Denaturants alter the flux through multiple pathways in the folding of PDZ domain

Zhenxing Liu[*,†] and D. Thirumalai[‡]

†*Department of Physics, Beijing Normal University, Beijing 100875, China*

‡*Department of Chemistry, The University of Texas at Austin, TX, 78712, United States*

E-mail: zxliu@bnu.edu.cn

Abstract

Although we understand many aspects of how small proteins (number of residues less than about hundred) fold, it is a major challenge to understand how large proteins self-assemble. To partially overcome this challenge, we performed simulations using the Self-Organized Polymer model with Side Chains (SOP-SC) in guanidinium chloride (GdmCl), using the Molecular Transfer Model (MTM), to describe the folding of the 110-residue PDZ3 domain. The simulations reproduce the folding thermodynamics accurately including the melting temperature (T_m), the stability of the folded state with respect to the unfolded state. We show that the calculated dependence of $\ln k_{obs}$ (k_{obs} is the relaxation rate) has the characteristic Chevron shape. The slopes of the Chevron plots are in good agreement with experiments. We show that PDZ3 folds by four major pathways populating two metastable intermediates, in accord with the kinetic partitioning mechanism. The structure of one of the intermediates, populated after polypeptide chain collapse, is structurally similar to an equilibrium intermediate. Surprisingly, the connectivities between the intermediates and hence, the fluxes through the pathways depend on the concentration of GdmCl. The results are used to predict possible outcomes for unfolding of PDZ domain subject to mechanical forces. Our study demonstrates that, irrespective of the size or topology, simulations based on MTM and SOP-SC offer a framework for describing the folding of proteins, mimicking precisely the conditions used in experiments.

Introduction

The most common way of initiating folding (unfolding) of proteins in ensemble and single molecule experiments is by decreasing (increasing) the concentration of denaturants. Thus, direct comparison with experiments is only possible if simulations are done using models that take the effects of denaturants into account.[1] Although atomic detailed simulations hold the promise of quantitative description of denaturant-induced folding or unfolding,[2–6] currently the only available method for obtaining the thermodynamics and folding kinetics of proteins,

even for proteins as large as GFP,[7] is the Molecular Transfer Model (MTM) in combination with coarse-grained SOP-SC representation of polypeptide chain.[8,9] The theoretical basis for the success of the MTM has been explained elsewhere.[10] Applications of MTM to probe folding of a variety of proteins have yielded quantitative agreement with experiments[7,8,11–13] which attests to the efficacy of the MTM.

One of the early applications of MTM was the demonstration that the Chevron plot of the 56-residue srcSH3 domain could be reproduced nearly quantitatively.[9] However, the extension of these calculations to proteins with more than hundred residues has been difficult even using the simplified coarse-grained SOP-SC models. One of the goals of this study is to overcome this challenge. The second problem that we address is to establish the folding mechanism as a function of denaturant concentration for a large single domain protein. If the folding mechanism involve parallel pathways, as theoretical and computational studies have firmly established,[14–17] then are the fluxes through the pathways modulated by changing the external conditions such as denaturant concentration or mechanical forces? Understanding the origin of parallel folding and unfolding pathways, and how they are altered by environmental changes, is important in establishing the generality of the protein folding mechanisms. Here, we investigate the denaturant-dependent folding and unfolding of PDZ3, a protein with110 residues.

PDZ domains, found in many cell junction-associated proteins, are a large family of globular proteins that mediate protein-protein interactions and play an important role in molecular recogniton.[18–21] The folding of PDZ3 domain, a member of this family, has been studied both by experiments and computations,[22–26] arriving at different conclusions. In GmdCl-induced equilibrium unfolding experiments, the folding transition appears to be highly cooperative, seemingly displaying a simple two-state behavior. In the Chevron plot of PDZ3 domain in GdmCl solution at pH= 6.3, both the folding and unfolding arms are linear functions of GdmCl concentrations $[C]$, indicating no detectable intermediate states at the ensemble level. However, native-state hydrogen exchange experiments reveal hidden intermediate states un-

der native conditions.[22] Interestingly, the addition of potassium formate at pH= 2.85 induces a rollover in the unfolding arm in the Chevron plot, suggestive of an intermediate.[24,25] This finding is reminiscent of the salt-induced detour found in the folding of the protein S6.[27] In the presence of potassium phosphate at pH= 7.5, there are two thermal unfolding transitions in DSC experiments, further demonstrating the existence of intermediate states, at least in the presence of salt.[26] Based on these experiments, we surmise that the PDZ3 domain folding must occur by multiple pathways, even if they are hard to detect in generic ensemble experiments. If there are multiple pathways, is it possible that the fluxes through these pathways could change by altering the concentration of denaturant in this large single domain as reported for a two-domain protein[28]? Here, we answer this question in the affirmative using MTM simulations in the folding of PDZ3 domain as a function of GdmCl concentration, after establishing that the SOP-SC simulations capture the key findings in ensemble experiments.

We combine simulations of coarse-grained off-lattice SOP-SC model[29–31] and the molecular transfer model[8–10] to decipher the folding mechanism of PDZ3 domain. The calculated fractions of molecules in the native basin of attraction(NBA), f_{NBA}, and the unfolded basin of attraction(UBA), f_{UBA} as a function of the denaturant concentration $[C]$ are in excellent agreement with experiments. In addition, we find that a small fraction, f_{IBA}, of an intermediate, not resolved in ensemble experiments, is populated in equilibrium The Tanford β parameters for the two transition state ensembles obtained in simulations are in quantitative agreement with those inferred from experiments.[24,25] Chevron plot calculated for the first time for a protein with over 100 amino acids, allows us to extract the locations of the transition state ensembles (TSEs) in terms of the Tanford β parameters. The calculated free energy profiles suggest that the at low (high) [GdmCl] a less (more) structured TSE is rate limiting. Folding trajectories both in aqueous and in denaturant solutions demonstrate directly the existence of the thermodynamic intermediate states as well as kinetic intermediate states. Our simulations vividly illustrate four parallel folding pathways in molecular detail.

We show that the fluxes between the assembly pathways can be modulated by varying the denaturant concentration. Many of our predictions are amenable to experimental test.

Methods

SOP-SC model:

Our simulations were carried out using the SOP-SC (Self-Organized Polymer-Side Chain) model for the protein.[9,10] Each residue is represented by two interaction centers, with one centered at the C_α position, and the other located at the center of mass of the side chain. The energy function of a conformation in the SOP-SC representation of the polypeptide chain is,

$$E_P(\{r_i\}) = V_{LJ}^{NAT} + V^{NEI} + V_{LJ}^{NN} + V_{FENE}. \tag{1}$$

The detailed functional form and the values of the parameters are described elsewhere.[9]

Molecular Transfer Model: In the MTM the effective free energy function for a protein in aqueous denaturant solution is

$$\begin{aligned} G_P(\{r_i\}) =& E_P(\{r_i\}) + \Delta G(\{r_i\}, [C]) \\ =& E_P(\{r_i\}) + \sum_i \delta g(i, [C]) \alpha_i / \alpha_{Gly-i-Gly}, \end{aligned} \tag{2}$$

where $\Delta G(\{r_i\}, [C])$ is the free energy of transferring a given protein conformation from water to aqueous denaturant solution, the sum is over all the interaction centers (i), $\delta g(i, [C])$ is the transfer free energy of interaction center i, α_i is the solvent accessible surface area (SASA), and $\alpha_{Gly-i-Gly}$ is the SASA of the ith interaction center in the tripeptide $Gly - i - Gly$. We used the procedure described in detail previously[9,10] to calculate the thermodynamic properties of proteins in the presence of denaturants.

Langevin Dynamics: We assume that the dynamics of the protein is governed by the

Langevin equation,

$$m\ddot{\mathrm{r}}_i = -\zeta\dot{\mathrm{r}}_i + F_c + \Gamma, \tag{3}$$

where m is the mass of a bead, ζ is the friction coefficient, $F_c = -\partial E_P(\{r_i\})/\partial r_i$ is the conformational force calculated using Eq. (1), Γ is the random force with a white noise spectrum. To enhance sampling, we used Replica-Exchange Molecular Dynamics (REMD)[32–34] to carry out thermodynamics sampling at low friction coefficient $\zeta_L = 0.05m/\tau_L$.[35] Here, τ_L is the unit of time (see below) for use in the computation of thermodynamic quantities, and m is the average mass of the beads. It is obvious that the precise values of these two quantities do not play a role in the determination of the equilibrium properties of interest. In the underdamped limit, we employ the Verlet leap-frog algorithm to integrate the equations of motion.

Brownian Dynamics: To obtain a realistic description of the kinetics of folding or unfolding, we set $\zeta_H = 50m/\tau_L$, which approximately corresponds to the value of the friction coefficient in water.[36] At the high ζ value where the inertial forces are negligible, we use the Brownian dynamics algorithm[37] to integrate equations of motion using

$$r_i(t+h) = r_i(t) + \frac{h}{\zeta}(F_c(t) + \Gamma(t)). \tag{4}$$

It is important to note for obtaining kinetic properties the systematic force, F_c in Eq. 4 is,

$$F_c = -\partial G_P(\{r_i\}/\partial r_i, \tag{5}$$

where $G_P(\{r_i\})$ is given in Eq. (2).

Time scales: In the high ζ limit the unit of time is $\tau_H \approx \frac{\zeta_H a^2}{k_B T_s} = \frac{(\zeta_H \tau_L/m)e_l}{k_B T_s}\tau_L$. Following Veitshans,[36] we chose $e_l = 1\ kcal/mol$, average mass $m = 1.8 \times 10^{-22}\ g$, $a = 4$ Å, which makes $\tau_L = 2\ ps$. For $\zeta_H = 50\ m/\tau_L$, we obtain $\tau_H = 164\ ps$. These estimates are used to obtain estimates of the folding times from our Brownian dynamics simulations.

In Langevin dynamics simulations, the integration time step, $h = 0.05\tau_L$, whereas in the Brownian dynamics simulations, $h = 0.1\tau_H$ (Eq. (4)).

Results

The structure of the $N = 110$ residue PDZ3 domain from PSD-95 is shown in Fig-1A (PDB ID: 1BFE). In contrast to the typical structure of a PDZ domain, it has one additional helix (α_3 and a two-stranded β sheet formed between two short strands, β_7 and β_8 at the C terminal. They are colored in red in Fig-1A.

Melting temperatures: The melting temperature, identified with the peak in the heat capacity $C_v(T)$ at $[C] = 0$ (black line in Fig-1B) is $T_m = 323K$, which is in good agreement with the experimentally measured $T_m = 344K$.[26] Similarly, by associating the melting temperatures with the peaks in the heat capacity at different values of $[C]$ (Fig-1 B) we determined the dependence of $T_m[C]$ on $[C]$. It is clear from Fig-1 C that $T_m[C]$ is linearly dependent on $[C]$ (see figure caption for the parameters of the fit).

Boundaries between the distinct thermodynamic states: To define the Native Basin of Attraction (NBA) and the basins corresponding to the unfolding (UBA) and intermediate states (IBA), we obtained the free energy profile $G(\chi)$ as a function of the structural overlap function, χ, which serves as an order parameter. The structural overlap function $\chi = 1 - \frac{N_k}{N_T}$ where,

$$\begin{aligned} N_k &= \sum_{i=1}^{N-3}\sum_{j=i+3}^{N} \Theta\left(\delta - |r_{i,j_bb} - r^o_{i,j_bb}|\right) + \sum_{i=1}^{N-3}\sum_{j=i+3}^{N} \Theta\left(\delta - |r_{i,j_ss} - r^o_{i,j_ss}|\right) \\ &+ \sum_{\substack{i=1,j=1 \\ |i-j|\geq 3}}^{N} \Theta\left(\delta - |r_{i,j_bs} - r^o_{i,j_bs}|\right) \end{aligned} \tag{6}$$

In Eq. (6), $\Theta(x)$ is the Heavyside function. If $|r_{i,j} - r^o_{i,j}| \leq \delta{=}2\mathring{A}$, there is a contact. N_k

is the number of contacts in the kth conformation and N_T is the total number in the folded state. The profile $G(\chi)$ at $T_m[0]$ (Fig-2A) shows that the conformations can be classified into three groups separated by the black vertical lines; $\chi_c^N = 0.73$ and $\chi_c^D = 0.84$ are the cutoff values separating the NBA, UBA, and IBA. If $\chi \leq \chi_c^N$, the corresponding conformation belongs to the NBA and if $\chi \geq \chi_c^D$, the conformation belongs to UBA, and all the other conformations are grouped into IBA.

GdmCl dependence of thermodynamic stability and m-value: In order to compare with experiments, we simulated the effects of GdmCl using the Molecular Transfer Model (MTM).[8] Following our previous studies,[8–10] we choose a simulation temperature, T_s, at which the calculated free energy of stability of the native state (N) with respect to the unfolded state (U), $\Delta G_{NU}(T_s)$ ($G_N(T_s) - G_U(T_s)$) and the measured free energy at T_E (=298K) $\Delta G_{NU}(T_E)$ coincide. For PDZ3, $\Delta G_{NU}(T_E = 298K) = -7.4 kcal/mol$ at $[C] = 0$,[22] which results in $T_s = 306K$, which is close to T_E. It is worth emphasizing that besides the choice of T_s no other parameter is adjusted to fit any experimental data.

With $T_s = 306K$ fixed, we calculated the dependence of the fraction of molecules in the NBA, $f_{NBA}([C], T_s)$, in the UBA, $f_{UBA}([C], T_s)$, and in the IBA, $f_{IBA}([C], T_s)$, on $[C]$ (Fig-2B). The midpoint concentration, C_m, obtained using $f_{UBA}([C_m], T_s) = 0.5$ is [C]=3.05M, which is very close to 2.90M, measured in experiments.[22] The agreement between the measured and simulated results for $f_{UBA}([C], T_s)$ versus $[C]$ is excellent (Fig. 2B), further establishing the efficacy of the MTM and SOP-SC simulations in capturing the folding thermodynamics of single domain proteins in general, and PDZ3 in particular.

The small value of $f_{IBA}([C], T_s)$, compared to $f_{NBA}([C], T_s)$ and $f_{UBA}([C], T_s)$, explains why the intermediate state is hard to detect in the equilibrium denaturation experiments.[22] Our finding that $f_{IBA}([C], T_s)$ is small is consistent with the observed protein-concentration dependent thermal unfolding in DSC experiments, where at low PDZ3 concentration only one transition is observed as the associated intermediate is not substantially populated.[26]

The native state stability with respect to U, $\Delta G_{NU}([C])(= G_N([C]) - G_U([C]))$, is calcu-

lated using $\Delta G_{NU}([C]) = -k_B T_s ln(\frac{f_{NBA}}{f_{UBA}})$. The linear fit, $\Delta G_{NU}([C]) = \Delta G_{NU}(0) + m[C]$, yields $\Delta G_{NU}([0]) = -6.18 kcal/mol$ and $m = 2.03 kcal/mol \cdot M$ (Fig. 2C), which is in reasonable agreement with experimental estimate $m = 2.50 kcal/mol \cdot M$.[22]

Free energy profiles as function of the order parameter χ: To illustrate how GdmCl changes the free energy landscape, we plotted the free energy profiles as functions of χ at different $[C]$ at $T_m[C]$ in Fig-3A and at a fixed temperature, $T_s = 306K$ in Fig-3B. We use χ in Eq. (6), the microscopic order parameter of the protein, to distinguish between the native, the unfolded and high-energy intermediate states.[38] Fig-3B shows that, at low [GdmCl], the first transition state ensemble, TSE1, is rate limiting for folding. However, at high [GmdCl] the second transition state ensemble, TSE2, is rate limiting. These findings agree qualitatively with the energy diagram for the folding reactions proposed elsewhere.[24] The movement of the transition states with changing concentration is in accord with the Hammond postulate.

Structures of the transition state ensembles (TSEs): The free energy profile as a function of χ in Fig. 2A, suggests that there are two barriers. The ensembles of conformations at their locations, grouped as TSE1 and TSE2, are shown by the shaded areas. The global characteristic of the TSE in ensemble experiments is usually described using the Tanford parameter, β. From the observed chevron plot, $(\beta_E^1, \beta_E^2) = (0.53, 0.87)$[24] or $(0.56, 0.90)$ for TSE1, TSE2 respectively.[25] It is generally assumed, that β is related to the buried solvent accessible surface area (SASA) in the TSE. For the TSE obtained in our simulations, we calculated the distribution $P(\Delta_R)$ (Fig-4A), where $\Delta_R = (\Delta_U - \Delta_{TSE})/(\Delta_U - \Delta_N)$ with Δ_U, Δ_{TSE}, Δ_N are the SASA in the DSE ($[C] = 8.0M$), TSE, and the NBA ($[C] = 0.0M$), respectively. We found that the average $< \Delta_R >= 0.41 \approx \beta_s^1$ for TSE1 and $< \Delta_R >= 0.74 \approx \beta_s^2$ for TSE2, which are in qualititative agreement with the experimentally measured values. The small deviations between simulations and experiments may be due to the following reasons. (1) The PDZ3 domain in the simulations has one additional helix and one β sheet, which is absent in the construct used in the experiments. These extra structural elements are

highly flexible even under native conditions (upper left in Fig-4B), which lowers $< \Delta_R >$. (2) When fitting the chevron plot to obtain β in experiments, both $(\beta_E^1, \beta_E^2) = (0.66, 0.93)$ and $(0.3, 0.8)$ give reasonable fits, indicating a range of β can describe the experimental data.[24] Given these observations, we surmise that the Tanford β calculated using the simulation data is in reasonable agreement with experimental values.

Fig-4B and Fig-4C show the contact maps obtained from the TSE1 and TSE2, respectively. It is clear that, relative to the native state (upper left) the TSE1 structures, populated at low $[C]$, is moderately structured. In contrast, the structures are ordered to a greater extent in the TSE2. The contacts between $\beta 1 - \beta 6$ have very low probabilities of formation indicated by the major blue region in Fig-4B, and have moderate formation probabilities indicated by the major green region in Fig-4C. One representative structure for TSE1 is shown on the left of Fig-4A, where we can see that β_2, β_3, β_4, β_5, β_6, α_1, α_2 and α_3 are packed loosely with β_1, β_7, β_8 forming no contact with the core of the protein. Note that β_7, β_8 and α_3 are the extra regions in our simulations compared to the typical structure of PDZ domain. A representative structure for TSE2 is shown on the right of Fig-4A, where the core of the native topology is well established except for β_1, β_7 and β_8. We should point out a minor discrepancy between our results and a previous study.[25] We find that β_1 is unstructured in the TSE2 but is found to be structured by Jemth.[25]

Folding kinetics and the Chevron Plot: We calculated the $[C]$-dependent folding (unfolding) rates from folding (unfolding) trajectories, which were generated from Brownian dynamics using the effective energy function, $G_P(\{r_i\})$ (see Methods). From sixty (one hundred for $[C] = 0$) folding trajectories, the fraction of unfolded molecules at time t, is computed using $P_u(t) = 1 - \int_0^t P_{fp}(s)ds$, where $P_{fp}(s)$ is the distribution of first passage times. We fit $P_u(t) \approx e^{-tk_f[C]}$ under folding conditions ($[C] < C_m$), from which $k_f[C]$ can be extracted. Similarly, a single exponential fit for unfolding conditions ($[C] > C_m$) yields $k_u[C]$. At high (low) $[C]$, we can approximate $k_{obs} = k_f([C]) + k_u([C])$ as $k_u([C])$ ($k_f([C])$). We globally fit the relaxation rate, k_{obs} using $\ln k_{obs} = \ln[k_f([0])e^{-m_f[C]/RT} + k_u([0])e^{m_u[C]/RT}]$,

where m_f (m_u) is the slope of the folding (unfolding) arm with $\ln k_f = \ln k_f(0) - m_f[C]/RT$ and $\ln k_u = \ln k_u(0) + m_u[C]/RT$.

A plot of $\ln k_{obs}$ as a function of $[C]$ over a wide concentration range ($0M \leq [C] \leq 8.0M$) shows a classic Chevron shape (Fig-5) observed in several experiments for a number of proteins. In the range $[C] \leq 1.5M$, $k_u \ll k_f$, so that $k_{obs} \approx k_f$ and similarly for $[C]$ above $4.5M$, $k_{obs} \approx k_u$. In the transition region ($2.0M \leq [C] \leq 4.0M$), the folding and unfolding rates are too small to be reliably calculated even using the SOP-SC simulations. Because the size of the PDZ3 domain is relatively large (110 amino acids) we could not generate folding and unfolding rates reliably around the midpoint even using the SOP-SC model. Comparison of the simulation and experimental results shows that the slopes (from the folding and unfolding arms) of the simulated chevron plot are qualitatively similar to the experimental values.

From the slope of the folding arm (simulation results in Fig-5), we obtain $m_f = 0.91 kcal/mol \cdot M$ and $m_u = 0.63 kcal/mol \cdot M$ from the unfolding arm. The corresponding experimental values are $m_f^E = 1.58 kcal/mol \cdot M$ and $m_u^E = 0.88 kcal/mol \cdot M$,[22] which shows fair agreement between experiments and simulations. Since the the fraction of molecules in IBA is negligibly small, both thermodynamics and kinetics simulations can be approximately described by a two-sate model and hence we expect $m \approx m_f + m_u$. From the simulated Chevron plot, we obtain $m \approx 1.54 kcal/mol \cdot M$, which differs by $\sim 24\%$ from $m = 2.03 kcal/mol \cdot M$ obtained from equilibrium $\Delta G_{NU}[C]$ calculations (Fig. 2C). In contrast, the relation $m = m_f + m_u = 2.46 kcal/mol \cdot M$, which is close to $m = 2.50 kcal/mol \cdot M$ found in equilibrium titration experiments. We conclude that our simulations capture only the qualitative features of the denaturant-dependent folding kinetics of PDZ3 domain.

Although MTM simulations reproduce the Chevron shape well, the dependence of $ln k_{obs}$ on $[C]$ does not agree quantitatively with experiments. For instance, $k_f[0]$ from simulations is 278 s^{-1}, which is only ≈ 1.6 times larger than the extrapolated value 170 s^{-1} for $[C] = 0$ from experiment. However, the unfolding rate at $[C] = 0$, $k_u[0]$ from simulations is 0.91

s^{-1} while $k_u[0]$ from experiment is 0.0022 s^{-1}. The simulations overestimate the unfolding rate by about 414 fold compared to experiments, even though the values of the slopes of the unfolding arms from simulations and experiments are reasonably close.

Fluxes through parallel pathways depend on the denaturant concentration: By analyzing the folding trajectories by using χ as the progress variable for the folding reaction, we find that PDZ3 folds along four distinct pathways. One representative trajectory for each pathway is shown in Fig-6, where χ is displayed as a function of t. In each pathway folding occurs in stages. In addition to the conformations in the UBA and the NBA, we identified two intermediate states ($KIN_1$and KIN_2), whose lifetimes vary greatly depending on the pathway. Arrows of each color in Fig-7 represent one folding pathway and the thickness of the arrows represents the probability of the pathway. At $[C] = 0$ (Fig-7A), the dominant pathway P1 is $D \rightarrow KIN_1 \rightarrow KIN_2 \rightarrow N$ (black arrows), through which $\sim$ 52% of the flux to the native state is channeled. In this pathway, β sheets between strands 1, 6, 4 form transiently in KIN_1 state, followed by the consolidation of core β sheets between strands 2, 3, 4 and α_2 in the KIN_2 state. The less probable alternative pathway P2 is $D \rightarrow KIN_2 \rightarrow N$ (red arrows), representing $\sim$ 38% of the trajectories, where folding occurs only through the KIN_2 state. Similarly, in the third probable pathway P3, $D \rightarrow KIN_1 \rightarrow N$ (green arrows), through which about $\sim$ 10% of the flux to the native state flows, folding occurs only through KIN_1 state.

The flux through P1, identified at $[C] = 0$ remains dominant at $[C] = 0.5M$ (Fig-7B, $\sim$ 52%) and $[C] = 1.0M$ (Fig-7C, $\sim$ 61%). The P2 and P3 pathways have lesser probabilities $\sim$ 36%, $\sim$ 6% for [C]=0.5M and $\sim$ 28%, $\sim$ 9% for [C]=1.0M, respectively. A direct pathway P4, $D \rightarrow N$(blue arrows) is observed with small probabilities ($\sim$ 6% for [C]=0.5M, $\sim$ 2% for [C]=1.0M). Thus, the PDZ3 domain folds through heterogeneous pathways. Most importantly, the populations of the folding pathways are sensitive to denaturant concentrations. Denaturant-modulated parallel pathways were also observed for adenylate kinase.[28] Interestingly, such parallel folding has been observed in the folding of both small proteins[39]

and larger proteins.[28,40,41]

Discussion

Post-Collapse kinetic intermediate is structurally similar to an equilibrium intermediate: To illustrate the relationship between the thermodynamically observed intermediate state (I_{EQ}) and kinetically observed intermediate states (KIN_1, KIN_2), we calculated the average fraction of native contacts at every residue f_Qs. The correlations between the f_Qs for the three states are shown in Fig-8A and Fig-8B. The correlation between I_{EQ} and KIN_1 is very low (correlation coefficient, R=0.3), indicating that at the early stages of folding a variety of compact but structurally diverse states are explored. The observation that in the initial stages of organization a heterogeneous mixture of states with small thermodynamic states are sampled is consistent with early atomic detailed simulations on cytochrome c.[42]

In contrast, the correlation between the calculated f_Q between I_{EQ} and KIN_2 is high (Fig-8B). with the correlation coefficient, R=0.98. A linear fit of the line in Fig-8B gives $y = A + Bx$ with $A = 0.008 \pm 0.007$, $B = 1.175 \pm 0.021$. Since the intercept, A, is close to zero, and the slope, B, is close to unity, we conclude that KIN_2 and I_{EQ} are the same intermediate species. Fig-8C shows, for KIN_1, I_{EQ} and KIN_2, the average fraction of native contacts at every residue f_Qs by color, vividly demonstrating the similarity between KIN_2 and I_{EQ}. These results show that in the later stages of folding the equilibrium and kinetic intermediates coincide, which is expected to hold for most, if not all, foldable proteins. In the later stages of folding, which corresponds to the stage after chain compaction, native-like interactions dominate, which was first shown using lattice models[43,44] for which precise computations can be performed. Because native-like structures, which have considerable Boltzmann weight, dominate it follows that the structural features of KIN_2 and I_{EQ} should coincide.

Denaturants alter the connectivity between the metastable intermediates:

The most interesting finding that denaturants alter the connectivity between the intermediates, and hence the fluxes through the distinct pathways, was already demonstrated in a most beautiful single-molecule fluorescence energy transfer (smFRET) experiment.[28] Using adenylate kinase (ADK), a 214-residue two (or three) domain protein, and by collecting a large number of trajectories Haran and coworkers showed that during the folding process six metastable intermediates are populated. Most importantly, they showed using Hidden Markov Model analysis of the smFRET trajectories that the pathways traversed by ADK depend on the concentration of GdmCl. Both sequential (state i is connected to its neighbor $i \pm 1$) as well as non-sequential transitions, leading to parallel folding routes, are found during the folding process. The findings for ADK in experiments are qualitatively reflected in the simulated folding pathways of the smaller single domain PDZ domain. In particular, at the three concentrations of GdmCl we find both sequential and non-sequential connectivities. Just like the ADK study, we also find that a minor population of unfolded states directly reach the NBA, as predicted by the kinetic partitioning mechanism (KPM).[45] Surprisingly, our simulations for PDZ show that KPM occurs only at 0.5M and 1.0M GdmCl but not in the absence of the denaturant. Overall, our simulations provide support to the discovery by Haran and coworkers[28] that fluxes through parallel folding pathways could be altered by changing the denaturant concentration.

It is interesting to compare this key finding[28] to the reports of parallel unfolding routes in I27 induced by denaturants[46] and SH3 domain by mechanical force (F).[47] In both these studies the $[C]$ or F dependence of $\ln k_u$ (k_u is the unfolding rate) exhibited upward curvature. Recently, we showed using theory that as long as the perturbation of the protein is linear in the external field (GdmCl or F) then upward curvature in the $[\ln k_u, [C]$ or $F]$ plot implies parallel unfolding pathways.[48] In light of these observations and the present study, it would be most instructive to use F to probe the folding and unfolding of the PDZ domain. Such experiments would clarify if denaturant-induced intermediates coincide with those found under tension, which should be the case if the perturbation is linear in $[C]$ and F. More

importantly, single molecule pulling experiments would provide insights into the nature of the TSEs and the modulation of fluxes through distinct pathways. Based on the free energy profiles in Fig. 3 we predict that $[\ln k_u, F]$ plot ought to reveal upward curvature. If so there are two possible distinct scenarios. In the first case, we expect that the free energy profile could be described by an effective one-dimensional reaction coordinate with an outer barrier dominating at low forces and an inner barrier becoming important at high forces.[49,50] The more interesting case is that the location of the transition state in terms of the molecular extension conjugate to the applied force is an increasing function of force just as found for the unfolding of the SH3 domain. Our finding based on the free energy profiles in Fig. 3 shows that TSE1 is dominant at low $[C]$ suggests that force-induced unfolding of the PDZ domain would also occur by parallel pathways. This prediction awaits single molecule pulling experiments.

Acknowledgement

ZL acknowledges financial support from the National Natural Science Foundation of China (11104015,11675017,11735005). DT is grateful to the National Science Foundation (CHE 16-36424), the National Institutes of Health (R01 GM089685), and the Collie-Welch Regents chair (F0019) for supporting this work.

References

(1) Thirumalai, D.; Liu, Z.; O'Brien, E. P.; Reddy, G. Protein folding: from theory to practice. *Curr. Opin. Struct. Biol.* **2013**, *23*, 22–29.

(2) Wallqvist, A.; Covell, D. G.; Thirumalai, D. Hydrophobic Interactions in Aqueous Urea Solutions with Implications for the Mechanism of Protein Denaturation. *J. Am. Chem. Soc.* **1998**, *120*, 427?428.

(3) Bennion, B.; Daggett, V. The molecular basis for the chemical denaturation of proteins by urea. *Proc. Natl. Acad. Sci.* **2003**, *100*, 5142–5147.

(4) O'Brien, E.; Dima, R.; Brooks, B.; Thirumalai, D. Interactions between hydrophobic and ionic solutes in aqueous guanidinium chloride and urea solutions: lessons for protein denaturation mechanism. *J. Am. Chem. Soc.* **2007**, *129*, 7346–7353.

(5) Canchi, D. R.; Paschek, D.; Garcia, A. E. Equilibrium Study of Protein Denaturation by Urea. *J. Am. Chem. Soc.* **2010**, *132*, 2338–2344.

(6) Hua, L.; Zhou, R.; Thirumalai, D.; Berne, B. Urea denaturation by stronger dispersion interactions with proteins t han water implies a 2-stage unfolding. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 16928–16933.

(7) Reddy, G.; Liu, Z.; Thirumalai, D. Denaturant-dependent folding of GFP. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109*, 17832–17838.

(8) O'Brien, E. P.; Ziv, G.; Haran, G.; Brooks, B. R.; Thirumalai, D. Effects of denaturants and osmolytes on proteins are accurately predicted by the molecular transfer model. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 13403–13408.

(9) Liu, Z. X.; Reddy, G.; O'Brien, E. P.; Thirumalai, D. Collapse kinetics and chevron plots from simulations of denaturant-dependent folding of globular proteins. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 7787–7792.

(10) Liu, Z. X.; Reddy, G.; Thirumalai, D. Theory of the Molecular Transfer Model for Proteins with Applications to the Folding of the src-SH3 Domain. *J. Phys. Chem. B* **2012**, *116*, 6707–6716.

(11) Maity, H.; Reddy, G. Folding of Protein L with Implications for Collapse in the Denatured State Ensemble. *J. Am. Chem. Soc.* **2016**, *138*, 2609–2616.

(12) Reddy, G.; Thirumalai, D. Collapse Precedes Folding in Denaturant-Dependent Assembly of Ubiquitin. *J. Phys. Chem. B* **2017**, *121*, 995–1009.

(13) Maity, H.; Reddy, G. Thermodynamics and Kinetics of Single-Chain Monellin Folding with Structural Insights into Specific Collapse in the Denatured State Ensemble. *J. Mol. Biol.* **2017**, *429*, xxx–xxx.

(14) Onuchic, J. N.; Wolynes, P. G. Theory of protein folding. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70–75.

(15) Thirumalai, D.; O'Brien, E. P.; Morrison, G.; Hyeon, C. Theoretical Perspectives on Protein Folding. *Annu. Rev. Biophys.* **2010**, *39*, 159–183.

(16) Dill, K.; Chan, H. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **1997**, *4*, 10–19.

(17) Dill, K. A.; MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science* **2012**, *338*, 1042–1046.

(18) Doyle, D. A.; Lee, A.; Lewis, J.; Kim, E.; Sheng, M.; MacKinnon, R. Crystal structures of a complexed and peptide-free membrane protein-binding domain: Molecular basis of peptide recognition by PDZ. *Cell* **1996**, *85*, 1067–1076.

(19) Harris, B. Z.; Lim, W. A. Mechanism and role of PDZ domains in signaling complex assembly. *J. Cell Sci.* **2001**, *114*, 3219–3231.

(20) Kim, E. J.; Sheng, M. PDZ domain proteins of synapses. *Nat. Rev. Neurosci.* **2004**, *5*, 771–781.

(21) Jemth, P.; Gianni, S. PDZ domains: Folding and binding. *Biochemistry* **2007**, *46*, 8701–8708.

(22) Feng, H. Q.; Vu, N. D.; Bai, Y. W. Detection of a hidden folding intermediate of the third domain of PDZ. *J. Mol. Biol.* **2005**, *346*, 345–353.

(23) Gianni, S.; Engstrom, A.; Larsson, M.; Calosci, N.; Malatesta, F.; Eklund, L.; Ngang, C. C.; Travaglini-Allocatelli, C.; Jemth, P. The kinetics of PDZ domain-ligand interactions and implications for the binding mechanism. *J. Biol. Chem.* **2005**, *280*, 34805–34812.

(24) Chi, C. N.; Gianni, S.; Calosci, N.; Travaglini-Allocatelli, C.; Engstrom, A.; Jemth, P. A conserved folding mechanism for PDZ domains. *FEBS Lett.* **2007**, *581*, 1109–1113.

(25) Calosci, N.; Chi, C. N.; Richter, B.; Camilloni, C.; Engstrom, A.; Eklund, L.; Travaglini-Allocatelli, C.; Gianni, S.; Vendruscolo, M.; Jemth, P. Comparison of successive transition states for folding reveals alternative early folding pathways of two homologous proteins. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 19241–19246.

(26) Murciano-Calles, J.; Cobos, E. S.; Mateo, P. L.; Camara-Artigas, A.; Martinez, J. C. An Oligomeric Equilibrium Intermediate as the Precursory Nucleus of Globular and Fibrillar Supramacromolecular Assemblies in a PDZ Domain. *Biophys. J.* **2010**, *99*, 263–272.

(27) Otzen, D.; Oliveberg, M. Salt-induced detour through compact regions of the protein folding landscape. *Proc. Natl. Acad. Sci.* **1999**, *96*, 11746–11751.

(28) Pirchi, M.; Ziv, G.; Riven, I.; Cohen, S. S.; Zohar, N.; Barak, Y.; Haran, G. Single-

molecule fluorescence spectroscopy maps the folding landscape of a large protein. *Nat. Commun.* **2011**, *2*, 493.

(29) Hyeon, C.; Thirumalai, D. Capturing the essence of folding and functions of biomolecules using coarse-grained models. *Nat. Commun.* **2011**, *2*, 487.

(30) Whitford, P. C.; Sanbonmatsu, K. Y.; Onuchic, J. N. Biomolecular dynamics: order–disorder transitions and energy landscapes. *Rep. Prog. Phys* **2012**, *75*, 076601.

(31) Pincus, D. L.; Cho, S. S.; Hyeon, C.; Thirumalai, D. Minimal Models for Proteins and RNA: From Folding to Function. *Progress in Molecular Biology and Translationanl Science* **2008**, *84*, 203–250.

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

(33) Zhou, R. H.; Berne, B. J.; Germain, R. The free energy landscape for beta hairpin folding in explicit water. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 14931–14936.

(34) Sanbonmatsu, K. Y.; Garcia, A. E. Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins* **2002**, *46*, 225–234.

(35) Honeycutt, J. D.; Thirumalai, D. The Nature of Folded States of Globular-Proteins. *Biopolymers* **1992**, *32*, 695–709.

(36) Veitshans, T.; Klimov, D.; Thirumalai, D. Protein folding kinetics: Timescales, pathways and energy landscapes in terms of sequence-dependent properties. *Fold. Des.* **1997**, *2*, 1–22.

(37) Ermak, D. L.; McCammon, J. A. Brownian dynamics with hydrodynamic interactions. *J. Chem. Phys.* **1978**, *69*, 1352–1369.

(38) Liu, Z.; Reddy, G.; Thirumalai, D. Folding PDZ2 Domain Using the Molecular Transfer Model. *J. Phys. Chem. B* **2016**, *120*, 8090–8101.

(39) Noe, F.; Schutte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 19011–19016.

(40) Li, W. F.; Terakawa, T.; Wang, W.; Takada, S. Energy landscape and multiroute folding of topologically complex proteins adenylate kinase and 2ouf-knot. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 17789–17794.

(41) Reddy, G.; Liu, Z. X.; Thirumalai, D. Denaturant-dependent folding of GFP. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 17832 –17838.

(42) Cárdenas, A. E.; Elber, R. Kinetics of cytochrome C folding: Atomically detailed simulations. *Proteins: Structure, Function, and Genetics* **2003**, *51*, 245–257.

(43) Camacho, C.; Thirumalai, D. Modeling the role of disulfide bonds in protein folding: Entropic barriers and pathways. *Proteins: Structure, Function, and Bioinformatics* **1995**, *22*, 27–40.

(44) Klimov, D. K.; Thirumalai, D. Multiple protein folding nuclei and the transition state ensemble in two-state proteins. *Proteins - Struct. Funct. Gene.* **2001**, *43*, 465–475.

(45) Guo, Z.; Thirumalai, D. Kinetics of Protein Folding: Nucleation Mechanism, Time Scales, and Pathways. *Biopolymers* **1995**, *36*, 83–102.

(46) Wright, C.; Lindorff-Larsen, K.; Randles, L.; Clarke, J. Parallel protein-unfolding pathways revealed and mapped. *Nat. Struct. Biol.* **2003**, *10*, 658–662.

(47) Jagannathan, B.; Elms, P. J.; Bustamante, C.; Marqusee, S. Direct observation of a force-induced switch in the anisotropic mechanical unfolding pathway of a protein. *Proc. Natl. Acad. Sci.* **2012**, *109*, 17820–17825.

(48) Zhuravlev, P. I.; Hinczewski, M.; Chakrabarti, S.; Marqusee, S.; Thirumalai, D. Force-dependent switch in protein unfolding pathways and transition-state movements. *Proc. Natl. Acad. Sci.* **2016**, *113*, E715–E724.

(49) Merkel, R.; Nassoy, P.; Leung, A.; Ritchie, K.; Evans, E. Energy landscapes of receptor-ligand bonds explored with dynamic force spectroscopy. *Nature* **1999**, *397*, 50–53.

(50) Hyeon, C.; Thirumalai, D. Multiple barriers in forced rupture of protein complexes. *J. Chem. Phys.* **2012**, *137*, 055103.

Figure Captions

Fig-1: (A) Cartoon representation of PDZ3 domain (PDB code: 1BFE). The sequence used in the simulations is given below using a one letter code for amino acids. (B) Heat capacity for different values of the GdmCl concentration, $[C]$. The values of $[C]$ measured in M from right to left are 0, 1, 3, 5 and 7. (C) The dependence of the melting temperatures, $T_m[C]$s on $[C]$.

Fig-2: (A) Free energy profile at the melting temperature T_m, as a function of the χ, structural oder parameter. χ_c^N and χ_c^D identify the boundaries between the states. The locations of the TSE1 between the NBA and IBA and TSE2 separating the IBA and UBA are labelled. (B) Fraction of molecules in the NBA (black, right scale), UBA (red, right scale) and IBA (green, right scale $\times 0.025$) as a function of GdmCl concentration, $[C]$. For comparison, experimentally monitored maximum wavelength of the fluorescence at different concentrations of GdmCl (blue, left scale) is also shown. Although it is not directly equal to f_{UBA}, it is correlated to f_{UBA}. (C) The $[C]$ dependence of the free energy of stability of the native state with respect to the unfolded state, ΔG_{NU}.

Fig-3: Free energy profiles as functions of χ at different $[C]$. (A) $T = T_m$. (B) $T = T_s$. The values of $[C]$ measured in molar units from top to bottom are 0, 1, 2, 3, 3.05, 4, 5, 6, 7, 8.

Fig-4: Transition state ensembles. (A) Distribution $P(\Delta_R)$ of the $\Delta_R = (\Delta_U - \Delta_{TSE})/(\Delta_U - \Delta_N)$, which is the fraction of buried solvent accessible surface area relative to the unfoled structures. The average $< \Delta_R >= 0.41$ for TSE1 and $< \Delta_R >= 0.74$ for TSE2. Snapshot of TSE1(TSE2) is shown on the left(right). (B) Contact maps of the native state ensemble (upper left) and the one for the TSE1 (lower right). (C) The same as B except for the TSE2.

Fig-5: Comparison of chevron plots obtained from simulations and experiment. The scale for the experimental plot for $\ln k_{obs}$ is on the left, and for the simulation is on the right.

Fig-6: Representative folding trajectories for the four folding pathways. The distinct routes to the folded states are given in terms of the time-dependent changes in χ. (A) P1, $D \rightarrow KIN_1 \rightarrow KIN_2 \rightarrow N$. (B) P2, $D \rightarrow KIN_2 \rightarrow N$. (C) P3, $D \rightarrow KIN_1 \rightarrow N$. (D) P4, $D \rightarrow N$.

Fig-7: Schematic of the folding pathways and the structural features of the populated states. (A) Folding pathways and fraction of native contacts of all states at $[C] = 0.0M$. Color code: Red, unstructured; Blue: fully structured. The three folding pathways that reach the native state are represented by colored arrows with the their widths representing the flux through the pathways. Different colors represent different folding pathways. (B, C) Same as A but at $[C] = 0.5M$ (B) and $1.0M$ (C) with the states represented by circles.

Fig-8: Correlation between equilibrium I_{EQ} and kinetic intermediates, KIN_1 and KIN_2 using the average residue-resolved fractions of native contacts, f_Qs. (A) Plot of f_Q between KIN_1 and I_{EQ}. The heterogeneous states populated during the early stages of folding results in little structural correlation between these states.Same as (A) except this plot shows correlation between KIN_2 and I_{EQ}. At late stages (post collapse) the structures of the intermediates at equilibrium and during kinetics coincide. (C) Sample structures for KIN_1, I_{EQ}, and KIN_2 are displayed. Red represents structures not present in the native state and blue is native-like.

Fig-9: Table of Contents (TOC) figure.

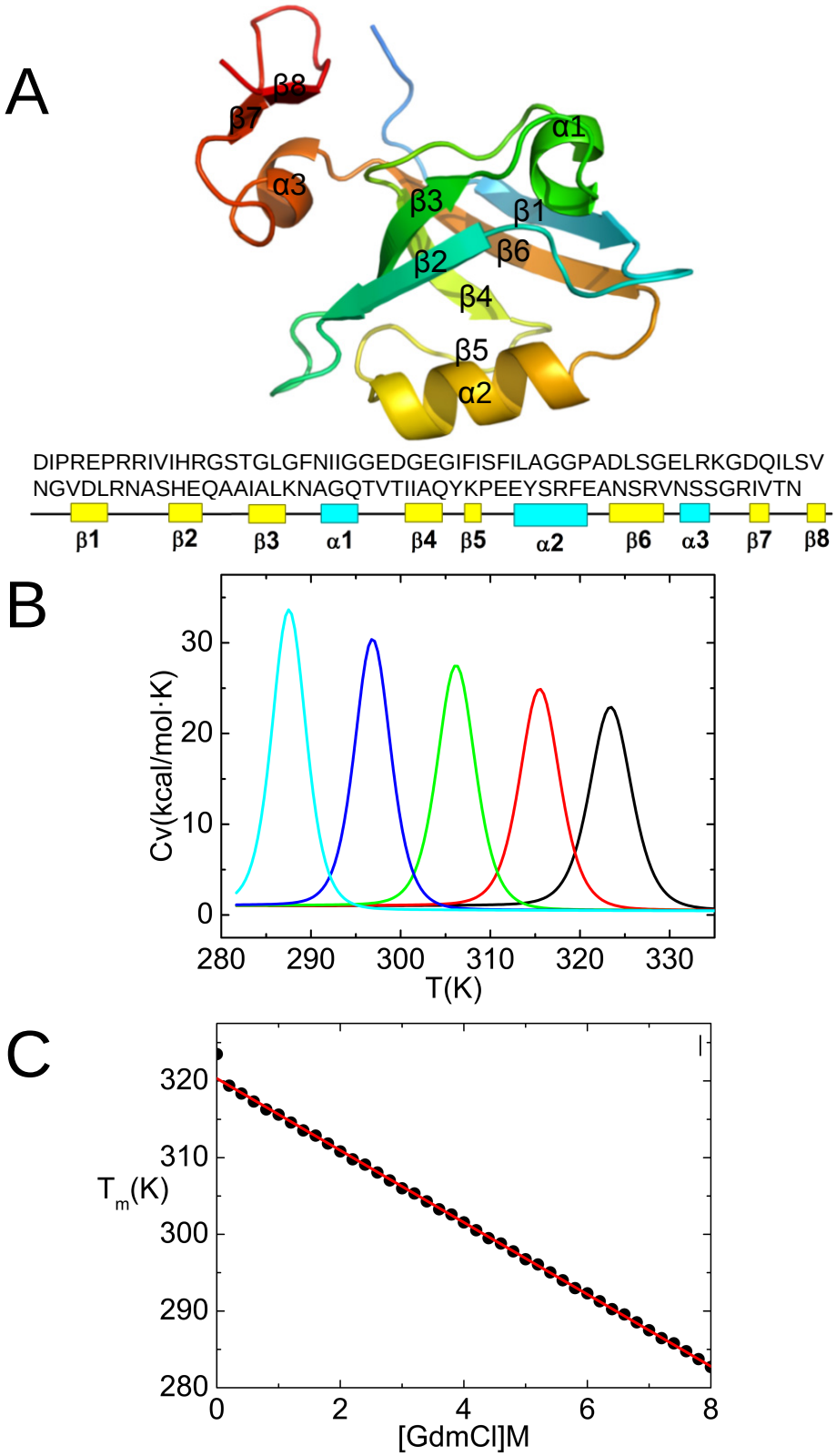
A
β8
β7
α3
α1
β3
β1
β6
β2
β4
β5
α2
DIPREPRRIVIHRGSTGLGFNIIGGEDGEGIFISFILAGGPADLSGELRKGDQILSV
NGVDLRNASHEQAAIALKNAGQTVTIIAQYKPEEYSRFEANSRVNSSGRIVTN
β1 β2 β3 α1 β4 β5 α2 β6 α3 β7 β8
B
Cv(kcal/mol·K)
T(K)
C
T_m(K)
[GdmCl]M

Figure 1

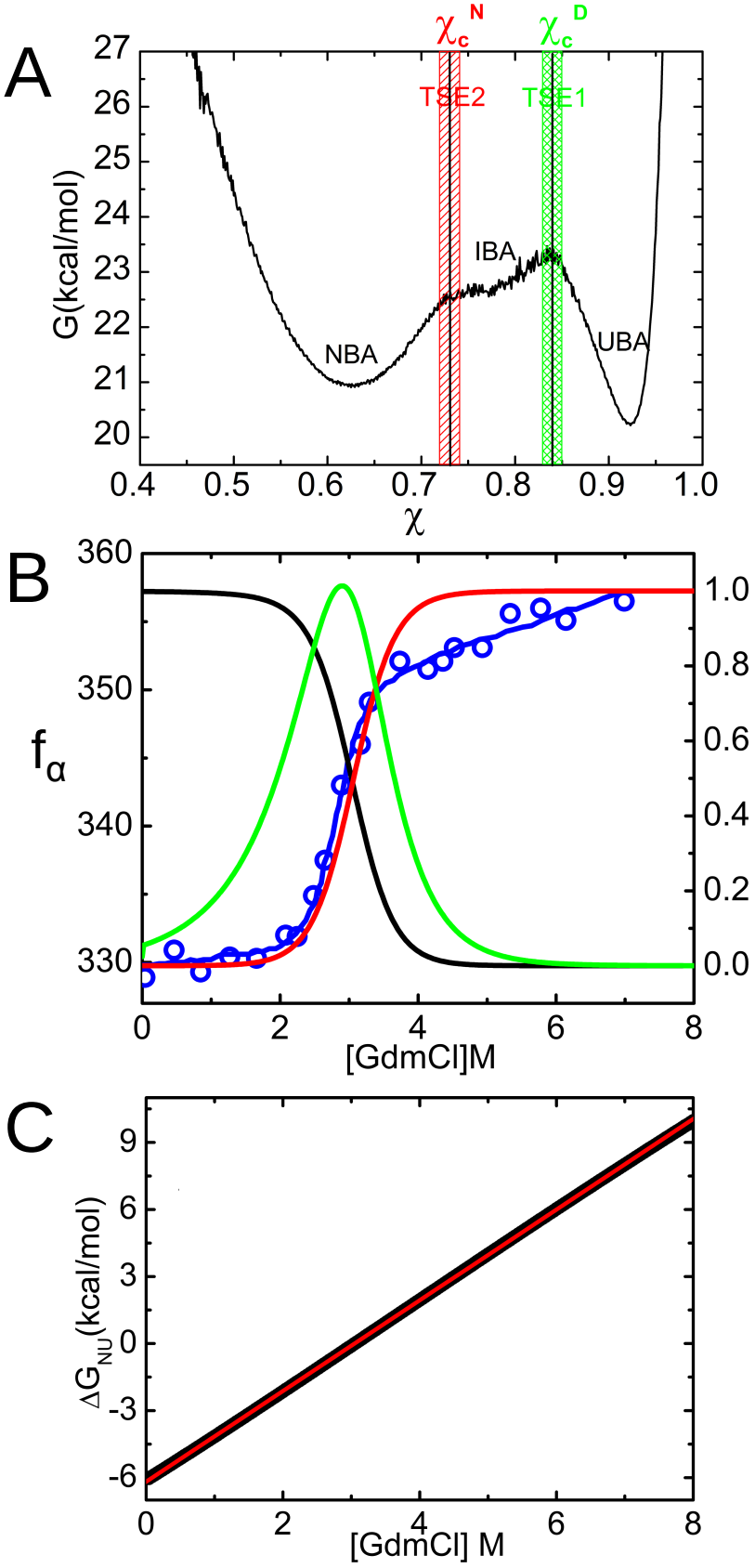
A
χ_c^N
χ_c^D
TSE2
TSE1
IBA
NBA
UBA
G(kcal/mol)
χ
B
f_α
[GdmCl]M
C
ΔG_{NU}(kcal/mol)
[GdmCl] M

Figure 2

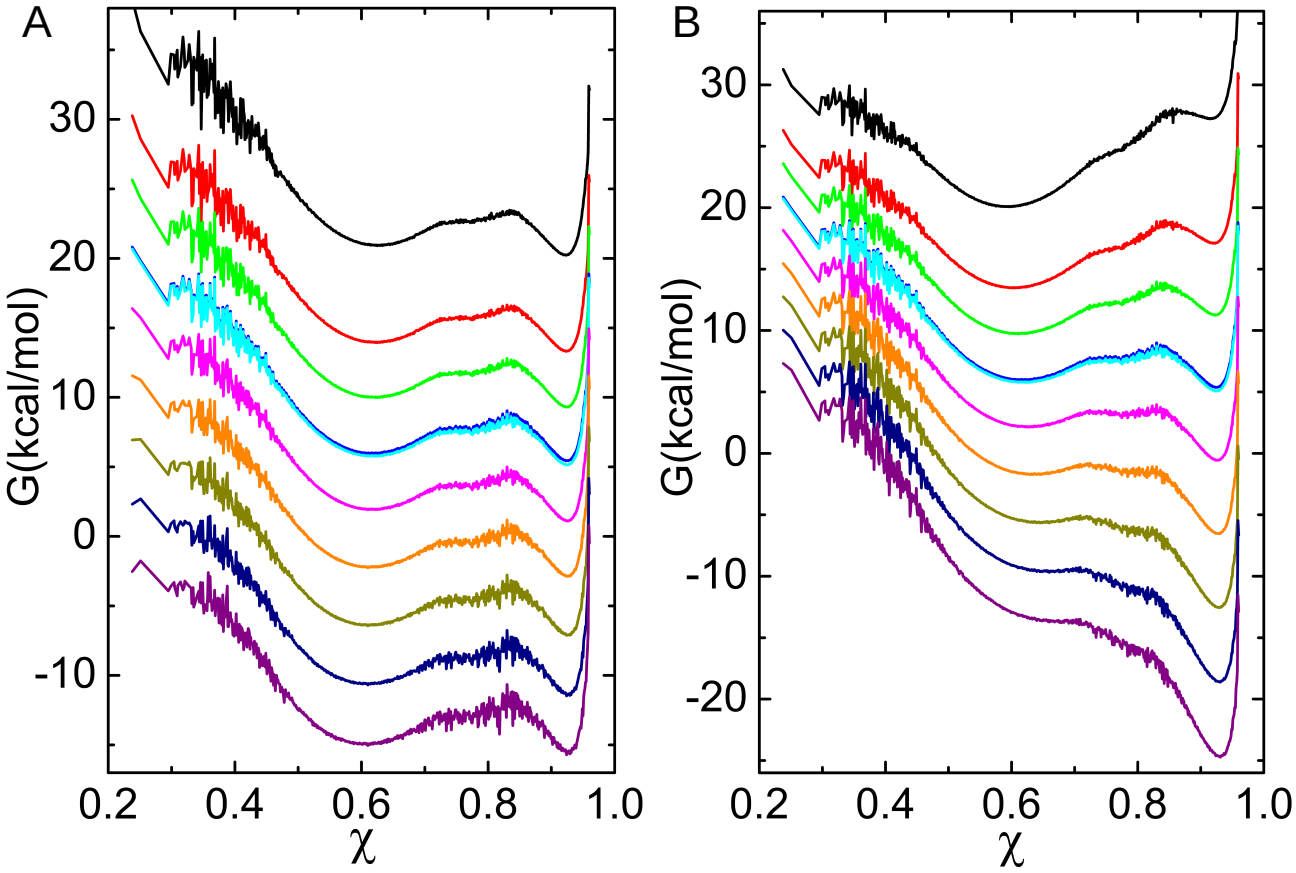

Figure 3

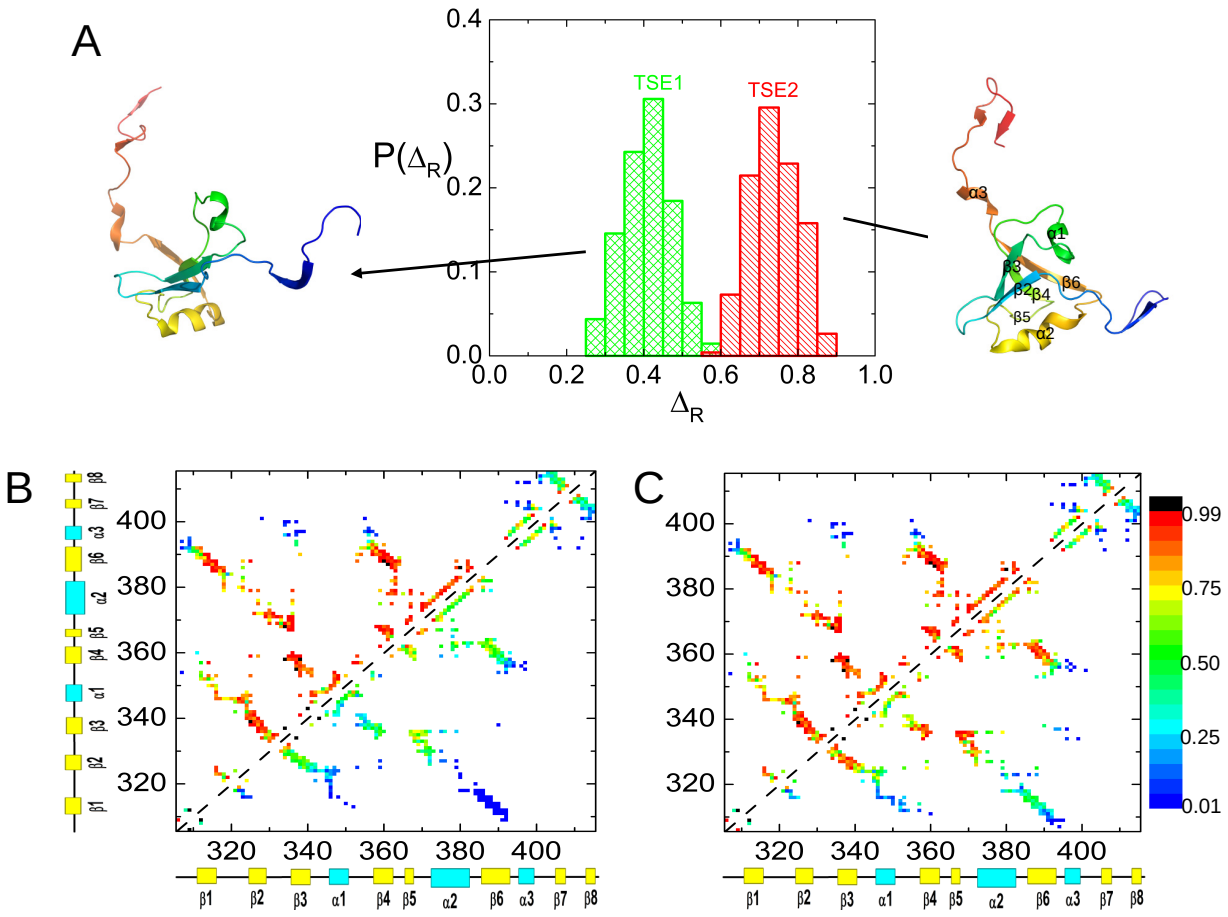
A
TSE1
TSE2
P(Δ_R)
Δ_R
0.0
0.1
0.2
0.3
0.4
0.0
0.2
0.4
0.6
0.8
1.0
α3
α1
β3
β2
β4
β6
β5
α2
B
C
320
340
360
380
400
β1
β2
β3
α1
β4
β5
α2
β6
α3
β7
β8
0.99
0.75
0.50
0.25
0.01

Figure 4

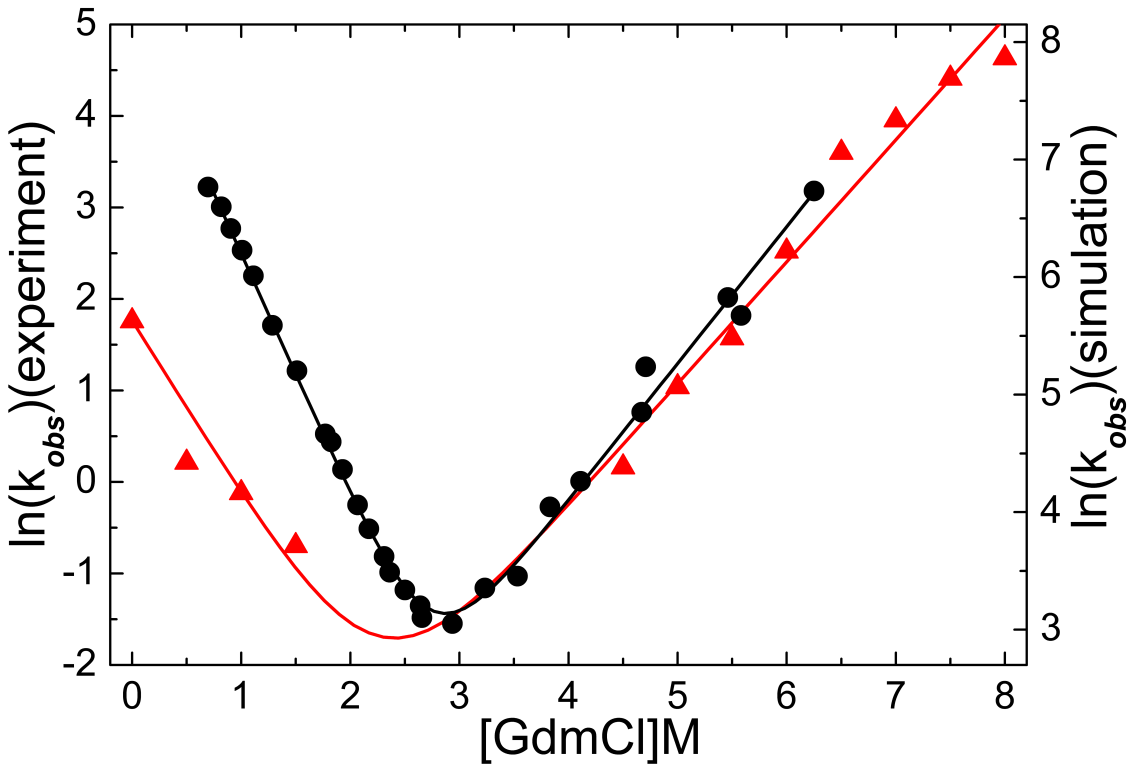
ln(k_obs)(experiment)
ln(k_obs)(simulation)
[GdmCl]M
5
4
3
2
1
0
-1
-2
8
7
6
5
4
3
0
1
2
3
4
5
6
7
8

Figure 5

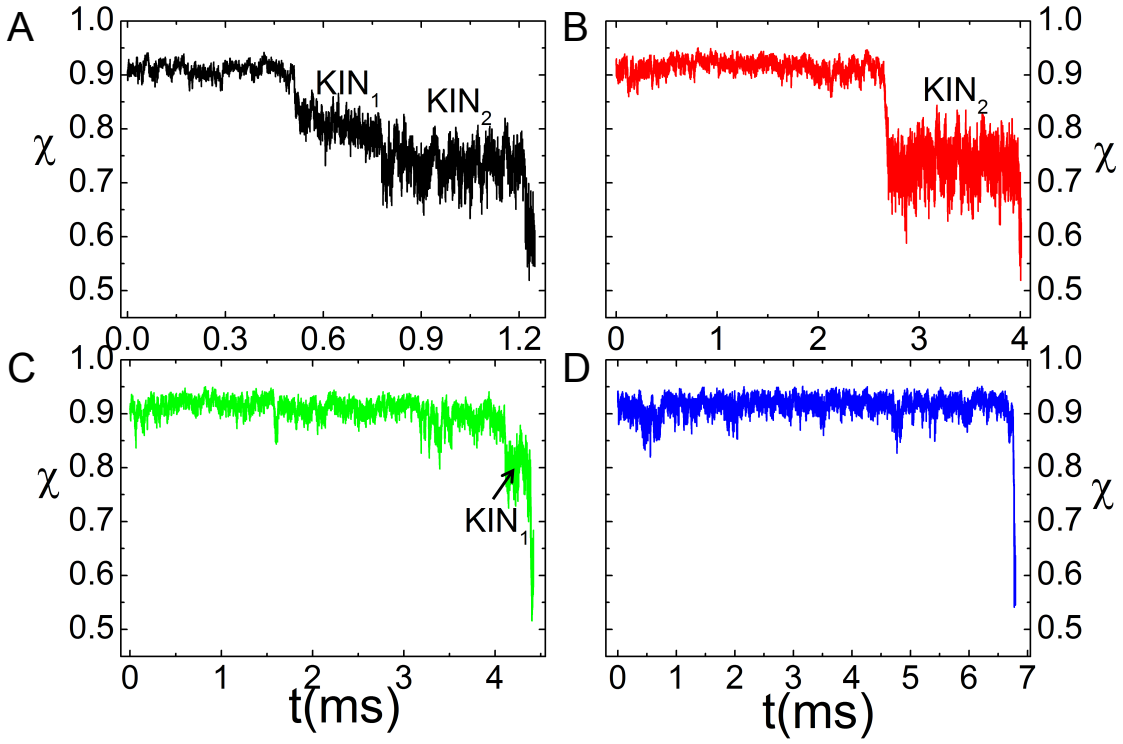
A
B
C
D
KIN1
KIN2
KIN2
KIN1
χ
χ
χ
χ
1.0
0.9
0.8
0.7
0.6
0.5
0.0
0.3
0.6
0.9
1.2
0
1
2
3
4
5
6
7
t(ms)
t(ms)

Figure 6

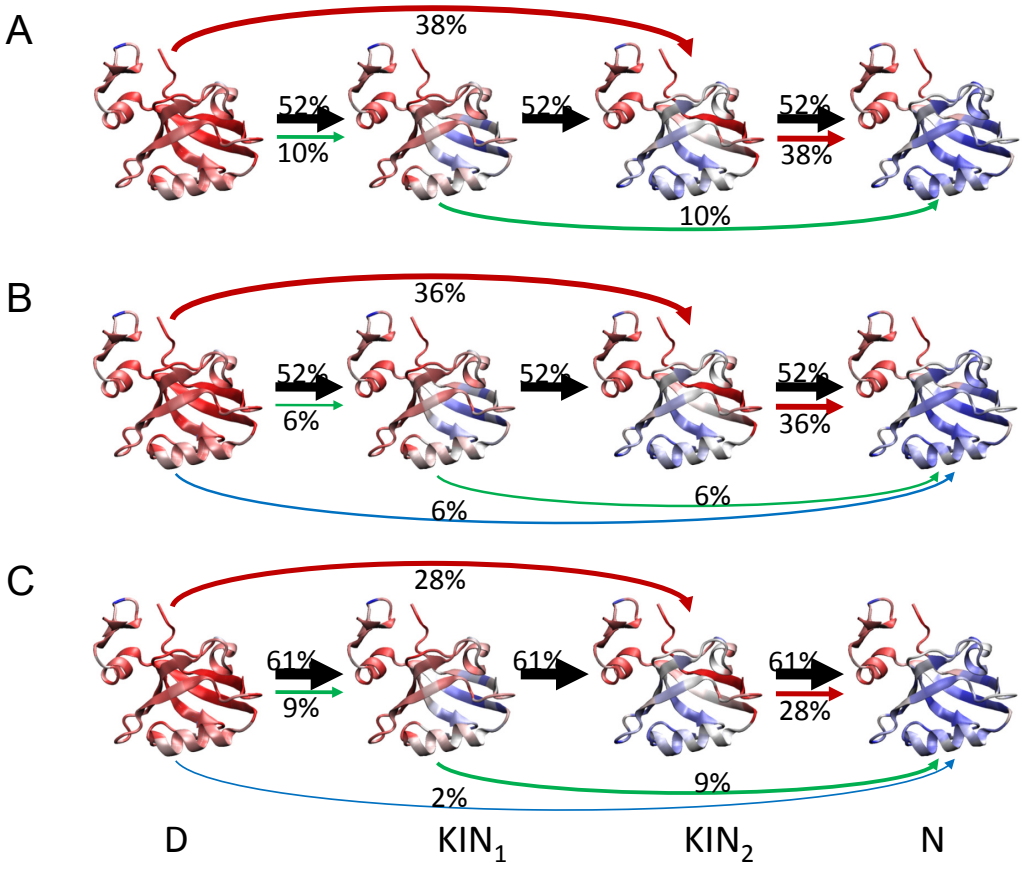
A
38%
52%
10%
52%
52%
38%
10%
B
36%
52%
6%
52%
52%
36%
6%
6%
C
28%
61%
9%
61%
61%
28%
9%
2%
D
KIN1
KIN2
N

Figure 7

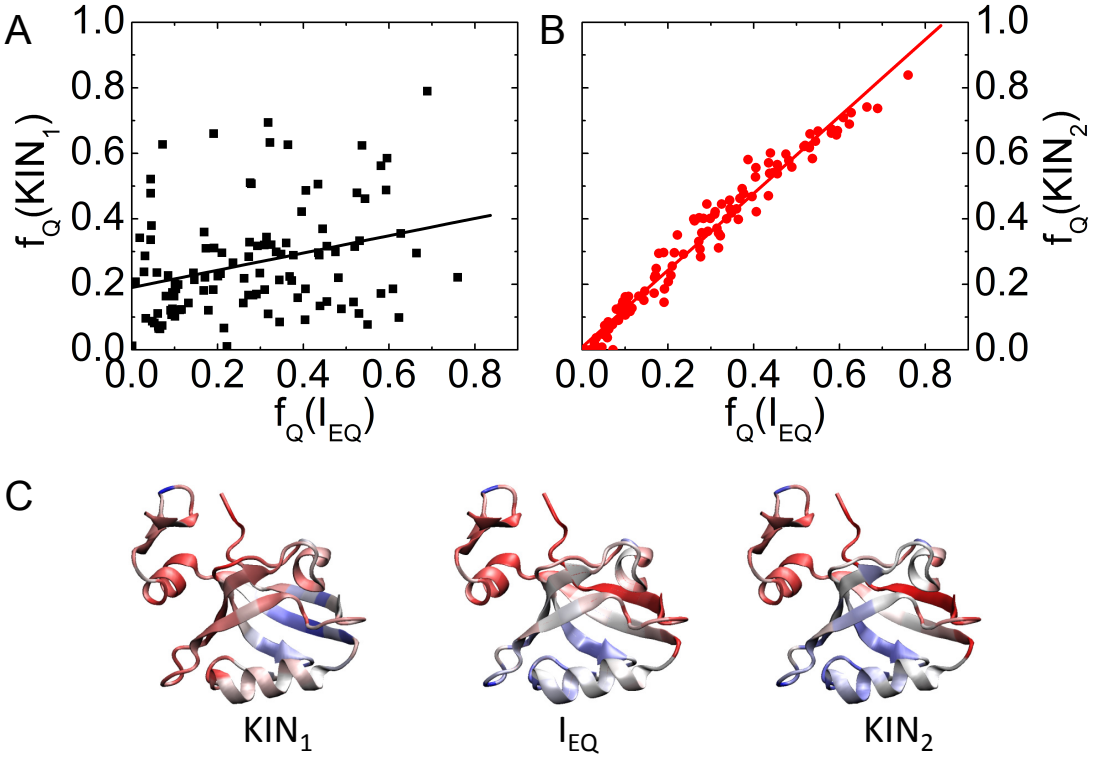
A
B
C
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
$f_Q(KIN_1)$
$f_Q(KIN_2)$
$f_Q(I_{EQ})$
KIN_1
I_{EQ}
KIN_2

Figure 8

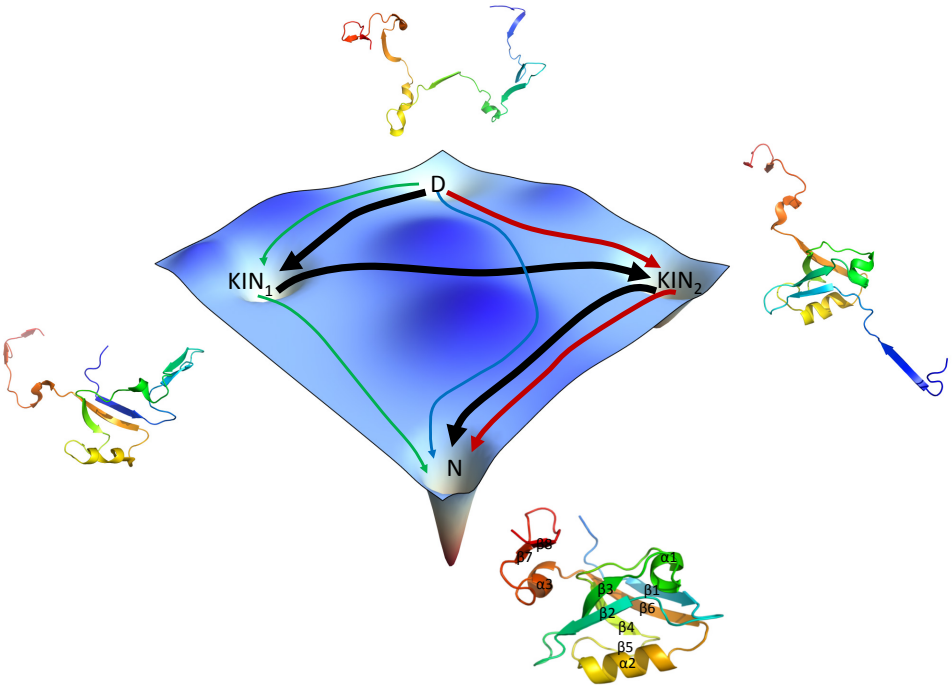

Figure 9