Compaction and Tensile Forces Determine the Accuracy of Folding Landscape Parameters from Single Molecule Pulling Experiments

Greg Morrison, Changbong Hyeon, Michael Hinczewski, and D. Thirumalai

Compaction and tensile forces determine the accuracy of folding landscape parameters from single molecule pulling experiments

Greg Morrison,[1,2] Changbong Hyeon,[3] Michael Hinczewski,[2] and D. Thirumalai[2]

[1]*School of Engineering and Applied Science, Harvard University, Cambridge, Massachusetts, 02138, USA*
[2]*Biophysics Program, Institute For Physical Science and Technology,*
University of Maryland, College Park, MD 20742
[3]*School of Computational Sciences, Korea Institute for Advanced Study, Seoul 130-722, Republic of Korea*

We establish a framework for assessing whether the transition state location of a biopolymer, which can be inferred from single molecule pulling experiments, corresponds to the ensemble of structures that have equal probability of reaching either the folded or unfolded states ($P_{fold} = 0.5$). Using results for the forced-unfolding of a RNA hairpin, an exactly soluble model and an analytic theory, we show that P_{fold} is solely determined by s, an experimentally measurable molecular tensegrity parameter, which is a ratio of the tensile force and a compaction force that stabilizes the folded state. Applications to folding landscapes of DNA hairpins and leucine zipper with two barriers provide a structural interpretation of single molecule experimental data. Our theory can be used to assess whether molecular extension is a good reaction coordinate using measured free energy profiles.

The response of biopolymers to mechanical force (f), at the single molecule level, has produced direct estimates of many features of their folding landscapes, which in turn has given a deeper understanding of how proteins and RNA fold. In particular, single molecule pulling experiments directly measure distribution of forces needed to rupture biomolecules, roughness and shapes of folding landscapes [1–4]. Such measurements have made it possible to decipher the molecular origin of elasticity and mechanical stability of the building blocks of life, which is the first step in describing how they interact to function in the cellular context. The major challenge is to provide a firm theoretical basis for interpreting the physical meaning and reliability of the folding landscape parameters that are extracted from trajectories that project dynamics in multi dimensional space onto one dimensional molecular extension, which is conjugate to f.

The key characteristics of the folding landscape of biomolecules that can be extracted from single molecule force spectroscopy (SMFS) measurements are f-dependent position of the transition state (TS), the distance ($\Delta x^{\ddagger} = x_{TS} - x_0$) from the ensemble of conformations that define the basin of attraction corresponding to the native states (NBA), and the free energy barrier ($\Delta F^{\ddagger}$). The assumption in the analysis of SMFS data is that molecular extension is a good reaction coordinate for RNA and proteins, which implies that a single degree of freedom accurately describes the behavior of the multiple degrees of freedom explored by the biomolecule. Structural meaning of $\Delta x^{\ddagger}$, a parameter that is unique to SMFS, has never been made clear. Despite many subtleties in determining $\Delta x^{\ddagger}$ from measurements [5–7], x_{TS} is most easily identified as a local maximum of the free energy profile at the transition mid-force f_m,

$F(R|f_m)$, which can be constructed by measuring the statistics of end-to-end distance, $P(R)$ at $f = f_m$ [1–4, 8, 9]. This method has been experimentally used to obtain sequence dependent folding landscapes of DNA hairpins [2], and more recently proteins [3]. In order to render physical meaning to $\Delta x^{\ddagger}$ we address two questions here: (1) Does $\Delta x^{\ddagger}$ describe the structures in the Transition State Ensemble (TSE)? The TSE describes a subset of structures that have equal probability of reaching the NBA or UBA staring from $\Delta x^{\ddagger}$. (2) Can a molecular tensegrity (short for tensional integrity) parameter [10] s, expressing balance between the internal compaction force $f_m = \Delta F_{UF}/\Delta x_{UF}$ and the applied tensile force $f_c = \Delta F^{\ddagger}(f_m)/\Delta x^{\ddagger}(f_m)(s = f_c/f_m)$, describe the adequacy of $\Delta x^{\ddagger}(f_m)$ in describing the TSE structures?

We use simulations of a RNA hairpin and an exactly soluble model, both of which are apparent two-state folders as indicated by $F(R|f)$, to answer the two questions posed above. The TS is a surface in the multidimensional folding landscape (stochastic separatrix [11]) across which the flux to the NBA and the Unfolded Basin of Attraction (UBA) is identical. This implies that the fraction of folding trajectories corresponding to $\Delta x^{\ddagger}$ that start from the TS should have equal probability ($P_{fold} \approx 0.5$ [12]) of reaching the NBA and UBA [12, 13]. At $f = f_m$, the mean dwell times in NBA and the basin of attraction corresponding to unfolded conformations (UBA) are identical, so that $\int_0^{x_{TS}} dR P(R) = \int_{x_{TS}}^{\infty} dR P(R) = 0.5$. However, it is unclear whether or not the barrier top position is consistent with the requirement $P_{fold} \approx 0.5$ in force experiments.

To ascertain, whether the barrier top position is consistent with the requirement $P_{fold} \approx 0.5$ in force experiments, we study folding of P5GA, a RNA hairpin,

FIG. 1. Relaxation dynamics of the ensemble of P5GA hairpins from the barrier top of $F(R|f_m)$. (a) $R(t)$ at $f = f_m$. TSE between $R = (4.1 - 4.2)$ nm is shown in the green box with a uniform R-distribution. (b) Starting from the TSE of hairpins from (a), 35 % of the trajectories (red) reach UBA and 65 % of the trajectories (blue) reach NBA. (c) Free energy profile at $f = f_m$. (d) TSE of P5GA has a broad distribution in the number of base pairs (N_{bp}).

for which the NBA and UBA are equally populated at $f_m = 14.7$ pN [8, 9]. Both free energy profiles and the kinetics predicted by Kramers' theory show excellent agreement with the simulation results [9], and formally establishes that extension R is a good reaction coordinate for describing hopping kinetics at $f \approx f_m$. In practice, experimental time traces that have a number of transitions as the one in Fig.1(a) can be used to estimate P_{fold}. With absorbing boundary conditions imposed at $R_N = 2$ nm and $R_U = 7.5$ nm (Fig.1(c)), we directly count the number of molecules from 47 points belonging to the TS region ($R = (4.1 - 4.2)$ nm (Fig.1(c)) that reaches $R < R_N$ (folded) and $R > R_U$ (unfolded). Although the R-distribution of the 47 points is uniform (Fig.1(a)) we obtained $P_{fold} \approx 0.74$. To determine P_{fold} using the ensemble method, we launched 100 trajectories from each of the 47 structures and monitored their evolution (Fig.1(b)) using Brownian dynamics simulations [14] with the multidimensional energy function for the hairpin [8]. We find that $P_{fold} = 0.65$ (Fig.1(c)), which is similar to the value obtained by analyzing the folding trajectory. An examination of the individual trajectories reveal that many molecules, initially with a gradient toward the UBA, re-cross the transition barrier to reach the NBA (blue trajectories in Fig.1(b)). Conversely, most of the molecules directly reach $R = R_N$ if they initially fall into the NBA, showing few recrossing events. Although the precise percentage of molecules reaching UBA or NBA depends on the particular value of the boundary (R_N and R_U), our simulations emphasize the importance of the *re-crossing dynamics*, which is known to cause significant deviations from the transition state theory.

Deviation from $P_{fold} \approx 0.50$ suggests that the global coordinate R alone is not sufficient to rigorously describe the hopping kinetics of P5GA. At least one other auxiliary coordinate is needed, and for structural reasons we take it to be the number, N_{bp}, of base pairs [2]. The broad asymmetric distribution of N_{bp} within the narrow TS region ($R = (4.1 - 4.2)$) implied by $F(R|f_m)$ (Fig.1(d)) shows that hopping kinetics at f_m should be described by multi (at least two) dimensional folding landscape even though the f-dependent rates of hopping between the NBA and UBA can be reliably predicted using $F(R|f_m)$ [9].

FIG. 2. Relaxation dynamics of GRM chains from the barrier top of $\beta F(R|f_m)$. (a) $\beta F(R|f_m)$ for the GRM with interior interaction. Parameters are listed in the text. Red symbols are from the simulation result sampled using the chain with Hamiltonian of Eq.1. Solid line is the theoretical fit. Arrow shows the position of barrier top. (b) The distribution of interior distance at $R = R_{TS}$. Black line is the theoretical prediction. (c) The initial distribution of R prior to the relaxation dynamics, which is very tight round $R \approx R_{TS} = 6$ nm. (d) The probability of being folded as a function of R. Open circles show the exact numerical solution for the sharp potential (Eq.1). Filled circles are from simulations. Solid black line shows the approximate solution using the smoothed potential (Eq.3).

To further illustrate if R-coordinate alone is sufficient to determine the TSE structures we consider an analytically solvable Generalized Rouse Model (GRM) [9] that has a single bond in the interior of the chain whose presence corresponds to the NBA. The GRM Hamiltonian [15]

$$\beta \mathcal{H} = \frac{3}{2a^2} \int_0^N ds\, \dot{\mathbf{r}}^2(s) - \beta \mathbf{f} \cdot \int_0^N ds\, \dot{\mathbf{r}}(s) + \beta V_c \left[\mathbf{r}(s_1) - \mathbf{r}(s_2) \right], \tag{1}$$

where $V_c[x] = k(x^2 - c^2)/2$ for $x \leq c$ and $V_c[x] = 0$ for $x > c$, describes a simple Gaussian chain under tension with an additional cutoff harmonic interaction at the interior points s_1 and s_2. The distribution of R for the GRM, with $\Delta s = |s_2 - s_1|$, is

$$P(R; \Delta s) \propto R \sinh \left(\beta f R \right) e^{-3R^2/2(N - \Delta s)a^2} \times \int_0^\infty dx\, x \sinh \left[\frac{3Rx}{(N - \Delta s)a^2} \right] e^{-3x^2 N/2\Delta s(N - \Delta s)a^2 - \beta V_c[x]}. \tag{2}$$

To obtain P_{fold} we set $N = 22$ for the number of bonds, $a = 0.545$ nm for their spacing, $\beta = 1/k_B T$, $\beta k = 1.65$ nm^{-2}, $c = 4$ nm for the strength and cutoff distance of the interior bond interaction, respectively, and $\Delta s = 18$. For this set of parameters, the transition mid-force (where $P(x < c) = P(x > c)$) is $f_m \approx 16.8$ pN. The two minima and the position of barrier top of the free energy $\beta F(R|f_m) = -\log[P(R)]$ are easily determined as $R_N \approx 3.91$ nm, $R_U \approx 9.39$ nm, and $R_{TS} = 6.00$ nm, respectively (Fig.2(a)).

To calculate P_{fold}, we prepared 5000 GRM chains with $R = R_{TS} = 6$ nm that corresponds to the maximum in $F(R|f_m)$ (Fig. 2a), and allowed the system to relax to either of the two basins of attraction, ending the simulation when the chain extension attains the value R_N or R_U. These simulations are performed using the Hamiltonian in Eq.1, and not on the simple one-dimensional profile $\beta F(R) = -\log[P(R)]$. We find 60 % (40 %) of

the initial chains from R_{TS} reach $R = R_N$ ($R = R_U$), which implies $P_{fold} = 0.6$. Similar to the P5GA, the GRM dynamics projected onto the R-coordinate using both sets of parameters exhibit a number of recrossing events.

To understand the relation between x (the structural coordinate which specifies the NBA and UBA in the GRM) and chain extension R (pulling coordinate that is conjugate to f) we approximate the sharp interaction in Eq.1 by a smoothed potential,

$$\beta V_c[x] \approx \beta V_S[x] = -\log\left(e^{-\beta k(x^2 - c^2)/2} + 1\right). \quad (3)$$

by taking advantage of the clear separation between the UBA and NBA (Fig. 2a). Defining $\mathcal{N}_k = \langle e^{-\beta k(x^2 - c^2)/2}\delta(|\mathbf{R}| - R)\rangle_0$, with $\langle\cdots\rangle_0$ denoting an average over the Gaussian backbone, we can compute approximately $P_R(x < c) = \int_0^c dx\langle\delta(|\mathbf{R}| - R)\rangle \approx \mathcal{N}_k/(\mathcal{N}_{k=0} + \mathcal{N}_k)$, with $\langle\cdots\rangle$ representing an average over the GRM potential in Eq.3. The probability of bond formation satisfies

$$P_R(x < c) \approx \left[1 + \lambda^{\frac{3}{2}}\exp\left(\frac{3R^2\kappa\sigma^2}{2Na^2\lambda} - \frac{\beta kc^2}{2}\right)\right]^{-1} \quad (4)$$

where $\kappa = Na^2\beta k/3$, $\sigma = \Delta s/N$, and $\lambda = 1 + \sigma(1-\sigma)\kappa$. At $R = 6$ nm, Eq.4 gives $P_R(x < c) \approx 0.402 \neq 0.5$. Thus, when the dynamics is initiated from the top of the apparent free energy barrier using the R variable, their internal coordinate (x) is populated primarily with unfolded conformations (Fig.2(b)). Despite the fact that the UBA is primarily populated at $R = R_{TS}$, we find that $P_{fold} \approx 0.6$, so that the NBA will be predominantly populated as the trajectories progress. The midpoint of transition ($P_R(x < c) = 0.5$) occurs at $R_{mid} = \left[(Na^2\lambda/3\kappa\sigma^2)(\beta kc^2 - 3\log\lambda)\right]^{1/2} = 5.91$ nm, which deviates slightly from $R_{TS} = 6.00$ nm. The results from $V_S[x]$, which are in excellent agreement with the simulations as well as the numerical results using $V_c[x]$ (Fig.2), also suggests that hopping kinetics in the GRM involves coupling between x and R. Thus, even in this simple system accurate location of the TSE should consider two dimensional free energy profiles (see [6] and [16]).

To answer the second question wes introduce a molecular tensegrity parameter, which is a ratio of tensile force and a force that determines the stability of the biopolymer. The limit of mechanical stability of the NBA is determined by the critical unbinding force $f_c = \Delta F^{\ddagger}(f_m)/\Delta x^{\ddagger}(f_m)$. At the midpoint force $f = f_m$, the tensegrity parameter $s = f_c/f_m$ determines whether the applied external tension is sufficient to overcome the stability of the NBA. Models, which approximate the free energy profiles using a cubic or cusp potential [7], could alter f_c or f_m from the form suggested above. However,

FIG. 3. P_{unfold} as a function of $s = \Delta F^{\ddagger}/f_m\Delta x^{\ddagger}$ for a variety of GRM parameters (colored circles). Red corresponds to small $\beta k \sim (0.1 - 1)nm^{-2}$, purple to intermediate $\beta k \sim (1 - 50)nm^{-2}$, and blue to high $\beta k \sim (50 - 500)nm^{-2}$. The point sizes indicate the magnitude of the midpoint force, with the smallest points at $f_m = 5$pN and the largest at $f_m = 25$pN. The large square indicates the results of the P5GA simulations with $P_{unfold} = 0.35$. Gray line is the theoretical prediction $P_{unfold}(s) = (1 + s)^{-1}/2$ for $s \gg 1$. Solid line in the inset shows the predicted $P_{unfold}(s)$ for $s \ll 1$ (Eq.5) using the DNA hairpin free energy profiles (not deconvolved) in [2]. Dashed line is for leucine zipper, which has two barriers, using data from [3]. As explained in the Supplementary Information, numbers 1 and 2 are for NBA→I and I→UBA transitions, respectively. Because of different k_U values (Eq.5) dashed and solid lines do not coincide.

because s involves the ratio of the two forces the precise numerical factors are not relevant. Barrier crossings between the NBA over the TS are governed by the competition between f_c and the applied force $f = f_m$. For $f_c \gg f_m$, barrier recrossing from the NBA to the TS will be extremely rare, while if $f_c \ll f_m$, barrier crossing events will be common. We would therefore expect as the ratio $s = f_c/f_m$ increases, barrier recrossings from within the NBA to the TS will decrease, and the probability of reaching $R = R_N$ to increase. Thus, P_{fold}, the structural link to $\Delta x^{\ddagger}$ should be determined by the experimentally measurable tensegrity parameter s.

To confirm this expectation, we ran 20 different sets of GRM parameters (1000 runs each), with f_m ranging between 5-25pN, βk ranging from 0.1 to 600 nm^{-2}, c between 0.26 and 6 nm, and Δs between $N/2$ and N. Fig. 3 shows that $P_{unfold} = 1 - P_{fold}$ decreases monotonically with $s = f_c/f_m$. The dependence of P_{unfold} on s (Fig. 3) can be derived by considering diffusion in a 1D version of the GRM potential in Eq.(1), $\beta U(x) = -fx + \frac{1}{2}k_ux^2 + V_c(x)$ with k_u being the curvature in the bottom of the UBA. Assuming constant diffusivity, $P_{unfold} = \int_{x_N}^c dx' e^{\beta U(x')} / \int_{x_N}^{x_U} dx' e^{\beta U(x')}$ [17], where the positions of the transition barrier, native, and unfolded wells are given by c, x_N, and x_U respectively. For $s \gg 1$, we obtain $P_{unfold} = (1 + s)^{-1}/2$, with the entire dependence on the energy landscape encoded in s. The analytical result, plotted as a solid curve in Fig. 3, captures the overall trend of the GRM and P5GA simulations.

For $s \ll 1$, P_{unfold} can be approximated as:

$$P_{unfold} = \frac{1}{2}\left[\frac{\Phi}{1 + \Phi} + \frac{32f_m^2(\Phi - 1)}{\pi k_U(\Phi + 1)^3}s\right]^{-1}, \quad (5)$$

where $\Phi = 4f_m/\sqrt{2\pi k_B Tk_U}$ and $k_U = \partial^2 U/\partial x^2$ at $x = x_U$. This relation can be used to predict the degree to which extension is a good reaction coordinate for

DNA hairpins. We calculated s from the measured energy landscapes in [2], and obtained P_{unfold} using Eq.(5). We took $f_m \approx 12.5$ pN, and $\beta k_U \approx 0.2$ nm^{-2}. The predicted P_{fold} inset in Fig. 3) varies between 0.54–0.61 for the DNA hairpin sequences A–D, close to the ideal value of $1/2$, indicating that extension is a reasonable coordinate except possibly for sequence C, which has a T:T base pair (bp) mismatch 7 bp from the stem. The limitation of extension as a reaction coordinate for this sequence is consistent with the observation that folding of sequence C has an intermediate as indicated by the three minima in $F(R|f_m)$ [2].

To illustrate that the theory is applicable when there are multiple barriers [4], we analyze the data for leucine zipper which unfolds by populating an intermediate. In this case there are two tensegrity parameters, s_1 (=0.05) and s_2 (=0.15) (see Supplementary Information). The predicted P_{unfold} values show that extension is a good reaction coordinate for the NBA→I transition but is less for the I→UBA transition.(Fig. 3).

Our results can be confirmed by analyzing long time folding trajectories $R(t)$ as long as multiple hopping events occur. Because $\Delta x^\ddagger$ is f-dependent, it follows that tensegrity can be altered by changing f. Thus, extension may be a good reaction coordinate over a certain range of f but may not remain so under all loading conditions.

[1] J. Liphardt et. al., Science **292**, 733 (2001); B. Onoa et. al., Science **299**, 1892 (2003); C. Cecconi, E. A. Shank, C. Bustamante, and S. Marqusee, Science **309**, 2057 (2005); C. Hyeon and D. Thirumalai, Proc. Natl .Acad. Sci. **100**, 10249 (2003); R. Nevo et. al., EMBO reports **6**, 482 (2005).

[2] M. T. Woodside et. al., Science **314**, 1001 (2006).

[3] J. C. M. Gebhardt, T. Bornschlögl, and M. Rief, Proc. Natl. Acad. Sci. **107**, 2013 (2010).

[4] M. Manosas, D. Collins, and F. Ritort, Phys. Rev. Lett. **96**, 218301 (2006).

[5] C. Hyeon and D. Thirumalai, Proc. Natl. Acad. Sci. **102**, 6789 (2005); C. Hyeon and D. Thirumalai, Biophys. J. **90**, 3410 (2006); O. K. Dudko et. al., Proc. Natl. Acad. Sci. **100**, 11378 (2003); H. J. Lin, H. Y. Chen, Y. J. Sheng, and H. K. Tsao, Phys. Rev. Lett. **98**, 088304 (2007); R. Merkel et. al., Nature **397**, 50 (1999); I. Derenyi, D. Bartolo, and A. Ajdari, Biophys. J. **86**, 1263 (2004).

[6] C. Hyeon and D. Thirumalai, J. Phys.: Condens. Matter **19**, 113101 (2007).

[7] O. K. Dudko, G. Hummer, and A. Szabo, Phys. Rev. Lett. **96**, 108101 (2006).

[8] C. Hyeon and D. Thirumalai, Biophys. J. **92**, 731 (2007).

[9] C. Hyeon, G. Morrison, and D. Thirumalai, Proc. Natl. Acad. Sci. **105**, 9604 (2008).

[10] D. E. Ingber, J. Cell. Sci. **116**, 1157 (2003).

[11] M. M. Klosek, B. J. Matkowsky, and Z. Schuss, Ber. Bunsenges. Phys. Chem. **95**, 331 (1991).

[12] R. Du et. al., J. Chem. Phys. **108**, 334 (1998).

[13] D. K. Klimov and D. Thirumalai, Proteins - Struct. Funct. Gene. **43**, 465 (2001).

[14] T. Veitshans, D. Klimov, and D. Thirumalai, Folding Des. **2**, 1 (1997).

[15] V. Barsegov, G. Morrison, and D. Thirumalai, Phys. Rev. Lett. **100**, 248102 (2008).

[16] Y. Suzuki and O. K. Dudko, Phys. Rev. Lett. **104**, 048101 (2010).

[17] N.G. van Kampen, *Stochastic Processes in Chemistry and Physics*, 3rd ed. (Elsevier, Amsterdam, 2007), pg. 305.
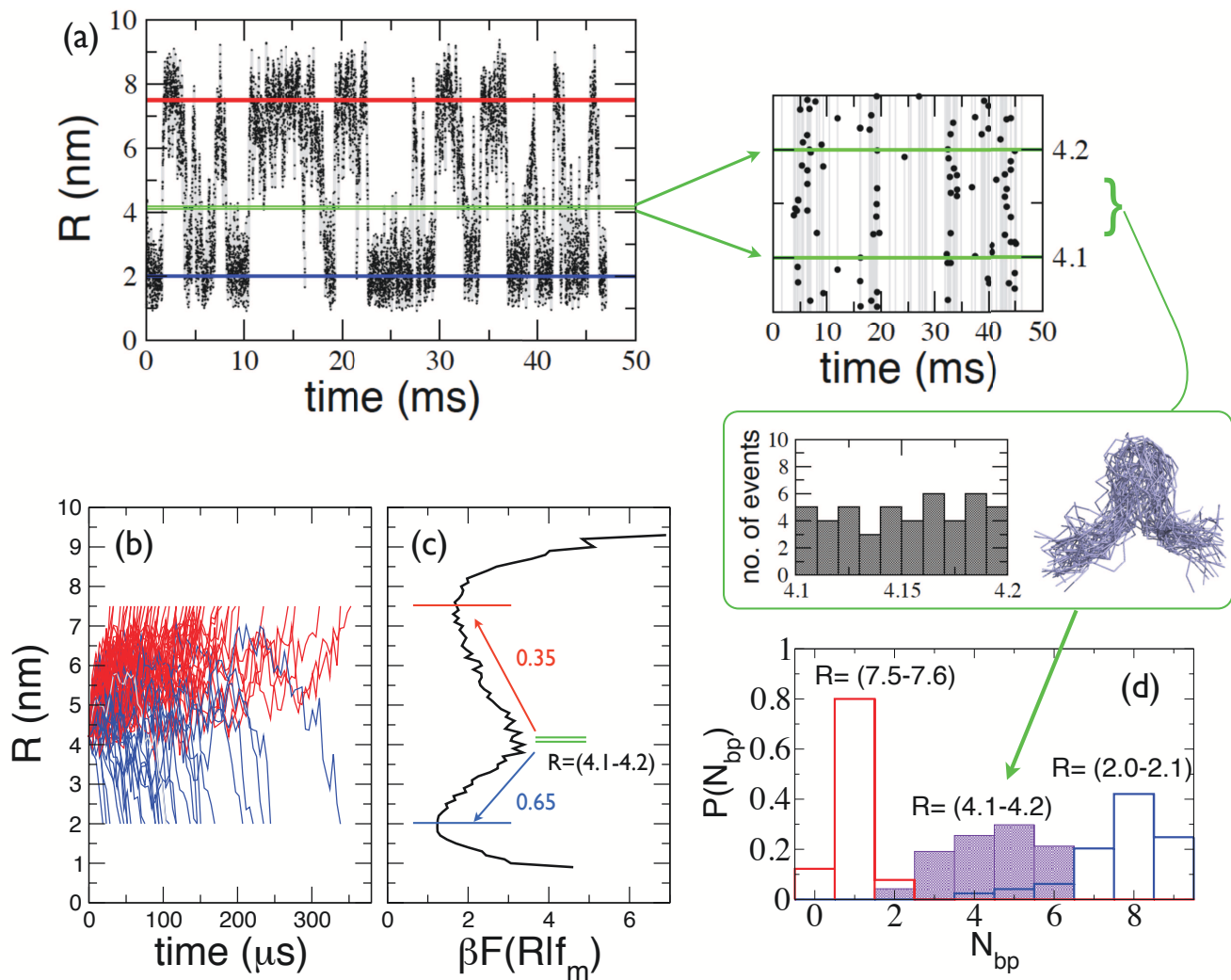
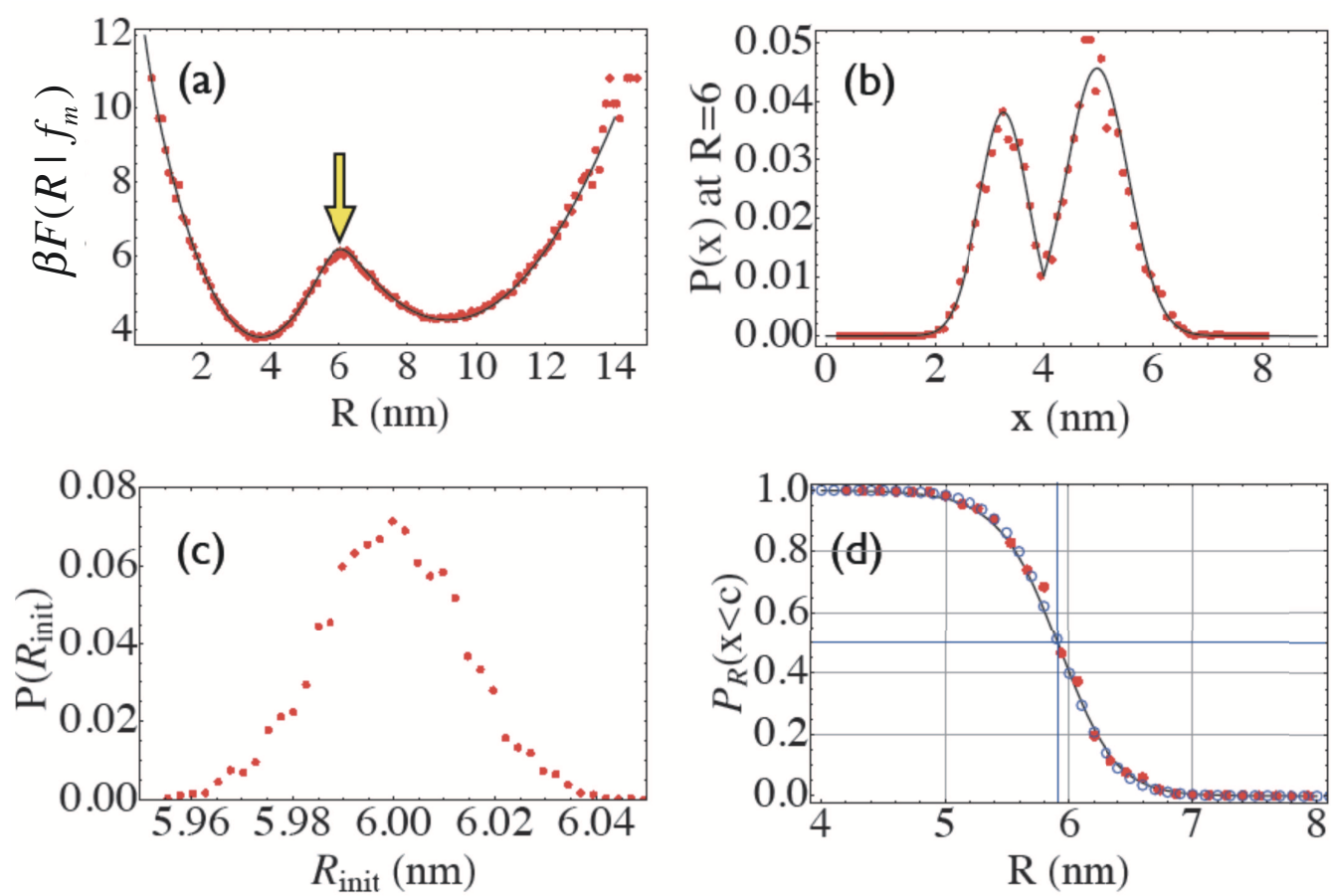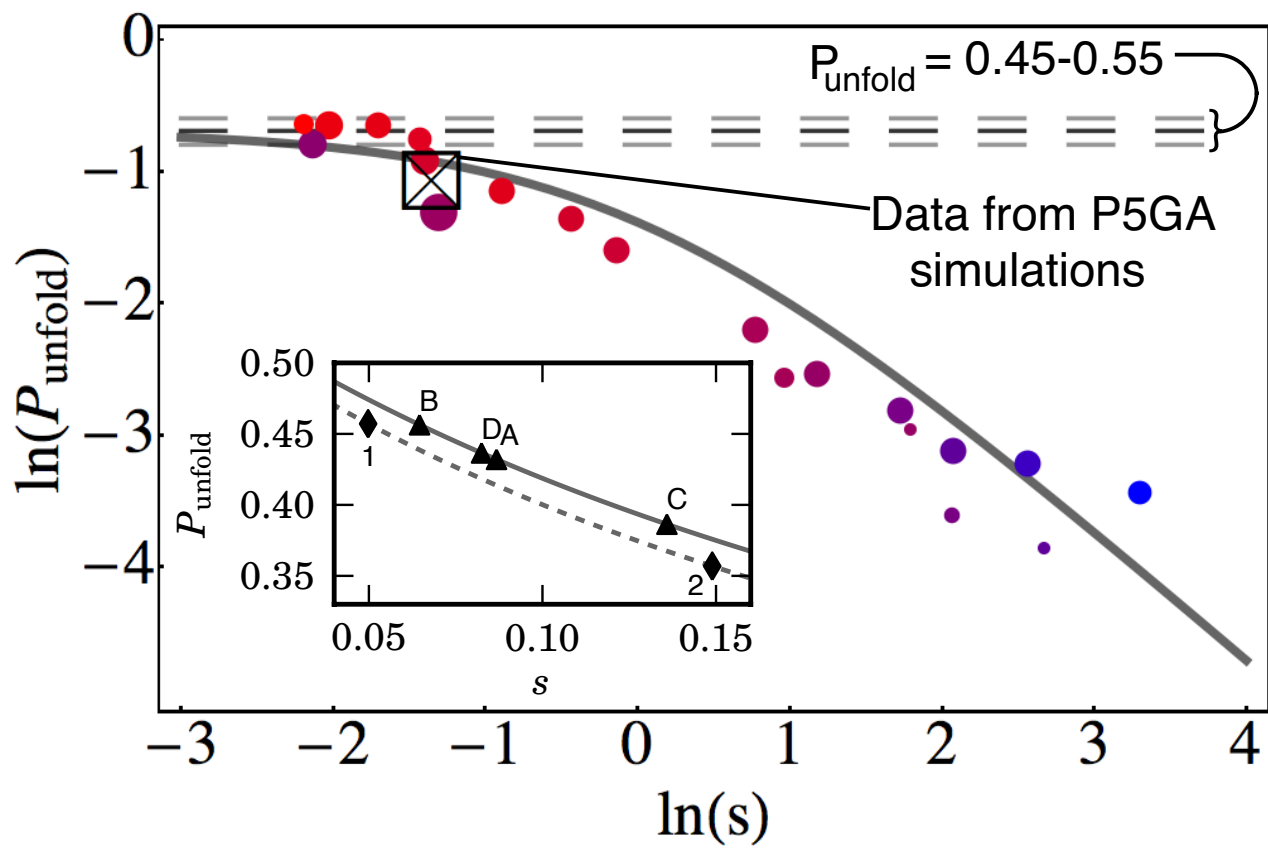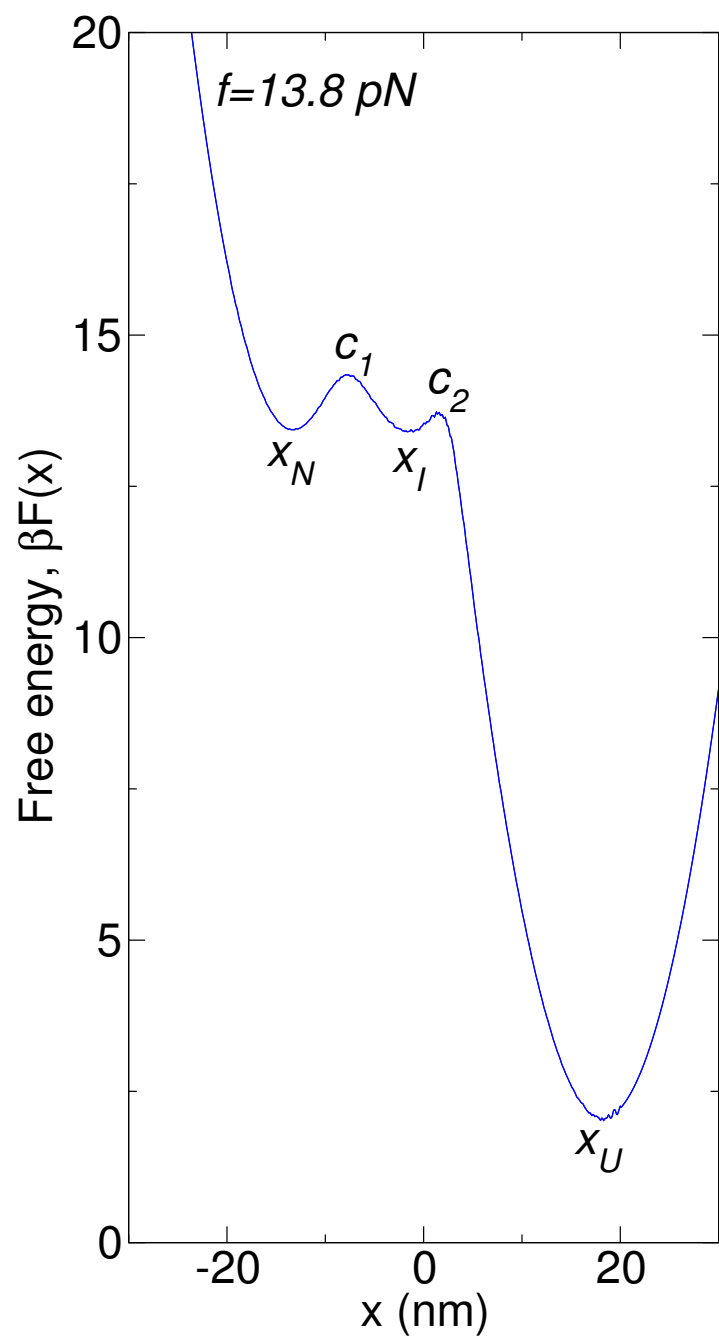Figure 1   LV12420   02Mar2011

Figure 2   LV12420   02Mar2011

Figure 3　　　LV12420　　02Mar2011

Figure 4   LV12420   02Mar2011