[Frontiers in Bioscience 14, 2801-2828, January 1, 2009]

Two-state protein folding kinetics through all-atom molecular dynamics based sampling

Peter G. Bolhuis

van't Hoff Institute for Molecular Sciences, University of Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands

TABLE OF CONTENTS

1. ABSTRACT

This review focuses on advanced computational techniques that employ all atom molecular dynamics to study the folding of small two state proteins. As protein folding is a rare event process, special sampling techniques are required to overcome high folding free energy barriers. Several biased sampling methods enable computation of the free energy landscape. Trajectory based sampling methods can assess the kinetics and the dynamical folding mechanisms. Proper sampling is only the first step, and further analysis is required to obtain the folding mechanisms reaction coordinate. Only a combination of several simulation techniques can solve the sampling problems connected with all-atom protein folding, and allow computation of experimental observables that can validate the force fields and simulation techniques. Several of the involved issues are illustrated with folding of small protein (fragments) such as beta hairpins and the Trp-cage mini protein.

2. INTRODUCTION

2.1 Defining the protein folding problem

Protein folding has received much attention in the physical sciences, largely because folding is the key to protein design. Predicting the native structure from only the linear sequence promises the creation of de novo designed proteins with specific function. In a living cell, after being synthesized in the ribosome, proteins fold in to a specific 3-dimensional native structure in order to become active (1). Already about fifty years ago it was recognized that long linear polypeptides chains can adopt this native structure starting from a random coil in a surprisingly short time (2). A random search cannot explain this behavior, as is expressed in the Levinthal paradox. Protein folding has since become a paradigm of a complex transition with aspects of phase transitions as well as of chemical reactions. Elucidating the origin of proteins stability and folding kinetics are important parts of the understanding of protein function in general. Moreover, the folding properties of proteins are important for the understanding of the basis of well known degenerative diseases, such as Alzheimer's, BSE, Creuzfeld-Jacob, ALS, Huntington's, Parkinson's disease, and many cancers and cancer-related syndromes (3).

Since the advance of molecular simulation in the 1970's, proteins have also received the attention of the simulation community. In particular the development of molecular dynamics (MD) in conjunction with accurate atomistic force fields, has had a great impact. In this review I will focus on the use of all atom atomistic molecular dynamics simulations to investigate folding properties. There are many factors that can be studied are with MD, for instance, the influence of the solvent, salt concentration, the temperature dependence, and non-equilibrium pulling, and the effect of other molecules in the environment such as denaturant, osmolytes, or even chaperones. I restrict myself to spontaneous equilibrium folding at ambient conditions in water. This is considered the canonical folding problem. For a review on other conditions I refer to an excellent review by Daggett (4). As spontaneous two state folding is a rare event, taking place on timescales that cannot be (easily) reached by straightforward MD, special rare event methods are required. First of all, free energy methods are indispensable to obtain insight in the equilibrium properties of proteins. After a short introduction on the background of protein folding, I will discuss several of such methods, notably umbrella sampling, Metadynamics, and replica exchange (5,6,7). Subsequently, I will discuss several methods developed to study the kinetic aspects of protein folding: parallel replica, high temperature MD (8,4). Several of the issues related to simulation of kinetics have been adequately described in a recent review (9). However, as I will argue, path based methods are necessary to get truly unbiased insight in the dynamics of protein folding in explicit solvent. Of these I will focus on the transition path methodology (10). The sampling is only a part of the solution, and analysis is just as crucial. Recent developments in reaction coordinate analysis allow insight in the mechanism and the extraction of important parameters. Thus, the theoretical study of protein folding by means of all-atom force fields depends not on a single simulation method, but rather on a combination of several complementary techniques.

2.2 Native state stability

Proteins spontaneously fold to their native state because that state has a lower free energy (11). In the unfolded (extended or denatured) state (In this review a state is defined as a thermodynamically (meta) stable state, consisting of many possible protein configurations.) the protein is solvated, i.e. the backbone and side groups form hydrogen bonds with the solvent. When folding proceeds these solvent hydrogen bonds are replaced with often energetically more favorable non covalent bonds within the protein, for instance hydrogen bonds or salt-bridges (11). Multiple intra-molecular hydrogen bonds and salt bridge stabilize secondary structures like helices and sheets. At the same time the configurational entropy of the proteins backbone reduces as it nears the native state. Instead of many possible conformations in the unfolded state (U), the native state (N) confines the protein in a single conformation. This negative entropic effect is offset by a gain in entropy of the solvent molecules. In addition, some of the hydrophobic side chains are buried in the protein during folding, reducing the energy (enthalpy) as well as increasing the solvent entropy. While each of these effects can be large (thousands of kJ/mole), all these effect taken together mostly cancel each other, and the free energy difference between the unfolded and native state is often only of a few tens of kJ/mol. This marginal stability imposes severe constraints on the accuracy of the molecular modeling. Both the interaction energies as well as the entropic contributions (configurations, vibrational, rotational) have to be taken into account accurately, in order to predict the folding behavior. Classical force fields can describe the energy of the protein as long as quantum mechanical processes do not play an important role, but the entropy can only be obtained from statistical Boltzmann sampling. Together with the high computational expense of modeling a protein by an all atom model this makes a statistical mechanical approach to protein folding a challenge.

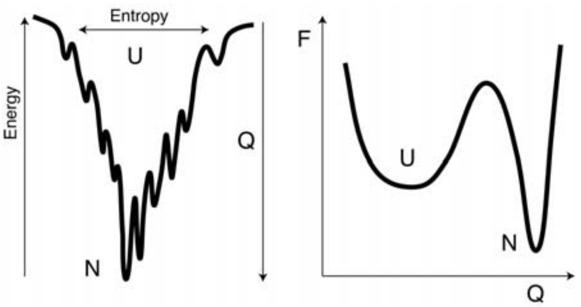

Figure 1. Left: The curved solid line depicts a schematic energy landscape funnel of the folding process. The x-axis represents the accessible configuration space. At high energies, many configurations are possible; hence the entropy is large, as indicated by a wide funnel. The landscape guides the protein to lower energies where the protein has less configuration freedom. During this process the reaction coordinate Q increases. The global minimum corresponds to the native state. Note that the transition state region does not appear naturally in this view. Right: Schematic free energy landscape of the protein folding process as a function of the reaction coordinate Q shows, two stable states separated by a barrier (the transition state). This picture combines the entropy and energy of the funnel, and views folding as a unimolecular reaction.

2.3. Two-state behavior

Because of its marginal stability, a protein in a native state can be forced to unfold relatively easy, e.g. by heat, adding denaturant or a change in pressure or pH. Such change of environment destabilizes the folded state or equivalently stabilizes the unfolded state. Anfinsen showed in 1973 that when reverting to the original conditions a single domain protein finds is way back to the folded state (2). In addition, it was found that such transitions in proteins exhibited two-state kinetics. This means that the relaxation toward the equilibrium population can be described by a single exponential with a single rate constant, obeying an Arrhenius-like temperature dependence (That proteins also often show non-Arrhenius behavior of protein folding highlights that it is not that simple). This finding established that the unfolded state and the native state are the most stable states, and all other possible states are at most metastable and are hardly populated. Assuming there is no interference of different protein molecules in the solution this concept identifies the kinetics of protein folding as that of a unimolecular chemical reaction (U↔N). On the other hand, the heat capacity peak that occurs around the folding temperature resembles that of first order phase transition. Indeed, this peak is caused by cooperative behavior of the protein (12). Both these observations allow the use of statistical mechanics and simulation techniques originally developed for chemical kinetics and first order phase transitions for investigation of protein folding.

2.4. Energy landscapes guide the folding mechanism

Folding proteins do not randomly search all conformations but are guided by a more or less funnel-shaped underlying energy landscape (see Figure 1). When proteins are prepared in the extended state, for instance at high temperature, they have much entropy (13). A sudden temperature decrease will lead to more population in the native state. During this process the protein, due to hydrophobic interaction, electrostatic and dispersion forces, first collapses into a molten globule: a metastable state in which there are much fewer conformations available. Only a few of those conformations lead to the native state in which the energy is low, but the entropy as well. Adapted by evolution the energy landscapes of natural proteins exhibit a single, stable native state, separated from any misfolded state by a relatively large energy gap. In contrast, random heteropolymers show a more glassy energy landscape with many degenerate misfolded states (14,15).

Before the protein can enter the native state it has to pass a transition state, a state with low entropy but without most energetically favorable contacts. Plotting the free energy as a function of the reaction coordinate Q (Figure 1), the transition state is on top of the free energy barrier to folding (16). Overcoming this barrier is the rate-limiting step in the folding mechanism. Thus, this free energy barrier towards folding is the origin of the two-state behavior.

For single domain two state folding proteins several decades of experimental, theoretical, and simulation studies have revealed two major qualitative folding mechanisms that explain this barrier (see Figure 2). The nucleation-condensation model (17,19) states that the protein in the transition sate forms a nucleus of native contacts which can further grow towards the native state. The diffusion-collision model (also known as the framework model) assumes a very fast secondary structure formation (18,19,20) followed by diffusion of these secondary structure elements until they collide into the correct tertiary structure (20). In recent years, these two mechanisms were combined in a unified view (21). Which of these two scenarios holds depends on the protein. Some two-state proteins, have an independently stable secondary structure, and fold according to the diffusion-collision mechanism, while other proteins, with a less stable secondary structure, fold cooperatively along the nucleation-condensation scenario. Two state folding behavior is prevalent in both cases as there is only one rate-limiting barrier.

The last step in the folding process towards the native state is often the expulsion of water molecules. When the protein is close enough to the native state to expel the interstitial water molecules and form a nucleus (22), the energy of stabilization (hydrogen bonds etc) together with the gain in translational entropy of the solvent (23) start to overcome the decrease in conformational chain entropy. In many cases the barrier to folding coincides with the water expulsion (4).

2.5. Folding rate determining factors

In the transition state theory (TST) framework the rate constant of folding k_f is exponentially dependent on the height of free energy barrier (11) .

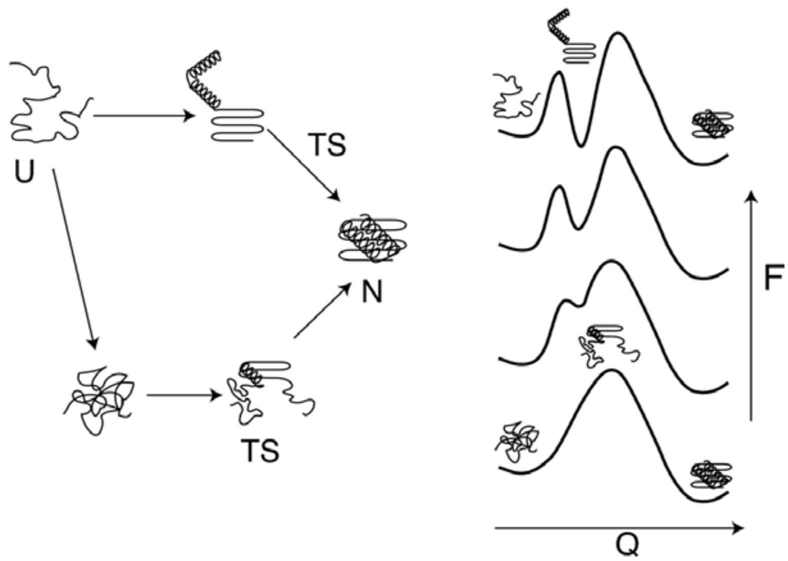

Figure 2. Left: Cartoon of the two main mechanisms of folding. The top route denotes the diffusion collision model. The secondary structure elements are stable and form fast, and diffuse until the native structure is found. The bottom pathway corresponds to the nucleation condensation mechanism. The secondary and tertiary structure can only fully develop into the native state after a folding nucleus (TS) has been formed first. Right: Unified global folding free energy landscape view for the diffusion-collision mechanism (top) can change gradually into that of a nucleation-condensation mechanism (bottom) depending on the stability of the secondary structure elements (21). In all cases two-state behavior is obeyed as the left barrier is lower than the highest barrier**.**

$$k_f \sim e^{-\beta \Delta G_{\ddagger}}, \quad (1)$$

where $\Delta G_{\ddagger}$ is the free energy difference between the barrier and the denatured state, and $\beta = 1/k_B T$ is the reciprocal temperature, with k_B Boltzmann's constant. However, there is an enormous variation of experimental folding rates for small single domain two state proteins While the fastest (small) proteins fold on the order of a microsecond (24), proteins like chymotrypsin inhibitor-2 fold on the order of seconds (11). The native state topology has been proposed as a major rate-determining factor. The topology can be partly characterized with contact order. A contact is formed when the alpha-carbons of two residues are within a certain cutoff distance (often 6 Å). Contact order measures how far contacts are separated along the sequence on average. Tertiary contacts have thus high contact order and alpha helices a low contact order. Contact order correlates well with folding rates (25). This finding is rationalized in the so-called topomer model, which states that the protein most of the time is searching for its native topology, after which the protein can fold relatively fast to the native state (25).

In contrast, Mirny and Shakhnovich argue that contact order is mostly not dominated by tertiary contacts, but in fact more by local contacts. In their view, alpha helices have a low contact order, while beta-sheets have more distant contacts and hence a higher contact order. Indeed helical proteins fold faster than beta-proteins, providing a rather trivial correlation between rate and contact order (19). Nevertheless, recent work by Dill *et al.* (26) suggests that the bottleneck in folding process is the search for the native state, thus lending credit to the topomer model.

2.6. Investigating the transition state with phi-analysis

The folding transition state of the rate-determining step can be studied experimentally with so-called phi analysis (11). Phi analysis consists of making single point mutations along the sequence and probing the effect on the kinetics. The phi value of this mutant is then defined as the ratio of the change of free energy of the transition due to mutation and the change of free energy of the native state due to the mutation.

$$\phi = \frac{\Delta\Delta G_{\ddagger}}{\Delta\Delta G_N} = \frac{\Delta G_{\ddagger}^m - \Delta G_{\ddagger}}{\Delta G_N^m - \Delta G_N}, \quad (2)$$

where the free energy differences ΔG is always measured with respect to the denatured state. ΔG_N is thus the free energy of the native state, or the folding free energy. $\Delta G_{\ddagger}$ denotes the folding free energy barrier with respect to the denatured state. The superscript m stands for the mutant. The $\Delta\Delta G$ is therefore the change in free energy difference of the transition state upon mutation. The phi value reveals the involvement of the mutated residue in the transition state structure. If the residue has a native-like structure in the transition state ensemble, then mutating it will alter the native as well as the transition state energetics, and the phi value will be close to unity. On the other hand, when the residue is completely unfolded in the transition state, changing it to different amino acid will not influences the energies of the transition state that much, and hence the phi value will be zero. Experimentally the phi values can be extracted from measured (un)folding rate constants assuming that transition state theory (TST) applies. In this way mutation studies allow for indirect structural interpretation of the transition state (11).

3. PROTEIN SIMULATIONS

3.1. Molecular simulation

Protein folding transition is ultimately governed by a balance between conformational entropy and stabilization energy, and requires a statistical mechanical description. In the last few decades molecular simulation methods has proved indispensable, amongst others, for studying phase behavior in complex systems, but also for obtaining microscopic insight in the structure and dynamics of proteins. Simulations are capable of sampling phase space without making any approximations besides the molecular interactions. The most well known simulation methods comprise molecular dynamics (MD), Langevin dynamics and Monte Carlo (MC). Monte Carlo is an importance sampling of phase space, using the criterion of detailed balance to converge to the correct canonical isothermal Boltzmann distribution. While MC is an important simulation method, able to simulate lattice models and other discrete models, it is less useful for the study of proteins using all atom models. Therefore, I limit myself in this review to molecular (and Langevin) dynamics.

In Molecular (or Langevin) Dynamics one integrates the deterministic Newtonian (or the stochastic Langevin) equation of motion using the instantaneous intermolecular forces between all atoms (27,5). This procedure leads to a time series of molecular configurations usually called a trajectory. This trajectory contains all dynamical information, and can be used for kinetics. Deterministic Hamiltonian dynamics conserves energy and, according to the ergodicity theorem, samples the micro-canonical (constant energy) ensemble in the limit of long time. Application of a thermostat, e.g. the Nose-Hoover thermostat, or Andersen thermostat (5) or the recent Bussi thermostat (28) allow for canonical sampling. In addition, a barostat can be used for the isobaric ensemble (29). Langevin integrators are necessary for situations in which many degrees of freedom have been integrated out. For instance, when the solvent is replaced by an effective medium, but also when using coarse grained models.

3.2. Molecular models for proteins

For proteins a host of (semi) empirical atomistic force fields have been developed including ENCAD(30-31),AMBER(32), CHARMM(33), GROMOS(34), and OPLSAA(35). Based on a potential form that includes bond, angle dihedral, van der Waals, and electrostatic terms, parameters in these force fields are fitted to *ab initio* results, and/or experimental data. The water model e.g. TIP3P or SPC is often included in the force field. Recent work has focused on including solvation free energies, and polarization (36,37). Currently, force fields can reproduce structural features reliable, although the prediction of accurate relative stability (i.e. free energy) is still elusive.

While by far not as computational intensive as quantum based simulation, all-atom MD is a relative expensive method, and much effort has been put into ways to speed up the code. Among the tricks that speed up the computation, are the fast Ewald summation for the electrostatics, bond constraints, multiple time-steps, and Verlet neighbor lists for the van der Waals interactions. There are many packages exist that can perform MD efficiently. Nevertheless, with current standard computer power, MD is roughly limited to 10^5 atoms and 1 microsecond

As for most proteins systems, more than 80 percent of the systems consist of water, it is convenient to replace the solvent with an implicit solvent model. Often used implicit solvents are the effective energy function (EEF1) (38) and the Generalized Born/surface area (GB/SA) model (39). EEF1 uses the solvation free energies of the amino acid side-chains, and treats electrostatic interaction via a distance-dependent dielectric constant. The GB/SA model employs a combination of the solute-solvent electrostatic polarization (GB) and the solvent accessible surface of the solute (SA) to approximate the cost of the creation of a solvent interface and the van der Waals interactions.

While this review is about accurate all-atom force fields, for many protein-folding studies such an approach is prohibitively expensive even when using implicit solvent. In that case, coarse-grained potentials are often used as a cheaper alternative. In a coarse-grained (CG) model several atoms are lumped together in a single particle. The coarse-grained particle interacts via a simplified potential with all the other CG particles. This simplification emerges because of the integrating out of degrees of freedom. Popular coarse-graining force fields include the basic Go model in all its guises, which is based on the native PDB state, as well as simplified off-lattice models see e.g. Refs (40-44). Most of these force fields have been devised to reproduce several properties. Other coarse graining force fields are obtained by actually carrying out the integrating out the degrees of freedom based on all atom force fields (45,46).

Even simpler than the above off-lattice coarse-grained potentials are the lattice models. Using a linear heteropolymer on a lattice, many statistical mechanical theories of protein folding have been tested (47,48, 16). Lattice models of proteins can be employed to test theories, and find generic folding behavior, including substrate induced folding, protein binding, disorder-order transitions associated with signaling proteins and even translocation (19,49,50). While lattice models undoubtedly yield global insight into the statistical mechanics of folding they are not probable candidates for giving molecular insight into the kinetics of protein folding, because there is not enough molecular information in these models, besides the linearity of the polymer and the interaction matrix.

3.3. Order parameters

The outcome of a molecular simulation is a sequence of system configurations. In case of MD and

Langevin Dynamics this is usually a molecular trajectory, and in case of MC it is an ensemble of configurations. The main point is that as long as these trajectories are a large file of numbers sitting on a hard disk, not much insight is to be gained. Visualization of the molecular trajectory is an option, and many visualization packages especially aimed at proteins have been developed (a well known example is VMD (51), but many others exist). Nevertheless, in many cases only visualizing a 3D representation of protein is insufficient., and a more quantitative analysis is desirable. Such quantification can be done with the help of relevant order parameters, which reduce the 3N dimensional configurations to just a few dimensions. In the literature many order parameters have been developed for proteins, amongst others, the number of native contacts (A contact is made when the alpha-carbons 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. the PDB structure), the proteins radius of gyration, the root means square deviation from the native state (Computing the average square atomic distance between the two structures yields the mean square distance. The RMSD follows from minimizing this means square distance with respect to translation and rotation, and finally taking the square root.), dihedral angles, solvent accessible surface, number of hydrogen bonds, distances between relevant groups, salt bridges, bonds distances, combinations of bonds, contact order. Phi values are hard to compute, as they involve the free energy barrier. However, a simple approximation of the phi value is through the ratio of the number of native contacts per residue (52). The rationale behind this is that similar to the phi value, the fraction of contacts gives the degree to which the environment of the residue is native-like (53). With this assumption the phi value can act as a simulation constraint to find transition states.

3.4. What can simulation do for us?

Performing a molecular dynamics simulation based on a classical force field, yields in principle, an accurate description of the system dynamics, within the accuracy of the force field. If the sampling of phase space is ergodic, it will give the equilibrium distribution of the systems of interest. This includes the stable states of the protein, as well as possible low populated intermediates. In addition, because the dynamics is unbiased, insight in folding kinetics and mechanism is in principle available. This insight also extents to the role of the solvent in the folding reaction, as well as chemical factors such hydrophobicity, electrostatics, hydrogen bonding and steric effects.

The goal of simulation is to provide new insight, interpret experimental results, test theories, and predict new phenomena (5). To do that MD must be able to predict experimentally accessible observables, such as thermodynamics, population, and the structure of the stable states and the intermediates. In addition there are kinetic quantities such as the folding rate, transition states, diffusion constants, Arrhenius factors, phi values etc. All this requires obviously an accurate force field, and one of the basic tasks in the simulation community is to validate such a force field by comparing predicted properties to experiments. While much progress have been made (54), different force fields predict slightly different results, and no force field is entirely perfect in correctly simulating protein dynamics (55).

Once a good force field exists, it can be used to predict experimental observables of novel proteins. In that case the MD approach is only hampered by the above-mentioned sampling problem. This (large) obstacle can be ameliorated by the development of faster computers (or, for instance, the use of graphics cards developed to improve efficiency of MD), or special accelerated software (such as assembly loops for fast computation of special functions, fast Fourier transforms etc). Here, we focus on the sampling algorithms. The interpretation of MD trajectories should be done, in fact, with statistical mechanics. After all, the predicted MD trajectory will not be realistic, i.e. will not correspond to the way each molecule reacts. We can only take MD results seriously by taking ensemble averages of the trajectories. For an ergodic system, this is a good way to obtain equilibrium properties of one stable state, e.g. the native state. For dynamic observables such as the rate, a good approach is to rerun the same simulation many times, with different initial conditions and average over these results (in fact this is the basis of the parallel replica methods, see next section). However, when the sampling of the reaction is very slow because of the existence of free energy barriers between the stable states, this might not be the most efficient method. The approach that is pursued here is to improve the sampling itself by using specially developed techniques.

4. EXPLORING THE FREE ENERGY LANDSCAPE

4.1. Free energy of folding

As folding is governed by thermodynamics, a key quantity of interest in a computational protein folding study is the free energy. To be more precise, one is interested in the free energy difference ΔF (or the Gibbs free energy ΔG at constant pressure) between the (meta) stable states in the folding process. The free energy can be related to the experimentally found equilibrium population. The important states in a protein obeying two-state kinetics are the native and unfolded states. Possible on-pathway intermediates, misfolded states, or other metastable states are supposed to be not well populated. The other important quantity is the free energy barrier to folding/unfolding, which is related to the experimental rate.

The free energy can be estimated via several molecular simulation techniques. A straightforward MD, Langevin Dynamics, or MC simulation can in principle yield all thermodynamic information. Keeping track of the population p_i of a state i during the simulation, the relative free energy of state i is $\beta F_i = -\ln p_i$. A usual tactics is to plot the free energy as a function of order parameters. Up to a constant this Landau free energy (5) is given by

$$\beta F(\{\lambda_i\}) = -\ln P(\{\lambda_i\}) + \text{const}$$

where $P(\{\lambda_i\})$ is the probability to find the system at certain value of the set of order parameters $\{\lambda_i\}$ This probability is

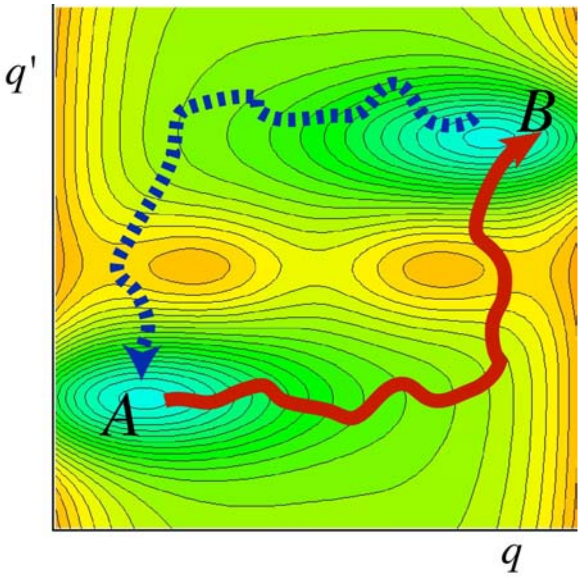

Figure 3. Schematic free energy contour plot of a two state system. The A and B minima are separated by a high barrier with three saddle points. Hysteresis can occur if the used order parameters (here represented by q) do not describe the reaction adequately. Solid line: Initiated in state A, the ensemble is biased along the q axis, until the transition to B occurs. Dashed line: the reverse biased sampling along q will take a different route. In both cases, the true transition state (in the center of the plot) is not sampled at all.

obtained by integrating out all degrees of freedom in the configurational partition function, except the set of order parameters $\{\lambda_i\}$ that are of interest:

$$P(\{\lambda_i\}) = \frac{\int dr \exp(-\beta U(r)) \prod_i (\delta(\lambda_i - \lambda_i(r)))}{\int dr \exp(-\beta U(r))} \quad . (4)$$

Here, *r* denotes the coordinate vector for all *N* particles, *U(r)* is the potential energy, and λ(r) the instantaneous value of the order parameter λ and δ(x) is the Dirac delta function. Note that this expression can be used irrespective of the simulation method as long as one samples the canonical distribution. Both straightforward MC and MD can thus serve to find the free energy by simply histogramming the values of the order parameters (5). While it might seem more clear that an MC simulation samples this distribution directly, according to the ergodicity theorem the probability also follows from straightforward from an long (in principle infinite) MD simulation.

The problem lies of course in the fact that the convergence of a free energy difference requires a very long MD trajectory, in fact, much longer than the reaction time of the problem. As most proteins fold beyond microseconds, a straightforward approach is not practical. These long timescales are due to high free energy barriers between the stable states. The problem is severe, because the time scale grows exponentially with the barrier height. Fortunately, one can make use of algorithms that are specially designed to overcome high free energy barriers.

4.2. Biased sampling

One of the oldest methods to tackle the problem of high barriers is the Umbrella Sampling (US) technique (5). This technique introduces a bias as a function of an order parameter into the partition function that enables a more even sampling across the order parameter range:

$$P_{bs}(\{\lambda_i\}) = \frac{\int dr \exp[-\beta U(r) + W(\{\lambda_i(r)\})] \prod_i \delta(\lambda_i - \lambda_i(r))}{\int dr \exp[-\beta U(r) + W(\{\lambda(r)\}]}$$

. (5)

To obtain the free energy one can apply equation 3 but with a correction for the bias

$$\beta F(\{\lambda_i\}) = -\ln P_{bs}(\{\lambda_i\}) - W((\{\lambda_i\}) + \text{const} \quad . (6)$$

To obtain an evenly distributed sampling P_{bs} is difficult because it requires a biasing function that is exactly the free energy function one is looking after. In practice, one can use a fixed potential like a simple harmonic potential to increase the likelihood of crossing the barrier. Multiple runs for the bias at different locations lead to histograms spanning the entire order parameter space. These can be joined together using the weighted histogram analysis method (56,57). Sampling more than one order parameter is possible using an adaptive scheme of US (57), but going beyond 2 order parameters has been unfeasible up to now.

An alternative to US is the Metadynamics method, which allows for sampling up to 6 dimensions (6). This method has proven fruitful for small proteins and will be discussed in the next section. Other recently developed methods focusing on multidimensional biasing functions include flooding (58), hyperdynamics (59), multicanonical sampling (60) and the adaptive biasing force method (61).

The efficiency of biasing depends very much on the choice of order parameter. The US method relies on a good sampling of the phase space in direction orthogonal on the order parameters. When there is an additional barrier in the direction perpendicular to the biasing variables, hysteresis can occur (see Figure 3). This hysteresis manifests itself when one tries to retrace the free energy back in the reverse direction. Suppose that unfolding a protein is forced by biasing the radius of gyration. Then the reverse biasing process from the unfolded state does not necessarily end in the native state and a different path is traced. Hence, a wrong free energy difference and barrier is estimated between the unfolded and the native states. Shea and Brooks (57) try to circumvent hysteresis problem partly by initiating an umbrella sampling run at 298 K with high temperature unfolding trajectories. However, this only works if the high temperature unfolding trajectories are sufficiently close to the folding process at ambient conditions. Still, such an approach is inherently biased by the high temperature trajectories, which can sample parts of phase space that would never be sampled at room temperature due to barriers in directions orthogonal to the biasing parameter. Umbrella sampling can only explore phase space locally. Because of this problem of the order parameter choice, parameter free methods like parallel tempering/replica exchange have become popular in the last decade (see section 4.4). Nevertheless, umbrella sampling is an important and versatile method that can be applied in many ways. (5,57).

4.3. Metadynamics

Related to umbrella sampling, the Metadynamics technique aims to explore the free energy landscape efficiently in a predefined multidimensional space of collective variables (6,62). The method forces the system to explore other parts of configuration space by employing a history dependent biasing potential that slowly pushes the system out of a free energy minimum. The bias potential is regularly updated based on the instantaneous position of the system:

$$V_{meta}(\vec{r},t) = \sum_i w_i \sum_{t'<t} \exp\left(-\frac{[\lambda_i(\vec{r}) - \lambda_i(\vec{r}(t'))]^2}{2\sigma_i^2}\right), \quad (7)$$

This time dependent bias potential places Gaussians of width σ_i and height w_i for each order parameter λ_i. As the simulation progresses, parts that have been visited are becoming less favorable, and the system is pushed to a different global state. Eventually, when the basin of that state is also filled with Gaussians, the system is forced to move back to the original state (provided there are only two states). In the limit of long time, the bias potential will converge to the negative of the free energy as a function of λ_i. In that case, the phase space will be sampled evenly. This result is similar to the case of umbrella sampling, with the difference that Metadynamics employs a highly adaptive bias. The Metadynamics method is also reminiscent of the Wang-Landau sampling method, which continually updates a bias function to achieve flat distribution sampling (63). The accuracy of the result is governed by the height and width of the Gaussians. The smaller these are, the more features of the free energy landscape can be resolved. In practice, the height and width of the Gaussian are adapted in the course of the Metadynamics simulation. Naturally, Metadynamics is very useful in combination with MD. While due to the time dependent bias the dynamics clearly will differ from the true dynamics of the system, Metadynamics is a powerful tool for the exploration of phase space and the computation of the free energy landscape. Still, the method, depends on an appropriate definition of the collective variables, and is in that sense not different to umbrella sampling. In the field of protein folding, Metadynamics has been applied to hairpins (64) and the Trp-cage protein (65). In both cases the sampling was enhanced by making use of replica exchange techniques.

4.4. Replica exchange/parallel tempering

The necessity and dependence on an *a priori* correct choice of order parameter is a severe drawback of the biased sampling methods described above. The Parallel Tempering (PT), a.k.a. the Replica Exchange Method (REM), has the large advantage of not requiring such a definition. REM is a Monte Carlo algorithm to sample free energy landscapes with many local minima (5,66,67). In effect, PT/REM heats up the system periodically in a simulation to help the system escape local minima, and then lower the temperature again to ambient condition. A naive implementation of heating/cooling cycles does not obey the Boltzmann distribution. The Monte Carlo algorithm connected to PT/REM does restore the detailed balance. The method requires the simultaneous running of many replicas at different temperature in parallel. Occasionally, a random swap between replicas is tried. This trial move is then accepted with a probability

$$P_{acc} = \min(1, \exp(\beta_{ij}\Delta U_{ij}), \quad (8)$$

where $\Delta\beta_{ij} = \beta_i - \beta_j$ is the difference in reciprocal temperature between the two replicas, and $\Delta U_{ij} = U_i - U_j$ is the difference in potential energy. This acceptance probability decreases exponentially for large temperature gaps and large energy differences (which scales linearly with the system size. That is why in practice many replicas are required to obtain a reasonable exchange probability. The temperature distribution can be chosen such that the swapping probability is optimal. During the REM run, the replicas thus diffuse through the temperature space, heat up and cool down. Still, all replicas are distributed according to the canonical distribution. Sugita and Okamoto (67) showed that REM can be combined with MD, and used to study proteins. Since then many such REMD studies have been conducted. For instance, Garcia and Onuchic (68) have applied the REMD approach to the folding of the tree helix bundle protein A. Lei *et al.* (69) performed a large scale REMD simulation of Villin headpiece in implicit solvent. They compute the FE landscape and predict the *ab initio* folding structure.

While replica exchange 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. In principle, REMD conserves the Boltzmann distribution, but its sampling efficiency depends on visiting both high and low temperature regions many times during the simulation and, more importantly the transition of interest should happen spontaneously at high temperature. The exchange of replicas should be fast enough to allow good diffusion of the replicas, but at the same time slow enough to allow the replicas to adapt. While a doubling of the temperature speeds up dynamics, for protein folding it often shifts the equilibrium to unfolding. This also negatively influences the convergence of PT/REM for protein folding in explicit solvent (70,71). Seibert *et al.* (72) estimated the convergence time takes about 200 ns per replica for a small beta-hairpin in explicit solvent. (72). Periole and Mark (73) concluded from a REMD simulation study on a heptapeptide, that REMD is more efficient than straightforward MD. However, they argue that it is not so such the direct speed up of the dynamics that can explain the efficiency. Rather, there is a large sorting effect due mixing of the replicas, which quickly seem to stabilize and can give a false sense of convergence (73). One way to improve on the slow convergence is to do away with the large energy of the explicit solvent. Berne and coworkers adapted the REMD scheme such that it only depends on the potential energy of the solute. A smaller potential energy allows for a reduction of the number of replicas, and an enhanced diffusion in temperature space leading to an enhancement of the sampling (74).

It is possible to combine the umbrella sampling and PT/REM by again adding a biasing function that is now a function of both the order parameter set $\{\lambda_i\}$ and the temperature T. The advantage of such a combination might be that even though the dynamics at high temperature might be faster, it is not a priori certain that all relevant states are well sampled at high temperature. This is in particular a problem for folding processes in which one state is stabilized by entropy, and the other by energetic interaction. Including a biasing function might be able to improve that, but, of course, assumes one already has a proper order parameter to bias in (75).

The PT/REM methods can also be combined with Frenkel's recently proposed scheme of waste recycling (76). In this scheme Monte Carlo trial moves are not just discarded as useless but actually contribute to the average. It is straightforward to extend this to PT (75), improving the statistics on the free energy dramatically. Note however, that the waste recycling scheme does not improve the sampling itself, but only the accuracy of the histograms. The PT/REM method is not limited to temperature exchange; other parameters can be used for the replica exchange (5). For instance, several approaches propose changing the Hamiltonian itself, instead of the temperature (77).

The REM has also been combined with Metadynamics (64), leading to improved sampling. Also, Piana and Laio (65) have developed a replica method in which the bias functions in collective variables are exchanged instead of the temperature. In this bias exchange method, several replicas bias in different order parameters, and occasionally exchanging replicas are tried. This approach seems particularly fruitful because the Metadynamics can in this sense be effectively extended to many dimensions.

5. PROTEIN FOLDING KINETICS AND MECHANISM

The previous section focused on extracting thermodynamic information from protein simulation, in particular the free energy. If the connection to the underlying dynamics of the system can be made, computer simulation can also access folding kinetics. Even lattice models can yield insight in folding kinetics, albeit of a more generic kind (47). When interested in the folding mechanism and rate of specific small proteins all-atom MD with explicit water seems to be the method of choice. The first reason is that it does obey realistic dynamics (but only as accurate as the force field), and contains all kind of molecular information such as side chain packing and dynamics. Moreover, it gives insight in the behavior of the solvent.

Short alpha-helical peptides fold reasonably fast (see e.g. Ref. (78)), but folding real proteins was for a long time out of reach. More than a decade ago, the tour de force microsecond MD simulation of Duan and Kollman of the 36-residue villin headpiece in explicit water showed only partial folding (79). Nowadays, standard computational power still only allows simulation on the order of a microsecond. Moreover, even if a folding event takes place within a microsecond of MD simulation time, it is only one possible pathway out of the many available to the system, and many folding and unfolding events are needed for an accurate estimate of the rate. Therefore, it seems not be possible to study accurately the kinetics of even very fast folding proteins by direct MD. Reproduction of many folding and unfolding events using direct MD, Langevin/Brownian dynamics or MC is currently only possible for coarse-grained models or atomistic models with implicit solvent. Nevertheless, there are several methods that aim to simulate the kinetics of folding for all-atom molecular simulation.

5.1. Parallel replica molecular dynamics

The first of these methods is the parallel replica method (not to be confused with REM). If one is interested in how often a protein passes through the transition region, one could start a MD simulation in a stable state, and just wait for it to make a transition over the barrier. The average time it takes the protein to cross the folding barrier is the so-called mean first passage time (MFPT) and is related to the folding rate by $k = 1/\tau_{MFPT}$. In general, it is extremely costly to calculate MFPT for atomistic models with explicit solvent by brute force. However, because for high free energy barriers the distribution of passage times is Poisson distributed, a few of the crossings take only a short time (these times are obviously offset by trajectories that take much longer that the MFPT). This observation is the basis for the parallel replica methods. This method runs many trajectories (replicas) in parallel each initiated from a different from an equilibrium ensemble (80). Clearly, to wait until all trajectories cross the barrier is as costly as running just a single trajectory. Therefore, as soon as one of the many replicas has crossed a barrier to a different stable state, all the replicas follow and are reinitialized in the new state (8). This procedure is then repeated until all possible transition has taken place. Pande *et al.* applied this method in combination with a distributed computing approach (15) using thousands of processors. For processes with an MFPT of microseconds, starting 10000 trajectories of a few nanoseconds from different initial conditions (but in the same initial state, e.g. the denatured sate) are likely to result in a few successful barrier crossings. However, there is no computational gain, because the other 10^4 trajectories remain just confined to the initial state. In addition, the crossing time of a few ns might be not representative, certainly not when the reaction has to occur via obligatory intermediate. Pande *et al* applied the parallel replica approach to the folding of villin headpiece (81) and of BBA5 (82). They found that the villin headpiece folding was governed by the formation of the hydrophobic core, while for BBA5 the (in fact experimentally very rapid) folding occurs due to fast secondary structure formation.

5.2. High temperature trajectories

In the replica exchange/parallel tempering approach raising the temperature increases kinetics dramatically. A doubling in temperature (e.g. from 300 to 600K) results in a kinetic enhancement, corresponding to taking the square root of the exponential Arrhenius factor. For an activation barrier of 10 k_BT the speed up is then around 150-fold, for 15 k_BT it is already 2000-fold. This brings the microsecond regime back to nanoseconds, and within reach of simulation. This effect is exploited in the high temperature simulations (see e.g. Daggett (4)). From high temperature MD simulations one can deduce qualitative knowledge about the folding mechanisms (83,84,85).

In a classic study Daggett and coworkers investigated high temperature unfolding of chymotrypsin inhibitor 2 (4,86), which is considered an archetypical single domain two state folding protein (11). This 64-residue protein has a native structure consisting of an alpha helix and a beta sheet. The protein folding kinetics has been studied by experiments, and the transition state has been studied by phi analysis. To obtain the phi values from first principles in an MD study requires the computation of either the folding rates of the wild type and the mutant, or the computation of the free energies. Both are fairly computational expensive, and Daggett *et al*, do not attempt this (4). Instead, they use a clustering analysis to compute the transition state ensemble. This led to a proposal of a single rate determining transition state ensemble. In silico mutation studies led to more insight in the structure and functioning of the transition states. In particular, several mutations for faster folding proteins could be predicted that were experimentally confirmed. The combination of experimental and theoretical work shows that the CI2 folds according to the nucleation condensation mechanism. The secondary structure by itself is not stable enough, and only when a nucleus of tertiary contact and secondary structure (small parts of the helix and the sheet) is formed, the protein collapses to the native state. It does so by expelling water from the interior, which is the last step to folding. While the simulated unfolding pathways have been performed at high temperature, it is claimed that the folding/unfolding pathways are relatively insensitive to temperatures for several proteins.

Daggett *et al.* also performed quenching studies of the CI2 TS ensemble (4). They find that after the quench the proteins collapse by expulsion of water, but what determines whether or not the protein refolds is the fact whether the water molecules in the protein core are bound to the main chain. Multiple quenched trajectories of a transition state also gave an indication of the refolding pathway. These and other (e.g. engrailed homeodomain) quenching studies indicate that the folding at room temperature is the reverse of unfolding at high temperature.

However, high temperature trajectories do not necessarily give the right mechanism of the reaction, and certainly do not yield the correct rate constant. While extrapolation of the rate behavior is possible in some cases using the Arrhenius expression, it is not necessarily reliable in the case of protein folding that can show non-Arrhenius behavior. Moreover, a high temperature trajectory is biased towards overcoming enthalpy barriers, whereas at low temperature the preferred pathway might be more entropy dominated. In addition, the force field is not parameterized for the high temperature, and as most simulations are done at constant volume, the pressures are unphysical high. In short, although there is much to be gleaned from high temperature trajectories one would prefer to do the simulation at ambient condition.

5.3. Rare event methods

5.3.1. Bennett-Chandler algorithm

Application of straightforward molecular dynamics at ambient conditions to protein folding is computational demanding because of the long times involved. These long timescales are directly related to the two-state folding barrier. The crossing of the folding barrier is a rare event, compared to the fundamental dynamical time step (usually femtoseconds). For this kind of barrier crossings one can use approaches such as the TST based Bennett-Chandler (BC) method (87,88). In this approach, the rate constant is equal to the equilibrium probability to be on top of the barrier, multiplied by a kinetic prefactor. The equilibrium probability is basically given by the barrier free energy maximum, with respect to the stable state. This free energy can be calculated in the usual manner, for instance with the methods mentioned in section 4. The second factor is the so-called transmission coefficient, and contains the dynamical information of the barrier crossing. This factor follows from starting many short trajectories from a constrained ensemble of configurations on the top of barrier. The transmission factor accounts for trajectories recrossing on the barrier: paths that initially seem to move into the direction of the final state but recross the barrier and return to the initial state. The value of the transmission coefficient depends on the quality of the order parameter used for reaction coordinate. When the reaction coordinate is a good one the transmission coefficient is close to unity and the BC approach works fine. However, for a complex problem for which the reaction coordinate is not so easily defined, the transmission coefficient can be vanishingly small, meaning that most configurations corresponding to the top of the barrier, are actually already in the basin of attraction of either the initial or the finals state. Thus the BC method suffers from the same hysteresis problem that was mentioned in the previous section: the so-called reaction coordinate problem.

5.3.2. The reaction coordinate problem

This reaction coordinate problem is prevalent in both simulation as well as experiments, because the barrier is not well populated, and cannot be observed in experiments as well as in simulation. Nevertheless, the rare crossing events are important for understanding the folding process. To get insight in folding, one requires knowledge of the reaction progression, e.g. when the transition state has been reached. A coordinate that does precisely this is the reaction coordinate. Moreover, such a coordinate should have a reasonably low dimension in order to understand the progress. For simple reactions such as between two atoms, the relative distance is likely to be a good order parameter. But for proteins the reaction coordinate is not so easily found. All atoms move simultaneously in a complex manner, and there is no

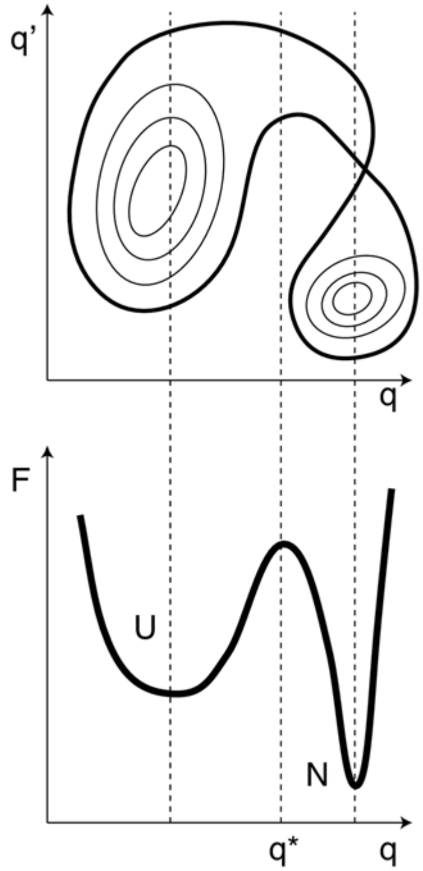

Figure 4. Generic free energy landscapes can illustrate the reaction coordinate problem in complex systems. If both q and q' are important ingredients of the reaction coordinate (top left panel), leaving out one of them, e.g. q', will lead to a wrong prediction of the transition state region (bottom panel), and hence a wrong mechanism and a statistically inaccurate rate constant. See text for more details.

obvious candidate for a description of the reaction. In principle, the reaction coordinate can be a complex combination of all atom coordinates. In the past, parameters such as native contacts ρ or RMSD have been proposed, but these turn out insufficient (71)

The problem of the reaction coordinate is depicted schematically in Figure 4. In this free energy contour plot all degrees of freedom are integrated out, except for the two true ingredients of the reaction coordinate *q* and *q*'. What remains are two stable states separated by a high free energy barrier. Because of the shape of the barrier, most trajectories crossing the barrier would follow on average a nonlinear path in the *q*-*q'*-plane. Clearly, to describe the barrier crossing both ingredients *q* and *q'* are required. However, *q'* does not sufficiently distinguish the initial from the final state, but *q* does If we would know nothing of the transition state, but we do know something of the initial and final state, then it would seem reasonable to call *q* the reaction coordinate. Using only *q* as the reaction coordinate the free energy maximum would lie at *q*,* as is depicted in the lower panel of Figure 4. While this free energy does show the expected two state features, a configuration belonging to *q** is already in the basin of attraction of U. On the other hand, the transition state in the top Figure belongs in the *q* space entirely to native state basin. This generic example illustrates that the reaction coordinate of complex processes, including protein folding, is generally difficult to define. Therefore in the past decade methods have been developed to overcome the folding barrier and sample the folding kinetics without the dependence of predefining the reaction coordinate. Transition path sampling is such a method (89).

5.4. Path sampling

5.4.1. Sampling the transition path ensemble

As mentioned above, reaction coordinates can be very elusive objects. A very long straightforward MD with many crossings would solve the reaction coordinate problem, but is computationally too expensive. However, this trajectory spends most of its time in the two stable states and very rarely leaves the stable state to cross the barriers. Hence, we know very much about the initial and final stable states. A potential solution to the rare event problem would hence be to focus only on the parts of the trajectories that are crossing the barriers. An infinite trajectory crosses the barrier an infinite number of times, thus forming an ensemble of crossing or *transition* paths. Because of the height of the barrier the time spent on crossing constitutes only a small fraction of total time. The *transition path ensemble* (TPE) is defined as the collection of dynamic, unbiased trajectories connecting the initial and final stable states (89,90,10). Obtaining this entire transition path ensemble would circumvent the problems related to defining a reaction coordinate. Transition path sampling was designed do just that by importance sampling. Based on the concept of sampling Markovian chains of states introduced by Pratt (91), TPS samples the TPE using a Monte Carlo scheme. Altering an existing pathway connecting initial and final state, and subsequently accepting or rejecting such a new trial way according to a proper acceptance rule results in a random walk through path space, and eventually leads to a representative TPE (90). TPS requires an initial trajectory to bootstrap the sampling. In the case of protein folding, high temperature unfolding pathways are easy to generate and act as valid initial pathways (71,95). Also, trajectories obtained from a biased sampling simulation might be used as input. In principle, the way one constructs such a path should not matter, as the importance sampling will relax the pathways to the equilibrium TPE.

5.4.2. The shooting algorithm

While any proper MC algorithm conserving detailed balance could sample the path ensemble, the shooting move turns out to be both simple and very efficient (10,90). This MC move starts by changing the momenta of a randomly chosen snapshot (called time slice) along an existing trajectory that connects the initial to final state. From this time slice the algorithm shoots a new trajectory 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 retained. 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 the properly weighted pathways representative for the TPE. Note that the TPS only *samples* the trajectories, and the resulting TPE should be analyzed for insight in the mechanism. 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.

5.4.3. Stochastic shooting algorithm

The shooting algorithm works because around the shooting point the new trial path stays close to the previous path, while at the same time the basins of attraction of the stable states make the trajectory commit to either state. Nevertheless, for long paths on rough energy landscape the TPS shooting algorithm can become less efficient because the old and the new trajectories will diverge completely before they can commit to the right stable basins. In that case, a small change in momenta will cause the forward as well as the backward trajectory to return to the same stable region, leading to a very low acceptance of trajectories. A stochastic implementation of TPS, which allows acceptance of a forward or backward shot independent of each other, enhances the acceptance dramatically (92). An Andersen-like thermostat ensures the stochastic nature of the trajectories by coupling the system to a heat bath (93). A small coupling constant related to a low enough frequency of re-initialization of the momenta of a randomly chosen molecule guarantees that the dynamics is equivalent to constant energy deterministic trajectories, while the paths can diverge fast enough for the stochastic path-sampling algorithm to work. The dynamics of the system is even less disturbed by only altering the momenta of the solvent molecules. The simplest implementation is then to only impose a new center of mass linear momentum, keeping the molecular angular momentum intact.

Strictly speaking we do not need to keep the path length fixed (94). In fact, only the part of the pathways that leave the initial stable state, cross the barrier and enter the final state are of importance. In addition, only shots from this barrier part have a reasonable chance of creating a new pathway. Hence, it is more efficient to stop integrating the equations of motion when a stable state has been reached. The assumption one makes is that the path will have a very low chance of recrossing after it has reached such as state. This assumption requires a slightly stricter definition of the stable states than needed in fixed path length TPS (94,95). On the other hand, the shifting moves, introduced in the original implementation of TPS are not necessary. The fluctuating path length requires a slightly adjusted Metropolis acceptance rule (94)

$$P_{acc}[o \to n] = h^n_{AB} \min\left[1, \frac{L^{(\mathrm{o})}}{L^{(\mathrm{n})}}\right] ;, (8)$$

where L is the path length and h_{AB} equals unity if the path connects the initial and final state, and zero otherwise. The subscript o and n refer to the old and new path respectively.

5.4.4. Rate constants

Within the TPS framework one can compute rate constants accurately. This involves a procedure akin to umbrella sampling computing the reversible work to constrain the path ensemble from completely free to the path ensemble of interest. The advantage of the TPS approach with respect to earlier techniques such as Bennett Chandler (87,88) or other TST based methods, is that TPS is much less sensitive to the choice of order parameter. The BC method is dependent on a correct choice of reaction coordinate, in order to yield a statistical meaningful result. As mentioned before, a poor choice will lead to a lower free energy barrier and hysteresis. The computation of the transmission coefficient is then impossible (90,94). The TPS rate calculation yields accurate results but is computationally demanding and, hence, a more efficient approach would be useful. The Transition interface sampling (TIS) method (94) is such an approach. TIS relies, just as the BC algorithm, on a factorization of the rate constant in a kinetic factor that measures the flux of leaving the initial state, and a crossing probability that measure the conditional probability that a trajectory reaches the final state, provided that it came directly from the initial state:

$$k_{AB} = f_A P_A(\lambda_B|\lambda_1). \quad (9)$$

The flux factor f_A is rather easy to calculate from a straightforward MD simulation. A trajectory will have a reasonable chance to leave the state, leading to an accurate estimate of the flux. The crossing probability $P_A(\lambda_B \mid \lambda_1)$ is very low, and is hence more difficult to estimate. The TIS method proceeds by dividing the space by *m* interfaces. For each interface *i* , a path sampling scheme is used to estimate the probability to reach the next interface *i+1*, under the condition that all trajectories cross the interface *i* and come directly from the initial state.

For the description of the interfaces a one-dimensional order parameter λ is used, which should be determined by the instantaneous configuration only. This order parameter is preferably identical to the reaction coordinate, but it is much less sensitive to deviation of the exact order parameter than the BC algorithm. Employing path sampling, the TIS algorithm computes for each interface the probability $P_A(\lambda_{i+1}| \lambda_{i+1})$ to reach λ_{i+1} while having crossed λ_i and coming from initial state A. The last interface to be crossed is that of final state B. The complete crossing probability $P_A(\lambda_B \mid \lambda_A)$ constant is than the product of all interface probabilities. Final multiplication with the flux yields the rate constant. In practice, all paths are binned in histograms as a function of λ for each interface path sampling. Subsequent matching of these histograms leads to the final crossing probability. For a fully detailed description of TIS I refer to Ref. (94).

Other techniques to sample paths connecting an initial and a finals state exist. For instance, the path action algorithms by Elber *et al.* employ the stochastic Onsager-Machlup action for the study of protein folding (96,97,98), These methods use a Monte Carlo scheme to sample the path ensemble, where the weight of each path is given by its action. While elegant and efficient, the large time-steps involved in the action prohibit a direct correspondence between the transition probabilities and the underlying dynamics. Hence, the action methods cannot quantitatively address the computation of the rate constant.

5.5. Coarse-graining kinetics

5.5.1. Partial paths and Milestoning

As mentioned above path sampling can be used to elucidate the folding process of small proteins and give an estimate of the rate constants (95). For large proteins the required path length might make a path sampling simply too expensive, not only due to system size, but also because large proteins are more likely to have multiple intermediate states. Metastable states that are stable on the nanosecond scale are already detrimental for an efficient implementation. However, the diffusive nature of the paths and the corresponding loss of correlation can be made into an advantage in some more recently developed path sampling methods. The partial path TIS technique (PPTIS) describes the rare event as a Markovian hopping process between the interfaces, with corresponding local hopping probabilities. Path sampling is then employed as a means to obtain these hopping probabilities. When the loss of correlation between three consecutive interfaces is justified, this method reproduces the kinetics in an accurate way (99).

The Milestoning (100) method by Elber and coworkers assumes that the protein diffuses through a localized 'tube' in configuration space. This assumption allows the translation of the rare folding process into a non-Markovian hopping between configuration space hyperplanes, the so-called 'milestones'. (Note the similarity between the PPTIS interfaces and the milestones, also stipulated in Ref. (100). The main difference is that they do not form a foliation, i.e. they can intersect. This is the reason why the method has to assume a tube-like process). An additional assumption is that configurations on each milestone are Boltzmann distributed. The Milestoning procedure itself consists of starting trajectories from an equilibrium ensemble on a milestone. The distribution of times to reach the next milestone in combination with a Kolgomorov equation then yields an accurate kinetic picture of the process of interest.

5.5.2. Markovian state models and stochastic road maps

The idea of correlation loss between intermediate states is the central concept in the Markovian state models (sometimes called stochastic road map (100,101,102,103) or equilibrium kinetic network (104). In these descriptions the folding pathways are represented by a chain of hopping events between a set of metastable states. A transition matrix gives the probabilities for hopping between these states. The other ingredient is the description of the states, which can be obtained by clustering of configurations or decomposition of state space. From an MSM one can compute overall rate constant, and committor distribution (see section 5.6.1). In a sense, the MSM is equivalent to a Kinetic Monte Carlo (105). Constructing the state space decomposition and the probability matrix are computational intensive (106). One approach to obtain the required ingredients for the MSM is to use clustering of conformations and measuring the transition probabilities from a TPS run (101). Krivov and Karplus used a network approach to study the beta-hairpin of the next section in an implicit solvent (106). Schütte *et al* have done much work in developing methods for the decomposition of phase space (107,108).

5.5.3. Mapping the kinetics on reaction coordinates

Notwithstanding the problems with free energy landscapes as a function of reaction coordinates, it is in principle possible to compute such a free energy landscape, and approximate the kinetics of a protein in solution by a Langevin description. Such a description can give correct results, but only if all the proper reaction coordinates have been identified. In addition, the effective diffusion along this coordinate is needed. When both the free energy and the diffusion are known, a mapping on an effective is possible (see e.g. Ref. (109)).

Recently it was shown that the free energy can be constructed from drift in the reaction coordinate, from straightforward MD simulation. Garcia *et al* applied that approach to replica exchange MD of a beta hairpin and obtained reasonable results for both the free energy and the kinetics (110).

5.6. Analyzing the path ensemble

5.6.1. Committors and the Transition State Ensemble

Sampling the transition path ensemble or the Markovian state networks is an important first step, but true insight in protein folding can only come from analyzing the myriad of pathways that constitutes the folding process. A protein can choose from many different paths via many different low populated intermediates or transition states, all leading eventually to the native state. A description of these intermediates and transition states would yield insight in the reaction mechanisms. Intermediates are characterized by long dwelling times on the molecular times scale, and can be analyzed for instance using the decomposition methods of Schütte *et al.* (107). Transition states are more difficult to define. As the path ensemble comprises many folding pathways, there is not just one transition state, but instead a whole ensemble of states: the transition state ensemble (TSE). But the question remains, what is the definition of the configurations comprising the TSE? In physical chemistry usually the transition state is defined as a saddle point in the potential energy landscape. At first sight it might seem useful to count all saddle points on the potential energy surface, but this makes only sense when the thermal energy is much smaller than most barriers in the energy surface. For solvated systems this is not the case. Besides, for complex systems the number of saddle points grows exponentially with the degrees of freedom. Counting saddles for solvated proteins would be unfeasible. Moreover, these saddles would give no insight in the reaction, because they are mostly not representative for folding. Another favorable option might seem to describe the transition state as a saddle point in the free energy landscape. A single saddle point in the free energy landscape corresponds then to many TS configurations, because all degrees of freedom except the reaction coordinates are integrated out. In addition, thermal fluctuations ensure that points close to the free energy saddle point also have a finite weight in the TSE ensemble. However, the free energy picture of the TSE depends entirely on the definition of the correct reaction coordinate. As discussed in section 5.3.2, this is just the problem we are trying to solve in the first place.

A definition of the transition state based on a configuration itself, independent of the free energy, which requires a priori knowledge of the reaction coordinate, would be helpful. Du *et al.* (111) proposed such a definition, making use of a general property of a transition state. For a configuration that is a transition state, a trajectory starting from that configuration with a random momentum has an equal probability to reach either the final state or the initial state. One can thus test an arbitrary configuration for this property by running many short trajectories from it, initiated with momenta from the Maxwell-Boltzmann distribution and measuring the probability to reach the final state p_B. This probability is, in general, called the commitment probability or 'committor' (90). For protein folding it is also known as p_{fold}. However, such a definition only marks the boundary between the basins of attraction of the initial and the final state: the separatrix. To obtain the TSE one would have to weight the configurations of the separatrix with the probability to occur. Suppose that we have a very long (in principle infinite) unbiased MD trajectory, visiting both stable states and crossing the barrier many times back and forth. By definition this trajectory has to pass through the TSE. In fact, the configurations along this trajectory that have a p_B=0.5 constitute the TSE. Such configurations lie precisely on the separatrix and are weighted correctly. It goes without saying that obtaining the TSE in this way is undo-able in practice. Computing the unbiased very long trajectory with many crossings was the very problem of rare events. Moreover, obtaining this trajectory is trivial compared to the effort one would need to perform in order to compute the committor values for each of the configurations along this trajectory. Nevertheless, the concept of the committor (or p_{fold}) can be used to extract the TSE from the transition path ensemble. As the TPE properly weights each pathway, the configurations along these paths obeying the pB=0.5 criterion belong to the TSE and, in fact constitute the TSE.

While Daggett found no commitment of CI2 within tens of nanoseconds (4), Pande has computed commitment probabilities for BBA5 in explicit water and found commitment times of less than 5 ns (82). Also for the Trp cage protein, commitment times were quite low (71). Rhee and Pande (112) compared the p_{fold} prediction for different simulation models and concluded that chemical detail is important for small protein folding kinetics, including the explicitness of the solvent. In particular, there is a large difference between the mechanism in the Go model and in the explicit solvent

The concept of the committor does yield the TSE, but does not give much insight in the reaction, as it only follows from performing many trajectories. Still, it can be used to analyze the validity of reaction coordinates, through committor distributions (90).

5.6.2. Testing reaction coordinates with committor distributions

For complex reactions such as protein folding the reaction coordinate is not trivial. As discussed above, from the transition path ensemble one can obtain the TSE by computing the committor values along the pathways. The structures with a committor half are considered a transition state and are part of the TSE. However, one would like to extract the reaction coordinate as well, or at least be able to test a candidate reaction coordinate.

A good reaction coordinate parameterizes the committor, that is, the reaction coordinate based on an instantaneous structure can predict what the commitment probability of that structure is. In principle, the committor itself is the perfect reaction coordinate by definition is, but does not contain information, i.e. it does not predict the committor value of other structures. A good reaction coordinate should be low dimensional and able to predict the committor. Moreover, structures with a particular reaction coordinate should have the same committor value. This enables a test for the quality of a proposed reaction coordinate. Suppose that a biased simulation (e.g. using umbrella sampling) samples a properly weighed equilibrium ensemble of configurations r confined to a particular value q of the proposed reaction coordinate function $q(r)$. For each member of this ensemble one can compute the committor p_B (where B denotes the final state, for instance, the native state). The probability distribution of the committor $P(p_B)$ reveals the ability of the reaction coordinate to represent the committor (113,114). If the proposed reaction coordinate is sufficient to predict the value of the committor, the committor distribution $P(p_B)$ will be peaked around committor value corresponding to the fixed reaction coordinate q. 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 the reaction coordinate can have different committors. In particular the committor distribution for free energy saddle points q^* is revealing. If $q(r)$ is a good reaction coordinate, all the configurations that have $q(r)=q^*$ are transition states and the committor distribution $P(p_B)$ is peaked around $pB=1/2$. When a committor distribution $P(p_B)$ turns out not unimodal at $p_B=1/2$, a correct characterization of the transition state ensemble requires degrees of freedom, additional to the proposed reaction coordinate (90). Initially introduced to study ionic dissociation in water (113), committor distribution analysis has subsequently applied to elucidate the mechanism of various complex biologically relevant reactions (114,115,116).

Pande *et al*. investigate the folding of the villin headpiece with MD and MSM (117). They also probed the role of water in the folding of villin using commitment probabilities. By keeping the structure fixed and the re-annealing the water molecules, they basically performed a committor analysis (with q the structure itself). In this way the authors could test the influence of the water molecules on the kinetics of the folding reaction. They conclude that the water structure does not the kinetics to a very large extent. Juraszek *et al.* come to the same conclusion in case of Trp-cage protein (71).

5.6.3. Genetic neural networks

A disadvantage of the committor distribution approach is that when a prospect reaction coordinate does not show the desired unimodal pB histogram, a new reaction coordinate should be tested by performing the computational expensive committor analysis again. To avoid the expensive recalculation of committor distributions, Ma and Dinner employed a genetic neural network (GNN) (118,119) that automatically selects the collective variable that best parameterizes the committor form among many possible reaction coordinate (115). The first step of this methods is the building of a database which contains for many different structures, the committor pB(x), as well as a list of all collective variables q(x) that are candidate reaction coordinates. As the method works best if all values of the committor are represented, taking the structure from a straightforward MD run is highly inefficient. A more practical way is to take the structures in the database from a transition path ensemble. Part of the database is then used as training set for the neural network, which optimizes its weights for certain combinations of collective variables. The quality of the network is then tested from the deviations between the predicted committor values and the previously computed committor, for the other part of the database. In principle, this optimization should be done for all combinations of reaction coordinate candidates, in order to find the one with smallest committor prediction error. As this is too expensive, Ma and Dinner employ a genetic algorithm to search for the best combination of collective variables. The GNN-method has been used to investigate the nature of the reaction coordinate for the isomerization of alanine dipeptide in vacuum and explicit solvent (115), but has not been applied yet to protein folding.

5.6.4. Bayesian path statistics

Hummer (120,121) proposed an alternative definition of transition states based on a Bayesian relation between the transition path ensemble and the equilibrium (canonical) ensemble. The transition path ensemble constrains pathways to connect A and B. Therefore the probability density of microscopic states P(x|TP) for these selected transition pathways (TP) differ from the equilibrium distribution ρ(x). The Bayesian expression for the probability P(TP|x), that a trajectory visiting phase space point x={r,p},with r the configuration, and p the momenta, is in fact on a transition path, reads

$$P(\mathrm{TP}|x) = \frac{P(x|\mathrm{TP})P(\mathrm{TP})}{\rho(x)}. \qquad (10)$$

Here, the factor P(TP) measures the overall probability to be on a transition states, and is equal to the fraction of time that a infinitely long trajectory spent on crossing the barrier from A to B. According to this expression the conditional probability P(TP|x) is maximized for structures with a low equilibrium weight, but with a high chance to occur on transition paths. This is true for transitions state, and thus this probability can be used to identify the TSE. For diffusive dynamics this probability becomes

$$P(\mathrm{TP}|r) = 2p_B(r)[1 - p_B(r)], \qquad (11)$$

which clearly has a maximum for p_B=0.5, which is indeed the definition of the TSE. In the following we assume that the dynamics is governed by diffusion, and equation 11 applies.

A practical application of the Bayesian relation involves the projection of reaction coordinate q. The probability P(TP|x) can be averaged over the constrained equilibrium ensemble q(x)=q, to yield P(TP|q), the probability that a configuration with a reaction coordinate q, is indeed on a transition path. Analogous to equation 10 this probability relates the probability density of q in the transition path ensemble P(q|TP) with the equilibrium ensemble P(q) according to

$$P(\mathrm{TP}|q) = \frac{P(q|\mathrm{TP})P(\mathrm{TP})}{p(q)}. , \qquad (12)$$

Here, *P(q) ∝ exp(-β F(q))* and *P(q|TP)* follow from equilibrium free energy (e.g. umbrella sampling) simulation and a path sampling simulation, respectively, *P(TP)* is again a normalizing factor.

Analogous to the committor distribution analysis, the application of eq. 12 for a good reaction coordinate reveals a peak in the probability *P(TP|q)* at the transition state value of *q*. This is because for all transition states, i.e. the states *x* with large probabilities *P(TP|x),* should correspond to approximately the same value of the reaction coordinate q(x). A poor choice of reaction coordinate q(x) will have no signification correlation with P(TP| x), and lead to a flat P(TP|q) . Best and Hummer used this approach to test reaction coordinates for the folding of a simple three-helix bundle protein using a Go like model (121) .

5.6.5. Likelihood Maximization

The likelihood maximization (LM) method of Peters and Trout (122) has the same aim as the GNN approach discussed above. The method screens many possible collective variables for fitness as a reaction coordinate. Where the GNN relies on the computation of committor values for a large set of TPS structures, the LM method uses the sampling data of the TPS itself only. Therefore it is in principle computationally more efficient than the GNN. The LM method is based on the notion that the shooting algorithm is itself based on a type of commitment probability. A forward shot from a particular structure leading to the correct final state reveals a tendency for a larger committor value than if the shot lead to the initial state. Hence, it should be (and in fact is) possible to use the acceptance data from a TPS run to find a description of the committor that best serves as a reaction coordinate.

The starting point of the method is the probability P(TP |r) that a certain structure *r* in configuration space leads to a transition path using the shooting algorithm, already introduced in the previous section. An unbiased estimate of that probability requires, in general a randomized momentum at the shooting point. Peters and Trout introduce therefore the aimless shooting algorithm that reinitializes the momenta (122). In addition, they bias the shooting point to the transition state region to enhance the acceptance ratio, but this is not strictly necessary. Each shot is a realization of a process that estimates P(TP|r), and which is strongly connected to a committor calculation. If the dynamics is rather diffusive, such as is the case with protein folding, the expression for the P(TP|r) is given by eq. 11. Any p_B(q) function that has a rough s-shaped curve, smoothly varying from 0 to 1, would give a function peaking at the transition state value of q and decaying to zero away from this peak, as is required for a good reaction coordinate (120). Peters and Trout choose the following dependence:

$$p_B(q) = (1 + \tanh[q])/2, \quad (13)$$

which leads to the probability

$$P(TP|q) = p_0(1 - \tanh^2[q]), \quad (14)$$

where p_0 is a constant prefactor. Note that the dependence of q(r) on the structure r is implicit. Just as in the Bayesian approach of Hummer *et al.* the goal is to find a reaction coordinate model q(r) that gives the best prediction of P(TP|q), the probability to be on a transition path for a certain value of q.

Next, Peter and Trout define a general simple dependence of the reaction coordinate q on a set of M collective variables $q_1, q_2, \ldots, q_M$

$$q = \alpha_0 + \sum_{k=1}^{M} \alpha_k q_k \quad , (15)$$

where α_k are the models' fitting parameters, to be optimized by the LM. The likelihood function L(α) gives the probability to observed the measured data, as a function of the model parameters α:

$$L(\alpha) = \prod_{r\in\text{acc}} P(\text{TP}|q(r)) \prod_{r\in\text{rej}} [1 - P(\text{TP}|q(r))]\,. \quad , \quad (16)$$

where the products run over, respectively, the accepted and rejected shooting points obtained by TPS. Maximizing (the logarithm of) the function L(α) with respect to the parameters α results in the best reaction coordinate given the model eq. 15, that describes the observed data,

In practice, the LM method is first tried on a large set of single order parameters. The largest likelihood gives then 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, given by the Bayesian criterion *ΔL= ln M*, then the new combination of parameters is adopted as the best reaction coordinate model. This procedure is repeated until no further significant improvement is found. Due to combinatorial explosion combination of 3-4 parameters is typically the maximum.

In a later paper, Peters and Trout note that knowledge of the fate of only half trajectories, that is, either forward (or backward) shots only, is enough to perform the analysis (123). They applied the LM to structural solid-solid transitions of terephtalic acid (124). The LM is particularly useful for screening complex reactions such as protein folding.

Note that the LM is entirely different from the often-used principal component analysis or essential dynamics (see e.g. Ref.(125)). These methods decompose movement of the protein into a high dimensional vector representing this motion, but do not reveal the low dimensional relevant reaction coordinate

6. APPLICATION OF PATH SAMPLING TECHNIQUES ON PROTEIN FOLDING

6.1. The GB1 beta-hairpin

6.1.1. Introduction

Advanced path sampling methods addressing the kinetics of protein folding have up to now only been applied to small proteins or even protein fragments. One of the earliest applications was the 16 residue C-terminal fragment (41-56) of protein G-B1 (sequence GEWTYDDATKTFTVTE). This short sequence forms a stable non-aggregating beta-hairpin in solution, and has become an experimental an theoretical model system for investigating beta sheet secondary structure formation. Eaton and coworkers (126,127) showed that the hairpin obeys a two state kinetics, with a reaction time of 6 microseconds. Many simulation studies followed using simplified models (128,129,130) full atom models in implicit solvent (131,132,104) or in explicit solvent (85,125,133,98,134,135,136).

Straightforward high temperature MD simulation in explicit solvent (85) and multi-canonical Monte Carlo sampling in implicit solvent (131) highlight the discrete nature of the folding process. The initial beta turn formation is followed by a collapse of the hydrophobic residues in the peptide into a hydrophobic core (H-state) and eventually formation of the backbone hydrogen bonds towards F and N states (see Figure 5). Garcia *et al.* (134) and Zhou *et al.* (135), determined the beta hairpin free energy landscape in explicit solvent. Bussi *et al*, use a combination of REMD and Metadynamics to improve sampling, and obtain a converged free energy landscape (64). Yang *et al.* extracted folding kinetics from a REMD simulation (110).

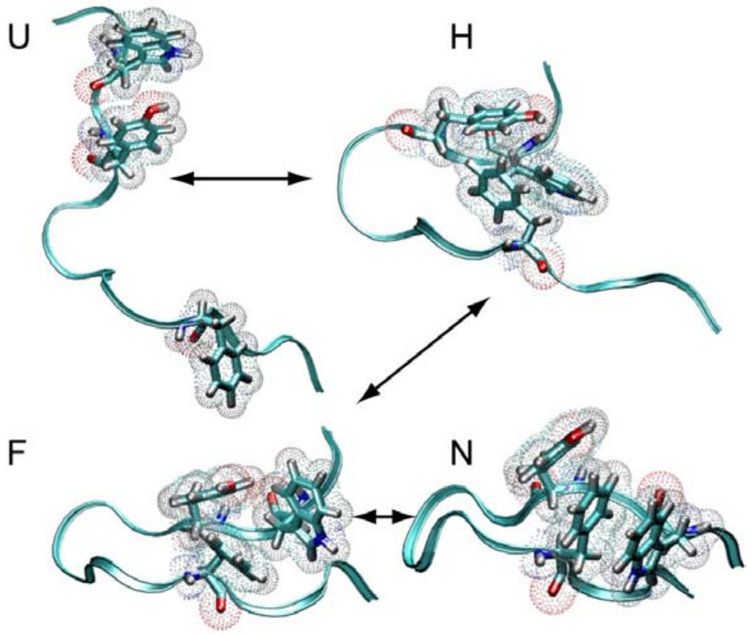

Figure 5. Top: Structures of the several (meta) stable states. The backbone is represented by a ribbon, the hydrophobic core in stick model with dots to indicate the size of the atoms. All other residues and solvent molecules are left out. (Figures are made with VMD (51)).

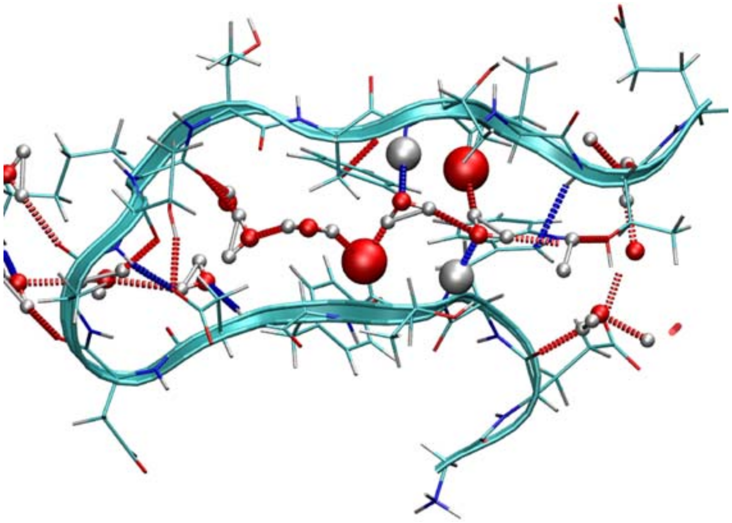

Figure 6. A typical F-H transition state configuration. A strip of water molecules separates the two strands. In particular, the water molecules bridge important backbone hydrogen bonds 3-4 (indicated by large spheres on the backbone oxygen and hydrogen atoms). (Figure is made with VMD (51))

6.1.2. TPS of GB1

Molecular dynamics runs at ambient conditions started from the native state indicate that there is a substantial barrier to unfolding (85,125,134,135). To be more precise, the H-F transition turns out to be the rate limiting step, or the highest barrier in the folding process of a two-state folder. Employing TPS techniques to the GB1-beta hairpin's H-F transition in explicit solvent at room temperature I could study the kinetic pathways and obtain the folding rates (92,95).

Computing the commitment probabilities p_A (folding) and p_B (unfolding) for representative pathways from the TPE (95), yielded true transition state ensembles for the rate-limiting step. Figure 6 shows that the transition state has a native like hairpin shape, with no native backbone hydrogen bonds formed, and a strip of water molecules bridging the two strands. Water acts as a lubricant to folding (57,137,138,139) bringing together the native backbone hydrogen bonds. Moreover, expulsion of water is the last step of folding, before collapsing to the

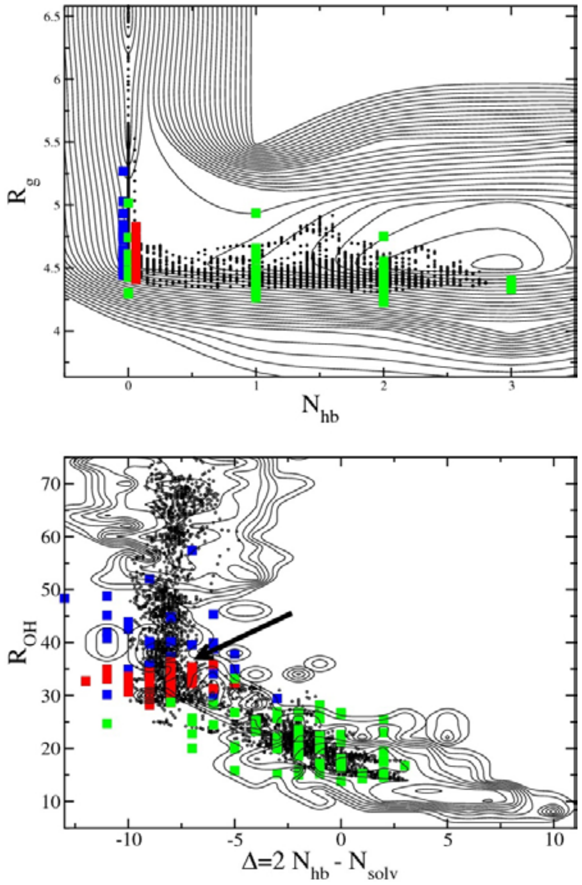

Figure 7. Representations of the folding event in 2 different order parameter planes. The free energy landscape from replica exchange is given by thin solid contour lines separated by 0.2 k_BT. A few smoothed paths in the F-H ensemble are denoted by a scatter plot (small dots). Each dot represents a time slice along a path. Also given are the different committor ensembles: pB<0.2 light gray, 0.4< pB<0.6 in dark gray and pB>0.9 in black. Arrows indicated the apparent transition state saddle points in the FE landscape.

native state. The major cause for the barrier to folding is thus the formation of this unfavorable transition state, which has a low conformational entropy of the protein as well as a low translational entropy of the bound solvent molecules (22,23).

Previous MD and REMD simulation focused on the number of backbone hydrogen bonds N_{hb} and the hydrophobic core radius of gyration R_g as governing reaction coordinates for the folding of GB1 (135). Analysis of the TSE obtained from the transition path sampling runs, shows the transition states plotted in these variables falls entirely within the basin of attraction of the unfolded H state (see Figure 7). This analysis shows that the TSE does not always correspond to saddle points in the free energy landscape, and it depends very much on the choice of reaction coordinate (90).

Inspection of the TSE structures suggests that the backbone solvation actually determines the F-H transition. A possible measure for solvation is the difference between native backbone hydrogen bonds N_{hb} and water molecules bound to water. We define Δ = 2 N_{hb} –N_{solv} where N_{solv} denotes the number of backbone-solvent hydrogen bonds. Another good order parameter describing the F-H transition is the inter-strand proximity, captured in the sum of distances between oxygen and hydrogen of the native backbone hydrogen bonds R_{OH}. Indeed, the free energy saddle point in the Δ -R_{OH} plane corresponds to the TSE p_B=0.5 ensemble, making a case for these as reasonable order parameters for the F-H transition.

From the TPS analysis the following general kinetic mechanism for the H-F transition of the GB1 hairpin arose. In the H state the hairpin has to search for the transition state, characterized by water molecules bridging the middle backbone hydrogen bonds (labeled 3-4). These hydrogen bonds are formed first as they are driven together by the nearby hydrophobic core. The vicinity of the core might also cause the expulsion of bridging waters.

6.1.3. Folding rate calculation

Using λ=R_{OH} the folding and unfolding rates as computed by TIS simulation were, respectively, $k_{H\text{-}F}$ = 0.20 μ$s^{-1,}$ and $k_{F\text{-}H}$ = 0.4 μs^{-1}, both in reasonable agreement with experiment. The TIS results suggested a barrier of more than 10 k_BT between the F and H states, whereas the free energy landscapes of the REMD revealed only a barrier of approximately 3-4 k_BT. The lower value in the REMD is caused by an overlap of the F and H state in the order parameters used, thus lowering the barrier. The low barrier also shows that TST based methods (87,88) must suffer from recrossings and give vanishing transmission coefficients to correct for the low free energy barrier. Indeed, a one dimensional overdamped Langevin description with a fixed diffusion constant on one-dimensional free energy landscape as function of R_{OH} yielded a rate constant that was orders of magnitude too high, indicating that the true free energy barrier must have been much higher than 3-4 k_BT (95).

This projection on a one dimensional Langevin equation (Fokker-Planck equation) was also done by Garcia and coworkers for the GB1 (see section 5.5.3 and Ref. (110). They determined the reaction coordinate (which was similar to R_{OH}) dependent diffusion and drift constant from straightforward MD and used these in a coarse-grained simulation. The kinetics they obtained was in reasonable agreement with the experimental rates, showing that recrossings result in a much slower diffusion at the barrier.

6.2. Trp-cage

6.2.1. Introduction

Following the success of small fragments like the GB1 hairpin, many small and fast folding proteins and protein fragments were discovered. Examples of these are the Villin headpiece, the BBA5, and the Trp-cage mini protein. These fast folding proteins have contributed much to the understanding of generic folding mechanisms

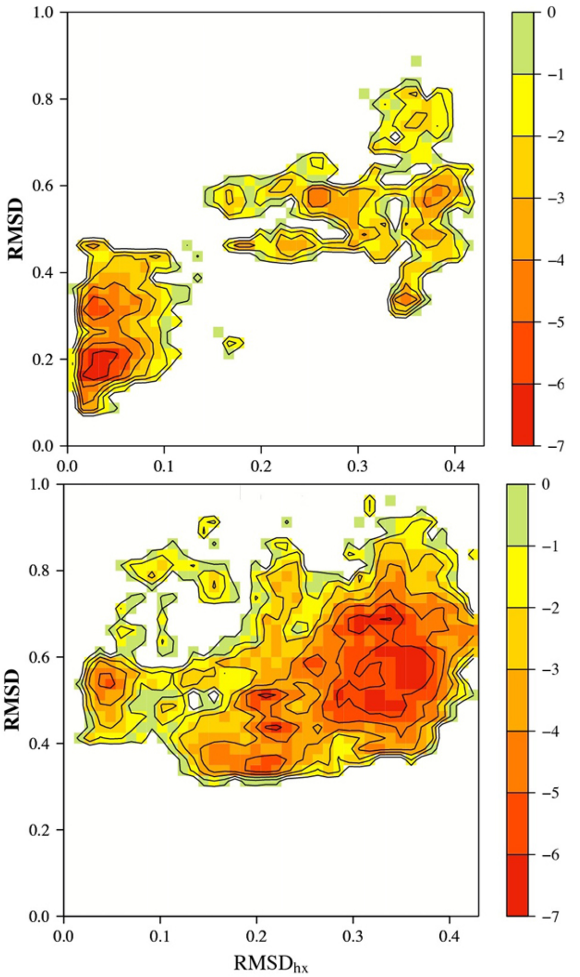

Figure 8. REMD free energy contour maps of Trp-cage in explicit water at 300K in the RMSD$_{hx}$ versus RMSD plane. The contours are separated by 1 k_BT. Left: results starting from the folded native state. Right: starting from the unfolded state. Note that the two free energy landscapes are not converged. Also note that for the unfolded REMD the transition to the native state is not observed at all.

because they bridge the gap between experiments and computer simulation. The designed 20-residue mini-protein Trp-cage (NLYIQ WLKDG GPSSG RPPPS) (14) is among the fastest folders with a room temperature relaxation time of 12 µs. The native state of Trp cage has an alpha-helix, a salt-bridge and a polyproline II helix shielding the central tryptophan from the surrounding solvent. Laser temperature-jump spectroscopy (141) showed that Trp cage is a two state folding, whereas fluorescent correlation spectroscopy (142) revealed an intermediate state with tryptophan solvent exposed. The folding mechanism, however, remained unclear. UV-resonance Raman spectroscopy measurements suggested early formation of the helix in the folding transition (143) . Molecular dynamics simulations investigated both thermodynamic stability of the protein and its possible folding pathways using all-atom models with implicit solvent (55,144,145,146,147) explicit solvent (148,65) or simplified models such as Go-models~ (149). Explicit solvent REMD studies simulations in (148) confirmed Trp-cage as a fast two state folder, with an intermediate state that contains two hydrophobic cores. A recent study by Piana and Laio, using a novel bias-exchange, reconfirms this finding(65). Recently, Paschek *et al.*(152) observe complete folding of Trp-cage in explicit solvent using REMD with the Amber force field. Earlier, all-atom implicit solvent MD (144,55,147) and a coarse-grained simulation (150) exhibited complete folding of the protein. Nevertheless, several observed misfolded states in the implicit solvent computations indicated a less reliable and efficient folding Trp-cage (147). The few studies that investigated the dynamics of the process either employed an implicit force field (FF) (55) or a simplified model (149). Because the solvent does play a role in protein folding (151) Juraszek *et al.* performed TPS of the folding of Trp-cage with an explicit (71). The initial and final states can be found by conducting a REMD simulation first. These simulations were done with the OPLSAA force field (154) using Gromacs (153). In the following sections we briefly describe the findings of that work (71).

6.2.2. Order parameters

Several order parameters turned out to be important: the protein radius of gyration using the alpha-carbons (R_g), the R_g including the side-chain atoms (R_{gsc}), the fraction of native contacts (ρ), the root mean square deviation from the native alpha-carbons structure (RMSD), the root mean square deviation of the alpha-helical residues (2-8) from an ideal helix (RMSD$_{hx}$), the solvent accessible surface (SASA) of the whole protein (153), the salt-bridge distance (sb) defined as the minimum distance between donors and acceptors in the hydrogen bond between Arg-16 and Asp-9 and the number of water molecules within 4 Å around Trp-6 (nw_W). Construction of free energy diagrams from REMD using these order parameters allows for the determination of the stable state definitions required for TPS.

6.2.3. Replica exchange MD

Conducting two independent 64-replica REMD simulations (148), one starting in the native and one starting in the unfolded state, can assess the convergence properties of the REMD. Figure 8 shows the free energy landscape for both REMD simulations plotted in the (Rg,ρ) plane. Both REMD runs did not converge in 36 ns per replica. Moreover, in the unfolded REMD the native state was not even reached. Cluster analysis revealed several different metastable states, shown in Figure 9. The largest cluster contains twisted hairpin like structures and bent loop structures, all having Trp-6 fully exposed to the solvent. The second largest cluster consisted of U-shape structures, with the tryptophan oriented correctly and packed in the center of the protein and one turn of the helix formed. The third cluster contained fully helical structures with the polyproline detached from the rest of the hydrophobic part, while the rest of the clusters could not be classified and were deemed molten globular structures. The three different groups correspond to FE minima in the RMSD$_{hx}$ vs RMSD plane of the unfolded REMD run (Figure 8). These groups immediately suggest two routes

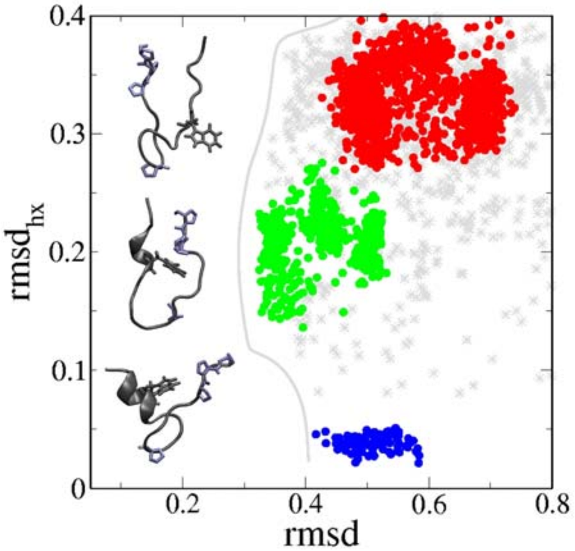

Figure 9. Three types of clusters found in the unfolded initiated REMD ensemble. The top cluster (red, 45%) is a set of several loop structures, with Trp-6 fully exposed to the solvent. The middle cluster (green, 15\) consists of the loop structures with Trp-6 positioned in between prolines. The bottom cluster (blue,3%) denotes fully helical structures. Typical structures belong to each group are shown on the left. Structures plotted as gray stars belong to clusters that did not reach the minimum abundance of 2%.

for the folding process: 1) initial loop formation followed by packing of the tryptophan between the proline residues.2) helix formation followed by correct packing of Trp-6. While both pathways seem possible, complete folding did not occur within 36 ns, because a substantial free energy barrier separates the native state from the intermediate. Still, the REMD simulations yield insight in intermediate structures and enabled a definition of the initial and final state for TPS.

Several simulation studies have observed spontaneous folding in MD (144) and in REMD (152). The difference with the results discussed here and those of Ref. (152) are almost certainly due to the force field. Piana and Laio performed a combination of replica exchange and Metadynamics (65), called bias exchange, to study the folding of the Trp-cage employing the OPLSAA force field. They find that the FE landscape starting from the extended state converges much better with the bias exchange, than by using PT-REMD. These authors conclude that the Trp-cage can fold via an intermediate in which the helix is fully formed, and the Trp-6 partially exposed. They do not find the L state, possibly due to the choice of biasing coordinates.

6.2.4. Transition path sampling

While in the TPS simulation of Ref (71) the folded state (A) was defined rather rigorously by RMSD, $\mathrm{RMSD}_{\mathrm{hx}}$, SASA, ρ and nw_{W}, the definition of B was less strict, only including parameters RMSD, ρ and nw_{W}. The $\mathrm{RMSD}_{\mathrm{hx}}$ was excluded because the order of unfolding was unknown and SASA was left out to allow molten globular structures with low SASA into the final state. The final region B thus includes both intermediate states found in the REMD simulation to avoid unnecessary long pathways in the TPS simulations.

The TPS simulation yielded several thousands room temperature pathways, broadly revealing two different main routes (Figure 10). Note that this path ensemble is as valid for the unfolding as for the folding process as the dynamical trajectories are microscopically time-reversible. Starting from the unfolded (intermediate) state each path in the ensemble reaches the native state within a few ns, but the folding barrier itself is of course substantial. During the folding process the protein can choose between a fast initial collapse to a loop state L and a helix formation. In the loop state the Trp will be still the solvent exposed. This hairpin resembling structure, also found in the REMD, is stabilized by tertiary contacts and hydrogen bonds. Incorporating the tryptophan in the protein center leads eventually to the native state. When a helix is formed as an initial step the protein will follow the I-N route. A small probability of misfolding exists to structures with a salt-bridge on the opposite side of the protein.

Unfolding is the reverse of folding. Starting from the native state, the trajectories in the path ensemble unfold by first partially expose of the tryptophan to the solvent. If the water penetrating the hydrophobic region between Trp-6 and Pro-12 leads to a thread of waters through the core, the protein will end up in the loop state L. This path preserves the compact form of the protein upon the desolvation. The other possibility is that hydrogen bonds between residues Gly10-Gly11 and Ser13-Ser14 impede solvation of this region. Subsequent solvation of Trp-6 and Pro-18 results in a detachment of the polyproline helix via the short-lived "Proline detached" (Pd) state. Thus, two hydrophobic clusters are formed separated by a layer of water molecules. Both the L and I state contain a salt bridge, which can break easily to form the unfolded state U. The N-L and N-I transitions thus contain indeed the rate limiting steps.

There are several reasons to think that these two global pathways are indeed the only possible folding routes for Trp cage. The first is that the L and I were also found as intermediates the REMD-unf simulations, but, in contrast to the TPS runs, the barrier towards the native state was not crossed within the 64 x 36 ns simulation time. The second reason is that several switches between the N-L to N-Pd-I pathways took place during the path sampling, mostly via an intermediate N-Pd-L pathway. The switching probability from N-Pd-I to N-L was about 4 times higher than the reverse switch, in agreement with the four times more likely L-N pathway in the ensemble. Such switches between qualitatively different pathways indicate that TPS is able to sample the path space adequately. Moreover, in the ensemble the least changed pathway diffuses over a large part of the RMSD-$\mathrm{RMSD}_{\mathrm{hx}}$ plane (71). Note that the TPS procedure does not impose the final L and I structures, but selects these from a loosely defined unfolded state B. The presence of two pathways can

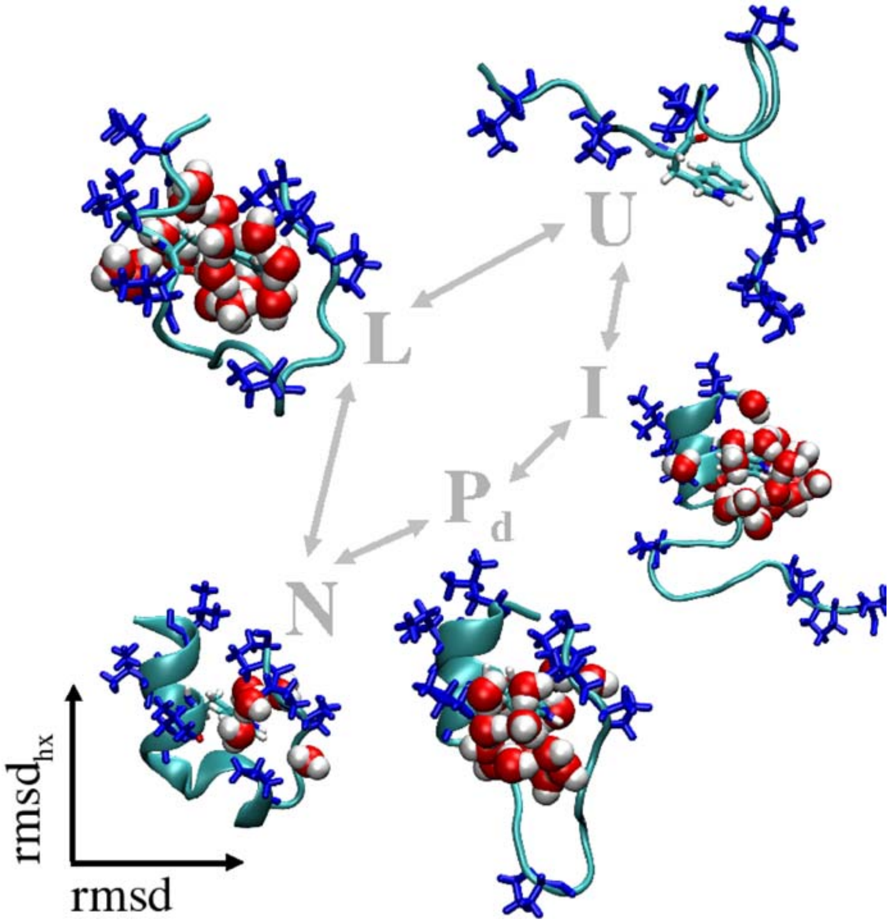

Figure 10. The two main folding routes in the path ensemble schematically depicted in the RMSD/RMSD$_{hx}$ plane. Protein backbone structures were rendered in cartoon style, hydrophobic residues in licorice and waters within 4 Å of the tryptophan in space-filling representations. At the first stage of folding (U) key hydrophobic contacts are formed and either Pro-12 or Pro-17/18/19 collapse on the helical residues. The first global route leads to the formation of two hydrophobic clusters. The alpha helix appears quickly in the larger cluster (I state) and eventually the smaller hydrophobic cluster approaches Trp-6 (Pd). Subsequently, the protein finds its native state (N) relatively easily. In the second scenario a loop structure (L) with correct tertiary contacts precedes the formation of the alpha helix. Both routes show water expulsion step in the last stage of folding (Structures made with VMD (51)).

explain the experimentally found existence of helical content (I state) in the early stages of folding (143) and the structurally restricted intermediate (142) (L-state).

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 alpha helix and the 3_{10}-helix. It is this double hydrogen-bonding that increases the residence time. For the N-Pd-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 alpha helix. The explicit modeling of solvent molecules can thus reveal crucial aspects of folding kinetics.

6.2.5. Transition states

The transition state (TS) ensemble follows from computing the committors along several trajectories of the path ensemble and consist of those time slices with the same commitment probability to unfold and to fold. Figure 11 shows a few of those configurations, that are in fact almost native like, but with a decreased number of native contacts. The N-L TS structure (Figure 11) has a largely dissolved helix and a entirely solvated tryptophan In the TS structures of the Pd-I pathway (Figure 11), the polyproline helix is perpendicular to the surface of the tryptophan aromatic ring, with 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 RMSDhx/RMSD plane in Figure 11 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/R_{gsc} representation, there

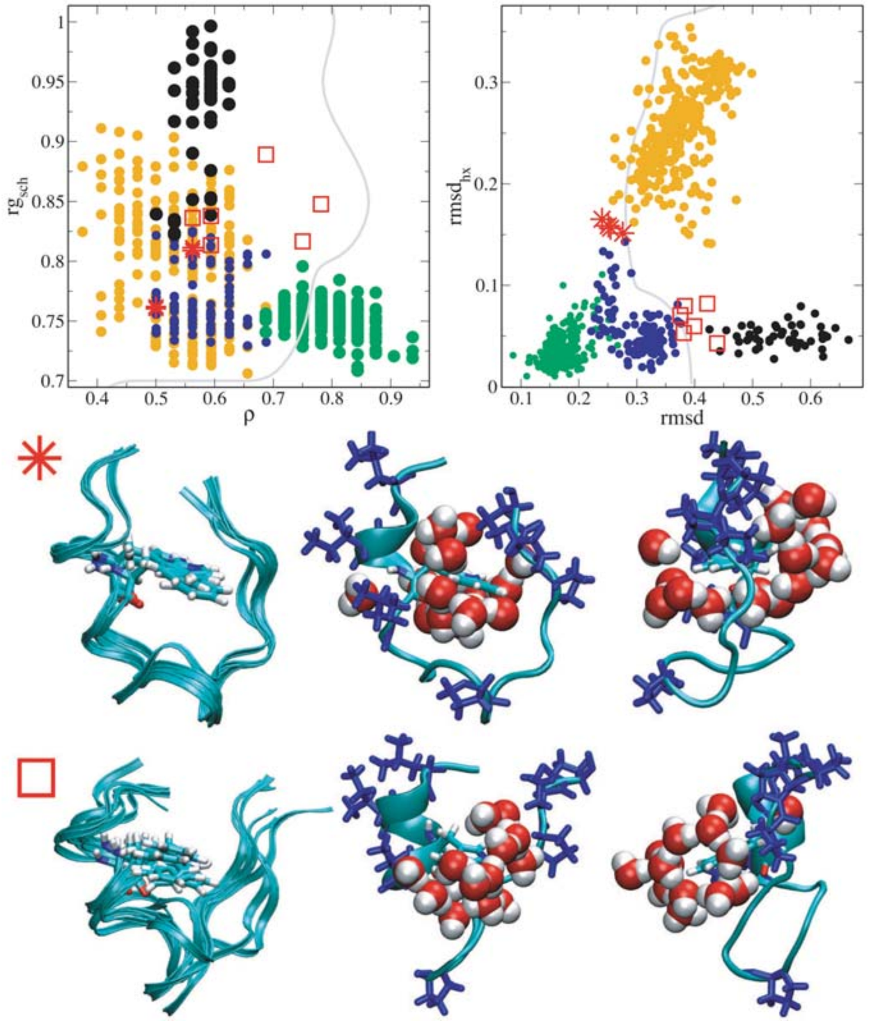

Figure 11. Stable states and transition states obtained by committor calculation, plotted in R_{gsc}- ρ (left top figure) and in RMSD-$RMSD_{hx}$ planes (right top figure). Transition states for the N-L and N-Pd transitions are shown as stars and squares respectively. Scatter points indicate the corresponding TPS trajectories. Native states (N) are plotted as green, loop structures (L) as orange, near-native structures with Pro-12 detached (Pd) as blue and the I-state as black points. The region to the left of the gray line on the $RMSD_{hx}$/RMSD (to the right on the ρ/R_{gsc}) plot is the part of the configuration space, which was not sampled, in the REMD-unf simulation. Lower structures: Left: superimposed TS structures for four different N-L (star) and six Pd-I paths (square). Middle and right: one of the TS structures for both routes and its side view plotted in cartoon representation. Hydrophobic residues are shown in licorice representation (blue), water molecules within 4 Å of Trp-6 in space-filling representation. (All structures made with VMD (51)).

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-unf simulation and suggests that the water expulsion transition is the rate-limiting step for both folding routes. The hypothesis of water expulsion or solvation as an important step was already put forward by Caflisch and Karplus in an early simulation study of barnase in explicit solvent (155), and in other computational studies (156, 139). Interestingly, folding simulations BBA5, a small protein stabilized by a hydrophobic core did not show this expulsion (151).

The water dynamics can have a profound contribution to the reaction coordinate, as was observed in ref (157) and later in Ref (158). To estimate the extent of this contribution, one can measure the effect of the water structure on the committor values of the transition states (112). After freezing the protein coordinates and randomizing the water structures, a new committor values 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 on 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 of the study by Ma and Dinner (115) that the solvent dynamics is coupled to alanine dipeptide isomerization. This contradiction might be explained by the fact that 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 in next to the protein. When the randomized solvent structure was only equilibrated for a 100 ps, the committor was significantly changed toward the unfolded state. This means that water molecules have not fully relaxed yet.

6.2.6. 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. These two different global routes are reminiscent of the two generic protein-folding mechanisms mentioned in section 2.4: the diffusion-collision mechanism (18) and the nucleation-condensation mechanism (17). Both mechanisms can be prevalent simultaneously in one protein (21,16). Further committor analysis revealed the water dynamics is not a part of the reaction coordinate. Nevertheless, the finding that some water molecules are strongly bound to the protein during the folding might lead to improved implicit solvent models.

7. PERSPECTIVE

The field of protein folding simulation has seen much progress in the past years, not in the least due to the discovery of fast folding small proteins that allowed a bridge between experiment and simulation. Many simulation studies have contributed to this progress, ranging from MC simulation of simple lattice models to full-blown all-atom brute-force MD simulation in explicit solvent. The latter direct simulation approach has profited from the development of increasingly accurate force fields, but still suffers from the time scale problem. The last few years have seen the rise of novel methodology that allows for more efficient sampling of phase space. Metadynamics has the potential to sample large barriers in many dimensions. Path sampling methods allow the prediction of mechanistic kinetic details that cannot be obtained otherwise. Nevertheless, for large proteins the path sampling computational effort due to both system size and long time-scales becomes prohibitive. Thus, there is still much room for improvement, in particular by using coarse-grained models based on atomistic detailed force fields, and by applying e.g. Markovian state models. Path sampling and related methods are only one part of the answer. The development of novel analysis techniques such as LM has enabled a relative cheap way to test reaction coordinate. I stress that there is not a single method that can solve the protein-folding problem. In contrast, one should use a combination of several complementary simulation methods in order to make progress. In the next few years we will almost certainly see the use of such combinations of methods to elucidate folding pathways, and predict final native structures. In this manner, simulation will be able to give better molecular understanding and interpretation of protein folding experiments.

8. ACKNOWLEDGEMENT

This work is part of the research programme of the 'Stichting voor Fundamenteel Onderzoek der Materie (FOM)', which is financially supported by the 'Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO)'.

9. REFERENCES

1. Alberts, B.; A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter: Molecular Biology of the Cell. 4th ed., Garland Publishing, NY (2002)

2. Anfinsen, C.B.: Principles that govern the folding of protein chains. *Science* 181: 223-230 (1973)

3. Dobson, C. M.: Principles of protein folding, misfolding and aggregation. *Sem. Cell Dev. Biol*. 15, 3-16 (2004)

4. Daggett, V.: Protein folding-simulation. *Chem. Rev.*, 106, 1898-1916 (2006)

5. Frenkel, D. & B. Smit: Understanding molecular simulation, 2nd ed., Academic Press, San Diego, CA (2002)

6. Laio, A. & M. Parrinello: Escaping free-energy minima. *Proc. Nat. Acad. Sci. USA*, 99, 12562-12566 (2002)

7.Earl, D. J. & M. W. Deem: Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.*, 7, 3910-3916 (2005)

8. Pande, V. S., 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: Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers*, 68, 91-109 (2003)

9. Snow, C. D., E. J. Sorin, Y. M. Rhee & V. S. Pande: How well can simulation predict protein folding kinetics and thermodynamics? *Annu. Rev. Biophys. Biomol. Struct.,* 34, 43-69 (2005)

10 Bolhuis, P. G., D. Chandler, C. Dellago & P. L. Geissler: Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.*, 53, 291-318 (2002)

11. Fersht, A.: Structure and Mechanism in Protein Science. Freeman, New York, (1999)

12. Grosberg, A. Statistical mechanics of protein folding, some outstanding problems. In: Computational Soft Matter: from Synthetic Polymers to Proteins, Vol.23 of NIC Series, edited by N. Attig, K.Binder, H. Grubmuller, and K. Kremer, Graphische Betriebe, Julich, (2004)

13. D. Wales, Energy Landscapes, Cambridge University Press, Cambridge, (2003)

14. Onuchic, J. N., Z. LutheySchulten & P. G. Wolynes: Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.*, 48, 545-600 (1997)

15. Pande, V. S., A. Y. Grosberg & T. Tanaka: Heteropolymer freezing and design: Towards physical models of protein folding. *Rev. Mod. Phys.*, 72, 259-314 (2000)

16. Dinner, A. R., A. Sali, L. J. Smith, C. M. Dobson & M. Karplus: Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.*, 25, 331-339 (2000)

17. Abkevich, V. I., A. M. Gutin & E. I. Shakhnovich: Specific Nucleus as the Transition-State for Protein-Folding - Evidence from the Lattice Model. *Biochemistry*, 33, 10026-10036 (1994)

18. Karplus, M. & D.L. Weaver: Protein folding dynamics. *Nature* 260: 404-406 (1976)

19. Mirny, L. & E. Shakhnovich: Protein folding theory: From lattice to all-atom models. *Annu. Rev. Biophys. Biomol. Struct.*, 30, 361-396 (2001)

20. Islam, S. A., M. Karplus & D. L. Weaver: Application of the diffusion-collision model to the folding of three-helix bundle proteins. *J. Mol. Biol.*, 318, 199-215 (2002)

21. Gianni, S., N. R. Guydosh, F. Khan, T. D. Caldas, U. Mayor, G. W. N. White, M. L. DeMarco, V. Daggett & A. R. Fersht: Unifying features in protein-folding mechanisms. *Proc. Nat. Acad. Sci. USA*, 100, 13286-13291 (2003)

22. Akmal, A. & V. Munoz: The nature of the free energy barriers to two-state folding. *Proteins Struct. Funct. Bioinf.,* 57, 142-152 (2004)

23. Harano, Y. & M. Kinoshita: Translational-entropy gain of solvent upon protein folding. *Biophys. J.*, 89, 2701-2710 (2005)

24. Kubelka, J., J. Hofrichter & W. A. Eaton: The protein folding 'speed limit'. *Curr. Opin. Struct. Biol.*, 14, 76-88 (2004)

25. Gillespie, B. & K. W. Plaxco: Using protein folding rates to test protein folding theories. *Annu. Rev. Biochem*, 73, 837-859 (2004)

26. Ghosh, K., S. B. Ozkan & K. A. Dill: The ultimate speed limit to protein folding is conformational searching. *J. Am. Chem. Soc.*, 129, 11920-11927 (2007)

27. Allen, M.P. & D.J. Tildesley: Computer Simulation of Liquids, Oxford University Press, Oxford (1987)

28. Bussi, G., D. Donadio & M. Parrinello: Canonical sampling through velocity rescaling. *J. Chem. Phys.*, 126, (2007)

29. Norberg, J. & L. Nilsson: Advances in biomolecular simulations: methodology and recent applications. *Q. Rev. Biophys.*, 36, 257-306 (2003)

30. Levitt M. Molecular dynamics of native protein. I. Computer simulation of trajectories. *J. Mol. Biol.*, 168,595-617 (1983)

31. Levitt, M., M. Hirshberg, R. Sharon & V. Daggett: Potential-Energy Function and Parameters for Simulations of the Molecular-Dynamics of Proteins and Nucleic-Acids in Solution. *Comput. Phys. Commun.*, 91, 215-231 (1995)

32. Cornell, W. D., P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell & P. A. Kollman: A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules. *J. Am. Chem. Soc.*, 117, 5179-5197 (1995)

33. MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin & M. Karplus: All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B*, 102, 3586-3616 (1998)

34. van Gunsteren, W.F. & H. Berendsen, Gromos-87 manual, Biomos BVV, Groningen, The Netherlands (1987)

35. Jorgensen, W. L., D. S. Maxwell & J. TiradoRives: Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118, 11225-11236 (1996)

36. Banks, J. L., G. A. Kaminski, R. H. Zhou, D. T. Mainz, B. J. Berne & R. A. Friesner: Parametrizing a polarizable force field from ab initio data. I. The fluctuating point charge model. *J. Chem. Phys*, 110, 741-754 (1999)

37. Yu, H. B. & W. F. van Gunsteren: Accounting for polarization in molecular simulation. *Comput. Phys. Commun.*, 172, 69-85 (2005)

38. Lazaridis, T. & M. Karplus: Effective energy function for proteins in solution. *Proteins Struct. Funct. Genet.*, 35, 133-152 (1999)

39. Qiu, D., P. S. Shenkin, F. P. Hollinger & W. C. Still: The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J. Phys. Chem. A.*, 101, 3005-3014 (1997)

40 Smith, A. V. & C. K. Hall: alpha-helix formation: Discontinuous molecular dynamics on an intermediate-resolution protein model. *Proteins Struct. Funct. Genet.*, 44, 344-360 (2001)

41. Head-Gordon, T. & S. Brown: Minimalist models for protein folding and design. *Curr. Opin. Struct. Biol.*, 13, 160-167 (2003)

42. Brown, S., N. J. Fawzi & T. Head-Gordon: Coarse-grained sequences for protein folding and design. *Proc. Nat. Acad. Sci. USA*, 100, 10712-10717 (2003)

43. Derreumaux, P. & N. Mousseau: Coarse-grained protein molecular dynamics simulations. *J. Chem. Phys.*, 126, (2007)

44. Marrink, S. J., H. J. Risselada, S. Yefimov, D. P. Tieleman & A. H. de Vries: The MARTINI force field: Coarse grained model for biomolecular simulations. *J. Phys. Chem. B*, 111, 7812-7824 (2007)

45. Liwo, A., M. Khalili & H. A. Scheraga: Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc. Nat. Acad. Sci. USA*, 102, 2362-2367 (2005)

46. Scheraga, H. A., M. Khalili & A. Liwo: Protein-folding dynamics: Overview of molecular simulation techniques. *Annu. Rev. Phys. Chem.*, 58, 57-83 (2007)

47. Dill, K. A., S. Bromberg, K. Z. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas & H. S. Chan: Principles of Protein-Folding - a Perspective from Simple Exact Models. *Protein Sci.*, 4, 561-602 (1995)

48. Chan, H. S. & K. A. Dill: Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics. *Proteins Struct. Funct. Genet.*, 30, 2-33 (1998)

49. Coluzza, I. & D. Frenkel: Designing specificity of protein-substrate interactions. *Phys. Rev. E*, 70, (2004)

50. Coluzza, I., S. M. van der Vies & D. Frenkel: Translocation boost protein-folding efficiency of double-barreled chaperonins. *Biophys. J.*, 90, 3375-3381 (2006)

51. Humphrey, W., A. Dalke & K. Schulten: VMD: Visual molecular dynamics. *J. Mol. Graphics.*, 14, 33-& (1996)

52. Lindorff-Larsen, K., M. Vendruscolo, E. Paci & C. M. Dobson: Transition states for protein folding have native topologies despite high structural variability. *Nat. Struct. Mol. Biol.,* 11, 443-449 (2004)

53. Periole, X., M. Vendruscolo & A. E. Mark: Molecular dynamics simulations from putative transition states of alpha-spectrin SH3 domain. *Proteins Struct. Funct. Bioinf.*, 69, 536-550 (2007)

54. Mackerell, A. D.: Empirical force fields for biological macromolecules: Overview and issues. *Journal of Computational Chemistry*, 25, 1584-1604 (2004)

55. Snow, C. D., B. Zagrovic & V. S. Pande: The Trp cage: Folding kinetics and unfolded state topology via molecular dynamics simulations. *J. Am. Chem. Soc.*, 124, 14548-14549 (2002)

56. Ferrenberg, A. M. & R. H. Swendsen: Optimized Monte-Carlo Data-Analysis. *Phys. Rev. Lett.*, 63, 1195-1198 (1989)

57. Shea, J. E. & C. L. Brooks: From folding theories to folding proteins: A review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem.*, 52, 499-535 (2001)

58. Grubmuller, H.: Predicting Slow Structural Transitions in Macromolecular Systems - Conformational Flooding. *Phys. Rev. E,* 52, 2893-2906 (1995)

59. Voter, A. F.: Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.,*78, 3908-3911 (1997)

60. Berg, B. A. & T. Neuhaus: Multicanonical Ensemble - a New Approach to Simulate 1st-Order Phase-Transitions. *Phys. Rev. Lett*, 68, 9-12 (1992)

61. Darve, E. & A. Pohorille: Calculating free energies using average force. *J. Chem. Phys.*, 115, 9169-9183 (2001)

62. Ensing, B., M. De Vivo, Z. W. Liu, P. Moore & M. L. Klein: Metadynamics as a tool for exploring free energy landscapes of chemical reactions. *Acc. Chem. Res.*, 39, 73-81 (2006)

63. Wang, F. G. & D. P. Landau: Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, 86, 2050-2053 (2001)

64. Bussi, G., F. L. Gervasio, A. Laio & M. Parrinello: Free-energy landscape for beta hairpin folding from combined parallel tempering and metadynamics. *J. Am. Chem. Soc.*, 128, 13435-13441 (2006)

65.Piana, S. & A. Laio: A bias-exchange approach to protein folding. *J. Phys. Chem. B*, 111, 4553-4559 (2007)

66. Mitsutake, A., Y. Sugita & Y. Okamoto: Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers*, 60, 96-123 (2001)

67. Sugita, Y. & Y. Okamoto: Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314, 141-151 (1999)

68. Garcia, A. E. & J. N. Onuchic: Folding a protein in a computer: An atomic description of the folding/unfolding of protein A. *Proc. Nat. Acad. Sci. USA*, 100, 13898-13903 (2003)

69. Lei, H. X., C. Wu, H. G. Liu & Y. Duan: Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc. Nat. Acad. Sci. USA*, 104, 4925-4930 (2007)

70. Sorin, E. J. & V. S. Pande: Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophys. J.*, 88, 2472-2493 (2005)

71. Juraszek, J. & P. G. Bolhuis: Sampling the multiple folding mechanisms of Trp-cage in explicit solvent. *Proc. Nat. Acad. Sci. USA*, 103, 15859-15864 (2006)

72. Seibert, M. M., A. Patriksson, B. Hess & D. van der Spoel: Reproducible polypeptide folding and structure prediction using molecular dynamics simulations. *J. Mol. Biol.*, 354, 173-183 (2005)

73. Periole, X. & A. E. Mark: Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent. *J. Chem. Phys.*, 126, (2007)

74. Liu, P., B. Kim, R. A. Friesner & B. J. Berne: Replica exchange with solute tempering: A method for sampling biological systems in explicit water. Proc. Nat. Acad. Sci. USA, 102, 13749-13754 (2005)

75. Coluzza, I. & D. Frenkel: Virtual-move parallel tempering. Chemphyschem, 6, 1779-1783 (2005)

76. Frenkel, D.: Speed-up of Monte Carlo simulations by sampling of rejected states. Proc. Nat. Acad. Sci. USA, 101, 17571-17575 (2004)

77. Faraldo-Gomez, J. D. & B. Roux: Characterization of conformational equilibria through Hamiltonian and temperature replica-exchange simulations: Assessing entropic and environmental effects. J. Comput. Chem., 28, 1634-1647 (2007)

78. Hummer, G., A. E. Garcia & S. Garde: Helix nucleation kinetics from molecular simulations in explicit solvent. Proteins Struct. Funct. Genet., 42, 77-84 (2001)

79. Duan, Y. & P. A. Kollman: Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science, 282, 740-744 (1998)

80. Ferrara, P., J. Apostolakis & A. Caflisch: Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations. *J. Phys. Chem. B*, 104, 5000-5010 (2000)

81. Zagrovic, B., C. D. Snow, M. R. Shirts & V. S. Pande: Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.,* 323, 927-937 (2002)

82. Snow, C. D., N. Nguyen, V. S. Pande & M. Gruebele: Absolute comparison of simulated and experimental protein-folding dynamics. *Nature*, 420, 102-106 (2002)

83.Beck, D. A. C. & V. Daggett: Methods for molecular dynamics simulations of protein folding/unfolding in solution. *Methods*, 34, 112-120 (2004)

84. Ferguson, N., R. Day, C. M. Johnson, M. D. Allen, V. Daggett & A. R. Fersht: Simulation and experiment at high temperatures: Ultrafast folding of a thermophilic protein by nucleation-condensation. *J. Mol. Biol.*, 347, 855-870 (2005)

85. Pande, V. S. & D. S. Rokhsar: Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein G. *Proc. Nat. Acad. Sci. USA*, 96, 9062-9067 (1999)

86. Li, A. J. & V. Daggett: Identification and characterization of the unfolding transition state of chymotrypsin inhibitor 2 by molecular dynamics simulations. *J. Mol. Biol*, 257, 412-429 (1996)

87. Chandler, D.: Statistical mechanics of isomerization dynamics in liquids and the transition state approximation. *J. Chem. Phys.* 68, 2959-2970 (1978)

88. Bennett, C.H., in: Algorithms for Chemical Computations, ACS Symposium Series No. 46, edited by R. Christofferson, American Chemical Society, Washington, D.C. (1977)

89. Dellago, C., P. G. Bolhuis, F. S. Csajka & D. Chandler: Transition path sampling and the calculation of rate constants. *J. Chem. Phys*, 108, 1964-1977 (1998)

90. Dellago, C., P. G. Bolhuis & P. L. Geissler: Transition path sampling. *Adv. Chem. Phys.,* 123. 1-78 (2002)

91. L.R. Pratt, A statistical method for identifying transition state in high dimension problems, *J. Chem. Phys:* 85, 5045-5048 (1986)

92 Bolhuis, P. G.: Transition path sampling on diffusive barriers. *J. Phys.Condens. Matter.*, 15, S113-S120 (2003)

93. H.C. Andersen: Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.,* 72, 2384-2393 (1980)

94. van Erp, T. S., D. Moroni & P. G. Bolhuis: A novel path sampling method for the calculation of rate constants. *J. Chem. Phys.*, 118, 7762-7774 (2003)

95. Bolhuis, P. G.: Kinetic pathways of beta-hairpin (Un)folding in explicit solvent. *Biophys. J.*, 88, 50-61 (2005)

96.Olender, R. & R. Elber: Calculation of classical trajectories with a very large time step: Formalism and numerical examples. *J. Chem. Phys.*, 105, 9299-9315 (1996)

97. Elber03 R. Elber, A. Ghosh, A. Cardenas, and H. Stern, Bridging the Gap Between Long Time Trajectories and Reaction Pathways. *Adv. Chem. Phys.* 126, 93 (2003)

98. Eastman, P., N. Gronbech-Jensen & S. Doniach: Simulation of protein folding by reaction path annealing. *J. Chem. Phys.*, 114, 3823-3841 (2001)

99. Moroni, D., P. G. Bolhuis & T. S. van Erp: Rate constants for diffusive processes by partial path sampling. *J. Chem. Phys.*, 120, 4055-4065 (2004)

100. Faradjian, A. K. & R. Elber: Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.*, 120, 10880-10889 (2004)

101. Singhal, N., C. D. Snow & V. S. Pande: Using path sampling to build better Markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.,*121, 415-425 (2004)

102. Amato, N. M. & G. Song: Using motion planning to study protein folding pathways. *J. Comput. Biol.*, 9, 149-168 (2002)

103 Amato, N. M., K. A. Dill & G. Song: Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. *J. Comput. Biol.*, 10, 239-255 (2003)

104 Krivov, S. V. & M. Karplus: Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Nat. Acad. Sci. USA*, 101, 14766-14770 (2004)

105. Bortz, A.B., M.H. Kalos & J.L. Lebowitz: A new algorithm for Monte Carlo simulation of Ising spin systems. *J. Comput. Phys.* 17, 10 (1975)

106. Chodera, J. D., W. C. Swope, J. W. Pitera & K. A. Dill: Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Model. Simul.*, 5, 1214-1226 (2006)

107. Schutte, C., A. Fischer, W. Huisinga & P. Deuflhard: A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys.*, 151, 146-168 (1999)

108. Deuflhard, P., W. Huisinga, A. Fischer & C. Schutte: Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra Appl.*, 315, 39-59 (2000)

109. Rhee, Y. M. & V. S. Pande: One-dimensional reaction coordinate and the corresponding potential of mean force from commitment probability distribution. *J. Phys. Chem. B*, 109, 6780-6786 (2005)

110. Yang, S. C., J. N. Onuchic, A. E. Garcia & H. Levine: Folding time predictions from all-atom replica exchange simulations. *J. Mol. Biol.*, 372, 756-763 (2007)

111. Du, R., V. S. Pande, A. Y. Grosberg, T. Tanaka & E. S. Shakhnovich: On the transition coordinate for protein folding. *J. Chem. Phys.*, 108, 334-350 (1998)

112. Rhee, Y. M. & V. S. Pande: On the role of chemical detail in simulating protein folding kinetics. *Chem. Phys.*, 323, 66-77 (2006)

113.Geissler, P. L., C. Dellago & D. Chandler: Kinetic pathways of ion pair dissociation in water. *J. Phys. Chem. B*, 103, 3706-3710 (1999)

114. Bolhuis, P. G., C. Dellago & D. Chandler: Reaction coordinates of biomolecular isomerization. *Proc. Nat. Acad. Sci. USA,* 97, 5877-5882 (2000)

115. Ma, A. & A. R. Dinner: Automatic method for identifying reaction coordinates in complex systems. *J. Phys. Chem. B*, 109, 6769-6779 (2005)

116. Quaytman, S. L. & S. D. Schwartz: Reaction coordinate of an enzymatic reaction revealed by transition path sampling. *Proc. Nat. Acad. Sci. USA*, 104, 12253-12258 (2007)

117. Jayachandran, G., V. Vishal & V. S. Pande: Using massively parallel simulation and Markovian models to study protein folding: Examining the dynamics of the villin headpiece. *J. Chem. Phys.*, 124, (2006)

118. So, S. S. & M. Karplus: Evolutionary optimization in quantitative structure-activity relationship: An application of genetic neural networks. *J. Med. Chem.*, 39, 1521-1530 (1996)

119. Dinner, A. R., S. S. So & M. Karplus: Statistical analysis of protein folding kinetics. In: Computational Methods for Protein Folding. *Adv. Chem. Phys.*, 120, 1 (2002)

120. Hummer, G.: From transition paths to transition states and rate coefficients. *J. Chem. Phys*, 120, 516-523 (2004)

121. Best, R. B. & G. Hummer: Reaction coordinates and rates from transition paths. *Proc. Nat. Acad. Sci. USA*, 102, 6732-6737 (2005)

121. Peters, B. & B. L. Trout: Obtaining reaction coordinates by likelihood maximization. *J. Chem. Phys*, 125, (2006)

123. Peters, B., G. T. Beckham & B. L. Trout: Extensions to the likelihood maximization approach for finding reaction coordinates. *J. Chem. Phys*, 127, (2007)

124. Beckham, G. T., B. Peters, C. Starbuck, N. Variankaval & B. L. Trout: Surface-mediated nucleation in the solid-state polymorph transformation of terephthalic acid. *Journal of the American Chemical Society*, 129, 4714-4723 (2007)

125. Roccatano, D., A. Amadei, A. Di Nola & H. J. C. Berendsen: A molecular dynamics study of the 41-56 beta-hairpin from B1 domain of protein G. *Protein Science*, 8, 2130-2143 (1999)

126. Munoz, V., P. A. Thompson, J. Hofrichter & W. A. Eaton: Folding dynamics and mechanism of beta-hairpin formation. *Nature*, 390, 196-199 (1997)

127. Munoz, V., E. R. Henry, J. Hofrichter & W. A. Eaton: A statistical mechanical model for beta-hairpin kinetics. *Proc. Nat. Acad. Sci. USA*, 95, 5872-5879 (1998)

128. Kolinski, A., B. Ilkowski & J. Skolnick: Dynamics and thermodynamics of beta-hairpin assembly: Insights from various simulation techniques. *Biophys. J.*, 77, 2942-2952 (1999)

129. Klimov, D. K. & D. Thirumalai: Mechanisms and kinetics of beta-hairpin formation. *Proc. Nat. Acad. Sci. USA* , 97, 2544-2549 (2000)

130. Wei, G. H., N. Mousseau & P. Derreumaux: Complex folding pathways in a simple beta-hairpin. *Proteins Struct. Funct. Bioinf.*, 56, 464-474 (2004)

131. Dinner, A. R., T. Lazaridis & M. Karplus: Understanding beta-hairpin formation. *Proc. Nat. Acad. Sci. USA,* 96, 9068-9073 (1999)

132. Zagrovic, B., E. J. Sorin & V. Pande: beta-hairpin folding simulations in atomistic detail using an implicit solvent model. *J. Mol. Biol.*, 313, 151-169 (2001)

133. Ma, B. Y. & R. Nussinov: Molecular dynamics simulations of a beta-hairpin fragment of protein G: Balance between side-chain and backbone forces. *J. Mol. Biol.*, 296, 1091-1104 (2000)

134. Garcia, A. E. & K. Y. Sanbonmatsu: Exploring the energy landscape of a beta hairpin in explicit solvent. *Proteins Struct. Funct. Genet.*, 42, 345-354 (2001)

135. Zhou, R. H., B. J. Berne & R. Germain: The free energy landscape for beta hairpin folding in explicit water. *Proc. Nat. Acad. Sci. USA*, 98, 14931-14936 (2001)

136. Tsai, J. & M. Levitt: Evidence of turn and salt bridge contributions to beta-hairpin stability: MD simulations of C-terminal fragment from the B1 domain of protein G. *Biophys. Chem.*, 101, 187-201 (2002)

137. Sheinerman, F. B. & C. L. Brooks: Calculations on folding of segment B1 of streptococcal protein G. *J. Mol. Biol.*, 278, 439-456 (1998)

138. Sheinerman, F. B. & C. L. Brooks: Molecular picture of folding of a small alpha/beta protein. *Proc. Nat. Acad. Sci. USA*, 95, 1562-1567 (1998)

139. Cheung, M. S., A. E. Garcia & J. N. Onuchic: Protein folding mediated by solvation: Water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Nat. Acad. Sci. USA*, 99, 685-690 (2002)

140. Neidigh, J. W., R. M. Fesinmeyer & N. H. Andersen: Designing a 20-residue protein. *Nat. Struct. Biol.*, 9, 425-430 (2002)

141. Qiu, L. L., S. A. Pabit, A. E. Roitberg & S. J. Hagen: Smaller and faster: The 20-residue Trp-cage protein folds in 4 mu s. *J. Am. Chem. Soc.*, 124, 12952-12953 (2002)

142.Neuweiler, H., S. Doose & M. Sauer: A microscopic view of miniprotein folding: Enhanced folding efficiency through formation of an intermediate. *Proc. Nat. Acad. Sci. USA*, 102, 16650-16655 (2005)

143. Ahmed, Z., I. A. Beta, A. V. Mikhonin & S. A. Asher: UV-resonance Raman thermal unfolding study of Trp-cage shows that it is not a simple two-state miniprotein. *J. Am. Chem. Soc.*, 127, 10943-10950 (2005)

144. Simmerling, C., B. Strockbine & A. E. Roitberg: All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.*, 124, 11258-11259 (2002)

145. Pitera, J. W. & W. Swope: Understanding folding and design: Replica-exchange simulations of "Trp-cage" fly miniproteins. *Proc. Nat. Acad. Sci. USA*, 100, 7587-7592 (2003)

146. Chowdhury, S., M. C. Lee & Y. Duan: Characterizing the rate-limiting step of Trp-cage folding by all-atom molecular dynamics simulations. *J. Phys. Chem. B*, 108, 13855-13865 (2004)

147. Ota, M., M. Ikeguchi & A. Kidera: Phylogeny of protein-folding trajectories reveals a unique pathway to native structure. *Proc. Nat. Acad. Sci. USA*, 101, 17658-17663 (2004)

148. Zhou, R. H.: Trp-cage: Folding free energy landscape in explicit water. *Proc. Nat. Acad. Sci. USA*, 100, 13280-13285 (2003)

149. Linhananta, A., J. Boer & I. MacKay: The equilibrium properties and folding kinetics of an all-atom G (o)over-bar model of the Trp-cage. *J. Chem. Phys.*, 122, (2005)

150. Ding, F., S. V. Buldyrev & N. V. Dokholyan: Folding Trp-cage to NMR resolution native structure using a coarse-grained protein model. *Biophys. J.*, 88, 147-155 (2005)
151. Rhee, Y. M., E. J. Sorin, G. Jayachandran, E. Lindahl & V. S. Pande: Simulations of the role of water in the protein-folding mechanism. *Proc. Nat. Acad. Sci. USA*, 101, 6456-6461 (2004)

152. Paschek, D., H. Nymeyer & A. E. Garcia: Replica exchange simulation of reversible folding/unfolding of the Trp-cage miniprotein in explicit solvent: On the structure and possible role of internal water. *J. Struct. Biol.*, 157, 524-533 (2007)

153. Lindahl, E., B. Hess & D. van der Spoel: GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.*, 7, 306-317 (2001)

154. Kaminski, G. A., R. A. Friesner, J. Tirado-Rives & W. L. Jorgensen: Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B*, 105, 6474-6487 (2001)

155. Caflisch, A. & M. Karplus: Acid and Thermal-Denaturation of Barnase Investigated by Molecular-Dynamics Simulations. *J. Mol. Biol.*, 252, 672-708 (1995)

156. Guo, W. H., S. Lampoudi & J. E. Shea: Posttransition state desolvation of the hydrophobic core of the src-SH3 protein domain. *Biophys. J.*, 85, 61-69 (2003)

157 ten Wolde, P. R. & D. Chandler: Drying-induced hydrophobic polymer collapse. *Proc. Nat. Acad. Sci. USA,* 99, 6539-6543 (2002)

158. Miller, T. F., E. Vanden-Eijnden & D. Chandler: Solvent coarse-graining and the string method applied to the hydrophobic collapse of a hydrated chain. *Proc. Nat. Acad. Sci. USA*, 104, 14559-14564 (2007)

Abbreviations: BC: Bennett-Chandler,CG: coarse-grained,GNN: genetic neural network,LM: likelihood maximization,MC: Monte Carlo,MD: molecular dynamics,MFPT: mean first passage time,MSM: Markovian state model,NVE: constant energy,PPTIS: partial path TIS,PT: parallel tempering,REM: replica exchange method,REMD: replica exchange molecular dynamics,RMSD: root mean square deviation,SASA: solvent accessible surface area,TIS: transition interface sampling,TP: transition path,TPE : transition path ensemble,TPS: transition path sampling,TS: transition state,TSE: transition state ensemble,TST: transition state theory,US: umbrella sampling

Key Words: Computer simulation, Rare Events, Rate Constant, Reaction Mechanism, Transition Pathway, Review

Send correspondence to: Peter G. Bolhuis, Van ‘t Hoff Institute for Molecular Sciences, Nieuwe Achtergracht 166, 1018 WV Amsterdam, Netherlands, Tel: 31-205256447, Fax 31-205255604, E-mail: bolhuis@science.uva.nl