Hydration layer dynamics and association mechanisms of food and antifreeze proteins
Brotzakis, Z.F.

Citation for published version (APA):
Hydration layer dynamics and association mechanisms of food and antifreeze proteins

A Molecular Dynamics and Transition Path Sampling study

Zacharias Faidon Brotzakis
Hydration layer dynamics and association mechanisms of food and antifreeze proteins
A Molecular Dynamics and Transition Path Sampling study

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam,
op gezag van de Rector Magnificus
prof. dr. ir. K. I. J. Maex
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel

op dinsdag 7 maart 2017, te 10.00 uur

door

Zacharias Faidon Brotzakis
geboren te Athene, Griekenland
Promotiecommissie

Promotor:

- prof. dr. P. G. Bolhuis (HIMS, Universiteit van Amsterdam)

Overige leden:

- prof. dr. H. J. Bakker (FOM institute AMOLF)
- prof. dr. D. Laage (École Normale Supérieure Paris)
- prof. dr. E. J. Meijer (HIMS, Universiteit van Amsterdam)
- prof. dr. L. Visscher (Vrije Universiteit Amsterdam)
- dr. I. K. Voets (Technische Universiteit Eindhoven)
- dr. J. Vreede (HIMS, Universiteit van Amsterdam)

Faculteit der Natuurwetenschappen, Wiskunde en Informatica


The research reported in this thesis was carried out at the Van ’t Hoff Institute for Molecular Sciences, Faculty of Science, University of Amsterdam (Science Park 904, 1098 XH, Amsterdam, The Netherlands) with financial support by NanoNextNL, a micro and nanotechnology consortium of the Government of The Netherlands and 130 partners.)
## Contents

1 Introduction .......................................................... 7
   1.1 Proteins in nature and technology ............................... 7
      1.1.1 Protein function ........................................... 7
      1.1.2 Food Proteins ................................................ 8
      1.1.3 Ice binding and anti-freeze proteins ....................... 8
   1.2 Protein structure and interactions .............................. 9
      1.2.1 Structure of proteins ....................................... 9
      1.2.2 Protein-protein interactions ................................ 10
   1.3 The role of solvent (water) and co-solutes (e.g. salts) on protein stability ........................................... 13
   1.4 Protein association ............................................... 16
   1.5 Molecular simulations ........................................... 20
   1.6 Aim of the thesis ................................................ 21
   1.7 Outline of the thesis ............................................ 22

2 Methods .................................................................. 25
   2.1 Molecular Dynamics ............................................... 26
      2.1.1 Molecular Dynamics of solvated proteins ................... 29
   2.2 Rare Event methods ............................................... 32
   2.3 Transition Path Sampling ......................................... 32
      2.3.1 Monte Carlo of trajectories ................................... 34
      2.3.2 Shooting move .................................................. 35

3 Dynamics of hydration water around native and misfolded α-lactalbumin ........................................... 41
   3.1 Introduction ......................................................... 42
   3.2 Methods ............................................................ 43
      3.2.1 Simulation setup ............................................... 43
3.2.2 Analysis ........................................... 44
3.2.3 Polarisation-resolved femtosecond infrared spectroscopy .... 46
3.3 Results and Discussion ................................. 47
3.3.1 Hydrogen bond dynamics .......................... 47
3.3.2 Reorientation dynamics ......................... 50
3.4 Conclusions ......................................... 60

Appendices 62
Appendix 3.A Water HB and reorientation dynamics analysis. .... 62

4 Correlation between water structure and dynamics in the hydration layer of a type III ocean pout anti-freeze protein 75
4.1 Introduction ......................................... 75
4.2 Methods ........................................... 77
4.2.1 Simulation setup ............................... 77
4.2.2 Analysis........................................ 78
4.3 Results and Discussion .......................... 81
4.3.1 Water reorientation dynamics slows down at surface .... 81
4.3.2 Water reorientation dynamics differs locally .... 82
4.3.3 Hydration water structure ....................... 83
4.3.4 Structure - reorientation time correlation .......... 84
4.3.5 Effect of dehydration ............................ 87
4.4 Conclusions ....................................... 89

Appendices 91
Appendix 4.A Water structure and dynamics for different temperatures and forcefields. ................................. 91

5 Stability and growth mechanism of self-assembling anti-freeze cyclic peptides 99
5.1 Introduction ......................................... 100
5.2 Methods ........................................... 101
5.2.1 System setup ................................... 101
5.2.2 Molecular dynamics ........................... 102
5.2.3 Transition Path Sampling ....................... 103
5.3 Results and Discussion .......................... 105
5.3.1 Nanotube equilibrium properties ................. 105
5.3.2 Mechanism of self-assembly by Transition Path Sampling .... 108
5.3.3 Analysis of the path ensemble ................ 111
Chapter 1

Introduction

1.1 Proteins in nature and technology

1.1.1 Protein function

Proteins are the workhorses of the living cells. They consist of linear polypeptide chains made out of amino acids linked together through covalent peptide bonds. In the cell the processes of transcription and translation convert DNA to mRNA, which subsequently serves as a template for the production of amino acid chains in the ribosomes [1]. Interestingly, there are 20,000-23,000 human genes that encode proteins, which with the above procedure, may produce hundreds of thousands distinct proteins of different functionality when expressed.

Proteins differ in amino acid sequence, structure, shape, and size, thus giving rise to different biological functions, including enzyme catalysis (enzymes), transport of ions through membranes (membrane transport proteins), regulation of cell activity by e.g. transmitting extracellular signals to the cell’s interior (signaling proteins), and translation of other proteins or organelles (molecular motors). An important point is that proteins are functional under physiological conditions at which they fold into a minimum free energy 3D configuration, their native state. Proteins with common sequence are likely to belong to a common ancestor (homologs), leading to similarities in structure, e.g. the similar X-ray crystallographic structure and similar sequence (43%) of myoglobin and hemoglobin. Notwithstanding this connection between sequence structure and function, there is still not a clear relation between the sequence, structure and functionality. Moreover, the mechanism of protein folding has not been completely understood, especially for larger proteins. Answering these questions would help de-
signing novel proteins of tailored stability and dynamics in applications such as vacci-

cines, drug delivery vesicles, new enzymes etc. The folding process is not 100 percent
accurate, and misfolding and/or aggregation of proteins can occur. This can lead to
neurodegenerative diseases like Alzheimer’s, Parkinson’s, Huntington, type 2 diabetes,
Bovine Spongiform Encephalopathy (BSE) [2–4]

1.1.2 Food Proteins

In this thesis we are mostly interested in proteins encountered in food applications.
Examples of protein rich food systems are soy, pea, gelatin, but also milk and dairy
products. In particular, we will mostly look at whey proteins, which are a substan-
tial component of milk. While clearly originating from biological resources (plants,
funghi, algae and animals), food proteins are exposed to a much higher range of tem-
perature (-20 to 100 °C), pH (1-10) and salinity conditions, related to food preparation
and application. High temperature or addition of co-solutes can cause protein denatura-
tion, which usually leads to aggregation. Food proteins, depending on the application,
encounter many different types of aggregates, such as 3D networks of aggregates span-
ning discrete space regions or even the entire space in the case of percolation of gels.
Food proteins can be used as surfactants in order to stabilize the interfaces of hetero-
genous food materials, e.g. foams and oil-water interfaces. Notably, protein-rich food
systems hardly ever consist of proteins alone, but often contain other hydrocolloids or
surface active molecules, controlling their aggregation [5].

1.1.3 Ice binding and anti-freeze proteins

Ice Binding Proteins (IBPs) occur in a variety of organisms and have a variety of func-
tions, such as blocking the growth of ice crystals, structuring ice [6, 7], inhibiting ice
recrystallization [8], and promoting adhesion [9]. Their unique capability to regulate
the ice crystal growth makes them great candidates for use in many applications in food
technology, medicine and material science.

For many freeze-avoiding species such as fish, insects etc, IBPs prevent the growth
of ice in their blood stream [10]. This subclass of IBPs is called anti-freeze pro-
tein (AFPs) as they predominantly lower the freezing point of ice below its melt-
ing point in a non-colligative manner. For instance, fish AFPs show a depression of
the freezing point by ≈ 2°C at serum levels of 30-40 mg ml⁻¹. Anti-freeze proteins
have various shapes and sizes and are grouped in fish and non fish AFP categories
(see Fig. 1.1.1). Fish AFPs are divided into type I,II,III and IV, with type I having
a repetitive aminoacid sequence and type II,III having a non-repetitive sequence and
globular structure. In this thesis we will investigate solvation properties of anti-freeze
protein of type $III$, as well as a new synthetic cyclic peptide that self-assembles into in an anti-freeze nanotube structure.

1.2 Protein structure and interactions

1.2.1 Structure of proteins

Apart from the primary structure-sequence of amino acids mentioned above, proteins exhibit secondary, tertiary and quaternary structure. Secondary structures are spatial arrangements of protein segments such as $\alpha$-helices, $\beta$-sheets and $\beta$-turns, which involve mainly backbone hydrogen bonds. The tertiary structure relates to the total 3D conformation of a folded polypeptide chain. Contrary to the secondary structure, the tertiary structure is stabilized by hydrophobic interactions, salt bridges, side chain hydrogen bonds and disulfide bridges. These interactions further stabilize the secondary structure. Finally, the last level of structural categorization is the quaternary structure. As proteins can be multimeric, the relative positions of the different domains form the quaternary structure (e.g. Hemoglobin).
1.2.2 Protein-protein interactions

Electrostatic repulsion and Van der Waals attraction: DLVO theory

Native globular proteins in solution can be viewed as colloidal particles suspended in an electrolyte solution. While in principle the underlying atomistic interactions govern the behaviour of such suspensions, it is possible to describe the interaction between such particles in terms of effective potentials between the centers of mass, by integrating out the microscopic degrees of freedom. While most of the interactions are of electrostatic nature, it is convenient to classify the effective interactions in different categories: screened electrostatic interactions, hydrophobic interactions, Van der Waals interaction, hydration interactions, and remaining molecular interactions such as salt bridges and hydrogen bonds.

Electrostatic and Van der Waals interactions are often treated together in the so-called DLVO theory. Here the particles interact via a sum of two body interactions: of the repulsive screened electrostatic and the Van der Waals attraction. For spherical particles (e.g. globular proteins) the interaction energy \( W \) becomes [12]:

\[
W = \frac{1}{2} ZRe^{-\kappa D} - \frac{AR}{12D},
\]  

(1.1)
Figure 1.2.2: Interaction energy of two spheres as a function of their distance, based on the DLVO theory. Figure taken from ref. [12]

where $Z$ is the interaction constant (see eq. 1.2), $R$ is the sphere radius, $D$ is the distance between the two spheres, $A$ is the Hamaker constant and $\kappa^{-1}$ is the Debye length. The interaction constant $Z$ is given by eq 1.2,

$$Z = 64 \pi \epsilon_0 \epsilon \left( \frac{k_B T}{e} \right)^2 \tanh^2 \left( \frac{ze\psi_0}{4k_BT} \right),$$

(1.2)

where $\epsilon_0$ is the dielectric permittivity in vacuum, $\epsilon$ is the dielectric constant of water, $e$ is the elementary charge, $T$ is the temperature and $k_B$ the Boltzmann constant, $z$ is the valence of an ion in the electrolyte solution and $\psi_0$ is the surface potential. The charges on the sphere attract counter ions from the solution, creating a double layer, which screens the electrostatic interaction. Note that eqs. 1.2, and the first term of 1.1 hold for symmetric electrolyte solutions (e.g. 1:1, 2:2). The Debye length is related to the electrostatic screening and depends on the ionic concentration as follows (eq 1.3),

$$\kappa^{-1} = \sqrt{\frac{\epsilon_0 \epsilon k_B T}{2N_A e^2 I}},$$

(1.3)

where $I$ stands for the ionic strength, and $N_A$ is Avogadro’s number. DLVO interactions scale with the radius of the interacting bodies. In Fig. 1.2.2 it is assumed that
particles have radius $R=100$ nm. Globular proteins of moderate size do not exceed a radius of a few nanometres, therefore the conclusions mentioned below are most evident in much larger size proteins or colloids. For low ionic strength (dilute electrolyte solutions) the Debye length increases, giving rise to an effective electrostatic repulsion, which forms a large energy barrier between 1 and 5 nm (see Fig. 1.2.2). Upon increasing the ionic concentration, the Debye length is decreased, therefore screening the electrostatic repulsion and reducing the barrier, while forming a secondary minimum. At very high salt concentration the effective screening is so large that the only force into play is the Van der Waals attraction. Increasing the salt concentration is therefore likely to destabilize the protein solution. Also changing the surface charge, e.g. by changing the pH, can induce aggregation. Through a change in protonation state of side chains the protein can reach its isoelectric point, where the electrostatic repulsion becomes minimal.

**Hydrophobicity**

Hydrophobic interaction is very common in biology and drives many processes such as folding and hydrophobic association. Water in the hydrogen bond network around hydrophobes, although more coordinated, is entropically hindered to restructure the network of hydrogen bonds [13]. However, for bigger hydrophobic objects the formation of a complete hydrogen bond network is geometrically impossible making hydrophobic hydration energetically and entropically unfavourable [14]. For spheres at a critical distance, this effect becomes prohibiting, and water at the interface evaporates. This gives rise to a dewetting transition [15, 16] and therefore to an effective attraction between the hydrophobic spheres.

**Other surface interactions**

Besides Van der Waals, screened electrostatic interactions, and hydrophobic attractions, there exist other non-DLVO forces that are present on protein association. The most obvious one is the direct hydrogen bonding between residues of different proteins. Further, salt bridges provide strong electrostatic interactions between proteins, thus providing a highly stabilizing role. Also, hydrophilic protein surfaces bind water hydration layers. When two proteins are approaching each other, the hydration layers of the two feel each other and cause repulsive hydration forces due to the change in the structuring of each individual layer. These hydrophilic hydration repulsion forces are oscillatory [12, 17, 18].

Other surface interactions that are crucial for protein association include water bridging interactions, where a water makes a double hydrogen bond with the protein. A
thermodynamic model by Ben-Naim predicts the importance of bridging waters in reducing the free energy of binding and driving the assembly towards tighter binding [19].

### 1.3 The role of solvent (water) and co-solutes (e.g. salts) on protein stability

The solvent mediates protein interactions as well as being an agent on its own, driving protein motion (hydration forces, H-bond network, and the hydrophobic/hydrophilic effect). Below, we briefly introduce the importance of water, its relevance as a solvent around proteins, and its relation to salts. Ions are very important in biology. Under physiological salt conditions (∼100 mM), ions participate in the osmotic stress regulation and other main processes [1]. In food related applications ions are used to alter the stability of colloids or proteins towards preferred phases (e.g by reducing protein-protein repulsion). Water by itself exhibits unique properties many of which can be attributed to the unique tetrahedral H-bond structure that is prevalent both in liquid and crystalline water. The water H-bond structure and dynamics is altered by the presence of a surface. In this thesis we investigate how the H-bond structure and dynamics is influenced by the presence of a protein surface.

**Importance of water around proteins (hydration layers)**

Water around proteins plays a significant role in many processes, from protein folding to enzyme function. Hydrophobic interactions are known to drive proteins to fold in order to increase the translational and orientational entropy of water [20, 21]. Moreover, water mediated hydrogen bond interactions assist in the stabilization of the secondary and tertiary structure [1]. Interestingly, proteins slow down the hydration water translation and orientation by a factor ∼ 3-5. Finally, the slow dynamics of water is important for enzymes such as metalloprotease MT1-MMP. Experiments and simulations showed that for this enzyme, the formation of the Michaelis complex, is driven by a slowdown of water reorientation towards the active site [22].

**Hofmeister series**

Addition of salt can stabilize or destabilize the protein solution. In the 19th century, the Czech protein scientist Hofmeister, investigated the influence of ions in the precipitation of yolk egg protein and other small colloids. The Hofmeister series (see eq. 1.4)
is an ordering of ions according to their propensity (measured by concentration) to precipitate proteins (salting out).

\[ CO_3^{2-} > SO_4^{2-} > S_2O_3^{2-} > H_2PO_4^- > F^- > Cl^- > Br^- > NO_3^- > I^- > ClO_4^- > SCN^- \]  

(1.4)

The origin of the Hofmeister series is not well understood and has been attributed to a number of causes. Jones and Dole [23] have shown that some ions affect water viscosity positively (kosmotropes) and others negatively (chaotropes). In other words, ions are structuring or disrupting the H-bond network of electrolyte solutions.

On the other hand, there has been an increasing number of studies suggesting the ion-protein interactions causing the precipitation/unfolding propensity, however without yet having reached a consensus on the mechanism of the series [24]. However, there is evidence that Hofmeister ions with a higher propensity have a greater affinity for amide groups and are excluded from hydrophobic groups [24].

**Water reorientation dynamics: Jump models**

Water H-bond structure and dynamics around proteins plays an important role in biology, such as in protein folding, ion channels and enzyme reactions. While in the Debye model [25] water motion was believed to be a barrierless random angular diffusion, more recent models indicate that water reorientation and motion is dictated by large-amplitude and sudden angular jumps. Here, we first briefly present some experimental techniques measuring water reorientation dynamics followed by a brief introduction to some of the recent models treating water reorientation and the hydrogen bond exchange mechanism.

IR pump-probe spectroscopy [26–28] has femtosecond time resolution and therefore can capture the very fast water reorientation. However, it suffers from the fact that it cannot address time delays larger than 10 ps. On the other hand nuclear magnetic resonance spectroscopy (NMR) [29, 30] can measure the spin-lattice relaxation time and give information about the average rotational relaxation but not its short time evolution.

One way to describe water reorientation dynamics is by the time correlation function of the OH bond vector. This function is given by eq. 1.5,

\[ C_n(t) = \langle P_n[u(0) \cdot u(t)] \rangle \]  

(1.5)

where \( P_n \) stands for the \( n^{th} \) order Legendre polynomial and \( u(t) \) is the OH vector at time \( t \). By plotting the time correlation function \( C_2 \) (see Fig. 1.3.1) one can extract the timescale of different mechanisms [31–33]. The fast (< 200 fs) librational motion
involves fast wobbling of the OH vector around the axis of the donor-acceptor H-bond (see Fig. 1.3.1ai,aii). A hydrogen bond jump event (see Fig. 1.3.1aiii and Fig. 1.3.1b) involves the elongation of the initial H-bond, while a new water oxygen acceptor is approaching. Once the two H-bond accepting oxygens are at the same distance, the OH jumps from the old to the new H-bond acceptor. In an effort to quantify the kinetics of the angular mechanism, Ivanov et. al. [32] developed the jump model. This model assumes constant amplitude jumps between hydrogen bonded water with frequency \(1/\tau_{\text{jump}}^n\).

However, in the absence of librational and frame tumbling motion (see Fig. 1.3.1ai and Fig. 1.3.1aiii respectively) in the jump model, an extended version was developed by Laage and Hynes [33], the extended jump model (EJM). There, the overall water reorientation timescale \(\tau_{\text{EJM}}^n\) is dictated by a hydrogen bond switching timescale \(\tau_{\text{jump}}^n\) and a frame tumbling timescale \(\tau_{\text{frame}}^n\), as shown in eq. 1.6.

Figure 1.3.1: a) Water reorientation mechanism associated with different time scales, b) schematic of the jump reorientation mechanism, c) transition state excluded volume area for dilute solutions and d) for concentrated solution. Figure taken from ref. [34].
When water solvates a solute molecule, its reorientation is retarded. In the EJM, this retardation can be attributed to the transition state excluded volume. The excluded volume in concentrated solutions creates convex environments (see Fig. 1.3.1d). The retardation factor due to excluded volumes ($\rho_v$) is larger than two as opposed to convex surfaces of dilute solutions (see Fig. 1.3.1c). While the presence of solutes can cause stronger water-solute H-bonds, the retardation of water reorientation dynamics can also be attributed to the relative difference in H-bond strengths of the pure bulk compared to the solution ($\rho_{HB}$). All the above effects define the overall retardation of the reorientation, shown in eq. 1.7

\[
\frac{1}{\tau_{EJM}} = \frac{1}{\tau_{n}^{jump}} + \frac{1}{\tau_{n}^{frame}}, \tag{1.6}
\]

\[
\tau_{n}^{jump} = \rho_v \rho_{HB} \tau_{bulk}^{jump}. \tag{1.7}
\]

1.4 Protein association

Association and self-assembly of bio-molecules involves many controlled processes occurring in biology, such as protein association, protein-ligand binding, signalling
Figure 1.4.2: Schematic representation of the free energy landscape as a function of a reaction coordinate for different pH for a dimer (D) to monomer (M) equilibrium.

and inter-cellular communication. Protein self-assembly and association is a controlled process not only in biology but in technology as well. Depending on the conditions self-assembly can result in a variety of different aggregates, which self-assemble/aggregate with different mechanisms (see Fig. 1.4.1). For instance, by controlling the conditions (e.g. pH) one can change the mechanism of self-assembly, and therefore grow different aggregates (fibrous or worm-like) having different properties (e.g. texture in food industry). Aggregation and fibrilization can also be an unwanted process, such as in the case of amyloidogenic peptides. In this thesis we will investigate how a new synthetic cyclic peptide self-assembles into a nanotube structure.

On a different note, the dimerization transition of native globular proteins is very important on the biological level as it is important in regulation networks as well as the first step of oligomerization, which later forms worm-like fibrils [36]. In food systems (e.g. milk), dimers can occur in thermodynamic equilibrium with monomers [37]. In the following section we introduce the thermodynamics and dynamics of dimer formation.

**Monomer-dimer equilibrium**

From a thermodynamic perspective, the free energy difference between the dimer state (D) and monomer state (M) is given by eq. 1.8.

\[-RT\ln(K) = \Delta G = \Delta H - T\Delta S\] (1.8)
Here, $K$ stands for the dimerization equilibrium constant, and $\Delta G$, $\Delta H$ and $\Delta S$ are the Gibbs free energy, enthalpy and entropy of dimer formation, respectively. $T$ is the temperature and $R$ is the gas constant. At pH values higher or lower than the isoelectric point $pI$, the electrostatic repulsion favours the monomer population and mainly contributes positively to the enthalpy $\Delta H$ of formation of dimers. Moreover, the reorientation and translation entropy of the proteins is higher in the monomer state, decreasing the $-T\Delta S$. Negative $\Delta G$, where $\Delta G = G_D - G_M$ corresponds to a driving force towards the dimer (see Fig. 1.4.2). When approaching the $pI$, the effective electrostatic screening repulsion becomes smaller, and at the expense of decreasing the translational and orientational entropy, proteins attract each other and the dimer formation is favoured. As mentioned above the effective attraction is dominated by van der Waals interactions, salt bridges, hydrogen bonds (decreasing $\Delta H$), and for some systems by hydration forces.

Protein association (e.g dimer association) processes are activated processes involving rare transitions between states that are separated by an free energy barrier. The thermodynamic perspective of dimerization was discussed above. Below we discuss reaction rate theories intending to quantify the kinetics of rare transitions, notably phenomenological rate equations. The Arrhenius equation [38] and the transition state theory (TST).

Using a statistical mechanics perspective, dimerization can be pictured as a transition between stable states or free energy minima N (native dimer) and U (unbound monomers), separated by an activation barrier (see Fig. 1.4.3). The rate of the transition between the two minima is inversely proportional to the exponent of the barrier height [39–41]. Consequently, this means that the dwell time in the stable states is much larger than the transition time, $\tau_{state} \gg \tau_{mol}$, the so-called separation of timescales. For a two state activated dynamical process, one can express the kinetics in terms of phenomenological rate equations shown in eqs. 1.9-1.10. The solution of these equations is given in eq. 1.11 and states that a system which has realized a density fluctuation ($\Delta C_N(0)$) of species N from the equilibrium density $\langle C_N \rangle$ at time zero, returns to its the equilibrium population exponentially according to the reaction time $\tau_{rxn}^{-1} = k_{NU} + k_{UN}$, where $k_{NU}$ and $k_{UN}$ are the forward and backward rate constants respectively.

\[
\frac{dC_N}{dt} = -k_{UN}C_N(t) + k_{NU}C_U(t) \quad (1.9)
\]
\[
\frac{dC_U}{dt} = k_{UN}C_N(t) - k_{NU}C_U(t) \quad (1.10)
\]
\[
\Delta C_N(t) = C_N(t) - \langle C_N \rangle = \Delta C_N(0) \exp \left( \frac{-t}{\tau_{rxn}} \right) \quad (1.11)
\]
The Arrhenius equation 1.12 addresses the kinetic rate constants of chemical reactions,

$$k = Ae^{-\frac{E_a}{k_BT}}$$

with $A$ the pre-exponential factor and $E_a$ the activation energy and $k_B$ the Boltzmann constant. The Arrhenius equation expresses the transition rate constant as a function of the temperature, the frequency of collisions $A$ and the probability $e^{-\frac{E_a}{k_BT}}$ to result in a reaction, given a collision. A more fundamental way to look at a reaction rate constant is the Eyring-TST equation \[39\]. It states that the rate of transformation from one free energy minima to another, e.g. $U$ (unbound state) to $N$ (native dimer) (see Fig. 1.4.3) reduces proportionally to the exponent of the free energy barrier height [39–41], and is given by equation 1.13,

$$k = \frac{k_B T}{\hbar} \kappa e^{-\frac{\Delta G^\ddagger}{k_B T}} = \frac{k_B T}{\hbar} \kappa e^{-\frac{\Delta H^\ddagger}{k_B T} + \frac{\Delta S^\ddagger}{k_B}}$$

with $\hbar$ being the Planck constant, $\kappa$ the transmission coefficient, $\Delta G^\ddagger$, $\Delta H^\ddagger$ and $\Delta S^\ddagger$ the Gibbs free energy, enthalpy and entropy of activation, respectively. The double dagger symbol denotes the values at the transition state. Note that the transition state theory is approximate, and that more elaborate theories or simulations are required to make progress.
The rate of protein association in liquids is bound from above by the diffusion of proteins in the available volume $V$ before they are close enough to associate. Such reactions, where the slow part of the reaction is bringing together the proteins, are denoted as diffusion controlled reactions. The bimolecular association rate constant $k_A$ for two uniformly interacting spherical particles is given by the Smoluchowski rate $k_A = 4\pi DR/V$, where $D = D_A + D_B$ and $R = R_A + R_B$. $D_A$ and $D_B$ are the diffusion constants of particles A and B and $R_A$ and $R_B$ are the radii of particles A and B respectively.

The kinetics and mechanisms of association, even on the dimer level, is poorly understood and varies with the nature of the proteins. Many proteins associate into homodimers with a (experimentally determined) rate constant $k_{on} \sim 10^6 M^{-1}s^{-1}$, which is three orders of magnitude slower than the diffusion limited Smoluchowski association rate constant predicts. This sounds reasonable because association of proteins into stable dimers requires very specific favorable interactions in the native (bound) dimer. However, an estimate of the association rate purely based on the geometric probability that a protein would randomly collide into the correct dimer state would be six orders of magnitude smaller than the diffusion limited rate [42]. Nonspecific interaction can enhance the rate [42, 43] even in the absence of strong hydrophobic and/or charged patches that could steer the assembly. The presence of strong hydrophobic or electrostatic interactions steering the association would result in a much faster association with a rate in the order of $10^9 M^{-1}s^{-1}$, such as in the barnase-barstar complex [44]. In the work by Ahmad et. al. it was shown that for the barnase-barstar complex, electrostatic steering drives the association and causes a very high association constant [44] $k_A=10^8 - 10^9 M^{-1}s^{-1}$, close to the diffusion limited association rate[42]. Nevertheless, the driving force and kinetic mechanism for protein association, even in the absence of such steering forces, is not well understood.

## 1.5 Molecular simulations

The laws of nature are expressed in equations which very scarcely can be solved analytically (without the use of a computer), e.g. Newtonian motion of three interacting bodies. Before computers, in order to predict material properties, scientists resolved to approximate theories (e.g VDW equation, Debye Huckel theory). However these theories are approximate and often are valid in few systems, thus lacking generality. Computer simulations/computer experiments serve as predictions which test/explain/validate the theory or the experiment. Here we make use of molecular dynamics simulations using all atom force fields. The MD simulations are limited in time and length scales. Especially for rare events, specific techniques need to be developed. One such technique is
Transition Path Sampling (TPS), which samples rare trajectories committing to stable states. Both MD and TPS are used in this thesis.

1.6 Aim of the thesis

The aim of this thesis is three-fold. Firstly, to better understand the role of water around single proteins in terms of structure and dynamics, secondly to understand the association and self-assembly mechanism of multiple proteins, and thirdly to understand the role of water in the association mechanism. Finally, as protein association is a rare event, we develop a new TPS shooting move which can efficiently sample rare transitions between states, separated by asymmetric barriers.

Understanding water structure and dynamics around proteins

Protein hydration water plays a significant role in determining the dynamics and structure of proteins, in facilitating protein folding, in recognizing ice crystal planes (anti-freeze proteins) and in mediating protein association. In this thesis we address the following research questions:

- What is the H-bond structure and dynamics of water around hydrophilic and hydrophobic groups. How does unfolding affect the overall dynamics of water?
- How does water H-bond structure correlate to water reorientation dynamics at the Ice Binding Site of an anti-freeze protein?

Understanding of dynamical mechanisms of association and self-assembly of multiple proteins, and the role of water in these processes

Protein association, especially in dilute conditions is by itself interesting as its mechanism is yet not clear. Moreover, protein association is pivotal in the first steps of self-assembly. Atomistic insight into the association mechanism and initial steps of self-assembly helps understanding and experimentally control these processes. As these processes are rare events, developing computational methods which can better sample these transition is very important. In order to address these processes, we ask the following questions:

- How do anti-freeze peptides self-assemble into nanotubes? What is the nanotube stability as a function of size?
How do globular proteins associate into dimers? Which interactions play a role? What is the role of water along the association?

Can we sample transitions with asymmetric barriers? Can we extract information about the transition state region?

1.7 Outline of the thesis

In this thesis we study the hydration layer dynamics and the association mechanisms of food and anti-freeze proteins. We do so by employing Molecular Dynamics and Transition Path Sampling. In chapter 2, we give a brief presentation of the MD and TPS methods. In chapter 3, we apply MD to address the dynamics of hydration water around native and misfolded α-lactalbumin. In chapter 4, we study the correlation between water structure and dynamics in the hydration layer of a type III ocean pout anti-freeze protein. In addition, in chapter 5, we investigate the stability and growth mechanism of self-assembling anti-freeze cyclic peptides. In chapter 6, we develop the spring shooting, a novel efficient transition path sampling move, and finally, in chapter 7 we elucidate the mechanism and role of solvent for β-lactoglobulin dimerization using Transition Path Sampling.
Bibliography


Molecular simulations are central to the computation of equilibrium and dynamic properties of classical many body systems, as well as in bridging microscopic with macroscopic observables [1–4] that could be of both experimental and basic importance.

The work of Metropolis et al. [5] in 1953 led to the foundation of Monte Carlo simulations (MC) and introduced the era of molecular simulations. A few years later application of the MC algorithm to a Lennard-Jones potential [6] could quantitatively capture thermodynamic properties. The direct comparison of these computed thermodynamic properties to liquid argon experiments manifested the predictive power of molecular simulations. A few years later in 1957 the Molecular Dynamics (MD) technique was introduced by Alder [7]; a powerful technique that could capture the dynamics of a system by integrating Newton’s equation of motion.

Nowadays Molecular Dynamics is routinely [3] used to explore the conformational dynamics and transitions of complex biological systems, e.g. protein folding, and protein association. Capturing the folding process by brute force Molecular Dynamics simulations of small proteins in the order of microseconds is nowadays tractable by using parallel computing or by use of GPUs [8, 9]. The development of the Anton machine enabled reaching long timescales. The Anton machine [10] is a massively parallel MD specific hardware capable of simulating several milliseconds per day. Many biological transitions occur on that timescale, such as drug binding to proteins, small protein folding [11, 12] etc.

However, natural systems undergo rare but important physical or chemical transitions (transformations) between stable states. For example, large protein folding, conformational changes of molecules, diffusion in solids, crystal nucleation are processes that occur at the millisecond to seconds timescales. To obtain the long time
dynamics of these systems one needs to capture a statistically important amount of these rare transitions, which is intractable even for the Anton machine. One way to overcome this problem is to perform rare event simulations that can adequately sample rare transitions. In this thesis we perform Transition Path Sampling to sample rare transitions between stable states, to identify on pathway intermediates, and understand the mechanism of the transition.

The chapter is organized as follows. We first introduce the Molecular Dynamics technique, an important tool of this thesis. Then we briefly present the Transition Path Sampling algorithm.

2.1 Molecular Dynamics

Classical Molecular Dynamics is a method that solves Newton’s deterministic equations of motion. It uses the gradient of the potential energy of a system $V(r)$ to evaluate the trajectory of $N$ atoms, given their initial conditions.

$$ F(t) = ma(t) \quad (2.1) $$

$$ u(t) = \frac{dr(t)}{dt} \quad (2.2) $$

$$ a(t) = \frac{du(t)}{dt} \quad (2.3) $$

Given the potential energy $V_i(r)$, for each atom $i$ of mass $m_i$, the force acting on the atom $i$ at time $t$ is given by the gradient of the potential $F_i(r(t)) = -\nabla V_i(r(t))$. Subsequently, one uses an integration algorithm to estimate the positions $r(t + \Delta t)$ and velocities $u(t + \Delta t)$ at time $t + \Delta t$.

There exist many integration algorithms to propagate Newton’s equation of motion [13, 14]. Since Newton’s equations of motion are time reversible, and preserve the phase space volume, good integrators are considered ones that are time reversible and cause minimal energy drift, therefore, conserving the equilibrium distributions in phase space. Integrators with those properties which are commonly used in Molecular Dynamics simulations are the Verlet, Velocity-Verlet and Leap-Frog algorithms.

The Verlet algorithm can be derived using a Taylor expansion about $r(t)$. One can sum up the Taylor expansions at $t + \Delta t$ and $t - \Delta t$ shown below to retrieve the equation for advancing the positions in time, as shown in eq 2.6. The advancement of velocities can be retrieved by subtracting eq. 2.4 and eq. 2.5 and is shown in eq. 2.7.
As shown in Fig. 2.1.1a, to propagate the positions, the Verlet algorithm uses the positions at time $t$ and $t - \Delta t$ and the forces at $t$. As shown in eq. 2.6 the Verlet algorithm propagates $r$ about $r(t)$ symmetrically in time thus making it time reversible [1], and simultaneously preserves linear momentum, thus conserving the total energy. However, the algorithm is subject to some computational imprecision since the velocities are calculated with an error of the order $\Delta t^2$. Similar to the Verlet algorithm is the Leap-Frog algorithm, shown in equations 2.8 and 2.9.

$$r(t + \Delta t) = r(t) + \dot{r} \Delta t + \frac{1}{2} \ddot{r} \Delta t^2 + \frac{1}{3} \dot{r} \Delta t^3 + O(\Delta t^4)$$ \hspace{1cm} (2.4)

$$r(t - \Delta t) = r(t) - \dot{r} \Delta t + \frac{1}{2} \ddot{r} \Delta t^2 - \frac{1}{3} \dot{r} \Delta t^3 + O(\Delta t^4)$$ \hspace{1cm} (2.5)

$$r(t + \Delta t) = 2r(t) - r(t - \Delta t) + \frac{F(t)}{m} \Delta t^2 + O(\Delta t^4)$$ \hspace{1cm} (2.6)

$$u(t + \Delta t) = \frac{1}{2\Delta t} \left( r(t + \Delta t) - r(t - \Delta t) + O(\Delta t^2) \right)$$ \hspace{1cm} (2.7)
\[
    r(t + \Delta t) = r(t) - u(t + \frac{1}{2} \Delta t) \Delta t \hspace{1cm} (2.8)
\]

\[
    u(t + \frac{1}{2} \Delta t) = u(t - \frac{1}{2} \Delta t) + \frac{F(t)}{m} \Delta t \hspace{1cm} (2.9)
\]

The positions at time \( t \) as well as the velocities at time \( t + \frac{\Delta t}{2} \) are used to propagate the positions, whereas the velocities at time \( t + \frac{\Delta t}{2} \) are calculated using the forces at time \( t \) and the velocities at time \( t - \frac{\Delta t}{2} \), as shown in Fig. 2.1.1b.

In the Leap-Frog algorithm, one needs to calculate the velocities at time \( t \) in an extra step, using \( u(t - \frac{1}{2} \Delta t) \) and \( u(t + \frac{1}{2} \Delta t) \). In order to avoid this extra step, the Velocity-Verlet algorithm stores the positions, velocities and forces at the same time step (equations 2.10 and 2.11) while minimizing the round off error.

\[
    r(t + \Delta t) = r(t) + u(t) \Delta t + \frac{F(t)}{2m} \Delta t^2 \hspace{1cm} (2.10)
\]

\[
    u(t + \Delta t) = u(t) + \frac{F(t + \Delta t) + F(t)}{2m} \Delta t \hspace{1cm} (2.11)
\]

All the above algorithms are time reversible and conserve phase space volume in the limit of small time steps. However, since there is no perfect integrating algorithm, there exists a small non-zero error in the calculation of the positions, velocities and forces. A small perturbation in the initial conditions can lead to a deviation from the ”true” trajectory that scales exponentially with time according to the Lyapunov exponent \( \lambda \). Although Molecular Dynamics does not predict the true physical trajectory, even not in principle, it provides statistically trustworthy trajectory predictions [2].

Molecular Dynamics simulations can simulate typically systems of thousands to millions of atoms, using boxes with a linear size of a few tens of nanometres. Such finite size systems cause artifacts on both dynamics and thermodynamics, because of the large fraction of atoms close to the boundaries. If the system had free or wall boundaries, the fraction of particles at the boundary would scale as \( N^{-\frac{1}{3}} \), leaving a significant amount of molecules being influenced by the finite size of the system, influencing its dynamic and thermodynamic properties. To simulate bulk properties, periodic boundary conditions (PBC) are used. These periodic boundaries mimic infinite bulk surroundings, where the periodic box is replicated an infinite amount of times into all directions [1, 2]. However, one needs to be careful in the choice of box size, since periodicity rules out fluctuations that are of longer length than the box size. This effect is critical in phase transitions, where long range density fluctuations come into play. In addition, to simulate for instance protein folding transitions or protein association,
one needs to construct a box with a size greater than the cutoffs of the short ranged interactions so that atoms do not interact with their own image, and at the same time large enough to accommodate the transition that one wants to study. For example, for the association of two middle sized proteins \((R_g \approx 2 \text{ nm each})\) in water, a cubic box of length 10 nm and about 100,000 atoms would be needed in order to capture dissociation events that require to visit a dissociated state with a minimum protein-protein distance of at least 1 nm. Moreover, when using PBCs short ranged atomic interactions are calculated only between nearest periodic neighbours (the minimum image convention). Since the slowest part of the simulation involves the force calculations which scales as \(N^2\), where \(N\) is the number of particles in the box, one wants to have the box size to the absolute minimum for more efficient calculations. Hence using truncated octahedron boxes that reduce the solvent molecules about 30% are commonly used in MD simulations.

### 2.1.1 Molecular Dynamics of solvated proteins

**All atom Force Field models**

The propagation of positions and velocities in the Verlet integration algorithm requires as input the force acting on each atom. Since the force on each atom can be obtained by the gradient of the potential energy at each atom, it follows that an atomic description of the potential energy is essential for MD. A force field is a set of classical inter-atomic semi-empirical potential energy functions that describes the interactions between atoms. Common force fields in biomolecular simulations are the AMBER [15], OPLSAA [16], CHARMM [17], and GROMOS [18]. The basic form of a force field is shown in eqs. 2.12-2.14.

\[
V_{tot} = V_{cov} + V_{noncov} = [V_{bond} + V_{angle} + V_{dih}] + [V_{vdw} + V_{el}] \tag{2.12}
\]

\[
V_{cov} = \sum_{i,j} \left( \frac{1}{2} k_{ij}^b (r_{ij} - r_{ij}^0)^2 \right) + \sum_{i,j,k} \left( \frac{1}{2} k_{ijk}^a (\theta_{ijk} - \theta_{ijk}^0)^2 \right) + \sum_{i,j,k,l} \left( \frac{1}{2} k_{ijkl}^d \left( 1 + \cos \left( n\varphi_{ijkl} - \varphi_{ijkl}^0 \right) \right) \right) \tag{2.13}
\]

\[
V_{noncov} = \sum_{i,j} k_{ij} \left[ \left( \frac{C_A}{r_{ij}} \right)^{12} - \left( \frac{C_B}{r_{ij}} \right)^6 \right] + \sum_{el} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \tag{2.14}
\]
The covalent bonds and angle terms are described harmonically, whereas the dihedrals by a periodic function. In eq 2.13, $r_{ij}$ is the bond length between atoms $i$ and $j$, $\theta_{ijk}$ is the valence angle between atoms $i$, $j$, $k$ and $\phi_{ijkl}$ is the dihedral angle between atoms $i,j,k,l$ and $n$ is the multiplicity. The corresponding bond, angle and dihedral equilibrium values are $r_{ij}^0$, $\theta_{ijk}^0$, $\phi_{ijkl}^0$ with force constants $k_{b}^{ij}$, $k_{a}^{ijk}$, $k_{d}^{ijkl}$ respectively. Non-covalent interactions consist of van der Waals interactions, Pauli repulsions and electrostatic interactions. The 6-12 Lennard-Jones potential parametrizes the Pauli repulsion with a term scaling $\propto r^{-12}$ and the van der Waals attraction with a term $\propto r^{-6}$. The parameters $C_A^{ij}$ and $C_B^{ij}$ are related to the repulsion and attraction between atoms $i,j$. The electrostatic interactions between atoms $i$ and $j$ are described with a Coulombic term. This term involves the partial atomic charges $q_i$, $q_j$, the dielectric permittivity of vacuum $\varepsilon_0$. All of the interactions except for the electrostatic, decay within the box size. However, since the electrostatic interactions scale $\propto r^{-1}$, truncation and tail correction of the electrostatic potential is not a good strategy, as the tail correction energy term will diverge for any potential that decays slower than $\propto r^{-3}$.

A way around this problem is provided by methods that split the sum of electrostatic interactions into a short and a long term. The short range term is solved in real space, and the long range term in reciprocal space. A commonly used method is the Particle Mesh Ewald approach in which the electrostatic interactions are split into short and long range terms. Up to a cutoff, the electrostatic contribution is calculated from the particle-particle interactions whereas for greater distances the charges are put on a grid, and from the grid density the Poisson equation is solved efficiently using Fast Fourier Transform techniques [13]. A similar often used technique is Particle Particle /Particle Mesh (PPPM) [13].

The accuracy of force-fields depends on the target data used to optimize the force-fields parameters. These data include QM calculations as well as experimental measurements. Bond and angle equilibrium constants as well as the multiplicity and phase are obtained from QM calculations of geometries. The vibrational spectra are used to adjust the force constants [19]. Optimization of the non bonded interactions is pivotal for reproducing the thermodynamic properties, as well as the structure of biomolecules. Charge determination methods such as the QM electrostatic potential (ESP) where the charges are optimized to reproduce the QM determined ESP is a standard approach to reproduce well the electronic properties of molecules. Lennard-Jones parameters are adjusted in order to reproduce experimentally determined targeted thermodynamic properties such as the heat of vaporization, density, compressibilities, heat capacities and free energy of solvation [20–22]. Van der Waals interactions can be also obtained using ab initio calculations [23].

In order to make molecular dynamic simulations computationally efficient, one
can constrain the fast vibrational degrees of freedom, thus allowing an increase in the MD time step. In MD simulations these constraints are usually enforced by using the method of Lagrange multipliers. The corresponding algorithms use a Lagrangian including constraints and differ in the way they solve the modified equations of motion to obtain the Lagrange multipliers and therefore the new positions and velocities. Some of the most common constraints algorithms are SHAKE [1], RATTLE [24], LINCS [25], and SETTLE [26]. Another way would be to use the multiple time step approach, by using short time step for the bond vibrations and longer time step for the long range attraction forces (e.g. RESP algorithm [27]).

A very time consuming part of the MD algorithm is the neighbour search of particle $i$ through all possible particles $j$, in order to determine the particles $j$ which reside within a cutoff distance from $i$ and carry out the force calculation $F_{ij}$. This procedure, scales as $N^2$. In order to reduce this computational cost, Verlet [28] suggested the use of neighbour lists. In this scheme, one does not update the neighbours of $i$ at every time step but with some lower frequency. Between the updates of the neighbour list the forces are calculated between particles in the stored neighbour list.

As mentioned earlier, Newton’s equation of motion conserve the total energy of the system also known as the Hamiltonian. Thus, MD samples the microcanonical ensemble (NVE). Constant temperature simulations are achieved by using thermostats, such as the v-rescale [29], Nose-Hoover [30] or Andersen [31] thermostats, that allow to sample the canonical ensemble (NVT). Moreover, constant pressure calculations are achieved by using barostats, such as the Parrinello-Rahman [32] barostat, that allow sampling of the isobaric isothermal ensemble (NPT).

For the MD simulations we used Gromacs as an MD engine due to its efficiency and documentation on biomolecular simulations [33].

**Water Force Field models**

Accurate treatment of the condensed aqueous environment is central in biomolecular simulations. This treatment is performed using explicit and implicit water force-fields. Some famous explicit water force-fields are the TIP3P [34], TIP4P [35], SPC [36] and TIP4P/2005 [37]. These force-fields are parametrized such that they reproduce static experimental properties such as the radial distribution function of water, as well as dynamic ones, such as the diffusion constant. SPC and TIP3P have interaction sites assigned to each of its atoms. Although SPC and TIP3P have slightly different tetrahedral angles (109.4° and 104.5° respectively), their charge distribution does not differ much. TIP4P has four interaction sites with an extra charge placed on a dummy atom, which improves the electrostatic distribution at the cost of computational time [38]. The TIP4P/2005 model has almost the same charge distribution and angle...
with TIP4P. Their difference is that the TIP4P has been parametrized to target the enthalpy of vaporization. Essentially, TIP4P/2005 estimates better all properties but the ones related to the gas phase, compared to TIP4P [38].

2.2 Rare Event methods

Many high dimensional complex systems in chemistry, physics and biology undergo rare physical or chemical transformations between metastable states. Such rare transitions include protein folding, protein association, ion dissociation, cluster isomerization, crystal nucleation diffusion in solids, etc.

A variety of methods exists that alleviate the exponential time scale problem (e.g. umbrella sampling [39], local elevation [40], conformational flooding [41], adaptive bias force [42], metadynamics [43]) and give access to an ergodic sampling of the phase space. However, these methods are based on implementing a bias along an order parameter, which alters the dynamics, gives up kinetic information and at the same time their efficiency depends on the quality of the order parameter as a reaction coordinate. On the other hand, path based methods do not require a reaction coordinate as an input, but only the state definitions and some initial chain of states connecting them. Methods such as the Nudge Elastic Band (NEB) [44] or the zero temperature string method [45] aim to find the minimum energy path (MEP) along one transition. Although these approaches give useful insight into the activation energy and the location of the saddle point, they do not contain dynamical information, since the evolution between states results from a minimization of an objective function. This problem was addressed in another category of path based methods, the so called action-based methods, which use the Hamilton’s principle of least action to obtain true dynamical trajectories connecting the two states [46, 47]. Finally, Transition Path Sampling [48, 49] is a trajectory based method, based on an importance sampling strategy on trajectory space, giving insight into both mechanistic and dynamical information. We will discuss this method in more detail in the following section.

2.3 Transition Path Sampling

Transition Path Sampling (TPS) is based on the idea that for a two state system, an infinitely long trajectory crosses the barrier infinite amount of times. The collection of all possible pathways that cross the barrier and connect the stable states is denoted the path ensemble and can be sampled by a MC random walk in the trajectory space. TPS, similar to many other trajectory based methods, needs a good definition of the
reactant and the product state but requires no prior information on the reaction coordinate. Given an initial path connecting the two states and using a Metropolis-Hastings criterion based on detailed balance one can generate a Markov chain of true dynamical pathways that maintain the equilibrium path ensemble distribution. There are many schemes to generate new pathways. In this thesis the one-way flexible shooting scheme has been employed and will be discussed later.

In TPS a path consists of a sequence of microstates at each time slice \( x(T) \equiv \{x_0, x_{\Delta t}, ..., x_{i\Delta t}, ..., x_T \} \), where \( x_{i\Delta t} = \{r_{i\Delta t}, p_{i\Delta t} \} \) comprises the coordinates \( r_{i\Delta t} \) and momenta \( p_{i\Delta t} \) of all the atoms of the system at each time step \( i = 0, 1, ..., N \) where \( N = \frac{T}{\Delta t} + 1 \) (Fig. 2.3.1). One can express the probability \( P(x(T)) \) of a pathway of duration \( T \) in a statistical mechanical way. \( P(x(T)) \) depends on the initial conditions and the underlying dynamics of the system according to eq 2.15.

\[
P(x(T)) = \rho_{x_0} \prod_{i=0}^{T-1} p(x_{i\Delta t} \rightarrow x_{i+1\Delta t}) / Z(T) \tag{2.15}
\]

\( p(x_{i\Delta t} \rightarrow x_{i+1\Delta t}) \) is the short time probability to evolve from \( x_{i\Delta t} \) to \( x_{i+1\Delta t} \) and \( \rho_{x_0} \) is the initial conditions distribution that scales according to the standard Boltzmann factor in the canonical ensemble. The normalization constant or partition function \( Z(T) \) is given by eq 2.16, where the path-integral \( \int Dx \) denotes summation over all pathways \( x \). For an infinite and discretized path this corresponds to summation over all phase space points.

\[
Z(T) \equiv \int P(x) Dx \tag{2.16}
\]

The transition path ensemble is the subset of the reactive trajectories, the ones connecting the stable states \( A \) and \( B \). Similarly, the probability of a path \( P_{AB}(T) \) is given by equations 2.17-2.18.

\[
P_{AB}(x) \equiv h_A(x_0)h_B(x_T)P(x)Z_{AB}^{-1} \tag{2.17}
\]

\[
Z_{AB}(T) \equiv \int h_A(x_0)h_B(x_T)P(x)Dx \tag{2.18}
\]

The characteristic functions \( h_A(x) \) and \( h_B(x) \) equal unity when the system at a configuration \( x \) is in stable state \( A \) or \( B \) respectively and equal zero elsewhere.

Parametrization of the stable states is based on projecting the high dimensional phase space into order parameter(s) which need to fulfil several criteria. Those are:
1) the state definitions should not overlap, 2) the order parameters should distinguish the states and 3) the states should be attractors and commit the trajectory beyond the barrier region with a high probability.
Figure 2.3.1: Schematic representation of the old \( \mathbf{x}^{(o)} \) depicted in solid line) and newly generated paths \( \mathbf{x}^{(n)} \) depicted with dashed lines) a) connecting the two states b,c) not connecting the states and using the two way shooting algorithm by perturbing the momenta at time \( t' \). In d) is described the one-way shooting algorithm where the new path is the old partial path (solid line) plus the newly generated forward partial path (dashed line).

2.3.1 Monte Carlo of trajectories

TPS is a random walk through trajectory space where by construction, trajectories are generated proportionally to their probability in the Transition Path Ensemble (TPE) (eq 2.17). This MC random walk is realized by generating a new trial trajectory and accepting or rejecting it according to the Metropolis criterion. A trial move consists of the generation of a new trajectory \( \mathbf{x}^{(n)} \) from an old trajectory \( \mathbf{x}^{(o)} \) with a probability \( P_{\text{gen}}(\mathbf{x}^{(o)} \rightarrow \mathbf{x}^{(n)}) \) and the acceptance of the newly generated trajectory with some probability \( P_{\text{acc}}(\mathbf{x}^{(o)} \rightarrow \mathbf{x}^{(n)}) \). Here the superscripts 'n' and 'o' abbreviate the 'new' and 'old' path, respectively. As detailed balance ensures that the probability of the forward move is equal to the probability of the reverse move \( (p(o \rightarrow n)=p(o \leftarrow n)) \), in the TPS framework this translates into eq 2.19

\[
P_{AB}(\mathbf{x}^{(o)})P_{\text{gen}}(\mathbf{x}^{(o)} \rightarrow \mathbf{x}^{(n)})P_{\text{acc}}(\mathbf{x}^{(o)} \rightarrow \mathbf{x}^{(n)}) = P_{AB}(\mathbf{x}^{(n)})P_{\text{gen}}(\mathbf{x}^{(n)} \rightarrow \mathbf{x}^{(o)})P_{\text{acc}}(\mathbf{x}^{(n)} \rightarrow \mathbf{x}^{(o)})
\] (2.19)
which states that the probability $P_{AB}$ to observe an old path multiplied with the conditional probability of generating and accepting a new path is equal to the probability of observing the old path multiplied with the probability of accepting and generating the old path from the new path. Using eq 2.17, the fact that $h_A(x_0^{(o)}) = h_B(x_T^{(o)}) = 1$ then the acceptance probability can be written as in eq 2.20,

$$
P_{acc}(x^{(o)} \rightarrow x^{(n)}) = h_A(x_0^{(n)})h_B(x_T^{(n)}) \frac{P(x^{(n)})P_{gen}(x^{(n)} \rightarrow x^{(o)})}{P(x^{(o)})P_{gen}(x^{(o)} \rightarrow x^{(n)})}
$$

(2.20)

where $P_{gen}$ is dependent on the specifics of the algorithm that generates the pathways. In order to fulfil the above equation one can apply a Metropolis rule, shown in eq 2.21

$$
P_{acc}(x^{(o)} \rightarrow x^{(n)}) = h_A(x_0^{(n)})h_B(x_T^{(n)}) \min \left\{ 1, \frac{P(x^{(n)})P_{gen}(x^{(n)} \rightarrow x^{(o)})}{P(x^{(o)})P_{gen}(x^{(o)} \rightarrow x^{(n)})} \right\}
$$

(2.21)

where the $\min$ function returns the smaller of its arguments. This means that if the acceptance probability is one then the move is accepted, whereas if it is lower than one the move is accepted only if the acceptance probability is larger than a random number between 0 and 1.

### 2.3.2 Shooting move

In the seminal TPS paper [48], the algorithm is based on the shooting move, where old reactive paths are perturbed to lead to new accepted paths. The shooting algorithm works as follows. First one selects a time slice $x_{t'}^{(o)}$ which belongs to the old path $x^{(o)}(T)$ with some probability $p_{sel}(t',x^{(o)})$. This time-slice is denoted as the shooting point and in the most straightforward version of the shooting algorithm the shooting point is uniformly selected (it has the same probability along the path). Then one can modify this shooting point by, for example, slightly changing the momenta, followed by a forward and a backward integration of Newton’s equations until $t = 0$ and $t = T$. Due to molecular chaos, the new path is expected to diverge exponentially from the old path. If the path connects the two states it can be accepted with a probability according to eq 2.21. The generation probability can be written analytically as the product of probabilities to select a shooting point $p_{sel}(t',x^{(o)})$, to modify it (e.g. changing the momenta) $p_{gen}(x_{t'}^{(o)} \rightarrow x_{t'}^{(n)})$ and the probability to create a partial path [50]. However, since 1) the dynamics conserves the stationary distribution $\rho_{st}(x)$, microscopic reversibility gives $p(x \rightarrow y)/p(y \rightarrow x) = \rho_{st}(y)/\rho_{st}(x)$ 2) the initial conditions belong to the stationary distribution, $\rho(x) = \rho_{st}(x)$ and 3) the generation and selection probabilities are symmetric, equation 2.21 simplifies to eq 2.22.
\[ P_{\text{acc}}(x^{(o)} \rightarrow x^{(n)}) = h^{(n)}_A(0)h^{(n)}_B(T) \min \left\{ 1, \frac{\rho^{(n)}_{k'}}{\rho^{(o)}_{k'}} \right\} \] (2.22)

Complex biomolecular systems often encounter long transition pathways. The length of these pathways is often much longer than the Lyapunov time \( \lambda \) it takes for the molecular chaos to appear. This means that although the dynamics is deterministic (e.g. Newtonian dynamics) the system essentially behaves diffusively. In these kind of systems, the two way shooting approach with a perturbation of the momenta will lead to a low acceptance ratio [51] as both the forward and backward paths have to reach the correct stable states, while the new pathway diverges quickly from the previous accepted pathway, and is likely to lose its reactivity. The one way shooting alleviates this problem as it only shoots one way per trial (forward or backward) thus increasing the acceptance ratio as only one partial path has to end in the correct state. The new pathway is accepted if the partial path ends in the correct state B for a forward path and A for a backward path. The new path consists of the old part and the new partial path, see Fig. 2.3.1d. The one-way shooting algorithm uses a stochastic thermostat, exploiting both the natural tendency to decorrelate due to the size of the system as well as the stochasticity introduced to the dynamics by the thermostat. No perturbation of momenta is done at the selected shooting point since this would only disrupt the natural NVE dynamics of the whole trajectory. As in the one-way shooting the new accepted path comprises the old and new partial path, the downside is that it takes more accepted paths before one reaches a path that shares no common points with the previous path, a so-called decorrelated path.

For reducing the computational cost of TPS, the flexible path length algorithm was introduced, where transition paths adjust to the length of the transition part, avoiding long dwelling in the stable states [50, 52]. As the selection probability for a shooting point is \( p_{\text{sel}}(t', x^{(o)}) = 1/L^{(o)} \), \( p_{\text{sel}}(t', x^{(n)}) = 1/L^{(n)} \), the acceptance probability becomes

\[ P_{\text{acc}}(x^{(o)} \rightarrow x^{(n)}) = h^{(n)}_A(0)h^{(n)}_B(T) \min \left\{ 1, \frac{L^{(o)}}{L^{(n)}} \right\} \] (2.23)

where \( L^{(o)} \) and \( L^{(n)} \) is the length of the old and new transition path respectively. In our implementation of the one-way shooting, the v-rescale [29] stochastic thermostat was used.
Bibliography


Chapter 3

Dynamics of hydration water around native and misfolded α-lactalbumin

As water is an essential ingredient in protein structure, dynamics and functioning, knowledge of its behavior near proteins is crucial. We investigate water dynamics around bovine α-lactalbumin by combining molecular dynamics simulations with polarization resolved femtosecond infrared (fs-IR) spectroscopy. We identify slowly reorienting surface waters and establish their hydrogen-bond lifetime and dynamical orientation relaxation dynamics, which we compare to the experimentally measured anisotropy decay. The calculated number of slow surface waters is in reasonable agreement with the results of fs-IR experiments. Slow waters form fewer hydrogen bonds compared to the bulk. At concave sites the protein-water hydrogen bonds break preferably via translational diffusion rather than via a hydrogen-bond jump mechanism. The reorientation of water molecules residing at these concave sites is slower than at convex water exposed sites. Protein misfolding leads to an increased exposure of hydrophobic groups, inducing relatively faster surface water dynamics. Nevertheless, the larger exposed surface slows down a larger amount of water. While for a native protein hydrating water is slower near hydrophobic residues than at hydrophilic sites, mainly due to stronger confinement, misfolding causes hydrophobic water to reorient relatively faster because the exposure of hydrophobic groups destroys concave protein cavities with a large excluded volume.
3.1 Introduction

The protein hydration shell is very significant and essential to protein structure, dynamics and function. The hydration shell dynamics relates to biochemical processes such as protein folding, molecular recognition and enzyme function[1–3].

Both experimental [3, 4] and simulation findings [5–7] agree on that a protein perturbs the water dynamics in its hydration shell. However, the root and the degree of this perturbation is not completely understood. Experimental techniques such as time-resolved fluorescence and NMR indicate a slowdown of water molecules close to the protein surface. NMR experiments [8], and molecular dynamics (MD) simulations[5, 7], show that most of the hydration shell water reorients 2-3 times slower compared to bulk water. The origin of this slowdown has been mostly attributed to the excluded volume effect due to topology and to a lesser extent to the strength of the hydrogen bond [5, 7].

Moreover, MD simulations [7] indicated that increasing the excluded volume and decreasing the hydration level by adding hexane molecules at the surface, led to shifted and broader reorientation time distributions. The effect of confinement on water dynamics has been assessed experimentally by polarization-resolved femtosecond infrared spectroscopy [9, 10], which showed that water near the surface of a (reversed) micelle is slow, while water in the core behaves very similar to bulk water. These experiments showed that confinement affects the dynamics of water, but only on very small length scales (sub nanometer), i.e. close to a surface. Femtosecond 2D infrared spectroscopy combined with polarization-resolved femtosecond infrared spectroscopy[11] showed that the slow component of water reorientation dynamics scales with the slow translational diffusion, indicating slow hydrogen bond dynamics near hydrophobic groups. Interestingly, in a MD study on a amyloidgenic αβ-disordered protein at room temperature, Jose et al. [12] found that water molecules around the protein exhibit faster reorientation and translation dynamics compared to water hydrating a native globular ubiquitin in water. Further studies conducted by Rahaman et al[13] observed also an acceleration of water dynamics around a temperature induced-unfolded thermophile protein compared to a folded one. Temperature induced unfolding or misfolding of the proteins alters the amount of confined water, and changes the hydrophobic/hydrophilic nature of the protein surface. The influence of this temperature induced misfolding on the water dynamics remains not well understood. In this work, by combining molecular dynamics simulations with infrared spectroscopy, we investigate the dynamics of water around the protein surface of α-lactalbumin, a major constituent of whey, and an important protein in the food industry[14]. As α-lactalbumin is a small (14 kDa) globular protein that is structurally quite similar to the widely studied
lysozyme [15], it forms an excellent system to study water dynamics in the protein hydration shell. We conducted MD simulations of several solvated $\alpha$-lactalbumin systems at ambient conditions. For a native solvated protein, we conducted two different concentrations and two different water models. For a misfolded solvated protein we conducted one concentration and one water model. In addition, we measured the water reorientation dynamics using polarization-resolved femtosecond infrared spectroscopy. The remainder of the paper is organized as follows. Section 3.2 describes the employed simulation, analysis and experimental methodology. In section 3.3 we present and discuss the analysis results, and compare to the experiments. We end with conclusions in section 3.4.

### 3.2 Methods

#### 3.2.1 Simulation setup

All molecular dynamics simulations were performed with the Gromacs 4.5.4 package [16]. The bovine $\alpha$-lactalbumin monomer structure, extracted from the PDB 1F6S [17] was solvated with water molecules (SPC/E or TIP4P/2005), resulting in dilute solutions of approximately 5 % w/w and more concentrated solutions of 8.5% w/w (denoted 5% SPC/E, 5% TIP4P/2005, 8.5% SPC/E and 8.5% TIP4P/2005 system, respectively).

The atomic interactions were defined by the amber99sb-ildn force field [18]. The protonation state of the amino acids corresponds to pH of 7, and five Na atoms were added to neutralize the system (pKl= 4-5). For each water force-field, after energy minimization, the system was equilibrated for 1 ns at ambient conditions (298 K and 1 atm) in the NPT ensemble. All bonds were constrained with the Lincs algorithm. A cutoff of 1 nm was used for the non-bonded Lennard-Jones interactions. The Particle Mesh Ewald method was used to calculate the electrostatic interactions with a Fourier spacing of 0.12 nm and a 1 nm cutoff for the short range electrostatic interactions. Neighbor lists were updated every 10 fs with a cutoff of 1 nm and the time step was 2 fs [18]. The leap-frog algorithm was used for integrating Newton’s equations of motion. In the NPT simulations the v-rescale thermostat[19] with a coupling time constant of 0.2 ps controlled the temperature, while the Parrinello-Rahman barostat[20] with a coupling time constant of 1.0 ps kept the pressure constant. From a 100 ns long MD trajectory in the NPT ensemble at ambient conditions (298 K and 1 atm) 10 frames were randomly selected from different parts of the trajectory. From each frame we initiated a short 1 ns production run in the NVE ensemble, by switching off the thermostat and barostat. This approach eliminated any unwanted influence from the thermostat or
barostat on the dynamics. To prevent energy drift we used a switching function for the non-bonded interactions from 0.8 -1.0 nm. The pair lists were updated every 5 fs with a cutoff of 1.2 nm and the time step was 1 fs. The frequency of the energy calculation was 10 fs, and the frames were saved every 100 fs in order to obtain sufficient data for the analysis.

The effect of protein misfolding on the water dynamics was investigated by analyzing two types of quenched high temperature systems. Starting from the last frame of the 100 ns ambient condition MD trajectories (for the 5% w/w SPC/E system), we ran another 100 ns at high temperature (700 K) followed by two types of quenching procedures. In the first, five high temperature frames were selected and relaxed at 298 K temperature (NPT) for 10 ns. Then, for each of these five room temperature MD runs, a frame corresponding to the average temperature, energy, and volume was selected and used to initiate a 1 ns NVE simulation for analysis, using the high frequency frame saving. This system is labelled *misfolded #1*. In the second procedure one frame corresponding to the average temperature and energy was selected from the high temperature run. Using this as an initial frame, a 100 ns simulation was performed at room temperature using a NPT simulation. From this simulation ten frames corresponding to the average energy and temperature and volume were selected, and from each, a 1 ns NVE run was conducted. This system is labelled *misfolded #2*. The ten 1 ns trajectories were subjected to dynamical analysis. We studied the misfolded protein at ambient conditions rather than elevated temperature where the unfolding occurs, in order to keep the bulk water behaviour as close as possible to the native state simulations, and focus entirely on the differences in dynamics induced by the altered protein surface.

### 3.2.2 Analysis

#### Hydrogen bond life times

For the rest of the chapter hydrogen bonds will be referred to measured as H-bonds. A H-bond was defined present if the distance between H-bond donor $D$ and acceptor atom $A$, $R_{DA} \leq 0.35$ nm, and the angle between the $DH$ and $HA$ vectors $\theta_{ADH} \leq 30^\circ$ [21, 22]. Here, either the donor and acceptor atom can belong to the protein or to water. The H-bond lifetime was defined as the time it takes for the bond to break and to form a stable bond with another acceptor (exchange), or to remain dangling (translational diffusion). If the bond reforms within 200 fs, it is considered not broken, in order to avoid counting fast recrossings events.

The recrossing time $\Delta \tau = 200$ fs was based on the fact that the vibrational characteristics of the water water H-bond occur in the far infrared and around $200 \text{ cm}^{-1}=166$
Different values of this recrossing time $\Delta \tau = 0$ fs and $\Delta \tau = 400$ fs were used to test whether this choice affects the results.

**Water reorientation dynamics**

Reorientation dynamics can be represented using the time correlation function [25, 26]

$$C_2(t) = \frac{2}{5} \langle P_2[u(0) \cdot u(t)] \rangle$$  \hspace{1cm} (3.1)

where $P_2$ denotes the second Legendre polynomial, $u$ is the unit vector characterizing the orientation of an OH group in a given frame and the angular brackets denote an ensemble average over all water molecules and all time origins. This correlation function can easily be related to anisotropy curves obtained from polarization-resolved femtosecond infrared spectroscopy and to orientation relaxation times from magnetic relaxation techniques [25].

From MD simulations we compute the reorientation dynamics for each individual molecule. We are particularly interested in the anisotropy decay of water molecules hydrating different protein sites. The reorientation dynamics of individual water molecules was investigated by following the dynamics of both OH bond vectors of each molecule $j$. For each bond unit vector $u_{jk}(t)$, we computed the time correlation function

$$c_j^m(t) = \frac{1}{5\ell} \sum_{t'=t_0^m}^{t_0^m+\ell \Delta t} \sum_{k=1}^2 P_2[u_{jk}(t') \cdot u_{jk}(t'+t)]$$  \hspace{1cm} (3.2)

where the sum over $k = 1, 2$ refers to the two OH bond vectors, and the time correlation is summed over the $m$th time interval $t_0^m < t' < t_0^m + \ell \Delta t$ that the molecule is within the hydration layer of the protein (defined as the water oxygen being within 4.4 from the protein heavy atoms). We used a buffer time of 2 ps to avoid counting fast non-essential excursions inside and outside the hydration layer [27]. For each curve $c_j^m(t)$ we extracted the reorientation decay time $\tau_j^m$ by fitting $c_j^m(t)$ to a single exponential fit in the interval $0 < t < 10$ ps (see Fig. 3.A.3). Each estimate $m$ was viewed as a separate measurement, since the water molecule can change its dynamics when leaving or entering the hydration layer. For each water molecule $j$ and for each $m$th interval of length $\ell$ the decay time $\tau_j^m$ was histogrammed with a weight $\ell$, leading to a probability distribution of decay times. The weight $\ell$ follows from the fact that the correlation in eq. 3.2 occurs $\ell$ times in eq. 3.1. The average decay time $\tau$ from these histograms thus should be close to the overall decay time of eq.3.1. The decay times of the individual water molecules allow us to establish the relation between the water structure and dynamics. We therefore divide the waters into several categories.
For each water molecule $j$ and for each $m$th interval $\ell$ the decay time $\tau_{m j}^{\ell}$ was histogrammed with a weight $\ell f_{HB}$ in the protein H-bonded category or with a weight $\ell f_{nHB}$ to the category of water that is non H-bonded to the protein. Here $f_{HB}$ and $f_{nHB}$ refer to the fraction of frames in the interval $\ell$ in which the water $j$ was hydrogen bonded to the protein or not.

A similar procedure was followed for categorizing and histogramming hydrophobic or hydrophilic water. For each water molecule $j$, the decay time $\tau_{m j}^{\ell}$ for each $m$th interval $\ell$ was histogrammed with a weight $\ell f_{HP}$ and $\ell f_{HF}$ for hydrophobic and hydrophilic water, respectively. Here $f_{HP}$ and $f_{HF}$ are the fraction of frames in the interval $\ell$ in which the water $j$ solvated respectively, a hydrophobic and hydrophilic group (i.e. the water oxygen was within 4.4 of, respectively an apolar or polar residue side chain).

In both cases the sub-population histograms add up to the total probability histograms. Errors in the mean were estimated from block averaging.

### 3.2.3 Polarisation-resolved femtosecond infrared spectroscopy

Experimentally, we measured the water reorientation dynamics using polarization-resolved femtosecond infrared spectroscopy [9–11, 24, 28–30]. Protein solutions were prepared by mixing bovine $\alpha$–lactalbumin (purity > 90%, Davisco foods) with isotopically diluted water, consisting of ultrapure milli-Q grade $H_2O$ and 4% D$_2$O (99.9 %D, Cambridge Isotope Laboratories). The deuterated hydroxyl (OD) groups in this solution absorb strongly at 2500 cm$^{-1}$, which corresponds to the stretch vibration of the OD group.

We excited and probed the OD stretch vibrations using resonant infrared pulses. These pulses were generated by frequency conversion of the output of a Ti:sapphire regenerative amplifier that produces 850 $\mu$J, 100 femtosecond pulses with a center wavelength of 800 nm at a repetition rate of 1 kHz. Part of the light was used to pump a BBO–based optical parametric amplifier. The 2 $\mu$m idler pulses produced by this amplifier were frequency–doubled in another BBO crystal and subsequently mixed with the remaining 800 nm light in a lithiumniobate crystal, yielding 10 $\mu$J pulses at 2500 cm$^{-1}$, with a pulse duration of 200 fs and a bandwidth of 100 cm$^{-1}$. These pulses were separated into pump, probe and reference beams by a wedged $CaF_2$ window and focused onto the sample by a parabolic mirror. The sample consisted of the protein solution sandwiched between two $CaF_2$ windows, held apart by a 50 $\mu$m spacer. After passing through the sample, the probe and reference beams were recollimated by a second parabolic mirror, dispersed in a grating-based spectrometer and detected with two lines of a 3x32 mercury–cadmium–telluride detector array.

The pump pulse induced transient absorption changes in the sample that we monitored with the probe pulse (the reference pulse was used for normalization to correct
for pulse to pulse fluctuations of the infrared light). We measured the transient absorption for probe pulses with their polarization parallel and perpendicular with respect to the pump polarization. Since the pump pulse most efficiently excites OD vibrations with their transition dipole moment parallel to the polarization direction of the light, initially the parallel absorption signal is higher, while after some time the orientation of the OD stretch vibration has randomized and the parallel and perpendicular signals become equal. From the parallel ($\alpha_{\parallel}$) and perpendicular ($\alpha_{\perp}$) transient absorption signals we constructed the anisotropy:

$$C_2(v,t) = \frac{\alpha_{\parallel}(v,t) - \alpha_{\perp}(v,t)}{(\alpha_{\parallel}(v,t) + 2\alpha_{\perp}(v,t))}$$ \hspace{1cm} (3.3)

The anisotropy $C_2$ of eq. 3.3 decays with the rate of molecular reorientation and is directly proportional to the second-order reorientation correlation function shown in equation 1.

Table 3.2.1: Simulation: Average protein water and bulk water H-bond lifetimes for six different systems 1) 5% protein in SPC/E water 2) 8.5% protein in SPC/E water 3) misfolded #1 4) misfolded #2 5) 5% protein in TIP4P/2005 water 6) 8.5% protein in TIP4P/2005 water. The number subscript indicates the error in the last digit.

<table>
<thead>
<tr>
<th>system #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_{bulk}$</td>
<td>2.45$_1$</td>
<td>2.47$_1$</td>
<td>2.45$_1$</td>
<td>2.47$_1$</td>
<td>2.90$_1$</td>
<td>2.95$_1$</td>
</tr>
<tr>
<td>$t_D$</td>
<td>3.27$_5$</td>
<td>3.41$_{13}$</td>
<td>3.43$_7$</td>
<td>3.61$_{15}$</td>
<td>3.93$_5$</td>
<td>3.84$_5$</td>
</tr>
<tr>
<td>$t_A$</td>
<td>4.62$_9$</td>
<td>4.70$_{14}$</td>
<td>4.53$_{15}$</td>
<td>4.67$_{13}$</td>
<td>5.72$_{10}$</td>
<td>5.61$_{21}$</td>
</tr>
<tr>
<td>$t_{D+A}$</td>
<td>3.99$_6$</td>
<td>4.23$_{14}$</td>
<td>4.01$_{14}$</td>
<td>4.17$_{14}$</td>
<td>4.84$_7$</td>
<td>4.74$_7$</td>
</tr>
<tr>
<td>$t_{directD}$</td>
<td>1.99$_5$</td>
<td>1.53$_5$</td>
<td>2.15$_2$</td>
<td>2.14$_5$</td>
<td>1.97$_5$</td>
<td>1.88$_3$</td>
</tr>
<tr>
<td>$S_{HB}$</td>
<td>1.62</td>
<td>1.71</td>
<td>1.63</td>
<td>1.69</td>
<td>1.67</td>
<td>1.60</td>
</tr>
</tbody>
</table>

### 3.3 Results and Discussion

#### 3.3.1 Hydrogen bond dynamics

The computed H-bond life times for the six studied systems are shown in Table 3.2.1. The average and standard deviation is taken from a block average of 10 sets of data per system. While the H-bond lifetime is sensitive to the selected recrossing time, the H-bond lifetimes follow the same trend when using different recrossing time $\Delta \tau$ [31]. We
Figure 3.3.1: Representation of the H-bond type. Left: a protein donating a H-bond to water, labelled directD and the accepting water donating a H-bond to another water labelled D. Right: a protein accepting a H-bond from water, labelled A. The classification includes both backbone and side-chain H-bonds.

distinguish between water molecules that are H-bonded to other water molecules in the bulk, and water molecules that are donating or accepting H-bonds from the protein, as illustrated in Fig. 3.3.1. A protein H-bond donor leads in fact to two types of H-bonds, labelled the directD and D, with the first one being more informative about the direct protein water interaction and the second one giving more relevant information on the reorientation of the OH vectors of the water accepting a H-bond from a protein donor (see Fig. 3.3.1 left). The H-bond type when the protein is accepting a hydrogen from water is labelled A (see Fig. 3.3.1 right). Table 3.2.1 shows that the overall average protein-water H-bond lifetime ($t_{D+A}$) is always longer than the H-bond lifetime in bulk water ($t_{bulk}$) for all systems in this study, mainly due to bonds between accepting protein groups and water ($t_A$). The relative increase of the protein-water H-bond lifetimes at the protein surface can be expressed as the ratio $S_{HB}$ of the average protein-water H-bond lifetime over the average lifetime in the bulk (see Table 3.2.1). The protein-water H-bond lifetime distribution is broad, including short H-bonds but also long-lived bonds, as indicated by the long tail in Fig. 3.A.1. The protein H-bond acceptors ($t_A$) form longer-lived bonds than the protein H-bond donating groups ($t_D$ and $t_{directD}$), which agrees well with the work of Sterpone et al[5, 32] who showed that the water dynamics is slowed down more at protein H-bond accepting groups. Upon misfolding the protein-water H-bond lifetimes does not change much compared to the folded native state. However, the average lifetimes of H-bonds between protein donating groups and water ($t_D$ and $t_{directD}$) increases upon misfolding, suggesting the existence of water mediated interactions between the protein residues in the misfolded systems, which are known to mediate folding of misfolded and unfolded systems [33]. Such a water mediated interaction is illustrated in Fig. 3.A.2. The H-bond dynamics in the 5% TIP4P/2005
system is substantially slower compared to the 5% SPC/E system. The larger standard deviation of protein water H-bond lifetimes upon misfolding is due to the different conformations in the ten analyzed windows of the misfolded system. The effect of protein concentration on the protein-water H-bond dynamics is marginal. Note that for the SPC/E system, the protein-water H-bond dynamics slightly decelerates with concentration, whereas it slightly accelerates for the TIP4P/2005 system.

To investigate the nature of the water-protein H-bond dynamics, we analyzed the H-bond breaking mechanism as a function of the protein site. For this process, Laage et al [34] introduced a framework involving large-amplitude angular jumps upon H-bond breakage. Figure 3.3.2 depicts the percentage of protein water H-bonds breaking via jumps as a function of the protein sites. This figure clearly shows a correlation between the curvature of the protein site (excluded volume) and the mechanism of H-bond breaking. For convex protein sites, protein water H-bonds break more likely via a jump mechanism, whereas for concave sites (occurring more in buried residues) H-bonds break via translational diffusion. This finding was also reported in the work of Brandeburgo et al [24], and can be rationalized by realizing that at concave sites there is less space for new hydrogen-bond partners, making it less likely that a water OH will find a new partner to jump to through an angular jump, thus increasing the chance of H-bonds breaking via diffusion.
3.3.2 Reorientation dynamics

Anisotropy decay of water

Figure 3.3.3(left) shows for the six simulated systems the computed anisotropy decay averaged over all water molecules, as well as the experimental anisotropy decay curve for a 5% w/w protein solution, measured by polarization-resolved femtosecond infrared spectroscopy. While the computed anisotropy curves for SPC/E are systematically lower than the experimental one, their slope is very similar. As the hydration water comprises only 6% of the overall water in the simulation box, these decay curves are dominated by the reorientation time of bulk water. Fig. 3.3.3(right) shows the simulated anisotropy decay for water molecules that are initially in the hydration layer (within 4.4 Å of the protein side chain heavy atoms). In accordance with the work of Sterpone et al. and Fogarty et al. [5, 7], for all systems the decay time of the hydration shell anisotropy is longer compared to the corresponding overall water anisotropy decay of Fig. 3.3.3(left). Thus, the water reorientation dynamics in the hydration shell of the protein is slowed down for all systems under study.
Slow waters

The anisotropy decay curve can be described at intermediate times ($0.5 < t < 7$ ps) with the expression

$$C_2(t) = R_0 e^{-t/t_r} + R_1$$  \hspace{1cm} (3.4)

where the constant $R_1$ represents the water molecules in the hydration layer of the protein that reorient very slow ($t > 7$ ps), while most of the water molecules reorient with bulk-like reorientation time $t_r$. We limit our analysis to the first 7 ps of the anisotropy decay because the standard deviation in the ultrafast infrared experiments increases strongly with increasing delay time. Reducing the water content by increasing the concentration leaves $t_r$ more or less constant. For the simulations $t_r$ is $2.417 \pm 0.031$ and $2.40 \pm 0.02$ for 5% w/w SPC/E and 8.5% w/w SPC/E respectively, close to the experimental value of bulk like water $t_r= 2.45 \pm 0.15$. The fraction of slow waters $f_{\text{slow}} = R_1/0.4$ (0.4 is the maximum value of the anisotropy) increases with the protein concentration, both in experiment and simulations. Figure 3.3.4 shows $f_{\text{slow}}$ as a function of protein concentration. The slope of the plot of $f_{\text{slow}}$ versus the concentration (in mol/kg water), multiplied with 55.25 (mol water molecules per kg) then yields the number of slow waters per protein molecule. The experimental data yield $323 \pm 16$ water molecules. For the SPC/E simulations, the number of slow water molecules is $350 \pm 33$. The SPC/E simulation data (bulk, 5% and 8.5%) in Fig. 3.3.4 fall within the error bars of the experimental values, and are hence in relatively good agreement with the experiment.
Analyzing the simulation results, we can also compute the amount of slow water molecules from the distribution of reorientation times as the fraction of the protein hydration shell water population reorienting slower than $\tau = 7$ ps (see Fig. 3.3.5 right). Multiplying this fraction with the average number of hydration waters gives the number of slow water molecules (again per protein molecule since there is only one protein in the simulation box), which can again be compared with the experiment. For the 5% SPC/E system the simulations yield $294 \pm 7$ slow water molecules, in fairly good agreement with the number of slow water found by experiments. The discrepancy between the two SPC/E estimates lies in the different ways of estimating the slow water content.

In contrast, for the 5% TIP4P/2005 system we find a higher number of slow water molecules compared to experimental findings. Calculating the number of slow water molecules from the slope of $f_{\text{slow}}$ and the reorientation time distribution we find $597 \pm 45$ and $519$ slow water molecules, respectively. The slower dynamics of TIP4P/2005, is to be expected, since Vega et al. [35] report a higher reorientation time for the TIP4P/2005 bulk water compared to the SPC/E bulk water by a factor of 1.2. The SPC/E model predicts the static dielectric constant better than TIP4P/2005 [35] for the bulk water, therefore, it is expected to reproduce the experiments better.

**Triexponential fit of hydration water relaxation**

As the experimental signal to noise ratio vanished for times larger than 7 ps and because the total anisotropy curve is dominated by bulk water relaxation, it does not make sense to fit to functions more complex than eq. 3.4 to the experimental data. However, the situation is different for the hydration layer water molecules as calculated from the simulations. As these waters are significantly slower than bulk water and experience a heterogeneous environment, it is worthwhile to fit the simulated hydration shell anisotropy decay (Fig. 3.3.3 right) to a triexponential function for times $0 < t < 70$ ps:

$$C_2(t) = a_0 e^{-t/\tau_0} + a_1 e^{-t/\tau_1} + a_2 e^{-t/\tau_2}$$ (3.5)

with $a_0, a_1, a_2$ the amplitudes of the three exponential decay modes. There are three characteristic decay times a sub-picosecond ($\tau_0$) a moderate ($\tau_1$) and a slow ($\tau_2$) time scale. The sub-picosecond reorientation time is due to water libration modes, while the moderate and slow component indicate slower water reorientation mechanisms, dependent on the interaction with the protein. This observation agrees with the work of Brandeburgo et al. [24]. The relaxation times are larger for the TIP4P/2005 systems compared to the SPC/E systems, as shown in Table 3.A.1, 3.A.2 and Fig. 3.3.3. Note also that, upon misfolding, the moderate and slow characteristic time scales decrease, indicating a speed up of hydration water.
Figure 3.3.5: Simulation: probability distribution of reorientation decay $\tau$ of water molecules based on a geometrical ensemble, initially outside of the hydration shell (left) and initially in the hydration shell (right), for the six systems of study: 5% SPC/E (red), 8.5% SPC/E (dark green), misfolded #1 (orange), misfolded #2 (yellow), 5% TIP4P/2005 (blue), 8.5% TIP4P/2005 (magenta).

**Distribution of reorientation times**

The reorientation time of a water molecule clearly depends on its environment. This heterogeneity in the dynamics is visible in Fig. 3.3.5 which shows the probability distribution of the reorientation time $\tau$ for bulk water (Fig. 3.3.5 left) and for water molecules that are initially in the protein hydration layer (Fig. 3.3.5 right). The distribution for hydration water is much broader and extending toward longer reorientation times than the bulk water distribution (Fig. 3.3.5 left) which is roughly Gaussian, centered around 2.5 ps for SPC/E water and around 2.9 ps for TIP4P/2005 water. Note that the percentage of the hydration shell water reorienting slower than 100 ps is really small, less than 1.5% out of the overall hydration shell water. Fig. 3.3.6 visualizes the water reorientation time as function of protein site. It can be seen that some buried residues are hydrated by very slow internal waters, whereas at the surface, the water reorientation dynamics is relatively faster. As indicated in section 3.3.1, water that is initially H-bonded to the protein at concave sites is more confined and more likely to break its bond (type A or D) through translational diffusion [24]. Due to the reduced possibility to break the H-bond via a jump mechanism through a bifurcated hydrogen-bonded transition state, these water molecules reorient on longer timescales.

**Splitting the distributions**

Figure 3.3.5(right) shows the probability distribution for reorientation times of the hydration layer waters. These distributions are non Gaussian due to the heterogeneous environments of the hydrating water molecules. In addition bulk-like water still con-
tributes to this distribution due to the definition of the hydration layer, and gives a peak at 2.5 ps, the characteristic reorientation time for bulk water. To obtain more insight into the origin of the heterogeneous reorientation time distribution in the hydration layer we categorize the water molecules based on their environment. For instance, we can subdivide the hydration water populations in hydrophobic and hydrophilic waters, based on the nearest protein residue (see methods). Figure 3.3.7 (first column) and shows such. Rather surprisingly, the hydrophilic and hydrophobic distributions do not show much difference in shape. This might be due to our weighting procedure, which mixes the hydrophilic and hydrophobic categories to some extent. Table 3.A.2 reports the average reorientation time of the categories for all systems. For the native state systems, the hydrophobic water is slightly slower than hydrophilic water, possibly due to the fact that hydrophilic side groups are slightly more accessible to bulk like water.

Another way we can categorize the water molecules is in terms of H-bonding to the protein. Figure 3.3.7 (right column) shows the distributions for this division a folded and misfolded system. The water H-bonded to the protein is slower than the water that is not H-bonded to the protein. Note, that the latter category can be either
Figure 3.3.7: Simulation: Number distributions of reorientation decay times of the hydration water for 5% SPC/E (top row) and misfolded #1 (bottom row). The first column depicts the overall hydration water (red), and its subdivision into molecules hydrating hydrophilic (green) and hydrophobic (blue) groups. The second column depicts the overall hydration water, (red), and its subdivision into molecules that are H-bonded to the protein (green) and molecules that are non H-bonded to the protein(blue).

hydrophobic or hydrophilic water which is not H-bonded to the protein. The protein H-bonded water distribution exhibits a less pronounced bulk-like water peak, and has a long tail while the non protein H-bonded water has a larger contribution of bulk-like water. Table 3.A.2 shows that indeed the average reorientation time for the protein H-bonded waters ($\tau_{HB}$) in the native state is significantly longer than for the non protein H-bonded waters ($\tau_{nHB}$).

One can further subdivide the water molecules that are H-bonded to the protein into waters that donate their hydrogen to a protein acceptor and waters that accept a bond from a protein donor. As shown in Table 3.A.2, in all systems water donating its hydrogen to a protein acceptor ($\tau_{HB_{acc}}$) reorients slower compared to water that is accepting a hydrogen from a protein hydrogen donor ($\tau_{HB_{don}}$). This observation is in accordance with the results of Table 3.2.1 which showed that protein H-bond acceptor ($t_A$) form longer-living bonds than the protein hydrogen donating groups ($t_D$ and $t_{directD}$). We
find that the strength of the H-bond, also plays a role in the reorientation dynamics, shown also by Usui et al \[36\].

The curves in Fig 3.3.7 suggests three reorientation time regimes for the surface population, a bulk like (0-4 ps) peaked at 2.5 a moderately slow (4-20 ps) and a very slow (> 20ps). Therefore we can also compare the different types of water based on their distribution over these time regimes (table 3.A.3).

Table 3.A.3 shows the population of hydration shell water molecules reorienting bulk-like, moderately slow or very slow according to its category: hydrophilic (HF), hydrophobic (HP), H-bonded to the protein (HB), non H-bonded to the protein (nHB), and finally all waters (all). While the hydrophobic/hydrophilic division hardly affects the fractions, clearly the waters that are H-bonded show a significantly larger fraction in the slow category, as expected. Again, TIP4P/2005 shows a slower behavior than SPC/E. Note that upon increasing the protein concentration (reducing the bulk water) for both SPC/E and TIP4P/2005 system, the hydration shell water reorients on average the same (see Table 3.A.2).

**Correlation of reorientation time with H-bond coordination number**

The dynamical heterogeneity of the protein hydration water is also observable experimentally. Fig 3.3.8 shows the measured anisotropy decay as a function of frequency at different picosecond delay times. For bulk water, the anisotropy decay is independent of frequency; but for water in α-lactalbumin solutions the anisotropy decay becomes frequency dependent, with a slower decay on the high frequency side. As the
frequency of the OD vibration depends on the H-bond strength, with strong H-bonds corresponding to low OD frequencies and weak H-bonds corresponding to high OD frequencies, this plot shows that water molecules with weaker H-bonds (high frequencies) reorient slower. This frequency dependence is not observed for small amphiphilic solutes [13, 33], and therefore likely originates from the protein structure. This is consistent with the observation that the slowest reorienting waters are located in concave sites on the structured protein surface (Fig. 3.3.6). It also suggests that there is a correlation between the position of a water molecule at the protein surface and its H-bond strength. To further investigate this, we analyze the H-bond coordination in the protein solutions, which directly influences the frequency of the OD stretch vibration [37].

Recently, Auer and Skinner [37] used Raman and VSF spectroscopy experiments to classify molecules in terms of H-bond structure in bulk and at a liquid vapour interface. In particular, they classified each water molecule by its number of H-bond acceptor $n_A$ and donors $n_D$. Here we adopt this classification of the H-bond structure, using a slightly different notation ($n_A-n_D$), where $n_A$ is the number of accepted H-bonds, and $n_D$ the number of donated H-bonds (see Fig. 3.3.9a for an illustration). Note that these numbers include bonds accepted from and donated to the protein. For bulk water, the population of the different H-bond classes was found to decrease in the order $(2-2) > (1-2) > (2-1) > (1-1) > (0-2)$. Our simulations show the same order for bulk water in the different protein systems (see Fig. 3.3.9 and Fig. 3.A.4). We find the average number of H-bonds of bulk water to be 3.54 and 3.62 for SPC/E 5% and 8.5%, respectively, in accordance with reported values of 3.59 for SPC/E water [38]. Fig. 3.3.9b shows that near a protein water interface, in this case for the 5% SPC/E system, the water H-bond class population decreases in the order $(2) > (1-2) > (2-1) > (1-1) > (0-2)$, which is the same as for bulk water. However, comparing protein hydration shell water with bulk water the populations $(1-2)$ and $(1-1)$ are enhanced and $(2-1)$ and $(2-2)$ are reduced in the hydration shell. This shift leads to an average of 3.31 H-bonds for the hydration shell water, which is less compared to bulk water (3.59 H-bonds). Similar results are found for the 5% TIP4P/2005, misfolded #1 and misfolded #2 systems with average H-bond numbers of 3.43, 3.29 and 3.29, respectively.

To relate the contribution from each of the H-bond classes to the water reorientation time distributions we partitioned these distributions based on the coordination labels ($n_A-n_D$). The resulting distributions are shown in Fig. 3.3.9c and Figs. 3.A.5 to 3.A.10.

The H-bond class of a water molecule does not seem to have a dramatic effect on the reorientation time distribution. To see the difference more clearly in Fig. 3.3.9d we separate for the 5% SPC/E system the distributions for the different populations into three reorientation time regimes: a bulk like (0-4 ps) peaked at 2.5, a moderately slow (4-20 ps) and a very slow (> 20ps). Even though the populations themselves
are widely spread, there is a clear correlation in this table: water with more donated H-bonds are slower, waters with more accepted H-bonds faster (increased bulk like population). This pattern occurs in all hydration water, but not in bulk, where there is no correlation between the water structures and the dynamics. This correlation is made even clearer in Table 3.A.4 which shows the fraction of water as a function of number of donated and accepted H-bonds, irrespective of class. The correlation however, does
not mean causation. Much more likely, the presence of the surface induces both effects: fewer accepted H-bonds, and slower reorientation dynamics.

**Effect of misfolding on reorientation dynamics**

There seems to be a negligible difference between the two types of misfolded systems in terms of the reorientation decay times, as the average $\tau$ values given in Table 3.A.2 are similar and the anisotropy decay curves in Fig. 3.3.3 are identical within the error. However, the hydration shell water reorientation time is faster compared to the native structure (see Table 3.A.1, 3.A.2 and Fig. 3.3.3right). This is in accordance with the work of Jose et al [12] who showed that the disordered $\text{A}_\beta_{1-42}$ amyloid protein exhibits faster water reorientation dynamics than a globular ubiquitin protein.

From the distributions of Fig. 3.3.7 and Fig. 3.A.8, the calculated number of slow water molecules for the misfolded systems are 373 and 401 for misfolded #1 and misfolded #2, respectively, higher than for the native 5% SPC/E system. This increase in the number of slow waters is due to the larger solvent accessible surface of a misfolded protein, with respect to the native state. This also follows from Fig. 3.3.7 which compares the number distributions for the native and misfolded protein as a function of decay time $\tau$ for surface waters near hydrophobic/hydrophilic groups. Similar to the native state results, both hydrophobic and hydrophilic water have a bulk like peak around 2.5 ps for the misfolded conformations. The largest difference between the native and misfolded state is the enhanced contribution of hydrophobic hydration water in the misfolded state. Not surprisingly, misfolding induces more hydrophobic residues to be exposed.

Table 3.A.3 compares the fraction of the bulk-like, moderate and slow hydration shell waters of the native and misfolded systems. The hydrophobic water in the misfolded systems exhibits a higher fraction of bulk-like water and a lower fraction of moderate and slow water compared to the native systems, and indeed, a smaller average reorientation decay time for the hydrophobic water (see Table 3.A.2). Interestingly, not only the water dynamics accelerates upon misfolding, but the hydrophobic water becomes actually slightly faster compared to the hydrophilic water. This larger fraction of non surface-perturbed waters was also observed in the work of Rahaman et al [13].

The contribution of water not H-bonded to the protein (nHB) to the decay time distributions in Fig. 3.3.7 and Figs. 3.A.5 to 3.A.10 is significantly higher for the misfolded state than for the native state. At the same time the contribution of water H-bonded to the protein (HB) only increases slightly between the native and misfolded state. This increased population of water non-H-bonded to the protein represents to a large extent the higher hydrophobic water population and similarly shows a slightly faster reorientation.
Table 3.A.1 and 3.A.2 show that the average hydration layer water in the misfolded state tends to be faster than in the native states. Water that is H-bonded to the protein remains slower than the water non-H-bonded to the protein, and as in the native systems, shows a less pronounced bulk-like peak. The H-bonded water is slightly faster in the misfolded state, unlike the H-bond lifetimes that remain unchanged. This suggests the strength of the bond does not change upon misfolding, but rather the decreased local excluded volume upon misfolding causes water to reorient faster (more bulk-like).

Interestingly, upon misfolding water that accepts a H-bond from a protein donor slightly slows down (see Table 3.A.2). In accordance with the increase of the protein donor - water H-bond lifetime \(t_D, t_{Direct}\) upon misfolding this bond does become stronger and therefore water reorients slower. This can be explained by the presence of water mediated H-bonds in a misfolded state (see Fig. 3.A.2). Fig. 3.3.6 shows projections of the water reorientation time on structural representations of the protein. Finally, Figs. 3.A.5 to 3.A.10 show that misfolding does not change the order of the population of H-bonding classes, but again that hydrophobic groups become more solvent accessible.

### 3.4 Conclusions

Using ultrafast IR spectroscopy experiments and molecular dynamics simulations we investigated the dynamics of hydration water in \(\alpha\)-lactalbumin solutions. We found good agreement between simulation and experiment for the number of slow waters around the protein surface. We showed that the slowly reorienting water molecules are the hydration shell waters that make on average fewer H-bonds with the surrounding molecules in comparison to bulk water, in particular those waters that accept fewer H-bonds from other water molecules or protein residues. In the native state the hydrophobic water on average exhibits slightly slower reorientation dynamics than hydrophilic water. These conclusions are corroborated by experimental evidence that the slow water fraction is characterized by weaker H-bonds. The simulations show that hydration waters donating more H-bonds to their surroundings are slower, in accordance with the finding that protein H-bond accepting (hydrophilic) groups slow down reorientation.

The H-bond breaking mechanism is also affected. Near convex sites the jump process dominates the breaking procedure, while near concave sites jumping becomes less frequent, thus slowing down the reorientation of water molecules. At concave sites the diffusive process (without jumps) becomes relatively more important.

Upon misfolding a larger amount of water reorients slowly, mainly due to the increase of the solvent accessible surface. However, the hydration shell water dynamics
slightly accelerates compared to the native hydration shell water, due to an increased fraction of bulk-like waters near hydrophobic groups. In the misfolded systems, the hydrophobic residues are more exposed to the bulk and are no longer part of a concave protein cavity. As a result, the reorientation of water molecules near hydrophobic groups becomes relatively faster compared to the folded systems. The protein-water H-bond lifetime does not change significantly upon misfolding, although waters accepting a H-bond from a protein donor reorient slower, probably due to water mediated protein interactions.
Appendix

3.A Water HB and reorientation dynamics analysis.

Table 3.A.1: Simulation: Fit of the anisotropy curve for the hydration shell water population to a sum of three exponentials (eq 3.5) for the six systems of study 1) 5% SPC/E 2) 8.5% SPC/E 3) misfolded #1 4) misfolded #2 5) 5% TIP4P/2005 6) 8.5% TIP4P/2005. The number subscript indicates the error in the last digit

<table>
<thead>
<tr>
<th>system #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>τ₀</td>
<td>0.20₁</td>
<td>0.20₁</td>
<td>0.19₁</td>
<td>0.18₁</td>
<td>0.21₁</td>
<td>0.19₁</td>
</tr>
<tr>
<td>τ₁</td>
<td>4.56₁</td>
<td>4.60₄</td>
<td>4.39₃</td>
<td>4.4₁₃</td>
<td>5.4₇₄</td>
<td>5.₃₄₄</td>
</tr>
<tr>
<td>τ₂</td>
<td>4₁0₈₅₅</td>
<td>4₀₈₅₆</td>
<td>3₁₅₁₂₅₁</td>
<td>3₄₄₄₄₆</td>
<td>4₅₃₅₆₃</td>
<td>4₀₂₀₄₆</td>
</tr>
</tbody>
</table>

Table 3.A.2: Simulation: Reorientation decay time τ of water molecules in ps, based on its category for the six systems of study 1) 5% SPC/E 2) 8.5% SPC/E 3) misfolded #1 4) misfolded #2 5) 5% TIP4P/2005 6) 8.5% TIP4P/2005. The number subscript indicates the error in the last digit

<table>
<thead>
<tr>
<th>system #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>τ_prot</td>
<td>4₀₇₄</td>
<td>4₀₇₄</td>
<td>3₈₅₉</td>
<td>3₉₂₆</td>
<td>4₉₃₆</td>
<td>5₀₀₆</td>
</tr>
<tr>
<td>τ_HF</td>
<td>4₁₀₂₁</td>
<td>4₁₀₅₁</td>
<td>3₉₈₂₀₉</td>
<td>4₀₉₁₄₁</td>
<td>4₈₀₉₁</td>
<td>4₈₉₉₁</td>
</tr>
<tr>
<td>τ_HP</td>
<td>4₁₇₉</td>
<td>4₁₀₉₄</td>
<td>3₇₅₇</td>
<td>3₇₉₃</td>
<td>₅₁₇₁₀₁</td>
<td>₅₂₂₁₆</td>
</tr>
<tr>
<td>τ_HB</td>
<td>5₀₅₉</td>
<td>₅₅₀₂₆</td>
<td>₄₆₆₁₅₁</td>
<td>₄₉₁₂₀₁</td>
<td>₆₂₃₁₁₁</td>
<td>₆₃₃₁₇</td>
</tr>
<tr>
<td>τ₅₀₆₆₆</td>
<td>3₅₃₅</td>
<td>₃₅₃₅</td>
<td>₃₄₅₄</td>
<td>₃₄₈₆</td>
<td>₄₂₂₈</td>
<td>₄₃₁₆</td>
</tr>
<tr>
<td>τ_HBW</td>
<td>₅₇₃₁₅₁</td>
<td>₅₉₈₂₆</td>
<td>₅₄₈₂₃₅</td>
<td>₅₆₉₂₆</td>
<td>₇₂₁₁₄</td>
<td>₇₃₇₈</td>
</tr>
<tr>
<td>τ_HBWsin</td>
<td>₆₂₃₂₃₅</td>
<td>₆₂₆₂₁₁</td>
<td>₅₈₈₃₈₆</td>
<td>₆₁₅₂</td>
<td>₈₂₅₁₀₁</td>
<td>₈₂₁₂₅</td>
</tr>
<tr>
<td>τ_HBdon</td>
<td>₄₂₅₁₄</td>
<td>₄₆₅₁₉</td>
<td>₄₄₂₂₃₅</td>
<td>₄₇₄₂₅</td>
<td>₅₁₁₂</td>
<td>₅₀₉₂₈</td>
</tr>
</tbody>
</table>
Figure 3.A.1: Simulation: Probability distribution of Protein water Hydrogen bond lifetimes for the four different systems of study a) 5% SPC/E (top left) b) 8.5% SPC/E (top right) c) misfolded #1 (middle left) d) misfolded #2 (middle right) e) 5% TIP4P/2005 (bottom left) f) 8.5% TIP4P/2005 (bottom right). With green is highlighted the protein water hydrogen bond lifetime distribution when the protein is accepting the hydrogen from water (t_A) and blue the protein water hydrogen bond lifetime distribution when the protein is donating the hydrogen to the water (t_D).
Figure 3.A.2: Water mediated protein interaction. Here we illustrate how a water of the type D is hydrogen mediating the contact of two protein residues.

Figure 3.A.3: Simulation: Each plot depicts the reorientation decays ($\log(P_2)$) of a single hydration shell water molecule from the simulation data (red) and the fitted curve (blue).
Table 3.A.3: Simulation: Fraction of hydration shell type of water molecules reorienting bulk-like, moderately slow or very slow.

<table>
<thead>
<tr>
<th>type</th>
<th>$\tau &lt; 4\text{ps}$</th>
<th>$4 \leq \tau &lt; 20\text{ps}$</th>
<th>$\tau \geq 20$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5% SPC/E</td>
<td>HF 0.28 0.65 0.07</td>
<td>HP 0.28 0.64 0.08</td>
<td>HB 0.17 0.69 0.14</td>
</tr>
<tr>
<td></td>
<td>sHB 0.35 0.62 0.03</td>
<td>all 0.28 0.64 0.08</td>
<td></td>
</tr>
<tr>
<td>8.5% SPC/E</td>
<td>HF 0.28 0.65 0.07</td>
<td>HP 0.30 0.62 0.08</td>
<td>HB 0.18 0.67 0.16</td>
</tr>
<tr>
<td></td>
<td>sHB 0.35 0.62 0.03</td>
<td>all 0.29 0.64 0.07</td>
<td></td>
</tr>
<tr>
<td>misfolded #1</td>
<td>HF 0.28 0.65 0.07</td>
<td>HP 0.33 0.63 0.08</td>
<td>HB 0.17 0.70 0.10</td>
</tr>
<tr>
<td></td>
<td>sHB 0.34 0.63 0.06</td>
<td>all 0.31 0.63 0.07</td>
<td></td>
</tr>
<tr>
<td>misfolded #2</td>
<td>HF 0.28 0.65 0.07</td>
<td>HP 0.32 0.62 0.06</td>
<td>HB 0.21 0.67 0.12</td>
</tr>
<tr>
<td></td>
<td>sHB 0.36 0.61 0.03</td>
<td>all 0.30 0.63 0.07</td>
<td></td>
</tr>
<tr>
<td>5% TIP4P/2005</td>
<td>HF 0.22 0.69 0.09</td>
<td>HP 0.20 0.69 0.1</td>
<td>HB 0.14 0.69 0.17</td>
</tr>
<tr>
<td></td>
<td>sHB 0.28 0.68 0.04</td>
<td>all 0.21 0.69 0.10</td>
<td></td>
</tr>
<tr>
<td>8.5% TIP4P/2005</td>
<td>HF 0.21 0.68 0.11</td>
<td>HP 0.20 0.69 0.11</td>
<td>HB 0.14 0.69 0.17</td>
</tr>
<tr>
<td></td>
<td>sHB 0.26 0.69 0.10</td>
<td>all 0.21 0.69 0.10</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.A.4: Simulation: Fraction of waters in hydration shell reorienting bulk-like, moderately slow or very slow as function number of donated and accepted H-bonds.

<table>
<thead>
<tr>
<th># donated h-bonds</th>
<th>$\tau &lt; 4\text{ps}$</th>
<th>$4 \leq \tau &lt; 20\text{ps}$</th>
<th>$\tau \geq 20$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.37 0.59 0.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.32 0.62 0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.27 0.65 0.08</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th># accepted h-bonds</th>
<th>$\tau &lt; 4\text{ps}$</th>
<th>$4 \leq \tau &lt; 20\text{ps}$</th>
<th>$\tau \geq 20$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.20 0.64 0.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.28 0.64 0.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.30 0.64 0.06</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 3.A.4: Simulation: Probability distribution of reorientation decay times of the bulk water for the six different systems of study for systems 5% SPC/E (top left), 8.5% SPC/E (top right), misfolding #1 (middle left), misfolding #2 (middle right), 5% TIP4P/2005 (bottom left), 8.5% TIP4P/2005 (bottom right).
Figure 3.A.5: Simulation: Number distributions of reorientation decay times of the hydration water for 5% SPC/E system. For the first row, the first column depicts the overall hydration water, and its subdivision into molecules hydrating hydrophilic and hydrophobic groups. The second column depicts the subdivision of hydrophilic population according to the hydrogen bond coordination. The third column depicts the subdivision of hydrophobic population according to hydrogen bond coordination. For the second row the first column depicts the overall hydration water, and subdivision into molecules that are hydrogen bonded to the protein (green) and molecules that are not (blue). The second column depicts the subdivision of h-bonded coordination according to hydrogen bond population. The third column depicts the subdivision of non-bonded population according to hydrogen bond coordination.
Figure 3.A.6: Simulation: Number distributions of reorientation decay times of the hydration water for 8.5 % SPC/E system. For the first row, the first column depicts the overall hydration water, and its subdivision into molecules hydrating hydrophilic and hydrophobic groups. The second column depicts the subdivision of hydrophilic population according to the hydrogen bond coordination. The third column depicts the subdivision of hydrophobic population according to hydrogen bond coordination. For the second row the first column depicts the overall hydration water, and subdivision into molecules that are hydrogen bonded to the protein (green) and molecules that are not (blue). The second column depicts the subdivision of h-bonded population according to hydrogen bond coordination. The third column depicts the subdivision of non-bonded population according to hydrogen bond coordination.
Figure 3.A.7: Simulation: Number distributions of reorientation decay times of the hydration water for the misfolded #1 system. For the first row, the first column depicts the overall hydration water, and its subdivision into molecules hydrating hydrophilic and hydrophobic groups. The second column depicts the subdivision of hydrophilic population according to the hydrogen bond coordination. The third column depicts the subdivision of hydrophobic population according to hydrogen bond coordination. For the second row the first column depicts the overall hydration water, and subdivision into molecules that are hydrogen bonded to the protein (green) and molecules that are not (blue). The second column depicts the subdivision of h-bonded population according to hydrogen bond coordination. The third column depicts the subdivision of non-bonded population according to hydrogen bond coordination.
Figure 3.A.8: Simulation: Number distributions of reorientation decay times of the hydration water for the misfolded #2 system. For the first row, the first column depicts the overall hydration water, and its subdivision into molecules hydrating hydrophilic and hydrophobic groups. The second column depicts the subdivision of hydrophilic population according to hydrogen bond coordination. The third column depicts the subdivision of hydrophobic population according to hydrogen bond coordination. For the second row the first column depicts the overall hydration water, and subdivision into molecules that are hydrogen bonded to the protein (green) and molecules that are not (blue). The second column depicts the subdivision of h-bonded population according to hydrogen bond coordination. The third column depicts the subdivision of non-bonded population according to hydrogen bond coordination.
Figure 3.A.9: Simulation: Number distributions of reorientation decay times of the hydration water for 5.5 % TIP4P/2005 system. For the first row, the first column depicts the overall hydration water, and its subdivision into molecules hydrating hydrophilic and hydrophobic groups. The second column depicts the subdivision of hydrophilic population according to the hydrogen bond coordination. The third column depicts the subdivision of hydrophobic population according to hydrogen bond coordination. For the second row the first column depicts the overall hydration water, and subdivision into molecules that are hydrogen bonded to the protein (green) and molecules that are not (blue). The second column depicts the subdivision of h-bonded population according to hydrogen bond coordination. The third column depicts the subdivision of non-bonded population according to hydrogen bond coordination.
Figure 3.A.10: Simulation: Number distributions of reorientation decay times of the hydration water for 8.5 % TIP4P/2005 system. For the first row, the first column depicts the overall hydration water, and its subdivision into molecules hydrating hydrophilic and hydrophobic groups. The second column depicts the subdivision of hydrophilic population according to the hydrogen bond coordination. The third column depicts the subdivision of hydrophobic population according to hydrogen bond coordination. For the second row the first column depicts the overall hydration water, and subdivision into molecules that are hydrogen bonded to the protein (green) and molecules that are not (blue). The second column depicts the subdivision of h-bonded coordination according to hydrogen bond coordination. The third column depicts the subdivision of non-bonded population according to hydrogen bond coordination.
Bibliography


Chapter 4

Correlation between water structure and dynamics in the hydration layer of a type III ocean pout anti-freeze protein

We report on a molecular dynamics study on the relation between the structure and (orientation and hydrogen bond) dynamics of hydration water around the ocean pout AFP III anti-freeze protein. We find evidence for an increasing ice-like structure from the area opposite to the ice binding site (IBS) towards the protein IBS, with the strongest ice-like structure around the THR-18 residue of the IBS. This ice-like structural signal correlates with increased reorientation decay times. Moreover, we find anti-correlation for several key residues that are not part of the IBS but are in its vicinity. These effects are enhanced at lower temperature. Finally, as AFP III anti-freeze protein is binding to ice crystal planes through a predominantly hydrophobic patch, we investigate the ice-like structure and dynamics of waters at partially dehydrated IBS. We find that upon dehydration the IBS becomes even more ice-like for the wild type, and that the water reorientation time becomes longer, but less so for the mutant T18N, which also has a higher hydration at the IBS. These results are in agreement with water-air VSFG spectroscopic experiments showing a reduced ice-like signal upon mutation at the IBS.

4.1 Introduction

The structure and dynamics of a protein’s hydration layer is crucial for its functioning and conformational dynamics. Hydration plays an active role in biological processes such as protein folding, ligand binding, and protein recognition [1–3]. Protein-
solvent interactions are especially important in antifreeze proteins (AFP), as they have to specifically recognize and bind — through their Ice Binding Site (IBS) — nucleating ice crystals in the excess of liquid water and prevent further growth of ice [4]. In spite of this seemingly tough goal AFPs show a large structural diversity in nature and are seen in many organisms such as fungi, bacteria, fish, insects, where their major function is to help these organisms survive at subzero temperatures [5]. In addition to their biological role, AFPs also have a variety of applications as ice avoiding agents, from organ preservation [6] to texture enhancers in food [7]. Many AFPs are characterized by an IBS consisting of regular β-sheets or spirals, commensurate with the ice-crystal planes. A counterexample is the ocean pout AFP III that does not exhibit a regular β-sheet ice binding site, raising the question of how the binding to ice is achieved instead.

Both simulations [8–10] and experiments [11–13] indicate that water is more structured at the IBS with respect to other non IBS sites of the AFP III protein. The type of interactions necessary for the affinity and specificity of the IBS to ice crystals have been addressed by mutation studies, X-ray crystallography [13, 14] and simulations [8, 9, 15]. There is an increasing consensus on the combined role of both polar groups (able to form hydrogen bonds) and apolar groups in ice binding by matching to ice-lattice oxygens as well as by hydrophobic ice-IBS interactions. AFP III’s ice-binding-site is relatively flat and hydrophobic, thus serving as a great candidate for surface sensitive interface experiments. In particular, VSFG studies [16] have shown that a single point mutation of a core IBS amino acid, from THR-18 into ASN-18, causes a loss in anti-freeze activity of the ocean pout AFP. The effect of this mutation on the water structure and dynamics is not known in full detail. Here, we employ all atom molecular dynamics (MD) simulations to obtain microscopic insight in the tetrahedral structure and reorientation dynamics of the protein hydration layer for the wild type and the T18N mutant. As these experiments were surface sensitive, we mimic the exposure of the protein to the water-air surface by studying a partially dehydrated IBS. For a β-helical anti-freeze protein, Nutt el al [17] pointed out that not only the IBS is important for ice recognition but also the non-IBS residues, which distort the ice-like water structure. Therefore, we also investigate here the correlation between reorientation and structure in different parts of the protein, and interpret this in terms of tetrahedral structure-making and -breaking ability of the residues.

The remainder of the paper is organized as follows. The next section describes the simulation setup, and analysis methodology. Subsequently, we present the analysis of the results, and discuss these in the light of the VSFG (and other) experiments. We end with conclusions.
4.2 Methods

4.2.1 Simulation setup

All molecular dynamics simulations were performed with the Gromacs 4.5.4 package [18]. The ocean pout wild type monomer structure was obtained from PDB 1HG7 [19] and the mutated ocean pout structure (THR-18 to ASN-18) was obtained from PDB 1JAB [20]. Each structure was solvated with (SPC/E or TIP4P/2005 molecules, resulting in solutions of approximately 5% w/w. The choice of two different water models serves, not so much as a comparative study of the two models, but as a consistency test. It is known that SPC/E water model is known to better reproduce the dielectric constant of water [21] than the TIP4P/2005. Therefore, in the main text, we report the results for the SPC/E water model unless stated otherwise. The TIP4P/2005 water model is being reported in the Appendix.

The atomic interactions were defined by the amber99sb-ildn force field [22]. The protonation state of the amino acids corresponds to pH between 4.25 and 10.53, since glutamic acid side chain is deprotonated (pKa=4.25) and the Lysine side chain is protonated (pKa=10.53). No ions needed to be added to the solution after assigning hydrogens to the pdb structures since the total charge of both structures was zero. After energy minimization, both water force-fields systems were equilibrated for 1 ns at a constant pressure of 1 atm and at several temperatures. These temperature were 298 K, 285 K, 270 K, 255 K, 225 K and 1 atm for the SPC/E systems and 298 K, 285 K, 270 K, 255 K for the TIP4P/2005 systems. All bonds were constrained with the Lincs algorithm. A cutoff of 1 nm was used for the non-bonded Lennard-Jones interactions. The Particle Mesh Ewald method was used to calculate the electrostatic interactions with a Fourier spacing of 0.12 nm and a 1 nm cutoff for the short range electrostatic interactions. Neighbour lists were updated every 10 fs with a cutoff of 1 nm and the time step was 2 fs [22]. The leap-frog algorithm was used for integrating Newton’s equations of motion. In the NPT simulations the v-rescale thermostat[23] with a coupling time constant of 0.2 ps controlled the temperature, while the Parrinello-Rahman barostat[24] with a coupling time constant of 1.0 ps kept the pressure constant.

For each temperature 10 frames corresponding to the average energy and volume were selected from a 100 ns long MD trajectory in the NPT ensemble. From each frame a 1 ns NVE run was conducted. This approach eliminated any unwanted influence from the thermostat or barostat on the dynamics. To prevent energy drift we used a switching function for the non-bonded interactions from 0.8-1.0 nm. The pair lists were updated every 5 fs with a cutoff of 1.2 nm and the time step was 1 fs. The frequency of the energy calculation was 10 fs, and the frames were saved every 100 fs in order to obtain sufficient data for the analysis.
In this paper the above systems of wild and mutant type water solutions for different temperatures will be denoted as fully hydrated systems. The effect of water dehydration on the water structure and dynamics was investigated by analyzing partially dehydrated systems. For each combination of temperature, water force field, and WT/mutant, we took the last frame of the 100 ns NPT MD trajectory and reduced the water content by about a factor of 10, leaving only 722 protein surface water molecules. We ran another 100 ns NVT at the same temperatures as the fully hydrated systems. From each of these NVT trajectories, 10 frames corresponding to the average energy and volume were selected, from which a 1 ns NVE run was conducted. The other parameters for the NVT and NVE of the dehydrated systems were identical to those of the fully hydrated systems.

4.2.2 Analysis

Hydrogen bond lifetimes

Throughout this paper hydrogen bonds will be referred to as H-bonds. An H-bond is defined present if the distance between H-bond donor \(D\) and acceptor atom \(A\), \(R_{DA} \leq 0.35\) nm, and the angle between the \(DH\) and \(HA\) vectors \(\theta_{ADH} \leq 30^\circ\) [25, 26]. Here, the donor or acceptor heavy atom can belong to either the protein or to water. The H-bond lifetime was defined as the time it takes for the bond to break and to form a stable bond with another acceptor (exchange), or to remain dangling (translational diffusion). If the bond reforms within 200 fs, it is considered not broken, in order to avoid counting fast recrossings events.

The recrossing time \(\Delta \tau = 200\) fs was based on the fact that the vibrational characteristics of the water water H-bond occur in the far infrared and around \(200\) cm\(^{-1}\)=166 fs [27, 28].

Water reorientation dynamics

Reorientation dynamics can be represented using the time correlation function [29, 30]

\[
C_2(t) = \frac{2}{5} \langle P_2[u(0) \cdot u(t)] \rangle
\]

(4.1)

where \(P_2\) denotes the second Legendre polynomial, \(u\) is the unit vector characterizing the orientation of an OH group in a given frame and the angular brackets denote an ensemble average over all water molecules and all time origins. This correlation function can easily be related to anisotropy curves obtained from polarization-resolved femtosecond infrared spectroscopy and to orientation relaxation times from magnetic relaxation techniques [29].
From MD simulations we compute the reorientation dynamics for each individual molecule. We are particularly interested in the anisotropy decay of water molecules hydrating different protein sites. The reorientation dynamics of individual water molecules was investigated by following the dynamics of both OH bond vectors of each molecule \( j \). For each bond unit vector \( \mathbf{u}_{jk}(t) \), we computed the time correlation function

\[
c_m^j(t) = \frac{1}{5\ell} \sum_{t' = t_0}^{t_0 + \ell \Delta t} \sum_{k=1}^{2} P_2[\mathbf{u}_{jk}(t') \cdot \mathbf{u}_{jk}(t' + t)]
\]

(4.2)

where the sum over \( k = 1, 2 \) refers to the two OH bond vectors, and the time correlation is summed over the \( m \)th time interval \( t_0^m < t' < t_0^m + \ell \Delta t \) that the molecule is within the hydration layer of the protein (defined as the water oxygen being within 4.4 Å from the protein heavy atoms). We used a buffer time of 2 ps to avoid counting fast non-essential excursions inside and outside the hydration layer [31]. For each curve \( c_m^j(t) \) we extracted the reorientation decay time \( \tau_{m}^j \) by fitting \( c_m^j(t) \) to a single exponential fit in the interval 0 < \( t \) < 10 ps. Each estimate \( m \) was viewed as a separate measurement, since the water molecule can change its dynamics when leaving or entering the hydration layer. For each water molecule \( j \) and for each \( m \)th interval of length \( \ell \) the decay time \( \tau_{m}^j \) was histogrammed with a weight \( \ell \), leading to a probability distribution of decay times. The weight \( \ell \) follows from the fact that the correlation in eq. 4.2 occurs \( \ell \) times in eq. 4.1. The average decay time \( \tau \) from these histograms thus should be close to the overall decay time of eq. 4.1. The decay times of the individual water molecules allow us to establish the relation between the water structure and dynamics. We therefore divide the waters into several categories.

For each water molecule \( j \) and for each \( m \)th interval \( \ell \) the decay time \( \tau_{m}^j \) was histogrammed with a weight \( \ell f_{\#AA} \) in the residue \#AA reorientation distribution. Here \( f_{\#AA} \) refers to the fraction of frames in the \( m \)th interval \( \ell \) in which the oxygen of water \( j \) is within 4.4 Å of any of the heavy atoms of residue \#AA.

**Water structure and dynamics at different areas of the protein.**

Based on mutation studies by Graether et al. [20] we define six different parts of the protein surface, the THR-18 and ASN-18, the IBS, the vicinity and the opposite face, the non-IBS residues (everything but the IBS), and the entire hydration shell which consists of the IBS and non-IBS (see for an illustration Fig. 4.2.1a,b). The vicinity and opposite face residues were selected to include key surface residues identified in Ref. 20. They include vicinity residues 39,42,46,47,61 and opposite residue 29. Figure Fig. 4.2.1a,b shows which residues of the protein surface belong to the IBS, the vicinity and the opposite group. Rendered in cyan are the IBS residues 9,10,12,
Figure 4.2.1: a), b): Partitioning of the protein into three distinct areas. The area IBS (denoted by cyan), vicinity (denoted by magenta) and opposite (denoted by green). THR-18 is denoted by black. c) $\theta$ angle

13,14,15,16,18,19,20,21,44. Depicted in magenta are the vicinity residues within 0.4 nm of the IBS (8, 23, 37, 39, 42, 43, 46, 47, 48, 50, 51, 61). The green opposite site residues are 1.5 nm away of the IBS and include residues 0, 1, 2, 26, 27, 28, 29, 30, 56.

**Water-water angular distribution function**

Sharp and coworkers [9, 32, 33] introduced a method to identify the tetrahedral structuring of water around amino acids based on the distribution of a water-water angle. In a given frame one computes for each amino acid the minimum water-water OOH angle $\theta$ (see Fig. 4.2.1c for a graphical definition) for all water-water pairs within 3.5 nm from each other and solvating that amino acid. The distribution $P(\theta)$ of these angles shows a bimodal distribution with a minimum at 30° (see Fig. 4.3.3a), distinguishing between tetrahedral water population (angles lower than 30°) and a perturbed H-bond network, mostly occurring around hydrophilic groups (angles higher than 30°). The tetrahedral structure parameter $S$ is now defined as the integral of $P(\theta)$ up to $\theta = 30°$.

It turns out that water around hydrophobic groups have a larger $S$ due to smaller H-bond angles $\theta$, inducing stronger water-water bonds, with bigger energy fluctuations and therefore a positive heat capacity of the solvating water. Reversely, the introduction of a hydrophilic group around water strains the water-water H-bond angle and shifts the angle distribution to higher values, and hence a lower $S$, thus decreasing the water-water bond energy and fluctuations which decreases the heat capacity of solvation [9, 32, 33].

Throughout the text we will associate tetrahedral/ice-like/structured water with a large $S$ value (high tetrahedral water population) and unstructured water with a low $S$ value (low tetrahedral water population). Unstructured water coinciding with slow reorientation dynamics will be labeled as glass-like water. Residues solvated by ice-
like tetrahedral water will be labeled as tetrahedral structure makers and the ones by glass-like waters, tetrahedral structure breakers.

**Structural parameter - water reorientation time correlation**

As mentioned above, each water $i$ surrounding one amino acid for an $m$th interval of length $\ell$, has a reorientation decay time $\tau_{i,m}$ associated with it. Additionally in this $m$th interval we calculated the distribution $P(\theta_{ij})$ of the angle $\theta_{ij}$ of waters $j$ within 3.5 Å of $i$ and hydrating the same amino acid. From the normalized distribution $P(\theta_i) = \sum_j P(\theta_{ij}) / \sum_{i,j} P(\theta_{ij})$ we obtain the structural parameter $S_{i,m}$ for each water $i$ hydrating a particular amino acid, by integrating from 0° to 30°. We bin the pair $\tau_{i,m}$ and $S_{i,m}$ for each residue in a 2D histogram, in order to investigate possible correlation between water tetrahedral structuring and reorientation dynamics. The $S - \tau$ correlation is quantified using Pearson’s product-moment correlation coefficient $R$.

### 4.3 Results and Discussion

![Figure 4.3.1](image)

Figure 4.3.1: Anisotropy decay of a) all water and b) the hydration shell water of the wild type and mutant SPC/E systems respectively.

#### 4.3.1 Water reorientation dynamics slows down at surface

From the MD simulations trajectories we calculated the anisotropy correlation function $C_2(t)$ using eq. (4.1) for all water molecules in the system and for the hydration shell water molecules only, for both wild type and mutant at different temperatures. Fig. 4.3.1 shows these correlation functions for the SPC/E model. As expected the
Figure 4.3.2: a) Water reorientation time, b) protein water H-bond lifetimes of hydration water for SPC/E systems of a wild type (WT), and mutant (T18N) at room temperature. Color coding as in Fig. 4.2.1a,b

anisotropy for all waters decays within a few ps for ambient conditions, but decays slower with decreasing temperature. This agrees with the work of Stirnemann et al [34], who found that the decay time becomes longer upon cooling. For the hydration water this effect is even more striking.

4.3.2 Water reorientation dynamics differs locally

Strikingly, the difference in overall hydration water reorientation dynamics between mutant and wild type is small (see Fig. 4.3.1). This indicates that the effect of the mutation on the water reorientation dynamics is either small and/or local, since the number of waters hydrating THR-18 and ASN-18 is small compared to the total amount of waters in the hydration shell, washing out any local effect. To study whether THR-18 and ASN-18 differ significantly in local water dynamics we therefore categorized the hydration water by their local environment (see Methods section 4.2.2). In Figure 4.3.2a, we plot the water orientation decay time for the six different parts of the protein. Xu et. al [35] discovered a small acceleration of the H-bond dynamics upon mutating ALA-16 to HIS-16. Here we find that water reorients significantly slower around THR-18 than around other regions, and also slower than at the mutated site ASN-18. This effect is enhanced at lower temperatures (see Fig. 4.A.1). However, the point mutation does not drastically accelerate the reorientation dynamics of the IBS, but only has a local effect. This local acceleration of water reorientation dynamics around ASN-18 coincides with the less stable water-ASN-18 hydrogen bonds, compared to the water-THR-18 ones (see Fig. 4.3.2b). This effect is again enhanced at lower temperatures (see Fig. 4.A.2).

X-ray crystallography mutation studies by Graether et al [20] and Neutron scatter-
Figure 4.3.3: a) OOH angle distribution of water pairs solvating the wild fully hydrated system IBS (cyan), THR-18 of the wild fully hydrated system (black), ASN-18 of the mutant fully hydrated system (yellow), bulk water (brown). b) Structure order parameter of water (S), for the different parts of the protein. All the above systems are in room temperature and for the SPC/E water systems. Color coding as in Fig. 4.2.1a,b

ing experiments by Howard et al [13], have indicated the central role of the THR-18 side chain OH in matching the ice plane and therefore in the anti-freeze activity of the molecule. Threonine’s side chain OH group actively participates through a H-bond with water. These findings are in agreement with our results which show that upon substitution of the THR-18 side chain to an ASN-18, the hydrogen bond and reorientation around ASN-18, becomes faster.

4.3.3 Hydration water structure

Following Sharp and Gallagher [9, 33] and other authors [10], we characterize the water structure by computing the OOH angle distribution for several parts of the protein (see Fig. 4.3.3a), from which we obtain the average water structural parameter S. Figure 4.3.3b (see also Fig. 4.A.3), shows the average water structural parameter S for the overall hydration shell, the IBS, the vicinity and the opposite region, and residue 18, for the WT and T18N systems. Interestingly, there seems to be a monotonic increase of water structure along the protein, i.e. the water structural parameters obey the relation $S_{opp} < S_{vicinity} < S_{IBS}$. This effect is slightly more pronounced at the wild type system, compared to the mutant system. Both the absolute values of S and the difference $S_{IBS} - S_{opposite}$ increases slightly upon cooling (see Fig. 4.A.3). This behaviour holds for both force fields.
Figure 4.3.4: Structural parameter S as a function of water reorientation decay time for selected residues of a) IBS (WT), b) IBS (T18N), c) THR-18 (WT), d) ASN-18 (T18N), e) ARG-39 (WT/vicinity), f) ALA-48 (WT/vicinity) at 298 K.

Water is more structured around the IBS and in particular residue 18 of the wild type, compared to the mutant, especially at low temperatures (see Fig. 4.A.3). Waters around residue 18, which reorient slower for the wild type compared to the mutant, have also a more ice-like structure for the wild type compared to the mutant. These findings partially agree with the VSFG experiments [16], which showed a drastic increase in ice like water signal for the wild type compared to the mutant system. While we observe a reduction of ice structure around THR-18 upon mutation, the reduction in the IBS ice-like water going from the wild type to the mutant is not as great as experiments suggest.

4.3.4 Structure - reorientation time correlation

We investigate the correlation between the water reorientation dynamics and tetrahedral structure by histogramming the hydrating waters as function of their $S_i$ and $\tau_i$ in which we weigh each entry with a weight $\ell$ (see Methods section). Figures 4.3.4 and 4.3.5
Figure 4.3.5: Structural parameter S as a function of water reorientation decay time for selected residues of a) IBS (WT), b) IBS (T18N), c) THR-18 (WT), d) ASN-18 (T18N), e) ARG-39 (WT/vicinity), f) ALA-48 (WT/vicinity) at 275 K.

show the 2D histograms of $S_i$ and $\tau_i$ of the IBS and of selected amino acids for, respectively, 298 K and 275 K for the wild type and mutant SPC/E system. At room temperature (Fig. 4.3.4a) the IBS of the wild type shows clearly a positive correlation between tetrahedral structure $S$ and water reorientation time ($\tau$). While this correlation is not perfect, a more structured water is more likely to exhibit a longer reorientation time. Upon mutation (Fig. 4.3.4b) the IBS $S - \tau$ distribution shifts to slightly lower $S$ and $\tau$, and the correlation becomes less pronounced. More strikingly, water hydrating the THR-18 (WT system) shows a higher values for $S$ and $\tau$ as well as a stronger $S - \tau$ positive correlation compared to ASN-18 (T18N system). (see Figs. 4.3.4c and 4.3.4d respectively). Clearly, more ice-like waters also reorient much slower. These positive correlations can be contrasted with the behavior of other key residues in the vicinity region of the IBS. For instance, the charged and hydrogen bond donating residue ARG-39 exhibits no or even a negative correlation (R=-0.55 at 275 K) between structure and dynamics (with a lower peak at S compared to THR-18). Here, a less structured water shows a longer reorientation time. Interestingly, water around ALA-48 is able to exhibit
Table 4.3.1: Pearson’s correlation coefficient between S and $\tau$ for water hydrating different IBS amino acids of WT and T18N at 275 K.

<table>
<thead>
<tr>
<th>#AA</th>
<th>WT</th>
<th>T18N</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLN 9</td>
<td>0.05</td>
<td>-0.38</td>
</tr>
<tr>
<td>LEU 10</td>
<td>0.39</td>
<td>-0.03</td>
</tr>
<tr>
<td>PRO 12</td>
<td>0.45</td>
<td>0.20</td>
</tr>
<tr>
<td>ILE 13</td>
<td>0.34</td>
<td>0.21</td>
</tr>
<tr>
<td>ASN 14</td>
<td>0.08</td>
<td>-0.05</td>
</tr>
<tr>
<td>THR 15</td>
<td>0.51</td>
<td>0.18</td>
</tr>
<tr>
<td>ALA 16</td>
<td>0.55</td>
<td>-0.18</td>
</tr>
<tr>
<td>THR/ASN 18</td>
<td>0.56</td>
<td>0.22</td>
</tr>
<tr>
<td>LEU 19</td>
<td>0.20</td>
<td>0.15</td>
</tr>
<tr>
<td>VAL 20</td>
<td>0.28</td>
<td>0.24</td>
</tr>
<tr>
<td>MET 21</td>
<td>0.46</td>
<td>-0.08</td>
</tr>
<tr>
<td>GLN 44</td>
<td>0.30</td>
<td>0.13</td>
</tr>
</tbody>
</table>

both ice like and glass-like structure. ALA-48 is next to ARG-47, which is a charged amino acid capable of donating hydrogen bonds.

Upon cooling the systems to 275 K, the WT (S-$\tau$) correlation at the IBS becomes stronger ($R=0.45$) compared to room temperature ($R=0.19$) (Fig. 4.3.5a). Since the S-$\tau$ correlation signals at 275 K are enhanced, we can more clearly distinguish between strong tetrahedral structure making and subtle tetrahedral structure making residues at the WT-IBS. Interestingly enough, as shown in Fig. 4.A.4 for the WT-IBS residues, there exist amino acids that show a very strong positive S-$\tau$ correlation (LEU-10, PRO-12, ILE-13, ALA-16, THR-15, THR-18, MET-21), as well as amino acids that show a subtle S-$\tau$ correlation (GLN-9, ASN-14, LEU-19, VAL-20 and GLN-44). At 275 K, the point mutation results in a weaker S-$\tau$ correlation for water around the T18N-IBS ($R=0.14$) (see Fig. 4.3.5b). As shown in Fig. 4.3.5c,d, Fig. 4.A.4 and Fig. 4.A.5 and summarized in Tab. 4.3.1, the origin of the decreased (S-$\tau$) correlation in the T18N IBS compared to the WT IBS, is predominantly attributed to the decreased (S-$\tau$) positive correlation of GLN-9, LEU-10, PRO-12, THR-15, ALA-16, ASN-18, MET-21. Again, as shown in Fig. 4.3.5e, the charged and hydrogen bond donating residues ARG-39 of the vicinity exhibit negative correlation (with a peak at lower S compared to THR-18), and water around ALA-48 exhibit both ice like and glass-like structure (Fig. 4.3.5e).

We can conclude that vicinity residues such as ARG-39 and ALA-48 reduce the structure of water, while their charges induce stronger protein-water hydrogen bonds, which slow down the water (reorientation) dynamics. Such residues can be labeled.
water 'tetrahedral structure breakers’. It follows that there exist "tetrahedral structure making" (hydrophobic) and “tetrahedral structure breaking” residues (charged/polar, H-bond makers) which influence the structure and dynamics of water near the IBS and at the vicinity of the IBS. The tetrahedral structure making residues should have a stronger affinity for ice, while the tetrahedral structure breaking residues would have a lower affinity to ice crystal. The point mutation reduces the $S_\tau$ correlation at the IBS and therefore the affinity of the IBS for ice. We speculate that water at the vicinity around the IBS should exhibit tetrahedral structure breaking behaviour in order for the Ice to bind only to the IBS and not to the surrounding parts, thus prohibiting crystal growth around the entire protein.

4.3.5 Effect of dehydration

The VSFG experiments that inspired this investigation probe the air water surface, where the symmetry of the system is broken. It is reasonable to assume that the T18N point mutation is more soluble than the wild type, due to its lower free energy of solvation [36] (10 kT lower), compared to THR-18. According to the hydrophobic character of residues [37], THR is a more hydrophobic residue, than ASN, and therefore the IBS surface of the WT will most likely be (part of) the more hydrophobic side of the protein, and therefore more likely to be adsorbed at the air-water interface. Hence, it is conceivable that the wild type AFP proteins have a different preferred orientation at the air water surface than the mutant, resulting in less hydration of the IBS for the wild
Figure 4.3.7: a) Water reorientation time, b) protein water H-bond lifetimes, and c) structural parameter $S$ of hydration water for SPC/E systems of a fully hydrated wild (wtf), mutant (mtf) and for dehydrated wild type (wtd) and mutant (mtd) at room temperature.

Type compared to the mutant IBS. To understand the effect of preferential adsorption at the air-water interface, we need to investigate the protein-water dynamics at a reduced hydration degree. We therefore also performed MD simulations at a lower hydration degree, by reducing the water content to 10% of the fully hydrated system. The remaining 722 waters initially hydrate the entire protein surface, but quickly lead to exposure of some parts to the gas phase due to the hydrophobicity of the surface. In particular the hydrophobic patch of the protein surface near the IBS became exposed (see Fig. 4.3.6 for snapshot). It is thus likely that this part of the surface becomes exposed at the air water interface. Nevertheless, the IBS is not completely water free, and we can analyze our system in the same way as for the fully hydrated system. Therefore, we measure again the structural parameter $S$, the reorientation time $\tau$ and the hydrogen bond life time $\tau_{HB}$, and compare them in Fig. 4.3.7). Clearly, the dehydrated WT shows a strong increase both in water structure and reorientation time at the THR-18, although not in $\tau_{HB}$. 
Comparing the 100 ns NVT simulations of the SPC/E dehydrated wild type and mutant systems at room temperature, on average there are 44 water molecules within 4.4 Å of the wild type IBS, compared to 50.41 for the mutant IBS. There are on average 4.6 water molecules present around THR 18 and on average 5.3 water molecules around ASN-18 of the mutant, indeed suggesting an increase of solvation upon mutation. Therefore, to assess whether differences in hydration contribute to the strong ice-like signal in VSFG experiments, one needs to compare the dehydrated wild type IBS, THR-18 water structure to the fully hydrated mutant IBS, ASN-18 water structure. Indeed, the dehydrated wild type $S_{IBS,18}$ is significantly higher than the relevant structural parameter $S$ of the fully hydrated mutant (see Fig. 4.3.7c).

### 4.4 Conclusions

Antifreeze proteins (AFPs) are believed to prevent ice formation by binding to specific ice crystal planes and blocking their growth. Their ice-binding recognition sites have remarkable solvation properties. Spectroscopic experiments[16] and analysis of crystal structures [20] of the ocean pout AFP III anti-freeze protein indicated that upon mutating THR 18 to ASN-18 water locally to the mutation reduces its tetrahedral structure dramatically, suggesting that THR18 plays a crucial role in the ice binding site, and is able to structure water, leading to an increased AFP III affinity to ice. The structure of a protein’s hydration layer is tightly related to its dynamics, but this relation has not been elucidated for AFPS. To investigate this relation we performed molecular dynamics simulations of AFP III. We investigate both dynamics and structure of the ice-binding surface and compared these to the non-ice-binding surfaces. We find that while the hydrogen bond dynamics is not remarkably altered, the water reorientation relaxation around THR18 is slower than that of the non ice-binding surface. This local slow reorientation relaxation is correlated with longer lived H-bonds between THR-18 and its hydrating water compared to ASN-18 in the mutant. Moreover, by comparing structural signatures we found, in agreement with experiments[16] and predictions [20], that upon replacing THR18 with ASN-18, water around the point mutation reduces its tetrahedral structure. Indeed, the THR18 water tetrahedral structure, exhibits a positive correlation with reorientation relaxation time, whereas this positive correlation is significantly reduced for the mutant as is for the entire IBS of the mutant system. This reduction of the $S-\tau$ correlation at the IBS of the mutant, enhanced at lower temperatures, could explain the smaller affinity of the mutant IBS to ice. In contrast, we find that ARG-39, a positively charged H-bond donating residue in the vicinity of the IBS, shows a negative correlation of water tetrahedral structure and reorientation while some others, can show both a positive and negative correlation (ALA-48). Overall the wa-
ter tetrahedral structure seems organized in a gradient fashion towards the IBS. In the VSFG experiments the proteins adsorbed at the surface contribute most to the signal. As these proteins experience a lower degree of hydration we also investigate the effect of hydration. We find that dehydration leads to a significant increase in the tetrahedral nature of water, as well as its reorientation dynamics.

Our findings show that amino acids can act as tetrahedral water structure makers and breakers. This may explain the mechanism of pout AFP to prevent engulfment within ice, since the charged residues such as ARG break the ice structure to a more glassy one around the IBS thereby preventing ice from covering the whole protein, which would make it inactive.
Appendix

4.A Water structure and dynamics for different temperatures and forcefields.

Figure 4.A.1: Reorientation time of hydration water for all fully hydrated systems of study, a) SPC/E and b) TIP4P/2005 water forcefield, wild (left) and mutant (right)
Figure 4.A.2: Hydrogen bond lifetimes of hydration water for all fully hydrated systems of study, a) SPC/E and b) TIP4P/2005 water forcefield, wild (left) and mutant (right)
Figure 4.A.3: Structural parameter $S$ of hydration water for all fully hydrated systems of study, a) SPC/E and b) TIP4P/2005 water forcefield, wild (left) and mutant (right).
Figure 4.A.4: Reorientation decay time of versus structural parameter S for selected residues of WT at 275 K.
Figure 4.A.5: Reorientation decay time of versus structural parameter S for selected residues of T18N at 275 K.
Bibliography


[18] S. Pronk, S. Päll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov,


Chapter 5

Stability and growth mechanism of self-assembling anti-freeze cyclic peptides

Cyclic peptides (CPs) that self-assemble in ice-binding nanotubes are great candidates for use as anti-freeze proteins. Based on cyclic peptide sequence, cyclo-[(L-LYS-D-ALA-L-LEU-D-ALA)]$_2$, which can stack into nanotubes, we propose an anti-freeze cyclic peptide (AFCP) sequence, cyclo-[(L-LYS-D-ALA)$_2$-(L-THR-D-ALA)$_2$] which contains THR-ALA-THR ice binding motifs. Using molecular dynamics simulations we investigate the stability of cyclic peptides and their growth mechanism. We find that dimers of the AFCP sequence dissociate more frequently and are less stable than dimers of the original CP sequence, while nanotubes consisting of more than two peptides are stable. This sudden increase in stability of nanotubes of the AFCP sequence may be explained by the formation of H-bonds between Threonine side-chains. The Threonine distances in the ice-binding motifs are similar to those in the ant-freeze protein of Christoneura fumiferana, suggesting good ice lattice matching, and a potential for depression of the freezing point. In addition, we investigated the nanotube growth process, i.e. the association/dissociation of a single CP to an existing AFCP nanotube, by Transition Path Sampling. We found a general dock-lock mechanism, in which a single CP first docks loosely before locking into place. Moreover, we identified several qualitatively different mechanisms for dissociation, involving different meta-stable intermediates, including a state in which the peptide was misfolded inside the hydrophobic core of the tube. We also find evidence for a mechanism involving non-specific association followed by 1D diffusion. Under most conditions, this will be the dominant pathway. The results yield insight in the mechanisms of peptide assembly, and might lead to improved design of self-assembling anti-freeze proteins.
5.1 Introduction

Anti-freeze proteins (AFPs) adsorb to the surface of ice crystals and prevent growth [1]. This lowers the freezing point temperature of the solution below its melting point, enabling the survival of many organisms living in subzero temperature environments. AFPs have been used in a variety of applications, such as in super-cooling organ preservation [2] by inhibiting ice-crystal growth and therefore preventing tissue damage, as well as in industry as texture enhancing agents [3].

Although all AFPs share anti-freeze activity, they show a significant diversity in this activity, in molecular weight, amino-acid sequence and structure. Many hyperactive AFPs have a β-helical rich structure with repetitive sequence motifs [4–6]. For example the spruce budworm AFP shows a β-sheet region where threonines (THR) are organized in a regular array of THR-Xaa-THR (TXT) motifs, where Xaa can be any amino-acid. Those motifs match both the basal and prism crystal plane of ice and therefore can bind to these planes [1]. In a molecular simulation study of the spruce budworm AFP (isoform CfAFP-501), Zhou et al.[7], showed that the THR O-O distances in the TXT motif of the spruce budworm anti-freeze protein as well as the THR O-O distances of neighbouring coils are similar to the ice prism plane O-O distances.

Hyperactive anti-freeze proteins that have a β-helical structure and a rich TXT motif can be mimicked using synthetic polypeptides. For instance, one proposed option for such mimicking is a nanotube made of cyclic peptides (CP) containing the TXT ice binding motif, as shown in Fig. 5.2.1 top left. Experiments [8, 9] and simulations [10, 11] indicate that cyclic peptides of alternating L and D amino acids self assemble into nanotubes under proper conditions. These nanotubes consist of an anti-parallel cyclic β-sheet hollow structure, stabilized by backbone hydrogen bonds between adjacent CPs and by side chain interactions [11] (see Fig. 5.2.1bottom). Vijayraj et al. [10, 11] investigated the number of required CPs for a stable nanotube, and showed the importance of alanine (ALA) amino acids in the stability and fluctuations of the nanotube. Introducing an ice binding motif in such cyclic peptides would lead to self-assembled nanotubes with great potential as an anti-freeze agent. Here, we employ molecular simulations to investigate the stability and fluctuations of anti-freeze cyclic peptide nanotubes (denoted AFCP nanotubes) comprising stacks of AFCP sequence cyclo-[(L-LYS-D-ALA)2-(L-THR-D-ALA)2] and of the experimentally self assembled nanotube [9] comprising the original CP sequence cyclo-[(L-LYS-D-ALA-L-LEU-D-ALA)2], (denoted original CP nanotube). After we have established that both the AFCP and the original CP nanotube are stable in solution, we continue with the question of the formation mechanism. The formation mechanism of the AFCP nanotube from dilute solution is poorly understood, but is believed to occur via association
of CPs, nucleation and growth. In the latter stage, growth is dominated by association (and dissociation) events to the end of a growing nanotube. This step can be seen as a rare event, and needs to be addressed with rare event simulation techniques. Here we employ the transition path sampling (TPS) technique, which allows us to harvest an ensemble of unbiased rare transition paths that give valuable information about the association process. Analyzing this ensemble gives insight in the different mechanisms. While we find several mechanisms of association/dissociation, all involve an intermediate docked state. Furthermore, we find evidence for a growth mode involving nonspecific binding to the nanotube, followed by a random walk of the CP along the nanotube until it finds one of the endpoints. This mechanism might have general implications for growing fibril structures in general.

The chapter is organized as follows. In section 5.2 we describe the used methods. In section 5.3 present and discuss the results. We end with concluding remarks.

5.2 Methods

5.2.1 System setup

Construction of the cyclic peptides and nanotubes

We construct two initial CP conformations, one for the original CP cyclo-[(L-LYS-D-ALA-L-LEU-D-ALA)\(_2\)], and the other for the AFCP, cyclo-[(L-LYS-D-ALA)\(_2\)-(L-THR-D-ALA)\(_2\)]. An initial linear structure for constructing these cyclic peptides was created using AmberTools [12]. The linear peptide was turned into a cyclic peptide by constructing a bond between the first and last residue and minimizing the structure using steepest descent in GROMACS[13]. The D-conformational orientation of the D-Alanine amino acids was obtained by flipping the L-Alanine amino acids in PyMOL [14] in order to obtain the original and AFCP sequences (see also Table 5.A.3 and Table 5.A.4). The geometry of each single CP unit was energy minimized using steepest descent while the Ramachandran dihedral angles (\(\Phi, \Psi\)) of all amino acids were restrained to their average anti-parallel \(\beta\)-sheet values as given in Table 5.2.1, in order to guarantee the planarity of the ring. Since the side chains of the cyclic peptides play a crucial role in the stability of the nanotubes[11], also the side chains of the rings were relaxed in a 10 ns NPT simulations.

The final geometry of the CPs was used to build various nanotubes of different size (e.g CPNT2, CPNT3, for stacks of two or three CP units). The CPs were stacked in an anti-parallel fashion, meaning that two adjacent CPs had opposite chain orientations. During the model building, necessary care was taken to align amino acids (L and D)
in the adjacent chains as shown in Fig 5.2.1 bottom and detailed Table 5.A.3 and Table 5.A.4 of the appendix).

### 5.2.2 Molecular dynamics

All energy minimization and molecular dynamics simulations were performed with the Gromacs 4.5.4 package[13]. Molecular dynamics simulations by Vijayaraj and Khurana [10, 15], suggest that using the amberf99sb [16] stabilizes the cyclic peptide nanotubes by formation of intramolecular hydrogen bonds. Therefore we choose here the same protein force-field, coupled with TIP3P water [17]. The protonation state of the amino acids corresponds to pH greater than 11 in order to compare to the experiments. All bonds were constrained with the Lincs algorithm. A cutoff of 0.8 nm was used for the non-bonded Lennard-Jones interactions. The Particle Mesh Ewald method was used to calculate the electrostatic interactions with a Fourier spacing of 0.12 nm and a 0.8 nm cutoff for the short range electrostatic interactions. Neighbor
Table 5.2.1: Average values for the dihedral angles $\Phi$, $\Psi$ and their estimated standard deviations $\sigma_{\Phi, \Psi}$ for anti-parallel $\beta$-strands [20].

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>$\Phi_{\text{aver}}$</th>
<th>$\sigma_{\Phi}$</th>
<th>$\Psi_{\text{aver}}$</th>
<th>$\sigma_{\Psi}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>L-ALA</td>
<td>-130.2</td>
<td>21.4</td>
<td>143.8</td>
<td>14.6</td>
</tr>
<tr>
<td>L-LEU</td>
<td>-115.2</td>
<td>15.8</td>
<td>131.8</td>
<td>13.5</td>
</tr>
<tr>
<td>L-LYS</td>
<td>-118.9</td>
<td>17.5</td>
<td>134.2</td>
<td>15.3</td>
</tr>
<tr>
<td>L-THR</td>
<td>-123.9</td>
<td>14.0</td>
<td>138.1</td>
<td>14.3</td>
</tr>
<tr>
<td>D-ALA</td>
<td>130.2</td>
<td>-</td>
<td>-143.8</td>
<td>-</td>
</tr>
</tbody>
</table>

lists were updated every 10 fs with a cutoff of 0.8 nm and the time step was 2 fs [16]. The leap-frog algorithm was used for integrating Newton’s equations of motion. In the NPT simulations the v-rescale thermostat[18] with a coupling time constant of 0.2 ps controlled the temperature, while the Parrinello-Rahman barostat[19] with a coupling time constant of 1.0 ps kept the pressure constant. After energy minimization the CPs were solvated with a dodecahedron box of TIP3P [17] water molecules extending 1.7 nm away from the solute atoms. The energy minimization, equilibration, and production runs of various nanotubes (CPNT) were carried out in different stages: (i) solvent equilibration for 10 ps by restraining the heavy atoms of the CPNT systems, (ii) a 1000 ps total system equilibration run by restraining the $C_\alpha$ atoms of the CPNTs and (iii) production MD for 100 ns with a 2 fs time step. From the production run, a frame for every 2 ps was collected for trajectory analysis. All the simulations were carried out in the NPT ensemble, except for the high temperature simulations and the pressure was maintained at 1 bar.

Table 5.2.2 gives information on the number of CP units in each CPNT system, total number of residues, water molecules, and initial volume of the periodic box.

5.2.3 Transition Path Sampling

The flexible one-way TPS algorithm

Transition Path Sampling [21, 22] (TPS) allows efficient sampling of infrequent transitions between two predefined stable states by harvesting an ensemble of trajectories that lead over a high free energy barrier, connecting the two stable states. Starting from an initial reactive path connecting the two stable states, TPS performs a Markov Chain Monte Carlo random walk in trajectory space by selecting a time frame of the current trajectory, changing the momenta slightly and shooting off a new trial trajec-
Table 5.2.2: Original CP sequence: Composition of systems containing CPNTs of sequence cyclo-[(L-LYS-D-ALA-L-LEU-D-ALA)$_2$]. AFCP sequence: Composition of systems containing nanotubes of AFCP sequence cyclo-[(L-LYS-D-ALA)$_2$-(L- THR-D-ALA)$_2$].

<table>
<thead>
<tr>
<th>System</th>
<th>#CP</th>
<th>#AA</th>
<th>#H$_2$O</th>
<th>#atoms</th>
<th>V (nm$^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPNT2</td>
<td>2</td>
<td>16</td>
<td>4093</td>
<td>12519</td>
<td>127.48</td>
</tr>
<tr>
<td>CPNT3</td>
<td>3</td>
<td>24</td>
<td>4519</td>
<td>13917</td>
<td>139.09</td>
</tr>
<tr>
<td>CPNT4</td>
<td>4</td>
<td>32</td>
<td>4847</td>
<td>15021</td>
<td>151.47</td>
</tr>
<tr>
<td>CPNT5</td>
<td>5</td>
<td>40</td>
<td>5549</td>
<td>17247</td>
<td>174.34</td>
</tr>
<tr>
<td>CPNT6</td>
<td>6</td>
<td>48</td>
<td>6280</td>
<td>19560</td>
<td>197.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System</th>
<th>#CP</th>
<th>#AA</th>
<th>#H$_2$O</th>
<th>#atoms</th>
<th>V (nm$^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPNT2</td>
<td>2</td>
<td>16</td>
<td>4035</td>
<td>12325</td>
<td>122.57</td>
</tr>
<tr>
<td>CPNT3</td>
<td>3</td>
<td>24</td>
<td>4100</td>
<td>12630</td>
<td>131.07</td>
</tr>
<tr>
<td>CPNT4</td>
<td>4</td>
<td>32</td>
<td>4662</td>
<td>14426</td>
<td>144.55</td>
</tr>
<tr>
<td>CPNT5</td>
<td>5</td>
<td>40</td>
<td>5417</td>
<td>16801</td>
<td>169.58</td>
</tr>
<tr>
<td>CPNT6</td>
<td>6</td>
<td>48</td>
<td>6267</td>
<td>19317</td>
<td>195.47</td>
</tr>
</tbody>
</table>

In this work we use the more efficient one way flexible shooting algorithm [22, 23]. First, a time frame $t_{sel}$ is uniform randomly selected from the current (old) path containing $N_o$ frames and the shooting direction (forward or backward) is randomly chosen with equal probability. A new partial trial trajectory of length $\tau_{part}$ is generated by molecular dynamics until the initial state (in case of backward shot) or the final state (in case of a forward shot) is reached. Because of the use of the stochastic $\gamma$-rescale thermostat[18] the new part of the trial trajectory will diverge from the old path. When the generated path ends in the wrong state, the entire move is rejected. The resulting
new partial path is glued to the complementary part of the previous (old) path to yield
the new trial path with a length \( N_n \). If performing a forward move, \( N_n = \tau_{\text{sel}} + \tau_{\text{part}} \), while
for a backward move, \( N_n = (N_o - \tau_{\text{sel}}) + \tau_{\text{part}} \). Next, to maintain detailed balance the
algorithm accepts the trial path according to \( P_{\text{acc}} = \min(1, N(o)/N(n)) \) [24]. In order
to prevent wasting computation time in very long paths connecting the two states, the
maximum allowed path length \( N_{\text{max}} = N(o)/\xi \) is computed in advance, where \( \xi \) is a
random number in the interval \([0, 1]\). This TPS algorithm has been previously used in
other protein systems [25, 26].

**Mechanistic analysis via the path density**

For more insight in the mechanism of the transitions we project the TPS ensemble on
two dimensional path plots. We construct the path density histograms by choosing
two order parameters (OP), and creating a 2D grid initialized to zero. Each path in
the ensemble is then projected on that grid. A bin in the 2D histogram is incremented
with the weight of that path, if the path visited that bin at least once. The path density
plots shows the existence of correlation between particular order parameters in the
mechanisms. It is more informative than a configuration projection which is usually
overwhelmed by intermediate states.

### 5.3 Results and Discussion

#### 5.3.1 Nanotube equilibrium properties

**Stability**

We performed a series of 100 ns MD simulations for each of the systems denoted in
Table 5.2.2. We define a system as being stable if it maintains an anti-parallel cyclic
\( \beta \)-sheet tube-like structure throughout the course of a 100 ns MD simulations. An
unstable system is, on the other hand, a system that deviates considerably from its
tube-like structure during the MD simulation. While this definition is rather heuristic,
it serves our purpose here to identify relative nanotube stability as a function of num-
ber of CPs. Analysis of the MD trajectories (see also Table 5.A.1 and 5.A.2 of the
appendix) shows that most of the CPNT systems maintain their anti-parallel \( \beta \)-sheet
tube-like structure throughout the simulation. The exception is the CPNT2 system of
the AFCP sequence, which exhibited strong deviations from a tube-like structure, lead-
ing in most cases to dissociation of the two CPs. The CPNT2 system of the AFCP
sequence is therefore clearly less stable than the corresponding CPNT2 system of the
original CP sequence. Figure 5.2.2 shows the root mean square fluctuations (RMSF) of
Figure 5.2.2: (a) RMSF of backbone atoms for the original CP sequence. (b) RMSF of backbone atoms for the AFCP sequence (b). Number of side chain hydrogen bonds ($N_{HB}$) between adjacent peptides.

backbone atoms within all CPNT systems of both sequences. The RMSF for CPNT2 is higher than those of longer CPNTs, as the carbonyl and amide groups of the CPs within the CPNT2 system are exposed to solvent molecules. The RMSF of the AFCP sequence CPNT2 is higher compared to the original CP sequence. The higher RMSF is a consequence of strong interactions of the hydroxyl groups of THR side-chains with solvent molecules. As strongly fluctuating backbone atoms perturb the specific backbone-backbone H-bonds of the AFCP sequence CPNT2, the latter dissociates easier than the original CP sequence CPNT2.

Figure 5.2.2a,b shows that CPNTs containing three or more CPs of either sequence have almost identical RMSFs. The sudden increase in stability of CPNTs of the AFCP sequence between CPNT2 and CPNT3 may be explained by the average number of H-bonds between side chains in the CPNTs (Figure 5.2.2)c, which is higher for the AFCP sequence than for the original CP sequence. This can be explained by the formation of H-bonds between hydroxyl groups in adjacent side-chains, as predicted by Vijayaraj et al. [11]. Indeed, when increasing the AFCP nanotube size from CPNT2 to CPNT3, we hypothesize that the hydroxyl groups of threonine side-chains interact less
strongly with water molecules while forming more side chain intramolecular hydrogen bonds, thus decreasing the RMSFs of AFCP sequence CPNTs similar to the original CP sequence RMSF levels.

CPNTs consisting of more than two CPs of both sequences form stable nanotube conformations. Vijayaraj et al. [10] reported, based on RMSD calculations of the whole tube, that larger oligomers (>CPNT3) show less fluctuation and more structural stability. The RMSD of the CPNT3 system should be slightly higher than that of larger nanotubes since two thirds of the system is composed of stronger fluctuating termini. Reanalyzing the data in Ref. [10], the RMSD decreases strongly from CPNT2 to CPNT3 and only slightly between CPNT3 and CPNT4, which suggest that the CPNT3 system, in fact, is stable. Finally, one of the outermost CPs of a CPNT6 system of the original CP sequence left its initial configuration in the course of a 100 ns MD run (Table 5.A.1), suggesting the original CP system might be less stable than the AFCP CPNT6 system, possibly due to additional stabilization of the CPNT by H-bonds between THR side chains, as was predicted by Vijayaraj et al. [10].

**Ice binding properties of AFCP nanotubes**

Figure 5.3.1 shows a) the intramolecular and b) the inter-molecular \( C_\alpha - C_\alpha \) distances of THR residues. For the stable CPNTs (>CPNT2) the average value of the intramolecular and the intermolecular THR \( C_\alpha - C_\alpha \) distances are, respectively, 7.036 and 4.875 Å. These distances are similar to the values reported by Li et al. [27] (respectively, 6.75 and 4.75 Å) and fall within the ranges reported by Zou et al. [7] (5.90-7.46 Å and 3.95-4.96 Å). The distances are very close to the ice Ih unit cell dimensions \( a = 4.518 \) Å and \( c = 7.356 \) Å [28] and as a result, lead to good lattice matching. Therefore, we expect that the AFCP sequence nanotubes exhibit similar or the same ice-binding ability as
### Table 5.3.1: Stable state definitions as a function of number of backbone hydrogen bonds, $C_{\alpha}$—RMSD of CP6 and minimum distance $d_{\text{min}}$.

<table>
<thead>
<tr>
<th>State</th>
<th>RMSD (nm)</th>
<th>$d_{\text{min}}$ (nm)</th>
<th>H-bonds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>min</td>
<td>max</td>
<td>min</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0.03</td>
<td>-</td>
</tr>
<tr>
<td>U</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>

CfAFP-501 spruce budworm isoform and are able to depress the freezing point.

#### 5.3.2 Mechanism of self-assembly by Transition Path Sampling

**The initial path**

We applied TPS to study the self-assembly process of the AFCP nanotube, in particular, the association step of a single CP to a growing nanotube, here chosen as the CPNT6. However, as the MD trajectories are time reversible we can also instead focus on the dissociation of a single CP from a stable nanotube. This dissociation transition occurs between two stable states, the bound state B and the unbound state U. The unbound state U consists of a CPNT5 and single CP, and can be simply defined by a minimum distance larger than 1 nm. In the bound state B all of the six CPs within the CPNT6 are part of an anti-parallel cyclic $\beta$-sheet nanotube structure. In this state the $C_{\alpha}$ RMSD of the outermost peptide CP6 from its natively bound state is less than 0.03 nm, and eight backbone hydrogen bonds are present between the outermost peptide and the rest of the nanotube. The stable state definitions for U and B are given in Table 5.3.1.

TPS requires an initial path between the two stable states, which for the association process is difficult to obtain with straightforward MD simulations, and might take many microseconds. Indeed the unbinding free energy difference of removing the top CP from a different hexamer sequence (cyclo-[(D-ALA-L-ALA)$_4$]) was found in the order of $\approx 7$ kcal/mol, giving rise to a microsecond timescale for the association [10]. While we indeed did not observe a spontaneous dissociation in the AFCP CPNT6 system in the stability MD trajectories, (Table 5.A.2), such a dissociation did occur in a high temperature (450 K) simulation. From this we constructed an equilibrated initial path at 300 K by performing a committor analysis, which consists of shooting off several room temperature trajectories for selected frames on the 450 K pathway to find a frame where the probability of returning to the initial state is similar to ending in the final state. The initial path is constructed by gluing two partial trajectories starting from the
Figure 5.3.2: The initial pathway for the TPS, as a function of a) backbone hydrogen bonds between CP6 and CP5 (N_{HB}), b) CP6 RMSD, and c) minimum distance (d_{min}) between peptide CP6 and the remainder of the tube.

same frame, and ending up in different states. The initial path is graphically illustrated as a function of time in Figure 5.3.2.

**Transition path sampling**

The TPS simulations consisted of 1084 trial one-way flexible shooting moves, which resulted in 325 accepted paths with an overall acceptance ratio of 0.30. The path tree is plotted in figure 5.3.3, and illustrates the decorrelation between the successive accepted shooting moves. Starting from the top each horizontal line indicates an accepted shooting attempt. Red lines indicate forward shots, green lines indicate backward shots. The thin black vertical lines indicate the shooting point location on the previous path. Each accepted new path thus consists of the newly formed green/red partial path, together with the complementary part of the previous path. A measure of decorrelation on the transition path ensemble is the number of decorrelated paths. A path is considered decorrelated when it shares no time slice with the previous decorrelated path. In our simulation we obtained 20 completely decorrelated paths. Inspection of the tree reveals a distribution of different path lengths with shorter and longer reactive pathways. Figure 5.3.4 shows this reactive path length distribution. The path length distribution
is peaked around 2 ns and has an average path length of 8.96 ns. The distribution is roughly Poissonian, with a long tail up to 40 ns. In addition, a second peak visible around 20 ns indicates that there are multiple mechanisms in this transitions.
5.3.3 Analysis of the path ensemble

Inspection of the transition path ensemble revealed three different dissociation mechanisms labeled I, II and III, with three on-pathway intermediate states, denoted iB, iA, and iC (see figure 5.3.5). Intermediate state iA is partly dissociated, but has still a few bound state backbone hydrogen bonds between CP6 and the nanotube intact, and can thus be seen as a ‘docked state’. Intermediate state iB is a misfolded state defined by a THR side chain of CP6 located inside the nanotube. Intermediate state iC is characterized by a CP6 peptide that has all bound state hydrogen bonds broken, but is still associated to the side of the tube. From the distribution of path length (figure 5.3.4) we can deduce that the average path length of each type of transition varies. The paths following mechanism III involving transitions BiAiCU (purple curve) are on average 10.76 ns, and are the origin for the long tail in the distribution. The much fast direct dissociation transition II via intermediate iA (BiAU) is much faster has an average length 3.28 ns (green curve). The dissociation transition I visiting on-pathway intermediate iB (BiAiBU) takes average 19 ns (blue curve). In addition, we identified several paths mixing these mechanisms such as BiAiBiCU and BiAiCiBU (see Figure 5.3.4).

To further understand the association/dissociation transition, we analyze the mechanisms of three selected pathways 3, 35, and 23 respectively corresponding to transition I, II and III. All dissociation/association pathways visit the docked intermediate state iA. Figure 5.3.5 shows that in intermediate iA CP6 has lost its planar β sheet conformation and has instead a V-shape, with only a fraction of the backbone hydrogen bonds

![Figure 5.3.4: Length distribution of the transition path ensemble trajectories. The total distribution is highlighted in red, and is split in the underlying distributions according to their transition mechanism: BiAU (green), BiAiBU (blue), BiAiCU (purple), BiAiBiCU (cyan), BiAiCiAU (grey), and BiAiCiBU (black).](image-url)
Figure 5.3.5: Mechanisms found in the Transition Path ensemble. The three corresponding mechanisms are: mechanism I ($B \rightleftharpoons iA \rightleftharpoons iB \rightleftharpoons U$), mechanism II ($B \rightleftharpoons iA \rightleftharpoons U$), mechanism III ($B \rightleftharpoons iA \rightarrow iC \rightarrow U$)

to CP5 being intact.

Fig. 5.3.6a show the path densities plotted as a function of several order parameters, the RMSD of CP6 with respect to the correctly bound nanotube configuration, the minimum distance CP6 to the nanotube, and the minimum distance between the CP6 and the center of mass of CP5. Here, the RMSD values around 0.3 correspond to the docked intermediate state iA. The three main mechanisms are visible as broad channels in the path densities.

**Mechanism of transition I (path 3)** Mechanism I consists of transitions $B \rightleftharpoons iA \rightleftharpoons iB \rightleftharpoons U$. The shape of the meta-stable intermediate iB resembles intermediate state iA, but with the side chain of THR-45 no longer H-bonded to the backbone of CP5, but moved towards the interior of the tube. Fig. 5.3.7bottom left shows that there are almost no hydrogen bonds formed between THR-45 side chains and the remainder
of the nanotube in the course of path 3. Fig. 5.3.7bottom right displays, moreover, that THR-45 has fewer H-bonds with water molecules than THR-43. The threonine side-chain is hold in place by a hydrophobic interaction between its methyl group and the interior of the nanotube. Threonine entry into the nanotube is also visible in the path density as a function of the minimum distance of the CP6 side chains and the center of mass of CP5 backbone (Fig. 5.3.6b). Clearly, when the minimum distance is around 0.1 nm, the system, visits the intermediate state iB. In addition THR-45 may be stabilized by the occasional formation of hydrogen bonds between its hydroxyl group and the remainder of the tube (Fig. 5.3.7 bottom left) or water molecules at the top of the tube (Fig. 5.3.7bottom right). As shown in Fig. 5.3.4, transition I is rare compared to transitions II and III.  

Mechanism of transition II (path 35) Mechanism II consists of transitions B ⇄ iA ⇄ U. Fig. 5.3.8 shows the transition through intermediate iA observed in path 35. Here, several intact backbone-backbone hydrogen bonds are broken at 4.5 ns, while the minimum distance increases and peptide CP6 dissociates without going through another intermediate state. Here the only hydrogen bonds formed between the CP6 and the tube are the backbone ones. Indeed, the red and black line of Fig. 5.3.8a coincide.

Figure 5.3.6: Path density plots of a) the RMSD of CP6 with respect to the correctly bound nanotube configuration vs. the minimum distance \(d_{\text{min}}(\text{top1}_{\text{sidechain}},\text{top2}_{\text{cm}})\) between the CP6 and the center of mass of CP5 and b) the minimum distance \(d_{\text{min}}(\text{top1}_{\text{rest}})\) of CP6 to the nanotube, vs. the minimum distance \(d_{\text{min}}\) between the CP6 and the center of mass of CP5. States B and U are highlighted by the red and black rectangles respectively.  

Mechanism of transition II (path 35) Mechanism II consists of transitions B ⇄ iA ⇄ U. Fig. 5.3.8 shows the transition through intermediate iA observed in path 35. Here, several intact backbone-backbone hydrogen bonds are broken at 4.5 ns, while the minimum distance increases and peptide CP6 dissociates without going through another intermediate state. Here the only hydrogen bonds formed between the CP6 and the tube are the backbone ones. Indeed, the red and black line of Fig. 5.3.8a coincide.
Figure 5.3.7: Time evolution of the order parameters of path 3. Top left: number of H-bonds \((N_{HB}(CP5 – CP6)_{bb})\) between backbone atoms of CP5 and CP6. Top right: RMSD in nm of CP6 (green) and minimum distance \((d_{min})\) in nm between CP6 and the remainder of the tube (red). Bottom left: number of H-bonds \((N_{HB}(Thr – 45_{side} – remainder))\) between the side chain of THR-45 of CP6 and the remainder of the tube. Bottom right: number of H-bonds \((N_{HB}(Thr – water)_{CP6})\) between THR side-chains and water molecules. The side chain of THR-45 is displayed in red and the side-chain of THR-43 in green.

**Mechanism of transition III (path 23)** Mechanism III consists of transitions \(B \rightleftharpoons iA \rightleftharpoons iC \rightleftharpoons U\). Meta-stable intermediate iC does not resemble intermediate iA and iB. Fig. 5.3.9a shows that all of the unique backbone-backbone CP5-CP6 hydrogen bonds are broken in the last part of path 23 where the system is in intermediate iC (red line). The H-bonds between CP6 and the nanotube (black line) are, on the contrary, frequently formed and broken, indicating that intermediate iC is probably not only stabilized by H-bonds, but also by hydrophobic interactions between CP6 and the nanotube. The frequent formation and breaking of these interactions suggests CP6 may slide along the tube. Indeed the RMSD in Fig. 5.3.9b increases, whereas the minimum distance remains small, indicating the peptide remains associated to the nanotube.

Fig. 5.3.7, Fig. 5.3.9 and Fig. 5.3.8 show all dissociation processes are initiated by breaking some of the stable inter-molecular backbone-backbone H-bonds. Once all \(\beta\)-sheet hydrogen bonds between backbone-atoms are broken, H-bonds and hydrophobic interactions between CP6 and the nanotube may exist. Once those are broken CP6
fully separates from the nanotube. The rate limiting step of the dissociation of CPs is
the breaking of all backbone-backbone H-bonds as suggested by Vijayaraj et al. [11].
During this process the system can get trapped in the meta-stable intermediate iB, due
to hydrophobic interactions between the methyl group of a THR and the interior of
the nanotube. The dissociation is realized by the disappearance of all side-chain and
backbone interactions between CP6 and the nanotube. The prediction of Vijayaraj et
[11] that the process is followed by the annihilation of various side-chain side-chain
interactions is, therefore, too simplified. Furthermore, we found that during dissocia-
tion the system may be trapped in intermediate iC, in which CP6 is associated to the
side of the tube. From the perspective of association, all TPS pathways are examples
of a dock-lock mechanism that has also been identified in protein aggregation studies
[29–32].

The existence of intermediate iC also explains why CPNTs align in a linear fashion
within their crystal structure. Due to the fact that a single CP interacts strongly with
the side of a nanotube, two nanotubes should also interact strongly with one another.
As a result solubility of the nanotube is low and crystals should early form as observed
in experiment [33]. This is because for nanotube solvation all interactions between the
nanotubes have to be broken.

5.3.4 Comparison of the association time scales

From the view point of the association process, a CP might therefore first bind non-
specifically to the tube, followed by a random walk along the nanotube before it docks
to end of the nanotube, and locks into place. Under certain conditions this one-dimensional diffusion along the nanotube might be faster than a random search for the end of the growing tube in a 3-dimensional volume. This can be argued as follows. The effective diffusion limited (Smoluchowski) association rate for a particle attaching to a specific point with a contact radius $\sigma$

$$k_{on} = \frac{4\pi D \sigma}{V}, \quad (5.1)$$

where $D$ is the translational diffusion constant, and $V$ is the volume (determined by the concentration of growing ends). The timescale connected to this rate is simply its inverse $\tau_{on} = k_{on}^{-1}$. The association rate constant of a CP associating to a nanotube of length $L$ is instead is given by [34]

$$k_{on}^{tube} = \frac{4\pi D L}{V \ln(2L/\sigma)} \quad (5.2)$$

The diffusion timescale to diffuse along the nanotube, in a quasi 1D random walk, is of the order of

$$\tau_1 = \frac{L^2}{3D_1}, \quad (5.3)$$

where $D_1$ is the translational diffusion constant along the nanotube. This leads to a total association timescale via pathway III of the order of

$$\tau_{on}' = \frac{V \ln(2L/\sigma)}{4\pi DL} + \frac{L^2}{3D_1} \quad (5.4)$$
This timescale can always be made smaller than the timescale of the direct mechanism, \( \tau_{on} \) because the first term is dominant at low concentration, and can be made smaller by considering longer \( L \).

To identify the crossover, we equate the two times scales \( \tau'_{on} = \tau_{on} \)

\[
\frac{V \ln(2L/\sigma)}{4\pi DL} + \frac{L^2}{3D_1} = \frac{V}{4\pi D\sigma} \quad \text{(5.5)}
\]

\[
\frac{\sigma \ln(2L/\sigma)}{L} + \frac{L^2 4\pi D\sigma}{3D_1 V} = 1 \quad \text{(5.6)}
\]

Clearly, the left hand side can be made arbitrarily small by increasing the volume \( V \) and/or increasing the length \( L \). For these condition, the indirect mechanism of association is preferred.

5.4 Conclusion

In summary, we have investigated the stability of nanotube structures formed by CPs with sequence cyclo-[\{(L-lys-D-ala-l-leu-d-ala)\}_2] and sequence cyclo-[\{(L-lys-D-ala)\}_2\cdot(l-thr-d-ala)\}_2] using molecular dynamics simulations. The AFCP sequence contains TXT ice binding motifs as they are found in anti-freeze protein CfAFP-501 which is the most active anti-freeze protein discovered thus far [7, 27].

MD simulations indicated that the AFCP sequence CPNT2 system is less stable than a original CP sequence CPNT2, because THR side-chains in the AFCP sequence CPNT2 system interact strongly with solvent water molecules, yielding strong fluctuations of backbone atoms and, therefore, perturbation of the inter-molecular H-bonds between the CP units. CPNTs of more than two peptides of both sequences maintained their stable anti-parallel \( \beta \)-sheet tube-like structure. The sudden increase in stability of CPNTs of the AFCP sequence (for sizes greater than two CP units) may be explained by the formation of H-bonds between THR side-chains. As a result, backbone atoms fluctuate less and AFCP CPNTs with three units become equally stable with respect to the original sequence CPNTs of the same size. Nanotubes of more than 3 AFCPs are slightly stabler than those of the original CP sequence due to the formation of H-bonds between side-chains in accordance with the prediction of Vijayaraj [11].

We compared the distances between THR in the constructed nanotube structures to THR distances reported for CfAFP-501 in literature. Comparison of the average distances between THR residues in the constructed nanotube structures of the AFCP sequence and CfAFP-501 showed that the distances are very similar. The distances are close to the O-O distances of ice and may lead to good lattice matching, suggesting
the nanotubes of the AFCP sequence show similar or the same ice-binding ability as CfAFP-501 and are able to depress the freezing point.

We studied the self-assembly and dissociation process of a single CP to a CPNT5 using Transition Path Sampling. The TPS simulations indicated that the dissociation process of a single CP from a stable nanotube is initiated by the breaking of native backbone-backbone H-bonds. Once all native intermolecular $\beta$-sheet H-bonds between backbone-backbone atoms are broken, alternative H-bonds and hydrophobic interactions between CP6 and the remainder of the nanotube tube may be still be present. When the latter also break, CP6 truly separates from the nanotube. The determining step in the dissociation of the CP is the breaking of all backbone-backbone H-bonds as reported by Vijayaraj et al., followed by the loss of all side-chain and backbone interactions between CP6 and the remainder of the tube. The observation of Vijayaraj et al. that during the dissociation various side-chain side-chain interactions are lost is thus too simple [11]. From the perspective of association, all TPS paths are examples of the dock-lock mechanism that has also been identified in protein aggregation studies [29–32].

We found evidence for at least three meta-stable on-pathway intermediates (iA, iB and iC) in the dissociation/associating process. The presence of intermediate iC indicates that CPNTs interact strongly with one another, in accordance with a proposed crystal structure of highly aligned nanotubes. This indicates that synthetic CP crystals should have a low solubility, which is consistent with experimental observation. [33].

Using a simple rate theory expression for the association timescale, we conclude that at low concentration of the AFCPs, association is should occur via a mechanism involving intermediate iC.

Insight in these mechanisms can lead to improved design of the cyclic systems. Current efforts are underway to do so [33].
Appendix

5.A Brute force MD and sequences of different nanotubes

<table>
<thead>
<tr>
<th>System</th>
<th>Run 1</th>
<th>Run 2</th>
<th>Run 3</th>
<th>Run 4</th>
<th>Run 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPNT2</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>CPNT3</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>CPNT4</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>CPNT5</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>CPNT6</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B → I</td>
</tr>
</tbody>
</table>

Table 5.A.1: Original sequence: Visiting states of different size CPNTs in the course of a 100 ns NPT simulations. B denotes the bound state and I the area of phase space which does not belong neither to B nor U.

<table>
<thead>
<tr>
<th>System</th>
<th>Run 1</th>
<th>Run 2</th>
<th>Run 3</th>
<th>Run 4</th>
<th>Run 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPNT2</td>
<td>B → I → U → I → U → I</td>
<td>B</td>
<td>B → I → U → I</td>
<td>B</td>
<td>B → I → B</td>
</tr>
<tr>
<td>CPNT3</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>CPNT4</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>CPNT5</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>CPNT6</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
</tbody>
</table>

Table 5.A.2: AFCP sequence: Visiting states of different size CPNTs in the course of a 100 ns NPT simulations. B denotes the bound state and U the unbound state. I denotes the area of phase space which does not belong neither to B nor U.
Table 5.A.3: Original CP sequence: schematic representation of the arrangement and chirality of the amino acids in the CP chains of sequence cyclo-[(L-LYS-D-ALA-L-LEU-D-ALA)\textsubscript{2}] within various CPNTs. The sequential chains are named as CP1 to CP6. The terminals are assumed to be bonded in order to form the cyclic structure.

<table>
<thead>
<tr>
<th>Chain Orientation</th>
<th>Chain Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>N → C</td>
<td>CP1</td>
</tr>
<tr>
<td>C ← N</td>
<td>CP2</td>
</tr>
<tr>
<td>N → C</td>
<td>CP3</td>
</tr>
<tr>
<td>C ← N</td>
<td>CP4</td>
</tr>
<tr>
<td>N → C</td>
<td>CP5</td>
</tr>
<tr>
<td>C ← N</td>
<td>CP6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chain Orientation</th>
<th>Chain Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>N → C</td>
<td>L LYS-1</td>
</tr>
<tr>
<td>C ← N</td>
<td>L LYS-9</td>
</tr>
<tr>
<td>N → C</td>
<td>L LYS-17</td>
</tr>
<tr>
<td>C ← N</td>
<td>L LYS-25</td>
</tr>
<tr>
<td>N → C</td>
<td>L LYS-33</td>
</tr>
<tr>
<td>C ← N</td>
<td>L LYS-41</td>
</tr>
</tbody>
</table>

Table 5.A.4: AFCP sequence: schematic representation of the arrangement and chirality of the amino acids in the CP Chains of sequence cyclo-[(L-LYS-D-ALA-)\textsubscript{2}-(L-THR-D-ALA)\textsubscript{2}] within various CPNTs. The sequential chains are named as CP1 to CP6. The terminals are assumed to be bonded in order to form the cyclic structure.

<table>
<thead>
<tr>
<th>Chain Orientation</th>
<th>Chain Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>N → C</td>
<td>L LYS-1</td>
</tr>
<tr>
<td>C ← N</td>
<td>L LYS-9</td>
</tr>
<tr>
<td>N → C</td>
<td>L LYS-17</td>
</tr>
<tr>
<td>C ← N</td>
<td>L LYS-25</td>
</tr>
<tr>
<td>N → C</td>
<td>L LYS-33</td>
</tr>
<tr>
<td>C ← N</td>
<td>L LYS-41</td>
</tr>
</tbody>
</table>

120
Bibliography


[33] Personal correspondence with dr. ilja k. voets, macromolecular and organic chemistry & institute for complex systems, eindhoven university of technology, the netherlands.

Chapter 6

Spring shooting, a novel efficient transition path sampling move

We present a novel transition path sampling shooting algorithm for efficient sampling of complex (biomolecular) activated processes with asymmetric free energy barriers. The method employs a fictitious potential that biases the shooting point toward the transition state. The method is similar in spirit to the aimless shooting technique by Peters and Trout [B. Peters and B. L. Trout, J. Chem. Phys. 125, 054108 (2006)], but is targeted for use with the one-way shooting approach, which has been shown to be more effective than two way shooting algorithms in systems dominated by diffusive dynamics. We illustrate the method on a 2D Langevin toy model, the association of two peptides and the initial step in dissociation of a β-lactoglobulin dimer. In all cases we show a significant increase in efficiency.

6.1 Introduction

Straightforward molecular dynamics of activated molecular processes, such as chemical reactions, biomolecular isomerizations, association processes, or phase transitions, is inefficient, as much time is wasted on sampling the stable states, in comparison to the transition itself. While many biasing methods exist that address the exponential time scale problem occurring in such activated processes, they typically rely on good reaction coordinates (e.g. umbrella sampling[1], flooding[2], local elevation[3], adaptive bias force[4], meta-dynamics[5]) or give up access to kinetic and mechanistic details (such as replica exchange molecular dynamics [6]). Transition path sampling was developed to bypass these problems by focusing on the transition paths between stable
The TPS algorithm employs importance sampling to generate an ensemble of unbiased dynamical trajectories that connect an initial with a final stable state, and is therefore independent of an a priori choice of order parameters or reaction coordinate that describes the transition. Rather, the reaction coordinate can be extracted from the sampled path ensemble. The method has been successfully applied to various activated processes (see ref. [10–12] for a review).

The standard TPS algorithm[8] takes an existing path, and creates a new trial path from it. The resulting trial path is accepted using a Metropolis-Hastings criterion based on detailed balance, designed to maintain the equilibrium path ensemble distribution. This generic TPS setup allows for much flexibility, as the way of generating a new trial path from an existing path is entirely up to the user. Even in the seminal publications on TPS[7], different approaches were considered, such as frame by frame moves, or dynamical schemes evolving the entire pathway simultaneously. The vanilla type, standard TPS algorithm relies on the so-called shooting move[9], which selects a random frame from an existing initial trajectory of fixed duration, alters the momenta slightly and integrates the equation of motion forward as well as backwards in time, until the total duration of the trajectory has been reached. For the shooting algorithm the acceptance criterion is simply determined by whether it connects the two stable states. (The original approach also included the shifting move. This move shifts the origin of the paths, but does not create new decorrelated pathways, at least for deterministic MD). The shooting moves are particularly effective since they allow for a quick decorrelation of the subsequent pathways, while still giving reasonable good acceptance probability as long as the trial shooting points are not too far from the transition states. While the shooting move has been employed for the majority of cases[10–12], other TPS approaches can be very useful as well, such as the noise history approach[13].

Following the introduction of the original shooting algorithm, many improvements were made. The transition interface sampling (TIS) [14] based on flexible shooting moves allowed restricting the path length to the necessary minimum, thus minimizing the computational effort. Biasing the shooting point toward the barrier region also is very useful, as it prevents ill-fated shots from near the stable states and improves the acceptance ratio[12, 15]. Another important development was the aimless shooting approach by Peters and Trout[16], which was originally developed to be applied in conjunction with a maximum likelihood based reaction coordinate analysis advocated by the same authors, but can be used independently as a valid shooting algorithm. The idea of the aimless shooting method is to keep the trial shooting point distribution close to the transition barrier. The reason for this goal is twofold. First, the acceptance ratio for the shooting is higher when the shooting point is close to the transition state. Second, the likelihood maximization relies on information close to the dividing surface. The aimless shooting method achieves these goals by basing the trial shooting point on
the previous shooting point, shifting it by a fixed number of frames in the forward or backward direction. When this shift happens to bring the shooting point closer to the transition state the acceptance becomes more likely, thus providing a natural tendency to stay close to the barrier, thus circumventing the need for predefined order parameters to bias in.

At the same time, it became clear that the standard two-way shooting move was less effective for complex systems, in particular, biomolecular systems, where the transition paths could become exceedingly long, much longer than the Lyapunov time scale. This means that (deterministic) dynamical pathways effectively become diffusive. In the original TPS publications it was already clear that for diffusive dynamics one had to employ the so-called one-way or stochastic shooting method[17]. This method attempts to create only a forward or a backward partial path from a uniformly randomly chosen trial shooting point rather than shooting in both directions. When the forward (backward) path is acceptable, that is, when it reaches the final (initial) state, the trial path replaces only part of the existing initial part, while the complementary part of the previous path is retained. While clearly several shots are needed to create a entirely new path, the one-way shooting algorithm is more efficient for diffusive or stochastic dynamics. Moreover, even in case of deterministic molecular dynamics (coupled to a stochastic thermostat) in which the molecular chaos sets in long before the average length of the path is reached, one-way shooting is preferred[18]. In fact, for biomolecular systems it is currently the standard approach. The effectiveness of the path sampling can be tracked using a so-called path tree [19].

The one-way shooting method has also several drawbacks. First, it requires more shots to decorrelate paths (although not more computer time). More seriously, it suffers in efficiency for asymmetric barriers. Asymmetric barriers occur, for instance, when the system on one side of the main barrier is trapped in an intermediate state, while it can easily reach the stable state on the other side. This means the paths on the trapped side become much longer (Note that this is, of course, also a problem for the two-way shooting algorithm). When uniform one-way shooting is used, this asymmetry leads to many more shooting attempts on one side of the barrier with respect to the other. Due to the asymmetry, most shooting points are chosen on the trapped side. A biasing function that selects shooting points around the transition state using an order parameter can ameliorate this situation. However, then one needs to know a priori what order parameter to use to parameterize the transition state, and TPS was developed just to avoid that. Clearly one would prefer to use an algorithm such as Aimless shooting, which automatically finds the transition state region. The current work provides such an algorithm, which we denote “spring shooting”. Inspired by Aimless shooting, the method chooses a new shooting point by shifting the old shooting point by a number of frames, and accepts or rejects this shift while applying a bias, the spring. This bias
is different for a forward with respect to a backward shot. For a forward shot the shooting point is pulled to lower frame number by the “spring”, for a backward shot the situation is the other way around. The opposing forces balance when the shooting point oscillates around the transition state.

The paper is organized as follows: we start with introduction and derivation of the algorithm. We illustrate method using a 2D Langevin model, and show that the efficiency improves with respect to the uniform shooting. We then further illustrate the new algorithms on the association reaction of two FF solvated dipeptides. We end with an application of a large system, the associating/dissociation reaction of the β-lactoglobulin dimer.

6.2 Methods

6.2.1 Derivation of the spring shooting algorithm

A TPS simulation performs a Monte Carlo random walk in trajectory space. A dynamical trajectory \( \mathbf{x}(L) = \{x_0, x_1, \ldots, x_L\} \) is discretized in time by a time-step \( \Delta t \) into \( L + 1 \) slices (or frames) \( x_\tau \). Each time frame \( x_\tau = \{\mathbf{r}_\tau, \mathbf{p}_\tau\} \) contains all positions \( r \) and momenta \( p \) of all \( N \) particles in the system at time \( t = \tau \Delta t \). The probability \( \mathcal{P}_{AB}[\mathbf{x}(L)] \) for a path of a fixed length \( L \) connecting the initial state \( A \) and the final state \( B \) is given by:

\[
\mathcal{P}_{AB}[\mathbf{x}(L)] = Z_{AB}^{-1} h_A(x_0) \mathcal{P}[\mathbf{x}(L)] h_B(x_L).
\] (6.1)

\( \mathcal{P}[\mathbf{x}(L)] \) is the dynamical probability of the discretized path \( \mathbf{x}(L) \), while \( h_\Omega(x) \) denotes an indicator function that is unity when \( x \in \Omega \), i.e. \( x \) is inside a stable state \( \Omega \), and zero otherwise. The product \( h_A(x_0)h_B(x_L) \) thus guarantees that the probability is non-zero only when a path starts from \( A \) and ends in \( B \). The normalization factor \( Z_{AB} \) is akin to a partition function. The dynamical path probability \( \mathcal{P}[\mathbf{x}(L)] \) is given by

\[
\mathcal{P}[\mathbf{x}(L)] = \rho(x_0) \prod_{\tau=0}^{L-1} p(x_\tau \rightarrow x_{(\tau+1)}),
\] (6.2)

where \( p(x_{\tau} \rightarrow x_{(\tau+1)}) \) represents the Markovian probability for transitions from a phase point at time frame \( \tau \) to one at time frame \( \tau + 1 \) (e.g. a delta function for molecular dynamics, or a Gaussian for Langevin dynamics [7]). All trajectories that connect the two defined stable states form the path ensemble, which plays a key role in TPS as the representation of the rare event process under study. A definition of reaction coordinate is not necessary in TPS, but the reactant \( A \) and product \( B \) regions should be defined properly by order parameters. These order parameters should be chosen such
that the two states can not only be distinguished from each other, but also that each state lies inside the basin of attraction of that stable state \[12\].

The shooting algorithm \[7\] can efficiently sample the path ensemble by selecting a random slice, changing the momenta slightly, and shooting off a new direction forward and backward in time. Each path that still connects A with B can be accepted in the Monte Carlo procedure. The general acceptance rule for the shooting move is\[8, 12\]

\[
P_{\text{acc}}[x^{(o)} \to x^{(n)} ] = h_A(x_0^{(n)}) h_B(x_L^{(n)}) \min \left[ 1, \frac{\mathcal{P}[x^{(n)}(L^{(n)})] P_{\text{gen}}[x^{(n)} \to x^{(o)}]}{\mathcal{P}[x^{(o)}(L^{(o)})] P_{\text{gen}}[x^{(o)} \to x^{(n)}]} \right] \quad (6.3)
\]

where the min function returns the lower of its arguments and \(P_{\text{gen}}[x^{(n)} \to x^{(o)}]\) denotes the generating probability for creating a new trajectory from an old one, making use of microscopic reversibility. One can show that for the shooting move this acceptance rule reduces to

\[
P_{\text{acc}}[x^{(o)} \to x^{(n)} ] = h_A(x_0^{(n)}) h_B(x_L^{(n)}) \min \left[ 1, \frac{P_{\text{sel}}(\tau', x^{(o)})}{P_{\text{sel}}(\tau, x^{(o)})} \right] \quad (6.4)
\]

where \(P_{\text{sel}}(\tau, x)\) denotes the probability to select the shooting point. In the simplest case, this selection probability is uniform, and is just determined by the length \(P_{\text{sel}}(\tau, x) = 1/L\). Note that when the generation of the trajectories is halted upon reaching a stable state, as is done in flexible shooting, the length of the paths varies and the acceptance criterion becomes dependent on the ratio \(L^{(o)}/L^{(n)}\)\[12, 14\].

However we are free to choose a different selection probability, for instance

\[
P_{\text{sel}}(\tau, x) = \frac{f(\lambda(x_\tau))}{\sum_{i=0}^{L} f(\lambda(x_i))} \quad (6.5)
\]

where the function \(f(\lambda(x_i))\) is based on the order parameter \(\lambda(x)\) which turn depends solely on the coordinates of the time frame.

To sample from this distribution one can generate the shooting points by uniform selection, do the shooting move and then accept with the criterion Eq. 6.4, or, alternatively, generate the shooting point directly from this distribution and perform the shooting, and accept with \(P_{\text{acc}}[x^{(o)} \to x^{(n)} ] = h_A(x_0^{(n)}) h_B(x_L^{(n)})\). To generate the shooting point from the distribution Eq. 6.5 one can use e.g. rejection (von Neumann) sampling, or other approaches[20, 21].

The shooting move is thus separated into two stages. The first is the generation of the shooting point, by a Metropolis Monte Carlo scheme, in which undesired shooting points can be quickly discarded, with a acceptance ratio \(P_{\text{acc}}^{\text{sp}}[\tau \to \tau'] = \)
The second stage is the creation and acceptance/rejection of the trial path. The total move acceptance ratio can be seen as product of the two stages

\[
P_{\text{acc}}[\tau \rightarrow \tau'; x^{(o)} \rightarrow x^{(n)}] = P_{\text{acc}}^p[\tau \rightarrow \tau'] P_{\text{acc}}^\text{shot}[x^{(o)} \rightarrow x^{(n)}]
\]

which is indeed the same as Eq. 6.4. In previous work such an approach has been used to great effect [15].

Eq. 6.6 holds also for the one-way shooting algorithm, in which one randomly chooses to generate a forward or a backward partial path starting from the shooting point. The complementary partial part is retained from the old trajectory. Note that one-way shooting requires some stochasticity in the dynamics, so that the shooting point itself does not have to be altered. Indeed, altering the shooting point like in standard two-way TPS would lead to unrealistic dynamics along the entire path.

The spring shooting algorithm is especially developed for use with the one-way algorithm. Similar to the aimless shooting algorithms, the spring shooting selects the shooting point not based on coordinates but only based on the time frame index:

\[
p_{\text{sel}}(\tau, x) = p_{\text{sel}}(\tau)
\]

The essence of the spring shooting algorithm is to treat the forward and backward shooting move as independent moves. Similar to Aimless shooting, the spring shooting start by shifting the shooting point with respect to the successful shooting point on the previous path by a certain number of frames. Whereas in Aimless shooting this shift is symmetric, in the spring method it is selected with an exponential function:

\[
p_{\text{sel}}(\tau) = c \exp(s k \tau)
\]

where \( \tau \) is (absolute) frame index, \( k \) a force constant determining the magnitude of the bias, and \( s \in \{-1, 1\} \) is determined by the direction of shooting. \( s = -1 \) for forward shooting, and \( s = 1 \) for backward shooting, and \( c \) is the normalization. Since we now allow all times (in principle) this normalization constant is given by

\[
c = \frac{1}{\sum_{b}^{\infty} \exp(s k \tau)}
\]

where the boundary \( b \) of the allowed time frame window be made arbitrarily large. While this normalization constant vanishes for \( b \rightarrow \infty \), it cancels out in the acceptance criterion itself\(^1\)

\[
P_{\text{acc}}^p[\tau \rightarrow \tau'] = \min\left[1, \frac{\exp(s k \tau')}{\exp(s k \tau)}\right] = \min[1, e^{s k \Delta \tau}]
\]

\(^1\)Note that here the limit is only taken after the construction of the acceptance criterion. An analogous
where $\Delta \tau = \tau' - \tau$ is the number of shifted frames from the previous shooting point. As a large $\Delta \tau$ either yields an exponentially small acceptance ratio or is likely to produce a failed shot, we limit the choice of $\Delta \tau$ between the interval $[-\Delta \tau_{\text{max}}, \Delta \tau_{\text{max}}]$, analogous to the maximum allowed displacement in a regular MC translational move. When the trial shooting point falls outside the current path the acceptance probability becomes zero, and the move is rejected.

As the spring shooting algorithm treats the forward and backward shots separately, the complete algorithm is as follows:

1. Select with equal probability a forward or a backward move. set $s = -1$ in case of forward move, $s = 1$ in case of backward.

2. Select a uniform shift $\Delta \tau$ in shooting point from the interval $[-\Delta \tau_{\text{max}}, \Delta \tau_{\text{max}}]$. Add this shift to the previous shooting point frame index $\tau' = \tau + \Delta \tau$. Reject the entire move if the index is outside of the current path.

3. Accept the trial shooting point $\tau'$ according to Eq. 6.9, otherwise reject the entire move.

4. Create partial path (forward or backward) employing molecular dynamics with a stochastic thermostat. Halt the path when a stable state is reached. Reject the move when the total trial path length exceeds a maximum length $L_{\text{max}}$. Glue the path to the complementary existing path, and accept according to the standard criterion $P_{\text{acc}}[x^{(o)} \rightarrow x^{(n)}] = h_A(x_0^{(n)})h_B(x_L^{(n)})$.

The advantage of this approach is that unfavorable shooting points are discarded without extra cost. The exponential shooting selection function can be viewed as a kind of external field pushing (or pulling) the shooting point forward (or backward). The shooting points for forward moves are preferably taken from frames before the previous shooting point, and the backward movers preferably after the previous shooting point. In this way pathways are decorrelated as much as possible, without wasting time creating partial paths that do not contribute to the decorrelation. Note that the acceptance criterion for the spring move is very simple, and in contrast to the standard one-way flexible path length algorithm, does not depend on the instantaneous (fluctuating) path length. Note also that the algorithm rejects trial paths which become longer

situation arises when deriving the acceptance ratio for a regular translational MC move (in an external field). The distribution function is an integral over the (possibly infinite) space, but since in the Metropolis acceptance criterion the two distributions have the same normalization constant the ratio remains finite. Note also that the choice of the $\Delta \tau_{\text{max}}$ is independent of $b$. The $\Delta \tau_{\text{max}}$ is analogous to the maximum allowed displacement in a regular MC translational move.
than $L_{\text{max}}$, which is set to prevent memory or storage problems, or as an indication that the path generation went awry, e.g. got stuck in an long-lived intermediate state.

The algorithm requires an initial shooting point location. While in principle it does not matter where this point is chosen, in practice it is more efficient to start close to the barrier. When the initial trajectory comes from gluing two trajectories together (e.g. obtained in a committor analysis), the initial shooting point naturally should be chosen close to the configuration from where these trajectories were initiated. In other cases it might be wise to establish a proxy for the barrier e.g. the maximum energy along the initial path. Even visual inspection might help to find a suitable shooting point location. Nevertheless, we stress that even when starting an entirely wrong shooting point, the algorithm ensures that the shooting point diffusive to the top of the barrier eventually.

The algorithm has two crucial parameters: the maximum frame shift $\Delta \tau_{\text{max}}$ and the spring constant $k$. In the examples we investigate how the performance of the algorithm depends on the choice of these two parameters.

### 6.2.2 Simulation details

#### 2D Langevin model

To test the performance and validity of the spring shooting algorithm we employ the method first on a toy model consisting of a particle evolving according to Langevin dynamics in a simple 2D asymmetric potential. We choose the potential to be

$$V(x,y) = -6e^{-(x+y)^2} - y^2 - 12e^{-3(x+y)^2} - y^2 + 0.0177778(0.0625x^4 + y^4)$$

**(6.10)**
This potential is visualized in Fig. 6.2.1 with a contour plot, and shows two minima, one at \( V\{x = -8.9554, y = 0\} = -4.78204 \) and one at \( V\{x = 3.9767, y = 0\} = -5.71887 \). A saddle point is located at \( V\{x = 7.90805, y = 0\} = 4.34544 \). For clarity the 1D projection \( \beta F(x) = -\ln \int dy \exp[-\beta V(x,y)] \) onto \( x \) is also shown, where the \( y \) variable is integrated out using Boltzmann averaging. The inverse temperature \( \beta = 1/k_B T \), with \( T \) the reduced (artificial) temperature and \( k_B \) the Boltzmann constant, determines how rare the transition between minima is. In this 2D potential a single particle moves according to full Langevin dynamics. To integrate the equation of motion we employ the BAOAB algorithm described in Ref[22]. We set time step \( \Delta t = 0.05 \), and a friction of \( \gamma = 1 \). The temperature was \( \beta = 10 \), chosen to ensure a low rate constant. Particles configurations are saved each 10th time step. Frames in the trajectories are thus separated by \( \Delta_{\text{frame}} = 0.5 \). The maximum path length is set to \( L_{\text{max}} = 10^5 \) frames.

**FF dimer**

All atomistic molecular dynamics simulations as well as the system preparation for the FF dimer were performed with the Gromacs 4.5.4 package [23]. All atom interactions were defined using the amber99sb-ildn [24] and TIP3P force fields [25]. The FF segment was taken from the sequence KLVFFA from the amyloid-beta peptide (AB residues 16-21) (PDB 2Y29 [26]) and capped with ACE and NME to render, respectively, the N and C terminal neutral. Two copies of the FF monomer were put at a distance of 1 nm from each other in a cubic box of 30 x 30 x 30, and energy minimized using the conjugate gradient method. After solvation of the box and a second energy minimization, we performed a short equilibration of water (10 ps NPT in ambient conditions) with the peptide positions restrained. The solvated system was equilibrated for 1 ns in ambient conditions in the NPT ensemble and thereafter was subjected to a long 200 ns NPT simulation. All bonds were constrained with the Lincs algorithm. A cutoff of 1 nm was used for the non-bonded interactions. The Particle Mesh Ewald method treated the electrostatic interactions, with a Fourier spacing of 0.12 nm and a 1 nm cutoff for the short range electrostatic interactions. Neighbour lists were updated every 10 fs with a cutoff of 1 nm and the time step was 2 fs [24]. Newton’s equations of motion were integrated with the leap-frog algorithm. In the NPT simulations the Bussi thermostat[27] with a coupling time constant of 0.2 ps controlled the temperature, while the Parrinello-Rahman barostat[28] with a coupling time constant of 1.0 ps kept the pressure constant.

An initial path was taken from the 200 ns MD run. TPS was performed with home written script capable of applying both one-way uniform and spring shooting, generating paths of flexible length. In both types of shooting, the maximum path length \( L_{\text{max}} \)
was set to 2000 frames, and frames were saved every 5 ps. The TPS algorithm generated trajectories in ambient conditions (T= 298 K, P= 1atm) in the NPT ensemble, with the above mentioned settings. Order parameters such as minimum distance and solvent accessible surface area (SASA) were obtained with the GROMACS analysis tools.

**β-lactoglobulin dimer**

All-atom molecular dynamics simulations as well as the system preparation for the β-lactoglobulin dimer were performed with Gromacs 4.6.7 package [23] using GPUs. All atom interactions were defined using the amber99sb-ildn [24] and TIP3P force fields [25]. The β-lactoglobulin (β-lac) dimer system was taken from the protein research databank (PDB 2AKQ). The dimer was put in a dodecahedral box and energy minimized using the conjugate gradient method. After solvation of the box with 20787 water molecules and a second energy minimization, we performed a short equilibration of water (10 ps NPT in ambient conditions) with the protein position restrained. The solvated system was equilibrated for 1 ns in ambient conditions in the NPT ensemble and thereafter was subjected to a long 200 ns NPT simulation. All other settings were as mentioned above in the FF case.

In the course of the 200 ns MD, the β-lactoglobulin dimer remained in its native bound state (see Fig. 6.A.1 of the Appendix). The native contacts were identified as those residue pairs that stayed within a minimum heavy atom distance of 0.4 nm for at least 90% in the 200 ns NPT trajectory. These native contact pairs are listed in Table 6.A.1 of the Appendix. Only 8 residue pairs are shown to fulfill the above criterion. These 8 residue pairs, as well as four native hydrogen bonds (between backbone NH and CO of residues 146-150,148-148,150-146) characterized by Sakurai et al [29] define the stable native (contact) state (NC).

In order to initialize the TPS simulation we need a single path undergoing the dissociation. Therefore we performed a meta-dynamics simulation using the PLUMED package [30] with the above MD settings. The collective variable for MetaD was the center of mass distance of the protein’s CAs, using a Gaussian hill with height 0.4 kJ/mol, and width of $\sigma = 0.1$ kJ/mol which are deposited every 2 ps. The resulting trajectory of the meta-dynamics run indeed undergoes a dissociation event. The trajectory was saved every 20 ps. For completeness we give the heavy atom minimum distance of the proteins along the trajectory in the Appendix (see Fig. 6.A.2). From this biased dimer dissociation trajectory we launched from a particular configuration (frame 108, corresponding to 2140 ps along the trajectory) a series of 10 trial trajectories with random velocities. These trajectories were performed at ambient conditions in the NPT ensemble, using the settings mentioned above. The particular frame 108 turned out to have a reasonable chance to end in the native state (NC), as well as escape the native
state. Two of these trajectories, ending in different stable states, were glued to yield the desired initial unbiased path. Ideally this trajectories would start in the native state and end in as in the unbound state (U), which we defined by a minimum distance between the two proteins larger than 1 nm. However, an MD trajectory from this particular configuration was extremely unlikely to provide trajectories ending in both state NC and state U (see Fig. 6.A.3 in the Appendix), as the system is trapped in an intermediate state, where only a fraction of native contacts is intact, thus preventing the full dissociation to state U. We therefore decided to first sample pathways between the NC state and an intermediate state NN with only one native contact formed.

We performed TPS simulations of the native state to non native state (NC ⇄ NN) at ambient conditions (T=300, P=1 atm) in the NPT ensemble, using the same home written scripts as for the FF dimer TPS, using three different type of one-way shooting schemes, a uniform shooting, Gaussian bias and a spring shooting. The maximum path length for this transition was set to \( L_{\text{max}} = 1500 \) and frames were saved every 10 ps. For the Gaussian bias shooting, we implemented a Gaussian bias in the selection of the shooting point at NC= 4 with a width \( \sigma = 1 \). For the spring shooting move TPS, we performed simulations of \( k = 0.1, 0.5, 5.0 \). For each value of \( k \) we perform a series of runs with frame shifts \( \Delta \tau_{\text{max}} = 2, 5, 8, 10, 15, 20, 25, 30, 50 \). In all cases the number of trial shots is \( 10^6 \). In addition we perform a simulation using uniform one-way

---

**Analysis of the path ensemble**

Home-written scripts analyzed the path sampling results to produce the path tree, the decorrelated path, the path length distribution, and the path density[15]. We construct the path density by choosing two order parameters (e.g. dipeptide minimum distance and solvent accessible surface area for the FF dimer) and binning each frame of each trajectory in the path ensemble to a 2D grid. Every path can only contribute to a specific bin once, even if visited multiple times. Note that accepted paths can occur multiple times in the ensemble, depending on whether the next trial moves have been rejected.

---

**6.3 Results and Discussion**

**6.3.1 2D Langevin model**

Using an initial path constructed from two trajectories initiated at the saddle point, and ending at state A and B, respectively, we conduct Transition Path Sampling simulations with the spring shooting algorithm. We perform a series of runs with different spring constants \( k = 0.01, 0.02, 0.05, 0.1, 1.0, 5.0 \). For each value of \( k \) we perform a series of runs with frame shifts \( \Delta \tau_{\text{max}} = 2, 5, 8, 10, 15, 20, 25, 30, 50 \). In all cases the number of trial shots is \( 10^6 \). In addition we perform a simulation using uniform one-way
The overall shooting acceptances are shown in Fig. 6.3.1. As expected, the acceptance ratio goes down when the frame shift is increased. Moreover, increasing the spring constant also decreases the acceptance ratio, since the biasing potential will reject shooting point trials more often. Still the overall acceptance never drops below 5%.

As both the backward and forward shooting trial moves contribute to the overall shooting acceptance, it is insightful to separate the contributions of both these sub-moves. In Fig. 6.3.2a the acceptance of both moves is plotted as a function of the instantaneous frame shift for a $k = 1.00$ and maximum frame shift $\Delta \tau_{\text{max}} = 10$. Here one clearly sees that the backward moves biases to positive frame shift, while the forward move biases to negative frame shift.

To show that the biasing potential in the spring shooting move has no influence on the resulting path ensembles, we plot the path length distribution of all simulations together in Fig. 6.3.2b. As a reference we also plotted the uniform one-way shooting result. Clearly, the distribution is the same for all settings, and only differs in accuracy due to sampling efficiency.

To establish the efficiency of the spring shooting algorithm, we need a measure of the decorrelation between paths. There is not a straightforward recipe for this. For the one way shooting algorithm a good measure is the amount of sampling that takes one from one decorrelated path to the other. This decorrelation path can be visualized using a path tree, see Fig. 6.3.3. This figure shows the path tree for a uniform one-way shooting algorithm. Starting from the top each horizontal line indicates an accepted
shooting attempt. Red lines indicate forward shots, green lines indicate backward shots. The thin black vertical lines indicate the shooting point location on the previous path. Each accepted new path thus consists of the newly formed green/red path, together with the complementary part of the previous path. The blue horizontal lines indicates the part of the paths that contribute to path decorrelation. Clearly, the longer the blue trajectory is, the better. Note also, that a vertical jump in the blue line means that paths that do not contribute to this decorrelation are skipped. In this case whole green and red parts are skipped when the vertical lines are long, indicating that the uniform shooting algorithm might be not so efficient.

The blue decorrelation trajectory is the part of the path ensemble that is the least changed during the path sampling. This trajectory effectively performs a random walk on top of the (diffusive) barrier. Clearly, the longer it is, the more it is able to explore path space, as subsequent paths are diffusing away from each other, eventually providing a representative sampling of the relevant path space. One might object that this choice of the decorrelation is rather arbitrary, and that it is more important to establish when the path ensemble becomes representative. Indeed, the decorrelation metric does not give information about how well the path ensemble is sampled. For instance, if the path ensemble oscillates between two narrow reaction channels, this will (most likely) not show up in the decorrelation measure. We stress, however, that this holds for any shooting algorithm. The aim of our analysis is to compare the efficiency of the different shooting algorithms. Besides, it is not straightforward to come up with a
system independent measure of the representative sampling of the path ensemble. We therefore believe that the least changed path is a reasonable system independent measure for decorrelation, to establish the difference in efficiency of the different sampling algorithms.

Taking the (horizontal) length of the blue path as a measure for decorrelation, we establish this length for all spring method simulations. The results are given in Fig. 6.3.4a. Here, clearly the decorrelation increases with the spring constant $k$. This is because the spring forces the shooting point into a direction that makes it more likely that the new trajectory contributes to the ‘blue’ decorrelation path. Note that increasing the spring constant above $k = 1.0$ does hardly alter the decorrelation curve, as then basically all trial moves in the ‘wrong’ direction will be rejected. For each spring constant the decorrelation curve behaves non-monotonically as a function of the maximum frame shift. While for small maximum frame shifts the decorrelation increases with $\Delta \tau_{\text{max}}$, the decorrelation reaches a maximum before starting to go down. This is because small maximum frames shifts are more likely to yield accepted new paths, but contribute only little, while large maximum frame shift contribute much more to decorrelation, but have a smaller acceptance probability. Clearly there is a $\Delta \tau_{\text{max}}$ at which the decorrelation is maximized. This maximum shifts to smaller $\Delta \tau_{\text{max}}$ for larger spring constant $k$. Note that the change in maximum is not that large: from $\Delta \tau_{\text{max}} \approx 20$ at $k = 0.01$ to $\Delta \tau_{\text{max}} \approx 10$ at $k = 5$. From this plot the largest decorrelation occurs for $k \geq 1$ and $\Delta \tau_{\text{max}} \approx 10$. Of course, this optimum will be system specific.
Figure 6.3.4: Efficiency of shooting algorithms. a) decorrelated path length as a function of maximum frame shift for different spring constants. b) CPU time in seconds. c) Efficiency as the decorrelation per CPU time. The curves correspond to spring constants $k = 0.01$ (black solid), 0.02 (red dotted), 0.05 (green dashed), 0.10 (blue long-dashed), 1.00 (yellow dot-dashed), 5.00 (brown dot-long dashed). Also shown are results for uniform shooting (grey horizontal dotted line), and for aimless shooting (purple dotted line).

Also shown in the plot is the decorrelation occurring in the uniform one way shooting algorithm (horizontal gray dotted line) and the aimless shooting method (horizontal purple dotted line). From this it seems that the aimless shooting always wins with respect to the spring shooting. This is because each successful aimless shot will contribute to the decorrelation, whereas in the spring shooting, there is still possibility to not contribute.

However, the path decorrelation alone is not the determining factor. Efficiency is also determined by CPU time. In Fig. 6.3.4b the CPU time for each simulation is plotted. Here the aimless shooting simulations are clearly the most expensive simulations, as for each shot both forward and backward paths need to be evaluated, which can both
Figure 6.3.5: Shooting point distribution as a function of the x-coordinate for the aimless shooting (black solid line), uniform shooting (red dotted line), spring shooting with constant k=0.05 (green dashed line) and k=1.00 (blue long dashed line). The aimless and spring algorithms are much more peaked, and closer to the saddle point compared to the uniform sampling.

be rather long if the shooting point is on the shallow side of the asymmetric barrier. The CPU time is already smaller for the uniform one-way shooting since then only half of the trajectory needs to be computed for every shot. The CPU time drastically decreases for the spring shooting move, becoming smaller with both $k$ and $\Delta \tau_{\text{max}}$. For the high values of $k \geq 1$ the CPU is only 25 % of the aimless shooting effort. This is caused by the early rejections, due to the spring bias, as well as due to the one-way nature of the shooting.

The most efficient algorithm shows the highest decorrelation per unit CPU time. In Fig. 6.3.4c we plot the ratio of the decorrelation and the CPU time. While the shape of the curves remain roughly as in Fig. 6.3.4a, now the aimless shooting efficiency drops substantially. For $k > 0.05$ the spring shooting algorithm already becomes more effective than aimless shooting (for this set up), and for the $k = 5$, $\Delta \tau_{\text{max}} = 10$ case, the efficiency is more than 3 times that of aimless shooting and more than 100 times as efficient as the uniform shooting move.

To understand the origin of the differences between the efficiency of the different shooting algorithms we compare the distributions of the x-value of shooting points in Fig. 6.3.5. Clearly the uniform one-way algorithm shoots mostly from areas far away from the transition state, thus leading to low decorrelation, and large non-contributing stretches of partial paths. The aimless shooting on the other hand focuses on the transition state. However, it is more expensive as it is a two-way shooting method, failing more often than a one-way shooting. The spring shooting algorithms fall in between these two extremes; closer to the aimless shooting distribution for $k = 1$ than
for $k = 0.05$. The conclusion is thus that one wants to be close to the aimless shooting distribution, which happens for relatively strong spring constant $k$, and a not too high maximum frame shift $\Delta \tau_{\text{max}}$.

An alternative way to identify the efficiency is to construct a correlation function of a property along the paths,

$$C(n) = \frac{\langle \delta g(0) \delta g(n) \rangle}{\langle \delta g(0) \delta g(0) \rangle}$$

where $n$ is the index of the shooting trial, the angular brackets denote average over the path ensemble taking each path as a time origin, and $\delta g(n) = g(x^n) - \bar{g}(x)$, with the overbar denoting the mean over all paths. This leads to an exponentially decaying correlation function. We extract the decay time of this correlation function by establishing the shooting trial index at which the correlation function reaches a value of $1/e$. For the correlation function of the maximum energy along the path [9, 31], the decay time is plotted in Fig. 6.3.6 for different $k$ values as function of the maximum frame shift $\Delta \tau_{\text{max}}$. Similar to the decorrelated path length, this function shows also an optimum for $k > 1$ and $\Delta \tau_{\text{max}} = 10$.

6.3.2 FF dimer association

In the next illustrative example we focus on the association dissociation reaction between two phenylalanine dipeptides. Because of their hydrophobicity these dipeptides
do attract each other, but also sometimes dissociate spontaneous, a state that is stabilized by entropy. We performed a 200 ns straightforward MD run, at 300 K and 1 atm. As shown in Fig. 6.3.7a in this time many transitions occurs between the bound and unbound state, indicating that the binding unbinding is not a rare event on that time scale. Still, while the relaxation time is on the order of a ns, most of the transitions are rather sharp and occur on a time scale of around 100 ps, thus characterizing them as activated processes on the ps timescale. Fig. 6.3.7b shows the free energy, obtained as the negative logarithm of the probability distribution along the minimum distance order parameter $r_{\text{min}}$. Indeed this plot shows a well defined minimum at the bound state around $r_{\text{min}} = 0.2nm$, and a very broad shallow minimum for the unbound state for $r_{\text{min}} > 0.6nm$, separated by a small barrier at $r_{\text{min}} = 0.4nm$ While the FE barrier is small, it is in principle asymmetric, with a steep binding force on the left side, and a shallow diffusive entropic force on the other side. The fact that the FE increases again
after the minimum of the unbound state U has been reached, is due to the finite size of the system. For larger system sizes the minimum would shift to larger distances. Moreover, the choice of the minimum distance as the order parameter also causes the free energy to increase already at smaller $r_{\text{min}}$ than if we would have chosen the center of mass distance. We define the stable states solely on the minimum distance $r_{\text{min}}$, with the bound state B by $r_{\text{min}} \leq 0.22$ and the unbound state U by $r_{\text{min}} \geq 1.2$.

Here our aim is to test the efficiency of the spring method. We perform a series of runs employing the spring shooting algorithm for different values of the spring constant $k$ as well as for different $\Delta \tau_{\text{max}}$. To initialize the TPS we extract a partial trajectory from the 200 ns straightforward MD run. This trajectory is plotted in Fig. 6.3.7c and connects state B ($r_{\text{min}} \leq 0.22$) and U ($r_{\text{min}} \geq 1.2$).

For the spring shooting move TPS, we performed simulations of multiple spring constants values $k=0.1, 1$ and $\Delta \tau_{\text{max}} = 10, 30, 50, 70, 90, 110, 150, 200, 250, 400$. As mentioned above for the 2D toy model, the spring shooting move should eventually sample the same equilibrium path ensemble, independent of its parameters. Therefore, properties of the path ensemble such as the distribution of path lengths and path densities should be the same independent of the employed shooting scheme. Fig. 6.3.8a illustrates the path length distribution remains unchanged upon changing the spring constant k and $\Delta \tau_{\text{max}}$’s, the equilibrium path distribution. The path length distribution is peaked at 200 ps and has an average path length of 340 ps. This distribution has a typical Poisson shape, with a long tail up to 2 ns.

At low spring $k=0.1$ and $\Delta \tau_{\text{max}} = 10$ the distribution deviates slightly from the
Figure 6.3.9: Path density plot of all atom dipeptide minimum distance versus solvent accessible surface for a) $k = 0.1, \Delta \tau_{\text{max}} = 10$, b) $k = 0.1, \Delta \tau_{\text{max}} = 100$. c) Difference between these path density plots.

As for the 2D toy model, the efficiency of the spring shooting method depends on the spring constant $k$ and the maximum allowed shift $\Delta \tau_{\text{max}}$. We define the efficiency of the spring shooting algorithm as the total decorrelated path length divided by the total number of computed MD steps (as a measure for the CPU time). For the FF dipeptide binding transition, the efficiency in terms of decorrelation length over the total number of MD steps of the TPS (including accepted and rejected forward or backward paths) is plotted in Fig. 6.3.8b. The efficiency shows a maximum at $\Delta \tau_{\text{max}} = 150$, corresponding to 750 ps path, since the frame saving frequency was 5 ps. Increasing the spring constant from $k = 0.1$ to $k = 1$ does not change the efficiency much, indicating that the efficiency is already saturated (see Fig. 6.3.4).

Fig. 6.3.8b shows that at small $\Delta \tau_{\text{max}}$ the efficiency is small, which is attributed to the reduced decorrelation of new trajectories, caused by the small allowed shooting point shifts. Upon increasing $\Delta \tau_{\text{max}}$ the efficiency increases as larger shooting point shifts are allowed in each trial move, and the chance of large contribution to decorrelation increases. The efficiency reaches its maximum around $\Delta \tau_{\text{max}} \approx 150 - 200$. In the limit of large $\Delta \tau_{\text{max}}$, the shooting efficiency decreases again, and becomes comparable to the uniform shooting.

In the spring shooting move, after accepting the newly generated shooting point (Eq. 6.9), the generated trajectory needs to end in the correct state in order for the trial move to be accepted within a reasonable amount of time, not exceeding a maximum
time frames $L_{\text{max}}$. As stated in the methods section, the maximum time was set to 10 ns, much longer that the path lengths occurring the sampled path ensemble.

The comparison between the efficiency uniform and spring shooting algorithm might be influenced by the fact that uniform shooting rejects paths based on the ratio of the path lengths, whereas the spring move does not. Therefore, when comparing the efficiency of the spring shooting with the uniform shooting, we report both the efficiency for the regular uniform shooting scheme (“uniform with Psel ratio”) as well as the efficiency for the uniform shooting where this rejection step is omitted (“uniform no Psel ratio”). The latter shooting scheme does not fulfill detailed balance and was implemented only to compare to the spring shooting algorithm. One would expect the uniform shooting without the path length ratio comparison to be much more efficient than the uniform shooting with the criterion as the latter rejects more trajectories ($P_{\text{acc}} = 40\%$) because of the length criterion compared to the former ($P_{\text{acc}} = 45\%$) leading to less decorrelation per CPU-time. However, we find that only 30\% of the rejected paths are rejected due to the length criterion. When the length criterion is off, the acceptance ratio drops only by 10\%. From this we deduce that the trajectories rejected due to the length criterion would result in a rejection anyway, by ending up in the wrong states. Indeed, the uniform shooting incorporating the length criterion has only a marginally smaller efficiency than the uniform shooting without the length criterion.

From the sampled path ensemble we can construct path density plots, shown in Fig. 6.3.9. The different spring shooting parameters do not alter the equilibrium properties of the transition. To show that we plotted the difference in distribution in Fig. 6.3.9c. The difference is of the order of 10\%. The bound state (B) is located below a minimum distance of 0.22 nm between the dipeptides. In the bound state there exist configurations of different hydration, i.e. ones with solvent accessible surface below 8.6 nm$^2$ (see Fig. 6.3.9b bottom left) and others with higher SASA. The unbound state (U) is characterized by a minimum dipeptide distance larger than 1.2 nm from each other. Here the solvent accessible surface area (SASA) is around 10.5 nm$^2$. The L shape of the path density plot indicates that when the dipeptides approach each other below a certain threshold the SASA starts to decrease, as expected for an association reaction. From the unbound state (top right area of the path density plot of Fig. 6.3.9) the peptides bind non-specifically through on-pathway intermediates with a SASA in the range of 9.5-10.5 nm$^2$, highlighted on the lower left side of the path densities Fig. 6.3.9. Thereafter, they reach a desolvated bound state (B) which has a low SASA ($< 8.6 \text{ nm}^2$ and is characterized by single phenyl-phenyl ring interactions, or both phenyl-phenyl interactions when peptides align in a parallel fashion, or by alignment of the peptides through hydrogen bond/hydrophilic interaction stabilization. The path density plots also suggest a dynamical bottleneck for the association located at $r_{\text{min}} = 0.4$ nm and a SASA of 10.5 nm$^2$. 

143
The efficiency, expressed as the decorrelation length over number of saved frames, of the uniform shooting move, the Gaussian bias shooting move and the transition. The efficiency, expressed as the decorrelation length over number of saved frames, of the uniform shooting move, the Gaussian bias shooting move and the spring shooting move, in terms of their sampling trees, shooting points and efficiency. A more detailed molecular interpretation of the TPS results will be discussed elsewhere.

In Fig. 6.3.10, we plot the path tree for the NC ⇔ NN transition. Although the sampling is not exhaustive, clearly the uniform shooting algorithm generates a higher abundance of accepted backward paths. This corresponds to a systematic selection of shooting points closer to the NC state (see Fig. 6.A.4). This is caused by the presence of an asymmetric barrier where the rate limiting step takes place for relatively low number of contacts, and the dimer spends long parts of the backward trajectory in near native states before finally locking into native bound state (NC).

The spring shooting move, ameliorates this effect and nearly equally creates both forward and backward paths. This is further corroborated by the shooting point as a function of accepted trials, where the selected shooting points are not driven by the natural entropy of the system, close to the native state, but are wandering to the middle of the transition. The efficiency, expressed as the decorrelation length over number of saved frames, of the uniform shooting move, the Gaussian bias shooting move and

**6.3.3 Dissociation of β-lactoglobulin dimer**

As for the FF dimer system the spring shooting algorithm is only marginally more efficient than the uniform shooting, in this subsection we consider a more complex protein-protein dissociation problem. We performed TPS on the first step in the dissociation process of the β-lactoglobulin dimer, the transition from the native state (NC) to an intermediate state (NN). The TPS simulation was initialized from a predefined dissociating path (see Methods).

Here, we discuss only the difference in sampling of the uniform shooting, the Gaussian bias shooting and the spring shooting, in terms of their sampling trees, shooting points and efficiency. A more detailed molecular interpretation of the TPS results will appear elsewhere.

In Fig. 6.3.10, we plot the path tree for the NC ⇔ NN transition. Although the sampling is not exhaustive, clearly the uniform shooting algorithm generates a higher abundance of accepted backward paths. This corresponds to a systematic selection of shooting points closer to the NC state (see Fig. 6.A.4). This is caused by the presence of an asymmetric barrier where the rate limiting step takes place for relatively low number of contacts, and the dimer spends long parts of the backward trajectory in near native states before finally locking into native bound state (NC).

The spring shooting move, ameliorates this effect and nearly equally creates both forward and backward paths. This is further corroborated by the shooting point as a function of accepted trials, where the selected shooting points are not driven by the natural entropy of the system, close to the native state, but are wandering to the middle of the transition. The efficiency, expressed as the decorrelation length over number of saved frames, of the uniform shooting move, the Gaussian bias shooting move and
the spring shooting move of $k=0.5$ and $\Delta \tau_{\text{max}}=70$ is $0.00066$, $0.01699$, and $0.0445$ respectively. Clearly the Gaussian bias shooting move and the spring shooting are two orders of magnitude more efficient than the uniform shooting. Comparing the Gaussian bias with the spring shooting of $k=0.5$ and $\Delta \tau_{\text{max}}=70$, we find that the spring shooting is about twice more efficient compared to the Gaussian bias shooting. Note that the choice of $k$ and $\Delta \tau_{\text{max}}$ might be not optimal, as no effort has been made to maximize efficiency for this transition.

6.4 Conclusion

We introduced a novel shooting algorithm for the transition path sampling framework that is aimed at efficiently sampling transitions in complex systems with diffusive behavior and asymmetric barriers. We showed that the new spring shooting is more efficient than uniform shooting, and also than the aimless algorithm. In addition, the new algorithms might also be useful to other path sampling methods such as Transition Interface Sampling [14], although there the efficiency gain is probably smaller because of the use of interfaces. Note however, that the efficiency analysis is not a general statement of superiority. There will be many cases in which other methods are more efficient. Nevertheless we believe that this new algorithm will help to improve the sampling of rare events in complex systems such as ligand binding and protein association reactions.
Appendix

6.A Initialization of the $\beta$-lactoglobulin dimer paths

In this section we provide additional information on the $\beta$-lactoglobulin dimer simulations. Table 6.A.1 lists the contacts found in the native dimer, together with their occupancy, based on the 200 ns trajectory. Here we defined a contact when the minimum heavy atom distance was less than 4.0 Å. The first 8 pairs are considered the native contact pairs. Fig. 6.A.1 plots the number native contact and the hydrogen bonds of the native dimer state in the 200 ns NPT trajectory.

Fig. 6.A.2 shows the minimum distance between the $\beta$-lactoglobulin proteins in a metadynamics simulation. Fig. 6.A.3 shows the number of native contacts for the ten simulations starting from frame 108 of the metadynamics run.
Table 6.A.1: Occupancy of residue contacts within the course of a 100 ns NPT MD.

<table>
<thead>
<tr>
<th>residue pair</th>
<th>occupancy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>150-146</td>
<td>100</td>
</tr>
<tr>
<td>148-148</td>
<td>100</td>
</tr>
<tr>
<td>146-150</td>
<td>100</td>
</tr>
<tr>
<td>148-147</td>
<td>99.98</td>
</tr>
<tr>
<td>147-148</td>
<td>99.84</td>
</tr>
<tr>
<td>149-146</td>
<td>99.56</td>
</tr>
<tr>
<td>146-149</td>
<td>99.44</td>
</tr>
<tr>
<td>33-33</td>
<td>90.74</td>
</tr>
<tr>
<td>150-29</td>
<td>79.46</td>
</tr>
<tr>
<td>29-150</td>
<td>78.66</td>
</tr>
<tr>
<td>137-141</td>
<td>75.96</td>
</tr>
<tr>
<td>151-29</td>
<td>70.66</td>
</tr>
<tr>
<td>141-137</td>
<td>69.78</td>
</tr>
</tbody>
</table>

Figure 6.A.2: Heavy atom minimum distance between the two proteins along a 20 ns meta-dynamics simulation.
Figure 6.A.3: The number of native contacts as a function of time for a series of ten 100 ns trajectories starting from the same configuration but with different momenta.

Figure 6.A.4: Number of native contacts of the shooting points of accepted reactive NC ⇔NN trajectories sampled by TPS, using spring shooting for $k = 0.5$ and $\Delta \tau_{max} = 70$ (red) and uniform shooting (blue).
Bibliography


Chapter 7

Elucidating the mechanism and role of solvent for \( \beta \)-lactoglobulin dimerization using Transition Path Sampling

Dimerization of proteins is a fundamental process in nature. While conceptually simple, the underlying association mechanism and the role of the solvent are poorly understood. Here we resolve these issues for the dimerization of \( \beta \)-lactoglobulin using transition path sampling of all atom molecular dynamics trajectories. The association process is found to occur via (at least) three distinct mechanisms: 1) aligned association to the native dimer interface, 2) misaligned association at non-native sites followed by hop towards the native state and 3) misaligned association followed by sliding of the protein towards the native state. We find that the native dimer state is stabilized by hydrogen bond bridging waters. Interestingly, water at the native interface can be found in two dynamical hydration states, a glassy one and a tetrahedral one. The crevice introduced upon binding increases the glassy populations as well as increases the average tetrahedrality of water, mainly at the vicinity of hydrophobic residues.

7.1 Introduction

The association of biomolecules, e.g. protein association or ligand binding to proteins, is important for biologically relevant processes, such as signalling, and inter-cellular communication, as well as for aggregation and self-assembly phenomena in (food) protein solutions. The kinetics and mechanisms of association, even on the dimer level, is poorly understood and varies with the nature of the proteins. Many proteins associate
into homodimers with a (experimentally determined) rate constant \( k_{on} \sim 10^6 M^{-1}s^{-1} \), which is three orders of magnitude slower than the diffusion limited Smoluchowski association rate constant predicts. This sounds reasonable because association of proteins into stable dimers requires very specific favorable interactions in the native (bound) dimer. However, an estimate of the association rate purely based on the geometric probability that a protein would randomly collide into the correct dimer state would be six orders of magnitude smaller than the diffusion limited rate [1]. Nonspecific interaction can enhance the rate [1, 2] even in the absence of strong hydrophobic and/or charged patches that could steer the assembly. The presence of strong hydrophobic or electrostatic interactions steering the association would result in a much faster association with a rate in the order of \( 10^9 M^{-1}s^{-1} \), such as in the barnase-barstar complex [3]. In the work by Ahmad et. al. it was shown that for the barnase-barstar complex, electrostatic steering drives the association and causes a very high association constant [3] \( K_a = 10^8 - 10^9 M^{-1}s^{-1} \), close to the diffusion limited association rate[1]. Nevertheless, the driving force and kinetic mechanism for protein association, even in the absence of such steering forces, is not well understood.

The first aim of this work is to investigate this kinetic mechanism for protein dimer formation using molecular simulations. Molecular simulation can assist in elucidating the kinetic mechanism of protein-association. However, as timescales of dissociation and association are on the order of microseconds to milliseconds, straightforward molecular dynamics (MD) using all atom force fields is unpractical. Transition Path Sampling (TPS) simulations enable focusing on the reactive association/dissociation trajectories, thus bypassing the long dwell times in the stable unbound and bound states. Here, we focus on the homo-dimerization of the \( \beta \)-lactoglobulin protein. \( \beta \)-lactoglobulin (\( \beta \)-lac) is a widely studied protein [4–7] due to its abundance in cow milk (0.2g/100ml), its globular shape and moderate size (18.4 kDa). Since \( \beta \)-lac is dimeric under ambient conditions it can act as a model system for improved understanding of protein-protein association, both from a fundamental as well as an applied perspective. We perform TPS simulations to elucidate the full association/dissociation mechanism of \( \beta \)-lac in solution.

A second aim is to understand the properties of the interfacial solvent in the association process. The protein hydration shell is very significant and essential to protein structure, dynamics and function. The hydration shell dynamics relates to biochemical processes such as protein folding, molecular recognition and enzyme function[8–10]. Both experimental [10, 11] and simulation findings [12–14] agree on that a protein perturbs the water dynamics in its hydration shell. Experimental techniques such as time-resolved fluorescence and NMR indicate a slowdown of water molecules close to the protein surface. NMR experiments [15], and molecular dynamics (MD) simulations[12, 14], show that most of the hydration shell water reorients 2-3 times...
slower compared to bulk water, due to excluded volume effects and to stronger hydrogen bonds [12, 14].

In the dimerization transition the interfacial water dynamics and structure might differ depending on the association state. Empirical models based only on the desolvation of polar and apolar areas underestimate the change in the heat capacity $\Delta C_p$ upon binding, compared to the isothermal titration calorimetry experiments [4]. This discrepancy was attributed to the neglected role of the buried waters at the contact interface, which cause a reduction in the heat capacity due to their solid-like character (loss of rotational and translational degrees of freedom). The enthalpy of interfacial water decreases upon association, further lowering the dimerization free energy. Including the entrapment of water and the enthalpic interactions with protein led to agreement of the model with the experimental heat capacity change upon association.

Nonspecific association of the proteins in a loosely bound dimeric state mediated by water molecules, allows reorientation of the proteins towards the correctly bound native structure [1]. A thermodynamic model by Ben-Naim [16] predicts the importance of water hydrogen bonded to both proteins – so-called bridging waters – in reducing the free energy of binding. Maximizing the number of bridging waters acts as a driving force for tighter binding. This finding was also supported by Ahmad et. al. [3] who showed an increased role for an adhesive water hydrogen bond network, stabilizing early intermediates and of increased water mediated interfacial connectivity upon binding.

Here, we analyze how the structure and dynamics of interfacial water varies at crucial points in the association trajectories obtained by Transition Path Sampling.

The chapter is organized as follows: in the next section we review the the simulation and analysis methodology. In Sec. 7.3 we discuss the results first for the association/dissociation mechanism, then for the water structure and dynamics. We end with concluding remarks.

### 7.2 Methods

#### 7.2.1 Molecular Dynamics

All molecular dynamics (MD) simulations as well as the system preparation were performed with Gromacs 4.6.7 package [17] using GPUs. Interactions were defined using the amber99sb-ildn [18] and TIP3P force fields [19]. The β-lactoglobulin (β-lac) dimer system was taken from the protein data bank (PDB:2AKQ) and put in a dodecahedral box and energy minimized using the conjugate gradient method. After solvation of the box with 20787 water molecules and a second energy minimization, we performed a
10 ps NPT short equilibration of water under ambient conditions with the protein position restrained. The solvated system was equilibrated for 1 ns in ambient conditions in the NPT ensemble and thereafter was subjected to a long 200 ns NPT simulation. All bonds were constrained with the Lincs algorithm. A cutoff of 1 nm was used for the non-bonded interactions. The Particle Mesh Ewald method treated the electrostatic interactions, with a Fourier spacing of 0.12 nm and a 1 nm cutoff for the short range electrostatic interactions. Neighbour lists were updated every 10 fs with a cutoff of 1 nm and the time step was 2 fs [18]. Newton’s equations of motion were integrated with the leap-frog algorithm. In the NPT simulations the v-rescale thermostat[20] with a coupling time constant of 0.2 ps controlled the temperature, while the Parrinello-Rahman barostat[21] with a coupling time constant of 1.0 ps kept the pressure constant.

7.2.2 Transition path sampling

Spring shooting algorithm

Transition Path Sampling [22, 23] (TPS) harvests an ensemble of rare trajectories that lead over a high free energy barrier, connecting two predefined stable states. Starting from an initial reactive path TPS performs a random walk in trajectory space by selecting a time frame, changing the momenta slightly and shooting off a new trial trajectory forward and backward in time by integrating the equations of motion. Acceptance or rejection of the trial trajectory is done according to the Metropolis rule [22, 23] which for the standard two way shooting move with fixed path length just checks if the trial path connects the two stable states. If not the trial path is rejected.

The more efficient one-way flexible shooting algorithm [23, 24] samples the minimal length pathways between stable states and has been previously used in other protein systems [25, 26]. The one-way shooting method has several drawbacks. First, it requires more shots to decorrelate paths (although not more computer time). Second, it suffers in efficiency for asymmetric barriers, which occur, for instance, when the system on one side of the main barrier is trapped in an intermediate state, while it can easily reach the stable state on the other side. This means the paths on the trapped side become much longer. When uniform one-way shooting is used, this asymmetry leads to many more shooting attempts on one side of the barrier with respect to the other. Due to the asymmetry, most shooting points are chosen on the trapped side, increasing the inefficiency.

The spring shooting algorithm is especially developed for use with the one-way algorithm[27]. It only differs in the way the shooting point is selected. Instead of uniform random selection, the spring shooting shifts the shooting point index with respect to the last successful shooting point, not in a symmetric but in an asymmetric
way according an acceptance criterion

\[ P_{acc}^{sp}[\tau \to \tau'] = \min \left[ 1, \frac{\exp(sk\tau')}{\exp(sk\tau)} \right] = \min[1, e^{sk\Delta\tau}], \quad (7.1) \]

where \( \Delta\tau = \tau' - \tau \) is the number of shifted frames from the previous shooting point \( \tau \), \( k \) denotes a force constant determining the magnitude of the bias, and \( s \in \{-1, 1\} \) is determined by the direction of shooting i.e. \( s = -1 \) for forward shooting, and \( s = 1 \) for backward shooting. The spring shooting algorithm thus treats the forward and backward shooting move as different types of moves. As a large \( \Delta\tau \) either yields an exponentially small acceptance ratio or is likely to produce a failed shot, in practice, we limit the choice of \( \Delta\tau \) between the interval \([-\Delta\tau_{\text{max}}, \Delta\tau_{\text{max}}]\), analogous to the maximum allowed displacement in a regular MC translational move. When the trial shooting point falls outside the current path the acceptance probability becomes zero, and the move is rejected. The remainder of the shooting move is identical to the uniform one-way shooting algorithm.

For completeness we summarize the entire algorithm here[27]:

1. Select with equal probability a forward or a backward move. Set \( s = -1 \) in case of forward move, \( s = 1 \) in case of a backward move.

2. Select a uniform shift \( \Delta\tau \) in shooting point from the interval \([-\Delta\tau_{\text{max}}, \Delta\tau_{\text{max}}]\). Add this shift to the previous shooting point frame index \( \tau' = \tau + \Delta\tau \). Reject the entire move if the index is outside the current path.

3. Accept the trial shooting point \( \tau' \) according to eq. 7.1, otherwise reject the entire move.

4. Create a partial path (forward or backward) employing molecular dynamics with a stochastic thermostat. Halt the path when a stable state is reached. Reject the move when the total trial path length exceeds a maximum length \( L_{\text{max}} \). Glue the partial path to the complementary existing path, and accept according to the standard criterion \( P_{acc}[x^{(o)} \to x^{(n)}] = h_A(x_0^{(n)})h_B(x_L^{(n)}) \).

The advantage of this approach is that unfavorable shooting points are discarded without extra cost. Pathways are decorrelated as much as possible, without wasting time creating partial paths that do not contribute to the decorrelation. Note also that the algorithm rejects trial paths which become longer than \( L_{\text{max}} \), which is set to prevent memory or storage problems, or as an indication that the path generation went awry, e.g. became trapped in an long-lived intermediate state.
Table 7.2.1: State definitions for the native state (NC) and the unbound state (U).

<table>
<thead>
<tr>
<th>State</th>
<th>Native contacts</th>
<th>Native hydrogen bonds</th>
<th>$r_{\text{min}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native state (NC)</td>
<td>$\geq 8$</td>
<td>$\geq 4$</td>
<td>-</td>
</tr>
<tr>
<td>Unbound state (U)</td>
<td>0</td>
<td>0</td>
<td>$&gt; 1 \text{ nm}$</td>
</tr>
</tbody>
</table>

Defining the stable states / Creating the initial path

As shown in Fig. 6.A.1, in the course of the 200 ns MD in the NPT ensemble in ambient conditions, the $\beta$-lactoglobulin dimer remained in its native bound state. The native contacts were identified as those residue pairs that stayed within a minimum heavy atom distance of 0.4 nm for at least 90% in the 200 ns NPT trajectory. These native contact pairs are listed in Table 6.A.1. Only 8 residue pairs are shown to fulfil this criterion (150-146, 148-148,146-150,148-147,147-148,149-146,146-149,33-33). All but the pair 33-33 are between beta sheets, and the 33-33 is a contact between the AB-loops of the protein [28]. These 8 residue pairs, as well as four native hydrogen bonds (between backbone NH and CO of residues 146-150,148-148,150-146), characterized by Sakurai et. al. [28], define the stable native contact state (NC). Since we are interested in the full dissociation transition, we define the unbound state (U) as the one where the minimum distance between monomer heavy atoms is greater than 1 nm ($r_{\text{min}} > 1 \text{ nm}$).

In order to initialize the TPS simulation we need a single path undergoing the dissociation. Therefore we performed Metadynamics using the PLUMED package [29] with the above MD settings. The collective variable for Metadynamics was the center of mass distance of the protein $C\alpha$ atoms, using a Gaussian hill with height 0.4 kJ/mol, and width of $\sigma= 0.1$ kJ/mol which are deposited every 2 ps. Frames were saved every 20 ps. The resulting Metadynamics trajectory indeed undergoes a dissociation event. For completeness we give the heavy atom minimum distance of the proteins along the trajectory in Fig. 6.A.2. From this biased dimer dissociation trajectory we launched from a particular configuration (frame 108, at 2140 ps along the trajectory) a series of 10 trial trajectories with random velocities. These trajectories were performed at ambient conditions in the NPT ensemble, using the settings of Sec. 7.2.1. The particular frame 108 turned out to have a reasonable chance to commit to the native state (NC), as well as escape the native state and to end in near-native states ($I_1$) (Fig. 6.A.3). However, since we are interested in the full dissociation transition from the native state to the unbound state, we tried to create an unbiased trajectory that connects the bound with the unbound state. To do so, we performed 10 NPT trajectories from the same frame with different velocities at 330 K. Increasing the temperature decreases the life-
time of the near-native bound dimer. Indeed, as shown in Fig. 7.A.1a, one path reaches the unbound state U, and one (Fig. 7.A.1b) reaches the native state (NC). These trajectories were glued together to yield the desired initial unbiased path. In order to confirm that at 330 K, the proteins are stable and do not undergo serious conformational changes, we plot the protein $C_\alpha$ RMSD from frame 108, of all committor trajectories performed both at 300 K and 330 K (see Fig. 7.A.2). No large conformational change takes place upon increasing the temperature. To further establish this, we calculated the Ramachandran plot (see Fig. 7.A.3) for the protein at 300 K and 330 K and found no significant structural differences between the two temperatures. Therefore, we will perform the TPS simulations at 330 K in order to keep the path length reasonable.

**TPS simulation settings**

We performed TPS simulations of the native state to non native state (NC $\rightarrow$ U) at $T = 330K$, $P = 1$ atm in the NPT ensemble using home written scripts encoding the spring shooting scheme. The maximum path length for this transition was set to $L_{\text{max}} = 7000$ and frames were saved every 10 ps. The spring shooting move parameters were set to $k = 5$ and $\Delta \tau_{\text{max}} = 200$. The spring constant ensures the shooting points remain close to the top of the very asymmetric barrier of $\beta$-lac dissociation [30]. The $\Delta \tau_{\text{max}}$ was chosen small compared to the maximum path length allowed (3%) so that the shooting point rejection as well as the whole trial move rejection was kept to a minimum.

We also investigated the B $\rightarrow$ U transition, where the bound state B is defined as when the protein protein interfacial area is $> 2 (nm^2)$. These simulations will be discussed elsewhere.

**Analysis of the path ensemble**

Home-written scripts analyzed the path sampling results to produce the path tree, the decorrelated path (LCP), the path length distribution, and the path density[26]. We construct the path density by choosing two order parameters (e.g. protein-protein minimum distance vs native patch vector angle) and binning each frame of each trajectory in the path ensemble to a 2D grid. Every path can only contribute to a specific bin once, even if visited multiple times. Note that accepted paths can occur multiple times in the ensemble, depending on whether the next trial moves have been rejected.

Another way of representing the transition path ensemble is the path tree (see Fig. 7.3.1). Horizontal lines depict backward and forward partial paths in red and green respectively, whereas the vertical lines correspond to the shooting points. The first horizontal blue path corresponds to the initial path. The least changed path (LCP), consisting of the stretches between successive alternating forward/backward shooting points
can be extracted and used as an approximation for the transition state ensemble[27].

Another diagnostic tool relating to the ruggedness of the barrier or to the different underlying channels of the transition is the path length distribution, where each accepted path of different length $L$ is histogrammed according to its weight in the path ensemble.

Since the protein orientation degrees of freedom might be important during the association/dissociation transition, we calculate the relative orientation of the two protein, characterized by an angle $\phi$. This angle is the sum of $\phi_1$ and $\phi_2$ angles, shown in Fig. 7.2.1a, where $\phi_1$ is formed between vectors $AB$ and $AA1$, and $\phi_2$ is formed between vectors $BA$ and $BB1$. Here, $A1$ and $B1$ are the centers of mass of the native contact residue patches (33,146,149,147,148,150) of protein A and B, respectively.

Water plays an important role in association. In order to address the solvent degrees of freedom, for each configuration we count the number of waters residing in a tube formed between the two proteins. The tube’s base centers are defined by the center of mass of each protein, a radius $r = 1.4$ nm and length $L$ being the center of mass distance between the two proteins.

### 7.2.3 Analysis of the water dynamics

**Launching NVE trajectories in the NC, I₁ and U states**

To obtain insight in the role of the solvent we analyze the structural and orientational dynamics [27] (see also chapters 3 and 4) of water in the dissociation transition. As the TPS simulations are done at 330 K, we first analyze the water structure and dynamics at
this temperature as follows. From a decorrelated reactive path in TPS ensemble (path #6) we initiated three sets of ten short 1 ns NVE runs using three different selected frames belonging to the native state NC, near native state I\(_1\) (1 < native contacts < 5) and unbound state U respectively. In the NVE simulations, in order to prevent energy drift we used a switching function for the non-bonded interactions from 0.8 to 1.0 nm. The pair lists were updated every 5 fs with a cutoff of 1.2 nm, and the time step was 1 fs. The frequency of the energy calculation was 10 fs, and the frames were saved every 100 fs in order to obtain sufficient data for analysis of the water dynamics. In order to identify whether water structure and dynamics changes significantly from 330 K to 300K, we repeat the same analysis at 300 K. Therefore, we first launch three sets of 1ns NPT runs at 300 K for NC, I\(_1\) and U respectively, with the protein position restrained, before launching ten 1 ns short NVE runs.

**Water reorientation dynamics**

Reorientation dynamics can be represented using the time correlation function [31, 32]

\[
C_2(t) = \frac{2}{5} \langle P_2[u(0) \cdot u(t)] \rangle, \quad (7.2)
\]

where \(P_2\) denotes the second Legendre polynomial, \(u\) is the unit vector characterizing the orientation of an OH group in a given frame and the angular brackets denote an ensemble average over all water molecules and all time origins. This correlation function can easily be related to anisotropy curves obtained from polarization-resolved femtosecond infrared spectroscopy and to orientation relaxation times from magnetic relaxation techniques [31].

From MD simulations we compute the reorientation dynamics for each individual molecule. Particularly interesting is the anisotropy decay of water molecules hydrating different protein sites. The reorientation dynamics of individual water molecules was investigated by following the dynamics of both OH bond vectors of each molecule \(j\). For each bond unit vector \(u_{jk}(t)\), we computed the time correlation function

\[
c_j^m(t) = \frac{1}{5\ell} \sum_{t'=t_0^m}^{t_0^m+\ell \Delta t} \sum_{k=1}^{2} P_2[u_{jk}(t') \cdot u_{jk}(t'+t)]], \quad (7.3)
\]

where the sum over \(k = 1, 2\) refers to the two OH bond vectors, and the time correlation is summed over the \(m\)th time interval \(t_0^m < t' < t_0^m + \ell \Delta t\) that the molecule is within the hydration layer of the protein (defined as the water oxygen being within 4.4 from the protein heavy atoms). We used a buffer time of 2 ps to avoid counting fast non-essential excursions inside and outside the hydration layer [33]. Note that \(\ell \Delta t\) is also
known as the residence time of the hydration water. For each curve $c^m_j(t)$ we extracted the reorientation decay time $\tau^m_j$ by fitting $c^m_j(t)$ to a single exponential fit in the interval $0 < t < 10$ ps. Each estimate $m$ was viewed as a separate measurement, since the water molecule can change its dynamics when leaving or entering the hydration layer. For each water molecule $j$ and for each $m$th interval of length $\ell$ the decay time $\tau^m_j$ was histogrammed with a weight $\ell$, leading to a probability distribution of decay times. The weight $\ell$ follows from the fact that the correlation in eq. 7.3 occurs $\ell$ times in eq. 7.2. The average decay time $\tau$ from these histograms thus should be close to the overall decay time of eq.7.2. The decay times of the individual water molecules allow us to establish the relation between the water structure and dynamics. We therefore divide the waters into several categories.

For each water molecule $j$ and for each $m$th interval $\ell$ the decay time $\tau^m_j$ was histogrammed with a weight $\ell f_\#AA$ in the residue $\#AA$ reorientation distribution. Here $f_\#AA$ refers to the fraction of frames in the $m$th interval $\ell$ in which the oxygen of water $j$ is within 4.4 Å of any of the heavy atoms in residue $\#AA$.

Water-Water angular distribution function

Sharp and coworkers [34–36] introduced a method to identify the tetrahedral structuring of water around amino acids based on the distribution of a water-water angle. In a given frame one computes for each amino acid the minimum water-water OOH angle $\theta$ (see Fig. 7.2.2 for a graphical definition) for all water-water pairs within 3.5 Å from each other and solvating that amino acid. The distribution $P(\theta)$ of these angles takes on a bimodal distribution with a minimum at 30°, distinguishing between tetrahedral water population (angles lower than 30°) and a perturbed H-bond network, mostly occurring around hydrophilic groups (angles higher than 30°). The tetrahedral structure parameter $S$ is defined as the integral of $P(\theta)$ up to $\theta = 30^\circ$.

Water around hydrophobic groups has a larger $S$ due to smaller H-bond angles $\theta$, inducing stronger water-water bonds, with larger energy fluctuations and therefore a positive heat capacity of the solvating water. In contrast, the introduction of a hy-
hydrophilic group around water strains the water-water H-bond angle and shifts the angle distribution to higher values, and hence a lower $S$, thus decreasing the water-water bond energy and fluctuations which decreases the heat capacity of solvation [34–36].

Throughout the text we will associate tetrahedral/ice-like/structured water with a large $S$ value (high tetrahedral water population) and unstructured water with a low $S$ value (low tetrahedral water population). Unstructured water coinciding with slow reorientation dynamics (as characterized by $\tau$) will be labeled as glass-like water. Residues solvated by ice-like tetrahedral water will be labeled as tetrahedral structure makers and the ones by glass-like waters, tetrahedral structure breakers.

**Structure parameter - water reorientation time correlation**

As mentioned above, each water $i$ surrounding one amino acid for an $m$th interval of length $\ell$, has a reorientation decay time $\tau_{i,m}$ associated with it. Additionally in this $m$th interval we calculated the distribution $P(\theta_{ij})$ of the angle $\theta_{ij}$ of waters $j$ within 3.5 of $i$ and hydrating the same amino acid. From the normalized distribution $P(\theta_{i}) = \sum_j P(\theta_{ij}) / \sum_{i,j} P(\theta_{ij})$ we obtain the structural parameter $S_{i,m}$ for each water $i$ hydrating a particular amino acid, by integrating from $0^\circ$ to $30^\circ$. We bin the pair $\tau_{i,m}$ and $S_{i,m}$ for each residue in a 2D histogram, in order to investigate possible correlation between water tetrahedral structuring and reorientation dynamics.

**Hydrogen bond bridge survival correlation function**

The hydrogen bond bridge correlation function (eq. 7.4) is a correlation function that traces the decay time of a hydrogen bond bridge between two intermolecular protein residues.

$$C_{Bridging}(\tau) = \frac{1}{C_{Norm}} \int \sum_{i,j}^{N_{res}} \sum_{N=1}^{N_{wat}} \left( \mathbb{1}_{N,i,j}^{Bridging}(t) \cdot \mathbb{1}_{N,i,j}^{Bridging}(t + \tau) \right) dt \quad (7.4)$$

Where $\tau$ is time, $C_{Norm}$ is a normalisation constant, such that $C_{Bridging}(0) = 1$. $i$ and $j$ are running over the residue number of proteins $A$ and $B$ respectively, $N_{res}$ is the total number of residues per protein and $N_{wat}$ is the total amount of waters in the simulation box. $\mathbb{1}_{N,i,j}^{Bridging}(t)$ is an indicator function at time $t$. The indicator function becomes one if water $N$ is hydrogen bond bridging residues $i$ and $j$ and zero otherwise.
7.3 Results-Discussion

7.3.1 Transition Path Sampling

TPS of the full NC ⇌ U transition

We performed several TPS runs at 330 K employing the spring shooting algorithm for the full association/dissociation mechanism between the unbound state and the native bound dimer state. The stable state definitions are given in Table 7.2.1. In total we performed 318 shooting trial moves, of which 20% was accepted. The average path length is 32.5 ns long and the path length distribution is depicted in Fig. 7.3.2. The path length distribution shows a significant population of paths having relatively short lengths (10-25 ns), and population of paths which have longer length (55-70 ns). Interestingly, the longer path population involves pathways belonging to the slide/misaligned and.
hop/misaligned mechanisms which will be described below.

To estimate the decorrelation between the trajectories in the path ensemble we plot the path tree in Fig. 7.3.1. The red horizontal lines correspond to forward partial paths, the green to backward partial paths. Grey vertical lines correspond to the relative positions of the shooting points and the cyan part corresponds to the least changed path (LCP), which roughly samples the barrier region. The dark blue horizontal line corresponds to the initial path of each of the three TPS simulations. The trees shown in Fig. 7.3.1b,c correspond to two independent TPS runs with an initial path #11 taken from the TPS ensemble of Fig. 7.3.1a. The path tree shows a variety of reactive path lengths.

A striking aspect of the trees is the presence of sequences of backward paths without any accepted forward path in between (e.g in Fig. 7.3.1a) from path 14 to 19, or from path 23 to 34) or the blocks of accepted forward paths without any accepted backward paths in between (Fig. 7.3.1a from path 8 to 10). To explain this, we plot the

<table>
<thead>
<tr>
<th></th>
<th>TPE(a)</th>
<th>TPE(b)</th>
<th>TPE(c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aligned</td>
<td>15-19</td>
<td>16-17</td>
<td>21-24</td>
</tr>
<tr>
<td>Slide misaligned</td>
<td>6-14,20-33,40</td>
<td>12-15,18-22</td>
<td>12-13,15-20</td>
</tr>
<tr>
<td>Hop misaligned</td>
<td>0-5,34-39</td>
<td>-</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 7.3.1: Pathways of each path ensemble belonging to different mechanisms. Note that the TPE b) and c) do not contain any paths with index below 11 because these ensembles were initiated from path #11 in paths ensemble a).
accepted shooting point (Fig. 7.3.2) shifts of the path ensembles shown in Fig. 7.3.1a. Using a high spring constant $k = 5$ biases the generation probability of the shooting points towards opposite sides. For forward trajectories nearly all accepted shooting points are frames with an index lower than the previous shooting point, i.e. closer to state NC. Analyzing the above mentioned blocks of backward paths (e.g. accepted paths 14 to 19), a forward trial path starting from a shooting point close to state NC is likely to become trapped at long lived intermediates states, and thus rejected. To avoid these traps, the shooting point has to shift to frames with higher index, i.e. toward the unbound state, which, due to the spring penalty, only backward shooting allows. Therefore, the consecutive series of backward paths can be seen as an effort of the TPS shooting move to shift the shooting points away from the many kinetic traps around the native state. Since the consecutive series of forward and backward partial paths coincides with the paths belonging to each of the three mechanisms (see Tab. 7.3.1), a second argument behind this monotonic shift of shooting points to the one or the other directions is that the shooting points are shifting from one transition to the other.
Mechanism of full association/dissociation

The paths sampled by TPS fall into three categories, representing three qualitatively distinct mechanisms of association (indicated in Fig. 7.3.3): a) *aligned association/dissociation* mechanism I (NC ⇔ I₁ ⇔ U), 2) an indirect, *misaligned/hopping association/dissociation* mechanism II (NC ⇔ I₁ ⇔ I₂ ⇔ I₃ ⇔ U), where the proteins first bind non-specifically, and then hop to a near native region before committing into the native state and 3) an indirect, *misaligned/sliding association/dissociation* mechanism (NC ⇔ I₁ ⇔ I₃ ⇔ U) in which the proteins bind non-specifically before rotating and sliding into the native state. Tab. 7.3.1 indicates which paths fall into each of the three categories. Note that switches between the different mechanisms frequently occur. To further identify these mechanisms, we plot the path densities of several order parameters (Fig. 7.3.4a,b,c,d) for each mechanism I, II, and III.

**Mechanism I**

In the *aligned association/dissociation* mechanism (ϕ < 50°) (Fig. 7.3.4a) the protein docks to a near native area (configuration I₁) of high alignment, and increases its dry surface area towards the native state (Fig. 7.3.4b). Interestingly, the water content in a protein-protein tube decreases sequentially as protein rolls (decreases the ϕ) towards the native state (NC) although there exist few pathways indicating the water content drastically decreases while protein docks without rotating (Fig. 7.3.4c). The hydrogen bond bridging waters are present in the near native state, adding an extra barrier to the association of the proteins (Fig. 7.3.4d).

**Mechanism II**

In the *misaligned/hopping association/dissociation* mechanism proteins first bind non specifically and misaligned (ϕ > 50°, configuration I₃) followed by a hop (configuration I₂) to a near native aligned configurations (I₁) before binding to the native state (NC). Misaligned configurations I₃ have a small dry area 1-2 nm², while the hop configuration I₂ has a zero dry area as proteins are dissociated before they bind and increase the dry area again in the near native configuration I₂ (see Fig. 7.3.4b). Note the important role of hydrogen bond bridging waters in stabilizing the misaligned configuration I₃ Fig. 7.3.4d.

**Mechanism III**

In the *misaligned/sliding association/dissociation* mechanism, proteins first bind at a misaligned configuration (I₃, ϕ > 50°) and then slide towards the near native con-
Figure 7.3.4: Path density plots for path ensembles according to the underlying mechanisms I (left column), II (center column), III (right column) as a function of several order parameters. On the y-axis is always \( \phi \). On the x-axis is a) the protein protein minimum distance, b) the dry contact area between the proteins, c) the number of waters in a tube of radius 1.4 nm between the proteins and d) the number of hydrogen bond bridging waters. The red and black rectangles highlight the NC and U state respectively.
Figure 7.3.5: Path density plots of $\phi$ as a function of a) protein-protein minimum distance, b) native contacts c) number of hydrogen bond bridging waters for the full transition path ensemble (full TPE) and least changed path ensemble (LCP) respectively.

figuration ($I_1$) before binding to NC Fig. 7.3.4a. Interestingly, along the sliding and aligning process (see Fig. 7.3.4b), the dry contact area is preserved (2-5 nm$^2$) with the hydrogen bond bridging waters being present along this alignment Fig. 7.3.4d. Note that the water in the tube between the two proteins is sequentially removed as proteins slide and roll towards the native state Fig. 7.3.4c.

**Transition State Ensemble**

In order to get further insight into the rate limiting steps of the transition, we plot the path density (Fig. 7.3.5) of several order parameters for both the full transition
Figure 7.3.6: Structure of an on pathway long-lived intermediate. In blue are highlighted the long-lived contacts 146-150, 33-40, 29-150.

path ensemble (full TPE) as well as for the least changed path ensemble (LCP). As discussed in previous work [30], the LCP ensemble approximately samples the rough barrier region serving as a proxy of the transition state ensemble (TSE).

In the LCP plot of Fig. 7.3.5a, it is evident that two transition state ensembles exist: one at $\phi < 40$ (TSE$_1$) and a second one at $\phi > 50$ (TSE$_2$). In mechanism I the paths pass only through TSE$_1$ as proteins associate in an aligned fashion, therefore posing mechanism I as a one step association dock$_1 \rightarrow$ lock. Mechanisms II and III exhibit first a misaligned association, passing the barrier at TSE$_2$ either by hopping or sliding towards the second barrier at TSE$_1$. Therefore mechanisms II and III are examples of a two step association process dock$_2 \rightarrow$ dock$_1 \rightarrow$ lock.

For paths in mechanism I, which pass only through TSE$_1$, the LCP plot Fig. 7.3.5b shows that upon binding paths have to pass through TSE$_1$, where proteins have partially formed native contacts ($1 < N_C < 5$) and the bottleneck is locally rearranging and aligning towards the formation of their native contacts. In mechanisms II and III, the proteins first dock at a misaligned state (I$_3$) with nearly all the native contacts broken and the bottleneck of TSE$_2$ is rearranging (by hopping or sliding) towards near native state I$_1$. Then proteins have to further locally rearrange to form all their native contacts. Because the rearrangement both in TSE$_1$ and TSE$_2$ is a bottleneck, we quantify these ensembles by calculating the contact pairs. All reactive paths in the path ensembles in Fig. 7.3.1 exhibit the transient but long-lived (occupancy $> 10$ns and heavy atom distance $< 0.4$ nm) ASP33-ARG40 double salt bridge contact between the carbonyl groups of ASP33 of one protein and the amide groups of ARG40 of the other protein. Other long-lived intermolecular contacts are preserved in a large number of paths, although not throughout the entire TPE, and are listed in Tab. 7.A.1. We highlight a
configuration containing the three most occurring contacts (ASP33-ARG40, HIS146-SER150, ILE29-SER150) in the path ensembles in Fig. 7.3.6. Experimental evidence by Sakurai et. al. [5] showed that the ASP33-ARG40 interaction occurring in every path in our TPE is important for the dimer formation process. Mutating any of these residues to an oppositely charged amino-acid drastically reduced the association constant. Also residues 146, 148 and 150 are important for dimerization, as follows from the decrease of the association constant when these residues are mutated to proline, thus breaking the β-sheet structure at the native binding site. This emphasizes the importance of interactions at the near native state in the formation of the dimer.

Notwithstanding the role of salt bridges, hydrophilic and minorly hydrophobic interactions, protein-water interactions play a big role. The LCP of Fig. 7.3.5c manifests that both in TSE₁ and TSE₂ formation of the hydrogen bond bridging waters between the proteins are dynamical bottlenecks towards the full association process. In mechanisms III and II, proteins bind non-specifically and misaligned (I₃) and either slide towards the aligned near native state (I₁) towards maximizing their bridging waters or lose these interactions again while hopping to the near native state (I₁) which has again abundant bridging waters. At the near-native state further formation of these bridging waters is a bottleneck (TSE₁) towards the native state, which has the maximum of bridging waters.

This observation is in agreement with the prediction of Ben-Naim and Northrup [1, 16] that water-mediated interactions drive or characterize the hydrophilic association. Note also that the number of bridging water reaches a maximum, before reaching the native state.

7.3.2 Hydration states of water

Since water plays an important role in association [1, 3, 16] finding the different hydration states of water at the native, near native and unbound state is pivotal for characterizing the bridging process. As mentioned in sec. 7.2.3, we analyse NVE trajectories of 330 K and 300 K, both in terms of structural parameter S and reorientational decay times τ. At 330 K, Fig. 7.3.7a shows that in the native state water around the native contact region – comprising residues 33, 146,147,148,149,150 — exhibits two hydration populations: a fast reorienting and more tetrahedrally structured (S > 0.4 and τ < 4) water population, that we label tetrahedral and a slowly reorienting, less structured water population that we label glassy (S < 0.4 and τ > 4). Note that the bulk water structure parameter is S= 0.38 at 330 K. Hydrophobic amino-acids such as ILE147 and LEU149 exhibit similar behaviour and show both populations (see Fig. 7.3.8a).

In contrast, some amino-acids exhibit only the glassy hydration state, such as the charged residue ASP33 (see Fig. 7.3.8d). The two hydrophobic amino acids (ILE147
Figure 7.3.7: Two dimensional histograms of the reorientation time \( \tau \) versus structural parameter \( S \) or water molecules residing at the native interface from NVE simulations performed in the native state, near native state, unbound state at 330 K (a,b,c) and at 300 K (d,e,f).

and LEU149) are solvated by both a tetrahedral and glassy water population because they are influenced by the hydration state of neighbouring charged amino acids 146HIS and 148ARG. Native contacts are supported by water mediated interactions: bridging waters hydrogen bonded to both proteins. The hydrogen bond bridge survival correlation function in Fig. 7.3.9 decays slower for waters in the native state compared to waters in the near-native state at both temperatures. This finding corroborates with the maximization of the native contacts towards the native state found in the TPE Fig. 7.3.5c.

In the unbound state (see Fig. 7.3.7c), the glassy water population (with \( \tau > 10 \) ps) decreases. Fig. 7.3.10 shows the residence time distributions of water belonging to each population, for the three states NC, I_1 and U. The residence time of the glassy water population in the native state is longer compared to that of the unbound state whereas the tetrahedral water residence times are rather insensitive, suggesting that the crevice formed upon association restricts waters to the glassy population formed
around charged amino acids. For example, the charged amino acid ASP33 exhibits an increase in the glassy population upon association (see Fig. 7.3.8d,e,f).

Upon association the tetrahedral population increase, as $S$ shifts from 0.4 to 0.5 (Fig. 7.3.7). Since the tetrahedral water population lives near the hydrophobic groups, we computed the $S$ – $\tau$ correlation plot for water around the hydrophobic amino acids ILE147 and LEU149 (Fig. 7.3.8a,b,c). Although in the unbound state these hydrophobic amino acids do not exhibit much tetrahedral water structure, in the native interface where these amino acids are opposite to each other, they exhibit a tetrahedral population. As the TPS paths were performed at 330 K we also did the analysis for this elevated temperature. To check whether lowering the temperature to room temperature would severely alter the results, we repeated the analysis for 300 K. Our findings indicate that the hydration states of water do not change qualitatively upon lowering the temperature from 330 K to 300 K.
Figure 7.3.9: Hydrogen bond bridge survival correlation function between any amino acid pair of the interface state for the native state (red) and the near native state (green) at a) 330 K and b) 300 K.

Figure 7.3.10: Residence time probability distribution of tetrahedral (red) and glassy (green) water for a) native contact state, b) near native state c) unbound state at 330 K, and d) native contact state, e) near native state f) unbound state at 300 K.
7.4 Conclusions

In this study we investigated the full association/dissociation mechanism of β-lactoglobulin to and from its native dimer state with Transition Path Sampling.

By combining the full Transition Path Ensemble (TPE) and the transition state ensemble (TSE) data, we found three association mechanisms describing the dimerization process: a one step aligned association (mechanism I), and a two step hop/misaligned (mechanism II) and a slide/misaligned (mechanism III) association. In agreement with experimental findings [5], salt bridges between charged amino acids were found to be present in all types of mechanism. The existence of many long-lived near native protein-protein interactions in all types of pathways signifies the increased role of local rearrangements in the near-native state before the proteins locks to the native dimer state. In agreement with the notion of water-mediated interactions such as water hydrogen bond bridges facilitating the docking and locking of the proteins into the native dimer [1, 16], we find that water hydrogen bond bridging interactions are a bottleneck towards aligning to the native state. This step can occur in a one step alignment (mechanism I) or a two step alignment (mechanism II or III). The LCP approximation of the TSE has been particularly useful in identifying the barrier regions of the process.

The hydration water structure and dynamics significantly changes during the association process. On the nanosecond long transition pathways, water adjusts to the protein surfaces still rather quickly (ps timescale). We found that the native state consists of two hydration populations, a tetrahedral and glassy one. The tetrahedral population being evident close to hydrophobic residues of the β-sheet and the glassy population close to charged and polar amino acids of the β-sheet and AB loop. The crevice introduced at the native dimer interface increased the glassy water hydration population while the hydrophobic amino acids coming into contact in the native state gave rise to the tetrahedral hydration population. Upon association both the water structural parameter increases, and the reorientational time increases. Thus the structure and dynamics of water indeed play a role in protein association. Moreover, we found that the bridging waters are maximal in the native state, thus providing evidence for their role on stabilization of the native state.

Our results give a first unbiased dynamical view of the mechanism of protein-protein association dissociation in explicit solvent, as well as an insight in the structural and dynamical role of the solvent in this process. Nevertheless, many issues remain. For instance, our path sampling only allowed dissociation and association trajectories of less than 70 ns. It might be that many more intermediates are visited before the system finally commits to either the bound or unbound state. Also, while there is no direct evidence that our findings generalize to arbitrary protein systems, β-lactoglobulin is not a special protein. Therefore we should investigate other protein systems.
Appendix

7.A Initialization of paths /Analysis

Figure 7.A.1: Committor analysis trajectories at 330 K as a function of a) minimum distance ($r_{\text{min}}$) between the two proteins and b) native contacts.
Figure 7.A.2: Root mean square deviation of the CA atoms of a) protein A at 300 K, b) protein B at 300 K, c) protein A at 330 K and c) protein B at 330 K.

Figure 7.A.3: Ramachandran angles for a) dimer in solution at 300 K and b) dimer in solution at 330 K.
<table>
<thead>
<tr>
<th>Contact</th>
<th>TPE (a) (40 paths)</th>
<th>TPE(b) (12 paths)</th>
<th>TPE(c) (13 paths)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIS146-SER150</td>
<td>30</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>ASP33- ARG40</td>
<td>28</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>ILE29-SER150</td>
<td>27</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>ARG40-ASP33</td>
<td>22</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>ASP28-ASN152</td>
<td>20</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>SER150-HIS146</td>
<td>19</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>ILE29-GLN155</td>
<td>17</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>GLN155-ILE29</td>
<td>17</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>SER150-ILE29</td>
<td>13</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>ILE29-PHE151</td>
<td>12</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>HIS146-ASN152</td>
<td>12</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LYS141-ASP130</td>
<td>11</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ASP 130-LYS141</td>
<td>11</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 7.A.1: Number of paths in which individual intermolecular contacts occur with a lifetime higher than 10 ns, for each path ensemble.
Bibliography


Appendix: Evaluating Science

Technological progress has been identified as the main factor boosting economic growth in the 20th century, as argued by the 1987 economy Nobel prize winner Robert Solow [1, 2]. Economists estimate the contribution of science to economic growth has been 20%-60% return on investments. Taking the example of basic research in the US in the post-war period, science has been primarily boosted by government funding. This model of primarily public funded research has been criticized during and after the 1990s by the New Public Management (NPM) administration of universities and research institutes.

According to NPM, the public sector would turn more efficient and effective under the evaluation and control by market forces [3]. This led to a severe change in the orientation of science and research away from primarily facilitating basic research, and towards an output performance strategy. According to a Dutch education ministry policy paper [4], valorization was introduced to describe the third main task of universities, the transfer of knowledge to society. Nowadays, NWO grants research proposals more favourably if they involve research with commercial benefits. Therefore, the question raised is whether this trend allows science to make progress and whether science should impact and benefit society. I will try to argue on how valorization as done nowadays often impinges on societal benefits and on science itself.

First of all, getting more and more driven by commercial needs, science poses low risk goals to itself, thus solving only short-term problems. However, basic research investigating different possible theories, involves high risk. During the 1920s, before Heisenberg introduced the uncertainty principle, Pauli wrote to a friend "For now, physics is again in confusion. I find physics too difficult for me and I would rather be an actor". A few months later and after Heisenberg’s paper, he wrote again to his friend saying "The type of mechanics brought up by Heisenberg gave me hope and joy for life". Scientific breakthroughs as the above signify the complex way science evolves, often without guarantee that a hypothesis will work, leaving the scientific debate open. Imagine what a tremendous mistake it would have been if Werner Heisenberg lived in the present having been unable to get a research grant only because his proposal wasn’t societally relevant or commercial enough.

Secondly, a crucial element of valorization is the existence of intellectual property rights (IPs). By introducing IPs, scientific knowledge and output is no longer freely available for sharing within the scientific community. This growing protection of science infrastructure and services endangers science, and hinders innovation, as scientists or innovators might find it costly to use protected material of other Institutes / Publishing houses / companies [5]. Thirdly valorization, introduces a tension between
principles. Scientists are made responsible to argue about the societal relevance of their research, thus altering their role from scientists having to solve basic problems to delegates of governments having to primarily deliver scientific products [6].

Notwithstanding the dangers of society and market driving research, I do believe that challenges entailed in the Millennium Development Goals (MDGs) [6] such as reduction of poverty, hunger and diseases are highly important for humankind. Therefore, public universities and institutions should indeed expand the definition of valorization towards societal goals, which are often no-commercial. Scientific knowledge and output should be openly shared and communicated.

However, science is not only about solving societal needs. It has autonomy and its own pace. According to Thomas Khun [7], conventional science is about solving puzzles, accumulating knowledge and making step-by-step progress which often leads to new technologies. However, the inability to make progress on these puzzles creates crises of existing theories and sometimes leads to a change of "epistemological paradigm", followed by a new theory. Such a change of paradigm has been Einstein’s theory on general relativity which generalized special relativity and Newton’s law of gravitation, providing a unified description of gravity as a geometry of space and time.

To conclude, science can have societal impact when not seen as a scientific factory. The possibility it gives to explore unexplored territories, the hope it creates to discover existing order in nature, the feeling of being useful, the curiosity to test the validity of established knowledge, are all virtuous driving forces. Once these ingredients are there, science has proved it can achieve breakthroughs with an increased societal impact.

Bibliography


List of publications

• Chapter 3

• Chapter 4

• Chapter 5

• Chapter 6

• Chapter 7
  Z. F. Brotzakis, P. G. Bolhuis, *in preparation*, Elucidating the mechanism and role of solvent for β-lactoglobulin dimerization using Transition Path Sampling
Summary

By the time the reader reads this line, billions of protein association events just occurred in our body, such as the ones regulating cell communication, signalling pathways, or in initiating a self-assembly processes, such as tissue fabrication, etc. The timescale of such transitions is slow, compared to atom vibrations and such events are termed rare, the reason being that protein or/and solvent interactions have to be disrupted and reformed in order for the transition to occur. Having an atomistic insight into rare transitions and their respective important interactions is pivotal for understanding and experimentally controlling such processes. Water is an important agent on its own in facilitating protein folding, recognizing ice crystal planes (anti-freeze proteins) and in mediating protein association. The aim of this thesis is threefold. First to better understand the role of water at the hydration shell of single proteins in terms of structure and dynamics, secondly to understand the association and first steps of self-assembly mechanisms of food and anti-freeze proteins, and thirdly to understand the role of water during the association mechanism. By performing Molecular Dynamics, we are able to investigate the H-bond structure and dynamics of water around hydrophilic and hydrophobic protein groups, as well as the effect of unfolding on water dynamics. We are able to correlate water reorientation dynamics with the H-bond structure at the hydration shell of anti-freeze proteins. Moreover, by employing Transition Path Sampling and Molecular Dynamics we study how anti-freeze peptides self-assemble into nanotubes, as well as their stability as a function of size. We further study the dimerization mechanism of globular proteins, the important interactions playing a role during the transition as well as the role of water. In order to do so, since the dimerization transition is rare, and the transition barrier asymmetric, we develop and employ a novel Transition Path Sampling shooting scheme that efficiently samples rare transitions with asymmetric barriers which simultaneously gives access to the transition state region.

Chapter 3: Dynamics of hydration water around native and misfolded α-lactalbumin. At first we investigate water dynamics around bovine α-lactalbumin by combining molecular dynamics simulations with polarization resolved femtosecond infrared (fs-IR) spectroscopy. We identify slowly reorienting surface waters and establish their hydrogen-bond lifetime and dynamical orientation relaxation dynamics, which we compare to the experimentally measured anisotropy decay. The calculated number of slow surface waters is in reasonable agreement with the results of fs-IR experiments. Slow waters form fewer hydrogen bonds compared to the bulk. At concave sites the protein-water hydrogen bonds break preferably via translational diffusion rather than
via a hydrogen-bond jump mechanism. The reorientation of water molecules residing at these concave sites is slower than at convex water exposed sites. Protein misfolding leads to an increased exposure of hydrophobic groups, inducing relatively faster surface water dynamics. Nevertheless, the larger exposed surface slows down a larger amount of water. While for a native protein hydrating water is slower near hydrophobic residues than at hydrophilic sites, mainly due to stronger confinement, misfolding causes hydrophobic water to reorient relatively faster because the exposure of hydrophobic groups destroys concave protein cavities with a large excluded volume.

Chapter 4: Correlation between water structure and dynamics in the hydration layer of a type III ocean pout anti-freeze protein. We report on a molecular dynamics study on the relation between the structure and (orientation and hydrogen bond) dynamics of hydration water around the ocean pout AFP III anti-freeze protein. We find evidence for an increasing ice-like structure from the area opposite to the ice binding site (IBS) towards the protein IBS, with the strongest ice-like structure around the THR-18 residue of the IBS. This ice-like structural signal correlates with increased reorientation decay times. Moreover, we find anti-correlation for several key residues that are not part of the IBS but are in its vicinity. These effects are enhanced at lower temperature. Finally, as AFP III anti-freeze protein is binding to ice crystal planes through a predominantly hydrophobic patch, we investigate the ice-like structure and dynamics of waters at partially dehydrated IBS. We find that upon dehydration the IBS becomes even more ice-like for the wild type, and that the water reorientation time becomes longer, but less so for the mutant T18N, which also has a higher hydration at the IBS. These results are in agreement with water-air VSFG spectroscopic experiments showing a reduced ice-like signal upon mutation at the IBS.

Chapter 5: Stability and growth mechanism of self-assembling anti-freeze cyclic peptides. Cyclic peptides (CPs) that self-assemble in ice-binding nanotubes are great candidates for use as anti-freeze proteins. Based on cyclic peptide sequence, cyclo-[(L-LYS-D-ALA-L-LEU-D-ALA)2], which can stack into nanotubes, we propose an anti-freeze cyclic peptide (AFCP) sequence, cyclo-[(L-LYS-D-ALA)2-(L-THR-D-ALA)2] which contains THR-ALA-THR ice binding motifs. Using molecular dynamics simulations we investigate the stability of cyclic peptides and their growth mechanism. We find that dimers of the AFCP sequence dissociate more frequently and are less stable than dimers of the original CP sequence, while nanotubes consisting of more than two peptides are stable. This sudden increase in stability of nanotubes of the AFCP sequence may be explained by the formation of H-bonds between Threonine side-chains. The Threonine distances in the ice-binding motifs are similar to those in the ant-freeze protein of Christoneura fumiferana, suggesting good ice lattice matching, and a potential for depression of the freezing point. In addition, we investigated the nanotube growth process, i.e. the association/dissociation of a single CP to an existing AFCP
nanotube, by Transition Path Sampling. We found a general dock-lock mechanism, in which a single CP first docks loosely before locking into place. Moreover, we identified several qualitatively different mechanisms for dissociation, involving different meta-stable intermediates, including a state in which the peptide was misfolded inside the hydrophobic core of the tube. We also find evidence for a mechanism involving non-specific association followed by 1D diffusion. Under most conditions, this will be the dominant pathway. The results yield insight in the mechanisms of peptide assembly, and might lead to improved design of self-assembling anti-freeze proteins.

Chapter 6: Spring shooting, a novel efficient Transition Path Sampling move. We present a novel transition path sampling shooting algorithm for efficient sampling of complex (biomolecular) activated processes with asymmetric free energy barriers. The method employs a fictitious potential that biases the shooting point toward the transition state. The method is similar in spirit to the aimless shooting technique by Peters and Trout [B. Peters and B. L. Trout, J. Chem. Phys. 125, 054108 (2006)], but is targeted for use with the one-way shooting approach, which has been shown to be more effective than two way shooting algorithms in systems dominated by diffusive dynamics. We illustrate the method on a 2D Langevin toy model, the association of two peptides and the initial step in dissociation of a β-lactoglobulin dimer. In all cases we show a significant increase in efficiency.

Chapter 7: Elucidating the mechanism and role of solvent for β-lactoglobulin dimerization using Transition Path Sampling. Dimerization of proteins is a fundamental process in nature. While conceptually simple, the underlying association mechanism and the role of the solvent are poorly understood. Here we resolve these issues for the dimerization of β-lactoglobulin using Transition Path Sampling of all atom molecular dynamics trajectories. The association process is found to occur via (at least) three distinct mechanisms: 1) aligned association to the native dimer interface, 2) misaligned association at non native sites followed followed by hop towards the native state and 3) misaligned association followed by sliding of the protein towards the native state. We find that the native dimer state is stabilized by hydrogen bond bridging waters. Interestingly, water at the native interface can be found in two dynamical hydration states, a glassy one and a tetrahedral one. The crevice introduced upon binding increases the glassy populations as well as increases the average tetrahedrality of water, mainly at the vicinity of hydrophobic residues.
Samenvatting

Tegen de tijd dat de lezer deze regel heeft gelezen, zijn er miljarden eiwit associatie processen gebeurd in ons lichaam, zoals bijvoorbeeld het reguleren van de communicatie van de cel, signaal transductie of het initiëren van zelf-assemblage processen, zoals het maken van celtissue. De tijdschaal van dit soort transities is langzaam, vergeleken met de vibraties van moleculen and zulke transities worden dus ook als zeldzaam beschouwd. De reden hiervoor is dat eiwit interacties moeten gebroken worden of gevormd voor de transitie om plaats te vinden. Atomair inzicht in zulke zeldzame transities is belangrijk voor het begrijpen van zulke processen. Water is een belangrijk op zich zelf in het faciliteren van eiwitvouwing, zoals het herkennen van kristal oppervlaktes (antifreeze eiwitten), and in het begeleiden van eiwit associatie. Het doel van deze thesis is drievoud. Ten eerste, het beter begrijpen van de rol van water in de hydratieschil van enkele eiwitten in termen van structuur en dynamica, ten tweede om de associatie en eerste stappen in de zelf-assemblage mechanisme van voedsel en antivriezeiwitten te begrijpen en ten derde de rol van water te bestuderen in dit mechanisme. Door het uitvoeren van moleculaire dynamica simulaties, zijn we in staat om de waterbrug netwerk en dynamica van water rond hydrofiele en hydrofobe eiwitgroepen, zowel als het effect van ontvouwing op de dynamica van water. We zijn in staat de water reorientatie dynamica van water te corrleren met het waterstofbrug netwerk van de hydratieschil van antivriezeiwitten. Bovendien, door het gebruik van Transition Path Sampling en moleculaire dynamica bestuderen we hoe antivries eiwitten zelf-assembleren in nanotubes en ook de stabiliteit als functie van de grootte de nanotubes. Verder, bestuderen we de dimerisatie mechanisme van globose eiwitten, de interacties die een belangrijke rol spelen in deze transitie alsook de rol van water. Om dit te doen, aangezien dit proces zeldzaam is, en de transitie barrière asymmetrisch, hebben we een nieuwe Transition Path Sampling techniek ontwikkeld wat efficient reactieve paden genereert voor systemen met een asymmetrische barrière.

Hoofdstuk 3: Dynamiek van hydratie water rond native en verkeerd gevouwen α-lactalbumin. Ten eerste hebben we de dynamica van water bestudeerd om dierlijk α-lactalbumin door het combineren van moleculaire dynamica simulaties en polarisatie opgeloste femtoseconde infrarood (fs-IR) spectroscopie. Wij identificeren langzaam oriënterende oppervlakte water moleculen en meten de levensduur van deze waterstofbruggen en de dynamische orientatie relaxatie dynamica, die we vergelijken met de gemeten anisotropieverval experimenten. De uitgerekende aantal langzame waters is in goeie overeenkomst met de resultaten van fs-IR experimenten. Langzame waters vormen minder waterstofbruggen vergeleken met bulk water. Bij concave sites
breken eiwit-water waterstofbruggen eerder via translatonale diffusie dan via water-
stofrug sprong mechanisme. De reorientatie van water moleculen zittend op deze
concave is langzamer dan convex water blootgelegde sites. Eiwit misvouwing leidt tot
een verhoogde blootstelling van hydrofobe groepen, wat leidt tot relatief snellere wa-
ter dynamica. Desondanks, de grotere blootgestelde oppervlakte vertraagt een grotere
hoeveelheid water. Terwijl voor native eiwit hydraterende water langzamer is dichtij
hydrofobe residuen dan bij hydrofiel sites, voornamelijk door sterkere opsluiting, mis-
vouwing zorgt ervoor dat hydrofoob water relatief sneller reorienteert doordat de bloot-
stelling van hydrofobe groepen de concave eiwit porien met een grote uitgesloten vol-
ume.

**Hoofdstuk 4: Correlatie tussen water structuur en dynamica in de hydratie laag,
van een type III amerikaanse puitaal antivries eiwit.** Wij geven een moleculair dy-
namica studie over de relatie tussen de structuur en (oriëntatie en waterstofbrug) dy-
namica van gehydrateerd water rondom de Amerikaanse puitaal AFP III antivries ei-
wit. We vinden bewijs voor een verhoogde ijsachtige structuur van het gebied tegen-
over de ijsbindend site (IBS) naar het IBS van het eiwit, met de sterkste ijsachtige
structuur rondom de THR-18 residue van het IBS. Dit ijsachtig structuur signaal corre-
leert met verhoogde reorientatie vervaltij. Bovendien, wij vinden een anticor-
relatie voor verschillende sleutel residuen die niet deel uitmaken van het IBS maar
dichtbij zitten. Deze effecten zijn verhoogd bij lagere temperaturen. Ten slotte, als het
AFP III antivries eiwit bindt met het ijsvlakte door voornamelijk hydrofobe stukken,
bestuderen we de ijsachtige structuur en dynamica van water bij partieel gehydrateerd
IBS. We vinden dat het IBS nog meer ijsachtig wordt in het wild-type wanneer de-
hydratie plaatsvindt, en dat de water reorientatie tijd groter wordt, maar minder zo
voor het T18N mutant, welke ook een hogere hydratatie heeft dan het IBS. Deze re-
sultaten komen overeen met water-lucht VSFG spectroscopie experimenten wat een
gereducteerd ijsachtig signaal bij mutatie van het IBS laat zien.

**Hoofdstuk 5: Stabiliteit en groei mechanisme van zelf-assemblerende antivries cy-
clische peptiden.** Cyclische peptiden (CP) die zelf-assembleren in ijsachtige nanotubes
zijn kandidaten om te gebruiken als antivries eiwitten. Gebaseerd op de cyclische pep-
tide sequentie, cyclo-[(L-LYS-D-ALA-L-LEU-D-ALA)2], welke kunnen stapelen in
nanotubes, stellen wij een antivries cyclische peptide (AFCP) sequentie voor, cyclo-
[(L-LYS-D-ALA)2-(L-THR-D-ALA)2] welke een THR-ALA-THR ijsbinding motief
bevat. Door moleculaire dynamica te gebruiken bestuderen wij de stabiliteit van cyclis-
che peptiden en het groei mechanisme. Wij vinden dat dimeren van de AFCP sequentie
dissociëren vaker en zijn minder stabiel dan dimeren van het originele CP sequentie,
terwijl nanotubes met meer dan twee peptiden stabiel zijn. Deze plotse sprong in sta-
biliteit van nanotubes van het AFCP sequentie kan worden uitgelegd door de formatie
van waterstofbruggen tussen Threonine en de zijketens. De Threonine afstanden in de
ijsbinding motieven zijn gelijk aan die in antivries eiwitten van Chriseoneura fumiferana, wat een goeie ijs-rooster matching oppert, en de mogelijkheid om het vriespunt te verlagen. Daarnaast bestuderen we het nanotube groei proces, i.e. de associate en dissociate van een enkele CP bij een reeds bestaande AFCP nanotube met Transition Path Sampling. We vinden een algemene dock-lock mechanisme, waar een enkele CP eerst licht bindt voordat het compleet bindt. Bovendien, we identificeren verschillende kwalitatieve verschillende mechanismes voor dissociatie, met verschillende meta-stabiele intermediaire toestanden, inclusief een toestand waar het peptide misvouwt binnen de hydrofobe kern van de tube. We vinden ook bewijs voor een mechanisme wat een non-specifieke associatie inhoudt gevolgd door 1D diffusie. Onder de meeste condities, zal dit het dominante pad zijn. De resultaten bieden inzicht in de mechanismes van eiwit assemblage, en kunnen ook leiden tot verbeterde ontwerp van zelf-assemblerende antivries eiwitten.

Hoofdstuk 6: Spring shooting, een nieuw efficient Transition Path Sampling move. We presenteren een nieuwe transition path sampling shooting algoritme voor efficient samplen van complexe (biomoleculair) geactiveerde processen met asymmetrische vrije energie barrières. Deze methode gebruikt een fictief potentiaal wat het shooting punt biased naar het transitiepunt. De methode is vergelijkbaar in geest met de aimless shooting move techniek van Peters en Trout [B. Peters en B. L. Trout, J. Chem. Phys. 125, 054108 (2006)], maar is gericht op het gebruik van one-way shooting, wat effectiever is gebleken dan de two-way shooting algoritmes in systemen die gedomineerd zijn door diffusie. We illusteren de methode met een 2D Langevin model, de associatie van twee peptiden en de initiële stap in de dissociatie van $\beta$-lactoglobulin dimer. In elke geval laten we zien dat de efficientie door het nieuwe spring shooting move verhoogd wordt.

Hoofdstuk 7: Mechanisme en rol van het oplosmiddel voor $\beta$-lactoglobulin dimerisatie door middel van Transition Path Sampling. Dimerisatie van eiwitten is een fundamenteel proces in de natuur. Al hoewel simpel, de onderliggende associatie mechanisme en rol van het oplosmiddel is niet volledig begrepen. Hier geven we verklaringen voor deze vragen over de dimerisatie van $\beta$-lactoglobulin door middel van Transition Path Sampling van moleculaire dynamica trajectoria. Het associatie proces vindt plaats via (ten minste) drie verschillende mechanismes: 1) directe associatie naar de native dimer, 2) associatie via non native sites gevolgd door hops naar de native toestand en 3) associatie gevolgd door het schuiven van het eiwit richting de native toestand. We vinden dat de native eiwit toestand gestabiliseerd wordt door waterstofbrug gebrugde waters. Waters dichtbij de native interface wordt gevonden in twee verschillende dynamische hydratatie toestanden, een glasachtige en een tetraedrale. De kloof geïntroduceerd bij de binding verhoogt de glasachtige populatie alsook de gemiddelde tetraedrale karakter van het water, voornamelijk in de buurt van hydrofobe structuren.
Acknowledgements

Few sentences in the preface of a thesis are by no means enough to express my sincere admiration and appreciation towards a brilliant scientist and an even more rare and noble man, my supervisor, Peter G. Bolhuis. During my PhD, it has been an honor collaborating with Peter who had always combined open mindedness to new ideas and possibilities with professional guidance and experience. Peter, working with you in Amsterdam helped me being academically more independent and much more analytical. It made me realize the importance of deep knowledge in chemical kinetics, thermodynamics and statistical physics. Finally, it elucidated how I can see when I am on the top of the barrier. Therefore I can do no less but express my deep gratefulness and luck having worked with you.

However things would not have carried on smoothly without the computational chemistry group as a whole. Scientific chats skewed by the constant exchange of arguments at mini-meetings, group meetings, coffee corner, launch had been giving their turn to socially and publicly oriented discussions of the post-modern unrested world of 2012-2016 and vice-versa. I will never forget the discussion with Evert Jan Meijer about water diffusion, with Jocelyne Vreede and David Swenson about the reaction coordinate analysis on two barriers in a row, the nucleic acid dolls and the extraordinary drinks of David’s. The discussions with David Dubbeldam about local pressure and the unbelievable Greek football European championship and Sheldon as well as with Bernd Ensing about metadynamics and Italy can’t be easily forgotten. Wagner Arthur and Ariana, you have been amazing colleagues, true milestones throughout my PhD. I will never forget how poorer and dull things would have been without you guys. The Leonardo Da Vinci move is especially devoted to you! Murat, I will never forget the warm welcome you gave me to the group and Juriaan I will never forget your advice on taking a break when things don’t work. Ran, I will never forget your snoring at Veldhoven, it has been quite and experience. Aatish and Kush, you have been quite a unique, patient and funny duet, Nazila, I will never forget the Persian similarities of Greek songs we shared, and Ivan since you moved to my office things got a more mathematic «twist». Verena, Ambuj, Andri, Nitish and Rakesh Jurn and Sander, I will definitely join more coffee-corner breaks before leaving, thanks for trying to give me a break from the PC. Aditya thanks for the awesome chats on reaction rates and you Arjun for being among other so funny. I wish you a fruitful PhD with Peter. Tomislav and Tatu going to the gym together and living in the same building has been an amazing sense of integration and feeling like home, especially in the early days of my PhD. Thank you! Talking about first steps of PhD, Petra, Cristel, Renate, Ineke, you have
been amazingly approachable and helpful. I will never forget your effort during my PhD. I would like to also thank my collaborators, Carien Groot and Huib Bakker for the experimental insight on water dynamics as well as Ilja Voets for her suggestions and comments based on her vast experience with anti-freeze proteins. I thank Mascha Gehre for the wonderful time we spent understanding nanotube self-assembly during her masters thesis. Finally, I express the endless appreciation and gratitude to Prof. Doros Theodorou and dr. George Boulougouris for introducing me to molecular simulations, rare events and Markov state models. Their conviction and belief in my skills have severely triggered my motivation and curiosity about science.

As we say in Greek “Some people are like lighthouses in your life”. Stefania, Iris and Dimitris Kokos, ἠσασταν αληθινό καταφύγιο στην αρχή του διδακτορικού μου, σας ευχαριστώ. Paolo e Giovanni, fino a Genoa per noi, amici miei. I thank the Greek political group Reinform for reminding me to act, to fight for our rights and to look forward. Elkan and Celita, you have been the best landlords I could have possibly imagined. Your house has been warm shelter I really felt home. I am grateful. Olima, Thomas, α/φοί Laourijsen, δε θα ζεγάσω ποτέ τις ρεμπέτικες βραδυές μας. Νικόλα και Νάνσυ, δε θα ζεγάσω ποτέ τις αιώνες νύχτες (αμφελο)φιλοσοφίας στο σαλόν μου. Pilar, no voy a olvidar nuestros passeos en Oosterpark. Carlinha, eu nunca vou esquecer os nossos passeios no Vondelpark. Ceci, eres una persona increíble y era muy afortunado de conocerte durante mi doctorado. Jimenaki, Virgi, Nora and Aoife (JVNA) I truly have no words here. Only that I adore your friendship to me and even more, love the friendship you have with each other. Joao and Tata, you been like distant little cousins. Thank you! Ezequiel and Lydia, you’ve been the most tango nuevo song I heard and saw in Amsterdam. We have known each other throughout most of my time in Amsterdam. I am grateful for your warm heart and open mind! Thanks to Academia de Tango I often remained on axis. Παρόλα αυτά, χάρειν του Πάνου και της Σοφίας απέκτησα τάνγκο άξονα και γι αυτό τους ευγνωμονώ απ τα βάθη της ψυχής μου μαζί με το ότι είναι φανταστικοί φίλοι. Ηλία, Ντέμπορα, Φοίβο, ἠσασταν το σημείο σύνδεσης του παρελθόντος με το παρόν. Σας ευχαριστώ! Ηλέκτρα, Βάλλυ, Χάρη, ζησαμε φιλοσοφικά πάρτυ και παραστάσεις αληθινές σοσιαλ οσο και σκηνικές. Σας ευχαριστώ! Θα ήθελα να ευχαριστήσω ιδιαίτερα το Δημητράκη για το υπέροχο εξώφυλλο αλλά και για τη μακρά φιλία που μας συνδέει. Θα ήθελα να πω και ό,τι λόγια για τους αμιάμπουλους μου, τους Θεοδωρή, Αγγέλλα, Κωνσταντίνο, Νόντα, Ορφέα, Ορέστη, Στέφανο, Αλέξανδρο, Κορίνα, Ελίζα (LB), Άλκη, Σοφοκλή, Γιάνγκο και Γιώργη. Παίδες σας ευχαριστώ που υπάρχετε, είστε φιλοσοφική πηγή δύναμης σε όλη μου τη ζωή. Κωστάκη, Νικόλα, Αλέξ, Πάρη, Νάσια, Ισμήνη, Νικόλα (Ξανα)γνωριμίτραξε και δέσαμε στα 20s μας. Σ évα δωματία στο Άμστερνταμ. Να εκδηλώσω την ευγνωμοσύνη μου στον Γιώργο, Δημητράκη, Θεοδωρή που στάθηκαν συναδέλφιτές και προπονητές στη Θάλασσα, σε όλους τους καιρούς.
Στη φύση, σε μη μηδενική θερμοκρασία (δηλαδή εκτός εργαστηρίου) υπάρχουν δύο τρόποι σταθεροποίησης, η ενθαλπία και η εντροπία, η ενέργεια και το χάος, η ἑλξη/άπωση και η αύξηση των δυνατών μικροκαταστάσεων. Συχνά, αυτές οι δυνάμεις είναι συμπληρωματικές, άλλες φορές ανταγωνιστικές. Εξαρτάται από το σύστημα η τελική κατάσταση ισορροπίας και ο ρυθμός προς αυτήν. Κι όμως, για να πάς από το Α στο Β χρειάζεσαι πάντα μία σπάνια τοπική διαταραχή. Μαζί σου έμαθα να απολαμβάνω αυτή τη διαταραχή χωρίς να ξέρω που θα καταλήξω. Οι δυνάμεις του δικού μας διπόλου δείχνουν περίτρανα το βέλος, Κλέο. Σε ευχαριστώ παράφορα για αυτό το ξεβόλεμα.

Την μεγαλύτερη ευγνωμοσύνη όμως τη χρωστώ στην πολυμερή και πολύπλευρη οικογένειά μου. Τους Ναυσικά, Μαρία, Κώστα, Αφροδίτη, Λευτέρη, Τζίνα, θεία Αφρό, ευχαριστώ θερμά. Αλλά τίποτα από όλα αυτά δε θα είχαν ξεκινήσει χωρίς την Νατάσσα και τον Παντελή που με γέννησαν, την Αννούλα και το Δημήτρη με τους οποίους εξισορρόπησα, τον αδελφό μου Γιάννη με τον οποίο μεγάλωσα.

Σα βγείς στον πηγαιμό για την Ιθάκη, να εύχεσαι να την μαχώνες από να μαχώνες, γεμάτος περιπέτειες, γεμάτος γνώσεις...

Τους Λαιστυγόνας και τους Κύκλωπας, τον άγριο Ποσειδώνα δε θα συναντήσεις, αν δεν τους κουβάνες μες στην ψυχή σου, αν την ψυχή σου δεν τους στήνει εμπρός σου.

Κι αν πτωχική την βρείς, η ιθάκη δεν σε γέλασε. Έτσι σοφός που έγινες, με τόση πέφα, ήδη θα το κατάλαβες οι Ιθάκες τι σημαίνουν.