Chemical Physics 396 (2012) 30–44

Contents lists available at ScienceDirect

Chemical Physics

journal homepage: www.elsevier.com/locate/chemphys

Transition path sampling of protein conformational changes

Jarek Juraszek, Jocelyne Vreede, Peter G. Bolhuis *

Van't Hoff Institute for Molecular Sciences, University of Amsterdam, P.O. Box 94157, 1090 GD Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 8 February 2011
In final form 28 April 2011
Available online 6 May 2011

Keywords:
Simulation
Rare events
Rate constant
Reaction mechanism
Transition pathway

ABSTRACT

Conformational changes in proteins often take place on long time scales compared to the molecular timescale. These long time scales, related to high free energy barriers, make such processes difficult to access with a straightforward molecular dynamics approach. The transition path sampling technique has been developed to overcome such timescale differences without assuming a predefined reaction coordinate. We review the transition path sampling methodology with the application of protein conformational change in mind. Using three case studies, based on previous work, we elucidate the strengths and pitfalls of the method. First, the extensive work on the folding of Trp-cage reveals how to sample parallel pathways, how to obtain rate constants, and how to extract reaction coordinates. The second case-study, on the folding of Trpzip4 β-hairpin, illustrates how to treat long-lived intermediates in a conformational change. The final example showcases the light-triggered conformational transition of Photo-active Yellow Protein into its signaling state, highlighting the wealth of insight that can be gathered from a transition path sampling approach, including new hypotheses for reaction mechanisms. We end with an outlook discussing future developments and application of the methodology.

© 2011 Elsevier B.V. All rights reserved.

1. Rare event simulations of proteins

1.1. Simulation of protein conformational changes

Proteins are the machines of life, by performing many different functions, such as for example catalysis by enzymes, sensing and signaling, regulation, maintaining cellular integrity by the formation of structural networks, and trafficking cargo by molecular motors [1]. Protein function requires the protein to undergo a conformational change in some form. Enzymes rearrange to accomodate substrate binding and catalysis of a reaction. Conformational changes affect the binding affinity of regulatory proteins, whereas allosteric transitions enable (membrane) receptor proteins to activate signaling cascades. It is clear that proteins are very dynamical objects rather than static structures and that their function strongly depends on their dynamics. This dynamical nature already becomes apparent at the start of a protein's life cycle. Following synthesis in the ribosome, the polypeptide chain emerges in a linear fashion into the cytoplasm, where it has to fold in a specific three-dimensional shape in order to become functional. Proteins that fold incorrectly can lead to neurodegenerative diseases [2]. The initial folding process of a protein can thus be considered as the archetypal conformational change.

The use of molecular simulations has given much insight in the stability and dynamics of proteins. Proteins are in fact only marginally stable, because the energetically favorable native contacts are offset by the loss of solvent interaction and conformational entropy [3]. Such marginal stability is also the reason why conformational transitions are important in signalling or regulation. A relatively small trigger such as the binding of a ligand, an electron transfer, or the absorption of a single photon by a chromophore, can cause a large conformational change in the protein. To describe this delicate balance between intra and intermolecular forces in a simulation requires an accurate force field. Most (semi) empirical atomistic protein force fields, e.g. ENCAD [4,5], AMBER [6], CHARMM [7,8], GROMOS [9,10], and OPLSAA [11], assume a potential form including the bond, angle, torsion, van der Waals, and electrostatic terms, for which the parameters are optimized using ab initio calculations, and/or experimental data. The water model e.g. TIP3P or SPC is often included in the force field. While much progress has been made [12] and current force fields can reliably reproduce structural features, the prediction of accurate relative stability (i.e. free energy) is still elusive [13].

1.2. Conformational changes as rare events

As the energy landscape of a solvated protein is rough, with a myriad of saddle points in the energy landscape, analysis of the potential energy is not practical. Moreover, the intricate interplay between conformational entropy and energy requires an approach based on statistical mechanics. Molecular dynamics (MD) provides a practical statistical mechanical sampling of phase space, and has become indispensable for obtaining microscopic insight in the

* Corresponding author.
E-mail address: p.g.bolhuis@uva.nl (P.G. Bolhuis).

0301-0104/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemphys.2011.04.032

structure and dynamics of proteins. MD numerically integrates the deterministic Newtonian equations of motion based on the instantaneous intermolecular forces between all atoms [14,15] yielding a time series of molecular configurations, or trajectory. This trajectory contains all equilibrium and dynamical information, and can be used to extract kinetics. MD has proved a very successful tool to study protein dynamics and even conformational changes [16].

The interpretation of MD trajectories should be done in terms of ensemble averages, because due to molecular chaos a very small change in the initial configuration will lead to entirely different trajectories after a few picoseconds. However, if the sampling of phase space is ergodic, MD will yield the equilibrium distribution of the system of interest. Still, as many conformational changes are rare events taking place on timescales that are not (easily) accessible for straightforward MD, even a single trajectory will not show the dynamics of interest. These long timescales are often due to the presence of high free energy barriers between (meta) stable states. The crossing of such a barrier is then a rare event, compared to the fundamental dynamical time step (usually femtoseconds). Because the probability for an event to occur within a certain time period obeys a stochastic distribution, one approach is to run many thousands of trajectories simultaneously, and then look for those in which the event actually took place [17]. However, as even the fastest computer currently available can only access simulation times up to a millisecond [18], overcoming high barriers corresponding to longer timescales requires special rare event methods. A first class of rare event methods aims at obtaining the static equilibrium properties of proteins by overcoming the free energy barriers between (meta) stable states through the introduction of artificial dynamics. This class of methods includes parallel tempering [19,20] or Replica Exchange MD (REMD) [21], which makes use of higher temperature to speed up dynamics, but also umbrella sampling and meta-dynamics, which bias the system to the rarely sampled part of phase space, and many others (e.g. temperature accelerated MD [22], flooding [23], hyperdynamics [24], local elevation [25]). In REMD a series of system replicas is run simultaneously at different temperatures ranging from low temperature where conformational changes are rare, to high temperatures at which sampling is much accelerated. A Monte Carlo algorithm allows exchanging temperature between replicas while maintaining the Boltzmann distribution [15,26,21,20]. As a result, REMD heats up the system periodically to help the system escape local minima, and then lowers the temperature again to the temperature of interest. While REMD allows one to sample a rugged energy landscape, for which the order parameters to bias in are not known, the convergence of the method is very slow [27].

Umbrella sampling tries to overcome the free energy barriers by imposing a fixed biasing potential [28]. Meta-dynamics applies a time dependent bias [29]. Such bias potentials need to be defined as a function of an order parameter that reduce the 3 N dimensional system configuration to just one or a few dimensions. Order parameters that are important in protein systems are e.g. the number of native contacts[1], the protein's radius of gyration, the root mean square deviation RMSD from the native state,[2] dihedral angles, solvent accessible surface, number of hydrogen bonds, distances between relevant groups, salt bridges, bonds distances, combinations of bonds, and contact order. However, when the order parameters used to determine the barrier are poor reaction coordinates, the free energy barriers might be severely underestimated, and the predicted molecular mechanism might even be wrong.

1.3. Path based methods

A second class of rare event methods aims at obtaining insight in the dynamics of protein conformational changes by the sampling of reactive pathways. To draw conclusions on conformational changes in proteins, a single MD trajectory is insufficient. Instead, the average behavior of an ensemble of rare event trajectories is needed to predict experimental dynamical observables such as the rate constant. Unfortunately, uncorrelated trajectories for rare events are difficult to obtain, precisely because of the long time scales involved. A solution to this problem is to consider only those parts of the trajectories leading from an initial conformational state to a final conformational state. While there are several techniques that aim to construct a path between two end points, such as milestoning [30], the string method [31], and path metadynamics [32], we focus here on a method that selects dynamical, unbiased rare event trajectories: the transition path sampling (TPS) methodology [33,34]. Sampling these rare paths constitutes only part of the challenge, as the analysis of the resulting path ensemble is equally important to obtain mechanistic insights. Recent developments in reaction coordinate analysis facilitate getting insight in the mechanism and the extraction of relevant order parameters.

1.4. Outline

In the remainder of the paper we give a brief overview of the TPS method (for more elaborate reviews we refer the reader to Refs. [34–37]), and discuss the different options for reaction coordinate analysis and the rate constant computation in the light of protein conformational changes in Sections 2–4, respectively. In Section 5 we present three case studies based on existing work that highlight the advantages of TPS, but also expose the challenges. We end with an outlook in Section 6.

2. Transition path sampling

2.1. The transition path ensemble

Rare event methods that rely on biasing along a predefined order parameter might yield an incorrect free energy, transition state and rate constant, when the chosen order parameter is not a good reaction coordinate (RC), i.e. does not capture the molecular mechanism. As reaction coordinates can be very elusive for complex systems, path based methods such as transition path sampling have been developed to overcome the folding barrier and sample the dynamics of conformational changes without the requirement of a priori knowledge of the reaction coordinate [33]. In principle, a very long straightforward MD with many crossings would be able to solve the rare event problem, but is computationally wasteful as the trajectory spends most of its time in one (meta) stable state and very rarely leaves this state to cross a barrier. While a straightforward MD trajectory yields much information on both the initial state, and – if one initiates the MD simulation there – also on the final state, the barrier region is hardly sampled. The basic idea of transition path sampling is to focus only on those parts of the trajectory that connect both the initial and final states, and hence are crossing the free energy barrier. As an infinite trajectory crosses the barrier an infinite number of times, these parts form an ensemble of crossing paths: the transition path ensemble (TPE). Due to the rarity of the event, this path ensemble constitutes only a small fraction of the infinite trajectory. The transition path ensemble is thus a collection of dynamically unbiased trajectories connecting

[1] A contact is made when the α-carbon atoms of non-adjacent residues are within a 6 Å distance. A native contact is a contact that also occurs in a reference configuration representing the native state, e.g. an experimentally determined structure.

[2] Computing the average square atomic distance between the two structures yields the mean square distance (MSD). The RMSD follows from minimizing this MSD with respect to translation and rotation, and finally taking the square root.

the initial and final stable states [33,35,34]. Because only the stable states themselves need to be defined, the path ensemble circumvents the problems related to defining a reaction coordinate. Transition path sampling samples the path ensemble by performing a Monte Carlo importance sampling scheme, in which an existing pathway connecting initial and final state is altered, followed by accepting or rejecting this new trial pathway according to a proper acceptance rule [35].

Out of many possible Monte Carlo algorithms that can sample the TPE the shooting move turns out to be both simple and efficient [34,35]. The shooting move starts by randomly selecting a certain time slice on the current trajectory that connects the initial to final state, and change its momenta slightly. From this time slice, known as the shooting point, a new trajectory of the same length is created by integrating the equations of motion both forward and backward in time. In the simplest implementation of the algorithm the new trial trajectory is accepted if it connects the initial with the final region. Otherwise it is rejected and the old path is kept. The shooting move is then repeated with a different shooting time slice. The resulting random walk in path space results eventually in a collection of properly weighted pathways representative for the TPE.

Mostly, we are interested in the parts of the trajectories that leave the initial stable state, cross the barrier and enter the final state. In fact, for complex systems only shots initiated from the barrier region rather than the stable state regions have a chance of creating acceptable pathways. Hence, it makes sense to stop integrating the equations of motion by MD when a stable state has been reached. As the trajectories are assumed to have a low probability to recross after reaching such a state, the stable states definitions should be (slightly) stricter [38,39]. On the other hand, the *shifting moves*, introduced in the original implementation of TPS [33] are redundant for this implementation of the variable path length TPS. The variable path length shooting move has the following Metropolis acceptance rule [38]:

$$P_{acc}[o \to n] = h_{AB}^{(n)} \min\left[1, \frac{L^{(o)}}{L^{(n)}}\right]. \quad (1)$$

where L is the path length and h_{AB} equals unity if the path connects the initial and final state, and zero otherwise. The subscripts o and n refer to the old and new path respectively. The flexible path length approach has the advantage of saving considerably on simulation time, as is clear from plotting a distribution of pathways (see Fig. 1).

Note that TPS only samples trajectories but does not lead to immediate insight in the mechanism. For that, the resulting TPE should be analyzed in more detail. The major advantage of TPS is that one does not have to impose reaction coordinates on the system, but rather extracts these from the simulation results.

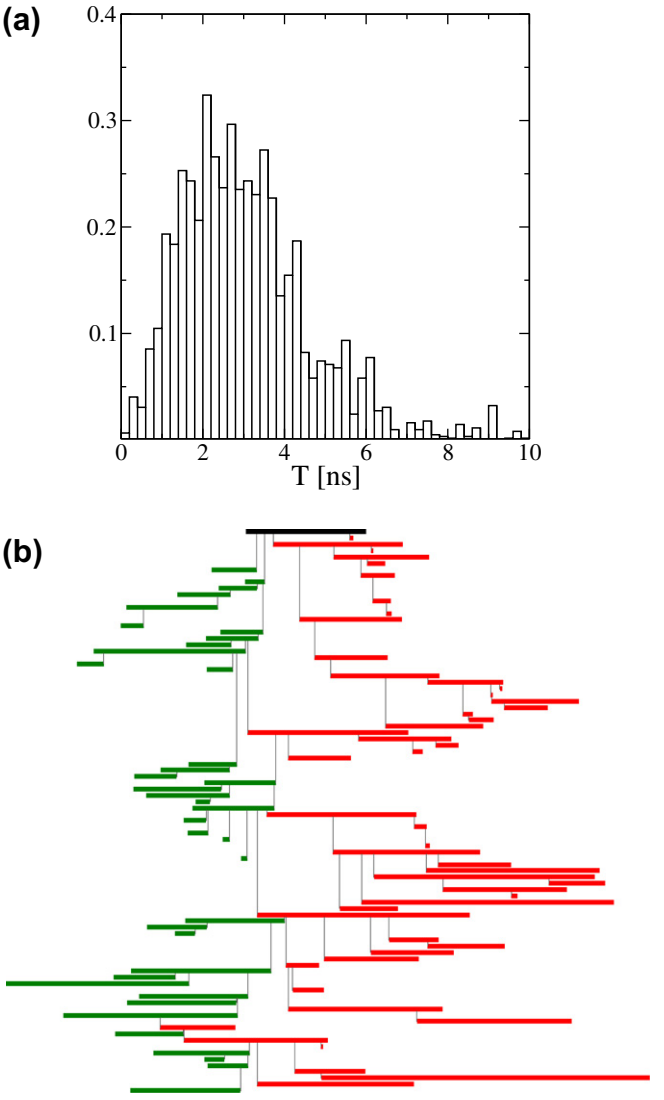

Fig. 1. Analysis of transition path sampling for the Trp-cage system described in Section 5.1. Left: the distribution of path lengths show a relatively long tail, indicating that the flexible path length shooting move is very efficient. Right: sampling trees can give an indication of the quality of the stochastic shooting. Starting with the initial path (top horizontal solid line), shooting points are indicated by thin vertical lines. The accepted paths are plotted as a horizontal line (forward shots to the right, backward shots to the left), and the Monte Carlo shooting moves proceed in the vertical direction. As only the path changes are shown, the quality of sampling follows from how often a backward-shot line starts from a forward-shot line and vice versa. The least changed path, connecting these shooting points gives an indication of the barrier sampling.

2.2. The one-way shooting algorithm

The shooting move works even on rough energy landscapes because the trial shooting point is only slightly different from the old one, and the new trial path stays close to the previous path, while the basins of attraction of the stable states make the trajectory commit to either state. Molecular chaos ensures that the paths are sufficiently different from each other to allow proper sampling of path space. However, for barriers that are both wide and have a rough energy landscape the TPS shooting algorithm can become less efficient because the old and the new trajectories will diverge completely before they are able to commit to the correct stable states. Even with a very small change in momenta on the order of machine precision both the forward and the backward trial path can return to the same stable region, leading to a very small number of accepted trajectories [40]. While in principle the motion is deterministic, in practice it is not possible to stay arbitrarily close to the old path using regular MD, a requirement for the efficiency of two-way shooting move. That is why Grünwald et al. proposed a scheme to keep a new trajectory arbitrarily close to the old one [41].

An implementation of the one-way stochastic dynamics shooting move [35] allows acceptance of a forward or backward shot independent of each other. The acceptance is much better (usually around 50%) than for the full two-way shooting because part of the path already connects one state [40]. Naturally, the price to pay is a slower decorrelation of paths along the Monte Carlo random walk. Moreover, in contrast to two-way shooting, decorrelation is not guaranteed, as some parts of the path might not or hardly change during the sampling. A posteriori analysis of the path ensemble is necessary to spot such behavior, e.g. by a sampling tree indicating which parts of the path have changed (see Fig. 1).

The one-way shooting algorithm only works for dynamics that is (partly) stochastic. One way of introducing stochasticity in an otherwise deterministic dynamics is a coupling between the system and a heat bath employing a Andersen-like thermostat [42]. This thermostat selects, at random times along the trajectory, a random atom and draws new momenta from a Maxwell–Boltzmann distribution. Note that while the thermostat keeps the temperature constant along the trajectory, it destroys the Hamiltonian nature of the dynamics. Also it does not preserve energy, or linear or angular momentum along the trajectory. Care should be taken that the coupling to the thermostat does not change the dynamical behavior of interest. Therefore, the coupling constant should be as small as possible [43]. A small coupling constant related to a sufficiently low frequency of reinitializing the momenta of a randomly chosen molecule guarantees that the dynamics is almost indistinguishable from the dynamics in constant temperature deterministic trajectories, while at the same the paths can diverge in a controlled way so that the stochastic path sampling algorithm can be applied. The dynamics of the system is even less disturbed by only altering the momenta of the solvent molecules. The easiest implementation is then to change only the linear momentum, keeping the molecular angular momentum intact [43]. Implementing the Andersen thermostat conserves the canonical distribution and hence the acceptance criterion is given by Eq. 1. If it is important to conserve linear momentum, one can implement the Lowe–Andersen thermostat [44]. Another thermostat that can be used to introduce stochasticity in the trajectories is the recently developed velocity-rescale thermostat [45].

2.3. Stable state definitions

TPS does not rely on the assumption of a reaction coordinate, but instead requires the definition of the initial and the final states in terms of order parameter(s) $\lambda(x)$, which reduces the high dimensional configuration x to a single quantity or number. The indicator function for a stable state is $h_A(x) = 1$ for $\lambda_{A,min} < \lambda(x) < \lambda_{A,max}$, and $h_A(x) = 0$ otherwise. The connectivity indicator in Eq. 1 is then $h_{AB} = h_A(x_0)h_B(x_L)$, with x_0 the configuration of the initial slice, and x_L the configuration of the final slice of the trajectory. Fortunately, this stable state definition does not have to be equivalent to the basin of attraction (the part of configuration space from which trajectories tend to relax to the corresponding stable state), as this requirement is difficult to fulfill. In fact, such a definition would not leave any barrier region for the TPS to sample. Hence, the stable states should parametrize a smaller region lying within the basins of attraction. When choosing stable state definitions it is important to obey the following criteria. (i) The state definitions should not overlap, i.e. no configuration can belong to both states simultaneously. (ii) The state definitions should distinguish *A* from *B*. that is, configurations belong to state *A* should lie in the basin of attraction of *A* and vice versa. (iii) A trajectory that reaches a state should be *committed* to that state, meaning that it is not likely to recross. (iv). A regular MD trajectory sampling in one of the stable states should visit the defined stable state region frequently on a molecular time scale.

While fulfilling criteria *i–iii* requires a smaller, stricter state definition, criterion *iv* demands a larger, less strict definition. In practice, the stable state can be defined by analyzing straightforward MD trajectories of state *A* and *B*. Some intuition is required for such an analysis, but much less than for a reaction coordinate. In fact, there is quite some freedom in the choice of stable state parametrization. For instance, one can use a combination of parameters, or even use different order parameters for each state [46]. The resulting TPE and the kinetic rate constants will not be very sensitive to the definition of the stable states, provided they fulfill the above criteria.

2.4. The initial path

Bootstrapping TPS requires an initial pathway. An initial trajectory can in principle be created by any means available, as long as it is a sequence of phase space points that connect *A* with *B*. By far the easiest way is to make use of the MD engine that produces the trial trajectories. For instance, if the event of interest become less rare at high temperature, one can perform a simulation at an elevated temperate, select the part of the trajectory that contains the event and rescale the momenta down to the temperature of interest. This trajectory, while not a true dynamical path, can serve as an initial pathway. Alternatively, one can perform a biased sampling using umbrella sampling or metadynamics along some approximate order parameter to create trajectories that overcome the barrier. Again, selecting the part of the biased trajectory that connects *A–B* can contain many suitable shooting points, and can serve as initial paths. When the rate constant is also needed, one might consider using transition interface sampling. Finally, one might create a non-equilibrium trajectory by applying targeted or steered MD [47].

The initial path is relaxed to the equilibrium TPE by repeatedly applying the shooting move. Naturally, while an initial path that resembles an unbiased dynamical trajectory will relax relatively fast to the equilibrium path ensemble, a poorly generated initial path might take a large number of shooting moves before equilibration is reached. If the initial path is created in such a way that it follows a mechanism or channel that is very different from the equilibrium one, the path sampling may even never relax to the correct channel. Therefore, whenever possible, the initial path should be generated at conditions close to the conditions of interest.

2.5. Multiple states, multiple routes

TPS assumes that there are only two major stable states, separated by one barrier. The existence of other long-lived intermediates between the stable states will force the path length to be of the order of the life time of the intermediate *I*. This might not present problems as long as the residence time in the intermediate is of the same order or smaller then the molecular time scale to cross a barrier. For longer residence times, the very long pathways will make TPS ineffective. One solution is to treat the transitions $(A \leftrightarrow I)$ and $(I \leftrightarrow B)$ as separate transitions. This naturally requires a state definition of the intermediate.

The number of transitions grows quadratically with the number of intermediates. The resulting network of transitions might be sampled with the multiple state version of TPS [48] in which the initial and final state can be any (predefined) stable state or intermediate.

3. Reaction coordinate analysis of the path ensemble

3.1. Committors and the transition state ensemble

Insight in the mechanism of protein conformational changes does not follow directly from sampling the transition path ensemble by TPS, but requires additional analysis. Protein conformational changes can take place via different pathways through different intermediates and transition states, all leading eventually to the final state. A description of these intermediates and transition states would yield insight in the reaction mechanisms. While intermediates can be characterized by long dwelling times on the molecular time scale, a transition state is less easy to identify for proteins. The usual interpretation of a transition state as a saddle point in potential or free energy is not applicable. Instead, for complex systems

with rough energy landscapes the concept of a transition state is usually based on the committor $p_B(x)$: the probability to reach the final state B from a certain configuration x [49–51].[3] The committor $p_B(x)$ can be estimated by initiating many short trajectories from the configuration x with randomized momenta and determining the fraction that reach final state B. A transition state is then a configuration that has an equal probability $p_A = p_B = 0.5$ to reach either state A or B. The collection of such equicommittor configurations defines the so-called separatrix: the boundary between the basins of attraction of the initial and the final state. The TPE sampled by TPS contains all properly weighted pathways leading from A to B, and hence also contains a respresentative sample of the transition states. This transition state ensemble (TSE) can thus be obtained by determining the equicommittor configurations in the TPE. Note however, that these committor calculations can be quite expensive for any system of real complexity.

3.2. Committor distribution analysis

The committor, being a function of the full 3 N dimensional configuration space, gives by itself not much insight without further analysis. To understand the physical process, the reaction coordinate (RC) should be expressed in a reasonably low number of dimensions. For simple reactions between two atoms, their relative distance is likely to be a good reaction coordinate. In proteins, where all atoms move simultaneously in a complex manner the reaction coordinate for conformational change is not as straightforward. Fortunately, the concept of the committor allows testing of an order parameter $q(x)$ as a candidate reaction coordinate. To act as a good reaction coordinate $q(x)$ should be able to parametrize the committor, and hence also the TSE. Consider the situation that q^* denotes the maximum of the free energy barrier between A and B. If $q(x)$ is a good reaction coordinate, a collection of configurations x at a fixed $q(x) = q^*$, for instance obtained by restrained MD, would have committors close to $p_B = 0.5$. The distribution of committor values $P(p_B)$ is hence peaked around the $p_B = 0.5$ For a poor reaction coordinate on the other hand, the committor distribution $P(p_B)$ will not be unimodal, as configurations with the same value of q can have different committors. In that case, a correct characterization of the TSE requires degrees of freedom, additional to the proposed reaction coordinate $q(r)$ [35]. This committor distribution testing procedure was introduced by Du et al. [49] to study protein folding and by Geissler et al. [52] to study ionic dissociation in water, and was used later on to elucidate the mechanism of various complex biologically relevant reactions [50,53–55]. A sampling error deconvolution [61] can make the committor test less expensive and also provides a coordinate error metric that is independent of the sampling error. A drawback of this approach is that suggestions for the improved reaction coordinate have to be tested again by the computationally expensive committor analysis. In the last few years several novel approaches have been proposed to avoid this expensive recalculation. For instance, Ma and Dinner employed a genetic neural network (GNN) [56,57] that automatically selects the collective variable that best parametrizes the committor form among many possible reaction coordinate [53]. This approach still requires the explicit computation of committors for a set of prospective transition state configurations as input for the GNN analysis. Recently, Qi et al. [58] exploited a network approach to esitimate the commitment probabilities.

Based on Bayesian statistics, Best and Hummer proposed a reaction coordinate analysis that combines the equilibrium (Boltzmann) distribution with transition path ensemble distributions [59,60]. The starting point is the Bayesian relation for the conditional probability $p(\mathrm{TP}|r(x))$ that a configuration with the reaction coordinate $r(x)$ lies on a transition path TP between states A and B:

$$p(\mathrm{TP}|r) = \frac{p(r|\mathrm{TP})p(\mathrm{TP})}{p_{eq}(r)}, \tag{2}$$

where $p_{eq}(r)$ denotes the equilibrium Boltzmann distribution as a function of r, $p(r(x)|\mathrm{TP})$ is the distribution of configurations x with a certain $r(x)$ visited along TPS pathways, and the normalizing factor $p(\mathrm{TP})$ is the overall likelihood to be on a transition path [60]. The conditional probability $p\ (\mathrm{TP}|r)$ is large for values of r that are often transversed in transition pathways but are rarely visited in equilibrium. Thus, a maximum in $p(\mathrm{TP}|r)$ allows one to identify transition states as the configurations that have the largest probability that trajectories passing through them are reactive [59]. For diffusive dynamics $p(\mathrm{TP}|r) = 2p_B(r)(1 - p_B(r))$, where $p_B(r)$ is the committor averaged over all configurations x with reaction coordinate $r(x)$. For a good reaction coordinate $r(x)$, all transition states – which have a maximum $p(\mathrm{TP}|r)$ – should correspond to approximately the same value of the reaction coordinate and thus $p(\mathrm{TP}|r)$ should be peaked around the transition state value of r. When r is a poor reaction coordinate, $p(\mathrm{TP}|r)$ will be more or less featureless, due to the lack of correlation between r and $p(\mathrm{TP}|r)$. The Best–Hummer approach requires a good estimate of the equilibrium distribution in the transition region, which can be difficult to obtain.

3.3. Likelihood maximization analysis

Peters et al. [61,62] proposed a committor analysis technique based on a likelihood maximization (LM) procedure requiring only input from a TPS simulation itself. The starting point is again the probability $P(TP|r(x))$ that a certain conformation x is on a transition path. The basic idea is that each trial shot in a TPS simulation is a realization of a process strongly connected to a committor calculation, which estimates $P(TP|r)$. For protein conformational changes the dynamics is mainly diffusive which simplifies this probability to $p(\mathrm{TP}|r) = 2p_B(r)(1 - p_B(r))$. Any $p_B(r)$ function with a rough sigmoidal shaped curve, smoothly varying from 0 to 1, would give a peak at the transition state value of r and decay to zero away from this peak, as is required for a good reaction coordinate [59]. In the LM analysis approach such a function is the tanh function [62]:

$$p_B(r(x)) = \frac{1}{2} + \frac{1}{2}\tanh\left(r[q(x)]\right). \tag{3}$$

The reaction coordinate $r(x)$ is approximated as a linear combination of different order parameters $q_i(x)$:

$$r(q(x)) = \sum_{i=1}^{n} a_i q_i(x) + a_0. \tag{4}$$

The LM optimizes the coefficients a_i in this linear combination of order parameters to obtain the model for the reaction coordinate r which best describes the committor data. Required as input for the LM procedure is the ensemble of M forward (or backward) shooting point configurations x_{sp} belonging to the accepted trajectories ending in the final state B ($x_{sp} \to B$) and the rejected shooting points ending in the initial state A ($x_{sp} \to A$). Using these configurations, the analysis proceeds by optimizing the log-likelihood:

$$\ln L = \sum_{x_{sp}\to B} \ln p_B\left(r(x_{sp})\right) + \sum_{x_{sp}\to A} \ln\left(1 - p_B\left(r(x_{sp})\right)\right). \tag{5}$$

The LM analysis method facilitates the screening for many different (combinations of) order parameters as candidate reaction coordinates. The likelihood maximization then gives the linear combination of order parameters which reproduce the shooting point data best. In practice the method is first applied to a large set of single

[3] In the protein folding literature, p_B is also known as p-fold.

order parameters. The largest likelihood gives then the collective variable that is the best reaction coordinate. Then, all combinations of two order parameters are tested. If a systematic improvement in the log likelihood is found with respect to the best single order parameter description the new combination of parameters is accepted. The minimal required improvement is given by the Bayesian Information Criterion $\Delta \ln L = 0.5 \ln M$. This procedure is repeated with higher order linear combinations until no improvement is found. Usually 3–4 parameters is the maximum. Non-linear reaction coordinates are also possible [61–63].

4. Rate constants of protein conformational changes

4.1. TPS rate

Traditionally, the transition state theory based Bennett–Chandler (BC) reactive flux method [64,65] computes the kinetic rate constants of rare events in complex systems by multiplying the equilibrium probability to be on top of the free energy barrier with a kinetic prefactor. The equilibrium probability is the exponential of the activation free energy with respect to the stable state. This free energy is usually obtained as a function of the reaction coordinate, for instance by biasing methods such as umbrella sampling. The second factor, the transmission coefficient, accounts for all the dynamical effects including recrossings. The efficiency of the BC approach depends very much on the choice of reaction coordinate, and a poor one might lead to unmeasurably small transmission coefficients.

In the TPS framework rate constants can be obtained by slowly constraining the path ensemble from a completely free ensemble to the path ensemble of interest in a series of path sampling simulations [35]. This calculation requires the re-introduction of an order parameter that slowly constrains the paths to the transition path ensemble. While this approach is much less sensitive to the choice of order parameter than the BC method, it is computationally demanding.

4.2. Transition interface sampling

More efficient is the transition interface sampling (TIS) method [38] which starts by defining a series of m non-intersecting interfaces parametrized by an order parameter $\lambda_i(x)$, with $0 < i < m$. Here, the first interface λ_0 is the boundary of stable state A and λ_m is the boundary of B. TIS relies, just as the BC algorithm, on a factorization of the rate constant in a kinetic prefactor and a probability:

$$k_{AB} = \phi_{10} P_A(\lambda_B|\lambda_1). \tag{6}$$

Here, ϕ_{10} measures the effective positive flux [38] of leaving the initial state and crossing interface λ_1, which is easily achieved in a straightforward MD simulation of the stable state, provided the first interface is not too far from state A. Each crossing of the first interface has to be followed by relaxation to the stable state, before the next crossing can be counted [66]. $P_A(\lambda_B|\lambda_1)$ is the – usually very low – *crossing probability* that measures the conditional probability that once a trajectory has crossed λ_1 it reaches the final state, provided that it came directly from the initial state. For each interface λ_i, path sampling is used to estimate the probability $P_A(\lambda_{i+1}|\lambda_i)$ to reach the next interface λ_{i+1}, under the condition that all trajectories cross the interface λ_i and come directly from the initial state A. Note that in TIS the paths are allowed to return to A, something that is excluded in TPS. The crossing probability

$$P_A(\lambda_B|\lambda_1) = \prod_{i=1}^{m-1} P_A(\lambda_{i+1}|\lambda_i), \tag{7}$$

is the product of all interface crossing probabilities. In practice, all paths are binned in histograms as a function of λ for each interface path sampling. Subsequent matching of these histograms (e.g. using WHAM [67,68]) leads to the desired crossing probability.

4.3. Forward flux sampling

Ten Wolde and coworkers [69,70] developed the forward flux (FFS) method to sample non-equilibrium stochastic dynamics in which the stationary phase space distribution is unknown a priori. TPS and TIS cannot be used for such processes as they demand microscopic reversibility [66]. While developed originally for non-equilibrium dynamics, FFS is also applicable to (stochastic) molecular dynamics simulations [70]. The central rate expression in FFS is that of TIS, Eq. 6, and the flux factor is computed in exactly the same way. The difference between the FFS and TIS methods is how the crossing probabilities are computed. While TIS computes crossing probabilities by importance sampling of the path ensemble, in FFS this ensemble is generated directly. The FFS method starts with an ensemble of phase points that cross the first interface λ_1, obtained from a flux calculation. From this ensemble forward trajectories are initiated until they hit the next interface λ_2 or go back all the way to A. The fraction of trajectories that reach λ_2 yields the crossing probability $P_A(\lambda_2|\lambda_1)$. The points that hit the next interface can be used to repeat the forward shooting procedure for the next interface, and so on, until the final state B is reached. Combining all the local crossing probabilities and the flux as in Eq. (6) yields the rate constant. In addition, gluing the successful shooting trajectories together gives the transition path ensemble from A to B. As the accuracy of the results strongly depends on the quality of the initial ensemble, FFS suffers from error propagation. Such error propagation does not occur in TIS, in which each interface ensemble is allowed to converge independently. Hence, FFS is more dependent on the choice of reaction coordinate than TIS. Berrero and Escobedo developed a least square method that optimizes the reaction coordinate based on an FFS path ensemble [71].

5. Application of path sampling on protein conformational changes

In this section we discuss the application of transition path sampling to protein conformational changes. Most of such work has been focused on the folding of small protein fragments; the GB1 hairpin [46] and its mutant Trpzip4 hairpin [72], the Trp-cage miniprotein [73,43] and the FBP WW-domain [74]. These fast folding proteins have contributed much to the understanding of generic folding mechanisms because they bridge the gap between experiments and computer simulation. Here we will recapitulate our work on Trpzip4 and Trp-cage. In the last section we briefly review the recent application of TPS to the biologically relevant conformational change in photo-active Yellow Protein PYP [75]. We chose these three systems because they clearly show the advantages TPS has to offer with respect to sampling protein conformational changes, but also because these three systems each posed a different challenge. Note that the goal here is to illustrate the highlights and pitfalls of applying TPS to protein systems, rather than to give an exhaustive overview. For a review on application of TPS on biological systems in general, see Ref. [76].

5.1. Trp-cage

5.1.1. Introduction

Neidigh et al. [77] designed Trp-cage (sequence NLYIQ WLKDG GPSSG RPPPS, PDB ID TC5B) to be a fast two-state folding mini-

protein, with a folding rate of $k \approx (4.1\ \mu s^{-1})$ [78]. The 20 residue peptide not only forms secondary structure elements, but also folds into a compact, globular conformation, burying a tryptophan residue in the center. The native structure of the 20-residue polypeptide contains an α-helix (residues 2–8), a 3_{10}-helix (residues 11–14), and a polyproline II helix (residues 17–19) (see Fig. 2). The three helices form a hydrophobic cavity in which Trp-6 is buried. This hydrophobic core is further stabilized by a salt bridge (between residues 9 and 16) as shown in Fig. 2(c).

Trp-cage has become an important model system for studying protein folding by molecular simulation, e.g. in all-atom implicit [79–83] and explicit solvent [84], all-atom Go models [85], and in a coarse-grained model [86]. Since Rhee et al. [87] suggested that the solvent does play a crucial role in protein folding, one that implicit solvent models are not able to capture, most kinetic studies employ explicit solvent. Employing distributed computing, Snow et al. directly accessed the kinetics by initiating many simultaneous simulations, of which a small percentage succeeds in crossing the barrier [80]. Explicit solvent REMD studies simulations in Ref [84] confirmed Trp-cage as a fast two state folder, with an intermediate state which contained two hydrophobic cores. A study by Piana and Laio, using a novel bias-exchange, further corroborated this finding [88]. Pascheck et al. [89] observed complete folding of Trp-cage in explicit solvent using REMD with the Amber force field. Earlier, all-atom implicit solvent MD simulations [79–81] and a coarse-grained model [86] observed complete folding of the protein. Nevertheless, several studies located misfolded states in the implicit solvent computations, indicating a less reliable and efficient folding [81]. Recently, Velez-Vega et al., studied Trp-cage unfolding using the FFS method [90]. In Ref. [73] we presented a TPS simulation of Trp-cage folding and unfolding using the OPLSAA force-field and explicit solvent. In the following we will review this work as a case study for the application of TPS on protein conformational changes.

5.1.2. Definition of stable states and the initial path

The first step in defining the endpoints of the TPS trajectories consists of exploring the free-energy landscape between the stable states, in case of Trp-cage-the native state N and the unfolded state *U*. We chose the REMD method to do this, as REMD does not require predefined order parameters to bias in. To find the collective variables most relevant for the description of stable states, we performed a cluster analysis based on the relative RMSD as a metric, for the room temperature ensemble from REMD, resulting in eight most populated clusters and a "background" ensemble, consisting of highly disordered and unstructured molten-globule configurations. The most populated cluster was the native state of Trp-cage, but we also found the "helical *I* state" and a number of *U*-shaped like structures, denoted collectively "*U* state". Finally, there were a few clusters that did not resemble Trp-cage topology and were clearly misfolded, globular configurations. These clusters were the most distant in terms of RMSD of the native state, whereas the *U*-shaped and the helical I cluster were much closer.

By examining these (meta) stables states with straightforward MD, we found that the RMSD to the native PDB state, the RMSD of the alpha-helix, the solvent accessible surface area, the number of native contacts and the number of water molecules were required to distinguish the native from the unfolded and intermediate states. The boundaries of the state definitions were determined using the criteria in Section 2. For instance, if the native state fluctuates around 1 Å RMSD of the PDB structure, we could use the RMSD values [0,0.8] Å to define the stable state *N*, if 0.8 Å is visited, say, roughly on average every 10 ps. We had to use slightly stricter values, in order to be sure that the correct transition is sampled [73].

A high temperature MD trajectory (500 K) leading from the native state, via the intermediate *I* state, to the fully extended conformation, served as initial path. We then equilibrated this trajectory in short TPS simulations, until the first entirely room-temperature trajectory was obtained. These trajectories were then used to seed the subsequent TPS simulations.

5.1.3. Interpretation of the TPS results

The TPS simulation reported in Ref [73] consisted of 3200 trial shootings with an acceptance ratio of 44%. The production part of the TPS procedure required 8.5 μs of simulation time, and resulted in an ensemble of 3200 pathways, on average 3.1 ns long (representing an aggregate simulation time of 10 us). Analysis of trees such as in Fig. 1 revealed that around 100 out of 3200 are independent uncorrelated trajectories. Thus, for Trp-cage, TPS is around two orders of magnitude more efficient than a straightforward MD simulation (apart from the trivial parallelizability of TPS).

Based on the obtained path ensemble we identified two major channels in the folding mechanism of Trp-cage: the $N - P_d - L$ and the $N - I$ folding routes, of which one (the $N - L$) is four times more likely than the other (see Fig. 3(a)). The difference between the corresponding free-energy barriers is indeed not very high as the paths interchange easily and mixed pathways occur from time to time. The barrier itself has a diffusive nature and mainly

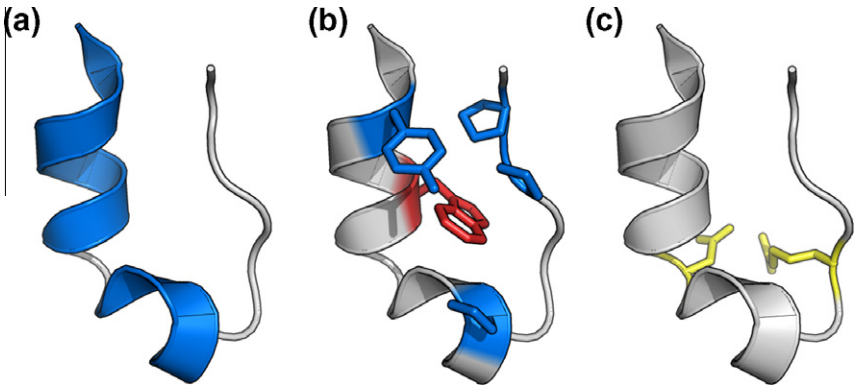

Fig. 2. Trp-cage mini protein (PDB ID TC5B): (a) protein backbone, with two helical regions colored in blue; (b) protein backbone in white with side-chains of amino-acids forming hydrophobic core shown as licorice. The central tryptophan residue is plotted in red, and the tryptophan pocket in blue; (c) salt-bridge between the residues 16 and 9 has been shown in yellow as licorice. Pictures rendered with Pymol. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

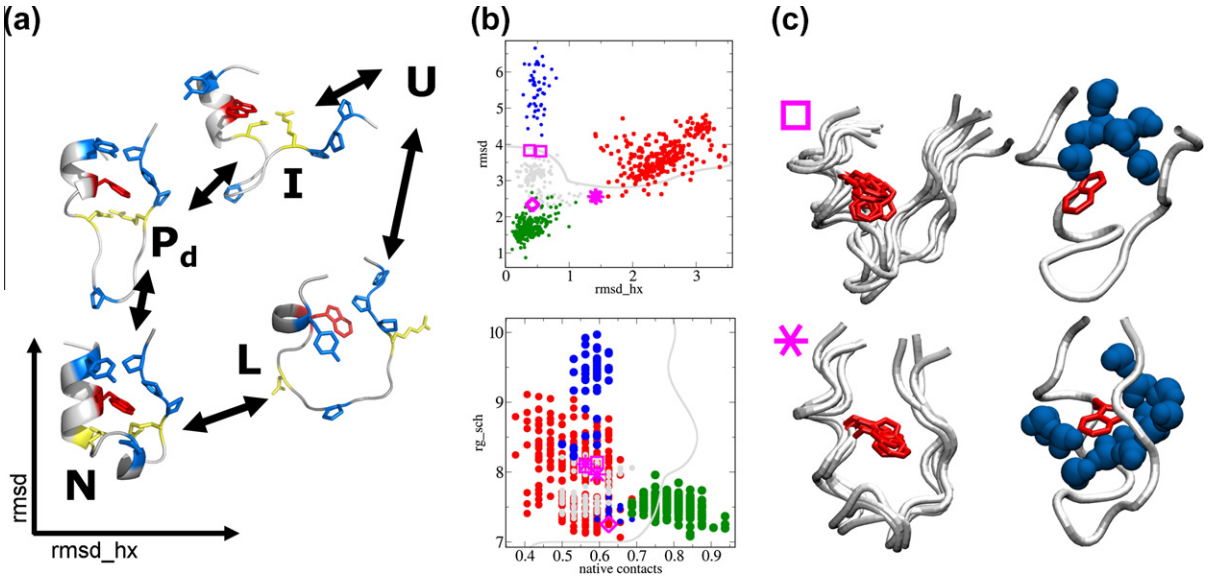

Fig. 3. (a) Schematic depiction of the folding mechanism of Trp-cage, elucidated with TPS simulations. (b) Stable states and transition states obtained by committor calculation, plotted as rg_{sch} versus ρ and *rmsd* versus $rmsd_{hx}$. Transition states for the *NL* and NP_d transitions are plotted in magenta as stars and squares respectively. The scatter points are taken from the corresponding TPS trajectories. Native states (*N*) are plotted in green, loop structures (*L*) in red, close to native structures with Pro-12 detached (P_d) in gray and the *I*-state in blue. The region on the bottom of the gray line on the $rmsd/rmsd_{hx}$ (to the right on the ρ/rg_{sch}) plot is the part of the configuration space not reachable by our REMD simulation starting from an unfolded configuration. (c) Left: superimposed TS structures for four different *NL* (star) and six for P_dI paths (square). Right: one of the TS structures for both routes in a side view plotted as a ribbon. Central tryptophan residue is plotted in licorice representation (red), water molecules within 5 Å of Trp-6 in space-filling representation blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

corresponds to desolvation of the internal tryptophan, which can occur in two ways.

From the unfolding perspective, for the $N-L$ pathway solvation happens between the tryptophan and Pro-12. The contact between Trp-6 and Pro-18 is maintained until the very last stages of unfolding and the protein core is solvated between Trp-6 and Pro-12 facilitating solvation of the helix and thus formation of the *L* (loop-like) state. Visual inspection of the path ensemble indicated that the system crosses the unfolding barrier after an hydrogen bond linked thread of water molecules penetrates the mini-protein between Trp-6 and Pro12, followed by solvation of the rest of the hydrophobic residues. The $N-P_d-I$ pathways significantly differ from this scenario. After the detachment of the Pro-12 from Trp-6 (the Pro-detached micro-state P_d) the previously mentioned thread of interconnected water molecules appears between the side-chain of Trp-6 and Ala-18, and thus the hydrophobic core is solvated from another side. The two mechanisms are clearly reflected in the transition state structures obtained from the committor analysis performed on paths from both subensembles (see Fig. 3(c)). In the $N-L$ case, the transition state shows a thread of waters below Tryptophan, while a thread of waters above tryptophan is present in the $N-I$ case.

Fig. 3b shows a few transition states structures for the $N-L$ and $N-I$ routes, obtained by committor analysis. These transition states are native like, but with a decreased number of native contacts. The $N-L$ TS structure has a largely dissolved helix and a entirely solvated tryptophan. In the TS structures of the P_d-I pathway the polyproline helix is perpendicular to the surface of the tryptophan aromatic ring, with a layer of water molecules in between. While the tryptophan is also fully solvated, there is no thread of water molecules penetrating the protein. Plotting the TSE in the $rmsd/rmsd_{hx}$ plane in Fig. 3(b) reveals that these order parameters are capable of describing the folding process, or at least distinguish the TSE from the stable states. In contrast, in the ρ/rg_{sch} representation, there is substantial overlap, and the location of the TSE is inside the native stable state, disqualifying these order parameters as proper reaction coordinates. The gray curve in the same figure is the projected area accessible to the REMD-unfolded simulation and suggests that the water expulsion transition is the rate limiting step for both folding routes.

The dynamics of the solvent is different in each of the routes. Most water-proteins contacts have a residence time of less than 50 ps, but water molecules around Trp-6 can stay bound much longer than 100 ps. In the $N-L$ route a quarter of the water molecules bound to Trp carbonyl oxygen remains trapped for longer than 100 ps, hydrogen bonded to both the α-helix and the 3_{10}-helix. It is this double hydrogen-bonding that increases the residence time.

For the $N-P_d-I$ route the water residence times are not as long, because water cannot bridge tryptophan and the glycine after the hydrophobic collapse, as the 3_{10}-helical part is separated from the α-helix. The explicit modeling of solvent molecules can thus reveal crucial aspects of folding kinetics.

The water dynamics can have a profound contribution to the reaction coordinate, as was observed in Ref. [91] and later in Ref. [92]. To estimate the extent of this contribution, we measured the effect of the water structure on the committor values of the transition states [93]. After freezing the protein coordinates, randomizing the solvent by performing MD at 400 K, and re-equilibrating for 1 ns at 300 K, a new committor value is estimated and tested for systematic deviation from 0.5. A strong deviation indicates that the solvent dynamics plays a large role in the reaction coordinate. Such an analysis for the transition states of Trp cage showed no significant change in committor. Therefore, the committor is independent of the dynamics of the water. The instantaneous water configuration still has a structural role by bridging several parts of the protein during folding. This finding seems to contradict the conclusion that the solvent dynamics is coupled to alanine dipeptide isomerization in a study by Ma and Dinner [53]. This contradiction might be explained by the fact that the dynamics of the Trp-cage backbone chain is orders of magnitude slower than that of both the dipeptide and the solvent, and its size larger than correlation lengths in water. The water molecules relax rather slowly next to the protein. When the randomized

solvent structure was only equilibrated for 100 ps, the committor was significantly changed toward the unfolded state. This difference in committor can be explained by the fact that waters can penetrate the hydrophobic core at high temperature, and need more than 100 ps to fully equilibrated at 300 K.

5.1.4. The folding and unfolding rate constants

Since the TIS method can only handle one barrier at the time, we choose the route that contributes most to the rate, i.e. the $L - N$ pathway (see Fig. 3). We performed two sets of TIS simulations, one for the $(N - L)$ unfolding (TIS-unf simulation) and another one for the folding $(L - N)$ transition (TIS-fol simulation) employing TIS interfaces in the helix RMSD order parameter $\lambda \equiv rmsd_{hx}$. While this order parameter is capable of distinguishing the N and L states, it does not distinguish between the N and I state. This is relevant for the TIS-fol simulation, for which the $L - I$ transition instead of $L - N$ may be observed. Also, the micro-state P_d which belongs to the $N - I$ route may attract some pathways for interfaces around the transition state, where the systems diffuses on the barrier and might switch the route. In those cases the TPS trial trajectory was rejected. As an initial pathway for the TIS sampling we used the TPS trajectories connecting both N and L.

Matching the results of TIS simulations for several interfaces, we obtained the crossing probability as a function of λ (see Fig. 4.) The final crossing probabilities are: $P(\lambda_L|\lambda_N) = 1.2 * 10^{-4}$ and $P(\lambda_N|\lambda_L) = 2.5 * 10^{-3}$. The flux factors ϕ_{10}^N and ϕ_{10}^L were calculated based from 10 ns long MD simulations in the native state (N) and in the loop state (L), respectively. The unfolding flux, corresponding to crossing of the $\lambda_1 = 0.06$ nm interface was $f_{10}^N = 6.7\ \mathrm{ns}^{-1}$. The folding flux through the $\lambda_1 = 0.23$ nm interface was $f_{10}^L = 1.0\ \mathrm{ns}^{-1}$. Together with the crossing probabilities these values give for the unfolding rate $k_{NL} = (1.2\ \mu\mathrm{s})^{-1}$ and for the folding rate $k_{LN} = (0.4\ \mu\mathrm{s})^{-1}$. The values lead to a free-energy difference between the folded N and the intermediate L state of $\Delta G_{NL} = k_B T \ln\left[\frac{k_{LN}}{k_{NL}}\right] \approx 1\ k_B T$.

The TIS rates are an order of magnitude higher than the experimentally measured rates [78]: $k_{unf}^{exp} = (12.7\ \mu\mathrm{s})^{-1}$ and $k_{fol}^{exp} = (4.1\ \mu\mathrm{s})^{-1}$. However, the experimental results are relative to the unfolded state, not the loop state. Assuming a simple steady state approximation for the L-state, we can deduce k_{UN} by multiplying k_{LN} with the exponent of the free-energy difference between the L and U state $\Delta G_{LU} \approx 1.5\ k_B T$ estimated by REMD [73,43], yielding $k_{UN} \approx k_{LN} * e^{-\Delta G_{LU}/k_B T} = (1.8\ \mu\mathrm{s})^{-1}$. This value differs only by a factor of 2 from the experimentally measured folding rate, which is probably within the statistical error. We note that because the other route $N - P_d - I - U$ is 4 times less likely, it will not influence the overall folding rate significantly. The discrepancy of the predicted unfolding rate with its the experimental value might be due to the OPLSAA force-field, which was also discussed in Ref. [88].

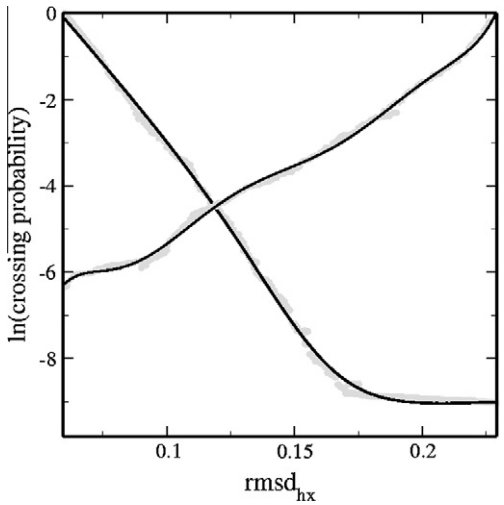

Fig. 4. Crossing probabilities $P(\lambda|\lambda_N)$ and $P(\lambda|\lambda_L)$ plotted as a function of the order parameter $\lambda = rmsd_{hx}$. Both curves reveal a plateau beyond (or below) a certain value of λ. The value of the plateau equals to the total crossing probability.

5.1.5. Reaction coordinate analysis

The ensemble of TPS shooting points naturally falls into two categories: one belonging to the $N - I$ and the other of $N - L$ pathways. Subjecting these two sub-ensembles separately to the likelihood maximization (LM) procedure [62] we could extract reaction coordinates. For the $N - I$ sub-ensemble the single most committor-correlated order parameter appeared to be $rmsd_{ca}$. No significant improvements were obtained for double combinations of trial order parameters. The resulting reaction coordinate is $rc_{NI} = -3.7 + 12 rmsd_{ca}$, where the RMSD is given in nm. For the $N - L$ sub-ensemble the helix RMSD $rmsd_{hx}$ yielded the maximum likelihood among the single order parameter. By adding another order parameter to our trial reaction coordinates, we were able to increase the maximum likelihood by a significant amount [62] for the combination of $rmsd_{hx}$ and $rmsd_{ca}$. Reaction coordinates of the third order did not result in significant improvement. The reaction coordinate for the $N - L$ route can thus be written as $rc_{NL} = -4.5 + 13 rmsd_{hx} + 8 rmsd_{ca}$.

From the shooting point ensemble we can extract the configurations that have $rc \approx 0$, corresponding of a predicted $p_B \approx 0.5$. Of course, the LM only *predicts* these structures to be transition states. To test this prediction we performed an additional full committor calculation for several of these structures, resulting in values between 0.2 and 0.5. Considering that the LM does not require an expensive commitment analysis, the method is reasonable, although not perfect. The fact that the committor value is not exactly 0.5 might be due to the limited number of shooting points. On the other hands, the LM approach could be improved by including more order parameters.

Our choice of TIS order parameter $\lambda = rmsd_{hx}$ for the $N - L$ transition could be the reason of some of the sampling problems in the folding TIS simulation. Although in principle TIS should not be very much dependent on the order parameter choice, including the $rmsd_{ca}$ in the order parameter would have been useful for the TIS folding rate calculation, as any $U - I$ transitions would have been rejected. Nevertheless, successful sampling was still possible using only $\lambda = rmsd_{hx}$.

5.1.6. Summary

To summarize, the use of TPS techniques has shed light on the folding mechanism of a small protein in explicit solvent. After a fast initial collapse, the Trp-cage can choose between two global routes. About 80% of the folding pathways first form the tertiary contact between Trp-6 and the polyproline part before the helix in the shape of a loop. The other 20% of the paths first form the helix before folding the tertiary structure. Committor analysis revealed the water dynamics is not a part of the reaction coordinate. A TIS simulation of the intermediate loop state L to the native state N transition revealed that while the folding $(L - N)$ rate, including a minor correction reasonably agrees with the experiment, the computed unfolding $(N - L)$ rate constant is one order of magnitude higher then the measured experimental value. This discrepancy is probably due to the quality of the OPLSAA force-field. The likelihood maximization method predicts that the best reaction coordinate for the $L - N$ transition is a combination of the $rmsd_{hx}$ and the $rmsd_{ca}$. The work on Trp-cage, neatly demonstrated what is possible within the TPS framework. Clearly, the reliability of the results is dependent on a correct identification

of the stable state, on sufficient sampling of the trajectories, but finally also on the correctness of the force field. In fact, the results kinetics might be used to improve these force fields, without having to rely on millisecond timescale simulations [18].

5.2. Trpzip4

5.2.1. A highly stable beta beta-hairpin

Among the few hairpins that are thermodynamically stable at ambient conditions, the 16-residue sequence at the *C*-terminal of the streptococcal protein G B1-domain (sequence GEWTYDDATKTFTVTE), is the most well studied. In the last decade it has become a model system to study hairpin formation experimentally, theoretically and by simulation [94–112]. Trpzip4 (sequence GEWTWDDATKTWTWTE), first designed by Cochran et al. [113], is among a number of mutants of the GB1 beta-hairpin that has been experimentally studied by several groups [113–115]. The sequence of Trpzip4 is based on the wild type GB1 hairpin, in which three out of four hydrophobic core residues were replaced by tryptophans (see Fig. 5). As a result of this strong mutation the entire hydrophobic core of the peptide consists of tryptophans. Trpzip4, while a fast folder (with a folding time of 15 s), is a very slow unfolder with an unfolding time of about 240 s [114]. The unfolding event will thus be difficult to simulate using straightforward MD, but can be studied with TPS [72]. Analysis of the pathways yields insight in how the strong mutation of the GB1 influences the folding landscape.

5.2.2. Stable state definitions and the initial path

Similar to the GB1 hairpin, REMD of Trpzip4 indicated the existence of the *U*, *H*, *F* and *N* states (see Fig. 5). It may seem that for systems as simple as beta hairpins, defining stable states is easy. When using the distance between the strands R_{OH} to describe both folded *N* and (partially) unfolded states *H* [46], the TPS simulations did not generate new paths that decorrelated from the initial path. While the acceptance ratio appeared sufficient, paths got trapped into an on-pathway metastable state *F*, which differs significantly from the native state *N* in the packing of the tryptophans, but not that much in the conformation of the backbone. Such problems with TPS sampling can be spotted by inspecting the sampling tree of TPS as schematically depicted in Fig. 6. For this reason, the path sampling was split into two parts: $U - F$ and $F - N$ path ensembles.

Inspection of the REMD and MD simulations suggested order parameters that can successfully distinguish between the relevant states: *N*, *F* and *H* (see Fig. 5). The radius of gyration, the RMSD, the solvent accessible surface areas and the number of native backbone hydrogen bonds were sufficient to distinguish the basin of attraction of the native state *N* from that of the intermediate *F*. The key to distinguish *F* from *N* was to define the *F* state by zero native hydrogen bonds. While most hydrogen bonds are still formed in the *F* state, they are much less stable than in the *N* state and subject to strong fluctuations, yielding regular configurations with zero hydrogen bonds within tens of picoseconds, without leaving the basin of attraction of the state *F*.

A high temperature unfolding trajectory originating from REMD simulations served as an initial pathway. For the $N - F$ transition, instead of gradually cooling it down by TPS, we performed a committor calculation at room temperature and determined points for which $p_N \in [0.4, 0.6]$, and glued together paths that lead to *N*, with paths that lead to *F*. The thus created initial trajectories were largely equilibrated at room temperature, except for the transition state region allowing for faster equilibration of the path ensemble.

5.2.3. Analysis of the TPS results

N–F transition: We performed in total 2500 trial shots with an aggregate simulation time of 4.8 μs. The average acceptance ratio was 27%, and the average accepted path length 2.9 ns. Analysis of the trees confirmed that the runs were sufficiently decorrelated [72]. Because the TPS trajectories are entirely time reversible we can interpret the results as unfolding or folding pathways. The $N - F$ transition is initiated at the termini by solvation of the most outer hydrogen bonds and seems to occur in several distinct steps. The major unfolding event is the termini moving apart. During unfolding Trp-14 is especially flexible and rotates to the other side of the hairpin in a joint move with Thr-13 splitting from Thr-4. Interestingly, the TPS ensemble separated into two minor channels, in which the radius of gyration *rg* differs. On the route via the lower *rg*, the side chain of Trp-14 detaches from the hydrophobic core and twists around the hairpin, settles down in a dry cavity formed by the bent hairpin backbone. This configuration persists for up to a couple of tens of picoseconds before the Trp-14 resolvates again. The fact that this microstate appears automatically in the TPS simulation means that it is an alternative on-pathway intermediate.

From the viewpoint of the folding process starting from the intermediate *F*-state, the system can choose between the two routes introduced above. The first route consists of twisting of the Trp-14 into a bent hairpin state with low *rg* followed by folding and straightening to *N*. The other route is more diffuse but goes straight from the *F* state to the native state at constant *rg*. During both routes the root mean square deviation *rmsd* and the strand distance R_{OH} decreases in correlated fashion. In the majority of paths the turn is formed before the termini come together. In the *F*-state, hydrogen bonds closest to the turn and the middle hydrogen bonds are already formed. The latter two hydrogen bonds need to be broken again during the transition to the *N* state because the peptide needs more flexibility for the rearrangement of the exterior tryptophan side chains (Trp-3,Trp-14). The presence of water bridges between these residues creates space for this

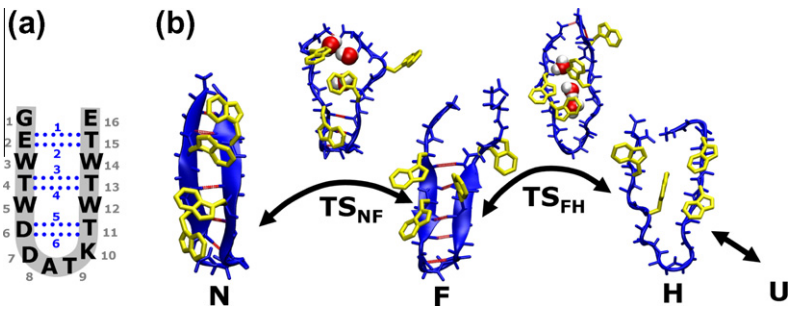

Fig. 5. (a) Schematic structure of the native state *N*, with the 6 hydrogen bonds indicated. (b) Folding and unfolding process of Trpzip4. The native state *N* and two intermediate state *F* and *H* are shown on the bottom row in blue backbone representation. Trp residues are shown in yellow stick model. The approximate transition states obtained with likelihood maximization analysis are shown on the top row. Bridging waters are shown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

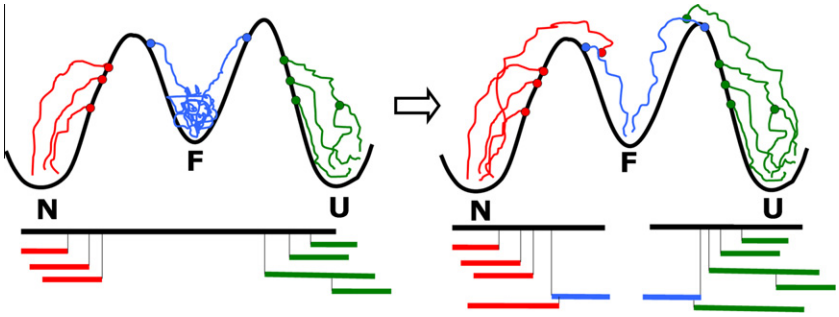

Fig. 6. Schematic depiction of the division of the TPS-ensemble because of intermediate state that hinders sampling, as in the case of Trpzip4. Shooting points have been schematically depicted as circles sitting on a free-energy landscape. In the case of one TPS simulation (left), all trajectories belonging to the basin of attraction of the intermediate get trapped and are not accepted. The resulting tree contains a large gap in the middle. After splitting the sampling in two parts (right), it is possible to generate new trajectories, since the paths ending in *F* are being accepted.

rearrangement which takes on average 800 ps (the actual barrier crossing time, not the rate constant). This process takes relatively long because not only the water bridges have to form but also the tryptophans must acquire their native position before the water can be expelled and the native state formed. Only when all tryptophans are repositioned correctly according to the zipper motif the middle hydrogen bonds fully form. This hydrogen bond formation order thus suggests a zipper-like folding mechanism [116]. However, the rate-limiting step is the $H-F$ transition, which is dominated by expulsion of water and formation of the middle hydrogen bonds.

F–H transition: The initial trajectory for the $H-F$ transition was prepared according to the same procedure as for the $F-N$ transition. We performed 520 trial shootings, of which 31% were accepted. The aggregate simulation time was 1.0μs and the average path length 1.9 ns. In the *F* state, the outer strand hydrogen bonds are fully solvated, and those in the turn partially. The unfolding starts by solvating the remaining hydrogen bonds while preserving the hydrophobic cluster of Trp-3, 5, and 12, preventing the hairpin strands to move apart. The number of broken hydrogen bonds increases from 1 to 3–4, while the strands are pulled further apart, changing the structure of the hydrophobic core and enlarging it slightly.

From the folding perspective, the strands are brought together during the formation of the Trp-3, 5 and 12 cluster. At the point when the Trp sidechains interact closely, the strands are on average 3.5 Å apart, and the hydrogen bonding starts. The middle hydrogen bonds appear as first, followed almost immediately by the formation of turn hydrogen bonds.

5.2.4. Reaction coordinate analysis

For the Trpzip4 $F-N$ transition, the RC that best correlates with the committor function is $rc_{NF} = -4.45 + 6.95rmsd + 0.8R_{OH}$, where both parameters are expressed in nanometers. The LM predicts that the protein structures in the shooting point ensemble characterized by $rc_{NF} \in [-0.1, 0.1]$ and an estimated $pB \in [0.4, 0.6]$ should be a reasonable approximation of the transition state ensemble. One of these structures is shown in Fig. 5. The turn region is intact during the transition, with backbone hydrogen bonds fluctuating but not being solvated. In contrast, the middle hydrogen bonds are well solvated and one or two water bridges exist between Thr-4 and Thr-13. The side chains of these two threonines are solvent exposed and not in contact with each other as in the native state. The terminal parts, including the Trp-3 and Trp-14 side chains, are also solvated. The order parameters that represent this dynamics did not show up in the reaction coordinate.

We also performed the LM analysis on the $H-F$ TPS ensemble of Trpzip4. The best candidate for the reaction coordinate was $rc_{FH} = -2.11 + 0.40 \times R_{OH}$. A typical Trpzip4 structure having the $rc_{FH} \in [-0.1, 0.1]$ and an estimated p_B value $\in [0.4, 0.6]$ is presented in Fig. 5. All structures fulfilling these conditions show Tryptophans 3, 5 and 12 forming a hydrophobic core while Trp-14 solvent exposed or twisted around the hairpin. The turn is natively shaped, not allowing for much solvation in this area. Water bridges appear between Glu-2 and Thr-15, Thr-4 and Thr-13, and also between Asp-6 and Thr-11. One or sometimes two bridges are observed for the same frame. We note that the exposed Trp-14 in the transition state might explain the slightly lower folding rate of Trpzip4 (15 μs^{-1}) compared to GB1 (6 μs^{-1}) [114].

5.2.5. Summary

The major conclusion is that the folding mechanism of Trpzip is more heterogeneous than for GB1 [46]. The metastable state *F* becomes more (meta) stable, due to a large increase of the $F-N$ barrier that "roughen" the free-energy landscape. Structure-wise, the largest difference is in the final folding step ($F-N$), which due to a twisted solvent exposed Trp residue in the F state is very different from its counterpart in GB1 and involves a rotation of the Trp around the backbone and a temporary breaking of the middle hydrogen bonds. Thus, TPS shows in atomistic detail how a three residue mutation leads to an altered folding landscape. The $F-N$ step involves partial unfolding, re-zipping of hydrogen bonds and rearrangement of the Trp-14 sidechain. For the rate limiting ($H-F$) step Trpzip4 desolvation is decoupled from strand closure. Nevertheless, likelihood maximization shows that the reaction coordinate for both hairpins remains the interstrand distance. We conclude that the folding mechanism of Trpzip4 is a combination of hydrophobic collapse and zipping of hydrogen bonds, in which one tryptophan is exposed to the solvent due to steric hindrance, making the folding mechanism more complex, and leading to an increased $F-N$ barrier. Thus, the results show in atomistic detail how a rather strong mutation leads to a different folding mechanism, and yields a more frustrated folding free-energy landscape. This work also clearly exposes the difficulties that might arise in the application of TPS when additional long-lived (meta) stable states show up unexpectedly in the mechanism.

5.3. Photoactive Yellow Protein

Photoactive Yellow Protein (PYP) is a water soluble blue-light photo receptor from *Halorhodospira halophila* [117,118] and consists of 125 amino acids and a covalently bound chromophore (para-coumaric acid, pCA) [119], which fold into a PAS core capped by an *N*-terminal domain [120,121]. Upon the absorption of a blue-light photon, PYP undergoes a number of conformational rearrangements, culminating in the formation of the signaling

state. The first steps in signaling state formation are the *trans* to *cis* isomerization of pCA, on a picosecond timescale [122], followed by a proton transfer, taking place within microseconds [123,124]. These reactions result in an intermediate state pB' with the chromophore in a strained configuration and a negative charge on a buried glutamate (Glu46) [125,126]. As a consequence, the protein partially unfolds to expose pCA and Glu46 to bulk water on a submillisecond time scale [127–129], completing the formation of the signaling state *pB*.

PYP is an excellent model signaling protein as it is easily accessible for the application of a wide variety of experimental techniques, and has been characterized in many different ways. The initial events of the photocycle have become much clearer with the use of ultra-fast spectroscopy techniques [130], time-resolved X-ray crystallography [131] and a combined approach of quantum mechanics with force field based simulations. A QM/MM study showed that proton transfer can only occur when the negative charge on Glu46 is sufficiently stabilized [132]. Still, revealing the structural and dynamical details of the poorly understood mechanism of the unfolding process at high resolution, e.g. by time-resolved NMR, remains elusive [128]. Replica Exchange MD (REMD) simulations indicated that the formation of *pB* involves losing the α-helical structure of region 43–51, and the solvent exposure of Glu46 and pCA [129]. The predicted signaling state remarkably resembled the NMR structure of the *pB* state of the N-terminally truncated mutant Δ_{25}-PYP [128,133].

Free energy profiles from previous REMD simulations suggested the presence of metastable states occurring during the formation of *pB* [133]. These intermediate states differ with respect to the conformation of region 43–52 and the location of Glu46 and the chromophore. Fig. 7 displays an artist impression of the reaction network study by TPS with representative snapshots of the minima. The label pB' indicates the fully folded state, I_α indicates a state in which the helical conformation of helix α3 has partially disappeared. U_α denotes a state in which helix α3 is completely unfolded, while Glu46 and pCA are still inside the protein. The labels S_E and S_X refer to states with a solvent exposed Glu46 and pCA, respectively. Finally, *pB* indicates the signaling state, with both pCA and Glu46 exposed to solvent. The REMD simulations provided the stable state definitions to use in the TPS simulations. As the various metastable states prevented sampling the pB' to *pB* transition directly, instead the unfolding transition was split up into four separate processes. These transitions are: (1) the loss of α-helical structure in helix α3 ($pB' - I_\alpha$), (2) solvent exposure of Glu46 ($U_\alpha - S_E$), (3) solvent exposure of pCA ($U_\alpha - S_X$), and (4) solvent exposure of pCA following Glu46 ($S_E - pB$).

5.3.1. Unfolding of helix α3

An initial TPS simulation of the $pB' - U_\alpha$ transition yielded very long pathways due to the presence of intermediate state I_α and therefore, we performed a TPS simulation of the $pB' - I_\alpha$ transition. Once the protein is in intermediate state I_α it relaxes to the U_α state in which the helix is completely unfolded within nanoseconds, indicating that the barrier separating I_α and U_α is lower than the barrier between pB' and I_α.

For PYP, about 75 order parameters could be involved in the signaling state formation, which were computed for the entire shooting point ensemble of the $pB' - I_\alpha$ TPS simulation. Using the LM analysis method yielded linear combinations of these order parameters that best describe the RC. When considering only single order parameters for the $pB' - I_\alpha$ transition $rmsd_\alpha$, the root mean square deviation with respect to the folded structure of region 43–51, is the best model for the reaction coordinate. Including more than one order parameter showed that adding the number of water molecules around Tyr42, nw_Y, resulted in a significant improvement. Using three order parameters showed again a significant improvement, when nw_Y is replaced by *dhb*2 and *dPA*, respectively the length of the backbone hydrogen bond between Ala44 and Asp48, and the distance between Pro54 and Ala44. Application of the Bayesian path statistical analysis [60] confirm these results.

The description for the optimal RC obtained with the LM analysis matches well with visual inspection of the $pB' - I_\alpha$ pathways. In the folded state Tyr42 is completely shielded from bulk water by Ala44 and Pro54. During the unfolding of helix α3, water molecules form hydrogen bonds to Tyr42, facilitated by a larger distance between Ala44 and Pro54 (*dPA*). This hydration is transient, as this

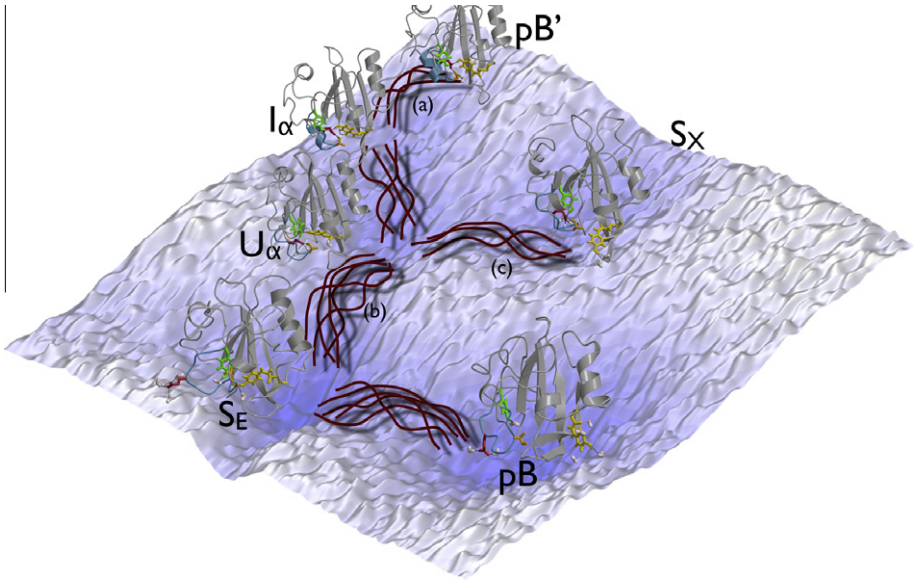

Fig. 7. An artistic rendering of the reaction network of the transition from pB' (top) to *pB* (bottom). In the background is the rough energy landscape with several (free energy) minima, indicated by the label. For each minimum a representative snapshot of the protein conformation is shown in grey cartoon style. The helix α3 is rendered blue, and the stick models represent relevant residues: pCA (yellow), Glu46 (red), Tyr42 (Green) and Thr50 (orange). Black dotted lines indicate hydrogen bonds. Water molecules are shown as pink and white stick models. The dark red curves between the minima depict the concept of transition path ensemble, sampled by TPS. The label (a), (b) and (c) refer to the transition states in Fig. 8. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

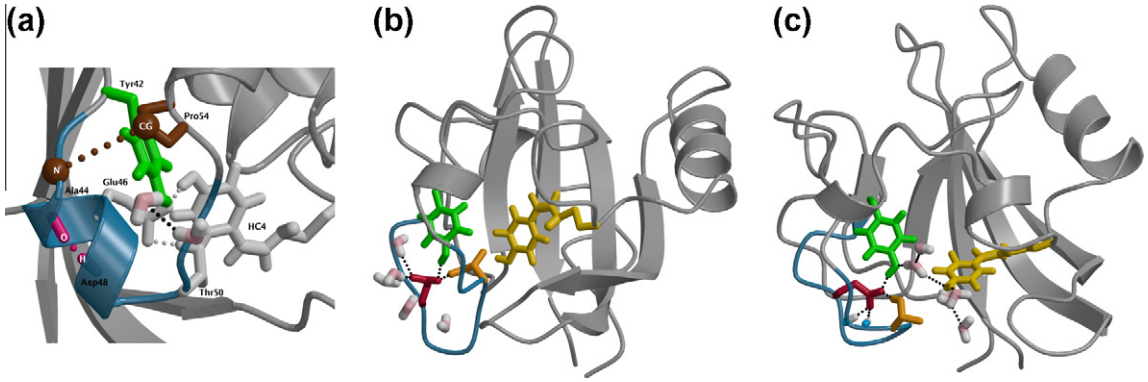

Fig. 8. (a) TS estimate of the $pB' - I_\alpha$ transition. Helix $\alpha 3$ is shown in blue, hydrogen bond 2 is depicted as a pink stick model. Atoms of Pro-54 and Ala-44 used for the calculation of *dPA* are shown as brown spheres. Hydrogen bond *dhb*2 is shown in magenta. Other residues making up the chromophore binding pocket are shown as light grey sticks. (b–c) TS estimates of solvent exposure of (b) Glu46 and (c) pCA. Representation is as described in Fig. 7. In (c) two backbone N–H groups are also shown as blue–white stick models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

distance decreases again when helix $\alpha 3$ is completely unfolded. The distance between Ala44 and Pro54 is not the only relevant order parameter, as the LM analysis also identified the hydrogen bond between Ala44 and Asp48 (*dhb*2) as an important contribution to the RC. Interestingly, including the helical hydrogen bonds between Asn43 and Gly47 (*dhb*1), and between Ala45 and Ile49 (*dhb*3) also showed a significant increase, but not as much as the Ala44–Asp48 hydrogen bond. These three hydrogen bonds are located at the solvent accessible side of helix $\alpha 3$, whereas the remaining two helical hydrogen bonds, located at the protein side of helix $\alpha 3$, do not contribute at all to the RC.

The LM analysis method predicts that structures with $r = 0$ are transition states. Extracting the configurations within the interval $r \in [-0.05, 0.05]$ from the $pB' - I_\alpha$ path ensemble shows an ensemble which is quite uniform. Explicit calculation of the committors of these configurations confirmed that 30% of the predicted structures are true transition states (with committor values between 0.3 and 0.7). The other conformations had committors close to either zero or unity indicating that the transition state ensemble for the $pB' - I_\alpha$ transition is thus very sensitive to small changes in the protein conformation, indicating it involves a relatively narrow barrier. Fig. 8(a) displays a typical configuration from this predicted TS ensemble with relevant order parameters that contributed most to the RC highlighted. Helix $\alpha 3$ is still in a helical conformation, indicating that the TS is very close to the folded state.

5.3.2. Solvent exposure of Glu46 and pCA

Once the helix is unfolded and the protein is in the U_α state, the next step in the formation of *pB* is the solvent exposure of both Glu46 and pCA. This step required two different TPS simulations at the least, sampling the solvent exposure of Glu46 ($U_\alpha - S_E$) or pCA ($U_\alpha - S_X$). Both processes proved difficult to sample, as exposing pCA or Glu46 involve long and diffusive paths.

LM analysis of the $U_\alpha - S_E$ shooting point ensemble reveals that *dXE*, the distance between Glu46 and pCA is the best RC for the exposure of Glu46 to the solvent. As pCA is an important hydrogen bond donor to Glu46, this distance parameter indicates that disruption of the hydrogen bond network around Glu46 inside the protein leads to its solvent exposure. Adding more order parameters does not significantly improve the RC. Fig. 8(b) shows one of the transition states configurations. This prediction agrees with visual inspection of the pathways, suggesting that the shift in the hydrogen bond network of the acidic side-chain of Glu46 is the main characteristic of the transition. In this transition Glu46 moves to the solvent, instead of water molecules entering the protein. On the U_α side of the barrier Glu46 interacts with Tyr42, Thr50 and the protein backbone. On the S_E side water molecules provide stabilization for Glu46. The TS conformation depicted in Fig. 8(b) shows a balance between these two hydrogen bonds networks, in which pCA does not participate.

LM analysis of the $U_\alpha - S_X$ simulation shows that dXY^{com}, the distance between Tyr42 and pCA, is the most significant single order parameter, indicating that the strength of the hydrogen bond between pCA and Tyr42 governs the solvent exposure of pCA. Adding *dXE* significantly improves the RC. Fig. 8(c) displays one of the predicted conformations of the TS ensemble for the pCA solvent exposure, with a partially solvated pCA and a water-mediated hydrogen bond between Tyr42 and pCA. Glu46 is still located inside the protein and does not interact with water molecules at all. Instead Glu46 forms hydrogen bonds to N–H groups in the protein backbone to stabilize its negative charge. Visual inspection of the $U_\alpha - S_X$ TPS trajectories show that Tyr42 and Glu46 interact, while pCA detaches from this hydrogen bond network and becomes (partially) solvated.

5.3.3. Final unfolding step toward pB

The results so far suggest two possible pathways, both leading from the U_α state to the unfolded signaling state *pB*, each with its own intermediate state, S_E and S_X. In the final state *pB* both Glu46 and pCA are solvent exposed. To obtain initial pathways that connect the *pB* state to the preceding intermediates, several conventional MD simulations were run at ambient and elevated temperatures. No additional solvent exposure events occurred at 300 K, indicating that both S_E and S_X are sufficiently metastable to prevent a fast escape from these states, and consequently, another barrier must be crossed to reach *pB*. Only at 500 K, Glu46 became solvent exposed after pCA did. Instead of reaching pB, this simulation actually sampled pCA going back into the protein core, thus returning the system to the U_α state. As the $S_X - pB$ transition does not occur even at high temperature, S_X is called a 'dead end'. In contrast, starting from the S_E state, solvent exposure of pCA occurs already at 425 K, indicating that it is easier for pCA to leave the protein once Glu46 is already solvent exposed, than vice versa. The results of a TPS simulation on the $S_E - pB$ transition showed that accepted pathways indeed reach *pB* and the end points overlap well with the experimental data [128]. Examination of the pathways reveals that in the barrier region pCA is still inside the protein and hydrated.

5.3.4. The most likely pathway is correlated with a salt bridge

When Glu46 moves into the solvent first, formation of the signaling state will follow, whereas if pCA goes out first, the protein

enters a 'dead-end' route. Comparison of the two TS estimates for $U_\alpha - S_E$ and $U_\alpha - S_X$ reveals that the region 43–51 has a different shape depending on whether Glu46 is exposed. Further differences constitute a salt bridge between Asp20 and Lys55, which also exists in the pB' and receptor states. This ionic bond is absent in the $U_\alpha - S_E$ TS, but is present in the TS of $U_\alpha - S_X$. Extending several of the TPS trajectories in the backward time direction shows that when coming from the $U_\alpha - S_X$ route, the salt bridge exists, whereas backwardly extended paths from the $U_\alpha - S_E$ trajectories have no salt bridge. This indicated that U_α splits into two substates with different states for the Asp20 and Lys55 interaction. These observations suggest that the absence of the DK salt bridge leads to Glu46 exposure, while its presence results in solvent exposure of pCA. As Asp20 is located in the *N*-terminal domain of PYP, the conformation of this region is correlated with the presence of the salt bridge, and thus with whether the protein will choose a dead-end or a productive pathway.

This work shows that TPS enables the sampling of millisecond timescale processes at room temperature and at atomistic detail. Additional LM analysis can predict relevant reaction coordinates for each conformational rearrangement in PYP. For the loss of α-helical content, we find that the best RC involves the overall deviation from the folded helix, but also a helical hydrogen bond and the distance between Pro54 and Ala44. RCs of the exposure of pCA and Glu46 involve the breaking of hydrogen bond interactions in the chromophore binding pocket, between pCA, Glu46 and Tyr42.

6. Outlook

As in other biological processes, it is often difficult to achieve detailed insight on mechanisms of protein conformational changes at the molecular level by experiments. The field of molecular simulation has advanced to a level in which complex biological systems can be modeled in atomistic detail, but is still suffering from the large differences in timescales between typical bond vibrations (fs) and biological processes (μs–s). The transition path sampling method is able to overcome such differences. In the last years we and others have studied biological systems with TPS [76]. In this paper we have given an brief overview of the application of transition path sampling to conformational dynamics in (small) proteins, and used three case studies to illustrate the strength of the methodology, but also expose the pitfalls.

While TPS simulations are powerful, there is much room for improvement. In particular, the existence of intermediate metastable states and the possibility of multiple reaction channels can be problematic. The first problem of having intermediate metastable states can be solved by dividing the reaction network into several sub-reactions, as we have done for the conformational transitions of Trpzip4 and PYP. However, in a more complex case one might want to resort to a Markovian state model (MSM) description [134]. A combination of the MSM approach and TPS would be very fruitful to study complex reaction networks involving high barriers. The second problem of the multiple reaction channels has in principle been solved using replica exchange TIS [135,136]. Combining TPS with the recently developed reweighted path ensemble (RPE) method [137] leads to a full description of both kinetics and equilibrium statistics. Moreover, the reweighted path ensemble allows for a non-linear reaction coordinate analysis, which would not be straightforward for conventional TPS [63].

Up to now, only proteins of modest size have been studied by TPS, with PYP of 125 residues in size being the largest of these. Larger protein systems usually also exhibit longer molecular time scales. While most of the transition paths presented in this paper were in the order of nanoseconds, pathways involving large conformational changes with a rate constant of seconds, possibly require hundreds of nanoseconds per path. Up to recently, it was not possible to consider such long pathways in TPS. However, the tremendous speed up in MD simulation achieved by Shaw et al. [18] opens up a whole new range of possibilities. By using this scale of computing power in combination with TPS, accessing large protein conformational changes on a scale of seconds might become accessible. We are therefore confident that in the next few years we will see many more TPS studies on large protein conformational changes.

References

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Molecular Biology of the Cell, Garland Publisher, 2002.
[2] C.M. Dobson, Semin. Cell Dev. Biol. 15 (1) (2004) 3.
[3] A.R. Fersht, Structure and Mechanism in Protein Science: a Guide to Enzyme Catalysis and Protein Folding, Freeman, New York, 1999.
[4] M. Levitt, J. Mol. Biol. 168 (1983) 595.
[5] M. Levitt, M. Hirschberg, R. Sharon, V. Daggett, Comput. Phys. Commun. 91 (1995) 215.
[6] W. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, J. Am. Chem. Soc. 117 (1995) 5179.
[7] A.D. MacKerell Jr., D. Bashford, M. Bellott, R.L. Dunbrack Jr., J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, J. Phys. Chem. B 102 (1998) 3586.
[8] B.R. Brooks, C.L. Brooks, A.D. Mackerell, L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R.M. Venable, H.L. Woodcock, X. Wu, W. Yang, D.M. York, M. Karplus, J. Comput. Chem. 30 (10) (2009) 1545.
[9] W. van Gunsteren, H. Berendsen, Gromos-87 Manual, Biomos BV, Groningen, The Netherlands, 1987.
[10] W.F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D.P. Geerke, A. Glaettli, P.H. Huenenberger, M.A. Kastenholz, C. Ostenbrink, M. Schenk, D. Trzesniak, N.F.A. van der Vegt, H.B. Yu, Angew. Chem. Int. Ed. 45 (25) (2006) 4064.
[11] W.L. Jorgensen, D.S. Maxwell, J. Tirado-Rives, J. Am. Chem. Soc. 118 (1996) 11225.
[12] A. Mackerell, J. Comput. Chem. 25 (13) (2004) 1584.
[13] P.L. Freddolino, S. Park, B. Roux, K. Schulten, Biophys. J. 96 (9) (2009) 3772.
[14] M.P. Allen, D.J. Tildesley, Computer Simulation of Liquids, Oxford University Press, Oxford, 1987.
[15] D. Frenkel, B. Smit, Understanding Molecular Simulation, second ed., Academic Press, San Diego, CA, 2002.
[16] B.J. Grant, A.A. Gorfe, J.A. McCammon, Curr. Opin. Struct. Biol. 20 (2) (2010) 142.
[17] V.S. Pande, I. Baker, J. Chapman, S.P. Elmer, S. Khaliq, S.M. Larson, Y.M. Rhee, M.R. Shirts, C.D. Snow, E.J. Sorin, B. Zagrovic, Biopolymers 68 (1) (2003) 91.
[18] D.E. Shaw, P. Maragakis, K. Lindorff-Larsen, S. Piana, R.O. Dror, M.P. Eastwood, J.A. Bank, J.M. Jumper, J.K. Salmon, Y.B. Shan, W. Wriggers, Science 330 (2010) 341.
[19] R. Swendsen, J. Wang, Phys. Rev. Lett. 57 (1986) 2607.
[20] D. Earl, M. Deem, Phys. Chem. Chem. Phys. 7 (23) (2005) 3910.
[21] Y. Sugita, Y. Okamoto, Chem. Phys. Lett. 314 (1–2) (1999) 141.
[22] M.R. Sørensen, A.F. Voter, J. Chem. Phys. 112 (2000) 9599.
[23] H. Grubmüller, Phys. Rev. E 52 (1995) 2893.
[24] A.F. Voter, Phys. Rev. Lett. 78 (1997) 3908.
[25] T. Huber, A. Torda, W. van Gunsteren, J. Comput. Aided Mol. Des. 8 (1994) 695.
[26] A. Mitsutake, Y. Sugita, Y. Okamoto, Biopolymers 60 (2) (2001) 96.
[27] X. Periole, A.E. Mark, J. Chem. Phys. 126 (2007) 014903.
[28] G.M. Torrie, J.P. Valleau, Monte Carlo free energy estimates using non-Boltzmann sampling, Chem. Phys. Lett. 28 (1974) 578.
[29] A. Laio, M. Parrinello, Proc. Natl. Acad. Sci. USA 99 (2002) 12562.
[30] A. Faradjian, R. Elber, J. Chem. Phys. 120 (2004) 10880.
[31] W. E, W. Ren, E. Vanden-Eijnden, Phys. Rev. B 66 (2002) 052301.
[32] D. Branduardi, F.L. Gervasio, M. Parrinello, J. Chem. Phys. 126 (5) (2007) 054103.
[33] C. Dellago, P.G. Bolhuis, F.S. Csajka, D. Chandler, J. Chem. Phys. 108 (5) (1998) 1964.
[34] P.G. Bolhuis, D. Chandler, C. Dellago, P.L. Geissler, Annu. Rev. Phys. Chem. 53 (2002) 291.
[35] C. Dellago, P.G. Bolhuis, P.L. Geissler, Adv. Chem. Phys. 123 (2002) 1.
[36] C. Dellago, P.G. Bolhuis, Adv. Polym. Sci. 221 (2009) 167.
[37] P.G. Bolhuis, D. Dellago, Trajectory-based rare event simulations, Reviews in Computational Chemistry, vol. 27, Wiley, New York, 2010.
[38] T.S. van Erp, D. Moroni, P.G. Bolhuis, J. Chem. Phys. 118 (17) (2003) 7762.
[39] P.G. Bolhuis, Biophys. J. 88 (1) (2005) 50.
[40] P.G. Bolhuis, J. Phys. Condens. Mat. 15 (1) (2003) S113.
[41] M. Grünwald, C. Dellago, P.L. Geissler, J. Chem. Phys. 129 (19) (2008) 194101.
[42] H.C. Andersen, J. Chem. Phys. 72 (1980) 2384.
[43] J. Juraszek, P.G. Bolhuis, Biophys. J. 95 (2008) 4246.

[44] C. Lowe, Europhys. Lett. 47 (2) (1999) 145.
[45] G. Bussi, D. Donadio, M. Parrinello, J. Chem. Phys. 126 (1) (2007) 014101.
[46] P.G. Bolhuis, Proc. Natl. Acad. Sci. USA 100 (2003) 12129.
[47] J. Hu, A. Ma, A.R. Dinner, J. Chem. Phys. 125 (11) (2006) 114101.
[48] J. Rogal, P.G. Bolhuis, J. Chem. Phys. 129 (22) (2008) 224107.
[49] R. Du, V.S. Pande, A.Y. Grosberg, T. Tanaka, E.S. Shakhnovich, J. Chem. Phys. 108 (1) (1998) 334.
[50] P.G. Bolhuis, C. Dellago, D. Chandler, Proc. Natl. Acad. Sci. USA 97 (11) (2000) 5877.
[51] W. E, E. Vanden-Eijnden, J. Stat. Phys. 123 (2006) 503.
[52] P.L. Geissler, C. Dellago, D. Chandler, J. Phys. Chem. B 103 (1999) 3706.
[53] A. Ma, A.R. Dinner, J. Phys. Chem. B 109 (14) (2005) 6769.
[54] J.E. Basner, S.D. Schwartz, J. Am. Chem. Soc. 127 (2005) 13822.
[55] S. Saen-oon, S. Quaytman-Machleder, V.L. Schramm, S.D. Schwartz, Proc. Natl. Acad. Sci. USA 105 (43) (2008) 16543.
[56] S. So, M. Karplus, J. Med. Chem. 39 (26) (1996) 5246.
[57] A. Dinner, S. So, M. Karplus, Adv. Chem. Phys. 120 (2002) 1.
[58] B. Qi, S. Muff, A. Caflisch, A.R. Dinner, J. Phys. Chem. B 114 (2010) 6979.
[59] G. Hummer, J. Chem. Phys. 120 (2) (2004) 516.
[60] R. Best, G. Hummer, Proc. Natl. Acad. Sci. USA 102 (2005) 6732.
[61] B. Peters, B.L. Trout, J. Chem. Phys. 125 (2006) 054108.
[62] B. Peters, G.T. Beckham, B.L. Trout, J. Chem. Phys. 127 (2007) 034109.
[63] W. Lechner, J. Rogal, J. Juraszek, B. Ensing, P.G. Bolhuis, J. Chem. Phys. 133 (17) (2010) 174110.
[64] D. Chandler, J. Chem. Phys. 68 (1978) 2959.
[65] C.H. Bennett, in: R. Christofferson (Ed.), Algorithms for Chemical Computations, ACS Symposium Series, vol. 46, American Chemical Society, Washington, DC, 1977.
[66] T.S. van Erp, P.G. Bolhuis, J. Comput. Phys. 205 (1) (2005) 157.
[67] A. Ferrenberg, R. Swendsen, Phys. Rev. Lett. 63 (1989) 1195.
[68] S. Kumar, D. Bouzida, R. Swendsen, P. kollman, J. Rosenberg, J. Comput. Chem. 13 (8) (1992) 1011.
[69] R. Allen, P.B. Warren, P. ten Wolde, Phys. Rev. Lett. 94 (2005) 018104.
[70] R.J. Allen, D. Frenkel, P.R. ten Wolde, J. Chem. Phys. 124 (2006) 024102.
[71] E.E. Borrero, F.A. Escobedo, J. Chem. Phys. 127 (16) (2007) 164101.
[72] J. Juraszek, P.G. Bolhuis, J. Phys. Chem. B 113 (50) (2009) 16184.
[73] J. Juraszek, P.G. Bolhuis, Proc. Natl. Acad. Sci. USA 103 (2006) 15859.
[74] J. Juraszek, P.G. Bolhuis, Biophys. J. 98 (4) (2010) 646.
[75] J. Vreede, J. Juraszek, P.G. Bolhuis, Proc. Natl. Acad. Sci. USA 107 (6) (2010) 2397.
[76] C. Dellago, P.G. Bolhuis, Top Curr. Chem. 268 (2007) 291.
[77] J. Neidigh, R. Fesinmeyer, H. Andersen, Nat. Struct. Biol. 9 (2002) 425.
[78] L. Qiu, S. Pabit, A. Roitberg, S. Hagen, J. Am. Chem. Soc. 124 (2002) 12952.
[79] C. Simmerling, B. Strockbine, A. Roitberg, J. Am. Chem. Soc. 124 (2002) 11258.
[80] C.D. Snow, B. Zagrovic, V. Pande, J. Am. Chem. Soc. 124 (2002) 14548.
[81] M. Ota, M. Ikeguchi, A. Kidera, Proc. Natl. Acad. Sci. USA 101 (2004) 17658.
[82] J. Pitera, W. Swope, Proc. Natl. Acad. Sci. USA 100 (2003) 7587.
[83] S. Chowdhury, M.C. Lee, Y. Duan, J. Phys. Chem. B 108 (2004) 13855.
[84] R. Zhou, Proc. Natl. Acad. Sci. USA 100 (2003) 13280.
[85] A. Linhananta, J. Boer, I. MacKay, J. Chem. Phys. 122 (2005) 114901.
[86] F. Ding, S. Buldyrev, V. Dokholyan, Biophys. J. 88 (2005) 147.
[87] Y.M. Rhee, E.J. Sorin, G. Jayachandran, E. Lindahl, V. Pande, Proc. Natl. Acad. Sci. USA 101 (2004) 6456.
[88] S. Piana, A. Laio, J. Phys. Chem. B 111 (2007) 4553.
[89] D. Paschek, H. Nymeyer, A. Garcia, J. Struct. Biol. 157 (2007) 524.
[90] C. Velez-Vega, E.E. Borrero, F.A. Escobedo, J. Chem. Phys. 133 (10) (2010) 105103.
[91] P.R. ten Wolde, D. Chandler, Proc. Natl. Acad. Sci. USA 99 (10) (2002) 6539.
[92] T.F. Miller, E. Vanden-Eijnden, D. Chandler, Proc. Natl. Acad. Sci. USA 104 (37) (2007) 14559.
[93] Y. Rhee, V. Pande, Chem. Phys. 323 (2006) 66.
[94] F.J. Blanco, G. Rivas, L. Serrano, Nat. Struct. Biol. 1 (1994) 584.
[95] F.J. Blanco, L. Serrano, Eur. J. Biochem. 230 (1995) 634.
[96] V. Munoz, P.A. Thomson, J. Hofrichter, W.A. Eaton, Nature 390 (1997) 196.
[97] V. Munoz, E.R. Henry, J. Hofrichter, W.A. Eaton, Proc. Natl. Acad. Sci. USA 95 (1998) 5872.
[98] S. Honda, N. Kobayashi, E. Munekata, J. Mol. Biol. 295 (2000) 269.
[99] A. Kolinski, B. Ilkowski, J. Skolnick, Biophys. J. 77 (1999) 2942.
[100] D. Klimov, D. Thirumalai, Proc. Natl. Acad. Sci. USA 97 (2000) 2544.
[101] A.R. Dinner, T. Lazaridis, M. Karplus, Proc. Natl. Acad. Sci. USA 96 (1999) 9068.
[102] B. Zagrovic, E.J. Sorin, V. Pande, J. Mol. Biol. 313 (2001) 151.
[103] V. Pande, D. Rokhsar, Proc. Natl. Acad. Sci. USA 96 (1999) 9062.
[104] D. Roccatano, A. Amadel, A.D. Nola, H.J.C. Berendsen, Protein Sci. 8 (1999) 2130.
[105] B. Ma, R. Nussinov, J. Mol. Biol. 296 (2000) 1091.
[106] P. Eastman, N. Gronbech-Jensen, S. Doniach, J. Chem. Phys. 114 (2001) 3823.
[107] A.E. Garcia, K. Sanbonmatsu, Prot. Struct. Func. Gen. 42 (2001) 345.
[108] R. Zhou, B.J. Berne, R. Germain, Proc. Natl. Acad. Sci. USA 98 (2001) 14931.
[109] J. Tsai, M. Levitt, Biophys. Chem. 101 (2002) 187.
[110] I. Daidone, M. D'Abramo, A. Di Nola, A. Amadei, J. Am. Chem. Soc. 127 (2005) 14825.
[111] S. Krivov, M. Karplus, Proc. Natl. Acad. Sci. USA 101 (2004) 14766.
[112] D. Evans, D. Wales, J. Chem. Phys. 121 (2004) 1080.
[113] A.G. Cochran, N.J. Skelton, M.A. Starovasnik, Proc. Natl. Acad. Sci. USA 98 (2001) 5578.
[114] D. Du, M.J. Tucker, F. Gai, Biochemistry 45 (2006) 2668.
[115] B. Ma, R. Nussinov, Protein Sci. 12 (2003) 1882.
[116] V. Munoz, R. Ghirlando, F.J. Blanco, G.S. Jas, J. Hofrichter, W.A. Eaton, Biochemistry 45 (2006) 7023.
[117] T. Meyer, G. Tollin, J. Hazzard, M. Cusanovich, Biophys. J. 56 (1989) 559.
[118] W. Sprenger, W. Hoff, J. Armitage, K. Hellingwerf, J. Bacteriol. 175 (1993) 3096.
[119] J. van Beeumen, B. Vreese, S.M. van Bun, K. Hoff, W.D. Hellingwerf, T. Meyer, D. McCree, M. Cusanovich, Protein Sci. 2 (1993) 1114.
[120] G. Borgstahl, D. Williams, E. Getzoff, Biochemistry 34 (1995) 6278.
[121] J. Pellequer, K. Wager-Smith, S. Kay, E. Getzoff, Proc. Natl. Acad. Sci. USA 95 (1998) 5884.
[122] T. Gensch, C. Gradinaru, I. van Stokkum, J. Hendriks, K. Hellingwerf, R. van Grondelle, Chem. Phys. Lett. 356 (2002) 347.
[123] T. Meyer, M. Cusanovich, G. Tollin, Arch. Biochem. Biophys. 306 (1993) 515.
[124] J. Hendriks, W. Hoff, W. Crielaard, K. Hellingwerf, J. Biol. Chem. 274 (1999) 17655.
[125] J. Hendriks, I. van Stokkum, K. Hellingwerf, Biophys. J. 84 (2003) 1180.
[126] D. Pan, A. Philip, W. Hoff, R. Mathies, Biophys. J. 86 (2004) 2374.
[127] A. Xie, L. Kelemen, J. Hendriks, B. White, K. Hellingwerf, W. Hoff, Biochemistry 40 (2001) 1510.
[128] C. Bernard, K. Houben, N. Derix, D. Marks, M. van der Horst, K. Hellingwerf, R. Boelens, R. Kaptein, N. van Nuland, Structure 13 (2005) 953.
[129] J. Vreede, W. Crielaard, K.J. Hellingwerf, P.G. Bolhuis, Biophys. J. 88 (2005) 3525.
[130] M.L. Groot, L. van Wilderen, D.S. Larsen, M.A. van der Horst, I.H. van Stokkum, K.J. Hellingwerf, R. van Grondelle, Biochemistry 42 (2003) 10054.
[131] S. Rajagopal, S. Anderson, H. Ihee, V. Srajer, M. Schmidt, R. Pahl, K. Moffat, Biophys. J. 86 (2004) 83A.
[132] E.J.M. Leenders, L. Guidoni, U. Röthlisberger, J. Vreede, P.G. Bolhuis, E.J. Meijer, J. Phys. Chem. B 111 (2007) 3765.
[133] J. Vreede, K.J. Hellingwerf, P.G. Bolhuis, Proteins: Struct. Funct. Bioinf. 72 (2008) 136.
[134] F. Noe, C. Schuette, E. Vanden-Eijnden, L. Reich, T.R. Weikl, Proc. Natl. Acad. Sci. USA 106 (45) (2009) 19011.
[135] T.S. van Erp, Phys. Rev. Lett. 98 (26) (2007) 268301.
[136] P.G. Bolhuis, J. Chem. Phys. 129 (11) (2008) 114108.
[137] J. Rogal, W. Lechner, J. Juraszek, B. Ensing, P.G. Bolhuis, The reweighted path ensemble, J. Chem. Phys. 133 (17) (2010) 174109.