UNIVERSITÀ DEGLI STUDI DI UDINE
hic sunt futura
www.uniud.it

PhD Course in Biomedical Science and Biotechnology

XXX Cycle

Molecular Dynamics simulations of amyloidogenic proteins. Unfolding, misfolding and aggregation

Author:
Cedrix J. Dongmo Foumthuim

Supervisor:
Prof. Federico Fogolari

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy (Ph.D)

Academic year 2016/2017

To

Corazón A. Dongmo
Angelo S. Dongmo
Gracia F. Dongmo

Abstract

Proteins are the main bulding blocks of biological systems. Their structure and function have been extensively studied so far both by experiments (Nuclear Magnetic Resonance, X-ray crystallography, Mass Spectrometry, etc.) and modeling strategies (Molecular Dynamics and Monte Carlo simulations, Density Functional Theory. etc.). *in vivo* in general and in solution in particular, they mostly adopt different and unique secondary and tertiary configurations, owing to their conformational freedom. The route and mechanism by which a specific shape is formed, i.e. the folding, which is not reversible in many cases, is not fully understood for several protein models nothwithstanding the fulgurant advances achieved in experimental and *in silico* techniques in the last decades. Under specific conditions (pH, temperature, concentration, etc.) such three-dimensional arrangement as well as the intra/inter-chains interactions can be lost and species such as disordered or fibrilar aggregates involved in several known human pathologies may develop.

In this thesis we probe the atomistic scale conformational dynamics of two amyloidogenic proteins, transthyretin and β2-microglobulin, using molecular dynamics simulations. We aim at understanding the major factors driving the misfolding and/or (un)folding of the latter specified proteins which play a precursor and prominent role in neurodegrative deseases. To this end the dynamics and dissociation of wild-type and mutant transthyretin is simulated. In particular the behaviour of a triple mutant (designed by Prof. R. Berni and coworkers to be monomeric) is studied. It comes out that the mutation considerably shifts the tetramer-folded monomer equilibrium towards the monomer, making this triple mutant a useful tool for structural and dynamical studies. The interaction of β2-microglobulin with hydrophobic surfaces is studied by molecular dynamics and the thermodynamics of the process is addressed using end-point free energy calculations. The results rationalize experimental observations reported in the literature.

Protein conformational dynamics and thermodynamics are currently experimentally probed by the backbone amide hydrogen exchange experiment (HX). The Bluu-Tramp experiment developed by prof. Esposito and coworkers allows the measurement of free energy, enthalpy and entropy of exchange in a single experiment. A proper comparison between experimental and simulation data require modeling of the process at atomic detail. Hence, we analyze also this aspect and try to relate the amide hydrogen protection observed in NMR experiments to various microscopic properties of the protein structure computed in the simulations. Using free energy calculations we aim at reproducing also the temperature dependence of the process.

Given the predominant role of protein association in most biological functions, we introduce a modeling approach to estimate the entropy loss upon complex formation, a contribution which is almost always neglected in many free energy calculation methodologies due to the high dimensionality of the degrees of freedom and adequate theoretical methods. The approach is applied to the case proteins considered in this thesis and an exact and approximate estimation of the full rotational-translational entropy are obtained in the context of nearest neighbor-based entropy formulation.

Overall, this thesis explores various aspects favouring the formation of misfolded and/or (un)folded protein species, ranging from dissociation of an homotetramer of transthyretin engineered *in silico*, through the interaction of β2-microglobulin with an hydrophobic surface model, to the backbone amide hydrogen exchange pattern of protection of the latter. Lastly and not the least, the thesis presents a computational methodology to address the roto-translational entropy loss upon complex formation of biomolecules.

Keywords : Molecular dynamics simulations, Free energy, Unfolding, Conformational changes, Amyloidogenic proteins.

Publication List

Some of the ideas and results discussed throughout this thesis have been printed before in :

1. F. Fogolari, O. Maloku, **C. J. Dongmo Foumthuim**; A. Corazza, G. Esposito. PDB2ENTROPY: entropy calculation from conformational ensembles, 2018, Submitted.
2. **C. J. Dongmo Foumthuim**; A. Corazza, G. Esposito and F. Fogolari. Molecular dynamics simulations of β2-microglobulin interaction with hydrophobic surfaces, *Mol. BioSyst.*, 2017, 13, 2625-2637.
3. M.A. Soler, A. Rodriguez, A. Russo, A. Feyisara Adedeji, **C. J. Dongmo Foumthuim**, C. Cantarutti, E. Ambrosetti, L. Casalis, A. Corazza, G. Scoles, D. Marasco, A. Laio, S. Fortuna. Computational design of cyclic peptides for the customized oriented immobilization of globular proteins, *Phys. Chem. Chem. Phys.*, 2017, 19, 2740-2748.
4. F. Fogolari; **C. J. Dongmo Foumthuim**; S. Fortuna; A. Soler; A. Corazza and G. Esposito. Accurate estimation of the entropy of rotation-translation probability distributions,*J. Chem. Theory Comput.*, 2016, 12 (1),1-8.

Acknowledgements

This piece of work could not have been completed without the help and assistance of some persons.

I would like to sincerely thank Prof. Fogolari F. for his guidance throughout the supervision of this work. It has been a very challenging scientific endeavours. Thanks to your human values, your `Science` and the approach to make things simple, we have almost always defeated the friction forces along the defined pathway.

I extend my thanks to Prof. Esposito G. and Dr. Corrazza A. for the constant support, assistance and friendly working environment and discussions.

I am grateful and cheerful to Prof. Scoles G., Dr. Fortuna S. and Dr. Soler M. A. whith whom the story began and for all the invaluable support, feedback and sympathy.

The world of computer simulation would probably has remained a mystery for me without the mentoring of Dr. McGrath J. Matthew who introcuded me through this field of research. Thanks once more, your support and feedbacks have helped me in shaping my research career. In the same regard Prof. Ghogomu N. Julius at the University of Dschang-Cameroon is also acknowledged.

I honestly appreciate the support and collaboration of my labmates for their team spirit and for the discussions we have had during our time here. I think of Cantarutti C., Maloku O., Dr. Vuano M. and Dr. Banelli T.

My beloved mother Kenfack N. Adrienne who has sacrificed all her life and makes every effort to see this dream becoming a fact is particularly acknowledged. Thanks mummy, we are near to the end.

I express my gratefulness to my wife Diffo Y. Joëlle for her constant, precious and invaluable emotional assistance.

My indebtedness also goes to my brothers and sisters Azangue B. Arthur, Ntsopmo N. Jourdelle, Donfack K. Lucresse and Tiofack K. Marc for their daily support and encouragements, I will never fail in the confidence you have placed in me.

I am also thankful to my entire family for their support and encouragements throughout my educational career. I especially think of Dongmo K. Stéphanie, Kitio K. Beaudelaire, Dongmo Clément, Mboufack Romain, Maazap Brigitte, Kom Frederic and Mbougue Odette.

All the persons that I may not have listed and who helped me in one way or in the other are also greatly thanked.

Contents

List of Figures

List of Tables

List of Abbreviations

MD	**M**olecular **D**ynamics
β**2m**	β2-microglobulin
TTR	Transthyretin
SSA	**S**enile **S**ystemic **A**myloidosis
HDX	**H**ydrogen **D**euterium **EX**change
NMR	**N**uclear **M**agnetic **R**esonance
a.a.	**A**mino **A**cid
C_α	**A**lpha **C**arbon
DNA	**D**eoxyribonucleic **A**cid
TEM	**T**ransmission **E**lectron **M**icroscopy
AFM	**A**tomic **F**orce **M**icroscopy
FTIR	**F**ourier **T**ransform **I**nfrared **S**pectroscopy
ssNMR	**s**olid-**s**tate **N**uclear **M**agnetic **R**esonance
ThT	**t**hioflavin-**T**
CR	**C**ongo **r**ed
CSF	**C**erebrospinal **F**luid
FAP	**F**amilial **A**myloid **P**olyneuropathy
FAC	**F**amilial **A**myloid **C**ardiomyopathy
DRA	**D**ialysis **R**elated **A**myloidosis
GBSA	**G**eneralized **B**orn-**S**uface **A**rea
OBC	**O**nufriev-**B**ashford-**C**ase
PBC	**P**eriodic **B**oundaries **C**onditions
PMF	**P**otential of **M**ean **F**orce
WHAM	**W**eighted **H**istogram **A**nalysis **M**ethod
REMD	**R**eplica **E**xchange **M**olecular **D**ynamics
FES	**F**ree **E**nergy **S**urface
CV	**C**ollective **V**ariables
RMSF	**R**oot **M**ean **S**quare **F**luctuations
RMSD	**R**oot **M**ean **S**quare **D**eviations
PDB	**P**rotein **D**ata **B**ank
MSM	**M**arkov **S**tate **M**odels
COM	**C**entre **O**f **M**ass
H-bonds	**H**ydrogen **B**onds
CLEANEX	**CLEAN C**hemical **EX**change
EX1	**M**onolecular **EX**change kinetics
EX2	**B**imolecular **EX**change kinetics
PF	**P**rotection **F**actor
LJ	**L**ennard-**J**ones

List of Symbols

F_i	force acting on the i^{th} particle	m_i	mass of atom i
U	potential energy	N	total number of particles
r_i	spatial coordinates	v_i	atomic velocities
a_i	particle acceleration	T	temperature
$\mathcal{P}$	probability/distribution function	P	pressure
p	momentum	V	volume
r_{cut}	cut-off distance	t	time
Θ	bond angle	Θ_0	equilibrium bond angle
k_r	stretching bond constant	k_Θ	bending force constant
k_ψ	height of the rotational barrier	ψ	proper dihedral angle
Υ	torsional angle phase	η	Fourier component
ω	improper dihedral	ω_0	equilibrium improper dihedral
q	partial charges	σ_{ij}	distance at which the LJ potential is null
ϵ_{ij}	LJ potential well depth	E_{vac}	vacuum potential energy
ΔG_{solv}	solvation free energy	ϕ	electrostatic potential
z_i	valence of the ions	C_i^b	bulk concentration
ρ	charge density	ϵ	local dielectric constant
α_i & α_j	effective Born radii	ϵ_0	vacuum permittivity
k_{B}	Boltzmann's constant	τ	rescaling factor
Δt	simulation time-step	ξ	reaction coordinates
$\mathcal{W}(\xi)$	weigthing term	$F(\xi)$	Helmholtz energy (PMF)
$\mathcal{H}$	total system's hamiltonian	$\mathcal{P}'(\xi)$	biased distribution
$V_G(S(\boldsymbol{r}),t)$	history-dependent bias potential	$F'(\xi)$	biased PMF
$s_i(t)$	i^{th} value for the CV	τ_G	stride of the Gaussians deposition
$\mathcal{N}$	Avogadro's number	Λ_A	thermal Debroglie wavelength
$\bar{V}_A$	partial solute's molar volume	r_A	solute's coordinates
m_A	mass of the solute	h	Planck's constant
P^0	standard pressure	r_S	solvent coordinates
U_{AS}	correlated solute-solvent energy	U_S	solvent potential
α, β & γ	OBC parametrers	ρ_i	atom's van der Walls radius
$\tilde{\rho}_i^{-1}$	pairwise summation function	r_i^{GB}	perfect GB radii
$R_{i,k}$	radius of the hypersphere	Γ	Gamma function
ς	Euler-Mascheroni constant	$\mathcal{R}$	rotation matrix
$\mathtt{Tr}$	trace operator	l	scaling length
ϕ	polar angle	ψ	azimuthal angle
θ	rotation axis	k_{int}	intrinsic exchange rate
k_{cl}	closing exchange rate	k_{op}	opening exchange rate
k_{exp}	experimental rate of exhange	$\rho g(r)$	pair correlation function

General Introduction

Proteins are found in almost all living systems, even in the simplest free-living organisms (Fraser et al., 1995) and they represent a large class of biomolecular assemblies made up by monomeric units of amino acid residues. After being synthesized in the cells, they may often adopt different conformational states to achieve their biological functions (Teague, 2003), and as such, they are the workhorses of biology (Jacobs et al., 2016). In most instances, their functionality is optimized in the near-native conformation, known to be a non trivial three-dimensional folded object, the exact shape of which being dependent upon the constituted amino acids primary sequence[1] and the physico-chemical parameters of their milieu (Anfinsen, 1972). This clearly states that, under specific conditions, for instance at a given pH, a primary considered sequence will always lead to firmly similar three-dimensional folds. Several factors come into play to maintain the folded structure of proteins, essentials of which are cohesive forces between the constituent amino acids such as van de Waals and coulombic interactions, hydrogen bonding, hydrophobic packing and so forth. As a matter of fact, with various degrees of flexibility, proteins are dynamics soft-matter moieties.

The flexible nature of protein assemblies enables them to achieve different functions *in vivo* including but not limted to substrate and cofactor binding, intracellular transport and enzymatic catalysis (Teague, 2003; Chevalier-Larsen and Holzbaur, 2006; Kamerlin and Warshel, 2010). This plasticity-like potential is even at the core of amyloidoses formation (Dobson, 2002). In fact, there is an intrinsic interplay between flexibility, stability and function and identifying the proper folding pathways may provide insightful knowledge about how protein's function is performed.

In the specific case of amylodogenic proteins, it is widely accepted that the onset of fibril formation is linked with partial unfolding and the many mutants produced and studied worldwide correlate protein instability and amyloid formation. To be concise, with amyloids we mean the insoluble fibrillar plaques deposited in living organisms as a result of globular proteins disruption. This proceeds through a conformational transition into a misfolded `cross-beta` form, leading to cellular dysfunction at the origin of many human neurological diseases ranging from Alzheimer to Parkinson through Huntington and Creutzfeldt-Jakob (Chiti and Dobson, 2006; Roberts, 2016; Chiti and Dobson, 2017). The interest in amyloid degeneration of proteins from the viewpoint of biophysics has been rather high over the last twenty years also due to the obvious link to protein folding.

[1] Apart from intrinsically disordered proteins

in vitro protein conformational (thermo)dynamics is mainly investigated through NMR spectroscopy (Wüthrich, 1986) and/or MS spectrometry (Chowdhury, Katta, and Chait, 1990) and to some extent also by X-ray crystallography (Kendrew et al., 1958). In the last years, we witnessed significant progress in quantitative experimental measurements and nowadays a range of techniques such as cryo-electron microscopy (Frank et al., 1995) and small-angle X-ray scattering (Svergun, 1999) are often used alongside with fluorescence spectroscopic techniques (Ha et al., 1999; Michalet, Weiss, and Jäger, 2006) to study the dynamics of proteins in real time. Most of them can provide highly resolved atomic-level clues on protein conformational changes and rugged free energy landscapes. However, in general, they are less informative in monitoring the fast kinetics events and/or short lifetime species[2]. Furthermore, several of them can only provide static frames on the systems of interest but no (or less) informations on the conformational dynamics. In this regards, with the recent developments, NMR spectroscopy constitutes a method of choice as it can access all time scales of protein dynamics (Kleckner and Foster, 2011; Rennella et al., 2012a; Rennella et al., 2012b; Palmer, 2015). Moreover, *in silico* methodologies such as molecular dynamics (MD) (Alder and Wainwright, 1957) have been proven to be relevant to address theoretically many of the issues just discussed.

MD simulations can adequately describe the dynamics of (bio)molecules at full atomistic scale representation with a much higher temporal and spatial resolution than most experimental techniques (Shea, and III, 2001; Karplus and McCammon, 2002; Scheraga, Khalili, and Liwo, 2007; Dror et al., 2010; Dror et al., 2012). Indeed, the latter condition (atomic scale resolution) was thought to be a necessary requirement in understanding the biological functions of biomacromolecules (Anfinsen, 1972). Nevertheless, even though the simulation timescales accessible by all-atom MD simulations have remarkably evolved to nearly milliseconds and beyond, they remain however, several orders of magnitude less significant than typical timescales for conformational changes relevant for key biochemical processes (Dror et al., 2010; Lane et al., 2013; Da et al., 2014).

Significant efforts have been made in understanding and elucidating the amyloid formation pathways and its relevance in the healthy-like and destructive behaviour of biological systems[3]. Nothwithstanding these achievements, still, a number of factors promoting the disruption of many globular protein structures and rendering them more amyloidgenic-prone have not yet been exhaustively elucidated[4]. Indeed, proteins in their natural cellular environments are considered as a component of a complex solution that may undergo, for either natural or traumatic events, a transition. In other words the whole thermodynamic system of the protein and the surrounding environment evolves from stability into metastability, a precarious condition which may explain the disruptive characteristics of the amyloid transition to restore a steady free energy minimum (Kitayama et al., 2013; Ikenoue et al., 2014; Lin et al., 2014).

[2] On the femto scale or so

[3] A clear differenciation is now been set between bad and good amyloids (Roberts, 2016)

[4] These broadly include the physico-chemical environment and the complexicity of tertiary and quaternary folds adopted by proteins under different conditions

This thesis explores the conformational dynamics of two paradigmatic protein models[5] for amyloidogenesis studies through all-atoms molecular dynamics simulations. Although they both belong to all-β structural class type, the first one, β2-microglobulin (β2m), adopts an Ig-like fold and is mainly involved in dialysis-related amyloidosis while the second one, transthyretin (TTR), with a prealbumin-like structural fold is known to form Senile Systemic Amyloidosis (SSA). Without the pretension of being exhaustive, the goal is the understanding of the major factors dragging the misfolding and/or (un)folding of these proteins.

To achieve this task, the dynamics and dissociation of wild-type and mutant homotetramer of TTR is simulated. More precisely, the behaviour of a triple mutant engineered *in silico* (no crystal structure was available when doing this work)[6] is studied. In addition, the interaction of β2m with hydrophobic surfaces is investigated and the thermodynamics of the binding process is addressed using end-point free energy calculations. Finally, the backbone amide hydrogen exchange (HDX) of the latter system is computed. The comparison of the amide hydrogen protection observed in NMR experiments[7] with a number of microscopic properties of the protein structure averaged in the simulations is attempted. By employing free energy calculations we aim at replicating the temperature dependence of the process as well.

On a more methodological perspective, a theoretical approach to estimate the entropy loss upon complex formation is presented. It is worth noting that this term is mostly unaccounted in many free energy calculation methodologies, obviously due to computational cost associated to the high dimensionality of the degrees of freedom and adequate sampling methods.

The conformational dynamics has provided clues into the molecular machinery of protein stability and function, thereby opening-up further design strategies with targeted responses (Childers and Daggett, 2017; Soler et al., 2017). Moreover, the structural diversity and intricate behaviour of amyloid fibril oligomers self-assembly under various circumstances confer to them a unique property potentially integrated in engineering novel responsive biopolymers and biomaterials (Knowles et al., 2007; Cherny and Gazit, 2008). Understanding the factors leading to the formation of misfolded/(un)folded protein species and later the mechanisms of precarious protein stability has technological and medical implications and could help rationalizing the development of novel therapeutic strategies with which to prevent and/or treat amyloidosis. An outstanding example is provided by medical devices which are currently used for removing β2-microglobulin from the blood of hemodialysis patients in medical devices (Suzuki, Shimazaki, and Kutsuki, 2003). Thus, our study may help enlighten our understanding of the complex machinery and the physico-chemical factors behind amyloidogenesis, thereby potentially inspire new design strategies.

[5] Both forming extracellular amyloid deposits
[6] The X-ray structure has just been released (Zanotti et al., 2017)
[7] Bluu-Tramp experiment (Rennella et al., 2012a; Rennella et al., 2012b)

Outline

This thesis is subdivided in three generic parts.

The first part is referred as a more theoretical one and covers two chapters. In Chapter 1, the fundamentals of protein structure in general and of amyloidogenic proteins in particular with emphasis on the case systems considered here, namely TTR and β2m are discussed. More in details, we state what is already known about amyloidogenic proteins in relation to both their structure, (thermo)dynamics and function. This chapter stresses on the relevance of the chosen systems and wraps up by an outlook on the actual work. In Chapter 2 we recall the theoretical backgrounds as far as molecular dynamics simulation is concerned. We aim not only at laying the background for the methodologies used, but also to clarify the readers about some concepts and terminology to be used thereafter.

The second part is a more methodological one. It essentially embodies Chapter 3, wherein a novel approach to address the full roto-translational entropy loss upon complex formation, deriving from the nearest neighgour entropy formalism is introduced. Theoretical foundations of the method are discussed and applications are made on the model systems considered in this thesis.

The third part comprises three chapters and is a more applicative one. In Chapter 4 the dynamics and dissociation pathways of human wild-type and of a particular triple mutant homotetramer of TTR is simulated. Chapter 5 deals with the interaction of β2m with a model cubic hydrophobic box and subsequently the thermodynamics of the binding process addressed by the mean of end-point free energy calculations. Chapter 6 presents the preliminary results of backbone amide hydrogen exchange of β2m and the comparison of the averaged simulated miscroscopic quantities with experimentally NMR measured amide protection factors. The thesis ends by the Concluding Remarks and Outlook which summarizes the findings of this work and frames some future outlooks.

Chapter 1

Amyloidogenic proteins: State-of-the-art

This chapter opens up by a short overview on the essential notions and nomenclature in relation to protein geometry in general. Afterwards, we outline the common features peculiar to amyloidogenic proteins that make them so diverse. We subsquently discuss the major amyloid formation mechanisms and pathways and the methods used for their detection as well as the correlated human diseases. The chapter closes up by targeting attention on the case systems selected for this thesis i.e. TTR and β2m emphasizing on their structure and function in relation to their amyloid propensity.

1.1 Protein structure: what a complex machinery?

Synthesized on the ribosome as unstructured linear chains of amino acid (a.a.) residues, proteins like other biological macromolecules such as carbohydrates, nucleic acids, etc. are fundamental pillars supporting the molecular foundation of life. Each of them play a cutting edge function in life's machinery, but what make proteins so unique is their capability to perform cellular functions and thus being ranked as the most versatile macromolecules. The a.a. residues building the protein structures are recruited amongst the 20 naturally occuring ones (Schulz and Schirmer, 1979b), each embedding two essentials chemical functions : amino (H_2N–) and carboxylic acid (HOOC–), the differences lying on the nature of the constituted side-chains groups (R) substituents on the alpha carbon (C_α).

It is worth mentioning that the 20 naturally occuring a.a. bulding block dictate 20^N random possible heteropolymer sequences of lenght N. However, in the midst of this innumerable ensemble of theoretical feasible protein-like sequences, only energetically favourable with relevant folding rates, i.e. thermodynamically and kinetically foldable on the relevant biological timescale will be observed in nature (Onuchic et al., 1997). Indeed, if proteins would have had to search through all their possible random conformations, their folding time would have been on the order of magnitude of the universe's age (Levinthal, Cyrus, 1968).

The composition of the side-chain groups could range from single hydrogen as in the case for glycine (the simplest a.a. residue) to aromatic rings through aliphatic moieties. They are routinely subdivided in four classes, nonpolar (hydrophobic), polar uncharged, negatively charged (basic) and positively charged (acidic) (see Fig. 1.1). In solution, a.a. residues are found essentially as dipolar ions, i.e. NH_2 becomes NH_3^+ after accepting a proton donated by COOH to become COO^-. Therefore, the primary structure of the protein corresponds to the stretched and untangled chain of a.a. residues connected through peptide bonds $\mathrm{—\overset{\overset{\mathrm{O}}{\|}}{C}-\overset{\overset{\mathrm{H}}{|}}{N}—}$ as a result of a condensation reaction between NH_3^+ tail of the first and COO^- tail of the second as schematised in Fig. 1.2. The C_α of all a.a. residues[1] is an asymmetric (chiral) centre, but only one enantiomer (L) is used to make proteins (Taylor et al., 2001).

At this stage, the protein primary structure encoded in the Deoxyribonucleic acid (DNA) has not yet acquired any shape but urges to do so to be biologically active. The pathways by which a protein goes from this messy-like assembly to its native state, also known as protein folding, is perhaps one of the most vital processes in biology. It is intriguing to note that amongst the large set of all possible heteropolymer conformations, each protein naturally encountered has a unique three-dimensional native fold dictated essentially by its sequence. Several factors come into play to maintain the folded globular threedimensional shape of the native state, essentials of which are cohesive forces between the constituent a.a. such as van der Waals and coulombic interactions, hydrogen bonding, hydrophobic packing, as well as covalent interactions including disulfide bridges (Schulz and Schirmer, 1979c).

The folding process is organized by levels of hierarchy (Fig. 1.3), even though intrisically disordered proteins partially keep their initial unstructured shape. Thus, in their native environment soon after the condensation of a.a. residues to set up a ramdom linear chain (Fig. 1.3-A), proteins adopt a well defined (unique) tertiary native structure (Fig. 1.3-C). Throughout the pathways, there is the appearance of canonical local shapes lay out by a.a. backbones connected in an order defining the protein main-chain. Not all the values of the backbone dihedral angles ϕ (C, N, C_α, C) and ψ (N, C_α, C, N) are allowed and each of these local folds known as secondary structures (Fig. 1.3-B) fall within a precise range of (ϕ, ψ) values (Hovmöller, Zhou, and Ohlson, 2002). The secondary structure is then the protein's intermediate level of structural organization and is mainly stabilized by backbone hydrogen bonding networks between spatially neighboring a.a. residues. Two main types of secondary structural motifs are usually adopted by proteins, α-helices and β-strands (see Fig. 1.3-C), that later assemble in tertiary structure. When a protein consists of more than one monomeric units as in the case of hemoglobin, its tertiary structure could further self-aggregate into a complex molecular assembly to form the quaternary structure (Fig. 1.3-D).

[1] Except for glycine

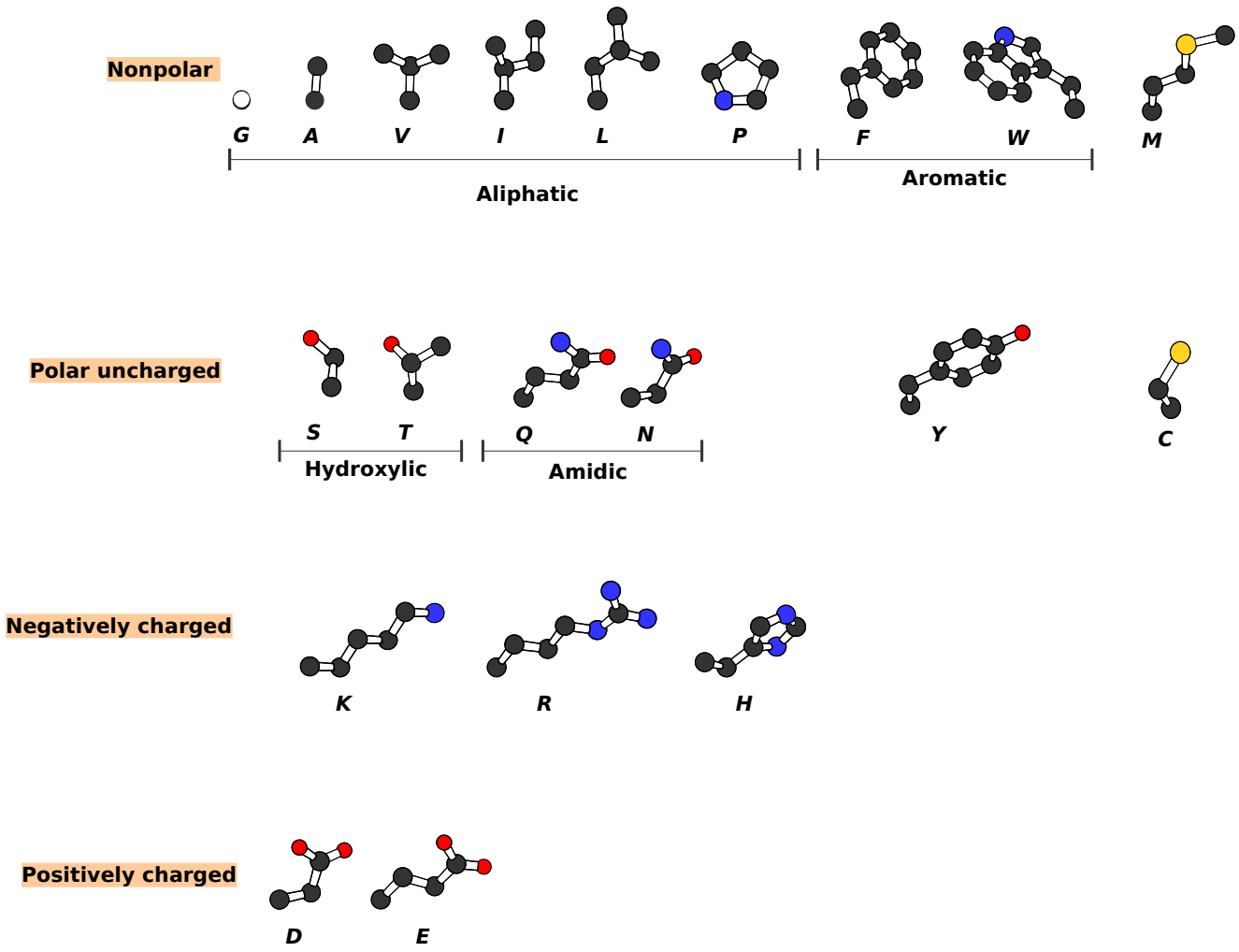


FIGURE 1.1: Representation of side-chains R for the 20 naturally occuring a.a. bulding block, adapted from (Koehl, 2006). The classification depends upon their chemical affinity with water. One-letter designation scheme for each a.a. is used. Red, blue, yellow and dark-grey represent oxygen, nitrogen, sulphur and carbon atoms respectively. Hydrogen atoms are omitted for the sake of lisibility.

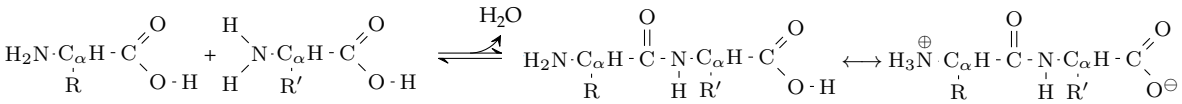


FIGURE 1.2: Illustration of a planar peptide bond resulting from the condensation between amino and carboxylic acid groups of two a.a. residues with side-chains R and R′ to put up a dipetide. The extension of this chemistry leads to a long heteropolymer sequence bulding up the protein primary structure.

Protein structures usually fall within one of these three classes : fibrous, membrane or globular. Fibrous proteins are insoluble in water and their secondary structures dictate almost all their shapes. Membrane proteins can be inserted into phospho-lipid bilayer membrane that surrounds the cells and many sub-cellular

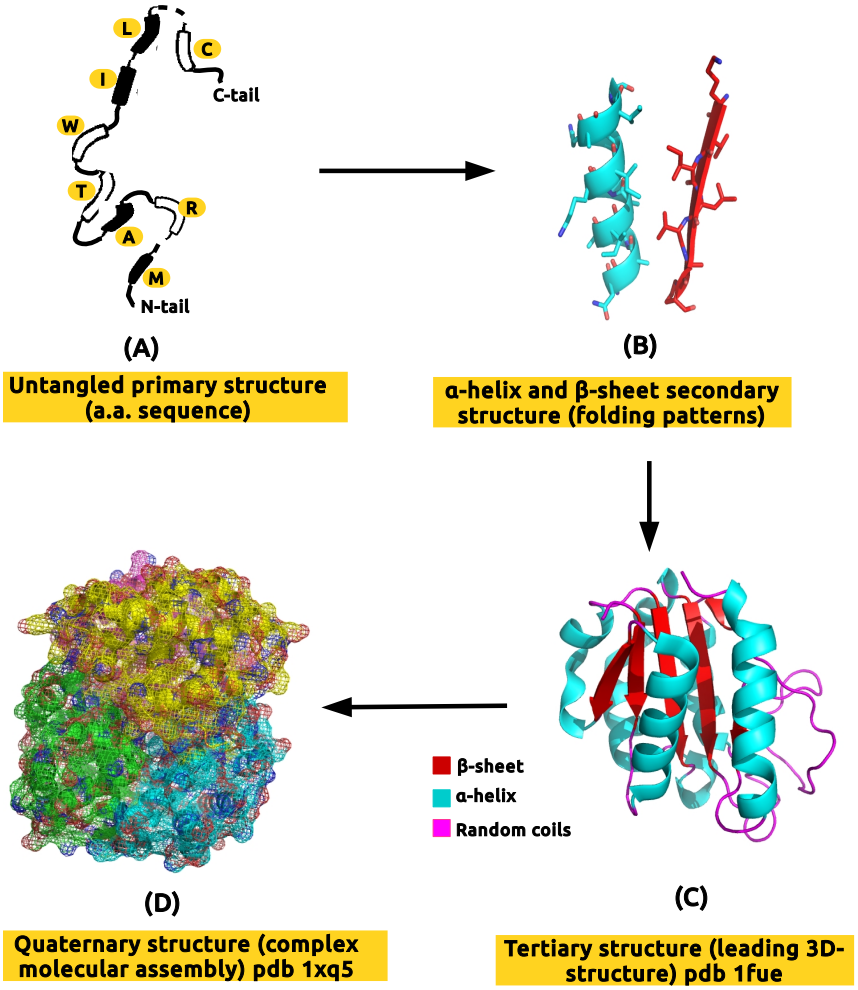


FIGURE 1.3: Schematic representation of folding pathway levels of hierarchy from structureless primary sequence to more complex well organised quaternary fold. Proteins are represented in cartoon view.

organelles. Globular proteins are soluble in water and usually fold into a precise compact block deriving from a non-repetitive a.a. sequence. Further differences among the protein classes can be found in the literature (Taylor et al., 2001; Koehl, 2006; Alberts et al., 2009; Schulz and Schirmer, 1979a). Whereas most globular proteins fold into a unique tridimensional shape, there are instances of structural fluctuation and mis-assembly leading to fibrillar aggregates. This is likely dictated either by an improper protein trafficking or denaturating conditions, albeit a number of proteins also form amyloid deposits under (near)-physiological conditions. More importantly, it is the conformational freedom of protein structures that confers to them such an ability to interconvert between several states.

These insoluble aggregated proteins may damage the cells and even the whole tissue. This precarious conformational state underlies a number of neurodegenerative disorders linked to cognitive impairments such as Huntington and Alzheimer diseases. It is worth recording that the intrinsic forces observed in

proteins should normally drive the correct folding without external factors. Nevertheless, it is seen that molecular chaperones (Ellis and Hartl, 1999; Hartl and Hayer-Hartl, 2002) strongly enhance the protein folding and considerably inhibit fibril formation.

1.2 Amyloids structure & characterization methods

Outstanding progresses have recently been made toward the structure elucidation of amyloid fibrils at a molecular level. As imaged *in vitro* by transmission electron microscopy (TEM) and/or atomic force microscopy (AFM), amyloid fibrils are filamentous assemblies of about 70-130 Å in diameter. They merely consist of nearly 2-6 protofilaments in number, with about 20-50 Å in diameter each that twist around each other to build a thread-like object (Sunde and Blake, 1997; Serpell et al., 2000). The evolvement of experimental protocols makes it possible to prepare samples of isolated amyloid fibrils from tissues, thereby allowing their characterization by a number of techniques such as X-ray diffraction (Eanes and Glenner, 1968; Bonar, Cohen, and Skinner, 1969) and NMR experiments (Lansbury et al., 1995).

It was showed that fibrils are molecular assemblies owning an extended cross-β conformation connected through hydrogen bonding networks displayed along the fibril axis in which their β-strands are arranged perpendicularly to the latter axis (Sunde and Blake, 1997). Several methods have confirmed this structural features of amyloid fibrils like Fourier transform infrared spectroscopy (FTIR) (Zandomeneghi et al., 2004), solid-state nuclear magnetic resonance (ssNMR) (Paravastu et al., 2008) and X-ray crystallography (Serpell et al., 2000). Moreover, as it is the case for many active biological compounds, amyloid fibers could potentially bind small molecule ligands[2] such as thioflavin-T (ThT) and Congo red (CR). Subsequently, upon binding, the well-ordered arrays parallel to fibril axis are imaged when interacting with polarized light (Nilsson, 2004). Therefore, the factors that underlie a protein aggregate as an amyloid fibril are typically the following : cross-β-like secondary structure, fibrillar morphology and characteristic binding abilities (Sipe and Cohen, 2000).

In this thesis we explore the conformatiomal dynamics of TTR (Hörnberg et al., 2000) and β2m (Saper, Bjorkman, and Wiley, 1991), given that conformational changes mediate amyloids formation. Both are well characterized and adopt well-defined tertiary folds in addition to form extracellular-like aggregates. Therefore, they are useful structural models for amyloidogenesis studies.

[2] Essentially dyes

1.3 TTR and β2m : a look into their structure and function

TTR and β2m are among the 37 proteins (Chiti and Dobson, 2017) whose ability and propensity to misfold and form fibrillar aggregates has been associated to human degenerative diseases. Their common pattern is mainly found in their canonical local shapes (secondary structure) as both belong to all-β-like structural class. Indeed, their amyloid potential seems to be inherited from this secondary structural content which has been established to drive amyloid fibrils formation.

1.3.1 Transthyretin

TTR-origin and function

Formerly known as prealbumin, TTR is a 55 kDa globular oligomeric protein made up by four identical monomeric units (I-IV, see Fig. 1.4-(b)) each composed by 127 a.a. residues. Before being released into plasma, TTR is principally produced in liver, choroids plexus and retina (Hamilton and Benson, 2001; Rodrigues et al., 2010; Ferreira et al., 2013; Reinés et al., 2014). It is mainly a carrier protein that binds to retinol binding protein (RBP) to transport vitamin A in plasma and represents the leading transporter of thyroxine (T4) in cerebrospinal fluid (CSF) (Jiang et al., 2001; Lim et al., 2013; Banerjee et al., 2013).

TTR-structure

The 3D-conformation of all TTR-tetramers display great order of similarities as shown at high resolution by Blake and coworkers (Blake et al., 1978) and Hörnberg et *al*. (Hörnberg et al., 2000). TTR structure is overall a rich-β-sheets content protein with a small α-helix domain between strands E and F (see Fig. 1.4-(b)). The structure is highly ordered with a flexible N-terminal region that could not be resolved in many crystallographic studies. In each TTR monomer, about 45 % of residues are arranged in a sandwich immunoglobulin-like topology made up by two four antiparallel-stranded β-sheets; the inner sheet DAGH opposed to the outer sheet CBEF. Its secondary structure also exhibits a short β-strand portion A* (see Fig. 1.4-(c)) which is antiparallel (folded back) to strand A through a π-turn (i+5) and which is involved in dimer-dimer contact (Haupt et al., 2014).

The dimer I-II (Fig. 1.4-(c & d)) also known as primary dimer is the first crystallographic asymmetric unit structurally symmetric to the dimer III-IV. It is stabilized by a network of six main chain hydrogen bond interactions established by three pairs of residues (A120(NH)⇆(OC)Y114, T118(NH)⇆(OC)Y116, Y116(NH)⇆(OC)T118) at the HH' interface (Fig. 1.4-(a)). Monomer I inner sheets DAGH form with its homologue H'G'A'D' in monomer II a kind of symmetric antiparallel pseudo-continous eight β-sheets like network centered on HH' (DAGH-H'G'A'D'), Fig. 1.4-(c) (Haupt et al., 2014; Cianci et al., 2015). Whereas at the opposite side, the pseudo-continuity in outer sheets CBEFF'E'B'C'(Fig. 1.4-(d)) is rather loose and only 4 backbone hydrogen bonds (F87(CO)⇆(HN)T96 and E89(NH)⇆(OC)V94)

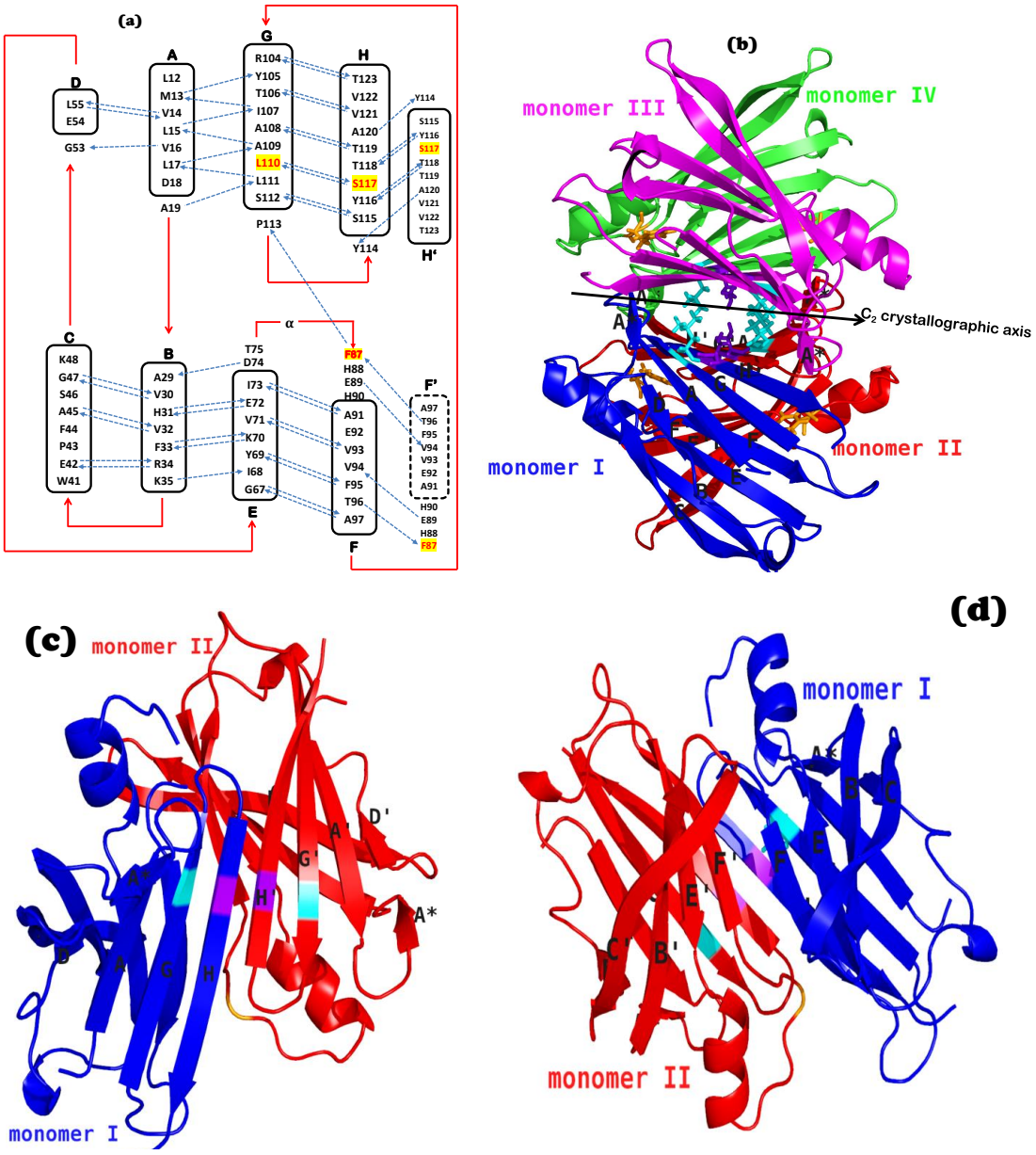


FIGURE 1.4: **(a)** Topology arrangement of main chain - main chain hydrongen bonds (blue lines) within WT-TTR monomeric subunits and at monomer - monomer interfaces H-H′ **(c)** and F-F′ **(d)**, where the letters A to H designate the β-strands in one monomer while H′ and F′ belong to the other monomer of the same dimer. Red arrows indicate the connection between strands and show the loop regions and blue ones go from donnor (NH) to acceptor (OC). Coloured residues[3](yellow) are the ones invloded in the mutation and α is the helical region between strands E and F. **(b)** Cartoon view of tetrameric assembly of WT-TTR on which the residues involved in the mutation (F87 (orange), L110 (cyan) & S117 (purpleblue)) are shown explicitely (stick representation) as well as the label of different β-strands schematised in **(a)**. PDB id 1F41 (Hörnberg et al., 2000)

are established (Fig. 1.4-(a)). While dimers are mainly kept together by hydrogen bonds at monomer-monomer interface, tetramers result to be the dimers of dimers essentially stabilized by hydrophobic contacts established within neigbouring subunits at the dimer-dimer interface via AB-GH loop interactions involving mainly residues L17, A19, V20, L110, P113, T119 and V121.

TTR-related amyloidoisis

TTR-related amyloidosis can be inherited in the case of genetic mutations or can be non-hereditary when it is due to wild-type (WT-TTR) (Dubrey, Ackermann, and Gillmore, 2015). WT-TTR amyloidosis also termed senile systemic amyloidosis (SSA) is the prevalent form of TTR amyloidosis which principally affects the heart. It develops with ageing and requires TTR tetramer dissociation and partial unfolding. Mutation-induced TTR amylodosis is instead associated to familial amyloid polyneuropathy (FAP) which impatcs essentially on nervous system and familial amyloid cardiomyopathy (FAC) with broad implication on heart (Banerjee et al., 2013; Dubrey, Ackermann, and Gillmore, 2015). The inter-connection between protein structure and stability and its capability to form amyloid fibrils has motivated several structural studies. To date, about 80 point mutations have been correlated to human inherited amyloidosis (Pasquato et al., 2007).

Motivations to triple mutant engineering mutation

WT-TTR displays a very stable tetrameric molecular assembly. It is established that its structure is resistant to dissociation at physiological concentrations within the pH range 5 - 7 (Lai, Colón, and Kelly, 1996). Several amyloid fibril formation models have been proposed and reviewed (Chiti and Dobson, 2006; Invernizzi et al., 2012; Chiti and Dobson, 2017). However, it is almost a consensus that common mechanism involves the dissociation of the native tetrameric TTR into unstable but folded monomers, followed by local unfolding of the latters into multiple non-native amyloidogenic intermediate states that self-assembled in solution (see Fig.1.5) (Chiti and Dobson, 2006; Rodrigues et al., 2010; Invernizzi et al., 2012; Lim et al., 2013; Greene et al., 2015; Chiti and Dobson, 2017). Furthermore, evidence has been provided pointing to the dissociation of the native tetramer as the rate-limiting step towards aggregation (Lashuel, Lai, and Kelly, 1998; Quintas, Saraiva, and Brito, 1999; Quintas et al., 2001; Jiang, Buxbaum, and Kelly, 2001; Jiang et al., 2001).

It arises therefore that knowledge of the dynamics and dissociation of TTR native structure is an important issue as tetramer dissociation, monomer misfolding and self-assembly of amyloidogenic monomers into amyloid and other aggregate morphologies is known to be linked to several human degenerative diseases. Specifically, TTR represents one of the few examples whereby disease intervention, through the stabilization of a native structure, has yielded a drug for TTR amyloidogenesis (Johnson et al., 2012), thus indicating that thermodynamics studies and modeling of tetramer dissociation is a very important target.

[3]More details in Chapter 4

To mimic the pathological situation, previous site-directed mutagenesis (mutations Phe87Met/Leu110Met (MT-TTR)) was carried out to promote the dissociations of the TTR tetramer into monomers (Jiang et al., 2001). To further shift the equilibrium towards monomers to obtain a more homogeneous dissociated species, an additional a.a. replacement has been introduced in MT-TTR molecule (Ser117Glu) by Berni and coworkers. Indeed, the latter mutant containing three mutations (3M-TTR) proved to be markedly more prone to *in vitro* aggregation in comparison with the double mutant MT-TTR (Zanotti et al., 2017).

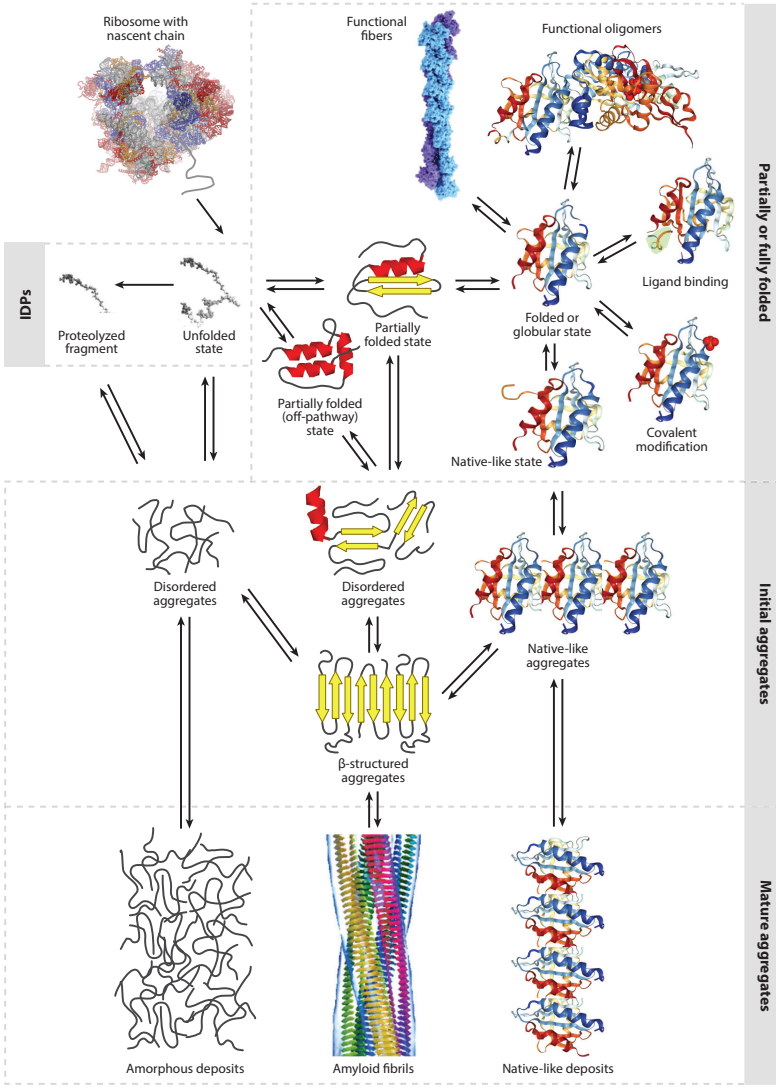


FIGURE 1.5: Schematic illustration of possible intermediate conformational states and their inter-conversion pathways adopted by various heteropolymer polypetides since their biosynthesis in the ribosome. From bottom left to bottom right respectively, common protein aggregation pathways can lead to the formation of amorphous deposits (as in the case of α-synuclein and islet amyloid polypeptide), amyloid fibrils (such as in TTR and β2m) or native-like deposits (as seen in prion protein and insulin). The figure is taken from (Chiti and Dobson, 2017).

TTR-amyloidosis therapeutics

As mentioned previously, the dissociation of TTR tetramer is the initial step into its fibrillation pathway (Kelly and Lansbury, 1994). Therefore, many strategies to TTR amyloidosis prevention exploit its ability to bind small molecules in the T4 binding channel (see Fig. 1.4-b), mimicking its hormone binding capability, thereby producing kinetic stabilization of the tetramer (Peterson et al., 1998; Almeida et al., 2005). Other clinical remedies to FAC and FAP amyloidosis often employ organs transplantation (e.g. liver), even though not all the affected organs can be transplanted (e.g. choroids plexus), where TTR is procuded as well (Ando and Ueda, 2012). Furthermore, additional treatments against TTR amyloidosis have been experimented, among which the exercise of gene therapy (Nakamura and Ando, 2004) and the resorption of amyloid deposits (Sebastião et al., 2000), just to cite two.

1.3.2 β2m

β2m-origin and function

β2m is a 99-residue subunit of the major histocompatibility complex class I (MHC I). With a molecular mass of about 12 kDa, it is a small β-sandwich globular protein interacting noncovalently with the human leukocyte antigen HLA-A2 through its α-chain. Thus, the conformation of α-chain is dependent on the presence of β2m. Therefore, being the subunit of the MHC I, its biological role appears to be more structural. Upon dissociation from MHC I, β2m is released in the blood and is essentially cleared by glomerular filtration followed by proximal tubular (in the kidneys) reabsorption and catabolism (Floege et al., 1991; Scarpioni et al., 2016). In renal insufficient patients undergoing long-term dialysis, it is the responsible for dialysis related amyloidosis (DRA) where insoluble amyloid fibrils of the protein are deposited in joints and connective tissue (Gejyo et al., 1985; Yamamoto and Gejyo, 2005; Naiki et al., 2016).

β2m-structure

The secondary structure of β2m consists of seven β-strands A to G (Fig. 1.6 (Fogolari et al., 2011)) assembled into two antiparallel pleated β-sheets (of 3+4 β-strands) connected by a central disulfide bridge (linking strands B & F) (see Fig. 1.7) highly resembling a β-sandwich immunoglobulin-like type (Ig) C1 domain (Becker and Reeke, 1985; Saper, Bjorkman, and Wiley, 1991; Verdone et al., 2002). No transmembrane domain is found in its structure and it holds a characteristic molecular assembly called a constant-1 Ig superfamily domain shared with other adaptative immune molecules including MHC I and II (Ohta et al., 2011).

β2m-related amyloidoisis

β2m is mainly responsible for dialysis-related amyloidosis (DRA). DRA is a common incidence of both chronic hemodialysis and peritoneal dialysis, resulting from the increase in the protein level in the serum of patients affected by renal dysfunction. This abnormal incresase in protein concentration leads to the

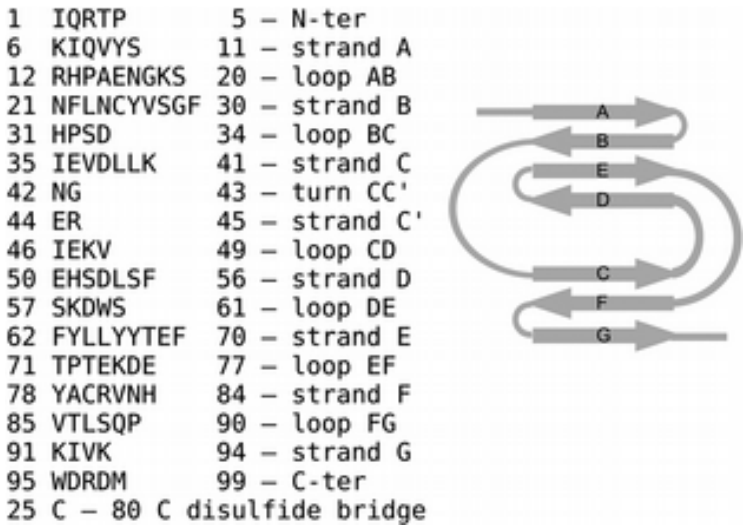


FIGURE 1.6: Schematic view of the topology of β2m alongside with the secondary structure description.

maturation of amyloid fibrils that accumulate principally in the osteoarticular tissues (ligaments, bone, muscle, etc.) and viscera, causing organs dysfunctions like carpal tunnel syndrome and bone cysts. It is worth noting that at pH 7 β2m structure is well folded and does not spontaneously forms amyloids (Bellotti et al., 1998; Myers et al., 2006), albeit its concentration is steady high in patients undergoing long-term haemodialysis (Gejyo et al., 1985). Therefore, the extrinsic factors that potentially trigger β2m amyloid formation have been investigated.

It was seen that interaction with metal ions such as Cu^{2+} destabilizes the folded state of the protein against thermal and urea denaturation, thus promoting amyloid fiber formation (Morgan et al., 2001; Calabrese and Miranker, 2007; Calabrese and Miranker, 2009). The acidic pH (McParland et al., 2000; Narang, Singh, and Mukhopadhyay, 2017) and the interaction with lysophosphatidic acid (Yamamoto et al., 2004) also induce the formation of β2m amyloid plaques under physiological conditions. One variant of β2m, ΔN6, a truncated conformer involved in DRA was found to be more stable than wild-type (2.5 $kcalmol^{-1}$ lower) (Esposito et al., 2000; Giorgetti et al., 2005; Eichner and Radford, 2011) and capable of triggering the fibrillogenesis process of the wild-type by promoting its direct conversion to a highly prone aggregation intermediate state, mimicking the pathogenic action of prion proteins (Eichner and Radford, 2009; Eichner et al., 2011). Nevertheless, in spite of the multiple hypothesis drawn, its detailled fibrillogenesis mechanism remains elusive. However, in general, the conformational dynamics of the natively folded wild-type favoured by the *cis* to *trans* His31-Pro32 shift to promote the non-native aggregation-competent intermediate species[4] is commonly accepted as the initiation step into the β2m amyloid formation pathway (Verdone et al., 2002; Eichner et al., 2011) in a process globally schematized in Fig. 1.5.

[4]Usually a partially unfolded state is a prerequisite into amyloid fibrils

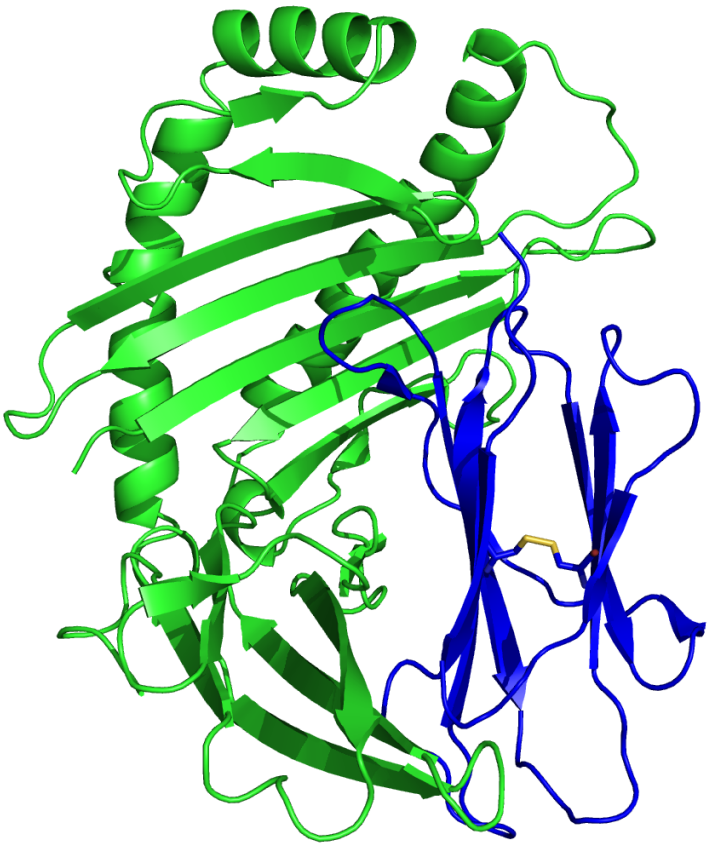


FIGURE 1.7: Cartoon view of β2m structure (blue) bounds to human histocompatibility antigen HLA-A2 (green). Residues Cys25 and Cys80 connecting strands B and F through disulfure bridge (yellow) are shown in stick representation. PDB id 3hla (Saper, Bjorkman, and Wiley, 1991)

Motivation to the *in silico* studies of β2m structure in this work

Our group has a long term expertise in characterizing[5] a number of amyloidogenic proteins including β2m (Esposito et al., 2000; Verdone et al., 2002; Esposito et al., 2005; Fogolari et al., 2007; Corazza et al., 2010; Fogolari et al., 2011). Besides, its structure is in overall well described, facilitating comparison with computer generated microscopic data. More specifically, β2m is a quite small system in size, therefore adequate model for molecular dynamics simulations. In particular, a NMR method, Bluu-Tramp, has recently been developed by Prof. Esposito and coworkers (Rennella et al., 2012a; Rennella et al., 2012b) that allows the estimation of (sub)-global unfolding thermodynamics parameters (free energy, enthalpy, entropy and heat capacity) at a single residue resolution in a single experiment using two temperature ramps. These data will serve to benchmark our

[5]Either *in silico* or *in vitro* using mainly NMR experiments

atomistic MD simulations and improve our knowledge of the temperature dependence of the free energy, namely the entropy, that still lacks full description at the atomic level. On the other side, β2m is a promising cancer therapeutic target (Shi et al., 2009) and is part of the FDA-Approved OVA1 panel's biomarkers (including also TTR) for ovarian cancer (Rein et al., 2011), striking to a detailed investigation of its molecular structure as well as a number of factors promoting its unfolding, typically at the atomistic scale resolution.

β2m amyloidosis therapeutics

Amyloids deposits usually impaired the normal regulation of vital organs. As for FAC and FAP amyloidosis, organs transplantation is a cornerstone therapeutical strategy. Therefore, renal replacement has been proposed for the prevention and treatment of DRA (Campistol, 2001). In addition, a number of dialysis methods such as hemofiltration and hemodiafiltration (Rabindranath et al., 2005) and the use of biocompatible and high-flux membranes (Hoshino et al., 2016) are often employed in this regard. Recently, *in vitro* and *in silico* investigations showed that the fibrilogenis of even a pathological variant of β2m, D76N, is hindered by citrate-gold nanoparticles (Cantarutti et al., 2017). Moreover, a specific nanobody Nb24 was proved to efficiently block the fibrillogenesis of the latter variant responsible for hereditary systemic amyloidosis of β2m (Raimondi et al., 2017). These latest findings open up promising therapeutic issues.

1.4 Summary

This chapter has covered the key notions behind protein assemblies in general and amyloidogenic proteins in particular, bearing emphasis on TTR and β2m, the two model systems explored in this thesis. Our understanding of protein conformational dynamics have greatly benefited from years of experimental endeavours, primarily using X-ray crystallography and/or NMR spectroscopy methods that have now been considerably evolved. However, the dissection of protein misfolding and aggretation at a molecular level raises a spectrum of intricate methodological issues mainly related to the size, heterogeneity and transient hallmark of the intermediates involved. In this regard, *in silico* techniques and mainly molecular dynamics (the focus of the forthcoming chapter) even on shorter timescale proved to be markedly of help, highlighting detailed mechanistic and structural insights to complement and benchmark experiments (Karplus and McCammon, 2002; Ma and Nussinov, 2006; Vaart, 2006; Fogolari et al., 2007; Scheraga, Khalili, and Liwo, 2007; Moroni, Scarabelli, and Colombo, 2009; Dror et al., 2010; Avila et al., 2011; Fogolari et al., 2011; Dror et al., 2012).

Chapter 2

Molecular dynamics : Overview and fundamentals of the method

This chapter emphasizes on the essential traits peculiar to computer simulation techniques in general and in molecular dynamics (MD) approach in particular. It starts by discussing the theoretical backgrounds of the method in the context of classical MD simulations, bearing interest in all-atoms representation of the system, using either explicit or implicit solvent models since both are employed in this thesis. Thereafter, some enhanced MD simulation strategies also applied to part of the MD simulations in this thesis are discussed. The accuracy of computer experiments depends on the reliability of the computational procedure also called algorithm. Nowadays, developing faster and more efficient algorithms occupies a very good place in computer-based research community. Another key aspect and perhaps the most important is the interaction potential often called empirical energy function or force-field, which describes the interaction between different atoms within the system. The availability of accurate force-fields whose parameters are tuned to compensate for the simplifications introduced in the model is a prerequisite for all reliable MD simulations. This chapter closes up by highlighting part of the force-field improvement initiated by using the Generalized Born-Suface Area (GB/SA) continuum solvation model of Onufriev-Bashford-Case (OBC) (Onufriev, Bashford, and Case, 2004), to refit the Born radii and OBC parameters of the amber99sb-ildn force-field (Lindorff-Larsen et al., 2010) to best reproduce the average solvation forces in comparison to the explicit solvent model. Albeit we could not foster much more attention on this latter aspect in this thesis, mainly because of time constraints, it shows up promising results that could be extended later on.

2.1 Classical molecular dynamics

MD simulations are largely employed in materials science and biophysics. Nowadays, the steady increase in computer powers, alongside with the heavily optimized and massively parallelized codes have extensively raised by several orders of magnitude the system sizes and timescales accessible in most MD computations. Subsequently, although short, atomic-level simulations can now sample the timescales relevant for fundamental biological processes such as small and large-scale conformational changes, protein folding, stability, flexibility and recognition (ligands, DNA, membranes). By providing highly spatial (atomistic) and temporal (femtosecond) resolution than most experimental techniques, MD simulation is becoming a computational microscope and an invaluable tool in molecular biology.

2.1.1 Basic ideas of MD simulation

Compared to electrons (which constitute with nuclei the building block of atoms), nuclei are heavy enough to be treated as classical point objects with mass m_i and are modelled independently from electrons[1]. Indeed, the mass of an atom is essentially concentrated in its nucleus. The motion of a set of atoms can then solely be related to their nuclei motions. Consequently, the time dependence of a conservative system should follow the Newtonian mechanics. In clear, MD simulation computes the time-based equilibrium and transport properties of a classical many-body system, by numerically solving the Newton's equations of motion (2$^{\text{nd}}$ Newton's law), see Eq. 2.1.

$$\mathbf{F}_i = m_i \mathbf{a}_i = m_i \frac{d^2 \mathbf{r}_i}{dt^2}, \quad \text{with} \quad i = 1, 2, \ldots, N \tag{2.1}$$

With all-atoms description of the system, the force acting on atoms can also be expressed as the negative first order derivative of the potential with respect to the spatial coordinates, i.e. the negative gradient of the potential energy function, $\mathbf{F}_i = -\boldsymbol{\nabla}_i U$. Thus,

$$m_i \frac{d^2 \mathbf{r}_i}{dt^2} = -\boldsymbol{\nabla}_i U(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) = -\frac{\partial}{\partial \mathbf{r}_i} U(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N) \tag{2.2}$$

In Eqs. 2.1 and 2.2, $\mathbf{F}_i$ is the force acting on the i^{th} particle of the system, m_i, r_i and $\mathbf{a}_i$ its mass, spatial coordinates and acceleration respectively. $U(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N)$ is the system potential energy. N is the total number of particles in the system and t is the time. Bold-written variables are vector quantities.

Since Eq. 2.2 describes the classical state of the system, it enables the complete determination of the matrix of atomic coordinates and velocities at any time, given

[1] Born-Oppenheimer approximation (Born and Oppenheimer, 1927)

the initial positions and velocities. However, in general, due to the high complexity nature of the potential energy function, it is not possible to provide an analytical solution to Eq. 2.2, apart from particular cases. Anyway, a unique solution to the latter equation relies on the initial set of coordinates and velocities, prerequisites for a MD computation. Therefore, an approximate solution to Eq. 2.2 is obtained step-by-step, based on numerical integrations. In all MD simulations, the trajectory consists of a suite of snapshots that represents each microstate of the system. Once at equilibrium and provided that the system has sampled enough microstates, one could then estimate the macroscopic ensemble[2].

The procedural workflow of a MD calculation embodies the followings :

1. Bulding up the system by defining the set of inputs including the starting coordinates/velocities ($\boldsymbol{r}_i/\boldsymbol{v}_i$), but also the system topology, the interaction potential ($U(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N)$), temperature ($T$), integration time-step (Δt), etc.;
2. Compute the forces exerted on different atoms in the system : $\boldsymbol{F}_i = -\frac{\partial}{\partial \boldsymbol{r}_i} U$;
3. Numerical integration of Newton's equations of movement : $\frac{d^2\boldsymbol{r}_i}{dt^2} = \frac{\boldsymbol{F}_i}{m_i}$, $\boldsymbol{v}_i = \frac{d\boldsymbol{r}_i}{dt}$, $\frac{d\boldsymbol{v}_i}{dt} = \frac{\boldsymbol{F}_i}{m_i}$;
4. Loop steps 2 and 3 until the chosen time-scale;
5. Estimate the average computed observables.

Bulding up the system

Independently of the type of MD computation one is interested in, one must always define the initial set of coordinates and velocities. As far as protein structures are concerned, the coordinates that could be either a single frame (when coming from X-ray) or a set of conformers (when coming from NMR solution structure) are often retrieved from the Protein Data Bank[3] (Berman et al., 2000). Atomic velocities v_i are usually assigned randomly based on the system temperature T, which is also an input term, according to the Maxwell-Boltzmann distribution function $\mathcal{P}(v_i)$, following the relation $\mathcal{P}(v_i) = \sqrt{\frac{m_i}{2\pi k_\mathrm{B} T}} exp\Big(- \frac{m_i v_i^2}{2k_\mathrm{B}T}\Big)$, where k_B is the Boltzmann's constant.

Forces computation

The force acting on each particle is related to the interaction energy function (U). Subsequently, the computation of forces involves the calculation of bonded (bond, angle, dihedral) and non-bonded (electrostatics and van der Walls) energy terms, see Molecular Mechanics Force Field. This step is a very time consuming one, notably for the computation of the non-bonded energy terms which could scale up to an order of magnitude of N^2 for each unit cell, N being the total number of atoms. In the simulation, one usually replicates the unit cell employing

[2]Ergodic principle : the numerical solution (time average) is statistically equivalent to the true solution (ensemble average) within a truncation error.

[3]Coordinates could also be obtained *in silico* by homology modeling

the periodic boundaries conditions (PBC) in order to avoid the finite system effect and to simulate the bulk system. In the PBC-like description, a particle that leaves the simulation cell along one face re-enters on the opposite face and only views the closest image of other atoms. This condition underlies the minimum image convertion that prevents a particle self-interaction, making non-physical models.

To circumvent the computational effort associated with the forces calculation, a number of approaches are introduced. The so-called Verlet list (Verlet, 1967) is often used to truncate the potential and to consider only interactions that fall within the truncated sphere with radius r_{cut}, see Eq. 2.3. However, this applies only for short-range interactions such as van der Walls. The cut-off scheme is simply not enough in treating the long-range interactions like coulombic ones. Thus, MD often employs the Ewald summation technique. It was recently implemented by the Particle-mesh Ewald algorithm (Darden, York, and Pedersen, 1993) which scales as *Nlog(N)* and results sufficiently faster than conventional Ewald approaches.

$$U^{\text{trunc}}(\boldsymbol{r}) = \begin{cases} U(\boldsymbol{r}) - U(\boldsymbol{r}_{\text{cut}}) & \boldsymbol{r} < \boldsymbol{r}_{\text{cut}} \\ 0 & \boldsymbol{r} \geq \boldsymbol{r}_{\text{cut}} \end{cases} \tag{2.3}$$

Step-by-step numerical integration

A solution to Eq. 2.1 is extremely difficult to be provided by a poorly intuitive computer solving routines. Conversely, computers readily solve simple equations such as algebraic ones. The second-order differential term in Eq. 2.1 must therefore be expressed as an algebraic expression, using the following Taylor series expansion, Eq. 2.4 :

$$x(t+\delta t) = x(t) + \delta t \frac{dx(t)}{dt} + \frac{1}{2!}\delta t^2 \frac{d^2x(t)}{dt^2} + \frac{1}{3!}\delta t^3 \frac{d^3x(t)}{dt^3} + \dots \tag{2.4}$$

In Eq. 2.4, if x does not significantly evolve with time, the higher order differential terms can be neglected for a suitable small value of the time-step δt. However, still, another expression of Taylor series expansion is required to approximate the second-order differential term in Eq. 2.1 :

$$x(t-\delta t) = x(t) - \delta t \frac{dx(t)}{dt} + \frac{1}{2!}\delta t^2 \frac{d^2x(t)}{dt^2} - \frac{1}{3!}\delta t^3 \frac{d^3x(t)}{dt^3} + \dots \tag{2.5}$$

By summing up side-by-side Eqs. 2.4 & 2.5, the second-order differential term can be given by :

$$\frac{d^2x(t)}{dt^2} = \frac{x(t+\delta t) - 2x(t) + x(t-\delta t)}{\delta t^2} + \vartheta(\delta t^4) \tag{2.6}$$

$\vartheta(\delta t^4)$ on the right-hand side in Eq. 2.6 defines the accuracy of the approximation, and in this case, all the terms higher than or equal to δt^4 are neglected. Eq. 2.6 could then be simply written as :

$$\frac{d^2x(t)}{dt^2} = \frac{x(t+\delta t) - 2x(t) + x(t-\delta t)}{\delta t^2} \tag{2.7}$$

Eq. 2.7 is known as the *central difference approximation*. Accordingly, the equation of the x-component of Newton's law of movement turns in the following :

$$x_i(t+\delta t) = 2x_i(t) - x_i(t-\delta t) + \frac{\delta t^2}{m_i} F_{xi}(t) \tag{2.8}$$

where $\boldsymbol{r}_i = r_i(x_i, y_i, z_i)$ are the molecular positions and $\boldsymbol{F}_i = (F_{xi}, F_{yi}, F_{zi})$ are the forces acting on particle i.

Similarly, equations for (y_i, z_i) components are easily derived. Given that Eq. 2.8 is a simple algebraic equation, the molecular positions at the next time-step can be estimated using the present and previous positions and the present forces. Eq. 2.8 does not require the velocity components for computing the atomic positions at the next time-step. It is termed the **Verlet Method** (Verlet, 1967). Nevertheless, from the trajectory, one could derive the velocity as follows :

$$\boldsymbol{r}_i(t+\delta t) = \boldsymbol{r}_i(t) + \dot{\boldsymbol{r}}_i(t)\delta t + \frac{\ddot{\boldsymbol{r}}_i(t)}{2!}\delta t^2 + \cdots = \boldsymbol{r}_i(t) + \boldsymbol{v}_i(t)\delta t + \frac{1}{2}\boldsymbol{a}_i(t)\delta t^2 + \ldots \tag{2.9}$$

$$\boldsymbol{r}_i(t-\delta t) = \boldsymbol{r}_i(t) - \dot{\boldsymbol{r}}_i(t)\delta t + \frac{\ddot{\boldsymbol{r}}_i(t)}{2!}\delta t^2 - \cdots = \boldsymbol{r}_i(t) - \boldsymbol{v}_i(t)\delta t + \frac{1}{2}\boldsymbol{a}_i(t)\delta t^2 - \ldots \tag{2.10}$$

By substracting hand-by-hand Eqs. 2.9 & 2.10, we obtained $\boldsymbol{v}_i(t)$ as in Eq. 2.11 :

$$\boldsymbol{r}_i(t+\delta t) - \boldsymbol{r}_i(t-\delta t) = 2\boldsymbol{v}_i(t)\delta t \Rightarrow \boldsymbol{v}_i(t) = \frac{\boldsymbol{r}_i(t+\delta t) - \boldsymbol{r}_i(t-\delta t)}{2\delta t} \tag{2.11}$$

It has previously been seen that the velocities are not required for estimating the positions at the next time step. Nonetheless, an approach using both atomic positions and velocities may be more adequate in keeping the system temperature constant. Such schemes include the **Velocity Verlet Method** that employs formulae displayed in Eq. 2.12 & Eq. 2.13 for determining the motion of molecules and provides significant stability and accuracy in the simulation and the Leap Frog algorithm.

$$\boldsymbol{r}_i(t+\delta t) = \boldsymbol{r}_i(t) + \delta t\boldsymbol{v}_i(t) + \frac{\delta t^2}{2m_i}\boldsymbol{F}_i(t) \tag{2.12}$$

$$\boldsymbol{v}_i(t+\delta t) = \boldsymbol{v}_i(t) + \frac{\delta t}{2m_i}[\boldsymbol{F}_i(t) + \boldsymbol{F}_i(t+\delta t)] \tag{2.13}$$

The **Leap Frog algorithm** evaluates the velocities at half-integer time steps and uses these velocities to compute the new positions using Eqs. 2.14 & 2.15 :

$$\boldsymbol{r}_i(t+\delta t) = \boldsymbol{r}_i(t) + \delta t \boldsymbol{v}_i(t+\frac{\delta t}{2}) \tag{2.14}$$

$$\boldsymbol{v}_i(t+\frac{\delta t}{2}) = \boldsymbol{v}_i(t-\frac{\delta t}{2}) + \frac{\delta t}{m_i}\boldsymbol{F}_i(r(t)) \tag{2.15}$$

This scheme is an alternative to the Verlet's one, and so, it leads to similar trajectories with the difference that the velocities are not defined at the same time as the positions. As a consequence, kinetic and potential energy are also not defined at the same time, and hence we cannot directly compute the total energy in the Leap Frog scheme.

2.1.2 Molecular Mechanics Force Field

MD simulations of biomolecular systems use force fields resulting from the fitting of experimental data or quantum mechanical and solvation free energies calculations to fit into simple functional forms. Accurate simulations therefore rely on the avalaibility and precision of the force field used to characterize the interactions. The force field, the mathematical model used to approximate the atomic-level forces acting on the simulated molecular system (Lindorff-Larsen et al., 2012) is a function of the potential terms that can be summed up as follows :

$$\begin{aligned} U = &\sum_{bonds} \frac{1}{2}k_r(r-r_0)^2 + \sum_{angles} \frac{1}{2}k_\Theta(\Theta-\Theta_0)^2 + \sum_{dihedrals}\sum_{n} \frac{1}{2}k_\psi[1+cos(\eta\psi+\Upsilon)] \\ &+ \sum_{impropers} \frac{1}{2}k_\omega(\omega-\omega_0)^2 + \sum_{i<j} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i<j}\frac{q_i q_j}{r_{ij}} \end{aligned} \tag{2.16}$$

In Eq. 2.16, the first four components are bonded potentials that describe bonds, angles, proper dihedrals and improper torsions of the covalent structure, respectively. The last two terms run over all pairwise atoms *i* and *j* separated from each other by a distance $r_{ij} = |r_j - r_i|$ and showcase the non bonded interactions. The bond stretching and bond bending (1^{st} and 2^{nd} terms) model the energetic change accompanying the deformations of the bond lengths *r* and bond angles Θ from their respective equilibrium values r_0 and Θ_0. These two interactions are mainly computed using an harmonic-like potential with force constants k_r and k_Θ. In the simulation, only a very small distortion is monitored in bond and angle fluctuations as they mostly oscillate around their equilibrium values. In clear, a high potential energy is associated to the deformation of bond stretching and bond bending.

The third and fourth terms account for energetics linked with bond rotations. Thus, k_ψ is the height of the rotational barrier associated to the proper dihedral angle ψ characterized by the torsional angle phase Υ for each Fourier component η (periodicity). Improper torsions ensure planarity in aromatic rings and allow distinguishing molecules from their mirror images in which k_ω is the force constant for the improper dihedral ω going up and down its equilibrium position ω_0. The last two components of Eq. 2.16 describe the van der Waals repulsive (at short distance, r^{-12} term) and attractive (at long distance, r^{-6} term) pairwise atomic forces between *i* and *j* shown as Lennard-Jones 12-6 potential and the electrostatic interactions. The variables q_i and q_j are the partial charges on pairwise atoms i and *j* separated by the distance r_{ij}, σ_{ij} is the distance at which the Lennard-Jones potential is zero and ϵ_{ij} is the well depth.

There are various levels of system force field representation, namely all-atoms in which individual atom interactions are accounted and mesoscopic models such as coarse-grained representations, in which a subset of atoms is clustered together into a bead centre. All the simulations in this work have employed an all-atoms type and thus, further description of coarse-grained paradigms will be overlooked here. The commonly used macromolecular empirical all-atoms potential functions include one the updated version of the following, AMBER (Pearlman et al., 1995), CHARMM (Brooks et al., 1983) and GROMOS (Gunsteren and Berendsen, 1987). The simulations described in Chapter 4 have employed the Charm27 version with CMAP correction (Bjelkmar et al., 2010) while in Chapter 5 & Chapter 6 the Amberff99SB (Lindorff-Larsen et al., 2010) have been used, both with an all-atom scheme. The latter force field representation could further be coupled either with an explicit solvent model (in Chapter 4 & Chapter 6) or with a continuum (implicit solvent) model (Chapter 5).

Explicit solvent

The explicit incorporation of water molecules in biomolecular simulations (and which usually accounts for > 90% of the total number of atoms) provides a more realistic-like environment, though it significantly increases the degrees of freedom and thereby the computational demand. Several water models have been developed in MD computations in spite of its small number of atom and size. Commonly used are SPC and SPC/E (Berendsen, Grigera, and Straatsma, 1987), TIP3P and TIP4P (Jorgensen et al., 1983) and TIP5P (Mahoney and Jorgensen, 2000). They mainly differ from one to another either by the number of interacting sites, the polarizability and/or the flexibility of covalent bonds. No extra-site is added in rigid 3-site models such as TIP3P and SPC. Besides, oxygen and hydrogen partial charges are indeed located on their respective ions. Meanwhile, in the rigid 4-site (TIP4P) and 5-site(TIP5P) water models, 1 and 2 virtual sites are added, respectively. In the former case, a dummy atom is placed in the bisector plane of the bond-bending angle H-O-H and in the latter case, partial oxygen charge is splitted in 2 virtual interacting sites for each lone electron pair to form a tetrahedral-like shape. In all the above models, the water molecule holds an effective dipole moment of nearly 2.3 D compared to 1.85 D for the experimental

gas-phase water model (Adcock and McCammon, 2006). Schemes accounting for the polarization effects (Sprik and Klein, 1988) and flexible bond stretching and bond bending terms (Ferguson, 1995) induce further computational costs, as a result, the most used ones are mainly limited to 3-site water models.

Implicit solvent

Outstanding lenght of simulation time is now been achieved (up to millisecond or so) in biomolecular simulations. This performance is mainly due to the steady increase in computer power and achitecture (CPU, GPU), massive parallelisation (MPI, openMP, ect.), enhanced sampling algorithms, and so on. However, one way to reduce the number of degrees of freedom (thereby accelerating the sampling) is the use of implicit solvent models (Roux and Simonson, 1999; Cramer and Truhlar, 1999; Fogolari, Brigo, and Molinari, 2002; Feig and Brooks, 2004; Onufriev, 2008; Kleinjung and Fraternali, 2014; Decherchi et al., 2015). Aqueous milieu is a more realistic biological environment, striking to the importance of solvent interactions in protein conformational changes, notably in accurate treatment of electrostatic effects. Thus, a core issue in implicit solvent is to find rigorous potential functions that better account for solvation effects and effective change in solute conformational free energy. This is achieved mainly in MD employing generalized Born (GB) formalisms (Still et al., 1990; Qiu et al., 1997; Bashford and Case, 2000).

MD simulations mainly strive to reproduce the potential of mean force of a solvated system that could easily fit in the following expression :

$$E_{tot} = E_{vac} + \Delta G_{solv} \tag{2.17}$$

where E_{vac} is the potential energy of the molecule in vacuum (easily computes by derivation of atomic coordinates) and ΔG_{solv} is the free energy change requires to transferring the molecule from vacuum into solvent, i.e. solvation free energy. The core action in Eq. 2.17 is then the modeling of solvation effects included in its second term. Most implicit models account for this by splitting the free energy of solvation into apolar and electrostatic solvation terms, see Eq. 2.18. The former is the solvation free energy of uncharged molecules i.e. in the context where all partial charges are canceled down and the latter is the work spent to charge the system reversibly to its final equilibrium state in the ionic atmosphere.

$$\Delta G_{solv} = \Delta G_{apolar} + \Delta G_{el} \tag{2.18}$$

ΔG_{apolar} is mostly assumed to be proportional to the solvent accessible surface area and is often neglected. Meanwhile, most of the effort is devoted to the computation of the time-consuming electrostatic contribution ΔG_{el}. A number of approaches accounted for electrostatic solvation effects (Roux and Simonson, 1999; Simonson, 2001) including the GB model which derived from the Poisson-Boltzmann (PB) theory (Simonson, 2003; Baker, 2005). For a homogeneous system with a local dielectric constant ϵ and charge density ρ, the electrostatic potential ϕ

is accurately computed by the Poisson equation, the latter referring to the Laplace equation, as follows :

$$\boldsymbol{\nabla} \cdot [\epsilon(\boldsymbol{r}) \boldsymbol{\nabla} \phi(\boldsymbol{r})] = \rho(\boldsymbol{r}) \tag{2.19}$$

The challenge in the above equation resides in the appropriate description of ionic charges and screening salt effects, at variance of solute charges suitably embedded in the molecular model used. Nonetheless, by making reasonable assumptions such as an approximate mean-field approach, the Poisson-Boltzmann overcomes the latter issue, leading to a non-linear partial differential equation. However, the linearized form is the mostly used one in biomolecular context as its solution approximates the PB solution even for non small values of the potential (Fogolari et al., 1999; Fogolari, Brigo, and Molinari, 2002; Onufriev, 2008), see Eq. 2.20 :

$$\boldsymbol{\nabla} \cdot [\epsilon(\boldsymbol{r}) \boldsymbol{\nabla} \phi(\boldsymbol{r})] = -\rho(\boldsymbol{r}) + \sum_i C_i^b z_i^2 \frac{q^2 \phi(\boldsymbol{r})}{k_{\mathrm{B}} T} \tag{2.20}$$

In Eq. 2.20 ϕ is the electric potential, ρ is the charge density, z_i and C_i^b are the valence and bulk concentration of ion i, k_{B} is the Boltzmann's constant, T is the temperature, q is the unit charge, ϵ is the local dielectric constant and ϵ_0 the vacuum permittivity. Eq. 2.20 is a differential equation, and as such, is mainly solved numerically. One solution is therefore given by the GB model (Simonson, 2003), one of the widely used formalism for the estimation of the electrostatic contribution to the free energy of solvation ΔG_{el}. GB models approximate the electrostatic component of ΔG_{solv} as a pairwise summation of interacting terms between atomic charges i and j as in Eq. 2.21 (Kleinjung and Fraternali, 2014) :

$$\Delta G_{el} \approx -\frac{1}{2}\left(\frac{1}{\epsilon_{in}} - \frac{1}{\epsilon_{out}}\right) \frac{q_i q_j}{\sqrt{r_{ij}^2 + \alpha_i \alpha_j e^{-r_{ij}^2/4\alpha_i\alpha_j}}} \tag{2.21}$$

where r_{ij} is the distance between atoms i and j, q_i and q_j their partial charges, ϵ is the relative dielectric constant, with subscripts *out* and *in* refering to solvent and solute, respectively. α_i and α_j are effective Born radii of the interacting sites i and j. These latter variables play a pivotal role in the implementation of GB theory and are often computed by approximating the electrostatic energy density due to the atom of interest by employing empirical partial differential formulae and subsequent integrations over an adequate volume (Hawkins, Cramer, and Truhlar, 1995; Hawkins, Cramer, and Truhlar, 1996; Onufriev, Bashford, and Case, 2000; Onufriev, Bashford, and Case, 2004) or less often, surface (Ghosh, Rapp, and Friesner, 1998; Fogolari, Corazza, and Esposito, 2013). An attempt to refit the Born radii and the Onufriev, Bashford and Case (OBC) parameters of amber99sb-ildn (Lindorff-Larsen et al., 2010) potential function to best reproduce the average solvation forces with respect to the explicit solvent model was initiated in this work, see Fitting parameters for GBSA.

2.1.3 Temperature and pressure control in MD computations

Most experiments are conveniently undertaken at constant pressure and temperature. Contrariwise, propagating the system through the phase space using classical mechanics (Newton's law of motion) as it is the case in MD theory directly leads to constant energy average observables, i.e. in the microcanonical ensemble (*NVE*). This raises the practical difficulty to link *NVE* averages of MD computations to easily accessible mascroscopic constant T and P experiments. Therefore, mimicking experimental conditions, many MD simulations, including those of this thesis, are rather performed under constant T and/or P, employing either a canonical ensemble (*NVT*) or a grand canonical (μVT) or canonical pressure (*NPT*) ensemble (further insights on these ensembles could be obtained in relevant textbooks as (McQuarrie, 1976; Frenkel and Smit, 2002)). Several algorithms namely thermostats and barostats have thus been developed to constraint thermodynamics variables along the simulations.

Thermostats

The simple way to preserve the system's temperature from large thermal variations is to couple it to an external bath. Even though MD computes time depedent average microcanonical quantities (constant energy), the system energy fluctuates[4] throughout the simulation path with discrete values, leading to temperature fluctuations. Thus, the probability of locating the system at a given energy microstate follows a Maxwell-Boltzmann distribution function. In particular, from the classical mechanics point of view, it is well described by the Maxwell-Boltzmann velocity distribution, i.e. the chance of a particle to have a momemtum $\boldsymbol{p}$, see Eq. 2.22 :

$$\mathcal{P}(\boldsymbol{p}) = \Big(\frac{\beta}{2\pi m}\Big)^{3/2} exp\Big(-\frac{\beta \boldsymbol{p}^2}{2m}\Big) \tag{2.22}$$

A set of *had doc* procedures are often used for temperature control. The most basic of which include velocity rescaling (Woodcock, 1971). It is naïvely achieved by multiplying/modifying the atomic velocities $\boldsymbol{v}_i$ in the current instantaneous temperature *T(t)* (Eq. 2.23) by a rescaling factor $\tau = \sqrt{\frac{T_{ref}}{T}}$, where T_{ref} is the reference temperature. Albeit the temperature is efficiently regulated in such an approach, the overall velocities distribution is not a Maxwell-Boltzmann like type.

$$T(t) = \frac{\sum_{i=1}^{N} m_i v_i^2(t)}{3k_{\mathrm{B}}N} \approx T_{ref} \tag{2.23}$$

with N being the total number of atoms in the system.

[4] This is due the spontaneous interconversion of kinetic and potential terms of the total energy

Following previous considerations, Berendsen proposed another way to rescale the velocities (Berendsen et al., 1984), which in turn is a weak temperature coupling algorithm. Atomic velocities are rescaled by the factor τ defined in Eq. 2.24. The Berendsen thermostat cancels out the kinetic energy fluctuations, in clear it does not sample a proper canonical ensemble. The simulations in this thesis have mostly employed the velocity rescaling of Bussi *et al.* (Bussi, Donadio, and Parrinello, 2007), a modification of Berendsen thermostat.

$$\tau = \sqrt{1 + \frac{\Delta t}{\tau_T}\left(\frac{T_{ref}}{T} - 1\right)} \tag{2.24}$$

In Eq. 2.24 T and T_{ref} are instantaneous and target temperatures respectively, Δt is the simulation time-step and τ_T is a coupling constant.

Away from more intuitive velocities rescaling based paradigms, stochastic like thermostats provide more intricate and perhaps more general approaches. In the Andersen thermostat formalism (Andersen, 1980), the system's temperature is kept unchanged by coupling it to a heat bath. More in details, particles are randomly selected at regular time-step to undergo collision and only for those selected, their velocities are resampled from a Maxwell-Boltzmann distribution, corresponding to the reference temperature. Moving a step ahead from stochastic to deterministic paradigms, Nosé-Hoover proposed a very efficient temperature coupling (Nosé, 1984; Hoover, 1985) based on an extended Lagrangian theory of classical mechanics in which two artificial degrees of freedom (a thermal reservoir and a friction term) are added to the atomic coordinate and velocity. Weak-coupling schemes adequately relax the system's temperature to the reference value of the thermostat but poorly sample the true canonical ensemble once at equilibrium. This artifact is corrected in the current Nosé-Hoover definition. This is also the case with the Langevin thermostat (Zwanzig, 1973) in which the canonical sampling is enabled by incorporating a friction term γ along with the previous highlighted stochastic (noise) term to the equation of movement.

Barostats

Simulations usually performed in explicit solvent imply the definition of boundaries conditions and the requirement for mechanical equilibrium, owing to volume fluctuations and thereby change in pressure. Therefore, as for thermal equilibrium originated from particles momenta, there is a need also to regulate the pressure. One of the most intuitive and week-coupling pressure schemes is the one of Berendsen (Berendsen et al., 1984) in which the reference pressure of the barostat P_{ref} is obtained by multiplying the atomic coordinates and box vectors every time-step Δt by the pressure scaling factor ζ as in Eq. 2.25 :

$$\zeta = \left\{1 - \frac{\Delta t}{\tau_P}\beta_c[P_{ref} - P]\right\}^{\frac{1}{3}} \tag{2.25}$$

P is the actual instantaneous pressure, β_c the isothermal compressibility of the system and τ_P is the relaxation pressure coupling, the larger it is the weaker the coupling is. Similarly to its thermostat settings, the Andersen barostat (Andersen, 1980) couples the system to an external fluid volume V with mass M whose expansion and compression affect the system in the same way a piston would do on a physical system. V is the system's volume and it is a dynamical variable defining the piston's coordinates while M refers to the strenght of the coupling. It is worth noting that whereas the Berendsen barostat successfully samples the expected average pressure, still it does not yield the exact *NPT* ensemble as it is seen for the Andersen's one.

Almost all the simulations (those in explicit solvent) in this work have employed a Parrinello-Rahman pressure coupling scheme (Parrinello and Rahman, 1981; Nosé and Klein, 1983), which from theoretical considerations should properly sample the canonical pressure *NPT* ensemble in a way rationalizing the definition of Nosé-Hoover for thermostat (Nosé, 1984; Hoover, 1985). In the Parrinello-Rahman barostat, the box vectors as depicted by the matrix $\boldsymbol{b}$ follow the matrix equation of motion as in Eq. 2.26 :

$$\frac{d^2\boldsymbol{b}}{dt^2} = V\boldsymbol{W}^{-1}\boldsymbol{b}'^{-1}(\boldsymbol{P} - \boldsymbol{P}_{ref}) \tag{2.26}$$

where V is the volume of the box, $\boldsymbol{W}$ a matrix parameter defining the strength of the coupling (mass parameter), $\boldsymbol{P}$ and $\boldsymbol{P}_{ref}$ the matrices of the instantataneous and target pressures, respectively.

Particle motions in this scheme are subsequently modified to fit into the following matrix differential equation :

$$\frac{d^2\boldsymbol{r}_i}{dt^2} = \frac{\boldsymbol{F}_i}{m_i} - \mathcal{M}\frac{d\boldsymbol{v}_i}{dt}, \qquad \text{with} \quad \mathcal{M} = \boldsymbol{b}^{-1}\left[\boldsymbol{b}\frac{d\boldsymbol{b}'}{dt} + \frac{d\boldsymbol{b}}{dt}\boldsymbol{b}'\right]\boldsymbol{b}'^{-1} \tag{2.27}$$

In Eq. 2.26 the mass parameter matrix $\boldsymbol{W}$ regulates the pressure coupling by coordinating the contraints applied on box vectors. Thus, by choosing the isothermal compressibilities terms β_c^{ij} and the pressure time constant τ_P in Gromacs, the coupling strength, dictated by the reciprocal of $\boldsymbol{W}$ is evaluated by the following equation (Eq. 2.28) in which $\boldsymbol{L}$ is the largest box matrix element.

$$(\boldsymbol{W}^{-1})_{ij} = \frac{4\pi^2\beta_c^{ij}}{3\tau_P^2\boldsymbol{L}} \tag{2.28}$$

2.2 Enhanced sampling approaches

Biological molecules, with extensive degrees of freedom sample rugged free energy landscapes that are often extremely tricky to explore even with the high

temporal resolution of most atomistic MD simulations. In many situations, conformational phase spaces are populated by local minima separated by low probability regions (free energy barriers), making the inter-connection almost a rare thermodynamic event in conventional timescales. These shortcomings could be circumvented either by modifying the potential energy function by adding external bias to drop the residual time the system get trapped in a local energy minimum well, and thereby speeding up the transition from one local minimum to another, or by increasing the kinetic energy via tempering techniques. In one way or in the other, this requires advanced sampling paradigms for capturing large-scale and long-time configurational changes (Lei and Duan, 2007; Adcock and McCammon, 2006; Schlick, 2009; Abrams and Bussi, 2014).

Umbrella sampling

Umbrella sampling can referred as a class of *non-Boltzmann* or *non-equilibrium* sampling methods (Kästner, 2011), in reference to collective variable (CV) biasing techniques such as Adaptive-Biasing Force (ABF), Temperature-Accelerated Molecular Dynamics (TAMD), etc. In the theory pioneered by Torrie and Valleau (Torrie and Valleau, 1974; Torrie and Valleau, 1977), one applies a biased potential on selected reaction coordinates to ride out the sampling of unfavorable regions then driving the system from one thermodynamic state to another. Along the reaction coordinates path, intermediate steps are covered by a series of independent simulations (windows) such as to ensure sufficient overlap among them. A biased potential acts thus as an *umbrella* that connects energetically separated regions in the phase space. In a more empirical formulae expression, one modifies the unbiased potential $U(\boldsymbol{r}^N)$ adding a weigthing term $\mathcal{W}(\xi)$ depending only on the reaction coordinates ξ such that :

$$U'(\boldsymbol{r}^N) = U(\boldsymbol{r}^N) + \mathcal{W}(\xi) \tag{2.29}$$

In Eq. 2.29 for the sake of simplicity $\mathcal{W}(\xi)$ is often expressed by a simple quadratic form as in Eq. 2.30, where K_ξ is the harmonic bias strength constant and ξ_0 the reference reaction coordinate state.

$$\mathcal{W}(\xi) = \frac{1}{2} K_\xi \Big(\xi - \xi_0\Big)^2 \tag{2.30}$$

Once a reaction coordinate is selected[5] one should in principle be able to estimate the probability density function $\mathcal{P}(\xi)d\xi$ of an unbiased canonical system along the reaction coordinate ξ by integrating over the system's degrees of freedom, with $\mathcal{P}(\xi)$ defines as in Eq. 2.31 :

$$\mathcal{P}(\xi) = \frac{\int \delta\left[\xi(\boldsymbol{r}^N) - \xi\right] exp\left[-\beta U(\boldsymbol{r}^N)\right] d\boldsymbol{r}^N}{\int exp\left[-\beta U(\boldsymbol{r}^N)\right] d\boldsymbol{r}^N} \tag{2.31}$$

[5]Including any order-paramater such as accessible solvent areas, but preferentially geometric factors like bending angles or bond stretchings are used

Note that Eq. 2.31 is free from momenta contributions (kinetic energy) since both potential energy $U(\boldsymbol{r}^N)$ and momenta $K(\boldsymbol{p}^N)$ terms of the total system's hamiltonian $\mathcal{H}(\boldsymbol{p}^N, \boldsymbol{r}^N) = K(\boldsymbol{p}^N) + U(\boldsymbol{r}^N)$ could be integrated separatelty and integration over the momenta space leads to a constant $\left(\frac{2\pi m}{\beta}\right)^{3N/2}$, thus not of interest.

Bearing in mind that the free-energy of a state underlies its equilibrium population, i.e. the normalized probability distribution of that state over a specific variable, one could deduce the free energy as in Eq. 2.32 :

$$F(\xi) = -k_{\mathrm{B}} T ln \mathcal{P}(\xi) \tag{2.32}$$

where $F(\xi)$ is the Helmholtz energy and represents the potential of mean force (PMF) (Roux, 1995).

However, by applying the modified potential to the simulated system, the resulting biased distribution $\mathcal{P}'(\xi)$ is not obeying to a Boltzmann's unbiased one. Assuming an ergodic system, the corresponding Boltzmann average distributions can be recovered from the non-equilibrium one following Eq. 2.33 :

$$\mathcal{P}(\xi) = \mathcal{P}'(\xi) exp[\beta \mathcal{W}(\xi)] \langle exp[-\beta \mathcal{W}(\xi)] \rangle \tag{2.33}$$

From Eq. 2.33 the free energy is easily derived as :

$$F(\xi) = F'(\xi) - \mathcal{W}(\xi) + \mathcal{C}, \quad \text{with} \quad \mathcal{C} = -k_{\mathrm{B}} T ln \langle exp[-\beta \mathcal{W}(\xi)] \rangle \tag{2.34}$$

In the above equation (Eq. 2.34) $F'(\xi)$ is the biased PMF computed from the simulation, $\mathcal{W}(\xi)$ is given analytically and the last term $\mathcal{C}$ is a constant independent of ξ. However, in general, there are methods employed to combine the results of different windows in umbrella sampling, given that sufficient overlap is achieved between each of them. Moreover, these methods provide a way to estimate the constant $\mathcal{C}$ (Ferrenberg and Swendsen, 1989) and the most relevant one is the Weighted Histogram Analysis Method (WHAM) (Kumar et al., 1992; Souaille and Roux, 2001).

Temperature Replica Exchange

In Replica Exchange Molecular Dynamics (REMD) (Sugita and Okamoto, 1999) (see scheme in Fig. 2.1), one submits a series of simultaneous and non-interacting (independent) simulations of the same system called replicas, started at different temperatures. At regular intervals, configurations may be swapped when exchange in replicas is attempted. The acceptance probability, $\mathcal{P}_{ij}$, that essentially ensures a canonical sampling, obeys a Monte Carlo-like Metropolis criterion, Eq. 2.35, allowing the final equilibrium distribution for each temperature to follow a Maxwell-Boltzmann distribution.

$$\mathcal{P}_{ij} = min\Big[1, exp(\beta_j - \beta_i)(E_j - E_i)\Big] \tag{2.35}$$

where E_i and E_j are the replica's i and j internal energies before and after the exchange with temperature T_i and T_j respectively. $\beta_i = 1/k_{\mathrm{B}}T_i$ and $\beta_j = 1/k_{\mathrm{B}}T_j$ with k_{B} being the Boltzmann's constant.

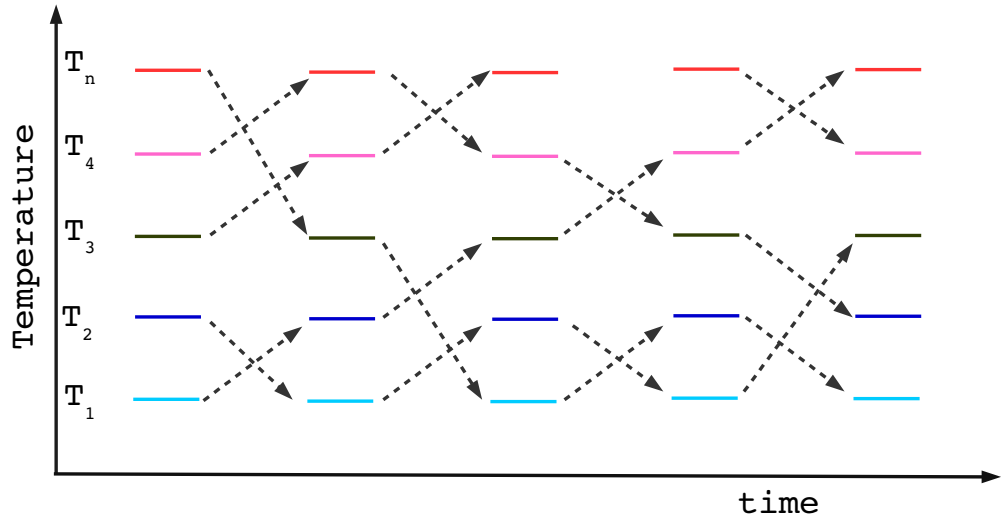


FIGURE 2.1: Basic schematic illustration of Replica Exchange Molecular Dynamics simulation technique. Horizontal lines depict individual replicas.

REMD is likely a tandem MD-random walk in temperature space. In clear, the kinetic trapping at lower temperatures (unbiased sampling) is gradually enhanced by exchanging conformations with higher temperature replicas able to cross the kinetic barriers faster.

Metadynamics

Like Umbrella sampling, metadynamics (Laio and Parrinello, 2002; Laio and Gervasio, 2008; Barducci, Bonomi, and Parrinello, 2011) is also an advanced MD technique modifying the potential and sampling a non-Boltzmann distribution. It provides a way to accelerate the dynamics and recover the free energy surface (FES) based on selected degrees of freedom or collective variables (CVs), thereby disfavoring the re-exploration of previously visited states in the defined CVs space. The simulation is then driven by an external time-based history-dependent bias potential $V_G(S(\boldsymbol{r}), t)$ constructed by summing up the Gaussians deposited every timestep τ_G along the trajectory. At a given time t, $V_G(S(\boldsymbol{r}), t)$ is expressed as follows :

$$V_G(S(\boldsymbol{r}), t) = \omega_G \sum_{\substack{t'<t \\ t'=\tau_G, 2\tau_G, \ldots}} exp\left(- \sum_{i=1}^{d} \frac{(S_i(\boldsymbol{r}) - s_i(t'))^2}{2\delta s_i^2} \right) \tag{2.36}$$

In Eq. 2.36, $s_i(t) = S_i(\boldsymbol{r}(t))$ is the instanteneous value for the i^{th} CV considered. τ_G is the stride of the Gaussians deposition with respective height or energy rate w_G and width δs_i. These latter are indeed of importance in determining the accuracy and efficiency of the free energy reconstruction.

It has been shown both empirically (Laio et al., 2005) and analytically (Bussi, Laio, and Parrinello, 2006) that in the limit of infinite time, and considering a system appropriately described by a Langevin mechanics, the bias potential $V_G(S(\boldsymbol{r}, t))$ converges to the minus of the free energy *F(S)*.

$$\lim_{t\to\infty} V_G(S(\boldsymbol{r}, t) = -F(S(\boldsymbol{r})) + \mathcal{C} \tag{2.37}$$

Most often, the great challenge in metadynamics simulations resides in the appropriate selection of CVs in order to minimize the errors in the free energy reconstruction and adequately sampling slow motions in the system. Nothwithstanding the power and strength of the approach, significant equilibration time is often required to properly reconstruct the FES of a high dimensional many-body system (like biomolecules or so), taking into account that the perfomance of the algorithm decreases when increasing the number of CVs. This issue is likely circumvented by "hybrid" techniques that combine either metadynamics with parallel tempering (Bussi et al., 2006) or with replica exchange such as bias exchange metadynamics (Piana and Laio, 2007).

2.3 Free energy from end-point MD simulation

The outcome of (bio)chemical processes and in general the thermodynamics of biological systems is dependent upon the sign of the change in Gibbs free energy. Its empirical formulation entails both enthalpic and entropic contributions, the latter being linearly scaled by the system temperature. For a well equilibrated system, the enthalpic counterpart can rationally be estimated as an ensemble average over a set of conformations from the MD trajectory. The bottleneck in the weighting of Gibbs free energy is the proper modeling of the entropic part, and in particular, of the solvation entropic effects. This state of facts underlines the interest in methods development to adequately account for solvent entropy and to a large extent for conformational entropy (Chapter 3). Here only a very short overview in relation to Gibbs free energy in the context of implicit solvent MD simulation is discussed, readers could however found further details in literature such as as (Gilson et al., 1997; Roux and Simonson, 1999; Wereszczynski and McCammon, 2012; Fogolari, Corazza, and Esposito, 2018).

The free energy of a state underlies its equilibrium population and can in turn be referred to its ensemble average as :

$$\Delta G = k_{\text{B}} T log\langle exp(\beta U)\rangle + C \tag{2.38}$$

where U is the potential energy of the system, C represents other constant terms that cancel upon comparison of different states, and the thermodynamic average, as denoted by $\langle\rangle$, is done over the ensemble. However, Eq. 2.38 is only readily applied when considering all the available configurational phase space. This constitutes therefore a severe limitation to most MD simulations since usually only states separated by few kcal from the equilibrated structure are visited. Going far beyond the equilibrium state would require either using denaturating conditions or biased sampling or even running very long simulation time.

For a solvated system, the standard molar free energy change of a solute can be written as a function of configurational integrals as follows (Gilson et al., 1997; Roux and Simonson, 1999; Wereszczynski and McCammon, 2012; Fogolari et al., 2015; Fogolari, Corazza, and Esposito, 2018) :

$$\Delta G_A^0 = -k_B T log\left(\left(\frac{1}{\Lambda_A}\right)^{3N} \int exp[-\beta(U(\boldsymbol{r}_A) + \Delta\mathcal{W}(\boldsymbol{r}_A, T))]d\boldsymbol{r}_A\right) + \frac{P^0 \Delta\bar{V}_A}{\mathcal{N}} \tag{2.39}$$

In Eq. 2.39 P^0 is the standard pressure, $\mathcal{N}$ is the Avogadro's number, N is the solute total number of atoms, $\bar{V}_A$ and $\boldsymbol{r}_A$ are the partial molar volume and coordinates for the solute, respectively. $\Lambda_A = \frac{h}{\sqrt{2\pi m_A k_B T}}$ represents the thermal Debroglie wavelength in which m_A is the solute mass and h is the Planck's constant. $P^0\Delta\bar{V}_A$ in the second term of Eq. 2.39 changes very little for biomolecular sytems and therefore is negligible upon conformational changes. The solute solvation potential of mean force $\mathcal{W}(\boldsymbol{r}_A, T)$ which accurately approximates the solvation free energy in implicit solvent models can be obtained by integrating out the strongly correlated solvent degrees of freedom as follows :

$$exp[-\beta\Delta\mathcal{W}(\boldsymbol{r}_A, T)] = \frac{\int exp[-\beta U_{AS}(\boldsymbol{r}_A, \boldsymbol{r}_S) + U_S(\boldsymbol{r}_S)]d\boldsymbol{r}_S}{\int exp[-\beta U_S(\boldsymbol{r}_S)]d\boldsymbol{r}_S} \tag{2.40}$$

in which $\boldsymbol{r}_S$ stands for the solvent coordinates, U_{AS} is the correlated solute-solvent energy and U_S is the solvent potential. By substiting Eq. 2.40 in Eq. 2.39, the standard molar free energy change of a solute is given by :

$$\Delta G_A^0 = -k_B T log\left(\int exp[-\beta(U(\boldsymbol{r}_A) + \Delta\mathcal{W}(\boldsymbol{r}_A, T))]d\boldsymbol{r}_A\right) \tag{2.41}$$

Eq. 2.41 explicitly highlights the temperature dependence of the potential of mean force $\mathcal{W}(\boldsymbol{r}_A, T)$. Furthermore, it should be noted that all the implicit models discussed so far (see Implicit solvent) implement the routines to compute $\mathcal{W}(\boldsymbol{r}_A, T)$.

The analysis of enthalpic and entropic contributions shows that the entropic contributions of the solvent are included in the ensemble average of $\Delta\mathcal{W}$, and that

the free energy can thus be expressed as:

$$\Delta G_A^0 = \langle U(\boldsymbol{r}_A) + \Delta \mathcal{W} \rangle - T\Delta S^{conf} \tag{2.42}$$

where ΔS^{conf} is the configurational entropy of the solute.
The above formula constitutes the basis of free-energy estimation from end-point simulations.

2.4 Fitting parameters for GBSA

This section briefly stresses on parameters tuning in GBSA continuum solvation model to better reproduce the average solvation forces as computed by the explicit solvent model counterpart. To do so, a set of simulations were runned on a dataset of 55 proteins (Tjong and Zhou, 2007a; Tjong and Zhou, 2007b) selected before to represent the variety of charges and shapes of proteins. Three models of Generalized Born (GB) are implemented in the package used (Gromacs) (Hess et al., 2008) namely Still (Still et al., 1990; Qiu et al., 1997), Hawkins, Cramer and Truhlar (HCT) (Hawkins, Cramer, and Truhlar, 1995; Hawkins, Cramer, and Truhlar, 1996) and Onufriev, Bashford, Case (OBC) (Onufriev, Bashford, and Case, 2004). In all models the calculation of GB radii is done by summation of pairwise contributions, which depend on atomic radii and the distance between atoms (Qiu et al., 1997). In this work we used the GB model to refit the Born radii and OBC parameters of amber99sb-ildn (Lindorff-Larsen et al., 2010) force fields to best reproduce the average solvation forces in comparison to the explicit solvent model.

In all the simulations, the protein was frozen by removing all the translational and rotational degrees of freedom, enabling only water to equilibrate (in the case of explicit solvent). The Generalized Born-Suface Area (GB/SA) continuum solvation model of Onufriev-Bashford-Case (OBC) (Onufriev, Bashford, and Case, 2004) and TIP3P water model (Jorgensen et al., 1983) respectively for implicit and explicit solvent simulations were applied. Simulations were carried out using Gromacs molecular package (Hess et al., 2008) with amber99sb-ildn (Lindorff-Larsen et al., 2010) force fields and each lasted 1 ns with coordinates and forces saved every 1 ps. The time-step used in the MD steps is 2 fs and velocity rescaling (modified Berendsen thermostat) (Bussi, Donadio, and Parrinello, 2007) with the coupling constant of 0.1 ps was used to maintain the temperature to the reference value of 300 K. The electrostatics interactions were treated with the smooth particle mesh Ewald summation (Essmann et al., 1995; Darden, York, and Pedersen, 1993) (in the case of explicit solvent). Short-range electrostatics and van der Waals interactions were truncated with a real space cut-off of 10 Å. The simulation started in explicit solvent with the protein kept fixed, the water is later replaced by the implicit solvation model of OBC, which is further replaced by the vacuum. This allows an estimation of the change in average solvation forces upon moving from explicit to implicit and from implicit to vacuum.

The van der Waals radii sets for Gromacs are available in the corresponding force field directory within the file `gbsa.itp`, see Table 2.1 and include all the atom types needed (C, N, O, S, H). Thus no extra atom type was added. We refit the parameters α, β and γ of OBC model and GB radii, which are defined as input, to obtain a better agreement between the average solvation forces computed using explicit and implicit solvent. The latter components are easily combined into this empirical formula for GB radius Eq. 2.43 (Fogolari, Corazza, and Esposito, 2015):

$$(r_i^{GB})^{-1} = \tilde{\rho}_i^{-1} - \rho_i^{-1} tanh(\alpha\Psi - \beta\Psi^2 + \gamma\Psi^3) \tag{2.43}$$

where ρ_i is the atom's van der Walls radius, $\tilde{\rho}_i^{-1} = (\rho_i - 0.09\text{Å})^{-1}$ and Ψ is the pairwise summation function over the atoms contributing to the GB radius. α, β and γ are the OBC components fitted to best reproduce the average solvation forces. We randomly alter the GB radii and OBC components obtained by least-square solution of this linear equation Eq. 2.44 :

$$\alpha\Psi - \beta\Psi^2 + \gamma\Psi^3 = tanh^{-1}\left(\frac{\tilde{\rho}_i^{-1} - (r_i^{GB})^{-1}}{\rho_i^{-1}}\right) \tag{2.44}$$

where r_i^{GB} are ***perfect radii*** approximated using the GBR6 surface integral formalism (Grycuk, 2003).

The average solvation force was obtained as the difference between the total force of both implicit and explicit solvent simulations with the total force obtained in vacuum, see Eq. 2.45 and Eq. 2.46. Table 2.1 summarizes the fitted parameters obtained compared to the default ones. The correlation coefficients between average implicit and explicit solvation forces and fitted linear equation were computed and outcomes are plotted in Fig. 2.2.

$$F_{x,y,z}^{solv}expl = F_{x,y,z}^{tot}expl - F_{x,y,z}^{tot}vac \tag{2.45}$$

$$F_{x,y,z}^{solv}impl = F_{x,y,z}^{tot}impl - F_{x,y,z}^{tot}vac \tag{2.46}$$

where $F_{x,y,z}^{solv}expl$, $F_{x,y,z}^{solv}impl$, $F_{x,y,z}^{tot}expl$, $F_{x,y,z}^{tot}impl$ and $F_{x,y,z}^{tot}vac$ stand for average solvation forces in explicit and implicit solvent model, average total forces in explicit and implicit solvent model and average total forces in vacuum.

TABLE 2.1: Original and fitted OBC parameters and GB radii.

Element	amber99sb-ildn GB radii		α, β, γ
	default	fitted	
H	0.105; 0.115; 0.125	0.113; 0.123; 0.133	$\alpha(1)$:1 ; $\beta(1)$: 0.8
C	0.1875; 0.190	0.1865; 0.189	$\gamma(1)$: 4.85
N	0.1625; 0.17063	0.1605; 0.16863	$\alpha(2)$: 0.52; $\beta(2)$: 0.28
O	0.148; 0.1535	0.152; 0.1575	$\gamma(2)$: 0.44
S	0.1775	0.1775	

(1) : default (2) : optimized

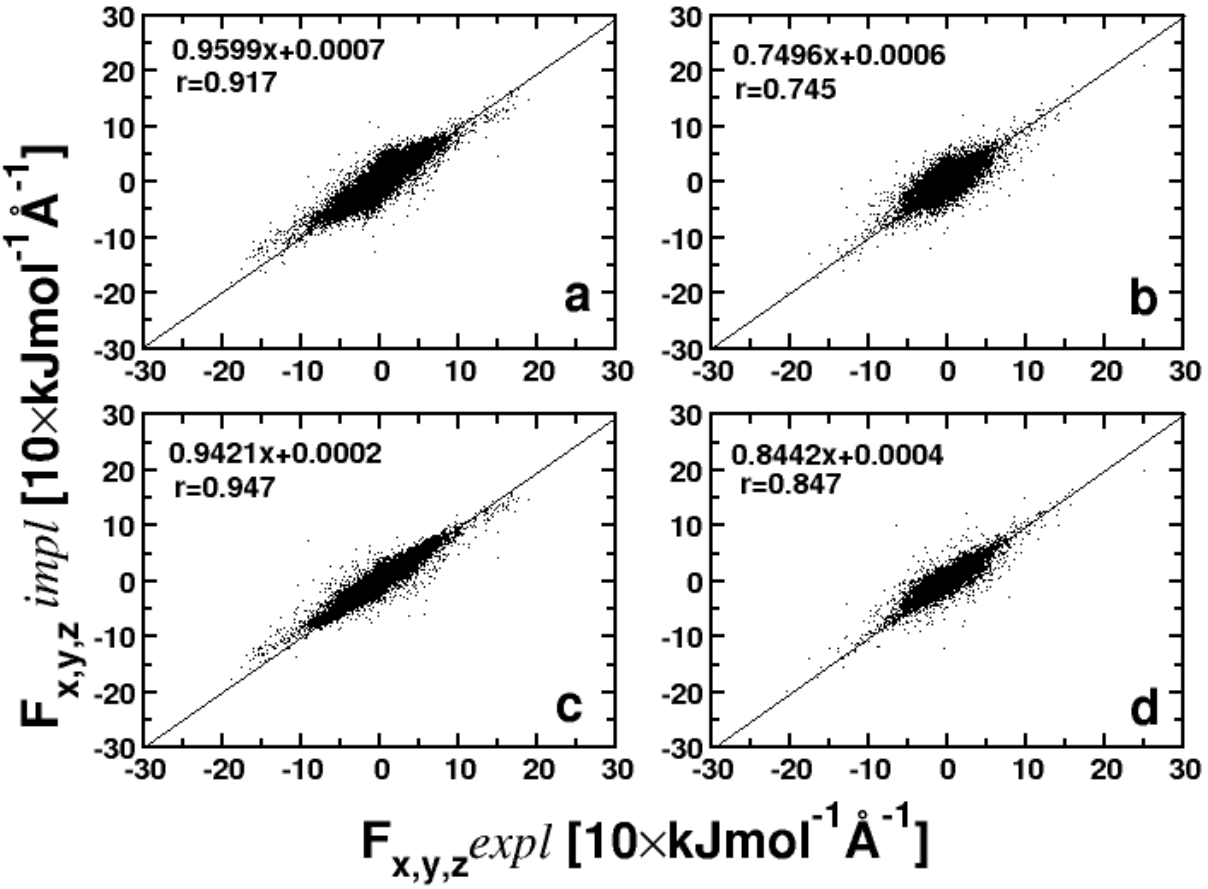


FIGURE 2.2: Total and Average solvation forces in implicit solvent compared to the ones obtained in explicit solvent. **[Left pannel] :** Total forces (a & c); **[Right pannel]** : Average solvation forces (b & d) ; **[Top pannel] :** Default parameters ; **[Bottom pannel] :** Fitted parameters.

Even though the above optimized parameters could not yet been readily applied to a real case folding simulation because they have not been thoroughly tested and calibrated, they result to an improvement of > 10 % in reproducing the average solvation forces in comparison to the default ones. It should be noted however that, at the first instance, using CHARMM force field to tune the same parameters to reproduce the PBSA average forces (Fogolari, Corazza, and Esposito, 2015), it was noticed an unfolding of the protein in the tens of nanoseconds and an apparent tendency to form helices. This underlines that CHARMM force field is not sufficiently accurate for such a job, at least as far as implicit solvent model is concerned.

2.5 Summary

MD simulation is increasingly given importance in the field of molecular biology and is nowadays able to sample timescales relevant for many biological functions including enzyme catalysis, conformational changes, protein folding, and so on. The high spatio-temporal resolution of most atomistic MD techniques has provided clues in the characterization and identification of crucial mechanisms underlying vital biological protein functions, not possible and/or limited otherwise by experimental methods. Therefore, by acting as a bridge between microscopic and macroscopic scales, MD could legitimately be considered as a microscope for biology, of course when applied to biological systems, since its range of application is extremely diverse. This chapter has recalled the essential concepts, notions and theories behind computational methods in general and MD simulation in particular. Moreover, it has covered the main simulation strategies employed in classical all-atoms MD with implicit or explicit solvent representations. Simulation routines developed to reduce the computational effort in sampling both long-time and large-scale conformational changes, as well as rugged free energy landscapes were discussed. In one way or in the other, all the approaches presented in this chapter were employed in the current thesis. Readers are now introduced with the technical concepts most often to be used thereafer. The chapter concluded by showing the GBSA parameters optimization initiated in this work. Even though we could not apply them any further in this work, the results show promising features that could probably be improved.

Chapter 3

Conformational entropy in biomolecular protein association[1] [2]

Extracting free energies from MD simulations is still a very challenging task, either because of the limited sampling or more importantly the highly correlated and numerous contributing terms. Conformational entropy is a component of the total entropy of a molecule and is accounted as one of these energy terms and represents an important component of the free energy change upon binding of a ligand to its target protein, and more in general upon complex formation of biomolecules. The conformational entropy of the solute remains a difficult problem to be tackled with correlations among degrees of freedom to be taken into account. Albeit advanced MD techniques could bias the potential thereby accelerating the sampling and recovering the free energy profile along some CVs, still it requires a considerable amount of computing time since equilibrium must be reached at each intermediate value of the CVs. This chapter introduces a computational methodology for the estimation of the rotational-translational entropy loss upon complex formation of biomolecules from end-point (free energy from MD trajectory). The approach is derived from the nearest-neighbor distance and overcomes the lack of correlation observed when taking rotation and translation entropy separately. Also, our theory enables an accurate approximation of full rototional-translational entropy, considering correctly the high dimensionality (6) of the problem which poses severe limitations to existing methods, like the histogram method (Edholm and Berendsen, 1984). The theory is now implemented in a C program freely distributed (Fogolari et al., 2017).

Keywords : MD, free energy, rotro-translational entropy.

[1] F. Fogolari; C. J. Dongmo Foumthuim; S. Fortuna; A. Soler; A. Corazza and G. Esposito. Accurate estimation of the entropy of rotation-translation probability distributions, *J. Chem. Theory Comput.*, 2016, 12 (1),1-8.

[2] F. Fogolari; O. Maloku; C. J. Dongmo Foumthuim; A. Corazza and G. Esposito. PDB2ENTROPY: entropy calculation from conformational ensembles, *Submitted*, 2018, xx, x-x

3.1 Introduction

The change in Gibbs free energy ΔG is often employed as a sensor in predicting the outcome of many biochemical processes. In its formulation, the enthalpic term ΔH is counterweighted by the entropic term ΔS, the latter being scaled by the system temperature. A proper estimation of ΔG would therefore imply an estimation of both terms. However, in most instances, the entropic term is neglected or simply assigned to the hydrophobic and solvation effects, thereby computing the enthalpic component as energy average, often by discarding the PV term. Notwithstanding the thermodynamics approximations imposed, legitimated by the lack of appropriate treatment methods, it is worth minding that, in the context of biomolecular proteins association, entropic contributions upon binding include also changes in conformational entropies of both molecules. In this context, an obvious contribution arises from the restriction in rotational and translational degrees of freedom. Estimates of both terms have been given based on heuristic considerations (Finkelstein and Janin, 1989) and in recent years based on the results of MD simulations (Gilson et al., 1997; Huggins, 2014; Huggins, 2015; Fogolari et al., 2015) defining the distributions of reference atoms in the ligand with respect to the target molecule.

Estimating the entropy of solute conformations with respect to a reference state constitutes a very intricate scenario (Gilson et al., 1997; Zhou and Gilson, 2009; Polyansky, Zubac, and Zagrovic, 2012) due to the large number of interacting bodies often strongly correlated. Moreover it would be desirable to understand what are the sources of entropy. In a heuristic approach applied for instance in docking studies the loss of roto-translational entropy upon binding is often neglected altogether because it is assumed to be similar for all docking poses. For sidechain entropy, the loss upon binding is often assumed to be proportional to the buried solvent accessible area which gives an idea of the restraints on sidechain motions. Recently an entirely different approach was proposed by Demchuk, Gilson and coworkers (Singh et al., 2003; Hnizdo et al., 2007; Hnizdo et al., 2008) who showed that entropy can be unbiasedly estimated based on nearest neighbours distances among different conformers. Although its significance, explicit computation of the rotational-translational entropy has been elusive and mostly weighted as a fixed contribution to the binding energy which is neglected when comparing the free energy of binding of different molecules to the same target (Ajay and Murcko, 1995).

The theory presented hereafter is derived from the nearest-neighbor approach and does not address the correlation of rotational and translational degrees of freedom with other degrees of freedom. Methods for dealing with such corrrelations could be found in literature (Killian, Kravitz, and Gilson, 2007; King and Tidor, 2009; King, Silver, and Tidor, 2012; Fenley et al., 2014). This chapter introduces the theoretical backgrounds of our theory and shows how it can be readily applied to MD samples of a homodimer of transthyretin and more in general to biomolecular complex formation.

The nearest neighbor approach

In the nearest-neighbor method an estimate of the local probability density $\mathcal{P}(\boldsymbol{x})$ is obtained by considering the distance to the k^{th} nearest neighbor from each sample, say $\boldsymbol{x}$, within an s-dimensional sphere of radius $\boldsymbol{R}$. If V is the volume of that sphere (i.e. the volume within the hypersphere $\sum_{i}^{s} \boldsymbol{x}_i^2 \leq \boldsymbol{R}^2$), $\mathcal{P}(\boldsymbol{x})$ is expressed from the number of neighbors k over total number of samples n as in Eq. 3.1 :

$$\hat{\mathcal{P}}(\boldsymbol{x})V = \frac{k}{n} \tag{3.1}$$

In Eq. 3.1 V depends on the metric used to define the space. For each sample, say $\boldsymbol{x}_i$, in an s-dimensional space, a sphere of radius $\boldsymbol{R}_{i,k}$ is equal to the distance of that sample to its k^{th} nearest neighbor. For euclidean distances the volume of the hypersphere with radius $\boldsymbol{R}_{i,k}$ is :

$$V_{i,k} = \frac{\pi^{\frac{s}{2}} \boldsymbol{R}_{i,k}^{s}}{\Gamma(\frac{s}{2}+1)} \tag{3.2}$$

where Γ is the Gamma function. If the metric is not euclidean the volume $V_{i,k}$ enclosing all points closer to sample i than $\boldsymbol{R}_{i,k}$ will be a different monotonic function of $\boldsymbol{R}_{i,k}$.

Since the probability density is known (from Eqs. 3.1 & 3.2), an estimate of the entropy of the distribution for a discrete number of states n (here the total number of sample or snapshots) is calculated according to Eq. 3.3:

$$\hat{S}_k^{(n)} = -k_{\text{B}} \sum_{i=1}^{n} \mathcal{P}(\boldsymbol{x}_i) log \mathcal{P}(\boldsymbol{x}_i) \approx -\frac{k_{\text{B}}}{n} \sum_{i=1}^{n} log \hat{\mathcal{P}}(\boldsymbol{x}_i) = \frac{k_{\text{B}}}{n} \sum_{i=1}^{n} log \left(\frac{n\pi^{\frac{s}{2}} \boldsymbol{R}_{i,k}^{s}}{k\Gamma(\frac{s}{2}+1)} \right) \tag{3.3}$$

Demchuk *et al.* (Singh et al., 2003) have introduced a correction term to the above heuristic treatment in order to eliminate the asymptotic bias, Eq. 3.4 :

$$\frac{\hat{S}_k^{(n)}}{k_{\text{B}}} = \sum_{i=1}^{n} log(nV_{i,k}) - L_{k-1} + \varsigma \tag{3.4}$$

where ς is the Euler-Mascheroni constant (0.5772...) and L_{k-1} is defined as : $L_0 = 0$, $L_n = \sum_{i=1}^{n} \frac{1}{i}$ to provide an unbiased estimator of the entropy of the probability distribution. Eqs. 3.3 & 3.4 are rather similar unless that *log(k)* is replaced by $L_{k-1} - \varsigma$. Moreover, $V_{i,k}$ has a unit of measure and therefore the numerical value of $\hat{S}_k^{(n)}$ depends on the latter unit. This rationalizes the fact that for the entropy $-k_{\text{B}} \sum \mathcal{P} log \mathcal{P}$, the probability density $\mathcal{P}$ has the units of the measure.

In the context of translational and rotational entropy, a formalism to compute both distances separately have been provided (Huynh, 2009; Huggins, 2015).

However, in practice, a know-how to adequately merge both of them including the correlation remains elusive and hard to tackle. In principle, after a metric has been chosen in rotation-translations' space, the rotational and translational distances for each sample to its k^{th} nearest neighbor, $\boldsymbol{R}_{i,k}$, can be computed followed by the computation of the volume of the hypersphere of radius $\boldsymbol{R}_{i,k}$.

Like previously stated, our theory follows exactly the approach of the nearest neighbor apart of globalising the metric compared to the heuristically introduced six-dimensional euclidean metric employed by Huggins to combine the rotational and translational distances (Huggins, 2015). We provide evidence that the method of Huggins is a perfect approximation to the correct treatment under the condition that the compound distances considered imply a rotation angle less than ca. 1 radian.

3.2 Theory formulation

3.2.1 The nearest neighbor method applied to rotation-translations

In the following, we will consider the volume of the sphere $V_{i,k}$ as dependent of $\boldsymbol{d}_{i,k}$, i.e. $V(\boldsymbol{d}_{i,k})$ with $\boldsymbol{d}_{i,k}$ being the distance of the i^{th} sample to its k^{th} nearest neighbor within the hypersurface of radius $\boldsymbol{d}_{i,k}$. Eq. 3.4 which gives the exact entropy formulation is then rewritten as in Eq. 3.5 to make evident the distance requirement :

$$\frac{\hat{S}_k^{(n)}}{k_{\text{B}}} = \sum_{i=1}^{n} log(nV(\boldsymbol{d}_{i,k})) - L_{k-1} + \varsigma \tag{3.5}$$

In order to efficiently apply Eq. 3.5, one needs to define a metric to describe the samples and to compute the volume of the hypersphere. However, this is not a routine task in rotations and rotation-translations spaces as it seems to be the case for euclidean spaces. Therefore, we will elaborate on the following fundamental points :

1. the description of rotation-translations;
2. the definition of a metric in rotation-translation space;
3. the computation of the volume of the hypersphere of radius $\boldsymbol{d}$ in that space, i.e. the function $V(\boldsymbol{d})$.

3.2.2 Description of rotation-translations

In the rotational-translational space, we will define our system by a translation vector ($\boldsymbol{t}$) of the origin with respect to a reference position and by a rotation matrix ($\boldsymbol{\mathcal{R}}$) underlining the rotation of the two vectors (and their vectors product) with respect to a reference coordinates system. Therefore, one needs to select a

fixed reference state of the system. The latter should in principle entail the most rigid part of the molecule. More precisely, we choose the reference coordinates in the system by selecting three atoms of the less flexible part of the molecule with origin centered at $\boldsymbol{x}_0$ and two additional orthogonal vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ belonging to the same part of the molecule (Gilson et al., 1997). It is worth noting that once the most rigid part of the system is identified, any set of three atoms could be used to define the reference state. However, one should select reasonable decoupled positions from other degrees of freedom in order to compute the entropy resulting from individual contributions.

3.2.3 Distance in rotation-translation space

The translation distance between two distinct states $(\boldsymbol{t}_a, \boldsymbol{\mathcal{R}}_a)$ and $(\boldsymbol{t}_b, \boldsymbol{\mathcal{R}}_b)$ without considering rotations is given $d_{\boldsymbol{t}} = \|\boldsymbol{t}_b - \boldsymbol{t}_a\|$, while in the absence of translation the rotation distance may be expressed as $d_{\boldsymbol{\mathcal{R}}} = \arccos\frac{\mathtt{Tr}(\boldsymbol{\mathcal{R}}_b^{-1}\boldsymbol{\mathcal{R}}_a)-1}{2}$, where $\mathtt{Tr}$ is the trace operator (Huynh, 2009; Huggins, 2015). The rotational distance defines the rotation angle of the matrix $\boldsymbol{\mathcal{R}}_b^{-1}\boldsymbol{\mathcal{R}}_a$ and is equivalent (up to a scaling factor) to 3 (Huynh, 2009; Huggins, 2015) and 6 (Huynh, 2009). When considering both translation and rotation, one has to merge $d_{\boldsymbol{t}}$ and $d_{\boldsymbol{\mathcal{R}}}$ components, taking into account the random scaling lenght l that multiplies the rotational distance, Eq. 3.6.

$$d^2 = d_{\boldsymbol{t}}^2 + l^2 d_{\boldsymbol{\mathcal{R}}}^2 = \|\boldsymbol{t}_b - \boldsymbol{t}_a\|^2 + l^2 \|\arccos\frac{\mathtt{Tr}(\boldsymbol{\mathcal{R}}_b^{-1}\boldsymbol{\mathcal{R}}_a) - 1}{2}\|^2 \tag{3.6}$$

In the rotation-translation space, relevant literature regarding the definition of metrics could be found in (Park and Ravani, 1997; Kuffner, 2004; Huynh, 2009). Furthermore, Huynh *et al*. (Huynh, 2009) have reviewed the computational facets of the most relevant metrics and Huggins (Huggins, 2015) has assessed their significance in nearest neighbors entropy estimation.

In the `compound` distance of Eq. 3.6, Huggins considers the scaling length l=1 Å. In theory this choice is arbitrary, however, in practice, it might lead to some meaningless issues, since the contributions of translational and rotational distances to the global one could be greatly unbalanced. For instance if $\boldsymbol{t}$ spans a volume of 1 Å^3 and l is chosen very small (say 0.01 Å, to consider an extreme case) the contribution of the rotational distance to global one will be effective only when the `compound` distance is comparable to 0.01 Å, implying a very fine sampling of the conformational space, which is unpractical. Therefore, the choice of l should reflect the relative weights of both components in Eq. 3.6.

3.2.4 Volume of the hypersphere in rotation-translations' space

Let's start by focusing on rotational distances without translations. In the case of random rotations, we consider the system's description in which the polar and

azimuthal angles ϕ and ψ of the rotation axis and the rotation angle θ are defined. In such a space, the probability function for a uniform sampling of states corresponds to the axis of rotation equally splitted over the solid angle of 4π (probability density $\mathcal{P}(\phi,\psi) = \frac{sin(\phi)}{4\pi}$) and the rotation angle θ spanning the range [0, π], with probability density expressed as in Eq. 3.7 (Miles, 1965) :

$$\mathcal{P}(\theta) = \frac{2}{\pi} sin^2\left(\frac{\theta}{2}\right) = \frac{1}{\pi}(1 - cos(\theta)) \tag{3.7}$$

Intiutive probability densities as for rotation angle uniformly covering the interval [0, π], i.e. $\mathcal{P}(\theta) = \frac{1}{\pi}$ do not lead to uniform distribution of rotations, because the reference system is rotating. Therefore, in the rotational space defined by (ϕ, ψ, θ), up to a scaling term, the only measure invariant under rotations is $\int\int\int sin(\phi)(1 - cos(\theta))d\phi d\psi d\theta$ (Miles, 1965). In our case, we used a metric that sets the distance d between two rotations described by 3×3 matrices $\boldsymbol{\mathcal{R}}_a$ and $\boldsymbol{\mathcal{R}}_b$ as in Eq. 3.8:

$$d = \arccos\frac{\mathtt{Tr}(\boldsymbol{\mathcal{R}}_b^{-1}\boldsymbol{\mathcal{R}}_a) - 1}{2} = \overline{\theta} \tag{3.8}$$

In Eq. 3.8 $\overline{\theta}$ is the rotation angle about the rotation axis for the composite rotation $\boldsymbol{\mathcal{R}}_b^{-1}\boldsymbol{\mathcal{R}}_a$. After the metric is setted, one should now compute the volume of the hypersphere of radius $\overline{\theta}$ in rotation space. In the latter space, such volume is expressed by the triple integral over $d\theta, d\phi$ and $d\psi$ which gives the measure in rotation-translation space over the solid angle for variables ϕ, ψ and from 0 to $\overline{\theta}$ for variable θ, see Eq. 3.9 :

$$V(d(\theta)) = \int_0^{\pi} d\phi \int_0^{2\pi} d\psi \int_0^{\overline{\theta}} sin(\phi)(1 - cos(\theta))d\theta = 4\pi(\overline{\theta} - sin\overline{\theta}) \tag{3.9}$$

It should be emphasized that in Eq. 3.9 only rotations scaling up to π values were integrated instead of 2π, to cancel down symmetry. Indeed, a rotation operation by an angle θ about the vector $\boldsymbol{\theta}$ is nothing but the rotation by an angle -θ about the vector -$\boldsymbol{\theta}$. Eq. 3.9 describes the volume in the rotation space without considering the translation. We will now estimate the volume including both terms.

The volume $V(d)$ of the hypersphere of radius d including rotations and translations, less than or equal to d from the chosen reference ($\boldsymbol{t} = \mathbf{0}$, θ = 0.0), is computed exactly as before (i.e. by direct integration). But, now, one should use the `coumpound` distance (Eq. 3.6). However, since the `coumpound` distance comprises both translation and rotation, the correlation makes it difficult to disentangle the integration limits, at variance of euclidean spaces.

The volume $V(d)$ is obtained by first computing the integral over the $d\phi$ and $d\psi$ variables (i.e. over the solid angle). This leads to a factor 4π. Subsequently, the translation integral over cartesian coordinates is converted in polar coordinates and the angular component is readily solved since it is not involved in the

coumpound distance, leading again to a factor 4π. This allows the global volume to be expressed as in Eq. 3.10 :

$$V(d) = (4\pi)^2 \int \int_{t^2+l^2\theta^2<d^2} t^2(1-cos(\theta))dtd\theta = (4\pi)^2 \int_0^d t^2 dt \int_0^{max(\pi, \frac{\sqrt{d^2-t^2}}{l})} (1-cos(\theta))d\theta \tag{3.10}$$

A solution to Eq. 3.10 is provided numerically employing the trapezoidal rule. We splitted the integration interval in equal bins and checked the convergence by halving the bins.

The analytical solution to Eq. 3.10 is obtained by making the assumption that, if the upper integration limit of (1 - cos(θ')) is restricted to a range where, say $\theta < 1$ rad, therefore $\int_0^\theta (1 - cos(\theta'))d\theta = \theta - sin(\theta) \approx \frac{\theta^3}{3!}$. By making this approximation one induces a relative error of at most 5% for θ = 1 rad and even much less for lower values of θ. This enables us to solve the volume integral *V(d)* of Eq. 3.10 by the expression shown in Eq. 3.11 :

$$V(d) \approx \int_0^d (4\pi)^2 t^2 \frac{(d^2-t^2)^{\frac{3}{2}}}{3!\, l^3} dt = \frac{\pi^3}{12} \frac{d^6}{l^3} \tag{3.11}$$

For arbritary values of θ and for $\frac{d}{l} \leq \pi$, one could include additional terms of the series expansion in Eq. 3.10 to obtain a very accurate estimate of the volume *V(d)*, see Eq. 3.12 :

$$\begin{aligned} V(d) &= \int_0^d (4\pi)^2 t^2 \left(\frac{(d^2-t^2)^{\frac{3}{2}}}{3!\, l^3} - \frac{(d^2-t^2)^{\frac{5}{2}}}{5!\, l^5} + \frac{(d^2-t^2)^{\frac{7}{2}}}{7!\, l^7} - \frac{(d^2-t^2)^{\frac{9}{2}}}{9!\, l^9} + \frac{(d^2-t^2)^{\frac{11}{2}}}{11!\, l^{11}} + ... \right) dt \\ &= \pi^3 \left(\frac{d^6}{12l^3} - \frac{d^8}{384l^5} + \frac{d^{10}}{23040l^7} - \frac{d^{12}}{2211840l^9} + \frac{d^{14}}{309657600l^{11}} + ... \right) \end{aligned} \tag{3.12}$$

The expression $\theta - sin(\theta)$ can still be expanded in series for $\frac{d}{l} > \pi$, however, to keep the integration easier, the rotational variable $d\theta$ should not span regions over than π rad. Note however that :

1. Eq. 3.11 exactly matches the volume of a hypersphere in a six-dimensional Euclidian space (see Eq. 3.2) and is the same as the formula used by Huggins (Huggins, 2014) (only a factor of 2 difference, attributed to the choice of reference state);
2. our equation provides a very accurate estimation for *V(d)* when $\theta < 1$ rad. and should be corrected, either by numerical integration or analytically by the truncated summation (Eq. 3.12) for larger distances;

3. evidently, the expression for the volume relies on the choice of the scaling factor l;

4. the correction added to the naive formulation of entropy (Eq. 3.3) by the nearest-neighbor theory implies however some approximations in order to properly address the Poisson binomial distribution when the total number of samples is too large ($n \to \infty$). The approximations made scale as $\frac{1}{n}$, which is merely similar to the ratio of the volume of the hypersphere of radius equal the distance to the k^{th} nearest-neighbor over the reference volume (as showed in `Theorem 8` by Demchuk *et al.* (Singh et al., 2003)). Consequently, one should expect a linear fitting between deviations of estimates/true entropy and the average distance of the k^{th} nearest-neighbors.

3.3 Application to MD simulations

In its present context, the theory as discussed above remains somehow `tasteless` until it is practically being applied to the post-processing of a MD simulation trajectory. Even though MD provides the thermodynamic samples for the complexes, a series of steps will however be required to readily apply the above formulae to extract the rotational- translational entropy loss upon complex formation of biomolecules. The procedural workflows, as discussed hereafter, are summarized in Fig. 3.1.

Among other order parameters, the atomic root mean square fluctuations (RMSF) provides a sensitive measure of the flexibility of the molecule. Furthermore, compared to global molecular properties like the center of mass motion, the RMSF has the advantage of decoupling the reference coordinates from other degrees of freedom, allowing in a more easy way the computation of individual contributing terms to the rotational and translational entropy loss upon binding.

The subset of reference atoms belonging to the 1^{st} molecule are used to align all MD snapshots in the reference frame of the first snapshot (or more in general to the most representative conformation). The rationale behind this step is to drop the random rotational- translational diffusion of the complex, the latter entering in the definition of a 1 M random reference state.

Once all the snapshots have been superimposed to the reference frame of the 1^{st} molecule, one ideally selects the most rigid part of the 2^{nd} molecule to obtain an optimal rotation-translation to align the first reference system with. Just after that, the rotational and translational coordinates $(\boldsymbol{\mathcal{R}}_1, \boldsymbol{t}_1, \boldsymbol{\mathcal{R}}_2, \boldsymbol{t}_2, ..., \boldsymbol{\mathcal{R}}_n, \boldsymbol{t}_n)$ between any pair of snapshots are extracted and used to compute the `coumpound` distance $d_{i,j}$ (see Eq. 3.6, with subscripts $(i, j) \equiv (a, b)$).

As already emphazised, the choice of l in the computation of the `compound` distance is arbitrary, but should be done with utmost care. Therefore, one chooses l in such a way that the translational and rotational terms are proportionally weighted in the resulting distance. Since the k^{th} nearest-neighbor distance is used in the formulation of entropy (Eq. 3.5), we take the ratio of the mean k^{th} nearest

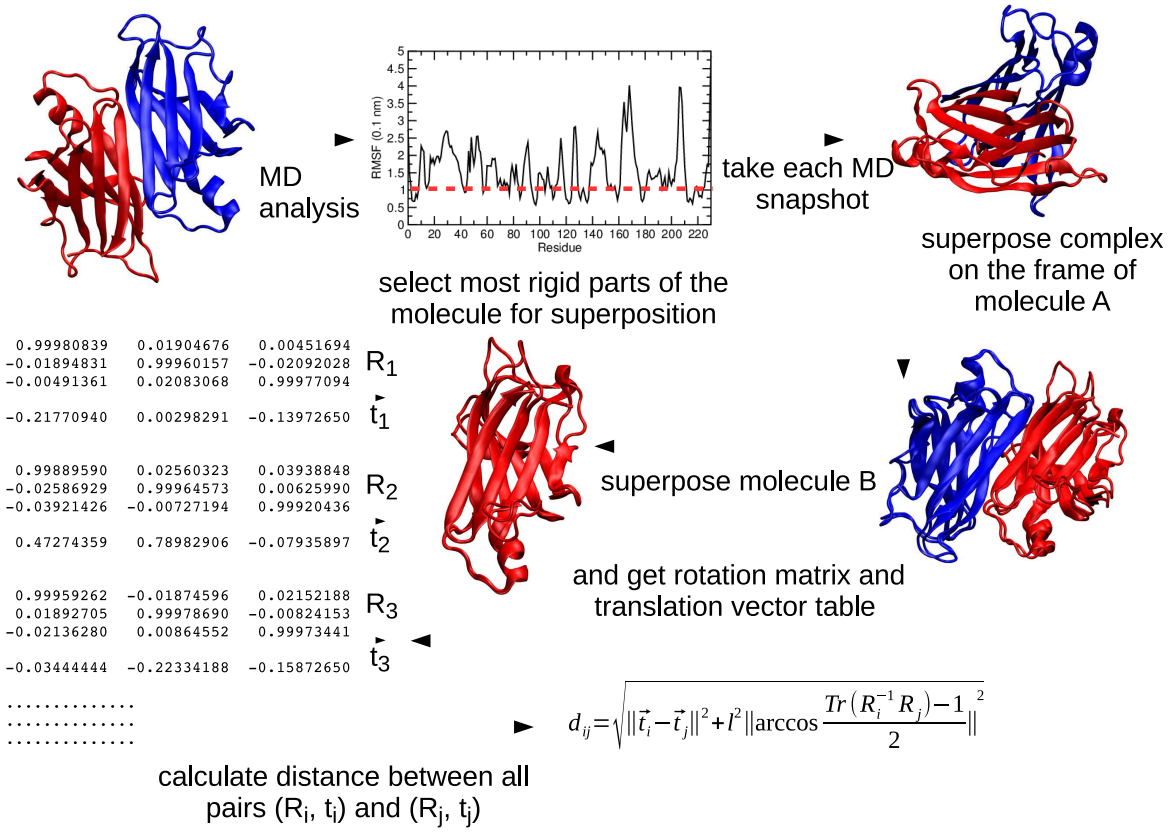


FIGURE 3.1: Application of our approach to a post-processing of a biomolecular MD simulation. Lower atoms' root mean square fluctuations (RMSF) are selected as reference state and used for superposition. Atoms of the 1st molecule are used to superimpose all frames on the starting complex. For all superimposed complexes the 2nd molecule is further rotated and translated to optimally align its lower RMSF atoms to the starting complex. Translations and rotations are listed for each frame and used to compute the distance between any pair of rotational-translational $(\boldsymbol{t}_i, \boldsymbol{\mathcal{R}_i})$ states of the 2nd molecule with respect to the 1st one.

neighbor translational distance to the mean k^{th} nearest-neighbor rotational distance, see Eq. 3.13.

$$l = \frac{\langle d^{\boldsymbol{t}}_{i,k} \rangle}{\langle d^{\boldsymbol{\mathcal{R}}}_{i,k} \rangle} \tag{3.13}$$

Practically, for each sample i the list of distances $d_{i,j}$ $(i \neq j)$ is sorted and the k^{th} shortest nearest-neighbor distance is noted as $d_{i,k}$. The volume $V(d_{i,k})$ is then computed according to Eq. 3.10 or its approximation following Eq. 3.12. Finally entropy is estimated from the total number of samples n, based on Eq. 3.5.

It is good recording that, in the explicit distance formulation of entropy (Eq. 3.5), the lengths are expressed in Angstroms and the implicit reference state corresponds to the concentration of 1 molecule in 1 Å^3 in random rotational state. A

conversion factor of -7.41 k_B is needed to obtain the entropy with the concentration of 1 M in random rotation space.

3.4 MD simulations of a homodimer of TTR

The X-ray crystallographic structure used as starting configuration is the wild-type TTR taken from the RCSB Protein Data Bank (PDB) (Berman et al., 2000) PDB id : 1F41 (Hörnberg et al., 2000), solved at the resolution of 1.3 Å. The structure has been found to crystallize as a tetramer made by two symmetric dimers with residues 1-9 and 126-127 not present. We removed all the crystallographic water molecules before running the simulation and isolated the dimer from the tetramer.

The MD simulation lasted 250 ns using Gromacs-4.6.2 (Hess et al., 2008). The molecular interactions were described using amber99sb-ildn force field (Lindorff-Larsen et al., 2010). To preserve the neutrality of the system, 10 Na^+ counter ions were added. The whole system was then solvated using the 3-site rigid water model TIP3P (Jorgensen et al., 1983; Jorgensen and Madura, 1985; Mahoney and Jorgensen, 2000). Protein atoms were placed at the center of a cubic box at a minimum distance of 10 Å from the edges. We used Periodic Boundary Conditions and the solvated system consisting of protein, 10 counter ions and about 15520 TIP3P water molecules.

The system was first minimized for a maximum of 2500 minimization steps using the steepest descent minimization algorithm. A minimization step size of 0.1 nm and a maximum convergence force of 1000.0 kJ $mol^{-1}nm^{-1}$ were employed. The equilibration phase was done in 2 steps; 100 ps in *NVT* ensemble followed by 100 ps in *NPT* ensemble. During the first equilibration stage the leap-frog integrator with integration timestep of 2 fs was used to update the changes in the system. Particle Mesh Ewald summation (Darden, York, and Pedersen, 1993; Essmann et al., 1995) accounted for long-range electrostatics interactions. The temperature was equilibrated to a reference value of 300 K using the velocity rescaling (modified Berendsen thermostat) (Bussi, Donadio, and Parrinello, 2007) with the coupling constant of 0.1 ps. Short-range electrostatics and van der Waals interactions were truncated with a 10 Å cut-off. All bonds were constrained with the LINCS alogorithm (Hess et al., 1997). In the *NPT* equilibration phase the same parameters were used and the pressure was stabilized to 1.0 bar using the Parrinello-Rahman pressure coupling (Parrinello and Rahman, 1981; Nosé and Klein, 1983), with the coupling constant of 2.0 ps. Finally, MD simulation lasted 250 ns in *NPT* ensemble. Snapshots were collected every 10 ps along the trajectory giving a total of 25000 snapshots, the last 24000 of which have been used in the analysis.

3.5 Results and discussion

We readily apply the equations discussed in Theory formulation to selected test and practical situations, representing well the features amended in the nearest-neighbour theory by our approach :

❶ the random rotation-translations entropy ;

❷ the correlated rotation-translations entropy ;

❸ the rotational-translational entropy of a homodimer of TTR.

For each of the points underlined above, only where explicitely indicated, the entropy was estimated with the implicit reference state for free rotations and translations randomly distributed over a cube of 1 $Å^3$. In this section, however, only the last point above (`item` ❸) will be discussed, the detailed consideration of the other two aspects can be found in the published materials (Fogolari et al., 2016). Among other reasons, the simple one is because the last point straightforwardly affords the step-by-step workflows presented above, Application to MD simulations. Nonetheless, it is consistent to highlight that, for the two points ❶ & ❷ above, the entropy was already known theoretically or through numerical integration, and using Eq. 3.4 with the volume computed by the integral of Eq. 3.10 we accurately reproduce the expected entropy values, therefore legitimating our approach. The same equations are then applied to estimate the rotational-translational entropy lost upon dimerization of a homodimer of TTR.

The motivations behind the choice of TTR in this thesis have been framed yet, see TTR Motivations. Moreover, its study is part of an ongoing project on the effect of mutations on the tetramerization of TTR, the focus of the forthcoming Chapter 4. It is well established that tetramer dissociation (associated with partial unfolding) is one, if not the first step into the amyloidogenesis pathways of several TTR variants (Quintas et al., 2001). In this context, besides other contributions, it is pertinent to perceive the effect of the relative mobility of monomers in the complex of wild-type and mutant TTRs, if any. Finally, biomolecular protein complexes offer more challenging situations compared to systems in which a ligand tightly binds to a receptor, thereby greatly reducing its rotational-translational space, making convergence of entropy calculation more obvious.

Entropy of TTR dimerization

MD snapshots taken at 1 ns intervals were superimposed on the starting structure (using backbone atoms) to identify the most rigid parts of both monomers (schematic views of TTR structures are displayed in Fig. 1.4). The list of residues with average RMSF < 1.2 Å was compiled and used to align all the trajectory snapshots (frames skipped every 10 ps). In particular, monomer I backbone atoms were first superimposed to the starting structure and then further rotation and translation were applied to superpose monomer II on the reference structure. The latter rotation-translation was taken as the rotational-translational state of monomer II in reference to the frame of monomer I, as done before by Fogolari

et al. (Fogolari et al., 2015). For consistency, the same steps were looped inverting the roles of monomers I and II.

The results obtained following the above steps are reported in Fig. 3.2, where the rototional-translational entropy lost upon binding of TTR dimer is plotted against the average k^{th} nearest-neighor distance, with different values of the scaling length l. It is seen in Fig. 3.2 that the best converging moieties are relative to the scaling length of 5 and 10 Å (last 2 continuous curves), consistent with the value of l computed according to Eq. 3.13 which is 9.7 Å. This result stresses on the importance in the choice of the scaling length (which is theoretically random) to cancel the domination effect of one of the two contributing terms entering in the definition of the `coumpound` distance.

The computed translational-rotational entropy lost upon binding with respect to the reference state is then estimated (by linear extrapolation) to -9.6 k_{B} (i.e -17.0 k_{B} with respect to 1 M concentration in random orientation state). Upon inverting the roles of monomers I & II, nearly similar values of entropy lost are obtained.

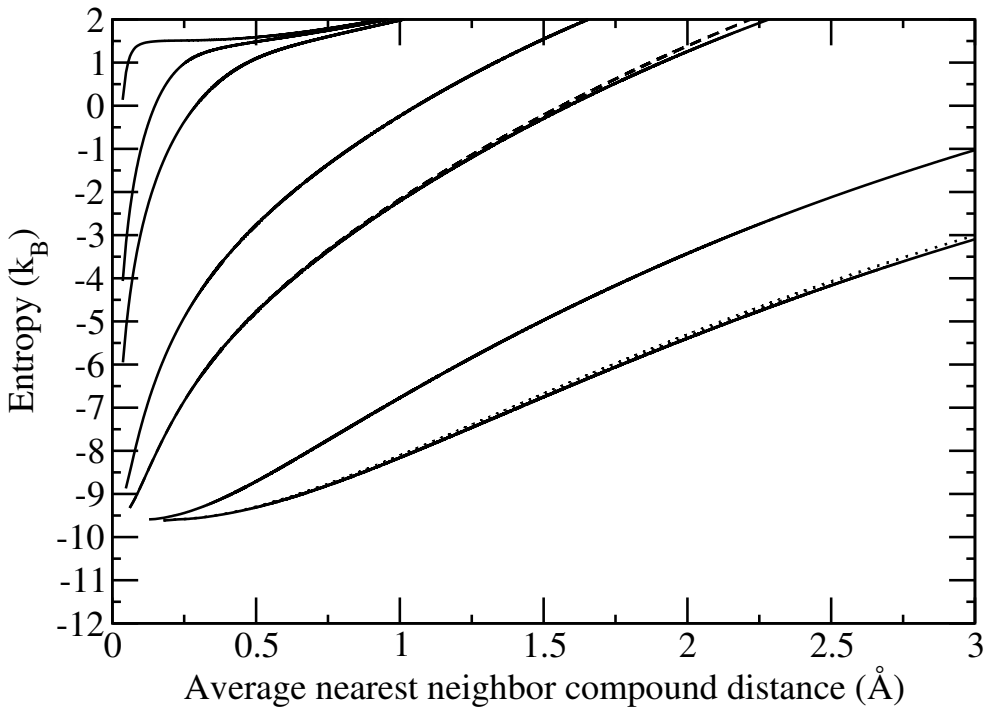


FIGURE 3.2: End-point rotational-translational entropy estimates of n=24000 samples of TTR homodimer from MD simulations. Entropy computed from the k^{th} nearest-neighbor is plotted against the average k^{th} nearest neighbor distance. Like it should be, distances are increasing with the number of neighbors k. The curves (continuous line) are relative to the different choices of scaling length l : (**from top**) 0.01 Å, 0.05 Å, 0.1 Å, 0.5 Å, 1.0 Å, 5.0 Å, 10.0 Å. The estimated value of l according to Eq. 3.13 is 9.7 Å (dotted line). The dashed line represents the entropy extimated following the Huggins' equation, its is reported for the sake of comparison.

Since the translational and rotational degrees of freedom are correlated, one can easily remark the effect of the scaling length l on the asymptotic behavior of the estimated entropy. However, this is less sensitive for random rotation-translations (point ❶ above, not shown). The entropy computed assuming rotations and translations to be uncorrelated is -9.6 k_B for rotations and 1.5 k_B for translations (overall -8.1 k_B). This implies an underestimation of the entropy loss by 1.5 k_B.

3.6 Summary

Conformational entropy in general and translational-rotational entropy in particular are often neglected upon binding of a ligand to its target, likely attributed to the inappropriate methods for dealing with large correlation involved and to the high dimensionality of the degrees of freedom. Thanks to the nearest-neighbor method recently implemented, whose rationale is estimating the local probability density around each sample by counting its number of neighbors within a hypersphere of radius equal the distance from that sample to its k^{th} nearest neighbors, part of this bottleneck can be handled over, albeit one still have to tackle the correlation effects when both translation and rotation are present. In order to properly address the entropy, important factors to be taken into account are the definition of a metric in rotational-translational space and the valuation of the volume of the hypherspere for the choosen radius. Using these parameters, it is possible to compute the distances of each sample to its k^{th} nearest neighbor. Huggins framed the issue and proposed a simplistic treatment for the computation of distances involving both translation and rotation in roto-translational space. Here, following its theory, we provide a more general `coumpound` distance to merge translations and rotations (Eq. 3.6). This latter formulation defines a general scaling factor l necessary for ensuring the equal weigthing of translation and rotation to the global distance. In a nutshell, this chapter has setted the theoretical backgrounds of our theory and showed how it can be applied to a realistic scenario of a post-processing of a MD trajectory. Our approach indeed, relieves an important part of the dimensionality problem facing binning methods like histogram and provides an accurate estimation of full rotational-translational entropy from samples of rotational-translational states.

Chapter 4

Molecular dynamics simulation of an *in silico* engineered mutation of transthyretin[1]

The evidence on the rigidity and stability of TTR structure that has been provided so far, has shown that the dissociation from native tetramer to unfolded monomer is strongly disfavored, preventing the amyloid fibril formation pathway. We present in this chapter MD simulations of an *in silico* engineered point mutation, Ser117Glu introduced by Berni and coworkers on a previously proposed double point mutant variant Phe87Met/Leu110Met. This mutation considerably shifts the tetramer-folded monomer equilibrium towards the monomer, making this triple mutant a useful tool for structural and dynamical studies. Here the structural basis of tetramer dissociation and the dynamics resulting from dissociation is assessed by MD simulations, performed on the wild-type (WT-TTR), double mutant (MT-TTR) and triple mutant (3M-TTR) TTR tetramer. Dissociation of the tetramer into dimers is observed for 3M-TTR and the transition is characterized in terms of dynamic and thermodynamic features.

Keywords: TTR, MD, point mutation, markov state models.

4.1 Introduction

TTR is a well folded homotetrameric assembly of about 55 kDa with each subunit entailing 127 aa. residues. Several variants of TTR have been found to be involved in misfolding diseases, thus providing reliable models for the study of amyloidosis formation pathways. Indeed, this justifies the numerous structural studies involving TTR, either truncated or in full length. Extensive description of

[1] C. J. Dongmo Foumthuim; S. Fortuna; A. Corazza, G. Esposito; F. Fogolari (2018). Molecular dynamics simulation of an *in silico* engineered mutation of transthyretin, **In preparation**.

structure, function and motivations behind the choice of TTR in this thesis have been discussed earlier in Chapter 1. In the present chapter, we used all-atom unbiased MD simulations in realistic-like experimental conditions, i.e. room temperature and pressure, neutral pH ionization state, physiological ion concentration and explicit water solvent, to probe the stability of an *in silico* engineered mutation in relation to WT-TTR PDB 1F41 (Hörnberg et al., 2000) and double point mutant, MT-TTR PDB 1GKO (Jiang et al., 2001). It is worth noting that no crystal structure was available when doing this work but very recently R. Berni and collaborators have released one (Zanotti et al., 2017). The mutation engineered on MT-TTR consists of mutating Ser117 residue located in edge strand H by Glu residue, i.e. Ser117Glu point mutation. Indeed, the resulting structure is a triple point mutant (Phe87Met/Leu110Met/Ser117Glu) variant of TTR (3M-TTR). Both Met110 located in strand G and Glu117 residues have their side chains pointing towards the central binding channel. Meanwhile, Met87 side chain located in between α-helix-F loop is pointing through the middle of F and H strands of the next monomer of the same crystallographic asymetric unit.

Using bioinformatic (Berrera, Molinari, and Fogolari, 2003; Schymkowitz et al., 2005) and statistical (Beauchamp et al., 2011) based-analysis tools, we have clearly showed that the triple point mutant variant (3M-TTR) is strongly monomeric prone even at standard and unbiased simulation conditions. While the process of fibril formation does comprise two broad steps, tetramer-to-folded monomer equilibrium, i.e. dissociation followed by monomer misfolding and self-assembly, this work has only covered the first step and future perspectives are directed towards complete description of the second one.

4.2 Simulation details

4.2.1 Molecular models

The X-ray structures used as starting configuration are the wild-type transthyretin (WT-TTR) taken from the RCSB Protein Data Bank (PDB) (Berman et al., 2000), PDB id : 1F41 (Hörnberg et al., 2000), solved at the resolution of 1.3 Å; the double point mutant (MT-TTR) F87M/L110M (PDB id: 1GKO (Jiang et al., 2001)) resolved at 2.10 Å and the *in silico* engineered triple mutant F87M/L110M/S117E (3M-TTR). The structure of triple mutant was obtained from that of MT-TTR mutating the Ser117 by Glu (S117E) using the protein modelling software Swiss-PdbViewer (Johansson et al., 2012). The N-terminal residues 1-9 and C-terminal 126-127 are not present in the structure and were not modeled. Tetramers were built from deposited asymmetric units (dimers) by applying crystal symmetry operations. All the crystallization water molecules were removed prior to run the simulations.

4.2.2 Molecular dynamics simulations

The molecular mechanics CHARMM27 all atoms force field with CMAP correction (Bjelkmar et al., 2010) accounted for characterization of molecular interactions. The protein atoms were placed at the center of a cubic box at a minimum distance of 10 Å from each edge. The whole system was then solvated using the 3-site rigid water model TIP3P (Jorgensen et al., 1983; Jorgensen and Madura, 1985; Mahoney and Jorgensen, 2000). To account for the neutrality of our systems, 20 (for WT-TTR and MT-TTR) and 24 (for 3M-TTR) Na^+ counterions were added to replace the equivalent number of water molecules and thus, equilibrate the negative charges. The initial set-up is therefore made by a solvated system consisting of about 7048 protein atoms, 24 counterions and 22573 TIP3P water molecules.

The systems were first minimized using the steepest descent minimization algorithm, with a minimization step size of 0.1 nm and a maximum convergence force of 1000.0 $kJmol^{-1}nm^{-1}$. The equilibration phase was done in 2 steps; 100 ps in *NVT* ensemble followed by 100 ps in *NPT* ensemble. During the first equilibration stage, the leap-frog integrator with integration timestep of 0.002 ps was used to update the changes in the system. Particle Mesh Ewald summation (Darden, York, and Pedersen, 1993; Essmann et al., 1995) accounted for long-range electrostatics interactions. The temperature was equilibrated to a reference value of 300 K using the velocity rescaling (modified Berendsen thermostat)(Bussi, Donadio, and Parrinello, 2007), with a coupling constant of 0.1 ps. Short-range electrostatics and van der Waals interactions were truncated with a 10 Å cutoff. All bonds were constrained with the LINCS alogorithm (Hess et al., 1997). In *NPT* equilibration stage, the previous parameters were still used and the pressure was stabilized to 1.0 bar using the Parrinello-Rahman pressure coupling (Parrinello and Rahman, 1981; Nosé and Klein, 1983), with a coupling constant of 2.0 ps. Finally, MD simulations lasted 100 ns in *NPT* ensemble. Snapshots were collected every 2 ps along the trajectory, giving a total of 50000 snapshots.

4.2.3 Molecular dynamics simulation analysis

MD trajectories were analyzed with available structural-based tools in *Gromacs*-5.0.4 (Hess et al., 2008; Abraham et al., 2014) and the thermodynamic stability of the systems was further processed using the Academic License version of the bioinformatics tool `Foldx` (Schymkowitz et al., 2005) ; In which the following commands were used `Stability` and `AnalyseComplex` respectively to gain information on protein stability and interacting interface free energies. In all the cases, the values were averaged over the whole simulation trajectory. `AnalyseComplex` command outputs the Gibbs interacting free energy of binding for a complex formation (folding), say AB ($A + B \rightarrow AB$), computed as $\Delta\Delta G_{\mathrm{AB}} = \Delta G_{\mathrm{AB}} - (\Delta G_{\mathrm{A}} + \Delta G_{\mathrm{B}})$ where ΔG is the free energy of folding.

For studying the dissociation mechanism of the triple mutant, we sampled and clustered its trajectory to construct a macrostate model using Markov State Models (MSMs) with MSMBuilder package, version 2.7 (Bowman, Huang, and Pande,

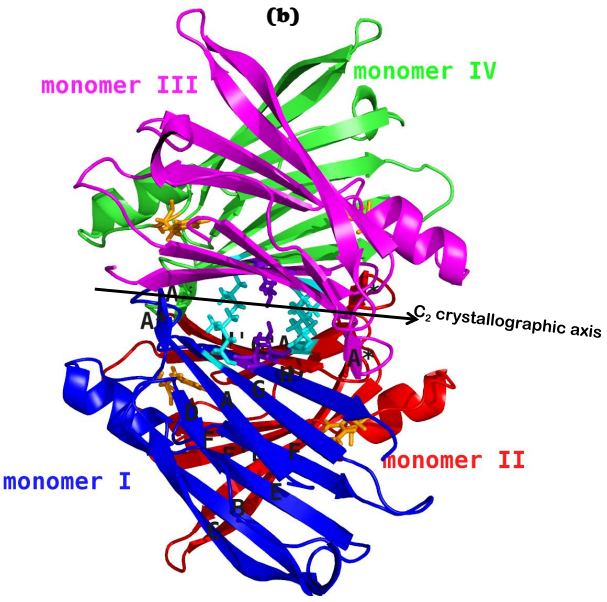
2009; Beauchamp et al., 2011; Bowman, 2014). We analysed the dissociation pathways using the build-in Transition Paths Theory module of MSMBuilder (`CalculateTPT`), assigning as initial states those macrostates resembling more the crystal structure (native-like states), i.e. those with smallest RMSD and as final state the one with largest RMSD. In the clustering phase, the RMSD metric using a hybrid *k*-centers/*k*-medoids clustering algorithm (Beauchamp et al., 2011) with a treshold intercluster distance of 1.6 Å and 250 iterations were used. The 50000 snapshots from MD simulation were initially clustered into 220 microstate centers which were further lumped into 76 coarse-grain model (macrostates) using the Bayesian Agglomerative Clustering Engine (BACE) (Bowman, 2012). Both models (micro and macro) were validated, i.e. whether or not they are Markovians, computing and analyzing the Implied Time Scale (ITS) plots (see Fig. A.1 in Appendix A). States were further visualized thanks to MSM Explorer package (Cronkite-Ratcliff and Pande, 2013).

Pictures presented in this chapter in particular and in the thesis as a whole were either collected with PyMOL (Delano, 2002) or VMD (Humphrey, Dalke, and Schulten, 1996), secondary structures were assigned using DSSP program (Kabsch and Sander, 1983) and hbonds occupancy was computed using the readHBmap.py tool reporting only hbonds with occupancy greater than 10 %.

4.3 Results and discussion

4.3.1 Tetramer dissociation

The TTR tetramer already reported in Fig. 1.4-(b) is placed here for an easily understanding of the discussion related to monomer nomenclatures :



In order to illustrate the structural differences and namely dissociation between the simulated tetramers and the starting one, we computed the centre of mass (COM) distances between the different monomers along the simulation timescale, Fig. 4.1. It is worth mentioning that Fig. 4.1-(d), (e) and (f) parallel well the profiles of root mean square deviation of backbone atoms (upon tetramer superposition) and protein radius of gyration and therefore the latters are not presented. Moreover, the structure of 3M-TTR is not at equilibrium. Thus the average data displayed below should be taken with due care and are considered as mostly qualitative.

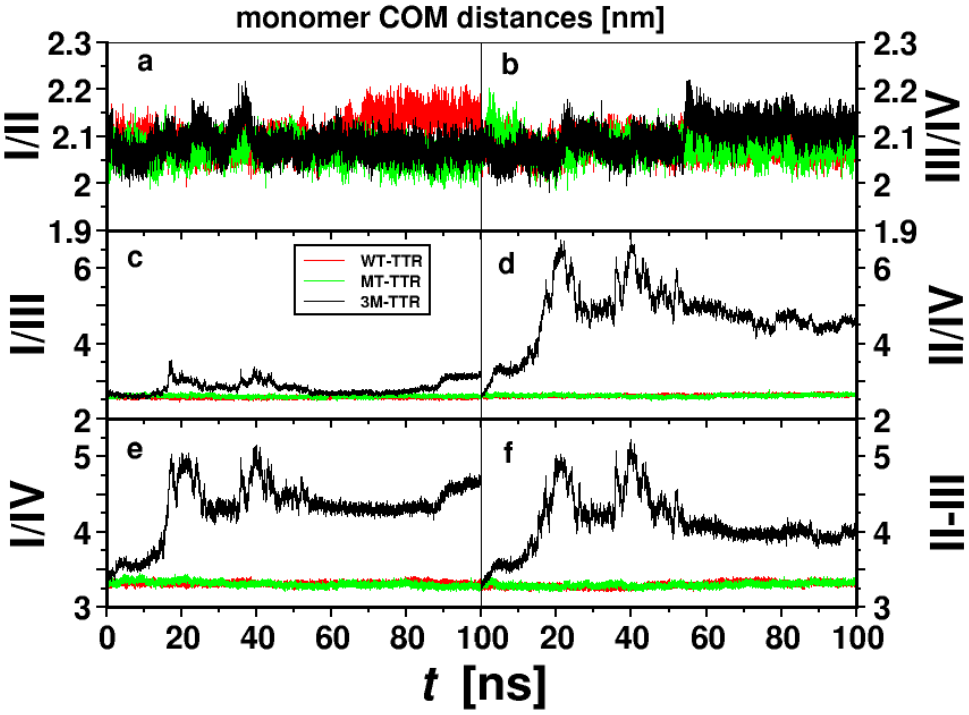


FIGURE 4.1: Centre of mass distances between the different monomers along the simulation time.

In Fig. 4.1 it is seen that the smallest separation distance between monomers is monitored in asymmetric units of both dimers I/II (a) and III/IV (b) for all simulated systems. The distance profiles between symmetric units II/IV (d) and asymmetric units I/IV (e) and II/III (f) of different dimers are quite similar. As a consequence of dissociation of 3M-TTR a large increase in center of mass distances, in relation to WT-TTR and MT-TTR, is observed at the interface of II/IV (d) on one hand and I/IV (e) and II/III (f) on the other hand, respectively. This analysis shows that 3M-TTR tetramer is largely undergoing structural block destructuration compared the others. Dissociation is initiated at the interface of mononers II/IV, that corresponds to the dimer-dimer contact, followed by detachment of monomer III, matching the experimental results by Foss et *al.*, (Foss, Wiseman, and Kelly, 2005).

WT-TTR and MT-TTR monomer's distances are the same (within fluctuations), showing that both structures are well preserved and overall superimposable (as

confirmed by RMSD analysis, not shown). This certainly is apparently inconsistent with the analytical ultracentrifugation experiments stating that no tetramer of MT-TTR is detected in solution over a concentration range of 7.2 μ to 0.1 mM (Jiang et al., 2001). The COM distance confirms the ease of dissociation of 3M-TTR structure compared to the other two. In addition, having shown that the dissociation is likely initiated at the dimer-dimer interface, namely in our simulation at II/IV, we further investigated the changes at the inner sheet-inner sheet (HGAD-HGAD) interface of symmetric units of different dimers (I-III and II-IV). This was done by computing the distance between symmetric residues along the simulation and compare to the reference wild type's X-ray structure, Fig. 4.2.

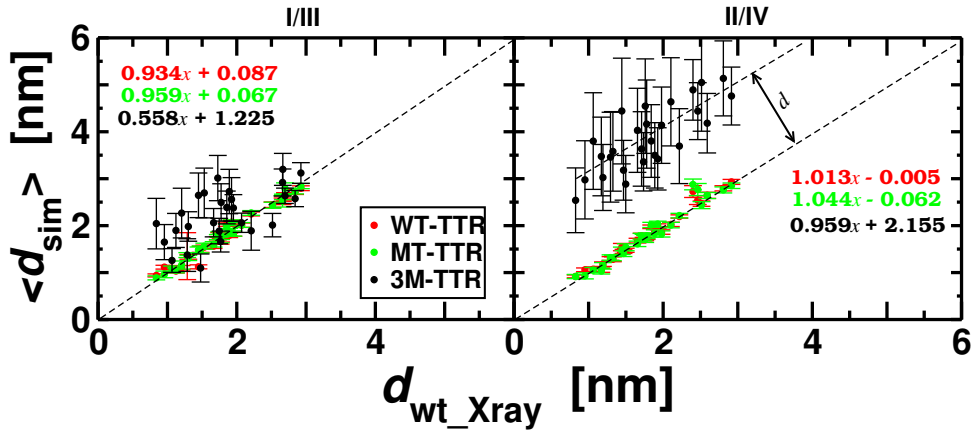


FIGURE 4.2: Average inner sheet - inner sheet distance between symmetric residues at dimer-dimer interface compared to initial (optimized) WT-TTR X-ray structure. Roman letters I to IV stand for monomeric units. Equations written on both graphs are the corresponding fitted linear equations. Each point represents the average distance between two symmetric inner sheet residues, with the corresponding standard deviation.

In Fig. 4.2 the inner sheet-inner sheet distances of correspondent residues of the two symmetric monomeric units (I-III and II-IV) at the dimer-dimer interface are in very good agreement for WT-TTR and MT-TTR compared to the wild-type crystal structure. Contrariwise, symmetric residues at the inner sheet-inner sheet interface of 3M-TTR undergo a significant distance increase, especially at the limit of monomers II and IV. The correlation coefficients computed between average simulation distances and WT-TTR Xray's one (R^2 = 0.983, 0.992, 0.595 at I-III and 0.988, 0.977, 0.779 at II-IV, respectively for WT-TTR, MT-TTR and 3M-TTR), reveal that the intrinsic separation gap, i.e. within the same tetrameric assembly, is $\sim$ 40 % (at I-III) and $\sim$ 20 % (at II-IV) bigger in 3M-TTR structure than in the other two ones. Furthermore, on a more global scale, i.e. variant-to-variant structure-based comparison, 3M-TTR displays a remarquable distance shift at the frontier of monomers II-IV, with a separation gap $d \approx 1.59$ nm (see Fig. 4.2) compared to the two others. These observations indicates a structural block dissociation in 3M-TTR.

4.3.2 Structural fluctuations

The detailed picture of residues mobility within TTR tetramer protein structures was visualised by computing the root mean square fluctuation (RMSF), Fig. 4.3. Analysis of RMSF plots clearly points out the higher structural flexibility of the loop regions in comparison with the rest of the protein structure. Large residue fluctuations are observed in loops B-C, F-G, α-helix-F loop and D-E loop.

WT-TTR and MT-TTR show nearly the same range of fluctuations with a mean value of 0.07±0.03 nm. The highest peaks in WT-TTR (Fig. 4.3-a) are found in monomer II (D38, 0.251 nm; E62, 0.162 nm), monomer IV (D99, 0.159 nm; S52, 0.112 nm) and monomer III (A81, 0.135 nm). In MT-TTR (Fig. 4.3-b), residues with highest fluctuations are closely the same as in WT-TTR and are located in monomer I (A37, 0.179 nm) and monomer II (D38, 0.177 nm; A81, 0.150 nm; S100, 0.139 nm). For 3M-TTR dissociation prevents this kind of analysis.

Monomers of 3M-TTR are not affected much by dissociation, indeed the plots in Fig. 4.3 (a′), (b′) and (c′) describing fluctuations within individual monomeric units, i.e. upon superposition of the single monomer, show very similar fluctuations in WT-TTR, MT-TTR and 3M-TTR.

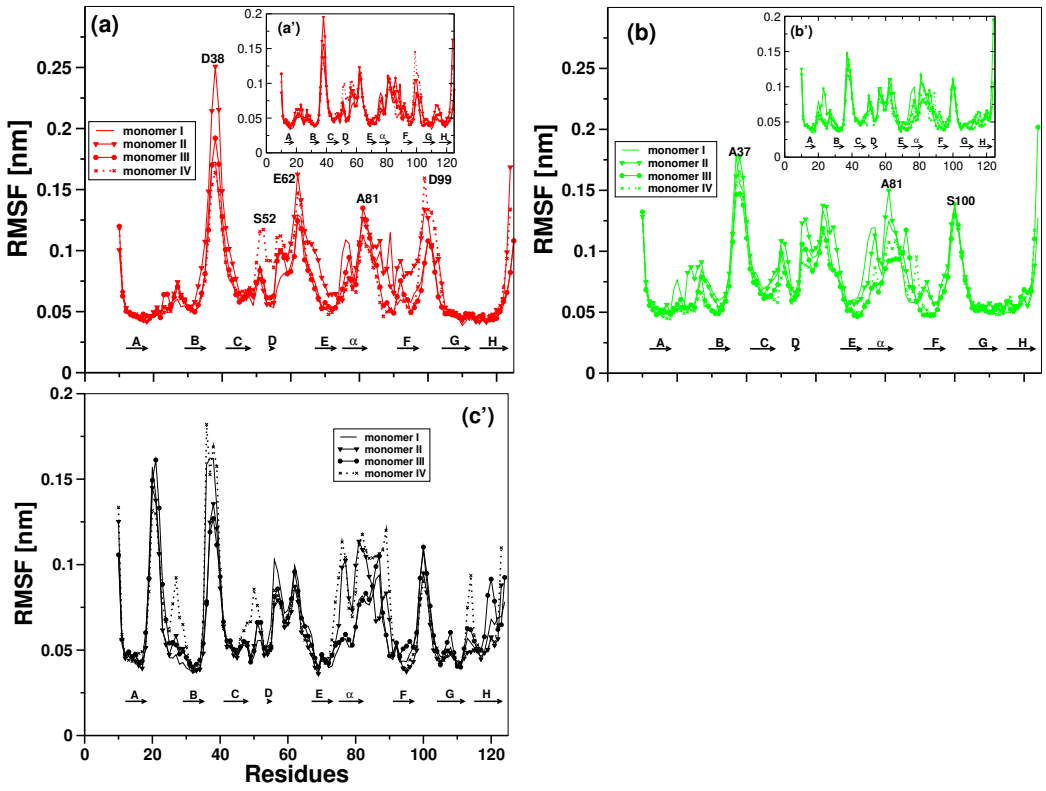


FIGURE 4.3: Residue average root mean square fluctuations of backbone atoms. From top to bottom WT-TTR (a, a′), MT-TTR (b, b′) and 3M-TTR (c′). While (a), (b) are averaged with respect to the center of mass of tetramers, (a′), (b′) and (c′) are averaged on individual monomers. Letters A-H represent the β-sheets and α is the short intermediate helical region (75-82) between stands E and F.

4.3.3 Localised structural transitions

Local conformational transitions were assessed by computing the dihedral phi (ϕ) and psi (ψ) angles of individual residues involved in the mutation and represented as Ramachandran plots, Fig. 4.4.

The X-ray ($\phi°, \psi°$) dihedral angles for residue 87 are (-84,-55), (-84,-40) and (-85,-41) in WT-TTR, MT-TTR and 3M-TTR respectively. The same values reported from simulation are as follow : (-75$\pm$12, -46$\pm$9), (-80$\pm$16, -44$\pm$13) and (-91$\pm$15, -38$\pm$15), Fig. 4.4-(a). These latter lie in the allowed regions for α-helical secondary secondary structure, $-89 < \phi < -39$ and $-66 < \psi < -16$ (Hovmöller, Zhou, and Ohlson, 2002), confirming that no secondary structural transition occured at this position.

The ϕ/ψ dihedral angles of L110 and M110 residues Fig. 4.4-(b) look even more similar than the previously observed ones. No major deviation was observed in individual monomers and the dihedral angles for residue 110 in each variant keep closed to the equilibrium value. The average reported simulation values are (-122$\pm$11, 122$\pm$8), (-116$\pm$12, 122$\pm$10) and (-117$\pm$11, 119$\pm$10) respectively in WT-TTR, MT-TTR and 3M-TTR. Indeed, these fall into the region corresponding to β-sheet secondary structure, $-180 < \phi < -45$ and $45 < \psi < 225$ (Hovmöller, Zhou, and Ohlson, 2002) and imply that no structural transition is seen in the process of point mutation Leu110Met.

In Fig. 4.4-(c), while both TTR variants are displaying nearly the same ϕ/ψ averages, (-137$\pm$10, 140$\pm$9), (-136$\pm$16, 134$\pm$12) and (-131$\pm$17, 135$\pm$11), it is seen that a distinct region is being accessed by mutant variants. The region is defined by $-40 < \phi < 70$ and $30 < \psi < 140$ corresponding to left handed helix conformation. The transition is observed only once for both MT-TTR and 3M-TTR within the first 10 ns of the simulation possibly as a transient adaptation to *in silico* mutation of residue 110.

4.3.4 Secondary structural changes

The illustration of dynamical processes in our systems during the simulations was done by computing the secondary structure, Fig. 4.5. The latter analysis was performed thanks to the DSSP program which determines the existence of hydrogen bonds as a criteria for the presence of secondary structure (Kabsch and Sander, 1983). While dihedral (ϕ, ψ) angles (Fig. 4.4) were useful for assessement of local structural changes, secondary structure is able to highlight the global structural changes.

In Fig. 4.5 we can see that the helical region exhibits large structural fluctuations and even get disrupted from time to time, like in monomer II for all the TTR variants. F-strand undergoes some structural fluctuations at its begining, particularly in monomer I, III and IV. In G-H loop, some part of the structure is being converted from bend to β-turn and conversely. D-strand displays quite high structural fluctuations in monomers II and IV of WT-TTR ; in monomers II

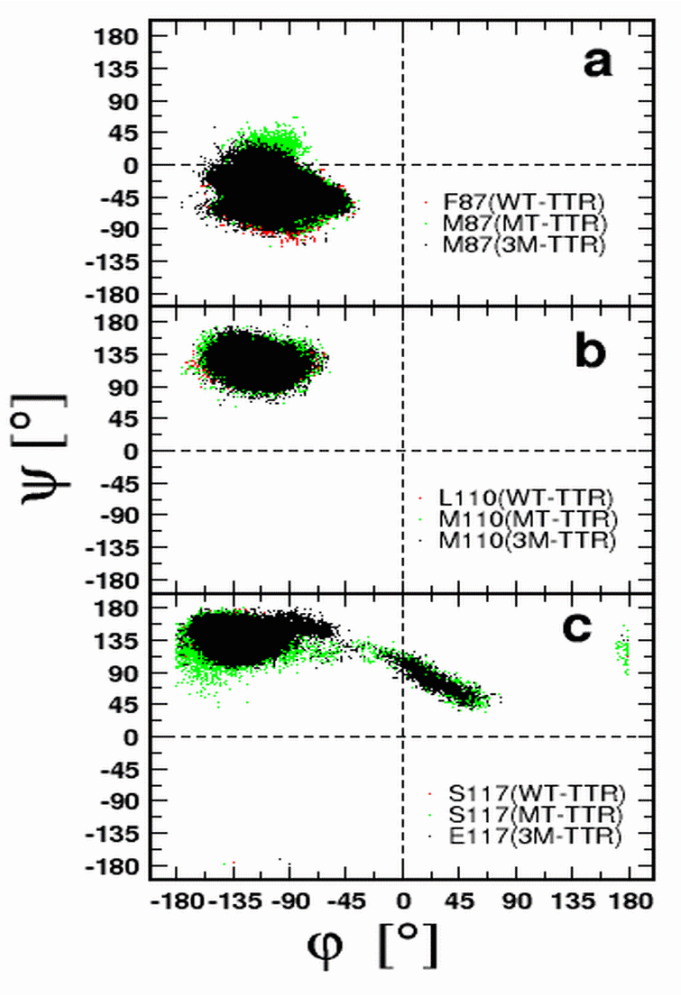


FIGURE 4.4: Ramachandran plots of individual residues at individual point mutations.

and III of MT-TTR and in all monomers for 3M-TTR. A- and C-strands make the most stable parts of the molecules. E-strand in the vicinity of α-helix is showing some fluctuations in monomers I, III & IV of all the variants. In spite of structural fluctuations exhibited by major parts of the tetrameric molecular assemblies, Fig. 4.5 clearly points out that monomers are nevertheless remaining folded, i.e. they preserve essentially their secondary structures. This observation is supported by the fact that no evident disruption and secondary structure conversion are seen in the core region (made by the β-barrels) of the different systems studied.

4.3.5 Intra and Interchains hydrogen bonds occupancy

In order to understand the mechanism of 3M-TTR dissociation we computed the H-bonds occupancy along the MD trajectory considering both main chain-main chain, main chain-side chain and side chain-side chain, at monomer-monomer and dimer-dimer interfaces. Fig. 4.6 summarizes the occupancies of the interactions that were identified to significantly perturb the tetrameric assemble of TTR

FIGURE 4.5: Time-based secondary structure analyses. From top to bottom the secondary structures of WT-TTR (a), MT-TTR (b) and 3M-TTR (c) are respectively displayed. Colours were used to distinguish between secondary structure types and letters A to H stand for β-strands and α is the helical portion between E and F strands.

upon mutations. It should be noted that for 3M-TTR the occupancy reflects an average of both (starting) associated and dissociated configuration.

In Fig. 4.6, seven major hbonds (with occupancy $\geqslant$ 80 %) can be seen almost symmetrically distributed in both dimers, three at I-II interface and four at III-IV interface. These are being lost in 3M-TTR with their average occupancy dropping below 40 %. These include the interactions Y114(CO)-A120(HN) and A120(HN)-Y114(O) (main chain-main chain) ; T119($H^{\gamma 1}$)-S115(O^{γ}) and S115($H^{\gamma 1}$)-T119($O^{\gamma 1}$) (side chain-side chain). Most of them are located in the H-strand and some in the G-H loop.

At the interface of symmetric units (I-III and II-IV) there is a significant loss in hbond occupancy in 3M-TTR, in particular at II-IV interface, confirming what we have seen in Fig. 4.2. The occupancies of 22GLY(O)-122VAL(HN) and 122VAL(HN)-22GLY(O) (main chain - main chain) located in H-strand and A-B loop lie under the 10 % (Fig. 4.6). Interestingly, it is worth mentioning that, in 3M-TTR, several side chain-side chain interactions are being formed following dissociation and subsequent temporary association. These involve many inner sheets residues like K15, R104 and E117, most of which show significant deviation from starting WT-TTR X-ray structure.

Analysis of hydrogen bonds connectivity confirms the pivotal role of such networks in preserving the protein integrity, typically in some dedicated portions,

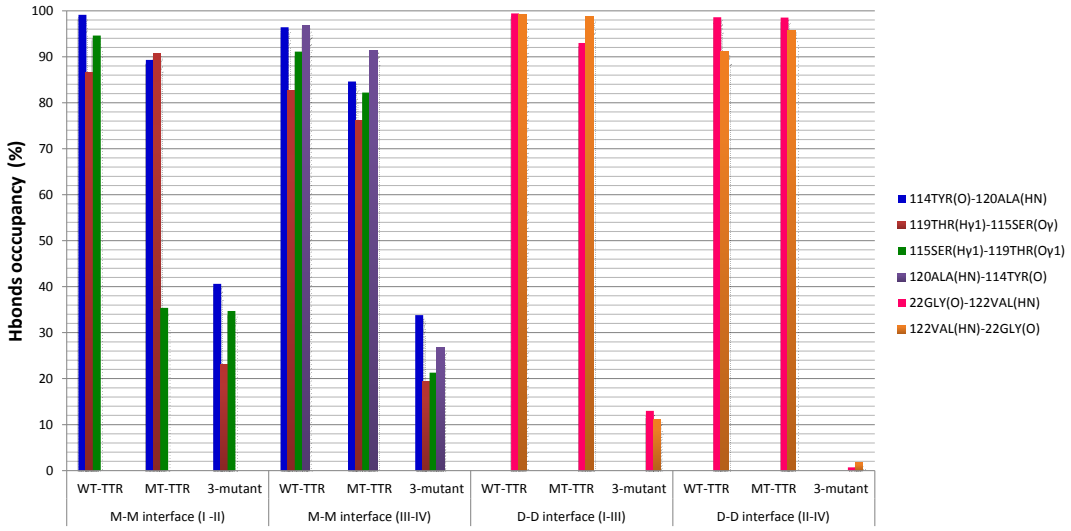


FIGURE 4.6: Summary of hbonds with their occupancies identified to destabilize the tetrameric structure of TTR. M and D stand respectively for monomer and dimer while I to IV are different monomers.

in this case H-strand, G-H and A-B loops. These susceptible portions of TTR that build up the momomer-monomer and dimer-dimer interfaces are definitely playing a key role in destabilization of tetrameric subunits of TTR, especially in the case of 3M-TTR. Indeed, the disappearance of hbonds in the previously mentioned domains along the timescale of the simulations brings a possible explanation on the integrity loss of tetrameric block in 3M-TTR.

4.3.6 Free energy along the dissociation route

The structural and thermodynamic stability of the mutant structures in relation to wild-type were further probed thanks to two independent bioinformatic tools, i.e. `Foldx` (Schymkowitz et al., 2005) (Fig. 4.7-A) and the statistical effective energy function `BMF` (Berrera, Molinari, and Fogolari, 2003) ((Fig. 4.7-B)). We computed the folding free energy differences between mutants and wild-type ($\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wild\text{-}type}$) throughout the tetramer-unfolded monomers equilibrium, Figs. 4.7. Some words of caution are due presenting the following data :

1. the energy functions used take into account implicitly the entropy loss from internal degrees of freedom, but not that arising from external degrees of freedom as implied by the tetramer to dimers and the dimers to monomers transitions.
2. the simulation of the 3M-TTR mutant is not at equilibrium, because of dissociation. For this reason the data should be regarded as qualitative.

Finally, notwithstanding the fluctuations and differences in underlying principles and datasets, the two approaches used here are consistent with each other as detailed below.

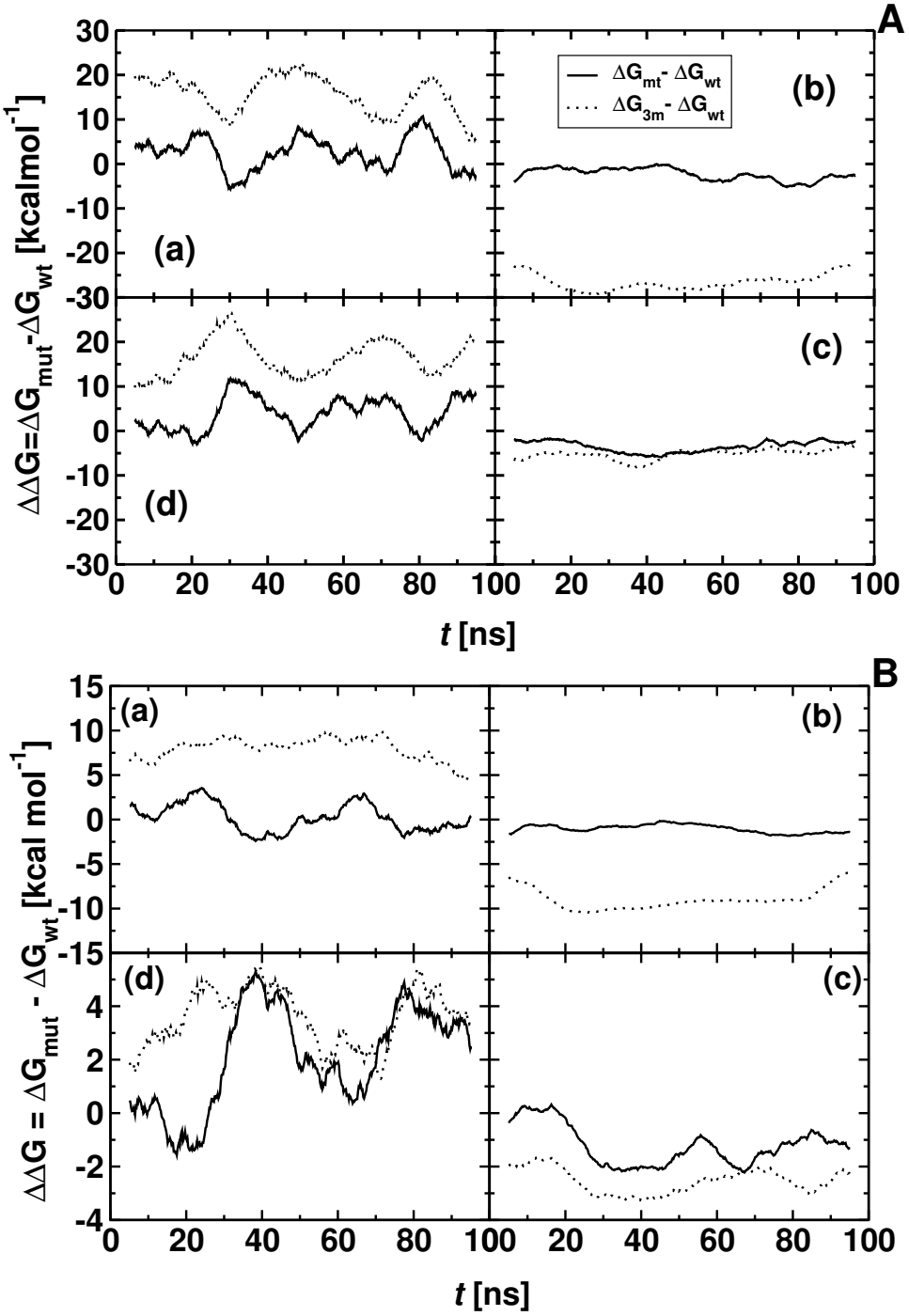


FIGURE 4.7: Folding free energy difference between mutants and native wild-type along the simulation trajectory and througout the tetramer-unfolded monomer equilibrium. In the case of tetramer (a), $\Delta\Delta G_T$ is the global stability free energy change of tetramers, i.e. the free energy required to fold the tetramers from their unfolded monomers. Subscripts *mut* and *wt* are respectively mutant (MT-TTR or 3M-TTR, with their simplified notations *mt* and *3m*) and wild-type (WT-TTR). $\Delta\Delta G_D$ (b) is the free energy necessary to form the dimers from tetramers (T→D). $\Delta\Delta G_M$ (c) is the free energy involved in the dissociation of dimers into folded monomers (D→M) and $\Delta\Delta G_M^{unf}$ (d) is the free energy of unfolding of individual monomers (M →M^{unf}).

The panels (a) in Figs. 4.7-(A & B) points out that both mutants have positive difference in free energy with respect to WT-TTR, thus are less stable. Their average free energy $\Delta\Delta G_T$ computed using `BMF` (`Foldx`) are +0.3 and +7.8 kcal/mol (+2.5 and +15.6 kcal/mol) respectively for MT-TTR and 3M-TTR, confirming the decreased tetramer thermodynamic stability upon mutations, significant in the case of 3M-TTR.

In order to track down all contributions, $\Delta\Delta G$'s values were computed and reported in subsets (b, c & d) of Fig. 4.7-(A & B) for all the steps in the equilibrium pathway from tetramers to unfolded monomers (required before monomers self-assemble into amyloid fibrils) through dimers and monomers. We called $\Delta\Delta G_D$, $\Delta\Delta G_M$ and $\Delta\Delta G_M^{unf}$ the free energy for the dissociation of tetramers into dimers (T→D), for the formation of folded monomers from dimers (D→M) and for the unfolding of individual monomers (M→M^{unf}) respectively. Based on simulations we guessed that tetramers first break down into dimers I/II and III/IV, i.e. along the C_2 crystallographic axis, instead of I/III and II/IV (See Fig. A.2 Appendix A). Thus, $\Delta\Delta G_D = \Delta G_D - \Delta G_T$ (dissociation free energy at I/II-III/IV dimer-dimer contact), $\Delta\Delta G_M = \Delta G_M - \Delta G_D$ (dissociation free energy at I-II and III-IV monomer-monomer contacts) and $\Delta\Delta G_M^{unf} = \Delta G_M^{unf} - \Delta G_M$.

$\Delta\Delta G_D$ using `BMF` (`Foldx`) average is -1.1 and -8.7 kcal/mol (-2.5 and -26.3 kcal/mol for MT-TTR and 3M-TTR respectively, with respect to WT-TTR. This means both equilibriums are shifted towards the right (subsets b in Figs. 4.7 A & B), i.e. the formation of dimers. The quite high value (absolute) displayed by 3M-TTR indicates the propensity of tetrameric assembly of the latter structure to dissociate into dimers I/II and III/IV with respect to wild-type. Considering the dimer-monomer equilibrium (subsets c in Figs. 4.7 A & B), $\Delta\Delta G_M$'s average using `BMF` (`Foldx`) is -1.2 and -2.5 kcal/mol (-3.3 and -5.2 kcal/mol) over the trajectory for MT-TTR and 3M-TTR respectively, relative to WT-TTR. The values are however rather limited compared to the overall stability computed using `BMF` for the process of four WT-TTR monomers associating into 2 dimers (-14.2 kcal/mol).

Finally the change with respect to WT-TTR in folding free energy of monomers has been computed. Both `BMF` and `Foldx` predict a shift towards the formation of more stable monomers for both mutants $\Delta\Delta G_M^{unf} > 0$, i.e. the equilibrium M⇄M^{unf} is shifted towards the folded form (panels d in Figs. 4.7 A & B). The unfavourable formation of unfolded monomers in both mutants strongly correlate with the secondary structure analysis displayed in Fig.4.5 showing that notwithstanding the enhanced dissociation of tetramers into monomers through dimers, monomers remain stable (folded). Besides, this result would strongly suggest that both monomers of our mutant variants are non-amyloidogenic, and this was known as far as MT-TTR is concerned (Jiang et al., 2001).

4.3.7 Mechanism of tetramer dissociation

The path of dissociation of dimers in 3M-TTR may be followed during the simulation (Fig. 4.8). The first step in conformational transition is the dissociation at the interface I/IV involving residues 17-24 and 110-123, including position 117

mutated to Glutamic acid in 3M-TTR at variance with MT-TTR where the corresponding residue is a Serine. The transition appears driven by the electrostatic repulsion of the pairs of acidic residues Glu 117 close in each dimer. First the interface is weakened (up to 15 ns) and disrupted (20 ns), then both I/II and III/IV dimers remain rigid. Dimer III/IV rotates for most of the simulation about a hinge centered on salt bridges Glu 51 (I) - Arg 104 (III), Arg 104 (I) - Glu 51 (III) and hydrogen bonds Glu 51 (III) - Thr 123 (I) and at the end only about the latter two interactions. The final (possibly transient) conformation is stabilized by salt bridges Glu 117 (I) - ARG21 (III) and Lys 15 - Glu 54 at interface I/III. Other interactions at interface II/III are mostly hydrophobic. No sign of dissociation or conformational rearrangement at dimers is observed during the simulation. For both dimers the proximity of Glu 117 acidic groups should result in repulsion which is however reduced by ionic interactions.

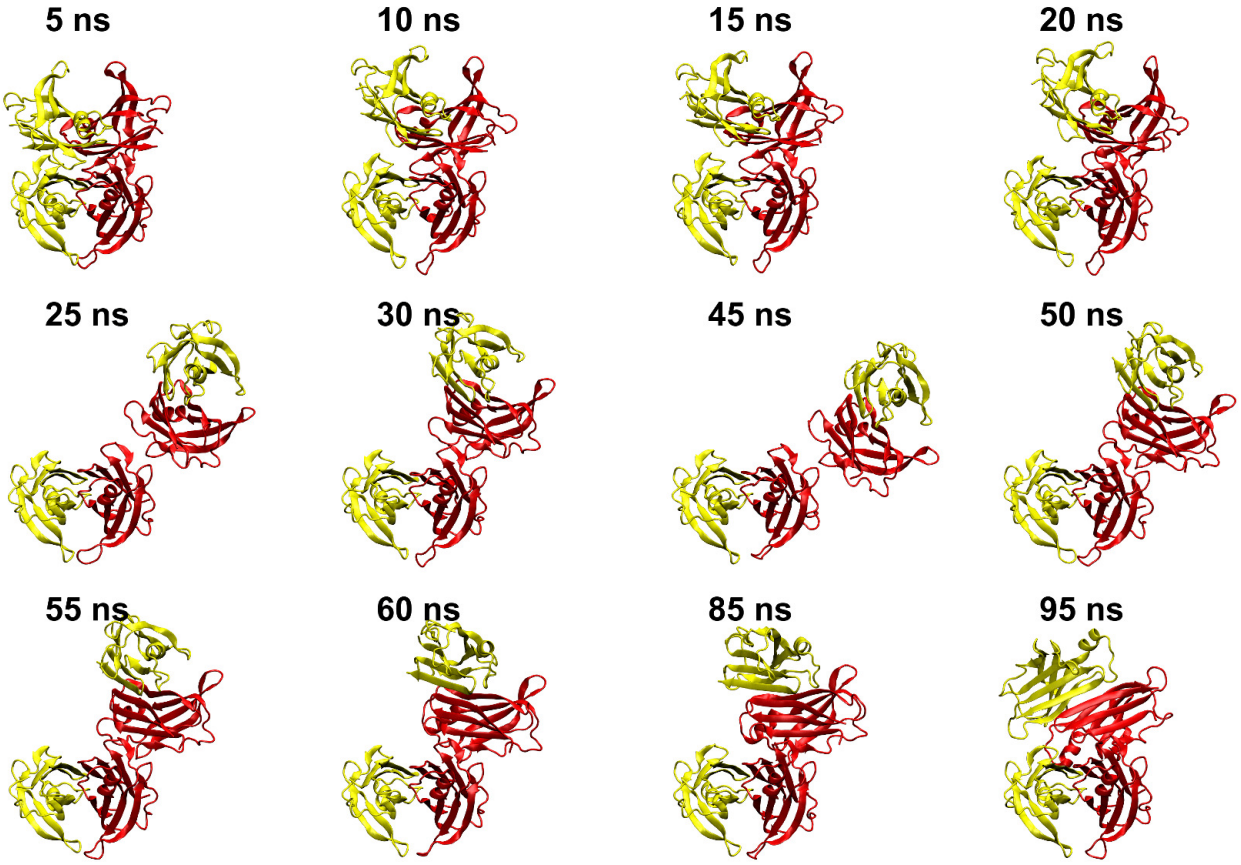


FIGURE 4.8: Snapshots from 3M-TTR 100 ns molecular dynamics simulations. Chains I and III are shown in red, chains II and IV are shown in yellow. Chains I and II, in the lower part of the tetramer are used for superposing all snapshots on the starting conformation.

4.4 Summary

Molecular Dynamics simulations were used in this work to probe the structural stability of an *in silico* engineered point mutation Ser117Glu on the double point mutant of transthyretin (TTR) MT-TTR (PDB 1GKO), in relation to wild-type (PDB 1F41) and MT-TTR itself. Molecular properties analysis such as secondary structure, solvent accessible surface area (SASA), hydrogen bonds, root mean square deviation (RMSD) clearly confirm the highly dissociative nature of that mutation. Furthermore, clustering and discretization of simulation trajectories using Markov State Models (MSMs) show that our microstate model poorly captures the native-like state in the case of our engineered structure (RMSD $\sim$ 8.2 Å) (see Fig. A.3-a). As far as MT-TTR is concerned, a very good structural agreement between the native-like microstate and X-ray structure was seen (RMSD $\sim$ 1.3 Å) (see Fig. A.3-b), revealing therefore that its structure was essentially preserved. It has been noticed that the thermodynamic equilibrium along the pathway from native tetramer to unfolded monomers is strongly shifted towards the formation of monomers. Assuming the thermodynamic underlines the kinetic, this would indeed imply the speeding up of the rate-limiting step just before monomers self-assemble into amyloid fibrils. Nonetheless, the free energy computed according to two independent approaches (`BMF` and `Foldx`) implies stable monomers (relative to the stability of WT-TTR) after tetramer dissociation, shifting therefore the last step of fibril formation towards the non-amyloidogenic mutant structures. The simulation data discussed throughout this chapter were completed well before a X-ray structure of the triple mutant been released. However, they nicely match the recently published experimental X-ray data on the triple mutant structure (Zanotti et al., 2017). Being the first *in silico* studies on this triple mutant variant of TTR, our simulations provide strong evidence of its high dissociation trend and essentially focused on the dissociation step from native tetramer to folded monomers. Future prospects are to provide complete description of molecular aggregation and self-assembly of monomeric units of this mutant variant.

Chapter 5

$\beta 2$-microglobulin interactions with hydrophobic surfaces : A molecular dynamics study[1]

In this chapter we present an extensive investigation on the interactions of β2-microglobulin, a paradigmatic protein model for amyloidogenic pathologies, with a model cubic hydrophobic box. A system made up with 27 copies of the protein is surrounded by the united atoms representation of methane hydrocarbons and simulated using the OBC-GBSA implicit solvent molecular dynamics. The results show that most proteins adsorb on the walls of the box without major distortions of local geometry, whereas free molecules maintain proper structure and fluctuations as observed in explicit solvent MD simulations. Thermodynamic analysis suggests that van der Waals interactions are much larger than all other contributions to the free energy of binding. Loss of secondary structure and local unfolding could have important physiological consequences suggesting that hydrophobic surfaces could thus act *in vivo* as promoters of partial unfolding and local clustering that are essential for seeding the formation of amyloid fibrils. Owing to the emerging role of biological interfaces occurring *in vivo* in general and in protein aggregation and fibril formation in particular, we believe our findings afford a general outlook complementary to the biochemical studies on this protein. Overall, albeit on shorter timescale, our simulation details a picture of the magnitude of the driving forces for β2m interaction with a model hydrophobic surface. These results could be significantly helpful in the design of medical devices that can be utilised for the removal of β2m from the blood of hemodialysis patients.

Keywords: Molecular dynamics, Implicit solvent, Protein-surface interaction, Adsorption.

[1] C. J. Dongmo Foumthuim; A. Corazza ; G. Esposito and F. Fogolari. Molecular dynamics simulations of β2-microglobulin interaction with hydrophobic surfaces, *Mol. BioSyst.*, 2017, 13, 2625-2637.

5.1 Introduction

Proteins in their natural cellular environments are considered as a component of a complex solution that may undergo, for either natural or traumatic events, a transition. One such event leading to a metastable condition may be a sudden change of solvation with formation of an aqueous/non aqueous interface, for instance at the boundaries of tissue compartments where interface with lipid membranes occurs. We used MD simulations to explore this type of events with β2m, a paradigmatic protein model for amyloidogenic pathologies. In particular, we address the interaction of β2m with a generic hydrophobic surface using implicit solvent MD simulations. It should be noted that, although much less frequent than anionic lipid surfaces, hydrophobic surfaces and interfaces are important and occur to some extent also *in vivo*, e.g. by perturbation of the cell membrane lipid bilayer structure or at interfaces in tissues. MD simulations studies have addressed the early steps of the process of β2m aggregation (Fogolari et al., 2007), the effect on stability and dynamics of denaturants and/or mutations (Ma and Nussinov, 2003; Park and Saven, 2006; Deng et al., 2006; Fogolari et al., 2011; Mendoza et al., 2011; Gümral et al., 2013; Chong et al., 2015; Chandrasekaran and Rajasekaran, 2016; Camilloni et al., 2016). But the effect of surfaces and interfaces on protein denaturation and amyloid formation has been much less explored.

To detail this picture we simulate a system where 27 copies of β2m are set in a cubic lattice within a box. The boundary of the box is constituted by a cube with each face made by hydrophobic united atoms arranged in a cubic lattice with each particle in van der Waals contact with nearest neighboring ones to provide a hydrophobic boundary and desolvation effects. We expected our simulation could detect adsorption on the wall and possibly 2D diffusion and encounters on the wall surface with induced unfolding and/or aggregation. Our aim is to detail the encounter with the surface, the effects induced on the protein, the driving forces and the strength of adsorption. The accuracy of the implicit solvent model used is checked by following the behavior of proteins non-interacting (`isolated`) with surfaces along with two independent control simulations of the protein alone in explixit water performed using `NAMD` and `AMBER` codes with slightly different protocols. For an extensive overview on β2m-surface interactions and protein-hydrophobic surface interaction in general, readers could consult the published materials (Dongmo Foumthuim et al., 2017).

5.2 Simulation details

5.2.1 System set-up

The starting structure was obtained by excising the X-ray coordinates of β2m (chain B) that composes the major human histocompatibility antigen HLA-A2 complex solved at 2.6 Å resolution (PDB id: 3HLA) (Saper, Bjorkman, and Wiley, 1991). All the external crystallographic water solvent were removed and missing

hydrogens were added using the `pdb2gmx` utility of the software package Gromacs (Abraham et al., 2014). 27 replicas of the starting structure were randomly oriented and placed at the nodes of a cubic grid (spacing 50 Å).

A cubic box comprising 41508 atoms was built assembling six square faces of apolar atoms which are assigned the methane united atom molecular mass 16 $gmol^{-1}$. Each square face is composed by three layers of atoms. The spacing between nearest atoms is 4.0 Å. The walls of the bounding box are meant to represent a generic hydrophobic surface. The density is lower than that of carbon atoms in liquid alkanes by a factor 1.5 to 2.1, which implies a lower number of wall atoms in the simulation.

The highly ordered structure of the bounding walls was restrained during the whole simulation. The final system included 27 identical protein chains (43848 atoms) and the apolar bounding box totaling 85356 atoms (Fig. 5.1).

The box was set large enough (200 Å/side) to enable individual proteins to fully rotate and no bias was used in the simulation. The choice of such high concentration (8 mM) system and the use of implicit solvent model enforces increasing probability of sampling the effects of the surface and of concentration thereby improving statistics. Compared to the advanced sampling techniques this system has the advantage that no CVs must be defined and no biasing potential must be applied. The set up allows in principle the observation of protein-wall and protein-protein contacts and their interplay and/or competition. The same approach has been previously used for protein-protein and protein-small molecule interactions (Fogolari et al., 2007; Fogolari et al., 2011; Moroni, Scarabelli, and Colombo, 2009; Abriata and Dal Peraro, 2015) and its features have been reviewed by Fogolari *et al*. (Fogolari et al., 2012).

It must be noted that in this settings hydrophobic surface and crowding effects (Politou and Temussi, 2015) are present at the same time and could be coupled. Crowding leads in general to stabilizing effects but also, depending on the interactions of the crowder (in this case the solute itself) with the solute, to destabilizing effects. As detailed in the Results and discussion, adsorption on a hydrophobic surface is a very fast and efficient process, and it takes place during the simulation time independently of possible protein-protein contacts and crowding. The description of β2m structure we will refer to in the discussion was done in β2m-structure sub-section.

5.2.2 Forcefield and implicit solvent model

The Generalised Born Surface Area (GBSA) continuum solvation model of Onufriev, Bashford and Case (OBC) (Onufriev, Bashford, and Case, 2004) with amber99sb-ildn force field (MacKerell et al., 1998; Lindorff-Larsen et al., 2010) was used to characterize the molecular interactions. Lennard-Jones parameters for the apolar wall atoms have been set to σ = 0.41 nm, slightly larger than typical single aliphatic carbons and ϵ = 0.625 kJ/mol, i.e. 1.4 times larger than the corresponding parameter for a methylene united atom in the force field OPLS (Jorgensen and

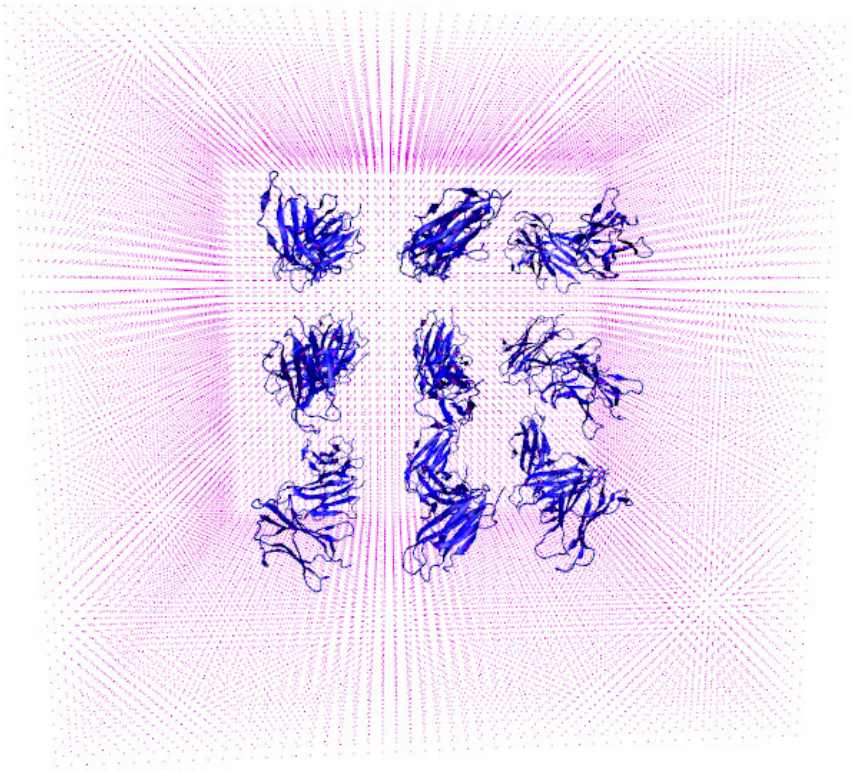


FIGURE 5.1: Initial system with 27 identical chains of β2m (blue) in apolar atoms box (magenta). The proteins are arranged in a cubic lattice. The view of the middle and back layers is partly covered by the front one.

Tirado-Rives, 1988), to partly compensate for the lower density of apolar interaction centers.

5.2.3 Implicit solvent MD simulation

All simulations were performed with gromacs-5.0.4 simulation code (Abraham et al., 2014) in implicit solvent model. The system was initialy minimized for a maximum of 1000 steps using the conjugated gradient minimization algorithm, a minimization step size of 0.05 nm and a maximum convergence force (tolerance) of 10 $kJmol^{-1}nm^{-1}$ with a steepest descent step performed every 50 steps. In the production phase, Brownian dynamics integrator with integration timestep of 1 fs and time constant of 0.2 ps was used to update the changes in the system. The velocity for Maxwell distribution temperature was set to 300 K. Short-range electrostatics and van der Waals interactions were truncated with a 16 Å cutoff and neighbor list was updated every 10 steps. All bonds were constrained with LINCS alogorithm (Hess et al., 1997) with order of matrix coupling expansion of 6 using up to 4 iterations and relative tolerance of 0.001 for SHAKE algorithm. All simulations were performed using the generalized Born surface area (GBSA) implicit solvent model. Generalized Born radii were computed using the Onufriev,

Bashford and Case GBSA algorithm (Onufriev, Bashford, and Case, 2004) using the default scale factors α, β and γ and updated every 5 steps with a cut-off of 16 Å. The implicit solvent dielectric constant was 78.54 and the surface tension 2.259 kJmol^{-1}nm^{-2} using the ACE-type approximation surface (Schaefer, Bartels, and Karplus, 1998). Snapshots were collected every 10 ps along the trajectory. The processes involving the rearrangement of proteins on hydrophobic surfaces are known to take place in timescales ranging from minutes to hours (Sethuraman and Belfort, 2005; Svaldo-Lanero et al., 2008). Nonetheless, albeit on short lengthscale (100 ns), our system set-up allows the proteins to sample the same phenomenon, thus reinforcing the confidence in our results. It is however relevant to precise that convergence for the adsorption process cannot be achieved and the discussion will take this fact into consideration.

5.2.4 Explicit solvent control MD simulation

For the sake of comparison and validation of the implicit solvent model used, a single β2m molecule was also simulated in explicit solvent. Two independent runs of 100 ns each were performed using the programs `AMBER` and `NAMD`. For both simulations the same starting structure of the implicit solvent simulation was used. For `AMBER` all crystallographic water solvent were removed and subsequently missing hydrogens, topology and input coordinates were built thanks to the `xLEaP` module of `AmberTools v15`. The `solvateoct` command of the `xLEaP` module was used to place the protein in a truncated octahedron box (198709.573 Å^3) at a minimum distance of 8 Å from the periodic box edge and fill the box with 5143 rigid 3-sites TIP3P water solvent (Jorgensen et al., 1983). For `NAMD` the protein was solvated in a cubic box with the edges at least 12.0 Å from the protein. 2 Na^+ counterions were added to neutralize the overall negative charge of the initial sytem for the `AMBER` simulation, whereas Na^+ and Cl^- ions where added to have a neutral system at 0.150 ionic strength for the `NAMD` simulation. The system was minimized, smoothly heated and equilibrated similar to the implicit solvent simulation described above. The pressure was kept around the reference value of 1.0 bar using the Berendsen barostat (Berendsen et al., 1984) with relaxation time of 2.0 ps. The particle mesh ewald (PME) accounted for long range electrostatics interactions in the `AMBER` simulation with default parameters, whereas a simple cutoff at 14.0 Å was applied in the `NAMD` simulation.

5.2.5 Free energy calculations

We consider a system with n solute degrees of freedom $(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ and m solvent degrees of freedom $(\boldsymbol{x}_{n+1}, ..., \boldsymbol{x}_{n+m})$, described by the energy function $U(\boldsymbol{x}_1, ..., \boldsymbol{x}_{n+m})$.

The OBC GBSA implicit solvent with the molecular mechanics forcefield is an approximation to the potential of mean force (PMF) $W(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ of the system where the degrees of freedom of water are integrated out in the configurational

integral. The accuracy of this approximation can be judged by the results of the simulation, but it seems appropriate in current context.

With due care, the thermodynamics of the system can be described in terms of solute degrees of freedom represented by the PMF (after the solvent degrees of freedom are integrated out). The difference in free energy between two macrostates A and B of the system, e.g. free and bound to the surface, is given in terms of the PMF, by the sum of the enthalpy and entropy terms of the system described by solute variables and the PMF, Eq. 5.1 :

$$\begin{aligned} \Delta G_{A\to B}(U, \boldsymbol{x}_1, ..., \boldsymbol{x}_{n+m}) &= \Delta G_{A\to B}(W, \boldsymbol{x}_1, ..., \boldsymbol{x}_n) \\ &= \Delta H_{A\to B}(W, \boldsymbol{x}_1, ..., \boldsymbol{x}_n) - T\Delta S_{A\to B}(W, \boldsymbol{x}_1, ..., \boldsymbol{x}_n) \end{aligned} \tag{5.1}$$

Some words of caution are due because the `enthalpy` terms now contain also implicitly solvent entropic contributions and contributions arising from the temperature dependent implicit solvent parameters. With these caveats, in the implicit solvent model used $\Delta H(W, \boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ is obtained as the difference in thermodynamic average of $W(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$, i.e.

$$\Delta H_{A\to B}(W, \boldsymbol{x}_1, ..., \boldsymbol{x}_n) = \langle W(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)\rangle_B - \langle W(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)\rangle_A \tag{5.2}$$

The conformational entropy may be estimated from the conformational ensembles obtained from simulations using the nearest neighbour method (Singh et al., 2003; Hnizdo et al., 2007; Hnizdo et al., 2008; Fenley et al., 2014; Huggins, 2015; Fogolari et al., 2015; Fogolari et al., 2016; Fogolari et al., 2017) which is able to provide an estimate of the probability density in conformational space about each sampled conformation. In practice it is necessary to collect a thermodynamic ensemble of conformational samples for each macrostate (e.g. free and bound proteins). Each conformation is converted in a list of torsion angles, i.e. changes in bond lengths and covalent angles are neglected. The distances among conformations of each residue are computed and used to estimate the entropy of each residue. Readers could have further insights in the cited literature.

A necessary approximation is to consider internal and rototranslational degrees of freedom as independent, which amounts to consider the entropy of these degrees of freedom in an additive way. Rotational and translational degrees of freedom may be treated together (Huggins, 2015; Fogolari et al., 2016) but correlation with internal degrees of freedom is neglected in this approximation. In case of limited sampling also rotation and translation must be treated independently, due to the high-dimensionality of the problem.

Neglected correlations among internal degrees of freedom of different residues may lead to < 20 % correction of the computed values (Fogolari et al., 2015). Notwithstanding all limitations, the method provides a quantitative estimate of the entropy difference between different macrostates (free and bound states).

Like pointed out already in Chapter 3, one needs to define a reference state. In the present case, backbone atoms of the disulfide-bridged residues Cys 25 and Cys 80 were identified as most rigid part of the molecule and used to define the rotational and translational reference state. We consider only restrictions in rotation and translation with respect to the axis perpendicular to the contacting wall. The reference state is 1M random orientation for the free molecule and we consider for the bound molecules parallel movement in the area of a face of a cube where a single molecule would be 1 M (i.e. 140.2 $Å^2$).

5.2.6 Contact analysis

In the analysis of contacts only heavy atoms are considered and two atoms are considered to be in contact when their distance is less or equal the sum of their van der Waals radii plus 1.0 Å. The latter cut-off was found to be the most effective for the definition of statistical contact potentials among a.a. (Berrera, Molinari, and Fogolari, 2003) and roughly corresponds on average to the uniform distance of 4.5 Å often used to define contacts between atoms.

5.3 Results and discussion

5.3.1 Implicit solvent validation

To avoid any artifact due to the use of implicit solvent model simulation, we started by performing a set of simulation runs. Firstly, the CHARMM forcefield (MacKerell et al., 1998) with CMAP correction (MacKerell, Feig, and Brooks, 2004) was used in combination with the OBC-GBSA model (Onufriev, Bashford, and Case, 2004) as implemented in `NAMD` (Kalé et al., 1999). The proteins show great trend to unfold and form helices in tens of nanoseconds. Similar bias were noticed also upon tuning the parameters of the model to accurately reproduce PBSA results (Fogolari, Corazza, and Esposito, 2015).

We tested and then used the AMBER forcefield with the OBC-GBSA model as implemented in `Gromacs` and we checked that the trajectory for `isolated` proteins (i.e. proteins which did not undergo extensive contacts with other proteins and/or with the walls of the bounding box) was stable but that also displayed fluctuations similar to those observed in explicit solvent simulations.

For comparison, two 100 ns simulations were perfomed in explicit solvent using the `NAMD` and `AMBER` MD simulation codes, with almost overlapping results. The RMSD from the starting conformation versus time of the only four proteins which remained `isolated` in the first 60 ns, upon superposition of secondary structures elements of the starting structure, and the corresponding RMSD from 100 ns explicit solvent `NAMD` simulation, is displayed in Fig. 5.2. The temporary increase of RMSD shown by one of the four trajectories is due to fluctuations at strand C' whose hydrogen bonds between Glu 44 and Lys 41 are lost and partly

regained. The overall overview of all the implicit and explicit solvent trajectories in terms of RMSD can be seen in Fig. B.1 Appendix B.

The `isolated` molecules fluctuations at each residue obtained upon superposition of secondary structure elements on the starting structure over all the trajectory show that conformational flexibility at the terminal and loop regions were larger (Fig. 5.3) and in a range consistent with explicit solvent MD simulations. Overall the AMBER forcefield and the OBC-GBSA implicit solvent model used were sufficiently accurate to exclude major artifacts in the simulation, at least on the timescale considered here.

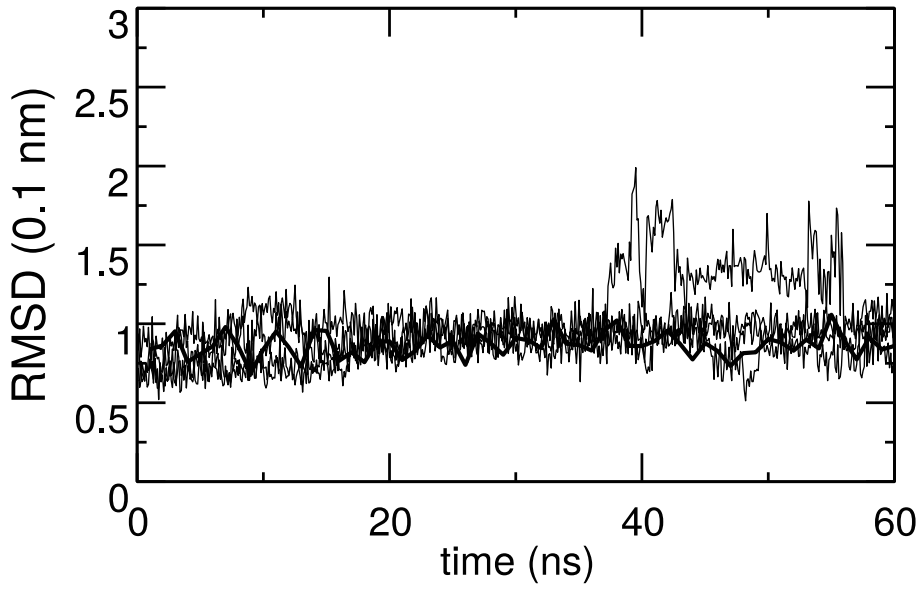


FIGURE 5.2: RMSD from the crystallographic structure versus simulation time for the four `isolated` proteins. RMSD is computed after optimal superposition on the starting structure of backbone atoms of secondary structure elements. The thick line is the corresponding RMSD from the NAMD explicit solvent simulation taken at 1 ns intervals.

5.3.2 Protein adsorption on hydrophobic walls

During the simulation 18 out of 27 molecules encounter and bind the walls of the bounding box within the first 35 ns whereas additional four molecules contact the walls after 80 ns. We refer to these 18 molecules as to the adsorbed molecules. The process, as discussed below, leads to large contact surface areas resulting from partial loss of structure.

We selected the snapshots (every 100 ps) where each of the proteins first encounter the wall. We then counted for each residue how many times it occurs in such first encounter snapshots (with no care of how many contacts each residue establishes with the walls). This analysis aims at understanding whether there is a prefered orientation as observed for the adsorption on gold nanoparticle (Brancolini et al., 2015). Note that since snapshots are taken at 100 ps intervals more

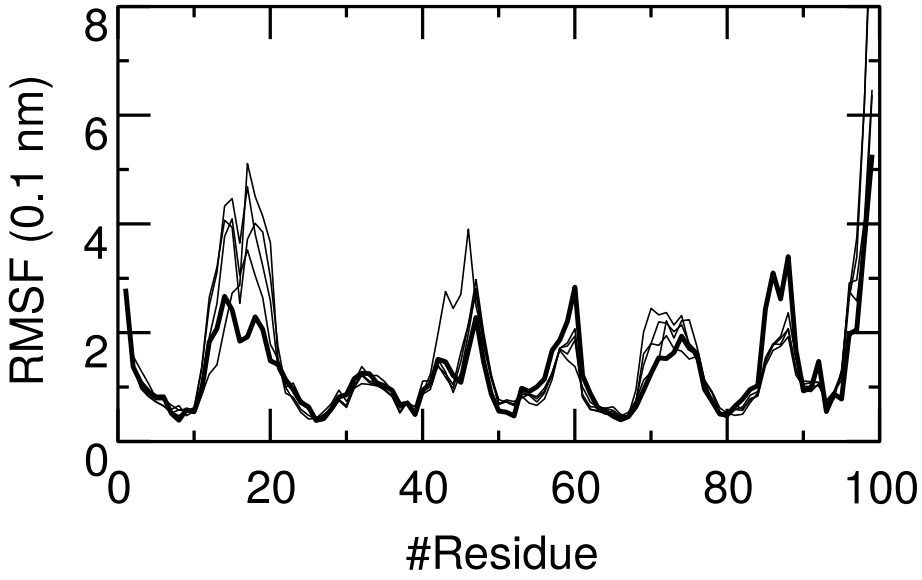


FIGURE 5.3: RMSF versus residue for the four `isolated` molecules in the considered ensemble. RMSF is computed at all residues after optimal superposition of backbone atoms of secondary structure elements on the starting structure. The thick line is the corresponding RMSF from 100 ns explicit solvent simulations.

than a residue may occur in each first encounter snapshot. The analysis of the first encounter snapshots for the 22 proteins contacting the wall shows that the N-terminal end (most frequently contacting residues: Ile 1, Trp 60) is a very localised site of encounter for half of the binding proteins, whereas a larger area of encounter is found at the C-terminal (most frequently contacting residues: Lys 75 and Met 99) (Fig. 5.4). The N-terminal region which is bound to the α-chain in the class I HLA complex was already pointed out to be involved in protein-protein aggregation (Fogolari et al., 2007) and protein gold nanoparticle binding (Brancolini et al., 2015). A role of C-terminal in aggregation was also found but less frequently (Fogolari et al., 2007).

By visual inspection of the simulation trajectories, the adsorption to the walls, that occurs very fast, is independent of protein-protein contacts, diverging from the scenario observed for gold nanoparticles (Cantarutti et al., 2017). This is most likely attributed to the high strength of association, as discussed in Thermodynamic analysis.

5.3.3 Irreversible adsorption on the simulated timescale

As soon as first contacts are formed the protein adapts to the surface and most frequently the number of protein-wall contacts steadily increases sometimes reaching a constant level. The dynamics on the surface then becomes slower. This is observed for most contacting molecules, albeit the rate of growth of contacts is not uniform (Fig. 5.5). Possibly as an effect of the Lennard-Jones parameters chosen for the bounding atoms, the diffusion on the wall is almost frozen

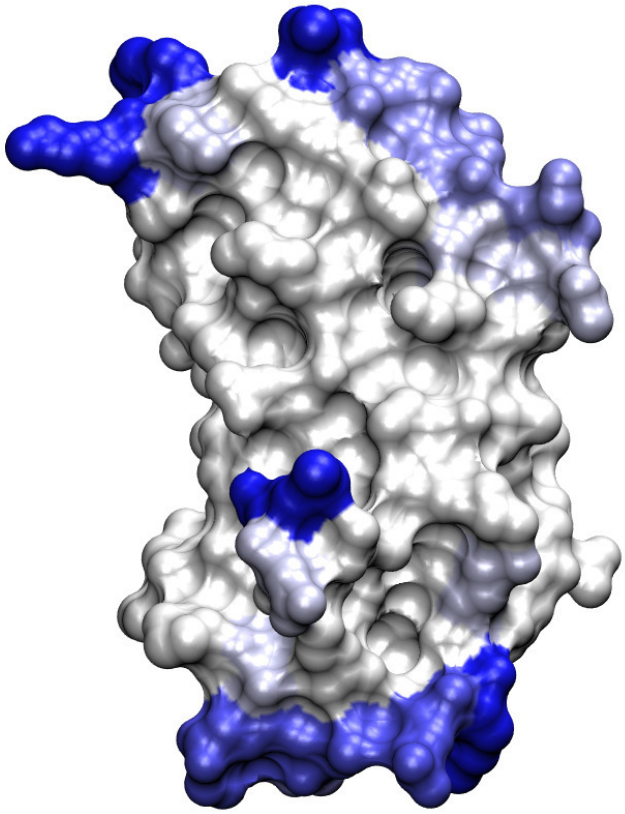


FIGURE 5.4: Most frequently contacting residues in first protein-bounding wall encounter snapshots are highlighted on the solvent accessible surface of β2m. Residues are coloured based on the number of first encounter snapshots in which they are found. Blue is saturated at 4 contacts. The N-terminal region is at the top of the figure.

as can be seen from the plot of the time course of the squared distance of the mass center with respect to the initial position. The effect is independent from protein-protein contacts, so it is attributable to the interaction with the surface. A similar effect on diffusion was not observed in the test simulations performed with the CHARMM forcefield with the walls made of position-restrained on all-atom molecules of methane. This shows that the simulation is rather sensitive to forcefield parameters and therefore to the nature of the hydrophobic surface. Since the density of methane (explicit or united) atoms is the same in the two simulations and lower than the density of real (e.g. graphite) hydrophobic surfaces, it appears that the AMBER forcefield using a slightly larger Lennard-Jones parameter should be more representative of a real scenario.

The plots reported in Fig. 5.5 show that convergence is not achieved in the simulation. Dynamics of adsorbed proteins becomes very slow and for many chains the number of contacts is still increasing at the end of simulation. Also free molecules are most likely to be adsorbed as soon as they contact the walls.

Notwithstanding this observation, it is worth noting that the maximum number of contacts is similar in most plots where a plateau is reached, which suggests that any further structural change should take place on a longer timescale. Experiments showed that such rearrangements take place in minutes to hours (Sethuraman and Belfort, 2005; Svaldo-Lanero et al., 2008) and that loss of structure, monitored through loss of function, at least for lysozyme, is also linked with crowding on the surface (Sethuraman and Belfort, 2005).

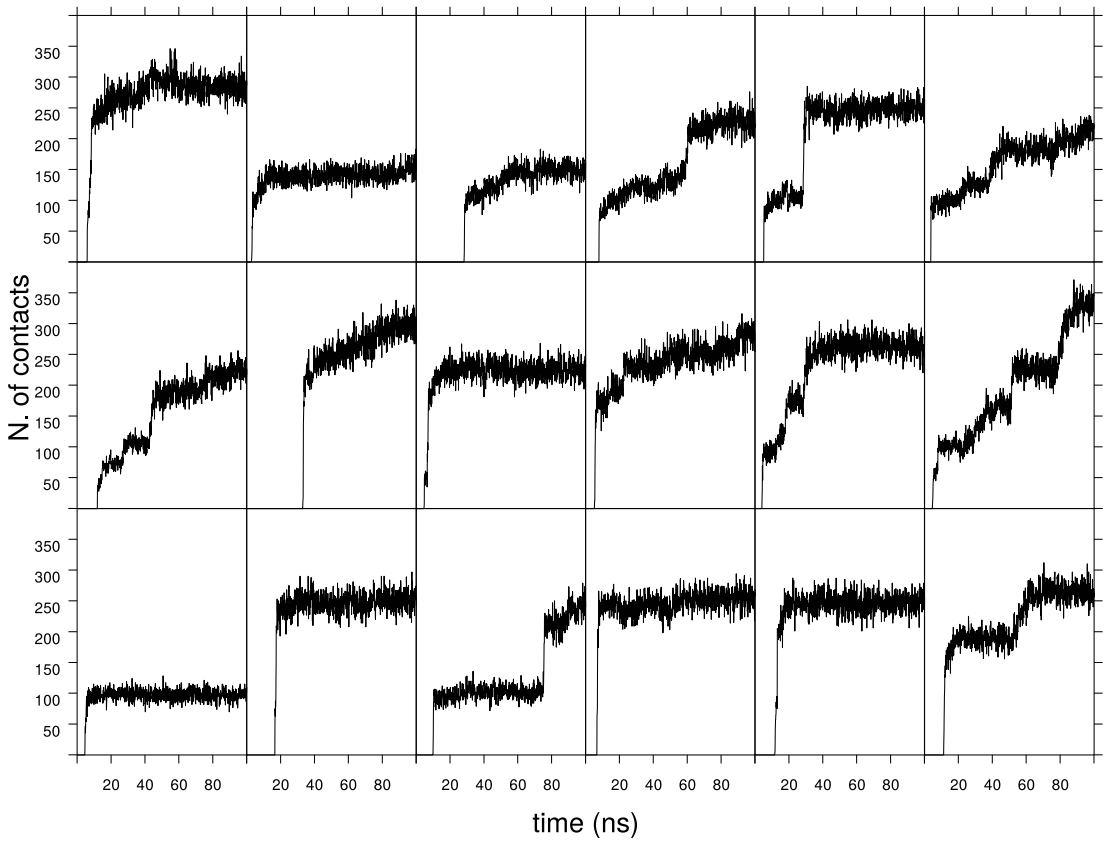


FIGURE 5.5: Number of contacts with the bounding walls versus time. Each of the 18 frames represents the contact number evolution of a single protein molecule of the considered pool. A steady increase of contacts following initial encounter is evident for the majority of the curves.

5.3.4 Adsorption leads to partial unfolding

A sudden change in structure is observed following adsorption. It is important to note that the change does not produce major changes in the expected peptide local conformation. Indeed the Ramachandran plot of the adsorbed proteins remains populated in the allowed regions, similar to the starting conformation. For this reason artifacts due to possibly overestimated interactions can be ruled out. The comparison of fluctuations for bound and unbound proteins shows that the major changes occur at the terminal strands A and G, and in the strands C,C',D, as shown in Fig. 5.6.

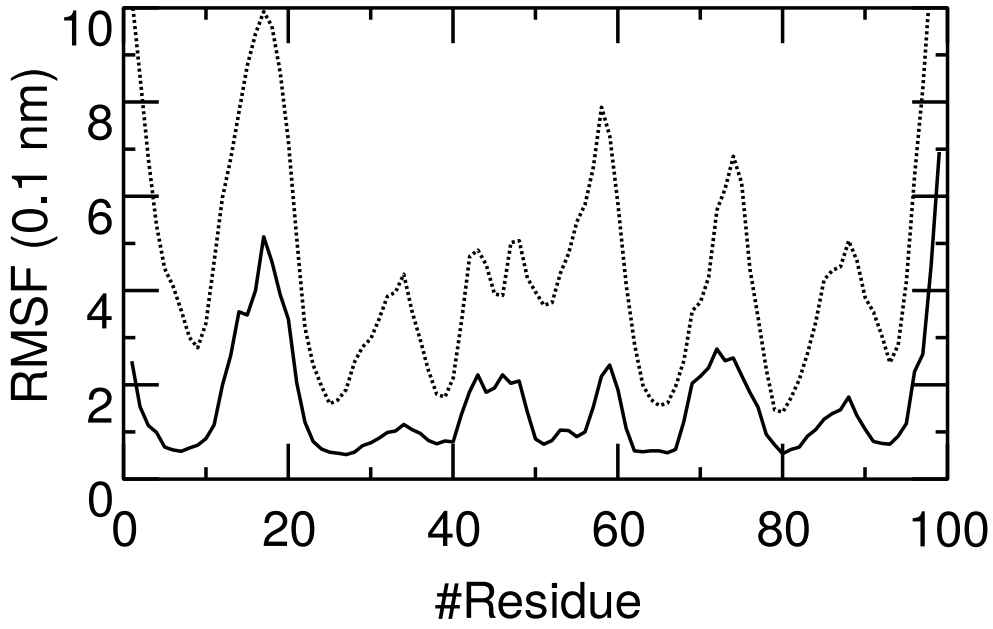


FIGURE 5.6: RMSF with respect to the crystallographic structure versus residue number for `isolated` proteins (continuous line) and wall-adsorbed proteins (broken lines). RMSF is computed on backbone atoms after optimal superposition on the starting conformation of backbone atoms of secondary structure elements. For the four `isolated` proteins four snapshots along the first 60 ns are considered, for the eighteen bound proteins only the last snapshot is considered.

Secondary structure analysis confirms and details this picture showing that the core of the protein entailing strands B and F, including disulfide bridged residues Cys 25 and Cys 80 and strands C and E remains fairly stable. The most labile secondary structure element is strand D which is little preserved in bound proteins, whereas the terminal strands A and G display tendency to detach from the immunoglobulin fold (Fig. 5.7). Also loops AB and EF show remarkable deviations from the starting structure. Conformational changes make proteins slowly increase their contacts with the bounding walls to adapt to the flat surface, as can be clearly seen in the last snapshot of the simulation (Fig. 5.11).

During adaptation to the flat surface native contacts are partially lost. We considered contact maps assigning to each contact a frequency according to the number of structures where the contact is found. In practice if a residue-residue contact is found in all molecules its frequency is 1, if it is found, e.g. in 5 out of 20 structures its frequency will be 0.25, independent of how many atoms are in contact. A plot of the contact map frequency for the four `isolated` proteins (4 equally spaced snapshots each during the trajectory) and the eighteen adsorbed proteins taken from the last snapshot is reported in Fig. 5.8. The comparison shows that in general the frequency of contacts in the adsorbed molecules is lesser. This is in particular true for the contacts involving strands A, D and G as mentioned above. This can be seen in the frequency difference plot (panel D of Fig. 5.8). The largest

FIGURE 5.7: Secondary structure versus time for the 18 chains adsorbed on the bounding walls.

relative changes summed over individual strands, loops, terminals and turn (i.e. difference in frequency divided by frequency in `isolated` proteins) are involving residues found in loops, but also in hydrogen-bonded strand-strand residue pairs. An average loss in relative frequency of contacts of about 25 % is found between strands D and E, and strands A and B.

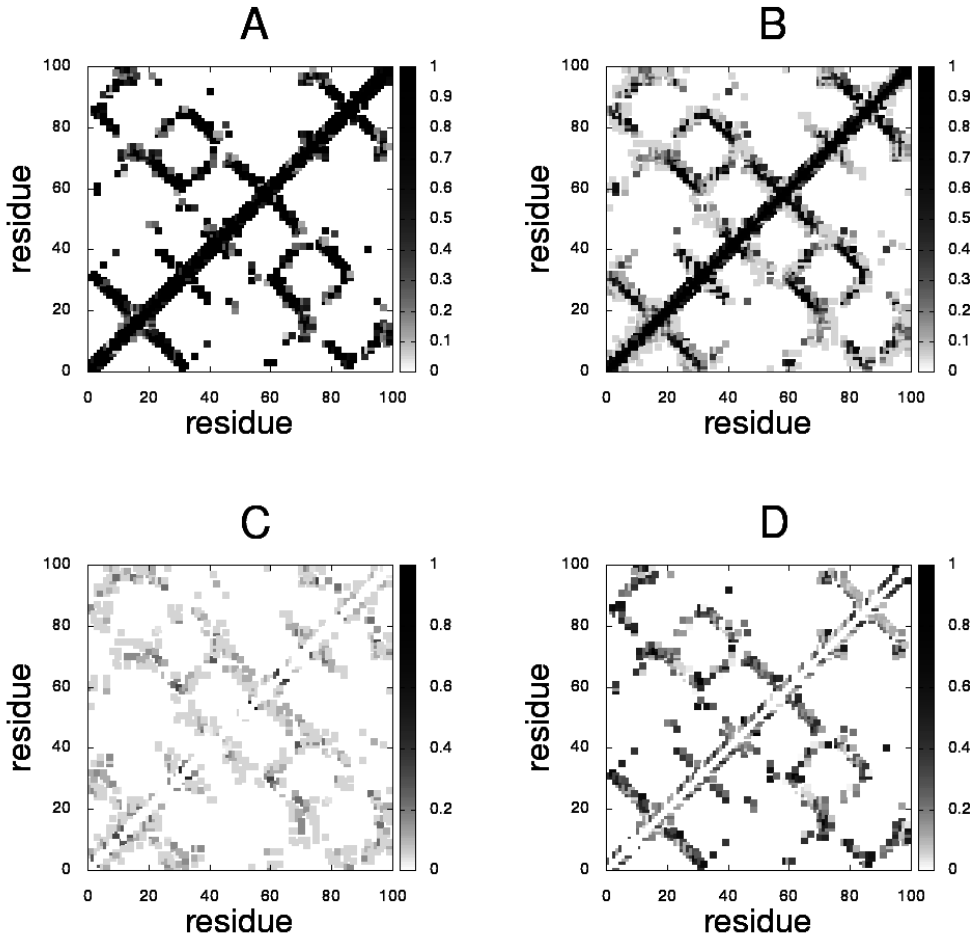


FIGURE 5.8: Residue-residue contact map frequencies for the four snapshots of each of the four `isolated` (panel A) and for the 18 adsorbed (panel B) proteins. The range of the z-axis is 0 to 1 to indicate the frequency of the contact. In the lower row the absolute value of the frequency difference between the two contact maps. Positive differences, i.e. for contacts (mostly non-native) which are more frequent in the adsorbed proteins compared to isolated ones, are displayed in panel C, whereas negative frequency differences, i.e. for native contacts which are less frequent in the adsorbed proteins, are displayed in panel D.

5.3.5 Adsorption surface characteristics

In order to better detail the interaction with the surface we computed the surface electrostatic potential at the protein surface interacting with the hydrophobic wall and at the opposite surface for the eighteen adsorbed proteins. The resulting picture (Figs. 5.9 & 5.10) confirms that there is an adaptation to the surface that in few cases results in a close to null potential at the interacting surface, strikingly contrasting the opposite polar surface. In most cases, however, the difference is smaller and less apparent to visual inspection. Hydrophobic interactions which

are formed upon protein adaptation to the surface are able to overcome electrostatic interactions, as detailed in the Thermodynamic analysis.

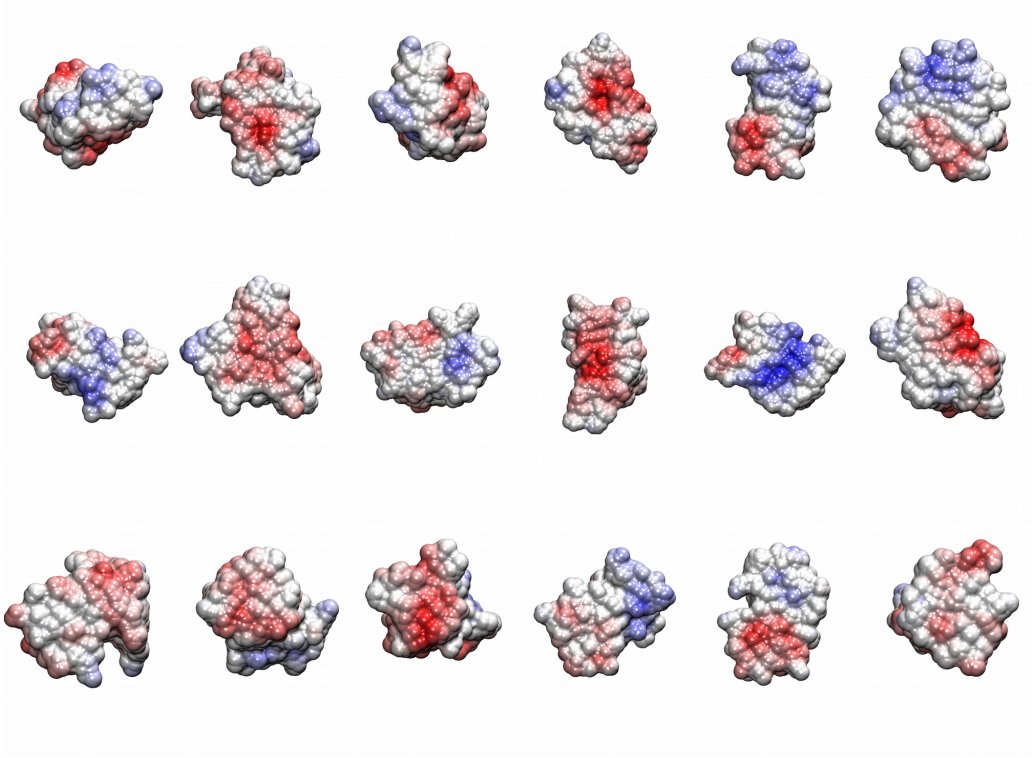


FIGURE 5.9: Electrostatic potential at the face contacting the wall.

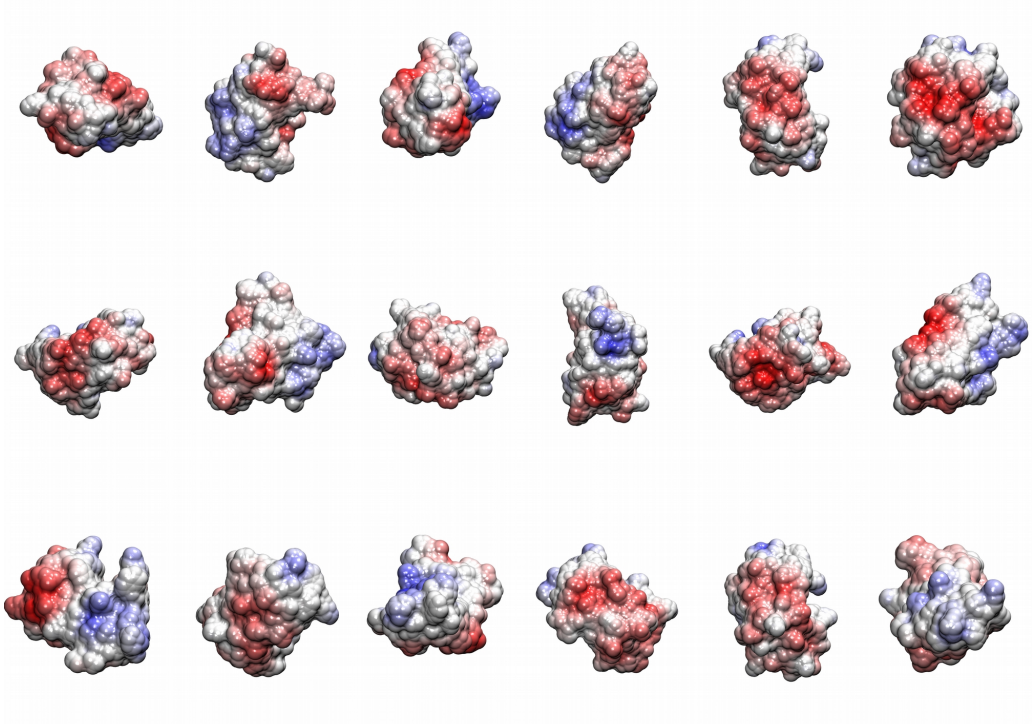


FIGURE 5.10: Electrostatic potential at the opposite face.

As a consequence of the polarity of the protein surface, before complete unfolding takes place, most contacts are made by polar and even charged residues. However, when the ratio of the number of contacts over solvent accessible surface area is considered the largest contributions are made by TRP and HIS residues, although other residues (SER, PRO, THR, ALA, GLU, ARG and ASP) display more contacts than expected based on their solvent accessible surface areas.

5.3.6 Thermodynamic analysis

The fast and irreversible adsorption of β2m on the bounding walls prompts for a thermodynamic analysis in order to understand the driving forces of the process.

β2m binding to the hydrophobic surface of the wall is accompanied with changes in

1. the bonded (bond, angle, dihedral) energy of the protein;
2. the nonbonded (Lennard-Jones and Coulombic) energy of both protein and walls (inclusive of their interactions);
3. the solvation (electrostatic and apolar) energy of both protein and walls;
4. the restraining energy of the wall;
5. the configurational entropy.

All contributions to the free energy of binding are computed as described in Free energy calculations.

Thermodynamic analysis poses the problem that the system has not reached equilibrium at the end of simulation, as mentioned above. For this reason all the following considerations must be taken with due care. To reduce the effects due to incomplete adsorption, we consider only the last 2.5 ns of the simulation, i.e. at the end of the non-completed adsorption process.

We evaluate the average of all contributions to the energy of the system containing each protein and the box for the eighteen adsorbed proteins. For free proteins we consider the only protein which is `isolated` at the end of the simulation. For the free bounding walls we consider as the unbound state the walls during the last 2.5 ns of the simulation before proteins are adsorbed. The subtraction of bound minus free average free energy, normalized for a single protein, is considered. In order to provide an error estimate the calculations were repeated at 60 and 80 ns.

The change in internal energy of the walls upon binding is estimated with reference to the equilibrated `isolated` box. It must be clear that since there are strong correlations among contributions the analysis must be taken with some care. Moreover given the limited sampling the analysis is mostly qualitative. We detail hereafter all contributions to the binding free energy in the order they are mentioned above (the complete table of contributing energy terms can be found in Table B.1, Appendix B).

Covalent terms were analysed and, as could be expected, the strong surface binding results in some distortion of protein covalent geometry leading to an average energy difference of 36±10 kcal/mol between bound and free molecules unfavorable to binding.

van der Waals interactions (computed as the Lennard-Jones interaction term) are seen to be very large and contribute most to the energy of binding by 273±11 kcal/mol favorable to binding, largely due to protein-wall interactions. The internal wall van der Waals interaction contributions to binding are negligible, whereas the internal protein van der Waals interactions are partly disrupted leading to 20±14 kcal/mol contribution unfavorable to binding.

Interaction with the wall disrupts also many electrostatic interactions in the protein resulting in 40±9 kcal/mol energy unfavorable to binding. This is accompanied however with larger interaction with solvent.

Non-polar solvation, proportional to the solvent accessible surface should favor binding, however as a result of partial unfolding further protein surface becomes accessible to the solvent resulting in a balance between the two contributions (0±3 kcal/mol). As an effect of partial unfolding the electrostatic solvation is also favoring binding by 32±5 kcal/mol, but when the sum of Coulombic and solvation terms is considered the average electrostatic contribution is unfavorable as could be expected and equal to 8 kcal/mol.

As a consequence of binding the walls' atoms are on average displaced from their equilibrium positions with a resulting average increase in the restraining energy of 53±43 kcal/mol.

The computation of protein conformational entropy was performed using the nearest neighbour method as reported in Free energy calculations. We performed two kind of analyses. First all non bound conformations were grouped together and the same was done for all bound conformations. Then, using the nearest neighbor method, the entropy for all residues and for the whole protein was computed (Singh et al., 2003; Hnizdo et al., 2007; Hnizdo et al., 2008; Fenley et al., 2014; Fogolari et al., 2015; Fogolari et al., 2016). Although flexibility seems reduced in bound with respect to free conformers, the largest variability in the way proteins are adsorbed on the box wall resulted in a larger conformational entropy reduction for the free rather than the bound set and consequently an entropic contribution of ca. 53 kcal/mol favorable to adsorption. To confirm that the interpretation was indeed correct we computed the entropy for different bound sets using single monomers separately. The results are only qualitative because the sampling on single monomers is rather poor, as checked by the dependence on the number of chosen neighbors for calculation. When single bound monomers are considered separately, the average conformational entropy of free and bound conformers is almost the same, and so it is not expected to contribute significantly to binding.

In other words on the timescale of the simulation the extent of the conformational space available to each free and each bound chain is similar with no loss or gain in conformational entropy. Binding implies rotation-translation entropy loss (with

respect to the 1 M random orientation reference state) which was also computed using the nearest neighbour method (Huggins, 2015; Fogolari et al., 2015; Fogolari et al., 2016). Due to limited sampling, we were forced to compute rotation and translation entropy separately. Neglection of correlation results in underestimation of the loss of entropy, although the expected systematic error should not affect the following considerations. With the choice of reference state and bound state described in Simulation details, the computed entropy change upon binding is on average -3.0±0.4 k_B for the translational contribution and -8.9±0.7 k_B for the rotational contribution.

If similar translational restrictions would take place in the two dimensions parallel to the adsorbing surface, e.g. arising from crowding, an additional loss of about 6.0 k_B could be estimated. Overall the rotation and translation entropic restrictions upon binding (resulting in 7 to 11 kcal/mol contribution unfavorable to binding) seem therefore comparable to the corresponding ones computed in protein-protein association (Fogolari et al., 2015; Fogolari et al., 2016).

In summary, the estimated energy favorable to binding for each adsorbed molecule is about 156 kcal/mol, i.e., upon dividing by the average number of residues contacting the walls, an average interaction energy of 4.0 kcal/mol for each interacting residue. The latter value is smaller but similar to the value of 11.1 kcal/mol estimated by Raffaini and Ganazzoli (Raffaini and Ganazzoli, 2010), using different forcefield and implicit solvent model. The entropic energy resulting from restriction in rotation and translations appears to be rather limited.

In spite of all possible inaccuracies the analysis shows that the adsorption of proteins on hydrophobic surfaces is a thermodynamic process with very large and favorable free energy, mostly determined by large van der Waals forces. The effect could obviously be enhanced by the choice of Lennard-Jones parameters for the hydrophobic surface, though the latters are in line with the parameters found in united atoms forcefields.

5.4 Summary

This chapter has focused on the extensive description of the interactions of β2m, a paradigmatic protein model for amyloidosis studies, with a model cubic hydrophobic box. In order to avoid any artifact due to adopted implicit solvent model and further to validate the tandem forcefield (Amber99sb-star-ildn)-implicit solvent used (OBC-GBSA), we carried out independently two sets of runs in explicit solvent and comparison was made to the selected chains called `isolated` which remained free i.e. unbound to the surface patches in the first 60 ns. This analysis showed that their structure was well preserved and they displayed fluctuations in the same range as those observed in explicit solvent simulations. This enabled us to confirm that the adopted implicit solvent model and forcefield is

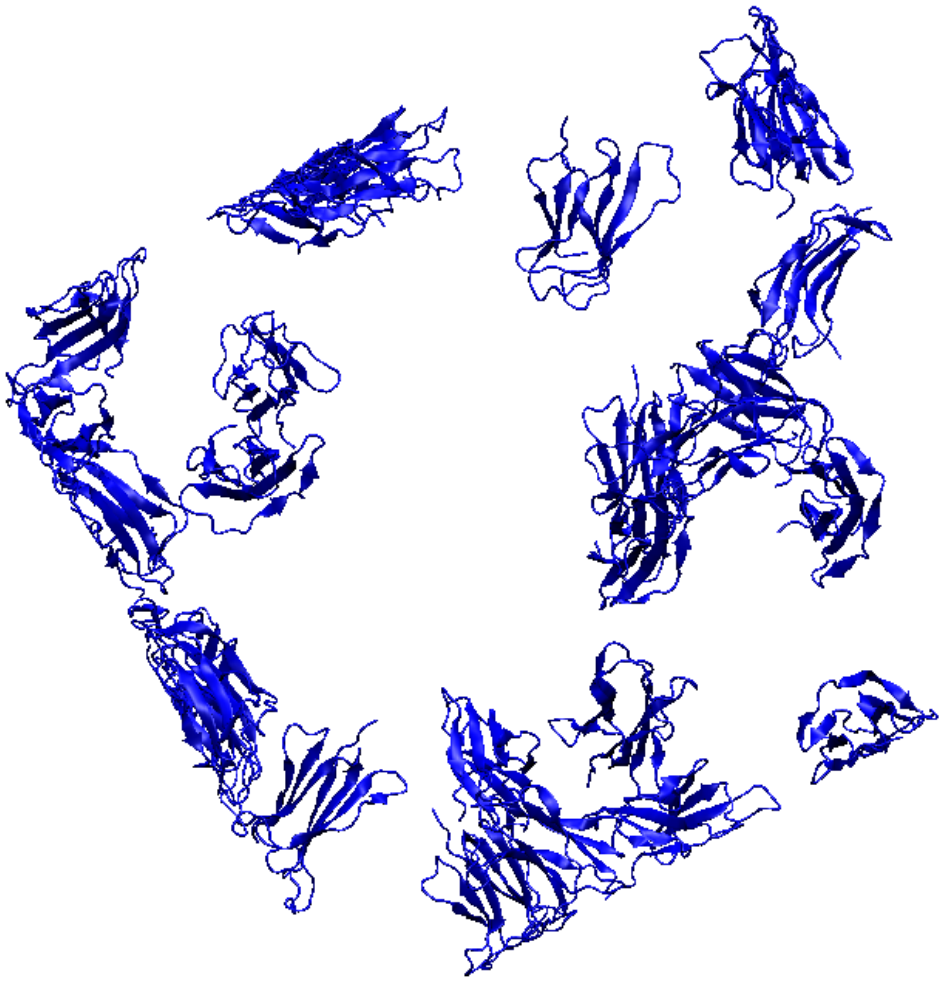


FIGURE 5.11: Snapshot of the system at 100 ns. The bounding box is not shown for the sake of clarity but it is clearly suggested by the proteins adsorbed at the interface.

adequate to describe the protein dynamics and thermodynamics. As seen experimentally with other proteins, β2m readily adsorbs on the surface on the simulated timescale. The adsorption appears to be a fast and irreversible and following adsorption the protein partly unfolds to adapt to the surface and maximize interactions. Partial unfolding is observed at the terminal strands A and G and large changes are observed also in strand D whose conformational variability has been shown both in simulations (Ma and Nussinov, 2003; Park and Saven, 2006; Fogolari et al., 2007) and in NMR and crystallographic structures (Verdone et al., 2002; Rosano, Zuccotti, and Bolognesi, 2005; Azinas et al., 2011). Also loops AB and EF display large changes with respect to the starting structure. These results imply that flexible or more labile structural domains are important for general association properties because of their ability to adapt to molecular interactions. The most encountered regions in proteins' first contacts with the walls entail residues close to the N- (ILE 1, TRP 60) and C-terminal (LYS 75 and MET 99) regions. This indicates a preference orientation for the molecular encounter with the hydrophobic surface. Furthermore, the region spatially close to

the N-terminal entails large hydrophobic residues such as TRP 60 and PHE 56. This region is contacting the walls of the box in the majority (14 out of 18) of the adsorbed proteins, substantiating its predominant role in association. No significant role of protein association is found in binding to the wall, i.e. association of proteins which is observed in the simulation because of the large concentration of molecules does not prevent nor favor association with the wall. Conversely, adsorbed proteins do not recruit other proteins for association during the simulation. Consistently, protein-protein interactions on the surface have been observed experimentally only on a longer timescale. The thermodynamic analysis shows that the protein is able to bind the surface without a large desolvation penalty because large rearrangements occur maintaining polar groups well solvated. The interaction energy appears to be very large, in line with previous similar computational and experimental studies, making the adsorption irreversible on the timescale simulated here. Besides all approximations in the modelization of the system and in the interactions, with a possible consequent overestimation of interactions, the order of magnitude of protein-hydrophobic surfaces interaction is realistic because it arises from a large number of interacting apolar groups. The main contribution to adsorption comes from van der Waals interactions, whereas other terms contribute much less in favorable or unfavorable ways to the binding. The very large interaction free energy suggests that, once adsorbed and unfolded, proteins could only favor fibril formation by acting as templates on the wall. A different scenario can be envisaged for the water-air interface which is strongly hydrophobic, but lacks van der Waals interactions. Indeed, in previous work by Bellotti *et al*. (Mangione et al., 2013) in the presence of air-water interface and agitation there was a dramatic increase in fibril formation of an amyloidogenic mutant of β2m compared to the same conditions where the interface was teflon-water, which has lower but similar surface tension. The work discussed throughout this chapter has provided a detailed picture of the magnitude of the driving forces for β2m interaction with a model hydrophobic surface which is consistent with previous experimental works.

Chapter 6

Amide hydrogen exchange in β2-microglobulin explored by MD simulations[1]

This chapter deals with the hydrogen/deteurium exchange (HDX) and aims to provide a detailed atomistic framework to the dynamics of amide hydrogens in globular proteins in general and in β2m in particular. Indeed, backbone HDX measurements are extensively used to describe the protein structures and conformational dynamics. Ideally, both NMR spectroscopy and mass spectrometry can serve as detection methods, yielding to HDX rate constant of individual amide NH site. However, such process reflects local and global dynamics which are accessed experimentally only through population averages. It is thus of fundamental importance to be able to correlate experimental data to MD averaged quantities. Moreover, in the case of β2m notwithstanding the recent advances in the topic (Rennella et al., 2012a; Rennella et al., 2012b), the full descriptive atomistic mechanism of local fluctuations is still elusive and poorly understood. A proper comparison between experimental and simulation data would therefore require modeling of the process at atomic detail resolution. Using free energy calculations we aim at reproducing also the temperature dependence of the process. Being preliminary results, the data discussed throughout this chapter are mostly qualitative. Nonetheless, they show promise since they consistently reproduce the general trend seen experimentally for different subdomains of β2m structure. Further analysis are currently been perfomed both with NMR (in relation to the fast-exchanging residues using a CLEANEX-like procedure (Hwang et al., 1997; Hwang, Zijl, and Mori, 1998)) and simulation, in order to rationally merge both macroscopic and microscopic facets and fully provide an atomistic free energy landscape description of the dynamics of amide hydrogens exchange in β2m as well as the temperature dependence of the thermodynamic parameters ($\Delta G, \Delta H, \Delta S, \Delta C_p$) accompanying the global and/or subglobal unfolding.

Keywords: Molecular dynamics, HDX, Free energy landscape, Umbrella sampling, Metadynamics.

[1] C. J. Dongmo Foumthuim; A. Corazza, G. Esposito and F. Fogolari. (2018). Amide hydrogen exchange in β2-microglobulin explored by molecular dynamics simulations, **In preparation**.

6.1 Introduction

Proteins usually adopt well defined tertiary shape (folded structure) that switches between several conformational states in solution, allowing the highly electropositive hydrogen sites bound to polar nitrogen and oxygen atoms of the buried residues to interact (exchange) with external solvent. Amongst these labile amide hydrogens, those of the backbone main chain (–C(O)–NH–) residues are significantly of interest compared to the side chain ones, because they build the hydrogen-bonded (H-bonded) network of the protein assembly[2]. Furthermore, their slower exchange kinetics makes them more easily accessible by many experiments. This makes the backbone amide hydrogen exchange (HDX) measurement one of the most widely used technique for investigating protein structure, dynamics and thermodynamics.

The behaviour of exchangeable NH groups (see Fig. 6.1) remains intricate and strongly differs upon moving from unstructured polypeptides to globular proteins. In unfolded polypeptides, the kinetics of NH-to-ND exchange i.e. the deuteration kinetics is largely dictated by the next nearest side chain groups to the amide considered. Furthermore, other factors at the primary chemical level have been proven to affect the latter measured kinetics, the most significant being the temperature, electrostatics (Fogolari et al., 1998), inductive and steric effects and the pH (Molday, Englander, and Kallen, 1972; Bai et al., 1993; Connelly et al., 1993). Bearing in mind that hydroxyde anions are less basic than amides, most exchange reactions in near-neutral pH are base-catalyzed and proceed with OD^- catalyst via the imidate ion intermediate R_1-C(O^-)=N-R_2 that afterwards interacts with the excess D_2O to form R_1-CO-ND-R_2 (see complete mechanism in Fig. 6.2).

The deuteuration rate therefore relies on the stability of the transient imidate intermediate, more it is stable more faster is the exchange. This is seen for positive charge density in the vicinity of the amide groups (by lowering the pKa) while the opposite is expected for negative charge groups (Molday, Englander, and Kallen, 1972; Fogolari et al., 1998; Abdolvahabi et al., 2014). This underlines the importance of inductive effects and to a lesser extent of steric factors (Bai et al., 1993) in the exhange kinetics. Nonetheless, it should be recorded that a successful NH-to-ND event arises upon diffusional collisions between the amide proton and the catalyst about hundreds of times, on average (Englander and Kallenbach, 1983).

The above mentioned factors have been accurately fitted into empirical rules that govern the observed or experimental exchange rate, k_{exp}, for a structureless small peptide in solution, based on the second-order rate constant they rely upon under specific conditions, also termed intrinsic rate constants, k_{int}, as :

$$k_{exp} = k_{int} \times [cat] = k_{int} \times [OD^-] \tag{6.1}$$

where $[cat] = [OD^-]$ is the concentration of the catalyst, in this case the base. This overall picture apparently simplistic for unstructured polypeptides is rather different for globular proteins.

[2]Except of proline residue which doesn't contain an NH group.

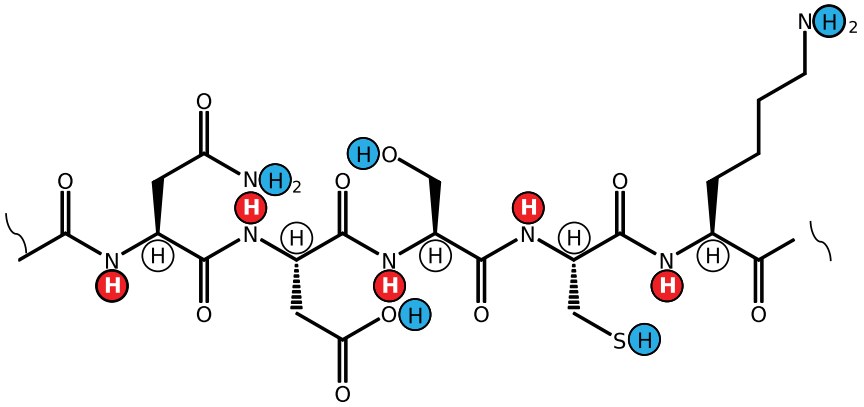


FIGURE 6.1: Schematic view of different amide hydrogens in an unfolded polypeptide chain with sequence Gly-Asn-Asp-Ser-Cys-Lys. Exchangeable amide hydrogens comprise both backbone main chain (in red) and side chain (in blue). Main chain hydrogens bound to carbon atoms (in white) are apolar and basically are not exchangeable. This picture is taken from (Jensen and Rand, 2016).

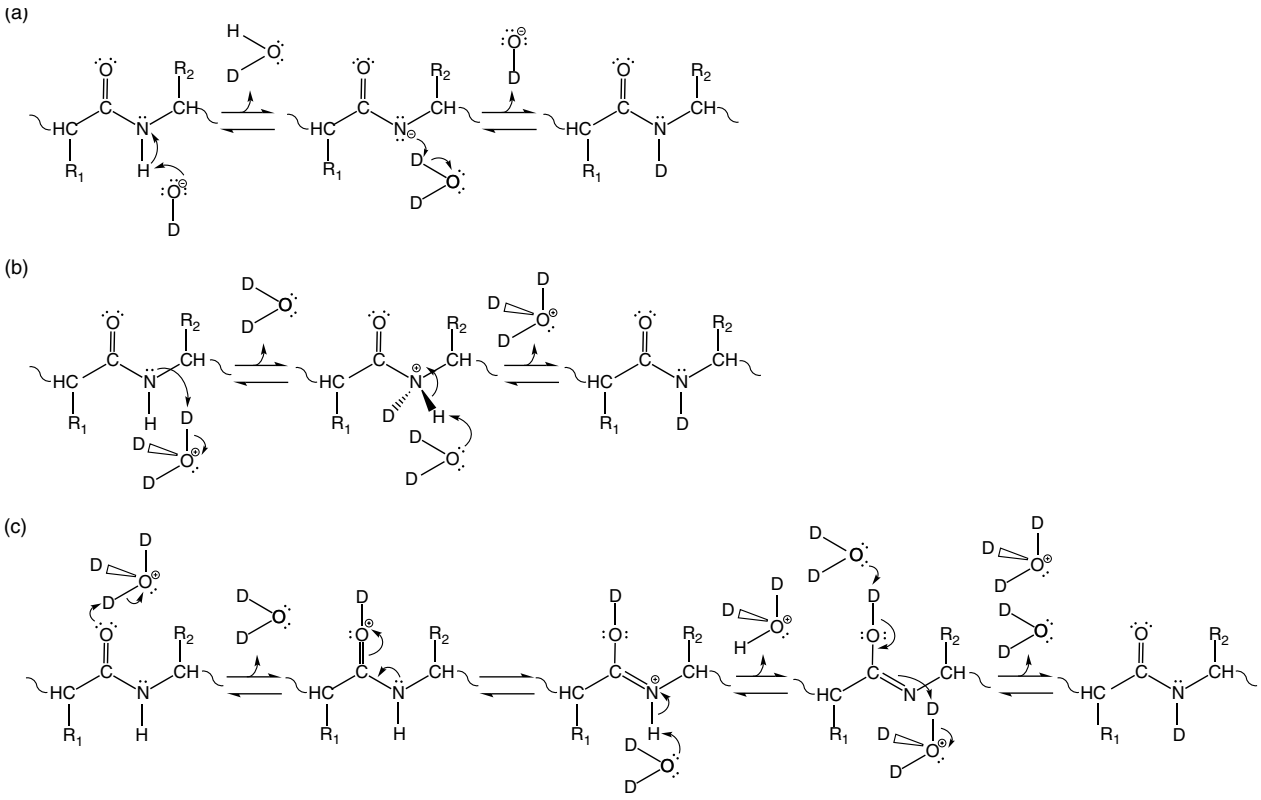


FIGURE 6.2: Possible reaction mechanims encountered in backbone amide hydrogen experiments. The most recurrent base-catalyzed process is shown in (a), while (b) and (c) reactions highlight the acid-catalyzed processes through N- and O-protonation, respectively. These schemes are taken from (Jensen and Rand, 2016).

Indeed, the dynamics of amide protons is significantly restrained in folded moieties thereby lowering the exchange kinetics. Moreover, the factors inducing the

protection of amide protons in folded regions are not well known. However, a number of them have been pointed out including the H-bonded patterns, the solvent accessibility, the crystallographically defined waters, and so forth, albeit a full rationale is far to be understood.

In many experiments, accessibility of amide protons to external solvent has been pinpointed as a key requirement to HDX exchange (Yan et al., 2007; Shan et al., 2013). This idea is rather under-weighted in front of the view that H-bonded networks essentially dictate the protection of amide protons, and primarily through main chain backbone NH$\cdots$OC contacts (Pan and Briggs, 1992; Skinner et al., 2012b). More intriguing is the case of H-bonded amides exhibiting strong pattern of protection even being at the protein surface (Skinner et al., 2012b). This state of facts rationalizes with the plastic nature of the protein assembly, in which, owing to conformational fluctuations (sometimes by breaking the H-bonds), the core of the protein, entailing the most buried and H-bonded amides, might undergo transient openings, exposing them to solvent, thereby allowing the HDX exchange.

It comes therefore that, HDX exchange in globular proteins relies on the cleavage of individual H-bonds and one assumes that the backbone amide protons (HN) exist either as closed (buried) or opened (exposed) state conformations. Closed states are likely those involved in intra-molecular H-bonds and significantly far from the aqueous solvent to participate in exchange (out-exchange). However, in a folded state conformation, each amide is involved in a dynamical equilibrium between a closed and an opened state (exchangeable-like state), from which exchange takes place at the constant rate k_{int} similarly as in structureless peptides, according to the scheme in Eq. 6.2 of Linderstrøm-Lang (Hvidt and Linderstrøm-Lang, 1954; Hvidt and Nielsen, 1966) :

$$(N\text{-}H)_{cl} \underset{k_{cl}}{\overset{k_{op}}{\rightleftarrows}} (N\text{-}H)_{op} \overset{k_{int}}{\rightarrow} (N\text{-}D)_{op} \tag{6.2}$$

where k_{cl}, k_{op} and k_{int} are the rates of closing, opening and the intrinsic or chemical exchange rate for individual amide site, respectively. Let's remind at any useful end that, not all structural distortions might expose the core of the protein to external solvent. This sometimes requires substantial dynamic structural excursion. Most HDX exchange relevant distortions can range from local fluctuations all the way to global unfolding through subglobal openings (Englander, Mayne, and Krishna, 2007).

The previous considerations chiefly account for the classic HDX exchange formalism. However, in general, there are atypical cases that are scarcely justified or even more, fall out of this model. This is seen for amide protons that are neither buried nor H-bonded, but have very slow exchange rates (Li and Woodward, 1999; Anderson, Hernández, and LeMaster, 2008; Skinner et al., 2012b). This troublesome issue has been rationalized by considering electrostatic effects that induce additional protection to these amides by modifying their pKa (Bai et al., 1993; Anderson, Hernández, and LeMaster, 2013). Albeit this pertinent

view goes beyond the primary structure effects usually taken into account, it is far from being comprehensive. Indeed, side hypothesis such as long-lived H-bonding to solvation waters, identified with crystallographic water molecules have been thought pertinent to justify the extent of protection of exposed amide sites (Skinner et al., 2012a; Skinner et al., 2012b).

The previous highlighted factors rationally provide explanations for different amide hydrogens encountered in globular proteins, without being exhaustive. This shows how intricate is the analysis of amide hydrogen exchange data (McAllister and Konermann, 2015). Furthermore, the traditional scheme of Linderstrøm-Lang is unable to unambiguously define the open and close state conformations. Also, it is worth noting that the study of HDX exchange has been merely achieved considering the NMR or X-ray structures (Pan and Briggs, 1992; Skinner et al., 2012a; Pirrone, Iacob, and Engen, 2015). This simply implies the static frames of the conformational ensemble which do not adequately sample the more rare thermally activated ensemble of the protein structure. The latter fact partly accounts for all the troublesome issues linked to the explanation of HDX exhange rates. Fortunately, these limitations can be circumvented by using computational approaches incorporating the dynamic nature of the protein and its surrounding components like MD computations (Sheinerman and Brooks, 1998; García and Hummer, 1999; Vendruscolo et al., 2003; Gsponer et al., 2006; Kieseritzky, Morra, and Knapp, 2006; Craig et al., 2011; Petruk et al., 2013; Hsu et al., 2013; Skinner et al., 2014; Xu et al., 2015; Persson and Halle, 2015). However, exchange rates are gathered experimentally on timescales much more longer than those achievable in many atomistic conventional MD simulations. Nonetheless, even on shorter lenghtscale, the resolution of most MD simulations should provide a general picture of the phenomenon of interest, with a much higher accuracy than static NMR or X-ray snapshots.

It was said previously that structurally protected amides exchange primarily via open-competent states. According to the scheme in Fig. 6.2, such a state is only transiently formed, thus very short-lived in principle. Therefore, it appears that the experimentally measured rate of exhange (k_{exp}) for individual amide NH depends on the rates of closing (k_{cl}), opening (k_{op}) and the intrinsic rate (k_{int}) given by Eq. 6.3 under steady-state conditions ($k_{cl} \gg k_{op}$ and $k_{cl} \gg k_{int}$) satisfied by most experiments as :

$$k_{exp} = \frac{k_{op} \times k_{int}}{k_{op} + k_{cl} + k_{int}} \tag{6.3}$$

Two assumptions are usually made to keep the conformational part of Eq. 6.3 separated from the intrinsic one, leading to the so-termed EX1 (monomolecular exchange) and EX2 (bimolecular exchange) regimes in reference to the pH value and conformational dynamics. Eq. 6.3 can be re-writen as follows :

$$\frac{k_{exp}}{k_{int}} = \frac{k_{op}}{k_{op} + k_{cl} + k_{int}}$$

$$\Rightarrow PF = \frac{k_{int}}{k_{exp}} = 1 + \frac{k_{int} + k_{cl}}{k_{op}} \tag{6.4}$$

In Eq. 6.4 PF is the protection factor that is an estimation of the decrease in the exchange rate of a residue in the protein structure in relation to its chemical or intrinsic rate in an unfolded state. In destabilizing conditions (high pH and high temperature), i.e. in the EX1 limits, with $k_{int} \gg k_{cl}$, Eq.6.4 becomes $PF = 1 + k_{int}/k_{op}$. Meanwhile in stabilizing conditions (low temperature and pH) i.e. in the EX2 regime with $k_{cl} \gg k_{int}$, Eq. 6.4 takes the form $PF = 1 + k_{cl}/k_{op}$. These equations for EX1 and EX2 can be further reduced respectively to $k_{exp} = k_{op}$ (provides kinetic information) and $PF = k_{cl}/k_{op} = K_{op}^{-1}$ in the case the system is sampling relatively stable conformations and $k_{cl} \gg k_{op}$. In the latter equation K_{op} is the thermodynamic equilibrium constant for the [close]→[open] equilibrium, i.e. the unfolding transition equilibrium constant required for exchange to proceed. The latter allows for the determination of the exchange free energy ΔG_{exp} under EX2 limits as in Eq. 6.5 :

$$\Delta G_{exp} = -RTlogK_{op} = -RTlog\left(\frac{k_{op}}{k_{cl}}\right) = RTlogPF \tag{6.5}$$

In the present work we would like to provide an atomistic framework to the dynamic of HDX in amyloidogenic proteins in general and β2m in particular for which the detailed mechanism remains elusive and poorly described. Open states were considered as snapshots having in the first coordination shell (r_{cut}= 2.76 and 3 Å) at least 2 interacting water molecules (see Fig. 6.8). This cut-off criterion, taken here as the distance between the amide proton and the water oxygen atom, nearly corresponds to the first minimum in the H(NH)$\cdots$O(H_2O) radial distribution function $\rho g(r)$ computed from the MD trajectory (Fig. 6.6) and closely come toward the optimal H-bond formation distance. The results show that, out of 93 protein residues considered, almost one half fulfilled the opening condition and were considered to exhange during the simulation. These so-considered open states have an average number of water of 2.02 and show for all of them an intense peak around 2 Å showing that water is well distributed around the first hydration shell of these amide backbone residues.

6.2 Simulation details

6.2.1 Metadynamic calculations

The calculation was done with plumed-2.1.1 plugin (Bonomi et al., 2009; Tribello et al., 2014) patched to gromacs-5.0.4 simulation code (Abraham et al., 2014). The molecular mechanics amber99sb-ildn (Lindorff-Larsen et al., 2010) force field accounted for the characterization of molecular interactions. The protein atoms were spatially placed at the center of a cubic box, at least at 10 Å from the box edges, and solvated by 10778 TIP3P water molecules (Jorgensen et al., 1983). One

Na^+ counterion was added to replace the equivalent number of water molecules and thus balance the overall excess charge of the system. PBC were applied in the calculations, mimicking a bulk system and thereby avoiding the finite size effect.

The system was initially minimized using the steepest descent minimization algorithm, with a minimization step size of 0.1 nm and a maximum convergence force of 1000.0 $kJmol^{-1}nm^{-1}$. The equilibration phase was done in 2 steps; 100 ps in *NVT* ensemble followed by 100 ps in *NPT* ensemble. During the first equilibration stage, the leap-frog integrator with integration timestep of 2 fs was used to update the changes in the system. Particle Mesh Ewald summation (Darden, York, and Pedersen, 1993; Essmann et al., 1995) accounted for long-range electrostatics interactions. The temperature was equilibrated to a reference value of 300 K using the velocity rescaling (modified Berendsen thermostat)(Bussi, Donadio, and Parrinello, 2007), with a coupling constant of 0.1 ps. Short-range electrostatics and van der Waals interactions were truncated with a 10 Å cutoff. All bonds were constrained with the LINCS algorithm (Hess et al., 1997). In *NPT* equilibration stage, using the previous parameters, the pressure was stabilized to 1.0 bar using the Parrinello-Rahman pressure coupling (Parrinello and Rahman, 1981; Nosé and Klein, 1983) with a coupling constant of 2.0 ps. Snapshots were collected every 10 ps along the trajectory.

Metadynamic calculation was perfomed using the global RMSD as collective variable (CV) with an optimal alignment type on the minimized reference X-ray structure. Gaussians with height of 1.2 kJ mol^{-1} and width of 0.30 nm were added every 500 time steps. Restraints on lower and upper walls were applied to the chosen CV to circumscribe the allowed sampled regions in the CV space. The value of the CV along with the metadynamics bias potential were printed out every 10 steps.

6.2.2 Umbrella sampling calculations

Starting from an equilibrated structure following the same protocol as described above, two calculations were performed, constant force pulling followed by umbrella sampling simulations. In the pulling calculations, the reaction coordinate chosen was the geometrical distance (all over the 3-components x, y and z) between the α-carbons of hydrogen-bonded residues Cys25 and Tyr66, i.e. the most stable hydrogen-bonded pair at the core of β-strands B and E, one the most buried part of the system. The simulation starts by pulling over ca. 0.8 nm [0.5-1.3 nm] along the reaction coordinate, using a reasonably low pulling rate, 1.2×10^{-3} nm ps^{-1} and linear potential force constant of 1000 kJ mol^{-1} nm^{-2}, in order to avoid a brute distortion of the system. Indeed large deviation from the equilibrated structure may lead to a considerable refolding-back time to the (near)equilibrated conformation. Nine conformations along the pulling trajectory were extracted (see Fig. C.1, Appendix C), nearly separated by 0.08 nm from each other and used as the initial structures for each umbrella sampling window.

Each umbrella sampling window was subsequently equilibrated in the canonical pressure ensemble *NPT* for 2 ns with the pressure held steady around 1.0 bar

using the Parrinello-Rahman pressure coupling (Parrinello and Rahman, 1981; Nosé and Klein, 1983) and a coupling constant of 1.0 ps. The temperature was already kept at 301 K using the Nose-Hoover thermostat (Nosé and Klein, 1983; Hoover, 1985), with a coupling constant of 0.5 ps. Particle Mesh Ewald summation (Darden, York, and Pedersen, 1993; Essmann et al., 1995) accounted for long-range electrostatics interactions, whereas short-range electrostatics and van der Waals interactions were truncated with a 14 Å cutoff. All bonds involving hydrogens were constrained with the LINCS alogorithm (Hess et al., 1997). The harmonic biasing force constant of 1000 kJ mol^{-1} nm^{-2} was imposed to the umbrella sampling calculations in each window. Afterwards, the proper dynamic followed in each of the equilibrated sampling window, using a time step of 2 fs and unconstraining previously restrained protein coordinates. The simulations last 50 ns, given a cumulative simulated time of 450 ns. The same procedure and parameters, as well as initial sampling windows were repeated at 305 K, 310 K and 315 K.

6.2.3 Trajectory analysis

The analysis of MD trajectory was done using available structural tools in Gromacs. The H(NH)$\cdots$O(H_2O) pair correlation function $\rho g(r)$ was obtained setting *r* as the distance between the backbone amide proton and the water oxygen atom. The number of water in the first solvation shell was integrated up to the cut-off r_{cut} of 2.7 and 3 Å from the corresponding $\rho g(r)$ plot, which closely matches its first minimum. No geometrical angular constraint was applied. Each individual residue in each frame (5717 snapshots in all for metadynamic simulation and 5000 for each umbrella sampling window) was analysed and average values were computed. The solvent accessible surface area for each individual residue was computed using a probe radius of 1.4 Å and averaged over the whole simulation trajectory.

The umbrella sampling simulations were analyzed using the whole trajectory of each window. The potential of mean force (PMF) profiles (see Fig. 6.9) were constructed using the weighted histogram analysis method (WHAM) (Kumar et al., 1992; Souaille and Roux, 2001) assuming sufficient overlap over the sampling windows (see Fig. C.2, Appendix C). The bootstrapping procedure implemented in the `g_wham` tool enables the error estimates (Hub, Groot, and Spoel, 2010). The block averaging analysis was also employed to estimate the statistical errors dividing the data pools in blocks and computing the mean and standard deviations of each block assuming they are sufficiently uncorrelated.

The protection factors shown in Fig. 6.8 and in tables C.1 & C.2 of Appendix C were computed considering the amide's open state conformations as those having within 2.76 Å and 3 Å in the H(NH)$\cdots$O(H_2O) radial distribution functions (see Figs. 6.6 & C.5 in Appendix C) at least 2 coordinating water molecules in the primary solvation shell. The protection factors were then estimated similarly as in Eq. 6.5 as the ratio $log\left(\frac{NF_{cl}}{NF_{op}}\right)$, where NF_{cl} and NF_{op} are repectively the number of frames in the close (buried) and open (exposed) state conformations and

assuming a bimolecular exchange mechanism, i.e. EX2 process. The same equation was employed by (Persson and Halle, 2015). The experimental protection factors were compiled from (Rennella et al., 2012a). All the proline together with ILE1 residues were excluded from the pool of backbone amides considered for analysis and discussion, leading thus to 93 amide residues.

6.3 Results and discussion

The stability of the global simulated system in the metadynamic calculation was first checked by computing the all-backbone-atoms root mean square deviation (RMSD) in comparison to the reference structure in the last 114 ns, see Fig. 6.3.

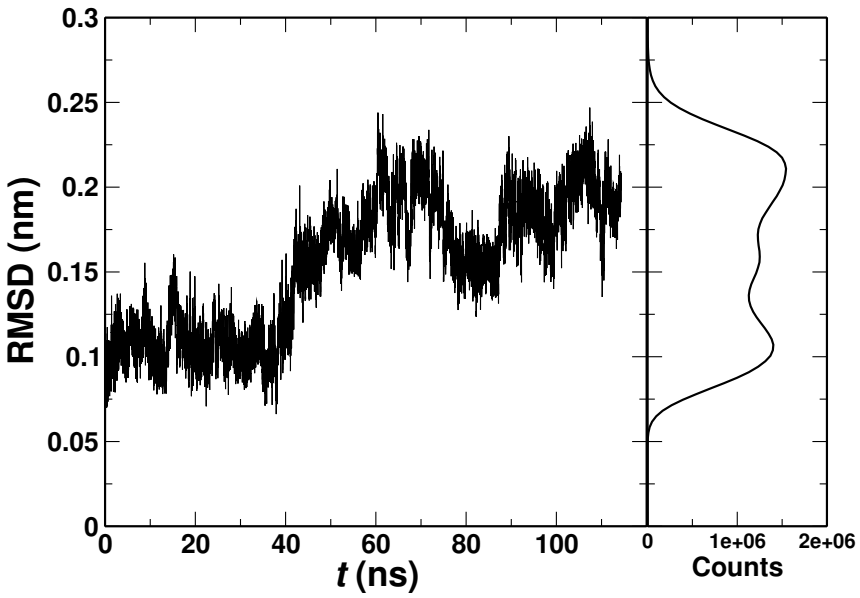


FIGURE 6.3: Time-based backbone root mean square deviation computed from the metadynamic trajectory in reference to the initial structure (left pannel) and the corresponding distribution (right pannel).

Even though the system is gradually forced to sample more region in the conformational space by the metadynamic bias potential, with RMSD values roughly distributed between 0.5 and 2.5 Å, the overall system results to remain however quite stable (folded). This picture is consistent with the high stability of β2m structure at near neutral pH conditions (Myers et al., 2006). Similar analysis (not shown here) on a single classical MD trajectory (50 ns) showed that the system was nearly fluctuating around 1.25 Å, thereby justifying why metadynamic simulation was used here. Thus, on the timescale simulated here the global unfolding-refolding structural transitions were not observed. Furthermore, residues mobility as depicted by atomic root mean square fluctuations indicates that, and as expected, large mobility arises from unstructured loop regions and terminal ends, with about 40 % more structural fluctuations (compared to the rest of the protein structure) at the C-terminal, entailing resdidues ARG97, ASP98 and MET99.

Some structural properties were also analyzed in umbrella sampling simulations in order to better perceive the different factors correlated to, or inducing the protection of backbone amide protons in β2m. At this stage of work progress, only the results of one sampling window (the one of the native conformation) will be discussed. As mentioned above, no global unfolding/refolding transition was seen also here. However, local parts of the protein structure occasionally undergo opening reactions, Fig. 6.4. Albeit the picture is not describing a deeply buried or a β-stranded region of the structure (longer simulation timescale or enhanced-induced fluctuations would be required) it exemplifies the idea as detailed in the scheme of Eq. 6.2.

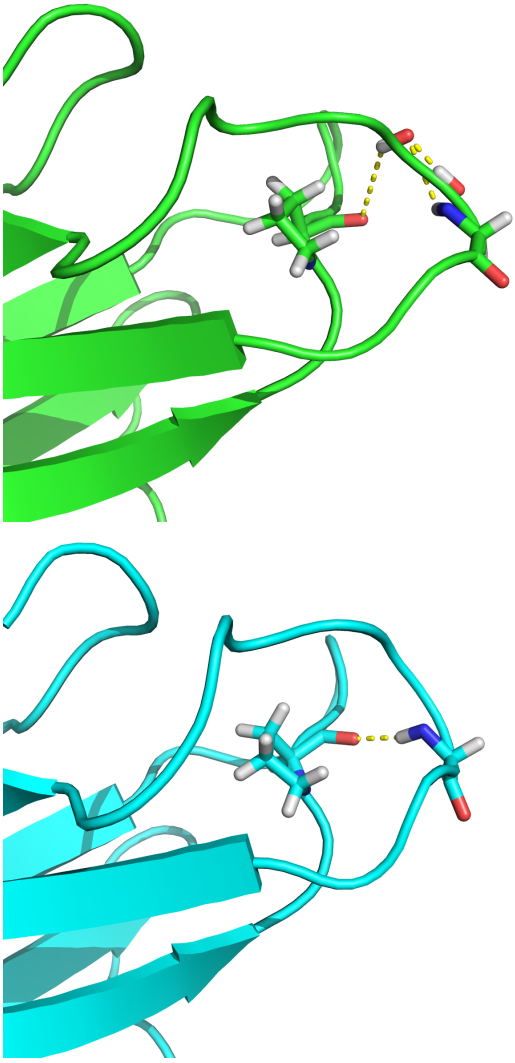


FIGURE 6.4: Overview of system transition between open (solvent exposed) in green and close (buried) in cyan conformational states. Stick representation distinguishes the amide protons of interest. At the top, the backbone NH of Gly18 is not involved in intramolecular hydrogen bonding but is instead coordinated to an external water solvent at 9 ns. At the bottom, the main-chain hydrogen bond with the carbonyl of Pro72 is re-established 4 ns later, prohibating the contact with external solvent.

Note however that, even being a priori from the less protected part of the structure, such opening/closing transitions are not recurrent throughout the trajectory. For instance, the above mentioned transition happened once and lasted more than 3 ns. A large set of opening events lead to very short life transiently formed species in general, lasting less than 120 ps on average, at least for what concerns this particular case and consistent with what was previously noticed by Persson and Halle for BPTI (Persson and Halle, 2015) (see Figs. C.3 & C.4 in Appendix C).

Atomic scale resolution provides more reliable indicators for the understanding of backbone amide hydrogens exchange. This is true as far as amide protons within the same structure displayed variegated behaviours, and each case seems to be particular (some already highlighted in Figs. C.3 & C.4). Furthermore, they do not always undergo cooperative opening/unfolding reactions since a residue can be highly protected whereas its nearest neighbours are not. This is seen for many residues at the limit of structured domains like for end-stranded residues. We therefore scrutinized various structural properties shown in Fig. 6.5, which displays the averages along the simulation time of the distance between individual amide site and its nearest carbonyl acceptor in the main chain, the NH solvent accessible surface area (SASA) and the NH···OC hydrogen bonding occupancies.

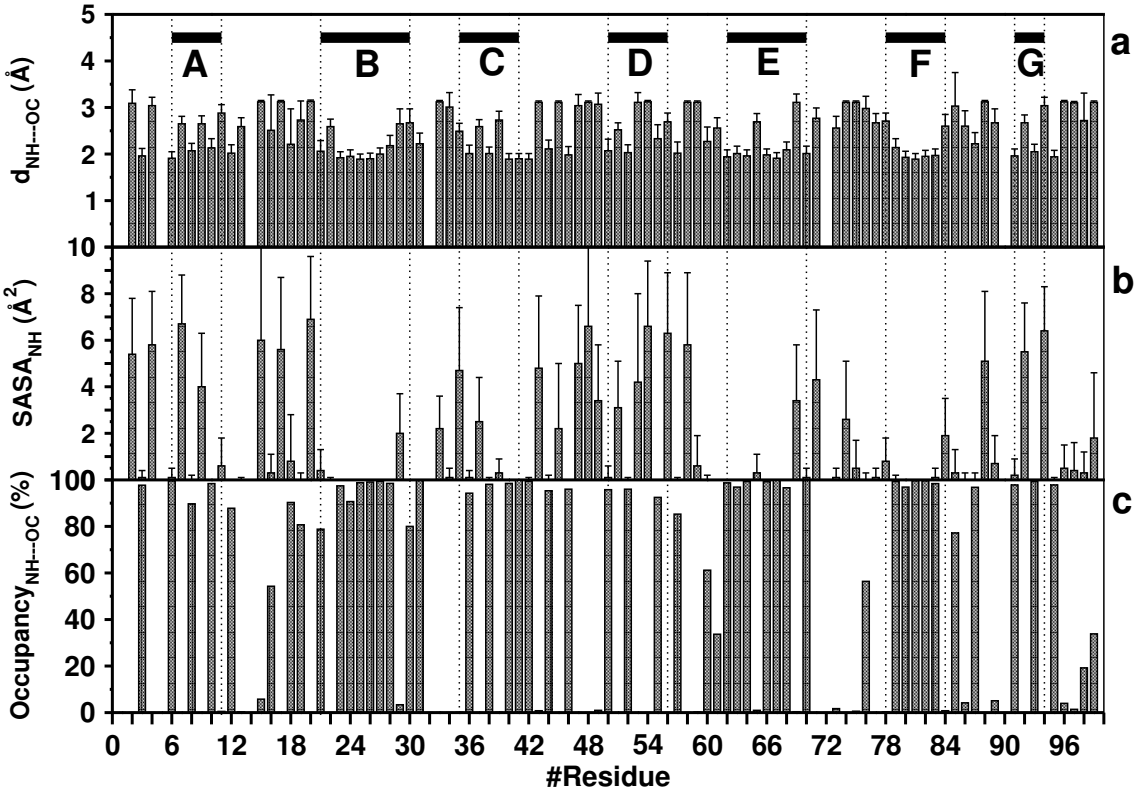


FIGURE 6.5: Structural properties analysis data averaged over the simulation time. (a) Distance between the amide hydrogen NH of each residue with its closest possible main-chain CO acceptor. (b) Solvent exposed accessible area for each NH moiety. (c) Hydrogen bonds occupancy of each main-chain NH···OC. Error bars are standard deviation and letters identify the stranded region of the protein structure.

The subsets in Fig. 6.5 that display some averaged structural properties along the simulation trajectory exhibit great traits of consistency. In fact, in Fig. 6.5-(a) the distance between each backbone amide and its closest possible carbonyl acceptor is plotted. Hydrogen bonds were considered to exist when the distance between the backbone amide proton and its nearest possible acceptor $d_{\mathrm{NH\,...OC}} \leq 0.25$ nm (Torshin, Weber, and Harrison, 2002) and the angle $\alpha_{\mathrm{H\text{-}N\text{-}O}} \leq 30°$. All the amides that fulfill this condition, in this case 43 amides, are mostly sampling the closed state conformation (see Figs. 6.7 & C.5 in Appendix C). This is the case for most β-stranded domains of β2m like β-strands B (ASN21, LEU23, ASN24, CYS25, TYR26, VAL27, SER28), E (PHE62, TYR63, LEU64, TYR66, TYR67, THR68, PHE70) and F (ALA79, CYS80, ARG81, VAL82, ASN83), the latters of which mostly exhibit smallest solvent accessible surface areas, 0.16 Å^2 on average for the 43 amides in that pool (Fig. 6.5-(b)), and high H-bond occupancies (Fig.6.5-(c)).

This confirms the great tendency that closed state conformations are mostly involved in intramolecular hydrogen bonds and likely have small SASA. However, the picture is not always simple as implied by the last case, since there are instances in which surface amides, even free from hydrogen bonds exhibit strong pattern of protection as discussed at the introduction. They are fortunately not seen in the current trajectory. Nonetheless, some amides exhibit less common behaviour, even though one of the 2 structural indicators used here could rationally explain their (de)protection. For instance, ILE35 (starting of β-strand E) which is H-bonded in principle ($d_{\mathrm{NH\,...OC}}$ 2.49 Å), displays high SASA (4.7 Å^2) and appears to be a fast exchanging residue (see Fig. 6.8). In addition, ASP76 and GLU77 (E-F loop) are strongly protected, which in turn is justified by their small SASA ($\leq$ 0.1 Å^2), but is not fully explained if considering that their main chain distances with the nearest carbonyl ($d_{\mathrm{NH\,...OC}}$ 2.98 & 2.67 Å respectively) imply no intramolecular H-bonds.

To further illustrate the miscellaneous behaviour of amide protons previously underlined, in Fig. 6.5-(c), the amide proton of ARG3 is establishing a H-bond with both carbonyl (most of the time, 97.7%) and amino groups (often, 18.1 %) of HIS31. The backbone NH of LYS6 is bounded to the CO of SER28 most of the time 99.5% and sometimes with the amino group of the same residue 33.6%, a fraction of interaction is also observed with the amine of GLY29 0.2%. LYS58 is not bounded to a main-chain carbonyl instead is often linked to the amine of TRP60 67.5%. TRP60 is both connected to CO of residues LYS57 (61.2%) and LYS58 (5.8 %). The same is seen for THR73 both interacting with CO of THR71 (1.7%) and ASP76 (0.3%). All these facts are perfectly matching what is observed upon probing the nearest solvation environment of the later backbone amides (see Fig. 6.7 and C.5 for details).

The ambiguous definition of open and close states in HDX remains perhaps the most troublesome issue and the core limitation of the Linderstrøm-Lang theory. In order to estimate the individual backbone amide protection factor from the umbrella sampling simulations, open states were considered as snapshots having in the first coordination shell (r_{cut}= 2.76 and 3 Å) at least 2 interacting water molecules (see Fig. 6.8). The same open-state definition was employed

by Persson and Halle (Persson and Halle, 2015) albeit different cut-off are used here. The cut-off criterion taken here as the distance between the amide proton and the water oxygen atom nearly corresponds to the first minimum in the H(NH)$\cdots$O(H_2O) radial distribution function $\rho g(r)$ computed from the MD trajectory (Fig. 6.6) and closely come toward the optimal H-bond formation distance.

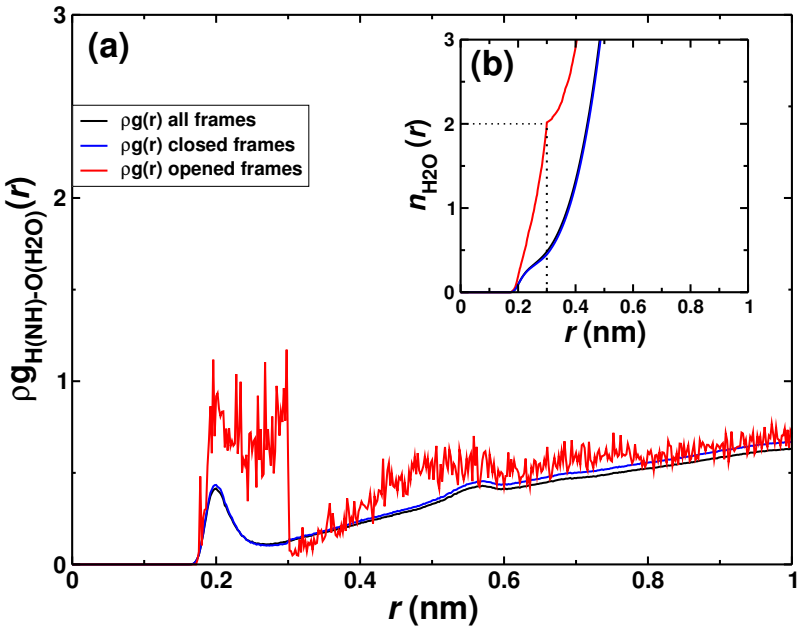


FIGURE 6.6: H(NH)$\cdots$O(H_2O) radial pair distribution functions in **(a)** and the average number of interacting water molecules in the first coordination shell (cut-off distance 3 Å) in **(b)**. Only metadynamic trajectory is displayed.

Overall, the global trajectory samples mostly the close states, consistent with what was already noted, namely that, no global unfolding/refolding event was recorded along the trajectory in the simulated timescale here.

The direct interacting number of water (Fig. 6.7) was then integrated from the corresponding $\rho g(r)$ plot up to the cut-off distance. It results that 45 residues (see Figs. 6.7 & C.5) sample the open state, i.e. have at least in one frame two coordinating water molecules in the primary solvation shell and within the cut-off distance imposed here. On average, the number of water in the first solvation sphere for the exchange-competent states is $n_{\mathrm{H2O}} \approx 2.02$ (Fig. 6.6-(b)). Meanwhile, the mean average of n_{H2O} is about 0.45 (Fig. 6.6-(b)) for the 93 residues sampling the close states (out-exchange) (Fig. 6.7 & C.5). Therefore, the number of open state conformations identified here (45), is in line with and rationalize those deduced from the main chain amide backbone distance previously discussed (50), using a different trajectory. This would indeed signify that, both criteria are adequate to tackle with the imprecise and intricate definition of the open and close state conformations.

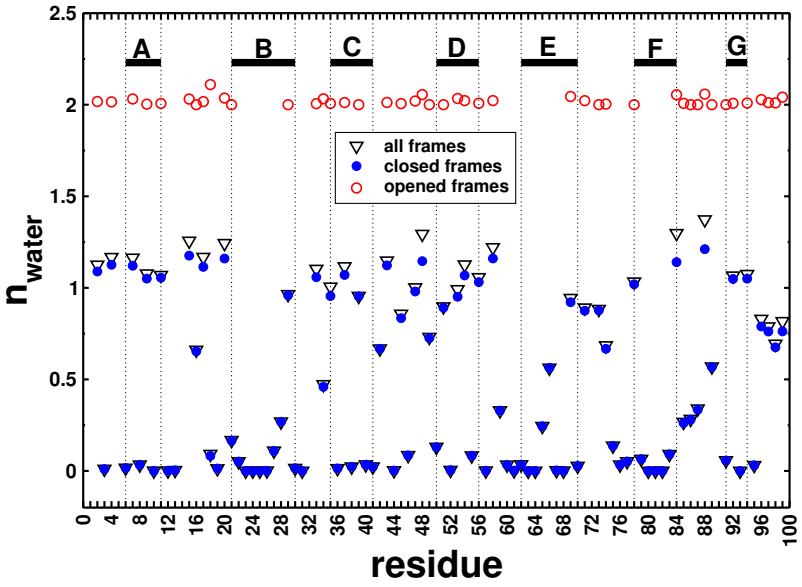


FIGURE 6.7: Average number of water molecules in the 1st solvation shell of each amide backbone NH residue. Open circles correspond to the average over the exchange-competent states whereas close circles and open triangles distinguish between the water average over the exchange-incompetent states and over both sets (global trajectory), respectively. Only metadynamic trajectory is shown.

Using the above definion for open and close states, the extent of protection was then estimated as explained in the simulation details for individual amide backbone NH hydrogens and comparison with experimental ones is attempted, see Fig. 6.8 and tables C.1 & C.2. The analysis stills required further consideration and thus, the values discussed here should be taken with some care, albeit the resulting profile falls well in the expected experimental trend.

It is seen in Fig. 6.8 that the cut-off of 2.76 Å results to provide a better estimation of protection factors compared to the experimental values although it captures less open-competent states than 3 Å. This confirms how important and sensitive is the proper choice of tresholds used in the definition of open/close states. Large cut-off range, no matter the nature of the structural property involved, would in general define a loose condition of exchange resulting in an unphysical EX1 monomolecular limits truly realized only when the amide is completely exposed to solvent as in global unfolding reactions, thereby providing wrong estimates of the protection factors, the latter being estimated assuming the EX2 kinetics. Common features in Fig. 6.8 include the following :

1. most fast exchanging residues, $logPF < 3$ are largely populating the protein unstructured regions like N- and C-tails, A-B and C-D loops. This would indeed be consistent with the experimental fact that large mobility induces more likely a fast exchange process ;

2. weakly protected residues $3 \leq logPF \leq 6$ show intermediate behaviour and are either in the loops or at the limit of β-stranded portions. SER11, ASP34 and VAL49 are some of them;

3. strongly protected residues $logPF > 6$ are for the majority located in the stranded regions of β2m, thereby rationalizing the view that secondary struture provides additional protection to backbone amide protons (Kieseritzky, Morra, and Knapp, 2006) ;

4. no experimental value is available for amides having protection lesser than 4, consistent with the fact that the determination of energy landscapes using HDX experiments in protein systems with respect to NMR resolution and timescale is only effective for $\Delta G > 5$ kcal mol^{-1} (Rennella et al., 2009).

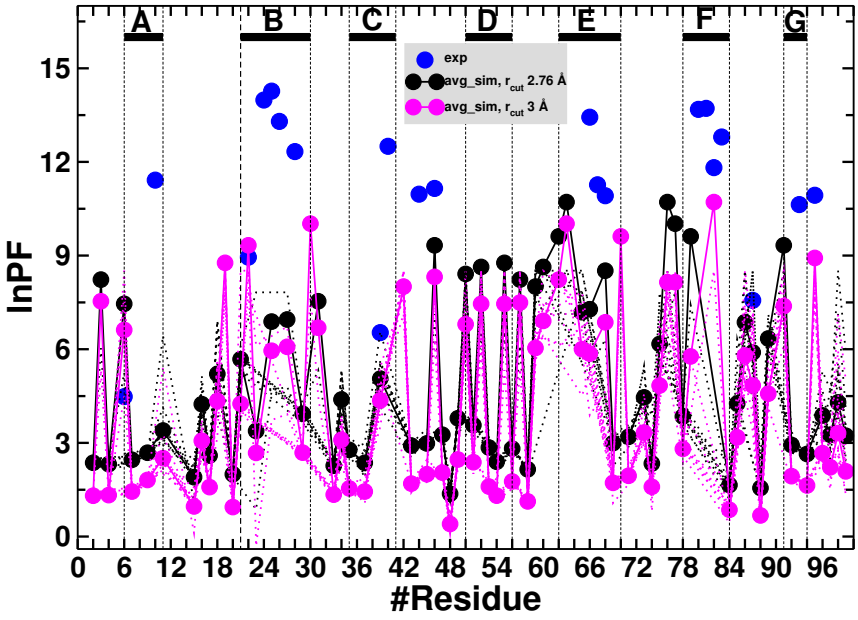


FIGURE 6.8: Protection factors averaged over all the umbrella sampling windows using the cut-off of 2.76 Å (in black) and 3 Å (in magenta). Experimental protection factors compiled from (Rennella et al., 2012a) are reported for comparison (in blue). Broken lines show the average over individual sampling window. Stranded regions are distinguished by the letters.

The data related to umbrella sampling simulation shown till now were averaged only in the reference window. In addition, Fig. 6.8 also reported the average protection factors considering the other sampling windows. It appears that, even though the average over the 9 sampling windows (continuous line) is closer to the experimental values than the average in individual window, the resulting computed data are still greatly underestimating the experimental measured ones. This could be ascribed in part to the short timescale and on the other hand to the inappropriate selection of global relevant reaction coordinates. These preliminary results remain qualitative and show however, a consistent picture compared to the overall expected trend. At the moment further NMR experiments are being performed. This will guide the simulation process and data analysis to efficiently predict the open and close state conformations, one of the bottleneck, if not the principal issue in the description of amide hydrogen exchange according to the well established Linderstrøm-Lang scheme.

The previous analysis have confirmed that the protein mainly samples the out-exchange conformational states along the trajectory, implying that the number of (sub)global opening-closing transitions are very limited. Umbrella sampling was indeed used to tackle this issue and promote the sampling of more (near)exchange-competent states. Fig. 6.9 displays the profiles of the potential of mean force obtained from the weighted histogram analysis, used to ensure the proper sampling of Boltzmann distribution and free energy at 4 relevant experimental temperatures.

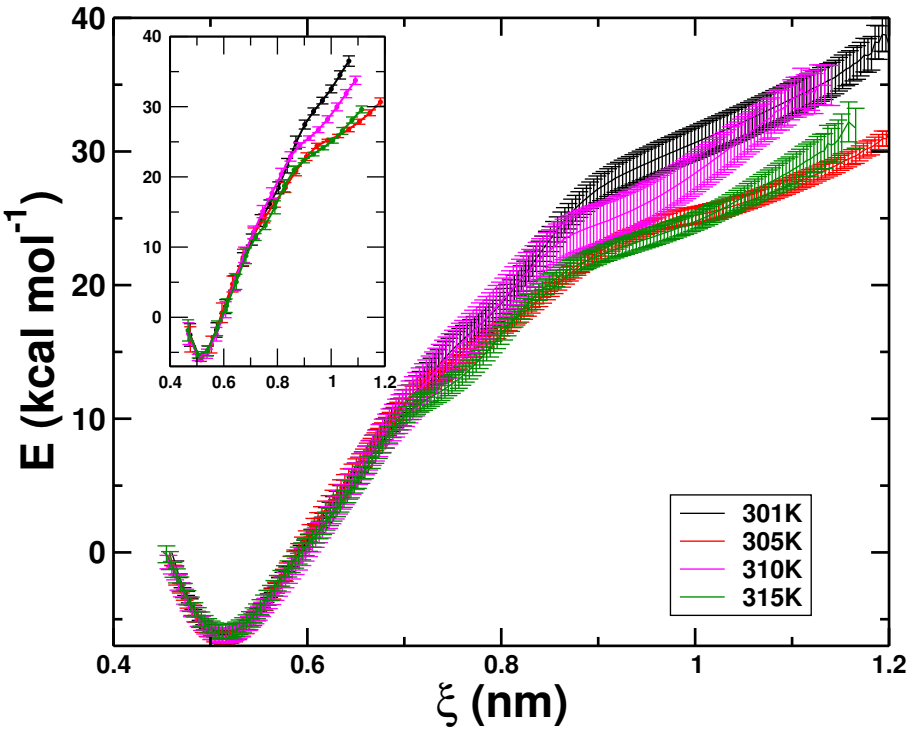


FIGURE 6.9: Potential of mean force associated with the opening of β2m structure as a function of the distance between the α-carbons of CyS25 and Tyr66 residues. Errors are statistical uncertainties obtained from 200 bootstraps. The inset shows the same graph with the errors computed using block averaging analysis.

As expected in Fig. 6.9, the PMF profiles are increasing along the selected reaction coordinates, consistent with the progressive local opening of β2m structure. Indeed, it was already showed that (Fig. C.1) moving along the CYS25-TYR66 Cα's distance starting from the reference sampling window, the H-bonds were gradually losed. Furthermore, within fluctuations, the PMF profiles are decreasing with temperature increase, confirming the picture that residue mobilities are enhanced at high temperatures and that these half-denaturing conditions induce large-scale dynamics, thereby decreasing the free energy gap required for (sub)global fluctuations. Albeit the resulting energy is still higher than the expected experimental values, perhaps attributed to short timescale and inadequate selection of reaction coordinates, the shape, however, follows the overall trend that should be envisioned from the global unfolding/opening reaction.

The statistical errors were computed using the bootstrapping scheme and thus the analysis should a priori be treated carefully as this procedure may significantly underestimate the errors if the data are correlated (Hub, Groot, and Spoel, 2010). For the sake of comparison and for a proper handling of correlation issues, error bars were also computed using the block average procedure, one of the most simple analysis scheme (inset of Fig.6.9). The result indicates that both error bars fall within consistent ranges and thus possible artifacts due to the use of a bootstrapping resampling scheme can be ruled out, albeit more sampling points should be considered to ensure a smooth convergence of the statistical errors.

The main purpose of this step and the use of different temperatures are to provide an atomistic free energy landscape description of the dynamic of backbone amide hydrogens in β2m as well as the temperature dependence of the thermodynamic parameters $\Delta G, \Delta H, \Delta C_p$ and precisely for ΔS accompanying the exchange process, in order to complement the qualitative analysis thoroughly discussed yet. This is still at the initial stage and will be pursued in the the next future.

6.4 Summary

The use of hydrogen/deuterium exchange (HDX) in the study and prediction of protein conformational dynamics and thermodynamics is a well recognized method that relies on the lability of high electropositive protons connected to polar O, S and N atoms. NMR and mass spectrometry methods are ideally used for the detection of such amides. The former exploits the difference in spins between H and D nuclei whereas the latter works on the difference in mass (D is 1 Da heavier than H) to measure the shift in *m/z* ratio. However, both methods, even though have significantly evolved in the recent years, remain limited in probing the fast exchangeable amide residues and mainly as far as atomic scale resolutions and timescales are concerned. Computational methods and namely atomistic molecular dynamic simulations can efficiently overcome this shortcoming notwithstanding the short timescales usually achievable in such methods. This chapter has recorded the principles and challenges associated to the study of amide hydrogen exchange in general and in β2m in particular. The ambiguous definition of open and close states in HDX remains perhaps the most troublesome issue and the core limitation of the Linderstrøm-Lang theory. Open states were considered here as frames along the trajectory having in the first coordination shell at least 2 interacting water molecules. This open state definition, albeit with different cut-off range was previously used and was proven to efficiently capture the relevant amide exchange dynamics and thermodynamics. The results discussed throughout this chapter even being more qualitative in general have provided essential and consistent description of the intricate and variegated nature of indidivual backbone amide protons in β2m. The computed protection factors result to greatly underestimate the experimental measured ones. This could be ascribed in part to the short timescale which does not allow for a proper system

equilibration and on the other hand to the inappropriate selection of global relevant reaction coordinates and/or the right definition of open/close states. However, in general, the picture emerging from the simulations shows great trait of consistency with experimental structural informations on β2m. Currently, further NMR experiments are being performed in relation to the fast exhanging residues using a CLEANEX-like protocol. This will guide the simulation process and data analysis to efficiently predict the open and close state conformations, thereby providing more accurate estimates of the thermodynamic parameters accompaning the (sub)global unfolding reaction as well as their temperature dependence.

Concluding Remarks and Outlook

Our understanding of protein conformational dynamics have greatly benefited from years of experimental endeavours, primarily using X-ray crystallography and/or NMR spectroscopy, methods that have now been considerably evolved. However, the dissection of protein misfolding and aggregation at a molecular level raises a spectrum of intricate methodological issues mainly related to the size, heterogeneity and transient hallmark of the intermediates involved. In this regard, *in silico* techniques and mainly atomistic molecular dynamics simulations, even on shorter timescale, proved to be markedly of help, highlighting detailed mechanistic and structural insights to complement and benchmark experiments.

The conformational dynamics has provided clues into the molecular machinery of protein stability and function, thereby opening-up further design strategies with targeted responses. Moreover, the structural diversity and intricate behaviour of amyloid fibril oligomers self-assembly under various conditions confer to them a unique property potentially useful for engineering novel responsive biopolymers and biomaterials. Understanding the factors leading to the formation of misfolded/(un)folded proteins species and later the mechanisms of precarious protein stability has technological and medical implications and could help rationalizing the development of novel therapeutic strategies with which to prevent and/or treat amyloidosis. An outstanding example is provided by medical devices which are currently used for removing β2-microglobulin from the blood of hemodialysis patients in medical devices. Thus, our study may help enlighten our understanding of the complex machinery and the physico-chemical factors behind amyloidogenesis, thereby potentially inspire new design strategies.

In this thesis we have probed the atomistic scale conformational dynamics of two paradigmatic protein models for amyloidogenesis studies, transthyretin and β2-microglobulin using molecular dynamics simulations. The principal aim was the understanding of the major factors driving the misfolding and/or (un)folding of the latter specified proteins which play a precursor and prominent role in neurodegrative deseases.

To achieve that, the dynamics and dissociation of wild-type and mutant homotetramer of transthyretin was simulated. In particular, the behaviour of a triple mutant engineered *in silico* (no crystal structure was available at the time this study was done) was studied. In addition, the interaction of β2-microglobulin with

hydrophobic surfaces was investigated and the thermodynamics of the binding process was addressed using end-point free energy calculations. Finally, the backbone amide hydrogen exchange (HDX) of the latter system was computed and the comparison of the amide hydrogen protection observed in NMR experiments with a number of microscopic properties of the protein structure averaged in the simulations was done. On a more methodological perspective, a theoretical approach to estimate the entropy loss upon complex formation was introduced.

The main conclusions can be summarized as follows :

1. **An *in silico* engineered mutation of transthyretin**

   (a) the mutation considerably shifts the tetramer-folded monomer equilibrium towards the monomer, making this triple mutant a useful tool for structural and dynamical studies ;

   (b) the clustering and discretization of simulation trajectories using Markov State Models (MSMs) after the equilibration phases show that the dissociation process in our engineered structure starts as fast as the dynamics starts, as confirmed by the most-native conformation (RMSD $\sim$8.2 Å starting from the production run) captured by our microstate model. Whereas as far as MT-TTR is concerned a very good structural agreement between the native-like microstate and X-ray structure was seen (RMSD $\sim$1.3 Å), revealing therefore that its structure was essentially preserved ;

   (c) the thermodynamic equilibrium along the pathway from native tetramer to unfolded monomers is strongly shifted towards the formation of monomers. Assuming that the thermodynamics underlines the kinetics, this would indeed imply the speeding up of the rate-limiting step just before monomers self-assemble into amyloid fibrils ;

   (d) the free energy computed according to two independent approaches (`BMF` and `Foldx`) implies stable monomers (relative to the stability of WT-TTR) after tetramer dissociation, shifting therefore the last step of fibrils formation towards the non-amyloidogenic mutant structures ;

   (e) the first step in conformational transition is the dissociation at the interface I/IV involving residues 17-24 and 110-123. The transition appears driven by the electrostatic repulsion of the pairs of acidic residues Glu 117 close in each dimer. Dimer III/IV rotates for most of the simulation about a hinge centered on salt bridges Glu 51 (I) - Arg 104 (III), Arg 104 (I) - Glu 51 (III) and hydrogen bonds Glu 51 (III) - Thr 123 (I) and at the end only about the latter two interactions. The final (possibly transient) conformation is stabilized by salt bridges Glu 117 (I) - ARG21 (III) and Lys 15 - Glu 54 at interface I/III.

2. **β2-microglobulin interactions with hydrophobic surfaces**

   (a) most proteins adsorb on the walls of the box (18 out of 27 within the first 35 ns and 4 additional molecules after 80 ns leading to 22 in all) without major distortions of local geometry, whereas free molecules

maintain proper structure and fluctuations as observed in explicit solvent MD simulations ;

(b) partial unfolding is observed at the terminal strands A and G and large changes are observed also in strand D whose conformational variability has been shown both in simulations and in NMR and crystallographic structures. Also loops AB and EF display large changes with respect to the starting structure. These results imply that flexible or more labile structural domains are important for general association properties because of their ability to adapt to molecular interactions ;

(c) the most encountered regions in proteins' first contacts with the walls entail residues close to the N- (ILE 1, TRP 60) and C-terminal (LYS 75 and MET 99) regions. This indicates a preference in orientation for the molecular encounter with the hydrophobic surface ;

(d) the thermodynamic analysis shows that the protein is able to bind the surface without a large desolvation penalty because large rearrangements occur maintaining polar groups well solvated. The interaction energy appears to be very large, in line with previous similar computational and experimental studies, making the adsorption irreversible on the timescale simulated here ;

(e) the very large interaction free energy suggests that, once adsorbed and unfolded, proteins could only favor fibril formation by acting as templates on the wall. A different scenario can be envisaged for the water-air interface which is strongly hydrophobic, but lacks van der Waals interactions.

3. **MD simulations of HDX in β2-microglobulin**

(a) no global unfolding-refolding structural transitions were observed on the timescale simulated here. Furthermore, residues mobility as depicted by atomic root mean square fluctuations indicates that, and as expected, large mobility arises from unstructured loop regions and terminal ends, with about 40% more structural fluctuations (compared to the rest of the protein structure) at the C-terminal, entailing residues ARG97, ASP98 and MET99. However, local parts of the protein structure occasionally undergo opening reactions, a large set of opening events leading to very short life transiently formed species in general, lasting less than 120 ps on average (for what concerns the particular case analysed) and consistent with what was previously noticed by Persson and Halle for BPTI ;

(b) 43 amides and mainly in β-stranded domains of β-2microglobulin like β-strands B (ASN21, LEU23, ASN24, CYS25, TYR26, VAL27, SER28), E (PHE62, TYR63, LEU64, TYR66, TYR67, THR68, PHE70) and F (ALA79, CYS80, ARG81, VAL82, ASN83) are mostly sampling the closed state

conformation, the latters of which mostly exhibit smallest solvent accessible surface areas, 0.16 Å^2 on average confirming the great tendency that closed state conformations are mostly involved in intramolecular hydrogen bonds and likely have small SASA ;

(c) open states were considered as frames along the trajectory having in the first coordination shell (r_{cut} = 2.76 and 3 Å) at least 2 interacting water molecules. 45 amides fall in that pool with the average number of water $n_{\text{H2O}} \approx 2.02$, in line with those deduced from the main chain hydrogen bonding (computed as the main chain amide backbone distance to the nearest carbonyl acceptor), using a different trajectory. This would indeed signify that, both criteria are adequate to tackle with the imprecise and intricate definition of the open and close state conformations ;

(d) the computed protection factors result to greatly underestimate the experimental measured ones. This could be ascribed in part to the short timescale, which does not allow for a proper system equilibration, and on the other hand to the inappropriate selection of global relevant reaction coordinates and/or the right definition of open/close states. However, in general, even being more qualitative, the picture emerging from the simulations shows great traits of consistency with experimental structural informations on β2-microglobulin ;

(e) the PMF profiles are increasing along the selected reaction coordinates, consistent with the free energy cost of progressive local opening of β2-microglobulin structure. Furthermore, within fluctuations, the PMF profiles are decreasing with temperature increase, confirming the picture that residue mobilities are enhanced at high temperatures and that these half-denaturing conditions induce large-scale dynamics, thereby decreasing the free energy gap required for (sub)global fluctuations.

4. **Conformational entropy in biomolecular protein association**

(a) conformational entropy in general and translational-rotational entropy in particular are often neglected upon binding of a ligand to its target, likely attributed to the inappropriate methods for dealing with large correlation involved and to the high dimensionality of the degrees of freedom ;

(b) with the nearest-neighbor method recently implemented, whose rationale is estimating the local probability density around each sample by counting its number of neighbors within a hypersphere of radius equal the distance from that sample to its k^{th} nearest neighbors, part of this bottleneck can be handled over, albeit one still have to tackle the correlation effects when both translation and rotation are present ;

(c) Huggins framed the issue and proposed a simplistic treatment for the computation of distances involving both translation and rotation in roto-translational space ;

(d) following its theory, we provide a more general coumpound distance to merge translations and rotations. Our formulation defines a general scaling factor necessary for ensuring the equal weigthing of translation and rotation to the global distance. Our approach indeed, relieves an important part of the dimensionality problem facing binning methods like histogram and provides an accurate estimation of full rotational-translational entropy from samples of rotational-translational states.

As perspective, we would like to provide an atomistic free energy landscape description of the dynamic of backbone amide hydrogens in β2-microglobulin which is still not fully described as well as the temperature dependence of the thermodynamic parameters $\Delta G, \Delta H, \Delta C_p$ and precisely for ΔS accompanying the exchange process in order to complement the qualitative analysis thoroughly discussed in Chapter 6. This is still at the initial stage and will be pursued in the the next future. Moreover further NMR experiments are being performed in relation to the fast exhanging residues using a CLEANEX-like protocol. This will guide the simulation process and data analysis to efficiently define the open and close state conformations, thereby providing more accurate estimates of the thermodynamic parameters accompaning the (sub)global unfolding reaction.

Moreover, detailing at atomic level what `open` and `close` states are will enable exact understanding of what HDX measured thermodynamic and kinetic quantities mean.

Appendix A

MD simulation of an *in silico* engineered mutation of TTR

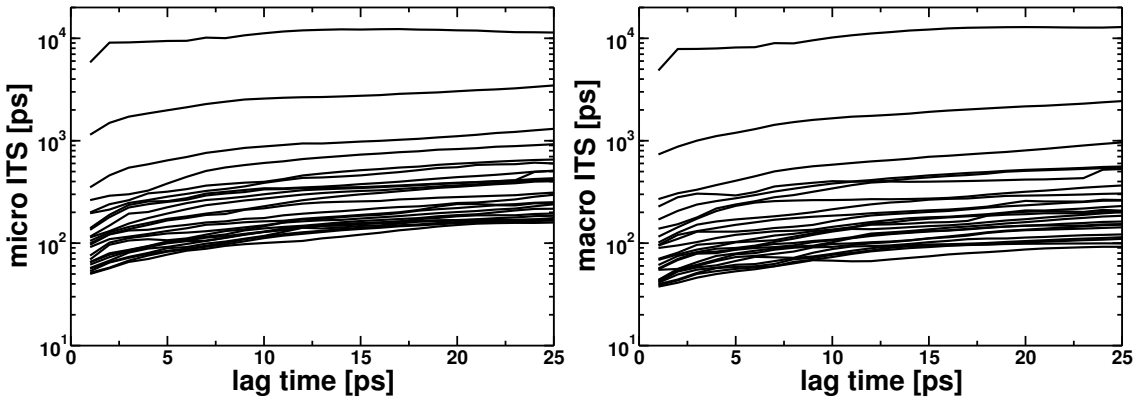


FIGURE A.1: Plots showing respectively from left to right the 25 slowest micro- and macro-Implied TimeScales validation of Markov State Models construction. Both plots display great level of similarities and the 3 slowest ITS level-off at ca. 3 steps proving on one hand that the models are Markovians and have reached convergence and on the other hand that the macrostate model is well representative of the microstate one since at least the 3 slowest ITS are preserved and both stay in the same range.

A.0.1 Insigth into the mesoscopic view of 3M-TTR

The overlay of the most-native like microstate in 3M-TTR, i.e. state with lowest RMSD compared to the crystal structure shows fairly good agreement, Fig. A.3-(a). The average RMSD of that microstate is 8.2 Å and this would mean that our microstate model does not efficiently capture the native one. On the other hand, the same analysis performed on MT-TTR (MSMs data not shown) leads to a very good structural agreement, Fig. A.3-(b) (average RMSD 1.3 Å). This would justify that, the 3M-TTR structure is largely undergoing structural changes (dissociation), making it difficult for the model to capture the native one.

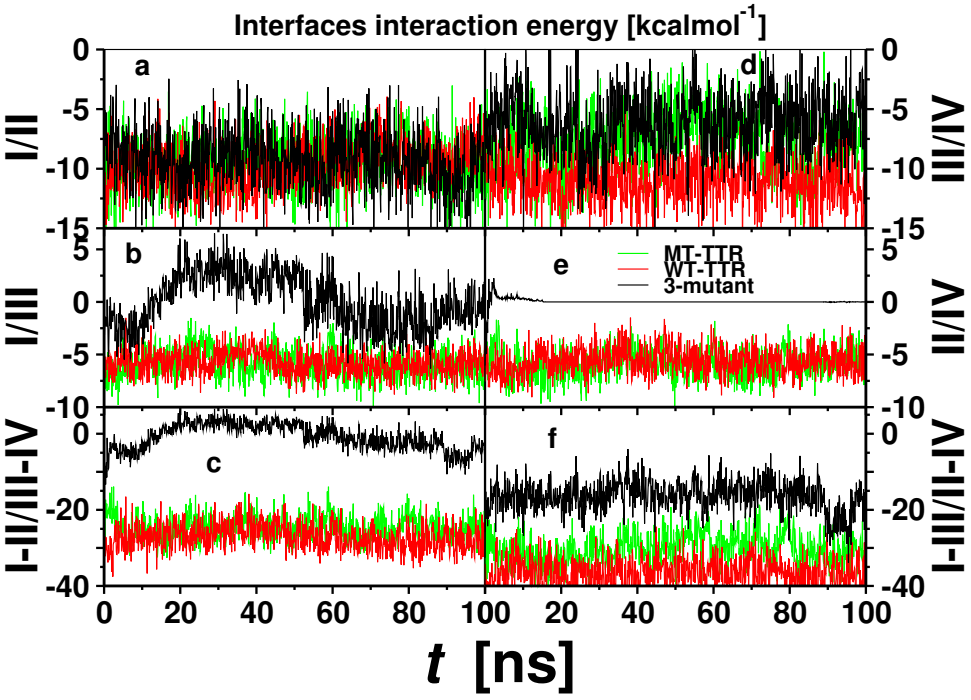


FIGURE A.2: The average interfaces interaction free energy within the different subunits of the tetrameric strucutres of TTR predicted by `Foldx`. a, b, d & e are monomer-monomer interfaces whereas c and f are dimer-dimer interfaces.

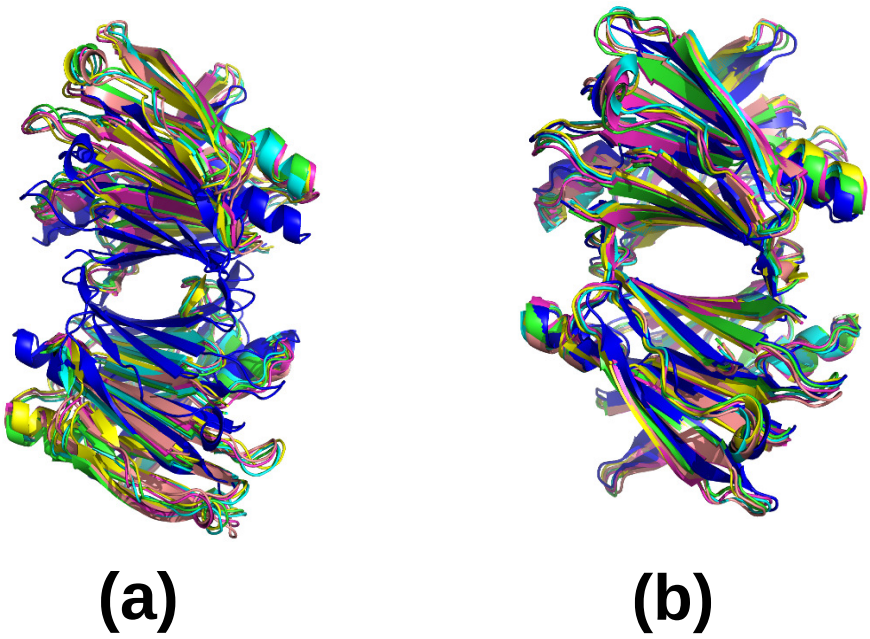


FIGURE A.3: Overlay of five randomly selected structures from the most native-like microstate with the crystal structure (shown in blue). We can see the fairly good structural agreement (RMSD average 8.2 Å) in 3M-TTR (a) and the very good structural alignment (average RMSD 1.3 Å) in MT-TTR (b).

Appendix B

$\beta 2$-microglobulin interactions with hydrophobic surfaces

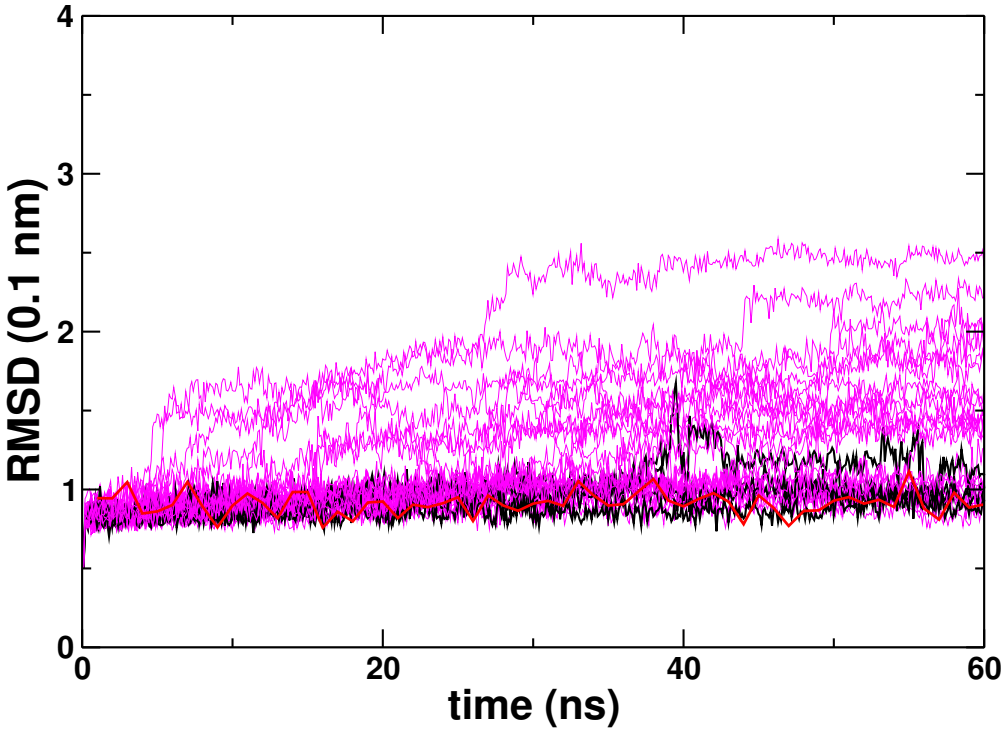


FIGURE B.1: RMSD of all implicit and explicit solvent simulations. Black lines highlight the four isolated implicit solvent trajectories within the first 60 ns and magenta the 23 remaining ones. Red trace shows the explicit solvent trajectory from NAMD simulation (frames collected at 1 ns intervals).

TABLE B.1: Summary of average energy terms in the last 2.5 ns. The average is normalized for the 26 bound-to-walls chains and the only `isolated` one at the end of the simulation. The table presents the average energy terms computed considering proteins alone on one hand and protein+hydrophobic walls on the other side (kJmol^{-1}).

	Bond	Angle	Proper-Dih.	Improper-Dih.	GB-Polarization	Nonpolar-Sol.	LJ-14	Coulomb-14	LJ-(SR)	Coulomb-(SR)	Position-Rest.	Potential	Total-Energy
contacting chains (single protein)	1498.49 ± 71.7705	3854.82 ± 102.635	4081.3 ± 52.9431	213.246 ± 197.777	-8310.76 ± 1026.57	273.858 ± 5.64234	1464.58 ± 33.3178	17469.1 ± 117.968	-2969.16 ± 51.7942	-27941.2 ± 1050.58	6.73329e+06 ± 201065	6.72292e+06 ± 201068	6.72292e+06 ± 201068
Non-contacting chains (single protein)	1487.43 ± 71.7233	3793.95 ± 99.8059	4006.98 ± 49.9469	210.079 ± 20.1853	-8163.41 ± 1000.52	273.078 ± 5.65175	1470.36 ± 33.639	17565.2 ± 110.991	-3052.97 ± 52.1952	-28217.4 ± 1012.96	3.09411e+07 ± 1.06766e+06	3.09305e+07 ± 1.06767e+06	3.09305e+07 ± 1.06767e+06
contacting chains (single protein+box)	1498.49 ± 71.7705	3854.82 ± 102.635	4081.3 ± 52.9431	213.502 ± 19.3128	-8768.82 ± 1027.82	171992 ± 5.64194	1464.58 ± 33.3178	17469.1 ± 117.968	-1.47306e+06 ± 646.233	-27941.2 ± 1050.58	7.02963e+06 ± 201260	5.72044e+06 ± 201254	5.72044e+06 ± 201254
Non-contacting chains (single protein+box)	1487.43 ± 71.7232	3793.95 ± 99.8061	4006.98 ± 49.9467	210.079± 20.1853	-8638.25± 1001.92	171991± 5.68563	1470.36± 33.639	17565.2± 110.99	-1.47192e+06± 642.606	-28217.4± 1012.95	3.12361e+07± 1.06726e+06	2.99278e+07± 1.06725e+06	2.99278e+07± 1.06725e+06

The standard deviation is computed according to the following equation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Appendix C

MD simulations of HDX exchange in β2-microglobulin

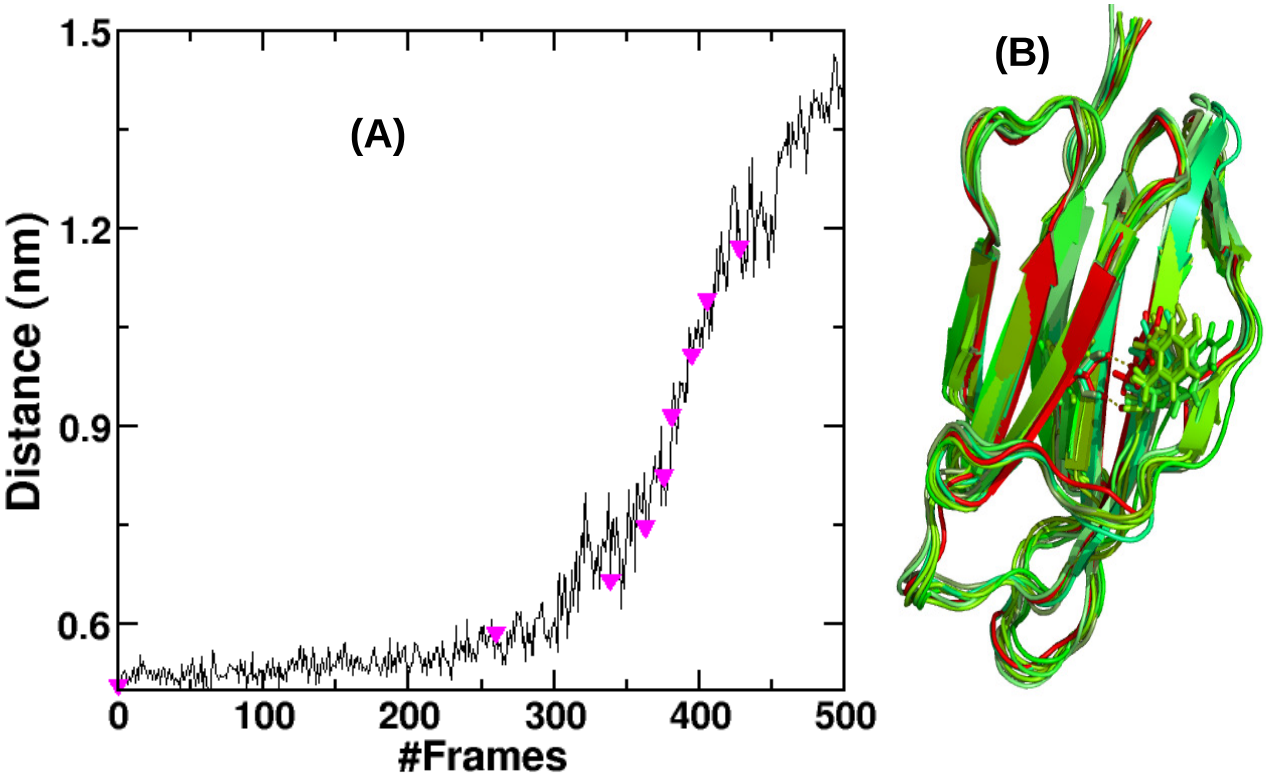


FIGURE C.1: In (A) the C_α (Cys25) - C_α (Tyr66) distance along the constant force pulling simulation is displayed in which triangles highlight the extracted conformations used as starting structures in the umbrella sampling simulations. In (B) the extracted frames are explicitely shown and aligned to the initial one in red. Cys25 and Tyr66 residues are shown in stick representation. The pulling progressively disrupts the H-bonds (that can be seen as well in (B)), inititially two to null.

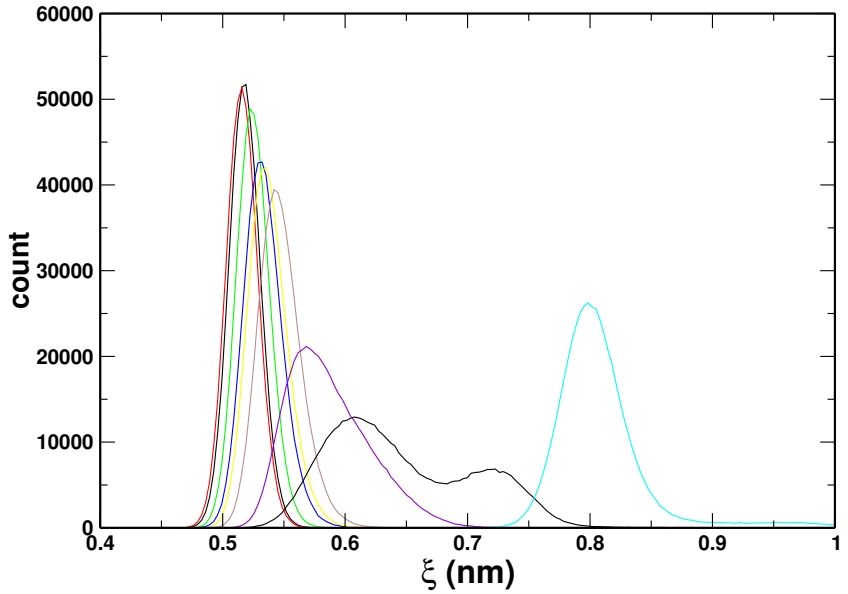


FIGURE C.2: Histograms distribution of the nine umbrella sampling windows considered in the simulation.

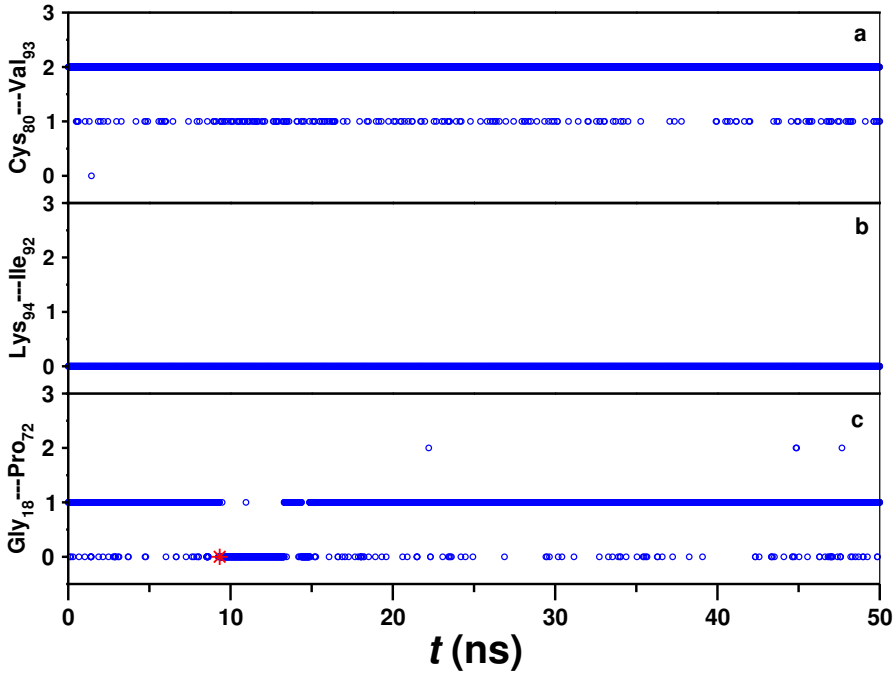


FIGURE C.3: Typical change in the number of main-chain intramolecular H-bonds of backbone amide protons with the nearest carbonyl acceptor. **(a)** NH (C80)···OC (V93) case ; Cys80 is always involved in intramolecular hydrogen bond. **(b)** NH (K94)···OC (I92) case ; no intra-molecular hydrogen bond with its closest acceptor. **(c)** NH (G18)···OC (P72) shows intermediate behaviour and oftenly alternate between close and open conformational states.

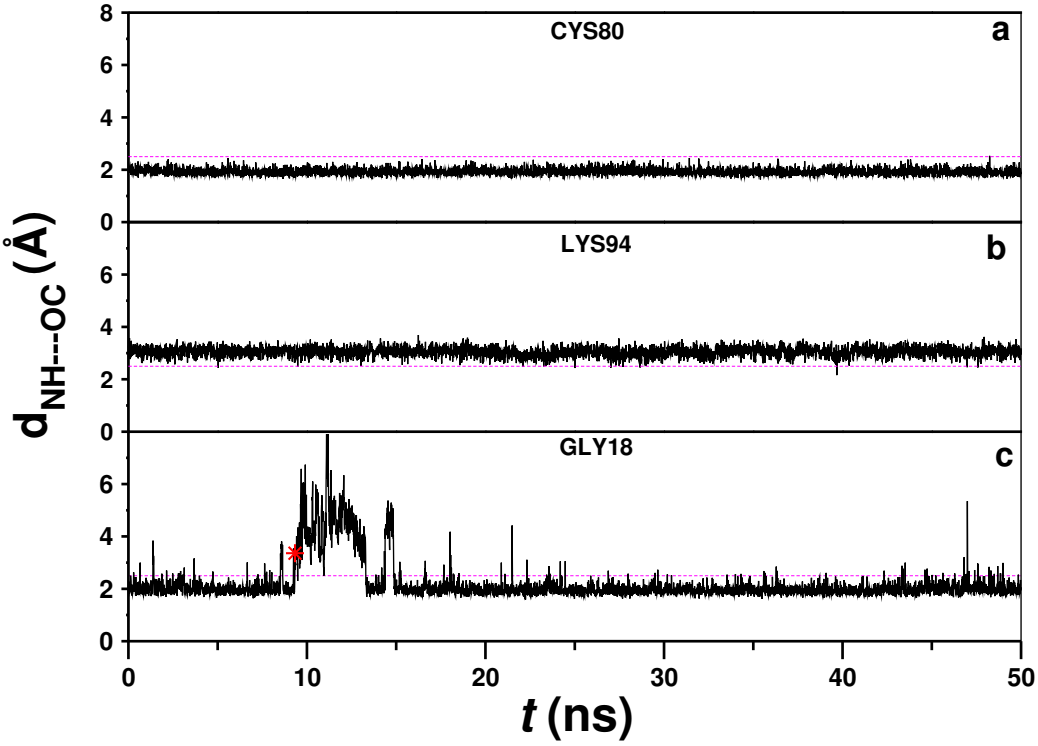
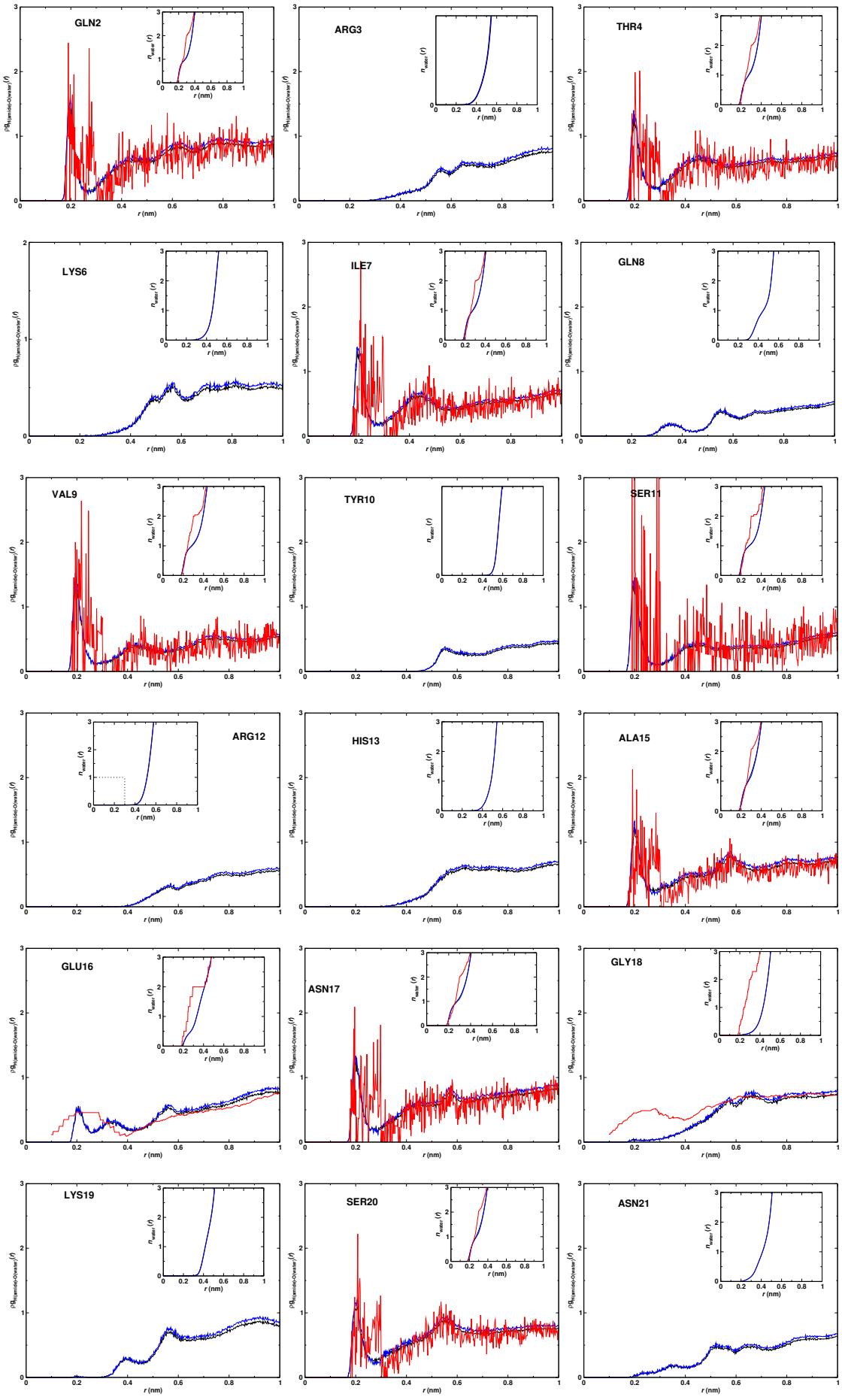
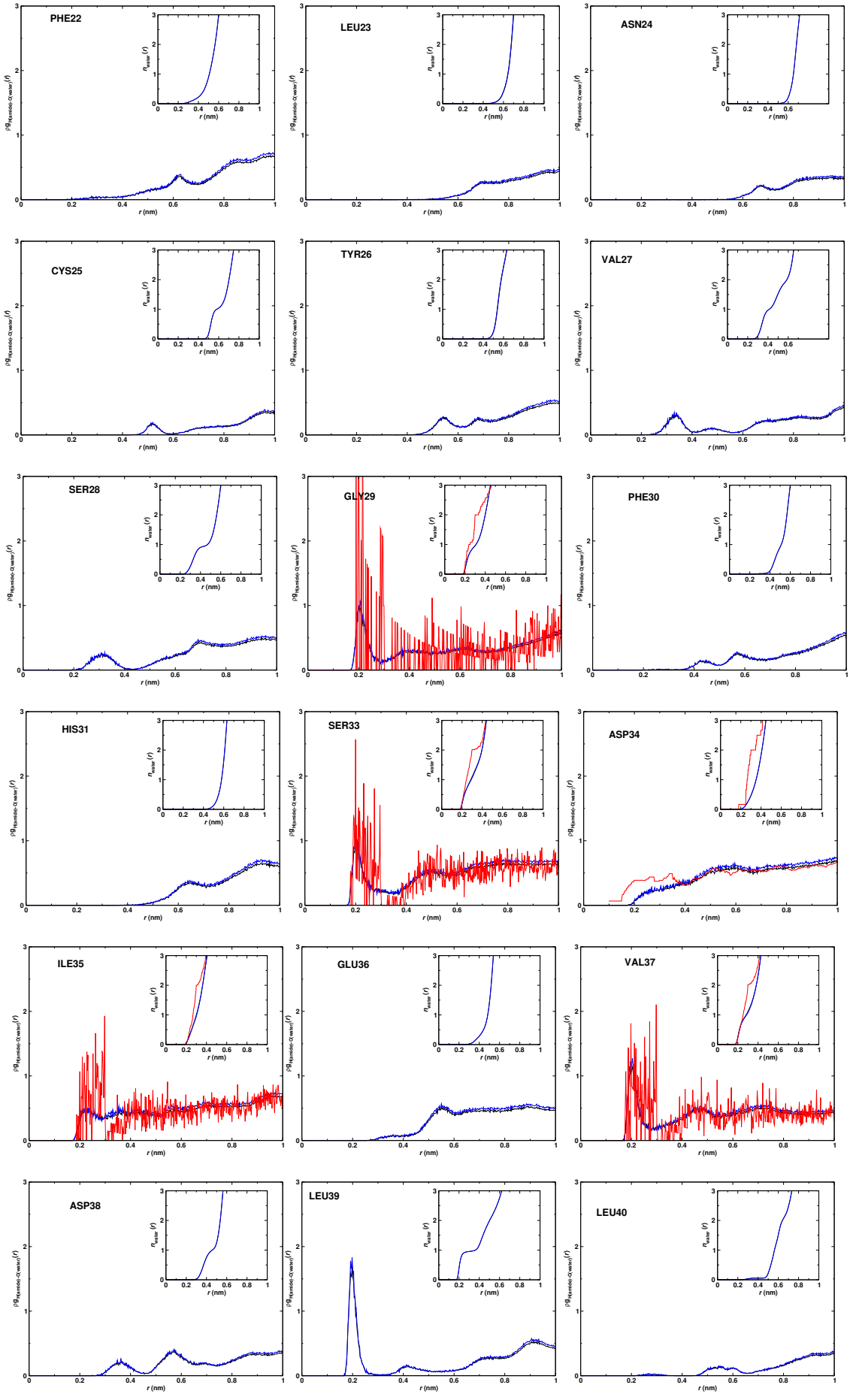
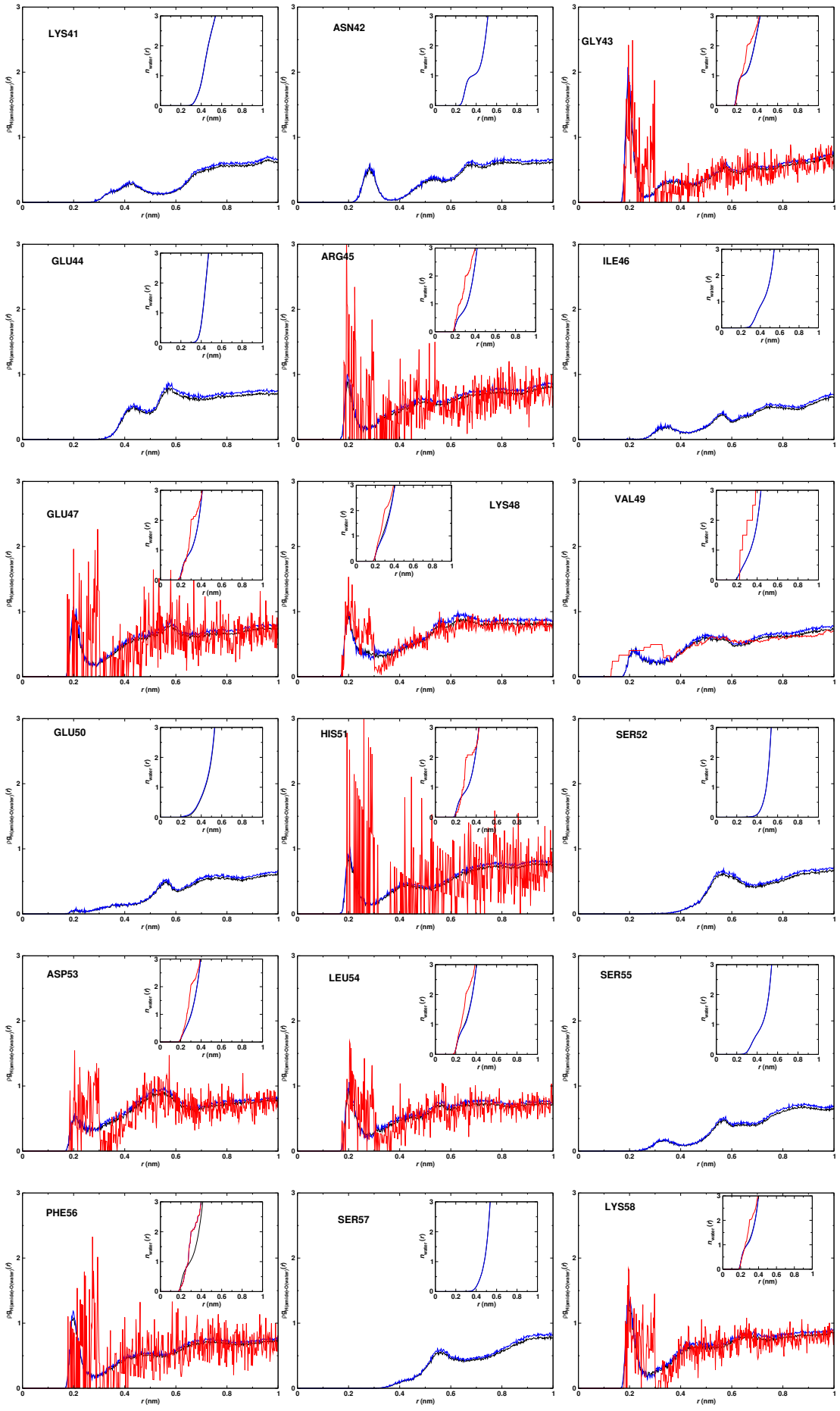
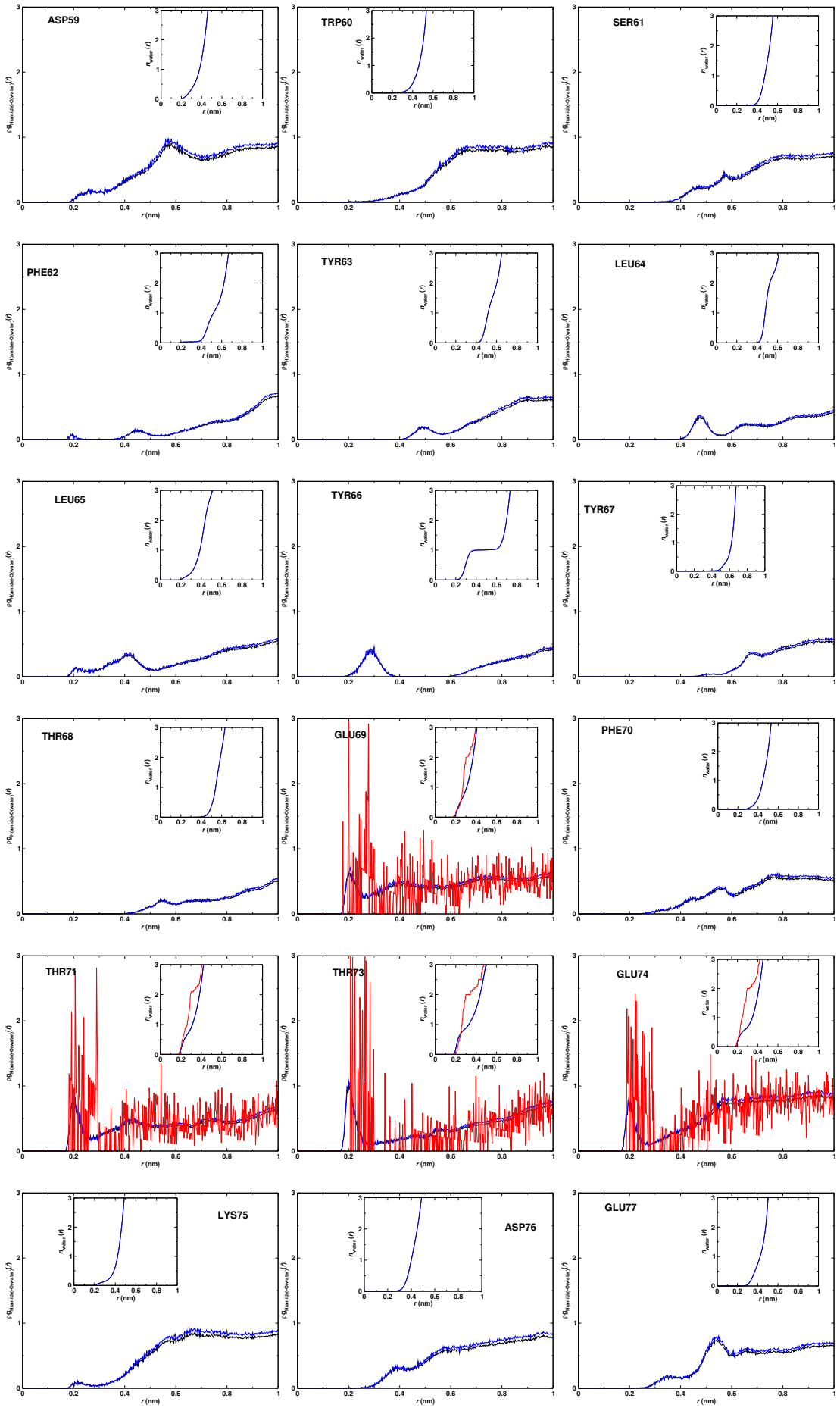


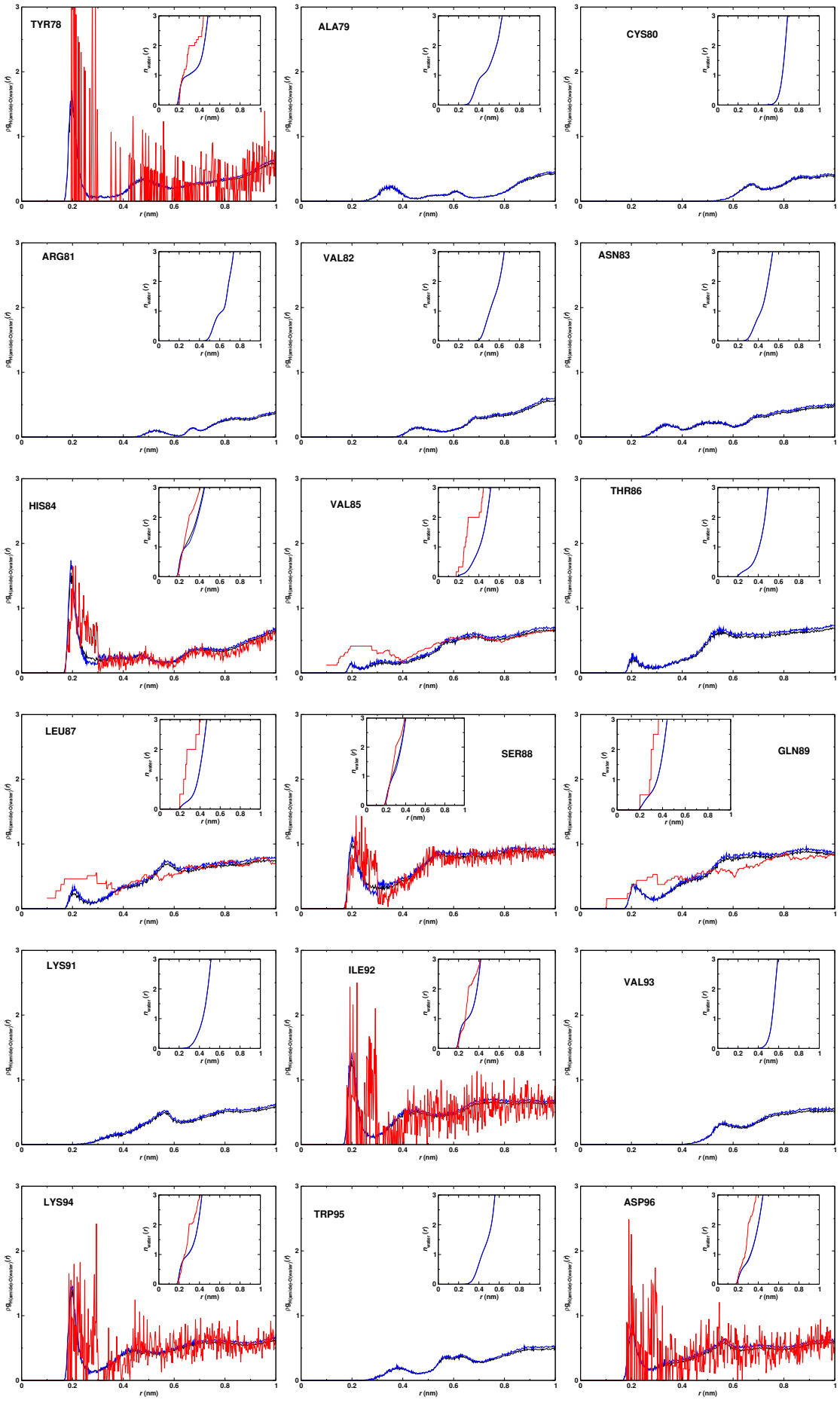
FIGURE C.4: variegated H-bond behaviours identified in the simulation. Each pannel represents the average separation distance between the backbone amide donor NH and its closest possible acceptor CO in the main-chain. Intercepted lines in magenta show the treshlod below which hydrogen bonds should exist and percentages account for occupancy of the latters along the simulation timescale. **(a)** NH (C80)$\cdots$OC (V93) case ; located at the core of the protein Cys80 is almost always bound to the CO of Val93 (96.9%) and less often to the amino group (12.7%) of the same residue. **(b)** NH (K94)$\cdots$OC (I92) case ; with no intra-molecular hydrogen bond with its closest acceptor (0%), this shows the typical case of highly solvent exposed residue. **(c)** NH (G18)$\cdots$OC (P72) case ; in spite of being most of the time (90.3%) connected to the carbonyl of Pro72, it is rarely interacting with the amino group of both Pro72 (0.7%) and Thr73 (1.8%). Transition open-close conformational states is seen in this case, mainly in between 9.57 - 16.63 ns.


GLN2
ARG3
THR4
LYS6
ILE7
GLN8
VAL9
TYR10
SER11
ARG12
HIS13
ALA15
GLU16
ASN17
GLY18
LYS19
SER20
ASN21
r (nm)


PHE22
LEU23
ASN24
CYS25
TYR26
VAL27
SER28
GLY29
PHE30
HIS31
SER33
ASP34
ILE35
GLU36
VAL37
ASP38
LEU39
LEU40
r (nm)


LYS41
ASN42
GLY43
GLU44
ARG45
ILE46
GLU47
LYS48
VAL49
GLU50
HIS51
SER52
ASP53
LEU54
SER55
PHE56
SER57
LYS58
r (nm)


ASP59
TRP60
SER61
PHE62
TYR63
LEU64
LEU65
TYR66
TYR67
THR68
GLU69
PHE70
THR71
THR73
GLU74
LYS75
ASP76
GLU77
r (nm)


TYR78
ALA79
CYS80
ARG81
VAL82
ASN83
HIS84
VAL85
THR86
LEU87
SER88
GLN89
LYS91
ILE92
VAL93
LYS94
TRP95
ASP96
r (nm)

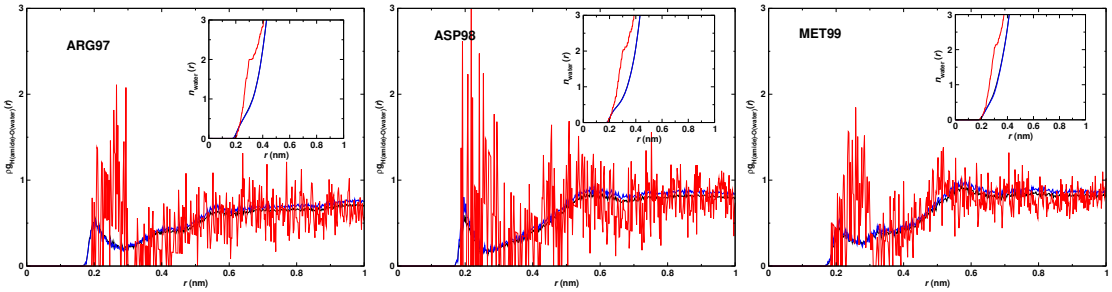


FIGURE C.5: Individual amide backbone residue H(NH)-O(H_2O) pair correlation functions. Black shows the average over all the frames, red represents the opened states with at least 2 water molecules within 3 Å and blue identifies all the closed states (< 2 waters within 3 Å). The average is done from the metadynamic trajectory.

TABLE C.1: Comparitive view of per residue protection factors computed taking as open states the backbone amides having within 2.76 Å at least 2 water molecules. The trajectory is from umbrella sampling simulations at 301 K. Indices 0, 260, 339, 363, 376, 381, 395, 405, 428 ; avg. and exp. are respectively the individual sampling windows, the average over the sampling windows and the experimental protection factors. PF values are computed using the relation $logPF = log\left(\frac{NF_{cl}}{NF_{op}}\right)$.

Residue	lnPF0	lnPF260	lnPF339	lnPF363	lnPF376	lnPF381	lnPF395	lnPF405	lnPF428	lnPF_avg	lnPF_exp
GLN2	2.343	2.295	2.376	2.465	2.657	2.305	2.521	2.371	2.440	2.455	—
ARG3	—	—	6.030	—	—	—	—	—	—	8.230	—
THR4	2.118	2.233	2.233	2.554	2.346	2.501	2.141	2.410	2.416	2.320	—
LYS6	—	—	—	—	—	—	—	—	5.254	7.456	4.483
ILE7	2.338	2.429	2.389	2.557	2.791	2.501	2.495	2.341	2.343	2.457	—
GLN8	—	—	—	—	—	—	—	—	—	—	—
VAL9	2.867	2.686	2.644	2.590	2.596	2.510	2.563	2.887	2.907	2.686	—
TYR10	—	—	—	—	—	—	—	—	—	—	11.417
SER11	4.119	3.028	3.435	6.436	3.122	3.378	3.403	2.928	3.210	3.395	—
ARG12	—	—	—	—	—	—	—	—	—	—	—
HIS13	—	—	—	—	—	—	—	—	—	—	—
ALA15	1.792	2.264	2.087	1.095	1.914	2.020	2.432	2.202	1.765	1.90	—
GLU16	4.300	3.608	5.144	4.927	3.944	3.832	4.679	4.481	4.226	4.24	—
ASN17	2.670	2.628	2.515	2.424	2.852	2.602	2.440	2.476	2.848	2.596	—
GLY18	4.846	5.144	5.293	6.907	6.907	6.570	5.377	5.623	3.955	5.208	—
LYS19	—	—	—	—	—	—	—	—	—	—	—
SER20	2.019	1.924	2.021	2.047	2.053	1.955	1.991	2.039	1.919	1.996	—
ASN21	5.254	5.806	5.623	5.334	5.742	5.742	6.117	5.216	—	5.681	—
PHE22	—	—	—	—	—	—	—	—	—	—	8.95
LEU23	—	—	—	—	—	—	7.824	—	0.870	3.382	—
ASN24	—	—	—	—	—	—	—	—	—	—	13.983
CYS25	—	—	—	—	—	—	—	—	4.680	6.885	14.267
TYR26	—	—	—	—	—	—	—	—	—	—	13.3
VAL27	—	—	—	—	—	—	7.824	—	4.796	6.952	—
SER28	—	—	—	—	—	—	—	—	—	—	12.333
GLY29	4.145	3.988	4.171	3.988	3.852	3.988	4.158	3.647	3.631	3.934	—
PHE30	—	—	—	—	—	—	—	—	—	—	—
HIS31	—	—	5.334	—	—	—	—	—	—	7.536	—
SER33	2.387	2.287	2.027	2.311	2.319	2.117	2.311	2.173	2.533	2.266	—
ASP34	5.254	4.820	5.334	5.254	4.314	3.767	4.330	3.988	3.882	4.389	—
ILE35	2.773	2.833	2.693	2.656	2.962	2.548	2.656	2.817	3.183	2.778	—
GLU36	—	—	—	—	—	—	—	—	—	—	—
VAL37	2.983	2.266	2.387	2.359	2.615	2.010	2.379	1.896	2.748	2.363	—
ASP38	—	—	—	—	—	—	—	—	—	—	—
LEU39	6.570	5.569	5.623	4.679	5.293	5.469	4.702	5.254	4.158	5.059	6.533
LEU40	—	—	—	—	—	—	—	—	—	—	12.5
LYS41	—	—	—	—	—	—	—	—	—	—	—
ASN42	—	—	—	—	—	—	—	—	—	—	—
GLY43	2.979	2.867	2.970	2.871	3.152	3.083	2.936	2.940	2.533	2.913	—
GLU44	—	—	—	—	—	—	—	—	—	—	10.967
ARG45	3.456	2.883	3.199	2.911	2.907	2.863	2.745	2.920	3.137	2.984	—
ILE46	—	7.130	—	—	—	—	—	—	—	9.328	11.15
GLU47	3.189	3.194	3.113	3.271	3.347	3.299	3.276	3.183	3.533	3.261	—
LYS48	1.498	1.494	1.562	1.042	1.234	1.224	1.482	1.238	1.493	1.372	—

VAL49	3.423	3.955	3.706	3.913	4.034	4.011	3.862	3.883	3.469	3.788	—
GLU50	—	7.130	—	8.517	—	—	7.824	7.418	—	8.412	—
HIS51	3.862	3.623	3.749	3.852	3.562	3.723	3.372	3.525	3.079	3.566	—
SER52	—	—	—	—	6.570	—	—	—	8.517	8.635	—
ASP53	2.784	2.851	2.762	2.996	2.825	2.612	2.759	3.032	3.122	2.850	—
LEU54	2.206	2.501	2.252	2.227	2.364	2.456	2.907	2.437	2.369	2.397	—
SER55	—	—	—	—	—	—	—	—	6.570	8.768	—
PHE56	2.895	2.612	2.932	2.602	3.098	2.879	2.924	2.676	2.825	2.816	—
SER57	—	—	—	8.517	—	—	—	—	6.117	8.229	—
LYS58	2.017	2.359	2.067	2.164	1.994	2.099	2.097	2.257	2.389	2.153	—
ASP59	7.418	8.517	8.517	8.517	—	8.517	8.517	7.824	6.907	8.006	—
TRP60	8.517	8.517	8.517	8.517	—	—	—	7.130	—	8.635	—
SER61	—	—	—	—	—	—	—	—	—	—	—
PHE62	—	7.824	8.517	—	—	—	—	—	—	9.616	—
TYR63	—	—	—	—	—	—	—	—	8.517	10.715	—
LEU64	—	—	—	—	—	—	—	—	—	—	—
LEU65	—	7.824	8.517	8.517	7.824	7.823	5.569	7.130	7.418	7.187	—
TYR66	—	—	—	—	—	—	6.436	—	5.377	7.280	13.433
TYR67	—	—	—	—	—	—	—	—	—	—	11.267
THR68	—	—	—	—	—	—	—	—	6.318	8.517	10.917
GLU69	3.158	3.299	3.103	3.305	3.152	3.299	2.777	3.041	2.266	2.993	—
PHE70	—	—	—	—	—	—	—	—	—	—	—
THR71	2.928	3.221	3.108	3.132	3.265	3.577	3.365	3.469	2.875	3.193	—
THR73	5.469	4.330	4.556	4.070	4.394	4.198	4.132	4.445	5.422	4.457	—
GLU74	2.557	2.518	1.917	2.287	2.462	2.599	2.443	3.872	1.542	2.339	—
LYS75	5.681	6.571	5.681	6.117	6.213	6.213	7.824	7.418	5.681	6.169	—
ASP76	—	—	—	8.517	—	—	—	—	—	10.715	—
GLU77	—	—	—	7.824	—	—	—	—	—	10.021	—
TYR78	4.314	3.944	4.158	3.656	3.813	3.616	3.832	3.497	4.046	3.845	—
ALA79	—	—	—	—	—	—	—	—	7.418	9.615	—
CYS80	—	—	—	—	—	—	—	—	—	—	13.683
ARG81	—	—	—	—	—	—	—	—	—	—	13.717
VAL82	—	—	—	—	—	—	—	—	—	—	11.817
ASN83	—	—	—	—	—	—	—	—	—	—	12.8
HIS84	1.824	1.710	1.566	2.416	1.729	1.186	1.536	1.521	1.651	1.651	—
VAL85	4.481	4.158	3.944	5.334	4.212	6.318	4.000	4.518	3.483	4.260	—
THR86	—	—	—	8.517	—	5.254	5.950	6.724	8.517	6.863	—
LEU87	—	—	8.517	—	8.517	4.463	4.702	5.569	—	5.900	7.567
SER88	1.383	1.500	1.436	2.492	1.678	1.320	1.449	1.382	1.687	1.557	—
GLN89	6.117	5.875	6.570	7.130	6.570	6.030	6.032	6.215	6.320	6.343	—
LYS91	8.517	—	—	—	—	—	—	7.418	—	9.328	—
ILE92	2.875	2.683	2.690	3.083	2.966	3.098	3.173	3.359	2.683	2.933	—
VAL93	—	—	—	—	—	—	—	—	—	—	10.633
LYS94	2.791	2.825	2.814	2.644	2.738	2.424	2.618	2.470	2.563	2.645	—
TRP95	—	—	—	—	—	—	—	—	—	—	10.933
ASP96	3.060	3.999	4.240	3.205	4.185	4.679	4.171	3.795	5.293	3.878	—
ARG97	4.796	2.829	2.492	3.862	3.163	3.390	2.911	3.215	4.702	3.258	—
ASP98	5.110	5.569	5.180	8.517	4.070	3.294	3.410	3.892	7.130	4.296	—
MET99	3.429	3.497	3.504	5.875	2.52	3.518	2.957	2.432	4.058	3.228	—

TABLE C.2: Comparitive view of per residue protection factors computed taking as open states the backbone amides having within 3 Å at least 2 water molecules. The trajectory is from umbrella sampling simulations at 301 K. Indices 0, 260, 339, 363, 376, 381, 395, 405, 428 ; avg. and exp. are respectively the individual sampling windows, the average over the sampling windows and the experimental protection factors. PF values are computed using the relation $logPF = log\left(\frac{NF_{cl}}{NF_{op}}\right)$.

Residue	lnPF0	lnPF260	lnPF339	lnPF363	lnPF376	lnPF381	lnPF395	lnPF405	lnPF428	lnPF_avg	lnPF_exp
GLN2	1.322	1.252	1.322	1.390	1.453	1.1584	1.341	1.262	1.298	1.930	
ARG3	—	—	5.334	—	—	—	—	—	—	8.229	—
THR4	1.178	1.330	1.272	1.506	1.328	1.476	1.227	1.316	1.353	1.994	
LYS6	7.824	—	7.418	—	8.517	—	—	—	4.517	7.187	4.483
ILE7	1.380	1.404	1.353	1.555	1.687	1.468	1.439	1.383	1.345	2.068	
GLN8	—	—	—	—	—	—	—	—	—	—	—
VAL9	1.966	1.844	1.877	1.735	1.696	1.692	1.654	1.924	1.972	2.491	
TYR10	—	—	—	—	—	—	—	—	—	—	11.417
SER11	3.259	2.171	2.569	5.422	2.141	2.627	2.309	1.993	2.472	3.111	
ARG12	—	—	—	—	—	—	—	—	—	—	—
HIS13	—	—	—	—	—	—	—	—	—	—	—
ALA15	0.794	1.269	1.172	0.108	0.992	1.031	1.573	1.231	0.784	1.766	
GLU16	3.183	2.572	3.749	3.271	2.875	2.666	3.372	3.122	3.276	3.466	
ASN17	1.614	1.638	1.556	1.399	1.784	1.590	1.471	1.455	1.799	2.188	
GLY18	4.314	4.499	4.463	5.569	5.469	5.623	4.198	4.284	3.183	4.902	
LYS19	7.824	8.517	—	—	—	—	8.517	8.517	7.824	8.768	
SER20	0.963	0.870	0.948	1.044	0.970	0.920	0.952	0.941	0.892	1.655	
ASN21	3.872	4.361	3.999	3.999	4.284	4.171	4.314	4.185	8.517	4.534	
PHE22	—	7.824	—	—	—	—	8.517	—	8.517	9.328	8.95
LEU23	—	—	—	—	—	—	6.117	—	-0.305	3.431	
ASN24	—	—	—	—	-	—	—	—	—	—	13.983
CYS25	—	—	—	—	—	—	—	—	3.731	6.450	14.267

TYR26	—	—	—	—	—	—	—	—	—	—	13.3
VAL27	—	—	—	—	—	—	6.724	—	3.923	6.619	
SER28	—	—	—	—	—	—	—	—	—	—	12.333
GLY29	2.752	2.759	2.817	2.647	2.581	2.713	2.810	2.560	2.584	3.066	
PHE30	—	—	7.824	-	—	—	—	—	—	10.021	
HIS31	—	—	4.481	—	—	—	—	—	—	7.248	
SER33	1.496	1.365	1.182	1.423	1.350	1.234	1.408	1.240	1.407	2.059	
ASP34	3.892	3.562	3.600	3.832	3.210	2.384	3.098	2.686	2.693	3.444	
ILE35	1.556	1.498	1.519	1.517	1.596	1.394	1.427	1.584	1.817	2.013	
GLU36	—	—	—	—	—	—	—	—	—	—	—
VAL37	1.906	1.289	1.468	1.459	1.609	1.191	1.408	1.069	1.739	2.139	
ASP38	—	—	—	—	—	—	—	—	—	—	—
LEU39	5.422	4.846	5.014	4.119	4.636	4.724	3.988	4.330	3.518	5.066	6.533
LEU40	—	—	—	—	—	—	—	—	—	—	12.5
LYS41	—	—	—	—	—	—	—	—	—	—	—
ASN42	7.824	7.824	—	8.517	7.130	7.824	8.517	7.824	8.517	8.006	
GLY43	1.768	1.692	1.716	1.647	1.762	1.844	1.710	1.757	1.409	2.158	
GLU44	—	—	—	—	—	—	—	—	—	—	10.967
ARG45	2.459	1.852	2.124	1.930	1.914	1.901	1.741	2.021	2.152	2.564	
ILE46	—	6.318	—	7.824	—	—	—	—	—	8.768	11.15
GLU47	2.006	1.957	1.939	2.112	2.180	1.966	2.051	1.970	2.302	2.480	
LYS48	0.509	0.368	0.697	0.101	0.229	0.334	0.498	0.262	0.662	1.397	
VAL49	2.319	2.637	2.473	2.581	2.434	2.660	2.530	2.501	2.164	2.825	
GLU50	6.724	6.436	7.130	8.517	7.824	6.724	6.318	6.213	7.130	7.025	
HIS51	2.533	2.563	2.440	2.650	2.392	2.376	2.222	2.369	2.010	2.803	
SER52	—	—	8.517	—	5.950	8.517	8.517	7.824	6.436	7.824	
ASP53	1.572	1.576	1.541	1.704	1.601	1.503	1.526	1.609	1.801	2.059	
LEU54	1.085	1.380	1.171	1.149	1.305	1.330	1.875	1.386	1.232	1.909	
SER55	—	—	—	—	8.517	—	8.517	—	5.334	7.770	
PHE56	1.870	1.636	1.877	1.527	1.959	1.875	1.752	1.632	1.810	2.316	
SER57	—	—	—	8.517	—	—	8.517	—	5.377	8.149	
LYS58	0.955	1.388	1.029	1.141	0.964	1.091	1.066	1.247	1.308	1.809	
ASP59	5.950	6.724	6.117	5.806	5.950	6.213	6.030	6.570	5.518	6.191	
TRP60	7.130	7.130	6.436	6.907	7.418	8.517	6.436	6.318	7.418	7.103	
SER61	—	—	—	—	—	—	—	—	—	—	—
PHE62	—	—	6.117	8.517	—	—	—	—	—	8.517	
TYR63	—	—	—	—	—	—	—	—	7.824	10.715	
LEU64	—	—	—	—	—	—	—	—	—	—	—
LEU65	7.824	7.130	6.724	6.570	6.213	7.418	4.517	6.030	5.950	6.369	
TYR66	—	—	—	—	—	—	5.216	—	3.882	6.138	13.433
TYR67	—	—	—	—	—	—	—	—	—	—	11.267
THR68	—	—	—	—	—	—	—	—	4.658	7.076	10.917
GLU69	1.804	1.924	1.836	1.968	1.877	1.952	1.540	1.744	1.072	2.153	
PHE70	8.517	—	8.517	—	—	8.517	—	—	—	9.616	
THR71	1.708	1.961	1.837	1.868	2.017	2.304	2.063	2.218	1.681	2.370	
THR73	4.575	3.194	3.562	3.055	3.254	3.045	2.970	3.127	4.171	3.746	
GLU74	1.799	1.779	1.175	1.544	1.641	1.789	1.663	2.940	0.811	2.427	
LYS75	4.463	5.180	4.536	4.595	4.955	5.144	5.875	6.907	4.082	5.148	
ASP76	7.824	—	8.517	—	7.824	8.517	7.418	—	7.130	8.229	
GLU77	—	7.823	—	6.724	—	—	8.517	—	7.130	8.316	
TYR78	3.168	2.911	3.083	2.710	2.773	2.578	2.821	2.402	3.046	3.288	
ALA79	—	—	—	—	—	—	5.681	—	3.680	5.792	
CYS80	—	—	—	—	—	—	—	—	—	—	13.683
ARG81	—	—	—	—	—	—	—	—	—	—	13.717
VAL82	—	—	—	—	—	—	—	8.517	—	10.715	11.817
ASN83	—	—	—	—	—	—	—	—	—	—	12.8
HIS84	0.958	1.004	0.635	1.763	0.850	0.435	0.723	0.806	0.788	1.844	
VAL85	3.436	3.189	2.647	4.034	3.199	4.984	2.875	3.449	2.498	3.618	
THR86	—	7.130	—	7.418	8.517	4.462	4.984	4.984	7.418	6.224	
LEU87	7.130	7.418	6.724	7.418	6.436	3.390	3.934	4.411	6.117	5.267	7.567
SER88	0.508	0.582	0.525	1.733	0.7412	0.448	0.520	0.510	0.764	1.635	
GLN89	4.378	4.145	4.747	5.623	4.445	4.362	4.658	4.411	5.144	4.769	
LYS91	7.418	—	7.130	8.517	8.517	7.824	7.130	6.117	7.824	7.536	
ILE92	1.905	1.773	1.732	1.957	1.932	2.101	2.175	2.224	1.721	2.503	
VAL93	—	—	—	—	—	—	—	—	—	—	10.633
LYS94	1.721	1.746	1.791	1.589	1.698	1.502	1.663	1.463	1.575	2.232	
TRP95	7.824	—	—	—	—	8.517	—	7.823	8.517	8.923	10.933
ASP96	2.250	2.615	2.741	2.326	2.780	3.237	2.584	2.395	3.913	3.067	
ARG97	3.188	1.839	1.511	2.856	2.182	2.184	1.950	2.180	3.127	2.733	
ASP98	3.842	4.556	4.269	7.130	2.840	2.553	2.437	2.755	5.950	3.809	
MET99	2.218	2.354	2.273	4.984	1.358	2.238	1.807	1.340	2.806	2.556	

Bibliography

Abdolvahabi, Alireza et al. (2014). "Metal-Ion-Specific Screening of Charge Effects in Protein Amide H/D Exchange and the Hofmeister Series". In: *Analytical Chemistry* 86.20, pp. 10303–10310. DOI: 10.1021/ac502714v.

Abraham, M. J. et al. (2014). *GROMACS User Manual version 5.0.4*. URL: www.gromacs.org.

Abrams, Cameron and Giovanni Bussi (2014). "Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration". In: *Entropy* 16.1, pp. 163–199. DOI: 10.3390/e16010163.

Abriata, Luciano A. and Matteo Dal Peraro (2015). "Assessing the potential of atomistic molecular dynamics simulations to probe reversible protein-protein recognition and binding". In: *Sci. Rep.* 5, p. 10549. DOI: 10.1038/srep10549.

Adcock, Stewart A. and J. Andrew McCammon (2006). "Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins". In: *Chemical Reviews* 106.5, pp. 1589–1615. DOI: 10.1021/cr040426m.

Ajay and Mark A. Murcko (1995). "Computational Methods to Predict Binding Free Energy in Ligand-Receptor Complexes". In: *Journal of Medicinal Chemistry* 38.26, pp. 4953–4967. DOI: 10.1021/jm00026a001.

Alberts, Bruce et al. (2009). "Essential cell biology". In: New York, NY 10016, USA, and 2 Park Square, Milton Park, Abingdon, OX14 4RN, UK: Garland Science, Taylor Francis Group.

Alder, B. J. and T. E. Wainwright (1957). "Phase Transition for a Hard Sphere System". In: *The Journal of Chemical Physics* 27.5, pp. 1208–1209. DOI: 10.1063/1.1743957.

Almeida, M. R. et al. (2005). "Small Transthyretin (TTR) Ligands as Possible Therapeutic Agents in TTR Amyloidoses". In: *Current Drug Targets - CNS Neurological Disorders* 4.5, pp. 587–596. DOI: 10.2174/156800705774322076.

Andersen, Hans C. (1980). "Molecular dynamics simulations at constant pressure and/or temperature". In: *The Journal of Chemical Physics* 72.4, pp. 2384–2393. DOI: 10.1063/1.439486.

Anderson, Janet S., Griselda Hernández, and David M. LeMaster (2008). "A Billion-fold Range in Acidity for the Solvent-Exposed Amides of Pyrococcus furiosus Rubredoxin". In: *Biochemistry* 47.23, pp. 6178–6188. DOI: 10.1021/bi800284y.

– (2013). "Assessing the chemical accuracy of protein structures via peptide acidity". In: *Biophysical Chemistry* 171.Supplement C, pp. 63 –75. DOI: 10.1016/j.bpc.2012.10.005.

Ando, Y. and M. Ueda (2012). "Diagnosis and Therapeutic Approaches to Transthyretin Amyloidosis". In: *Current Medicinal Chemistry* 19.15, pp. 2312–2323. DOI: 10.2174/092986712800269317.

Anfinsen, C B (1972). "The formation and stabilization of protein structure". In: *Biochemical Journal* 128.4, pp. 737–749. DOI: 10.1042/bj1280737.

Avila, Cesar L. et al. (2011). "Multiscale Molecular Dynamics of Protein Aggregation". In: *Current Protein Peptide Science* 12.3, pp. 221–234. DOI: 10.2174/138920311795860205.

Azinas, S. et al. (2011). "D-strand perturbation and amyloid propensity in β-2 microglobulin". In: *FEBS J.* 278, pp. 2349–2358. DOI: 10.1111/j.1742-4658.2011.08157.x.

Bai, Yawen et al. (1993). "Primary structure effects on peptide group hydrogen exchange". In: *Proteins: Structure, Function, and Bioinformatics* 17.1, pp. 75–86. DOI: 10.1002/prot.340170110.

Baker, Nathan A (2005). "Improving implicit solvent simulations: a Poisson-centric view". In: *Current Opinion in Structural Biology* 15.2, pp. 137 –143. DOI: 10.1016/j.sbi.2005.02.001.

Banerjee, Avik et al. (2013). "Conserved water mediated H-bonding dynamics of Ser117 and Thr119 residues in human transthyretin–thyroxin complexation: Inhibitor modeling study through docking and molecular dynamics simulation". In: *Journal of Molecular Graphics and Modelling* 44, pp. 70 –80. DOI: 10.1016/j.jmgm.2013.04.010.

Barducci, Alessandro, Massimiliano Bonomi, and Michele Parrinello (2011). "Metadynamics". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.5, pp. 826–843. DOI: 10.1002/wcms.31.

Bashford, Donald and David A. Case (2000). "Generalized Born Models of Macromolecular Solvation Effects". In: *Annual Review of Physical Chemistry* 51.1, pp. 129–152. DOI: 10.1146/annurev.physchem.51.1.129.

Beauchamp, Kyle A. et al. (2011). "MSMBuilder2: Modeling Conformational Dynamics on the Picosecond to Millisecond Scale". In: *Journal of Chemical Theory and Computation* 7.10, pp. 3412–3419. DOI: 10.1021/ct200463m.

Becker, J W and G N Reeke (1985). "Three-dimensional structure of β2-microglobulin". In: *Proceedings of the National Academy of Sciences* 82.12, pp. 4225–4229. URL: http://www.pnas.org/content/82/12/4225.abstract.

Bellotti, Vittorio et al. (1998). "β2-microglobulin can be refolded into a native state from ex vivo amyloid fibrils". In: *European Journal of Biochemistry* 258.1, pp. 61–67. DOI: 10.1046/j.1432-1327.1998.2580061.x.

Berendsen, H. J. C., J. R. Grigera, and T. P. Straatsma (1987). "The missing term in effective pair potentials". In: *The Journal of Physical Chemistry* 91.24, pp. 6269–6271. DOI: 10.1021/j100308a038.

Berendsen, H. J. C. et al. (1984). "Molecular dynamics with coupling to an external bath". In: *The Journal of Chemical Physics* 81.8, pp. 3684–3690. DOI: 10.1063/1.448118.

Berman, Helen M. et al. (2000). "The Protein Data Bank". In: *Nucleic Acids Research* 28.1, pp. 235–242. DOI: 10.1093/nar/28.1.235.

Berrera, Marco, Henriette Molinari, and Federico Fogolari (2003). "Amino acid empirical contact energy definitions for fold recognition in the space of contact maps". In: *BMC Bioinformatics* 4.1, p. 8. DOI: 10.1186/1471-2105-4-8.

Bjelkmar, Pär et al. (2010). "Implementation of the CHARMM Force Field in GROMACS: Analysis of Protein Stability Effects from Correction Maps, Virtual Interaction Sites, and Water Models". In: *Journal of Chemical Theory and Computation* 6.2, pp. 459–466. DOI: 10.1021/ct900549r.

Blake, C.C.F. et al. (1978). "Structure of prealbumin: Secondary, tertiary and quaternary interactions determined by Fourier refinement at 1.8 Å". In: *Journal of Molecular Biology* 121.3, pp. 339 –356. DOI: 10.1016/0022-2836(78)90368-6.

Bonar, Laurence, Alan S. Cohen, and Martha M. Skinner (1969). "Characterization of the Amyloid Fibril as a Cross- Protein". In: *Proceedings of the Society for Experimental Biology and Medicine* 131.4, pp. 1373–1375. DOI: 10.3181/00379727-131-34110.

Bonomi, Massimiliano et al. (2009). "PLUMED: A portable plugin for free-energy calculations with molecular dynamics". In: *Computer Physics Communications* 180.10, pp. 1961 –1972. DOI: 10.1016/j.cpc.2009.05.011.

Born, M. and R. Oppenheimer (1927). "Zur Quantentheorie der Molekeln". In: *Annalen der Physik* 389.20, pp. 457–484. DOI: 10.1002/andp.19273892002.

Bowman, Gregory R. (2012). "Improved coarse-graining of Markov state models via explicit consideration of statistical uncertainty". In: *The Journal of Chemical Physics* 137.13, p. 134111. DOI: 10.1063/1.4755751.

– (2014). "A Tutorial on Building Markov State Models with MSMBuilder and Coarse-Graining Them with BACE". In: *Protein Dynamics: Methods and Protocols*. Ed. by Dennis R. Livesay. Totowa, NJ: Humana Press, pp. 141–158. DOI: 10.1007/978-1-62703-658-0_8.

Bowman, Gregory R., Xuhui Huang, and Vijay S. Pande (2009). "Using generalized ensemble simulations and Markov state models to identify conformational states". In: *Methods* 49.2, pp. 197 –201. DOI: 10.1016/j.ymeth.2009.04.013.

Brancolini, Giorgia et al. (2015). "Probing the Influence of Citrate-Capped Gold Nanoparticles on an Amyloidogenic Protein". In: *ACS Nano* 9.3, pp. 2600–2613. DOI: 10.1021/nn506161j.

Brooks, Bernard R. et al. (1983). "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations". In: *Journal of Computational Chemistry* 4.2, pp. 187–217. DOI: 10.1002/jcc.540040211.

Bussi, Giovanni, Davide Donadio, and Michele Parrinello (2007). "Canonical sampling through velocity rescaling". In: *The Journal of Chemical Physics* 126.1, p. 014101. DOI: 10.1063/1.2408420.

Bussi, Giovanni, Alessandro Laio, and Michele Parrinello (2006). "Equilibrium Free Energies from Nonequilibrium Metadynamics". In: *Phys. Rev. Lett.* 96 (9), p. 090601. DOI: 10.1103/PhysRevLett.96.090601.

Bussi, Giovanni et al. (2006). "Free-Energy Landscape for β Hairpin Folding from Combined Parallel Tempering and Metadynamics". In: *Journal of the American Chemical Society* 128.41, pp. 13435–13441. DOI: 10.1021/ja062463w.

Calabrese, Matthew F. and Andrew D. Miranker (2007). "Formation of a Stable Oligomer of β-2 Microglobulin Requires only Transient Encounter with Cu(II)". In: *Journal of Molecular Biology* 367.1, pp. 1 –7. DOI: 10.1016/j.jmb.2006.12.034.

Calabrese, Matthew F. and Andrew D. Miranker (2009). "Metal binding sheds light on mechanisms of amyloid assembly". In: *Prion* 3.1, pp. 1–4. DOI: 10.4161/pri.3.1.8601.

Camilloni, C. et al. (2016). "Rational design of mutations that change the aggregation rate of a protein while maintaining its native structure and stability". In: *Sci. Rep.* 6, p. 25559. DOI: 10.1038/srep25559.

Campistol, Josep M. (2001). "Dialysis-Related Amyloidosis After Renal Transplantation". In: *Seminars in Dialysis* 14.2, pp. 99–102. DOI: 10.1046/j.1525-139x.2001.00038.x.

Cantarutti, Cristina et al. (2017). "Citrate-stabilized gold nanoparticles hinder fibrillogenesis of a pathological variant of β2-microglobulin". In: *Nanoscale* 9 (11), pp. 3941–3951. DOI: 10.1039/C6NR09362K.

Chandrasekaran, P. and R. Rajasekaran (2016). "A systematic molecular dynamics approach to the structural characterization of amyloid aggregation propensity of β2-microglobulin mutant D76N". In: *Mol. BioSyst.* 12 (3), pp. 850–859. DOI: 10.1039/C5MB00759C.

Cherny, Izhack and Ehud Gazit (2008). "Amyloids: Not Only Pathological Agents but Also Ordered Nanomaterials". In: *Angewandte Chemie International Edition* 47.22, pp. 4062–4069. DOI: 10.1002/anie.200703133.

Chevalier-Larsen, Erica and Erika L.F. Holzbaur (2006). "Axonal transport and neurodegenerative disease". In: *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* 1762.11. Molecular Basis of Amyotrophic Lateral Sclerosis, pp. 1094 –1108. DOI: 10.1016/j.bbadis.2006.04.002.

Childers, Matthew Carter and Valerie Daggett (2017). "Insights from molecular dynamics simulations for computational protein design". In: *Mol. Syst. Des. Eng.* 2 (1), pp. 9–33. DOI: 10.1039/C6ME00083E.

Chiti, Fabrizio and Christopher M. Dobson (2006). "Protein Misfolding, Functional Amyloid, and Human Disease". In: *Annual Review of Biochemistry* 75.1, pp. 333–366. DOI: 10.1146/annurev.biochem.75.101304.123901.

– (2017). "Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade". In: *Annual Review of Biochemistry* 86.1, pp. 27–68. DOI: 10.1146/annurev-biochem-061516-045115.

Chong, S. H. et al. (2015). "Structural and Thermodynamic Characteristics of Amyloidogenic Intermediates of β2-Microglobulin". In: *Sci. Rep.* 5, p. 13631. DOI: 10.1038/srep13631.

Chowdhury, Swapan K., Viswanatham Katta, and Brian T. Chait (1990). "Probing conformational changes in proteins by mass spectrometry". In: *Journal of the American Chemical Society* 112.24, pp. 9012–9013. DOI: 10.1021/ja00180a074.

Cianci, Michele et al. (2015). "Structural evidence for asymmetric ligand binding to transthyretin". In: *Acta Crystallographica Section D* 71.8, pp. 1582–1592. DOI: 10.1107/S1399004715010585.

Connelly, Gregory P. et al. (1993). "Isotope effects in peptide group hydrogen exchange". In: *Proteins: Structure, Function, and Bioinformatics* 17.1, pp. 87–92. DOI: 10.1002/prot.340170111.

Corazza, Alessandra et al. (2010). "Native-unlike Long-lived Intermediates along the Folding Pathway of the Amyloidogenic Protein β2-Microglobulin Revealed

by Real-time Two-dimensional NMR". In: *Journal of Biological Chemistry* 285.8, pp. 5827–5835. DOI: 10.1074/jbc.M109.061168.

Craig, Patricio O. et al. (2011). "Prediction of Native-State Hydrogen Exchange from Perfectly Funneled Energy Landscapes". In: *Journal of the American Chemical Society* 133.43, pp. 17463–17472. DOI: 10.1021/ja207506z.

Cramer, Christopher J. and Donald G. Truhlar (1999). "Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics". In: *Chemical Reviews* 99.8, pp. 2161–2200. DOI: 10.1021/cr960149m.

Cronkite-Ratcliff, Bryce and Vijay Pande (2013). "MSMExplorer: visualizing Markov state models for biomolecule folding simulations". In: *Bioinformatics* 29.7, pp. 950–952. DOI: 10.1093/bioinformatics/btt051.

Da, Lin-Tai et al. (2014). "Application of Markov State Models to Simulate Long Timescale Dynamics of Biological Macromolecules". In: *Protein Conformational Dynamics*. Ed. by Ke-li Han, Xin Zhang, and Ming-jun Yang. Cham: Springer International Publishing, pp. 29–66. DOI: 10.1007/978-3-319-02970-2_2.

Darden, Tom, Darrin York, and Lee Pedersen (1993). "Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems". In: *The Journal of Chemical Physics* 98.12, pp. 10089–10092. DOI: 10.1063/1.464397.

Decherchi, Sergio et al. (2015). "Implicit solvent methods for free energy estimation". In: *European Journal of Medicinal Chemistry* 91.Supplement C, pp. 27 –42. DOI: 10.1016/j.ejmech.2014.08.064.

Delano, Warren L. (2002). *The PyMOL Molecular Graphics System, Version 1.7.3.0 Schrödinger, LLC*. URL: http://www.pymol.org/.

Deng, Nan-Jie et al. (2006). "Molecular basis for the Cu^{2+} binding-induced destabilization of β2-microglobulin revealed by molecular dynamics simulation". In: *Biophysical Journal* 90.11, pp. 3865 –3879. DOI: 10.1529/biophysj.105.064444.

Dobson, Christopher M. (2002). "Protein-misfolding diseases: Getting out of shape". In: *Nature* 418.6899, pp. 729–730. DOI: 10.1038/418729a.

Dongmo Foumthuim, Cedrix J. et al. (2017). "Molecular dynamics simulations of β2-microglobulin interaction with hydrophobic surfaces". In: *Mol. BioSyst.* 13, pp. 2625–2637. DOI: 10.1039/C7MB00464H.

Dror, Ron O. et al. (2010). "Exploring atomic resolution physiology on a femtosecond to millisecond timescale using molecular dynamics simulations". In: *The Journal of General Physiology* 135.6, pp. 555–562. DOI: 10.1085/jgp.200910373.

Dror, Ron O. et al. (2012). "Biomolecular Simulation: A Computational Microscope for Molecular Biology". In: *Annual Review of Biophysics* 41.1, pp. 429–452. DOI: 10.1146/annurev-biophys-042910-155245.

Dubrey, Simon, Elizabeth Ackermann, and Julian Gillmore (2015). "The transthyretin amyloidoses: advances in therapy". In: *Postgraduate Medical Journal* 91.1078, pp. 439–448. DOI: 10.1136/postgradmedj-2014-133224.

Eanes, E. D. and G. G. Glenner (1968). "X-ray Diffraction Studies On Amyloid Filaments". In: *Journal of Histochemistry & Cytochemistry* 16.11, pp. 673–677. DOI: 10.1177/16.11.673.

Edholm, O. and H.J.C. Berendsen (1984). "Entropy estimation from simulations of non-diffusive systems". In: *Molecular Physics* 51.4, pp. 1011–1028. DOI: 10.1080/00268978400100661.

Eichner, Timo and Sheena E. Radford (2009). "A Generic Mechanism of β2-Microglobulin Amyloid Assembly at Neutral pH Involving a Specific Proline Switch". In: *Journal of Molecular Biology* 386.5, pp. 1312 –1326. DOI: 10.1016/j.jmb.2009.01.013.

– (2011). "Understanding the complex mechanisms of β2-microglobulin amyloid assembly". In: *FEBS Journal* 278.20, pp. 3868–3883. DOI: 10.1111/j.1742-4658.2011.08186.x.

Eichner, Timo et al. (2011). "Conformational Conversion during Amyloid Formation at Atomic Resolution". In: *Molecular Cell* 41.2, pp. 161 –172. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2010.11.028.

Ellis, R John and F Ulrich Hartl (1999). "Principles of protein folding in the cellular environment". In: *Current Opinion in Structural Biology* 9.1, pp. 102 –110. DOI: 10.1016/S0959-440X(99)80013-X.

Englander, S. Walter and Neville R. Kallenbach (1983). "Hydrogen exchange and structural dynamics of proteins and nucleic acids". In: *Quarterly Reviews of Biophysics* 16.4, 521–655. DOI: 10.1017/S0033583500005217.

Englander, S. Walter, Leland Mayne, and Mallela M.G. Krishna (2007). "Protein folding and misfolding: mechanism and principles". In: *Quarterly Reviews of Biophysics* 40.4, 1–41. DOI: 10.1017/S0033583508004654.

Esposito, G. et al. (2000). "Removal of the N-terminal hexapeptide from human β2-microglobulin facilitates protein aggregation and fibril formation". In: *Protein Science* 9.5, pp. 831–845. DOI: 10.1110/ps.9.5.831.

Esposito, Gennaro et al. (2005). "Solution structure of β2-microglobulin and insights into fibrillogenesis". In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1753.1, pp. 76 –84. DOI: 10.1016/j.bbapap.2005.07.003.

Essmann, Ulrich et al. (1995). "A smooth particle mesh Ewald method". In: *The Journal of Chemical Physics* 103.19, pp. 8577–8593. DOI: 10.1063/1.470117.

Feig, Michael and Charles L Brooks (2004). "Recent advances in the development and application of implicit solvent models in biomolecule simulations". In: *Current Opinion in Structural Biology* 14.2, pp. 217 –224. DOI: 10.1016/j.sbi.2004.03.009.

Fenley, Andrew T. et al. (2014). "Correlation as a Determinant of Configurational Entropy in Supramolecular and Protein Systems". In: *The Journal of Physical Chemistry B* 118.24, pp. 6447–6455. DOI: 10.1021/jp411588b.

Ferguson, David M. (1995). "Parameterization and evaluation of a flexible water model". In: *Journal of Computational Chemistry* 16.4, pp. 501–511. DOI: 10.1002/jcc.540160413.

Ferreira, Priscila et al. (2013). "Structure-Based Analysis of A19D, a Variant of Transthyretin Involved in Familial Amyloid Cardiomyopathy". In: *PLOS ONE* 8.12. DOI: 10.1371/journal.pone.0082484.

Ferrenberg, Alan M. and Robert H. Swendsen (1989). "Optimized Monte Carlo data analysis". In: *Phys. Rev. Lett.* 63 (12), pp. 1195–1198. DOI: 10.1103/PhysRevLett.63.1195.

Finkelstein, Alexei V. and Joël Janin (1989). "The price of lost freedom: entropy of bimolecular complex formation". In: *Protein Engineering, Design and Selection* 3.1, p. 1. DOI: 10.1093/protein/3.1.1.

Floege, J et al. (1991). "Clearance and synthesis rates of β2-microglobulin in patients undergoing hemodialysis and in normal subjects." In: *The Journal of laboratory and clinical medicine* 118 2, pp. 153–65. URL: https://www.ncbi.nlm.nih.gov/pubmed/1856578.

Fogolari, F., A. Brigo, and H. Molinari (2002). "The Poisson–Boltzmann equation for biomolecular electrostatics: a tool for structural biology". In: *Journal of Molecular Recognition* 15.6, pp. 377–392. DOI: 10.1002/jmr.577.

Fogolari, F. et al. (2012). "Studying interactions by molecular dynamics simulations at high concentration". In: *J. Biomed. Biotechnol.* 2012. DOI: 10.1155/2012/303190.

Fogolari, F. et al. (2017). "PDB2ENTROPY: entropy calculation from conformational ensembles". In: *Submitted* xx.x, pp. xx–xx.

Fogolari, Federico, Alessandra Corazza, and Gennaro Esposito (2013). "Generalized Born forces: Surface integral formulation". In: *The Journal of Chemical Physics* 138.5, p. 054112. DOI: 10.1063/1.4789537.

– (2015). "The Accuracy of Generalized Born Forces". In: *Computational Electrostatics for Biological Applications: Geometric and Numerical Approaches to the Description of Electrostatic Interaction Between Macromolecules*. Ed. by Walter Rocchia and Michela Spagnuolo. Cham: Springer International Publishing, pp. 143–155. DOI: 10.1007/978-3-319-12211-3_7.

– (2018). "Free energy, enthalpy and entropy from implicit solvent end-point simulations". In: *Front. Mol. Biosci.* Pp. x –xx. DOI: 10.3389/fmolb.2018.00011.

Fogolari, Federico et al. (1998). "pKa Shift Effects on Backbone Amide Base-Catalyzed Hydrogen Exchange Rates in Peptides". In: *Journal of the American Chemical Society* 120.15, pp. 3735–3738. DOI: 10.1021/ja963133m.

Fogolari, Federico et al. (1999). "Biomolecular Electrostatics with the Linearized Poisson-Boltzmann Equation". In: *Biophysical Journal* 76.1, pp. 1 –16. DOI: 10.1016/S0006-3495(99)77173-0.

Fogolari, Federico et al. (2007). "Molecular Dynamics Simulation Suggests Possible Interaction Patterns at Early Steps of β2-Microglobulin Aggregation". In: *Biophysical Journal* 92.5, pp. 1673 –1681. DOI: 10.1529/biophysj.106.098483.

Fogolari, Federico et al. (2011). "Molecular dynamics simulation of β2-microglobulin in denaturing and stabilizing conditions". In: *Proteins: Structure, Function, and Bioinformatics* 79.3, pp. 986–1001. DOI: 10.1002/prot.22940.

Fogolari, Federico et al. (2015). "Distance-Based Configurational Entropy of Proteins from Molecular Dynamics Simulations". In: *PLOS ONE* 10.7, pp. 1–26. DOI: 10.1371/journal.pone.0132356.

Fogolari, Federico et al. (2016). "Accurate Estimation of the Entropy of Rotation–Translation Probability Distributions". In: *Journal of Chemical Theory and Computation* 12.1, pp. 1–8. DOI: 10.1021/acs.jctc.5b00731.

Foss, Ted R., R. Luke Wiseman, and Jeffery W. Kelly (2005). "The Pathway by Which the Tetrameric Protein Transthyretin Dissociates". In: *Biochemistry* 44.47, pp. 15525–15533. DOI: 10.1021/bi051608t.

Frank, Joachim et al. (1995). "A model of protein synthesis based on cryo-electron microscopy of the E. coli ribosome". In: *Nature* 376.6539, pp. 441–444. DOI: 10.1038/376441a0.

Fraser, Claire M. et al. (1995). "The Minimal Gene Complement of Mycoplasma genitalium". In: *Science* 270.5235, pp. 397–404. DOI: 10.1126/science.270.5235.397.

Frenkel, Daan and Berend Smit (2002). In: *Understanding Molecular Simulation (Second Edition)*. Ed. by Daan Frenkel, and Berend Smit. Second Edition. San Diego: Academic Press. DOI: 10.1016/B978-012267351-1/50003-1.

García, Angel E. and Gerhard Hummer (1999). "Conformational dynamics of cytochrome c: Correlation to hydrogen exchange". In: *Proteins: Structure, Function, and Bioinformatics* 36.2, pp. 175–191. DOI: 10.1002/(SICI)1097-0134(19990801)36:2<175::AID-PROT4>3.0.CO;2-R.

Gejyo, F. et al. (1985). "A new form of amyloid protein associated with chronic hemodialysis was identified as β2-microglobulin". In: *Biochemical and Biophysical Research Communications* 129.3, pp. 701 –706. DOI: 10.1016/0006-291X(85)91948-5.

Ghosh, Avijit, Chaya Sendrovic Rapp, and Richard A. Friesner (1998). "Generalized Born Model Based on a Surface Integral Formulation". In: *The Journal of Physical Chemistry B* 102.52, pp. 10983–10990. DOI: 10.1021/jp982533o.

Gilson, M.K. et al. (1997). "The statistical-thermodynamic basis for computation of binding affinities: a critical review". In: *Biophysical Journal* 72.3, pp. 1047 – 1069. DOI: 10.1016/S0006-3495(97)78756-3.

Giorgetti, Sofia et al. (2005). "β2-Microglobulin isoforms display an heterogeneous affinity for type I collagen". In: *Protein Science* 14.3, pp. 696–702. DOI: 10.1110/ps.041194005.

Greene, Michael J. et al. (2015). "Cooperative Stabilization of Transthyretin by Clusterin and Diflunisal". In: *Biochemistry* 54.2, pp. 268–278. DOI: 10.1021/bi5011249.

Grycuk, Tomasz (2003). "Deficiency of the Coulomb-field approximation in the generalized Born model: An improved formula for Born radii evaluation". In: *The Journal of Chemical Physics* 119.9, pp. 4817–4826. DOI: 10.1063/1.1595641.

Gsponer, Joerg et al. (2006). "Determination of an ensemble of structures representing the intermediate state of the bacterial immunity protein Im7". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.1, pp. 99–104. DOI: 10.1073/pnas.0508667102.

Gunsteren, W. F. van and H. J. C Berendsen (1987). "Gromos-87 manual". In: Biomos BV: Nijenborgh 4, 9747 AG Groningen, The Netherlands.

Gümral, Devrim et al. (2013). "Reduction of conformational mobility and aggregation in W60G β2-microglobulin: assessment by ^{15}N NMR relaxation". In: *Magnetic Resonance in Chemistry* 51.12, pp. 795–807. DOI: 10.1002/mrc.4018.

Ha, Taekjip et al. (1999). "Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism". In: *Proceedings of the National Academy of Sciences* 96.3, pp. 893–898. DOI: 10.1073/pnas.96.3.893.

Hamilton, J.A. and M.D. Benson (2001). "Transthyretin: a review from a structural perspective". In: *Cellular and Molecular Life Sciences CMLS* 58.10, pp. 1491–1521. DOI: 10.1007/PL00000791.

Hartl, F. Ulrich and Manajit Hayer-Hartl (2002). "Molecular Chaperones in the Cytosol: from Nascent Chain to Folded Protein". In: *Science* 295.5561, pp. 1852–1858. DOI: 10.1126/science.1068408.

Haupt, Melina et al. (2014). "Binding site asymmetry in human transthyretin: insights from a joint neutron and X-ray crystallographic analysis using perdeuterated protein". In: *IUCrJ* 1.6, pp. 429–438. DOI: 10.1107/S2052252514021113.

Hawkins, Gregory D., Christopher J. Cramer, and Donald G. Truhlar (1995). "Pairwise solute descreening of solute charges from a dielectric medium". In: *Chemical Physics Letters* 246.1, pp. 122 –129. DOI: 10.1016/0009-2614(95)01082-K.

– (1996). "Parametrized Models of Aqueous Free Energies of Solvation Based on Pairwise Descreening of Solute Atomic Charges from a Dielectric Medium". In: *The Journal of Physical Chemistry* 100.51, pp. 19824–19839. DOI: 10.1021/jp961710n.

Hess, Berk et al. (1997). "LINCS: A linear constraint solver for molecular simulations". In: *Journal of Computational Chemistry* 18.12, pp. 1463–1472. DOI: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

Hess, Berk et al. (2008). "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation". In: *Journal of Chemical Theory and Computation* 4.3, pp. 435–447. DOI: 10.1021/ct700301q.

Hnizdo, Vladimir et al. (2007). "Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules". In: *Journal of Computational Chemistry* 28.3, pp. 655–668. DOI: 10.1002/jcc.20589.

Hnizdo, Vladimir et al. (2008). "Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods". In: *Journal of Computational Chemistry* 29.10, pp. 1605–1614. DOI: 10.1002/jcc.20919.

Hoover, William G. (1985). "Canonical dynamics: Equilibrium phase-space distributions". In: *Phys. Rev. A* 31 (3), pp. 1695–1697. DOI: 10.1103/PhysRevA.31.1695.

Hoshino, Junichi et al. (2016). "Significance of the decreased risk of dialysis-related amyloidosis now proven by results from Japanese nationwide surveys in 1998 and 2010". In: *Nephrology Dialysis Transplantation* 31.4, pp. 595–602. DOI: 10.1093/ndt/gfv276.

Hovmöller, Sven, Tuping Zhou, and Tomas Ohlson (2002). "Conformations of amino acids in proteins". In: *Acta Crystallographica Section D* 58.5, pp. 768–776. DOI: 10.1107/S0907444902003359.

Hsu, Yuan-Hao et al. (2013). "Fluoroketone Inhibition of Ca^{2+}-Independent Phospholipase A_2 through Binding Pocket Association Defined by Hydrogen/Deuterium Exchange and Molecular Dynamics". In: *Journal of the American Chemical Society* 135.4, pp. 1330–1337. DOI: 10.1021/ja306490g.

Hub, Jochen S., Bert L. de Groot, and David van der Spoel (2010). "g_wham-A Free Weighted Histogram Analysis Implementation Including Robust Error

and Autocorrelation Estimates". In: *Journal of Chemical Theory and Computation* 6.12, pp. 3713–3720. DOI: 10.1021/ct100494z.

Huggins, David J. (2014). "Estimating Translational and Orientational Entropies Using the k-Nearest Neighbors Algorithm". In: *Journal of Chemical Theory and Computation* 10.9, pp. 3617–3625. DOI: 10.1021/ct500415g.

Huggins, DavidJ. (2015). "Quantifying the Entropy of Binding for Water Molecules in Protein Cavities by Computing Correlations". In: *Biophysical Journal* 108.4, pp. 928 –936. DOI: 10.1016/j.bpj.2014.12.035.

Humphrey, William, Andrew Dalke, and Klaus Schulten (1996). "VMD: Visual molecular dynamics". In: *Journal of Molecular Graphics* 14.1, pp. 33 –38. DOI: 10.1016/0263-7855(96)00018-5.

Huynh, Du Q. (2009). "Metrics for 3D Rotations: Comparison and Analysis". In: *Journal of Mathematical Imaging and Vision* 35.2, pp. 155–164. DOI: 10.1007/s10851-009-0161-2.

Hvidt, Aase and K. Linderstrøm-Lang (1954). "Exchange of hydrogen atoms in insulin with deuterium atoms in aqueous solutions". In: *Biochimica et Biophysica Acta* 14.Supplement C, pp. 574 –575. DOI: 10.1016/0006-3002(54)90241-3.

Hvidt, Aase and Sigurd O. Nielsen (1966). "Hydrogen Exchange in Proteins". In: ed. by C.B. Anfinsen et al. Vol. 21. Advances in Protein Chemistry Supplement C. Academic Press, pp. 287 –386. DOI: 10.1016/S0065-3233(08)60129-1.

Hwang, Tsang-Lin, Peter C.M. van Zijl, and Susumu Mori (1998). "Accurate Quantitation of Water-amide Proton Exchange Rates Using the Phase-Modulated CLEAN Chemical EXchange (CLEANEX-PM) Approach with a Fast-HSQC (FHSQC) Detection Scheme". In: *Journal of Biomolecular NMR* 11.2, pp. 221–226. DOI: 10.1023/A:1008276004875.

Hwang, Tsang-Lin et al. (1997). "Application of Phase-Modulated CLEAN Chemical EXchange Spectroscopy (CLEANEX-PM) to Detect WaterProtein Proton Exchange and Intermolecular NOEs". In: *Journal of the American Chemical Society* 119.26, pp. 6203–6204. DOI: 10.1021/ja970160j.

Hörnberg, Andreas et al. (2000). "A comparative analysis of 23 structures of the amyloidogenic protein transthyretin". In: *Journal of Molecular Biology* 302.3, pp. 649 –669. DOI: 10.1006/jmbi.2000.4078.

Ikenoue, Tatsuya et al. (2014). "Heat of supersaturation-limited amyloid burst directly monitored by isothermal titration calorimetry". In: *Proceedings of the National Academy of Sciences* 111.18, pp. 6654–6659. DOI: 10.1073/pnas.1322602111.

Invernizzi, Gaetano et al. (2012). "Protein aggregation: Mechanisms and functional consequences". In: *The International Journal of Biochemistry Cell Biology* 44.9, pp. 1541 –1554. DOI: 10.1016/j.biocel.2012.05.023.

Jacobs, T. M. et al. (2016). "Design of structurally distinct proteins using strategies inspired by evolution". In: *Science* 352.6286, pp. 687–690. DOI: 10.1126/science.aad8036.

Jensen, Pernille Foged and Kasper D. Rand (2016). "Hydrogen Exchange". In: *Hydrogen Exchange Mass Spectrometry of Proteins*. John Wiley Sons, Ltd, pp. 1–17. DOI: 10.1002/9781118703748.ch1.

Jiang, Xin, Joel N. Buxbaum, and Jeffery W. Kelly (2001). "The V122I Cardiomyopathy Variant of Transthyretin Increases the Velocity of Rate-Limiting Tetramer Dissociation, Resulting in Accelerated Amyloidosis". In: *Proceedings of the National Academy of Sciences of the United States of America* 98.26, pp. 14943–14948. DOI: 10.2307/3057390.

Jiang, Xin et al. (2001). "An Engineered Transthyretin Monomer that Is Nonamyloidogenic, Unless It Is Partially Denatured". In: *Biochemistry* 40.38, pp. 11442–11452. DOI: 10.1021/bi011194d.

Johansson, Maria U. et al. (2012). "Defining and searching for structural motifs using DeepView/Swiss-PdbViewer". In: *BMC Bioinformatics* 13.1, p. 173. DOI: 10.1186/1471-2105-13-173.

Johnson, Steven M. et al. (2012). "The Transthyretin Amyloidoses: From Delineating the Molecular Mechanism of Aggregation Linked to Pathology to a Regulatory-Agency-Approved Drug". In: *Journal of Molecular Biology* 421.2, pp. 185 –203. DOI: 10.1016/j.jmb.2011.12.060.

Jorgensen, William L. and Jeffry D. Madura (1985). "Temperature and size dependence for Monte Carlo simulations of TIP4P water". In: *Molecular Physics* 56.6, pp. 1381–1392. DOI: 10.1080/00268978500103111.

Jorgensen, William L. and Julian Tirado-Rives (1988). "The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin". In: *Journal of the American Chemical Society* 110.6, pp. 1657–1666. DOI: 10.1021/ja00214a001.

Jorgensen, William L. et al. (1983). "Comparison of simple potential functions for simulating liquid water". In: *The Journal of Chemical Physics* 79.2, pp. 926–935. DOI: 10.1063/1.445869.

Kabsch, Wolfgang and Christian Sander (1983). "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features". In: *Biopolymers* 22.12, pp. 2577–2637. DOI: 10.1002/bip.360221211.

Kalé, Laxmikant et al. (1999). "NAMD2: Greater Scalability for Parallel Molecular Dynamics". In: *Journal of Computational Physics* 151.1, pp. 283 –312. ISSN: 0021-9991. DOI: 10.1006/jcph.1999.6201.

Kamerlin, Shina C. L. and Arieh Warshel (2010). "At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis?" In: *Proteins: Structure, Function, and Bioinformatics* 78.6, pp. 1339–1375. DOI: 10.1002/prot.22654.

Karplus, Martin and J. Andrew McCammon (2002). "Molecular dynamics simulations of biomolecules". In: *Nat Struct Mol Biol* 9.9, pp. 646–652. DOI: 10.1038/nsb0902-646.

Kelly, Jeffery W. and Peter T. Lansbury (1994). "A chemical approach to elucidate tin mechanism of transthyretin and β-protein amyloid fibril formation". In: *Amyloid* 1.3, pp. 186–205. DOI: 10.3109/13506129409148451.

Kendrew, J. C. et al. (1958). "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis". In: *Nature* 181.4610, pp. 662–666. DOI: 10.1038/181662a0.

Kieseritzky, Gernot, Giulia Morra, and Ernst-Walter Knapp (2006). "Stability and fluctuations of amide hydrogen bonds in a bacterial cytochrome c: a molecular

dynamics study". In: *JBIC Journal of Biological Inorganic Chemistry* 11.1, pp. 26–40. DOI: 10.1007/s00775-005-0041-1.

Killian, Benjamin J., Joslyn Yundenfreund Kravitz, and Michael K. Gilson (2007). "Extraction of configurational entropy from molecular simulations via an expansion approximation". In: *The Journal of Chemical Physics* 127.2, p. 024107. DOI: 10.1063/1.2746329.

King, Bracken M., Nathaniel W. Silver, and Bruce Tidor (2012). "Efficient Calculation of Molecular Configurational Entropies Using an Information Theoretic Approximation". In: *The Journal of Physical Chemistry B* 116.9, pp. 2891–2904. DOI: 10.1021/jp2068123.

King, Bracken M. and Bruce Tidor (2009). "MIST: Maximum Information Spanning Trees for dimension reduction of biological data sets". In: *Bioinformatics* 25.9, p. 1165. DOI: 10.1093/bioinformatics/btp109.

Kitayama, Hiroki et al. (2013). "A common mechanism underlying amyloid fibrillation and protein crystallization revealed by the effects of ultrasonication". In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1834.12, pp. 2640 –2646. DOI: 10.1016/j.bbapap.2013.09.016.

Kleckner, Ian R. and Mark P. Foster (2011). "An introduction to NMR-based approaches for measuring protein dynamics". In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1814.8, pp. 942 –968. DOI: 10.1016/j.bbapap.2010.10.012.

Kleinjung, Jens and Franca Fraternali (2014). "Design and application of implicit solvent models in biomolecular simulations". In: *Current Opinion in Structural Biology* 25.Supplement C, pp. 126 –134. DOI: 10.1016/j.sbi.2014.04.003.

Knowles, Tuomas P. et al. (2007). "Role of Intermolecular Forces in Defining Material Properties of Protein Nanofibrils". In: *Science* 318.5858, pp. 1900–1903. DOI: 10.1126/science.1150057.

Koehl, Patrice (2006). "Protein Structure Classification". In: *Reviews in Computational Chemistry*. John Wiley Sons, Inc., pp. 1–55. DOI: 10.1002/0471780367.ch1.

Kuffner, J. J. (2004). "Effective sampling and distance metrics for 3D rigid body path planning". In: *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*. Vol. 4, pp. 3993–3998. DOI: 10.1109/ROBOT.2004.1308895.

Kumar, Shankar et al. (1992). "THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method". In: *Journal of Computational Chemistry* 13.8, pp. 1011–1021. DOI: 10.1002/jcc.540130812.

Kästner, Johannes (2011). "Umbrella sampling". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.6, pp. 932–942. DOI: 10.1002/wcms.66.

Lai, Zhihong, Wilfredo Colón, and Jeffery W. Kelly (1996). "The Acid-Mediated Denaturation Pathway of Transthyretin Yields a Conformational Intermediate That Can Self-Assemble into Amyloid". In: *Biochemistry* 35.20, pp. 6470–6482. DOI: 10.1021/bi952501g.

Laio, Alessandro and Francesco L Gervasio (2008). "Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science". In: *Reports on Progress in Physics* 71, p. 126601. DOI: 10.1088/0034-4885/71/12/126601.

Laio, Alessandro and Michele Parrinello (2002). "Escaping free-energy minima". In: *Proceedings of the National Academy of Sciences* 99.20, pp. 12562–12566. DOI: 10.1073/pnas.202427399.

Laio, Alessandro et al. (2005). "Assessing the Accuracy of Metadynamics". In: *The Journal of Physical Chemistry B* 109.14, pp. 6714–6721. DOI: 10.1021/jp045424k.

Lane, Thomas J et al. (2013). "To milliseconds and beyond: challenges in the simulation of protein folding". In: *Current Opinion in Structural Biology* 23.1. Folding and binding / Protein-nucleic acid interactions, pp. 58 –65. DOI: 10.1016/j.sbi.2012.11.002.

Lansbury, Peter T. et al. (1995). "Structural model for the β-amyloid fibril based on interstrand alignment of an antiparallel-sheet comprising a C-terminal peptide". In: *Nat Struct Mol Biol* 2.11, pp. 990–998. DOI: 10.1038/nsb1195-990.

Lashuel, Hilal A., Zhihong Lai, and Jeffery W. Kelly (1998). "Characterization of the Transthyretin Acid Denaturation Pathways by Analytical Ultracentrifugation: Implications for Wild-Type, V30M, and L55P Amyloid Fibril Formation". In: *Biochemistry* 37.51, pp. 17851–17864. DOI: 10.1021/bi981876+.

Lei, Hongxing and Yong Duan (2007). "Improved sampling methods for molecular simulation". In: *Current Opinion in Structural Biology* 17.2, pp. 187 –191. DOI: 10.1016/j.sbi.2007.03.003.

Levinthal, Cyrus (1968). "Are there pathways for protein folding?" In: *J. Chim. Phys.* 65, pp. 44–45. DOI: 10.1051/jcp/1968650044.

Li, Renhao and Clare Woodward (1999). "The hydrogen exchange core and protein folding". In: *Protein Science* 8.8, pp. 1571–1590. DOI: 10.1110/ps.8.8.1571.

Lim, Kwang Hun et al. (2013). "Localized Structural Fluctuations Promote Amyloidogenic Conformations in Transthyretin". In: *Journal of Molecular Biology* 425.6, pp. 977 –988. DOI: 10.1016/j.jmb.2013.01.008.

Lin, Yuxi et al. (2014). "Solubility and Supersaturation-Dependent Protein Misfolding Revealed by Ultrasonication". In: *Langmuir* 30.7, pp. 1845–1854. DOI: 10.1021/la403100h.

Lindorff-Larsen, Kresten et al. (2010). "Improved side-chain torsion potentials for the Amber ff99SB protein force field". In: *Proteins: Structure, Function, and Bioinformatics* 78.8, pp. 1950–1958. DOI: 10.1002/prot.22711.

Lindorff-Larsen, Kresten et al. (2012). "Systematic Validation of Protein Force Fields against Experimental Data". In: *PLOS ONE* 7.2, pp. 1–6. DOI: 10.1371/journal.pone.0032131.

Ma, Buyong and Ruth Nussinov (2003). "Molecular dynamics simulations of the unfolding of β2microglobulin and its variants". In: *Protein Engineering, Design and Selection* 16.8, pp. 561–575. DOI: 10.1093/protein/gzg079.

– (2006). "Simulations as analytical tools to understand protein aggregation and predict amyloid conformation". In: *Current Opinion in Chemical Biology* 10.5, pp. 445 –452. DOI: 10.1016/j.cbpa.2006.08.018.

MacKerell, A. D., M. Feig, and C. L. Brooks (2004). "Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations". In: *J. Comput. Chem.* 25, pp. 1400–1415. DOI: 10.1002/jcc.20065.

MacKerell, A. D. et al. (1998). "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins". In: *The Journal of Physical Chemistry B* 102.18, pp. 3586–3616. DOI: 10.1021/jp973084f.

Mahoney, Michael W. and William L. Jorgensen (2000). "A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions". In: *The Journal of Chemical Physics* 112.20, pp. 8910–8922. DOI: 10.1063/1.481505.

Mangione, P. Patrizia et al. (2013). "Structure, Folding Dynamics, and Amyloidogenesis of D76N β2-Microglobulin: ROLES OF SHEAR FLOW, HYDROPHOBIC SURFACES, AND α-CRYSTALLIN". In: *Journal of Biological Chemistry* 288.43, pp. 30917–30930. DOI: 10.1074/jbc.M113.498857.

McAllister, Robert G. and Lars Konermann (2015). "Challenges in the Interpretation of Protein H/D Exchange Data: A Molecular Dynamics Simulation Perspective". In: *Biochemistry* 54.16, pp. 2683–2692. DOI: 10.1021/acs.biochem.5b00215.

McParland, Victoria J. et al. (2000). "Partially Unfolded States of β2-Microglobulin and Amyloid Formation in Vitro". In: *Biochemistry* 39.30, pp. 8735–8746. DOI: 10.1021/bi000276j.

McQuarrie, Donald A. (1976). In: *Statistical Mechanics*. Ed. by Stuart A. Rice. 10 East 53rd Street, New York, N.Y. 10022: Harper & Row Publisher Inc.

Mendoza, Vanessa Leah et al. (2011). "Structural Insights into the Pre-Amyloid Tetramer of -2-Microglobulin from Covalent Labeling and Mass Spectrometry". In: *Biochemistry* 50.31, pp. 6711–6722. DOI: 10.1021/bi2004894.

Michalet, Xavier, Shimon Weiss, and Marcus Jäger (2006). "Single-Molecule Fluorescence Studies of Protein Folding and Conformational Dynamics". In: *Chemical Reviews* 106.5, pp. 1785–1813. DOI: 10.1021/cr0404343.

Miles, R. E. (1965). "On random rotations in R3". In: *Biometrika* 52.3-4, p. 636. DOI: 10.1093/biomet/52.3-4.636.

Molday, R. S., S. W. Englander, and R. G. Kallen (1972). "Primary structure effects on peptide group hydrogen exchange". In: *Biochemistry* 11.2, pp. 150–158. DOI: 10.1021/bi00752a003.

Morgan, Charles J et al. (2001). "Kidney dialysis-associated amyloidosis: a molecular role for copper in fiber formation11Edited by R. Ebright". In: *Journal of Molecular Biology* 309.2, pp. 339 –345. DOI: 10.1006/jmbi.2001.4661.

Moroni, Elisabetta, Guido Scarabelli, and Giorgio Colombo (2009). "Structure and sequence determinants of aggregation investigated with molecular dynamics". In: *Frontiers in Bioscience* 14, pp. 523–539. DOI: 10.2741/3260.

Myers, Sarah L. et al. (2006). "A Systematic Study of the Effect of Physiological Factors on β2-Microglobulin Amyloid Formation at Neutral pH". In: *Biochemistry* 45.7, pp. 2311–2321. DOI: 10.1021/bi052434i.

Naiki, Hironobu et al. (2016). "Molecular pathogenesis of human amyloidosis: Lessons from β2-microglobulin-related amyloidosis". In: *Pathology International* 66.4, pp. 193–201. DOI: 10.1111/pin.12394.

Nakamura, Masaaki and Yukio Ando (2004). "Applications of gene therapy for familial amyloidotic polyneuropathy". In: *Expert Opinion on Biological Therapy* 4.10, pp. 1621–1627. DOI: 10.1517/14712598.4.10.1621.

Narang, Dominic, Anubhuti Singh, and Samrat Mukhopadhyay (2017). "Stepwise unfolding of human β2-microglobulin into a disordered amyloidogenic precursor at low pH". In: *European Biophysics Journal* 46.1, pp. 65–76. DOI: 10.1007/s00249-016-1138-x.

Nilsson, Melanie R (2004). "Techniques to study amyloid fibril formation in vitro". In: *Methods* 34.1. Investigating Protein Folding, Misfolding and Nonnative States: Experimental and Theoretical Methods, pp. 151 –160. DOI: 10.1016/j.ymeth.2004.03.012.

Nosé, Shuichi (1984). "A molecular dynamics method for simulations in the canonical ensemble". In: *Molecular Physics* 52.2, pp. 255–268. DOI: 10.1080/00268978400101201.

Nosé, Shuichi and M.L. Klein (1983). "Constant pressure molecular dynamics for molecular systems". In: *Molecular Physics* 50.5, pp. 1055–1076. DOI: 10.1080/00268978300102851.

Ohta, Yuko et al. (2011). "Primordial Linkage of β2-Microglobulin to the MHC". In: *The Journal of Immunology* 186.6, pp. 3563–3571. DOI: 10.4049/jimmunol.1003933.

Onuchic, José Nelson et al. (1997). "Theory Of Protein Folding: The Energy Landscape Perspective". In: *Annual Review of Physical Chemistry* 48.1, pp. 545–600. DOI: 10.1146/annurev.physchem.48.1.545.

Onufriev, Alexey (2008). "Chapter 7 - Implicit Solvent Models in Molecular Dynamics Simulations: A Brief Overview". In: ed. by Ralph A. Wheeler and David C. Spellmeyer. Vol. 4. Annual Reports in Computational Chemistry Supplement C. Elsevier, pp. 125 –137. DOI: 10.1016/S1574-1400(08)00007-8.

Onufriev, Alexey, Donald Bashford, and David A. Case (2000). "Modification of the Generalized Born Model Suitable for Macromolecules". In: *The Journal of Physical Chemistry B* 104.15, pp. 3712–3720. DOI: 10.1021/jp994072s.

– (2004). "Exploring protein native states and large-scale conformational changes with a modified generalized born model". In: *Proteins: Structure, Function, and Bioinformatics* 55.2, pp. 383–394. DOI: 10.1002/prot.20033.

Palmer, Arthur G. (2015). "Enzyme Dynamics from NMR Spectroscopy". In: *Accounts of Chemical Research* 48.2, pp. 457–465. DOI: 10.1021/ar500340a.

Pan, Yinquan and Martha S. Briggs (1992). "Hydrogen exchange in native and alcohol forms of ubiquitin". In: *Biochemistry* 31.46, pp. 11405–11412. DOI: 10.1021/bi00161a019.

Paravastu, Anant K. et al. (2008). "Molecular structural basis for polymorphism in Alzheimer's -amyloid fibrils". In: *Proceedings of the National Academy of Sciences* 105.47, pp. 18349–18354. DOI: 10.1073/pnas.0806270105.

Park, F. C. and Bahram Ravani (1997). "Smooth Invariant Interpolation of Rotations". In: *ACM Trans. Graph.* 16.3, pp. 277–295. DOI: 10.1145/256157.256160.

Park, Sheldon and Jeffery G. Saven (2006). "Simulation of pH-dependent edge strand rearrangement in human β-2 microglobulin". In: *Protein Science* 15.1, pp. 200–207. DOI: 10.1110/ps.051814306.

Parrinello, M. and A. Rahman (1981). "Polymorphic transitions in single crystals: A new molecular dynamics method". In: *Journal of Applied Physics* 52.12, pp. 7182–7190. DOI: 10.1063/1.328693.

Pasquato, Nicola et al. (2007). "Acidic pH-induced Conformational Changes in Amyloidogenic Mutant Transthyretin". In: *Journal of Molecular Biology* 366.3, pp. 711 –719. DOI: 10.1016/j.jmb.2006.11.076.

Pearlman, David A. et al. (1995). "AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules". In: *Computer Physics Communications* 91.1, pp. 1 –41. DOI: 10.1016/0010-4655(95)00041-D.

Persson, Filip and Bertil Halle (2015). "How amide hydrogens exchange in native proteins". In: *Proceedings of the National Academy of Sciences* 112.33, pp. 10383–10388. DOI: 10.1073/pnas.1506079112.

Peterson, Scott A. et al. (1998). "Inhibiting transthyretin conformational changes that lead to amyloid fibril formation". In: *Proceedings of the National Academy of Sciences* 95.22, pp. 12956–12960. DOI: 10.1073/pnas.95.22.12956.

Petruk, Ariel A. et al. (2013). "Molecular Dynamics Simulations Provide Atomistic Insight into Hydrogen Exchange Mass Spectrometry Experiments". In: *Journal of Chemical Theory and Computation* 9.1, pp. 658–669. DOI: 10.1021/ct300519v.

Piana, Stefano and Alessandro Laio (2007). "A Bias-Exchange Approach to Protein Folding". In: *The Journal of Physical Chemistry B* 111.17, pp. 4553–4559. DOI: 10.1021/jp0678731.

Pirrone, Gregory F., Roxana E. Iacob, and John R. Engen (2015). "Applications of Hydrogen/Deuterium Exchange MS from 2012 to 2014". In: *Analytical Chemistry* 87.1, pp. 99–118. DOI: 10.1021/ac5040242.

Politou, Anastasia and Piero Andrea Temussi (2015). "Revisiting a dogma: the effect of volume exclusion in molecular crowding". In: *Current Opinion in Structural Biology* 30.Supplement C, pp. 1 –6. DOI: 10.1016/j.sbi.2014.10.005.

Polyansky, Anton A., Ruben Zubac, and Bojan Zagrovic (2012). "Estimation of Conformational Entropy in Protein–Ligand Interactions: A Computational Perspective". In: *Computational Drug Discovery and Design*. Ed. by Riccardo Baron. New York, NY: Springer New York, pp. 327–353. DOI: 10.1007/978-1-61779-465-0_21.

Qiu, Di et al. (1997). "The GB/SA Continuum Model for Solvation. A Fast Analytical Method for the Calculation of Approximate Born Radii". In: *The Journal of Physical Chemistry A* 101.16, pp. 3005–3014. DOI: 10.1021/jp961992r.

Quintas, Alexandre, Maria João M. Saraiva, and Rui M. M. Brito (1999). "The Tetrameric Protein Transthyretin Dissociates to a Non-native Monomer in Solution: A Novel Model For Amyloidogenesis". In: *Journal of Biological Chemistry* 274.46, pp. 32943–32949. DOI: 10.1074/jbc.274.46.32943.

Quintas, Alexandre et al. (2001). "Tetramer Dissociation and Monomer Partial Unfolding Precedes Protofibril Formation in Amyloidogenic Transthyretin Variants". In: *Journal of Biological Chemistry* 276.29, pp. 27207–27213. DOI: 10.1074/jbc.M101024200.

Rabindranath, Kannaiyan S. et al. (2005). "Comparison of hemodialysis, hemofiltration, and acetate-free biofiltration for ESRD: Systematic review". In: *American Journal of Kidney Diseases* 45.3, pp. 437–447. DOI: 10.1053/j.ajkd.2004.11.008.

Raffaini, Giuseppina and Fabio Ganazzoli (2010). "Protein Adsorption on a Hydrophobic Surface: A Molecular Dynamics Study of Lysozyme on Graphite". In: *Langmuir* 26.8, pp. 5679–5689. DOI: 10.1021/la903769c.

Raimondi, Sara et al. (2017). "A specific nanobody prevents amyloidogenesis of D76N β2-microglobulin in vitro and modifies its tissue distribution in vivo". In: *Sci. Rep.* 7, p. 46711. DOI: 10.1038/srep46711.

Rein, Brandon J. D. et al. (2011). "Potential Markers for Detection and Monitoring of Ovarian Cancer". In: *Journal of Oncology*, p. 475983. DOI: 10.1155/2011/475983.

Reinés, Juan Buades et al. (2014). "Epidemiology of transthyretin-associated familial amyloid polyneuropathy in the Majorcan area: Son Llàtzer Hospital descriptive study". In: *Orphanet Journal of Rare Diseases* 9.1, p. 29. DOI: 10.1186/1750-1172-9-29.

Rennella, Enrico et al. (2009). "Equilibrium Unfolding Thermodynamics of β-Microglobulin Analyzed through Native-State H/D Exchange". In: *Biophysical Journal* 96.1, pp. 169–179. DOI: 10.1529/biophysj.108.142448.

Rennella, Enrico et al. (2012a). "Determining the Energy Landscape of Proteins by a Fast Isotope Exchange NMR Approach". In: *Journal of the American Chemical Society* 134.10, pp. 4457–4460. DOI: 10.1021/ja209004q.

Rennella, Enrico et al. (2012b). "Single-shot NMR measurement of protein unfolding landscapes". In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1824.6, pp. 842 –849. DOI: 10.1016/j.bbapap.2012.04.002.

Roberts, Roland G. (2016). "Good Amyloid, Bad Amyloid—What's the Difference?" In: *PLOS Biology* 14.1, pp. 1–3. DOI: 10.1371/journal.pbio.1002362.

Rodrigues, J. Rui et al. (2010). "Potentially amyloidogenic conformational intermediates populate the unfolding landscape of transthyretin: Insights from molecular dynamics simulations". In: *Protein Science* 19.2, pp. 202–219. DOI: 10.1002/pro.289.

Rosano, Camillo, Simone Zuccotti, and Martino Bolognesi (2005). "The three-dimensional structure of β2-microglobulin: Results from X-ray crystallography". In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1753.1, pp. 85 –91. DOI: 10.1016/j.bbapap.2005.07.010.

Roux, Benoît (1995). "The calculation of the potential of mean force using computer simulations". In: *Computer Physics Communications* 91.1, pp. 275 –282. DOI: 10.1016/0010-4655(95)00053-I.

Roux, Benoıt and Thomas Simonson (1999). "Implicit solvent models". In: *Biophysical Chemistry* 78.1, pp. 1 –20. DOI: 10.1016/S0301-4622(98)00226-9.

Saper, M.A., P.J. Bjorkman, and D.C. Wiley (1991). "Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 Å resolution". In: *Journal of Molecular Biology* 219.2, pp. 277 –319. DOI: 10.1016/0022-2836(91)90567-P.

Scarpioni, R. et al. (2016). "Dialysis-related amyloidosis: challenges and solutions". In: *Int. J. Nephrol. Renovasc. Dis.* 9, pp. 319–328. DOI: 10.2147/IJNRD.S84784.

Schaefer, Michael, Christian Bartels, and Martin Karplus (1998). "Solution conformations and thermodynamics of structured peptides: molecular dynamics simulation with an implicit solvation model1 1Edited by A. F. Fersht". In: *Journal of Molecular Biology* 284.3, pp. 835 –848. DOI: 10.1006/jmbi.1998.2172.

Scheraga, Harold A., Mey Khalili, and Adam Liwo (2007). "Protein-Folding Dynamics: Overview of Molecular Simulation Techniques". In: *Annual Review of Physical Chemistry* 58.1, pp. 57–83. DOI: 10.1146/annurev.physchem.58.032806.104614.

Schlick, Tamar (2009). "Molecular dynamics-based approaches for enhanced sampling of long-time, large-scale conformational changes in biomolecules". In: *F1000 Biol Reports* 1.51. DOI: 10.3410/B1-51.

Schulz, Georg E. and R. Heiner Schirmer (1979a). In: *Principles of Protein Structure*. New York, NY: Springer New York.

– (1979b). "Amino Acids". In: *Principles of Protein Structure*. New York, NY: Springer New York, pp. 1–16. DOI: 10.1007/978-1-4612-6137-7_1.

– (1979c). "Noncovalent Forces Determining Protein Structure". In: *Principles of Protein Structure*. New York, NY: Springer New York, pp. 27–45. DOI: 10.1007/978-1-4612-6137-7_3.

Schymkowitz, Joost et al. (2005). "The FoldX web server: an online force field". In: *Nucleic Acids Research* 33.suppl$_2$, pp. 382–388. DOI: 10.1093/nar/gki387.

Sebastião, Maria Paula et al. (2000). "The molecular interaction of 4'-iodo-4'-deoxydoxorubicin with Leu-55Pro transthyretin 'amyloid-like' oligomer leading to disaggregation". In: *Biochemical Journal* 351.1, pp. 273–279. DOI: 10.1042/bj3510273.

Serpell, Louise C et al. (2000). "The protofilament substructure of amyloid fibrils11Edited by F. E. Cohen". In: *Journal of Molecular Biology* 300.5, pp. 1033 – 1039. DOI: 10.1006/jmbi.2000.3908.

Sethuraman, Ananthakrishnan and Georges Belfort (2005). "Protein Structural Perturbation and Aggregation on Homogeneous Surfaces". In: *Biophysical Journal* 88.2, pp. 1322 –1333. DOI: 10.1529/biophysj.104.051797.

Shan, Yibing et al. (2013). "Transitions to catalytically inactive conformations in EGFR kinase". In: *Proceedings of the National Academy of Sciences* 110.18, pp. 7270–7275. DOI: 10.1073/pnas.1220843110.

Shea, Joan-Emma, and Charles L Brooks III (2001). "From Folding Theories To Folding Proteins: A Review and Assessment of Simulation Studies of Protein Folding and Unfolding". In: *Annual Review of Physical Chemistry* 52.1, pp. 499–535. DOI: 10.1146/annurev.physchem.52.1.499.

Sheinerman, Felix B. and Charles L. Brooks (1998). "Molecular picture of folding of a small α/βprotein". In: *Proceedings of the National Academy of Sciences* 95.4, pp. 1562–1567. eprint: http://www.pnas.org/content/95/4/1562.full.pdf.

Shi, Chunmeng et al. (2009). "β2-Microglobulin: emerging as a promising cancer therapeutic target". In: *Drug Discovery Today* 14.1, pp. 25 –30. DOI: 10.1016/j.drudis.2008.11.001.

Simonson, Thomas (2001). "Macromolecular electrostatics: continuum models and their growing pains". In: *Current Opinion in Structural Biology* 11.2, pp. 243 –252. DOI: 10.1016/S0959-440X(00)00197-4.

– (2003). "Electrostatics and dynamics of proteins". In: *Reports on Progress in Physics* 66.5, p. 737. DOI: 10.1088/0034-4885/66/5/202.

Singh, Harshinder et al. (2003). "Nearest Neighbor Estimates of Entropy". In: *American Journal of Mathematical and Management Sciences* 23.3-4, pp. 301–321. DOI: 10.1080/01966324.2003.10737616.

Sipe, Jean D. and Alan S. Cohen (2000). "Review: History of the Amyloid Fibril". In: *Journal of Structural Biology* 130.2, pp. 88 –98. DOI: 10.1006/jsbi.2000.4221.

Skinner, John J. et al. (2012a). "Protein dynamics viewed by hydrogen exchange". In: *Protein Science* 21.7, pp. 996–1005. DOI: 10.1002/pro.2081.

– (2012b). "Protein hydrogen exchange: Testing current models". In: *Protein Science* 21.7, pp. 987–995. DOI: 10.1002/pro.2082.

Skinner, John J. et al. (2014). "Benchmarking all-atom simulations using hydrogen exchange". In: *Proceedings of the National Academy of Sciences* 111.45, pp. 15975–15980. DOI: 10.1073/pnas.1404213111.

Soler, Miguel A. et al. (2017). "Computational design of cyclic peptides for the customized oriented immobilization of globular proteins". In: *Phys. Chem. Chem. Phys.* 19 (4), pp. 2740–2748. DOI: 10.1039/C6CP07807A.

Souaille, Marc and Benoıt Roux (2001). "Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations". In: *Computer Physics Communications* 135.1, pp. 40 –57. DOI: 10.1016/S0010-4655(00)00215-0.

Sprik, Michiel and Michael L. Klein (1988). "A polarizable model for water using distributed charge sites". In: *The Journal of Chemical Physics* 89.12, pp. 7556–7560. DOI: 10.1063/1.455722.

Still, W. Clark et al. (1990). "Semianalytical treatment of solvation for molecular mechanics and dynamics". In: *Journal of the American Chemical Society* 112.16, pp. 6127–6129. DOI: 10.1021/ja00172a038.

Sugita, Yuji and Yuko Okamoto (1999). "Replica-exchange molecular dynamics method for protein folding". In: *Chemical Physics Letters* 314.1, pp. 141 –151. DOI: 10.1016/S0009-2614(99)01123-9.

Sunde, Margaret and Colin Blake (1997). "The Structure of Amyloid Fibrils by Electron Microscopy and X-Ray Diffraction". In: *Advances in Protein Chemistry* 50, pp. 123 –159. DOI: 10.1016/S0065-3233(08)60320-4.

Suzuki, Kazuo, Masami Shimazaki, and Hidetoshi Kutsuki (2003). "β2-Microglobulin-selective AdsorbentColumn (Lixelle) for the Treatment of Dialysis-related Amyloidosis". In: *Ther. Apher. Dial.* 7.1, pp. 104–107. DOI: 10.1046/j.1526-0968.2003.00019.x.

Svaldo-Lanero, Tiziana et al. (2008). "Nanopatterning by protein unfolding". In: *Soft Matter* 4 (5), pp. 965–967. DOI: 10.1039/B800621K.

Svergun, D.I. (1999). "Restoring Low Resolution Structure of Biological Macromolecules from Solution Scattering Using Simulated Annealing". In: *Biophysical Journal* 76.6, pp. 2879 –2886. DOI: 10.1016/S0006-3495(99)77443-6.

Taylor, William R et al. (2001). "Protein structure: geometry, topology and classification". In: *Reports on Progress in Physics* 64.4, p. 517. DOI: 10.1088/0034-4885/64/4/203.

Teague, Simon J. (2003). "Implications of protein flexibility for drug discovery". In: *Nat Rev Drug Discov* 2.7, pp. 527–541. DOI: 10.1038/nrd1129.

Tjong, Harianto and Huan-Xiang Zhou (2007a). "GBr^6: A generalized Born method for accurately reproducing solvation energy of the nonlinear Poisson-Boltzmann equation". In: *The Journal of Chemical Physics* 126.19, p. 195102. DOI: 10.1063/1.2735322.

Tjong, Harianto and Huan-Xiang Zhou (2007b). "GBr^6: A Parameterization-Free, Accurate, Analytical Generalized Born Method". In: *The Journal of Physical Chemistry B* 111.11, pp. 3055–3061. DOI: 10.1021/jp066284c.

Torrie, Glenn M. and John P. Valleau (1974). "Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid". In: *Chemical Physics Letters* 28.4, pp. 578 –581. DOI: 10.1016/0009-2614(74)80109-0.

Torrie, G.M. and J.P. Valleau (1977). "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling". In: *Journal of Computational Physics* 23.2, pp. 187 –199. DOI: 10.1016/0021-9991(77)90121-8.

Torshin, Ivan Y., Irene T. Weber, and Robert W. Harrison (2002). "Geometric criteria of hydrogen bonds in proteins and identification of 'bifurcated' hydrogen bonds". In: *Protein Engineering, Design and Selection* 15.5, pp. 359–363. DOI: 10.1093/protein/15.5.359.

Tribello, Gareth A. et al. (2014). "PLUMED 2: New feathers for an old bird". In: *Computer Physics Communications* 185.2, pp. 604 –613. DOI: 10.1016/j.cpc.2013.09.018.

Vaart, Arjan van der (2006). "Simulation of conformational transitions". In: *Theoretical Chemistry Accounts* 116.1, pp. 183–193. DOI: 10.1007/s00214-005-0004-1.

Vendruscolo, Michele et al. (2003). "Rare Fluctuations of Native Proteins Sampled by Equilibrium Hydrogen Exchange". In: *Journal of the American Chemical Society* 125.51, pp. 15686–15687. DOI: 10.1021/ja036523z.

Verdone, Giuliana et al. (2002). "The solution structure of human β2-microglobulin reveals the prodromes of its amyloid transition". In: *Protein Science* 11.3, pp. 487–499. DOI: 10.1110/ps.29002.

Verlet, Loup (1967). "Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules". In: *Phys. Rev.* 159 (1), pp. 98–103. DOI: 10.1103/PhysRev.159.98.

Wereszczynski, Jeff and J. Andrew McCammon (2012). "Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition". In: *Quarterly Reviews of Biophysics* 45.1, 1–25. DOI: 10.1017/S0033583511000096.

Woodcock, L.V. (1971). "Isothermal molecular dynamics calculations for liquid salts". In: *Chemical Physics Letters* 10.3, pp. 257 –261. DOI: 10.1016/0009-2614(71)80281-6.

Wüthrich, Kurt (1986). "NMR of Proteins and Nucleic Acids (Baker Lecture Series)". In: New York: Wiley.

Xu, Jia et al. (2015). "Phosphorylation in the Catalytic Cleft Stabilizes and Attracts Domains of a Phosphohexomutase". In: *Biophysical Journal* 108.2, pp. 325 –337. DOI: 10.1016/j.bpj.2014.12.003.

Yamamoto, Suguru and Fumitake Gejyo (2005). "Historical background and clinical treatment of dialysis-related amyloidosis". In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1753.1, pp. 4 –10. DOI: 10.1016/j.bbapap.2005.09.006.

Yamamoto, Suguru et al. (2004). "Glycosaminoglycans Enhance the Trifluoroethanol-Induced Extension of β2-Microglobulin–Related Amyloid Fibrils at a Neutral

pH". In: *Journal of the American Society of Nephrology* 15.1, pp. 126–133. DOI: 10.1097/01.ASN.0000103228.81623.C7.

Yan, Xuguang et al. (2007). "Deuterium exchange and mass spectrometry reveal the interaction differences of two synthetic modulators of RXR LBD". In: *Protein Science* 16.11, pp. 2491–2501. DOI: 10.1110/ps.073019707.

Zandomeneghi, Giorgia et al. (2004). "FTIR reveals structural differences between native -sheet proteins and amyloid fibrils". In: *Protein Science* 13.12, pp. 3314–3321. DOI: 10.1110/ps.041024904.

Zanotti, Giuseppe et al. (2017). "Structural and dynamics evidence for scaffold asymmetric flexibility of the human transthyretin tetramer". In: *PLOS ONE* 12.12, pp. 1–14. DOI: 10.1371/journal.pone.0187716.

Zhou, Huan-Xiang and Michael K. Gilson (2009). "Theory of Free Energy and Entropy in Noncovalent Binding". In: *Chemical Reviews* 109.9, pp. 4092–4107. DOI: 10.1021/cr800551w.

Zwanzig, Robert (1973). "Nonlinear generalized Langevin equations". In: *Journal of Statistical Physics* 9.3, pp. 215–220. DOI: 10.1007/BF01008729.