Chemical Physics Letters 471 (2009) 310–314

Contents lists available at ScienceDirect

Chemical Physics Letters

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

Analysis of single molecule folding studies with replica correlation functions

Peter Lenz [a,*], Samuel S. Cho [b,1], Peter G. Wolynes [b]

[a] Fachbereich Physik, Philipps-Universität Marburg, D-35032 Marburg, Germany
[b] Center for Theoretical Biological Physics and Department of Chemistry and Biochemistry, University of California at San Diego, La Jolla, CA 92093, United States

ARTICLE INFO

Article history:
Received 13 November 2008
In final form 18 February 2009
Available online 21 February 2009

ABSTRACT

Single molecule experiments that can track individual trajectories of biomolecular processes provide a challenge for understanding how these stochastic trajectories relate to the global energy landscape. Using trajectories from a native structure based simulation, we use order parameters that accurately distinguish between protein folding mechanisms that involve a simple, single set of pathways versus a complex one with multiple sets of competing pathways. We show how the folding dynamics can be analyzed with replica correlation functions in a way that is compatible with single molecule experiments.

It is now well established that proteins fold via an ensemble of states in which the energy landscape is globally funneled or directed towards a structurally well-defined native state [1–4]. At the level of individual trajectories of protein folding, one can envisage multiple pathways from the denatured ensemble descending down a funneled energy landscape until they join up as they approach the native state. For proteins in the laboratory the nature of the discrete trajectories of protein molecules and the degree of similarity of one path of approach to another as the molecule goes to the folded state is still an open question.

Extraordinary progress has been made in the past few years in the development of experimental methods sensitive enough to study the dynamical properties of single molecules [5]. It is now possible to follow, in part, the trajectories of individual molecules to track the dynamical time-dependence of features of their conformation in space [6–9]. The random sequence of events during transitions of the folding of proteins and other biomolecules at the individual molecular level, intrinsically gives stochastic data. As such, statistical physics tools are needed to quantify the pathways traversed in a set of individual trajectories and map the statistical properties of such single molecule data to the topography of the energy landscape [10].

In this Letter, we demonstrate that statistical physical tools can richly quantify, in principle, the pathways from a collection of trajectories. Specifically, using trajectories from simulations of proteins from perfectly funneled energy landscapes, we show that appropriately computed replica correlation functions can diagnose whether folding occurs largely via a single set of pathways or whether several distinct sets of competing routes to the native state actually coexist. This is the first application of replica correlation functions to a concrete problem after their introduction in the context of protein folding in Ref. [11].

Ordinarily, partially folded protein conformations in simulations are characterized with the help of appropriately chosen order parameters measuring the similarity to the final native structure. On the other hand, replica order parameters measure the similarity between the different routes taken to the final product. To define such a quantity a measure in phase space is required to quantify how similar two microscopically distinct configurations are to each other. For proteins and random polymers, the overlap function between two conformations is a useful object because it correlates with pair interaction energies. The overlap between conformation i and j is then an explicit function $q^{ij} = \Phi(\{\mathbf{r}^{(i)}\}, \{\mathbf{r}^{(j)}\})$ of the atomic coordinates $\{\mathbf{r}^{(i)}\}$ and $\{\mathbf{r}^{(j)}\}$. On a somewhat coarse-grained scale, an appropriate choice for the function Φ is the fraction of contacts between specific residues in the macromolecule which occur in both configurations. When one of the configurations is the native state (with atomic coordinates $\{\mathbf{r}^{(f)}\}$) then $Q_i = \Phi(\{\mathbf{r}^{(i)}\}, \{\mathbf{r}^{(f)}\})$ is simply the fraction of native-like tertiary contacts of conformation i.

This global quantity q^{ij} determines the overall shape similarity, but would be difficult to directly measure. Experimentally, the corresponding pair specific quantity is accessible by means of fluorescence quenching studies. If residue m contains an energy donor that can be excited, while residue n has a chromophore to act as an energy acceptor, then the time-dependence of the mn contact can be experimentally resolved. If this contact is made in configuration i and in configuration j of another trajectory then $q^{ij}_{mn} \neq 0$. The global overlap is then given by the sum over all possible contacts, i.e., $q^{ij} = \sum_{m,n} q^{ij}_{mn}$.

With this reaction coordinate the replica correlation function between two paths α and β (which both fold into the native state) can be defined as [11]

* Corresponding author.
E-mail address: peter.lenz@physik.uni-marburg.de (P. Lenz).

[1] Present Address: Institute for Physical Sciences and Technology, Department of Chemistry and Biochemistry, University of Maryland, College Park, MD 20742, United States.

0009-2614/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.cplett.2009.02.054

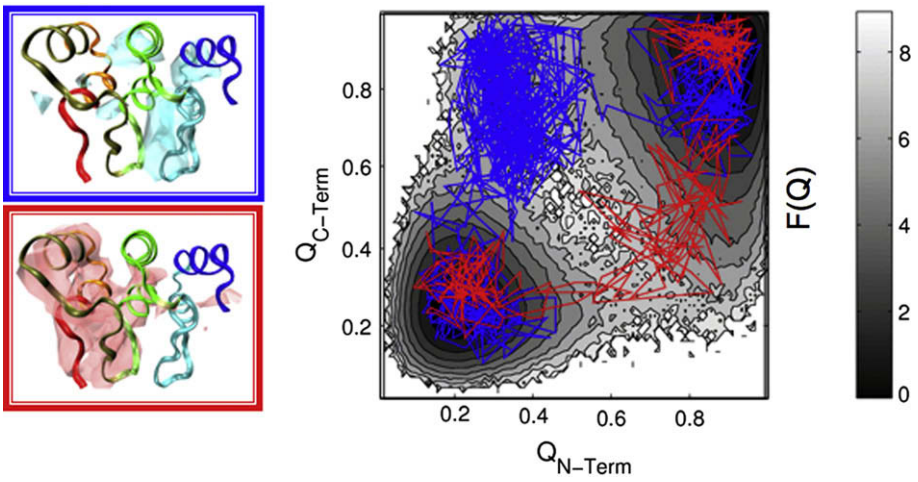

Fig. 1. Protein folding with multiple sets of pathways. (Left) A ribbon diagram is shown of 1N0Q with clouds representing the folded ensemble of the pathway where the c-terminal (top) or n-terminal (bottom) folds first. (Right) A free energy profile is projected to the fraction of native contacts of the n-terminal [Q(n-term)] and the c-terminal [Q(c-term)] with two example trajectories overlaid.

$$q^{\alpha\beta}(t_{p1}, t_{p2}, t_{l1}, t_{l2}) = \left\langle \Phi(\{\mathbf{r}_i^{\alpha}, t_{l1}\}, \{\mathbf{r}_i^{\beta}, t_{l2}\})\theta(N^{\alpha} - N_t) \times \theta(N^{\beta} - N_t) j^{\alpha}(t_{p1} + t_{l1}) j^{\beta}(t_{p2} + t_{l2}) \right\rangle, \quad (1)$$

where $\langle \ldots \rangle$ denotes an ensemble average, $\theta(x)$ is Heaviside's function, t_p and t_l are the preparation times (i.e., the time elapsed before the observation is started) and lookback time, respectively, N_t is the value of the reaction coordinate at the transition state, and $j^{\alpha}(t)$ and $j^{\beta}(t)$ are the (normalized) flux in trajectory α and β into the product state at time t. Generally, special care is required in choosing the adequate reaction coordinate and associated transition state. However, for the purpose of this study it is sufficient to simply choose the transition state in such a way that once the system leaves the transition state (in direction of the product) it is committed to react [19,20]. Quantities similar to these replica correlation functions have been studied in the theory of disordered systems such as glasses and spin glasses [12–16].

The replica correlation provides a measure of how similar two trajectories are to each other at all times. However, to compare trajectories, one has to take into account that different trajectories traversing the same route might spend different amounts of time in the vicinity of a given region in phase space such as, e.g., the transition state. Even if the two trajectories lead to the folded state via the same intermediate conformations, the sequence of events takes different amounts of time. The Laplace-transform provides a natural means of comparing trajectories of different duration, and we use it here to quantify the similarity of our molecular trajectories. This general analysis reveals, in particular, the near equilibrium behavior (for $s \to 0$) and fast motions ($s \to \infty$).

The simulation data were obtained from a C_α native topology-based model, which is described in detail in Ref. [17]. Briefly, a single bead centered on the C_α position represents a residue and bond and angle potentials string together the beads to their neighbors along the protein chain. The dihedral potential encodes the secondary structures. The protein's native topology defines the network of favorable long-range tertiary interactions, while all other non-bonded interactions are repulsive.

The network of native contact pairs was determined using the CSU (Contacts of Structural Units) software [18]. Multiple trajectories with numerous unfolding/folding transitions were collected and analyzed using the weighted histogram analysis method (WHAM) to calculate the free energy surface projected onto the fraction of native contacts Q (defined as in Ref. [19]). The folding temperature (T_f) was identified as the peak of a specific heat versus temperature profile.

To analyze the transitions between the unfolded and folded states, we performed multiple constant temperature simulations ($T = T_f$) of the src-SH3 protein and the designed ankyrin repeat protein (PDB Codes: 1SRL and 1N0Q, respectively). Each constant temperature trajectory consists of multiple transitions between the unfolded and folded ensembles. The trajectories were then combined to calculate the free energy profiles with respect to Q (see Fig. 1 as an example for 1N0Q).

From long trajectories, we extracted for further analysis only those portions where the transitions between the unfolded and folded states occurred. The unfolded and folded states were chosen as the values of Q at which the free energy is $1k_BT$ above the appropriate free energy minimum. The Q values that demarcate the unfolded and folded ensembles for 1SRL are 0.26 and 0.81, respectively, while for 1N0Q the values are 0.15 and 0.81. Three hundred and ninety-nine trajectories for 1SRL which went through 40 790 different conformations and 126 trajectories for 1N0Q with 51 570 different conformations were used for analysis.

Previous studies indicate that 1SRL predominantly folds via a single correlated route, while 1N0Q, clearly possesses distinct sets of competing folding pathways [19]. One clear difference between these systems is the distribution of the folding times (i.e., number of steps required to reach the folded state for the first time starting from the unfolded one). These are shown in Fig. 2. While the distribution for 1SRL is unimodal, the one for 1N0Q is wide and skewed, consistent with two peaks (as expected for 1N0Q which has at least two different folding pathways with different folding times).

A protein with several distinct sets of folding pathways should encounter a more diverse ensemble of conformations along its folding trajectory. To see this, we have discretized the set of folding trajectories and analyzed the distribution of q_{ij} (defined as in Ref. [19]) for all states i and j which have a given Q (and thus a given similarity with the folded state) for different Q. In doing so, we projected all of the conformations onto $N = 30$ different states where state i (with $1 \leqslant i \leqslant 30$) represents all microscopic conformations having Q between $(i-1)/N$ and i/N. A complete folding trajectory is a sequence of microscopic conformations $\{\mathbf{r}(t)\}$, but in this discretization scheme it becomes a sequence $\{\rho(t)\}$ of integers. A coarse-grained folding pathway is a sequence of transitions between the $N = 30$ different discrete states starting at the unfolded state ($\rho(t = 0) = 1$) and ending in the native state ($\rho(t = t_f) = N$).

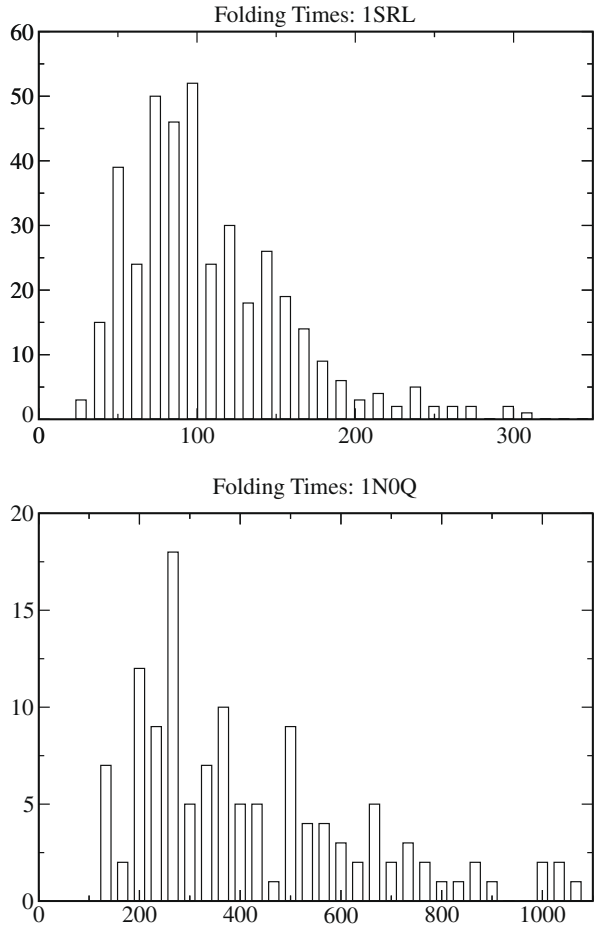

Fig. 2. Histogram of folding times for 1SRL (top) and 1N0Q (bottom).

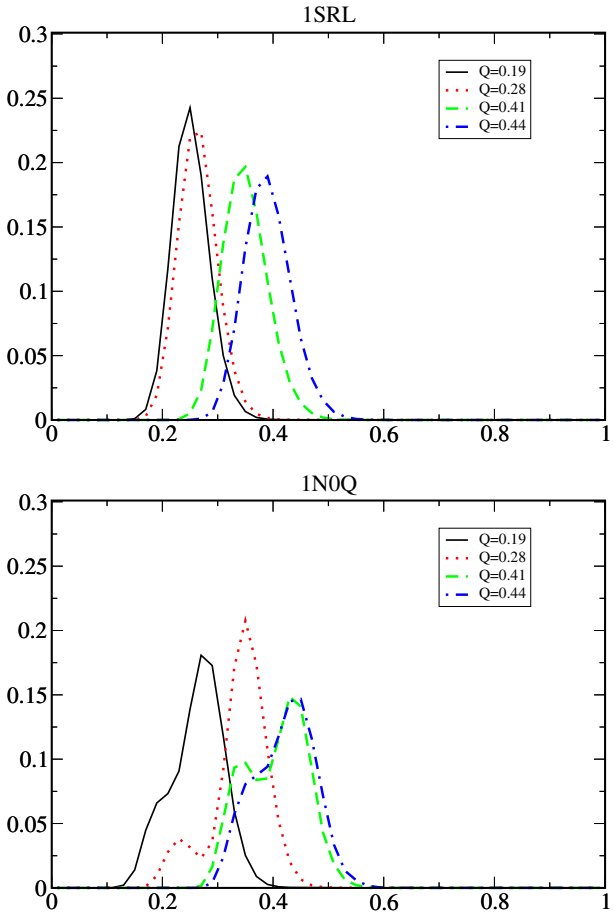

Fig. 3. The distribution of q as function of Q. Data shown are for bins of width 0.02 (continuously interpolated) centered around $Q = 0.19, 0.28, 0.41, 0.44$ for 1SRL (top) and for 1N0Q (bottom). For 1N0Q the distribution starts to be bimodal around $Q = 0.24$ (data not shown). For $Q > 0.44$ the distribution becomes unimodal. The q-distribution of 1SRL is unimodal for all $0.26 < Q < 0.81$.

The comparison of all N_Q conformations having a given Q requires $\sim N_Q^2$ numerical operations. We therefore restricted the analysis to a subset of the 1N0Q trajectories. Fig. 3 shows the distribution of q for four different Q values for all 399 1SRL trajectories and 34 randomly picked 1N0Q trajectories (out of the total 126 trajectories). Generally, we find that the distribution of q is unimodal for 1SRL for all Q, indicating that the conformations of the folding pathways are very similar. For 1N0Q however, the distribution of relative q is distinctly bimodal for the range $0.24 \leqslant Q \leqslant 0.44$.

To check how folding times influence the q-distribution, we can partition the folding trajectories into slow, medium, and fast folding pathways[2]. As shown in Fig. 4, for the fast trajectories the bimodality is more pronounced at smaller $Q(Q \simeq 0.28)$, while for slower trajectories the q-distribution becomes bimodal only at larger $Q(Q \simeq 0.44)$. All of the 1SRL trajectories (slow, fast and medium) are unimodal (data not shown).

The bimodality of the q-distribution implies that the two subsets of the protein conformations have very few common structural features. In the context of protein folding, the key event leading to the bifurcation into the two subsets occurs upon reaching the transition state ensemble. Thus, in Fig. 3 those conformations that have reached the transition state have large structural differences from those that have not yet reached the transition state, leading to the small q-values. The unimodal distribution for 1SRL and the bimodal distribution for 1N0Q agree with our expectations from their distinct pathways patterns. For 1N0Q, the fast and slow trajectories take different routes in configuration space. For the fast trajectories the transition state is reached for small $Q(Q \simeq 0.28)$, while the slow trajectories reach the transition state only for larger $Q(Q \simeq 0.44)$. Thus, the transition state is reached by conformations with significant structural differences implying that (at least) two different transition states exist that lead to different folding times.

As in the earlier analysis of pathways on random landscapes, we analyze the Laplace-transform of the replica correlation function [11]. The numerically calculated Laplace-transform $q^{\alpha\beta}$ of Eq. (1) is shown in Fig. 5 as function of the single Laplace variable s associated with the lookback time (for $t_{l1} = t_{l2}$). The Laplace variables s_{p1} and s_{p2} associated with the preparation times can be set to zero since our trajectories are well equilibrated. As shown in Fig. 5, for small $s(0 < s < 1)q^{\alpha\beta}$ decays algebraically for both 1SRL and 1N0Q. For larger $s > 1, q^{\alpha\beta}(s)$ decays exponentially (data not shown) as one expects for a set of data with discrete time steps (with the dominant contribution coming from the correlations between nearly denaturated states). Note, $q^{\alpha\beta}(s)$ is generally larger for 1SRL than for 1N0Q reflecting the fact that the conformations encountered during folding are generally more similar for 1SRL than for 1N0Q.

This characteristic s-dependence reflects the full dynamics of the folding transition (from the transition state ensemble to the native state). To illustrate this connection we now consider a simple protein folding model that describes the folding transition as a

[2] For 1N0Q (1SRL), fast trajectories find the folded states within 330 steps (100 steps) and slow ones need more than 660 steps (200 steps).

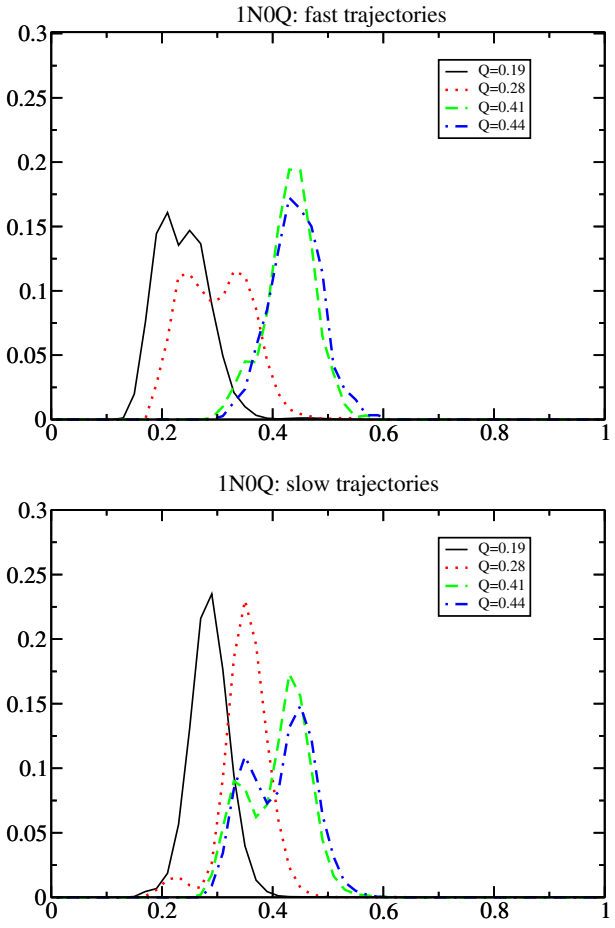

Fig. 4. Distribution of q for given Q for 1N0Q. Data are shown for fast trajectories (top) and slow trajectories (bottom). Fast trajectories find the folded states within 330 steps, slow ones need more the 660 steps. As one sees the q-distribution is sharper for the fast trajectories.

simple sequence of reactions between well-separated states. More specifically, we generalize the model introduced in Ref. [11] to a system with two different ensembles of transition states. The states of the reactant ensemble Ω_D reach the transition states with energy $E_{i,\alpha}$ of ensemble $\alpha(\alpha = 1, 2)$ with rates $k_{d,\alpha}$, the reverse reaction has a rate $k_{0,\alpha}e^{\beta E_{i,\alpha}}$, while the transition from ensemble α to the folded state occurs with rate $\kappa_{F,\alpha}$. Interconversion between the states of ensemble α occurs with rate $\omega_{0,\alpha}/\Omega^{\dagger}_{\alpha}e^{\beta E_{i,\alpha}}$, where $\Omega^{\dagger}_{\alpha}$ is the number of states in ensemble α.

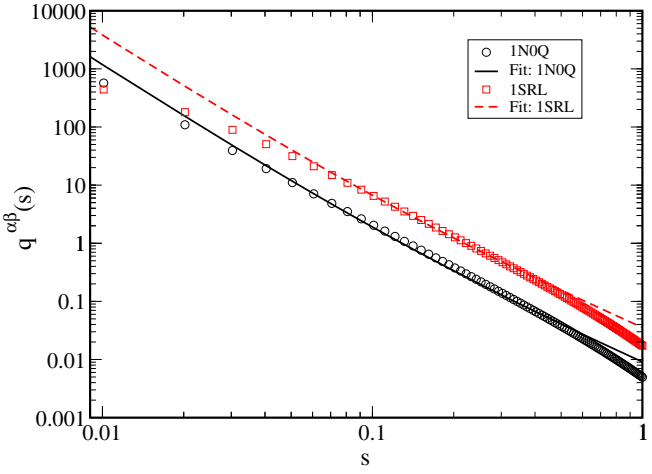

Fig. 5. The Laplace-transform of the replica correlation function $q^{\alpha\beta}(s)$ for 1N0Q (solid black) and 1SRL (dashed red) as determined numerically from the simulation data (time unit = 1 simulation step). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The occupation probabilities P_D (of the reactant states) and $P_{l,\alpha}$ (of transition state l of ensemble α) can be explicitly calculated

$$\widetilde{P}_D(s)\left(s+\sum_{\alpha}\Omega^{\dagger}_{\alpha}k_{d,\alpha}\right)=\sum_{\alpha}k_{0,\alpha}f_{i,\alpha}, \tag{2}$$

$$\sum_{l}\widetilde{P}_{l,\alpha}=Z^{\dagger}_{\alpha}k_{d,\alpha}\widetilde{P}_D(s)+\frac{1}{s+\omega_{i,\alpha}}+\frac{Z^{\dagger}_{\alpha}\omega_{0,\alpha}}{\Omega^{\dagger}_{\alpha}}f_{i,\alpha}, \tag{3}$$

where the Laplace-transformed quantities are denoted by $\tilde{P}$. Furthermore,with $\tilde{s}=s+\kappa_{F,\alpha}$

$$f_{i,\alpha}=\frac{k_{d,\alpha}\widetilde{P}_D(\Omega^{\dagger}_{\alpha}-Z^{\dagger}_{\alpha}\tilde{s})+1-\tilde{s}/(s+\omega_{i,\alpha})}{\tilde{s}Z^{\dagger}_{\alpha}\omega_{0,\alpha}/\Omega^{\dagger}_{\alpha}+k_{0,\alpha}}, \tag{4}$$

$$\omega_{i,\alpha}=\kappa_{F,\alpha}+(k_{0,\alpha}+\omega_{0,\alpha})e^{\beta E_{i,\alpha}}, \tag{5}$$

$$Z^{\dagger}_{\alpha}(s)=\sum_{i}\frac{1}{s+\omega_{i,\alpha}}. \tag{6}$$

The Laplace-transformed replica correlation function becomes

$$q^{\alpha\beta}(s)=\sum_{\alpha}\sum_{i}\kappa^2_{F,\alpha}\left(\sum_{k}\widetilde{P}^{(i)}_{k,\alpha}(s_l)\widetilde{P}^{eq}_{i,\alpha}(s_p)\right)^2, \tag{7}$$

where $\widetilde{P}^{(i)}_{k,\alpha}(s_l)$ is the (Laplace-transformed) occupation of state k in transition state ensemble α assuming that at $t_l = 0$ only state i is populated and $\widetilde{P}^{eq}_{i,\alpha}(s_p)$ is the occupation of state i of transition state ensemble α assuming that at $t_p = 0$ only the reactant ensemble is populated.

If the internal relaxation in the transition state ensembles can be neglected ($\omega_{0,\alpha} = 0$), then (in leading order)

$$q^{\alpha\beta}(s)\sim\sum_{\alpha}\sum_{i}\frac{1}{(s+\omega_{i,\alpha})^2}, \tag{8}$$

while for a system with forward reaction only ($k_{0,\alpha} = 0$)

$$q^{\alpha\beta}(s)\sim\sum_{\alpha}\frac{\Omega_{\alpha}}{(s+\kappa_{F,\alpha})^2}. \tag{9}$$

With the last 2 formulas, it is not possible to fit the numerical data of Fig. 5, which decays as $q^{\alpha\beta}\sim s^{-2.5}$ (data not shown). In Eqs. (8) and (9) the decay of $q^{\alpha\beta}$ with s is too slow and reasonable fits require that either $\omega_{i,\alpha} < 0$ or $\kappa_{F,\alpha} < 0$ (data not shown). The characteristic s-dependence of $q^{\alpha\beta}$ can only be explained if higher order corrections are taken into account in Eq. (8).

The data can be interpreted more directly by taking the dynamics of the folding transition explicitly into account. For this purpose, it is sufficient to describe the folding transition as a 1-dimensional diffusion process in a potential. To keep the analysis analytically tractable we focus here on a linear potential.

Here, the transition state is assumed to be at $x = 0$, the folded state at $x_f > 0$ (which is in accordance with our above choice of the transition state). In the presence of a linear potential $V(x)$ the probability distribution $P(x, t)$ obeys the Fokker–Planck equation

$$\partial_t P(x,t)=\frac{D}{k_BT}\partial_x(V'(x)P(x,t))+D\partial^2_x P(x,t), \tag{10}$$

where $V'(x)=\partial_x V(x)=-k_BT/a$. Upon Laplace-transforming $P(x, t)$ one has for the initial condition $P(x, 0) = \delta(x - x_0)$ for $x_f > x_0 \geqslant 0$

$$sP(x,s)-DP''(x,s)+\frac{D}{a}P'(x,s)=\delta(x-x_0). \tag{11}$$

One can easily show that this equation has the solution

$$P_G(x,x')=-\int_{-s}^{\infty}dk\frac{\tau}{ak}f(k+s)e^{-f(k+s)(x+x')/2a}, \tag{12}$$

where $f(k+s)=\sqrt{1+4\tau^2(s+k)}$ and $\tau=a^2/D$. With $q^{\alpha\beta}(s)\sim\int_0^\infty dx_0 P_G^2(x_f,x_0)$, one then obtains in leading order for the replica correlation function

$$q^{\alpha\beta}(s)\sim\frac{\tilde{\tau}^{-2}}{s^2}+C\frac{\tilde{\tau}^{-3}}{s^3} \tag{13}$$

with a constant C and $\tilde{\tau}=\tau x_f/a$. The fit in Fig. 5 corresponds to $\tilde{\tau}^{-1}=0.13$ (1N0Q) and $\tilde{\tau}^{-1}=0.11$ (1SRL), which in turn corresponds to energy differences $\Delta F\simeq 7.7k_BT$ and $\Delta F\simeq 8.9k_BT$ between the transition state and the folded state. For 1N0Q this compares well with $\Delta F\simeq 7k_BT$ from the simulations, while for 1SRL the estimate for the barrier is off by a factor of 2 (simulations: $\Delta F\simeq 4k_BT$). This implies that the description of the folding dynamics as diffusion in a linear potential works better for 1N0Q than for 1SRL.

In this study, we have sought to demonstrate how static and dynamic replica correction functions can be used to analyze single molecule experiments. These tools allow one to characterize quantitatively how large the accessed phase space is during a complex reaction. The s-dependence of the Laplace-transformed replica correlation function $q(s)$ provides information about the multiplicity of routes taken to the folded state.

Acknowledgements

The authors are grateful for helpful discussions with Koby Levy. This work was supported by National Institutes of Health Grant 5R01 GM44557 and the Center for Theoretical Biological Physics through National Science Foundation Grants PHY0216576 and PHY0225630. S.S.C. is supported by a Ruth L. Kirschstein National Research Service Award from the National Institutes of Health. P.L. is supported by the Fonds der Chemischen Industrie.

References

[1] J.N. Onuchic, P.G. Wolynes, Curr. Opin. Struct. Biol. 14 (2004) 70.
[2] M. Oliveberg, P.G. Wolynes, Quart. Rev. Biophys. 38 (2005) 245.
[3] P.G. Wolynes, J.N. Onuchic, D. Thirumalai, Science 267 (1995) 1619.
[4] N.D. Socci, J.N. Onuchic, P.G. Wolynes, Proteins 32 (1998) 136.
[5] H.P. Lu, L. Xu, X.S. Xie, Science 282 (1998) 1.
[6] E.L. Florin, V.T. Moy, H.E. Gaub, Science 264 (1994) 415.
[7] S.B. Smith, Y.J. Cui, C. Bustamante, Science 271 (1996) 795.
[8] B. Schuler, W.A. Eaton, Curr. Opin. Struct. Biol. 18 (2008) 16.
[9] A. Engel, D.J. Müller, Nat. Struct. Biol. 7 (2000) 715.
[10] G. Hummer, A. Szabo, Proc. Natl. Acad. Sci. USA 98 (2001) 3658.
[11] J.N. Onuchic, J. Wang, P.G. Wolynes, Chem. Phys. 247 (1999) 175.
[12] L.F. Cugliandolo, J. Kurchan, Phys. Rev. Lett. 71 (1993) 173.
[13] M. Mezard, E. Parisi, M.A. Virasaro, Spin Glass Theory and Beyond, World Scientific Press, Singapore, 1986.
[14] J.P. Bouchaud, J. Phys. I 2 (1992) 1.
[15] C. Monthus, J.P. Bouchaud, J. Phys. A 29 (1996) 3.
[16] H. Sompolinsky, A. Zippelius, Phys. Rev. Lett. 47 (1981) 359.
[17] C. Clementi, H. Nymeyer, J.N. Onuchic, J. Mol. Biol. 298 (2000) 937.
[18] V. Sobolev, A. Sorokine, J. Prilusky, E.E. Abola, M. Edelman, Bioinformatics 15 (1999) 327.
[19] S. S Cho, Y. Levy, P.G. Wolynes, Proc. Natl. Acad. Sci. USA 103 (2006) 586.
[20] H. Nymeyer, N.D. Socci, J.N. Onuchic, Proc. Natl. Acad. Sci. USA 97 (2000) 634.