Statistical Mechanical Theory of Protein Folding in Water Environment

Alexander V. Yakubovich, Andrey V. Solov'yov and Walter Greiner

Abstract We present a statistical mechanics formalism for the theoretical description of the process of protein folding↔unfolding transition in water environment. The formalism is based on the construction of the partition function of a protein obeying two-stage-like folding kinetics. Using the statistical mechanics model of solvation of hydrophobic hydrocarbons we obtain the partition function of infinitely diluted solution of proteins in water environment. The calculated dependencies of the protein heat capacities upon temperature are compared with the corresponding results of experimental measurements for staphylococcal nuclease.

1 Introduction

Proteins are biological polymers consisting of elementary structural units, amino acids. Being synthesized at ribosome proteins are exposed to the cell interior where they fold into their unique three dimensional structure. The process of forming the protein's three dimensional structure is called the process of protein folding. The correct folding of protein is of crucial importance for the protein's proper functioning. Despite numerous works devoted to investigation of protein folding this process is still not entirely understood. The current state-of-the-art in experimental and theoretical studies of the protein folding process are described in recent reviews and references therein [1–5].

A. V. Yakubovich (✉) · A. V. Solov'yov · W. Greiner
Frankfurt Institute for Advanced Studies, Goethe University, 60438 Ruth-Moufang Str. 1, Frankfurt am Main, Germany
e-mail: yakubovich@fias.uni-frankfurt.de

A. V. Yakubovich
Author A.Y. on leave from A.F. Ioffe Physical Technical Institute, Polytechnicheskaya 26, 194021 Saint-Petersburg, Russia

W. Greiner (ed.), *Exciting Interdisciplinary Physics*,
FIAS Interdisciplinary Science Series, DOI: 10.1007/978-3-319-00047-3_39,

In this chapter we develop a novel theoretical method for the description of the protein folding process which is based on the statistical mechanics principles. Considering the process of protein folding as a first order type phase transition in a finite system, we present a statistical mechanics model for treating the folding↔unfolding phase transition in single-domain proteins. The suggested method is based on the theory developed for the helix↔coil transition in polypeptides discussed in [6–14]. A way to construct a parameter-free partition function for a system experiencing α-helix↔random coil phase transition *in vacuo* was studied in [6]. In [8] we have calculated potential energy surfaces (PES) of polyalanines of different lengths with respect to their twisting degrees of freedom. This was done within the framework of classical molecular mechanics. The calculated PES were then used to construct a parameter–free partition function of a polypeptide and to derive various thermodynamical characteristics of alanine polypeptides as a function of temperature and polypeptide length.

In this chapterr we construct the partition function of a protein *in vacuo*, which is the further generalization of the formalism developed in [9], accounting for folded, unfolded and prefolded states of the protein. This way of the construction of the partition function is consistent with nucleation-condensation scenario of protein folding, which is a very common scenario for globular proteins [15] and implies that at the early stage of protein folding the native-like hydrophobic nucleus of protein is formed, while at the later stages of the protein folding process all the rest of amino acids also attain the native-like conformation.

For the correct description of the protein folding in water environment it is of primary importance to consider the interactions between the protein and the solvent molecules. The hydrophobic interactions are known to be the most important driving forces of protein folding [16]. In the present work we present a way how one can construct the partition function of the protein accounting for the interactions with solvent, i.e. accounting for the hydrophobic effect. The most prominent feature of our approach is that it is developed for concrete systems in contrast to various generalized and toy-models of protein folding process.

We treat the hydrophobic interactions in the system using the statistical mechanics formalism developed in [17] for the description of the thermodynamical properties of the solvation process of aliphatic and aromatic hydrocarbons in water.

However, accounting solely for hydrophobic interactions is not sufficient for the proper description of the energetics of all conformational states of the protein and one has to take electrostatic interactions into account. In the present work the electrostatic interactions are treated within a similar framework as described in [18].

We have applied the developed statistical mechanics model of protein folding for a globular protein, namely staphylococcal nuclease. This protein has simple two-stage-like folding kinetics and demonstrate two folding↔unfolding transitions, refereed as heat and cold denaturation [19, 20]. The comparison of the results of the theoretical model with that of the experimental measurements shows the applicability of the suggested formalism for an accurate description of various thermodynamical characteristics in the system, e.g. heat denaturation, cold denaturation, increase of the reminiscent heat capacity of the unfolded protein, etc.

Our chapter is organized as follows. In Sect. 2.1 we present the formalism for the construction of the partition function of the protein in water environment and justify the assumptions made on the system's properties. In Sect. 3 we discuss the results obtained with our model for the description of folding↔unfolding transition in staphylococcal nuclease. In Sect. 4 we summarize the chapter and suggest several ways for a further development of the theoretical formalism.

2 Theoretical Methods

2.1 Partition Function of a Protein

To study thermodynamic properties of the system one needs to investigate its potential energy surface with respect to all the degrees of freedom. For the description of macromolecular systems, such as proteins, efficient model approaches are necessary.

The most relevant degrees of freedom in the protein folding process are the twisting degrees of freedom along its backbone chain [6, 7, 9–11, 13, 14, 21, 22]. These degrees of freedom are defined for each amino acid of the protein except for the boundary ones and are described by two dihedral angles φ_i and ψ_i (for definition of φ_i and ψ_i see e.g. [6, 7, 9–11, 13, 14]).

The degrees of freedom of a protein can be classified as stiff and soft ones. We call the degrees of freedom corresponding to the variation of bond lengths, angles and improper dihedral angles as stiff, while degrees of freedom corresponding to the angles φ_i and ψ_i are soft degrees of freedom [6]. The stiff degrees of freedom can be treated within the harmonic approximation, because the energies needed for a noticeable structural rearrangement with respect to these degrees of freedom are about several eV, which is significantly larger than the characteristic thermal energy of the system (kT), being at room temperature equal to 0.026 eV [12–14, 23–25].

A Hamiltonian of a protein is constructed as a sum of the potential, kinetic and vibrational energy terms. Assuming the harmonic approximation for the stiff degrees of freedom it is possible to derive the following expression for the partition function of a protein *in vacuo* being in a particular conformational state j [6]:

$$Z_j = A_j (kT)^{3N-3-\frac{l_s}{2}} \int_{\varphi \in \Gamma_j} \ldots \int_{\psi \in \Gamma_j} e^{-\varepsilon_j(\{\varphi,\psi\})/kT} \mathrm{d}\varphi_1 \ldots \mathrm{d}\varphi_n \mathrm{d}\psi_1 \ldots \mathrm{d}\psi_n, \quad (1)$$

where T is the temperature and k is the Boltzmann constant. N in Eq. (1) is the total number of atoms in the protein, l_s is the number of soft degrees of freedom. A_j in Eq. (1) is defined as follows:

$$A_j = \frac{V_j \cdot M^{3/2} \cdot \sqrt{I_j^{(1)} I_j^{(2)} I_j^{(3)}} \prod_{i=1}^{l_s} \sqrt{\mu_i^s}}{(2\pi)^{\frac{l_s}{2}} \pi \hbar^{3N} \prod_{i=1}^{3N-6-l_s} \omega_i}. \quad (2)$$

A_j is a factor which depends on the mass of the protein M, its three main momenta of inertia I, specific volume V, the frequencies of the stiff normal vibrational modes ω_i and on the generalized masses μ^s corresponding to the soft degrees of freedom [6]. ε in Eq. (1) describes the potential energy of the system corresponding to the variation of soft degrees of freedom.The contribution of the kinetic energy associated with soft degrees of freedom is also accounted for in Eq. 1. Integration in Eq. (1) is performed over a certain part of a phase space of the system (a subspace Γ_j) corresponding to the soft degrees of freedom φ and ψ. The form of the partition function in Eq. (1) allows one to avoid the multidimensional integration over the whole coordinate space and to reduce the integration only to the relevant parts of the phase space. ε_j in Eq. (1) denotes the potential energy surface of the protein as a function of twisting degrees of freedom in the vicinity of protein's conformational state j. Note, that in general the proper choice of all the relevant conformations of a protein and the corresponding set of Γ_j is not a trivial task.

One can expect that the factors A_j in Eq. (1) depend on the chosen conformation of the protein. However, due to the fact that the values of specific volumes, momenta of inertia and frequencies of normal vibration modes of the system in different conformations are expected to be close [9, 26], the values of A_j in all conformations become nearly equal, at least in the zero order harmonic approximation, i.e. $A_j \equiv A$. Another simplification of the integration in Eq. (1) comes from the statistical independence of amino acids. We assume that within each conformational state j all amino acids can be treated statistically independently, i.e. the particular conformational state of i-th amino acid characterized by angles $\varphi_i \in \Gamma_j$ and $\psi_i \in \Gamma_j$ does not influence the potential energy surface of all other amino acids, and vice versa. This assumption is well applicable for rigid conformational states of the protein such as native state. Indeed, for the ensemble of bound harmonic oscillators the partition function of the system depends only on the oscillators masses and spring strength, but not on the particular way of bindings in the system. Therefore the partition function of the protein in the rigid native state (i.e. all atoms vibrate harmonically) will depend only on the shape of the potential energy surfaces of amino acids in the vicinity of their minima, and not on the inter-amino acid interactions. For the native state of a protein all atoms of the molecule move in harmonic potential in the vicinity of their equilibrium positions. However, in unfolded states of the protein the flexibility of the backbone chain leads to significant variations of the distances between atoms, and consequently to a significant variation of interactions between atoms. Accurate accounting (both analytical and computational) for the interactions between distant atoms in the unfolded state of a protein is extremely difficult (see Ref. [27] for analytical treatment of interactions in unfolded states of a protein). In this work we assume that all amino acids in unfolded state of a protein move in the identical mean field created by all the amino acids and leave the corrections to this approximation for further considerations.

With the above mentioned assumptions the partition function of a protein Z_p (without any solvent) reads as:

$$Z_p = A \cdot (kT)^{3N-3-\frac{l_s}{2}} \sum_{j=1}^{\xi} \prod_{i=1}^{a} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \exp\left(-\frac{\varepsilon_i^{(j)}(\varphi_i, \psi_i)}{kT} \right) \mathrm{d}\varphi \mathrm{d}\psi, \qquad (3)$$

where the summation over j includes all ξ statistically relevant conformations of the protein, a is the number of amino acids in the protein and $\varepsilon_i^{(j)}$ is the potential energy surface as a function of twisting degrees of freedom φ_i and ψ_i of the i-th amino acid in the j-th conformational state of the protein. The exact construction of $\varepsilon_i^{(j)}(\varphi_i, \psi_i)$ for various conformational states of a particular protein will be discussed below. We consider the angles φ and ψ as the only two soft degrees of freedom in each amino acid of the protein, and therefore the total number of soft degrees of freedom of the protein $l_s = 2a$.

Partition function in Eq. (3) can be further simplified if one assumes (i) that each amino acid in the protein can exist only in two conformations: the native state conformation and the random coil conformation; (ii) the potential energy surfaces for all the amino acids are identical. This assumption is applicable for both the native and the random coil state. It is not very accurate for the description of thermodynamical properties of single amino acids, but is reasonable for the treatment of thermodynamical properties of the entire protein. The judgement of the quality of this assumption could be made on the basis of comparison of the results obtained with its use with experimental data. Such comparison is performed in Sect. 3 of this work.

Amino acids in a protein being in its native state vibrate in a steep harmonic potential. Here we assume that the potential energy profile of an amino acid in the native conformation should not be very sensitive to the type of amino acid and thus can be taken as e.g. the potential energy surface for an alanine amino acid in the α-helix conformation [8]. This assumption is well justified for proteins with the rigid helix-rich native structure. The staphylococcal nuclease, which we study here has definitely high α-helix content. Using the same arguments the potential energy profile for an amino acid in unfolded protein state can be approximated by e.g. the potential of alanine in the unfolded state of alanine polypeptide (see Ref. [8] for discussion and analysis of alanine's potential energy surfaces).

Indeed, for an unfolded state of a protein it is reasonable to expect that once neglecting the long-range interactions all the differences in the potential energy surfaces of various amino acids arise from the steric overlap of the amino acids's side chains. This is clearly seen on alanine's potential energy surface at values of $\varphi > 0°$ presented in Ref. [8]). But the part of the potential energy surface at $\varphi > 0°$ gives a minor contribution to the entropy of amino acid at room temperature. This fact allows one to neglect all the differences in potential energy surfaces for different amino acids in an unfolded protein, at least in the zero order approximation.

For the description of the folding $\leftrightarrow$ unfolding transition in small globular proteins obeying simple two-state-like folding kinetics we assume that the protein can exist in one of three states: completely folded state, completely unfolded state and partially folded state where some amino acids from the flexible regions with no prominent secondary structure are in the unfolded state, while other amino acids are in the

folded conformation. With this assumption the partition function of the protein reads as:

$$Z_p = Z_0 + \sum_{i=a-\kappa}^{a} \frac{\kappa!}{(i-(a-\kappa))!(a-i)!} Z_i, \tag{4}$$

where Z_i is defined in Eq. (1), Z_0 is the partition function of the protein in completely unfolded state, a is the total number of amino acids in a protein and κ is the number of amino acids in flexible regions. The factorial term in Eq. (4) accounts for the states in which various amino acids from flexible regions independently attain the native conformation. The summation in Eq. (4) is performed over all partially folded states of the protein, where $a-\kappa$ is the minimal possible number of folded amino acids. The factorial term describes the number of ways to select $i-(a-\kappa)$ amino acids from the flexible region of the protein consisting of κ amino acids attaining native-like conformation.

Omitting the contribution of protein's center of mass motion, the partition function of the system can be written as:

$$Z_p = \tilde{Z}_p \cdot A(kT)^{3N-3-a}, \tag{5}$$

where

$$\tilde{Z}_p = Z_u^a + \sum_{i=a-\kappa}^{a} \frac{\kappa! Z_b^i Z_u^{a-i} \exp(i \cdot E_0/kT)}{(i-(a-\kappa))!(a-i)!} \tag{6}$$

$$Z_b = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \exp\left(-\frac{\varepsilon_b(\varphi,\psi)}{kT}\right) \mathrm{d}\varphi \mathrm{d}\psi \tag{7}$$

$$Z_u = \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \exp\left(-\frac{\varepsilon_u(\varphi,\psi)}{kT}\right) \mathrm{d}\varphi \mathrm{d}\psi. \tag{8}$$

Here $\varepsilon_b(\varphi,\psi)$ (b stands for *bound*) is the potential energy surface of the amino acid in the native conformation and $\varepsilon_u(\varphi,\psi)$ (u - *unbound*) is the potential energy surface of amino acid in the random coil conformation. E_0 is the energy difference between the two conformational states of the amino acid. The potential energy profile of an amino acid is calculated as a function of its twisting degrees of freedom φ and ψ. Let us denote by ε_b^0 and ε_u^0 the global minima on the potential energy surfaces of the amino acid in folded and in unfolded conformations respectively. The absolute value of the energy of the amino acid can be written as $\varepsilon^0 + \varepsilon(\varphi,\psi)$. By E_0 in Eq. (8) is denoted the energy difference between global minima of the potential energy surfaces of the

amino acid in folded and in unfolded conformations, i.e. $E_0 = \varepsilon_u^0 - \varepsilon_b^0$. We count the absolute value of the potential energy from the energy of the global minima on the potential energy surface of the amino acid in unfolded conformation, $\varepsilon_u^0 = 0$ and each potential energy function $\varepsilon_b(\varphi, \psi)$ and $\varepsilon_u(\varphi, \psi)$ is counted form a reference energy value equal to ε_b^0 and ε_u^0 respectively.

The potential energy surfaces for amino acids as functions of angles φ and ψ were calculated and thoroughly analyzed in [8].

In nature proteins perform their function in the aqueous environment. The correct theoretical description of the folding↔unfolding transition in water environment should account for solvent effects.

2.2 *Partition function of a protein in water environment*

In this section we evaluate E_0 and construct the partition function for the protein in water environment.

The partition function of the infinitely diluted solution of proteins Z can be constructed as follows:

$$Z = \sum_{j=1}^{\xi} Z_p^{(j)} \cdot Z_{\text{water}}^{(j)}, \tag{9}$$

where $Z_{\text{water}}^{(j)}$ is the partition function of all water molecules in the j-th conformational state of a protein and $\tilde{Z}_p^{(j)}$ is the partition function of the protein in its j-th conformational state, in which we further omit the factor describing the contribution of stiff degrees of freedom in the system. This is done in order to simplify the expressions, because stiff degrees of freedom provide a constant contribution to the heat capacity of the system since the heat capacity of the ensemble of harmonic oscillators is constant, i.e. $\tilde{Z}_p \equiv Z_p$.

There are two types of water molecules in the system: (i) molecules in pure water and (ii) molecules interacting with the protein. We assume that only the water molecules being in the vicinity of the protein's surface are involved in the folding↔unfolding transition, because they are affected by the variation of the hydrophobic surface of a protein. This surface is equal to the protein's solvent accessible surface area (SASA) of the hydrophobic amino acids. The number of interacting molecules is proportional to SASA and include only the molecules from the first protein's solvation shell. This area depends on the conformation of the protein. The main contribution to the energy of the system caused by the variation of the protein's SASA associated with the side-chains of amino acids because the contribution to the free energy assosiated with solvation of protein's backbone is small [28]. Thus, in this work we pay the main attention to the accounting for the SASA change arising due to the solvation of side chains.

We treat all water molecules as statistically independent, i.e. the energy spectra of the states of a given molecule and its vibrational frequencies do not depend on a particular state of all other water molecules. Thus, the partition function of the whole system Z can be factorized and reads as:

$$Z = \sum_{j=1}^{\xi} Z_p \cdot Z_s^{Y_c(j)} Z_w^{N_0 - Y_c(j)}, \tag{10}$$

where ξ is the total number of states of a protein, Z_s is the partition function of a water molecule affected by the interaction with the protein and Z_w is the partition function of a water molecule in pure water. Y_c is the number of affected water molecules in the j-th conformational state of a protein. N_0 is the total number of water molecules in the system. Since we are focused on the *change* of the thermodynamic properties of the system in the course of protein unfolding, in the partition function we do not account for water molecules that do not interact with the protein in any of its conformational states, i.e. $N_0 = max\{Y_c(j)\}$.

To construct the partition function of water we follow the formalism developed in [17] and refer only to the most essential details of that work. The partition function of a water molecule in pure water reads as:

$$Z_s = \sum_{l=0}^{4} \xi_l f_l \exp(-E_l/kT), \tag{11}$$

where the summation is performed over 5 possible states of a water molecule (the states in which water molecule has 4, 3, 2, 1 and 0 hydrogen bonds with the neighboring molecules). E_l are the energies of these states and ξ_l are the combinatorial factors being equal to 1, 4, 6, 4, 1 for $l = 0, 1, 2, 3, 4$, respectively. They describe the number of choices to form a given number of hydrogen bonds. f_l in Eq. (11) describe the contribution to the partition function arising to to the translation and libration oscillations of the molecule. In the harmonic approximation f_l are equal to:

$$f_l = \left[1 - \exp(-h\nu_l^{(T)}/kT)\right]^{-3} \left[1 - \exp(-h\nu_l^{(L)}/kT)\right]^{-3}, \tag{12}$$

where $\nu_l^{(T)}$ and $\nu_l^{(L)}$ are translation and libration motions frequencies of a water molecule in its l-th state, respectively. These frequencies are calculated in Ref. [17] and are presented in Table (1). The contribution of the internal vibrations of the water molecule is not accounted for in Eq. (11) since frequencies of these vibrations can be considered as equal in all energetic states of the molecule.

The partition function of a water molecule from the protein's first solvation shell reads as:

Table 1 Parameters of the partition function of water

Number of hydrogen bonds	u	1	2	3	4
Energy level, E_i (cal/mol)	6670	4970	3870	2030	0
Translational frequencies, $\nu_i^{(T)}$, cm^{-1}	26	86	61	57	210
Librational frequencies, $\nu_i^{(L)}$, cm^{-1}	197	374	500	750	750

$$Z_s = \sum_{l=0}^{4} \xi_l f_l \exp(-E_l^{\text{shell}}/kT), \tag{13}$$

where f_l are defined as in Eq. (12) and E_l^{shell} denotes the energy levels of a water molecule interacting with aliphatic hydrocarbons of protein's amino acids. For simplicity we treat all side-chains of the hydrophobic core of a protein as being consisted of aliphatic hydrocarbons since most of the protein's hydrophobic amino acids consists of aliphatic-like hydrocarbons.

In our theoretical model we also account for the electrostatic interaction of protein's charged groups with the water.

The presence of electrostatic field around the protein leads to the reorientation of H_2O molecules in the vicinity of the charged groups due to the interaction of the H_2O molecules dipole moments with the electrostatic field. The corresponding factor in the partition function of H_2O molecules reads as:

$$Z_{\text{elec}} = \left(\frac{1}{4\pi} \int \exp\left(-\frac{E \cdot d \cos\theta}{kT} \right) \sin\theta \mathrm{d}\theta \mathrm{d}\varphi \right)^{\alpha}, \tag{14}$$

where E is the strength of the electrostatic field, d is the absolute value of the H_2O molecule dipole moment and α is the effective number of such molecules that are affected by electrostatic interaction. Note that the effects of electrostatic interaction turn out to be more pronounced in the folded state of the protein. This happens because in the unfolded state of a protein opposite charges of amino acid's side chains are in average closer in space due to the flexibility of the backbone chain, while in the folded state the positions of the charges are fixed by the rigid structure of a protein.

Taking the integral in Eq. (14), the correction to the partition function of a water molecule in pure water reads as:

$$Z'_w = \left(\sum_{l=0}^{4} \left[\xi_l f_l \exp(-E_l/kT) \right] \right) \left(\frac{kT \sinh\left[\frac{Ed}{kT} \right]}{Ed} \right)^{\alpha}. \tag{15}$$

This equation shows how the electrostatic field enters the partition function. In general, E depends on the position in space with respect to the protein. However,

here we neglect this dependence and instead we treat the parameter E as an average, characteristic electrostatic field created by the protein.

Having constructed the partition function of the system, we can evaluate with its use the thermodynamic characteristics of the system, such as entropy, free energy, heat capacity, etc. In this work we focus on the analysis of the dependence of protein's heat capacity on temperature and compare the predictions of our model with available experimental data.

3 Results and Discussion

In this section we calculate the dependencies of the heat capacity on temperature for a globular protein staphylococcal nuclease and compare the results obtained with experimental data from [19, 20].

Staphylococcal nuclease is relatively small globular protein consisting of 149 amino acids. It is a relatively nonspecific enzyme that digests single-stranded and double-stranded nucleic acids, but is more active on single-stranded substrates [31]. The structure native structure of the protein in shown in the centre in Fig. 1. Under certain experimental conditions (salt concentration and pH) the staphylococcal nuclease experience two folding↔unfolding transitions, which induce two peaks in the dependency of heat capacity on temperature (see Fig. 1). The peaks at lower temperature are due to the cold denaturation of the proteins. The peaks at higher temperatures arise due to the ordinary folding↔unfolding transition. The availability of experimental data for the heat capacity profiles of the mentioned protein, the presence of the cold denaturation and simple two-stage-like folding kinetics are the reasons for selecting this particular protein as case study for the verification of the developed theoretical model.

To calculate the SASA of staphylococcal nuclease in the folded state the 3D structure of the protein was obtained from the Protein Data Bank [32] (PDB ID 1EYD). Using CHARMM27 [25] forcefield and NAMD program [33] we performed the structural optimization of the protein and calculated SASA with the solvent probe radius 1.63 Å.

The value of SASA of the side-chains in the folded protein conformation is equal to 6858 $Å^2$. In order to calculate SASA for an unfolded protein state all the angles φ and ψ have been taken to be equal to 180°, i.e. as in a fully stretched conformation. Then, the optimization of the structure with the fixed angles φ and ψ was performed. The optimized geometry of the stretched molecule has a minor dependence on the value of dielectric susceptibility of the solvent, therefore the value of dielectric susceptibility was chosen to be equal to 20, in order to mimic the screening of charges by solvent. SASA of the side-chains in the stretched conformation of the protein tuned out to be equal to 15813 $Å^2$. The volume of one mole of water is $18\,cm^3$ therefore the volume of one molecule is $\sim$30 $Å^3$. To estimate the width of the solvation shell we consider the water as being in dense honeycomb hexagonal packing on the surface of the solute with a probe radius of the water molecule being equal to 1.4 Å [34]. The width

of the first solvation shell of the solute equals to ~3.26Å. Knowing the volume of a single water molecule, and the width of the solvation shell, one can obtain the total number of water molecules those interaction with the protein changes during the protein unfolding as follows:

$$N_{H_2O} = \frac{S_{unf} - S_{fol}}{V_0 / h_w}, \tag{16}$$

where $\frac{15813-6858}{30/(2\cdot 1.63)} \approx 973$. Therefore, there are 973 "pure" water molecules in the system with the folded protein that interact with the hydrophobic surface of the inner amino acids when the protein gets unfolded.

To account for the effects caused by the electrostatic interaction of water molecules with the charged groups of the protein it is necessary to evaluate the strength of the electrostatic field E in Eq. (15). The strength of the field can be estimated as $E \cdot d = kT/2$, where d is the dipole moment of a water molecule, k is Bolzmann constant and T = 300 K is the room temperature. According to this estimate the energy of electrostatic interaction of water molecules is equal to half of the thermal energy per degree of freedom of a molecule. This value of the field corresponds to the distance ~10 Åwith respect to a singly-charged atom. This distance was calculated using the distance-dependent dielectric susceptibility derived in Ref. [35] as follows:

$$\varepsilon(r) = (\varepsilon_\infty - \varepsilon_0)\left(\coth(\alpha(r - r_0)) - \frac{1}{\alpha(r - r_0)}\right) + \varepsilon_0, \tag{17}$$

where $\varepsilon(r)$ is the distance-dependent dielectric susceptibility, r is the distance from the charge, ε_∞ and ε_0 are dielectric susceptibilities of water (78) and protein (2) respectively, $\alpha = 0.45\ \text{Å}^{-1}$ and $r_0 = 2.9$ Å. These values of α and r_0 were proposed in [35] for the treatment of biomolecules.

The number of water molecules that interact with the electrostatic field can be obtained as a number of the molecules being in the sphere around an amino acid's charge site. For simplicity we do no account for the spatial distribution of the charge in the charged amino acid and treat it point-like. The dipole moments of H_2O molecules within the first and the second solvation shells of a charged atom are strongly polarized by the atom's electrostatic field [36]. Therefore, the orientation of the dipole moments of these H_2O molecules does not depend on the conformational state of the protein. The number of water molecules, N_w, which *change* their orientation during the of protein folding process can be calculated as follows:

$$N_w = \frac{1}{\rho}\left(\frac{4}{3}\pi R^3 - \frac{4}{3}\pi(r_{\text{ion}} + 2 \cdot 2r_w)^3\right), \tag{18}$$

where r_{ion}=1 Åis the radius of the charged atom, r_w=1.63 Åis the radius of a water molecule, ρ=30 nm^{-3} is the density of water molecules and R is the radius of the sphere around the charged atom. This radius can be calculated as follows. The change

of the free energy of a water molecule ΔF_E associated with the electrostatic interaction of H_2O molecules can be calculated from Eq. (15):

$$\Delta F_E = -kT \ln \left(\frac{kT \sinh \left[\frac{E(r)d}{kT} \right]}{E(r)d} \right), \tag{19}$$

where the strength of electrostatic field $E(r)$ depends on the distance r from the charged atom. The averaged over the sphere ΔF_E should be equal to the ΔF_E calculated for water molecules at the distance of ~10 Åform a singly-charged atom. Thus,

$$\frac{3 \int_{r_{\min}}^{R} \Delta F_E(r) r^2 dr}{R^3 - r_{\min}^3} = -kT \ln \left(\frac{kT \sinh \left(\frac{E(r_0)d}{kT} \right)}{E(r_0)d} \right), \tag{20}$$

where $r_{\min} = r_{\text{ion}} + 2 \cdot 2r_w$ and $r_0 = 10$ Å. R can be calculated from Eq. (20) using series expansion $\ln \left(\frac{\sinh(x)}{x} \right) \approx x^2/6$. The largest real root of Eq. (20) equals to 20.5 Å. Substituting the radius of sphere $R = 20.5$ Åto Eq. (18) one obtains the number of water molecules per charged residue of the protein $N_w \approx 1200$. Staphylococcal nuclease has 8 charged residues at physiological conditions [37], thus there are 9600 H_2O molecules in the system that interact with the electrostatic field of the folded protein. The value of α in Eq. (15) is calculated as a ratio of the number of electrostatically interacting H_2O molecules to the number of molecules that interact with the hydrophobic surface of the protein: $\alpha = \frac{9600}{973} \approx 10$.

Note that number of molecules interacting with the electrostatic field N_w and the strength of the electrostatic field E should be considered as the *effective* parameters of our model. In this work we do not perform accurate accounting for the spatial dependence of the electrostatic field. Instead, we introduce the parameters α and E that can be interpreted as effective values of the number of H_2O molecules and the strength of the electrostatic field correspondingly. Let us stress that the number of water molecules α and the strength of the field E are not independent parameters of our model because by choosing the higher value of E and smaller α or vice versa one can derive the same heat capacity profile. Therefore, below we focus on the investigation of the dependence of the protein heat capacity on E at the fixed value of α equal to 10.

Another important parameter of the model is the energy difference between the two states of the protein normalized per one amino acid, E_0 introduced in Eq. (8). This parameter describes both the energy loss due to the separation of the hydrophobic groups of the protein which attract in the native state of the protein via Van-der-Waals interaction and the energy gain due to the formation of Van-der-Waals interactions of hydrophobic groups of the protein with H_2O molecules in the protein's unfolded state. Also, the difference of the electrostatic energy of the system in the folded and unfolded states is accounted for in E_0. The difference of the electrostatic energy

Table 2 Values of E_0 for staphylococcal nuclease at different values of pH of the solvent

pH value	7.0	5.0	4.5	3.88	3.23
E_0 (kcal/mol)	−1.0635	−1.07	−1.077	−1.093	−1.2

may depend on various characteristics of the system, such as concentration of ions in the solvent and its pH, on the exact location of the charged sites in the native conformation of the protein and on the probability distribution of distances between charged amino acids in the unfolded state. Thus, exact calculation of E_0 is a rather difficult and separate task which we do not intend to solve in this work. Instead, in the current study the energy difference between the two phases of the protein is considered as a parameter of the model. We treat E_0 as being dependent on external properties of the system, in particular on the pH value of the solution.

Another characteristic of the protein folding↔unfolding transition is its cooperativity. In the model it is described by the parameter κ in Eq. (4). κ describes the number of amino acids in the flexible regions of the protein. The staphylococcal nuclease possesses a prominent two-stage folding kinetics, therefore only 5–10% of amino acids is in the protein's flexible regions. Thus, the value of κ for this protein is small can be estimated as being equal to $149 \cdot 7\,\% \approx 10$ amino acids.

The values of E_0 for staphylococcal nuclease at different values of pH are presented in Table 2. Note E_0 is negative similarly to hydrophobic molecule butane (see Ref. [17] for details).

The dependence of heat capacity on temperature calculated for staphylococcal nuclease at different pH are presented in Fig. 1 by solid lines.

The results of experimental measurements form Ref. [19] are presented by symbols. From Fig. 1 it is seen that staphylococcal nuclease experience two folding ↔ transitions in the range of pH between 3.78 and 7.0. At the pH value 3.23 no peaks in the heat capacity is present. It means that the protein exists in the unfolded state over the whole range of experimentally accessible temperatures.

In our calculation we have adjusted the absolute value of the heat capacity. However, the absolute value of the heat capacity is not a parameter of the model because the experimentally measured absolute value of the heat capacity depends not entirely on the properties of the protein but also on the properties of the solution, ion concentration, etc.

Comparison of the theoretical results with experimental data shows that our theoretical model reproduces experimental behavior better for the solvents with higher pH. The heat capacity peak arising at higher temperatures due to the standard folding↔unfolding transition is reproduced very well for pH values being in the region 4.5–7.0. The deviations at low temperatures can be attributed to the inaccuracy of the statistical mechanics model of water in the vicinity of the freezing point.

The accuracy of the statistical mechanics model for low pH values around 3.88 is also quite reasonable. The deviation of theoretical curves from experimental ones

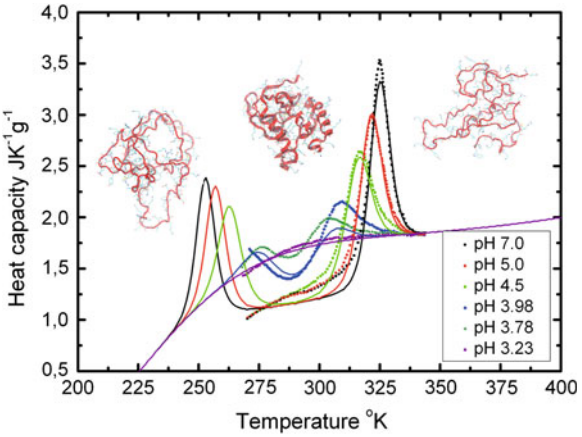

Fig. 1 Dependencies of the heat capacity on temperature for staphylococcal nuclease (PDB ID 1EYD) at different values of pH. *Solid lines* show results of the developed theoretical model. Symbols present experimental data from Ref. [19]. Structure of the protein in native and unfolded conformations are shown in temperature regions where the corresponding conformation exists.

likely arise due to the alteration of the solvent properties at high concentration of protons or due to the change of partial charge of amino acids at pH values being far from the physiological conditions.

Despite some difference between the predictions of the developed model and the experimental results arising at certain temperatures and values of pH the overall performance of the model can be considered as extremely good for such a complex process as structural folding transition of a large biological molecule.

4 Conclusions

We have developed a novel statistical mechanics model for the description of folding↔unfolding processes in globular proteins obeying simple two-stage-like folding kinetics. The model is based on the construction of the partition function of the system as a sum over all statistically significant conformational states of a protein. The partition function of each state is a product of partition function of a protein in a given conformational state, partition function of water molecules in pure water and a partition function of H_2O molecules interacting with the protein.

The introduced model includes a number of parameters responsible for certain physical properties of the system. The parameters were obtained from available experimental data and three of them (energy difference between two phases, cooperativity of the transition and the average strength of the protein's electrostatic field) were considered as being variable depending on a particular protein and pH of the solvent.

We have compared the predictions of the developed model with the results of experimental measurements of the dependence of the heat capacity on temperature for staphylococcal nuclease. The experimental results were obtained at various pH of solvent. The suggested model is capable to reproduce well within a single framework a large number of peculiarities of the heat capacity profile, such as the temperatures of cold and heat denaturations, the corresponding maximum values of the heat capacities, the temperature range of the cold and heat denaturation transitions, the difference between heat capacities of the folded and unfolded states of the protein.

The good agreement of the results of calculations obtained using the developed formalism with the results of experimental measurements demonstrates that it can be used for the analysis of thermodynamical properties of many biomolecular systems. Further development of the model can be focused on its advance and application for the description of the influence of mutations on protein stability, analysis of assembly and stability of protein complexes, protein crystallization process, etc.

Acknowledgments A.Y. thanks Stiftung Polytechnische Gesellschaft Frankfurt am Main for financial support.

References

1. V. Muñoz, Conformational dynamics and ensembles in protein folding. Annu. Rev. Biophys. Biomol. Struct. **36**, 395–412 (2007)
2. K.A. Dill, S.B. Ozkan, M.S. Shell, T.R. Weikl, The protein folding problem. Annu. Rev. Biophys. **37**, 289–316 (2008)
3. J.N. Onuchic, P.G. Wolynes, Theory of protein folding. Curr. Op. Struct. Biol. **14**, 70–75 (2004)
4. E. Shakhnovich, Protein folding thermodynamics and dynamics: Where physics, chemistry, and biology meet. Chem. Rev. **106**, 1559–1588 (2006)
5. N.V. Prabhu, K.A. Sharp, Protein-solvent interactions. Chem. Rev. **106**, 1616–1623 (2006)
6. A. Yakubovich, I. Solov'yov, A. Solov'yov, W. Greiner, Ab initio theory of helix↔coil phase transition. Eur. Phys. J. D. **46**, 215–225, (2007)(arXiv:0704.3079v1 [physics.bio-ph])
7. A. Yakubovich, I. Solov'yov, A. Solov'yov, W. Greiner, Ab initio description of phase transitions in finite bio-nano-systems. Europhys. News **38**, 10 (2007)
8. I. Solov'yov, A. Yakubovich, A. Solov'yov, W. Greiner, α-helix↔random coil phase transition: analysis of ab initio theory predictions. Eur. Phys. J. D. **46**, 227–240 (2008) (arXiv:0704.3085v1 [physics.bio-ph])
9. A. Yakubovich, I. Solov'yov, A. Solov'yov, W. Greiner, Phase transition in polypeptides: a step towards the understanding of protein folding. Eur. Phys. J. D **40**, 363–367 (2006)
10. A. Yakubovich, I. Solov'yov, A. Solov'yov, W. Greiner, Conformational changes in glycine tri- and hexapeptide. Eur. Phys. J. D **39**, 23–34 (2006)
11. A. Yakubovich, I. Solov'yov, A. Solov'yov, W. Greiner, Conformations of glycine polypeptides. Khimicheskaya Fizika (Chemical Physics) (in Russian) **25**, 11–23 (2006)
12. I. Solov'yov, A. Yakubovich, A. Solov'yov, W. Greiner, On the fragmentation of biomolecules: fragmentation of alanine dipeptide along the polypeptide chain. J. Exp. Theor. Phys. **103**, 463–471 (2006)
13. I. Solov'yov, A. Yakubovich, A. Solov'yov, W. Greiner, Ab initio study of alanine polypeptide chain twisting. Phys. Rev. E **73**, 021916 (2006)
14. I. Solov'yov, A. Yakubovich, A. Solov'yov, W. Greiner, Potential energy surface for alanine polypeptide chains. J. Exp. Theor. Phys. **102**, 314–326 (2006)

15. B. Noetling, D.A. Agard, How general is the nucleation condensation mechanism? Proteins **73**, 754–764 (2008)
16. S. Kumar, C.-J. Tsai, R. Nussinov, Maximal stabilities of reversible two-state proteins. Biochemistry **41**, 5359–5374 (2002)
17. J. H. Griffith, H. Scheraga, Statistical thrmodynamics of aqueous solutions. i. water structure, solutions with non-polar solutes, and hydrophobic ineractions. J. Mol. Struc. **682**, 97–113 (2004)
18. A. Bakk, J.S. Høye, A. Hansen, Apolar and polar solvation thermodynamics related to the protein unfolding process. Biophys. J. **82**, 713–719 (2002)
19. Y. Griko, P. Privalov, J. Aturtevant, S. Venyaminov, Cold denaturation of staphyloccocal nuclease. Proc. Natl. Acad. Sci. U.S.A **85**, 3343–3347 (1988)
20. P. Privalov, Thermodynamics of protein folding. J. Chem. Thermodyn. **29**, 447–474 (1997)
21. S. He, H.A. Scheraga, Macromolecular conformational dynamics in torsion angle space. J. Chem. Phys. **108**, 271–286 (1998)
22. S. He, H.A. Scheraga, Brownian dynamics simulations of protein folding. J. Chem. Phys. **108**, 287–300 (1998)
23. W. Scott, W. van Gunsteren, The GROMOS Software Package for Biomolecular Simulations, in *Methods and Techniques in Computational Chemistry: METECC-95*, ed. by E. Clementi, G. Corongiu (STEF, Cagliari, 1995), pp. 397–434
24. W. Cornell, P. Cieplak, C. Bayly et al., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J. Am. Chem. Soc. **117**, 5179–5197 (1995)
25. A. MacKerell, D. Bashford, R. Bellott et al., All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B **102**, 3586–3616 (1998)
26. S. Krimm, J. Bandekar, Vibrational analysis of peptides, polypeptides, and proteins v. normal vibrations of β-turns. Biopolymers **19**, 1–29 (1980)
27. M. Cubrovic, O. Obolensky, A. Solov'yov, Semistiff polymer model of unfolded proteins and its application to nmr residual dipolar couplings. Eur. Phys. J. D. **51**, 41–49 (2009)
28. A. Finkelstein, O. Ptitsyn, *Protein Physics: A Course of Lectures* (Elsevier Books, Oxford, 2002)
29. G. Makhatadze, P. Privalov, Contribution to hydration to protein folding thermodynamics. I. The enthalpy of hydration. J. Mol. Biol. **232**, 639–659 (1993)
30. W. Humphrey, A. Dalke, K. Schulten, Vmd - visual molecular dynamics. J. Molec. Graphics **14**, 33–38 (1996)
31. F.A. Cotton, E.E. Hazen Jr, M.J. Legg, Staphylococcal nuclease: Proposed mechanism of action based on structure of enzyme-thymidine 3',5'-bisphosphate-calcium ion complex at 1.5-a resolution. Proc. Natl. Acad. Sci. U.S.A **76**, 2551–2555 (1979)
32. Protein data bank, www.rcsb.org/ (2009)
33. J.C. Phillips, R. Braun, W. Wang et al., Scalable molecular dynamics with NAMD. J. Comp. Chem. **26**, 1781–1802 (2005)
34. R. Wade, M.H. Mazor, J.A. McCammon, F. Quiocho, A molecular dynamics study of thermodynamic and structural aspects of the hydration of cavities in proteins. Biopolymers **31**, 919–931 (1991)
35. J. Mazur, R.L. Jernican, Distance-dependent dielectric constants and their application to double-helical DNA. Biopolymers **31**, 1615–1629 (1991)
36. J.H. Griffith, H. Scheraga, Statistical thermodynamics of aqueous solutions ii.alkali halides at infinite dilution. J. Mol. Struc. **711**, 33–48 (2004)
37. H.-X. Zhou, Residual charge interactions in unfolded staphylococcal nuclease can be explained by the gaussian-chain model. Biophys. J. **83**, 2981–2986 (2002)
38. J.P. Collman, R. Boulatov, C.J. Sunderland, L. Fu, Functional analogues of cytochrome c oxidase, myoglobin, and hemoglobin. Chem. Rev. **104**, 561–588 (2004)
39. D. Shortle, M.S. Ackerman, Persistence of native-like topology in a denatured protein in 8 m urea. Science **293**, 487–489 (2001)