Proteins in action: simulations of conformational changes in small proteins

Juraszek, J.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 4

Trp-cage mini-protein in explicit solvent: Calculation of the rate constant

We perform rate calculations of the rate limiting folding and unfolding processes of Trp-cage mini-protein in explicit solvent using Transition Interface Sampling. Previous Transition Path Sampling simulations revealed that in this (un)folding process the protein maintains its compact configuration, while a (de)increase of secondary structure is observed[14]. The calculated folding rate agrees reasonably with experiment, while the unfolding rate is 10 times higher. We discuss possible origins for this mismatch. We recomputed the unfolding rate with the Forward Flux Sampling method, and found a discrepancy of 4 orders of magnitude. This discrepancy is explained by the fact that the FFS is much more sensitive to the choice of order parameter than TIS. Finally, we used the previously computed TPS ensemble [14] to screen combinations of many order parameters for the best model for the reaction coordinate by employing likelihood maximization. We found that a combination of the root mean square deviation of the helix and of the entire protein provided the best reaction coordinate description.

4.1 Introduction

Trp-cage is a model mini-protein (NLYIQ WLKD GPSSG RPPPS) designed by Neidigh et al [19] to reach the limits of folding speed but preserving the secondary and tertiary structure. This 20-residue polypeptide in the native state contains an α-helix (residues 2-8), a 3_{10}-helix (residues 11-14) and a polyproline II helix (residues 17-19) (Figure 4.1). The three helices form a hydrophobic cavity, further stabilized by a salt bridge (between residues 9 and 16), in which Trp-6 is buried.

Trp-cage has been the subject of many experimental studies in the last six years. Laser temperature-jump spectroscopy experiments by Qiu et al. [26] revealed two state folding mechanism with the folding rate $k \approx (4.1 \mu s)^{-1}$. This hypothesis has been recently reex-

---

1This chapter is based on J. Juraszek and P. G. Bolhuis Rate constant and reaction coordinate of Trp-cage folding in explicit water. submitted to Biophysical Journal
 determined by Neuweiler [20]. The fluorescent correlation spectroscopy shows that the protein (un)folds in a more complicated manner via an intermediate molten globule-like state, characterized by exposure of the tryptophan to the solvent. It remains unclear at what stage of folding the helix is being formed. The correlation between tryptophan fluorescence and circular dichroism (CD) melting data was proposed as evidence of simultaneous breaking of the hydrophobic core and helix solvation during (un)folding [26]. UV-resonance Raman spectroscopy measurements show some evidence of a helical structure in the denatured state of Trp-cage, thus suggesting an early formation of the helix is possible [1]. The helix melting curve is also broader than usual, the $\alpha$-helix is stable until 30°C and melts between 40 and 70 °C [1].

Several computational studies on Trp-cage exist. REMD simulations in explicit solvent [30] confirmed the two state nature of Trp-cage (un)folding. An intermediate state structure, containing two hydrophobic cores, was proposed as the reason of Trp-cage being such fast folder. Folding events of Trp-cage have been observed in all-atom implicit solvent MD simulations [27, 28, 22], in a coarse-grained model [10] and also recently in the explicit solvent REMD simulation in the AMBER FF by Paschek at all [23]. Two different folding routes, the predominant one in agreement with the previously mentioned model, have also been predicted by an all-atom Go model [18]. In Chapter 3 we studied Trp-cage with Transition Path Sampling and found that it follows two major (un)folding routes [14] each representing a generic protein folding mechanism: nucleation-condensation (NC) and diffusion-collision (DC) (see Section 1.2.3). Along one route ($N - I$) the polypeptide is first forming the main secondary structure - the alpha helix, followed by the appearance of the tertiary contacts (DC). On the second pathway ($N - L$) the tertiary contacts precede the formation of the secondary structure elements (NC).

In this chapter, we employ Transition Interface Sampling [29] to calculate the rate constants for the folding and unfolding of Trp-cage mini-protein in explicit solvent. We choose the most abundant pathway of the two possible folding routes, the nucleation-condensation pathway [14]. On this pathway, the so-called $N - L$ route (see Figure 4.1), the protein maintains its native state tertiary contacts, while the secondary structures are solvated.

### 4.2 System Preparation

The 262-atoms protein NMR structure (PDB entry 1LE3) was solvated in 2030 SPC water molecules in a rhombic dodecahedral box of the diameter of 45 Å. The box size was changed to 45.23 Å upon equilibration at ambient conditions of 1bar and 300 K for 10 ns, preceded by energy minimization and a protein position restraint run of 100 ps. The MD runs were performed using Nosé-Hoover thermostat. The pressure coupling in the equilibration run was done using Berendsen box scaling method. All of the following MD simulation in this paper were performed at constant volume (box diameter 45.23 Å) using the Gromacs molecular simulation package [17] together with OPLSAA FF [15] and SPC model of water [17]. In all of our simulations the time-step was 2 fs, dodecahedral periodic boundary conditions were applied, long range electrostatic interaction were treated by Fast Particle-Mesh Ewald [7, 11] with a grid spacing of 1.2 Å and the Nosé-Hoover thermostat [21, 13] ensured a constant temperature.
4.3 METHODS

4.3.1 Molecular Dynamics

All of our Transition Interface Sampling simulations are performed using Molecular Dynamics with a particular version of Andersen temperature coupling, applied only to the center of mass motion of water molecules. The coupling we employ is very weak, so the trajectories do not diverge significantly from their deterministic counterparts [12]. We choose the coupling constant in such a way that the diffusion of water molecules remains unchanged compared to a simulation using the Nosé-Hoover thermostat. For the box sizes we use, the coupling frequency appears to be slightly less than 1 water molecule per MD step. We have tested this procedure on a sample of pure water, in order to check if the system couples correctly to the desired temperature and if the velocity distribution is preserved compared to the Nosé-Hoover MD. The simulations were carried out on a system of 2303 SPC water
CHAPTER 4. TRP-CAGE RATE CONSTANT

Figure 4.2: Comparison between the simulations using our version of the Andersen thermostat and the standard Nosé-Hoover thermostat. (up) Distributions of the kinetic energy. (bottom) Distributions of the particle velocity in the x-direction.

molecules in a rhombic dodecahedral box, using the same settings as for all the simulations of Trp-cage described in Section 4.2. Figure 4.2 shows that this procedure yields velocity distributions identical to the ones obtained using Nosé-Hoover MD and that the equipartition of energy is fulfilled.

The velocity distributions of the center of mass of water molecules are Gaussian with the same standard deviation $\sigma = 0.373$ for both Nosé-Hoover and our version of Andersen thermostat. The equipartition of energy is fulfilled, meaning that the kinetic energy is equally distributed among rotation of water molecules and the motion of their center of mass. Both kinetic energy contributions are depicted in Figure 4.2 and are identical for both simulations.

After insertion of the protein, we observed that the system relaxes to a temperature higher by about 2 degrees than desired. This phenomenon is caused by a slight energy drift of the temperature-uncoupled polypeptide. Many factors cause the uncoupled system to overheat but the single precision MD of Gromacs is probably the most significant one. The protein couples to the room temperature through exchange of energy with the solvent, and as in the folded state the solvent-protein contact area is the lowest, the exchange of energy is slower than in the unfolded state. Even though the difference of 2 degrees would not have a significant impact on the (un)folding rate, we compensate for the mild overheating of the protein by decreasing the imposed temperature of water by about 4 degrees. We could as well increase the Andersen coupling constant in order to remove heat faster from the water
shell surrounding the protein, but this in return would cause the diffusion of waters to drop down below the Nosé-Hoover MD reference.

### 4.3.2 Transition Interface Sampling

While Trp-cage is considered one of the fastest folding polypeptides forming secondary and tertiary structure elements, the free-energy barrier separating unfolded from the folded state is still too high to observe a folding event in a regular MD simulation in explicit solvent. The experimental folding time is $\tau_{\text{fol}} = 4.1 \mu s$ [26], and one would have to run on average $4.1 \mu s$ MD to observe one folding event. We know from a previous study that transition paths crossing the rate limiting barrier are on average $\tau_{\text{trans}} = 3$ ns long - three orders of magnitude shorter than the folding time. As the probability to find the protein on a transition paths is about $7 \times 10^{-4}$, it would be a waste of computational effort to examine the folding rare events with a straightforward MD. Moreover, $4.1 \mu s$ is still far beyond today’s computational limits. Instead we use Transition Path Sampling to sample transition pathways effectively.

**Transition Path Sampling.**

Transition Path Sampling (TPS) [8, 6, 9] comprises a set of techniques designed for sampling the transition path ensemble, connecting two stable states, without prior knowledge of the TS ensemble nor the reaction coordinate. New transition paths are sequentially generated starting from the initial pathway using the shooting algorithm.

The transition path sampling method relies on the shooting algorithm [9], which alters a time slice on an existing path randomly and integrates a new trial path forward and backward in time. The deterministic shooting algorithm runs into problems for 'long' diffusive folding trajectories (long compared to the time-step, i.e. longer than a few ps). While a random shooting point might seem to lie in the barrier region (i.e. outside of the stable state definitions) it can in fact already be completely committed to one of the stable states. In that case the acceptance ratio will be extremely low. Only when shooting from points around the true transition state region, we can expect a reasonable acceptance. To alleviate this problem we employ the stochastic shooting algorithm [4] allowing shooting in one direction, either forward or backwards. Application of deterministic MD to generate stochastic trajectories requires the introduction of a small amount of stochasticity in the trajectories, for instance by the Andersen thermostat. As mentioned above, the Andersen coupling constant can be made small enough so that there is no noticeable difference from completely deterministic dynamics [12].

The advantage of stochastic sampling is an improved acceptance ratio of about 50%. However, on the other hand, we have to wait several successful shots before an entirely new pathway is generated, because a single successful shooting replaces only one part of the trajectory. In case of Trp-cage in explicit solvent, as examined in this chapter, the method was two orders of magnitude more efficient than regular MD [14], in finding the uncorrelated transition pathways.
Transition Interface Sampling - the algorithm

The TIS algorithm, contrived for the calculation of rate constants [29], adopts the same concept of sampling transition paths by employing the shooting algorithm as in TPS. The method does not depend on the choice of an order parameter, and thus in principle can be applied to complicated systems in which the reaction coordinate is not a priori known. The only requirement is that an order parameter $\lambda$ can successfully distinguish between the initial and final states and can be used to divide the configuration space of the system in a number of subspaces by introducing $N$ interfaces $\lambda = \lambda_i$, where $i = 0, \ldots, N$.

The calculation of the rate constant is then reduced to the subsequent calculation of the conditional crossing probabilities $P_A(\lambda_{i+1}|\lambda_i)$ that a trajectory starting in state $A$ crosses interface $\lambda_{i+1}$, provided that it has passed through interface $\lambda_i$. The crossing probabilities together with the effective positive flux through the first interface $f_A^{1,0}$, as defined in Ref [29], enable to calculate the rate constant $k_{AB}$:

$$k_{AB} = f_A^{1,0} \prod_{i=1}^{N-1} P_A(\lambda_{i+1}|\lambda_i) \quad (4.1)$$

The flux factor $f_A^{1,0}$ in equation (4.1) can be calculated by performing an MD simulation in the initial state $A$ and counting the number of recrossings events of the first interface $\lambda_1$ per unit of time. The conditional probabilities $P_A(\lambda_{i+1}|\lambda_i)$ can be determined by performing a TIS simulation.

We employ the stochastic version of the shooting algorithm, meaning that upon shooting we do not modify either positions nor momenta in the system. The stochasticity introduced by the mild Andersen coupling, will cause the trajectory to diverge from the initial one, taking care of the constant temperature at the same time. To increase the efficiency of sampling we introduce a bias causing the shooting points to be drawn around the interface, with a Gaussian distribution, to assure that almost each trajectory will cross the interface. This bias is introduced by assigning to each timeslice $\tau$ a non uniform weight $w(\tau)$, that depends on the values of the order parameter $\lambda(\tau)$ and $\lambda_i$. The probability of accepting a given timeslice as a shooting point can be written as:

$$p_{sp}(\tau) = w(\tau) / \sum_{\tau=0}^{N} w(\tau) \quad (4.2)$$

and becomes $1/N$ in case no bias is introduced. In our version of the TIS algorithm $p_{sp}(\tau)$ equals to:

$$p_{sp}^{\text{gauss}}(\lambda_i, \sigma)(\tau) = \frac{e^{\exp(-(\lambda_{\tau} - \lambda_i)^2/2\sigma^2)}}{\sum_{\tau=0}^{\tau=F} e^{\exp(-(\lambda_{\tau} - \lambda_i)^2/2\sigma^2)}} \quad (4.3)$$
where \( \sigma \) is of the order of picosecond fluctuations of the parameter \( \lambda \) around interface \( \lambda_i \). Instead of using a fixed path length, we rely on a flexible path length definition [29]. From an existing path with \( N^o \) time slices we choose a random time slice \( \tau \) as our shooting point. We randomly choose either the forward or backward direction for shooting and reverse the momenta for a backward shot. We then integrate the equations of motion using Andersen coupled MD until after time \( \tau_f \) we reach either region A or B. A trial path is constructed in which the newly shot trajectory replaces a part of the old path starting at \( \tau \). In case of the backward shot all momenta are reversed again. The new trial path has a path length \( N^n = \tau + \tau_f \) in case of forward trial shot and \( N^n = (N^o - \tau) + \tau_f \) in case of a backward shot. If the trial path does not connect A with B it is rejected straightaway, otherwise, in order to obey detailed balance it may be accepted with the Metropolis acceptance ratio \( P_{acc} = \min(1, W^o/W^n) \), where the min function returns the smaller of its arguments and \( W^i \) is a sum of all weights \( w(\tau) \) of trajectory \( i \) (a path weight). To avoid having to reject paths that do connect A and B but are too long, in practice we choose a random number \( \xi \in (0, 1] \) and determine the maximum path weight \( W_{max} = W^o/\xi \) in advance. The MD integration can then be halted if the total trial path weight exceeds \( W_{max} \). Note that our implementation of TPS differs from the implementation in Ref [5] in that the beginning and the end of the paths are always at the boundary of respectively A and B. It also differs from the standard TIS implementation [29] because of the use of the stochastic algorithm. We believe that our implementation here is an efficient path sampling algorithm for diffusive processes.

In summary, our stochastic TIS algorithm consists of the following steps:

1. Select uniformly random time slices on the current trajectory \( o \), with the probability given by Equation 4.3, until a random number \( \alpha \in [0, 1] \) is smaller than \( p_{gauss}(\lambda_i, \sigma) \).
   Time slice \( \tau_{sp} \) fulfilling this condition is the new shooting point for new trajectory \( n \).

2. Draw a random number \( \xi \in (0, 1] \) and calculate the maximum allowed sum of weights \( W^n_{max} \) for the new path from \( W^n_{max} = W^o/\xi \).

3. With probability \( p_{2way} = 0.2 \) start two molecular dynamics simulations: for a forward move from unchanged timeslice \( \tau_{sp} \); for a backward move reverse the momenta\(^2\). Go to point 5.

4. If two way shooting is not performed, decide to shoot forward or backward with the same probability and reverse the momenta of the shooting point in case backward shooting was accepted.

5. Continue the MD simulation, with Andersen temperature coupling as defined in Molecular Dynamics section, of Andersten thermostat and an Nosé-Hoover MD. until the sum of weights of the resulting trajectory (after gluing with a part of the old trajectory \( o \) in case of one way shooting or gluing backward and forward trajectories in case of the two-way shooting) exceeds \( W^n_{max} \) or one of the two stable states is reached.

\(^2\) In case of the Leap Frog algorithm, the reversal of the momenta consists of integrating the system half timestep in order to obtain \( v(t + \frac{1}{2} \Delta t) \). The reversed timeslice is of the form \( (x(t), -v(t + \frac{1}{2} \Delta t)) \).
6. In case the resulting trajectory is not of the form \( A \rightarrow \lambda_i \rightarrow A \) or \( A \rightarrow \lambda_i \rightarrow B \) we reject it. Otherwise, the trajectory is accepted and trajectory \( n \) becomes the current trajectory. In both cases, we restart the whole procedure with point 1.

4.3.3 Analyzing Reaction Coordinates

An order parameter can be considered a good reaction coordinate if it describes the progress of a reaction. A function, that by construction does so, is the committor \(^{[9]}\). Calculating the committor function consists of shooting a high number of trial trajectories from the timeslices along a transition path \(^{[9]}\). This procedure, also known as \( p\)-fold analysis is computationally extremely expensive. Secondly, even though the committor function yields the transition states, it is an abstract coordinate that does not bring any insight in the mechanism of the reaction. This is the reason why it is insightful to look for the order parameter, that would approximate the committor well enough, but would still be a simple function of the coordinates of the system.

Peters et al. \(^{[25, 24]}\) have recently formulated an approach, allowing to choose the order parameters best reproducing the commitment probability. In the algorithm, known as Maximum Likelihood Estimation (MLE), a number of linear combinations of the available order parameters are tested for the best correlation with the committor function, based on transition path sampling ensemble. For the existing TPS ensemble, the method yields insight in the reaction coordinate and allows to approximate the transition states at no significant additional computational expense.

As an input one can extract from the path ensemble the forward shooting point configurations \( x \) belonging to trajectories ending in the final states. Using these configurations the MLE predicts the committor from the reaction coordinate \( r \) by a \( \tanh \) function \( pB(x) = \frac{1}{2} + \frac{1}{2}\tanh(r[q(x)]) \). The trial reaction coordinate is estimated as a linear combination of a maximum of 3 order parameters \( q, r(q(x)) = \sum_{i=1}^{n} a_i q_i(x) + a_0 \), where \( n = 1, 2, 3 \). For the shooting point configurations \( x \) we compute all of the parameters defined in Section 4.4 and try any combination of up to 4 OP's.

4.3.4 Forward Flux Sampling

Forward Flux Sampling (FFS) \(^{[2]}\) is a method designed for sampling stochastic dynamical pathways connecting two stable states separated by a free-energy barrier and calculate the rate constant for the transition. The method is similar to TIS in the sense of using the interfaces. However, the methods differ in the way the new trajectories are constructed. In FFS one shoots only forward, and the integration is ceased whenever the trajectory reaches the initial state or the next interface. In contrast, TIS trajectories are always anchored in the stable states. More details on FFS can be found in Section 2.7.

4.4 Order Parameters

All TPS-based algorithms, including TIS rely on the proper definition of the stable states. The order parameters used in these definitions should not only distinguish between the stable states but also should be representative for these states \(^{[9, 3, 5]}\). We obtain the
4.5. RESULTS AND DISCUSSION

<table>
<thead>
<tr>
<th>OP</th>
<th>$N_{\text{min}}$</th>
<th>$N_{\text{max}}$</th>
<th>$L_{\text{min}}$</th>
<th>$L_{\text{max}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rmsd (nm)</td>
<td>0</td>
<td>0.25</td>
<td>0.45</td>
<td>0.8</td>
</tr>
<tr>
<td>rmsd$_{hx}$ (nm)</td>
<td>0</td>
<td>0.05</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>sas (nm$^2$)</td>
<td>17</td>
<td>18.5</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.75</td>
<td>0.90</td>
<td>0.20</td>
<td>0.50</td>
</tr>
<tr>
<td>$nw_{trp}$</td>
<td>0</td>
<td>7</td>
<td>12</td>
<td>25</td>
</tr>
</tbody>
</table>

Table 4.1: Order parameters (OP) defining the upper (max) and lower (min) boundaries of the stable state N (native) and L (loop).

set of state-defining order parameters from straightforward MD and REMD simulations. For Trp-cage, in all of the simulations, we monitor the following order parameters: the protein radius of gyration using the $\alpha$-carbons only ($rg$), the fraction of native contacts ($\rho$), the root mean square deviation from the native $\alpha$-carbons structure ($rmsd$), the root mean square deviation of the $\alpha$-helical residues 2-8 from an ideal helix ($rmsd_{hx}$), RMSD of the hydrophobic core: the tryptophan and the prolines 12 and 17-19 ($rmsd_{core}$), the solvent accessible surface (sas) of the whole protein, the distance ($sb$) between donors and acceptors in the hydrogen bonds of the salt-bridge between Arg-16 and Asp-9 and the number of water molecules around tryptophan ($nw_{trp}$). All distances are expressed in nm.

4.5 Results and Discussion

4.5.1 Rate Constant Calculation

As we can tackle only one barrier at a time with TIS, we have chosen the most likely one of the two possible unfolding pathways of Trp-cage, namely the $N \rightarrow L$ route (Figure 4.1). On this route the protein unfolds the helix and water solvates the core while the overall U-shape, tertiary contacts and small size are preserved. Based on our TPS ensemble, we know, that this route has a barrier lower than the $N \rightarrow I$ transition by approximately $1 \ k_B T$, and there are no additional intermediates on the route. We have performed two sets of TIS simulations, one for the unfolding ($N \rightarrow L$) (TIS-unf simulation) and another one for the folding transition ($L \rightarrow N$) (TIS-fol simulation). The stable state definitions are given in Table 4.1. The order parameter we chose to describe the interfaces was the helix RMSD: $\lambda \equiv rmsd_{hx}$. This OP distinguishes the two states. The flux factor $f^{1,0}_N$ (Equation 4.1) was calculated based on 10 ns long MD simulations in the native state. The flux factor $f^{1,0}_L$ was estimated on the same amount of molecular dynamics as for the folded state. Ten structures for the loop state flux calculation were picked randomly from the endpoints of the TIS trajectories. When we calculate the flux, we count only the crossings on the way from the stable state through the given interface: the effective positive flux [29]. After each recrossing event we check whether the trajectory relaxes back to the stable state (crosses through $\lambda_0$), before a new crossing event can be counted. The procedure yielded a forward flux $f^{1,0}_N = 6.7[nm^{-1}]$ through the interface $\lambda_1 = 0.06[nm]$ for the native state and a reverse
process \( f_L^{\lambda,0} = 1.0 [n s^{-1}] \) through the interface \( \lambda_i = 0.23 [\mu m] \) for the loop state.

For the calculation of the crossing probability \( P(\lambda_L | \lambda_N) \), we defined the following interfaces
\( \lambda_i = rmsd_{hx} = 0.06, 0.08, 0.10, 0.13, 0.15 \) and 0.17. For the reverse transition crossing probability \( P(\lambda_N | \lambda_L) \) we chose \( \lambda_i = rmsd_{hx} = 0.23, 0.19, 0.17, 0.15, 0.12 \) and 0.10. For each of the above interfaces we have performed a TIS simulation, resulting in an ensemble of trajectories of the form \( N \rightarrow \lambda_i \rightarrow N \) or \( N \rightarrow \lambda_i \rightarrow L \) for the unfolding process and \( L \rightarrow \lambda_i \rightarrow L \) or \( L \rightarrow \lambda_i \rightarrow N \) for the folding. The statistics of all ensembles are presented in Table 4.2.

For each of the interfaces we can plot the crossing probability as a histogram of \( \lambda \). By matching and reweighting these histograms we obtain the total crossing probability curve (Figure 4.4). When plotted on a log scale, the functions \( P(\lambda | \lambda_N) \) and \( P(\lambda | \lambda_L) \) both reveal a plateau beyond certain value of \( \lambda \). The appearance of the plateau is a consequence of having crossed the transition state. Beyond (or below) certain value of \( \lambda \), the trajectories are committed to the final state, and thus the crossing probability becomes constant. The value of the plateau is equal to the total crossing probability. From the TIS simulations \( P(\lambda_L | \lambda_N) = 1.2 \times 10^{-4} \) and \( P(\lambda_N | \lambda_L) = 2.5 \times 10^{-3} \). These results give the following rates for unfolding and folding:

\[
  k_{NL} = (1.2 \mu s)^{-1} \quad k_{LN} = (0.4 \mu s)^{-1}
\]

These values can be directly compared to the experiment:

\[
  k_{unf}^{exp} = (12.7 \mu s)^{-1} \quad k_{fol}^{exp} = (4.1 \mu s)^{-1}
\]

The error of these numbers is difficult to estimate, but should not be higher than of the order of 1 \( k_B T \). The calculated rate constants yield the free-energy difference of \( \Delta G_{NL} = ln \left[ \frac{k_{NL}}{k_{LN}} \right] \approx 1 k_B T \) between the folded and the intermediate state. This value is the same as the free-energy difference between the native and unfolded state \( \Delta G_{NU}^{exp} = 1 k_B T \) as seen in experiments. However, we have to keep in mind, that the experimental results are
4.5. RESULTS AND DISCUSSION

Figure 4.4: Left: crossing probabilities for both $N - L$ and $L - N$ transitions as a function of the TIS order parameter ($rmsd_{hx}$) [nm]. The data points were fitted with polynomials of the order 7. Right: schematic representation of the calculated $N - L - U$ unfolding route (black solid line), compared to experimental measurements (dotted line). The calculated unfolding rate is lower by about $2k_B T$ than the experimental one. The folding rate differs with experimental measurement only by $0.5 k_B T$.

relative to the unfolded, not the loop state. From our Replica Exchange simulation of Trp-cage [14] the free-energy difference between the loop and unfolded states was estimated to be $\Delta G_{UL} \approx 1.5 k_B T$. Using this value, the computed free-energy difference between the folded and unfolded state equals to $\Delta G_{NU} = \Delta G_{NL} - \Delta G_{LU} \approx -0.5 k_B T$. The discrepancy of $1.5 k_B T$ with the experimental value might be due to the OPLSAA forcefield. We speculate that the lower stability of the native state of Trp-cage in the OPLSAA force-field ($2.3 k_B T$ difference with experiment) may be an OPLSAA force-field issue. Interestingly, the folding seems to agree with experimental measurements. After correcting the rate constant $k_{LN}$ by the free-energy difference between the $L$ and $U$ state $\Delta G_{UL}$, we obtain $k_{UN} = k_{LN} e^{-\Delta G_{UL}} = (1.8 \mu s)^{-1}$ (see Figure 4.4). This value differs only by a factor of 2 from the experimentally measured folding rate. This is presumably within the error of the computation.

4.5.2 Multiple Pathways

During the TIS simulation we have encountered several problems, related to the following observations:

- There are two distinct pathways for the (un)folding process, and when we start the TIS simulation for an interface close to the initial state, there exists a non-negligible probability that the system will choose the other pathway, that we want to exclude.

- Parameter $\lambda = rmsd_{hx}$ does not distinguish between the native state and the other intermediate state I. This problem is persistent to the folding TIS simulation, started in state $L$, may easily end up in state I.

- There is a close-to-native metastable state ($P_d$), which is on-pathway for the $N - I$, but not for the $N - L$ transition. For the interfaces $\lambda < 1 \text{ Å}$ the system is sometimes attracted to this metastable state $P_d$, rather than to the native state $N$. 

Figure 4.5: The TPS ensemble of the $N \rightarrow L$ transition (a,d) versus the TIS ensembles of the $N \rightarrow L$ (b,e) and $L \rightarrow N$ (c,f) routes for their extreme interfaces respectively in two representations: $\text{rmsd}_{hx}$, $\text{rmsd}_{ca}$ (d,e,f) Red color bin means at least 70% of pathways were crossing through that bin, white - no pathways. Interfaces have been demarked with vertical lines for the TIS ensembles. The black thick solid line in the middle of the plots connect the native state, characterized by $\text{rmsd}_{hx} = 0.05$, $\text{nw}_{\text{trp}}$ 9 and $\text{rmsd}_{ca} = 0.19$ with the $L$ state, having unfolded helix ($\text{rmsd}_{hx} = 0.23$), more waters within the cut-off distance of the Trp-6 ($\text{nw}_{\text{trp}}$ 15) and $\text{rmsd}_{ca}$ 0.35. Additionally in plot (d) $\text{RC}_{NL} = \text{const}$ lines have been shown (see Section 4.6). The FFS ensemble is presented in the same representations in plot (g,h), together with the black lines indicating reaction coordinates. Unfolding crossing probability curves obtained for the order parameter $\lambda_i = \text{rmsd}_{hx}$ for FFS (red points) and TIS (black line) are plotted for comparison (i). In all plots, RMSD parameters are expressed in nm.
4.5. RESULTS AND DISCUSSION

| \( i \) | \( \lambda_i \) | \( \lambda_{i+1} \) | \( P(\lambda_{i+1}|\lambda_i) \) | \( N_{\text{traj}} \) |
|---|---|---|---|---|
| 0 | 0.06 | 0.08 | 0.1914 | 789 |
| 1 | 0.08 | 0.10 | 0.0829 | 1810 |
| 2 | 0.10 | 0.11 | 0.1112 | 2698 |
| 3 | 0.11 | 0.12 | 0.1250 | 1108 |
| 4 | 0.12 | 0.13 | 0.1472 | 1019 |
| 5 | 0.14 | 0.16 | 0.1708 | 896 |
| 6 | 0.16 | 0.17 | 0.6085 | 493 |
| 7 | 0.17 | 0.18 | 0.7979 | 376 |
| 8 | 0.18 | 0.19 | 0.8779 | 374 |
| 9 | 0.19 | 0.20 | 0.8761 | 347 |
| 10 | 0.20 | 0.22 | 0.7143 | 35 |

Table 4.3: The summary of FFS results: Crossing probabilities \( P(\lambda_i|\lambda_{i+1}) \) and the total number of generated trajectories

We tried to circumvent the above mentioned problems by carefully monitoring our TIS simulations. In case a TIS run switched to sample a different free-energy barrier we rejected it and restarted at the previous step. We also use TPS trajectories connecting both states via \( N – L \) route as an input for each of the interfaces, to ensure the sampling on the correct barrier.

4.5.3 Forward Flux vs Transition Interface Sampling

We performed an FFS simulation for Trp-cage starting from the native state, using \( rmsd_{hx} \) as the order parameter (\( \lambda \)). The interfaces we used are presented in Table 4.3. This set of interface values was obtained recurrently by trial and error. In case we were not able to reach the subsequent trial interface often enough, we decreased the gap in \( \lambda \), until the desired minimal ratio was approximately reached. Our arbitrary choice was \( P_{\min}(\lambda_i|\lambda_{i+1}) \approx 0.1 \). When the probability of crossing through the next interface was higher than 0.1, we continued the simulation with the next interface. The conditional probabilities \( P(\lambda_i|\lambda_{i+1}) \) for the resulting set of interfaces are presented in Table 4.3. The goal of this study was to try sampling the \( N – L \) transition with a potentially faster method than TIS, while calculating the rate constant at the same time. The transition path ensemble in the \( rmsd_{hx} – nw_{\text{trp}} \) and \( rmsd_{hx} – rmsd_{ca} \) planes together with the crossing probability curve are presented in Figure 4.5. The FFS crossing probability is \( 4 \times 10^{-3} \) lower than the one obtained with TIS. This factor increases the unfolding free-energy barrier by \( 8 k_B T \) resulting in the rate constant \( k_{NL}^{\text{FFS}} = (632 \mu s)^{-1} = 49 \times k_{\text{uns}}^{\text{exp}} \). This dramatic decrease of the rate constant (2400 \( \approx e^8 \) fold) arises because FFS did not sample the correct barrier. By increasing the \( rmsd_{hx} \) the system was biased to unfold the \( \alpha \)-helix (see Figure 4.5-g,h). This did not occur via the lowest free-energy path possible. On the contrary, the barrier crossed was higher by \( 8 k_B T \) than the one found with TIS. In some cases, the protein completely unfolded without even visiting the \( L \) state, showing that indeed, direct \( N – U \) transitions are possible, although very improbable. None of the trajectories ended up in the \( L \)-state, and the provisional p-fold calculation showed their endpoints are either committed to the \( U \) or \( N \) state.
It is interesting to note the differences in the transition path ensembles for the FFS and TIS/TPS methods (Figure 4.5). On the FFS pathways, the $\alpha$-helix unfolds from the N-terminus. Even when the whole helix is solvated the Trp-6 still stacks in between the Proline residues. This is the cause of the low values of $nw_{trp}$, which are essentially constant on the pathways. In contrast the TIS ensemble showed a slow but steady solvation of the hydrophobic core. Even though the two simulations were started from the same equilibrated PDB structure (the TPS was started from an unfolding pathway initiated with this configuration), it seems that the initial state for the two cases is different. Pathways belonging to the TPS ensemble decrease the overall RMSD slower than the ones from the FFS ensemble. The FFS pathways are all anchored in the initial configuration, while the TPS paths relax the initial state within the allowed definition. It seems that the last step of folding of Trp-cage is a final rearrangement of parts of the backbone not belonging to the helix, causing the 1 Å $rmsd_{ca}$ difference between the initial configurations of the TPS ensemble and FFS pathways. Nonetheless, FFS does not allow the pathways to increase the $rmsd_{ca}$ at the beginning of the simulation. This problem may be solved by moving first interface further from the initial state, but then the FFS method would become much less efficient. Both methods are in principle sampling the same ensemble. In principle the FFS method should eventually relax to the proper transition path ensemble. but this might be problematic if there exist two “valleys”, separated by a free-energy barrier in an orthogonal direction to the order parameter used. If this is the case, and the order parameter $\lambda$ is not the best reaction coordinate, then the FFS method might channel all pathways to the nearest valley, even if the free-energy barrier will eventually turn out higher. Our implementation of the TIS algorithm does not have this problem as we guide our ensemble in the right valley, using initial TPS trajectories, anchored in the final and initial states. On the other hand, TIS is less efficient than FFS in the sense of the generation of the number of trajectories. the TIS trajectories are more decorrelated from each other than the FFS ones.

4.6 Analysis of Reaction Parameters

We have divided the whole TPS path ensemble in two parts: one consisting of the $N-I$ and the other of $N-L$ pathways, and we then subjected both subensembles to the MLE procedure [24]. For the N-I subensemble the single most committor-correlated OP appeared to be $rmsd_{ca}$ No significant improvements were obtained for combinations of two trial order parameters. The resulting reaction coordinate is $rc_{NI} = -3.7 + 12rmsd_{ca}$, where the RMSD is given in nm. For the $N-L$ subensemble, the highest scoring single OP was the helix RMSD $rmsd_{hx}$. By adding another OP to our trial reaction coordinates, we were able to increase the maximum likelihood by a significant amount [24] for the combination of $rmsd_{hx}$ and $rmsd_{ca}$. The reaction coordinates of the 3-rd order did not result in significant improvement. The reaction coordinate for the $N-L$ route can thus be written as $rc_{NL} = -4.5 + 13rmsd_{hx} + 8rmsd_{ca}$.

The TIS ensemble density maps are plotted in Figure 4.5. Although they overlap with the TPS ensemble connecting both states, the interfaces furthest from the respective initial states ($\lambda = 0.17$ for A and $\lambda = 0.10$ for B) do not exactly coincide with the transition state region ($rmsd_{hx} \approx 0.15 \pm 0.025$). This is because the TS-ensemble is quite broad in $rmsd_{hx}$. 
4.7 Conclusions

The form of the reaction coordinate for the N-L transition is probably the reason of our problems with sampling the folding TIS ensembles, as we did not include the $rmsd_{ca}$ in the parameter $\lambda$. This information would have been useful, especially for the folding rate calculation as with this choice ($rc_{NL}$) we would exclude the $U-I$ transitions. Nonetheless successful sampling was still possible using a single $\lambda$.

4.7 Conclusions

The unfolding rate constant calculated with the OPLSAA force-field is one order of magnitude higher then the measured experimental value, while the folding rate, including a minor correction agrees well with the experiment. This discrepancy is most probably an OPLSAA force-field related issue. The native state appears to be less stable than the unfolded state with a free-energy difference of about $2k_B T$. The lower stability of the native state of Trp-cage in the OPLSAA force-filed has already been observed previously [16].

The TPS and TIS ensembles follow the pathways corresponding to the lowest free-energy barriers. In contrast, Forward Flux Sampling resulted in serious overestimation of the free energy barriers and hence underestimation of the rate constant, because of the channeling of paths into the wrong direction. This is not caused by the fact that FFS is in principle wrong, but in practice more sensitive to the choice of order parameter that TIS. Application of Likelihood Maximization for the the TPS ensemble revealed that the reaction coordinate is a combination of the $rmsd_{hx}$ and the $rmsd_{ca}$. Using this rc instead of only the $rmsd_{hx}$ might improve the sampling of TIS, and will almost certainly improve the FFS results.

A future study might test the proposed reaction coordinate thoroughly by committer analysis. Also, TIS can be used to compute the rate for the other transitions in the Trp-cage system, i.e. the $N-I$, $I-U$, $L-U$ and possible transitions to misfolded states.

As a final remark, we plan to apply the methodology to other proteins.

Bibliography


[10] Ding, F., Buldyrev, S., and Dokholyan, V. Folding Trp-cage to NMR resolution native struc-
ture using a coarse-grained protein model. *Bio-


[16] Laio, A. personal communication.


