of the most popular ones, replica exchange molecular dynamics (REMD), is based on multiple parallel MD simulations that are performed with replicas of a system at increasing temperatures T_1, T_2, etc. Periodic Monte Carlo exchange moves are attempted, aiming to allow conformations to exchange temperature ensembles with a probability that depends on their potential energy and temperature difference. Thus, conformations are simulated at all temperatures and escape local minima with the kinetic energy provided at higher temperatures, while Boltzmann distributions are generated at all temperatures.

REMD has been successful in $ab-initio$ folding of a variety of small peptides and proteins (up to 20-30 residues). However, with larger proteins, the overlap of potential energy distributions diminishes, since the potential energy and its fluctuation scale with fk_BT, respectively with $\sqrt{f}k_BT$, where f is the number of degrees of freedom of the system, k_B Boltzmann's constant and T the temperature. Consequently, the related Monte Carlo exchange probability and number of exchanges in a simulation are also diminished. This is generally compensated by choosing smaller temperature intervals between replicas and thereby increasing the number of necessary replicas (as well as the computational cost of the simulation) to cover a given temperature range. An additional problem relates to explicit solvent simulations, in which solvent to solvent interactions account for the largest part of the total potential energy. Consequently, explicit solvent REMD simulations almost exclusively sample solvent degrees of freedom. These two limitations have lead to the development of REMD protocols for large explicit solvent systems that are based on exchange probabilities computed with subsystem (e.g. protein only) potential energy functions, allowing for a targeted sampling of protein degrees of freedom and a reduction of the computational effort. In this thesis, this approximation is tested by implementing its simplest variation that entirely neglects solvent-solvent interaction as a new REMD protocol termed $REMDpe$ (Chapter 2). Possible REMD limitations for large explicit solvent systems are tested (Chapter 3) with $REMDpe$, which is further applied to perform a thorough and comparative investigation of prion (Chapter 4) and doppel (Chapter 5) protein misfolding.

In Chapter 2, the practical validity of the *REMDpe* approximation is assessed with simulations of the prion protein, a system that is too large (i.e. 103 residues) to allow for efficient simulations using traditional REMD over the necessary temperature range. A first validation consists in testing whether protein and total potential energy distributions are consistent with their analogs from straightforward reference MD simulations. Second, the overlap pattern of the total potential energy distributions are characterized at different temperatures and show that the exchanges in the *REMDpe* simulations are performed according to a Boltzmann weight. Native structures are found to have the lowest protein and total potential energies, as compared to higher energies found for various unfolded structures. Although no obvious bias is detected in the three validations, the conformational landscapes of the *REMDpe* simulation at low temperatures progressively shift to non-native regions of the free energy surface. In Chapter 3, this phenomenon is quantified, and its origin identified in insufficient low temperature residence times required for refolding native-like structures. REMD is based on the assumption that systems have to be decorrelated between exchange attempts. Increasing inter-exchange times accordingly would allow for decorrelation and sufficient low temperature residence times but is practically impossible to achieve for large protein simulations, highlighting a major limitation and possible source of bias for present-day REMD simulations.

We have chosen the prion as a test case because of its link to transmissible spongiform encephalopathies. Diseases of this category are believed to be caused by a rare prion protein (PrP) misfold leading from the cellular, monomeric, soluble, α-helical PrP^C isoform to a pathogenic, aggregated, insoluble, β-rich PrP^{Sc} isoform of unknown structure. Gaining experimental knowledge of the PrP^{Sc} structure has remained elusive, and aroused interest in predictions supplied by computer simulations. REMD provides a powerful tool allowing to explore a diversity of misfolds and select stable ones that accumulate at lower temperatures. In Chapter 4, we describe a PrP *REMDpe* simulation in which rare new β-strands are formed and arrange into a multitude of different β-sheets, reproducing the α-helix → β-sheet conversion observed with circular dichroism spectra. The α-helical and β-sheet propensities along the sequence can thus be computed. We develop and apply the *β contact map clustering* (*bcmc*) protocol to identify the most frequent β-sheet pattern defining β-rich folds. 10 new β-rich folds are found and compared to recent experimental data characterizing PrP^{Sc}, providing atomistically detailed models for putative monomeric precursors of PrP^{Sc} or β-oligomeric conformations.

In Chapter 5, an analogous simulation is performed with doppel, a structural homolog of prion (with an identical three α-helix, two β-strand fold) originating from the same gene family, but characterized by a different sequence (only 25% sequence homology), expression pattern and physiological function. Unrelated to amyloid neurodegenerative diseases, doppel supplies the perfect test system to investigate the misfolding of a non-amyloidogenic protein. Prion and doppel misfolding are compared in their monomeric form in the quest to identify prion-specific features that might reveal the mechanism of conversion to PrP^{Sc}. In agreement with experiments, we find a lower thermal stability for doppel. Surprisingly, we also observe β-rich forms for doppel. However, the β-rich

folds of the two proteins are very different. Moreover, a major difference is found in the free energy barriers leading from the native structure to such conformations as well as to non-native conformations in general: These barriers are low for prion and can already be crossed at 300K, while for doppel they are at least 3 times higher. This difference suggests an intrinsic misfolding and β-enrichment propensity for the monomeric form of prion as compared to doppel.

Keywords: Replica-exchange molecular dynamics (REMD), solute energy based REMD, protein potential energy, Boltzmann distribution, non-native structure, conformational landscape, free energy surface, structural clustering, prion, β-rich folds, scrapie, β-oligomer, doppel, transmissible spongiform encephalopathies.

Résumé

Les simulations de dynamique moléculaire (MD) ont contribué de façon croissante à comprendre les processus biomoléculaires, permettant de prédire des propriétés thermodynamiques et structurales. Malheureusement, la frustration des surfaces d'énergie libre des protéines, avec de faibles différences d'énergie libre séparant les conformations natives repliées des conformations dénaturées, a vite déçu les espoirs suscités par cette technique dans le domaine de la prévision de structures de protéines. La multitude de minima énergétiques de ces surfaces emprisonnent fréquemment les simulations MD contemporaines de protéines. Afin de permettre aux simulations d'échapper à ces minima et explorer des portions plus larges du paysage conformationnel, des techniques d'échantillonnage accéléré ont été développées. L'une des plus courantes, *replica exchange molecular dynamics* (REMD), est basée sur plusieurs simulations MD parallèles effectuées avec des copies (*replicas*) à des températures croissantes T_1, T_2, etc. A une fréquence régulière, des échanges de conformations sont tentés selon la méthode de Monte Carlo, avec un probabilité qui dépend de leur différences d'énergie potentielle et de température. Ainsi, les conformations sont simulées à toutes les températures et peuvent échapper des minima locaux grâce à l'énergie cinétique fournie à haute température, alors que des distributions canoniques sont générées à toutes les températures.

La méthode REMD a pu être utilisée pour prédire des structures tri-dimensionnelles à partir de la séquence de peptides et petites protéines (jusqu'à 20-30 aa). Toutefois, pour des protéines plus grandes, les recoupement des distributions d'énergie potentielle diminuent: Pour un système à f degrés de liberté, l'énergie potentielle est proportionnelle fk_BT alors que ses fluctuations sont proportionnelles à $\sqrt{f}k_BT$, où k_B est la constante de Boltzmann et T la température. Par conséquent, la probabilité d'échange de Monte Carlo et le nombre d'échanges dans une simulation diminuent également. Ceci est généralement compensé par un choix de températures moins espacées entre les copies du système, ce qui accroît le nombre de copies requises (et la puissance de calcul nécessaire) pour couvrir un intervalle de températures donné. Un problème additionnel survient avec les simulations faites en solvent explicite, dans lesquelles les interactions solvent-solvent constituent de loin la contribution la plus importante à l'énergie potentielle totale. Cette contribution majoritaire fait que les simulations REMD effectués en solvent explicite échantillonnent avant tout les degrés de liberté du solvent. Ces deux limitations ont conduit au développement de protocoles REMD pour simulations de grands systèmes en solvent explicite. Ces protocoles sont basés sur des probabilités d'échange qui sont calculés avec l'énergie potentielle d'un sous-système (par exemple, la protéine seule), permettant de concentrer l'échantillonnage sur les degrés de liberté de la protéine et de réduire la puissance de calcul requise. Dans

la présente thèse, cette approximation est étudiée dans sa variante la plus simple, qui néglige entièrement les interactions solvent-solvent. Cette variante est implémentée dans un nouveau protocole REMD, dénommée *REMDpe* (Chapitre 2). Les améliorations et limitations apportées par la méthode REMD à l'étude de grands systèmes en solvent explicite sont évaluées avec le protocole *REMDpe* (Chapitre 3). Ce protocole est ensuite utilisé pour effectuer une analyse comparative de la dénaturation de deux protéines structurellement homologues, prion (Chapitre 4) et doppel (Chapitre 5).

La validité pratique de l'approximation *REMDpe* est évaluée au Chapitre 2 avec une simulation de la protéine prion, un système qui est trop gros (103 résidus) pour être simulé de manière efficace avec la méthode REMD traditionnelle (sans approximation) sur le même intervalle de températures. Une première validation consiste à tester si les distributions d'énergies de protéine et totales sont équivalentes à leurs analogues de simulations MD de référence. Deuxièmement, les chevauchements d'énergie potentielle totale sont caractérisés pour différentes températures et montrent que les échanges sont effectués selon une pondération canonique. Les énergies potentielles de protéine et totales les plus faibles sont trouvées pour les structures natives, par rapport à des énergies plus élevées trouvées pour diverses structures dénaturées. Bien qu'aucun artéfact évident ne soit identifié pour l'approximation *REMDpe* par ces trois validations, le paysage conformationnel des basses températures converge progressivement vers des régions non-natives de la surface d'énergie libre. Au Chapitre 3, ce phénomène est quantifié, et son origine trouvée dans un temps de résidence à basse température qui s'avère insuffisant pour replier les structures dénaturées en une conformation native. La méthode REMD est fondée sur l'hypothèse que les structures sont décorrélées entre des tentatives d'échange de température. Un accroissement du temps de simulation entre ces échanges permettrait cette décorrélation, ainsi que des temps de résidence suffisants à basse température, mais est pratiquement impossible à réaliser pour des simulations de grandes protéines, mettant en évidence un bias majeur des simulations REMD courantes.

Nous avons choisi le prion comme système d'essai à cause de son lien avec les encéphalites spongiformes transmissibles. Les causes de ces maladies demeurent mal comprises, et l'hypothèse la plus courante met en scène une conversion rare du prion (PrP), menant à de la forme cellulaire normale PrP^{C} monomérique, soluble et riche en hélices α à une forme pathogène PrP^{Sc}, oligomérique, insoluble, riche en feuillets β et de structure inconnue. Les tentatives de résolution structurales de PrP^{Sc} sont restées infructueuses à ce jour, reportant les espoirs dans les prévisions structurales informatiques. La méthode REMD fournit un puissant outil permettant d'explorer diverses formes dénaturées et de sélectionner celles qui sont stables et qui s'accumulent à basse température. Au Chapitre 4, nous décrivons une simulation *REMDpe* de prion dans laquelle de rares nouveaux feuillets β sont formés et s'agencent en une multitude d'arrangements, reproduisant la conversion hélice $\alpha \rightarrow$ feuillet β observée dans les expériences de dichroïsme circulaire. Ainsi, nous pouvons calculer les propensions de formation d'hélices α et de feuillets β en fonction de la séquence. Nous développons et appliquons le protocole β *contact map clustering* (*bcmc*), permettant d'identifier les feuillets β les plus fréquents, utilisés pour définir des repliements β (définis ici comme un ensemble de structures riches en feuillets β contenant

le même agencement tri-dimensionnel de feuillets). 10 nouveaux repliements β sont ainsi trouvés et comparés à de récentes données expérimentales caractérisant PrP^{Sc}, fournissant des modèles de résolution atomique pour d'hypothétiques précurseurs monomériques de PrP^{Sc} ou de β-oligomères.

Au Chapitre 5, une simulation analogue est effectuée avec doppel, un homologue structural du prion (avec un repliement identique, caractérisé par 3 hélices α et 2 feuillets β) provenant de la même famille de gènes mais différant par la séquence (seulement homologue à 25%), le profile d'expression génétique et la fonction physiologique. Sans lien aux maladies neurodégénératives amyloïdes, doppel fournit le parfait système d'essai pour simuler une dénaturation non amyloïde. Les dénaturations simulées de prion et de doppel sont comparées dans la quête de caractéristiques propres au prion susceptibles de révéler le mécanisme de conversion de PrP^{C} à PrP^{Sc}. En accord avec les expériences, la stabilité thermique de doppel est plus faible que celle du prion. Étonnamment, des structures enrichies en feuillets β sont aussi trouvées pour doppel. Toutefois, une différence majeure est trouvée dans les barrières d'énergie libre menant à ces conformations ainsi qu'à des conformations non-natives: Ces barrières sont basses pour le prion et peuvent déjà être franchies dans des simulations standard à 300K, alors qu'elles sont au moins 3 fois plus élevées pour doppel. Cette différence suggère que la forme monomérique du prion possède, contrairement à doppel, une propension naturelle à la dénaturation et aux conversions menant à des structures riches en feuillets β.

Mots-clé: *Replica-exchange molecular dynamics* (REMD), REMD basé sur l'énergie du soluté, énergie potentielle de protéine, distribution canonique, structure non-native, paysage conformationnel, surface d'énergie libre, regroupement structural, prion, repliements riches en feuillets β, *scrapie*, β-oligomères, doppel, encéphalite spongiforme transmissible.

Chapter 1

Molecular dynamics and replica-exchange molecular dynamics

1.1 Molecular dynamics

The behavior and energetics of molecules are fundamentally quantum mechanical. However, quantum mechanical models are computationally demanding and severely limit the system size and time scale of the simulation. Approximating atomic and molecular interactions with classical mechanics, based on the classical motion of nuclei (treated as point charges in a framework that totally neglects electrons), allows to increase simulation size and time scale limits, while preserving accuracy for the description of a number of properties that do not depend on the electronic distribution in a molecule. Molecular dynamics (MD) simulations are becoming an increasingly powerful tool for the prediction of molecular properties, including protein structure and dynamics [1, 2, 3]. MD algorithms iteratively compute the solution of the classical equations of motion for a group of atoms. Positions of individual atoms are updated after every iteration, or MD step Δt, and the ensemble of MD steps describes a trajectory for all the atoms, as well as for the system they form. Typically, Δt is determined by the fastest motions in the system, which are bond vibrations. One has to use a smaller Δt to ensure the accuracy of the numerical integration. Typical choices are Δt $\sim$ 1-2 fs for methods based on empirical force fields. The force $\mathbf{f}_i$ acting on atom i with position r_i and mass m_i is given by Newton's law:

$$\mathbf{f}_i(t) = -\frac{\partial}{\partial \mathbf{r}_i} E(\mathbf{r}(t)) = m_i \frac{d^2\mathbf{r}_i}{dt^2} \tag{1.1}$$

where E is the potential energy and $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_n)$ is the collective array of all particle i positions. The force f_i results from the sum of interactions with all the other atoms of the system according to E. The associated equations of motion are solved with a numerical integration scheme, yielding new atomic positions at each time step. One of the most stable and common algorithms is the velocity-Verlet algorithm, which makes use of a Taylor expansion truncated beyond the quadratic terms for the coordinates [4].

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \frac{\mathbf{f}_i(t)}{2m_i}\Delta t^2 \tag{1.2}$$

The update of the velocity $\mathbf{v}_i$ of particle i is given by:

$$\mathbf{v}_i(t+\Delta t) = \mathbf{v}_i(t) + \frac{\mathbf{f}_i(t) + \mathbf{f}_i(t+\Delta t)}{2m}\Delta t \tag{1.3}$$

MD trajectories are generated and analyzed in order to study stable conformations of a system, as well as transitions between these conformations. Using a thermostat and/or barostat, a canonical or isobaric-isothermal ensemble can be sampled by MD simulations. The ergodic hypothesis states that a system of particles sampling a given portion of phase space for a sufficiently long time will access every single available microstate. This allows to relate properties derived from the ensemble of microscopic states to observable macroscopic properties A_{obs}. If the ensemble is large enough (i.e. as obtained with a sufficiently long MD trajectory $\Gamma(t)$), the latter is equal to the former:

$$A_{obs} = \langle A \rangle_{ens} = \lim_{T \to \infty} \frac{1}{T} \int_0^T A\left[\Gamma(t)\right] dt \tag{1.4}$$

1.1.1 Empirical force fields

Different models and levels of theory can be applied to an MD scheme in order to describe E and compute the corresponding forces in (1.1). Chemistry deals with the formation and destruction of chemical bonds, which depend on the interactions of electrons. Such phenomena can only be described with quantum mechanical models that allow for very accurate descriptions of a system. However, the explicit treatment of electrons requires an intense computational effort that sets limits to the system sizes and time scales that can be modelled. Molecular mechanics (MM) methods renounce to an explicit treatment of electrons, and model the potential energy and derived forces of the system as a function of the nuclear coordinates only, with predefined atomic charges located on the nuclei in predefined molecular topologies (describing covalent bonds, which will never change during the simulation). With this approximation, system sizes and time scales can be significantly increased, allowing to model many of the conformational changes that underly biochemical processes. The typical force field, or atomistic potential energy function ($E(r)$) used in most biomolecular simulations is of the form:

$$\begin{aligned} E(r) &= \sum_{\text{bonds}} K_r (r - r_{eq})^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} K_\phi [1 + \cos(n\phi - \gamma)] \\ &\quad + \sum_i \sum_{\substack{j<i \\ j \notin \text{excl}(i)}} \left[\frac{C_{ij}^{(12)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^{6}} + \frac{1}{4\pi\epsilon_r\epsilon_0} \frac{q_i q_j}{r_{ij}} \right] \end{aligned} \tag{1.5}$$

Bond lengths r and angles θ are treated harmonically with reference points r_{eq} and θ_{eq} and harmonic constants K_r and K_θ, respectively. Dihedral angles are restrained with a periodic term bearing n minima, and scaled by the constant K_ϕ. Van der Waals interactions are modelled by a Lennard-Jones (LJ) 12-6 function with atom-pair specific parameters $C^{(6)}$ and $C^{(12)}$. The last term of (1.5) represents the electrostatic interaction between two atoms i and j with charges q_i and q_j, separated by a distance r_{ij}. The relative dielectric permittivity, ϵ_r is typically set to 1 in explicit solvent simulations.

The concept of force field refers to the set of values obtained for the parameters of (1.5) (atom and bond specific r_{eq}, θ_{eq}, K_r, K_θ, $C_{ij}^{(12)}$, $C_{ij}^{(6)}$, charges, etc.), derived from quantum mechanical calculations of small molecules (i.e. AMBER, [5]), experiments (i.e. free enthalpies of solvation measured with small organic molecules for GROMOS [6]) or a mix of the two (i.e. OPLS [7], for which covalent bonded force-field contributions are taken from AMBER). Thus, force fields generally all adopt a form that is similar to (1.5) but differ in the parameters. Although most force-fields are not directly parameterized for the description of experimental protein structures, many perform this task surprisingly well.

Lennard-Jones and Coulomb interactions between atoms that are directly bonded to each other via one or two bonded atoms are mainly driven by quantum effects that cannot be approximated as in the rightmost term of (1.5). The corresponding atom pair interactions are therefore excluded from this term and accounted for in the corresponding bond, angle and dihedral terms.

The number of long range or nonbonded pairwise interactions in (1.5) increases as N^2, where N is the number of interaction sites of the system. These interactions represent the major part of the computational burden. Therefore, their computation for a given atom are generally reduced to the interactions with atoms that are within a given cutoff distance. To further reduce the computational effort, a twin range cutoff scheme can be applied, in which interactions within a short range cutoff are computed every time step, while the remaining interactions up to a long range cutoff are only updated every few steps. In order to account for the non negligible effect of electrostatic interactions beyond the long range cutoff, the most popular approaches are the Particle Mesh Ewald [8] and the Generalized Reaction Field [9] methods. The first approach makes use of the exact periodicity of the system and includes all electrostatic interactions using lattice sum techniques based on the Ewald summation. The second approach uses a reaction field correction which accounts for the mean polarization effect outside the cutoff region.

1.2 Replica-Exchange Molecular Dynamics

Unfortunately, when applied to the protein folding problem, classical molecular dynamics simulations encounter three major limitations: (i) MD simulation times that can be performed on today's computers correspond to a physical time that is in the order of a few hundreds of nanoseconds only [10, 1]. This is far from sufficient to describe folding processes taking place on time scales that have been experimentally determined to be of the order of microseconds to milliseconds. (ii) The free energy difference between folded and misfolded, denatured or random-coil conformations is often very small and might be beyond the accuracy of a force field. (iii) Sampling of phase space by MD is often insufficient to probe the immense conformational space of proteins. This conformational freedom implies a very rugged free energy surface, with many local minima in which MD simulations can get trapped.

To alleviate this problem, different enhanced sampling techniques have been developed [11, 12, 13, 14, 15, 16]. Their common goal is to enhance barrier crossing rates, enabling the simulation to explore a broader portion of phase space and to escape local minima. This is equivalent to increasing the physical time scale of the simulation, thus accessing time scales that are otherwise prohibitively expensive in terms of computer time. Some of these techniques aim at the enhanced sampling of a specific portion of the total system (e.g. Canonical Adiabatic Free Energy Sampling (CAFES) [11]), while others aim towards the enhanced sampling of the entire system along one or more reaction coordinates (e.g. metadynamics [12]). More general approaches include simulated annealing [13], stochastic tunneling [14, 17], conformational flooding [15] and generalized ensemble methods [18, 16]. The latter were devised to extend sampling by performing Monte Carlo moves into other thermodynamic states (differing by temperature or other parameters) by performing a random walk in potential energy space that allows to escape energetic minima. A number of powerful simulation algorithms have been developed from this basic idea, such as multi canonical simulation, simulated tempering and replica exchange (reviews can be found in ref. [18, 16]).

In replica exchange, multiple simulations are performed in different thermodynamic states, defining an extended system as generalized ensemble. Monte Carlo exchange moves consist in exchanging two systems a and b (replicas), simulated at thermodynamic conditions A and B. Accepted moves result in a new extended system in which system a and b are now at thermodynamic conditions B and A. This algorithm was first applied to spin glass simulations [19]. Independent systems or replicas were simulated with Monte Carlo moves for regular time intervals. Between each of these intervals, exchange Monte Carlo moves, involving the exchange of the two systems' respective thermodynamic conditions, were attempted. Sugita and Okamoto have derived a molecular dynamics version of this algorithm [20]. The fundamental idea of their replica exchange molecular dynamics (REMD) scheme is presented in this section, along with the final development that leads to the Monte Carlo probability function which determines whether exchange attempts will be accepted.

Given a system x in state i, this state can be described by $x^i = (p^i, q^i)$, where q^i stands for the entire set of atomic coordinates and p^i for the atomic momenta. The Hamiltonian $H(q, p)$ of the system is the sum of the kinetic energy $K(p)$ and the potential energy $E(q)$. In a canonical ensemble, the Boltzmann factor (or probability to find the system i in given state) is given by:

$$W(x^i) = \frac{1}{Z} exp \left\{ -\beta H(p^i, q^i) \right\} \tag{1.6}$$

where Z is the partition function, $Z = tr\left(e^{-\beta H}\right)$.

In replica exchange, a generalized ensemble $X(..., x_m^i, ..., x_n^j, ...)$ is constructed with M copies, or replicas, of the system x, at different states i,j, etc. and temperatures m,n, etc. With non-interacting replicas, the Boltzmann factor of this REMD generalized ensemble is given by the product of the Boltzmann factors of each replica:

$$W_{REMD}(X) = \frac{1}{Z'} exp \left\{ -\sum_{m=1}^{M} \beta_m H(q^{i(m)}, p^{i(m)}) \right\} \tag{1.7}$$

where $i(m)$ stands for a given state of the system at temperature m and Z' is the partition function for the generalized ensemble. Now consider exchanging a pair of replicas in the generalized ensemble. An exchange of replicas i and j which are at temperatures T_m and T_n can be written as:

$$X(..., x^i_m, ..., x^j_n, ...) \rightarrow X'(..., x^j_m, ..., x^i_n, ...) \tag{1.8}$$

The kinetic energy of a replica is determined by the momenta p of its N atoms k:

$$K(p) = \sum_{k=1}^{N} \frac{p_k^2}{2m_k} = \frac{3}{2} N k_N T \tag{1.9}$$

By exchanging replicas, we assign them to new temperatures. The most natural way to do this is to rescale velocities uniformly by the square root of the ratio of the two temperatures, so that the kinetic energy (1.9) is rescaled according to the ratio of the two temperatures:

$$\begin{aligned} p^{i'} &= \sqrt{\frac{T_n}{T_m}} p^i & p^{j'} &= \sqrt{\frac{T_m}{T_n}} p^j & \quad (1.10) \\ (p^{i'})^2 &= \frac{T_n}{T_m} (p^i)^2 & (p^{j'})^2 &= \frac{T_m}{T_n} (p^j)^2 & \\ K(p^{i'}) &= \frac{T_n}{T_m} K(p^i) & K(p^{j'}) &= \frac{T_m}{T_n} K(p^j) & \quad (1.11) \end{aligned}$$

Imposing the detailed balance condition on the transition probability $w(X \rightarrow X')$ will force the exchange process to converge to an equilibrium distribution:

$$\begin{aligned} W_{REMD}(X) \quad w(X \rightarrow X') &= W_{REMD}(X') \quad w(X' \rightarrow X) \\ \frac{w(X \rightarrow X')}{w(X' \rightarrow X)} &= \frac{W_{REMD}(X')}{W_{REMD}(X)} \end{aligned} \tag{1.12}$$

Replacing the numerator $W_{REMD}(X')$ and denominator $W_{REMD}(X)$ by their expression according to (1.7), the terms involving all the replicas cancel out, with the exception of those that are to be exchanged:

$$\begin{aligned} \frac{W_{REMD}(X')}{W_{REMD}(X)} &= \exp\Big\{ -\beta_m \left[K(p^{j'}) + E(q^j) \right] - \beta_n \left[K(p^{i'}) + E(q^i) \right] \\ &\quad + \beta_m \left[K(p^i) + E(q^i) \right] + \beta_n \left[K(p^j) + E(q^j) \right] \Big\} & (1.13) \\ &= \exp\left\{ -\beta_m \frac{T_m}{T_n} K(p^j) - \beta_n \frac{T_n}{T_m} K(p^i) + \beta_m K(p^i) + \beta_n K(p^j) \right. \\ &\quad \left. -\beta_m \left[E(q^j) - E(q^i) \right] - \beta_n \left[E(q^i) - E(q^j) \right] \right\} & (1.14) \\ &= \exp\left\{ [\beta_n - \beta_m] \left[E(q^i) - E(q^j) \right] \right\} = \exp(-\Delta) & (1.15) \end{aligned}$$

where (1.13) is rewritten with the velocity rescaling from (1.11), resulting in (1.14), which is further simplified by

$$\beta_m \frac{T_m}{T_n} = \frac{1}{k_B} \frac{T_m}{T_n} = \frac{1}{k_B T_n} = \beta_n \quad \text{and} \quad \beta_n \frac{T_n}{T_m} = \frac{1}{k_B} \frac{T_n}{T_m} = \frac{1}{k_B T_m} = \beta_m \tag{1.16}$$

removing the momenta from the final ratio and obtaining the same result (1.15) as for the original algorithm [19], in which the sampling between exchanges was performed by Monte Carlo moves (devoid of momenta).

Although the probability $w(X \rightarrow X')$ remains unknown, we can set it to any value satisfying the detailed balance condition (1.12). For example, with the choice of Metropolis et al. [21], we set:

$$w(X \rightarrow X') = \frac{W_{REMD}(X')}{W_{REMD}(X)} = \exp(-\Delta) \tag{1.17}$$

$$\mathrm{P}_{ex} = \begin{cases} 1, & \text{for}\Delta \leq 0, \\ \exp(-\Delta), & \text{for}\Delta > 0 \end{cases} \tag{1.18}$$

For each exchange attempt, the exchange probability P_{ex} is computed according to (1.18). The potential energy difference in the second term of (1.15) shows that high exchange probabilities require the potential energy of a replica at T_{n+1} to be low enough to be part of the overlap of potential energy distributions of T_n and T_{n+1}. The exchange is then automatically accepted if the energy of the replica at T_{n+1} is lower than the one at T_n ($\Delta \leq 0$ and $P_{ex} = 1$) and accepted with a probability P_{ex} otherwise. In this case, P_{ex} is compared to a random number, and if the former is greater than the latter, the exchange is accepted. The momenta of the atoms of two exchanged replicas are rescaled according to (1.10). In the exchange, the replica at T_{n+1} becomes a member of the T_n canonical distribution (and the one at T_n, a member of the T_{n+1} canonical distribution). Thus, the major advantage of REMD is that it builds canonical distributions at all temperatures, and excludes unlikely high energy structures from low temperature ensembles.

In multi temperature REMD, N replicas, or copies of a given molecular system, undergo parallel MD simulations at different temperatures. These temperatures T_1, T_2,...,T_{n-1}, T_n cover a given range, in which the lowest temperature T_1 is representative of a “ground state“ (i.e. the temperature at which the force field was parameterized) and the higher temperatures allow to escape any energetic minima the simulation encounters. At given times, all the MD simulations are stopped and exchanges are attempted among adjacent pairs of replicas that are simulated at neighboring temperatures T_n and T_{n+1}. Once exchanges have been attempted on all replica pairs and the respective temperature reassignments performed where relevant, the parallel MD simulations are restarted. An REMD simulation consists in an iteration of parallel MD simulations and exchange attempts. In the latter step, the replicas can move in ”temperature space” and thus escape energy minima. This scheme only holds if replicas are fully decorrelated between exchange attempts, i.e. if the MD time between two exchange attempts allows for the full decorelation of a replica exchanged in an exchange attempt and the same replica before the next attempt.

Different exchange schemes are possible. The most common consist in alternating exchange schemes E_1 and E_2 at the end of an MD interval, with E_1 attempting to exchange all n / $n+1$ pairs, and E_2, all $2n$ / $2n+1$ pairs. Exchange strategies are not restricted to adjacent temperature intervals, but in practice, exchange attempts performed on pairs of replicas with larger temperature differences result in very low exchange probabilities and are never accepted.

1.3 References

[1] Y. Duan and P. A. Kollman. “Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution.” *Science*, **282**, (1998) 740–744.

[2] W. F. van Gunsteren, R. Buergi, C. Peter, and X. Daura. “The Key to Solving the Protein-Folding Problem Lies in an Accurate Description of the Denatured State”. *Angew Chem Int Ed Engl*, **40**, (2001) 351–355.

[3] S. Jang, E. Kim, S. Shin, and Y. Pak. “Ab initio folding of helix bundle proteins using molecular dynamics simulations.” *J Am Chem Soc*, **125**, (2003) 14 841–14 846.

[4] L. Verlet. “Computer ”experiments” on classical fluids. i. thermodynamical properties of lennard-jones molecules”. *Physical Review*, **159**, (1967) 98–103.

[5] T. CheathamIII, P. Cieplak, W. Cornell, and P. Kollman. “A modified version of the cornell et al. force field with improved sugar pucker phases and helical repeat”. *Journal of Biomolecular Structure & Dynamics*, **16**, (1999) 845–862.

[6] W. van Gunstern, S. Billeter, A. Eising, P. Huenenberger, P. Krueger, A. Mark, W. Scott, and I. Tironi. *Biomolecular Simulation: The GROMOS96 Manual and User Guide* (Hochschulverlag AG, Zuerich, 1996).

[7] W. Jorgenson and J. Tiradorives. “The opls potential functions for proteins - energy minimizations for crystals of cyclic-peptides and crambin”. *J Am Chem Soc*, **110**, (1988) 1657–1666.

[8] U. Essmann, L. Perera, M. Berkowitz, T. Darden, H. Lee, and L. Pedersen. “A smooth particle mesh ewald method”. *J. Chem. Phys.*, **103**, (1995) 8577–8593.

[9] I. G. Tironi, R. Sperb, P. E. Smith, and W. F. van Gunsteren. “A generalized reaction field method for molecular dynamics simulations”. *J Chem Phys*, **102**, (1995) 5451–5459.

[10] M. M. Seibert, A. Patriksson, B. Hess, and D. van der Spoel. “Reproducible polypeptide folding and structure prediction using molecular dynamics simulations.” *J Mol Biol*, **354**, (2005) 173–183.

[11] J. vandeVondele and U. Rothlisberger. “Canonical adiabatic free energy sampling (cafes): A novel method for the exploration of free energy surfaces”. *J Phys Chem B*, **106**, (2002) 203–208.

[12] M. Iannuzzi, A. Laio, and M. Parrinello. “Efficient exploration of reactive potential energy surfaces using car-parrinello molecular dynamics.” *Phys Rev Lett*, **90**, (2003) 238 302.

[13] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. “Optimization by simulated annealing”. *Science*, **220**, (1983) 671–680.

[14] W. Wenzel and K. Hamacher. “Stochastic tunneling approach for global minimization of complex potential energy landscapes”. *Phys Rev Lett*, **82**, (1999) 3003–3007.

[15] Grubmller. “Predicting slow structural transitions in macromolecular systems: Conformational flooding.” *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, **52**, (1995) 2893–2906.

[16] Y. Okamoto. “Generalized-ensemble algorithms: enhanced sampling techniques for monte carlo and molecular dynamics simulations.” *J Mol Graph Model*, **22**, (2004) 425–439.

[17] A. Schug, T. Herges, and W. Wenzel. “Reproducible protein folding with the stochastic tunneling method.” *Phys Rev Lett*, **91**, (2003) 158 102.

[18] A. Mitsutake, Y. Sugita, and Y. Okamoto. “Generalized-ensemble algorithms for molecular simulations of biopolymers.” *Biopolymers*, **60**, (2001) 96–123.

[19] K. Hukushima and K. Nemoto. “Exchange monte carlo method and application to spin glass simulations”. *J. Phys. Soc. Japan*, **65**, (1996) 1604–1608.

[20] Y. Sugita and Y. Okamoto. “Replica-exchange Molecular Dynamics Method for Protein Folding”. *Chem. Phys. Lett.*, **314**, (1999) 141–151.

[21] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. N. Teller, and E. Teller. “Equation of state calculations by fast computing machines”. *J Chem Phys*, **21**, (1953) 1087–1092.

Chapter 2

Testing the validity of solute energy based replica-exchange molecular dynamics

2.1 Abstract

A well known limitation of REMD simulations for large explicit solvent systems are the low exchange probabilities. Ad hoc protocols have been devised to circumvent this problem. A frequent approach consists in computing a reduced potential energy function used for the computation of increased exchange probabilities, while the standard total potential energy E_{tp} is used do drive MD between exchanges. This reduced potential energy function is usually obtained by adding all solute terms and an approximation of the solvation energy. In order to assess the validity of this approximation, we test its simplest implementation: We construct a reduced potential energy function E_{pp} in which we retain all solute energy terms and half of the solute-solvent interaction energy terms, discarding all solvent-solvent interaction energy terms. This approximation is tested by running a REMD simulation of the 103-residue prion protein. The practical validity of this REMD partial energy (*REMDpe*) simulation is evaluated by (i) comparing potential energy distributions (E_{tp} and E_{pp}) to the ones we obtain from reference MD simulations, (ii) with the energy distribution overlaps, used to verify that *REMDpe* exchanges are being performed according to a Boltzmann weight and (iii) by comparing E_{pp} and E_{tp} averages obtained for different structural clusters (i.e. native, collapsed and non-native structures), in order to check whether the relative energy ordering of the different groups are reproduced with both potential energy functions.

2.2 Introduction

Classical molecular dynamics simulations are becoming a more and more powerful tool for the prediction of molecular properties, including protein structure and dynamics [1, 2, 3]. Unfortunately, when applied to the study of systems that are subject to extensive conformational changes (e.g. protein folding), current MD simulations correspond to physical times in the order of a few hundreds of nano seconds [4, 1], which is far from sufficient to describe folding processes of the order of microseconds to milliseconds [5]. This is in part due to the high number of degrees of freedom involved in protein folding, implying a very rugged free energy surface with many local minima in which MD simulations can get trapped.

To alleviate this problem, different enhanced sampling techniques have been developed [6, 7, 8, 9, 10, 11], among which replica-exchange molecular dynamics (REMD) [12] is one of the most popular. In REMD, N independent replicas of a given molecular system undergo parallel MD simulations at different temperatures. These temperatures T_1, T_2,...,T_{n-1}, T_n cover a given range, in which the lowest temperature T_1 is representative of a “ground state“ (e.g. the temperature at which the force field was parameterized) and the higher temperatures allow for fast barrier crossing rates ensuring the escape from local energetic minima. At given times, all MD simulations are stopped and exchanges are attempted among pairs of replicas that are simulated at neighboring temperatures T_n and T_{n+1}. In an exchange, the conformation that was so far simulated at T_n is assigned to the new temperature T_{n+1}, and the one at T_{n+1} to T_n. Exchanges are accepted or refused according to a Monte-Carlo exchange probability P_{ex} that is constructed so that exchanges generate Boltzmann distributions. The joint probability distribution of an extended system composed of multiple copies of an original system can be written under the assumption that each temperature samples a Boltzmann distribution. The exchange probability P_{ex} can be obtained by solving the equations describing this joint probability [12]:

$$P_{ex} \quad = \quad min\left(1, exp\left\{\left[\frac{1}{k_B T_n} - \frac{1}{k_B T_{n+1}}\right]\left[E(x^i_{T_n}) - E(x^j_{T_{n+1}})\right]\right\}\right) \qquad (2.1)$$

where k_B is Boltzmann's constant, T_n and T_{n+1} two adjacent temperatures of the input temperature range, and $E(x^j_{T_{n+1}})$ and $E(x^i_{T_n})$ the potential energies of replicas i and j, with the corresponding sets of coordinates $x^j_{T_{n+1}}$ and $x^i_{T_n}$. The potential energy difference in the second term shows that high exchange probabilities require the potential energy of a replica at T_{n+1} to be low enough to be part of the overlap of potential energy distributions of T_n and T_{n+1}. The exchange is then automatically accepted if the energy of the replica at T_{n+1} is lower than the one at T_n and accepted with a probability P_{ex} otherwise. In this case, P_{ex} is compared to a random number, and if the former is greater than the latter, the exchange is accepted. Once exchanges have been attempted on all replica pairs and the respective temperature re-assignments performed where relevant, the parallel MD simulations are restarted. This scheme only holds if replicas are fully decorrelated between exchange attempts, i.e. if the MD time between two exchange attempts allows for the full decorelation of the replicas between exchanges.

Ideally, a replica should cross the full temperature range several times during an REMD simulation. Every exchange allows to change temperature, and potentially escape a local minimum and sample new conformations. When large systems (i.e. proteins solvated in an explicit solvent) are simulated, a small temperature difference has to be chosen for adjacent replicas in order to obtain a reasonable overlap of their respective potential energy distributions. The number of replicas required to cover a given REMD temperature range efficiently (i.e. sufficient exchange probabilities and motion in temperature space) actually scales with $\sqrt{f} ln\left[\frac{T_{max}}{T_{min}}\right]$, where f is the number of degrees of freedom of the system, T_{min} the lowest REMD temperature and T_{max} the highest [13]. For large, biomolecular systems, f and the required number of replicas (and CPUs) tend to get very high [14, 15]. This problem mainly appears in explicit solvent simulations, where the large number of water molecules required to solvate a system of interest increases f dramatically.

One way to circumvent this problem are Hamiltonian REMD strategies. A generalized formulation of REMD allows for parallel trajectories to be performed on other variables than temperature. Hamiltonian REMD schemes based on atomic position overlaps [13], hydrophobicity parameters [13], force field parameters [16] and scaled potential energies [17] generally reduce the number of required replicas. While any transformation of the Hamiltonian can be applied to the "sampling" simulations, a “target“ simulation is performed with the correct Hamiltonian and can only be accessed by physically acceptable configurations, as ensured by the exchange criterion. The "replica exchange with solute tempering" (REST) Hamiltonian REMD scheme was specifically designed to reduce the number of replicas in explicit solvent REMD simulation based on multi-temperature simulations [18]. This approach is promising for small systems [18] but can fail for larger ones [19].

On the other hand, several ad hoc methods for explicit solvent REMD have been developed based on the concept to use the potential energy of a subsystem for the evaluation of the exchange probabilities to decrease the number of replicas needed to cover a given temperature range [20, 21, 22]. Potential energies of a subsystem can be chosen to be more representative of the protein conformational space and the simulation will not mainly sample "environment" (i.e. solvent) degrees of freedom. Moreover, subsystem potential energies have a larger overlap leading to increased exchange probabilities. The time spent on sampling mere solvent rearrangements was demonstrated with a heptapeptide explicit solvent REMD simulation in which specific exchanges were primarily dominated by fluctuations of the solvent-solvent interactions [23]. One such "partial REMD“ (PREMD) scheme uses 2 thermostats for each replica, the first controlling the temperature of the subsystem (with different temperatures for each replica providing for the enhanced sampling of the subsystem) and the second, the one of the environment (same temperature for all replicas) [20]. Thus, the subsystem potential energy is the total potential energy of a system for which only the subsystem is at a different (higher) temperature.

A related strategy consists in using a single target temperature total potential energy (E_{tp}) to drive MD, with a reduced potential energy function computed specifically at every exchange trial, where it is used to determine the exchange probability. This reduced potential energy includes all potential energy terms of the subsystem (usually, the solute, e.g. the solvated protein) and uses an approximation to compute the potential energy of solvent-solvent and solvent-solute interactions. In the so called hybrid REMD, one can choose to include the first layers of water molecules solvating the solute of interest into the subsystem, while a generalized Born energy term gives an estimate of the subsystem to solvent and solvent to solvent interactions in the reduced potential energy function [21]. In practice, technical complications arise in the choice of the number of explicit solvent molecules and in the migration of waters and ions from the first subsystem solvation shells to outer "bulk" water. Such migrations can lead to artificial discontinuities in the solute potential energy. REMD hybrid Poisson-Boltzmann exploits the same idea, except that no explicit solvent molecules are considered and the solvation energy is approximated by the Poisson-Boltzmann equation [22]. These techniques are very recent and have not yet been extensively tested. Although these methods are appealing, they are not based on a rigorous physical justification and do not guarantee that Boltzmann distributions are maintained.

In this work, we opted for one of the simplest possible implementation that neglects solvent-solvent contributions entirely and test the validity of this first order approximation for large protein simulations. Thus, the reduced potential energy function includes the total potential energy of the protein and half of the protein-solvent interaction energy. In this way, we renounce to introduce any approximation of the solvent-solvent interaction energy. This approach is based on the assumption that solvent to solvent interactions play a minor role in conformational changes of the solute. Thus, this method is termed REMD partial energy, or *REMDpe*, illustrating the fact that the reduced (or partial) potential energy function (E_{pp}) is obtained by discarding less important terms of the total potential energy:

$$E_{pp} = E_{solute}^{bonded} + E_{solute}^{non-bonded} + \frac{1}{2} E_{solute-solvent}^{non-bonded} \tag{2.2}$$

$E_{solute-solvent}^{non-bonded}$ is scaled down by a factor of $\frac{1}{2}$ in order to give more weight to protein energy terms.

We have chosen to test this approximation with the 103-residue prion protein (PrP). A misfolded, β-rich conformer, PrP^{Sc}, is involved in various forms of spongiform encephalopathies [24, 25, 26, 27, 28]. Experiments [29, 30, 31, 32, 33, 34, 35, 36] and simulations [37, 38, 39] have shown that this protein possesses a very rich folding landscape, providing an interesting test case for an enhanced sampling method.

A NVT *REMDpe* run is performed in order to test the practical validity of the method and detect possible introduced artefacts. The averages of the 32*88.4 ns long *REMDpe* simulations are first exploited to supply a quantitative decomposition of the total potential energy (E_{tp}) in terms of protein and solvent contributions. The validity of the approximate exchange protocol based on E_{pp} only is then assessed by following three tests: (i) A comparison of the potential energy distributions (E_{tp} and E_{pp}) to the ones computed from reference MD simulations, representing the “true“ canonical distributions at a given temperature, (ii) The potential energy histogram overlaps, that can be used to verify that *REMDpe* exchanges are being performed according to a Boltzmann weight and (iii) a comparison of E_{pp} and E_{tp} averages obtained for different structural clusters (i.e. native, collapsed and non-native structures), in order to investigate whether the relative preferences are retained. To anticipate our results, the three validations suggest that *REMDpe* approach does not seem to introduce any visible artifacts and can thus be used to study large proteins in an explicit solvent.

2.3 Methods

All molecular dynamics (MD) simulations were performed with the GROMACS-3.3.0 package [40], the GROMOS96 force field [41], the SPC water model [42] and an MD timestep of 1.5 fs with constraints on covalent bonds involving hydrogen atoms (tolerance of 10^{-4} kJ(mol nm)$^{-1}$) applied via the SHAKE algorithm [43]. The starting configuration was based on the mouse PrP NMR structure (res 124-226, PDB code 1AG2 [44]). The protonation states of ionizable side chains were predicted by finite-difference Poisson-Boltzmann calculations [45, 46] in order to mimic a pH 4 environment favoring possible conformational changes [47]. The DELPHI program [48] supplied with the WHATIF package [49] was used to solve the Poisson-Boltzmann equation. The protonation states of HIS, ASP and GLU side chains were consequently set as follows: HIS17; p, ASP21; d, GLU23; d, ASP24; d, GLU29; p, ASP44; p, HIS54; p, ASP55; d, HIS64; p, GLU73; d, GLU77; p, ASP79; d, GLU84; p, GLU88; p and GLU98; p (where p stands for protonated and d for deprotonated). A rhombic dodecahedral box with 14076 water molecules was constructed around the protein. 8 Cl^- ions were added to neutralize protein charges.

All simulations, including the equilibration of the NMR structure, as well as the $REMDpe$ equilibration and production runs, were performed with 2 Nosé-Hoover thermostats (one for the protein and one for solvent and counter ions, with respective time-coupling constants of 0.4 and 1.6 ps) [50, 51]. Coulombic interactions were treated using a twin-range cutoff, in which interactions within 1.0 nm and between 1.0 and 1.4 nm were computed every MD step, respectively every 5 MD steps. Electrostatic interactions beyond 1.4 nm were approximated with a generalized reaction field [52] generated by a dielectric continuum with dielectric constant of 66.

The NMR structure was first minimized by 500 steps of steepest descent. The equilibration was performed in the NPT ensemble (Berendsen barostat [53] with a time coupling constant of 1 ps). First, the solvent was equilibrated for 225 ps at 300K with protein atom position restraints of $25*10^3$ $kJmol^{-1}nm^{-2}$. Second, the system was gradually heated to 300K with 6 successive 50 ps MD runs, for which the increasing temperatures were 50, 100, 150, 200, 250 and 300K and the decreasing protein atom position restraints were $25*10^3$, $5*10^3$, $3.75*10^3$, $2.5*10^3$ and $1*10^3$ $kJmol^{-1}nm^{-2}$. In the final stage, 3 ns of 300K MD with no position restraints supplied the starting configuration of the $REMDpe$ simulations.

The REMDpe protocol was implemented in GROMACS-3.3.0. We have modified the source code so that we can sum up the following potential energy terms: solute bonded (E_{solute}^{bonded}), solute non-bonded ($E_{solute}^{non-bonded}$) and half of the solute to solvent non-bonded ($\frac{1}{2}E_{solute-solvent}^{non-bonded}$), into the partial potential energy function E_{pp} (equation 2.2). The REMD routine of GROMACS was then adapted accordingly in order to compute exchange probabilities with this reduced potential energy function.

Trial $REMDpe$ simulations were performed with different temperature distributions for the replicas, in order to ensure similar exchange probabilities at all temperatures. The following temperature distribution was selected for the 32 replicas and used for the production run, performed in the NVT ensemble: 300.0, 306.0, 312.1, 318.2, 324.3, 330.4, 336.6, 342.8, 349.1, 355.3, 361.6, 368.0, 374.3, 380.7, 387.1, 393.6, 400.1, 406.6, 413.1, 419.7, 426.3, 432.9, 439.6, 446.3, 453.0, 459.8, 466.5, 473.3, 480.2, 487.1, 494.0 and 500.9K. With this distribution, we obtained roughly uniform $REMEpe$ exchange frequencies of ~30%.

The equilibration of the $REMDpe$ replicas consisted in 32 parallel MD runs linearly heating (linear increase of the thermostat target temperature) 32 copies of the equilibrated initial configuration described above to the different $REMDpe$ temperatures within 300 ps. MD was continued for another 200ps at these temperatures, generating the $REMDpe$ starting replicas. The $REMDpe$ production run was carried out for 88.4 ns (total aggregate time of 2.8 μs), with exchange attempts performed every 60 ps on adjacent temperature replicas x and x+1 for odd exchange trial numbers, and 2x and 2x+1 for even ones. Structures, energies and temperatures were saved every 1.5 ps.

2.4 Results and Discussion

2.4.1 Protein contributions to the total potential energy

In Table 2.1, a list of all contributions to the potential energy, computed as time average over the entire 88.4 ns *REMDpe* at 300K, is given. The largest contribution (97%) to E_{tp} originates from solvent-solvent non-bonded interactions. Although a certain number of solvent to solvent interactions in the direct vicinity of the protein might contribute to the folding process, we have chosen to test the approximation consisting in excluding solvent-solvent contributions entirely from E_{pp} used to calculate the *REMDpe* exchange probabilities. The total protein energy arises from the sum of stabilizing and destabilizing interactions, reflected in the negative and positive signs of the potential energy contributions, respectively. -68%, 95.6% and 72.5% of the E_{pp} originate from, respectively, protein bonded interactions, protein non-bonded interactions and half of the protein-solvent non-bonded interaction.

2.4.2 Potential energy distributions

We performed two reference MD simulations at NVT, one at 300K (328.4 ns) and one at 500K (36.2 ns). The validation of the *REMDpe* run consisted in verifying whether the E_{tp} and E_{pp} distributions at 300K and 500K resulting from the *REMDpe* simulations are identical to those of the reference simulations (Figure 2.1). The superposition shows consistent distributions for the four cases, with no obvious shifts between *REMDpe* and reference MD distributions.

2.4.3 Energy histogram overlaps

Energy histogram overlaps allow to assess whether *REMDpe* exchanges are being performed according to a Boltzmann weight. Equation 2.3 gives the probabilities $P_i(E)$ and $P_{i+1}(E)$ to find a conformation of energy E at two neighboring NVT REMD temperatures T_i and T_{i+1}. β_i is the reciprocal temperature, defined as $\frac{1}{k_B T_i}$, where k_B is the Boltzmann constant and Z_i is the partition function at temperature T_i.

$$P_i(E) = \frac{1}{Z_i} e^{-\beta_i E} \quad \text{and} \quad P_{i+1}(E) = \frac{1}{Z_{i+1}} e^{-\beta_{i+1} E} \tag{2.3}$$

This equation only holds if both REMD temperatures exchange structures according to a Boltzmann weight. The ratio $\frac{P_{i+1}(E)}{P_i(E)}$ can be computed in the overlap of the E_{tp} distributions obtained with two REMD temperatures:

Table 2.1: *Potential energy contributions from solvent and solute atoms. The numbers given are the averages obtained from 88.4 ns NVT* REMDpe *at 300K.* E^{bonded}_{solute}, $E^{non-bonded}_{solute}$ *and* $E^{non-bonded}_{solute-solvent}$ *are defined in Equation 2.2,* $E^{non-bonded}_{solvent-solvent}$*; potential energy contribution due to solvent-solvent non-bonded interactions,* Coul*; Coulombic electrostatic interaction,* SR*; "short range" interaction energy between atom A and B within the primary cutoff of 1.0 nm (twin-cutoff scheme),* LR*; "Long range" interaction energy between atom A and B located between the primary and the secondary cutoff (of 1.4 nm),* LJ*; Lennard-Jones interaction,* 14*; Special interactions between bonded atoms n and n+3,* E_p*; potential energy contribution,* E_{tp}*; total potential energy and* E_{pp}*; partial potential energy used to compute exchange probabilities.*

Interaction group	Interaction	E$_p$	% of E_{tp}	% of E_{pp}
E^{bonded}_{solut}	Bonds	1151.595	-0.2	-18.0
	Angles	1820.453	-0.3	-28.5
	Proper Dihedrals	768.159	-0.1	-12.0
	Improper Dihedrals	613.970	-0.1	-9.6
$E^{non-bonded}_{solut}$	Coul-SR[1]	-25210.687	4.2	181.6
	LJ-SR	-3772.628	0.6	59.0
	Coul-LR	-44.652	0.0	0.7
	Coul-14	9129.828	-1.5	-142.7
	LJ-14	190.174	0.0	-3.0
$E^{non-bonded}_{solut-solv}$	Coul-SR[2]	-8382.983	1.4	65.5
	LJ-SR[2]	-917.753	0.2	7.2
	Coul-LR[2]	22.487	0.0	-0.2
$E^{non-bonded}_{solv-solv}$	Coul-SR	-676559.668	113.2	
	LJ-SR	100073.388	-16.7	
	Coul-LR	-2740.947	0.5	
Total	E$_{pp}$	-6397.729	1.1	
	E$_{tp}$	-597461.534		

[1]RF-excl (GROMACS reaction field correction for excluded atom pairs) is included in Coul-SR of E_{tp}, but not included in E_{pp}.
[2]$E^{non-bonded}_{solute-solvent}$ terms of E_{pp} are weighted by one half (details in text).

$$\frac{P_{i+1}(E)}{P_i(E)} = \frac{Z_i}{Z_{i+1}} \times e^{-\beta_{i+1}E+\beta_i E}$$

$$\ln\left(\frac{P_{i+1}(E)}{P_i(E)}\right) = (\beta_i - \beta_{i+1})E + C \quad \text{where} \quad C = \ln\left(\frac{Z_i}{Z_{i+1}}\right) \text{is a constant.} \tag{2.4}$$

Equation 2.4 describes a line $y = a\ x + b$. Plotting $y = ln\left(\frac{P_{i+1}(E)}{P_i(E)}\right)$ as a function of E, one should therefore obtain a line of slope $\beta_i - \beta_{i+1}$. This method has been used to verify that REMD exchanges were being performed according to a Boltzmann weight [54, 55].

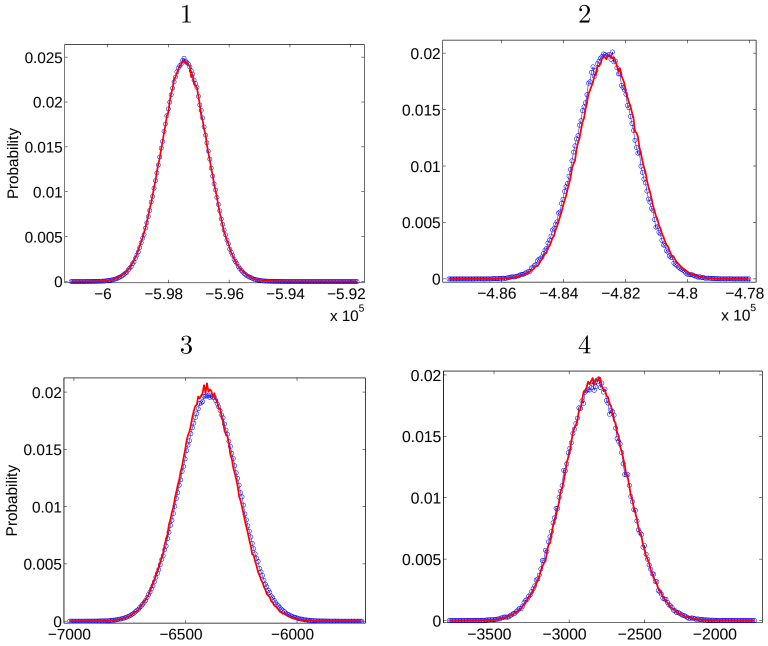


Figure 2.1: *Superpositions of reference MD and* REMDpe *potential energy distributions. Blue lines and circles; reference MD simulation, red thick line; REMDpe. Panels 1 and 3: 300K, panels 2 and 4: 500K, panels 1 and 2:* E_{tp}*, panels 3 and 4:* E_{pp}*. The distributions were all computed with 200 bins and normalized in order to compare* REMDpe *and reference MD trajectories of different lengths.*

We computed $ln\left(\frac{P_{i+1}(E)}{P_i(E)}\right)$ values and plotted them as a function of energy for all replicas, along with linear fits yielding an r^2 of at least 0.96 (worst fit, obtained for temperatures 324/330K)(Figure 2.2). Furthermore, the slopes of the fitted lines were in agreement with the theoretical value of $\beta_i - \beta_{i+1}$ (panel 6), indicating that the *REMDpe* exchanges are indeed being performed according to a Boltzmann weight.

2.4.4 Relative average energies of conformational clusters

In this test, we want to probe if *REMDpe* introduces any artificial biases in the sampling of closely related and more distant structural groups. In the context of the prion test case, an obvious choice of structural groups constitutes in folded, native states as opposed to non-native ones. We used three separate geometrical properties to define three closely related native groups: *Native - RMSD* (structures with an RMSD value $\leq$ 0.44 nm to the NMR reference structure), *native -fraction of native contacts* (structures with a fraction of native contacts $\geq$ 0.64, using the same reference) and *native - secondary structure* (structures with at least 38 α-helical and 4 β-sheet residues). Structures were permitted to belong to one, two or three different native groups. Two distant structural groups were defined for the non-native state (structures that did not belong to any of the three groups above): The *collapsed* group, arbitrarily defined by non-native structures with a radius of

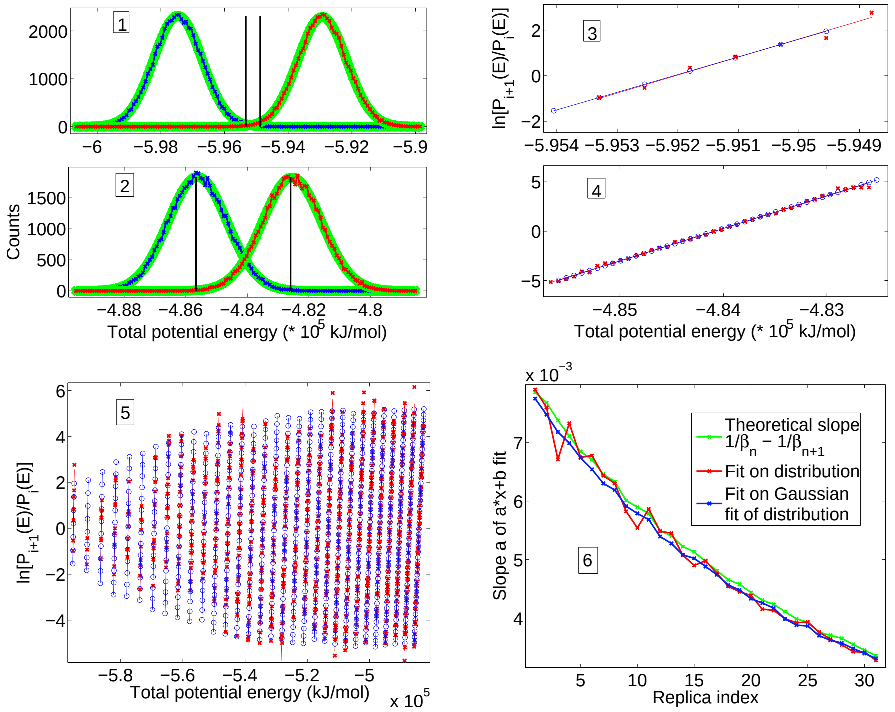


Figure 2.2: *Energy histogram overlaps.* E_{tp} *distribution overlaps of temperatures 300/306K (blue and red, lowest temperature pair, panel 1) and 494/501K (blue and red, highest temperature pair, panel 2). Gaussian functions (green) fit the distributions well* ($r^2 > 0.99$)*.* $P_i(E)$ *and* $P_{i+1}(E)$ *points of the overlaps (delimited by black vertical lines) between distributions of temperature pairs* T_i *and* T_{i+1} *are used to compute the values of* $ln[P_{i+1}(E)/P_i(E)]$ *shown in panels 3 (temperatures 300/306K), 4 (temperatures 494/501K) and 5 (all temperature pairs). Overlaps and* $ln[P_{i+1}(E)/P_i(E)]$ *were computed twice, once starting from the raw distributions (red crosses in panels 3 to 5) and once starting from points of the Gaussian functions that were fitted to the raw distributions (blue circles in panels 3 to 5).* $ln[P_{i+1}(E)/P_i(E)]$ *points should form a line of slope* $\beta_i - \beta_{i+1}$*, according to equation 2.4. This was tested by fitting lines to the* $ln[P_{i+1}(E)/P_i(E)]$ *points that were computed from the raw distributions (red lines) and to the* $ln[P_{i+1}(E)/P_i(E)]$ *points that were computed from the fitted Gaussians (blue lines) in panels 3 to 5. Panel 6 shows the slopes of these fitted lines for all replica pairs: Red indicates the slopes computed from the raw distribution, blue the ones computed from the fitted Gaussians and green the theoretical slopes* $\beta_i - \beta_{i+1}$ *one should obtain.*

gyration ≤ 1.3 nm and *other*, with non-native structures that had larger radii of gyration. In summary, we defined three native structure groups (*native - RMSD*, *native - fraction of native contacts* and *native - secondary structure*) and two non-native structure groups (*collapsed* and *other*). Typical examples of these structural groups are shown in Figure 2.3.

In order to ensure that the E_{pp} and the E_{tp} favor the same conformational clusters, distributions were computed for both potential energy functions, for each of the 5 structural groups,at all temperatures, and for 22.1 ns consecutive portions of the simulation (Figure 2.4). The average energy differences between the groups is very small, especially when compared to the overlap of the distributions (error bars). This is consistent with the marginal energetic stability of proteins, characterized by folding free energies of -20 to -60 kJ/mol [56] and much larger thermal fluctuations ($\sim$ 300 kJ/mol [57]). In order to identify the minimal potential energy structural group, relative potential energies E^r_{xp} (with x = t for total or p for protein) were defined as the difference between the average E_{xp} of a given structural group and the average E_{xp} of the native group (fraction of native contacts definition). The native group was chosen as reference because it is supposed to occupy the potential energy global minimum. In other words, if the E^r_{xp} of a given structural group is positive, that structural group has a higher average potential energy than the one of the native group.

The E^r_{xp} were computed for 22.1 ns consecutive portions of the simulation as a function of structural group and temperature (Figure 2.5). One can thus rank the structural groups according to their energetic stability for every temperature, obtaining temperature dependent structure-energy profiles. Our third validation consists in checking how well the trends (qualitative agreement) of the $\bar{E}^r_{pp}$ and $\bar{E}^r_{tp}$ profiles match to test if any bias is introduced by the $REMDpe$ approximation that would favor different structural groups. Despite the large overlap of error bars, a similar profile emerges for both E^r_{tp} and E^r_{pp}, with the following ranking of groups as a function of decreasing stability: The three native groups (in varying order), followed by the two non-native ones. A difference appears in the ranking of these two non-native groups: While the E_{tp} seems to favor *other* over *collapsed*, the opposite ranking is found for the E_{pp}, suggesting that the E_{pp} might over stabilize collapsed states for non native structures. However, the temperature dependent structure-energy profiles are globally not very different, and show that the native structure occupies the minimum of both E_{tp} and E_{tp}. Therefore, no obvious bias is introduced by our approximation, suggesting that normal REMD (with exchange probabilities based on E_{tp}) would favor the same structural groups, that would eventually accumulate at lower temperatures. Finally, the profiles do not significantly differ from one 22.1 ns simulation interval to another.

2.5 Conclusions

In order to test the validity of REMD performed with a reduced potential energy function, we evaluate the practical validity of the most simple implementation of this strategy; an REMD method in which exchange probabilities are computed based on a partial potential energy (E_{pp}) term that completely discards solvent to solvent interactions. A number of tests are used to validate the performance of this $REMDpe$ approach: (i) For 300K and 500K, we obtain the same E_{tp} and E_{pp} distributions as with reference MD simulations at the same temperature; (ii) Energy histogram overlaps show that $REMDpe$ exchanges are being performed according to Boltzmann weights, and (iii) Similar relative stabilities are predicted for native and non-native structural groups using E_{pp} and E_{tp}, with the native

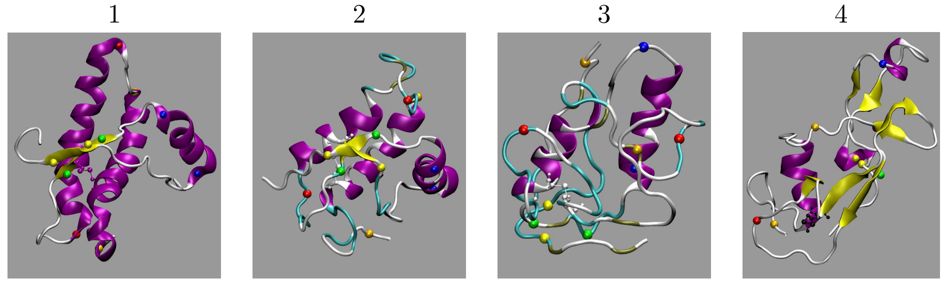


Figure 2.3: *Examples of different structural groups that were identified in the simulation and for which potential energy averages were computed. Panels 1 and 2 illustrate native structures: Panel 1; the NMR structure and panel 2; a partially unfolded structure with a fraction of native contacts* $\sim$ *64% placing it close to the limit of our* native - fraction of native contacts *group. Panels 3 and 4 show non-native structures: Panel 3; a* collapsed *structure (radius of gyration* $\leq$ *1.3 nm, non native) and panel 4; a structure that does not belong to any of these groups (group* other*). Helices are colored in purple,* β*-sheets in yellow.*

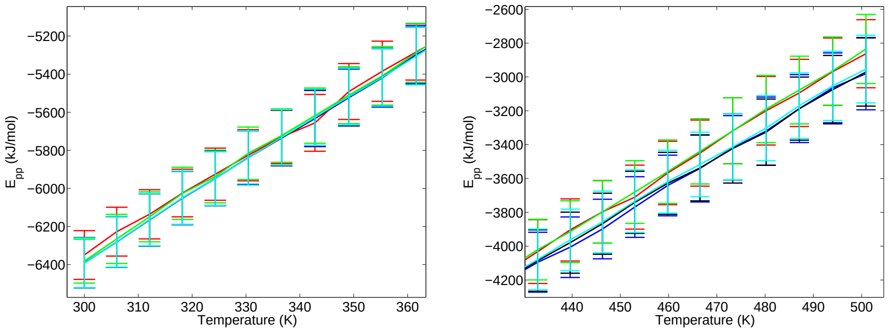


Figure 2.4: E_{pp} *distributions obtained in the first 22.1 ns (first fourth) of the simulation. The distributions are shown with averages (colored lines) and errorbars (+/- one standard deviation) for different structural groups at the first (left panel) and last (right panel) eleven temperatures. The structural groups were defined as follows: (i)* Native - RMSD, *structures that have an RMSD of* $\leq$ *0.44 nm of the NMR structure (dark blue) (ii)* Native - fraction of native contacts, *structures that have* $\geq$ *64% of the contacts found in the NMR structure (cyan), (iii)* Native - high secondary structure content, *structures that have at least 38* α*-helical and 4* β*-sheet residues (black) (iv)* Collapsed, *structures that have a radius of gyration* $\leq$ *1.3 nm and that do not belong to any of the native groups (red) (v)* Other, *structures that do not belong to any of the first four groups (green).*

structure corresponding to the energy minimum in both cases. Explicit solvent REMD simulations are problematic for large systems. Therefore, alternative REMD protocols are likely to become a necessity for larger, biomolecular simulations. In the present work, we show that solute energy based exchange probabilities do not introduce any obvious errors or biases for all the tested properties, and that this result is already achieved with a zero

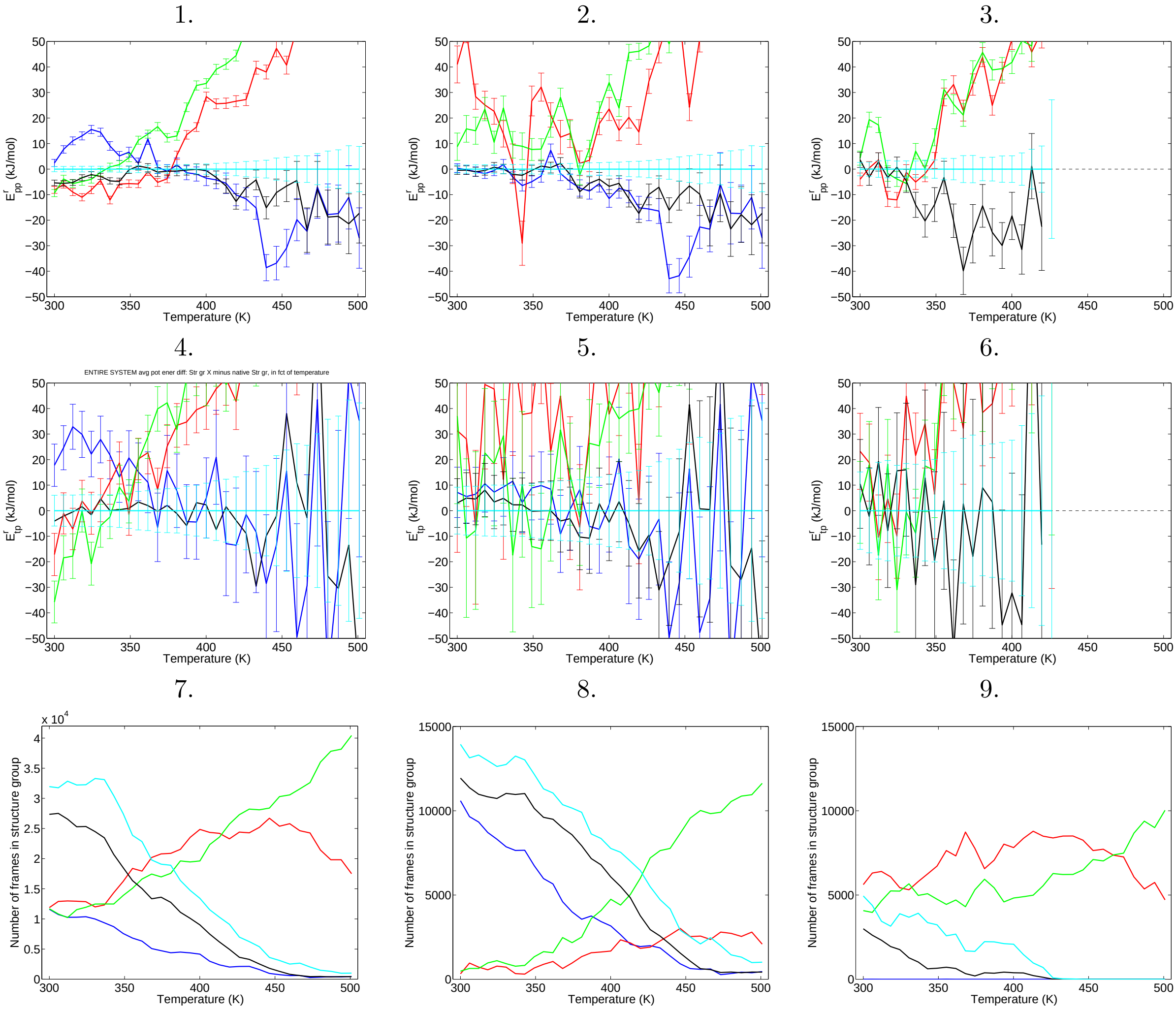


Figure 2.5: *E^r_{xp} (with $x = p$ or t, for protein or total, and p = potential energy), defined as the difference between the average E_{xp} of a given structural group and the average E_{xp} of the native group (fraction of native contacts definition). The colors of the lines refer to the different structural groups as in Figure 2.4. Panels 1, 4 and 7 present averages over the entire simulation (88.4 ns), while panels 2, 5 and 8 present averages for the first fourth (first 22.1 ns), and panels 3, 6 and 9 averages of the last fourth (last 22.1 ns). Panels 7, 8 and 9 show the population sizes of the different structural groups in the different simulation intervals. The E^r_{pp} shown in Panel 2 is related to Figure 2.4.*

order approximation that totally neglects solvent-solvent interactions.

2.6 References

[1] Y. Duan and P. A. Kollman. “Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution.” *Science*, **282**, (1998) 740–744.

[2] W. F. van Gunsteren, R. Buergi, C. Peter, and X. Daura. “The Key to Solving the Protein-Folding Problem Lies in an Accurate Description of the Denatured State”. *Angew Chem Int Ed Engl*, **40**, (2001) 351–355.

[3] S. Jang, E. Kim, S. Shin, and Y. Pak. “Ab initio folding of helix bundle proteins using molecular dynamics simulations.” *J Am Chem Soc*, **125**, (2003) 14 841–14 846.

[4] M. M. Seibert, A. Patriksson, B. Hess, and D. van der Spoel. “Reproducible polypeptide folding and structure prediction using molecular dynamics simulations.” *J Mol Biol*, **354**, (2005) 173–183.

[5] J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen. “Designing a 20-residue protein.” *Nat Struct Biol*, **9**, (2002) 425–430.

[6] J. vandeVondele and U. Rothlisberger. “Canonical adiabatic free energy sampling (cafes): A novel method for the exploration of free energy surfaces”. *J Phys Chem B*, **106**, (2002) 203–208.

[7] M. Iannuzzi, A. Laio, and M. Parrinello. “Efficient exploration of reactive potential energy surfaces using car-parrinello molecular dynamics.” *Phys Rev Lett*, **90**, (2003) 238 302.

[8] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. “Optimization by simulated annealing”. *Science*, **220**, (1983) 671–680.

[9] W. Wenzel and K. Hamacher. “Stochastic tunneling approach for global minimization of complex potential energy landscapes”. *Phys Rev Lett*, **82**, (1999) 3003–3007.

[10] Grubmüller. “Predicting slow structural transitions in macromolecular systems: Conformational flooding.” *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, **52**, (1995) 2893–2906.

[11] Y. Okamoto. “Generalized-ensemble algorithms: enhanced sampling techniques for monte carlo and molecular dynamics simulations.” *J Mol Graph Model*, **22**, (2004) 425–439.

[12] Y. Sugita and Y. Okamoto. “Replica-exchange Molecular Dynamics Method for Protein Folding”. *Chem. Phys. Lett.*, **314**, (1999) 141–151.

[13] H. Fukunishi, O. Watanabe, and S. Takada. “On the hamiltonian replica-exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction”. *J Chem Phys*, **116**, (2002) 9058–9067.

[14] W. Li, J. Zhang, and W. Wang. "Understanding the folding and stability of a zinc finger-based full sequence design protein with replica exchange molecular dynamics simulations." *Proteins*, **67**, (2007) 338–349.

[15] A. D. Simone, A. Zagari, and P. Derreumaux. "Structural and hydration properties of the partially unfolded states of the prion protein." *Biophys J*.

[16] R. Affentranger, I. Tavernelli, and E. E. di Iorio. "A novel hamiltonian replica exchange md protocol to enhance protein conformational sampling". *J. Chem. Theory Comput.*, **2**, (2006) 217–228.

[17] S. Jang, S. Shin, and Y. Pak. "Replica-exchange method using the generalized effective potential." *Phys Rev Lett*, **91**, (2003) 058 305.

[18] P. Liu, B. Kim, R. A. Friesner, and B. J. Berne. "Replica exchange with solute tempering: a method for sampling biological systems in explicit water." *Proc Natl Acad Sci U S A*, **102**, (2005) 13 749–13 754.

[19] X. Huang, M. Hagen, B. Kim, R. A. Friesner, R. Zhou, and B. J. Berne. "Replica exchange with solute tempering: efficiency in large scale systems." *J Phys Chem B*, **111**, (2007) 5405–5410.

[20] X. Cheng, G. Cui, V. Hornak, and C. Simmerling. "Modified replica exchange simulation methods for local structure refinement." *J Phys Chem B Condens Matter Mater Surf Interfaces Biophys*, **109**, (2005) 8220–8230.

[21] A. Okur, L. Wickstrom, M. Layten, R. Geney, K. Song, V. Hornak, and C. Simmerling. "Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/Implicit Solvation Model". *J. Chem. Theory Comput.*, **2**, (2006) 420–433.

[22] Y. Mu, Y. Yang, and W. Xu. "Hybrid hamiltonian replica exchange molecular dynamics simulation method employing the poisson-boltzmann model." *J Chem Phys*, **127**, (2007) 084 119.

[23] X. Periole and A. E. Mark. "Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent." *J Chem Phys*, **126**, (2007) 014 903.

[24] S. Prusiner. "Novel proteinaceous infectious particle causes scrapie". *Science*, **216**, (1982) 136–144.

[25] B. Oesch, D. Westaway, M. Wälchli, M. P. McKinley, S. B. Kent, R. Aebersold, R. A. Barry, P. Tempst, D. B. Teplow, and L. E. Hood. "A cellular gene encodes scrapie prp 27-30 protein." *Cell*, **40**, (1985) 735–746.

[26] B. Chesebro, R. Race, K. Wehrly, J. Nishio, M. Bloom, D. Lechner, S. Bergstrom, K. Robbins, L. Mayer, and J. M. Keith. "Identification of scrapie prion protein-specific mrna in scrapie-infected and uninfected brain." *Nature*, **315**, (1985) 331–333.

[27] K. Basler, B. Oesch, M. Scott, D. Westaway, M. Wälchli, D. F. Groth, M. P. McKinley, S. B. Prusiner, and C. Weissmann. “Scrapie and cellular prp isoforms are encoded by the same chromosomal gene.” *Cell*, **46**, (1986) 417–428.

[28] K. M. Pan, M. Baldwin, J. Nguyen, M. Gasset, A. Serban, D. Groth, I. Mehlhorn, Z. Huang, R. J. Fletterick, and F. E. Cohen. “Conversion of alpha-helices into beta-sheets features in the formation of the scrapie prion proteins.” *Proc Natl Acad Sci U S A*, **90**, (1993) 10 962–10 966.

[29] I. V. Baskakov, G. Legname, M. A. Baldwin, S. B. Prusiner, and F. E. Cohen. “Pathway complexity of prion protein assembly into amyloid.” *J Biol Chem*, **277**, (2002) 21 140–21 148.

[30] I. V. Baskakov, G. Legname, Z. Gryczynski, and S. B. Prusiner. “The peculiar nature of unfolding of the human prion protein.” *Protein Sci*, **13**, (2004) 586–595.

[31] O. V. Bocharova, L. Breydo, A. S. Parfenov, V. V. Salnikov, and I. V. Baskakov. “In vitro conversion of full-length mammalian prion protein produces amyloid form with physical properties of PrP(Sc).” *J Mol Biol*, **346**, (2005) 645–659.

[32] A. Tahiri-Alaoui and W. James. “Rapid formation of amyloid from alpha-monomeric recombinant human PrP in vitro.” *Protein Sci*, **14**, (2005) 942–947.

[33] L. Redecke, M. von Bergen, J. Clos, P. V. Konarev, D. I. Svergun, U. E. A. Fittschen, J. A. C. Broekaert, O. Bruns, D. Georgieva, E. Mandelkow, N. Genov, and C. Betzel. “Structural characterization of beta-sheeted oligomers formed on the pathway of oxidative prion protein aggregation in vitro.” *J Struct Biol*, **157**, (2007) 308–320.

[34] K.-W. Leffers, H. Wille, J. Stöhr, E. Junger, S. B. Prusiner, and D. Riesner. “Assembly of natural and recombinant prion protein into fibrils.” *Biol Chem*, **386**, (2005) 569–580.

[35] F. Eghiaian, T. Daubenfeld, Y. Quenet, M. van Audenhaege, A.-P. Bouin, G. van der Rest, J. Grosclaude, and H. Rezaei. “Diversity in prion protein oligomerization pathways results from domain expansion as revealed by hydrogen/deuterium exchange and disulfide linkage.” *Proc Natl Acad Sci U S A*, **104**, (2007) 7414–7419.

[36] K. Kuwata, H. Li, H. Yamada, G. Legname, S. B. Prusiner, K. Akasaka, and T. L. James. “Locally disordered conformer of the hamster prion protein: a crucial intermediate to prpsc?” *Biochemistry*, **41**, (2002) 12 277–12 283.

[37] M. L. DeMarco and V. Daggett. “Molecular mechanism for low ph triggered misfolding of the human prion protein.” *Biochemistry*, **46**, (2007) 3045–3054.

[38] M. S. Shamsir and A. R. Dalby. “One gene, two diseases and three conformations: molecular dynamics simulations of mutants of human prion protein at room temperature and elevated temperatures.” *Proteins*, **59**, (2005) 275–290.

[39] R. I. Dima and D. Thirumalai. “Probing the instabilities in the dynamics of helical fragments from mouse prpc.” *Proc Natl Acad Sci U S A*, **101**, (2004) 15 335–15 340.

[40] D. V. D. Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. “GROMACS: fast, flexible, and free.” *J Comput Chem*, **26**, (2005) 1701–1718.

[41] W. van Gunstern, S. Billeter, A. Eising, P. Huenenberger, P. Krueger, A. Mark, W. Scott, and I. Tironi. *Biomolecular Simulation: The GROMOS96 Manual and User Guide* (Hochschulverlag AG, Zuerich, 1996).

[42] H. Berendsen, J. Postma, W. van Gunsteren, and J. Hermans. *Intermolecular Forces* (Reidel, Dordrecht, 1981).

[43] J. P. Ryckaert, G. Ciccotti, and H. Berendsen. “Numerical integration of cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes”. *J Comp Phys*, **23**, (1977) 327–341.

[44] R. Riek, S. Hornemann, G. Wider, M. Billeter, R. Glockshuber, and K. Wüthrich. “NMR structure of the mouse prion protein domain PrP(121-321).” *Nature*, **382**, (1996) 180–182. Exp, nmr.

[45] J. Antosiewicz, J. A. McCammon, and M. K. Gilson. “Prediction of pH-dependent properties of proteins.” *J Mol Biol*, **238**, (1994) 415–436.

[46] M. Davis and J. McCammon. “Electrostatics in biomolecular structure and dynamics”. *Chem Rev*, **90**, (1990) 509–521.

[47] W. Swietnicki, R. Petersen, P. Gambetti, and W. K. Surewicz. “ph-dependent stability and conformation of the recombinant human prion protein prp(90-231).” *J Biol Chem*, **272**, (1997) 27 517–27 520.

[48] A. Yang, M. Gunner, R. Sampogna, R. Sharp, and B. Honig. “On the calculation of pKa in proteins”. *Proteins*, **15**, (1993) 252–256.

[49] G. Vriend. “WHAT IF: a molecular modeling and drug design program.” *J Mol Graph*, **8**, (1990) 52–6, 29.

[50] S. Nosé. “A unified formulation of the constant temperature molecular dynamics methods”. *J. Chem. Phys.*, **81**, (1984) 511–519.

[51] W. Hoover. “Canonical dynamics-equilibrium phase space distribution”. *Phys. Rev. A.*, **31**, (1985) 1695–1697.

[52] I. G. Tironi, R. Sperb, P. E. Smith, and W. F. van Gunsteren. “A generalized reaction field method for molecular dynamics simulations”. *J Chem Phys*, **102**, (1995) 5451–5459.

[53] H. Berendsen, J. Postma, W. van Gunsteren, A. Dinola, and J. Haak. “Molecular dynamics with coupling to an external bath”. *J Chem Phys*, **81**, (1984) 3684–3690.

[54] K. Y. Sanbonmatsu and A. E. García. “Structure of met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics.” *Proteins*, **46**, (2002) 225–234.

[55] D. Sindhikara, Y. Meng, and A. E. Roitberg. “Exchange frequency in replica exchange molecular dynamics.” *J Chem Phys*, **128**, (2008) 024 103.

[56] H. Savage, C. Elliot, C. Freeman, and J. Finney. “Lost hydrogen-bonds and buried surface-area: Rationalizing stability in globular-proteins”. *J Chem Soc Faraday Trans*, **89**, (1993) 2609–2617.

[57] H. B. Callen. *Thermodynamics* (John Wiley, 1960).

Chapter 3

Exploring protein conformational space with replica exchange molecular dynamics: possibilities & limitations

Abstract

Replica-exchange molecular dynamics (REMD) has become a very powerful tool to perform enhanced sampling of protein conformational space. However, applications to larger proteins (> 30 aa) remain rare, because the intrinsically low exchange probabilities prevent efficient diffusion of replicas in temperature space. In the present work, we apply a variant REMD technique to the study of the 103-residue prion protein to test the possibilities and limitations of this method for the extensive sampling of systems with complex conformational landscapes. As expected, the sampled conformational space is substantially increased. However, comparison with straightforward reference simulations at 300, 400 and 500K shows that the free energy landscape in the REMD simulations is severely distorted. Although native-like structures have the lowest potential energy, new non-native conformations, obtained at high temperature, frequently exchange to low temperatures. Thus, the low temperature population is progressively shifted towards non-native portions of the free energy surface. The origins of this effect can be traced back to a slow regeneration of native structures at low temperatures, which cannot compete with the fast high temperature unfolding dynamics. This bias is induced by the use of typical inter-exchange times of 1-10 ps that turn out to be much shorter than the intrinsic characteristic structural decorrelation times of the system that are of the order of 100 ns. Therefore, unbiased REMD studies for large systems are likely to require unfeasibly long inter-exchange times.

3.1 Introduction

Replica-exchange molecular dynamics (REMD) has become a very powerful method in the mechanistic study of peptide and small protein folding. In fact, in a number of cases, REMD techniques have allowed to identify the native fold as the lowest free energy minimum [1, 2, 3, 4]. This result is also remarkable because the free energy difference between folded and misfolded, denatured or random-coil structures is often very small and can possibly lie beyond the accuracy of a force field.

In REMD, N replicas, or copies of a given molecular system, undergo parallel MD simulations at different temperatures. These temperatures T_1, T_2,...,T_{n-1}, T_n cover a given range, in which the lowest temperature is representative of a “ground state“ (e.g. the temperature at which the force field was parameterized) and the higher temperatures allow to escape from energetic minima. At given times, all the MD simulations are stopped and exchanges are attempted among adjacent pairs of replicas that are simulated at neighboring temperatures T_n and T_{n+1}. Exchange attempts consist in calculating a Monte-Carlo exchange probability. This probability is compared to a random number, and if the former is greater than the latter the exchange is accepted. When an exchange is made, the conformation that is at T_n is assigned to the new temperature T_{n+1}, and the one at T_{n+1} to T_n. Once exchanges have been attempted on all replica pairs and the respective temperature re-assignments have been performed, where relevant, the parallel MD simulations are restarted.

For canonical ensembles (NVT), the exchange probability is constructed so that exchanges generate Boltzmann distributions. The joint probability distribution of an extended system composed of multiple copies of an original system can be written under the assumption that each replica samples a Boltzmann distribution at its given temperature. The exchange probability P_{ex} is obtained by solving the equations describing this joint probability [5]:

$$P_{ex} \quad = \quad min\left(1, exp\left\{\left[\frac{1}{k_B T_n} - \frac{1}{k_B T_{n+1}}\right]\left[E(x^i_{T_n}) - E(x^j_{T_{n+1}})\right]\right\}\right) \tag{3.1}$$

where k_B is Boltzmann's constant, T_n and T_{n+1} two adjacent temperatures of the input temperature range, $E(x^j_{T_{n+1}})$ and $E(x^i_{T_n})$ the potential energies of replicas i and j, with the corresponding sets of coordinates $x^j_{T_{n+1}}$ and $x^i_{T_n}$. The potential energy difference in the second term shows that high exchange probabilities require the potential energy of a replica at T_{n+1} to be low enough to be part of the overlap of potential energy distributions at T_n and T_{n+1}. The exchange is then accepted, and the replica exchanged to T_n, where it becomes a member of the canonical distribution at T_n. Thus, this probability function builds up a canonical distribution at each temperature, and in particular at the minimal temperature. This scheme only holds if replicas fully decorrelate between exchange attempts (Markov chain), i.e. if the MD time between two exchange attempts allows for the full decorelation of an exchanged replica.

Ideally, a replica should cross the full temperature range several times during an REMD simulation. Every exchange allows to change temperature, and potentially escape a local minimum and sample new conformations. When large systems (i.e. proteins solvated in an explicit solvent) are simulated, the number of degrees of freedom is very high, and adjacent temperatures have to be very close in order to ensure a sufficient overlap of potential energy distributions. Many replicas and an increased computational power are thus required, making the method unpractical. Recently, alternative, approximative methods have been developed to decrease the number of replicas needed to cover a given temperature range [6, 7, 8, 9, 10, 11, 12]. Although these variant protocols are very promising in extrapolating the benefits of REMD to larger proteins, such studies have remained rare. Indeed, most of the studies have aimed towards the validation of these variant protocols on smaller systems.

In the present study, we probe the efficiency of REMD for larger proteins using the *REMDpe* variant protocol (Chapter 2). In *REMDpe*, exchange probabilities are computed with a partial potential energy function E_{pp}, that only retains potential energy terms due to protein-protein and protein-solvent interactions:

$$E_{pp} = E^{bonded}_{solute} + E^{non-bonded}_{solute} + \frac{1}{2}E^{non-bonded}_{solute-solvent} \tag{3.2}$$

where $E^{non-bonded}_{solute-solvent}$ is multiplied by $\frac{1}{2}$ in order to give more weight to protein energy terms. This partial potential energy function focuses on the conformation of the solute, discarding solvent degrees of freedom (that often occupy the largest part of the sampling time), while the total potential energy E_{tp} continues driving the MD part.

In Chapter 2, we have validated the *REMDpe* protocol for a prion protein (PrP) simulation showing that (i) exchanges were being performed according to a Boltzmann weight, (ii) *REMDpe* 300K and 500K trajectories yielded the same potential energy distributions (for E_{tp} as well as for E_{pp}) as reference MD simulations at the same temperatures and (iii) different protein structural groups (native, non- native, collapsed and other structures) were ranked in the same way according to their decreasing average E_{pp} and E_{tp}, showing that the native structure is the minimum for both E_{pp} and E_{tp}. These three tests thus suggested that REMD exchange probabilities computed with E_{tp} should give comparable results to the ones of *REMDpe* computed with the E_{pp} and no obvious bias is introduced by the approximate protocol.

Here, we use a 88.4 ns, 32 replica (total aggregate time of 2.8 μs) *REMDpe* simulation of the prion protein to test the efficiency of the method in characterizing possible misfolded conformations. Compared to a 315 ns long 300K MD reference trajectory, the sampling is clearly enhanced in the 300K *REMDpe* ensemble. However, we also show that, although the simulation was started with equilibrated native structures for all replicas, the native fold is progressively lost even at the lowest temperature simulations. The main competitive misfold is identified in collapsed structures, which are formed at high temperature and progressively flood the low temperature ensembles. Comparing different population histograms computed as a function of the fraction of native contacts and the radius of gyration, we can show that a quasi-barrierless diffusion traps all the high temperature trajectories into a deep, non-native minimum. When structures of this minimum exchange back to low temperatures, the regeneration of native contacts is in principle possible, but depends on the crossing of energetic barriers that would require much longer low temperature residence times than the typical applied inter-exchange intervals of the order of a few ps. The global picture that emerges from this study is that the sampling is clearly enhanced, but that inter-exchange times that are inferior to the characteristic structural decorrelation times of the system progressively bias the conformational landscape towards high temperature like non-native populations, eventually decreasing the native population of the 300K ensemble to essentially zero.

3.2 Methods

All molecular dynamics (MD) simulations were performed with the GROMACS-3.3.0 package [13], the GROMOS96 force-field [14], the SPC water model [15] and an MD time step of 1.5 fs with constraints on covalent bonds involving hydrogen atoms applied via the SHAKE algorithm [16] with a tolerance of 10^{-4} kJ mol^{-1} nm^{-1}. Temperture was controlled with two Nosé-Hoover thermostats (one for the protein and one for solvent and counter ions, with respective time-coupling constants of 0.4 and 1.6 ps) [17, 18]). Coulombic interactions were treated using a twin-range cutoff, in which interactions within 1.0 nm and between 1.0 and 1.4 nm were computed every MD step, respectively every 5 MD steps.

Electrostatic interactions beyond 1.4 nm were approximated with a generalized reaction field [19] generated by a dielectric continuum with dielectric constant of 66. The starting conformation was the mouse PrP NMR structure (res 124-226, PDB code 1AG2 [20]). The protonation states of ionizable side chains were predicted by finite-difference Poisson-Boltzmann calculations [21, 22] in order to mimic a pH 4 environment that is known to favor conformational changes [23]. The DELPHI program [24] supplied with the WHATIF package [25] was used to solve the Poisson-Boltzmann equation. The protonation states of HIS, ASP and GLU side chains were consequently set as follows: HIS17; p, ASP21; d, GLU23; d, ASP24; d, GLU29; p, ASP44; p, HIS54; p, ASP55; d, HIS64; p, GLU73; d, GLU77; p, ASP79; d, GLU84; p, GLU88; p and GLU98; p (where p stands for protonated and d for deprotonated). A rhombic dodecahedron box with 14076 water molecules was constructed around the protein. 8 Cl^- ions were added to neutralize protein charges.

The equilibration of the NMR structure and of the *REMDpe* replicas were performed as described in Chapter 2. The *REMDpe* production run was performed in the NVT ensemble with the following temperature distribution: 300.0, 306.0, 312.1, 318.2, 324.3, 330.4, 336.6, 342.8, 349.1, 355.3, 361.6, 368.0, 374.3, 380.7, 387.1, 393.6, 400.1, 406.6, 413.1, 419.7, 426.3, 432.9, 439.6, 446.3, 453.0, 459.8, 466.5, 473.3, 480.2, 487.1, 494.0 and 500.9K. With this distribution, we obtained roughly homogenous *REMDpe* exchange frequencies of $\sim$30%. The *REMDpe* production run was carried out for 88.4 ns (total aggregate time of 2.8 μs), with exchange attempts performed every 60 ps on adjacent temperature replicas n and n+1 for odd exchange trial numbers, and 2n and 2n+1 for even ones. Structures, energies and temperatures were saved every 1.5 ps. Three MD reference simulations were performed at 300, 400 and 500K with respective simulation times of 314.8, 346.5 and 64 ns. For the two high temperature simulations (400 and 500K), the temperature was first raised to the target temperature with 300 ps of temperature rescaling.

Secondary structure elements were computed with the DSSP [26] algorithm in the GROMACS-3.3.0 interface. An in-house program was used to compute contact maps and fractions of native contacts (Q_{fr}). Two residues were considered in contact if the shortest distance between two atoms of these residues was inferior to 0.45 nm, and the fraction of native contacts of a given conformation was defined as the percentage of contacts of the NMR structure that were still present. All the other analysis were performed with GROMACS-3.3.0 [13] routines.

Free energies G' (with respect to an arbitrairily chosen zero of energy) were computed from probability distributions projected on the fraction of native contacts (Q_{fr}) and the radius of gyration (R_g):

$$\text{G'}_{Q_{fr},R_g} = -k_B T \ln\left(\frac{N(Q_{fr}, R_g)}{M}\right) \tag{3.3}$$

where $N(Q_{fr}, R_g)$ is the number of conformations in the interval $[Q_{fr} - \delta Q_{fr}, Q_{fr} + \delta Q_{fr}]$ and $[R_g - \delta R_g, R_g + \delta R_g]$ and M is the total number of structures.

In order to assess structural decorrelation times, autocorrelation functions $C(t)$ were computed for chosen properties $x(t)$ along the reference trajectories:

$$C(t) = \frac{\sum_{k}^{M}(x(t_k) - \langle x \rangle)(x(t_k + t) - \langle x \rangle)}{\sum_{k}^{M}(x(t_k) - \langle x \rangle)^2} \tag{3.4}$$

where M is the total number of frames in the trajectory, t_k is the time at frame k and $\langle x \rangle$ is the average of the property x over the entire trajectory.

3.3 Results and Discussion

3.3.1 Extent of conformational sampling in the reference MD simulation

A 315 ns MD simulation was performed at 300K starting from the equilibrated NMR structure to provide a reference for the conformational space sampled in a straightforward MD run. In order to investigate the diffusion time of the system in the conformational space we computed the time series of the fraction of native contacts (Q_{fr}) and radius of gyration (R_g)(Figure 3.1, Panel 1). During the first 90 ns of dynamics, the protein remains very close to the starting configuration (equilibrated NMR structure, with $Q_{fr} \sim 0.8$ and $R_g \sim 1.4$). This behavior is consistent with the picture of the system in thermal fluctuation around its (global) free energy minimum. However, around 90 ns, Q_{fr} starts to decrease, in parallel to an increase in R_g (Figure 3.1, Panel 1). This concerted trend carries on for 30 ns, leading to a new structural ensemble characterized by a Q_{fr} of $\sim$ 65% and a R_g of $\sim$ 1.47 nm (Figure 3.6, Panels 2 and 3), in which the simulation stays for the remaining 195 ns.

Analysing the trajectory reveals that the increase in R_g was caused by the unfolding of the second helix (H2) (Figure 3.1, Panel 4). A principle component analysis (PCA) of the trajectory shows that the lowest-frequency mode is indeed related to the unfolding of H2 (Figure 3.1, Panels 2 and 3), which is known to be the least stable of the three helices both from experimental [27, 28, 29, 30] and theoretical [31, 32] studies. Surprisingly, residues of the unfolded H2 form new β-sheets (Figure 3.1, Panel 4). Obtaining native-like structure with alternative secondary structure is a rather unusual outcome for a straightforward low temperature (300K) MD simulation at first sight, and might raise possible concerns about the validity of the underlying force field. However, we would like to stress that no such event was obtained for a 315 ns control simulation performed under the same conditions with doppel protein, a structural homolog of prion with only 25% amino-acid sequence homology, but an identical fold, with three helices and two short β-sheets [33]. This suggests that this change in native structure is most probably due to an intrinsic property of the system. This is also supported by the fact that alternative prion

β-rich folds have indeed been observed at physiological conditions in complete absence of denaturing agents [34, 35, 36]. In view of the fact that this structure is still very close to the native one ($Q_{fr} \sim 65\%$), we will include this second minimum in the native population pool, and set a Q_{fr} threshold of 64% to separate native from non-native structures.

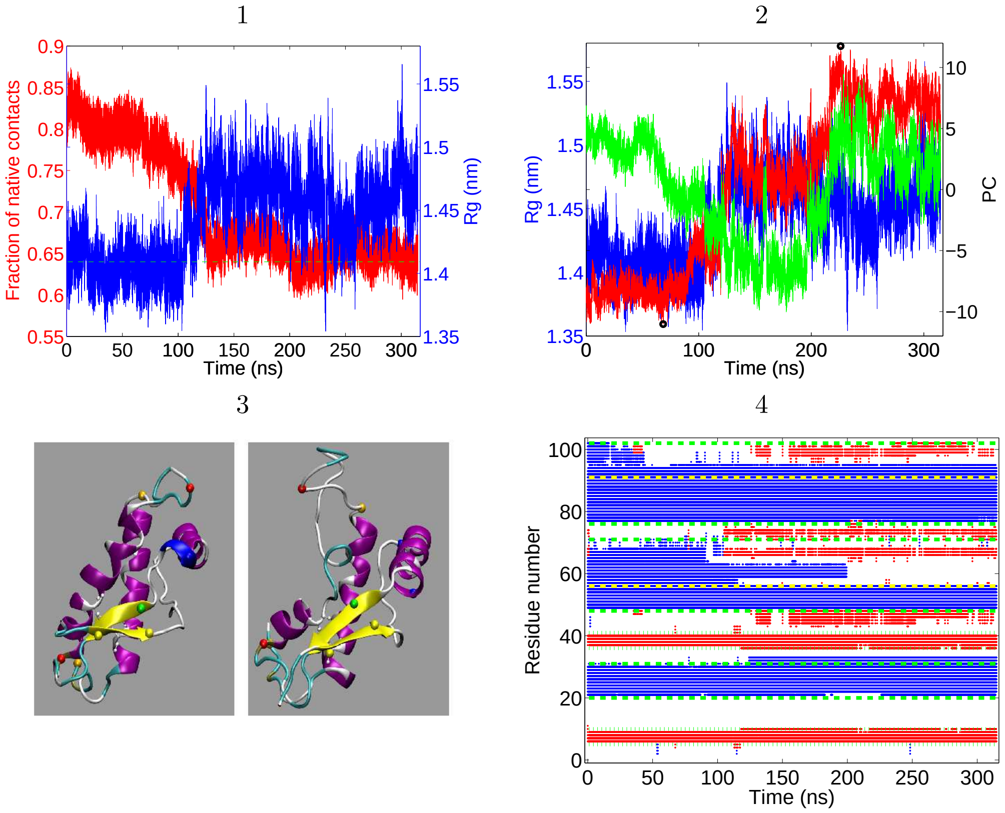


Figure 3.1: *The 300K MD reference simulation. Time series of the radius of gyration (R_g, blue line in panels 1 and 2), of the fraction of native contacts (red, panel 1) and of the first (red) and second (green) principal components (PC, panel 2). In panel 1, the horizontal green dashed line shows our choice of minimal fraction of native contacts defining native structure. The minimum and maximum first PC trajectory frames are highlighted with a black circle in panel 2, and the corresponding structures are respectively shown in the left and right halves of panel 3 (color coding as in Figure 3.4). Panel 4: Secondary structure (DSSP) of all residues as a function of time. Color coding for the secondary structure: Red; β-sheet and blue; α-helix. Native (NMR) secondary structure elements are delimited by dashed and dotted green horizontal lines: Dotted; β-sheets (in sequence order: S1 and S2), dashed; α-helices (in sequence order: H1, H2 and H3). Yellow dashed horizontal lines: Cys forming the disulfide bridge. Residues were numbered starting from the first residue of the NMR PDB file.*

3.3.2 Extent of conformational sampling in the *REMDpe* simulation

The maximal extent of the sampled conformational spaces (projections on Q_{fr} and R_g) of the reference MD 300K and the *REMDpe* 300K simulations are superimposed in Figure 3.6 Panel 2. The native minimum (equilibrated NMR structure) is the common starting point of the two simulations ($Q_{fr} \sim 0.8$), which both evolve to lower Q_{fr} regions of conformational space. REMD clearly enhances the sampling of conformational space, although it cannot access the second minimum of the 300K MD reference simulation described above. However, the *REMDpe* 300K free energy surface (FES) reveals the presence of very similar conformations, characterized by a practically unfolded H2 (Figure 3.4, Panels 2 and 3). Small differences exist mainly in the loop formed by H2 residues, which refolds into tight loops and β-sheets in *REMDpe* and extends into the solvent in the reference MD, explaining the higher R_g values. The same trend is observed for other residues (i.e. the H3 C-terminus) which do not form secondary structure: their packing to the protein core is tighter in *REMDpe*, resulting in a lower R_g. Obtaining a very similar quasi non-native intermediate with the 300K reference MD and *REMDpe* further validates the latter method and highlights prion's intrinsic ability to adopt multiple conformations at room temperature.

The *REMDpe* simulation was initiated with the equilibrated NMR structure at all temperatures (initial native population of 100%). A low temperature interval was arbitrarily defined within the 300 to 336K range (first 7 temperatures). We monitored a decreasing fraction of native structures in this low temperature interval along the entire simulation (Figure 3.2, left panel). From 0 to 28.5 ns of simulation (38 averaging windows of 750 ps and 500 structures each, ending at the first green vertical dashed line in Figure 3.2), the native population decreased from 100% to 50%, and remained around this plateau value of 50% for another 47.25 ns (ending at window 101, second green vertical dashed line in Figure 3.2) before sharply decreasing to a quasi vanishing fraction within the last 12 ns. Thus, the simulation appears to have undergone three different phases: a decrease of native population, a plateau phase characterized by an equilibrium of native and misfolded structures and a terminal phase, characterized by the further decrease and quasi entire depletion of the native population.

This lead us to seek for the predominant competitive misfold, which could be identified as a group of collapsed structures. A “collapsed structure” group was therefore arbitrarily defined by structures with a $R_g \leq 1.3$ nm (R_g native state ~ 1.41 nm) and a $Q_{fr} < 64\%$. Following the time series of the fraction of the total population belonging to this “collapsed structure” group indeed revealed that it increased along time at low temperatures, competitively replacing the native structures (Figure 3.2, right panel). The trend of this increase follows the decrease of the fraction of native structures described above, although the end of the initial sharp increase, as well as the end of the plateau phase occurred a bit earlier (respectively around 17.25 ns and 68.25 ns). In the plateau phase of the simulation, the collapsed structures roughly accounted for 25% of the total configurations and the native structure group, for 50%. In the terminal phase of the simulation, these numbers are reversed. One can thus deduce that the collapsed structure accounted for 50% and 75% of the non-native population (fraction of native contacts of the NMR structure < 64%) in, respectively, the plateau phase and the terminal phase.

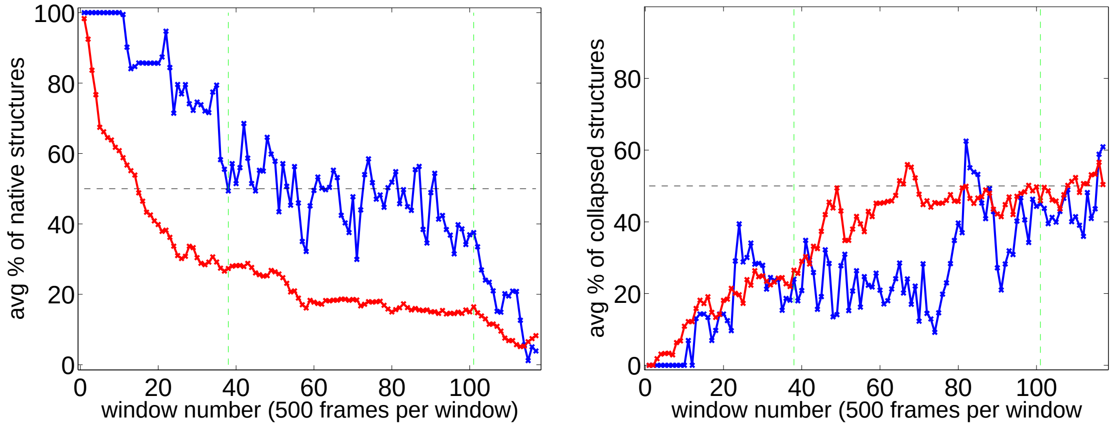


Figure 3.2: *Evolution of the fraction of native structures (left panel) and of the fraction of collapsed structures (right panel) along the 88.4 ns of the simulation. Structures with* $\geq$ *64% of the contacts found in the NMR structure were defined as native, while the remaining non-native structures with a radius of gyration* $\leq$ *1.3 nm were defined as collapsed. Averages were computed for windows of 500 frames (750 ps, structures saved every 1.5 ps) at each temperature, and the plotted fractions were obtained by averaging the averages of windows with the same time index on the following temperature intervals: 300 to 336K (low temperature); dark blue line, 300 to 500K (all temperatures); red line. Dark dashed line; 50% of structures. The green vertical dashed lines delimit a plateau in the evolution of the fraction of native structures at low temperature.*

Thus, the main structural transformations are well described by a correlated decrease of both R_g and Q_{fr}. The conformational landscapes clearly show this correlation (Figure 4.6). At 300K, structures starting from the native basin ($Q_{fr} \sim 0.8$) misfold and collapse, or visit a very rare extended, misfolded conformation ($R_g \sim 1.5$ and $Q_{fr} \sim 0.44$). Conformations belonging to the different high probability regions in the plane spanned by R_g and Q_{fr} are illustrated in Figure 3.4. This misfolding and simultaneous collapse to compact structure is in agreement with the experimental results of Kuwata et al., where high pressure unfolding of prion (monitored by NMR) is concomitant to protein collapse around molecular voids the authors identify within H2 and H3 [27]. Since the *REMDpe* simulations were performed in the NVT ensemble, higher pressures build up at elevated temperatures. For both *REMDpe* and reference MD runs, the pressure was $\sim 1 \pm 200$, 1650 ± 350 and 3635 ± 400 bar at respectively 300, 400 and 500K.

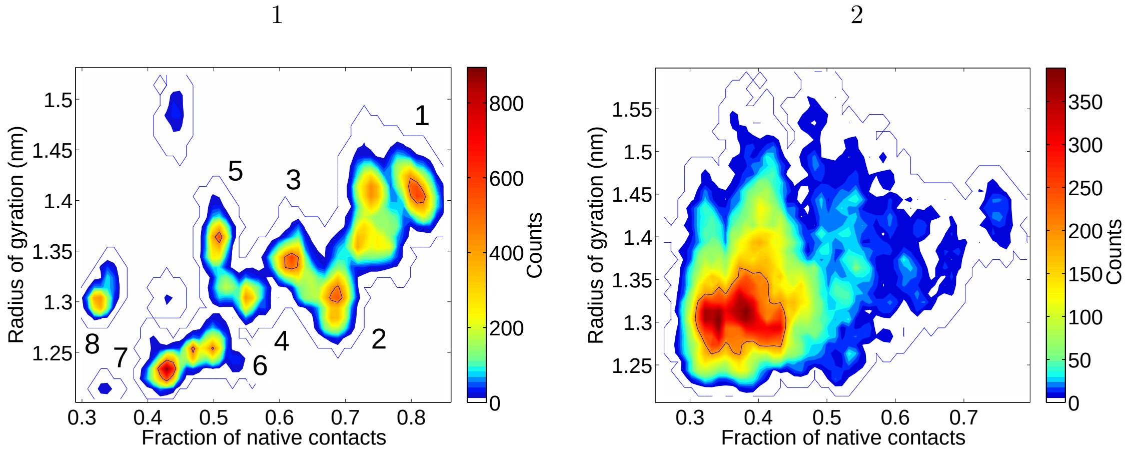


Figure 3.3: *Population histograms computed as a projection on the fraction of native contacts and the radius of gyration. Panel 1:* REMDpe *at 300K (with numbers highlighting high probability regions of which structures are presented in Figure 3.4) and Panel 2;* REMDpe *at 500K. Two solid countour lines are added to Panels 1 and 2: an outer one to encompass the maximal extent of non zero probability regions ($\geq$ 1 count) and an inner one at half of the maximum count observed in order to highlight high probability regions.*

3.3.3 High temperature non-native collapsed structures

Population histograms as a function of R_g are plotted for all temperatures in the right panel of Figure 3.5, highlighting different R_g intervals that delimit high probability regions. These histograms clearly show that structures collapse at higher temperature and are exchanged to the lower ones, replacing native conformations. Correlating the time series of temperatures visited by a replica with its times series of R_g indeed shows that most replicas, independent of their starting temperature, eventually collapse at high temperature and rarely revert to higher R_g values, whichever temperature they visit thereafter (data not shown).

In order to understand the mechanism of collapse taking place at high temperature, we investigated the evolution of R_g along two high temperature reference MD simulations, one at 400K and one at 500K. In both simulations, R_g clearly decreases as a function of time (Figure 3.5, left panel) and this decrease is much faster at 500K: the 400K reference MD had to be extended beyond 300 ns in order to obtain the same collapse. The 500K simulation shows that R_g values as low as 1.27 nm can already be obtained in the first 5 ns, which proves that a replica can indeed "collapse" very quickly, after spending some time in the 400-500K temperature range. These results suggest that a "collapsed" state is a natural outcome for high temperature MD performed with the prion protein and the GROMOS force field in the NVT ensemble.

Thus, obtaining collapsed structures around temperatures from $\geq$ 400K on is an intrinsic feature and not a possible artifact of the *REMDpe* protocol. High temperature MD

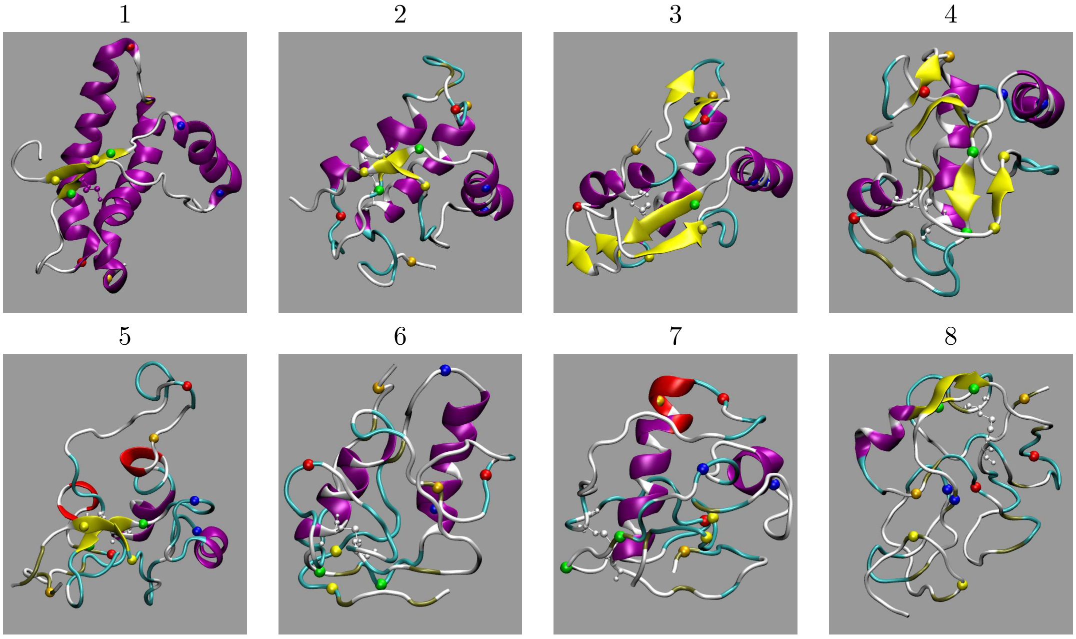


Figure 3.4: *Conformations populating the high probability regions of the 300K* REMDpe *population histogram shown in Figure 4.6 (left panel), where numbers indicating the regions refer to the numbers of the panels showing the corresponding conformations in this figure. Helices are colored in purple, β-sheets in yellow. In order to highlight the sequence-positions of structural re-arrangements, sequence portions spanning NMR secondary structure elements are highlighted with CPK sphere representations of the C-alpha atoms of the residues delimiting β-strands S1 (yellow), S2 (green) and helices H1 (blue), H2 (red) and H3 (orange).*

reference trajectories actually show a transition from a narrow native conformational basin to a broader non-native basin characterized by structural collapse (Figure 3.5 left Panel and Figure 3.6 Panel 3). High temperature *REMDpe* ensembles progressively converge to the same non-native basin (Figure 3.3, right Panel, Figure 3.6 Panels 1 and 4, Figure 3.7 Panel 4) and can only revert to the native one via structure exchanges, the frequency of such reversions diminishing with simulation time. This process resembles a barrierless diffusion, as suggested by Figure 3.6 Panel 4: Replicas moving to higher temperatures eventually end up in the non-native basin.

3.3.4 Too short inter-exchange times cause strong sampling biases

On average, collapsed structures have a higher potential energy than native structures, but the difference between the average values of the corresponding potential energy distributions is very small and the overlap of the fluctuations at all temperatures large (Chapter 2). This enables non-native structures sampled at high temperature to exchange with na-

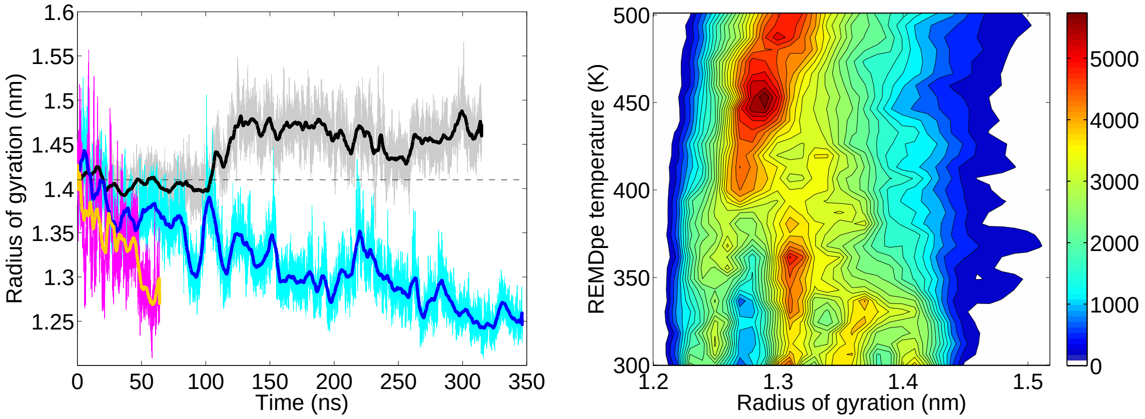


Figure 3.5: *High temperatures lead to collapsed structures. Left panel; Radius of gyration (R_g) time series for the three reference MD simulations. For each simulation, the first line shows the total data and the second one, the running average computed on windows of 500 structures: Grey and black; 300K, cyan and dark blue; 400K, magenta and orange; 500K. Right panel;* REMDpe *Population histograms as a function of R_g for all temperatures, with colors coding for the number of structures (according to the colorbar).*

tive structures at lower temperatures. While a correct application of the REMD protocol with sufficiently long decorrelation times between exchanges will guarantee a Boltzmann sampling at all temperatures, commonly used short inter-exchange times of $\sim$ 1-10 ps (60 ps in the present study) can lead to a strongly biased sampling at low temperature. In particular, because of the short inter-exchange time ($<<$ relaxation time at low temperature), structures from the high temperature ($\sim$ 450K) pool characterized by a R_g of $\sim$ 1.28 nm in Figure 3.5 remain trapped in regions of the phase space characterized by small values of R_g even at low temperatures. The populations around R_g of 1.25, 1.3 and 1.35 nm obtained at 300K are obtained by a fast succession of exchanges from high temperature and cannot diffuse back to the basin of the native fold ($\sim$ 1.4 nm) because of the high free energy barriers. Therefore, these populations are clearly overweighted.

Monitoring the time series of the fraction of native contacts per replica shows that recovery of native contacts from non-native states is rare at all temperatures (data not shown). The FES sampled at high temperature shows a smooth, barrierless character leading to a deep non-native minimum that promotes an irreversible fast trapping of all trajectories (unfolding rate k_u). Consequently, the process of unfolding/misfolding is particularly favored both thermally and kinetically at high temperature. For the replicas that reach the low temperature range, the FES appears too rugged, and the barriers too high to allow for refolding to the native structure with a much smaller rate (i.e. $k_f << k_u$). With a 60 ps interval between exchanges, the average MD time at the lowest temperature is far from sufficient to allow for such refolding. Consequently, our low temperature ensemble slowly shifts towards low Q_{fr}/low R_g value regions. FES locations of the initial, intermediate and final population of $REMDpe$ 300K (Figure 3.7 Panels 1 to 3) reveal this transition. In any 2.5 ns time interval of the last 14.8 ns of the simulation, the 300K ensemble explores the entire final low Q_{fr}/low R_g conformational landscape presented in Panel 3, suggesting that the simulation has converged to this conformational landscape. This conclusion goes beyond our particular application to the prion protein and is very likely to apply to any high temperature non-native minimum and REMD setup of a relatively large protein with inter-exchange MD intervals that are much shorter than characteristic decorrelation times.

In order to compare the rugged nature of our low temperature FES to the smoother ones at high temperatures, we have also computed the diffusion coefficient for both 300K and 500K reference MD simulations (which do not suffer from the bias of structure exchanges). The diffusive harmonic model has been applied to characterize funnel-like FES obtained from protein simulations [37]. From the times series of Q_{fr}, we obtained the autocorrelation time τ_{corr} computed from an exponential fit (the autocorrelation time decays exponentially at long times) of the autocorrelation function (Methods, equation 3.4) and the mean square instantaneous fluctuation σ^2. The model gives us $D(T) = \sigma(T)^2/\tau_{corr}(T)$ for a given temperature T. With $D(300K) \sim 1.8\times10^{-4}\ {Q_{fr}}^2/ns$ and $D(500K) \sim 1.5\times10^{-3}$ ${Q_{fr}}^2/ns$, the 300K FES indeed shows a more rugged nature that compromises fast refolding to native conformations.

Autocorrelation times provide a good measure of structural decorrelation times and were computed for other properties of the same reference MD 300 and 500K trajectories. (Table 3.1). The two structural properties analyzed (R_g and Q_{fr}) show long decorrelation times of 50-100 ns at 300K, which are reduced to $\sim$ 12 ns at 500K, but still way beyond our 60 ps exchange attempt frequency. For comparison, structural decorrelation times of $\sim$ 5 ns were already found for pentapeptides in an implicit solvent [38], corroborating the long decorrelation times we find for the 103-residue prion protein in explicit solvent. This is also demonstrated by the partial potential energy (E_{pp}) that contains all protein-protein and protein-solvent interaction terms. Consistently, its decorrelation time is also long (18 and 4.8 ns at respectively 300 and 500K), reflecting structural decorrelation. At contrast, the total potential energy mainly reflects solvent-solvent interactions, and its decorrelation time is much shorter (maximum of 20 ps at 300K). In practice, one should therefore choose exchange with trial frequencies which are much longer than the 60 ps used in this work (as suggested by Rhee et al., who experimented trial frequencies of 1 ns [39]), but this is not feasible with the present computational resources. To our knowledge, a majority of the REMD studies published so far report exchange-trial frequencies of $\sim$ 1-5 ps and might suffer from the bias reported in the present work.

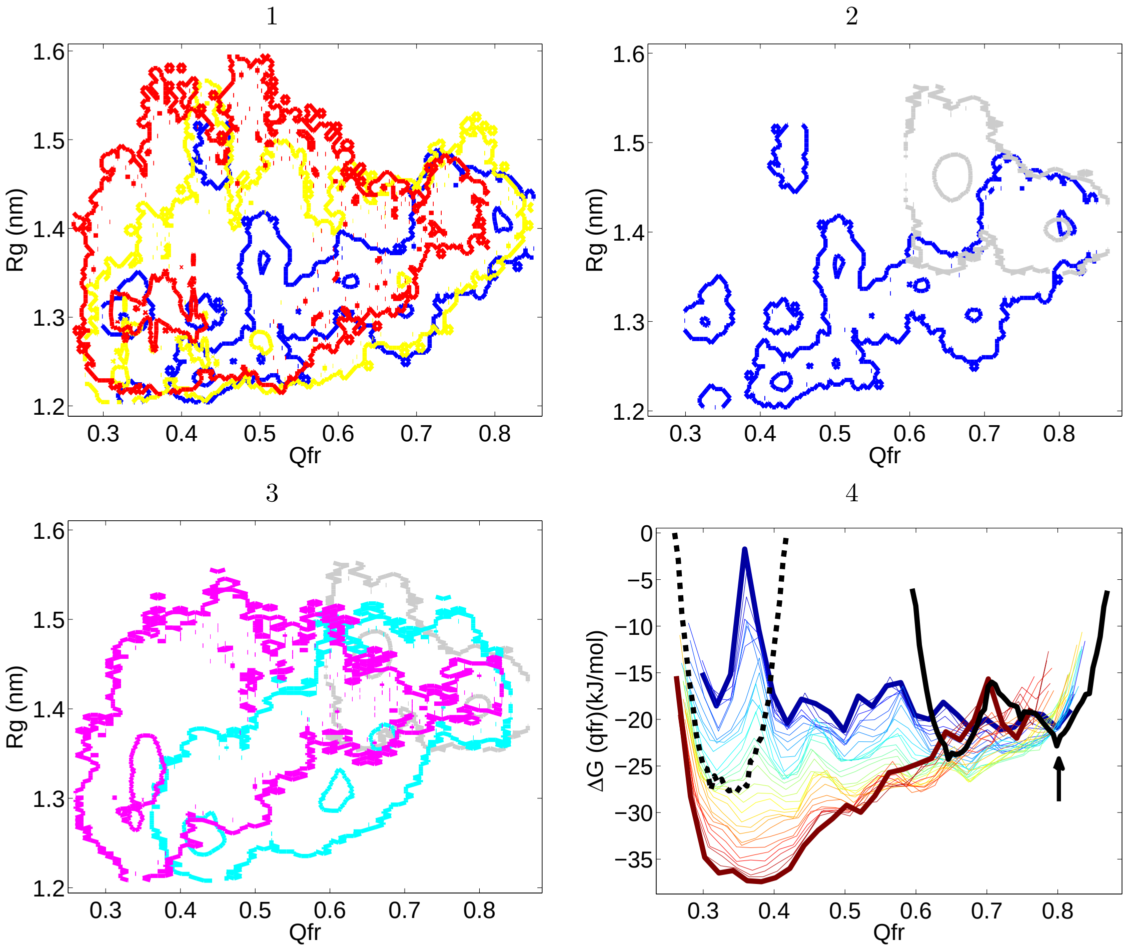


Figure 3.6: *Free energy surfaces (FES) computed as a projection on the fraction of native contacts (Q_{fr}) and the radius of gyration (R_g). The maximal extent (outer contour lines) and minima (inner contour lines drawn at half of the FES minimum) of the following FES are superposed in panels 1, 2 and 3: Panel 1;* REMDpe *300K (blue), 400K (yellow) and 500K (red), Panel 2;* REMDpe *300K (blue) and reference MD 300K (grey) and Panel 3; reference MD 300K (grey), 400K (cyan) and 500K (magenta). FES computed as a projection on the sole Q_{fr} for all* REMDpe *temperatures are shown in Panel 4 , with bold lines highlighting those at 300K (blue) and 500K (red). The corresponding FES of the reference MD simulations are shown with a bold black line: Solid line; 300K (with an arrow locating the NMR structure), dashed line; final minimum of the 500K simulation (48.549 ns to final 64.044 ns).*

Table 3.1: *300K and 500K reference MD trajectory decorrelation times (ns) found for the radius of gyration (R_g), the fraction of native contacts (Q_{fr}), and the total (E_{tp}) and partial (E_{pp}) potential energies. For each property, the autocorrelation function was computed (Methods, equation 3.4), and decorrelation times approximated with the characteristic time of the slowest exponential component of the fit to the autocorrelation function.*

T(K)	R_g	Q_{fr}	E_{tp}	E_{pp}
300	50	97.5	<0.01	18
500	12.5	11.5	<0.01	4.8

One of the largest explicit solvent systems for which "ab initio" folding was achieved with an REMD simulation starting from an extended structure is the Trp-cage, a synthetic, fast folding, 20 amino-acid (aa) miniprotein [4]. Explicit solvent REMD simulations performed with larger proteins are less common, and only provide a detailed description of unfolding or, at best, of partial regeneration of native structure achieved starting from non-native high temperature configurations that still possess native contacts. Recent studies include the Zinc finger artificial construct, 28 aa [40], protein A, 46 aa [41] and another PrP study [32]. None of these simulations exceeded 30 ns per replica, so that it is hard to assess whether they suffer from the bias we describe with longer time scales, and whether this bias also prevents extensive refolding to the native structure.

3.4 Conclusions

In the present work, we have assessed benefits and limitations of REMD applications to relatively large explicit solvent protein systems. Such studies are only possible with variant protocols, such as *REMDpe*, that has been validated in Chapter 2. We show that *REMDpe* clearly enhances the sampled conformational space. However, despite the lower potential energy of the native structure, we have observed its extensive unfolding in the low temperature *REMDpe* ensemble, revealing a strong limitation that is related to the slow regeneration of native contacts at low temperature, which cannot compete with fast and irreversible high temperature unfolding. The origin of this effect are the inter-exchange times that are much shorter than the structural relaxation times. The *REMDpe* simulation eventually leads to the complete unfolding of the entire native structure pool: (i) High temperature ensembles converge to a non-native basin, with unfolding as fatal issue for replicas introduced by exchange, (ii) Exchanges leading to lower temperatures can only select structures from a non-native pool at high temperature that (iii) explore new low temperature non-native minima, progressively shifting the original low temperature native conformational population to non-native portions of the FES. In this last stage, low temperature native contact regeneration is inhibited: At low temperatures, a replica is eventually exchanged to higher temperatures (where unfolding dominates) by a new, non-native, low energy structure before substantial reformation of native contacts can even

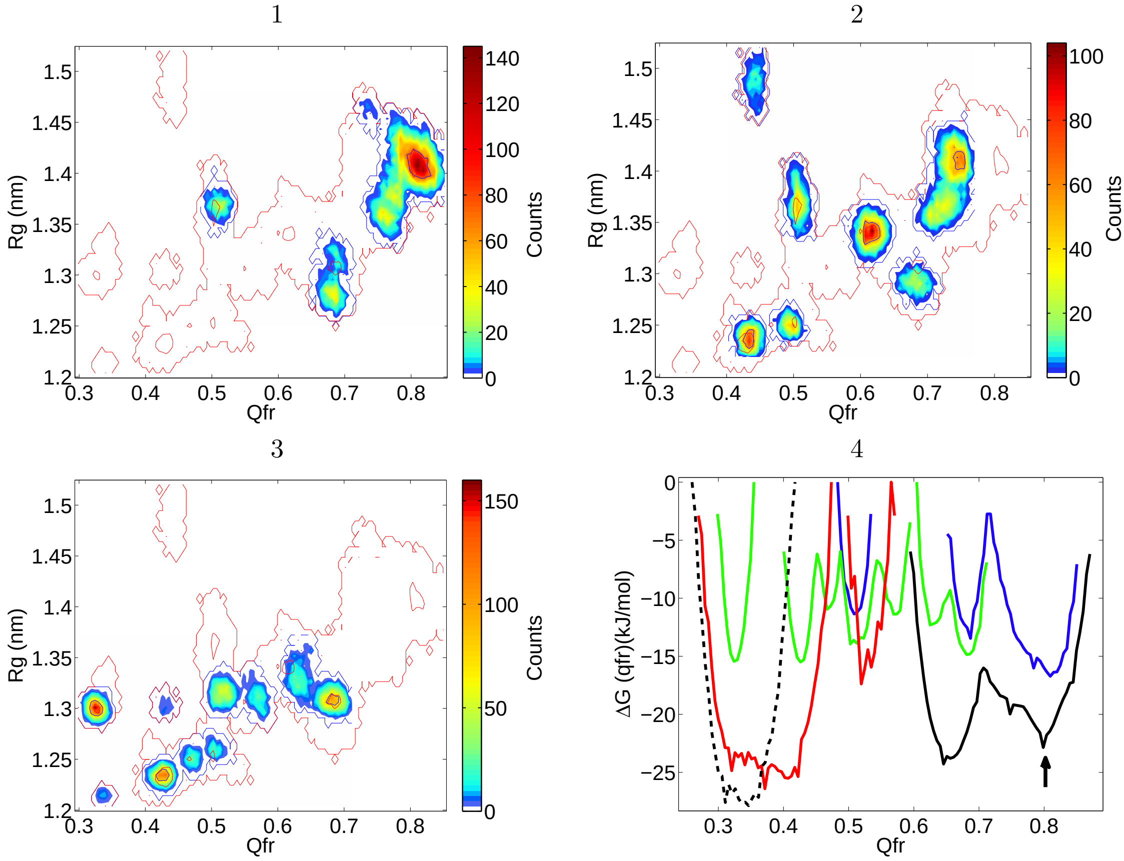


Figure 3.7: *Conformational landscape (projected on the fraction of native contacts Q_{fr} and the radius of gyration R_g) sampled during different intervals of time. Location of* REMDpe *300K populations of the first, third and last sixth (three 14.8 ns intervals) are respectively shown in Panels 1, 2 and 3. Solid contour lines are added, representing these time-dependant conformational landscapes (blue lines) and the total conformational landscape sampled in the 88.4 ns of the simulation (red lines). In both cases, outer and inner contour lines respectively delimit the maximal extent of the conformational landscape and the high probability regions sampled. Panel 4: Location of different populations on the FES computed as a projection on the sole Q_{fr}:* REMDpe *300K first (blue) and last (green) sixth,* REMDpe *500K last sixth (red), reference MD 300K full trajectory (black solid line, with an arrow locating the NMR structure) and reference MD 500K last sixth (black dashed line).*

begin. Refolding to the native structure would require longer low temperature MD times between two subsequent exchanges, of the order of the typical decorrelation times of ∼ 100 ns we found for the prion protein. Indeed, the unbiased character of the Markov chain implied by an REMD process requires this time scale for inter-exchange times. This is much longer than the 60 ps inter-exchange time we use, which is itself already ∼ 10 longer than the most frequent choices used in the literature. This effect is intrinsic to REMD, and is presented as a limitation to be taken into consideration when setting up REMD simulations of large systems. Nevertheless, REMD remains a very powerful tool, allowing to access new portions of the FES, but the relative weight of non-native minima of the FES may be strongly biased by the choice of short inter-exchange times.

3.5 Appendix

In this section, some additional results, supporting the findings of Chapter 3, are presented.

3.5.1 Average protein and total potential energy per structural group monitored in 30 successive trajectory windows

The average E_{xp} and the E_{xp}^{r} (where r = relative, $x = t$ for total or p for protein and p = potential) were computed for 30 successive trajectory windows of 2.94 ns each, at all temperatures and for the 5 structural groups defined in Chapter 2. The E_{xp}^{r} is defined as the difference between the average E_{xp} of a given structural group and the average E_{xp} of the native structural group (fraction of native contacts definition). This measure shows which of the structurals groups has a lower potential energy for a given window. The results are presented in Figure 3.8 and show that the minimal energy structural group changes from one window to another, with no clear trend. Nevertheless, the results confirm that, at all temperatures, the native structure group occupies the minimum of both E_{pp} and E_{tp} for at least half or more of the trajectory windows.

3.5.2 Average protein and total potential energy in function of the location on the conformational landscape

Computing average E_{pp} and E_{tp} for regular Q_{fr} and R_g intervals allows to map potential energy surface minima and maxima (E_{pp} and E_{tp}) on the conformational landscape as projected on the Q_{fr} and R_g reaction coordinates. If $REMDpe$ is a valid approximation, the mapped E_{pp} and E_{tp} minima and maxima should coincide. The results are presented for 300, 400 and 500K, using averages obtained from the entire trajectory (Figure 3.9) or from 4 successive 22.1 ns trajectory windows at 300K (Figure 3.10), 400K (Figure 3.11) and 500K (Figure 3.12) in order to analyse the dynamics in the different stages of the non-converged simulation. The mapped E_{tp} generally appears flat (with diffuse minima and maxima), contrarily to the mapped E_{pp}. In most cases, there is a qualitative agreement between the mapped E_{tp} and E_{pp}. The mapped E_{pp} confirms that non-native collapsed structures (group *Collapsed*) are more stable (lower average E_{pp}) than extended non-native structures (group *Other*), and that this difference is more pronounced than it is in the mapped E_{tp}. Similarly, native structures appear to be more stable in the mapped E_{pp} than in the mapped E_{tp}, suggesting that the E_{pp} might actually be more efficient at preserving native structures from exchanging to higher temperatures. Free energy minima (as revealed by the conformational landscape maxima in the figures) frequently coincide to mapped E_{pp} minima at 300K, while this is almost never the case at 400 and 500K.

Thus, entropy dominates at high temperature.

3.5.3 Exchange attempt statistics as a function of structural groups involved

In order to ensure that exchange attempts leading non-native collapsed or non-native extended structures to lower temperatures are on average not more successful than the inverse exchanges, the following three exchange attempts were monitored. They are identified with the following syntax: C; *Collapsed*, N; *Native*; O; *Other* (structural groups, as defined in Chapter 2) and *vs*; versus, comparing an exchange type (referred to as “considered exchange“, at the left of the double arrow) with the inverse exchange (at the right of the double arrow), in order to assess which of the two is more successful on average.

1) $[(T_{n+1}, C), (T_n, N)] \leftrightarrow [(T_{n+1}, N), (T_n, C)]$ vs $[(T_{n+1}, N), (T_n, C)] \leftrightarrow [(T_{n+1}, C), (T_n, N)]$ The inverse exchange shows a slightly (although not significantly) higher fraction of successful exchanges than the considered exchange (Figure 3.13).

2) $[(T_{n+1}, O), (T_n, N)] \leftrightarrow [(T_{n+1}, N), (T_n, O)]$ vs $[(T_{n+1}, N), (T_n, O)] \leftrightarrow [(T_{n+1}, O), (T_n, N)]$. The inverse exchange shows a slightly (although not significantly) higher fraction of successful exchanges than the considered exchange (Figure 3.14).

3) $[(T_{n+1}, C), (T_n, O)] \leftrightarrow [(T_{n+1}, O), (T_n, C)]$ vs $[(T_{n+1}, O), (T_n, C)] \leftrightarrow [(T_{n+1}, C), (T_n, O)]$ The considered exchange shows a slightly (although not significantly) higher fraction of successful exchanges than the inverse exchange (Figure 3.15).

Thus, exchange attempts leading non-native structures to lower temperatures are not more successful on average. Moreover, similar results were obtained when performing the same analysis on the first and second halves of the simulation. Finally, these results are in agreement with the relative stabilities per structural group as assessed with protein and total potential energy averages.

3.5.4 Correlation of protein energy and radius of gyration or fraction of native contacts

In order to check if non-native and collapsed structures correspond to lower potential energies, correlations of potential energy with i) the protein radius of gyration R_g and ii) the protein fraction of native contacts Q_{fr} were computed at 300K and 500K. Different potential energy functions were tested in this correlation. Neither the total nor the protein potential energy functions were correlated to any of the two reaction coordinates. However, the Protein-Protein Lennard Jones (LJ) contributions correlate positively with R_g and Q_{fr} (a decrease of which leads to a decrease of the LJ interaction energy), while all the non-bonded protein-solvent interactions correlate negatively with R_g and Q_{fr}.

3.5.5 Doppel protein *REMDpe* simulation: Comparable results

Chapter 5 describes an *REMDpe* simulation performed with doppel, a structural homologue of prion. The relative stabilities per structural group as assessed by protein and total potential energy averages provided similar results (on the entire trajectory, on four 22.1 ns intervals and on 30 2.94 ns intervals), as well as the statistics on exchange attempts.

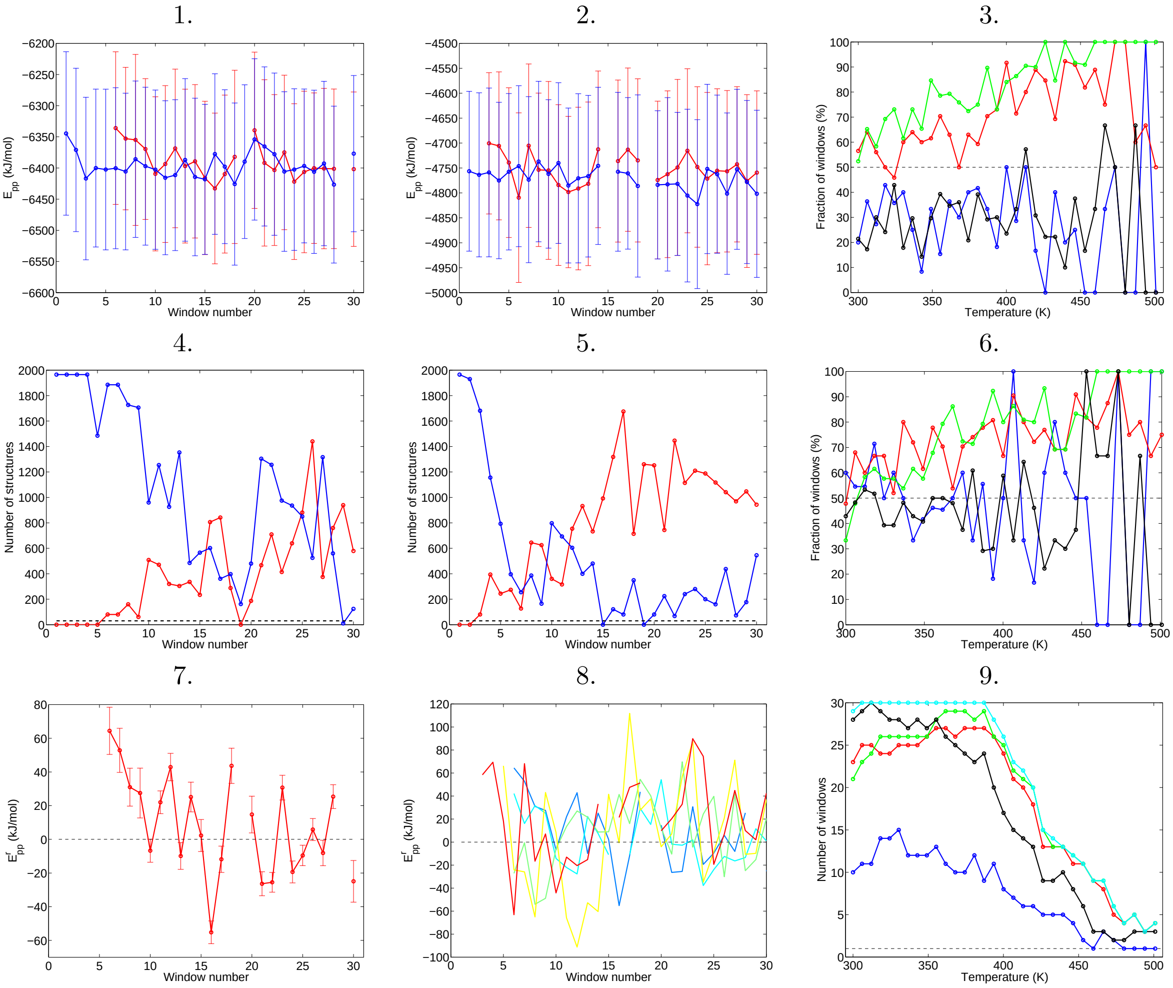


Figure 3.8: *E_{xp} and E^r_{xp} computed for different structural groups in 30 successive trajectory windows of 2.94 ns each. The E^r_{xp} (where r = relative, x = p for protein or t for total, and p = potential) is defined as the difference between the average E_{xp} of a given structural group and the average E_{xp} of the native structural group (fraction of native contacts definition). Panels 1, 2, 4 and 5 compare the collapsed (red) and native (fraction of native contacts definition) (blue) structural groups, at 300K (panels 1 and 4) and at 393K (panels 2 and 5). Panels 1 and 2: Average E_{pp} per trajectory window (errorbars: one standard deviation). Panels 4 and 5: Population size per structural group and trajectory window. Average E_{pp} were only computed for a given structural group for windows containing at least 30 (horizontal black dashed line) structures of that structural group and at least 30 structures of the native group. Panel 7: E^r_{pp} of the collapsed structure group, per 300K trajectory window. Panel 8: E^r_{pp} of the collapsed structure group, per trajectory window, at 300 (blue, data from Panel 7), 318 (cyan), 342 (green), 368 (yellow) and 393K. Panel 3: Fraction of windows for which the E^r_{pp} is below zero for the structural group considered (identified with the same color coding as in Figure 2.4). Panel 6: As Panel 5, for the E^r_{tp}. An E^r_{xp} that is lower than zero reveals that, for a particular window, the average E_{xp} of the native structure group (fraction of native contacts definition) is lower than the one of the structural group considered. Panel 9: Number of windows from which the fractions of Panels 3 and 5 were deduced (same color coding). Black horizontal dashed line: 1 window.*

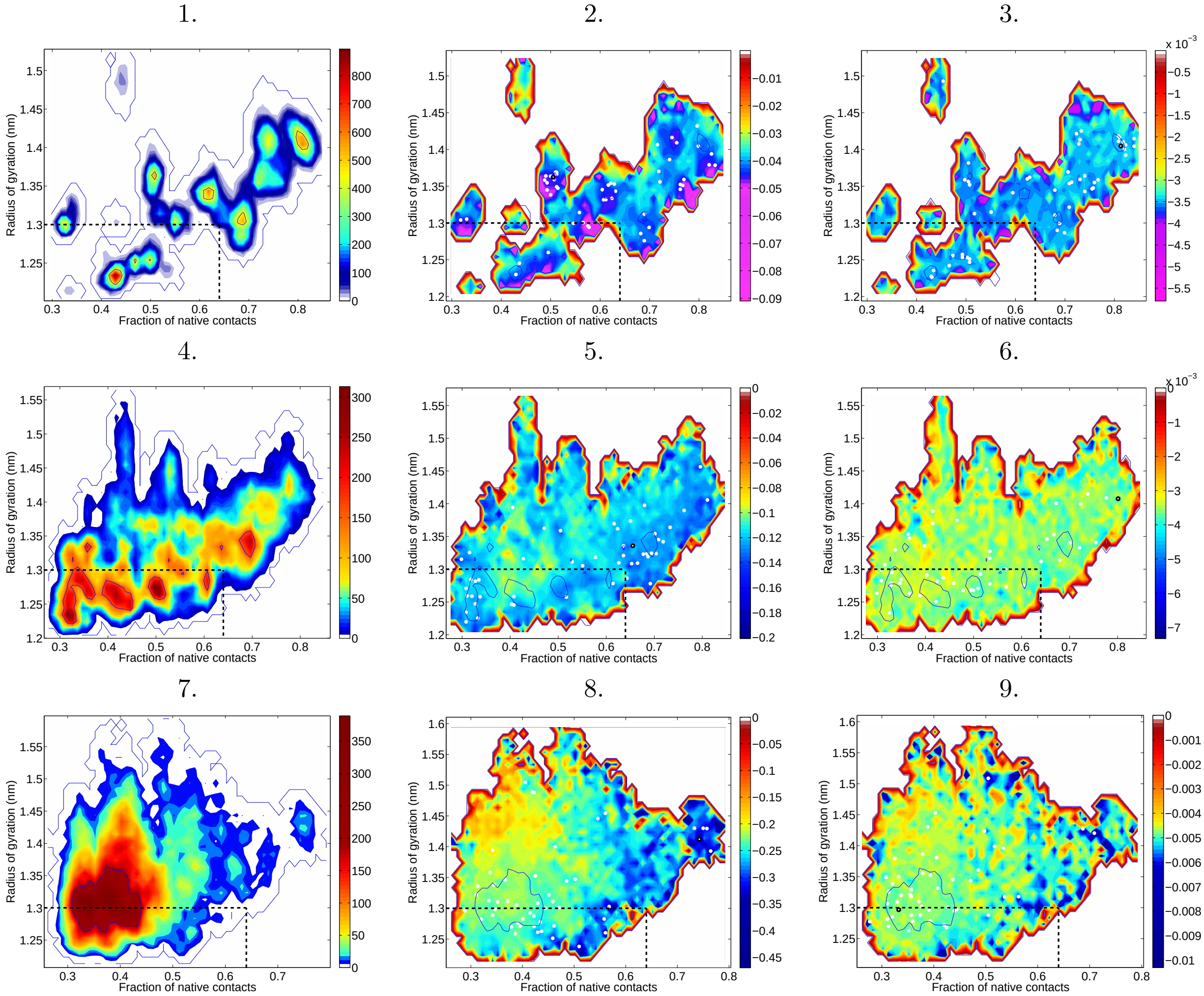


Figure 3.9: *Scaled average* E_{xp} *(or* E^{s}_{xp}*, with* $x = p$ *for protein or* t *for total and* $s =$ *scaled) in function of the location on the conformational landscape (as projected on* Q_{fr} *and* R_g*). Panels 1, 4 and 7: Conformational landscapes (colorbar: number of structures). Panels 2, 5 and 8:* E^{s}_{pp}*, Panels 3, 6 and 9:* E^{s}_{tp}*. In Panels 2, 5, 8, 3, 6 and 9, the* E^{s}_{xp} *values are given in the colorbar, while 50 white circles highlight the 50 lowest* E_{xp} *structures (lowest* E_{xp} *structure: black circle). Averages and conformational landscapes were computed from the entire 88.4 ns trajectory at 300K (Panels 1, 2 and 3), 400K (Panels 4, 5 and 6) and 500K (Panels 7, 8 and 9).* E^{s}_{xp} *were defined with* $E^{s}_{xp} = 1 - \bar{E}_{xp}/max(\bar{E}_{xp} + 1)$*, where* $\bar{E}_{xp}$ *is the average* E_{xp} *of all structures of a given portion of the conformational landscape, in order to improve the contrast of the contour plots. In all panels, boundaries of the portion of the conformational landscape defining the non-native* Collapsed *structural group are indicated by a horizontal (*$R_g \leq 1.3$*) and a vertical (*$Q_{fr} \geq 0.64$*) black dashed line.*

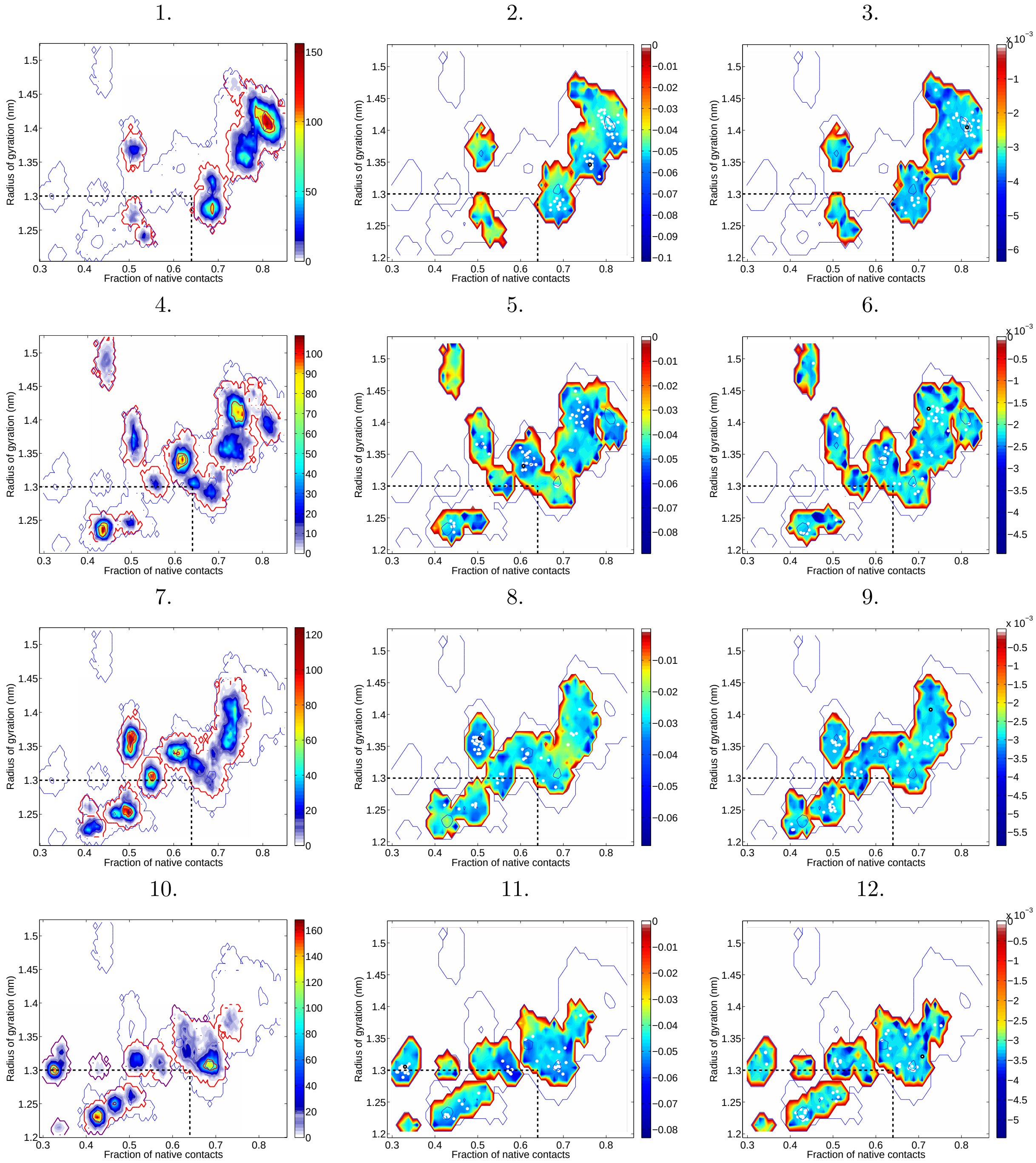


Figure 3.10: *E^s_{xp} (as defined in Figure 3.9) in function of the location on the conformational landscapes (as projected on Q_{fr} and R_g) of the four fourths of the 300K trajectory. Panels 1, 4, 7 and 10: Conformational landscapes (colorbar: number of structures). Panels 2, 5, 8 and 11: E^s_{pp}, Panels 3, 6, 9 and 12: E^s_{tp}. In Panels 2, 5, 8, 11, 3, 6, 9 and 12, the E^s_{xp} values are given in the colorbar, while 50 white circles highlight the 50 lowest E_{xp} structures (lowest E_{xp} structure: black circle). Averages and conformational landscapes were computed for the four fourths of the simulation, with each fourth spanning 22.1 ns: First (Panels 1, 2 and 3), second (Panels 4, 5 and 6), third (Panels 7, 8 and 9) and fourth (Panels 10, 11, 12) fourths of the simulation. In all panels, boundaries of the portion of the conformational landscape defining the non-native* Collapsed *structural group are indicated by a horizontal ($R_g \leq 1.3$) and a vertical ($Q_{fr} \geq 0.64$) black dashed line.*

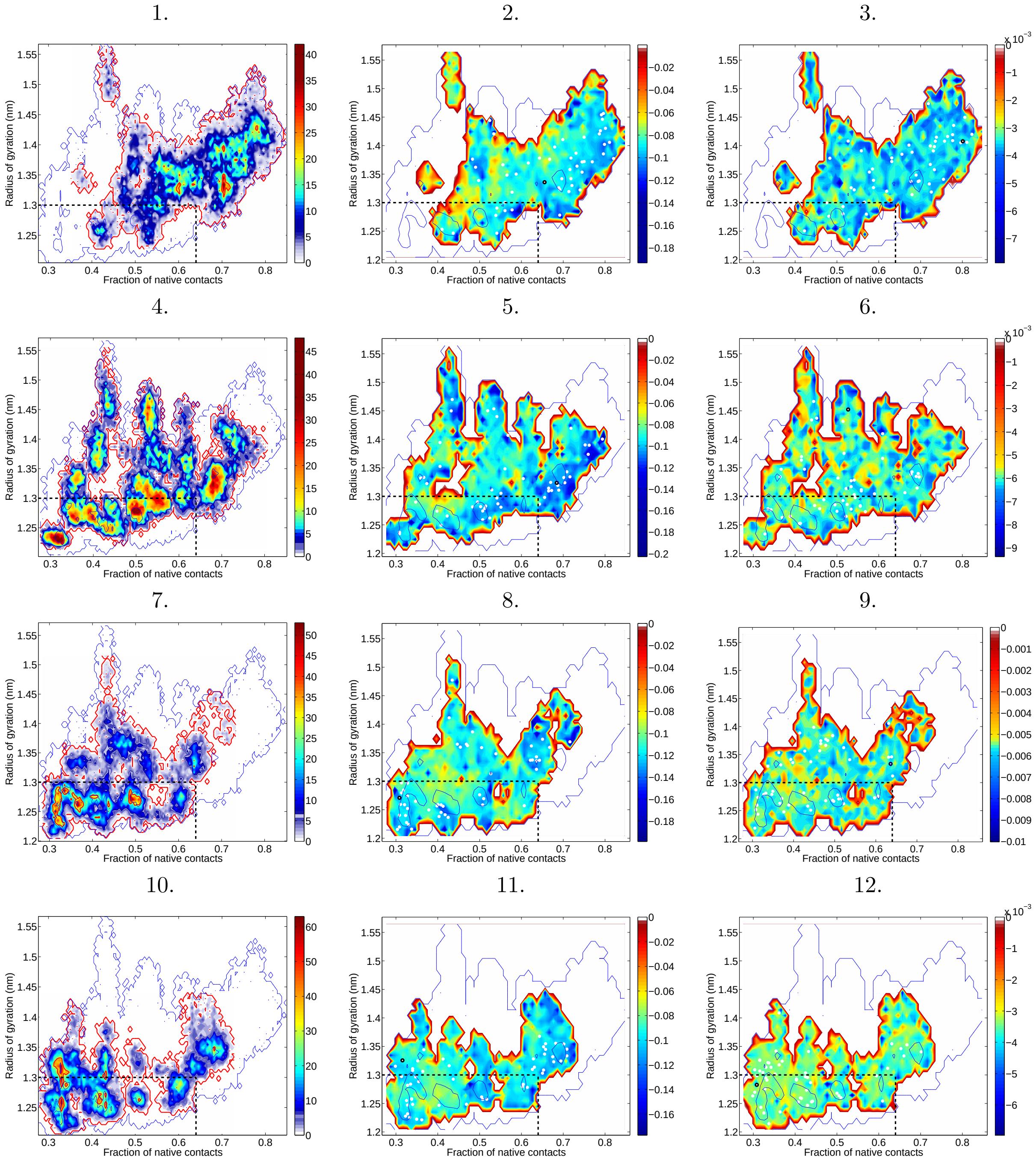


Figure 3.11: *E^s_{xp} (as defined in Figure 3.9) in function of the location on the conformational landscapes (as projected on Q_{fr} and R_g) of the four fourths of the 400K trajectory. Panels 1, 4, 7 and 10: Conformational landscapes (colorbar: number of structures). Panels 2, 5, 8 and 11: E^s_{pp}, Panels 3, 6, 9 and 12: E^s_{tp}. In Panels 2, 5, 8, 11, 3, 6, 9 and 12, the E^s_{xp} values are given in the colorbar, while 50 white circles highlight the 50 lowest E_{xp} structures (lowest E_{xp} structure: black circle). Averages and conformational landscapes were computed for the four fourths of the simulation, with each fourth spanning 22.1 ns: First (Panels 1, 2 and 3), second (Panels 4, 5 and 6), third (Panels 7, 8 and 9) and fourth (Panels 10, 11, 12) fourths of the simulation. In all Panels, boundaries of the portion of the conformational landscape defining the non-native* Collapsed *structural group are indicated by with a horizontal ($R_g \leq 1.3$) and a vertical ($Q_{fr} \geq 0.64$) black dashed line.*

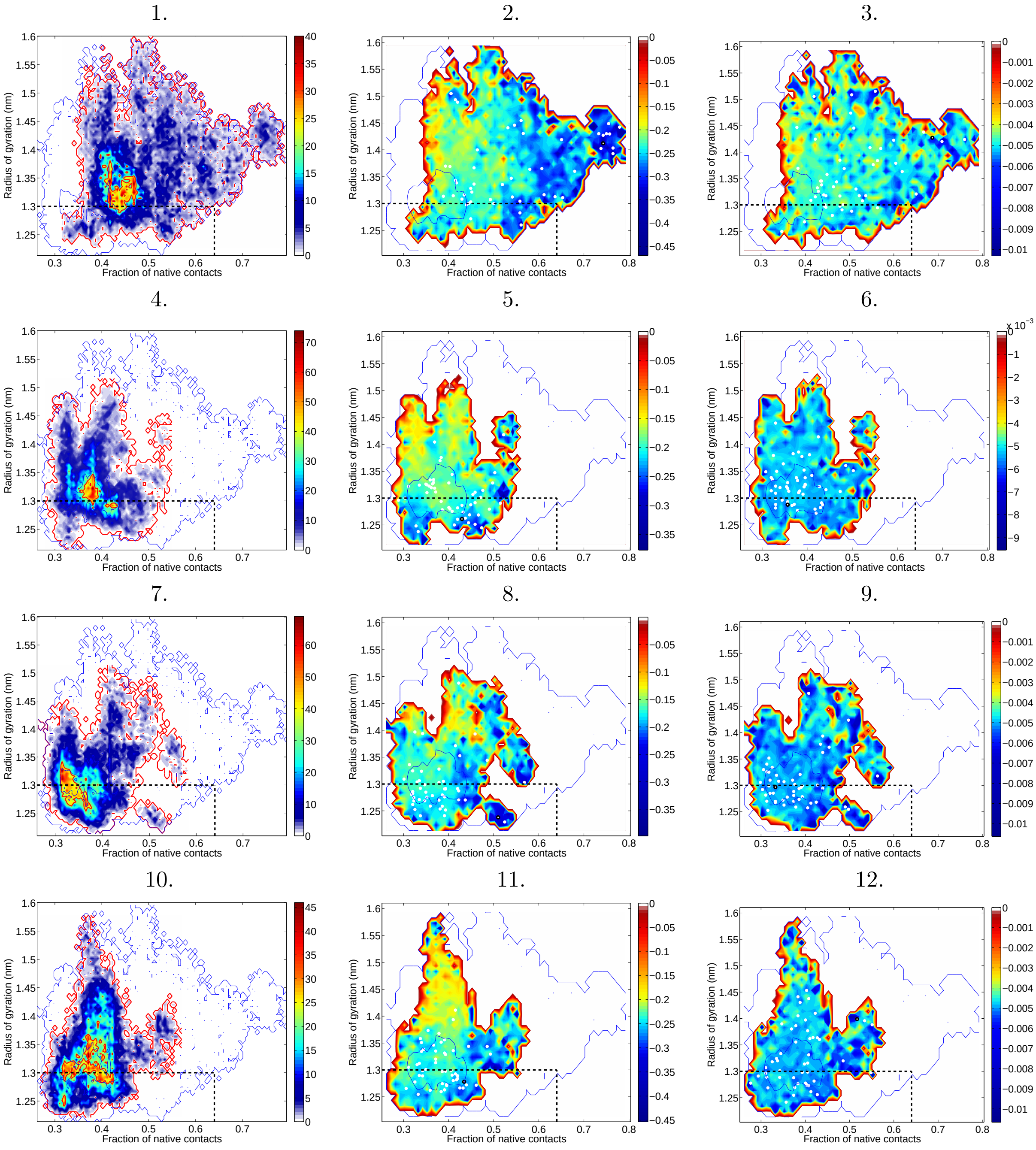


Figure 3.12: *E^s_{xp} (as defined in Figure 3.9) in function of the location on the conformational landscapes (as projected on Q_{fr} and R_g) of the four fourths of the 500K trajectory. Panels 1, 4, 7 and 10: Conformational landscapes (colorbar: number of structures). Panels 2, 5, 8 and 11: E^s_{pp}, Panels 3, 6, 9 and 12: E^s_{tp}. In Panels 2, 5, 8, 11, 3, 6, 9 and 12, the E^s_{xp} values are given in the colorbar, while 50 white circles highlight the 50 lowest E_{xp} structures (lowest E_{xp} structure: black circle). Averages and conformational landscapes were computed for the four fourths of the simulation, with each fourth spanning 22.1 ns: First (Panels 1, 2 and 3), second (Panels 4, 5 and 6), third (Panels 7, 8 and 9) and fourth (Panels 10, 11, 12) fourths of the simulation. In all Panels, boundaries of the portion of the conformational landscape defining the non-native* Collapsed *structural group are indicated by a horizontal ($R_g \leq 1.3$) and a vertical ($Q_{fr} \geq 0.64$) black dashed line.*

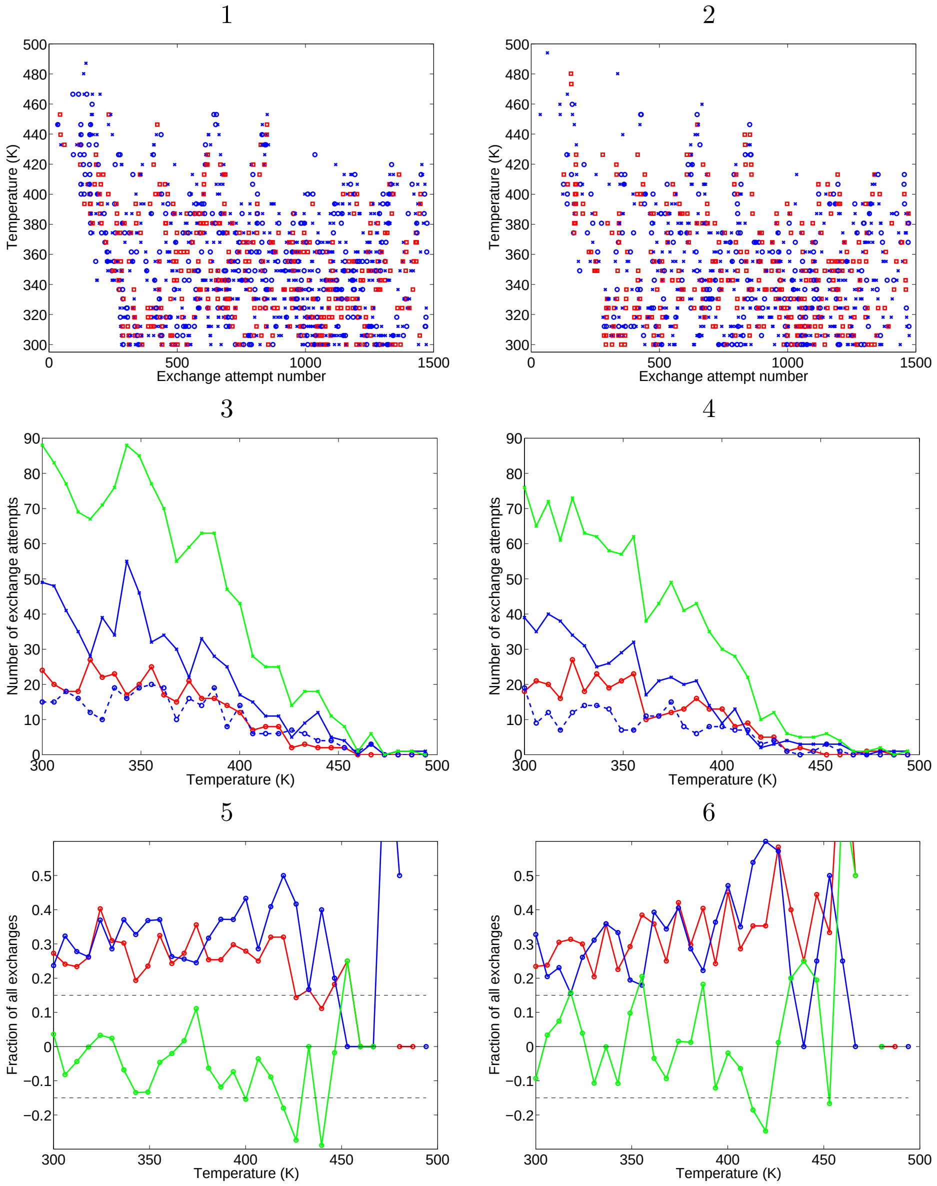


Figure 3.13: *REMD exchange attempts of structures of the non-native* Collapsed *and* Native *(fraction of native contacts definition) structure groups. Panels 1 and 3: Exchange attempts between* Collapsed *structures at* T_{n+1} *and* Native *structures at* T_n*, Panels 2 and 4: Exchange attempts between* Native *structures at* T_{n+1} *and* Collapsed *structures at* T_n*. Panels 1 and 2: Temperatures and times for all exchange attempts with a probability P i)* $P < 0.5$ *(blue "x"), ii)* $0.5 \leq P < 1$ *(blue circles) and iii)* $P \geq 1$ *(direct exchanges, red squares). Panels 3 and 4: Number of exchange attempts with a probability P i)* $P < 0.5$ *(solid blue line), ii)* $0.5 \leq P < 1$ *(dashed blue line), iii)* $P \geq 1$ *(direct exchanges, solid red line) and total number of exchange attempts (green). Panel 5: Fraction of direct exchanges (number of direct exchanges divided by the number of all exchange attempts), Panel 6: Fraction of exchange attempts with a probability P in the interval* $0.5 \leq P < 1$ *(number of such exchange attempts divided by the number of exchange attempts with* $P < 1$*). In Panels 5 and 6, fractions of exchange attempts between* Collapsed *structures at* T_{n+1} *and* Native *structures at* T_n *are depicted with a red line, and fractions of exchanges attempts between* Native *structures at* T_{n+1} *and* Collapsed *structures at* T_n *are indicated by a blue line. The green line is the difference of the former and the latter.*

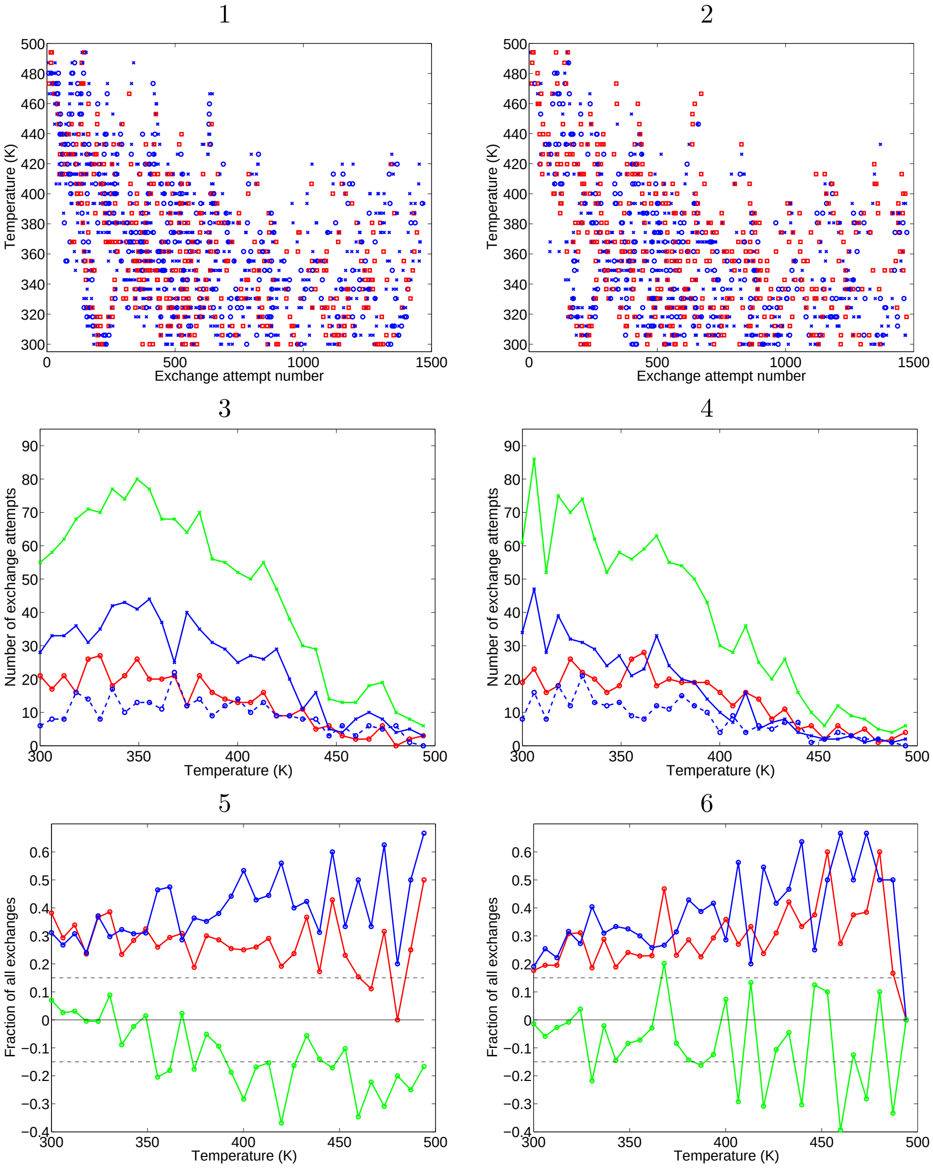


Figure 3.14: *REMD exchange attempts of structures of the non-native, non-collapsed* Other *and the* Native *(fraction of native contacts definition) structure groups. Panels 1 and 3: Exchange attempts between* Other *structures at* T_{n+1} *and* Native *structures at* T_n*, Panels 2 and 4: Exchange attempts between* Native *structures at* T_{n+1} *and* Other *structures at* T_n*. Panels 1 and 2: Temperatures and times for all exchange attempts with a probability* P *i)* $P < 0.5$ *(blue "x"), ii)* $0.5 \leq P < 1$ *(blue circles) and iii)* $P \geq 1$ *(direct exchanges, red squares). Panels 3 and 4: Number of exchange attempts with a probability* P *i)* $P < 0.5$ *(solid blue line), ii)* $0.5 \leq P < 1$ *(dashed blue line), iii)* $P \geq 1$ *(direct exchanges, solid red line) and total number of exchange attempts (green). Panel 5: Fraction of direct exchanges (number of direct exchanges divided by the number of all exchange attempts), Panel 6: Fraction of exchange attempts with a probability* P *in the interval* $0.5 \leq P < 1$ *(number of such exchange attempts divided by the number of exchange attempts with* $P < 1$*). In Panels 5 and 6, fractions of exchange attempts between* Other *structures at* T_{n+1} *and* Native *structures at* T_n *are depicted with a red line, and fractions of exchanges attempts between* Native *structures at* T_{n+1} *and* Other *structures at* T_n *are indicated by a blue line. The green line is the difference of the former and the latter.*

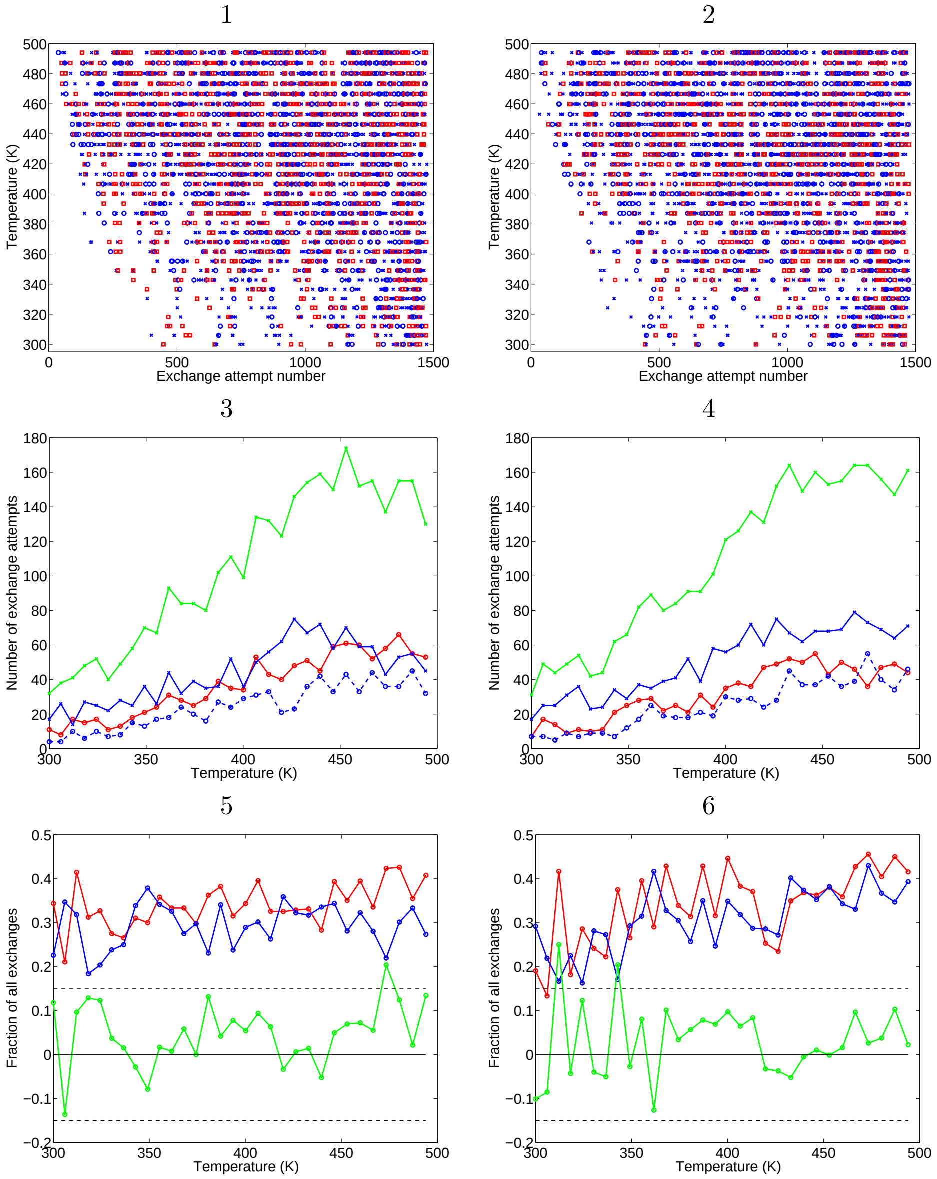


Figure 3.15: *REMD exchange attempts of structures of the non-native* Collapsed *and non-native non-collapsed* Other *structure groups. Panels 1 and 3: Exchange attempts between* Collapsed *structures at T_{n+1} and* Other *structures at T_n, Panels 2 and 4: Exchange attempts between* Other *structures at T_{n+1} and* Collapsed *structures at T_n. Panels 1 and 2: Temperatures and times for all exchange attempts with a probability P i) $P < 0.5$ (blue "x"), ii) $0.5 \leq P < 1$ (blue circles) and iii) $P \geq 1$ (direct exchanges, red squares). Panels 3 and 4: Number of exchange attempts with a probability P i) $P < 0.5$ (solid blue line), ii) $0.5 \leq P < 1$ (dashed blue line), iii) $P \geq 1$ (direct exchanges, solid red line) and total number of exchange attempts (green). Panel 5: Fraction of direct exchanges (number of direct exchanges divided by the number of all exchange attempts), Panel 6: Fraction of exchange attempts with a probability P in the interval $0.5 \leq P < 1$ (number of such exchange attempts divided by the number of exchange attempts with $P < 1$). In Panels 5 and 6, fractions of exchange attempts between* Collapsed *structures at T_{n+1} and* Other *structures at T_n are depicted with a red line, and fractions of exchanges attempts between* Other *structures at T_{n+1} and* Collapsed *structures at T_n are depicted with a blue line. The green line is the difference of the former and the latter.*

3.6 References

[1] R. Zhou, B. J. Berne, and R. Germain. "The free energy landscape for beta hairpin folding in explicit water." *Proc Natl Acad Sci U S A*, **98**, (2001) 14 931–14 936.

[2] M. M. Seibert, A. Patriksson, B. Hess, and D. van der Spoel. "Reproducible polypeptide folding and structure prediction using molecular dynamics simulations." *J Mol Biol*, **354**, (2005) 173–183.

[3] X. Periole and A. E. Mark. "Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent." *J Chem Phys*, **126**, (2007) 014 903.

[4] D. Paschek, H. Nymeyer, and A. E. García. "Replica exchange simulation of reversible folding/unfolding of the trp-cage miniprotein in explicit solvent: on the structure and possible role of internal water." *J Struct Biol*, **157**, (2007) 524–533.

[5] Y. Sugita and Y. Okamoto. "Replica-exchange Molecular Dynamics Method for Protein Folding". *Chem. Phys. Lett.*, **314**, (1999) 141–151.

[6] H. Fukunishi, O. Watanabe, and S. Takada. "On the hamiltonian replica-exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction". *J Chem Phys*, **116**, (2002) 9058–9067.

[7] S. Jang, S. Shin, and Y. Pak. "Replica-exchange method using the generalized effective potential." *Phys Rev Lett*, **91**, (2003) 058 305.

[8] X. Cheng, G. Cui, V. Hornak, and C. Simmerling. "Modified replica exchange simulation methods for local structure refinement." *J Phys Chem B Condens Matter Mater Surf Interfaces Biophys*, **109**, (2005) 8220–8230.

[9] P. Liu, B. Kim, R. A. Friesner, and B. J. Berne. "Replica exchange with solute tempering: a method for sampling biological systems in explicit water." *Proc Natl Acad Sci U S A*, **102**, (2005) 13 749–13 754.

[10] X. Huang, M. Hagen, B. Kim, R. A. Friesner, R. Zhou, and B. J. Berne. "Replica exchange with solute tempering: efficiency in large scale systems." *J Phys Chem B*, **111**, (2007) 5405–5410.

[11] A. Okur, L. Wickstrom, M. Layten, R. Geney, K. Song, V. Hornak, and C. Simmerling. "Improved Efficiency of Replica Exchange Simulations through Use of a Hybrid Explicit/Implicit Solvation Model". *J. Chem. Theory Comput.*, **2**, (2006) 420–433.

[12] Y. Mu, Y. Yang, and W. Xu. "Hybrid hamiltonian replica exchange molecular dynamics simulation method employing the poisson-boltzmann model." *J Chem Phys*, **127**, (2007) 084 119.

[13] D. V. D. Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. "GROMACS: fast, flexible, and free." *J Comput Chem*, **26**, (2005) 1701–1718.

[14] W. van Gunstern, S. Billeter, A. Eising, P. Huenenberger, P. Krueger, A. Mark, W. Scott, and I. Tironi. *Biomolecular Simulation: The GROMOS96 Manual and User Guide* (Hochschulverlag AG, Zuerich, 1996).

[15] H. Berendsen, J. Postma, W. van Gunsteren, and J. Hermans. *Intermolecular Forces* (Reidel, Dordrecht, 1981).

[16] J. P. Ryckaert, G. Ciccotti, and H. Berendsen. “Numerical integration of cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes”. *J Comp Phys*, **23**, (1977) 327–341.

[17] S. Nosé. “A unified formulation of the constant temperature molecular dynamics methods”. *J. Chem. Phys.*, **81**, (1984) 511–519.

[18] W. Hoover. “Canonical dynamics-equilibrium phase space distribution”. *Phys. Rev. A.*, **31**, (1985) 1695–1697.

[19] I. G. Tironi, R. Sperb, P. E. Smith, and W. F. van Gunsteren. “A generalized reaction field method for molecular dynamics simulations”. *J Chem Phys*, **102**, (1995) 5451–5459.

[20] R. Riek, S. Hornemann, G. Wider, M. Billeter, R. Glockshuber, and K. Wüthrich. “NMR structure of the mouse prion protein domain PrP(121-321).” *Nature*, **382**, (1996) 180–182. Exp, nmr.

[21] J. Antosiewicz, J. A. McCammon, and M. K. Gilson. “Prediction of pH-dependent properties of proteins.” *J Mol Biol*, **238**, (1994) 415–436.

[22] M. Davis and J. McCammon. “Electrostatics in biomolecular structure and dynamics”. *Chem Rev*, **90**, (1990) 509–521.

[23] W. Swietnicki, R. Petersen, P. Gambetti, and W. K. Surewicz. “ph-dependent stability and conformation of the recombinant human prion protein prp(90-231).” *J Biol Chem*, **272**, (1997) 27 517–27 520.

[24] A. Yang, M. Gunner, R. Sampogna, R. Sharp, and B. Honig. “On the calculation of pKa in proteins”. *Proteins*, **15**, (1993) 252–256.

[25] G. Vriend. “WHAT IF: a molecular modeling and drug design program.” *J Mol Graph*, **8**, (1990) 52–6, 29.

[26] W. Kabsch and C. Sander. “Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.” *Biopolymers*, **22**, (1983) 2577–2637.

[27] K. Kuwata, H. Li, H. Yamada, G. Legname, S. B. Prusiner, K. Akasaka, and T. L. James. “Locally disordered conformer of the hamster prion protein: a crucial intermediate to prpsc?” *Biochemistry*, **41**, (2002) 12 277–12 283.

[28] K. Kuwata, Y. O. Kamatari, K. Akasaka, and T. L. James. "Slow conformational dynamics in the hamster prion protein." *Biochemistry*, **43**, (2004) 4439–4446.

[29] J. Ziegler, H. Sticht, U. C. Marx, W. Mller, P. Rsch, and S. Schwarzinger. "Cd and nmr studies of prion protein (prp) helix 1. novel implications for its role in the prp^c → prp^{Sc} conversion process." *J Biol Chem*, **278**, (2003) 50 175–50 181.

[30] X. Lu, P. L. Wintrode, and W. K. Surewicz. "Beta-sheet core of human prion protein amyloid fibrils as determined by hydrogen/deuterium exchange." *Proc Natl Acad Sci U S A*, **104**, (2007) 1510–1515.

[31] R. I. Dima and D. Thirumalai. "Probing the instabilities in the dynamics of helical fragments from mouse prpc." *Proc Natl Acad Sci U S A*, **101**, (2004) 15 335–15 340.

[32] A. D. Simone, A. Zagari, and P. Derreumaux. "Structural and hydration properties of the partially unfolded states of the prion protein." *Biophys J.*

[33] H. Mo, R. C. Moore, F. E. Cohen, D. Westaway, S. B. Prusiner, P. E. Wright, and H. J. Dyson. "Two different neurodegenerative diseases caused by proteins with similar structures." *Proc Natl Acad Sci U S A*, **98**, (2001) 2352–2357.

[34] K. Post, M. Pitschke, O. Schfer, H. Wille, T. R. Appel, D. Kirsch, I. Mehlhorn, H. Serban, S. B. Prusiner, and D. Riesner. "Rapid acquisition of beta-sheet structure in the prion protein prior to multimer formation." *Biol Chem*, **379**, (1998) 1307–1317.

[35] K. Jansen, O. Schäfer, E. Birkmann, K. Post, H. Serban, S. B. Prusiner, and D. Riesner. "Structural intermediates in the putative pathway from the cellular prion protein to the pathogenic form." *Biol Chem*, **382**, (2001) 683–691.

[36] K.-W. Leffers, H. Wille, J. Stöhr, E. Junger, S. B. Prusiner, and D. Riesner. "Assembly of natural and recombinant prion protein into fibrils." *Biol Chem*, **386**, (2005) 569–580.

[37] N. D. Socci, J. N. Onuchic, and P. G. Wolynes. "Diffusive dynamics of the reaction coordinate for protein-folding funnels". *J Chem Phys*, **104**, (1996) 5860–5868.

[38] E. Lyman and D. M. Zuckerman. "The structural de-correlation time: A robust statistical measure of convergence of biomolecular simulations", (2006).

[39] Y. M. Rhee and V. S. Pande. "Multiplexed-replica exchange molecular dynamics method for protein folding simulation." *Biophys J*, **84**, (2003) 775–786.

[40] W. Li, J. Zhang, and W. Wang. "Understanding the folding and stability of a zinc finger-based full sequence design protein with replica exchange molecular dynamics simulations." *Proteins*, **67**, (2007) 338–349.

[41] A. E. García and J. N. Onuchic. "Folding a protein in a computer: an atomic description of the folding/unfolding of protein a." *Proc Natl Acad Sci U S A*, **100**, (2003) 13 898–13 903.

Chapter 4

A replica-exchange molecular dynamics study of prion misfolding and β-rich folds

Abstract

In the present study, we perform a replica exchange molecular dynamics simulation corresponding to a 2.8 μs total time for the extensive enhanced sampling of the conformational space of mouse Prion (PrP) 124-226 monomer in explicit-atom aqueous solution at pH 4. 1.3% of the conformations sampled display a high β content ($\geq$ 19 residues), allowing to assess β-propensities along the sequence and highlight labile hot spots. A clustering algorithm is applied to sort the structures of this pool in function of the topology of their β-contacts. 10 β-rich folds are thus defined and analyzed in regard to their topology, accumulation temperatures and structural characteristics. At contrast to models derived from previous MD simulations, we present putative structural models for monomeric precursors of PrP^{Sc} and PrP β-oligomers that are characterized by a C-terminal β-rich core which is consistent with recent experiments.

4.1 Introduction

According to the widely accepted protein-only hypothesis, prion diseases are caused by a misfold of the prion protein (PrP) frequently referred to as the scrapie isoform PrP^{Sc}. PrP^{C} could be in equilibrium with a metastable intermediate, designated PrP*, that catalyzes the conversion [1]. This process is devoid of any chemical change but involves profound conformational changes from the soluble, predominantly α-helical PrP^{C} to the insoluble, aggregated β-rich PrP^{Sc} [2].

The first experimental PrP^{C} structure was obtained by NMR for mouse PrP (residues 124 to 226) [3]. The glycoprotein shows an unstructured N-terminal (res 21-123) and a structured C-terminal domain (res 124-226). PrP C-terminal domains of different species reveal highly conserved structures, comprised of two short β-strands, 3 α-helices H1, H2 and H3, a disulphide bridge connecting H2 and H3 and two glycosylation sites (one in H2 and one in the H2-H3 loop) (Figure 4.3, panel 11).

The insolubility of PrP^{Sc} has thwarted attempts to investigate its structure by either x-ray or NMR spectroscopy. Circular dichroism (CD) of the full length protein has revealed that PrP^{C} (6% β and 43% α) undergoes dramatic secondary structure changes in the conversion to PrP^{Sc} (43% β and 30% α)[2]. For PrP27-30 [2, 4, 5], a subfragment of PrP^{Sc} obtained with proteinase K digestion via cleavage around residues 87 to 91 (depending on the PrP strain), the corresponding values are respectively 47-54% β and 17-25% α. Consequently, if PrP27-30 monomeric forms exist, they probably contain at least $\sim$ 17 α-helical residues.

A consensus is slowly emerging on the picture of a very complex PrP folding landscape, allowing for a large variety of misfolding pathways. The multitude of possible conformations observed in-vitro, under different perturbing environments (redox potential changes [6], chemical unfolding agents [7, 8, 9, 10], detergent removal of protein solution [11, 12, 13], high temperature [14] and pressure [15]) includes several soluble β-oligomeric conformations (some of which were detected at physiological conditions), amyloidogenic conformations and amyloid fibrils.

The smallest number of monomers forming PrP^{Sc} or an infectious unit has remained elusive. Silveira and colleagues showed that small oligomers composed of less than 6 units were noninfectious in Syrian hamsters, while particles harboring the highest infectivity were non-fibrillar structures composed of 14-28 units of PrP^{Sc} [16]. Octamers [17], 12mers and 36mers [18], as well as 15mers [19] have also been observed in in-vitro oligomerization experiments. These questions have lead to the quest of a monomeric PrP^{Sc} or PrP^{*} precursor form, as well as to the investigation of PrP^{C} changes leading to such a form [14]. In the present theoretical study, we address these two issues.

We restrict our study to the structured C-terminal domain of monomeric PrP^{C}. This choice is justified by several independent experiments showing that the C-terminal structured domain might by itself be able to promote the pathology: (i) PrP27-30 is still infectious [20]. (ii) The structured C-terminal domain (residues 124 to 226) can also undergo complex misfolding conversions leading to β-sheet rich oligomers [21] and amyloid aggregates [22]. (iii) Limited proteolysis of fibrils formed in vitro with recombinant mouse PrP has generated an even smaller C-terminal domain PrP fragment that has been reported to support fibril propagation in vitro [23]. This unusual protease-resistant core encompasses residue 152 or 162 (C-terminus of H1 or S2) to 226 and has also been found in a novel form of sporadic CJD [24]. (iv) Lu et al., who performed hydrogen exchange and mass spectroscopy experiments on recombinant human PrP amyloid fibrils, identified a β-rich core comprising mainly H2 and H3 residues [25]. (v) The same β-rich core was found for recombinant human PrP fibrils formed in vitro with site-directed spin labeling experiments combined with EPR spectroscopy [26].

A number of atomistic models of monomeric misfolded PrP or PrP^{Sc} have been derived from molecular dynamics (MD) simulations. The so called spiral model proposed by Daggett et al. [27, 28, 29, 30] is in agreement with a number of experimental data, but the suggested β-rich region N-terminal to H1 is inconsistent with the recently determined location of the β-core involving H2-H3 [23, 25, 26]. In high temperature MD studies, a number of β-enriched conformations have transiently been observed [31, 32]. Finally, an REMD study has been performed to investigate early changes in the solvation shell and subdomain motions of murine PrP [33], but no substantial β structure enrichment was observed with the limited simulation time (30 ns) and temperature range (320-370K).

In the present work, we perform an extensive 2.8 μs (total time) REMD [34] simulation with the monomeric C-terminal part of PrP^{C} (residues 124-226). Our aim is to assess β-sheet formation propensities along the sequence and investigate major possible β-rich rearrangements. In order to increase the probability of observing PrP^{Sc} related changes, we have chosen to perform the simulations at pH 4, mimicking an acid environment favoring the PrP^{C} to PrP^{Sc} conversion [35]. One of the issues we address is the possibility that some of the new β-sheets might already be formed prior to aggregation, and allow for further aggregation driven formation of inter-monomer β-sheets. The observed β-rich conformations can serve as structural models for putative monomeric precursors of PrP^{Sc} and/or β-oligomers, allowing to characterize early stages of the pathology which, due to insolubility and aggregation, remain poorly understood.

4.2 Methods

All molecular dynamics (MD) simulations were performed with the GROMACS-3.3.0 package [36], the GROMOS96 force-field [37], the SPC water model [38] and an MD timestep of 1.5 fs. The length of covalent bonds involving hydrogen atoms was constrained by the SHAKE [39] algorithm with a tolerance of 10^{-4} kJ(mol nm)$^{-1}$. Temperature control and electrostatic interactions were performed as described in Chapter 2. The starting conformation was the mouse PrP NMR structure (res 124-226, PDB code 1AG2 [3]). In order to mimic a low pH environment favoring the PrP^{C} to PrP^{Sc} conversion [35], the protonation states of ionizable side chains were assigned for a pH of 4 by finite-difference Poisson-Boltzmann calculations [40, 41]. The DELPHI program [42] supplied with the WHATIF package [43] was used to solve the Poisson-Boltzmann equation. The protonation states of HIS, ASP and GLU side chains were consequently set as follows: HIS17; p, ASP21; d, GLU23; d, ASP24; d, GLU29; p, ASP44; p, HIS54; p, ASP55; d, HIS64; p, GLU73; d, GLU77; p, ASP79; d, GLU84; p, GLU88; p and GLU98; p (where p stands for protonated and d for deprotonated). A rhombic dodecahedron box with 18575 water molecules was constructed around the protein. 9 Cl^{-} ions were added to neutralize protein charges.

In order to effectively simulate this large, explicit solvent system, the simulations were carried out with an adapted protocol, REMD - partial energy (*REMDpe*) (Chapter 2). The equilibration of the NMR structure and of the 32 *REMDpe* replicas were performed as described in Chapter 2. The *REMDpe* production run was performed in the NVT ensemble with the following temperature distribution: 300.0, 306.0, 312.1, 318.2, 324.3, 330.4, 336.6, 342.8, 349.1, 355.3, 361.6, 368.0, 374.3, 380.7, 387.1, 393.6, 400.1, 406.6, 413.1, 419.7, 426.3, 432.9, 439.6, 446.3, 453.0, 459.8, 466.5, 473.3, 480.2, 487.1, 494.0 and 500.9K. With this distribution, we obtained *REMEpe* exchange frequencies of ~30%. The *REMDpe* production run was carried out for 88.4 ns (total aggregate time of 2.8 μs). Structures, energies and temperatures were saved every 1.5 ps. A 315 ns-long MD reference simulations was performed at 300K.

Secondary structure elements were assigned with the DSSP [44] algorithm via the GROMACS-3.3.0 interface. An in-house program was used to compute contact maps and fractions of native contacts. Two residues were considered in contact if the shortest distance between two atoms of these residues was inferior to 0.45 nm, and the fraction of native contacts of a given conformation was defined as the percentage of contacts of the NMR structure that were still present. All the other analysis were performed with GROMACS-3.3.0 [36] routines.

A pool of β-rich structures was constructed with all the structures containing $\geq$ 19

β-residues. In order to cluster the different β-rich folds as a function of the topology of the new β-sheets, we developed and applied the *β contact map clustering* (*bcmc*) protocol. *bcmc* comprises the 3 following steps: (I) β contact maps (termed *bcm*) were computed with all β residues (according to the DSSP definition). Each β-rich structure was thus represented by a 103*103 *bcm* matrix (where 103 is the sequence length of the PrP structure) containing, in row i and column j, the minimal inter-residue backbone H-bond distance whenever i and j were β residues and "0" if either i or j were not β residues. Minimal inter-β residue distances were only retained as β contacts if (i) β residues i and j were at least 2 residues apart in sequence, defining a possible contact between two distinct β-strands and (ii) their distance was $\leq$ 0.35nm. For non-retained distances, the corresponding position of the *bcm* was set to "0". The *bcm* obtained with this procedure describe the pairs of β-strands forming sheets (native and novel) that characterize the folds found in the β-rich structures. *bcm* were computed for all the β rich structures (containing at least 19 β residues). (II) In order to group the β contacts, the sequence was subdivided into intervals approaching the optimum target size of I_r residues. The number of intervals s was obtained by rounding $103/I_r$ up to the next integer. Every residue i was assigned to an interval of which the identifier I_i number was obtained by rounding $(i/103)*s$ up to the next integer. A simplified $s*s$ *sbcm* matrix was constructed by setting $sbcm(I_i, I_j)$ to "1" whenever any pair of residues $i \in I_i$ and $j \in I_j$ would form a β contact ($bcm(i,j) = 1$) and to zero if no such pair could be found. (III) MATLAB [45] K-means non-hierarchical clustering [46] was applied to vectorial representations of the *sbcm* obtained for all β-rich structures.

A number of trials were performed, varying the target number of clusters, I_r and the size of the simplified matrix in order to obtain the smallest possible number of clusters containing the same *sbcm* for all members. This lead to $I_r = 15$, $s = 7$ and to a target number of clusters of 4500, of which a majority were of small size. Clusters were assigned to group $\alpha+$ whenever they contained at least 1 structure with 17 α-helical residues (the minimal α-helical content of PrP^{Sc}) and to group $\alpha-$ otherwise. In order to limit the analysis to frequent β-sheet pairing patterns and stable folds, cluster population size threshholds were set as low as 20 for $\alpha+$ clusters (so that very rare, but potentially PrP^{Sc} related folds could be considered) and to 182 for $\alpha-$ clusters. Finally, similar clusters (similar *bcm* and 3D structures) with populations sizes exceeding these threshholds were grouped into 8 $\alpha+$ (at least one $\alpha+$ cluster) and 2 $\alpha-$ (no $\alpha+$ cluster) folds. Thus, each fold contained clusters sharing a common group of β contacts, with each cluster showing an additional alternative β contact.

4.3 Results and Discussion

Statistics were collected from an REMD run with 32 trajectories of 88.4 ns each, distributed over a temperature range of 300-500K (total of 2.8 μs and 1'885'856 sampled conformatins). We first analyzed the secondary structure propensity along the sequence, defined as the fraction of simulation time spent per residue in α and β conformations (Figure 4.1). At all temperatures, H1 is found to be the most stable helix, followed by H3 and H2, consistent with experimental stability studies [15, 47, 48, 25] and simulations [49, 33]. These results suggest that H1 might remain intact in PrP^{Sc} and add up to the minimal observed α-helical content of 17 residues. This is also supported by the decrease of α-helical content and unchanged β-content observed upon PK digestion of recombinant PrP27-30 amyloid leading to a smaller protease resistant core and cleaving/degrading residues N-terminal to the C-terminus of H1 [23].

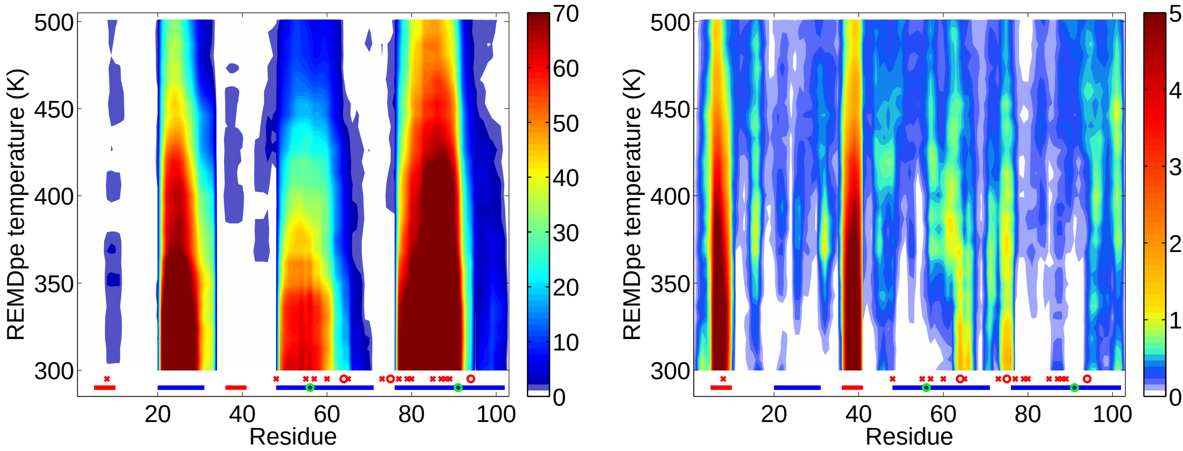


Figure 4.1: *The secondary structure propensity per residue was computed as the frequence of time spent per residue in α (left window) or β (right window) conformation at all REMDpe temperatures. The first (i) and second (ii) colored rulers just above the x-axis show, respectively: (i) Solid lines; NMR α-helices (blue) and β-sheets (red), with green circles indicating the location of the disulphide bridge forming Cys residues. (ii) Mutations favoring prion diseases [50]: Red "x" (mutations increasing hydrophobicity) and red "o" (mutations decreasing hydrophobicity). Residues were numbered starting from the first residue of the NMR PDB file.*

Surprisingly, besides mere loss of native structure at elevated temperatures, new β-sheets form occasionally. In fact, 38.3%, 6.75% and 1.29% of all structures (at all temperatures) have a β content exceeding 7, 13 and 18 residues, respectively. Remarkably, every residue of the sequence adopts a β conformation at least once during the simulation, generating a variety of β-rich folds (Figure 4.2, Panel 11) and revealing the intrinsic propensity of PrP124-226 to adopt β-rich conformations even in its monomeric form. Furthermore, the new β-sheets are essentially formed by residues belonging to H2 and H3 in the native structure (Figure 4.1), corroborating the suggestion from independent experiments that the "β-rich core", or minimal common structure of most PrP^{Sc} variants, is located in the H2-H3 sequence interval [23, 25, 26].

We collected all structures with a β content exceeding 18 residues (24'428 structures, $\sim$ 1.29% of the total) into a β-rich pool. In order to identify the major β-rich folds, we developed and applied the β *contact map clustering* (*bcmc*) protocol, presented in detail

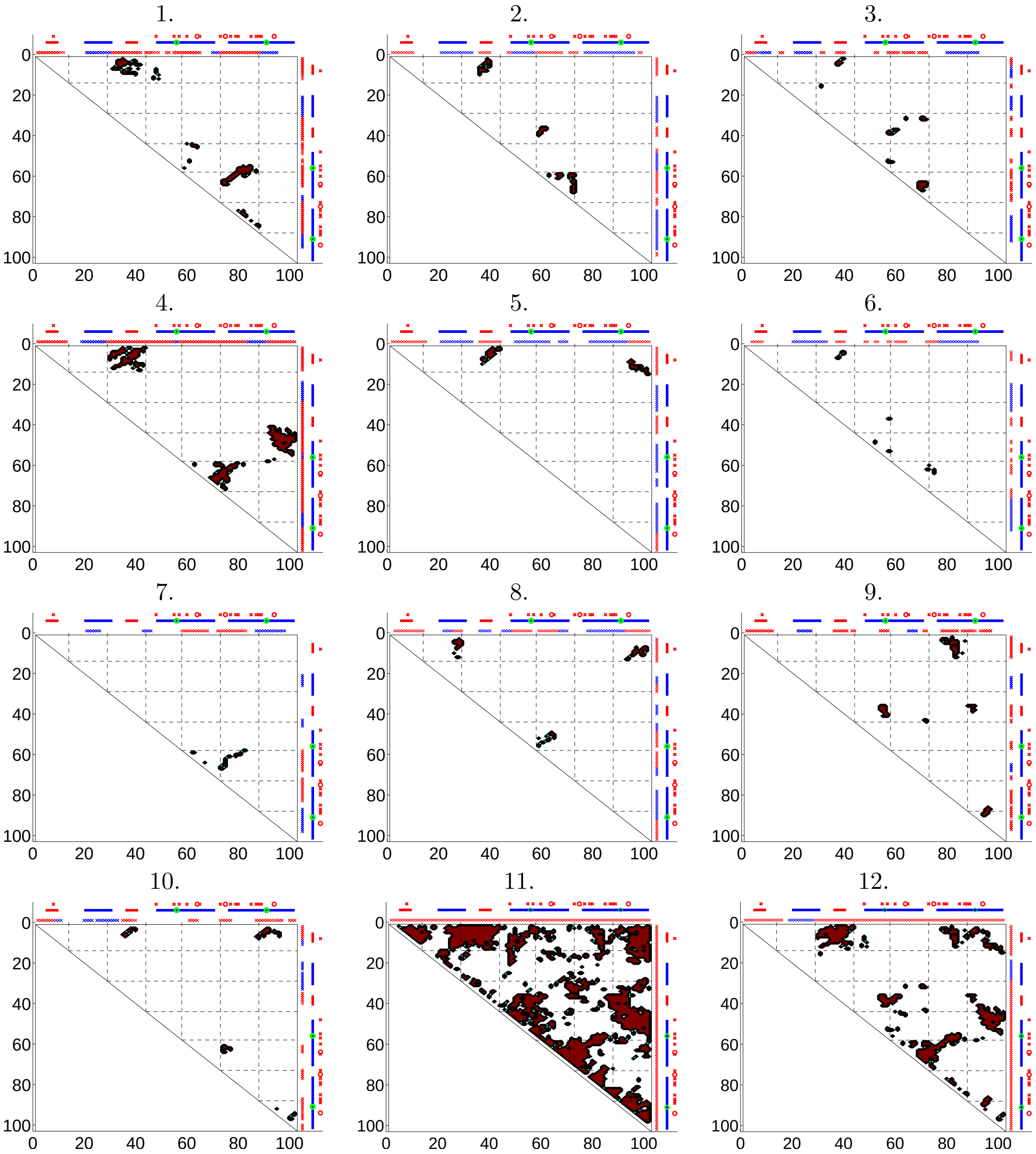


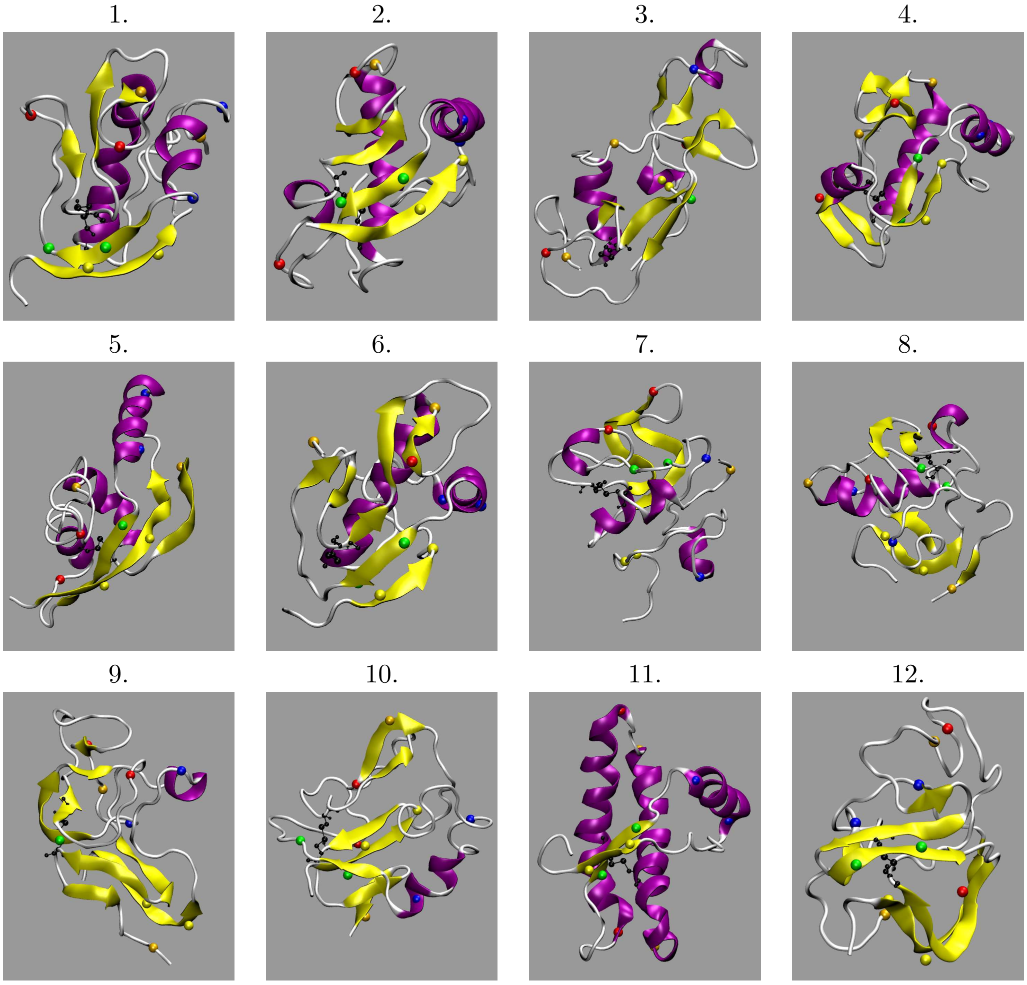
Figure 4.2: *Superpositions of the bcm of all the structures of: 1.-8.; $\alpha+$ folds (defined as containing at least one structure with ≥ 17 α residues) 1-8, 9. and 10.; $\alpha-$ folds (defined as containing no $\alpha+$ structures) 9 and 10, 11; all β-rich structures and 12; all the major fold (at least 182 members per cluster: folds 1-5, 9 and 10) structures. The 3 colored rulers above and to the right of the bcm represent, from the innermost to the outermost: (i) Red "x"; residues in β conformation, blue "x"; residues in α-helix conformation. Each line of red and blue "x" shows, for a given fold, the projection of the per-residue secondary structure of all member structures, with red "x" superposing blue ones, in order to display the maximal extent of the β-sheets. (ii) Solid lines; NMR structure α-helices (blue) and β-sheets (red), with green circles indicating the location of the disulphide bridge forming Cys residues. (iii) Mutations favoring prion diseases [50]: Red "x" (mutations increasing hydrophobicity) and red "o" (mutations decreasing hydrophobicity). β-contacts are depicted in the half bcm matrix. Dashed black lines separate the cells of the simplified bcm matrix used in the clustering. Residues were numbered starting from the first residue of the mouse NMR structure (residue 124 is our initial residue 1).*

in the Methods section. Folds were termed $\alpha+$ whenever they contained at least one structure with ≥ 17 α-helical residues (the minimal α-helical content of PrP^{Sc}) and α-otherwise. The 10 folds defined with this procedure contained 56.7% of the β-rich pool (Table 4.1) and 74.9% and 51.1% of all $\alpha+$ and $\alpha-$ structures, respectively. In Figure 4.3, one representative conformation is shown for each fold. The *bcm* of the folds are shown in Figure 4.2 and describe, for every β-strand; sequence location and hydrogen-bonding partner β-strands.

Table 4.1: *Folds obtained via bcmc clustering of the β-rich pool (24'428 structures with ≥ 19 β-residues). $\alpha+$; folds containing at least 1 $\alpha+$ structure with ≥ 17 α-helical residues, $\alpha-$; folds that only have structures with less α-helical residues, F; fold number. A minimal size of 182 members defines clusters forming the main folds (all except 6-8), while a smaller threshhold of 20 members is used to identify small $\alpha+$ clusters (containing at least 1 $\alpha+$ structure), grouped into additional folds 6-8. max β; maximum number of β-residues observed in a structure of the fold. %; fraction of β-rich pool per fold.*

	F	%	*max* β
$\alpha+$	1	17	31
	2	0.93	21
	3	8.9	25
	4	16.58	30
	5	0.77	28
	6	0.08	22
	7	0.45	21
	8	0.11	25
$\alpha-$	9	9.9	36
	10	2.19	26

All β rich structures form at high temperatures, but fold 2 and fold 4 are the only ones to accumulate at room temperature, while all other folds remain marginally stable at mid or high temperatures only (Figures 4.4 and 4.5). Fold 2 is characterized by the native β-sheet and two new β-strands; S3, formed at mid-H2 and S4, at the H2 N-terminus, forming the 4-stranded β-sheet S1-S2-S3-S4. Fold 4 is characterized by 3 distinct, non-interacting β-sheets: (i) the native S1-S2, (ii) S3, at the S2-H2 loop, hydrogen-bonded to S6, at the protein C-terminus and (iii) S4, at the N-terminus of H2, hydrogen-bonded to S5, at the H2-H3 loop. Fold 4 can also be accessed with straightforward 300K MD on a 100 ns time scale, certifying that the energetic barrier is indeed very low and reachable at low pH within the thermal fluctuations at 300K (data not shown). Furthermore, fold 4 is one of the most frequent β-rich folds (~17%, only comparable in abundance to fold 1, Table 4.1). In contrast, folds 1, 3 and 5-10 possibly never exist at room temperature, or require aggregation to larger multimers in order to stabilize. Folds 2 and 4 could therefore be monomeric precursors on the pathway to a stable and soluble PrP conformation, a β-



1. 2. 3. 4. 5. 6. 7. 8. 9. 10. 11. 12.

Figure 4.3: *Representative structures of 1.-8.; $\alpha +$ folds (defined as containing at least 1 structure with ≥ 17 α residues), 9. and 10.; $\alpha -$ folds (defined as containing no $\alpha +$ structure), 11.; NMR structure, 12. maximum β content structure obtained in simulation (38 β-residues). Helices are colored in purple, β-sheets in yellow. In order to highlight the sequence-positions of structural re-arrangements, sequence portions spanning NMR secondary structure elements are highlighted with CPK sphere representations of the C-alpha atoms of the residues delimiting S1 (yellow), S2 (green), H1 (blue), H2 (red) and H3 (orange). The figures were all oriented as the NMR structure of panel 11.*

oligomeric form that has been reported to form in the time scale of hours to days in stock solutions without prior denaturing treatment [8]. Such a form has even been suggested as the free energy minimum in aqueous solution [11, 12, 17].

The different folds are also characterized as a function of their location on the conformational landscape computed as a projection on the fraction of native contacts (Q_{fr}) and radius of gyration (R_g)(Figure 4.6). Fold 3 is an elongated conformation, containing structures with R_g as high as 1.6 nm (Figure 3, green and very rare red dots at $R_g \sim$

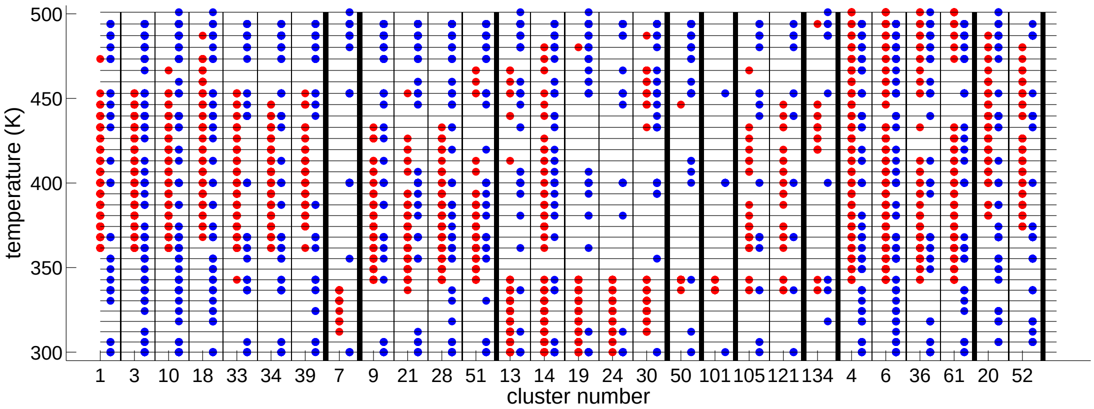


Figure 4.4: *Summary-plot presenting, for each of the 28 clusters of the 10 folds, the temperatures at which they were observed (red dots) and the replicas that sample their conformation (blue dots indicating the initial temperature of the replica). The vertical thick black lines separate the 10 folds, from fold 1 (clusters 1,3,10,18,33,34 and 39) to the far left to fold 10 (clusters 20 and 52) to the far right. Folds 2 and 4 accumulate at low temperatures. The high-temperature occurrences at cluster 13, 14 and 30 represent a small fraction of fold 4 population. All other folds accumulate at mid- to high temperatures. In addition, it appears that most clusters and folds are sampled by more than one replica.*

1.5 nm), while all other folds range from the native structure NMR R_g (1.4 nm), as fold 4, to highly collapsed structures, such as folds 1 and 10. Only folds 2, 4 and 5 contain structures with a Q_{fr} as high as 60 to 80 %. While folds 2 and 4 accumulate at lower temperature, fold 5 accumulates around 343K. In other words, the only β-rich folds that can accumulate at low temperature are native-like. The lowest average potential (total or protein) energy average is indeed found for the native structure group, as compared to corresponding values for other structural groups, such as collapsed, unfolded, etc. (data not shown). Finally, the main unfolding pathway we observe is related to a collapse of the protein (Figure 4.6), which is in agreement with the results of Kuwata et al., where high pressure unfolding is concomitant to protein collapse around molecular voids the authors identify within H2 and H3 [15]. Our simulations are performed in the NVT ensemble, building high pressures at high temperature (up to 3635 bar at 500K).

The β-rich core of any of our 10 *bcmc* folds consists in the sum of all the residues that can adopt a β conformation in at least 1 of the structures assigned to that fold and are presented in Figure 4.7. Although most of them are compatible with the recently determined location of the β-core involving H2-H3 [23, 25, 26], we have computed them from a monomer simulation and not from the fibrils used in these experiments. They could nevertheless form the seed allowing monomers to form inter-monomer β-sheets and aggregate, and are therefore a good measure of β propensity and monomer structural lability. The next step will obviously consist in assessing the aggregation potential of each one of our *bcmc* folds. Although these experimental β-rich core constraints are relatively loose, they match most of our folds, with the exception of fold 5. The presence of a conserved native

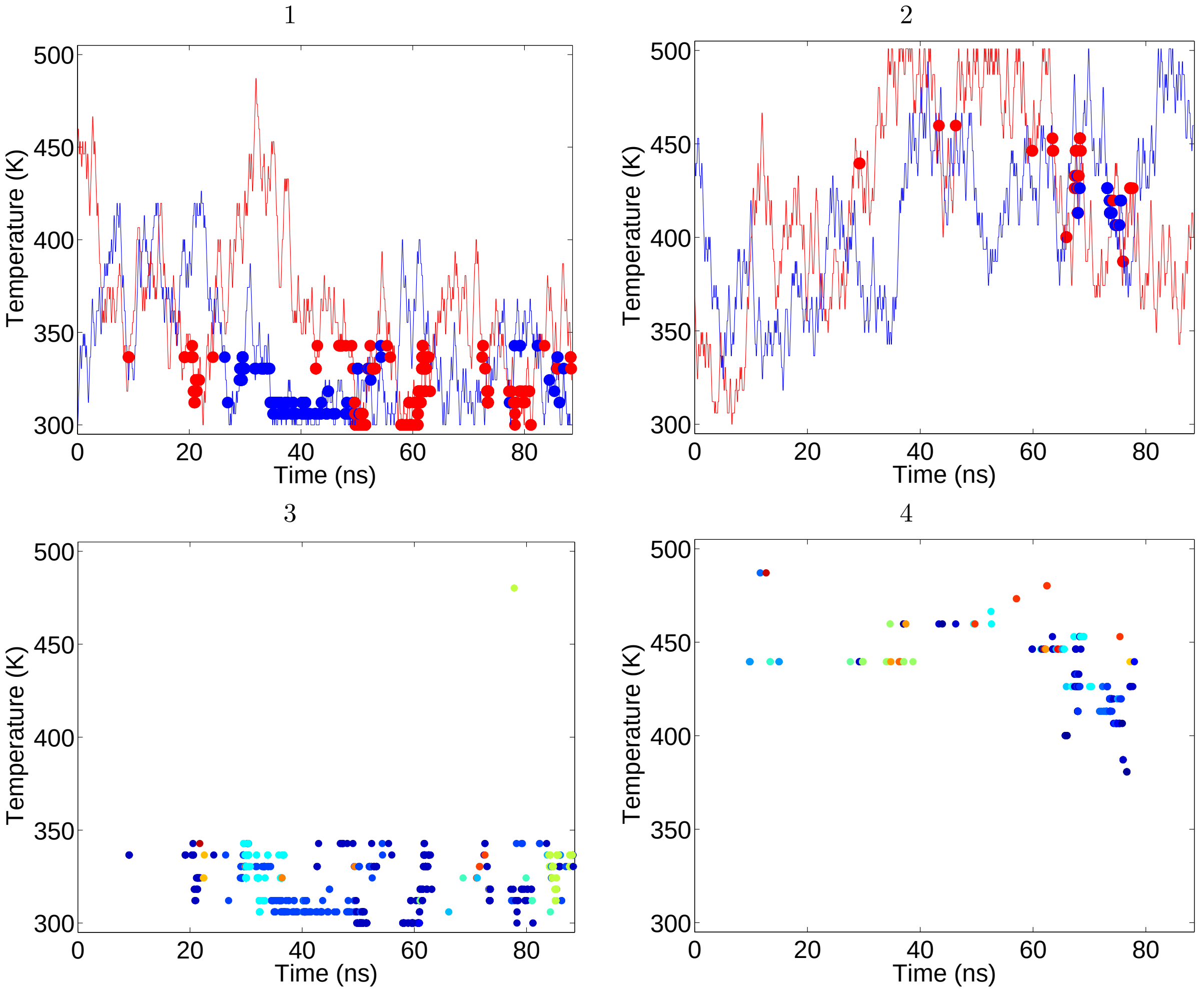


Figure 4.5: *Temperature-space trajectories of replicas converting into structures of given* bcmc *clusters. Examples of cluster 19 (fold 4, Panels 1 and 3) and 20 (fold 10, Panels 2 and 4). Panels 1 and 2 depict the trajectories (lines) followed by the two replicas yielding most structures of the cluster (dots, colored as the line representing the trajectory that generated them). Panels 3 and 4 present all structures (dots) of the clusters. Structures (dots) that were generated by a same* REMDpe *replica have the same color. Fold 4 accumulates at the lowest, physiological temperatures and is readily obtained in a 300K 315 ns reference calculation. Fold 10 never exchanges down to lower temperatures.*

S1 in 9 *bcmc* folds of 10 is not necessarily contradictory: For these residues, Lu et al. did not observe high deuterium exchange protection in a control experiment performed with PrP^{C} bearing an intact S1 [25], possibly because of its larger solvent accessible surface area.

The extent of the β-rich cores of the 10 folds also provides an idea of their maximal β content, and can be compared to experimental values. CD experiments have shown that PrP27-30 contains 47% of β structure, or 66 β residues for the 141 residue fragment [2, 4, 5]. Assuming the helical content of 17 residues found in PrP^{Sc} mainly originates from

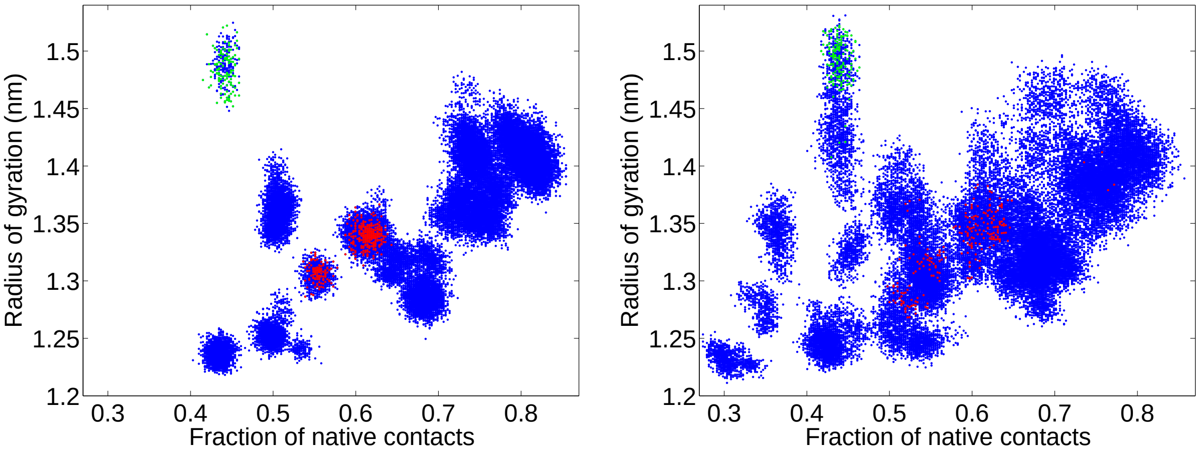


Figure 4.6: *Conformational landscape computed as a projection on the fraction of native contacts and the radius of gyration. $\alpha+$ (red dots) and $\alpha-$ (green dots) are β-rich structures (β residue content ≥ 19) with respective α-helix residue content of ≥ 17 and < 17. Left: 300K, Right: 361K.*

native helices, a conversion leaving H1 (11 residues) and H3 C-terminus (4-10 residues) intact would yield the observed α-helical content and allow for every residue of the β-rich core to adopt a β conformation. With the larger β-rich core of Cobb et al. [26], this would result in a maximal β content of 60 residues. Most of the β content observed in our simulation originates from residues of these experimental β-rich cores, and further extension of the β-sheets we observe might occur via aggregation.

The highest β contents we observe in our simulation is 38 residues for a rare $\alpha-$ structure that was not assigned to a fold (Figure 4.3 Panel 12), and 31 residues for an $\alpha+$ fold (fold 1, Table 4.1). This result is comparable to the maximal β content of 38 residues achieved in the MD simulations that lead to the spiral model [27, 28, 30]. However, the corresponding β-rich core is located in the N-terminal part of the protein, in contradiction to the experimentally found C-terminal β-rich cores. The spiral model might nevertheless reside on the conformational landscape of the monomer. With fold 5, we observe some of its features (all helices partially preserved, an extended native β-sheet and a new β-strand located in the S2-H1 loop), but this *bcmc* fold is one of the rarest we obtain and is only present at high temperature.

4.4 Conclusions

Our REMD simulation is the first MD study that extensively samples the formation of diverse new β sheets in the monomeric PrP C-terminal domain. It provides putative structural models for the recent experiments suggesting that PrP^{Sc} is characterized by a H2-H3 sequence interval β-rich core. Although rare, β-rich conformations could be sampled in

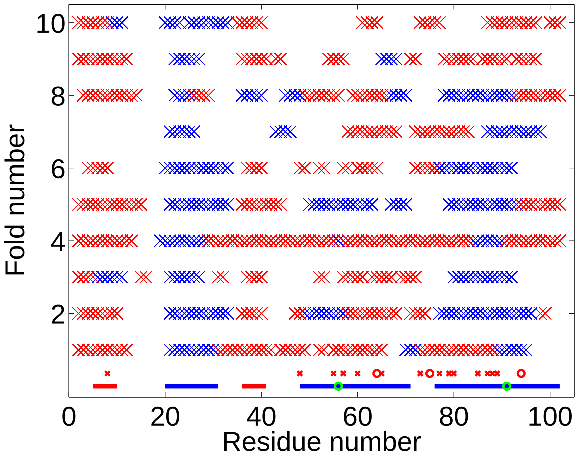


Figure 4.7: *Secondary structure of the 10 folds identified via bcmc. Red "x"; residues in β conformation, blue "x"; residues in α-helix conformation. Each horizontal line of red and blue "x" shows, for a given fold (identifier number on the y-axis), a projection of the per-residue secondary structure of all member structures, with red "x" superposing blue ones in order to display the maximal extent of the β-sheets and highlight the β-rich core. The first (i) and second (ii) colored rulers just above the x-axis show: (i) Solid lines; NMR α-helices (blue) and β-sheets (red), with green circles indicating the location of the disulphide bridge forming Cys residues. (ii) Mutations favoring prion diseases [50]: Red "x" (mutations increasing hydrophobicity) and red "o" (mutations decreasing hydrophobicity). Residues were numbered starting from the first residue of the mouse NMR structure (residue 124 is our initial residue 1).*

sufficient amount to estimate β propensity along the sequence and to allow for a very large number of different β-sheet arrangements. β-rich folds identified by the most frequently observed patterns of new β-sheets are proposed as possible models for precursors of a β-oligomeric conformation or PrP^{Sc}. Only two of these folds are found to accumulate at low temperature, showing structural stabilization and suggesting a possible relation to stable β-oligomers.

4.5 References

[1] S. Prusiner. “Novel proteinaceous infectious particle causes scrapie”. *Science*, **216**, (1982) 136–144.

[2] K. M. Pan, M. Baldwin, J. Nguyen, M. Gasset, A. Serban, D. Groth, I. Mehlhorn, Z. Huang, R. J. Fletterick, and F. E. Cohen. “Conversion of alpha-helices into beta-sheets features in the formation of the scrapie prion proteins.” *Proc Natl Acad Sci U S A*, **90**, (1993) 10 962–10 966.

[3] R. Riek, S. Hornemann, G. Wider, M. Billeter, R. Glockshuber, and K. Wüthrich. “NMR structure of the mouse prion protein domain PrP(121-321).” *Nature*, **382**, (1996) 180–182. Exp, nmr.

[4] B. W. Caughey, A. Dong, K. S. Bhat, D. Ernst, S. F. Hayes, and W. S. Caughey. “Secondary structure analysis of the scrapie-associated protein PrP 27-30 in water by infrared spectroscopy.” *Biochemistry*, **30**, (1991) 7672–7680.

[5] M. Gasset, M. A. Baldwin, R. J. Fletterick, and S. B. Prusiner. “Perturbation of the secondary structure of the scrapie prion protein under conditions that alter infectivity.” *Proc Natl Acad Sci U S A*, **90**, (1993) 1–5.

[6] L. Redecke, M. von Bergen, J. Clos, P. V. Konarev, D. I. Svergun, U. E. A. Fittschen, J. A. C. Broekaert, O. Bruns, D. Georgieva, E. Mandelkow, N. Genov, and C. Betzel. “Structural characterization of beta-sheeted oligomers formed on the pathway of oxidative prion protein aggregation in vitro.” *J Struct Biol*, **157**, (2007) 308–320.

[7] I. V. Baskakov, G. Legname, M. A. Baldwin, S. B. Prusiner, and F. E. Cohen. “Pathway complexity of prion protein assembly into amyloid.” *J Biol Chem*, **277**, (2002) 21 140–21 148.

[8] I. V. Baskakov, G. Legname, Z. Gryczynski, and S. B. Prusiner. “The peculiar nature of unfolding of the human prion protein.” *Protein Sci*, **13**, (2004) 586–595.

[9] O. V. Bocharova, L. Breydo, A. S. Parfenov, V. V. Salnikov, and I. V. Baskakov. “In vitro conversion of full-length mammalian prion protein produces amyloid form with physical properties of PrP(Sc).” *J Mol Biol*, **346**, (2005) 645–659.

[10] A. Tahiri-Alaoui and W. James. “Rapid formation of amyloid from alpha-monomeric recombinant human PrP in vitro.” *Protein Sci*, **14**, (2005) 942–947.

[11] K. Post, M. Pitschke, O. Schäfer, H. Wille, T. R. Appel, D. Kirsch, I. Mehlhorn, H. Serban, S. B. Prusiner, and D. Riesner. “Rapid acquisition of beta-sheet structure in the prion protein prior to multimer formation.” *Biol Chem*, **379**, (1998) 1307–1317.

[12] K. Jansen, O. Schäfer, E. Birkmann, K. Post, H. Serban, S. B. Prusiner, and D. Riesner. “Structural intermediates in the putative pathway from the cellular prion protein to the pathogenic form.” *Biol Chem*, **382**, (2001) 683–691.

[13] K.-W. Leffers, H. Wille, J. Stöhr, E. Junger, S. B. Prusiner, and D. Riesner. "Assembly of natural and recombinant prion protein into fibrils." *Biol Chem*, **386**, (2005) 569–580.

[14] F. Eghiaian, T. Daubenfeld, Y. Quenet, M. van Audenhaege, A.-P. Bouin, G. van der Rest, J. Grosclaude, and H. Rezaei. "Diversity in prion protein oligomerization pathways results from domain expansion as revealed by hydrogen/deuterium exchange and disulfide linkage." *Proc Natl Acad Sci U S A*, **104**, (2007) 7414–7419.

[15] K. Kuwata, H. Li, H. Yamada, G. Legname, S. B. Prusiner, K. Akasaka, and T. L. James. "Locally disordered conformer of the hamster prion protein: a crucial intermediate to prpsc?" *Biochemistry*, **41**, (2002) 12 277–12 283.

[16] J. R. Silveira, G. J. Raymond, A. G. Hughson, R. E. Race, V. L. Sim, S. F. Hayes, and B. Caughey. "The most infectious prion protein particles." *Nature*, **437**, (2005) 257–261.

[17] I. V. Baskakov, G. Legname, S. B. Prusiner, and F. E. Cohen. "Folding of prion protein to its native alpha-helical conformation is under kinetic control." *J Biol Chem*, **276**, (2001) 19 687–19 690.

[18] H. Rezaei, F. Eghiaian, J. Perez, B. Doublet, Y. Choiset, T. Haertle, and J. Grosclaude. "Sequential generation of two structurally distinct ovine prion protein soluble oligomers displaying different biochemical reactivities." *J Mol Biol*, **347**, (2005) 665–679.

[19] B.-Y. Lu and J.-Y. Chang. "Isolation and characterization of a polymerized prion protein." *Biochem J*, **364**, (2002) 81–87.

[20] E. Flechsig, D. Shmerling, I. Hegyi, A. J. Raeber, M. Fischer, A. Cozzio, C. von Mering, A. Aguzzi, and C. Weissmann. "Prion protein devoid of the octapeptide repeat region restores susceptibility to scrapie in prp knockout mice." *Neuron*, **27**, (2000) 399–408.

[21] S. Hornemann and R. Glockshuber. "A scrapie-like unfolding intermediate of the prion protein domain PrP(121-231) induced by acidic pH." *Proc Natl Acad Sci U S A*, **95**, (1998) 6010–6014.

[22] S. M. Martins, D. J. Frosoni, A. M. B. Martinez, F. G. D. Felice, and S. T. Ferreira. "Formation of soluble oligomers and amyloid fibrils with physical properties of the scrapie isoform of the prion protein from the C-terminal domain of recombinant murine prion protein mPrP-(121-231)." *J Biol Chem*, **281**, (2006) 26 121–26 128.

[23] O. V. Bocharova, L. Breydo, V. V. Salnikov, A. C. Gill, and I. V. Baskakov. "Synthetic prions generated in vitro are similar to a newly identified subpopulation of prpsc from sporadic creutzfeldt-jakob disease." *Protein Sci*, **14**, (2005) 1222–1232.

[24] W.-Q. Zou, S. Capellari, P. Parchi, M.-S. Sy, P. Gambetti, and S. G. Chen. "Identification of novel proteinase k-resistant c-terminal fragments of prp in creutzfeldt-jakob disease." *J Biol Chem*, **278**, (2003) 40 429–40 436.

[25] X. Lu, P. L. Wintrode, and W. K. Surewicz. "Beta-sheet core of human prion protein amyloid fibrils as determined by hydrogen/deuterium exchange." *Proc Natl Acad Sci U S A*, **104**, (2007) 1510–1515.

[26] N. J. Cobb, F. D. Snnichsen, H. McHaourab, and W. K. Surewicz. "Molecular architecture of human prion protein amyloid: a parallel, in-register beta-structure." *Proc Natl Acad Sci U S A*, **104**, (2007) 18 946–18 951.

[27] D. O. Alonso, S. J. DeArmond, F. E. Cohen, and V. Daggett. "Mapping the early steps in the ph-induced conformational conversion of the prion protein." *Proc Natl Acad Sci U S A*, **98**, (2001) 2985–2989.

[28] M. L. DeMarco and V. Daggett. "From conversion to aggregation: protofibril formation of the prion protein." *Proc Natl Acad Sci U S A*, **101**, (2004) 2293–2298.

[29] M. Demarco, J. Silveira, B. Caughey, and V. Daggett. "Structural properties of prion protein protofibrils and fibrils: An experimental assessment of atomic models." *Biochemistry*, **45**, (2006) 15 573–15 582.

[30] M. L. DeMarco and V. Daggett. "Molecular mechanism for low ph triggered misfolding of the human prion protein." *Biochemistry*, **46**, (2007) 3045–3054.

[31] S. Colacino, G. Tiana, and G. Colombo. "Similar folds with different stabilization mechanisms: the cases of prion and doppel proteins." *BMC Struct Biol*, **6**, (2006) 17.

[32] M. S. Shamsir and A. R. Dalby. "One gene, two diseases and three conformations: molecular dynamics simulations of mutants of human prion protein at room temperature and elevated temperatures." *Proteins*, **59**, (2005) 275–290.

[33] A. D. Simone, A. Zagari, and P. Derreumaux. "Structural and hydration properties of the partially unfolded states of the prion protein." *Biophys J*.

[34] Y. Sugita and Y. Okamoto. "Replica-exchange Molecular Dynamics Method for Protein Folding". *Chem. Phys. Lett.*, **314**, (1999) 141–151.

[35] W. Swietnicki, R. Petersen, P. Gambetti, and W. K. Surewicz. "ph-dependent stability and conformation of the recombinant human prion protein prp(90-231)." *J Biol Chem*, **272**, (1997) 27 517–27 520.

[36] D. V. D. Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. "GROMACS: fast, flexible, and free." *J Comput Chem*, **26**, (2005) 1701–1718.

[37] W. van Gunstern, S. Billeter, A. Eising, P. Huenenberger, P. Krueger, A. Mark, W. Scott, and I. Tironi. *Biomolecular Simulation: The GROMOS96 Manual and User Guide* (Hochschulverlag AG, Zuerich, 1996).

[38] H. Berendsen, J. Postma, W. van Gunsteren, and J. Hermans. *Intermolecular Forces* (Reidel, Dordrecht, 1981).

[39] J. P. Ryckaert, G. Ciccotti, and H. Berendsen. "Numerical integration of cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes". *J Comp Phys*, **23**, (1977) 327–341.

[40] J. Antosiewicz, J. A. McCammon, and M. K. Gilson. "Prediction of pH-dependent properties of proteins." *J Mol Biol*, **238**, (1994) 415–436.

[41] M. Davis and J. McCammon. "Electrostatics in biomolecular structure and dynamics". *Chem Rev*, **90**, (1990) 509–521.

[42] A. Yang, M. Gunner, R. Sampogna, R. Sharp, and B. Honig. "On the calculation of pKa in proteins". *Proteins*, **15**, (1993) 252–256.

[43] G. Vriend. "WHAT IF: a molecular modeling and drug design program." *J Mol Graph*, **8**, (1990) 52–6, 29.

[44] W. Kabsch and C. Sander. "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." *Biopolymers*, **22**, (1983) 2577–2637.

[45] T. MathWorks. *MATLAB* (2007a).

[46] G. A. F. Seber. *Multivariate Observations* (Wiley, 1984).

[47] K. Kuwata, Y. O. Kamatari, K. Akasaka, and T. L. James. "Slow conformational dynamics in the hamster prion protein." *Biochemistry*, **43**, (2004) 4439–4446.

[48] J. Ziegler, H. Sticht, U. C. Marx, W. Mller, P. Rsch, and S. Schwarzinger. "Cd and nmr studies of prion protein (prp) helix 1. novel implications for its role in the prp^{c} → prp^{Sc} conversion process." *J Biol Chem*, **278**, (2003) 50 175–50 181.

[49] R. I. Dima and D. Thirumalai. "Probing the instabilities in the dynamics of helical fragments from mouse prpc." *Proc Natl Acad Sci U S A*, **101**, (2004) 15 335–15 340.

[50] I. B. Kuznetsov and S. Rackovsky. "Comparative computational analysis of prion proteins reveals two fragments with unusual structural properties and a pattern of increase in hydrophobicity associated with disease-promoting mutations." *Protein Science*, **13**, (2004) 3230–3244.

Chapter 5

What makes doppel different?

Abstract

The discovery of the doppel (Dpl) protein has questioned the limits of the structure-function relationship paradigm of biophysics since 1999. Indeed, while Dpl and prion (PrP) proteins share an identical fold (three helices and two short β-strands), they differ in sequence (only 25% of homology), function and disease-related β-rich conformations that occur for PrP only. In a previous study, we have investigated the misfolding and rare, β-rich folds of monomeric PrP with replica-exchange molecular dynamics simulations performed at pH 4 in order to mimic an acid environment favoring disease-related structural conversions. In the present work, we perform analogous simulations for Dpl with the aim of comparing the two systems and characterizing possible specificities of PrP for misfolding and amyloidogenesis. In agreement with experiments, we find a lower thermal stability for Dpl. Extensive conformational sampling is achieved for both proteins. Surprisingly, β-rich forms are found for both proteins. However, a main difference is found in the free energy barriers leading to such conformations as well as to non-native conformations: These barriers are low for PrP and can already be crossed at 300K within a 100 ns reference MD simulation, while they are at least three times higher for Dpl. This difference suggests an intrinsic misfolding and high β-enrichment propensity for PrP as compared to Dpl. Per residue secondary structure propensities reveal that novel β-sheets of both PrP and Dpl are formed by residues belonging to the helices that are the least stable in the respective native structure: H2 and H3 for PrP (in agreement with experimental data) and H1 for Dpl. These results further corroborate experimental data obtained from PrP. Seven β-rich folds could be characterized for PrP and five for Dpl, which are clearly distinct and share only one single quasi-common fold.

5.1 Introduction

Creutzfeldt-Jakob disease (CJD), bovine spongiform encephalopathies (BSE), scrapie and other transmissible spongiform encephalopathies (TSE) are related to the misfolding of the endogenous prion protein (PrP), as revealed by the isolation [1] and sequencing [2, 3, 4] of PrP^{Sc}, an aggregated, infectious PrP isoform found in central nervous system cells of infected organisms. The insolubility of PrP^{Sc} has thwarted attempts to investigate its atomistic structure by either x-ray or NMR spectroscopy. The conversion leading from the normal, cellular PrP^{C} isoform to PrP^{Sc} is a rare event devoid of any chemical change but involving profound conformational changes from a soluble, predominantly α-helical to an insoluble, aggregated predominantly β fold [5]. Circular dichroism has indeed revealed that PrP^{C} (6% β and 43% α structure) undergoes dramatic secondary structure changes in the conversion to PrP^{Sc}, characterized by 43% β (i.e. seven times more than in the native structure) and 30% α (i.e. three times less than in the native structure)[5]. These values are respectively 47-54% β and 17-25% α for PrP27-30 [5, 6, 7], a subfragment of PrP^{Sc} obtained with proteinase K digestion via cleavage around residues 87 to 91 (depending on the PrP strain). According to the widely accepted protein-only hypothesis, PrP^{Sc} catalyzes the conversion of PrP^{C} to PrP^{Sc}.

The first experimental PrP^C structure was obtained by NMR for mouse PrP (residues 121 to 231) [8]. The 231-residue glycoprotein shows an unstructured N-terminal domain (res21-123) and a structured C-terminal domain (res124-231). The former contains four putative Cu^{2+}-binding octarepeat units [9] followed by a hydrophobic region spanning residues 89-140 that is amyloidogenic and believed to mediate neurotoxicity. PrP C-terminal domains of different species reveal highly conserved structures (Figure 5.8, Panel "p NMR"), comprising two short β-strands (S1:128-131 and S2:161-164), three α-helices (H1:144-154, H2:179-193 and H3:200-217), a disulphide bridge connecting H2 and H3 (Cys 178 and Cys 213) and two glycosylation sites (Asn181 in H2 and Asn197 in the H2-H3 loop) [8]. The structured C-terminal domain (residues 124 to 231) alone can also undergo complex misfolding conversions leading to β-sheet rich oligomers [10] and amyloid aggregates [11]. Moreover, it has been shown that it can also bind copper [12, 13, 14].

Despite numerous efforts, the physiological functions of the cellular PrP as well as the pathogenic effects of the scrapie isoform remain enigmatic. However, the ability of the PrP to bind Cu^{2+} in vitro and in vivo suggests a role in copper homeostasis [9]. Furthermore, the analysis of infected brain tissues revealed elevated amounts of free radicals, significant perturbations of metal ion levels, as well as dramatically increased oxidative damage of proteins following prion infection [15]. PrP was also shown to interact with the neural cell adhesion molecule (NCAM) at the neuron cell surface, suggesting a function in NCAM-mediated signaling [16, 17]. An alternative strategy aimed towards the identification of possible clues to the normal function of PrP is the creation of knockout mouse lines deficient for PrP ($PrP^{0/0}$). The first two $PrP^{0/0}$ lines - Zürich I [18] and Edinburgh [19] - were viable and phenotypically normal, suggesting that PrP function was not indispensable, or can be overtaken by another protein. However, animals of a third (*Ngsk* $PrP^{0/0}$, [20]) and fourth (*Rcm0* $PrP^{0/0}$, [21]) $PrP^{0/0}$ line developed late onset ataxia accompanied by Purkinje cell degeneration. A third strategy applied to the elucidation of PrP function consisted in searching for related genes. Large cosmid clones containing the PrP gene of different species were investigated, and a candidate was eventually found 16 kb downstream of the mouse PrP gene *Prnp*. The new gene, encoding a 179-residue protein with $\sim$ 25% sequence identity with all known prion proteins, was called doppel (*PrnD* gene and Dpl protein), a German synonym for "double" [21].

Like PrP, Dpl mRNA was found to be expressed during embryogenesis but, at contrast to PrP, it was found to be expressed at low levels in the adult central nervous system (CNS) and at high levels in the testis [21], where it was later shown to be indispensable in spermatogenesis and male reproduction [22, 23]. Even more intriguing was the finding that Dpl mRNA expression is upregulated in the CNS of the two $PrP^{0/0}$ lines developing late-onset ataxia and Purkinje cell degeneration but not in the two other healthy $PrP^{0/0}$ lines. A comparison of the 4 PrP knockout allele sequences finally explained why only the two former ones lead to neurodegeneration: The insertion/deletion used to construct the two former lines did not disrupt *Prnd* mRNA expression, allowing for the upregulation of expression and associated neurodegeneration, while *Prnd* expression was disrupted in the two latter, disabling this upregulation [21].

PrP knockout cells have been shown to undergo Dpl-induced apoptosis in a dose dependant manner [24]. Alternatively, Dpl toxicity might be related to oxidative stress, since Dpl was shown to induce expression of the nitric oxide synthase (NOS) [25, 15] and Dpl toxicity could be blocked by a NOS inhibitor [25]. Dpl-induced ataxia leads to a pathology that differs from the PrP^{Sc}-induced TSE and is devoid of amyloid fibrils. Surprisingly, re-introduction of *Prnp* transgenes in $PrP^{0/0}$ lines suppressed the toxic effects caused by Dpl [26], suggesting an antagonistic function of PrP to block the neurotoxicity of Dpl. The physical binding of the two proteins was demonstrated with an ELISA assay in which PrP^{C}, bound to the plate, could bind Dpl [25]. PrP and Dpl were also shown to co-patch at the plasma membrane and co-internalize in neuroblastoma cells [27].

An NMR structure was obtained for the full-length mouse Dpl26-157, intriguingly revealing a fold that is very similar to the one of the C-terminal domain of mouse PrP121-231, with the three helices and two short β-strands described above [28](Figure 5.8, Panels "d NMR“ and "p & d NMR"). Small differences can be found in the β-strands (shorter in Dpl), as well as in a kink of the second Dpl helix that contributes to a triangular hydrophobic pocket. Further similarities to PrP include the two glycosylation sites (Asn99 and Asn111, with sequence positions that differ from those of PrP and link different oligosaccharides) and a GPI anchor at Gly155 [29]. However, in comparison with PrP, Dpl contains an extra disulphide bond. The first disulphide bond (Cys109 and Cys143) is analogous to the single one of PrP (Cys178 and Cys213) and connects H2 to H3. The second one connects the S2-H2 loop to a Cys of the flexible C-terminus of the protein (Cys95 and Cys148). The NMR structure of human Dpl is very similar, and the high degree of amino acid conservation suggests structural similarity between all mammalian Dpl proteins [30]. Although the 4 PrP copper binding octarepeat domains are absent in Dpl, one or two copper binding sites were found in the loop connecting H2 and H3 [31, 32].

PrP in vitro unfolding studies have been performed with the aim of understanding the misfolding process leading to the pathogenic PrP^{Sc} isoform, highlighting a number of misfolding pathways and β-rich conformations [33, 34, 35, 36, 37, 38, 39, 40, 41]. Exploring the corresponding pathways of Dpl is a promising technique that might help to characterize the specificities of PrP misfolding and amyloidogenesis. The relative free energy ΔG^u between the native and unfolded states can be estimated at 300K by applying the modified Gibbs-Helmholtz equation [42] to temperatures of transition midpoints T_m and van't Hoff enthalpies ΔH^m at T_m obtained from thermal unfolding experiments. Surprisingly, Dpl with two disulfide bridges was found to be less stable ($T_m \sim 53-58$°C, $\Delta H^m \sim 192-205$ kJ/mol and $\Delta G^u \sim 12$ kJ/mol [43, 44]) than PrP with one disulfide bridge ($T_m \sim 70$°C, $\Delta H^m \sim 292$ kJ/mol and $\Delta G^u \sim 26$ kJ/mol [45, 46]). These results were confirmed with chemical unfolding experiments, resulting in a ΔG^u of 12.6 kJ/mol for Dpl and a ΔG^u of 19.3 kJ/mol for Prp [43, 47]. Nicholson et al. observed superprotection in H2-H3 in PrP (suggesting a partially structured unfolded state) but not in Dpl [47]. Moreover, neither the wild type Dpl, nor a single disulphide bridge mutant Dpl could be induced to exhibit the α to β transition that is typical of PrP to PrP^{Sc} conversion [43]. A slight increase of the β-content was however obtained by co-incubating Dpl and negligible amounts of PrP106-126 (too small to produce any CD spectra) [25].

In a previous study, we have applied replica-exchange molecular dynamics (REMD) to investigate PrP misfolding (Chapter 4). These simulations allowed us to identify a pool of rare β-rich structures that were clustered into 10 different folds according to the topology of the newly formed β-sheets. The aim of the present study is to apply the same simulation protocol to Dpl. The relative thermodynamic stabilities found for the two proteins are in agreement with unfolding experiments. Surprisingly, β-rich configurations are found for both proteins and not only for PrP. However, the free energy barriers separating the native structures to such states, as well as to other non-native states, are at least 3 times higher for Dpl at low temperature. The characterization of Dpl misfolding and the characterization of its β-rich folds highlights conformational changes of a structural homolog that has, to the best of our knowledge, no direct link to amyloidogenesis. Furthermore, assessing Dpl to PrP sequence differences that lead to differences in β-sheet/α-helical propensities or to different local folds in the framework of a same global fold sheds new light in the twilight zone separating standard unfolding from disease related aggregation and amyloidogenesis.

5.2 Methods

All molecular dynamics (MD) simulations were performed with the GROMACS-3.3.0 package [48], the GROMOS96 force-field [49], the SPC water model [50] and an MD timestep of 1.5 fs. The length of covalent bonds involving hydrogen atoms was constrained by the SHAKE [51] algorithm with a tolerance of 10^{-4} kJ(mol nm)$^{-1}$. Temperature was controlled with 2 Nosé-Hoover thermostats [52, 53], one for the protein and one for solvent and counter ions, with respective time-coupling constants of 0.4 and 1.6 ps. Electrostatic interactions were performed as described in Chapter 2. The starting conformation was the mouse Dpl NMR structure (res 51-157, PDB code 1I17 [28]). The solvated Dpl system was constructed analogous to our previous PrP simulations (Chapter 2). The protonation states of ionizable side chains were assigned for a pH of 4 (as for PrP in Chapter 2, where this pH was chosen in order to mimic a low pH environment favoring the PrP^{C} to PrP^{Sc} conversion [54]) by finite-difference Poisson-Boltzmann calculations [55, 56]. The DELPHI program [57] supplied with the WHATIF package [58] was used to solve the Poisson-Boltzmann equation. The protonation states of HIS, ASP and GLU were consequently set as follows: GLU4; d, ASP18; p, ASP20; d, GLU24; d, ASP38; d, GLU43; d, GLU47; p, GLU53; p, GLU70; d, GLU74; d, ASP77; p, HIS81; p, GLU91; d, HIS97; p, ASP99, p and GLU103; p (where p stands for protonated and d for deprotonated). A rhombic dodecahedron box with 18575 water molecules was constructed around the protein. 9 Cl^{-} ions were added to neutralize protein charges. In order to effectively simulate this large, explicit solvent system, the simulations were carried out with REMD using partial energy, a special protocol adapted for large, explicit solvent systems ($REMDpe$) (Chapter 2).

The equilibration of the NMR structure and of the 32 $REMDpe$ replicas were performed as described in Chapter 2. The $REMDpe$ production run was performed in the

NVT ensemble with the following temperature distribution: 300.0, 306.0, 312.1, 318.2, 324.3, 330.4, 336.6, 342.8, 349.1, 355.3, 361.6, 368.0, 374.3, 380.7, 387.1, 393.6, 400.1, 406.6, 413.1, 419.7, 426.3, 432.9, 439.6, 446.3, 453.0, 459.8, 466.5, 473.3, 480.2, 487.1, 494.0 and 500.9K. With this distribution, we obtained roughly homogeneous $REMEpe$ exchange frequencies of ∼30%. The $REMDpe$ production run was carried out for 56.4 ns (corresponding to a total time of 1.8 μs). Structures, energies and temperatures were saved every 1.5 ps. A 315 ns-long MD reference simulations was performed at 300K.

Secondary structure elements were assigned with the DSSP [59] algorithm via the GROMACS-3.3.0 interface. An in-house program was used to compute contact maps and fractions of native contacts. Two residues were considered in contact if the shortest distance between two atoms of these residues was inferior to 0.45 nm, and the fraction of native contacts of a given conformation was defined as the percentage of contacts of the NMR structure that were still present. All the other analysis were performed with GROMACS-3.3.0 [48] routines. A pool of β-rich structures was constructed with all the structures containing ≥ 19 β-residues. The β-contact map clustering (*bcmc*) protocol was applied to identify the main β-rich folds. This protocol is described in Chapter 4, where it was used to cluster PrP β-rich folds with the same clustering parameters.

5.3 Results and Discussion

5.3.1 Reference MD simulations

315 ns reference MD simulations were performed for both proteins PrP and Dpl. Dpl shows a very high stability, characterized by the conservation of all native secondary structure elements, and remains unchanged throughout the simulation (Figure 5.1, left Panel). At stark contrast, in the PrP simulation helices partially unfold between 45 and 200 ns, yielding a new misfold that is stabilized by 3 β-sheets (2 new β-sheets adding up to the conserved native one)(Figure 5.1, right Panel). H3 C-terminal residues begin unfolding around 45 ns and form the first new β-sheet involving S2-H2 loop residues. H2 C-terminal residues begin unfolding around 92 ns and form the second new β-sheet with H2-H3 loop residues. H2 further unfolds until 200 ns, after which the new misfold remains unchanged until the end of the simulation. We have chosen to define native structure as structures with a fraction of native contacts $\geq$ 64%. With an average fraction of native contacts around 65%, this misfold is close to non-native. The low stabilities observed for H2 and H3 are in agreement with experimental data [41, 60, 61, 62] and simulations [63, 64].

Although this conversion takes place on a $\sim$ 50 ns time scale, it is a single event observed in one single trajectory, and the $REMDpe$ simulations should allow to provide more statistics for this misfold. Indeed, it was also observed in the PrP $REMDpe$ simulation (PrP β-rich fold p4), where it forms at high temperature, as all other β-rich folds, but is one of the only 2 β-rich folds that accumulate at lower temperature (Chapter 4). Thus, although rare, this PrP fold can be obtained by both thermal fluctuations at 300K (after 100-200 ns) and as lowest energy β-rich fold in $REMDpe$, suggesting it is indeed a low energy state that might accumulate at certain conditions. Therefore, it supports the notion of relatively low free energy barriers separating folded from misfolded PrP configurations at 300K. At contrast, the Dpl 300K reference MD simulation can not access any misfolded conformations, suggesting a much higher kinetic stability. Consequently, non-native configurations of Dpl can only be observed with enhanced sampling methods such as *REMDpe* simulations.

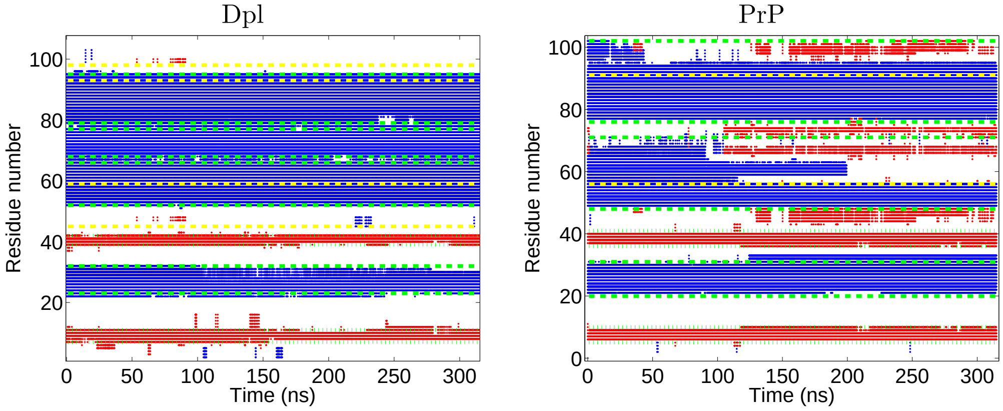


Figure 5.1: *Dpl and PrP 300K MD reference simulations: Secondary structure (DSSP) of all residues as a function of time. Color coding for the secondary structure: Red; β-sheet and blue; α-helix. Native (NMR) secondary structure element locations are delimited by dashed and dotted green horizontal lines: Dotted; β-sheets (in sequence order: S1 and S2), dashed; α-helices (in sequence order: H1, H2a, H2b and H3 for Dpl and H1, H2 and H3 for PrP). Yellow dashed horizontal lines: Cys forming disulfide bridges. Residues were numbered starting from the first residue of the NMR PDB file.*

5.3.2 *REMDpe* simulations

For a more comprehensive picture of the overall conformational landsccape of the two systems, we performed *REMDpe* simulations. Thus, the sampling of the conformational space of both proteins could be extended to new regions, as revealed by the probability distributions for the 0-56.4 ns simulation time interval (entire length of Dpl simulation and $\sim$ 2/3 of the PrP simulation) projected on the fraction of native contacts and the radius of gyration (Figure 5.2). The high temperature probability distributions of PrP and Dpl are very similar (Figure 5.2, Panel 4) and converge to a new, non-native and collapsed high probability region. At contrast, the 300K probability distributions differ for the two proteins: While PrP reveals a succession of connected high probability regions leading from the native state to a non-native, collapsed one (Figure 5.2, Panels 1 and 3), the corresponding Dpl high probability regions are totally separated (Figure 5.2, Panels 2 and 3). This indicates high free energy barriers that are only crossed with the help of the enhanced sampling provided by *REMDpe*. In particular, the native Dpl population is isolated by such barriers on the free energy surface (FES), suggesting that the protein must be particularly stable at 300K, in the absence of denaturing conditions (high temperature, detergents, etc.). We estimate that the Dpl barriers are 3 times higher than the corresponding barriers of PrP. Thus, in agreement with the reference MD 300K simulations, the *REMDpe* 300K simulations also show a higher kinetic stability for Dpl compared to PrP.

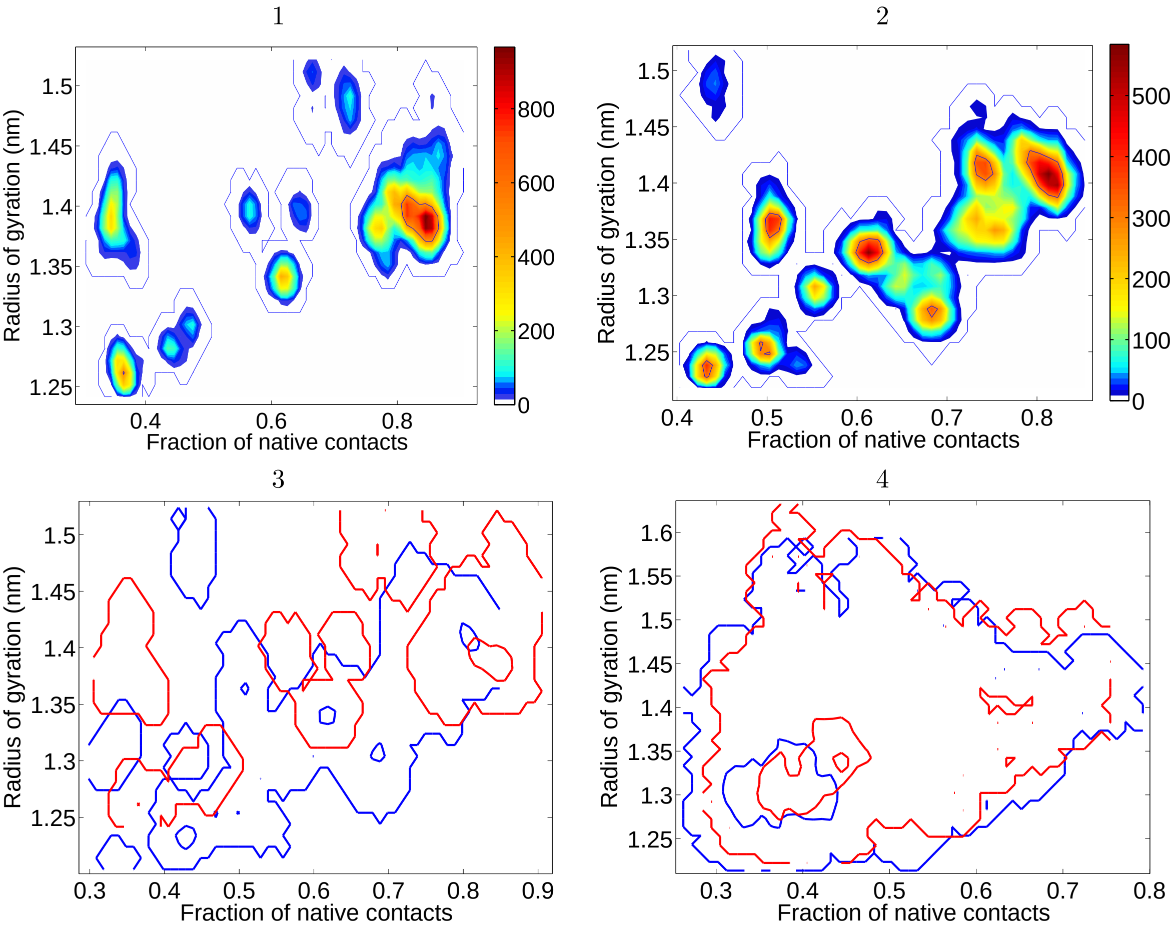


Figure 5.2: *Probability distributions computed as a function of the fraction of native contacts and the radius of gyration for the 0-56.4 ns trajectory intervals of the PrP and Dpl* REMDpe *simulations (entire length of Dpl simulation and ∼ 2/3 of the PrP simulation). Panel 1: Dpl 300K, Panel 2: PrP 300K. Colorbars in Panels 1 and 2 report the number of structures. Two solid countour lines are added to Panels 1 and 2: an outer one to encompass the maximal extent of non zero probability regions (≥ 1 count) and an inner one at half of the maximum count observed in order to highlight high probability regions. Analog contour lines are drawn in Panels 3 (300K) and 4 (500K) in order to compare the superposed probability distributions of Dpl (red) and PrP (blue).*

We also compared the thermal stability of the two proteins, assessed by the fraction of native versus non-native structures present at different *REMDpe* temperatures. For both proteins, native structures were defined as structures with ≥ 64% native contacts of the respective NMR structure. The average fractions of native structures per temperature show a higher thermal stability for PrP (Figure 5.3), in agreement with experiments [43, 44, 45, 46, 47]. An exception is found at at very low temperature ($\sim$ 300K), for which the fraction of native structures are slightly lower for PrP than for Dpl, due to the fact that PrP can loose native-like structure at very low temperature in favor of β-enriched conformations. The global picture emerging from our data shows a higher thermal stability for PrP, in agreement with experiments, and a higher kinetic stability for Dpl at lower temperature. The lower kinetic stability of PrP at 300K allows for alternative folds to occur under normal conditions and provides clues to interpret the experimental observation of a variety of PrP misfolds at physiological conditions (i.e. β-oligomers observed in-vitro and in the absence of denaturing conditions [65]).

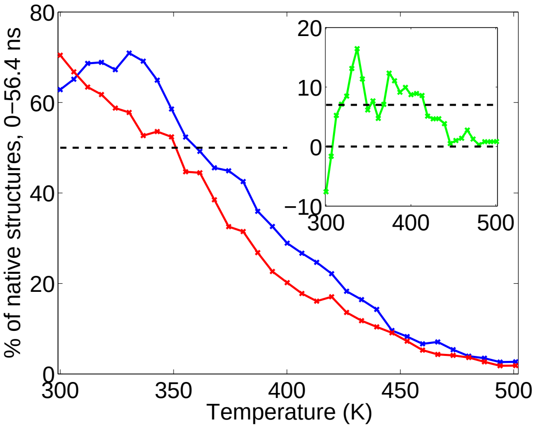


Figure 5.3: *Average fractions of native structures, at different temperatures computed in the time interval 0 - 56.4 ns of the PrP (blue) and Dpl (red) simulation (entire length of Dpl simulation and $\sim$ 2/3 of the PrP simulation). Dark dashed line; 50% of native structures. Inset: Differences between the fractions of native structures for PrP and Dpl at different temperature. Lower black dashed line: 0%, Upper one: 7% (average difference in the 300-439K temperature range). Structures with a threshold of $\geq$ 64% of the native contacts are defined as native.*

Surprisingly, besides mere loss of native structure at elevated temperatures, new β-sheets form at rare occasions for both proteins: Of all the sampled conformations (at all temperatures), 1.3% of PrP and 2.7% of Dpl structures have a β-content $\geq$ 18 residues. Although β-structure enrichment was not observed in Dpl thermal unfolding studies [43], β-rich conformations might be favored by other denaturing conditions. Indeed, a number of proteins that are not involved in amyloidogenesis related diseases show β-rich states under a variety of denaturing conditions [66, 67, 68, 69, 70]. Furthermore, the structural similarity of Dpl and PrP also suggests that β-rich states are plausible for Dpl, and that their absence at physiological conditions is in fact due to the difference in barrier height separating native from unfolded conformations, as suggested by our 300K reference MD and *REMDpe* simulations. Interestingly, the location of the β-rich structures on the conformational landscape projected on the fraction of native contacts and radius of gyration reveal that β-rich configurations cluster into a number of different well distinct misfolds (Figures 5.4 and 5.5).

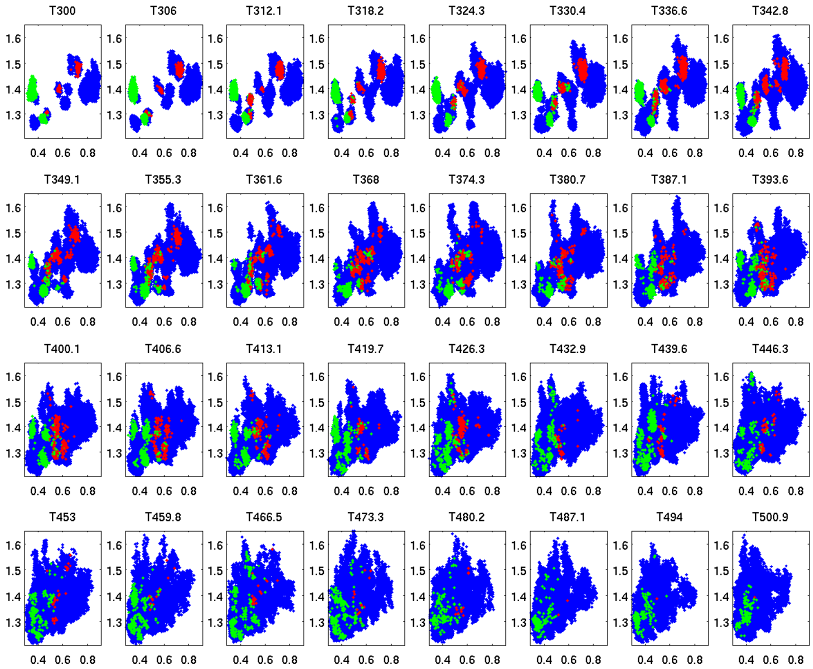


Figure 5.4: *Location of Dpl β-rich structures (≥ 19 β-residues) on the conformational landscape obtained as a projection on the fraction of native contacts (x-axis) and the radius of gyration (y-axis), for all temperatures. Red and green dots indicate β-rich structures with respectively ≥ and < 17 α-helical residues, while blue dots show all other structures.*

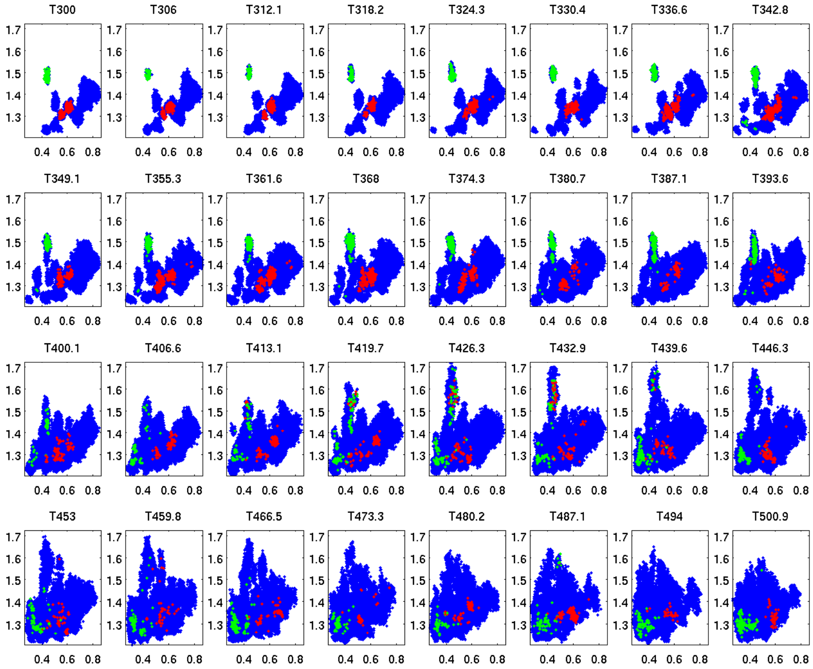


Figure 5.5: *As Figure 5.4, for the 0-56.4 ns interval of the PrP simulation (corresponding to the total length of the Dpl simulation).*

5.3.3 Secondary structure propensities

In order to compare the secondary structure propensity of the PrP sequence to that of Dpl, secondary structure propensities along the sequence were defined by the fraction of simulation time that each residue spent in respectively α-helical and β-sheet conformations, at a given temperature (Figure 5.6). For PrP, H2 is found to be the least stable helix at all temperatures, followed by H3 and H1, consistent with our 300K reference MD simulation (Subsection 5.3.1), experimental stability studies [41, 60, 61, 62] and predictions from other simulations [63, 64]. These results suggest that H1 might remain intact in PrP^{Sc} and constitute part of the minimal α-helical content of 17 residues observed in circular dichroism experiments of PrP27-30 [5, 6, 7]. This is also supported by the decrease of α-helical content and unchanged β-content observed upon PK digestion of recombinant PrP27-30 amyloid leading to a smaller protease resistant core and cleaving/degrading residues N-terminal to the C-terminus of H1 [71]. For Dpl, the pattern differs strongly; in fact H1 is the least stable helix, followed by H2 and H3. Salt bridges have been shown to stabilize H1 in PrP [72], and are indeed observed in the simulation (data not shown). These salt bridges are absent in Dpl H1, providing a possible rationale for its lower stability.

Also the location of the newly formed β-sheets differs for the two proteins (Figure 5.6). In PrP, β-sheets are mainly formed by residues belonging to H2 and H3 in the native structure. This sequence interval contains most of the disease promoting mutations [73], as well as the "β-core" of residues for which three different independent experiments suggest an involvement in a PrP^{Sc} β-sheet scaffold [71, 62, 74]. New β-sheets in Dpl are mainly formed in the sequence interval delimited by the protein N-terminus and the native β-strand S2, where residues of the unstable H1 helix become available, at contrast to PrP, where these residues remain in helical conformation most of the time.

5.3.4 β-rich folds

In Chapter 4, we have introduced the *bcmc* protocol, developed to identify the main β-rich folds in the β-rich pool of structures with at least 19 residues in β-conformation in the PrP *REMDpe* simulation. In the present work, we apply an identical *bcmc* protocol to the β-rich pool observed during the Dpl *REMDpe* simulation, and compare main β-rich folds (obtained from *bcmc* clusters with a population of at least 182 members) of PrP and Dpl. Figure 5.7 shows the *bcm* of the main PrP and Dpl folds, while representative structures are shown in Figure 5.8. Seven, respectively five main β-rich folds were found for PrP and Dpl. The seven main PrP folds have been discussed previously in relation to recent PrP misfolding experiments (Chapter 4). A comparison of the PrP β-rich folds with those of Dpl shows that there are very little common trends: Only Dpl fold 2 (d2) and PrP fold 4 (p4) resemble each other to some degree.

p4 is presented in subsection 5.3.1, where we show that it is also formed during the PrP reference MD simulation at 300K and that it is one of the two only PrP β-rich folds that accumulates at low temperature. As the NMR, the d2/p4 common structural motif

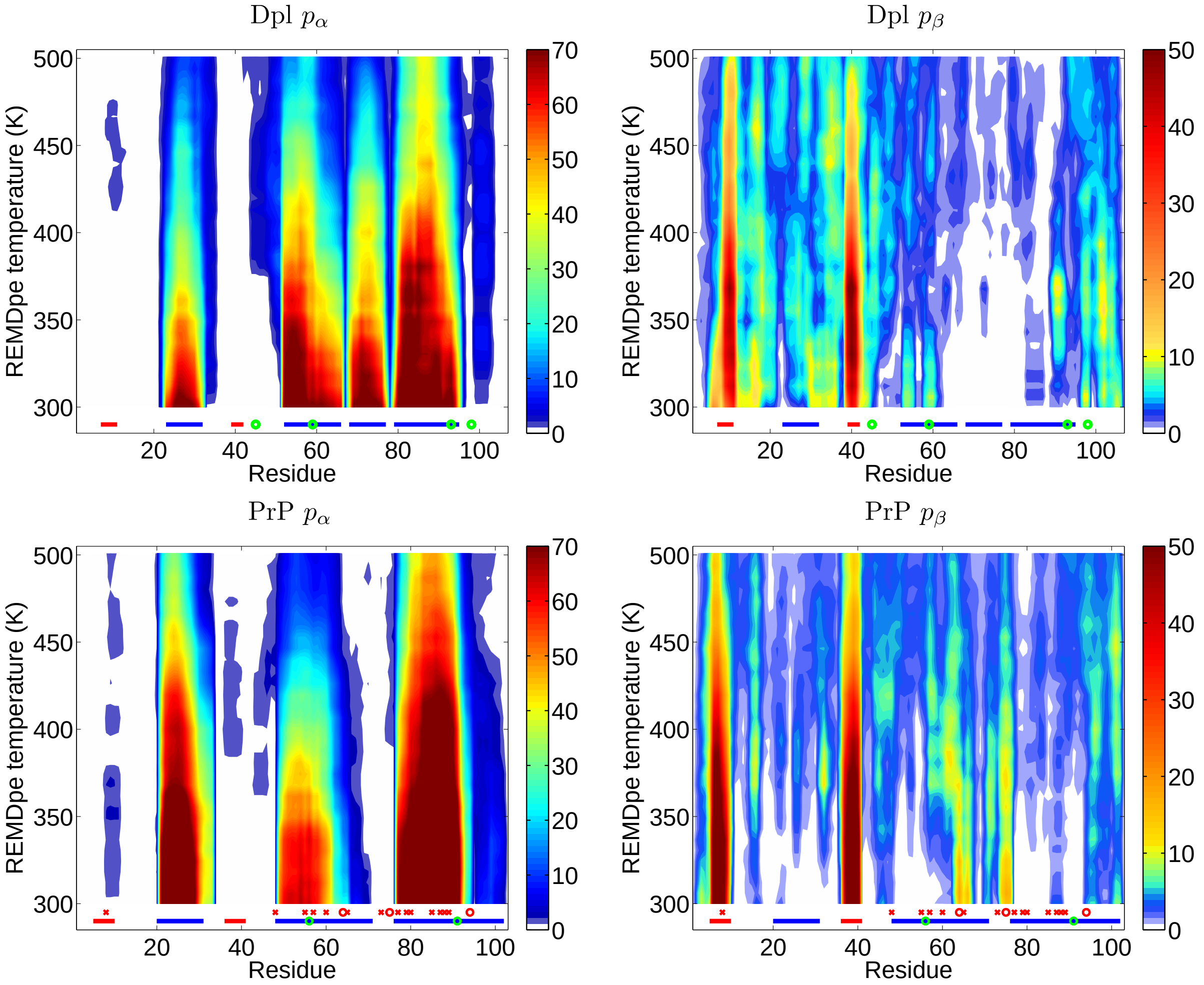


Figure 5.6: *Per residue α-helical propensity (p_α) and per residue β sheet propensity (p_β), computed as the % (colorbar) of time spent by each residue at different REMDpe temperatures in α/β conformation. The first (i) and second (ii) colored rulers just above the x-axis show: (i) Solid lines; NMR α-helices (blue) and β-sheets (red), with green circles indicating the location of the disulphide bridge forming Cys residues. (ii) Mutations favoring prion diseases [73]: Red "x" (mutations increasing hydrophobicity) and red "o" (mutations decreasing hydrophobicity) (shown for PrP only). Residues were numbered starting from the first residue of the NMR PDB file.*

can therefore be formed by two sequences that only share 25% of homology. In Chapter 4, this stability lead us to suggest that p4 might be related to a precursor of the β-oligomeric form, a stable and soluble PrP conformation that has been reported to form in the time scale of hours to days in stock solutions without prior denaturing treatment [34]. It has even been suggested that this structure corresponds to the free energy minimum in aqueous solution [37, 38, 65]. Although we do not have sufficient experimental guidelines to allow for a clear identification of a PrP β-rich fold as the most likely PrP^{Sc} monomeric precursor form, the finding that Dpl can access to a p4-like fold provides further reasons to believe that p4/d2 is not related to the pathway leading to PrP^{Sc}.

Table 5.1 summarizes the simulation length, size of the β-rich pool, total number of different β-contacts and number of *bcmc* folds obtained by $REMDpe$ simulation for PrP and Dpl. The fraction of all simulated structures that form the β-rich pool of Dpl is 2.7% for 56.4 ns of simulation, practically the double of the corresponding fraction in PrP (1.3%), obtained in 88.4 ns of simulation. Despite this difference, PrP allows for a larger diversity of β-sheet arrangements, with more different β-contacts and β-rich folds. Although this result only applies to β-rich conformations, it is in agreement to previous simulations comparing Dpl and PrP. Colacino et al. showed that the network of interactions stabilizing the native fold of PrP was already disrupted at 350K, allowing for multiple misfolding pathways, while the corresponding network in Dpl was stronger, suggesting a limited number of misfolding pathways [77]. The simulations of Settiani et al. also suggest that there are more misfolding pathways available to PrP than to Dpl [78].

In chapter 4, we also analyzed whether the PrP β-rich folds contained a sufficient amount of α-helical residues in order to be consistent with experimental determinations of the α-helical content of PrP^{Sc}. β-rich folds containing at least one structure with ≥ 17 α-helical residues were termed $\alpha+$, while β-rich folds with no such structure were termed α-. Although such a distinction is irrelevant for Dpl, most of its β-rich structures belong to a α- fold, contrarily to PrP (Table 5.2). This finding is consistent with the thermal stability (related to the helical content), which was found to be lower for Dpl than for PrP (subsection 5.3.2). It also supports the idea that PrP can progressively unfold into β-rich states, that are therefore more likely at normal conditions than for other proteins, while Dpl would need to overcome a high energy barrier (as shown in Figure 5.2) that disrupts most of the helical content to access to a β-rich state.

Figure 5.9 shows a superposition of all the Dpl and PrP *bcm*s from (i) all the structures in the β-rich pool (Panel 1) and (ii) all the structures that were assigned to a *bcmc* fold (Panel 2). Panel 2 shows stable and frequent occurring β-sheets that are present in Dpl and PrP, or in one of the proteins only, and define their β-rich folds. We refer to the regions of the *bcm* with sequence intervals in which residues are numbered starting from the first residue of the NMR PDB file. Dpl only shows one region of the *bcm* that cannot be accessed by PrP. This region contains all the β-sheets that are formed by residues of the sequence interval 15-35, hydrogen-bonding to residues of the same interval, and involves residues of the unstable H1 (subsection 5.3.3). H1 hardly unfolds in PrP β-rich folds and consequently such β-sheets cannot form. In comparison, there are 3 regions of the *bcm* that only PrP β-rich folds can access: (i) Sequence intervals 5-20, hydrogen-bonded to sequence interval 80-95, (ii) 55-65, hydrogen bonded to 75-95, and (iii) 15-35, hydrogen bonded to 60-80. The first two regions are related to unfolded H3 in PrP, while H3 hardly unfolds in Dpl β-rich folds. Thus, the stability of the helices, related to the sequence (e.g. salt bridges stabilizing H1 for PrP) appears as a determinant of the β-propensity and β-folds that can be formed. This hypothesis is supported by the Dpl and PrP simulations of Colacino et al., in which “native” residue interaction cores contribute to the stabilization of the native fold and are mainly formed by H1 and H3 residues for PrP and H2 and H3 residues for Dpl [77].

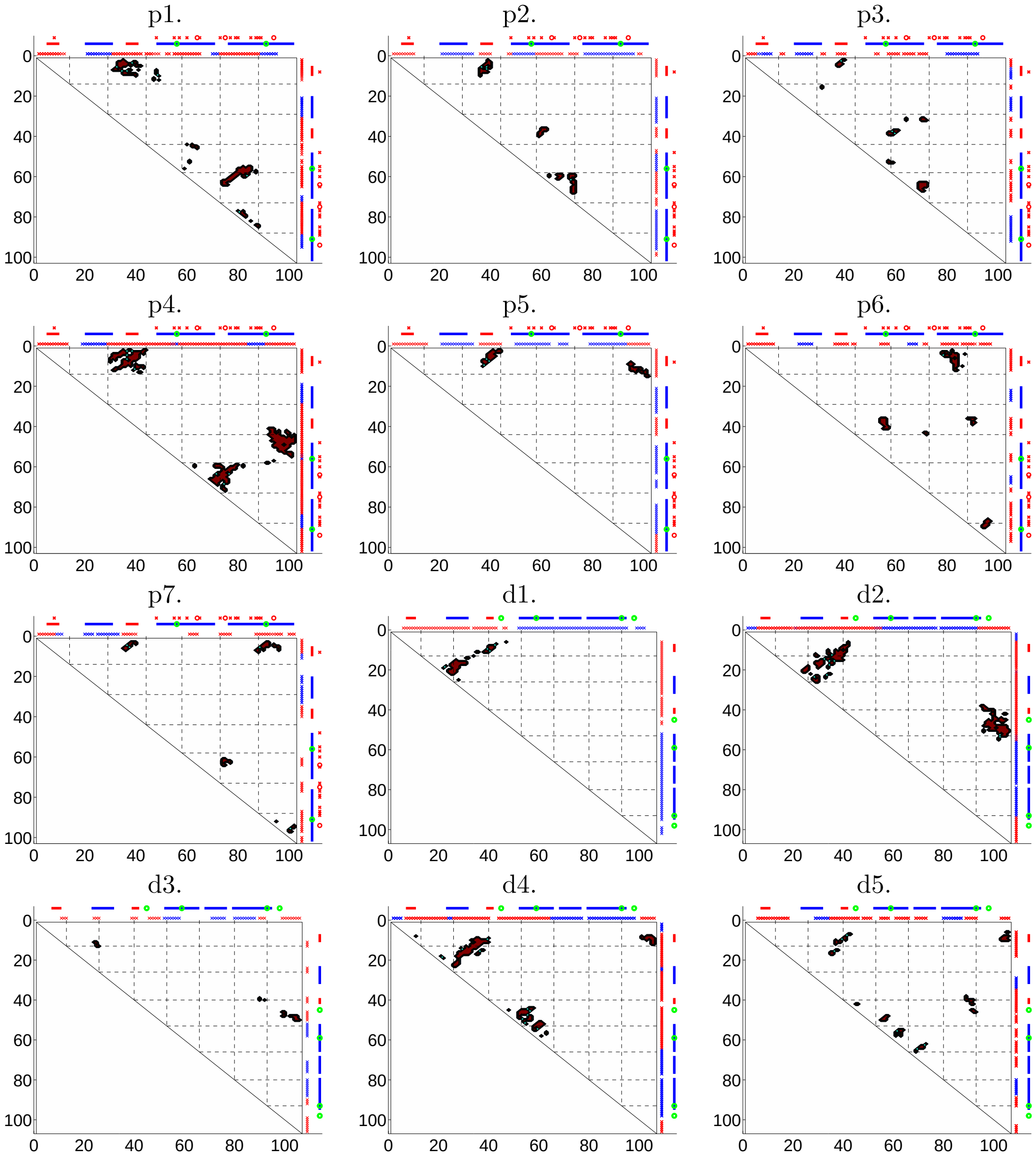


Figure 5.7: *Superposition of the* bcm *of all structures of the main* bcmc *folds of PrP (p1 to p7) and Dpl (d1 to d5). The black dashed lines delimit the sequence intervals used in the* bcmc *procedure (see section 5.2). The colored bars at the top and right hand side of the plots depict, from the innermost (i) to the outermost (iii): (i) Red "x" (β-sheet) and blue "x" (α-helix), showing a superposition of all the per-residue secondary structure conformations of all the structures of the fold (with red "x" systematically superposing blue ones), (ii) native structure α-helices (blue solid lines) and β-sheets (red solid lines), green circles; location of the disulphide bridge forming Cys residues and (iii) Mutations favoring prion diseases [73]: Red "x" (mutations increasing hydrophobicity) and red "o" (mutations decreasing hydrophobicity) (shown for PrP only). Residues were numbered starting from the first residue of the NMR PDB file.*

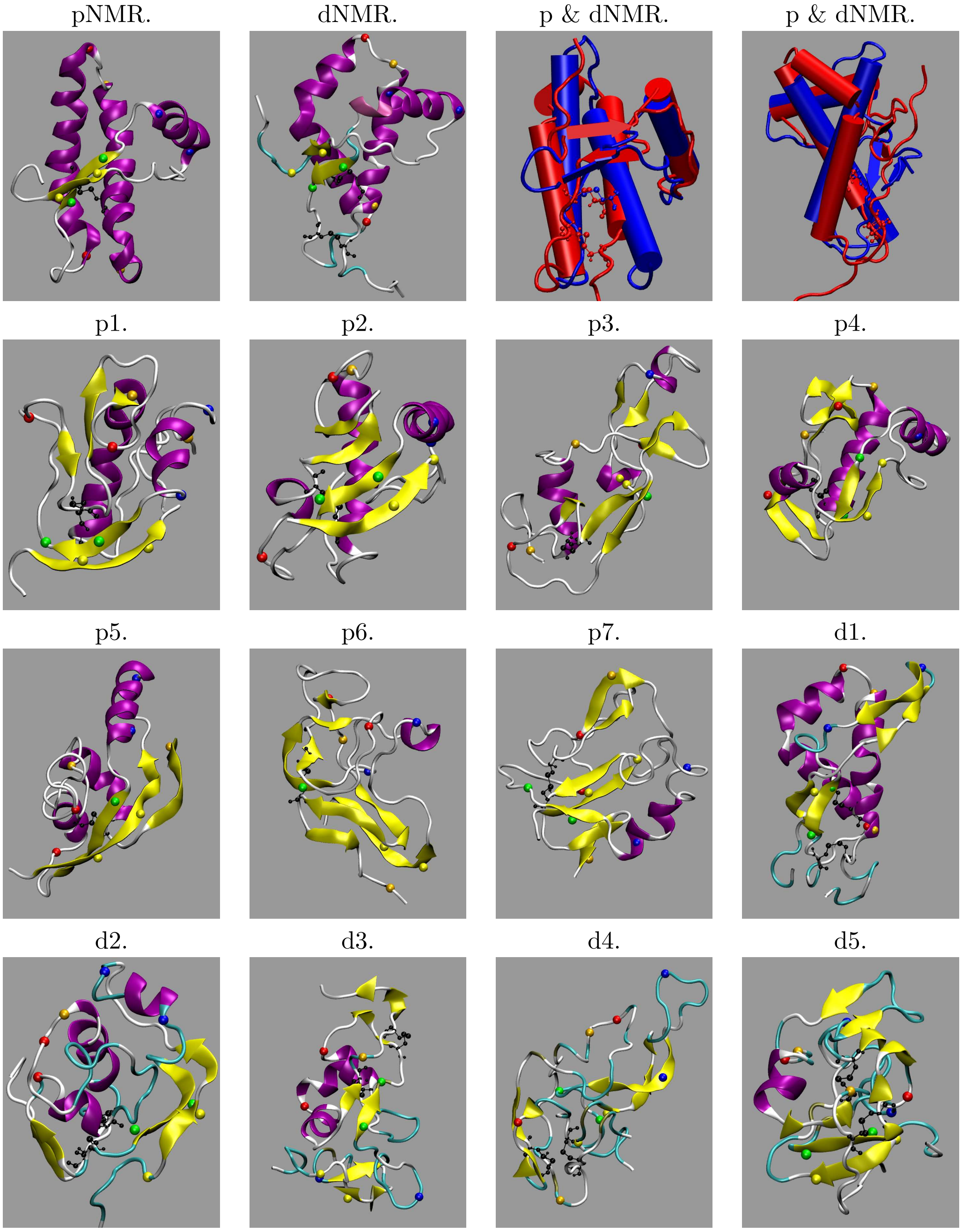


Figure 5.8: *Typical structures of the main* bcmc *folds of PrP (p1-p7) and Dpl (d1-d5). Panels pNMR and dNMR show the NMR structures of PrP and Dpl. The two p & dNMR panels both show the structural superposition (90° rotated views) of these NMR structures (blue; PrP, red; Dpl), computed with the STAMP [75] combined structure superposition-sequence alignment algorithm implemented in VMD [76]. In all panels except the two p & dNMR ones: (i) Helices are colored in purple and β-sheets in yellow, (ii) in order to highlight the sequence-positions of structural rearrangements, sequence portions spanning NMR secondary structure elements are highlighted with sphere representations of the C-alpha atoms of the residues delimiting S1 (yellow), S2 (green), H1 (blue), H2 (red) and H3 (orange), (iii) the disulphide bridge forming Cys residues are shown with black sphere representations for all the atoms.*

Table 5.1: *Comparison of the Dpl and PrP simulations: Total time, number of different β-contacts, size of the β-rich pools and % thereof assigned to main folds.*

	Dpl (107 res)	Prp (103 res)
Total simulation time (ns, per replica)	56.4	88.4
% of all structures in β-rich pool	2.7	1.3
Number of structures in β-rich pool	32746	24428
Total number of different β contacts in β-rich pool	1378	1609
% of β-rich pool in main *bcmc* folds	66	56
Total number of different β contacts in main *bcmc* folds	315	469
Number of main *bcmc* folds	5	7

Table 5.2: *Fractions (%) of β-rich pools found in the main* bcmc *folds (F) of Dpl and PrP. α+ refers to the folds that contain at least one structure with $\geq$ 17 α-helical residues and α-, to folds that comprise no such structure.*

	Dpl		Prp	
	F	%	F	%
α+	1	5.7	1	17
	2	5.5	2	0.9
	3	0.7	3	8.9
			4	16.6
			5	0.8
α-	4	39.9	6	9.9
	5	14.9	7	2.2

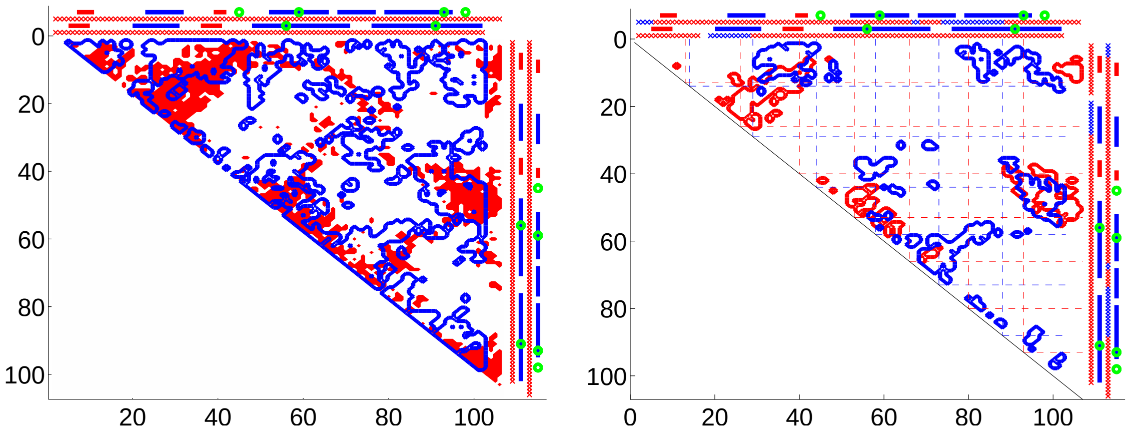


Figure 5.9: *Superposition of all Dpl (red) and PrP (blue) β-rich structure* bcm *(left Panel) and of all those thereof that were attributed to main* bcmc *folds (right Panel, same colors). The four colored bars at the top and right hand side of the plot depict, from the innermost (i) to the outermost bar (iv): (i) and (iii); Red ”x” (β-sheet) and blue “x“ (α-helix) showing a superposition of all the per-residue secondary structure conformations (with red ”x“ systematically superposing blue ones) for PrP (i) and Dpl (iii), (ii) and (iv); Solid lines; NMR α-helices (blue) and β-sheets (red), with green circles indicating the location of the disulphide bridge forming Cys residues for PrP (ii) and Dpl (iv). The red (Dpl) and blue (PrP) dashed lines of the right Panel delimit the sequence intervals used in the* bcmc *procedure (see Section 5.2). Residues were numbered starting from the first residue of the NMR PDB file.*

5.4 Conclusions

In the present work, *REMDpe* simulations of misfolding and of rare β-rich conformations of PrP and of its non-pathogenic structural homolog Dpl are compared, with the aim of highlighting PrP-specific misfolding characteristics that might relate to PrP pathologies. In agreement with experiments, we find a higher thermal stability for PrP than for Dpl. However, for Dpl, the free energy barriers leading to non-native and β-rich states are at least 3 times higher than for PrP, suggesting a higher kinetic stability for the former. Indeed, although both proteins can access β-rich conformations via thermal misfolding (high *REMDpe* temperatures), only PrP can readily convert into the β-rich misfold p4 via long ($\sim$ 100 ns) straightforward reference MD simulations at physiological temperature (300K), whereas β-rich folds can only be observed in enhanced sampling simulations of Dpl. This difference suggests an increased intrinsic misfolding and β-enrichment propensity for PrP compared to Dpl.

The "β-cores“ observed in the β-rich folds for both PrP and Dpl are formed by residues belonging to the helices that are the least stable in the corresponding native structures: H2 and H3 for PrP and H1 for Dpl. Thus, the stability of the helices, related to the sequence (e.g. salt bridges stabilizing H1 for PrP) appears as a determinant of the β-propensity and β-folds that can be formed. Seven β-rich folds are found for PrP and five for Dpl, with one single quasi-common fold, p4/d2, that accumulates at low temperature in the PrP *REMDpe* simulation and is also formed in the PrP 300K reference MD simulation. This stable β-rich misfold is therefore accessible to two different amino-acid sequences, suggesting a sequence-independent stabilization process and a possible relation to soluble PrP β-oligomers formed at certain experimental conditions and found to be even more stable than the native structure [37, 38, 65]. Finally, the fact that there are practically no common PrP/Dpl β-rich folds suggests that if Dpl β-rich folds are at all possible under physiological conditions (which is in contradiction to the high kinetic stability we find), they are not related to amyloidogenic pathologies. At contrast, one or more of the PrP specific β-rich folds may represent PrP* or a PrP^{Sc} monomeric state.

5.5 Appendix: Sequence alignments

5.5.1 Structure based sequence alignment of PrP of different species

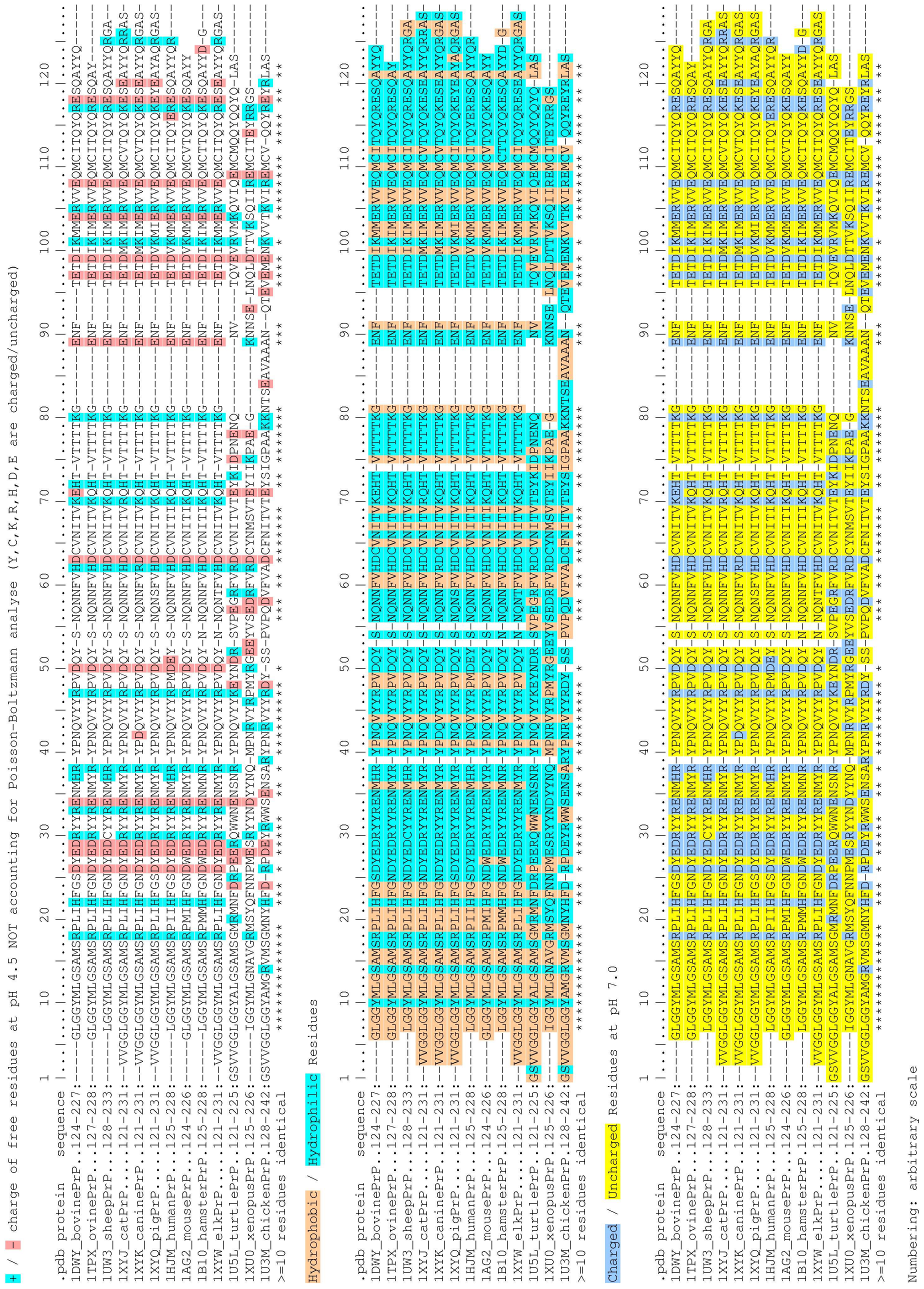
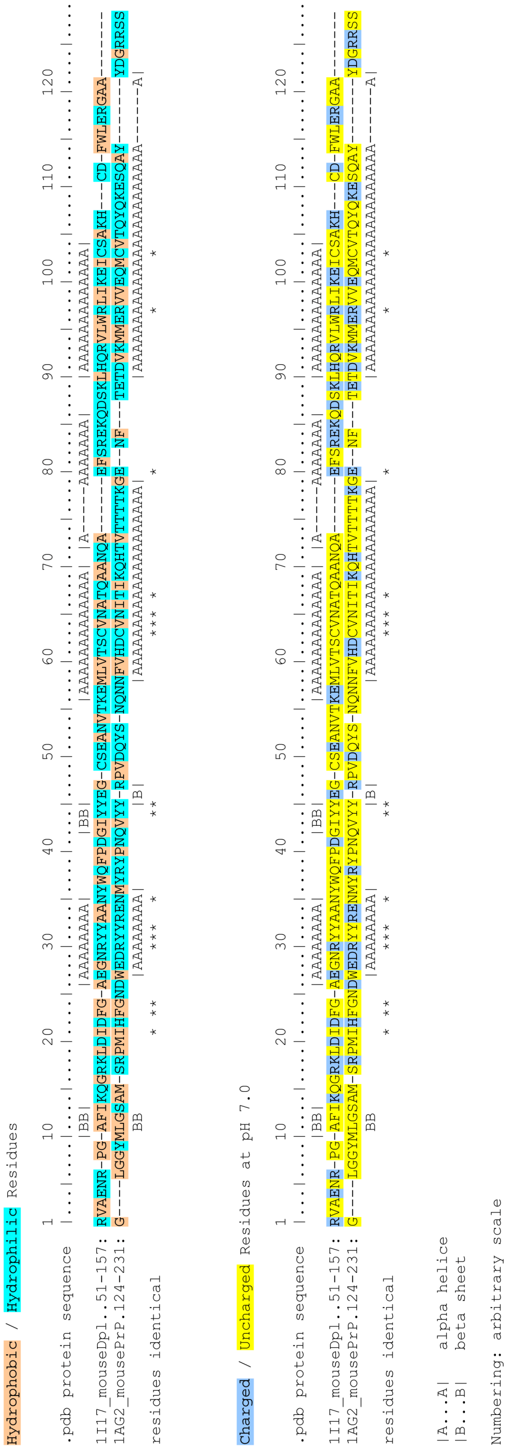


Figure 5.10: *Sequence alignment of PrP from different species for which the protein structure was experimentally resolved. The alignment was computed with the STAMP [75] combined structure superposition-sequence alignment algorithm implemented in VMD [76]). The protonation states presented refer to free residues in a pH 4 solution.*

5.5.2 Structure based sequence alignment of mouse PrP and Dpl



Hydrophobic / Hydrophilic Residues

```
                       1        10        20        30        40        50        60        70        80        90        100       110       120
.pdb protein sequence  |...|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....
                                |BB|            |AAAAAAAA|      |BB|          |AAAAAAAAAAAAA| |A-----AAAAAAA|   |AAAAAAAAAAAAA|
1I17_mouseDpl..51-157: RVAENR-PG-AFIKQGRKLDIDFG-AEGNRYYAANYWQFPDGIYYEG-CSEANVTKEMLVTSCVNATQAANQA------EFSREKQDSKLHQRVLWRLIKEICSAKH---CD-FWLERGAA-------
1AG2_mousePrP.124-231: G----LGGYMLGSAM-SRPMIHFGNDWEDRYYRENMYRYPNQVYY-RPVDQYS-NQNNFVHDCVNITIKQHTVTTTTKGE--NF---TETDVKMMERVVEQMCVTQYQKESQAY-------YDGRRSS
                                 BB              |AAAAAAAA|        |B|          |AAAAAAAAAAAAAAAAAAAA|          |AAAAAAAAAAAAAAAAAAAAAAA-------A|
residues identical                         * **     ***  *        **                 *** *            *                *     *
```

Charged / Uncharged Residues at pH 7.0

```
                       1        10        20        30        40        50        60        70        80        90        100       110       120
.pdb protein sequence  |...|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....
                                |BB|            |AAAAAAAA|      |BB|          |AAAAAAAAAAAAA| |A-----AAAAAAA|   |AAAAAAAAAAAAA|
1I17_mouseDpl..51-157: RVAENR-PG-AFIKQGRKLDIDFG-AEGNRYYAANYWQFPDGIYYEG-CSEANVTKEMLVTSCVNATQAANQA------EFSREKQDSKLHQRVLWRLIKEICSAKH---CD-FWLERGAA-------
1AG2_mousePrP.124-231: G----LGGYMLGSAM-SRPMIHFGNDWEDRYYRENMYRYPNQVYY-RPVDQYS-NQNNFVHDCVNITIKQHTVTTTTKGE--NF---TETDVKMMERVVEQMCVTQYQKESQAY-------YDGRRSS
                                 BB              |AAAAAAAA|        |B|          |AAAAAAAAAAAAAAAAAAAA|          |AAAAAAAAAAAAAAAAAAAAAAA-------A|
residues identical                         * **     ***  *        **                 *** *            *                *     *
```

```
|A...A|   alpha helice
|B...B|   beta sheet

Numbering: arbitrary scale
```

Figure 5.11: *Sequence alignment of mouse PrP and Dpl, computed with the STAMP [75] combined structure superposition-sequence alignment algorithm implemented in VMD [76]. The structural superposition partially computed from the sequence alignment is presented in Figure 5.8.*

5.6 References

[1] S. Prusiner. “Novel proteinaceous infectious particle causes scrapie”. *Science*, **216**, (1982) 136–144.

[2] B. Oesch, D. Westaway, M. Wälchli, M. P. McKinley, S. B. Kent, R. Aebersold, R. A. Barry, P. Tempst, D. B. Teplow, and L. E. Hood. “A cellular gene encodes scrapie prp 27-30 protein.” *Cell*, **40**, (1985) 735–746.

[3] B. Chesebro, R. Race, K. Wehrly, J. Nishio, M. Bloom, D. Lechner, S. Bergstrom, K. Robbins, L. Mayer, and J. M. Keith. “Identification of scrapie prion protein-specific mrna in scrapie-infected and uninfected brain.” *Nature*, **315**, (1985) 331–333.

[4] K. Basler, B. Oesch, M. Scott, D. Westaway, M. Wälchli, D. F. Groth, M. P. McKinley, S. B. Prusiner, and C. Weissmann. “Scrapie and cellular prp isoforms are encoded by the same chromosomal gene.” *Cell*, **46**, (1986) 417–428.

[5] K. M. Pan, M. Baldwin, J. Nguyen, M. Gasset, A. Serban, D. Groth, I. Mehlhorn, Z. Huang, R. J. Fletterick, and F. E. Cohen. “Conversion of alpha-helices into beta-sheets features in the formation of the scrapie prion proteins.” *Proc Natl Acad Sci U S A*, **90**, (1993) 10 962–10 966.

[6] B. W. Caughey, A. Dong, K. S. Bhat, D. Ernst, S. F. Hayes, and W. S. Caughey. “Secondary structure analysis of the scrapie-associated protein PrP 27-30 in water by infrared spectroscopy.” *Biochemistry*, **30**, (1991) 7672–7680.

[7] M. Gasset, M. A. Baldwin, R. J. Fletterick, and S. B. Prusiner. “Perturbation of the secondary structure of the scrapie prion protein under conditions that alter infectivity.” *Proc Natl Acad Sci U S A*, **90**, (1993) 1–5.

[8] R. Riek, S. Hornemann, G. Wider, M. Billeter, R. Glockshuber, and K. Wüthrich. “NMR structure of the mouse prion protein domain PrP(121-321).” *Nature*, **382**, (1996) 180–182. Exp, nmr.

[9] C. S. Burns, E. Aronoff-Spencer, C. M. Dunham, P. Lario, N. I. Avdievich, W. E. Antholine, M. M. Olmstead, A. Vrielink, G. J. Gerfen, J. Peisach, W. G. Scott, and G. L. Millhauser. “Molecular features of the copper binding sites in the octarepeat domain of the prion protein.” *Biochemistry*, **41**, (2002) 3991–4001.

[10] S. Hornemann and R. Glockshuber. “A scrapie-like unfolding intermediate of the prion protein domain PrP(121-231) induced by acidic pH.” *Proc Natl Acad Sci U S A*, **95**, (1998) 6010–6014.

[11] S. M. Martins, D. J. Frosoni, A. M. B. Martinez, F. G. D. Felice, and S. T. Ferreira. “Formation of soluble oligomers and amyloid fibrils with physical properties of the scrapie isoform of the prion protein from the C-terminal domain of recombinant murine prion protein mPrP-(121-231).” *J Biol Chem*, **281**, (2006) 26 121–26 128.

[12] G. M. Cereghetti, A. Schweiger, R. Glockshuber, and S. V. Doorslaer. "Electron paramagnetic resonance evidence for binding of cu(2+) to the c-terminal domain of the murine prion protein." *Biophys J*, **81**, (2001) 516–525.

[13] G. M. Cereghetti, A. Schweiger, R. Glockshuber, and S. V. Doorslaer. "Stability and cu(ii) binding of prion protein variants related to inherited human prion diseases." *Biophys J*, **84**, (2003) 1985–1997.

[14] M. C. Colombo, J. Vandevondele, S. V. Doorslaer, A. Laio, L. Guidoni, and U. Rothlisberger. "Copper binding sites in the c-terminal domain of mouse prion protein: A hybrid (qm/mm) molecular dynamics study." *Proteins*, **70**, (2008) 1084–1098.

[15] B. S. Wong, D. R. Brown, T. Pan, M. Whiteman, T. Liu, X. Bu, R. Li, P. Gambetti, J. Olesik, R. Rubenstein, and M. S. Sy. "Oxidative impairment in scrapie-infected mice is associated with brain metals perturbations and altered antioxidant activities." *J Neurochem*, **79**, (2001) 689–698.

[16] M. P. Hornshaw, J. R. McDermott, and J. M. Candy. "Copper binding to the n-terminal tandem repeat regions of mammalian and avian prion protein." *Biochem Biophys Res Commun*, **207**, (1995) 621–629.

[17] G. Schmitt-Ulms, G. Legname, M. A. Baldwin, H. L. Ball, N. Bradon, P. J. Bosque, K. L. Crossin, G. M. Edelman, S. J. DeArmond, F. E. Cohen, and S. B. Prusiner. "Binding of neural cell adhesion molecules (n-cams) to the cellular prion protein." *J Mol Biol*, **314**, (2001) 1209–1225.

[18] H. Beler, M. Fischer, Y. Lang, H. Bluethmann, H. P. Lipp, S. J. DeArmond, S. B. Prusiner, M. Aguet, and C. Weissmann. "Normal development and behaviour of mice lacking the neuronal cell-surface prp protein." *Nature*, **356**, (1992) 577–582.

[19] J. C. Manson, A. R. Clarke, P. A. McBride, I. McConnell, and J. Hope. "Prp gene dosage determines the timing but not the final intensity or distribution of lesions in scrapie pathology." *Neurodegeneration*, **3**, (1994) 331–340.

[20] S. Sakaguchi, S. Katamine, N. Nishida, R. Moriuchi, K. Shigematsu, T. Sugimoto, A. Nakatani, Y. Kataoka, T. Houtani, S. Shirabe, H. Okada, S. Hasegawa, T. Miyamoto, and T. Noda. "Loss of cerebellar purkinje cells in aged mice homozygous for a disrupted prp gene." *Nature*, **380**, (1996) 528–531.

[21] R. C. Moore, I. Y. Lee, G. L. Silverman, P. M. Harrison, R. Strome, C. Heinrich, A. Karunaratne, S. H. Pasternak, M. A. Chishti, Y. Liang, P. Mastrangelo, K. Wang, A. F. Smit, S. Katamine, G. A. Carlson, F. E. Cohen, S. B. Prusiner, D. W. Melton, P. Tremblay, L. E. Hood, and D. Westaway. "Ataxia in prion protein (prp)-deficient mice is associated with upregulation of the novel prp-like protein doppel." *J Mol Biol*, **292**, (1999) 797–817.

[22] K. Peoc'h, C. Serres, Y. Frobert, C. Martin, S. Lehmann, S. Chasseigneaux, V. Sazdovitch, J. Grassi, P. Jouannet, J.-M. Launay, and J.-L. Laplanche. "The human

"prion-like" protein doppel is expressed in both sertoli cells and spermatozoa." *J Biol Chem*, **277**, (2002) 43 071–43 078.

[23] A. Behrens, N. Genoud, H. Naumann, T. Rülicke, F. Janett, F. L. Heppner, B. Ledermann, and A. Aguzzi. "Absence of the prion protein homologue doppel causes male sterility." *EMBO J*, **21**, (2002) 3652–3658.

[24] A. Sakudo, D. chan Lee, I. Nakamura, Y. Taniuchi, K. Saeki, Y. Matsumoto, S. Itohara, K. Ikuta, and T. Onodera. "Cell-autonomous prp-doppel interaction regulates apoptosis in prp gene-deficient neuronal cells." *Biochem Biophys Res Commun*, **333**, (2005) 448–454.

[25] T. Cui, A. Holme, J. Sassoon, and D. R. Brown. "Analysis of doppel protein toxicity." *Mol Cell Neurosci*, **23**, (2003) 144–155.

[26] N. Nishida, P. Tremblay, T. Sugimoto, K. Shigematsu, S. Shirabe, C. Petromilli, S. P. Erpel, R. Nakaoke, R. Atarashi, T. Houtani, M. Torchia, S. Sakaguchi, S. J. DeArmond, S. B. Prusiner, and S. Katamine. "A mouse prion protein transgene rescues mice deficient for the prion protein gene from purkinje cell degeneration and demyelination." *Lab Invest*, **79**, (1999) 689–697.

[27] M. L. Massimino, C. Ballarin, A. Bertoli, S. Casonato, S. Genovesi, A. Negro, and M. C. Sorgato. "Human doppel and prion protein share common membrane microdomains and internalization pathways." *Int J Biochem Cell Biol*, **36**, (2004) 2016–2031.

[28] H. Mo, R. C. Moore, F. E. Cohen, D. Westaway, S. B. Prusiner, P. E. Wright, and H. J. Dyson. "Two different neurodegenerative diseases caused by proteins with similar structures." *Proc Natl Acad Sci U S A*, **98**, (2001) 2352–2357.

[29] G. L. Silverman, K. Qin, R. C. Moore, Y. Yang, P. Mastrangelo, P. Tremblay, S. B. Prusiner, F. E. Cohen, and D. Westaway. "Doppel is an n-glycosylated, glycosylphosphatidylinositol-anchored protein. expression in testis and ectopic production in the brains of prnp(0/0) mice predisposed to purkinje cell loss." *J Biol Chem*, **275**, (2000) 26 834–26 841.

[30] T. Lührs, R. Riek, P. Güntert, and K. Wüthrich. "NMR structure of the human doppel protein." *J Mol Biol*, **326**, (2003) 1549–1557.

[31] K. Qin, J. Coomaraswamy, P. Mastrangelo, Y. Yang, S. Lugowski, C. Petromilli, S. B. Prusiner, P. E. Fraser, J. M. Goldberg, A. Chakrabartty, and D. Westaway. "The prp-like protein doppel binds copper." *J Biol Chem*, **278**, (2003) 8888–8896.

[32] G. M. Cereghetti, A. Negro, E. Vinck, M. L. Massimino, M. C. Sorgato, and S. V. Doorslaer. "Copper(ii) binding to the human doppel protein may mark its functional diversity from the prion protein." *J Biol Chem*, **279**, (2004) 36 497–36 503.

[33] I. V. Baskakov, G. Legname, M. A. Baldwin, S. B. Prusiner, and F. E. Cohen. “Pathway complexity of prion protein assembly into amyloid.” *J Biol Chem*, **277**, (2002) 21 140–21 148.

[34] I. V. Baskakov, G. Legname, Z. Gryczynski, and S. B. Prusiner. “The peculiar nature of unfolding of the human prion protein.” *Protein Sci*, **13**, (2004) 586–595.

[35] O. V. Bocharova, L. Breydo, A. S. Parfenov, V. V. Salnikov, and I. V. Baskakov. “In vitro conversion of full-length mammalian prion protein produces amyloid form with physical properties of PrP(Sc).” *J Mol Biol*, **346**, (2005) 645–659.

[36] A. Tahiri-Alaoui and W. James. “Rapid formation of amyloid from alpha-monomeric recombinant human PrP in vitro.” *Protein Sci*, **14**, (2005) 942–947.

[37] K. Post, M. Pitschke, O. Schäfer, H. Wille, T. R. Appel, D. Kirsch, I. Mehlhorn, H. Serban, S. B. Prusiner, and D. Riesner. “Rapid acquisition of beta-sheet structure in the prion protein prior to multimer formation.” *Biol Chem*, **379**, (1998) 1307–1317.

[38] K. Jansen, O. Schäfer, E. Birkmann, K. Post, H. Serban, S. B. Prusiner, and D. Riesner. “Structural intermediates in the putative pathway from the cellular prion protein to the pathogenic form.” *Biol Chem*, **382**, (2001) 683–691.

[39] K.-W. Leffers, H. Wille, J. Stöhr, E. Junger, S. B. Prusiner, and D. Riesner. “Assembly of natural and recombinant prion protein into fibrils.” *Biol Chem*, **386**, (2005) 569–580.

[40] F. Eghiaian, T. Daubenfeld, Y. Quenet, M. van Audenhaege, A.-P. Bouin, G. van der Rest, J. Grosclaude, and H. Rezaei. “Diversity in prion protein oligomerization pathways results from domain expansion as revealed by hydrogen/deuterium exchange and disulfide linkage.” *Proc Natl Acad Sci U S A*, **104**, (2007) 7414–7419.

[41] K. Kuwata, H. Li, H. Yamada, G. Legname, S. B. Prusiner, K. Akasaka, and T. L. James. “Locally disordered conformer of the hamster prion protein: a crucial intermediate to prpsc?” *Biochemistry*, **41**, (2002) 12 277–12 283.

[42] W. J. Becktel and J. A. Schellman. “Protein stability curves”. *Biopolymers*, **26**, (1987) 1859–1877.

[43] S. M. Whyte, I. D. Sylvester, S. R. Martin, A. C. Gill, F. Wopfner, H. M. Schtzl, G. G. Dodson, and P. M. Bayley. “Stability and conformational properties of doppel, a prion-like protein, and its single-disulphide mutant.” *Biochem J*, **373**, (2003) 485–494.

[44] K. Lu, W. Wang, Z. Xie, B. S. Wong, R. Li, R. B. Petersen, M. S. Sy, and S. G. Chen. “Expression and structural characterization of the recombinant human doppel protein.” *Biochemistry*, **39**, (2000) 13 575–13 583.

[45] W. Swietnicki, R. B. Petersen, P. Gambetti, and W. K. Surewicz. “Familial mutations and the thermodynamic stability of the recombinant human prion protein.” *J Biol Chem*, **273**, (1998) 31 048–31 052.

[46] H. Rezaei, Y. Choiset, F. Eghiaian, E. Treguer, P. Mentre, P. Debey, J. Grosclaude, and T. Haertle. "Amyloidogenic unfolding intermediates differentiate sheep prion protein variants." *J Mol Biol*, **322**, (2002) 799–814.

[47] E. M. Nicholson, H. Mo, S. B. Prusiner, F. E. Cohen, and S. Marqusee. "Differences between the prion protein and its homolog doppel: a partially structured state with implications for scrapie formation." *J Mol Biol*, **316**, (2002) 807–815.

[48] D. V. D. Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen. "GROMACS: fast, flexible, and free." *J Comput Chem*, **26**, (2005) 1701–1718.

[49] W. van Gunstern, S. Billeter, A. Eising, P. Huenenberger, P. Krueger, A. Mark, W. Scott, and I. Tironi. *Biomolecular Simulation: The GROMOS96 Manual and User Guide* (Hochschulverlag AG, Zuerich, 1996).

[50] H. Berendsen, J. Postma, W. van Gunsteren, and J. Hermans. *Intermolecular Forces* (Reidel, Dordrecht, 1981).

[51] J. P. Ryckaert, G. Ciccotti, and H. Berendsen. "Numerical integration of cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes". *J Comp Phys*, **23**, (1977) 327–341.

[52] S. Nosé. "A unified formulation of the constant temperature molecular dynamics methods". *J. Chem. Phys.*, **81**, (1984) 511–519.

[53] W. Hoover. "Canonical dynamics-equilibrium phase space distribution". *Phys. Rev. A.*, **31**, (1985) 1695–1697.

[54] W. Swietnicki, R. Petersen, P. Gambetti, and W. K. Surewicz. "ph-dependent stability and conformation of the recombinant human prion protein prp(90-231)." *J Biol Chem*, **272**, (1997) 27 517–27 520.

[55] J. Antosiewicz, J. A. McCammon, and M. K. Gilson. "Prediction of pH-dependent properties of proteins." *J Mol Biol*, **238**, (1994) 415–436.

[56] M. Davis and J. McCammon. "Electrostatics in biomolecular structure and dynamics". *Chem Rev*, **90**, (1990) 509–521.

[57] A. Yang, M. Gunner, R. Sampogna, R. Sharp, and B. Honig. "On the calculation of pKa in proteins". *Proteins*, **15**, (1993) 252–256.

[58] G. Vriend. "WHAT IF: a molecular modeling and drug design program." *J Mol Graph*, **8**, (1990) 52–6, 29.

[59] W. Kabsch and C. Sander. "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." *Biopolymers*, **22**, (1983) 2577–2637.

[60] K. Kuwata, Y. O. Kamatari, K. Akasaka, and T. L. James. "Slow conformational dynamics in the hamster prion protein." *Biochemistry*, **43**, (2004) 4439–4446.

[61] J. Ziegler, H. Sticht, U. C. Marx, W. Müller, P. Rösch, and S. Schwarzinger. "Cd and nmr studies of prion protein (prp) helix 1. novel implications for its role in the $\text{prp}^c \rightarrow \text{prp}^{Sc}$ conversion process." *J Biol Chem*, **278**, (2003) 50 175–50 181.

[62] X. Lu, P. L. Wintrode, and W. K. Surewicz. "Beta-sheet core of human prion protein amyloid fibrils as determined by hydrogen/deuterium exchange." *Proc Natl Acad Sci U S A*, **104**, (2007) 1510–1515.

[63] R. I. Dima and D. Thirumalai. "Probing the instabilities in the dynamics of helical fragments from mouse prpc." *Proc Natl Acad Sci U S A*, **101**, (2004) 15 335–15 340.

[64] A. D. Simone, A. Zagari, and P. Derreumaux. "Structural and hydration properties of the partially unfolded states of the prion protein." *Biophys J*.

[65] I. V. Baskakov, G. Legname, S. B. Prusiner, and F. E. Cohen. "Folding of prion protein to its native alpha-helical conformation is under kinetic control." *J Biol Chem*, **276**, (2001) 19 687–19 690.

[66] J. I. Guijarro, M. Sunde, J. A. Jones, I. D. Campbell, and C. M. Dobson. "Amyloid fibril formation by an sh3 domain." *Proc Natl Acad Sci U S A*, **95**, (1998) 4224–4228.

[67] F. Chiti, P. Webster, N. Taddei, A. Clark, M. Stefani, G. Ramponi, and C. M. Dobson. "Designing conditions for in vitro formation of amyloid protofilaments and fibrils." *Proc Natl Acad Sci U S A*, **96**, (1999) 3590–3594.

[68] F. Chiti, N. Taddei, M. Bucciantini, P. White, G. Ramponi, and C. M. Dobson. "Mutational analysis of the propensity for amyloid formation by a globular protein". *The EMBO journal*, **19**, (2000) 1441–1449.

[69] M. Ramirez-Alvarado, J. S. Merkel, and L. Regan. "A systematic exploration of the influence of the protein stability on amyloid fibril formation in vitro." *Proc Natl Acad Sci U S A*, **97**, (2000) 8979–8984.

[70] K. Yutani, G. Takayama, S. Goda, Y. Yamagata, S. Maki, K. Namba, S. Tsunasawa, and K. Ogasahara. "The process of amyloid-like fibril formation by methionine aminopeptidase from a hyperthermophile, pyrococcus furiosus." *Biochemistry*, **39**, (2000) 2769–2777.

[71] O. V. Bocharova, L. Breydo, V. V. Salnikov, A. C. Gill, and I. V. Baskakov. "Synthetic prions generated in vitro are similar to a newly identified subpopulation of prpsc from sporadic creutzfeldt-jakob disease." *Protein Sci*, **14**, (2005) 1222–1232.

[72] J. O. Speare, T. S. Rush, M. E. Bloom, and B. Caughey. "The role of helix 1 aspartates and salt bridges in the stability and conversion of prion protein." *J Biol Chem*, **278**, (2003) 12 522–12 529.

[73] I. B. Kuznetsov and S. Rackovsky. "Comparative computational analysis of prion proteins reveals two fragments with unusual structural properties and a pattern of increase in hydrophobicity associated with disease-promoting mutations." *Protein Science*, **13**, (2004) 3230–3244.

[74] N. J. Cobb, F. D. Snnichsen, H. McHaourab, and W. K. Surewicz. “Molecular architecture of human prion protein amyloid: a parallel, in-register beta-structure.” *Proc Natl Acad Sci U S A*, **104**, (2007) 18 946–18 951.

[75] R. B. Russell and G. J. Barton. “Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels.” *Proteins*, **14**, (1992) 309–323.

[76] W. Humphrey, A. Dalke, and K. Schulten. “Vmd: visual molecular dynamics.” *J Mol Graph*, **14**, (1996) 33–8, 27–8.

[77] S. Colacino, G. Tiana, and G. Colombo. “Similar folds with different stabilization mechanisms: the cases of prion and doppel proteins.” *BMC Struct Biol*, **6**, (2006) 17.

[78] G. Settanni, T. X. Hoang, C. Micheletti, and A. Maritan. “Folding pathways of prion and doppel.” *Biophys J*, **83**, (2002) 3533–3541.

Chapter 6

Outlook

The following additional analysis might provide further understanding of the misfolding mechanisms, as well as a better characterization of potential PrP^{Sc} monomeric precursors.

6.1 Protein energy as a function of α and β content

The total potential energies (E_{tp}) or partial potential energies (E_{pp}) introduced in Chapter 2 could be used to evaluate the contribution of secondary structure to global stability. In Chapter 2, averages were computed for high secondary content structures only, but they could also be computed for any secondary structure content using the same protocol. A first step would consist in mapping potential energy as a function of the quantity of α and/or β residues present. This would supply a description of the average energetic contribution of the hydrogen bonding involved in a pair of residues forming α or β secondary structure, and could be analyzed with a 3-D plot containing (i) the number of α residues on the x-axis, (ii) the number of β residues on the y-axis, and (iii) the average potential energy obtained for a particular combination of (i) and (ii) on the z-axis. A further step would then consist in subdividing these averages as a function of the sequence position of the secondary structure, allowing to understand sequence and/or tertiary structure effects involved in the folding process.

6.2 Dissecting the mechanisms of protein unfolding

One of the main differences observed between the unfolding of Dpl and PrP relates to differences in helix stabilities, with the high stability of H1 in PrP and H3 in Dpl (Chapter 5). Arg, Glu and Asp residues of PrP H1 allow for stabilizing salt bridges between side chains that are not present in Dpl H1. We have monitored the evolution of the distances characterizing these salt bridges in two high temperature 500K test runs and observed that in the first simulation, the disruption of some of the salt bridges occurred simultaneously with the unfolding of H1, whereas in the second simulation, intact salt bridges correlated with an intact H1 (data not shown). A deeper investigation of the first simulation should allow to understand if salt bridge disruption or unfolding of helix H1 is the primary event. Similarly, such studies would be very instructive for all the major secondary structure unfolding events observed, as well as for the formation of major new β-sheets. If such events occur in PrP but not in Dpl or vice-versa, the explanation must reside in sequence differences that would be highlighted, as opposed to sequence differences that are irrelevant for the major differences in unfolding behaviors of PrP and Dpl.

6.3 Electrostatic properties and aggregation propensities of β-rich folds

The conformational changes of monomeric PrP investigated in Chapter 4 are most probably related to a first step in pathogenesis, with subsequent steps involving further rearrangements, followed by aggregation to multimers and eventually to amyloid fibrils. In Chapter 4, we propose several monomeric β-rich folds. An obvious next step consists in characterizing the aggregation propensities of these conformations. Favorable interactions between apolar or charged surface residues will help orient monomers into a stable monomer-monomer binding mode. We have started to characterize the electrostatic potentials (ESP) of the PrP β-rich folds by mapping them onto the solvent accessible surface area, but have not yet detected any obvious pattern suggesting a protein binding interface. These studies could be extended to the solvent and/or counter ions, which might interact with protein cavities and enhance the ESP locally. Another possibility would be to test aggregation propensities with automated docking software.

Acknowledgments

First of all, I am deeply indebted to Ursula for having offered me the possibility to move into the theoretical field. I am extremely grateful for the liberty she gave me in the choice of my projects and scientific questions addressed, as well as in the methological approaches I chose to try to solve them. Ursula was always available for discussions, and gave me access to computer power I would never have dreamed of. Next on the list of my enlightened guides, Ivano offered me the opportunity to work in the field of REMD that he had quit by the end of his PhD, promising me that his code was ready, and that with little investment, I would get some results from a short side project. The side project became my main project, and here I am, 4 years later, extremely grateful to have been cast into the developer's field by problems related to REMD limitations with larger, biomolecular systems. I could also walk into Ivano's office anytime, with numerous statistical physics, folding free energy or FORTRAN problems!

This excellent supervision would never have sufficed without the extremely knowledgeable technical support provided by the following wonderful persons with unlimited generosity. Back in 2003, in AMD-Athlon prehistory (when proteins used to scramble around at a ridiculous speed of $\sim$ 100 ps per day), Patrick and Anatole had set up the computer clusters in Zürich and Lausanne. 2 years later, Christian took over, rebuilding the cluster, then based on a pool of new AMD-Opteron computers. By the end of my stay at LCBC, Matteo jumped in to rebuild things on an entirely renewed pool of Intel Woodcrest dualcore machines. By then, with this last upgrade, proteins were soaring high at 3000 ps/day, and calculations that used to take months at the beginning of my PhD could now be solved within days.. Although this will soon grow obsolete as well, the present thesis and computational effort would never have been possible without the hard work invested by this marvelous crew day and night, ensuring an extremely stable and fast cluster that was permanently on the world's top-end of computational power. Last but not least in the hardware miracle list, Matteozinho's help with the “ptpool“ cluster and with the Linux partition of my personal laptop, that he used to fix in a couple of seconds during his coffee breaks and escapes of what he used to call the ”cleaning lady attacks”.

Two further invaluable contributions to my work came from Mitch and Maurício. Mitch assisted me from the very beginning to the very end of my PhD in 4 very different issues: First, as the MATLAB guru who converted me to his religion and finally allowed me to analyse tons of parallel data efficiently. Second, as the physicist who was permanently available and willing to bring his expertise into my statistical mechanical and generalized reaction field problems. Third, as the indispensable GROMACS expert assisting me in my REMD hacks. And fourth, as my patriotic guide, transforming my military service into a 5-month internship at the Swiss Institute for Bioinformatics! All that expertise was complemented by a friendly and funky philospher who could get me into hip-hop clubs at Amsterdam despite my utterly inappropriate clothing... Maurício built up (and baptized!!) the “ptpool“ cluster from scratch, allowing for a considerable speedup of my simulations and a decisive turn to my PhD. Delightful days of Brazilian chatting, in the ”Nordestino” accent of our childhoods... With everlasting thanks, Maurício, believe me, I would still be fixing that machine had you not been there.. Very precious scientific contributions also came from numerous discussions with the other Brazilian “Nordestino” of the lab, Roberto, and from Michele, Geoff (another amyloid-protein REMD freak!), Marilisa, Maria-Carola and Julian. I would also like to thank Professors Requeña, Dal Peraro, Michielin and Helm for agreeing to be examiners of this thesis. Jakob and Dominique, my two semester students, also provided very interesting contributions to the doppel project. Last but not least, Karin’s extremely efficient support in organizing trips and conferences and settling any administrative problem within days.

The LCBC was, is and will remain an utmost interesting international Babel- computational beehive! Ursula’s open minded group building policy has yielded an unbelievably rich and everchanging blend of scientific backgrounds and cultures. The perfect place to enjoy very diverse views on political, historical and sometimes even philosophical issues, with Håkan, Ivano, Enrico, Maria Carola, Maurício, Roberto, Michele, Anatole, Mitch, Denis, Fanny, Geoff, Christian, Sam, I-Chun and Julian. The party crew, Denis, Fanny, Geoff, Matteo, Enrico, Pablo, Patrick and Julian provided for some very intense moments of my life. Other summits were experienced with Christian and Anja, who brought me back (3 times!) to 4000 m high alpine Heavens I had not revisited since teenage. Among the other gifts of nature and social delights shared with LCBC were the group outings (winter skiing and summer hiking), chemistry department rooftop grillparties, wine tasting at Roberto and Theresa’s place, cooking of Italian specialties at Michele’s place or at our Christmas party, enjoying a raclette at Karin’s place, sunsets with the “LCBC balcony smoker’s club” and countless other events! And of course, the daily cheer up (and face cleaning) by Topas and Ruby (always happy to see me.. 20 m away)!

Mysteries of life somehow brought me back to Lausanne for my PhD, back to the place where I had started my studies (that I had finished in Zürich). An extremely happy consequence was the opportunity to meet again with friends from earlier life. Starting with the one without whom, at a time I was mainly looking for positions in Zürich, Basel or abroad, I might not have returned to Lausanne. Back in 2003, Eric told me about a young professor who was starting her theoretical group right next to his organic chemistry lab. In our common years of PhD, Eric was a weekly support, with whom I shared interdisciplinary science, mountain tours and a month round trip of Central Asia. Next, Olivier, an old highschool and biology studies friend, doing practically the same PhD (computer simulations and very similar statistics) in a different area, modeling the geographical distributions of endangered and invasive plant species. Among our arguments, along the weeks of these 4 years, politics, science, and.. the question of determining which of us had most betrayed our original ideal of biology! A good pretext for a month trip in India..

Numerous other friends had remained on the campus, in areas ranging from law to stem cell biology, passing through history, economy, politics and religion, and providing for marvelous moments of shared science and daily life: Laurent, Léonard, Friederike, Christophe, Fahd, Nicolas, Ivan, Christelle, Yves and Laure.. Other intense and delightful moments, supporting me along the efforts of research, were spent with Isabelle, Jacques, Alessandro (the “Zab Attak cocktail tasting club”, hmmmm?), Caro, Vincent and the Swiss Alpine Club.

Beyond all what words can tell, the last dedication goes to my family for the support provided throughout my studies. I could never have dreamed of a better education and cultural opening, growing up with two Swiss national languages and a childhood spent at an American School in Brazil.. the best mix to fully enjoy the unique LCBC blend! Along and deep in my heart, Nathalie, bringing daily sunshine into my life...

Curriculum Vitae

Pascal Baillod

Av. du Galicien 3	Swiss nationality
CH-1008 Prilly, Switzerland	Born August 29, 1977
pascal.baillod@epfl.ch	Single
tel. +41(0)79 351 26 42	

Education

- Ph.D. studies, Computational Structural Biochemistry *Oct 2003 - June 2008*
  - Swiss Federal Institute of Technology (EPFL), Laboratory of Chemical and Biochemical Computation, thesis supervisor: Professor U. Röthlisberger.
  - Development of alternative simulation methods for large structures, as well as conformation clustering algorithms. Application to the study of structural conversions of the Prion protein, with a focus on the identification of rare conformations that might play a role in the Creutzfeldt-Jakobs disease.
- Biochemistry studies. Diploma in Biochemistry. *Oct 1997- Mar 2002*
  - University of Lausanne and Swiss Federal Institute of Technology (ETHZ), Zürich.

Working experience

- Research fellow, Swiss Institute of Bioinformatics (5 months) *2005, 2006*
- Teaching (EPFL, during Ph.D.) *2003-2005*
  - Biomolecular modeling exercises, 1st year Chemistry. Elaborating and teaching exercises for a new course.
  - Biomolecular modeling research projects, 3rd year Chemistry and Physics. Teaching students how to apply advanced biomolecular tools to solve open questions in a research project.
- Teaching (Quito, Ecuador and Morges, Switzerland) *Apr 2002 - Jun 2003*
  - Teaching French, Mathematics, Geography and Arts to 8-12 year old children in the government school of Morges, Switzerland
  - Volunteering in schools and hospitals of Quito, Ecuador. Teaching English and Science to children (6 to 18 years old). Fundraising in Europe and organization of surgeries and treatments of indigent children.
- Consulting, writing articles and secretary work for APTE (2 months) *2001*

  - APTE association: Consulting and promotion of microtechnics in industry, Zürich. Publication of a popularization article, interviews and company database update.

Computational skills

- Programming and modeling
  - Classical Molecular Dynamics simulations, Amber and Gromacs packages.
  - Homology modeling (Swiss Model, Modeller).
  - Scripting languages: matlab, tcl, perl, bash.
  - Programming: Fortran, C++ (basic level).
- Administration
  - Partial administration of an AMD opteron 32 compute-node cluster. Firmware update, installation of new compute nodes, software compilation. Rocks (CentOS) linux distribution.
  - Management of the network backup of two terabyte-servers.

Languages

- English: Read, written and spoken fluently.
- French: Mother tongue, main language of studies.
- Swiss-German: Mother tongue.
- German: Read and spoken fluently. Writing: Average level.
- Portuguese: Read and spoken fluently. Writing: Average level.
- Spanish: Read and spoken fluently. Writing: Weak.

Interests

- Sports: Trekking, jogging, biking, skiing, ski-touring.
- Hobbies: Lecture (history, politics, economy, science), traveling, music.