Hydration layer dynamics and association mechanisms of food and antifreeze proteins

A Molecular Dynamics and Transition Path Sampling study

Brotzakis, Z.F.

Citation for published version (APA):
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 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 $\beta$-lactoglobulin ($\beta$-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

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

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 $[-\Delta\tau_{\max}, \Delta\tau_{\max}]$, 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 $[-\Delta\tau_{\max}, \Delta\tau_{\max}]$. 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 $L_{\max}$. 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 $L_{\max}$, 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>$&gt; 1$ 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, 148-146, 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 ($r_{\text{min}} > 1$ 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 \( \text{U} \), 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 \( \rightleftharpoons \text{U} \)) at \( T = 330 K, P = 1 \text{ atm} \) 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 \( \text{B} \rightleftharpoons \text{U} \) transition, where the bound state \( \text{B} \) is defined as when the protein protein interfacial area is \( > 2(\text{nm}^2) \). 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, I$_1$ 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

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

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 \( c_j^m(t) \) 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. 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. 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 \( \tau_j^m \) 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-
drophilic 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 $P(\theta_i) = \sum_j P(\theta_{ij})/\sum_i \sum_j P(\theta_{ij})$ 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.

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

Where $\tau$ is time, $C_{\text{Norm}}$ is a normalisation constant, such that $C_{\text{Bridging}}(0) = 1$, $i$ and $j$ are running over the residue number of proteins $A$ and $B$ respectively, $N_{\text{res}}$ is the total number of residues per protein and $N_{\text{wat}}$ is the total amount of waters in the simulation box. $\mathbb{1}_{N_{i,j}^{\text{Bridging}}(t)}$ 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).
Figure 7.3.3: Network of transitions during the full association process. We identify three type of mechanisms. Transition I (NC ⇔ I₁ ⇔ U), transition II (NC ⇔ I₁ ⇔ I₂ ⇔ I₃ ⇔ U), and transition III (NC ⇔ I₁ ⇔ I₃ ⇔ U). These three mechanisms are indicated by the gray arrows.

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.

configuration (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
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 TSE$_1$ and TSE$_2$ 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 (I$_3$) and either slide towards the aligned near native state (I$_1$) towards maximizing their bridging waters or lose these interactions again while hopping to the near native state (I$_1$) which has again abundant bridging waters. At the near-native state further formation of these bridging waters is a bottleneck (TSE$_1$) 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 $\tau$. 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 $\tau < 4$) water population, that we label tetrahedral and a slowly reorienting, less structured water population that we label glassy ($S < 0.4$ and $\tau > 4$). 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 and LEU149)...
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

![Committor analysis trajectories at 330 K as a function of a) minimum distance ($r_{min}$) between the two proteins and b) native contacts.](image)

Figure 7.A.1: Committor analysis trajectories at 330 K as a function of a) minimum distance ($r_{min}$) 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></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>1</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


