Effect of hydrogen bond networks on the nucleation mechanism of protein folding

Y. S. Djikaev* and Eli Ruckenstein†

Department of Chemical and Biological Engineering, SUNY at Buffalo, Buffalo, New York 14260, USA

(Received 30 July 2009; revised manuscript received 23 October 2009; published 29 December 2009)

We have recently developed a kinetic model for the nucleation mechanism of protein folding (NMPF) in terms of ternary nucleation by using the first passage time analysis. A protein was considered as a random heteropolymer consisting of hydrophobic, hydrophilic (some of which are negatively or positively ionizable), and neutral beads. The main idea of the NMPF model consisted of averaging the dihedral potential in which a selected residue is involved over all possible configurations of all neighboring residues along the protein chain. The combination of the average dihedral, effective pairwise (due to Lennard-Jones-type and electrostatic interactions), and confining (due to the polymer connectivity constraint) potentials gives rise to an overall potential around the cluster that, as a function of the distance from the cluster center, has a double-well shape. This allows one to evaluate the protein folding time. In the original NMPF model hydrogen bonding was not taken into account explicitly. To improve the NMPF model and make it more realistic, in this paper we modify our (previously developed) probabilistic hydrogen bond model and combine it with the former. Thus, a contribution due to the disruption of hydrogen bond networks around the interacting particles (cluster of native residues and residue in the protein unfolded part) appears in the overall potential field around a cluster. The modified model is applied to the folding of the same model proteins that were examined in the original model: a short protein consisting of 124 residues (roughly mimicking bovine pancreatic ribonuclease) and a long one consisting of 2500 residues (as a representative of large proteins with superlong polypeptide chains), at *p*H =8.3, 7.3, and 6.3. The hydrogen bond contribution now plays a dominant role in the total potential field around the cluster (except for very short distances thereto where the repulsive energy tends to infinity). It is by an order of magnitude stronger for hydrophobic residues than for hydrophilic ones. The range of "residue-cluster" distances, at which the hydrogen bond effect exists, is twice as long for hydrophobic residues as for hydrophilic ones.

DOI: 10.1103/PhysRevE.80.061918 PACS number(s): 87.15.Cc, 87.15.hm, 82.30.Rs, 82.70.Uv

I. INTRODUCTION

A well-defined three-dimensional structure [1,2] is necessary for a protein molecule to carry out a specific biological function. The formation of the native structure of a biologically active protein constitutes the core of the so-called "protein folding problem" [3,4]. Many thermodynamic and kinetic aspects of the protein folding process remain unexplained [5–8].

It is believed that initially an unfolded protein transforms quickly into a compact (but not native) configuration [9,10]. One of the pathways for the transition from a compact configuration to the native one is similar to nucleation, i.e., once a critical number of (native) tertiary contacts is established the native structure is formed rapidly without passing through any detectable intermediates [6,9–12].

We have recently developed [13,14] a model for the nucleation mechanism of protein folding (NMPF) in terms of ternary nucleation by using the first passage time analysis. A protein was considered as a random heteropolymer [15–17] consisting of hydrophobic, hydrophilic, and neutral beads with all the bonds in the heteropolymer having the same constant length and all the bond angles equal and fixed . The ionizability of some of protein residues (aspartic and glutamic acids, lysine, arginine, and histidine) was taken into account by considering the hydrophilic residues to be of three subtypes, namely, those which cannot be ionized at all and those which can (depending on the *p*H of the surrounding solution) be either positively or negatively ionized.

The main idea underlying the NMPF model consists of representing the overall potential field around a cluster of native protein residues (i.e., field in which a non-native residue performs a chaotic motion) as a combination of the mean dihedral potential, effective pairwise potential, and confining potential. The latter is due to the polymer connectivity that confines the protein residues around its center. The effective pairwise potential (for pairwise interactions of a selected residue with those in the cluster) for an ionized residue contains an electrostatic contribution. The mean dihedral potential (in which a selected residue is involved) is calculated by averaging it over all possible configurations of all neighboring residues along the protein chain. As a function of the distance from the cluster center, the overall potential field has a double-well shape, which allows one to develop a self-consistent kinetic theory for the nucleation mechanism of protein folding and evaluate its characteristic time (as well as the temperature dependence of the latter).

In that model [13,14] hydrogen bonding (hb) was taken into account indirectly through its effect on the diffusion coefficients of protein residues. On the other hand, hydrogen bonding plays a crucial role in the formation, stability, and denaturation of the native structure of a biologically active protein, i.e., its folding, stability, and unfolding [1,3,4,18,19]. In addition to hydrogen bonds between the atoms of the

*idjikaev@buffalo.edu

†Corresponding author. FAX: (716) 645-3822; feaeliru@buffalo.edu

polypeptide backbone (main-chain–main-chain N-H···O bonds), which are responsible for the formation of the secondary structure of proteins (both α helices and β sheets), many hydrogen-bonded interactions are provided by the polar groups of the side chains. Moreover, the biological activity of proteins appears to depend on the formation of a two-dimensional hydrogen-bonded network spanning most of the protein surface and connecting all the surface hydrogen-bonded water clusters [20–22].

In order to improve our kinetic model for protein folding, it appeared logical to find a way for explicitly taking into account the hydrogen bonding "water-water" and "water-protein residue." As a first step, we developed a probabilistic model for the effect of hydrogen bond networks of water molecules around two solute particles (immersed in water) on their interaction [23,24]. The probabilistic hydrogen bond (PHB) approach was applied to the solvent-mediated interaction of (a) two spherical hydrophobic solutes [23] and (b) two infinite parallel plates whereof the surfaces facing each other have a composite hydrophobic-hydrophilic character [24] (i.e., they are covered with uniformly distributed hydrophobic and hydrophilic sites).

When two hydrophobic particles sufficiently approach each other, the disruption of boundary water-water hydrogen bonds in their first hydration layers can give rise to an additional contribution to their overall interaction potential. According to numerical evaluations, in the interplay between a decrease in the number of boundary bonds per water molecule (as a result of the proximity to the foreign hydrophobic particle) and the possible enhancement [25–27] of such a bond the former effect is predominant because the larger number of weaker bulk hydrogen bonds provide a more negative contribution to the free energy than the smaller number of stronger boundary hydrogen bonds (BHBs). Consequently, our model suggests that the disruption of the boundary hydrogen bonds, which occurs when the first two hydration shells of two particles overlap, results in an attractive contribution between the particles. This attraction is naturally short range, appearing only when the separation between two particles becomes smaller than four lengths of a hydrogen bond.

To implement the PHB approach into the NMPF model, it is necessary to slightly modify the former to adapt it to the particular situations encountered on the nucleation pathway of protein folding. Indeed, the folded cluster of the protein consists of three kinds of residues; hence, its surface can be expected to have a composite (hydrophobic-hydrophilic) character. On the other hand, a single residue in the unfolded part of the protein is either hydrophobic or hydrophilic (neutral residues can be treated as hydrophobic as far as their hydrogen-bonding ability is concerned). Thus, the PHB model needs to be modified to examine the solvent-mediated interaction of a spherical particle of composite nature (modeling a folded cluster of a protein) with (1) a spherical hydrophobic particle (modeling hydrophobic and neutral protein residues) and (2) a spherical hydrophilic particle (modeling hydrophilic protein residues). The additional contributions to the interaction potentials, arising due to the disruption of hydrogen bond networks around the interacting particles, will then have to be added to the overall potential

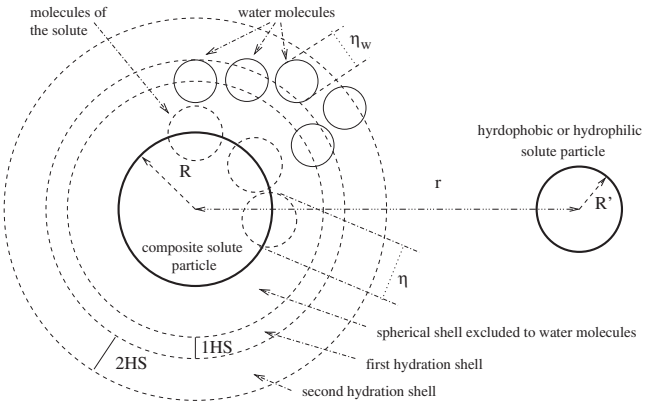

FIG. 1. A schematic representation of the first two hydration layers two spherical solutes of radii R and R' at a distance r between their centers. The surface of each particle is shown as a thick solid line. The circles of diameter η_w represent water molecules. The dashed circle of radius η represents a molecule of the solute particle.

field around a cluster in the NMPF model. As a result, the water-water hydrogen bonding will be taken *explicitly* into account in a kinetic model for the nucleation mechanism of protein folding.

The paper is structured as follows. In Sec. II we extend the PHB model to the cases where the spherical interacting particles have different nature (one composite the other either hydrophobic or hydrophilic) in addition to being of different radii. In Sec. III we present a modified version of our NMPF model based on the first passage time analysis in the framework of a ternary nucleation theory. The modified NMPF model will explicitly implement the effect of water-water hydrogen bond network by means of the PHB model. The numerical results of the application of the modified NMPF model to the folding of two model proteins are presented in Sec. IV. The results are discussed and conclusions are summarized in Sec. V.

II. PROBABILISTIC MODEL FOR THE EFFECT OF WATER-WATER HYDROGEN BONDING ON THE INTERACTION OF SOLUTE PARTICLES

Let us consider two solute particles (1 and 2) of spherical shape in water. The radii of the particles will be denoted by R and R' and the distance between them (i.e., between their centers) by r (Fig. 1). The radii of the particles are determined by the locus of the outermost molecules that constitute them. If the smaller solute consists of a single molecule, its radius R' is equal to zero. (This case would represent a single residue in the unfolded part of the protein.) The characteristic distance of pairwise interactions between molecules constituting particles R and R' will be denoted by η.

In order to make the model applicable to the problem of protein folding, we will consider the particle R to have a composite hydrophobic-hydrophilic character so that its surface is covered with both hydrophobic and hydrophilic sites uniformly distributed over the whole surface. The smaller particle R' will be either completely hydrophobic or com-

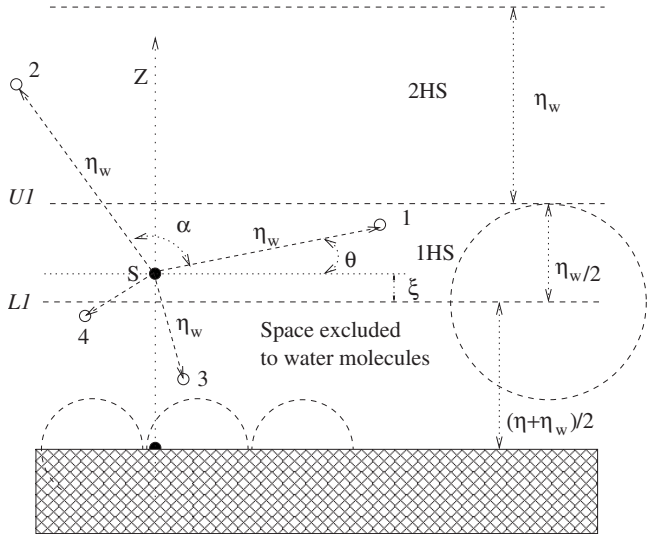

FIG. 2. A schematic representation of a water molecule in the first hydration shell (1HS) of the planar substrate surface (shadowed area). The lower and upper boundaries of the 1HS are marked as the planes $L1$ and $U1$ and shown as long-dashed lines. The molecule, shown as a disk S, is at the distance ξ from the lower boundary (closest to the hydrophobic surface) of 1HS. The four hb arms of the molecule are shown as short-dashed lines with the empty circles as the arm tips. Arms 1 and 2 are in the plane of the figure, whereas arms 3 and 4 are located out of the figure plane (one of them under it, the other above it). The angle between any two hb arms is α. The angle Θ is the angle formed between the hb arm and its tangential projection (parallel to the lower boundary of the 1HS). In this figure, the origin of the Cartesian coordinate system lies at the hydrophobic surface with the z axis being normal thereto. It is assumed that $-\pi/2 \leq \Theta \leq \pi/2$ with $\Theta < 0$ if the z coordinate of the hb-arm tip is greater than $(\eta+\eta_w)/2+\xi$ and $\Theta > 0$, otherwise. The large dashed circle represents a water molecule, while the smaller dashed semicircles represent the substrate molecules.

pletely hydrophilic. A water molecule is capable of forming a hydrogen bond with a hydrophilic site of the surface but not with a hydrophobic one. The probabilities ω_b and ω_l that a selected water molecule, adjacent to the particle R, will be in contact with hydrophobic or hydrophilic sites, respectively, are given by

$$\omega_b = \frac{A_b}{A}, \quad \omega_l = \frac{A_l}{A}, \tag{1}$$

where A_b and A_l are the total areas covered by hydrophobic and hydrophilic sites, respectively, and $A=A_b+A_l$ is the total surface area of the solute. Clearly, $\omega_b+\omega_l=1$.

The location of a water molecule will be determined by the location of its center. The length of a water-water hydrogen bond (i.e., the distance between the centers of two bonded water molecules) will be denoted by η_w. Water molecules located in the layer of thickness $\frac{1}{2}\eta_w$ at distances to the particle surface from $\frac{1}{2}(\eta+\eta_w)$ to $\frac{1}{2}(\eta+\eta_w)+\frac{1}{2}\eta_w$ will be considered to belong to the first hydration shell (hereafter referred to as 1HS; see Figs. 1 and 2). The second hydration shell (hereafter referred to as 2HS) of thickness η_w is formed by water molecules at distances to the particle surface from $\frac{1}{2}(\eta+\eta_w)+\frac{1}{2}\eta_w$ to $\frac{1}{2}(\eta+\eta_w)+\frac{3}{2}\eta_w$. Hydrogen bonding being short ranged, the other layers are not affected by the presence of the particle surface.

Clearly, the properties of a hydrogen bond between a water molecule and a hydrophilic site on the solute may depend on the nature of the site. In a rough approximation, however, its length will be considered to be the same for all hydrophilic sites and equal to η. The characteristic length of pairwise interactions between a water molecule and that of the hydrophobic particle will be assumed to be equal to $\frac{1}{2}(\eta+\eta_w)$.

A water molecule itself is modeled to have four arms each capable of forming a single hydrogen bond. The configuration of the four hb arms is completely symmetric with the angle between any of them equal to $\alpha=109.47°$. This tetrahedral configuration is assumed to be rigid (independent of whether the water molecule is located in the bulk or in the 1HS). Each hb arm can adopt a continuum of orientations subject to the constraint of tetrahedral rigidity. A water molecule forms a hydrogen bond with another molecule when the tip of any hb arm of the first molecule exactly coincides with the second one. The length of an hb arm is thus equal to the length of a hydrogen bond, η_w.

The water-water hydrogen bonds are treated along the lines of the Müller-Lee-Graziano (MLG) model [28,29]. Some of the hb arms of water molecules in the 1HS cannot form bonds because of the proximity to the hydrophobic regions of the solute. The bonds that such molecules do form (hereafter referred to as "boundary hydrogen bonds") can be somewhat stronger than the bulk ones [25,29], although such an enhancement is still a subject of discussion (see, e.g., Ref. [27] and references therein). However, some 1HS water molecules can form hydrogen bonds with the hydrophilic sites of the particle. The hydrogen bonds of such molecules are not altered compared to the bulk water bonds. Nevertheless, adopting a probabilistic approach one can consider the whole network of hydrogen bonds involving all 1HS molecules as a BHB network. A water-water hydrogen bond is enhanced (compared to its bulk value) if at least one of the two water molecules belongs to the 1HS and it does not form a hydrogen bond with the surface of the solute. The probability of the latter is ω_b, while the probability that a 1HS water molecule forms a hydrogen bond with the solute is ω_l. Therefore, one can assume that the properties of such a composite hydrogen bond network involving 1HS molecules can be obtained by averaging the properties of the corresponding networks for purely hydrophobic and purely hydrophilic particles with weights ω_b and ω_l, respectively.

To some extent, our model for water-water hydrogen bonding is a selective combination of a three-dimensional analog of the Mercedes-Benz (MB) water model [25] with the MLG model [28,29]. In the original MB model [25] (which is two dimensional) the water molecules are modeled as Lennard-Jones (LJ) disks in a donor-acceptor approximation, with three hydrogen bonding arms (whereof the completely symmetric configuration resembles the Mercedes-Benz logo, hence the name of the model). The interaction potential between two water molecules is the sum of two terms, with one representing the LJ interaction of disks and the other representing their hydrogen bonding ability. Al-

though the latter is considered to be continuous, a hydrogen bond is modeled to be optimal at a specified distance and a relative orientation of the two molecules involved: if at this distance one hb-forming arm of one molecule aligns itself with a hb arm of the other molecule, then the hydrogen bond energy has its minimum value. Deviations from this lowest-energy hydrogen bond configuration (in distance and mutual orientation) are assumed to have a Gaussian distribution with a single width parameter for all degrees of freedom. Although the MB model allows continuous variations of the separation and orientation of the water molecules (disks), it is also consistent with the concept of bimodal character of the energetics of hydrogen bonds in the MLG model [28,29]. The discrete three-dimensional model presented in Refs. [23,24] can be regarded as a particular case of a three-dimensional version of the hydrogen bonding feature of the MB model where the Gaussian distribution would be infinitely narrow. However, the analytical treatment that we pursued in our model would be much more difficult to carry out if we considered the hydrogen bonding ability of water molecules to have a continuous (with respect to the separation and mutual orientation of molecules involved) character.

A. Networks of water-water hydrogen bonds around solutes

Let us now examine how the BHB network affects the pairwise interaction potential Φ between two particles. Clearly, this effect can be neglected if the second hydration layers (2HS) around the two solutes do not overlap. However, if the two particles, R and R', are sufficiently close to each other, they have to share some parts of their BHB networks. The overlap of the 2HS of one particle with the 2HS of the other leads to a decrease in the total number of 2HS molecules but does not affect the total number of 1HS molecules. The latter decreases only as a result of the overlap of the 1HSs of the two particles. The overlap of the BHB networks of the two particles thus causes their mutual disruption.

If the distance r between the two particles (i.e., between their centers) is greater than $\tilde{r} \equiv R+R'+\eta+4\eta_w$, there is no overlapping of the first two hydration shells of the two particles so that they do not share their BHB networks and there is no contribution to the potential $\Phi=\Phi(r)$. However, when the particles are sufficiently close to each other, so that $r<\tilde{r}$, they share some parts of their first two hydration shells. Thus, with decreasing r (at $r<\tilde{r}$), the total volume of the first two hydration shells decreases, which leads to the decrease in the total number of molecules in these shells. This, in turn, results in a decrease in the total number of boundary hydrogen bonds in the first two hydration shells of the two particles, hereafter denoted by $N_s \equiv N_s(r)$. The total number of BHBs in the first two hydration shells of the two solutes at a distance r between them will be denoted by $\nu_s \equiv \nu_s(r)$. A decrease in N_s results in a decrease in ν_s (both occurring because of the overlapping of the first two hydration shells of the two particles): the corresponding quantities are given by $N_s^{r^\infty} \equiv N_s^{r^\infty}(r)=N_s(\infty)-N_s(r)$ and $\nu_s^{r^\infty} \equiv \nu_s^{r^\infty}(r)=\nu_s(\infty)-\nu_s(r)$, respectively. Clearly, $\nu_s^{r^\infty}(r)=0$ and $N_s^{r^\infty}(r)=0$ for $r \geq \tilde{r}$, while $\nu_s^{r^\infty}(r)>0$ and $N_s^{r^\infty}(r)>0$ for $r<\tilde{r}$.

The $N_s^{r^\infty}$ molecules which left the 1HS and 2HS of the two solutes pass into the bulk water where they can form new hydrogen bonds with the energy $\epsilon_b<0$ per bond. The average number of bonds per molecule of bulk water is denoted by n_b. The total energy Φ_b of the newly formed bonds is a function of r (because so is $N_s^{r^\infty}$): $\Phi_b \equiv \Phi_b(r)=\epsilon_b n_b N_s^{r^\infty}(r)$. Denoting the number density of water molecules in the 1HS and 2HS by ρ_w, one can rewrite Φ_b as

$$\Phi_b(r) = \epsilon_b n_b \rho_w V_o(r), \tag{2}$$

where $V_o(r)$ is the volume of the region resulting from the overlap of the first two hydration shells of the two solutes (note that ρ_w may differ from the bulk water density ρ_w^l; see Sec. IV). The explicit expression for V_o depends on whether the smaller particle R' is hydrophobic or hydrophilic (see Appendix A).

On the other hand, the same $N_s^{r^\infty}$ molecules were involved in $\nu_s^{r^\infty}$ BHBs before leaving the first two hydration shells of particles R and R'. Denoting the energy of a single BHB by $\epsilon_s<0$, one can write the total energy of these $\nu_s^{r^\infty}$ bonds as $\Phi_s \equiv \Phi_s(r)=\epsilon_s \nu_s^{r^\infty}(r)$.

The contribution to the interaction potential between solutes R and R', arising from the disruption of the BHB networks in their vicinities because of their overlap, is given as

$$\phi^{\mathrm{hb}} = \Phi_b - \Phi_s. \tag{3}$$

This contribution is a function of r (because so are Φ_s and Φ_b) and ω_b (because so is Φ_s), i.e., $\phi^{\mathrm{hb}} \equiv \phi^{\mathrm{hb}}(r,\omega_b)$.

The evaluation of Φ_s is complicated by the composite character of the solute particle R because the water molecules belonging to its 1HS can form hydrogen bonds with the hydrophilic sites thereupon. According to the adopted model, one water molecule belonging to the 1HS can form only a single hydrogen bond with the particle surface; this happens when one of its hb arms is almost perpendicular to the surface and its tip pointing to a hydrophilic site of the latter (the probability of this event is ω_l). In such a situation, the water molecule forms the same number of hydrogen bonds as in the bulk (n_b) with the same (bulk) energy per bond (ϵ_b). Otherwise, the number n_e of bonds per 1HS water molecule will be smaller, $n_e<n_b$, but the bonds may be energetically enhanced (with the energy per bond $\epsilon_e<\epsilon_b$) [25–27].

Explicit expressions for Φ_s can be obtained in various ways differing in their accuracy. For the solvent-mediated interaction of two infinitely large parallel plates of composite nature [24], we used a linear interpolation of Φ_s as a function of ω_b. At $\omega_b=0$ the function $\Phi_s(\omega_b)$ reduced to Φ_b, whereas at $\omega_b=1$ the function $\Phi_s(\omega_b)$ provided Φ_e, the total energy of hydrogen bonds in which the $N_s^{r^\infty}$ molecules would have been involved if the surface of both interacting plates had been completely hydrophobic. For the solvent-mediated interaction between two spherical solutes (either both composite or one composite the other hydrophobic or hydrophilic), in this (linear) approximation we would have $\Phi_s(\omega_b)=\omega_b\Phi_e+(1-\omega_b)\Phi_b$, where Φ_b is given by Eq. (2) and the energy Φ_e is given by the expression

$$\Phi_e = \epsilon_e n_e \{f_{1,2}(1-\chi)[V_o(r) - V_m(r)] - 0.5\chi V_m(r)\}, \quad (4)$$

with n_e being the average number of hydrogen bonds formed by a 1HS water molecule; ϵ_e being the energy per such a bond; $\chi \approx 0.4$ being the probability that a hydrogen bond, formed by a selected 1HS molecule, is a bond of type 1 (when both water molecules involved belong to the 1HS; see Ref. [23]); $V_m(r)$ being the overlap volume of the 1HSs of the two solutes as a function of r (see Appendix A); and coefficient $f_{1,2}$ relating the average density of BHBs of type 2 (between molecules of the 1HS and 2HS) in the overlap region of 2HS's of the two solutes to the number density of water molecules in the 1HS (see Ref. [23] for details).

Hereafter, we will adopt another more accurate approximation for Φ_s. This energy as a function of r can be found as

$$\Phi_s(r) = \epsilon_s \nu_s^{r\infty}(r), \quad (5)$$

where ϵ_s is the average energy per each hydrogen bond formed by a water molecule in the 1HSs of solutes R and R'. Let us denote the average energy of such a bond in the 1HS of composite solute R by ϵ_s^c. On the other hand, the energy of a bond formed by a molecule in the 1HS of a hydrophobic solute is ϵ_e, while the water molecules around the hydrophilic solute do not form BHBs at all (all hydrogen bonds there are the same as in bulk). Therefore, if a water molecule belongs to the overlap region of the 1HSs of the two particles, there is some uncertainty regarding the energy of water-water bonds that such a molecule forms in the case where the composite solute R interacts with a hydrophobic solute R': this energy is either ϵ_s^c or ϵ_e, with equal probabilities. Adopting a most simple approach to get around this uncertainty, one can assume that ϵ_s is the arithmetic average of BHBs in the 1HSs of particles R and R' separately,

$$\epsilon_s = \begin{cases} \frac{1}{2}(\epsilon_s^c + \epsilon_e) & \text{(for a hydrophobic solute } R') \\ \epsilon_s^c & \text{(for a hydrophilic solute } R'). \end{cases} \quad (6)$$

We can then apply the linear interpolation with respect to ω_b (as described above) to ϵ_s^c as follows:

$$\epsilon_s^c = \omega_b \epsilon_e + (1-\omega_b)\epsilon_b. \quad (7)$$

The quantity $\nu_s^{r\infty}$, a decrease in the total number of BHBs in the first two hydration shells at a distance r between solutes, can be determined if one knows the average number n_s of hydrogen bonds formed by one molecule of water in the overlap region of the 1HSs of particles R and R'. If the latter is hydrophobic, then n_s can be estimated to be the arithmetic mean of n_e and n_s^c, with n_s^c being the average number of such bonds in the 1HS of the composite particle R. If the solute R' is hydrophilic, then n_s simply equals n_s^c. Thus,

$$n_s = \begin{cases} \frac{1}{2}(n_s^c + n_e) & \text{(for a hydrophobic solute } R') \\ n_s^c & \text{(for a hydrophilic solute } R'). \end{cases} \quad (8)$$

Again, an explicit expression for $\nu_s^{r\infty}$ will depend whether the solute R' is hydrophobic or hydrophilic (see Appendix A).

In order to determine n_s^c, consider a water molecule m_1^c in the 1HS of the composite particle R. By analogy with Eq. (11) of Ref. [23] and Eq. (13) of Ref. [24], one can represent n_s^c as the sum

$$n_s^c = p_1 + p_{2(1)} + p_{3(2,1)} + p_{4(3,2,1)}, \quad (9)$$

where p_1 is the probability that one of the hb arms of molecule m_1^c forms a hydrogen bond, $p_{2(1)}$ is the probability that a second hb arm forms a hydrogen bond subject to the restriction that one of the hb arms has already formed the bond, $p_{3(2,1)}$ is the probability that a third hb arm forms a hydrogen bond subject to the restriction that two of the hb arms have already formed bonds, and $p_{4(3,2,1)}$ is the probability that the fourth hb arm forms a bond subject to the restriction that three of the hb arms have already formed bonds. If the solute surface were purely hydrophilic, then these probabilities would be b_1, b_1^2, b_1^3, and b_1^4, respectively, where b_1 is the probability that a bulk water molecule forms a hydrogen bond. If the solute surface were completely hydrophobic, then these probabilities would be equal to s_1, $s_{2(1)}$, $s_{3(2,1)}$, and $s_{4(3,2,1)}$, respectively, related to b_1 by $s_1 = k_1 b_1$, $s_{2(1)} = k_2 b_1^2$, $s_{3(2,1)} = k_3 b_1^3$, $s_{4(3,2,1)} = k_4 b_1^4$, with the coefficients $k_1 \simeq 0.521\,694$, $k_2 \simeq 0.433\,148$, $k_3 \simeq 0.304\,122$, $k_4 \simeq 0.006\,433$ (see Refs. [23,24]). Taking into account the definition of the probabilities s_1, $s_{2(1)}$, $s_{3(2,1)}$, and $s_{4(3,2,1)}$, one can find their ω_b dependence by means of a linear interpolation between their values at $\omega_b = 0$ and $\omega_b = 1$, which provides

$$p_1 = K_1(\omega_b) b_1, \quad p_{2(1)} = K_2(\omega_b) b_1^2, \quad p_{3(2,1)} = K_3(\omega_b) b_1^3,$$

$$p_{4(3,2,1)} = K_4(\omega_b) b_1^4, \quad (10)$$

with the coefficients

$$K_i(\omega_b) = 1 - \omega_b + \omega_b k_i \quad (i = 1,2,3,4). \quad (11)$$

The probability b_1 is unambiguously determined by the thermodynamic state of the bulk water (temperature, pressure, etc.) as the solution of the equation $n_b = b_1 + b_1^2 + b_1^3 + b_1^4$ satisfying the constraint $0 < b_1 < 1$ (the bulk quantity n_b, whereof the dependence on thermodynamic conditions is well enough documented, is assumed to be given).

Substituting Eqs. (2) and (5) into Eq. (3) and taking into account Eqs. (6)–(11) (as well as those in Appendix A), one can calculate the solvent-mediated contribution to the interaction between composite and hydrophobic or hydrophilic particles, contribution arising because of the overlap of the boundary hydrogen bond networks around the two particles.

B. Case of protein folding via nucleation

It is clear that in order to apply the probabilistic approach to the nucleation mechanism of protein folding, the role of the composite particle should be attributed to the cluster of native residues (i.e., with correct tertiary contacts as they exist in the native protein). The radius of this cluster evolves (increases) as folding advances. On the other hand, the role of the smaller particle is played by the residues of the protein unfolded part. Their radius is thus constant and equal to $R' = 0$. In terms of hydrogen-bonding ability, both hydrophobic and neutral protein residues can be considered as "hydro-

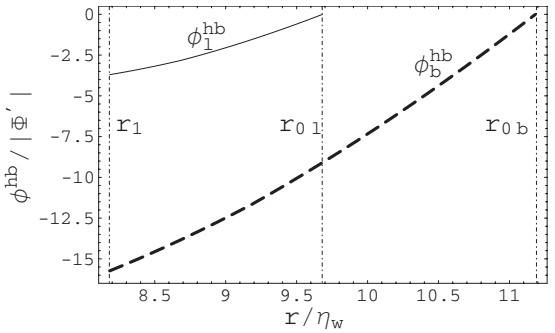

FIG. 3. The boundary hydrogen bond networks contribution ϕ^{hb} to the total interaction potential between two spherical solutes, one of which is a composite particle of radius $R=5\eta_w$ (and of hydrophobic surface fraction $\omega_b=0.4$) and the other is either a hydrophobic (thick dashed curve) or hydrophilic (thin continuous curve) particle of radius $R'=0$ (which corresponds to a single protein residue). The potential is plotted as $\phi^{hb}/|\Phi'|$ vs r/η_w, where $\Phi'=\epsilon_b n_b \rho_w \eta_w^3$ is twice the energy of water-water hydrogen bonds in the volume η_w^3 of bulk water. The leftmost dotted-dashed vertical line indicates the location of r_1, whereas the middle and rightmost lines indicate the location of r_0 for hydrophilic and hydrophobic beads, respectively. The solvent (water) is under such conditions that $n_b=3.65$.

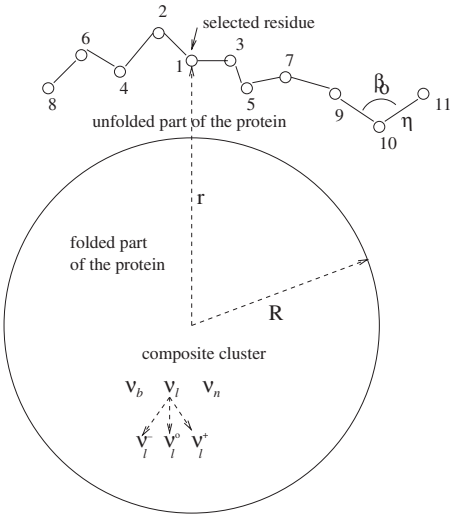

FIG. 4. A piece of a heteropolymer chain around a spherical cluster consisting of ν_b hydrophobic, ν_l hydrophilic, and ν_n neutral beads. Among the hydrophilic beads themselves ν_l^o are uncharged, ν_l^- are negatively charged, and ν_l^+ are positively charged. Bead 1 is in the plane of the figure, whereas other beads may all lie in different planes. All bond angles are equal to β_0 and their lengths are equal to η. The radius of the cluster is R and the distance from the selected bead 1 to the cluster center is r

phobic" (i.e., unable to form hydrogen bonds with water molecules).

The water-water hydrogen bond contribution ϕ^{hb} to the interactions of hydrophobic and hydrophilic residues with a protein folded cluster of radius $R=5\eta_w$ is presented in Fig. 3. The composition of the cluster was taken to be that of a bovine pancreatic ribonuclease (BPR), whereof the nucleation mechanism of folding we previously modeled [14]. The surface of the protein cluster was assumed to have the same composition as the whole protein so that the fraction of hydrophobic sites thereupon was taken to be $\omega_b=(N_b+N_n)/(N_b+N_l+N_n)$, where $N_b=40$, $N_l=81$, and $N_n=3$ are the numbers of hydrophobic, hydrophilic, neutral residues in BPR. The protein residues themselves are assumed to have the characteristic length of pairwise interaction $\eta=1.3\eta_w$. Water was assumed to be under such thermodynamic conditions that $n_b=3.65$. Quantitatively, the energetic enhancement of boundary hydrogen bonds (in the 1HS of a hydrophobic surface) can be characterized by the ratio ϵ_e/ϵ_b, where $\epsilon_e<0$ is the energy of a hydrogen bond involving at least one 1HS molecule and $\epsilon_b<0$ is the energy of a bond between two bulk molecules. In the PHB formalism [23,24], the ratio ϵ_s/ϵ_b is allowed to take on any positive value, i.e., $0<\epsilon_s/\epsilon_b<\infty$, although it is expected to be close to unity. As suggested in Ref. [29], the enhancement ratio ϵ_s/ϵ_b was taken to be 1.1.

The contribution ϕ_l^{hb} to the "cluster-hydrophilic residue" interactions in Fig. 3 is presented by the upper (thin solid curve), whereas the contribution ϕ_b^{hb} to the "cluster-hydrophobic residue" interactions is plotted as a thick dashed (lower) curve. The potentials are plotted as $\phi^{hb}/|\Phi'|$ vs r/η_w, where $\Phi'=\epsilon_b n_b \rho_w \eta_w^3$ is twice the energy of water-water hydrogen bonds in the volume η_w^3 of bulk water. The effect of BHB networks around the cluster and residues on "cluster-residue" interactions is clearly much (by an order of magnitude) stronger for the hydrophobic residues than for the hydrophilic ones. The range of $\phi^{hb}(r)$ for the former is twice as large as that for the latter.

III. MODIFIED MODEL FOR THE NUCLEATION MECHANISM OF PROTEIN FOLDING

A. Ternary heteropolymer as a protein model

The NMPF model [13,14] considered a protein polypeptide chain as a heteropolymer that consists of N connected beads which can be thought of as representing the α carbons of various amino acids (see Fig. 4). The heteropolymer consists of hydrophobic (b), hydrophilic (l), and neutral (n) beads. Besides, some (not all) hydrophilic beads are ionizable (some negatively and some positively), because 5 out of 11 hydrophilic amino acids in real proteins are ionizable. Two adjacent beads are connected by a covalent bond of a fixed length η. This model (and its variants) has been shown [6,10,12] to be able to capture the essential characteristics of protein folding even though it contains only some of the features of a real polypeptide chain.

The total energy of the heteropolymer (polypeptide chain) can contain three contributions of different types. First, the contribution from repulsive and attractive forces between pairs of nonadjacent beads (these can be, e.g., of Lennard-Jones or other types) that are at least three links apart (the interaction between nearest neighbors is taken into account by the link of constant length between them, while the interactions between next-nearest neighbors are taken into account by the rigidity of the angle between neighboring links). Next is a contribution from the harmonic forces due to the oscillations of the bond angles. Finally, a contribution from

the dihedral angle potential due to the rotation around the peptide bonds.

The bond angle forces are believed [6,10] to play a minor role in the protein folding and unfolding; hence, all bond angles can be set to be equal to β_0. It was shown [6,10] by molecular dynamics (MD) simulations, which employed low friction Langevin dynamics, that a proper balance between the remaining two contributions to the total energy of the heteropolymer ensures that the heteropolymer folds into a well-defined β-barrel structure. It was also found [6,10] that the balance between the dihedral angle potential, which tends to stretch the molecule into a state with all bonds in a *transconfiguration*, and the attractive hydrophobic potential is crucial to induce folding into a β-barrel-like structure upon cooling. If attractive forces are excessively dominant they make the heteropolymer fold into a globulelike structure, while an overwhelming dihedral angle potential forces the chain to remain in an elongated state (even at low temperatures) with bonds mainly in the transconfiguration.

Protein folding via nucleation is modeled to occur as the formation and evolution of a cluster of native residues within the protein. The cluster is assumed to have a spherical shape all the time. If the ionizability of (some) hydrophilic residues is taken into account, a cluster of native residues, involved in the nucleation mechanism of protein folding, should be characterized by five independent variables and it would be necessary to use the formalism of a five-component nucleation. Such a model would require extremely lengthy numerical calculations when applied to real proteins. To avoid this difficulty, one can assume that, as the protein folds via nucleation, the "hydrophilic" mole fractions of positive and negative residues in the cluster remain equal to those in the whole protein. Under such assumptions, the cluster can be characterized by only three independent variables and a model for the nucleation mechanism of protein folding can be again developed in terms of a ternary nucleation theory. In the framework of a three-component heteropolymer representation, the protein folding process can be regarded as a ternary phase transition with the ternary cluster (of native residues) within a ternary mother phase (unfolded part of the protein).

B. Hydrogen bond contribution to the potential field around the folded part of a protein

In our NMPF model we considered the folded part of the protein to be a cluster of spherical shape immersed in a ternary fluid mixture (whereof the role is played by the unfolded part of the protein). Let us denote the number of molecules (beads) of component i in such a ternary spherical cluster by ν_i $(i=b,l,n)$. The radius of the cluster will be denoted by R implying that it plays the role of a composite solute particle discussed above. The total number of beads of component i in the protein is denoted by N_i. In the original applications of the first passage time analysis to nucleation [30–32] a molecule of component i (in our model $i=b,l,n$) located in the surface layer of the cluster was considered to perform a thermal chaotic motion in a spherically symmetric potential well $\phi_i(r)$ resulting from the pair interactions (say, of LJ type) of this molecule with those in the cluster [r is the distance between the center of the cluster and the molecule; here, it is the distance between the centers of the cluster and selected bead (see bead 1 in Fig. 4)].

The total potential $\psi_i(r)$ for a residue of type i around the cluster previously (i.e., in the original version of the NMPF model) had three different constituents, $\phi_i(r)$, $\bar{\phi}_i^{\delta}(r)$, and ϕ_{cp}, which represented the effective LJ and electrostatic (for pairwise interactions of the selected residue with those in the cluster), average dihedral angle, and confining (representing the external boundary of the volume available to the unfolded residues) potentials, respectively,

$$\psi_i(r) = \phi_i(r) + \phi_{\mathrm{cp}}(r) + \bar{\phi}_i^{\delta}(r) \quad (i=b,l,n). \tag{12}$$

Complete details concerning the physical nature and calculation of $\phi_i(r)$, $\bar{\phi}_i^{\delta}(r)$, and ϕ_{cp} are given in Refs. [13,14].

In the original NMPF model hydrogen bonding was taken into account just indirectly via the diffusion coefficients of amino-acid residues. We will hereafter improve that model by combining it with the above-presented BHB model for the solvent-mediated interactions of solute particles. The improvement consists of augmenting the overall potential field around the cluster, $\psi_i(r)$, by an additional term $\phi_i^{\mathrm{hb}}(r)$ arising because of the disruption of the BHB networks around that residue and the cluster. As a result, instead of Eq. (12) the overall potential $\psi_i(r)$ will be now

$$\psi_i(r) = \phi_i(r) + \phi_{\mathrm{cp}}(r) + \bar{\phi}_i^{\delta}(r) + \phi_i^{\mathrm{hb}}(r) \quad (i=b,l,n), \tag{13}$$

where $\phi_i^{\mathrm{hb}}(r)$ is given by Eq. (3) and auxiliary equations in Sec. II with $R'=0$ and $R=[3v(\nu_b+\nu_l+\nu_n)/4\pi]^{1/3}$.

As with Eq. (12) (representing the original NMPF model [13,14]), the combination of potentials in Eq. (13) gives rise to a double potential well around the cluster with a barrier between the two wells. Figure 5 presents typical shapes of the constituents $\phi_i(r)$ and $\bar{\phi}_i^{\delta}(r)$ as functions of the distance from the cluster center, as well as the overall potential well $\psi_i(r)$ itself (for details regarding the numerical calculations, see Refs. [13,14]). The contribution $\phi_i(r)$, arising from the pairwise interactions, has a form reminiscent of the underlying Lennard-Jones potential, while the contribution from the average dihedral potential has a rather remarkable behavior. Indeed, starting with a maximum value at the cluster surface, it monotonically decreases with increasing r until it becomes constant for large enough values of r. Thus, except very short distances from the cluster surface where $\phi_i(r)$ sharply decreases from ∞ to its global minimum, the potential $\psi_i(r)$ is shaped by two competing terms, $\phi_i(r)$ and $\bar{\phi}_i^{\delta}(r)$, which increase and decrease, respectively, with increasing r, and by the confining potential ϕ_{cp}. The double-well shape of the overall potential $\psi_i(r)$ is of crucial importance to the NMPF model (both original and modified) because it allows one to use the mean first passage time analysis [33,34] for the determination of the rates of both absorption and emission of beads by the cluster.

Once the rates of emission and absorption are found as functions of cluster independent variables, one can develop a self-consistent kinetic theory for the nucleation mechanism

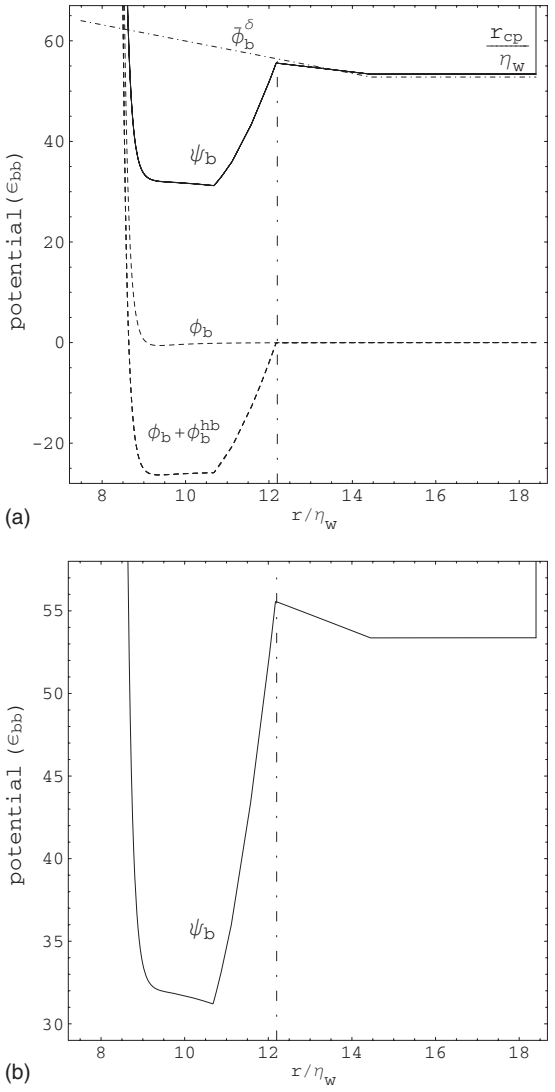

FIG. 5. (a) The potentials $\phi_b(r)+\phi_b^{\mathrm{hb}}$ (lower dashed curve), $\phi_b(r)$ (upper dashed curve), $\bar{\phi}_b^{\delta}(r)$ (dotted-dashed curve), and $\psi_b(r)$ (solid curve), for a hydrophobic bead around a cluster with $\nu_b=23$, $\nu_l=41$, $\nu_l^-=4.5$, $\nu_l^+=6.5$, $\nu_n=2$ at $p\mathrm{H}=7.3$ [$\phi_b(r)$, $\bar{\phi}_b^{\delta}(r)$, ϕ_{cp}, ϕ^{hb}, and $\psi_b(r)$ are the effective pairwise, average dihedral angle, confining, solvent-mediated (hydrogen bond), and total potentials for a hydrophobic bead around the cluster of folded protein]. The confining potential is shown as a solid vertical line at $r_{\mathrm{cp}}\simeq 18.5\eta_w$ whereas the hydrogen bond contribution is plotted separately in Fig. 3. All the potentials are given in units of ϵ_{bb}, the energy parameter in the Lennard-Jones potential of pairwise interactions between two hydrophobic beads. (b) shows the curve $\psi_b(r)$ alone for the better visualization of its double-well shape.

of protein folding and evaluate the folding time and its temperature dependence. The time necessary for the protein to fold is evaluated as the sum of the times necessary for the appearance of the first nucleus and the time necessary for the nucleus to grow to the maximum size of the folded protein in the native state (a brief description of the whole procedure is presented in Appendix B, while the reader is referred to Refs. [13,14] for details).

It is worth emphasizing that the double-well character of $\psi_i(r)$ is not related to the two-state (native-unfolded) nature of proteins. In the NMPF model, the latter is reflected in the function $G(\nu_b,\nu_l,\nu_n)$ (see Appendix B), which is the analog of the free energy of formation of a cluster ν_b,ν_l,ν_n and has the shape of a hyperbolic paraboloid (in a four-dimensional space of variables ν_b,ν_l,ν_n,G). Its path of the steepest descent has at least one maximum at the point with the coordinates $\nu_{bc},\nu_{lc},\nu_{nc},G(\nu_{bc},\nu_{lc},\nu_{nc})$ corresponding to the critical cluster. The initial point ($\nu_b=0$, $\nu_l=0$, $\nu_n=0$) of that path corresponds to a completely unfolded protein and its final point ($\nu_b=N_b$, $\nu_l=N_l$, $\nu_n=N_n$) to a completely folded (native) structure. The height of the barrier between the unfolded and folded states, provided by $G_c\equiv G(\nu_{bc},\nu_{lc},\nu_{nc})$, plays a crucial role in the kinetics of protein folding and determines the time of this process.

The path of the steepest descent of $G(\nu_b,\nu_l,\nu_n)$ can have additional local maxima that can lie on either side of the global maximum. (Such a situation would arise if the rates of emission and absorption of residues by a cluster were nonmonotonic functions of cluster variables. This could be possible if the diffusion coefficients of protein residues in the inner potential well (ipw) and outer potential well (opw) were dependent on cluster variables.) The local minima would then correspond to intermediate (partially folded) metastable states of the protein [1,4,5]. The kinetics of the transition from the initial (unfolded) state into the first intermediate one, transitions between intermediate states, and transition from the last one into the final (completely folded native) state are governed mostly by the heights of the barriers between them. The latter [as well as the whole shape of the function $G(\nu_b,\nu_l,\nu_n)$] are determined by the emission and absorption rates as functions of the variables ν_b,ν_l,ν_n (see Appendix B).

Our models of a protein and its folding have so far always provided a single maximum on the path of the steepest descent of height from $26k_BT$ to $33k_BT$ (k_B being the Boltzmann constant and T being the temperature). This is a clear indication of a two-state nature of the model proteins considered, at least under external conditions at which they were studied. In principle, the character of the protein can change if under different external conditions the height of the barrier can decrease or increase. If it decreases down to a few k_BT's (or disappears at all), the protein would fold in a virtually (or purely) barrierless way, which is characteristic of fast-folding proteins [35]. On the other hand, this (i.e., virtually or purely barrierless folding) can be a result of some particular values for the set of parameters in the model interaction potentials. One can also expect that under favorable circumstances the existence of intermediate (metastable) states of a protein on a particular folding trajectory can significantly decrease protein folding time compared to the time on another folding trajectory (of the same protein) with just one global maximum (assuming that its height is comparable to that of the global maximum of the path with intermediate states). This can take place if all intermediate barriers between previous and next local minima are significantly smaller than the height of the global barrier.

IV. NUMERICAL EVALUATIONS

As an illustration of the above theory, we have applied it to the folding of two model proteins, composed of 124 and

2500 amino acid residues. The first roughly mimics a BPR and was previously studied in Refs. [14] in the framework of the original NMPF model. According to the classification of Ref. [36], BPR consists of $N_b=40$ hydrophobic, $N_l=81$ hydrophilic, and $N_n=3$ neutral amino acids (with the total of $N=124$). We have considered a model heteropolymer consisting of the same numbers of hydrophobic, hydrophilic, and neutral beads: $N_b=40$ hydrophobic, $N_l=81$ hydrophilic, and $N_n=3$ neutral ones. Out of 81 hydrophilic beads, only 28 are ionizable, of which $\tilde{N}_l^-=10$ can be charged negatively and $\tilde{N}_l^+=18$ positively. The presence of other solute molecules and ionized bead counterions was not taken into account explicitly. The second model protein was chosen as a representative of large proteins. As in Ref. [13], it was a heteropolymer consisting of a total of 2500 residues, with the same mole fractions of hydrophobic, hydrophilic, neutral, and positively and negatively ionizable residues as for a short protein, that is, $N_b=807$, $N_l=1633$, $N_n=60$, and $\tilde{N}_l^+=363$ and $\tilde{N}_l^-=202$ (with a total of 565 ionizable residues). We intentionally modeled such a large protein (relatively rare in nature) to demonstrate that our method provides reasonable folding times on a reasonable time scale even for the largest proteins. Straightforward Monte Carlo (MC) or MD simulations of folding and unfolding behavior of such large proteins are currently impossible.

The actual numbers N_l^- and N_l^+ of negatively and positively ionized residues in the protein can be smaller than $\tilde{N}_l^-$ and $\tilde{N}_l^+$, respectively, with the average dissociation coefficients $k^-=N_l^-/\tilde{N}_l^-$ and $k^+=N_l^+/\tilde{N}_l^+$ depending on the pH of the surrounding medium and on the average dissociation coefficients K_a and K_b of the acidic and basic residues in the protein. In the BPR, there are five residues of aspartic acid (*asp*), five of glutamic acid (*glu*), ten of lysine (*lys*), four of arginine (*arg*), and four of histidine (*his*) so that K_a and K_b in BPR were taken to be $K_a=(5K_{asp}+5K_{glu})/(5+5)$ and $K_b=(10K_{lys}+4K_{arg}+4K_{his})/(10+4+4)$, where the dissociation's constants K_{asp}, K_{glu}, K_{lys}, K_{arg}, and K_{his} are available in the literature [36,37] (although they exhibit some scatters). At any given pH, the effective dissociation coefficients k^- and k^+ are completely determined by $pK_a=-\log_{10}K_a$, $pK_b=-\log_{10}K_b$, respectively, as $k^-=1/(1+10^{-p\mathrm{H}+pK_a})$, $k^+=1/(1+10^{-14+p\mathrm{H}+pK_b})$. We have carried out numerical calculations for pH=7.3 (roughly that of living cells) at which $N_l^-/N_l^+\simeq 0.69$, as well as for pH=6.3 at which $N_l^-/N_l^+\simeq 0.38$ and for pH=8.3 at which $N_l^-/N_l^+\simeq 1.59$. Note that the standard values [36,37] of the dissociation constants of ionizable residues are suitable for modeling unfolded states only. However, the corrections to K_a and K_b for the native state can be neglected because they affect only the electrostatic contributions to the total potential field around the cluster $\psi_l(r)$ which themselves constitute small corrections to the other terms in $\psi_l(r)$.

The contribution from the BHB networks to the overall potential was calculated as described in Sec. II with $R'=0$ and $R=[3v(\nu_b+\nu_l+\nu_n)/4\pi]^{1/3}$. The electrostatic interactions between a selected charge $q_\pm$ at $\mathbf{r}$ and the elementary charge at point $\mathbf{r}'$ (within the cluster) is given in the Debye-Hückel approximation for the screened Coulomb potential by

$$w_\pm(|\mathbf{r}'-\mathbf{r}|)=\frac{k}{\varepsilon}\frac{eq_\pm}{|\mathbf{r}'-\mathbf{r}|}\mathbf{e}^{-\kappa|\mathbf{r}'-\mathbf{r}|}, \tag{14}$$

where k is the electrostatic constant (1 in cgs units and $1/4\pi\varepsilon_0$ in SI units, with $\varepsilon_0=8.854\times10^{-12}$ F/m being the dielectric permittivity of vacuum), ε is the relative permittivity of the medium wherein the protein is present, and κ is the inverse Debye length (in our numerical calculations $1/\kappa\simeq 0.3$ nm which roughly corresponds to an electrolyte concentration of 0.3 M, typical for living cells [36]). The total electrostatic potential $u_\pm(r)$ of a residue carrying a charge $q_\pm=\pm e$ around the cluster (at a distance r from its center) was calculated as

$$u_\pm(r)=\int_V d\mathbf{r}'\ \rho_q(r')w_\pm(|\mathbf{r}'-\mathbf{r}|),$$

where $\mathbf{r}$ is the coordinate of the selected bead with a charge $q_\pm$ and the "number density" of charged residues at point $\mathbf{r}'$ within the cluster is assumed to be uniform, i.e., $\rho_q\equiv\rho_q(r')=(\nu_l^+-\nu_l^-)/(4\pi R^3/3)$. The density ρ_q can be negative, in which case the total charge of the cluster is negative. The integration in this equation has to be carried out over the whole volume of the system, but the contribution from the unfolded part is assumed to be small owing to the smaller density. The nonelectrostatic interactions between any two nonadjacent beads were modeled by Lennard-Jones (LJ) potentials whereas the potential due to the dihedral angle δ was represented by the expression $\phi_\delta=\epsilon'_\delta(1+\cos\delta)+\epsilon''_\delta(1+\cos 3\delta)$, where ϵ'_δ and ϵ''_δ are energy parameters which depend on the nature and sequence of the four beads involved in the dihedral angle δ. The parameters of the LJ and electrostatic potentials were chosen as in Ref. [14], but the parameters of the dihedral angle potential needed a significant adjustment as explained below. The typical total densities ρ_f and ρ_u of protein residues in the folded and unfolded (but compact) states and the diffusion coefficients of residues D^{iw} and D^{ow} in the ipw and opw were taken the same as in Ref. [14]: $\rho_f\eta^3=0.57$, $\rho_u=0.2\rho_f$, $D^{\mathrm{iw}}\rho_f=D^{\mathrm{ow}}\rho_u$, $D^{\mathrm{iw}}=D_i^{\mathrm{iw}}$ $(i=b,l,n)$, $D^{\mathrm{ow}}=D_i^{\mathrm{ow}}$ $(i=b,l,n)$, with D^{iw} assumed to vary between 10^{-6} and 10^{-8} cm^2/s (because of the lack of reliable data on the diffusion coefficient of a residue in a protein chain)

For hydrophobic, neutral, and uncharged hydrophilic beads the potential fields around the cluster are not affected by the charged residues in the cluster. For negatively and positively charged hydrophilic residues, the overall potentials around the cluster have the electrostatic contributions, $u_-(r)$ and $u_+(r)$, respectively. Included in the effective pairwise potential $\phi_j(r)$ $(j=+,-)$ in Eqs. (12) and (13), they are equal to each other in absolute values but have opposite signs, i.e., $u(r)\equiv u_+(r)=-u_-(r)$, so that $\phi_\pm(r)=\phi_o(r)\pm u(r)$, where $\phi_o(r)$ is the effective pairwise potential of uncharged hydrophilic beads (due exclusively to LJ interactions). As previously [14], the effective Lennard-Jones and average dihedral angle potentials have been considered to be independent of whether a hydrophilic residue is positively or negatively charged or noncharged.

Figure 5 presents typical shapes of the overall potential well $\psi_b(r)$ and its different constituents as functions of distance from the cluster center. All the potentials are given in units of the energy parameter ϵ_{bb} of the Lennard-Jones potential of pairwise "hydrophobic-bead–hydrophobic-bead" interactions. The curves shown are for a hydrophobic bead, but the curves for neutral and hydrophilic (both uncharged and positively and negatively charged) beads look very similar to those in Fig. 5. The results shown are for $p\mathrm{H}=7.3$ and a cluster $\nu_b=23$, $\nu_l^o=41$, $\nu_l^-=4.5$, $\nu_l^+=6.5$, $\nu_n=2$ carrying a total charge $q_\nu=+2e$.

Note that the pH of the medium (surrounding the protein) and the charge of the cluster do not affect the potentials (both overall and its constituents) for hydrophobic, neutral, and hydrophilic uncharged beads. The pairwise potential for hydrophilic charged residues, $\phi_j(r)$ $(j=+,-)$, is sensitive to the charge of the cluster because it contains the electrostatic contribution $\pm u(r)$ from the "charged-bead–charged-cluster" interactions. For example, for a positively charged cluster $\phi_o(r)+u_-(r)<\phi_o(r)<\phi_o(r)+u_+(r)$, i.e., the potential well (due to the well of the LJ potential) becomes shallower for positively charged hydrophilic beads and deeper for negatively charged ones as compared to uncharged ones [14]. However, the electrostatic contribution $u(r)$ to $\phi_j(r)$ $(j=+,-)$ is much weaker (by an order of magnitude) than the LJ contribution, i.e., $|u(r)/\phi_o(r)|\ll 1$. On the other hand, the entire effective pairwise contribution to ψ_j is much weaker than the BHB networks contribution, i.e., $|\phi_j(r)/\phi_j^{\mathrm{hb}}|\ll 1$. Therefore, the effective electrostatic potential $u(r)$ affects the overall potential field ψ_j very weakly. Nevertheless, this weak effect results in the pH dependence of the protein folding time because it is magnified by the formalism of the first passage time analysis (involving the exponentials of the overall potential ψ_j).

As seen in Fig. 3, the BHB network contribution ϕ^{hb} is a continuous negative function of r (its first derivative has finite discontinuities at some $r=r_1$ and $r=r_0>r_1$). However, the slope of ϕ^{hb} as a function of r is clearly discontinuous at $r=r_1\equiv R+(\eta+\eta_w)$ and $r=r_0$ with $r_0\equiv R+\eta+2\eta_w$ for hydrophilic beads and $r_0\equiv R+\eta+4\eta_w$ for hydrophobic ones. This is an artifact of the model and arises because of the sharp boundaries assumed for the 1HS and 2HS. When r decreases from ∞ to r_0, the 1HS and 2HS of the particles are not affected; hence, $\phi^{\mathrm{hb}}(r)=0$ for $r\geq r_0$. When r decreases from r_0 to r_1, 2HS and 1HS molecules are removed from the first two hydration shells of the particle(s), which leads to the decrease in ϕ^{hb} from zero to its minimum value (of strongest attraction) attained at r_1. When r becomes smaller than r_1, virtually no water molecules can fit in between the particles and the change in r practically does not lead to any change in the solvent-mediated interaction of the particles which hence remain constant (corresponding to the strongest interaction between particles). As a result, the attractive force between the particles is piecewise continuous with finite discontinuities at $r=r_1$ and $r=r_0$. (In Fig. 3, the location of r_1 is shown by the leftmost dotted-dashed vertical line, whereas the middle and rightmost lines indicate the location of r_0 for hydrophilic and hydrophobic beads, respectively). Thus, ϕ^{hb} represents an additional attraction between a cluster of native residues of the protein and a residue in the protein unfolded part at distances between their centers in the range $\eta<r<r_0$. Although this range is rather short, its strength is much higher than that of the effective pairwise potential at not too short distances between cluster and residue.

The effective pairwise potential $\phi_b(r)$ arising from elementary pairwise interactions (LJ for hydrophobic, uncharged hydrophilic, and neutral beads and LJ+electrostatic for positively and negatively charged hydrophilic ones) has a form reminiscent of the Lennard-Jones potential. Although this contribution to $\psi_b(r)$ is much weaker than $\bar{\phi}_b^{\delta}(r)$ and ϕ_b^{hb} at $r\geq\eta$, it tends to ∞ as $r\to 0$ thus forming the inner (closer to the cluster) boundary of the inner potential well. Therefore, it cannot be neglected on the right-hand side of Eq. (13) (likewise, one cannot neglect the confining potential that forms the outer boundary of the outer potential well, although it is zero everywhere at $r\leq r_{\mathrm{cp}}$).

The average dihedral potential (assigned to a selected bead) $\bar{\phi}_b^{\delta}(r)$ has a maximum value at the cluster surface and decreases monotonically with increasing r until it becomes constant for $r\geq R+\tilde{d}$, where $\tilde{d}$ is the maximum [13,14] distance between beads 1 and 6 (or beads 1 and 7) which depends on η and β_0. Such a behavior of $\bar{\phi}_i^{\delta}(r)$ can be interpreted as a consequence of the decrease in entropy (hence an increase in the free energy) of the heteropolymer chain as the selected bead 1 approaches the cluster surface for $r<R+\tilde{d}$ which, in turn, is due to a decrease in the configurational space available to the neighboring beads (beads 2–7).

The overall potential $\psi_b(r)$ depends very much on the parameters for the potential associated with a single dihedral angle. Previously [13,14], their values were chosen so that the overall potential field around a cluster had a double-well shape. Using the same values for the dihedral angle potential parameters in Eq. (13), corresponding to a modified (i.e., combined with the PHB model) NMPF model, would result in a potential field that has a single-well shape. This, in turn, would prevent us from applying the first passage time analysis to determining the absorption rate W_i^+ of the cluster [see Eq. (B3) in Appendix B]. Then, it remains rather unclear how one can determine W_i^+ and eventually evaluate the protein folding time.

In order to avoid this difficulty and conserve the double-well shape of the potential field around the cluster in the modified NMPF model, it is necessary to properly adjust the energy parameters in the dihedral angle potential. This adjustment must be carried out subject to a reasonable physical criterion, such as—for instance—the requirement that the predicted folding times must be in the range of experimentally observed ones. To satisfy this requirement, we had to increase the two energy parameters (which, for simplicity, are taken to be equal to each other) in the dihedral angle potential by a factor of 12 compared to their values used in Refs. [14]. With such values, the potential field around the cluster again has a double-well shape (see Fig. 5), although significantly different from the original NMPF model (because of the BHB networks contribution): the inner well (ipw) is separated by a potential barrier from the outer well (opw). The geometric characteristics of the wells (widths, depths, etc.) and the height and location of the barrier between them depend on the interaction parameters. The bar-

rier has different heights for the ipw and opw beads. For a given protein the location r_{cp} of the outer boundary of the opw is determined by the size of the cluster and densities ρ_f and ρ_u. The existence of an opw allows one to consider the absorption of a bead by the cluster as an escape of the bead from the opw by crossing over the barrier into the ipw. One can therefore use the mean first passage time analysis to determine not only the emission rate but also the rate of absorption of beads by the cluster.

For the model proteins considered (a short BPR of $N=124$ residues and a large protein of $N=2500$ residues) with the above choices of the system parameters our model estimates show that the time of protein folding is determined mainly by the time necessary for the first nucleation event to occur, as expected and as was the case in the original NMPF model. With the diffusion coefficient $D^{iw}=10^{-6}$ cm^2/s, the modified NMPF model estimates the characteristic time of folding of the short protein to be about 6 s at pH=8.3, 5 s at pH=7.3, and 8 s at pH=6.3, while the folding times for the long protein are about 165 s at pH=8.3, 140 s at pH=7.3, and 220 s at pH=6.3. For $D^{iw}=10^{-8}$ cm^2/s, the folding times of the short protein are predicted to be about 600, 500, and 800 s at pH=8.3, 7.3, and 6.3, respectively, while for the long protein they are about 16 500, 14 000, and 22 000 s at pH=8.3, 7.3, and 6.3, respectively. The folding times for the model BPR are in a good agreement with the experimentally observed folding times of real BPR (see Ref. [38], where the BPR folding time was reported to be on the order of 1000 s, and references therein). This suggests that the smaller value of the diffusion coefficient of protein residues in the unfolded state, $D^{iw}=10^{-8}$ cm^2/s, is more appropriate to model the folding of proteins whereof the sequence and structure are similar to those of BPR. On the other hand, faster folding proteins are probably better characterized by faster diffusion of their residues in the unfolded state and hence are better modeled with $D^{iw}=10^{-6}$ cm^2/s. (Note again that, besides linearly depending on $1/D^{iw}$, the folding times predicted by the above model are also sensitive to the energy parameter in the dihedral angle potential, the only interaction parameter of adjustable character). For the short protein the effect of pH on the protein folding time is significantly less pronounced in the modified model than in the original one. This was also expected because the contribution of the electrostatic interactions to the overall potential field is less important in the modified model compared to the original one. As previously [14], among all three pH's considered, the physiological pH provides the lowest folding time. Clearly, one cannot necessarily conclude that the folding time as a function of pH has a minimum at 7.3, but one can suggest that this function does have a minimum at some $6.3<p\mathrm{H}<8.3$. To more accurately determine the location of this minimum, it is necessary to calculate the folding times for more values of pH. The sensitivity of the folding time to pH is stronger for the large protein, which can be accounted for by larger net charges that the cluster of native residues can have during protein folding. Our results allow us to make some meaningful comparison of the folding times for two similar proteins differing (mainly) in the total number N residues in the polypeptide chain (with all the mole fractions of different kind of residues being the same). The dependence of the folding time on the protein length (i.e., N) suggested by our model is in a qualitatively good agreement with the results previously reported in Refs. [39–41] and obtained by using extensive Monte Carlo simulations of lattice model proteins. According to the latter (and several other theoretical studies), the folding time as a function of N can be approximated by a power-law function with the exponent ranging from 3 to 6. Our results would suggest the value of 1.1 for this exponent, but more detailed studies of several proteins with different lengths (i.e., N's) are needed for more accurate conclusions.

V. CONCLUSIONS

We have recently developed [13,14] a kinetic model for the nucleation mechanism of protein folding (NMPF) in terms of ternary nucleation by using the first passage time analysis. The main idea underlying the NMPF model consists of averaging the dihedral potential in which a selected residue is involved over all possible configurations of all neighboring residues along the protein chain. The combination of the average dihedral potential with the effective pairwise potential of a selected residue and with a confining potential caused by the bonds between the residues provides an overall potential around the cluster. As a function of the distance from the cluster center, the overall potential field has a double-well shape. This allows one to develop a self-consistent kinetic theory for the nucleation mechanism of protein folding and evaluate its characteristic time.

In the original NMPF model hydrogen bonding was not taken into account explicitly. To improve the NMPF model, we have developed a probabilistic hydrogen bond (PHB) model for the effect of hydrogen bond networks of water molecules around two solute particles (immersed in water) on their interaction [23,24]. That model suggests that the disruption of the boundary hydrogen bonds, which occurs when the first two hydration shells of two solute particles overlap, results in a short-ranged but strong attraction between the particles.

In this paper we have combined the PHB [23,24] and NMPF [13,14] models by slightly modifying the former to adapt it to protein folding. The folded cluster of the protein consists of three kinds of residues; hence, its surface can be expected to have a composite (hydrophobic-hydrophilic) character. On the other hand, a single residue in the unfolded part of the protein is either hydrophobic or hydrophilic (neutral residues are treated as hydrophobic). The PHB model has been extended to the solvent-mediated interaction of a spherical particle of composite nature (modeling a cluster of folded protein) with (1) a spherical hydrophobic particle (modeling hydrophobic and neutral protein residues) and (2) a spherical hydrophilic particle (modeling hydrophilic protein residues). In such a way, the water-water hydrogen bonding is taken explicitly into account in a modified kinetic model for the nucleation mechanism of protein folding. The additional contributions to the interaction potentials, arising due to the disruption of hydrogen bond networks around the interacting particles, have been thus added to the overall potential field around a cluster in the NMPF model.

For a numerical illustration we have again applied the model to the folding of two model proteins, one mimicking

bovine pancreatic ribonuclease, protein consisting of $N=124$ residues whereof $N_b=40$ are hydrophobic, $N_l=81$ hydrophilic (of which ten are negatively and 18 positively ionizable), and $N_n=3$ neutral, and the other—representing large proteins—consisting of $N=2500$ residues with $N_b=807$, $N_l=1633$, $N_n=60$ (with 202 negatively and 363 positively ionizable residues). Numerical calculations, performed at pH $=8.3$, 7.3, and 6.3, show that in the modified NMPF model the effect of pH on the protein folding time is less pronounced than in the original one (and the smaller the protein, the smaller this effect). This was expected because the contribution of the electrostatic interactions to the overall potential field is less important in the modified model compared to the original one. The hydrogen bond contribution now plays a dominant role in the total potential around the cluster (except for very short distances from the cluster surface where the LJ-type repulsion between the cluster and a residues increases infinitely thus giving rise to the inner wall of the double-well potential field). This contribution is by an order of magnitude stronger for hydrophobic residues than for hydrophilic ones. Besides, for the former the range of residue-cluster distances, at which the contribution exists at all, it is twice as large as for the latter.

In conclusion, one can note that, in principle, a similar approach can be used to develop a model for protein folding in a barrierless way, much like it was used in the model for barrierless protein denaturation [42]. The only difference would be that thermal denaturation is characteristic of most proteins, while fast-folding proteins (i.e., proteins folding in a barrierless way) are rather rare, so it is not straightforward to determine whether the model proposed above would capture this peculiar feature of folding of such rare proteins. A significant computational effort would be required to demonstrate that some specific proteins with specific sets of interaction parameters under specific conditions exhibit barrierless folding.

APPENDIX A: EXPLICIT EXPRESSIONS FOR $V_o(r)$, $V_m(r)$, AND $\nu_s^{r\infty}(r)$

In Eq. (2) the volume of the region resulting from the overlap of the first two hydration shells of the solutes R and R' depends on whether the latter (i.e., the smaller solute) is hydrophobic or hydrophilic. If it is hydrophobic, then its 1HS is a spherical layer with the inner and outer radii $R'+\frac{1}{2}(\eta+\eta_w)$ and $R'+\frac{1}{2}(\eta+\eta_w)+\frac{1}{2}\eta_w$, respectively, whereas its 2HS is a spherical layer with the inner and outer radii $R'+\frac{1}{2}(\eta+\eta_w)+\frac{1}{2}\eta_w$ and $R'+\frac{1}{2}(\eta+\eta_w)+\frac{3}{2}\eta_w$, respectively. On the other hand, if the smaller solute is hydrophilic, then it does not have 1HS and 2HS, but it has an exclusion sphere of radius $R'+\frac{1}{2}(\eta+\eta_w)$ wherein no water molecules can be located. In both cases, the 1HS and 2HS of the composite solute are determined by three concentric spheres of radii $R+\frac{1}{2}(\eta+\eta_w)$, $R+\frac{1}{2}(\eta+\eta_w)+\frac{1}{2}\eta_w$, and $R+\frac{1}{2}(\eta+\eta_w)+\frac{3}{2}\eta_w$. Therefore, the overlap volume $V_o(r)$ can be found as

$$V_o(r)=\begin{cases} V_{bo}(r)\equiv V_{ex}\left(r,R+\frac{1}{2}[\eta+\eta_w]+\frac{3}{2}\eta_w,R'+\frac{1}{2}[\eta+\eta_w]+\frac{3}{2}\eta_w\right) & \text{(hydrophobic } R') \\ V_{lo}(r)\equiv V_{ex}\left(r,R+\frac{1}{2}[\eta+\eta_w]+\frac{3}{2}\eta_w,R'+\frac{1}{2}[\eta+\eta_w]\right) & \text{(hydrophilic } R'), \end{cases} \tag{A1}$$

with

$$V_{ex}(r,a,b)=\frac{\pi}{12r}(a+b-r)(r^2+2ra-3a^2+2rb-3b^2+6ab), \tag{A2}$$

determining the volume of the region resulting from the overlap of two spheres of radii a and b as a function of the distance r between them (i.e., between their centers). Likewise the overlap volume $V_m(r)$ (of the 1HSs of the solutes) in Eq. (4) is determined as

$$V_m(r)=\begin{cases} V_{bm}(r)\equiv V_{ex}\left(r,R+\frac{1}{2}[\eta+\eta_w]+\frac{1}{2}\eta_w,R'+\frac{1}{2}[\eta+\eta_w]+\frac{1}{2}\eta_w\right) & \text{(hydrophobic } R') \\ V_{lm}(r)\equiv V_{ex}\left(r,R+\frac{1}{2}[\eta+\eta_w]+\frac{1}{2}\eta_w,R'+\frac{1}{2}[\eta+\eta_w]\right) & \text{(hydrophilic } R'). \end{cases} \tag{A3}$$

In order to calculate the decrease in the total number of BHBs in the first two hydration shells at a distance r between the solutes, it is convenient to classify them as follows. In the vicinity of a hydrophilic solute all hydrogen bonds are assumed to be the same as in the bulk water. For a hydrophobic particle, a BHB bond can be of type 1 (when both water molecules involved belong to the 1HS) or of type 2 (when one of the water molecules involved belongs to the 2HS). For the composite solute, in addition to types 1 and 2, a BHB bond can be also of type 3 when it is formed with a hydro-

philic site on the solute surface. The densities of types 1 and 2 BHBs are different for solutes R and R' not only because they are of different sizes but also because they are of different nature (the smaller solute R' cannot give rise to type 3 bonds at all). They can be all calculated by following the same procedure as in Refs. [23,24]. The simplest way to estimate the average density of type 2 bonds in the 2HSs of the two solutes and the average density of type 1 bonds in the 1HSs of the particles is to take their arithmetic means. One can thus obtain expressions for $\nu_s^{r\infty}$. If the solute R' is hydrophobic,

$$\nu_s^{r\infty} = \frac{1}{2} n_e \rho_w \left\{ \left(\frac{n_s}{n_e} \chi^{(2)} \frac{V_1}{V_2} + (1-\chi) \frac{V_1'}{V_2'} \right) [V_{bo}(r) - V_{bm}(r)] + \left(\frac{n_s}{n_e} (0.5\chi^{(1)} + \chi^{(3)}) + 0.5\chi \right) V_{bm}(r) \right\}, \tag{A4}$$

whereas for the hydrophilic solute R' we have

$$\nu_s^{r\infty} = n_e \rho_w \left\{ \frac{n_s}{n_e} \chi^{(2)} \frac{V_1}{V_2} [V_{lo}(r) - V_{lm}(r)] + \frac{n_s}{n_e} (0.5\chi^{(1)} + \chi^{(3)}) V_{lm}(r) \right\}, \tag{A5}$$

In these expressions, V_1 and V_2 are the volumes of the 1HS and 2HS of the composite solute; V_1' and V_2' are the volumes of the 1HS and 2HS of the hydrophobic solute R'; χ is the probability that a hydrogen bond, formed by a 1HS molecule in the vicinity of a hydrophobic solute, is a bond of type 1 (previously [23,24], it was estimated to be $\chi \simeq 0.083\ 168\ 5$); whereas $\chi^{(1)}$, $\chi^{(2)}$, and $\chi^{(3)}$ are the probabilities that a hydrogen bond, formed by a 1HS molecule in the vicinity of a composite solute, is of type 1, 2, or 3, respectively. The latter three probabilities can be calculated as

$$\chi^{(1)} = \frac{1}{C} \int_0^{0.5} d\kappa \int_{-\Theta_n(\kappa)}^{\Theta_x(\kappa)} d\Theta \sin\Theta, \quad C = \int_0^{0.5} d\kappa \int_{-\Theta_o(\kappa,\omega_b)}^{\pi/2} d\Theta \sin\Theta, \tag{A6}$$

$$\chi^{(2)} = \frac{1}{C} \int_0^{0.5} d\kappa \int_{\Theta_x(\kappa)}^{\pi/2} d\Theta \sin\Theta, \quad \chi^{(3)} = \frac{1}{C} \int_0^{0.5} d\kappa \int_{-\Theta_o(\kappa,\omega_b)}^{-\Theta_n(\kappa)} d\Theta \sin\Theta, \tag{A7}$$

with $\kappa = \xi/\eta_w$ (ξ being the distance of the selected molecule from the inner boundary of the 1HS (see Fig. 2), $\Theta_n(\xi) = \arcsin\kappa$, $\Theta_x(\xi) = \arcsin(0.5-\kappa)$, and $\Theta_x(\xi) = \arcsin(0.5-\kappa)$. Numerically, these expressions provide $\chi^{(1)} = 0.057\ 898\ 6$, $\chi^{(2)} = 0.638\ 261$, and $\chi^{(3)} = 0.303\ 84$.

It should be noted that in Eqs. (A4) and (A5) it is assumed that when the 2HSs of both particles overlap, a single molecule in the shared regions of their 2HSs can form type 2 BHBs only with the 1HS molecules of one solute. Although this assumption may lead to some inaccuracies in the expressions for $\nu_s^{r\infty}$, its effect on ϕ^{hb} can be expected to be significant only at large distances $r > 4\eta$ where the potential ϕ^{hb} itself is relatively small. With decreasing r the effect of this assumption on ϕ^{hb} will also decrease. Indeed, because of strong orientational restriction on the 1HS molecules, the water molecules belonging to the 1HS of one plate can hardly form type 2 BHBs with the molecules belonging to the 1HS of the other plate. Besides, when one of the particles itself (with an excluded spherical shell of thickness η) overlap with the 2HS of the other particles, water molecules from the latter are removed without replacement hence type 2 BHBs that they were involved in are broken without the probability of being reformed. Therefore, Eqs. (A4) and (A5) can be considered to provide reasonably good estimates for $\nu_s^{r\infty}$ as a function of r.

APPENDIX B: EVALUATION OF THE PROTEIN FOLDING TIME IN THE FRAMEWORK OF TERNARY NUCLEATION FORMALISM

Let us use $W_i^- \equiv W_i^-(\nu_b, \nu_l, \nu_n)$ and $W_i^+ \equiv W_i^+(\nu_b, \nu_l, \nu_n)$ ($i = b,l,n$) to denote the rates of emission and absorption, respectively, of beads of type i by a cluster containing ν_b hydrophobic, ν_l hydrophilic, and ν_n neutral residues. These rates represent the fundamental kinetic characteristics of the protein folding and unfolding processes. At any given temperature both functions W^- and W^+ can be determined by using the first passage time analysis (the method was first [30–32] applied to calculating $W^- = W^-(\nu)$ in unary nucleation and later extended [13,14] to both W^- and W^+ in protein folding and unfolding).

Consider a heteropolymer bead (i.e., a protein residue) of type i ($i=b,l,n$) performing a chaotic motion in a spherically symmetric potential well $\psi(r)$ with one boundary infinitely high (say, at $r=r_a$) and the other one of finite height (say, at $r=r_b$). Te mean first passage time τ necessary for the molecule to escape from the well is

$$\tau = \frac{1}{Z} \frac{1}{D} \int_{r_a}^{r_b} dr\, r^2 e^{-\Psi(r)} \int_r^{r_b} dy\, y^{-2} e^{\Psi(y)} \int_{r_a}^{y} dx\, x^2 e^{-\Psi(x)}, \tag{B1}$$

where D is the diffusion coefficient of a residue, $\Psi(r) = \psi(r)/k_B T$, and

$$Z = \int_{r_a}^{r_b} dr\, r^2 e^{-\Psi(r)}. \tag{B2}$$

The expression for τ was derived by solving a single-molecule master equation for the probability distribution function of a surface layer molecule (residue or bead) moving in a potential field $\psi(r)$ [13,14]. The diffusive motion of the bead is assumed to be governed by the Fokker-Planck equation [33,34]. The Fokker-Planck equation reduces to the Smoluchowski equation (which involves diffusion in an external field) if the relaxation time for the velocity distribution function of the molecule is very short and negligible compared to the characteristic time scale of the passage process.

The rates of emission and absorption of beads of type i by the cluster (i.e., the numbers of residues of type i escaping

from the ipw into the opw and from the opw into the ipw, respectively, per unit time) are provided by

$$W_i^- = \frac{N_i^-}{\tau_i^-}, \quad W_i^+ = \frac{N_i^+}{\tau_i^+}, \tag{B3}$$

where N_i^- and N_i^+ denote the numbers of molecules in the ipw and opw, respectively, and τ_i^- and τ_i^+ are the mean first passage times for the transition of a bead of type i from the opw into the ipw and from the ipw into the opw, respectively. Applying Eqs. (B1)–(B3) to calculate W^-, the locations of the boundaries of the ipw must be used, that is, $r_a=R$ and $r_b=R+\lambda^-$, with R being the radius of the cluster and λ^- being the width of the ipw; the diffusion coefficient must be taken to be D^{iw}. One the other hand, in calculating W^+, the locations of the boundaries of the opw must be used in Eqs. (B1)–(B3), that is, $r_a=R+\lambda^-+\lambda^+$ and $r_b=R+\lambda^-$, with λ^+ being the width of the opw; the diffusion coefficient must be taken to be D^{ow}. The quantities N_i^- and N_i^+ can be calculated as the product of the "volume × number density,"

$$N_i^- = \frac{4\pi}{3}[(R+\lambda^-)^3 - R^3]\rho_f,$$

$$N_i^+ = \frac{4\pi}{3}[(R+\lambda^-+\lambda^+)^3 - (R+\lambda^-)^3]\rho_u,$$

where ρ_f and ρ_u are the number densities of residues in the folded and unfolded parts of the protein.

Knowing the emission and absorption rates as functions of ν_b, ν_l, ν_n, one can find the nucleation rate J_s and estimate the time t_f necessary for the protein to fold via nucleation. Roughly speaking, the protein folding (via nucleation) consists of two stages. During the first stage, a critical cluster (nucleus) of native residues is formed (nucleation proper). Until the nucleus forms, the emission rate W^- is larger than W^+, but the cluster still can attain the critical size and composition by means of fluctuations. At the second stage the nucleus grows via regular absorption of native residues which dominates their emission, $W^- < W^+$. Thus, the folding time is given by

$$t_f \simeq t_n + t_g, \tag{B4}$$

where t_n is the time necessary for one critical cluster to nucleate within a compact (but still unfolded) protein and t_g is the time necessary for the nucleus to grow up to the maximum size, i.e., the size of the entirely folded protein. The time t_n of the first nucleation event can be estimated as

$$t_n \simeq 1/[J_s V_0], \tag{B5}$$

where J_s is the steady-state rate of ternary nucleation and V_0 is the volume of the unfolded protein in a compact configuration.

The nucleation rate J_s is found by solving the steady-state version of the kinetic equation of ternary nucleation governing the temporal evolution of $g(\nu_b, \nu_l, \nu_n, t)$, the distribution of clusters with respect to their three independent variables of state at time t. This equation has to be solved in the vicinity of the saddle point of the function $G(\nu_b, \nu_l, \nu_n) = -k_B T \ln[g_e(\nu_b, \nu_l, \nu_n)/(\rho_{bu}+\rho_{lu}+\rho_{nu})]$, where $g_e(\nu_b, \nu_l, \nu_n)$ is the equilibrium distribution of clusters, which can be constructed once the emission and absorption rates $W_i^- = W_i^-(\nu_b, \nu_l, \nu_n)$ and $W_i^+ = W_i^+(\nu_b, \nu_l, \nu_n)$ are known as functions of ν_b, ν_l, ν_n (for more details, see Refs. [13,14]). Note that in the first passage time analysis based approach to the kinetics of multicomponent nucleation the function $G(\nu_b, \nu_l, \nu_n)$ plays a role similar to the free energy of cluster formation in the classical nucleation theory (CNT). It determines a surface in a four-dimensional space which, under appropriate conditions (i.e., high enough metastability of the initial phase), is expected to have the shape of a hyperbolic paraboloid with at least one "saddle" point at the coordinates $\nu_{bc}, \nu_{lc}, \nu_{nc}$ (hereinafter the subscript "c" marks quantities at the saddle point). The steady-state kinetic equation of nucleation has to be solved subject to two boundary conditions, with one expressing the assumption that small clusters are in equilibrium and the other expressing the absence of too large clusters,

$$J_s = \frac{|\lambda_0|/2\pi k_B T}{\sqrt{-\det(\mathbf{G}''/2\pi k_B T)}} g_e(\nu_{bc}, \nu_{lc}, \nu_{nc}), \tag{B6}$$

where $\mathbf{G}''$ is the matrix of second derivatives of the function $G(\nu_b, \nu_l, \nu_n)$ with respect to ν_b, ν_l, ν_n and λ_0 is a negative eigenvalue of the matrix $\mathbf{A}\cdot\mathbf{G}''$, with the elements of the diagonal matrix $\mathbf{A}$ of the absorption rates given by $A_{ij} = \delta_{ij} W_i^+$ $(i,j=b,l,n)$, with δ_{ij} being the Kronecker delta. Note that although the form of the expression for J_s in Eq. (B6) is identical to that in Ref. [43], the latter was obtained (and applied to binary and ternary nucleation) in the framework of CNT. The crucial difference is hidden in the method for obtaining the equilibrium distribution $g_e(\nu_b, \nu_l, \nu_n)$. In the kinetic approach to a nucleation theory (originally proposed in Refs. [30–32]) it is obtained by using the mean first passage time analysis and the principle of detailed balance, while in CNT (whereupon its use had been previously based) the equilibrium distribution would have the form $(\rho_{bu}+\rho_{lu}+\rho_{nu})\exp[-F(\nu_b, \nu_l, \nu_n)]$, where $F(\nu_b, \nu_l, \nu_n)$ is the free energy of formation of a cluster, derived by using the concept of surface tension.

The growth time t_g is provided by the integral

$$t_g \simeq \int_{\nu_{bc}}^{N_b} \frac{d\nu}{W_b^+(\nu_b, \nu_l(\nu_b), \nu_n(\nu_b)) - W_b^-(\nu_b, \nu_l(\nu_b), \nu_n(\nu_b))}. \tag{B7}$$

Here, the functions $\nu_l = \nu_l(\nu_b)$ and $\nu_n = \nu_n(\nu_b)$ determine the growth path in parametric form and can be found as the solution of a couple of simultaneous differential equations (see Refs. [13,14])

$$\frac{d\nu_l}{d\nu_b} = \frac{W_l^+(\nu_b, \nu_l, \nu_n) - W_l^-(\nu_b, \nu_l, \nu_n)}{W_b^+(\nu_b, \nu_l, \nu_n) - W_b^-(\nu_b, \nu_l, \nu_n)},$$

$$\frac{d\nu_n}{d\nu_b} = \frac{W_n^+(\nu_b, \nu_l, \nu_n) - W_n^-(\nu_b, \nu_l, \nu_n)}{W_b^+(\nu_b, \nu_l, \nu_n) - W_b^-(\nu_b, \nu_l, \nu_n)}.$$

[1] T. E. Creighton, *Proteins: Structure and Molecular Properties* (W. H. Freeman, San Francisco, 1984).
[2] L. Stryer, *Biochemistry*, 3rd ed. (W. H. Freeman, San Francisco, 1988).
[3] C. B. Anfinsen, Science **181**, 223 (1973).
[4] C. Ghelis and J. Yan, *Protein Folding* (Academic Press, New York, 1982).
[5] B. Nölting, *Protein Folding Kinetics* (Springer-Verlag, Berlin, 2006).
[6] J. D. Honeycutt and D. Thirumalai, Biopolymers **32**, 695 (1992).
[7] J. S. Weissman and P. S. Kim, Science **253**, 1386 (1991).
[8] T. E. Creighton, Nature (London) **356**, 194 (1992).
[9] A. R. Fersht, Curr. Opin. Struct. Biol. **7**, 3 (1997).
[10] Z. Guo and D. Thirumalai, Biopolymers **36**, 83 (1995).
[11] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Biochemistry **33**, 10026 (1994).
[12] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **84**, 7524 (1987); J. Phys. Chem. **93**, 6902 (1989); Biopolymers **30**, 177 (1990).
[13] Y. S. Djikaev and E. Ruckenstein, J. Phys. Chem. B **111**, 886 (2007).
[14] Y. S. Djikaev and E. Ruckenstein, J. Chem. Phys. **126**, 175103 (2007); **128**, 025103 (2008); Phys. Chem. Chem. Phys. **10**, 6281 (2008); Adv. Colloid Interface Sci. **146**, 18 (2009).
[15] E. I. Shakhnovich and A. M. Gutin, J. Phys. A **22**, 1647 (1989).
[16] D. Bratko, A. K. Chakraborty, and E. I. Shakhnovich, J. Chem. Phys. **106**, 1264 (1997).
[17] Z. Konkoli, J. Hertz, and S. Franz, Phys. Rev. E **64**, 051910 (2001).
[18] W. Kauzmann, Adv. Protein Chem. **14**, 1 (1959).
[19] P. L. Privalov, Crit. Rev. Biochem. Mol. Biol. **25**, 281 (1990).
[20] A. Oleinikova, N. Smolin, I. Brovchenko, A. Geiger, and R. Winter, J. Phys. Chem. B **109**, 1988 (2005).
[21] M. Koizumi, H. Hirai, T. Onai, K. Inoue, and M. Hirai, J. Appl. Crystallogr. **40**, s175 (2007).
[22] I. Brovchenko, A. Krukau, N. Smolin, A. Oleinikova, A. Geiger, and R. Winter, J. Chem. Phys. **123**, 224905 (2005).
[23] Y. S. Djikaev and E. Ruckenstein, J. Chem. Phys. **130**, 124713 (2009).
[24] Y. S. Djikaev and E. Ruckenstein, J. Colloid Interface Sci. **336**, 575 (2009).
[25] K. A. T. Silverstein, A. D. J. Haymet, and K. A. Dill, J. Chem. Phys. **111**, 8000 (1999).
[26] W. Blokzijl and J. B. F. N. Engberts, Angew. Chem., Int. Ed. Engl. **32**, 1545 (1993).
[27] E. C. Meng and P. A. Kollman, J. Phys. Chem. **100**, 11460 (1996).
[28] N. Müller, Acc. Chem. Res. **23**, 23 (1990).
[29] B. Lee and G. Graziano, J. Am. Chem. Soc. **118**, 5163 (1996).
[30] G. Narsimhan and E. Ruckenstein, J. Colloid Interface Sci. **128**, 549 (1989).
[31] E. Ruckenstein and B. Nowakowski, J. Colloid Interface Sci. **137**, 583 (1990).
[32] B. Nowakowski and E. Ruckenstein, J. Colloid Interface Sci. **139**, 500 (1990).
[33] S. Chandrasekhar, Rev. Mod. Phys. **15**, 1 (1943).
[34] C. W. Gardiner, *Handbook of Stochastic Methods* (Springer, New York, 1983).
[35] J. Kubelka, J. Hofrichter, and W. A. Eaton, Curr. Opin. Struct. Biol. **14**, 76 (2004); F. Huang, S. Sato, T. D. Sharpe, L. M. Ying, and A. R. Fersht, Proc. Natl. Acad. Sci. U.S.A. **104**, 123 (2007).
[36] W. H. Elliott and D. C. Elliott, *Biochemistry and Molecular Biology* (Oxford University Press, New York, 2003).
[37] C. L. A. Schmidt, P. L. Kirk, and W. K. Appleman, J. Biol. Chem. **88**, 285 (1930).
[38] G. Xu, M. Narayan, and H. A. Scheraga, Biochemistry **44**, 9817 (2005); L. Pradeep, H.-C. Shin, and H. A. Scheraga, FEBS Lett. **580**, 5029 (2006).
[39] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, Phys. Rev. Lett. **77**, 5433 (1996).
[40] M. S. Li, D. K. Klimov, and D. Thirumalai, J. Phys. Chem. B **106**, 8302 (2002).
[41] M. Cieplak, T. X. Hoang, and M. S. Li, Phys. Rev. Lett. **83**, 1684 (1999).
[42] Y. S. Djikaev and E. Ruckenstein, J. Chem. Phys. **131**, 045105 (2009).
[43] H. Trinkaus, Phys. Rev. B **27**, 7372 (1983).