


# Quantitative Prediction of Protein Folding Behaviors from a Simple Statistical Model

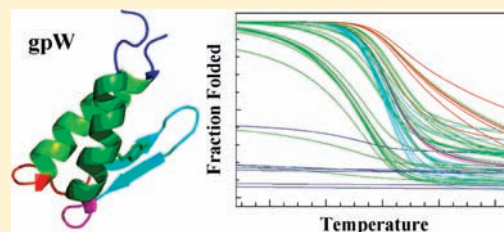
Pierpaolo Bruscolini<sup>\*,†</sup> and Athi N. Naganathan<sup>\*,‡</sup>

<sup>†</sup>Departamento de Física Teórica & Instituto de Biocomputación y Física de Sistemas Complejos (BIFI), Universidad de Zaragoza, Zaragoza, Spain

<sup>‡</sup>Barcelona Supercomputing Center, Barcelona, Spain

 Supporting Information

**ABSTRACT:** The statistical nature of the protein folding process requires the use of equally detailed yet simple models that lend themselves to characterize experiments. One such model is the Wako–Saitô–Muñoz–Eaton model, that we extend here to include solvation effects (WSME-S), introduced via empirical terms. We employ the novel version to analyze the folding of two proteins, gpW and SH3, that have similar size and thermodynamic stability but with the former folding 3 orders of magnitude faster than SH3. A quantitative analysis reveals that gpW presents at most marginal barriers, in contrast to SH3 that folds following a simple two-state approximation. We reproduce the observed experimental differences in melting temperature in gpW as seen by different experimental spectroscopic probes and the shape of the rate-temperature plot. In parallel, we predict the folding complexity expected in gpW from the analysis of both the residue-level thermodynamics and kinetics. SH3 serves as a stringent control with neither folding complexity nor dispersion in melting temperatures being observed. The extended model presented here serves as an ideal tool not only to characterize folding data but also to make experimentally testable predictions.



## INTRODUCTION

The three-dimensional (3D) structure of proteins is characterized by an intricate network of a large number of noncovalent interactions. The chemically diverse amino-acidic residues have at once enabled satisfying the functional, thermodynamic, and kinetic features required of a protein. In spite of this complexity, many proteins are observed to fold in a two-state-like fashion with the (de)population of just two apparent macro-states—folded and unfolded.<sup>1,2</sup> This has led to the justifiable popularity of the chemical two-state model. While it has been quite successful in analyzing protein folding data, it fails to make any testable predictions and lacks structural details thus limiting its applicability. Moreover, with the identification of one-state downhill folding proteins<sup>3–5</sup> and those that fold over marginal barriers<sup>6–8</sup> it is imperative to move away from a two-state description of the folding process. A simple but detailed model that captures the statistical nature of folding would be ideal. There have been, however, very few models that have lived up to the scrutiny of the protein folding community.

Of particular interest are native-centric models of protein folding.<sup>9–11</sup> One such treatment is that developed by Wako and Saitô (WS)<sup>12,13</sup> and later independently by Muñoz and Eaton (ME).<sup>10</sup> Hereby, we call it the WSME model. Instead of two macro-states, it presents a configuration space of  $2^N$  microstates, where  $N$  is the protein length. The assumption is that these configurations can encode for all the relevant parts of the energy landscape. The price to pay is that the greater complexity opens

the way to two complementary approaches in dealing with the model. The first methodology aims at reproducing the experimental results as precisely as possible, and is ready to sacrifice the formal rigor in the process. Examples of this approach are the so-called single or double sequence approximation (SSA, DSA). Here, the complexity of the configuration space is significantly reduced from  $2^N$  to  $\sim N^2$  or  $\sim N^4$  by restricting it to those configurations with just a single stretch (in SSA) or two stretches (in DSA) of native residues, with the rest unfolded. They have therefore been successfully used in juxtaposition with experimental data to quantitatively study the folding behaviors of  $\beta$ -hairpins,<sup>14</sup> to predict folding rates from 3D structure,<sup>10</sup> to identify barrier-less transitions in BBL,<sup>3</sup> and in the detailed analysis of the folding of Villin headpiece domain.<sup>15</sup> Also, this is the only accessible approach when the WSME model is extended to account for interactions across loops.<sup>16–18</sup>

The second approach addresses the fact that the thermodynamics of the model can be studied without any reduction in the configuration space. In their original paper, Wako and Saitô implemented an exact solution (ES) calculation that enables computing the contribution to the energetics from all possible stretches of native residues (i.e.,  $2^N$  species) based on a transfer-matrix formalism.<sup>12,13</sup> An exact solution to the same problem has since been developed independently by Bruscolini and Pelizzola,<sup>19,20</sup>

**Received:** December 3, 2010

**Published:** March 18, 2011

and Henry and Eaton.<sup>16</sup> The distribution of partially structured species and unfolded states are better captured in this approach, and therefore it has been used to characterize the folding process of several proteins<sup>21–32</sup> at a semiquantitative level. Interestingly, most of the previous studies save for that on BBL,<sup>3</sup> have ignored energetic contributions from solvation effects, despite the fact that a good estimate of the solvation heat capacity change upon unfolding is required, for instance, to reproduce the ubiquitous observation of cold denaturation in protein thermodynamics.<sup>33–35</sup> The reason is that it is not trivial to estimate the solvation contributions for the innumerable partially structured species. The exact solution has therefore always been coupled to a very simplified form of the energy function, with the implicit assumption that such a simple function, together with a better description of the configuration space, is sufficient to reproduce experimental data. However, given the previous successes and the information content that one can extract from this model, it would be highly desirable to combine a more accurate parametrization of the energy function to the more precise ES approach enabling a thorough study of folding processes.

In the standard WSME model, the free-energy of each microstate is determined by the simplest form of the Gibbs free-energy function. In other words, it includes an enthalpic term that is determined by the number of contacts between the native residue pairs and an entropic cost for fixing residues in native conformations. Both these terms are assumed to be temperature-independent, and the entropic cost is also taken to be sequence-independent, even if a dependence on secondary structure has been previously considered.<sup>10</sup> In this work, we parametrize the WSME model adopting Freire's empirical treatments of the heat capacities of folded and unfolded states<sup>36,37</sup> to introduce solvation effects (labeled WSME-S with S for solvation). In this novel version, the free-energy of each microstate is similar to the standard model but includes temperature dependence on both enthalpy and entropy apart from sequence-specific entropic costs. A simple empirical formula developed by Freire<sup>37</sup> is employed to recast the heat capacity of any model configuration in terms of the accessible/buried surface areas. The surface areas are in turn estimated by invoking a correlation between the degree of burial/exposure and the number of contacts gained/lost to overcome the improbable time required in calculating the surface areas of  $2^N$  species. We then employ the WSME-S model to analyze the folding of gpW<sup>7</sup> and  $\alpha$ -spectrin SH3,<sup>38</sup> two proteins of similar size and stability but with more than three orders of difference in the relaxation rates. Our results predict a consistent difference in the thermodynamic and kinetic behavior of the two proteins with gpW displaying downhill-like characteristics while SH3 is two-state-like in very good agreement with experiments. This work also highlights the importance and advantages of a quantitative analysis with statistical models to characterize and predict folding behaviors.

## MODEL

In this section we briefly review the WSME model and outline the changes introduced in the WSME-S to include solvation and obtain more realistic predictions. A detailed description can be found in the Supporting Information (SI). The WSME model is a G $\bar{o}$ -like model<sup>9</sup> which considers a protein as a sequence of  $N$  aminoacids described by  $N$  binary variables, denoted by  $m_k$ ,  $k = 1, 2, \dots, N$ . These variables are associated to three-dimensional protein conformations in that  $m_k$  takes the value 1 when the  $k$ th residue is in a native-like conformation and 0 when it is unfolded.

The residues are independent from one another, so that a total of  $2^N$  configurations are possible; there is no intrinsic bias between the state of any two consecutive residues, and the probability of a configuration is determined just by its effective free-energy derived below.

The mapping between protein conformations and model states is such that many different partially unfolded protein conformations are represented by the same model configuration. This is because for each residue, the set of unfolded conformations is larger than the native one; for this reason, an entropic cost  $q_k > 0$  is introduced for ordering a residue  $k$ . The main feature of the model is that two amino acids interact only if they are in contact in the native state (non-native interactions are neglected, in the spirit of G $\bar{o}$ -like models) and if all the peptide bonds between them are native-like (that is, the corresponding dihedral angles  $\phi$ ,  $\psi$  assume their native values). The latter is a drastic assumption which makes the model amenable to analytic treatments, up to the exact solution of the equilibrium.

The effective free energy (sometimes called "effective Hamiltonian" in the physics literature) of the model is written as

$$H = \sum_{i < j} \varepsilon_{ij} \Delta_{ij} \prod_{k=i}^j m_k - RT \sum_{k=1}^N q_k (1 - m_k) \quad (1)$$

where  $R$  is the gas constant and  $T$  the absolute temperature;  $\Delta$  is the contact matrix: its  $(ij)$  element takes the value 1 if aminoacids  $i$  and  $j$  are in contact in the native state and 0 otherwise. An alternative common definition is to take  $\Delta_{ij}$  as the number of contacts between heavy atoms of residues  $i$  and  $j$ . Here and in the following, we will write  $\sum_{i < j}$  as a short form for  $\sum_{i=1}^{N-1} \sum_{j=i+1}^N$ . Usually, the contact energies  $\varepsilon_{ij}$  and the entropic parameters  $q_k$  are considered homogeneous:  $\varepsilon_{ij} = \varepsilon < 0$ ,  $q_i = q$  for each  $i, j$ , even if different choices have also been considered in the literature.<sup>10,14,16,17,39</sup>

The parameter values are then fitted to reproduce some experimental signal such as the fraction of folded protein or the midpoint temperature. Despite its simplicity, the model has been shown to correctly reproduce the main features of experimental equilibrium and kinetics, even if only in a semiquantitative fashion. To improve the quantitative agreement between theoretical predictions and experimental results, we modify the model, allowing the parameters to be temperature dependent as presented below. To guarantee that an exact solution is still possible, we ask that the effective energy of the WSME-S model inherits the same structure from WSME:

$$\mathcal{H}(\mathbf{m}, T) = \varphi(T) + H_c(\mathbf{m}, T) \quad (2)$$

where  $H_c(\mathbf{m}, T) = \sum_{i < j} h_{ij}(T) m_{ij}$ , with  $m_{ij} \equiv \prod_{k=i}^j m_k$ , and we look for an expression of the temperature-dependent parameters  $\varphi$ ,  $h_{ij}$ , that is consistent with the experimental finding on how the interactions between residues depend on external parameters, such as the temperature. Unfortunately, the interaction energies are not experimental observables, and hence it is not easy to find a phenomenological expression for them. Instead, we resort to the phenomenological expression proposed in ref 37 for the heat capacity of any protein conformation:

$$C_p(T) = c_1 MW + c_2 BS + c_3 A_{AP} + c_4 A_P \quad (3)$$

as a function of the molecular mass (MW), the total buried surface area (BS), the polar ( $A_P$ ) and apolar ( $A_{AP}$ ) accessible surface areas. The coefficients  $c_i \equiv c_i(T)$  are linear or quadratic polynomials in  $T$  (see Supporting Information). This expression suggests that if we

are able to associate to each configuration  $\mathbf{m} \equiv \{m_i, i = 1, \dots, N\}$  of the model a heat capacity  $C(\mathbf{m}, T)$  consistent with eq 3, then by integration of  $C(\mathbf{m}, T)$  we will obtain an explicit expression for the effective energy (eq 2 above), so as to have a realistic estimate of the free-energy associated to any given configuration  $\mathbf{m}$ . Such  $C(\mathbf{m}, T)$  will represent the heat capacity associated with the excitation of all the physical degrees of freedom which are still free when fixing the residues to their native/unfolded states (e.g., vibrations of covalent bonds and angles, side-chain movements, solvent degrees of freedom, fast backbone fluctuations in the unfolded regions, etc.). The total heat capacity, derived from the knowledge of the effective energy (eq 2), will contain a contribution related to  $C(\mathbf{m}, T)$ , plus a contribution from the fluctuations among different conformations  $\mathbf{m}$ , as we will see below.

To use eq 3, we need to estimate surface areas. This is easy to do for the native and the fully extended denatured states, that can be mapped to the two extremal states of the model: all  $m_i = 1$  and all  $m_i = 0$ , respectively. The problem, though, is to evaluate the surface areas of all the other configurations, both because we lack detailed geometric information on partially folded structures, and because their number is exponential in  $N$ . Therefore, we introduce a key approximation that  $A_X(\mathbf{m})$ ,  $X = \{P, AP\}$ , is linear in the number of native contacts  $n_X^c(\mathbf{m})$  involving atoms of type  $X$  and varies between the areas of unfolded  $A_X(U)$  and folded  $A_X(N)$  species:

$$A_X(\mathbf{m}) = A_X(U) - \phi_X n_X^c(\mathbf{m}) \quad (4)$$

where  $\phi_X$  is the area density per contact of type  $X$ . Then, resorting to eqs 4 and 7–11 in ref 37 and assuming that the buried surface area of  $\mathbf{m}$  is simply the portion of the total surface area that is not exposed in configuration  $\mathbf{m}$ , we propose the following expression for the heat capacity of any model configuration:

$$C(\mathbf{m}, T) = B'(T) + \sum_{i < j} d_{ij}(T) m_{i,j} + a + b(T - T_0) \quad (5)$$

which preserves the same dependence on  $m_{i,j}$  as eq 2. Here  $T_0$  is a reference temperature and  $B'(T)$  and  $d_{ij}(T)$  are quadratic polynomials whose coefficients are derived from eqs 4 and 7–11 and Table 5 in ref 37. The linear term  $a + b(T - T_0)$  is introduced to account for any experimental concentration errors which could shift the absolute estimate of eq 5, and to account for the difference between constant-pressure and constant-volume heat capacity of the protein solution (see Supporting Information).

To proceed toward eq 2 we have to calculate the enthalpy and entropy of any configuration  $\mathbf{m}$ , which we obtain in the standard fashion by integrating the heat capacity:

$$\mathcal{U}(\mathbf{m}, T) = \mathcal{U}(\mathbf{m}, T_0) + I(T, C) \quad (6)$$

$$\mathcal{S}(\mathbf{m}, T) = \mathcal{S}(\mathbf{m}, T_0) + I(T, C/T) \quad (7)$$

where  $I(T, x)$  indicates the integral of  $x$  between  $T_0$  and  $T$ , easily calculated from eq 5, while the integration “constants” at  $T_0$  actually depend on the configuration  $\mathbf{m}$  in an unknown way. Finding an appropriate expression for such quantities is the second crucial step in our development. To this end, we resort to eqs 4–6 in ref 40, that relate the enthalpic difference between the native and unfolded state  $\Delta H(60^\circ\text{C})$  to the solvation of polar and apolar areas, and their entropic difference  $\Delta \mathcal{S}(T)$  to a solvation and a (residues-dependent) configurational term.

Inspired by those equations, and in order to preserve the same functional dependence on the  $\mathbf{m}$  as in eq 2, we propose the expressions:

$$\mathcal{U}(\mathbf{m}, T_0) = \sum_{i < j} (\varepsilon \Delta_{i,j} + w_{i,j}^0) m_{i,j} \quad (8)$$

$$\mathcal{S}(\mathbf{m}, T_0) = \sum_{i=1}^N \alpha (q_i - q_i^B m_i) - \sum_{i < j} (\tau_{i,j} + \alpha Q_{i,j}) m_{i,j} \quad (9)$$

in which  $w_{i,j}^0$  and  $\tau_{i,j}$  come from the solvation terms and  $q_i$ ,  $q_i^B$ ,  $Q_{i,j}$  are related to the configuration entropy term in ref 40. In addition to those, we have introduced a WSME-like interaction  $\varepsilon \Delta_{i,j}$  and a prefactor  $\alpha$  for the configuration entropy, as a correction to the above phenomenological terms, to account for the fact that the above definitions eqs 8 and 9 are not the only possible choice, and they are derived from approximated phenomenological expressions. The combination of eqs 6 and 7 in  $\mathcal{F}(\mathbf{m}, T) = \mathcal{U}(\mathbf{m}, T) - T\mathcal{S}(\mathbf{m}, T)$ , yields eq 2, where the  $h_{i,j}(T)$  has the form of a cubic polynomial in the temperature plus a  $T \ln T$  term, whose coefficients depend on the adjustable parameters  $a$ ,  $b$ ,  $\varepsilon$ ,  $\alpha$  (see Supporting Information for details). With the explicit knowledge of all the coefficients appearing in eq 2, we are now in the position to study the equilibrium and kinetics of the WSME-S model. Notice that, if we could trust completely our construction and its approximations, and remove the four adjustable parameters that we have introduced as corrections, the above development and the resulting coefficients  $\varphi(T)$ ,  $h_{i,j}(T)$  would contain no free parameter and could be applied to any protein. In the above form, though, we need a way to fix  $\varepsilon$ ,  $\alpha$ ,  $a$  and  $b$  for each protein we study: the natural way to do so is to fit the parameters by matching the model prediction for the heat capacity to the experimental DSC signal for the protein considered, without baseline subtraction (the unfolded and native baselines can be calculated a posteriori from eq 5 setting all  $m_i = 0$  or 1, respectively). Because of the nontrivial temperature-dependence of the energy (eq 2), the predicted heat capacity is the sum of two contributions, the first accounting for the average heat capacity  $\langle C(\mathbf{m}, T) \rangle$  at fixed configuration, and the second for the average of the enthalpy fluctuations  $\langle \Delta \mathcal{U}^2(\mathbf{m}, T) \rangle$  due to probability fluctuations in the configuration space. Such equilibrium averages can be exactly and efficiently calculated with the same techniques used for the standard WSME.<sup>19,20,25</sup>

## METHODS

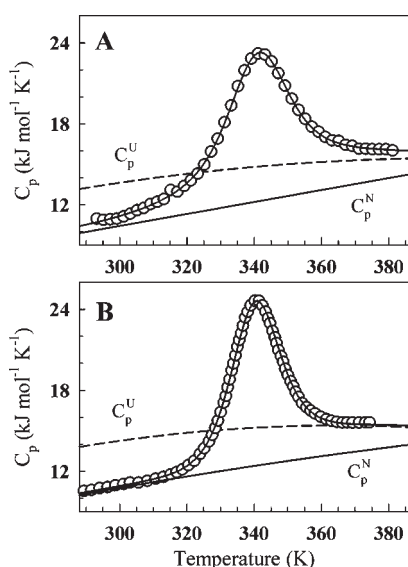
We characterize the equilibrium by considering the following observables besides the heat capacity: the free-energy profiles

$$\mathcal{G}(m, T) = -RT \ln \left( \sum_m' e^{-\beta H_\varepsilon(m, T)} \right) \quad (10)$$

(here  $\sum'$  indicates that the sum is restricted to the configurations with exactly  $m$  native residues) and the probability

$$p_{i,j} = \left\langle \prod_{k=i}^j m_k \right\rangle \quad (11)$$

of finding the region between residues  $i$  and  $j$  completely structured. Single residue probabilities as reported in Figure 2C and Figure 2F are a particular case of the latter equation with  $i = j$ . The fraction of secondary structure content is estimated from the residue probabilities, in conjunction with the available secondary structure assignment of the PDB structures.



**Figure 1.** Fits (continuous lines) to the experimental DSC thermograms (circles) of gpW (A) and SH3 (B). The predicted native (N) and unfolded (U) baselines are also shown.

Folding and unfolding kinetics are studied performing T-jumps to different final temperatures, either from 430 K (for the folding case) or 240 K (for the unfolding). The evolution is simulated either by Monte Carlo (MC) dynamics or by a one-dimensional master equation for the probabilities along the reaction coordinate  $m = \sum_i m_i$ , with rates dictated by the free energy profiles  $\mathcal{G}(m, T)$ . In the former case, an individual residue flip between 0 and 1 is accepted according to a standard Metropolis criterion on the change in the effective energy  $H_e(\mathbf{m}, T)$ . Initial configurations are sampled from the equilibrium distribution at the initial temperature; then, single-molecule evolutions are generated at the final temperature. Groups of 400 trajectories are averaged to obtain the ensemble signal  $p(t) \doteq \langle N^{-1} \sum_{i=1}^N m_i(t) \rangle$ , that is fitted with one- or two-exponential functions (the latter usually giving the best results):

$$p(t) = p_\infty + a_1 e^{-k_1 t} + a_2 e^{-k_2 t} \quad (12)$$

where  $p_\infty$  is calculated resorting to the equilibrium averages  $\langle m_i(t \rightarrow \infty) \rangle = p_{i,j}$  defined in eq 11. The process is repeated 10 times, to estimate average and errors on rates and amplitudes. The second approach resorts to writing a master equation from a diffusion equation for the probabilities  $f(m, t)$ , following the lines of eqs A3 and A5 of ref 41, with a diffusion coefficient  $D$  independent of  $m$ . The system is prepared in the equilibrium distribution at the initial temperature, and at  $t = 0$  the temperature is set to the final one. The evolution of the quantity  $p(t) = \sum_{m=0}^N m f(m, t)$  is followed, and then fitted to the single or double exponential functions above. Experimental amplitudes corresponding to small T-jumps (between  $T$  and  $T + \Delta T$ ) are predicted as the derivatives of the simulated slow phase amplitudes.

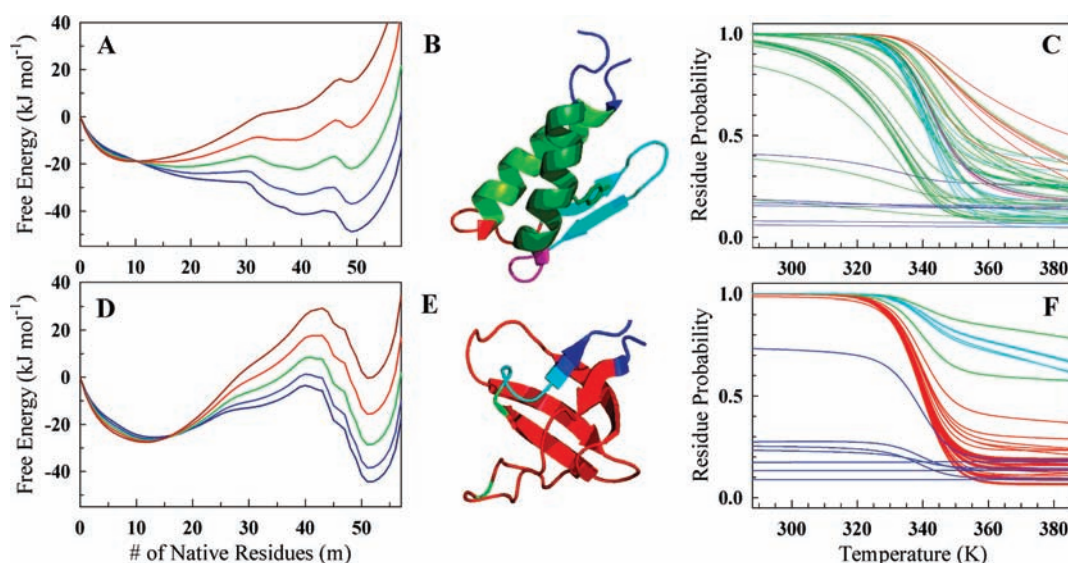
## RESULTS AND DISCUSSION

**Equilibrium.** The WSME-S model has just four adjustable parameters (see Model), that are obtained from a direct least-squares fit of the model heat capacity function to the experimental thermogram (without baseline subtraction). It reproduces the thermograms of both proteins quite accurately and is of comparable quality to a 6-parameter two-state fit (Figure 1). The predicted folded and unfolded baselines for SH3 coincide with those that could be estimated a priori from linear extrapolations of the experimental pre- and post-transition baselines. The behavior is quite different for gpW which has the same peak

temperature as SH3 but only a slightly broader thermogram. The folded and unfolded baselines are slightly downshifted and cannot be guessed from an inspection of the experimental thermogram. As heat capacity is a measure of energy fluctuations,<sup>42,43</sup> this immediately suggests that gpW presents a great plasticity under all conditions, sampling several conformations even at the lowest accessible temperatures, which yields an extra contribution to the heat capacity on top of the native baseline. The presence of large conformational fluctuations in the native-ensemble of gpW has also been predicted previously from a G $\bar{0}$ -model analysis.<sup>44</sup> On the contrary, the ability to identify a clear pre-transition region for SH3 hints at a two-state like behavior with minor residual fluctuations. It is important to note that both proteins have similar sizes (58 and 57 residues for gpW and SH3, respectively) and accessible surface areas (4037 and 3866 Å<sup>2</sup>) thus factoring out their contribution in this comparison. The stark differences in experimental thermograms, predicted baselines, and hence in conformational fluctuations suggest different folding mechanisms at work in these systems. Notice that the standard WSME model, upon the addition of independently estimated baselines, can reproduce the thermograms reasonably well (Figure S5 in Supporting Information), but is not able to reach the same level of quantitative agreement as WSME-S; this is especially true for the gpW case. This in turn affects the magnitude of the predicted barriers as small differences in widths of thermograms translate to large differences in barriers. In other words, to estimate barriers from DSC experiments the fit has to be perfect. In more detail, we observe that the heat capacity of the standard model (with constant parameters  $\epsilon$ ,  $q$ ) at very low and very high temperatures, when the configurational fluctuations fade out, is zero. The larger heat capacity of the unfolded states is therefore not captured well by the standard model (note the difference between experimental high-temperature heat capacity curve and the fit in Supporting Information, Figure S5B for SH3) due to the lack of solvation terms. This also implies that the baselines must be estimated directly from the experimental signal (as in Supporting Information, Figure S5), which is a very difficult task for gpW, for the reasons discussed above. Both these issues are directly addressed in the WSME-S model thus making it superior to earlier versions.

The rich conformational behavior of gpW is further confirmed by the predicted free-energy profiles (Figure 2A and Figure 2D). SH3 conforms to a two-state approximation, with a pronounced barrier at all temperatures, a substantially fixed unfolded minimum and a weak shift of the barrier position according to the Hammond postulate. On the other hand, gpW profiles present three minima separated by two small bumps of less than 1.SRT; either barrier becomes a shoulder at low or high temperatures, and there is a pronounced shift of the unfolded minimum between less and more structured configurations. In other words, gpW presents a weak three-state like behavior that is in fact compatible with downhill folding especially at temperatures either side of the midpoint, in agreement with previous experimental observations.<sup>7</sup> Even if a one-dimensional projection does not guarantee a faithful description of the real free-energy topography a priori, a 7-fold ratio of the barriers at  $T_m$  of the two proteins should be reflected in the rates (see below).

The possibility of calculating exactly the probability that residue  $i$  is native at a given temperature is exploited to make quantitative predictions on which parts of the proteins unfold first (Figure 2C and Figure 2F). In both proteins there are regions, typically at the N- and C-termini, that are little structured at all temperatures; and regions that are likely to preserve some structure even at high

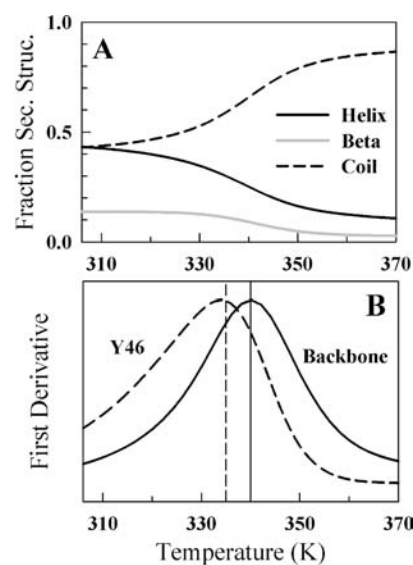


**Figure 2.** Equilibrium behavior of gpW (upper row) and SH3 (lower row). (A, D) The 1D free-energy surfaces at different temperatures of 301–381 K in steps of 20 K (dark blue to dark red; the surfaces in green correspond to the  $T_m$ ). (B, E) Native structures of gpW and SH3. (C, F) Residue unfolding probabilities colored according to the secondary structure and sequence location.

temperatures, as for instance the Proline-rich helical turn before strand 5 of SH3. In SH3 most of the signals can be superimposed almost perfectly, while a great heterogeneity of behaviors is found for gpW. Most importantly, a clear hallmark of cooperativity can be found for SH3 by looking at the midpoint temperature for the individual probabilities, defined as the position of the peak in the first derivative of the signal. Independent of the degree of structure at low or high temperatures, their  $T_m$  are spread within 1.5 K around the mean  $T_m$ , pointing toward a collective unfolding of all residues in the protein. On the contrary, individual residues in gpW behave quite independently from one another, and the spread in their midpoint temperatures is around 11 K (see Supporting Information Figure S3), something that can in principle be tested by performing an atom-by-atom analysis using NMR techniques.<sup>45</sup> The consistency between the magnitude of barriers and the spread in melting temperatures supports the use of multiprobe experiments to distinguish between folding mechanisms.<sup>45,46</sup>

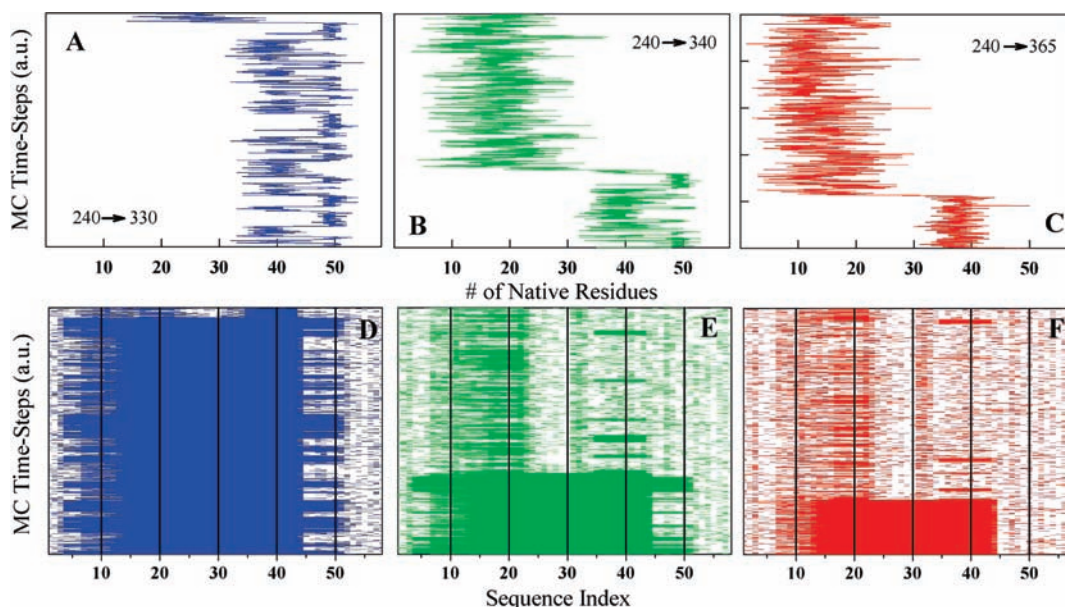
The recent experimental study on gpW unfolding<sup>7</sup> points to  $\sim 5$  K difference in  $T_m$  when monitored by fluorescence (335 K) and far-UV CD (340 K), two different spectroscopic probes that report on the environment around tyrosine and the secondary structure content, respectively. Figure 3A plots the fraction of secondary structure as a function of temperature as predicted by the model. The loss of helical secondary structure can be approximated as a predictor of the far-UV CD signal at 222 nm, the derivative of which agrees nicely with the experimental estimate of 340 K (Figure 3B). Moreover, it is possible to reproduce the lower midpoint temperature of 335 K for tyrosine 46 from just the residue probability. The ability of the model to predict spectroscopic observables that were not used in the fitting procedure further highlights the robustness of the method.

**Single-Molecule Behavior.** Using the parameters obtained from the fit to the DSC data, we simulate the relaxation kinetics after a temperature-jump (see Methods). Figure 4 reports representative single-molecule trajectories after a simulated T-jump to different final temperatures both in terms of the average behavior (Figure 4A–C) and at the residue-level (Figure 4D–F). Though single-molecule trajectories can differ



**Figure 3.** Predicting spectroscopic signals of gpW folding. (A) Temperature dependence of the secondary structure content from the model. (B) First derivative of the melting curves from the model compared to the experimental midpoint temperature (vertical lines) from fluorescence of tyrosine 46 (dashed curves) and far-UV CD at 222 nm that monitors the helical content (continuous curve).

significantly from one another in the same experimental conditions, Figure 4 clearly suggests the succession of the unfolding events, and the parts of the protein involved. We can see that after a T-jump the average fraction of native residues shifts from the native basin ( $\sim 50$  native residues) to an intermediate value ( $\sim 40$  residues). Then, after a stochastic unfolding time, the system reaches the unfolded state, whose amount of residual structure strongly depends on the temperature, as expected from the free-energy profiles (see Figure 2). Moreover, the conformational behavior before unfolding is strongly temperature-dependent: at  $T_m = 340$  K there is a pre-equilibration with frequent oscillations



**Figure 4.** Single-molecule behavior of gpW from MC simulations. Three representative trajectories are reported, at different final temperatures upon  $T$ -jumps from 240 K: (A–C) average number of native residues as a function of time; (D–F) corresponding detailed distributions of native and unfolded residues along the protein, as functions of time. Structured regions appear as colored blocks.

between the native and intermediate fractions and at  $T = 330$  these two states appear as better resolved, and the oscillations between them can be considered as representative of the equilibrium. On the contrary, at  $T = 365$  K, well inside the unfolding region, the jump to the intermediate fraction is immediate, and the protein hardly visits the native basin again before reaching the unfolded state.

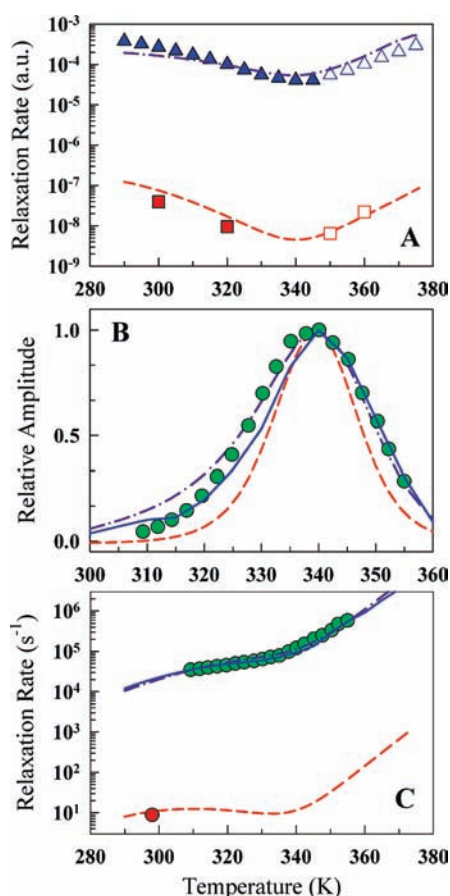
From a structural point of view, the bottom panels of Figure 4 reveal that the native basin corresponds to a core spanning all regions of the gpW secondary structure, from the beginning of the first helix to the end of the second. The intermediate presents a core structure encompassing residues 13 to 44 (from the C-terminal part of the first helix to the N-terminal half of the second helix), though the structure can include most of the first helix, depending on the temperature. Considering the WSME model's intrinsic tendency to enhance cooperativity,<sup>18,47</sup> this picture is consistent with a native-basin that moves with temperature which in the 1D projection appears as a two wells separated by a small barrier. The unfolded basin is characterized by the loss of  $\beta$ -strand structure, and presents non-negligible residual native-like structure in all the loops and in the C-terminal part of the first helix.

**Kinetics.** Single-molecule analysis is very helpful to understand mechanisms, but owing to the essential randomness of each trajectory it does not allow extrapolations to ensemble properties. To compare the predicted rates to the experimental ones, we average ensembles of 400 single-molecule trajectories (see Methods). The evolution of the average native fraction at different final temperatures reveals the existence of two phases for gpW: a slow one defining the folding/unfolding time and corresponding to the final relaxation to equilibrium, and a faster one, weakly dependent on the final temperature, and  $\sim 2$  orders of magnitude faster than the former (see Supporting Information Figure S4). The fast rate corresponds to an initial rearrangement toward intermediate values of the reaction coordinate, and is less significant than the slower one for the reasons discussed below. The slower relaxation rates of gpW from MC kinetics display a

shallow chevron-like behavior (Figure 5A). The rates estimated from a diffusive analysis on the free-energy surfaces of gpW agree reasonably well with MC results both in the shape of the rate versus temperature plot and in the observation of two phases. For SH3, MC kinetics was performed only at a few chosen temperatures and with a reduced ensemble (200 molecules), as it involved very long simulation times. The resulting MC rates agree quite well with those from the diffusive analysis, revealing a significantly slower folding and steeper chevron (just a single phase was observed).

The comparison of the normalized experimental amplitudes of gpW<sup>7</sup> (Figure 5B) with those from both computational methods shows a remarkable agreement justifying that the relevant kinetic information is contained in the slower phase. It also provides a stringent check to the intrinsic consistency of the model; that is, after fitting the model parameters to an equilibrium signal, we are able to make correct predictions on the kinetics. The predicted amplitude for SH3 is sharper than that of gpW, in tune with the experimental observation of a sharper thermogram. The same holds true, though in a weaker sense, for the rescaled rates (Figure 5C). To compare with experimental results,<sup>7,38</sup> we scale the predicted rates with a phenomenological Arrhenius-like temperature dependence on the diffusion coefficient with an activation term of  $\sim 1$  kJ/mol per residue for the diffusive analysis and  $\sim 1.2$  kJ/mol per residue for the MC kinetics.<sup>48</sup> This method also reproduces the observed relaxation rate of SH3 at 298 K without invoking further assumptions.

The results from the kinetic analysis of the two proteins have several important consequences. First, the agreement between MC and diffusive analysis suggests that the number of native residues  $m$  is a good reaction coordinate for this protein. Second, the diffusion along  $m$ , with coordinate-independent diffusion coefficient, describes well the motion along the optimal pathway. Third, broad thermograms and hence small barriers result in a signature behavior on both the kinetics (observed as a shallow chevron) and the amplitudes (transition spans a broader



**Figure 5.** Kinetic analysis. (A) Temperature dependence of the slower rates of gpW (filled triangles, high- $T$  to low- $T$  jumps; open triangles, low- $T$  to high- $T$  jumps) and SH3 (filled and open squares). The rates from a 1D diffusive analysis on the free-energy surfaces for gpW and SH3 are also shown (dash-dotted and dashed lines, respectively). (B) gpW experimental amplitudes from 11 K  $T$ -jumps (circles) together with predicted amplitudes, following the same color code as panel A. The continuous line is the MC-kinetics result. (C) Scaled rates vs temperature from different schemes for gpW and SH3 following the color code of panel B. Filled red circle, experimental rate of SH3 at 298 K.

temperature range). Finally, both single-molecule and kinetic analysis (Figure 4 and Figure 5) suggest that the fast rate in gpW is associated to a rearrangement of the population in the vicinity of the initial basin, affecting the shallow intermediate; this in turns agrees with the picture emerging from Figure 2A that reveals great plasticity of both the folded and unfolded basins upon temperature changes. The fact that the faster rate was not observed in experiments suggests that this may be a model-specific behavior, hinting that the estimated barrier for gpW is an upper estimate.

## CONCLUSIONS

We have developed an extension of the WSME model with a twofold goal: on the one hand, we wanted to see how much we could “push” this simple model toward quantitative agreement with experimental results, and to provide the experimentalists with a simple and quantitative tool to interpret their data. On the other hand, we wanted to apply the model to the difficult case of the analysis of a weakly cooperative protein, to which the standard models are of doubtful applicability, and to compare the performance and predictions with those obtained for a standard two-

state protein. It is important, therefore, to comment on the scope and limits of the present approach, while reviewing the main results obtained.

The model neglects non-native interactions, so that its application to proteins where a specific non-native interaction represents a crucial step in the folding pathway is questionable from the very beginning. However, it is well recognized that Gō-models give reliable predictions in general, and that the introduction of nonspecific, non-native interactions of moderate strength does not affect the overall predictions (see ref 49 for a recent review of the literature on the subject). This supports the view that the funnel paradigm holds and that native interactions play a fundamental role in determining mechanisms, while non-native interactions just “fine-tune” the rates or the cooperativity. We feel that this should hold good for the WSME-S model, as well.

The main strength of the model, that makes it outstanding among similar ones, is that every equilibrium average can be calculated exactly, in less than a second; predictions that would require decades of computer time with molecular dynamics or Monte Carlo simulations, that employ more realistic atomic potentials, can be made within a minute on a desktop PC. However, such strength is intimately related to the main weakness of the model: the fact that it just considers interactions within a native island of continuous residues. This approximation is expected to become more problematic for proteins where the folding nucleus is sparse along the chain, and entails the formation of nonlocal contacts, encompassing sequence regions that are still substantially unstructured; so, exceptions apart, we expect that the model will have worse performance for longer proteins, where it is more likely that nonlocal contacts enclosing unfolded regions have an important role in the folding process. We are studying possible improvements to the model to eliminate such limitation, but for the moment we suggest that WSME-like models should not be used to interpret data when there is strong experimental evidence for nonlocal interactions to play a key role in the folding mechanism of a protein.

Despite the above considerations, it is evident that the extended WSME-S model and the resulting analysis presented here provide a surprising amount of information on the folding mechanisms of two small proteins of similar size and thermodynamic stability. It is important to note that the model was used to fit only the DSC data of gpW and SH3, which carry information on the partition function of the system under study: the rest of our results are predictions, some of which need to be experimentally verified. In parallel, we have shown that the various experimental criteria empirically developed as evidence for downhill folding—broad thermograms, dispersion in melting temperatures from equilibrium probes, flat chevrons and hence weak temperature dependence of relaxation rates, broader kinetic amplitudes—emerge naturally from a predictive model, in those cases where at most marginal barriers are present, and therefore highlight the importance of a quantitative analysis employing statistical models, and calls for further improvements in this direction.

## ASSOCIATED CONTENT

**S Supporting Information.** A more detailed presentation of the theory is presented, together with the values of the model parameters, and figures corresponding to the predictions of the standard WSME model, dispersion of the midpoint temperatures

for individual residues, and the fast and slow relaxation rates of gpW. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## AUTHOR INFORMATION

### Corresponding Author

pier@unizar.es; anarayan@bsc.es

## ACKNOWLEDGMENT

P.B. acknowledges E. Freire for giving him access to his program to calculate ASAs, and A. Velázquez for illuminating explanations on DSC techniques. P.B. is supported by Spanish (MICINN) Grant FIS2009-13364-C02-01. A.N.N. is supported by the Juan de la Cierva fellowship from the Spanish Ministry of Science and Innovation. The numerical calculations were run with in-house software on the BIFI computer cluster.

## REFERENCES

- (1) Jackson, S. W. M.; Brandts, J. F. *Biochemistry* **1970**, *9*, 2294–2301.
- (2) Jackson, S. E. *Fold. Des.* **1998**, *3*, R81–R91.
- (3) Garcia-Mira, M. M.; Sadqi, M.; Fischer, N.; Sanchez-Ruiz, J. M.; Muñoz, V. *Science* **2002**, *298*, 2191–2195.
- (4) Li, P.; Oliva, F. Y.; Naganathan, A. N.; Muñoz, V. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 103–108.
- (5) Liu, F.; Gruebele, M. J. *Mol. Biol.* **2007**, *370*, 574–584.
- (6) Yang, W. Y.; Gruebele, M. *Nature* **2003**, *423*, 193–197.
- (7) Fung, A.; Li, P.; Godoy-Ruiz, R.; Sanchez-Ruiz, J. M.; Muñoz, V. *J. Am. Chem. Soc.* **2008**, *130*, 7489–7495.
- (8) Naganathan, A. N.; Li, P.; Perez-Jimenez, R.; Sanchez-Ruiz, J. M.; Muñoz, V. *J. Am. Chem. Soc.* **2010**, *132*, 11183–11190.
- (9) Taketomi, H.; Ueda, Y.; Go, N. *Int. J. Protein Pept. Res.* **1975**, *7*, 445–459.
- (10) Muñoz, V.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11311–11316.
- (11) Alm, E.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11305–11310.
- (12) Wako, H.; Saito, N. *J. Phys. Soc. Jpn.* **1978**, *44*, 1931–1938.
- (13) Wako, H.; Saito, N. *J. Phys. Soc. Jpn.* **1978**, *44*, 1939–1945.
- (14) Muñoz, V.; Thompson, P. A.; Hofrichter, J.; Eaton, W. A. *Nature* **1997**, *390*, 196–199.
- (15) Kubelka, J.; Henry, E. R.; Cellmer, T.; Hofrichter, J.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 18655–18662.
- (16) Henry, E. R.; Eaton, W. A. *Chem. Phys.* **2004**, *307*, 163–185.
- (17) Chung, H. S.; Tokmakoff, A. *Proteins* **2008**, *72*, 488–97.
- (18) Nelson, E. D.; Grishin, N. V. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 1489–93.
- (19) Bruscolini, P.; Pelizzola, A. *Phys. Rev. Lett.* **2002**, *88*, 258101.
- (20) Pelizzola, A. *J. Stat. Mech., Theory Exp.* **2005**, 11010, 11010.
- (21) Zamparo, M.; Pelizzola, A. *J. Stat. Mech., Theory Exp.* **2006**, P12009.
- (22) Zamparo, M.; Pelizzola, A. *Phys. Rev. Lett.* **2006**, *97*, 068106.
- (23) Imparato, A.; Pelizzola, A.; Zamparo, M. *Phys. Rev. Lett.* **2007**, *98*, 148102.
- (24) Bruscolini, P.; Pelizzola, A.; Zamparo, M. *Phys. Rev. Lett.* **2007**, *99*, 038103.
- (25) Bruscolini, P.; Pelizzola, A.; Zamparo, M. *J. Chem. Phys.* **2007**, *126*, 215103.
- (26) Imparato, A.; Pelizzola, A. *Phys. Rev. Lett.* **2008**, *100*, 158104.
- (27) Zamparo, M.; Pelizzola, A. *J. Chem. Phys.* **2009**, *131*, 035101.
- (28) Caraglio, M.; Imparato, A.; Pelizzola, A. *J. Chem. Phys.* **2010**, *133*, 065101.
- (29) Itoh, K.; Sasai, M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 13865–13870.
- (30) Itoh, K.; Sasai, M. *J. Chem. Phys.* **2009**, *130*, 145104.
- (31) Itoh, K.; Sasai, M. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 7775–7780.
- (32) Faccin, M.; Bruscolini, P.; Pelizzola, A. *J. Chem. Phys.* **2011**, *134*, 075102.
- (33) Makhatazde, G. I.; Privalov, P. L. *Adv. Protein Chem.* **1995**, *47*, 307–425.
- (34) Sanchez-Ruiz, J. M. In *Subcellular Biochemistry*; Biswas, B. R. S., Ed.; Plenum Press: New York, 1995; Vol. 24, pp 133–176.
- (35) Robertson, A. D.; Murphy, K. P. *Chem. Rev.* **1997**, *97*, 1251–1267.
- (36) Freire, E. In *Protein Stability and Folding. Theory and Practice*; Shirley, B. A., Ed.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 1995; Vol. 40; p 191.
- (37) Gómez, J.; Hilser, V. J.; Xie, D.; Freire, E. *Proteins* **1995**, *22*, 404–412.
- (38) Viguera, A. R.; Martinez, J. C.; Filimonov, V. V.; Mateo, P. L.; Serrano, L. *Biochemistry* **1994**, *33*, 2142–2150.
- (39) Muñoz, V.; Henry, E. R.; Hofrichter, J.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5872–5879.
- (40) Luque, I.; Mayorga, O. L.; Freire, E. *Biochemistry* **1996**, *35*, 13681–13688.
- (41) Lapidus, L. J.; Steinbach, P. J.; Eaton, W. A.; Szabo, A.; Hofrichter, J. *J. Phys. Chem. B* **2002**, *106*, 11628–11640.
- (42) Cooper, A. *Proc. Natl. Acad. Sci. U.S.A.* **1976**, *73*, 2740–2741.
- (43) Muñoz, V.; Sanchez-Ruiz, J. M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 17646–17651.
- (44) de Sancho, D.; Rey, A. *J. Comput. Chem.* **2008**, *29*, 1684–1692.
- (45) Sadqi, M.; Fushman, D.; Muñoz, V. *Nature* **2006**, *442*, 317–321.
- (46) Naganathan, A. N.; Muñoz, V. *Biochemistry* **2008**, *47*, 6752–6761.
- (47) Abe, H.; Wako, H. *Phys. Rev. E* **2006**, *74*, 011913.
- (48) Naganathan, A. N.; Doshi, U.; Muñoz, V. *J. Am. Chem. Soc.* **2007**, *129*, 5673–5682.
- (49) Hills, R. D.; Brooks, C. L. *Int. J. Mol. Sci.* **2009**, *10*, 889–905.