Supporting Information for "Collapse Precedes Folding in Denaturant-Dependent Assembly of Ubiquitin"

Govardhan Reddy*,† and D. Thirumalai‡

†*Solid State and Structural Chemistry Unit, Indian Institute of Science, Bangalore, Karnataka, India 560012*

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

E-mail: greddy@sscu.iisc.ernet.in

Protein Model and Simulation Methods

Hamiltonian for the Self Organized Polymer-Side Chain (SOP-SC) model: The coarse-grained model for the protein Ubiquitin (Ub) is identical to the model used in our earlier work where we studied the temperature effects on Ub folding.[1] The protein is coarse-grained using the native-centric Self Organized Polymer - Side Chain model (SOP-SC).[2,3] In this model each residue is represented using two beads, one for the backbone atoms and the other for the side chain (SC). In the absence of denaturants, $[C] = 0$, the coarse-grained force-field of a protein conformation given by the co-ordinates $\{\mathbf{r}\}$ is:

$$E_{CG}(\{\mathbf{r}\}, 0) = E_B + E_{NB}^N + E_{NB}^{NN} + \lambda E^{el}. \tag{S1}$$

The bond between covalently linked beads is modeled using FENE potential, E_B, given by

$$E_B = -\sum_{i=1}^{N_B} \frac{k}{2} R_o^2 \log\left(1 - \frac{(r_i - r_{cry,i})^2}{R_o^2}\right), \tag{S2}$$

where N_B(=151) is the total number of bonds between the beads in the coarse grained model of Ub.

The non-bonded native interactions, E_{NB}^N, in Equation S1 is,

$$\begin{aligned} E_{NB}^N = &\sum_{i=1}^{N_N^{bb}} \epsilon_h^{bb} \left[\left(\frac{r_{cry,i}}{r_i}\right)^{12} - 2\left(\frac{r_{cry,i}}{r_i}\right)^{6}\right] + \sum_{i=1}^{N_N^{bs}} \epsilon_h^{bs} \left[\left(\frac{r_{cry,i}}{r_i}\right)^{12} - 2\left(\frac{r_{cry,i}}{r_i}\right)^{6}\right] \\ &+ \sum_{i=1}^{N_N^{ss}} 0.5(0.7 - \epsilon_i^{ss})(300.0 k_B) \left[\left(\frac{r_{cry,i}}{r_i}\right)^{12} - 2\left(\frac{r_{cry,i}}{r_i}\right)^{6}\right] \end{aligned} \tag{S3}$$

where N_N^{bb}(=177), N_N^{bs}(=486), and N_N^{ss}(=204) are the numbers of backbone-backbone, backbone-sidechain, sidechain-sidechain native interactions, respectively, k_B is the Boltzmann constant, r_i is the distance between the i^{th} pair of residues, and $r_{cry,i}$ is the corresponding distance in the crystal structure. For Ub the numbers are mentioned in the parentheses. The Betancourt-

Thirumalai statistical potential[4] is used for interaction strength between a pair of side chain beads i, ϵ_i^{ss}.

The non-native interactions, E_{NB}^{NN}, in Equation S1 is taken to be

$$E_{NB}^{NN} = \sum_{i=1}^{N_{NN}} \epsilon_l \left(\frac{\sigma_i}{r_i}\right)^6 + \sum_{i=1}^{N_{ang}^{bb}} \epsilon_l \left(\frac{\sigma^{bb}}{r_i}\right)^6 + \sum_{i=1}^{N_{ang}^{bs}} \epsilon_l \left(\frac{\sigma_i^{bs}}{r_i}\right)^6 \tag{S4}$$

where N_{NN}(=10159 in Ub) is the total number of non-native interactions, N_{ang}^{bb}(=74 in Ub) is the number of pairs of backbone beads separated by 2 bonds in the SOP-SC model, and N_{ang}^{bs}(=150 in Ub) is the number of pairs of backbone and side chain beads separated by 2 bonds in the SOP-SC model. σ^{bb} is the diameter of the backbone beads, and σ_i^{bs}(= $f[\sigma^{bb} + \sigma_i^{sc}]/2.0$) is the sum of the radii of the backbone and the side chain in the i^{th} pair of angular interactions scaled by a factor $f = 0.8$. The radii for side chains of amino acids are given in Table S2 in Ref.[3] The values of the interaction parameters used in the energy function are given in Table S1.

At neutral pH we expect electrostatic interactions to be important in Ub as it has 23 charged residues (Figure 1A). The electrostatic effects in Equation S1, E^{el} are modeled using the screened Coulomb potential,

$$E^{el} = \sum_{i=1}^{N_c-1} \sum_{j=(i+1)}^{N_c} \frac{q_i q_j \ exp(-\kappa r_{ij})}{\epsilon r_{ij}}, \tag{S5}$$

where N_c is the number of charged residues, q_i and q_j are the charges on the side chains of the i^{th} and j^{th} residues respectively, κ is the inverse Debye length, and r_{ij} is the distance between the beads located at the centers of mass of side chains i and j. The electrostatic interaction between two bonded residues is ignored. The value of q_i, measured in unit of electron charge, is +1 for positively charged residue and is -1 for negatively charged residues. The dielectric constant ϵ is $10\epsilon_o$ (ϵ_o is vacuum permitivity). In implicit solvent simulations of proteins a range of dielectric constants with values from 2-20 are typically used.[5] We used

a value of $10\epsilon_o$ (ϵ_o is vacuum permitivity), which gave a reasonable radius of gyration of the protein in the unfolded state. We calculated κ assuming a monovalent salt of 10 millimolar is present in the solution. The parameter λ in Eq. S1 is intended to account for pH effects. At neutral pH $\lambda = 1.0$. In acidic pH, E^{el} is not as relevant and hence we set $\lambda = 0$.

Simulations: In order to sample the conformations of the polypeptide chain efficiently we used low friction Langevin dynamics simulations[6] to calculate the thermodynamic properties. The equation of motion used in the Langevin dynamics simulations is $m\ddot{\vec{r_i}} = -\zeta\dot{\vec{r_i}} + \vec{F_c} + \vec{\Gamma}$, where m is the mass of the protein beads, ζ is the friction coefficient, $\vec{r_i}$ is the position of the bead i, $\vec{F_c} = \frac{\partial E_{TOT}}{\partial \vec{r_i}}$, $\vec{\Gamma}$ is the random force with a white noise spectrum. In the discretized form the autocorrelation function of the random force is $\langle \Gamma(t)\ \Gamma(t+nh) \rangle = \frac{2\zeta k_B T}{h}\delta_{0,n}$, where $n = 0, 1, ...$ and $\delta_{0,n}$ is the Kronecker delta function. The Langevin equation is integrated using the velocity Verlet algorithm.[6,7] We used $\zeta = 0.05\ m/\tau_L$ and $h = 0.005\ \tau_L$, where τ_L is the unit of time to compute the thermodynamic properties.

To provide a realistic description of the folding kinetics we performed Brownian dynamics simulations and the equations of motion are integrated using the Ermak-McCammon algorithm,[8] $\vec{r_i}(t+h) = \vec{r_i}(t) + \frac{h}{\zeta}\vec{F_c} + \vec{\Gamma}$. Here $\vec{\Gamma}$ is a random force with a Gaussian distribution with mean zero and variance $\langle \Gamma(h)^2 \rangle = \frac{2k_B T h}{\zeta}$. The friction coefficient $\zeta = 50\ m/\tau_H$ approximately corresponds to the value in water and $h = 0.005\ \tau_H$. In the simulations, the characteristic unit of length $a = 1$ Å, energy $\epsilon = 1\ kcal/mol$, and mass $m = 1.8 \times 10^{-22}\ g$ (typical mass of the bead). The unit of time in Langevin dynamics simulations is $\tau_L (= \sqrt{ma^2/\epsilon}) = 0.51\ ps$. In Brownian dynamics, simulation time is mapped into real time, τ_H using $\tau_H \approx \frac{\zeta_H a^2}{k_B T} = \frac{(\zeta_H \tau_L / m)\epsilon}{k_B T}\tau_L \approx 43\ ps$.

References

(1) Reddy, G.; Thirumalai, D. Dissecting ubiquitin folding using the self-organized polymer model. *J. Phys. Chem. B* **2015**, *119*, 11358–11370.

(2) Hyeon, C.; Dima, R. I.; Thirumalai, D. Pathways and kinetic barriers in mechanical unfolding and refolding of RNA and proteins. *Structure* **2006**, *14*, 1633–1645.

(3) Liu, Z.; 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.

(4) Betancourt, M.; Thirumalai, D. Pair potentials for protein folding: Choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Prot. Sci.* **1999**, *8*, 361–369.

(5) Fogolari, F.; Brigo, A.; Molinari, H. Protocol for MM/PBSA molecular dynamics simulations of proteins. *Biophys. J.* **2003**, *85*, 159–166.

(6) 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.

(7) Swope, W.; Andersen, H.; Berens, P.; Wilson, K. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: application to small clusters. *J. Chem. Phys.* **1982**, *76*, 637–649.

(8) Ermak, D. L.; Mccammon, J. A. Brownian Dynamics with hydrodynamic interactions. *J Chem Phys* **1978**, *69*, 1352–1360.

(9) Wintrode, P.; Makhatadze, G.; Privalov, P. Thermodynamics of ubiquitin unfolding. *Proteins* **1994**, *18*, 246–253.

(10) Ibarra-Molero, B.; Loladze, V.; Makhatadze, G.; Sanchez-Ruiz, J. Thermal versus guanidine-induced unfolding of ubiquitin. An analysis in terms of the contributions from charge-charge interactions to protein stability. *Biochemistry* **1999**, *38*, 8138–8149.

(11) Camacho, C. J.; Thirumalai, D. Kinetics and thermodynamics of folding in model proteins. *Proc. Natl Acad Sci USA* **1993**, *90*, 6369–6372.

Table S1: Parameters for the SOP-Side Chain model

Parameters	Protein
R_o	2.0 Å
k	20 $kcal/mol/Å^2$
R_c	8 Å
ϵ_h^{bb}	0.5 $kcal/mol$[a]
ϵ_h^{bs}	0.5 $kcal/mol$[a]
ϵ_l	1.0 $kcal/mol$[a]
σ^{bb}	3.8 Å
ϵ	10.0

[a] Values are chosen such that the protein melting temperature in simulations is approximately in agreement with the experiments.[9,10]

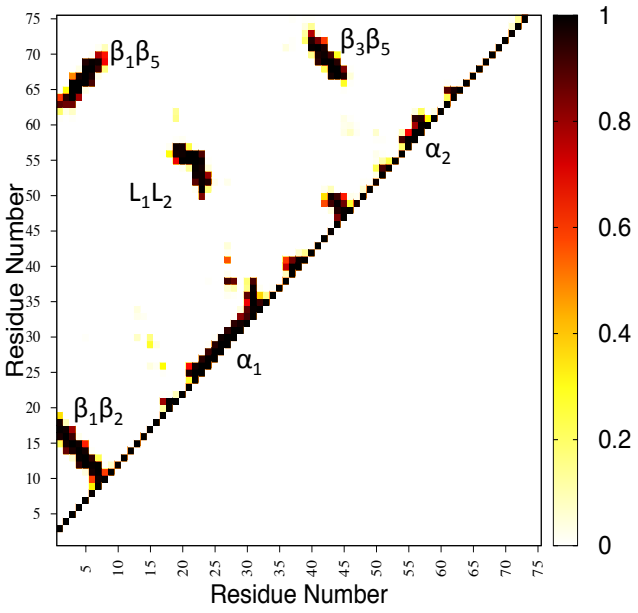

Figure S1: The C_α contact map of Ub shows interactions between various secondary structural elements in the folded state.

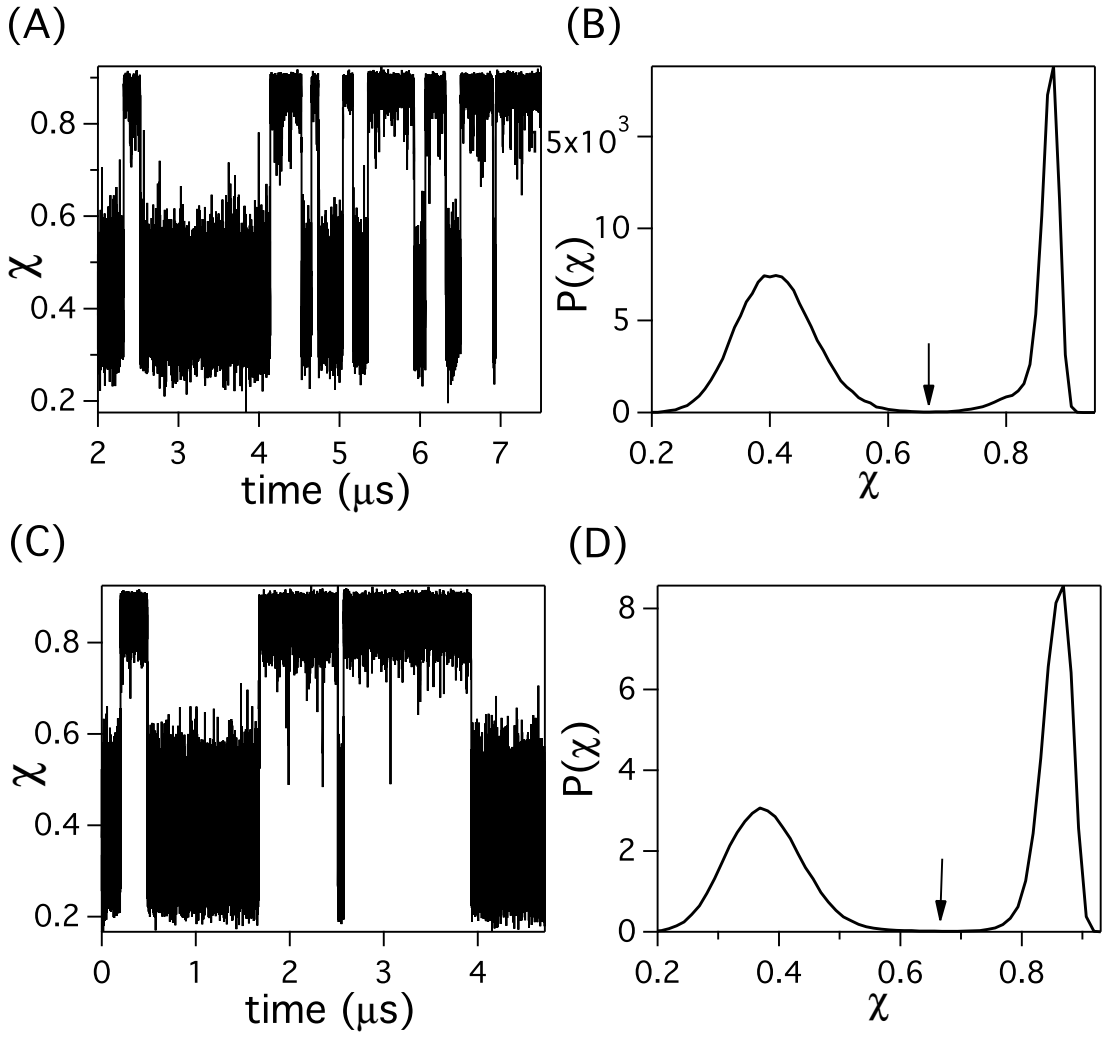

Figure S2: (A) The structural overlap function,[11] χ, as a function of time for Ub at low pH at $T_m = 353\ K$. Ub undergoes multiple transitions between the UBA and NBA basins. (B) The probability distribution of χ, $P(\chi)$ at T_m. The arrow is the value of $\chi = 0.67$ separating the UBA and NBA basins. (C) χ as a function of time for Ub at neutral pH at $T_m \approx 355\ K$. (D) The probability distribution of χ, $P(\chi)$ at T_m and neutral pH. The arrow at $\chi = 0.68$ separates the UBA and NBA.

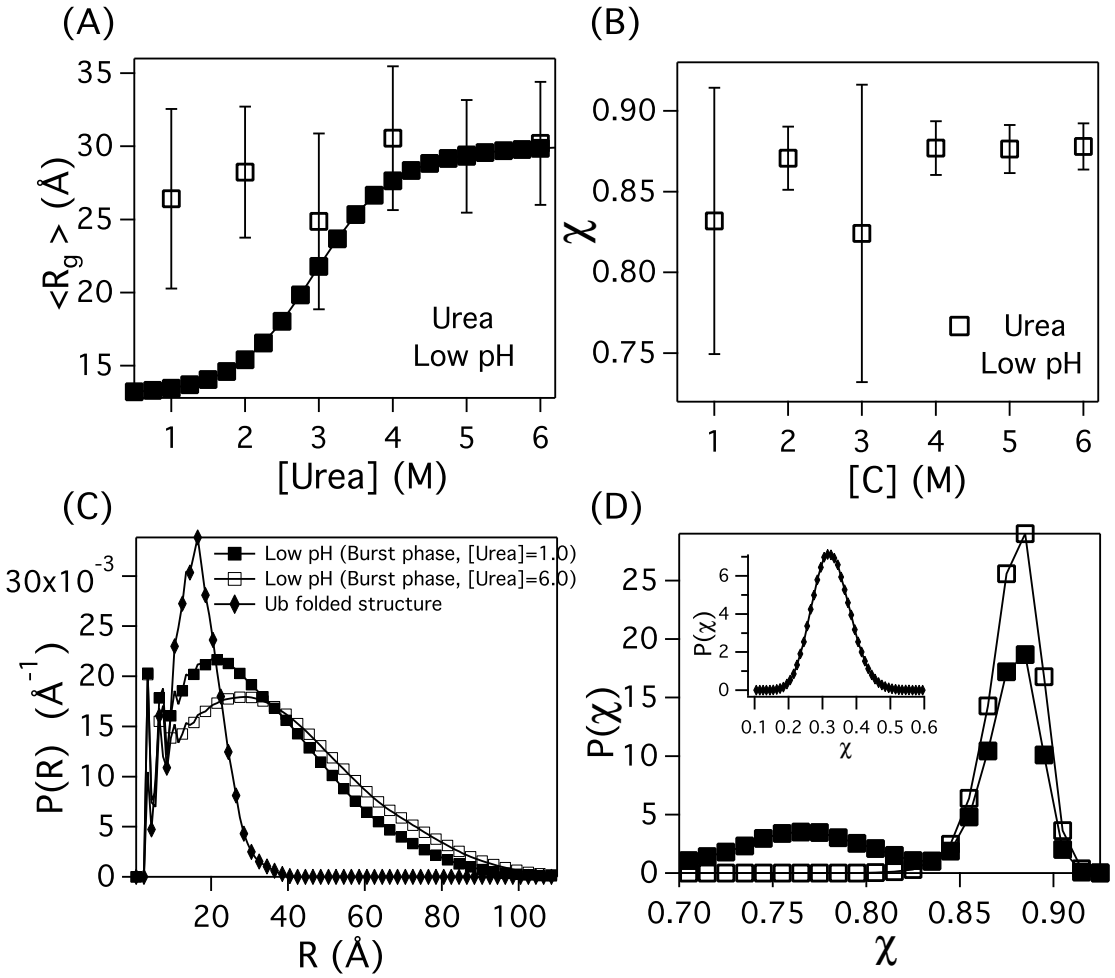

Figure S3: (A) Radius of gyration, R_g, as a function of denaturant concentration. Equilibrium $\langle R_g \rangle$ from coarse-grained simulations at low pH as a function of $[Urea]$ is in solid squares. Data in empty squares (low pH, $T = 332$ K, $[Urea]$) is $\langle R_g \rangle$ during the burst phase of Ub folding. (B) Structural overlap function, χ, plotted as a function of $[C]$. The empty squares in the plot represent the same conditions described in panel-(A). (C) Pair distance distribution function, $P(R)$, plotted as a function of distance, R, for the Ub native structure, and during the burst phase of Ub folding. (D) Probability distribution of χ, $P(\chi)$. The symbols represent the same conditions described in panel-(C). Inset shows $P(\chi)$ for the folded structure at neutral pH, $T = 335$ K and $[C] = 0$ M conditions.

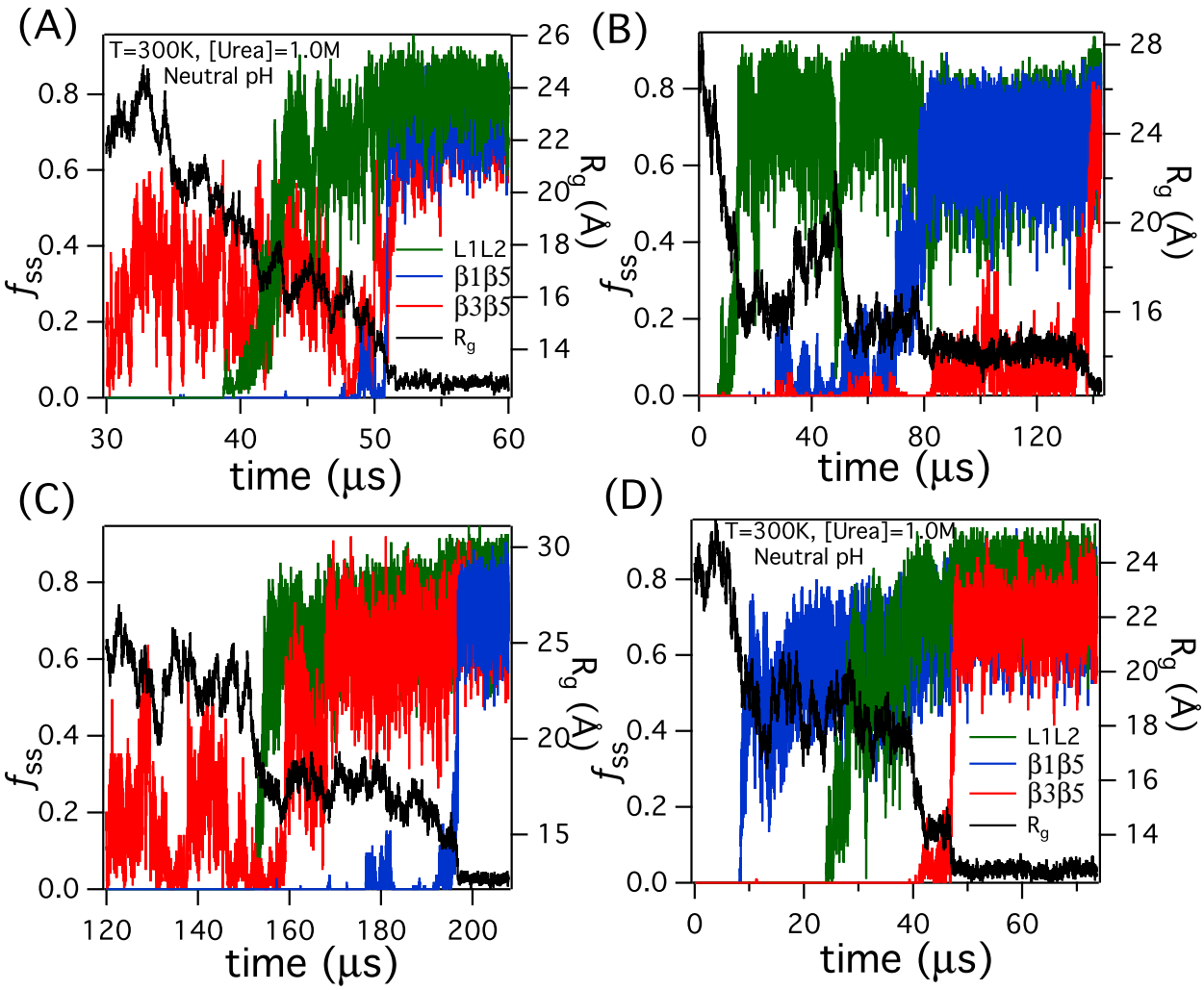

Figure S4: Fraction of native contacts in various secondary structural elements ($\beta_1\beta_5$, $\beta_3\beta_5$, and L_1L_2), f_{ss} and R_g as a function of time in folding trajectories at neutral pH, $T = 300\ K$ and $[Urea]$=1.0 M. The plots in four panels are for trajectories labeled KIN1, KIN2, KIN3 and KIN4 in Figure 9A. The green, blue and magenta colors correspond to secondary structures L_1L_2, $\beta_1\beta_5$, and $\beta_3\beta_5$, respectively. Radius of gyration, R_g as a function of time is shown in black.

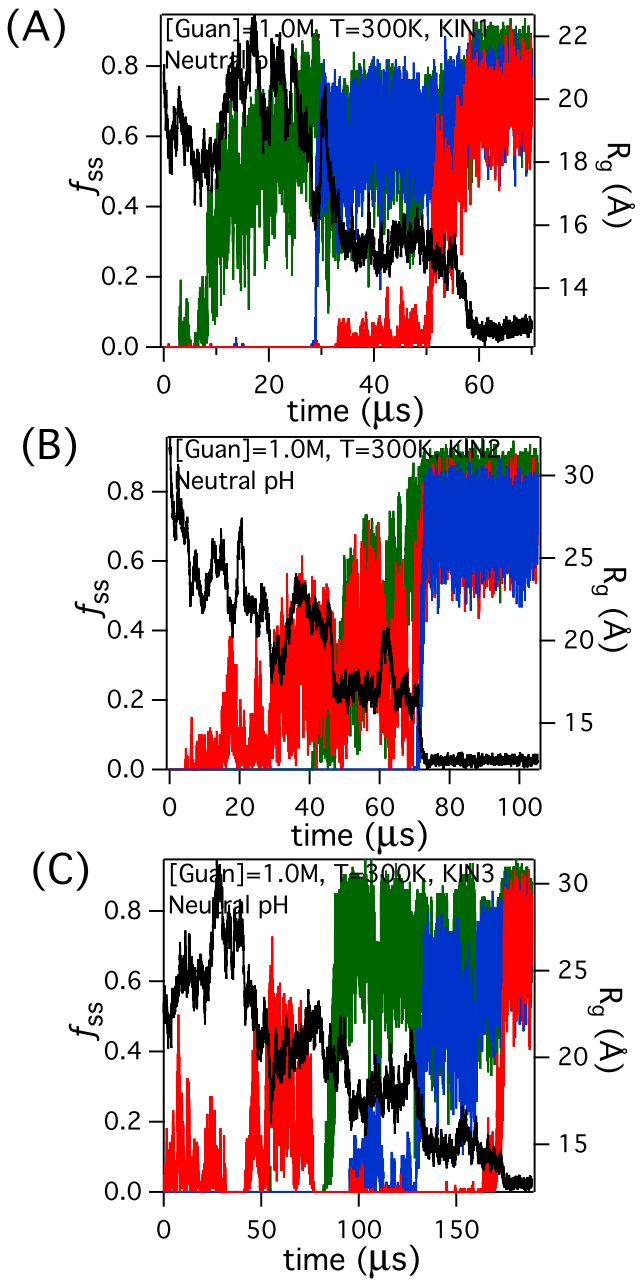

Figure S5: Fraction of native contacts in various secondary structural elements ($\beta_1\beta_5$, $\beta_3\beta_5$, and L_1L_2), f_{ss} and R_g as a function of time in folding trajectories at neutral pH, T = 300 K and $[GdmCl]$=1.0 M. The plots in the three panels are for trajectories labeled KIN1, KIN2, and KIN3 in Figure 9B. The green, blue and magenta colors correspond to secondary structures L_1L_2, $\beta_1\beta_5$, and $\beta_3\beta_5$, respectively. Radius of gyration, R_g as a function of time is shown in black.

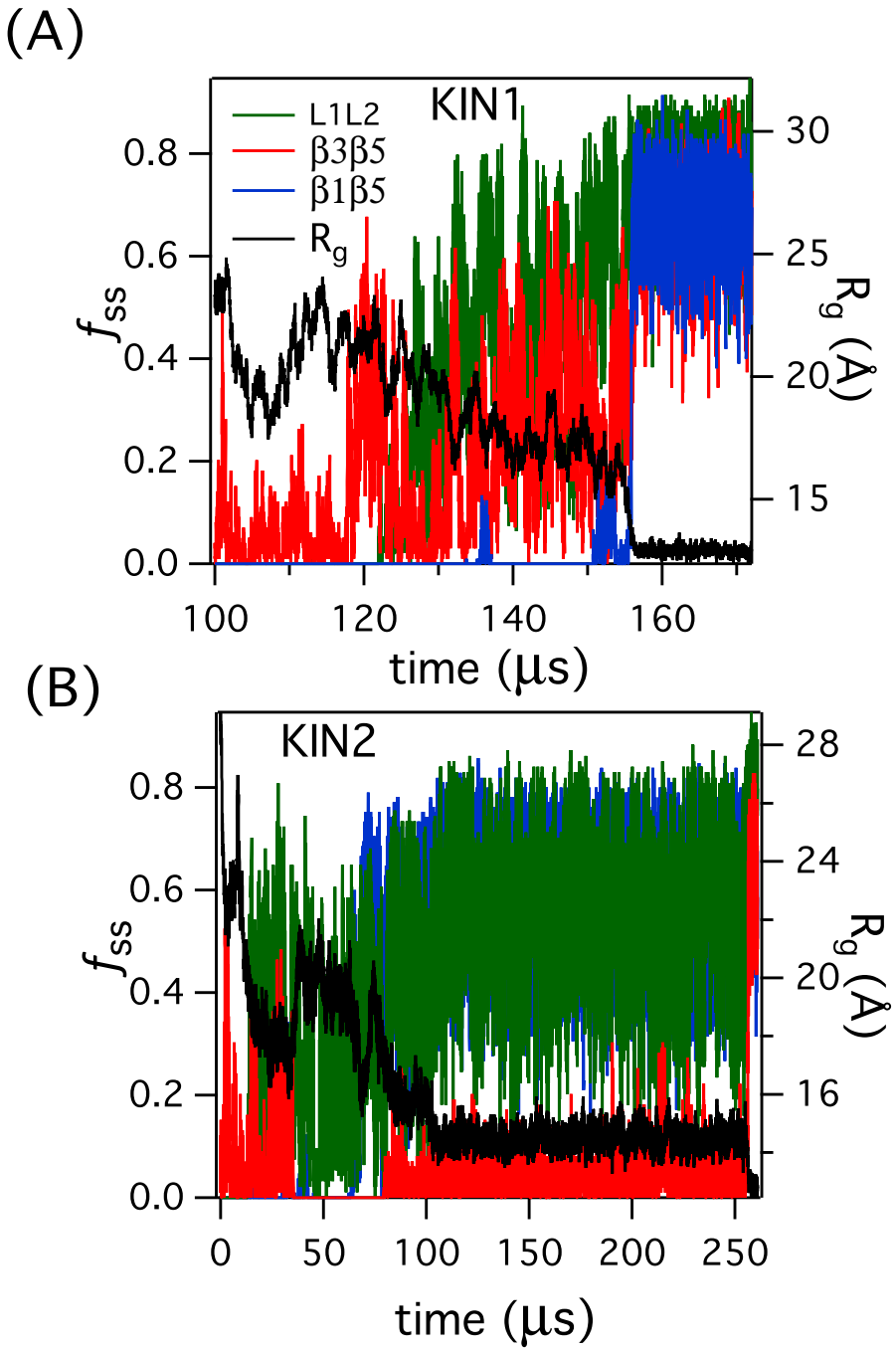

Figure S6: Fraction of native contacts in various secondary structural elements ($\beta_1\beta_5$, $\beta_3\beta_5$, and L_1L_2), f_{ss} and R_g as a function of time in folding trajectories at neutral pH, $T = 335\ K$ and $[Urea]$=1.0 M. The plots in four panels are for trajectories labeled KIN1 and KIN2 in Figure 9C. The green, blue and magenta colors correspond to secondary structures L_1L_2, $\beta_1\beta_5$, and $\beta_3\beta_5$, respectively. Radius of gyration, R_g as a function of time is shown in black.

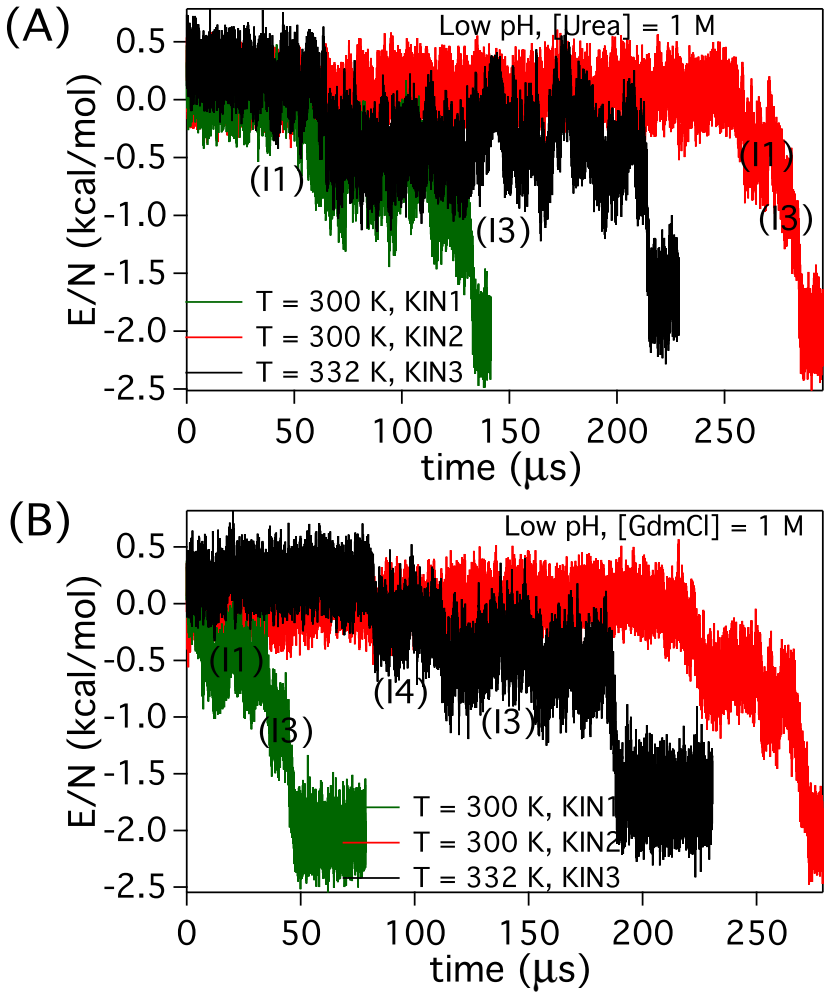

Figure S7: Ub folding kinetics at low pH. Energy per residue, E/N, is plotted as a function of time at conditions (A) $[Urea]$=1.0 M, T = 300 K & 332 K. (B) $[GdmCl]$=1.0 M, T = 300 K & 332 K. Kinetic intermediates I1, I3 and I4 are identified in the folding pathways. Representative structures of I1, I3 and I4 are shown in Figure 9.

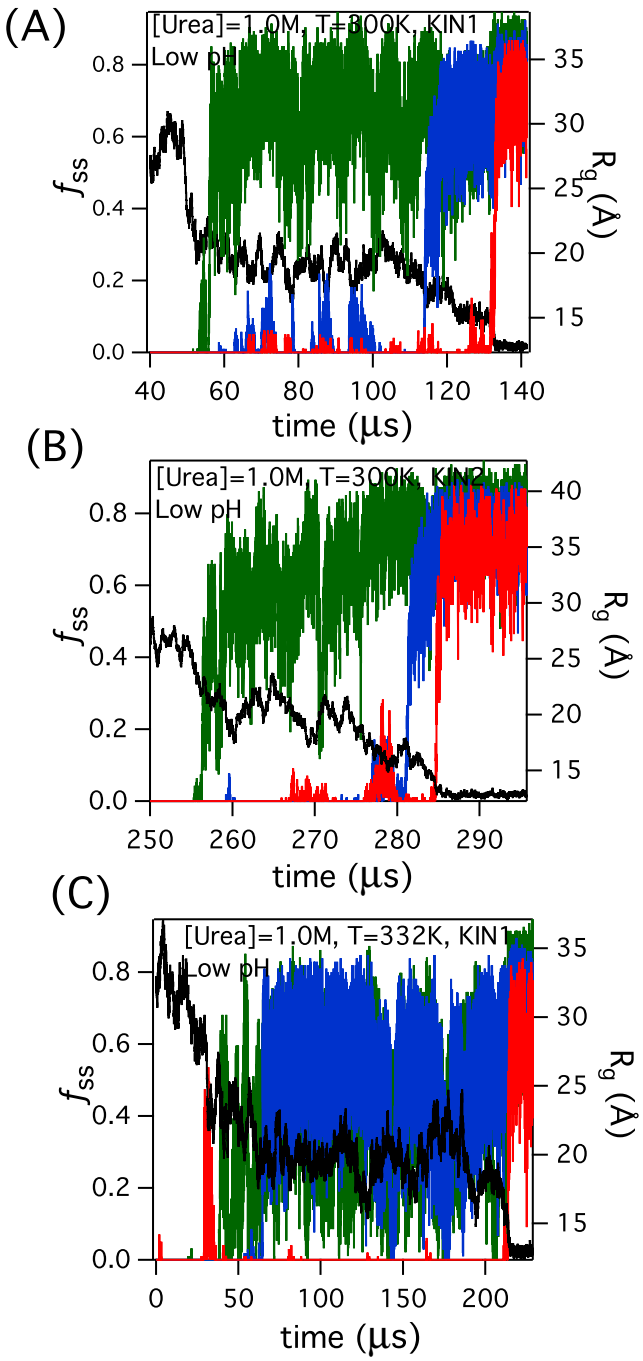

Figure S8: Fraction of native contacts in various secondary structural elements ($\beta_1\beta_5$, $\beta_3\beta_5$, and L_1L_2), f_{ss} and R_g as a function of time at low pH folding trajectories. (A) Trajectory KIN1 at $T = 300\ K$; $[Urea] = 1.0$ M (B) Trajectory KIN2 at $T = 300\ K$; $[Urea] = 1.0$ M and (C) Trajectory KIN3 at $T = 332\ K$; $[Urea] = 1.0$ M The plots in three panels are for trajectories labeled KIN1, KIN2, and KIN3 in Figure S7A. The green, blue and magenta colors correspond to secondary structures L_1L_2, $\beta_1\beta_5$, and $\beta_3\beta_5$, respectively. Radius of gyration, R_g as a function of time is shown in black.

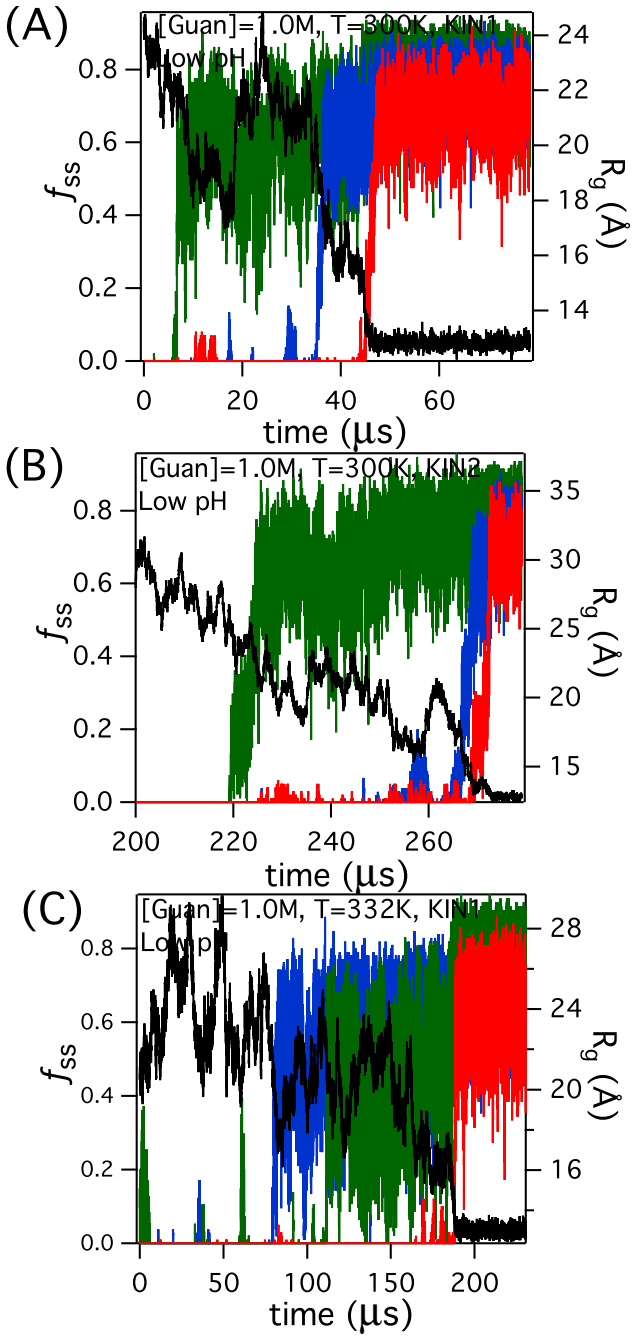

Figure S9: Fraction of native contacts in various secondary structural elements ($\beta_1\beta_5$, $\beta_3\beta_5$, and L_1L_2), f_{ss} and R_g as a function of time at low pH folding trajectories. (A) Trajectory KIN1 at $T = 300\ K$; $[GdmCl] = 1.0$ M (B) Trajectory KIN2 at $T = 300\ K$; $[GdmCl] = 1.0$ M and (C) Trajectory KIN3 at $T = 332\ K$; $[GdmCl] = 1.0$ M The plots in three panels are for trajectories labeled KIN1, KIN2, and KIN3 in Figure S7B. The green, blue and magenta colors correspond to secondary structures L_1L_2, $\beta_1\beta_5$, and $\beta_3\beta_5$, respectively. Radius of gyration, R_g as a function of time is shown in black.