Coexistence of Phases in a Protein Heterodimer

Andrey Krokhotin,[1,a)] Adam Liwo,[2,b)] Antti J. Niemi,[1,3,c)] and Harold A. Scheraga[4,d)]
[1] *Department of Physics and Astronomy and Science for Life Laboratory, Uppsala University, P.O. Box 803, S-75108 Uppsala, Sweden*
[2] *Faculty of Chemistry, University of Gdansk, ul. Sobieskiego 18, 80-952 Gdansk, Poland*
[3] *Laboratoire de Mathematiques et Physique Theorique CNRS UMR 6083, Fédération Denis Poisson, Université de Tours, Parc de Grandmont, F37200, Tours, France*
[4] *Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, New York 14853-1301, USA*

(Received 27 April 2012; accepted 22 June 2012; published online 17 July 2012)

A heterodimer consisting of two or more different kinds of proteins can display an enormous number of distinct molecular architectures. The conformational entropy is an essential ingredient in the Helmholtz free energy and, consequently, these heterodimers can have a very complex phase structure. Here, it is proposed that there is a state of proteins, in which the different components of a heterodimer exist in different phases. For this purpose, the structures in the protein data bank (PDB) have been analyzed, with radius of gyration as the order parameter. Two major classes of heterodimers with their protein components coexisting in different phases have been identified. An example is the PDB structure 3DXC. This is a transcriptionally active dimer. One of the components is an isoform of the intra-cellular domain of the Alzheimer-disease related amyloid precursor protein (AICD), and the other is a nuclear multidomain adaptor protein in the Fe65 family. It is concluded from the radius of gyration that neither of the two components in this dimer is in its own collapsed phase, corresponding to a biologically active protein. The UNRES energy function has been utilized to confirm that, if the two components are separated from each other, each of them collapses. The results presented in this work show that heterodimers whose protein components coexist in different phases, can have intriguing physical properties with potentially important biological consequences. © *2012 American Institute of Physics*. [http://dx.doi.org/10.1063/1.4734019]

I. INTRODUCTION

Protein oligomers constitute the majority of functional units in living cells as, e.g., the mitochondrial complex III (Ref. 1) or complex I.[2] The mutual interactions usually influence the structure of each of the components substantially. Natively-unfolded proteins,[3] which attain a defined structure only when bound to a protein counterpart,[4] are an extreme example. This process resembles the transition from a liquid to a solid phase. Furthermore, protein aggregation into large β-sheet fibrils can cause conformational diseases termed amyloidoses.[5,6] The carcinogenic outcome of aberrant expression of some proteins is also a result of a change of the pattern of the interactions with their partner proteins.[7]

The problem of protein oligomerization and aggregation has received great attention,[8–10] starting from the very beginning of research on protein structure, and there now exists an enormous body of experimental[11–14] and theoretical[15–18] research on the subject. Different oligomers/aggregates are mainly treated from the point of view of the type of fold[8,9] or specific interactions that stabilize them,[19,20] rather than in terms of the identification of their universal, more quantitative features.

On the other hand, the thermodynamic and spectroscopic data of proteins in solution are often interpreted in terms of the two-state model, according to which the ensemble of protein molecules is a superposition of the folded and the unfolded phase.[21] Thus, a protein in solution can be regarded as two coexistent phases, a concept quite familiar in polymer research. Protein heterodimers can thus be treated by the conventional mean-field Flory-Huggins theory,[22–25] which enables one to compute the fraction of each monomer in the collapsed (folded) phase. The conventional Flory-Huggins theory,[22,23] however, assumes that individual protein chains can be modeled in terms of random walks. But biologically active proteins are mainly in a collapsed phase. Moreover, the description in terms of volume fraction of the folded and the unfolded conformation cannot account for the difference in the stability of different protein segments. Thus, a mean-field approach[22,23] is not appropriate to describe individual proteins or protein oligomers at a detailed level.

The phase structure of a system in thermal equilibrium is determined by the minima of the thermodynamic Helmholtz free energy F, which is a sum of the internal energy U and the entropy S [Eq. (1)]. For a protein, this general principle has been formulated by Anfinsen[26] as the *thermodynamic hypothesis*, according to which the native structure of a protein is the global minimum of its Helmholtz free energy under physiological conditions.

$$F = U - TS. \tag{1}$$

[a)] Electronic mail: Andrei.Krokhotine@cern.ch.
[b)] Electronic mail: adam@chem.univ.gda.pl.
[c)] Electronic mail: Antti.Niemi@physics.uu.se.
[d)] Electronic mail: has5@cornell.edu.

The entropy is a measure of conformational complexity. In general, an increase in the volume of the conformational space also leads to an increased phase complexity.[24,27] Consequently, macromolecular complexes such as protein heterodimers should have a much richer phase structure than an ensemble of essentially structureless point-like molecules. Various factors can influence the volume of the available conformational space, such as protein architecture and the degree of polymerization. Even for a pair of two distinct proteins, the number of all possible conformations can be enormous.[24,27]

In this work, the structures of protein heterodimers are analyzed in terms of coexisting polymer phases. Instead of a conventional description, based on the notions of collapsed or not collapsed states, a more refined analysis is performed. For this the dependence of the radius of gyration on chain length to distinguish different phases, which differ by chain compactness, is utilized as the order parameter. The statistics of oligomeric proteins in the protein data bank (PDB) (Ref. 28) is analyzed first and then, as a concrete example, focus is placed on the behavior of a dimer that originates from the transmembrane amyloid precursor protein (APP) by proteolytic cleavages.[29,30] The APP has several isoforms that can be present in many organs; however, its exact physiological function remains under debate.[30–32] Full understanding of the proteolytic processing of APP is also lacking.[30–32] But there appears to be two different pathways for proteolytic cleavage, a nonamyloidogenic, and an amyloidogenic one. The latter gives rise to the extracellular Aβ42 peptides that may be involved in Alzheimer's disease.[30,33] Both pathways also give rise to an isoform of the intracellular APP (AICD).[34] In isolation, AICD is presumed to be an intrinsically unstructured protein.[30–32] There are no structural data available in the protein data bank that could be used to elucidate its physical properties. But it can bind with the Fe65 family of nuclear multidomain adaptor proteins.[30,34–37] Upon binding, AICD assumes a regular form that can be analyzed by x-ray crystallography. In the present article, the x-ray structure that is described in the PDB under the code 3DXC (Ref. 38) is treated. It is a complex of a 28-residue segment of AICD with Fe65, that has 130 residues. The closely related 3DXD and 3DXE complexes, also exist. These can be analyzed similarly and with similar conclusions. It is found that the 3DXC complex has very interesting physical properties that sets it apart from all but a very few protein complexes. This AICD/Fe65 complex is an example of an apparently previously unrecorded but seemingly systematic phenomenon in the context of protein research that is termed *protein phase coexistence* in this work: Like ice together with water, the two proteins in this complex are apparently in two different phases. An oligomer that displays this phenomenon of phase coexistence under physiological conditions is for sure an interesting object for future research. But since the AICD/Fe65 complex has the supplementary potential of being an important piece in the puzzle to understand Alzheimer's disease, there are many good reasons to investigate its physical properties.

This paper is organized as follows: In Subsection II A, the use of the radius of gyration and its variation with the number of residues as a measure of compactness is reviewed. Subsections II B–II D develop the mathematical formalism to study the structure and dynamics of proteins: In Subsection II B, it is explained how the C$^\alpha$ backbone can be described in terms of its virtual-bond and virtual-torsion angles. In Subsection II C, the description of protein geometry in the collapsed phase in terms of a soliton solution of the discretized nonlinear Schrödinger equation (DNLS) is outlined. Finally, in Subsection II D, the coarse-grained UNRES model[39–42] of polypeptide chains is described briefly. Use of this coarse-grained approach elongates the time scale by over three orders of magnitude with respect to all-atom simulations,[43,44] thus enabling us to observe significant conformational changes in comparatively short simulation time.

In Subsection III A, oligomers in the PDB are analyzed. Two general classes of phase coexistent oligomers are identified. It is observed that the Alzheimer's disease related AICD/Fe65 dimer is an interesting example, with neither of the two component chains in the collapsed phase.

In Subsection III B, the structure of AICD in the AICD/Fe65 dimer with PDB code 3DXC is analyzed on general grounds. It is shown that, despite being composed of two solitons of the DNLS equation, the AICD is in a linear rod-like phase. It is found that there are two natural positions for the first soliton, and it is suggested that this could lead to a genetic switch.

Finally, in Subsection III C, Langevin dynamics simulations with the UNRES energy function[39–42] are used to confirm that, in isolation, both AICD and Fe65 of the 3DXC dimer collapse into the space-filling phase of biologically active proteins.

A brief overview of the results in Sec. IV concludes the article.

II. METHODS

A. Radius of gyration as an order parameter

It has been known for a long time[22,23,25] that a linear polymer chain such as a single-chain protein has a non-trivial phase structure that depends both on the chemical properties and on the temperature of the polymer-solvent system. In a good solvent environment, such as aqueous denaturating agents, the interactions between the polymer segments and the solvent molecules cause the polymer to expand and the polymer behaves like a self-avoiding random walk. In a poor solvent environment, such as water, the polymer-polymer self-interactions dominate, and the polymer collapses into a space-filling conformation. These two regimes are separated by the Θ-point, at which the polymer can be modeled by an ideal chain. Moreover, in the thermodynamical limit where the number N of amino-acid residues is very large, certain aspects of the phase structure become *universal* .[45–48] An example of a universal quantity is the compactness index ν. To define this quantity, the radius of gyration R_g (Refs. 22–25) is introduced.

$$R_g^2 = \frac{1}{2N^2}\sum_{i,j}(\mathbf{r}_i - \mathbf{r}_j)^2, \tag{2}$$

where $\mathbf{r}_i$ are the coordinates of the atoms in the polymer. For a protein, for simplicity, $\mathbf{r}_i$ may be restricted to run over the coordinates of only the backbone C$^\alpha$ atoms. The compactness index ν governs the large-N asymptotic form of Eq. (2). When the number N of amino-acid residues becomes very large, then[49]

$$R_g^2 \overset{N \text{ large}}{\longrightarrow} R_0^2 N^{2\nu}(1 + R_1 N^{-\delta_1} + \cdots), \tag{3}$$

where R_0 is the form factor that characterizes the length scale (in ångstrøms) in the large-N limit, δ_1, etc. are critical exponents, R_1, etc. are the amplitudes; the terms with δ_1, etc. and R_1, etc., account for finite-size corrections. Besides the compactness index ν, the critical exponents δ_1, etc., are universal quantities,[49] but the amplitudes R_0, R_1, etc., are not universal.[49]

As a universal quantity, ν is independent of the detailed atomic structure. Different values of ν correspond to the different phases of the protein.[22–25,27] The four mean-field values of ν are given by Eq. (4)[22–25,27,49,50]

$$\nu = \begin{cases} 1/3 \\ 1/2 \\ 3/5 \\ 1 \end{cases}. \tag{4}$$

Under physiological, i.e., poor solvent conditions, in which a single-chain protein collapses into the space-filling conformation,[22,23] the mean field exponent $\nu = 1/3$. For an ordinary random-walk chain, the mean field value is $\nu = 1/2$. This phase appears at the Θ-point[22,23] that separates the collapsed phase from the high-temperature self-avoiding random walk phase for which the Flory value $\nu = 3/5$ is found. Finally, when $\nu = 1$, the protein loses its inherently fractal structure and scales like a rigid rod.[23,51] Examples of polypeptide structures that can represent this phase are monotonic α-helices or polyproline II helices.

The mean-field values of the critical exponents ν, δ_1, etc., in Eq. (3) are usually modified by fluctuations. For example, in the universality class of the self-avoiding random walk, the modified values are[52,53]

$$\begin{aligned} \nu &= 0.5880 \pm 0.0015, \\ \delta_1 &= 0.47 \pm 0.03. \end{aligned} \tag{5}$$

A subsequent numerical Monte Carlo evaluation of the critical exponents of Eq. (5), gave very similar values,[49]

$$\begin{aligned} \nu &= 0.5877 \pm 0.0006, \\ \delta_1 &= 0.56 \pm 0.03. \end{aligned} \tag{6}$$

In the case of the collapsed phase where $\nu \approx 1/3$ the value of δ_1 is not known to us. But from (5) and (6) one can estimate that even in the case of a relatively short polypeptide chain, such as the 28-residue segment of AICD considered here, the finite-size correction to the radius of gyration can not be very large. With (6), one estimates that in the $\nu \approx 1/3$ phase when $N = 28$

$$\begin{aligned} R_g &\approx R_0 \cdot 28^{1/3}\left(1 + \frac{1}{2} R_1 \cdot 28^{-0.56}\right) \\ &\approx 3.03 \cdot R_0 (1 + 0.08 \cdot R_1). \end{aligned} \tag{7}$$

Consequently, with $R_1 \sim \mathcal{O}(1)$, only the leading term needs to be retained, even for protein chains with not more than around 28 or so residues.

The pre-factor R_0 and the finite-size correction coefficient R_1 can, in principle, be computed for a protein. These factors contain all the effects of temperature and chemical microstructure, and all the atomic level details of the protein. Consequently, for a protein, relation, Eq. (3) becomes valuable only when the numerical value of R_0 is either unique or can assume no more than a small number of clearly identifiable different values. It is found that this is indeed the case, i.e., for proteins, the values of R_0 are very restricted, but different, for collapsed proteins and for proteins that are not in the collapsed phase. Moreover, it appears that, when N increases, the detailed amino acid structure of a protein becomes increasingly irrelevant in determining the value of the radius of gyration, i.e., for long protein chains, any inhomogeneity of the amino-acid sequence can be treated essentially as a finite size correction in Eq. (3).

In detailed studies of proteins, the radius of gyration has until now been utilized only sparsely.[50,54–57] It appears to us that, commonly, the value of R_g has been viewed as nothing more than a rough measure, with a systematic change in its value indicating that protein collapse has taken place. In the present article, another point of view is presented. It is proposed that the seemingly quite small diversity in the values that R_0 assumes for proteins, both in the collapsed and in the other available phases, suggests that Eq. (3) could assume a much wider rôle in protein research. The goal sought here is to scrutinize this potential usefulness of the radius of gyration as a quantitative order parameter. It is aimed to develop both ν and R_0 here into practical tools to understand protein conformations and phase structure, not only for monomers but also for heterodimers.

B. Protein backbone geometry

We set the stage by explaining how protein geometry is described in terms of the C$^\alpha$ backbone. For this purpose, the positions $\mathbf{r}_i$ of the backbone C$^\alpha$ atoms in Eq. (2) are utilized to introduce the unit tangent vectors

$$\mathbf{t}_i = \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{|\mathbf{r}_{i+1} - \mathbf{r}_i|}, \tag{8}$$

the unit binormal vectors

$$\mathbf{b}_i = \frac{\mathbf{t}_{i-1} \times \mathbf{t}_i}{|\mathbf{t}_{i-1} \times \mathbf{t}_i|}, \tag{9}$$

and the unit normal vectors

$$\mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i. \tag{10}$$

The orthogonal triplet $(\mathbf{n}_i, \mathbf{b}_i, \mathbf{t}_i)$ determines a frame at the positions $\mathbf{r}_i$ of the backbone. The discrete virtual-bond angles and the discrete virtual-torsion angles are defined by Eqs. (11) and (12), respectively.

$$\theta_i \equiv \theta_{i+1,i} = \arccos(\mathbf{t}_{i+1} \cdot \mathbf{t}_i), \tag{11}$$

$$\gamma_i \equiv \gamma_{i+1,i} = \sigma \times \arccos(\mathbf{b}_{i+1} \cdot \mathbf{b}_i), \tag{12}$$

with

$$\sigma = \mathrm{sgn}[(\mathbf{b}_{i-1} \times \mathbf{b}_i) \cdot \mathbf{t}_i]. \tag{13}$$

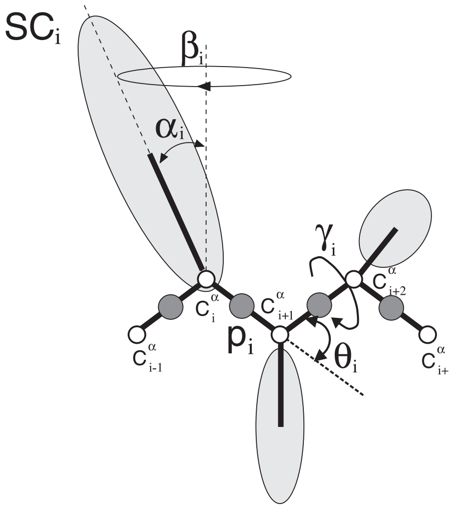

FIG. 1. Definitions of the variables of the UNRES model. The virtual-bond angle θ_i is determined by the three C$^\alpha$ carbons at sites $i, i+1, i+2$ and is defined as the angle between the C$^\alpha_i \cdots$ C$^\alpha_{i+1}$ virtual-bond vector and the C$^\alpha_{i+1} \cdots$ C$^\alpha_{i+2}$ virtual-bond vector [Eq. (11)]. It should be noted that, for consistency with the notation of Sec. II B, the angles θ used in this work are complements of the original angles θ, i.e., $\pi - \theta$ (see, e.g. Ref. 39). The C$^\alpha$ carbon atoms are represented by small open circles. The virtual-bond-dihedral angle γ_i it the angle between the two planes, determined by the C$^\alpha$ at sites (i, $i+1$, $i+2$) and ($i+1$, $i+2$, $i+3$) [Eqs. (12) and (13)]. In addition, the UNRES energy function [Eqs. (32)–(35)] involves the following structure, shown in the Figure: The interaction sites are peptide-bond centers (p), and side-chain ellipsoids of different sizes (SC) attached to the corresponding α-carbons with different virtual-bond lengths b_{SC}. The UNRES energy is also a function of the coordinates of the SC and p sites which are functions of ($\theta, \gamma, \alpha, \beta$) and also contains terms that depend explicitly on these angles.

A graphic definition of these angles, and all the other geometric quantities that are used in this paper, is given in Figure 1.

Conversely, if the angles (θ_i, γ_i) are known, Eq. (14) can be used to construct the frame at position $i+1$ iteratively from the frame at position i. Once all the frames are computed, the Cartesian coordinates of the entire C$^\alpha$-trace can be calculated from Eq. (15).

$$\begin{pmatrix} \mathbf{n_{i+1}} \\ \mathbf{b}_{i+1} \\ \mathbf{t}_{i+1} \end{pmatrix} = \begin{pmatrix} \cos\theta\cos\gamma & \cos\theta\sin\gamma & -\sin\theta \\ -\sin\gamma & \cos\gamma & 0 \\ \sin\theta\cos\gamma & \sin\theta\sin\gamma & \cos\theta \end{pmatrix}_{i+1,i} \begin{pmatrix} \mathbf{n_i} \\ \mathbf{b}_i \\ \mathbf{t}_i \end{pmatrix}$$

$$\equiv \mathcal{R}_{i+1,i} \begin{pmatrix} \mathbf{n_i} \\ \mathbf{b}_i \\ \mathbf{t}_i \end{pmatrix}, \tag{14}$$

$$\mathbf{r}_k = \sum_{i=0}^{k-1} |\mathbf{r}_{i+1} - \mathbf{r}_i| \cdot \mathbf{t}_i, \tag{15}$$

where $\mathbf{r}_k$ is the vector of the Cartesian coordinates of the kth C$^\alpha$ atom. With no loss of generality $\mathbf{r}_0 = 0$ and $\mathbf{t}_0$ can be set to point into the direction of the positive z-axis.

It should be noted that Eq. (15) does not involve the vectors $\mathbf{n}_i$ and $\mathbf{b}_i$. Consequently, any linear combination of these two vectors can be selected in constructing the backbone. For this, the frame ($\mathbf{n}_i, \mathbf{b}_i$) is rotated by an angle Δ_i leaving $\mathbf{t}_i$ intact,

$$\begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_i \to e^{\Delta_i T^3} \begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_i = \begin{pmatrix} \cos\Delta_i & \sin\Delta_i & 0 \\ -\sin\Delta_i & \cos\Delta_i & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_i, \tag{16}$$

where the tensor $(T^i)_{jk}$ is defined by Eq. (17).

$$(T^i)_{jk} = \epsilon_{ijk},$$
$$[T^i, T^j] = \epsilon_{ijk} T^k, \ i = 1, 2, 3; \ j = 1, 2, 3; \ k = 1, 2, 3, \tag{17}$$

with

$$\epsilon_{ijk} = \begin{cases} 1 & \text{if } \{ijk\} \text{ is an even permutation of } \{1, 2, 3\}, \\ -1 & \text{if } \{ijk\} \text{ is an odd permutation of } \{1, 2, 3\}. \end{cases} \tag{18}$$

If the vectors $\mathbf{n}$ and $\mathbf{b}$ are combined into the complex vector

$$\mathbf{n} + i\mathbf{b}$$

Equation (16) can be recast into Eq. (19).

$$\mathbf{n}_i + i\mathbf{b}_i \to e^{i\Delta_i}(\mathbf{n}_i + i\mathbf{b}_i) \equiv \mathbf{e}^1_i + i\mathbf{e}^2_i. \tag{19}$$

A direct computation shows that the frame rotation of Eq. (16) induces the following transformations in Eq. (14),

$$\theta_i \, T^2 \ \to \ e^{\Delta_i T^3}(\theta_i T^2)\, e^{-\Delta_i T^3}, \tag{20}$$

$$\gamma_i \ \to \ \gamma_i + \Delta_{i-1} - \Delta_i. \tag{21}$$

In what follows, Eqs. (20) and (21) are utilized to extend the fundamental range $[0, \pi]$ of the virtual-bond angle θ_i into $[-\pi, \pi]$ $mod(2\pi)$, and the range of γ_i to $[-\pi, \pi)$ $mod(2\pi)$. This extension is compensated for in the fundamental range of θ_i by introducing the following transformation

$$\begin{aligned} \theta_k &\to -\theta_k \qquad \text{for all } k \geq i \\ \gamma_i &\to \gamma_i - \pi. \end{aligned} \tag{22}$$

It should be noted that regular protein secondary structures correspond to definite values of (θ_i, γ_i). For example, for the standard α-helix

$$\alpha - \text{helix}: \begin{cases} \theta \approx \frac{\pi}{2} \\ \gamma \approx 1 \end{cases}, \tag{23}$$

and for the standard β-strand

$$\beta - \text{strand}: \begin{cases} \theta \approx 1 \\ \gamma \approx \pi \end{cases} \tag{24}$$

with the angles in radians. All the other regular secondary structures such as 3/10 helices, left-handed helices, etc., can be described in a similar way.

For the protein backbone with all peptide groups in a planar trans conformation, the following can be set:

$$|\mathbf{r}_{i+1} - \mathbf{r}_i| = d \approx 3.8\,\text{Å}. \tag{25}$$

For the present purposes, deviations from the value of Eq. (25) are negligible even if the peptide bond involves proline,[58] if only the trans peptide groups are considered. The excluded-volume (steric) constraint [Eq. (26)] can also be imposed.

$$|\mathbf{r}_i - \mathbf{r}_k| \geq 3.8\ \text{Å} \quad \text{for} \ \ |i-k| \geq 2. \tag{26}$$

C. Soliton description of protein-backbone geometry

Despite the apparent complexity of interactions that are described by the various molecular dynamics force fields and realistic coarse-grained energy functions, collapsed proteins display a surprisingly small variety in their shapes. There seems to be a self-organizing principle at work, that strongly limits the diversity among the biologically active protein structures. This is also reflected in Figure 2(a) that the values of the *a priori* highly variable R_0 in Eq. (3) are very restricted. Indeed, the presence of a universal self-organizing principle in protein folding is manifested in the PDB structures.[28] For example, thus far the structural classification scheme SCOP (Ref. 59) has identified around 1.400 unique folds in the PDB while, in CATH,[60] there are currently around 1.300 topologies. These numbers have remained largely unchanged during the last 3–4 years, suggesting that the number of different protein folds is quite limited, and probably most of them have already been found.[59–62] The success of SCOP and CATH and other approaches such as FSSP (Ref. 63) in classifying proteins confirms that proteins are built in a modular fashion, from a relatively small number of elemental components.

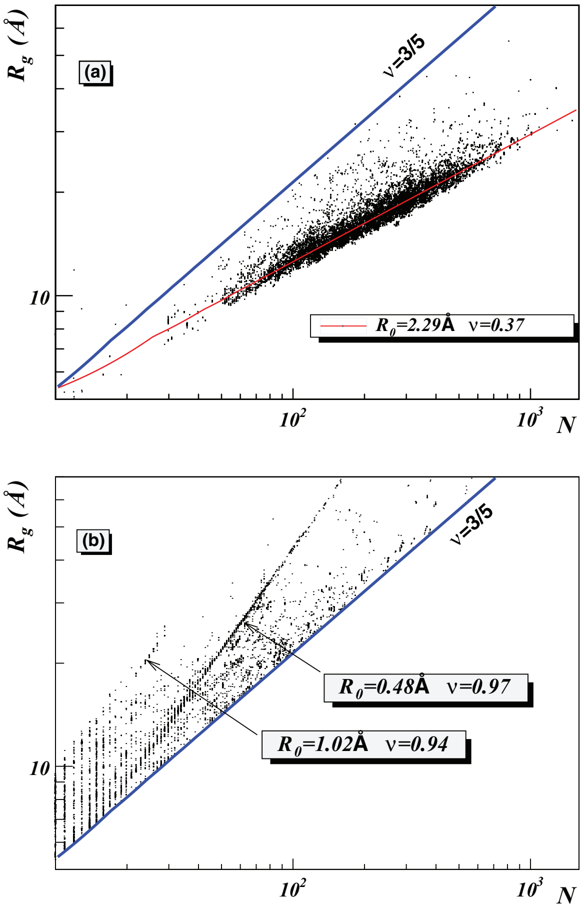

FIG. 2. (a) The (N, R_g) distribution of all individual single-chain PDB proteins with resolution less than 2.0 Å and with less than 30% homology. The lower line ($\nu = 0.37$) describes mostly α-helical proteins and the top line is a $\nu = 3/5$ Flory line. There are practically no single-chain proteins above the Flory line. (b) The (N, R_g) distribution for all multi-chain proteins from the current PDB that are above the Flory line. The two clusters given by equation (38), are clearly visible. The values of R_0 and ν shown in the graphs were determined by linear regression.

In Ref. 64, it has been observed that the self-similar modular structure of proteins can be described mathematically in terms of a soliton solution[65,66] to a discrete version of the nonlinear Schrödinger (DNLS) equation.[67–70] The original DNLS equation already shares a long history with protein research. It was introduced by Davydov[67] who also showed that the equation supports a soliton solution. He proposed that the soliton explains how a localized deformation (and its energy excitation) propagates without dissipation along the α-helix. He also argued that a soliton that becomes trapped could evoke a deformation of the protein shape.

A soliton is the archetype structural self-organizer.[65,66] A soliton arises when non-linear interactions merge elementary constituents such as atoms into a localized collective excitation that is stable against small perturbations and cannot decay, unwrap, or disentangle. Solitons are extremely widely studied objects that can appear in many practical and theoretical scenarios. For example, solitons can be deployed for data transmission in transoceanic cables, for conducting electricity in organic polymers and describing chemical energy transportation in proteins.[66] Many phenomena such as the formation of the morning glory cloud in the atmosphere, the Meissner effect in superconductivity and dislocations in liquid crystals[65,66] can be explained in terms of solitons. Solitons also model hadronic particles, cosmic strings, and magnetic monopoles in high energy physics.[65]

Following the justification in Ref. 71, a soliton, which is a solution to the equation of motion of the backbone Helmholtz free energy of Eq. (27) for the virtual-bond and virtual-torsion angles, is considered.

$$F = -\sum_{i=1}^{N-1} 2\,\theta_{i+1}\theta_i + \sum_{i=1}^{N}\left\{2\theta_i^2 + q\cdot\left(\theta_i^2 - m^2\right)^2 + \frac{d}{2}\,\theta_i^2\gamma_i^2 \right.$$
$$\left. -\,a\gamma_i + \frac{c}{2}\gamma_i^2\right\}, \tag{27}$$

where θ_i is ith virtual-bond angle, γ_i is ith virtual-bond-dihedral (torsion) angle, and a, c, d, q, and m are parameters.[64,71–77] Equation (27) has been derived and motivated in detail in Refs. 64 and 71–77. Here it suffices to state that Eq. (27) can be shown to bethe long-distance limit that describes the full microscopic energy of a folded protein in the universal sense of Refs. 45–48. As such, it does not explain the details of the (sub)atomic level mechanisms that give rise to protein folding.

The soliton solution is constructed by seeking the minimum of Eq. (27).[64,71,75,76] The necessary condition for the minimum is finding the zero of the gradient of F in the virtual-bond angles θ and in the virtual-bond dihedral angles γ. The solution of this problem is the solution of a system of $2N - 5$ nonlinear equations in $2N - 5$ unknowns (where N is the number of residues). In order to obtain this solution, the virtual-bond-dihedral angles γ are first expressed as functions of the virtual-bond angles θ, as given by Eq. (28).[64,71,75,76]

$$\gamma_i[\theta] = \frac{a}{c + d\,\theta_i^2} \equiv \frac{u}{1 + v\,\theta_i^2} \tag{28}$$

with $u = a/c$ and $v = d/c$. By inserting Eq. (28) into Eq. (27), the virtual-bond-dihedral angles γ are eliminated and a system of Eqs. (29) for the motion of the virtual-bond angles θ is obtained.

$$\theta_{i+1} = 2\theta_i - \theta_{i-1} + \frac{dV_{pot}[\theta]}{d\theta_i^2}\theta_i \quad (i = 1, \ldots, N), \tag{29}$$

where $\theta_0 = \theta_{N+1} = 0$ and

$$V_{pot}[\theta] = \frac{a}{c + d\,\theta^2} + 2(1 - qm^2)\theta^2 + q\,\theta^4. \tag{30}$$

Here the familiar structure of the DNLS equation is recognized.[67,68] The only difference is in the first term on the right-hand side in Eq. (30). However, because this term contains θ in the denominator, its variation with θ is not that pronounced as the variation of the other two terms, which are proportional to the second and the fourth power of θ, respectively. Moreover, because $|\theta| > 1$ radian for proteins, it turns out that the first term is small in value compared to the other terms.

The profile of the so-called dark soliton solution[67–70] to Eq. (29) can be constructed numerically by following the iterative procedure of Ref. 71. But its explicit form until now has not been found in terms of elementary functions. However, an *excellent* approximation is obtained by *naively* discretizing the continuum dark nonlinear Schrödinger equation (NLSE) soliton[76]

$$\theta_i = \frac{(\mu_1 + 2\pi M_1)\cdot\exp[\sigma_1(i-s)] - (\mu_2 + 2\pi M_2)\cdot\exp[-\sigma_2(i-s)]}{\exp[\sigma_1(i-s)] + \exp[-\sigma_2(i-s)]}, \tag{31}$$

where s is a parameter that determines the center of the soliton. Equation (31), together with Eq. (28), represents a profile of θ and γ that is variable from site to site, as characterized by a loop in a protein. The $\mu_{1,2} \in [0, \pi]$ are parameters, whose values are determined by the adjacent helices and strands. M_1 and M_2 constitute the integer parts of $\mu_{1,2}$, and for simplicity $M_1 = M_2 \equiv M$ is taken. To satisfy the monotonic character of the profile of Eq. (31), the experimental values of θ_i have to vary monotonically along the amino-acid sequence. Otherwise, a multiple of 2π is added to the experimental values. This does not affect the backbone geometry because θ_i's are defined mod (2π). The integer M determines how many times θ_i covers its fundamental domain $[-\pi, \pi)$ when the soliton is traversed once. It should be noted that negative values of θ_i are related to positive values of θ_i by Eq. (22). Finally, *only* σ_1 and σ_2 are intrinsically specific parameters for a given loop. But they specify only the length of the loop, not its shape which is determined by the functional form of Eq. (31) and, as in the case of $\mu_{1,2}$, they are combinations of the parameters in Eq. (30).

The virtual-torsion angles $\gamma_i, i = 1, 2, \ldots, N-3$ are computed by substituting Eq. (31) into Eq. (28). Since there are only two independent parameters u and v in Eq. (28), the profile of γ_i is dependent entirely on θ_i, and on the structure of the adjacent regular secondary structures.

D. UNRES model of polypeptide chains

In the UNRES model,[39–42] a polypeptide chain is represented as a sequence of α-carbon atoms (C$^\alpha$s) with attached side chains (SCs) and peptide groups positioned halfway between the consecutive C$^\alpha$s. Only the side chains and the peptide groups are interaction sites, while the C$^\alpha$s assist in geometry definition. The side chains are represented by ellipsoids of revolution, each with dimensions that characterize the amino acid in question. All definitions are summarized in the legend of Figure 1: The center of the ith side-chain ellipsoid is located a distance $b_{SC}(i)$ from the corresponding C$^\alpha$, in the direction $[\alpha_{SC}(i), \beta_{SC}(i)]$ where $\alpha_{SC}(i)$ and $\beta_{SC}(i)$ are the positional and torsional angles in a spherical coordinate system that is centered at the corresponding C$^\alpha$. Among the dynamical variables that are not accounted for explicitly are the solvent degrees of freedom, the angles of rotation about side-chain bonds, the angles of rotation of the peptide groups about the C$^\alpha\cdot\cdot\cdot$C$^\alpha$ virtual bonds etc. These are all included effectively, in the numerical values of the various parameters.

Schematically, the UNRES energy can be grouped into a sum of three terms,

$$E_{UNRES} = E_1 + E_2 + E_3 \tag{32}$$

E_1 itself is a sum of three terms that involve only the backbone virtual-bond and virtual-torsion angles,

$$E_1 = w_b \sum_i U_b(\theta_i) + w_{tor}\cdot f_2(T)\sum_i U_{tor}(\gamma_i) + w_{tord}\cdot f_3(T)\sum_i U_{tord}(\gamma_i, \gamma_{i+1}), \tag{33}$$

where $f_2(T)$ and $f_3(T)$ are temperature-dependent multipliers. The first term models virtual-bond-angle bending, the other two are torsional and double-torsional energies. Next,

$$E_2 = w_{SC}\sum_{i<j} U_{SC_iSC_j} + w_{SC_p}\sum_{i\neq i} U_{SC_{ipj}} + w_{pp} f_2(T)\sum_{i<j-1} U_{pipj} + w_{rot}\sum_i U_{rot}(\alpha_{SC_i}, \beta_{SC_i}\theta_i) + w_{bond}\sum_i U_{bond}(d_i). \tag{34}$$

These are terms that describe the following interactions, respectively: Side-chain-side-chain, side-chain-peptide,

peptide-peptide, side-chain-local, and virtual-bond distortion. Finally,

$$E_3 = \sum_{m=3}^{6} w_{corr}^{(m)} f_m(T)\, U_{corr}^{(m)} + w_{SS} \sum_i U_{SS;i} \tag{35}$$

is included. These two terms describe multi-body (correlation) interactions and the formation of disulfide bonds, respectively. The temperature-dependent multipliers are Bolzmann-like,

$$f_m(T) = \frac{\ln(e + e^{-1})}{\ln\left\{\exp\left[\left(\frac{T}{T_0}\right)^{m-1}\right] + \exp\left[-\left(\frac{T}{T_0}\right)^{m-1}\right]\right\}} \tag{36}$$

with $T_0 = 300$K. The ws are the weights of the respective effective energy terms, which are determined by force-field calibration to reproduce the structure and folding thermodynamics of selected training protein(s). Langevin dynamics was implemented[43,44] with UNRES to study protein folding pathways; later,[78,79] the replica-exchange[80] and multiplexed replica exchange[81] algorithms were implemented to simulate ensembles of the proteins under study. In this work, the treatment is restricted to canonical Langevin simulations with the use of the force field calibrated with the 1GAB protein.[40]

III. RESULTS AND DISCUSSION

A. Phase co-existence in protein oligomers

We now proceed to inspect the phases of some protein oligomers, the structures of which are available from the PDB, in terms of the radius of gyration, with the compactness index ν in Eq. (3) as the order parameter. The possibility of a phase co-existence in oligomers is of particular interest.

The following line of arguments proposes that it does indeed make sense to employ Eq. (3) to study the phase structure of individual proteins: The data in SCOP (Ref. 59) and CATH (Ref. 60) and the results reported in Ref. 64 establish that most proteins have a self-similar structure over macromolecular distance scales. Since R_0 is the only relevant length scale in the limit of long proteins, the large scale self-similarity implies that R_0 can have only very restricted values. It is found that these different values are characteristics of the different fold types, such as mostly-α, mostly-β *etc.*

In Figure 2(a), the radius of gyration R_g is plotted as a function of the number of residues for all those individual single-chain proteins in the PDB that have resolution less than 2.0 Å and sequence homology less than 30%. With a few exceptions, the values of R_g assemble around a line described by Eq. (3) with $\nu \approx 1/3$ for a collapsed protein. In Figure 2(a), the line along which the mostly-α structures are clustered (that are dominant in the PDB data) is colored red. This line is given by

$$R_g^{\alpha} \approx R_0 \cdot N^{\nu} \approx 2.29 \cdot N^{0.37} \quad (\text{Å}). \tag{37}$$

For the other subclasses such as α-and-β, mostly-β etc., the numerical values of R_0 and ν are slightly different. In the limit $N \to \infty$, it is expected that ν converges towards a unique value, while R_0 may assume a small number of different values that are determined by the corresponding subclasses. Similar results were obtained by Dewey[50] and by Dima and Thirumalai,[56] see also Ref. 57.

It is notable that in the PDB data displayed in Figure 2(a) the non-linear, finite size corrections in Eq. (7) are not visible. This confirms that the numerical value of the coefficient R_1 is small.

When the analysis is extended to include individual chains in protein oligomers, two clearly visible line-like clusters are found in the PDB that to our knowledge have not been reported previously. These two line-like clusters are identified in Figure 2(b) by using a clusterwise linear regression, as

$$\left.\begin{aligned} R_g^{(2)} &\approx 0.48 \cdot N^{0.973} \\ R_g^{(3)} &\approx 1.02 \cdot N^{0.94} \end{aligned}\right\} (\text{Å}). \tag{38}$$

In both clusters, the value of ν is very close to 1. Consequently these two clusters are in the same $\nu = 1$ universality class, the slight difference in the values of ν being a finite size effect.[82]

The line-like cluster $R_g^{(2)}$ includes several membrane proteins and viral capsomers. An example of the latter is the HIV envelope glycoprotein with PDB code 1AIK. The cluster $R_g^{(3)}$ is populated mainly by collagen proteins like 2CUO in the PDB. In both clusters, mostly oligomers that consist of several similar or nearly similar subchains are found, and each subchain is located in the same cluster in Eq. (38). Apparently, the interactions between the different subchains override the effects of the poor solvent environment, preventing the individual subchains from collapsing into the $\nu \approx 1/3$ cluster.

In Figure 3, the spectrum of the virtual-bond and virtual-torsion angles defined by Eqs. (11) and (12), respectively, is shown as an example, for both 1AIK and 2CUO. The self-similar structure, which is consistent with a rigid rod-like geometry, is clearly visible in both spectra (i.e., the values of the virtual-bond angles θ and the virtual-bond-dihedral angles γ vary little along the sequence). It should be noted that, for 1AIK, the (θ, γ) spectrum coincides with that of the α-helix [Eq. (23)]. For 2CUO, the virtual-bond angle has the β-strand value [Eq. (24)], while the virtual-torsion angle repeats a pattern in which two values are in the vicinity around –1.4–1.5 radians, followed by one value which is about –2.0 radians. This corresponds to the Gly-Pro-X backbone structure which is characteristic of a collagen.

Remarkably, it is found that there are also protein complexes in the clusters of Figure 2(b) that do not follow the structural patterns of membrane proteins, namely viral capsomers and collagens. In particular, a small number of hetero-oligomers that are composed of two or more proteins are identified, each on a *different* cluster defined by Eq. (3). Moreover, a number of dimers, in which one of the subchains is in the Θ-point cluster $\nu \approx 1/2$ with

$$R_g^{\theta} \approx 1.23 \cdot N^{0.508} \quad (\text{Å}), \tag{39}$$

while another subchain is in the $R_g^{(2)}$ cluster [Eq. (38)] of Figure 2(b), have been found here. To the extent that ν can be interpreted as an order parameter that characterizes different phases of a single polymer chain, the present observation suggests that these dimers are examples of proteins that

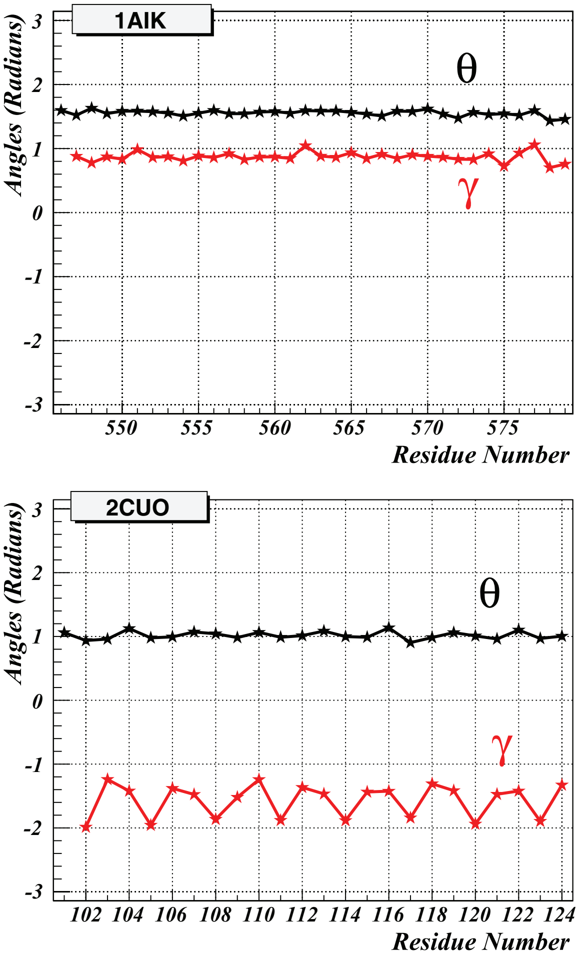

FIG. 3. The spectrum of the virtual-bond and virtual-torsion angles for 1AIK (top) and 2CUO (bottom), using PDB indexing. The black lines and symbols correspond to the virtual-bond angles θ and the red lines and symbols correspond to the virtual-bond dihedral angles γ, respectively.

display *phase coexistence*. As demonstrated by the emergence of two clearly visible separate line-like clusters in Figure 2(b), with a sparsely populated intervening region, there is indeed a sharp cross-over transition between the different classes. The cross-over is reminiscent of a phase transition, in the thermodynamical limit.

Two different families of such phase coexistent heterodimers in the set of PDB data used here have been found. The first family consists of proteins whose multiple subchains are in different phases, and are stable only in oligomeric state. The PDB codes for the 11 proteins of this family are listed in Table I. The second family consists of oligomers whose multiple subchains can also exist as stable proteins independently of each other. There are 16 different complexes in this data set and their PDB codes are listed in Table II.

TABLE I. First class of phase co-existent heterodimers. Single proteins but with multiple subchains that are in different phases.

1WDC	1G72	1GOT	1HTR	1LTS	2FP7
2RIV	3ABK	3ARC	3CX5	3DBO	

TABLE II. Second class of phase co-existent heterodimers; more than one protein with subchains in different phases.

1L2W	1JDH	1TH1	2F8X
2EPV	2PRR	2BFX	2D7C
2VGO	2K8F	2QKH	3EGG
3HTU	3HPW	3IXS	3DXC

In each of these phase coexistent proteins, there is a difference in the lengths of the subchains. The subchains in the $\nu \approx 1$ cluster $R_g^{(2)}$ are systematically shorter than the subchains in the $\nu \approx 1/2$ cluster, reflecting the respective loss of conformational entropy in complex formation.

It is possible that there are also heterodimers with other combinations of phase coexistence. For example one long subchain component could remain in the collapsed $\nu \approx 1/3$ phase while a more restrained short subchain is in one of the phases with $\nu \approx 1/2$, $\nu \approx 3/5$, or $\nu \approx 1$. However, no clear examples of these patterns are found in the present PDB data.

In Figure 4, the distribution of the individual subchains in the second class of phase coexistent complexes in the (N, R_g) plane is shown. The figure clearly displays how the longer subchains are located around the $\nu \approx 1/2$ cluster while the shorter chains are located on the $R_g^{(2)}$ cluster in Eq. (38).[83,84]

From a biological point of view, there are two notable hetero-oligomers in this second class. They are 2K8F and 3DXC (with the identical structures 3DXD, 3DXE). The first is the "molecular interpreter"[85] p300 in the cluster $R_g^{(2)}$ and the tumor suppressing protein p53 in the $\nu \approx 1/2$ cluster. The second is the Alzheimer disease related AICD/Fe65 complex[30,34–37] with the 28-residue AICD in the cluster $R_g^{(2)}$ and the Fe65 with 130 residues in the $\nu \approx 1/2$ cluster. In Sec. III A the apparently exceptional physical properties of the AICD/Fe65 complex 3DXC are considered.

B. Structure of AICD in the AICD/Fe65 dimer

The structure of the shorter AICD subchain of the AICD/Fe65 dimer is analyzed here in detail. Its crystal structure is shown in Figure 5. The spectrum of the

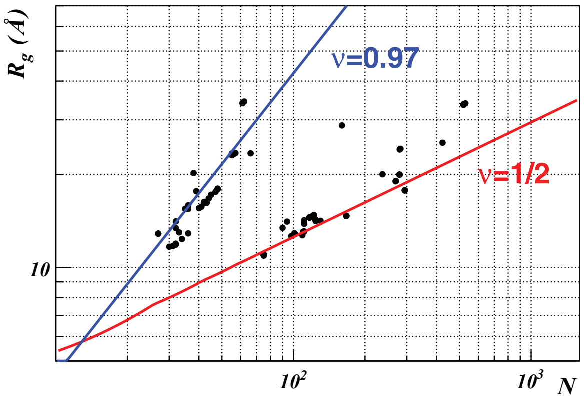

FIG. 4. The distribution of individual chains on the (N, R_g) plane in the second class of phase coexistent hetero-oligomers found here. The data clearly accumulate around the top line that describes the cluster $R_g^{(2)}$ and the bottom line, the latter describing a Θ-point cluster with best-fit values $R_0 = 1.234$ and $\nu = 0.508$.

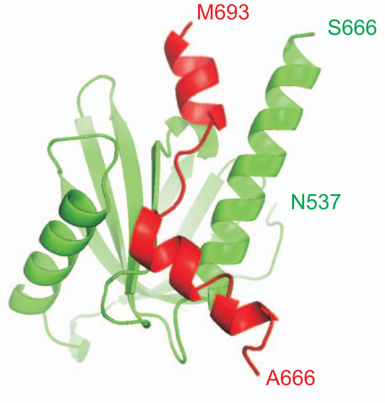

FIG. 5. Cartoon representation of the experimental structure of the AICD/Fe65 complex (PDB: 3DXC). Green: The Fe65 (longer) chain. Red: The AICD (shorter) chain. The first and the last residues of each chain are marked. Residue numbers have been taken from the 3DXC structure.

virtual-bond and virtual-torsion angles that describe its backbone are first inspected. The C$^\alpha$ coordinates $\mathbf{r}_i$ in the PDB structure 3DXC, in combination with Eqs. (8)–(12), are used to arrive at a (θ_i, γ_i) profile that is displayed in Figure 6(a). The AICD is located in the cluster $R_g^{(2)}$ with $\nu \approx 1$. However, in contrast to the spectra in Figure 3, its (θ_i, γ_i) profile does not display any self-similar structure. Instead, there appears to be a highly irregular behavior, in particular in γ_i, between the sites 678–685. In order to examine this behavior, the transformation defined by Eq. (22) is implemented to arrive at the profile shown in Figure 6(b). The profile for θ suggests that it can be considered as two successive hyperbolic-tangent-like profiles such as Eq. (31), i.e., as two consecutive solitons. This figure reveals that AICD appears to consist of a pair of two very closely located loops which are both preceded and followed by α-helices.

The (θ_i, γ_i) profile in Figure 6(b) is reminiscent of the soliton profiles that have been studied extensively in Ref. 64. There, the DNLS soliton is shown to describe the modular building blocks of folded proteins that are in the $\nu \approx 1/3$ universality class. It is found that over 92% of the PDB proteins can be modeled in terms of the soliton profile [Eqs. (28) and (31)] that has been introduced in Ref. 71, in terms of no more than 200 different sets of numerical parameters. Consequently, a question arises as to whether the (θ_i, γ_i) profile that appears in Figure 6(b) could also be described in terms of the parameters of the DNLS soliton.

For protein 3DXC, it is found that the profile of each of the two loops in Figure 6(b) can be described by using the Ansatz given by Eqs. (28) and (31). The parameter values are listed in Table III. The corresponding (θ_i, γ_i) profiles computed from these equations describe the first loop at sites 676–683 (in terms of PDB indexing) with RMSD precision of 0.29 Å and the second loop at sites 681–688 with RMSD precision of 0.17 Å (see Figure 7). Both of these precisions are substantially below the Debye-Waller fluctuation distances that can be computed from the experimental B-factors in the PDB data by using Eq. (40).

$$\langle \mathbf{r}^2 \rangle_{DW} = \frac{B}{8\pi^2}\ (\text{Å}^2). \qquad (40)$$

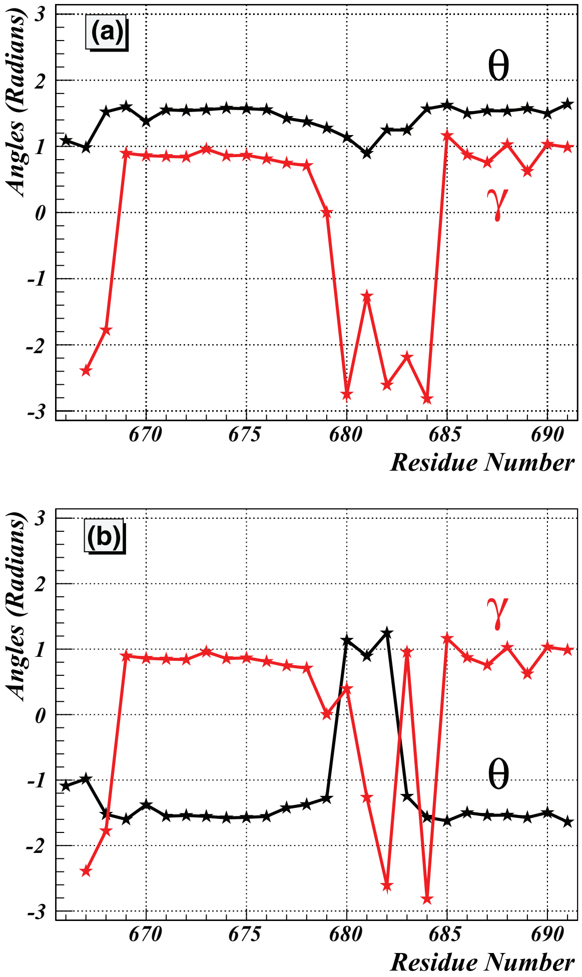

FIG. 6. (a) The spectrum of backbone virtual-bond angles θ_i (black line) and virtual-torsion angles γ_i (red line) for the AICD component of 3DXC (chain B). (b) The same spectrum after application of the transformation of Eq. (22) to reveal the soliton structure. The sites are indexed with residue numbers from the PDB.

TABLE III. Parameter values for the two solitons implied in Figure 6(b). For virtual-bond angles, Eq. (31) is used. For virtual-torsion angles Eq. (28) is used. It should be noted that the values of both θ and γ in these two equations are defined mod (2π). The large values of M enable us to describe the irregular structures in Figure 6(b). These irregularities are due entirely to multi-valuedness of the angular variables.

	loop	
Parameters of equation (31)	676–683	681–688
$\mu_1 + 2\pi N$	51.517	39.274
$\mu_2 + 2\pi N$	51.766	38.617
σ_1	2.984	3.327
σ_2	2.983	3.347
s	679.91	682.17
$u \times 10^{-4}$	345.9832942	0.4445445
$v \times 10^{4}$	− 1.6625	− 6.3344
number of matches	177	896

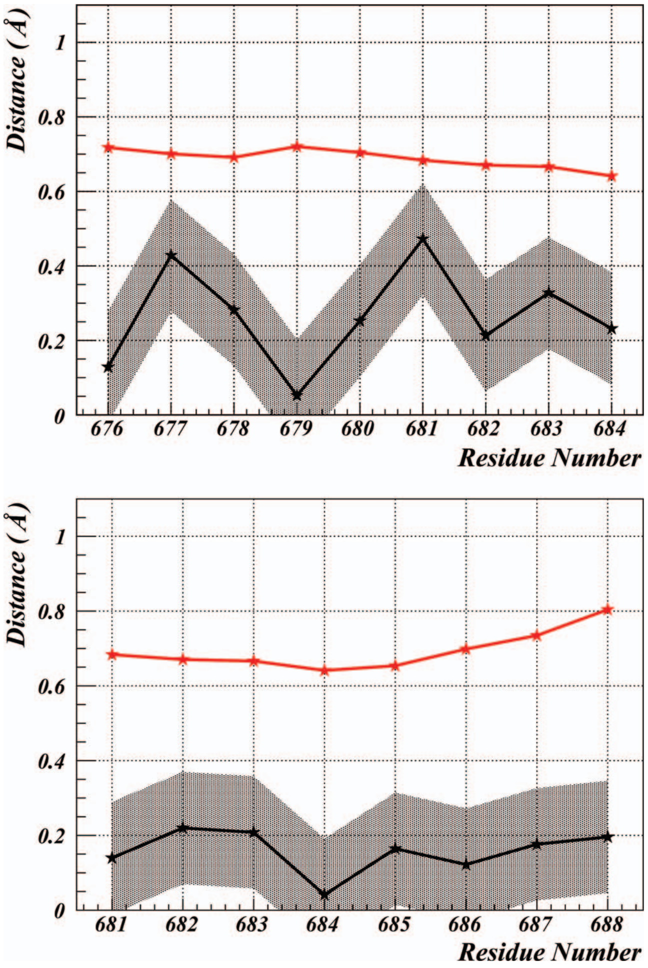

FIG. 7. The two solitons of 3DXC. Residues are indexed with the numbers from the PDB structure. The black line denotes the residue-wise difference between the coordinates computed from the soliton and those computed from the PDB conformation. The red line denotes the Debye-Waller (one standard deviation) fluctuation distance, computed from the B-factors in the PDB. The grey area describes the estimated 0.15 Å zero-point fluctuation distance around the solitons.

Consequently, there is a very high quality description of the backbone in terms of two DNLS solitons. In Figure 7, it is shown how the soliton structures compare to the 3DXC backbone and experimental Debye-Waller fluctuations around it.

In Table III, the number of times each of the two loops in 3DXC appears in the PDB are listed. From these, any two loops are identified whenever the mutual RMSD distance falls below 0.5 Å. This represents a typical value of the fluctuation distance in very high resolution x-ray structures. It should be noted that the Debye-Waller fluctuation distances for *every* site in the two loop regions of 3DXC are all above 0.65 Å (Figure 7).

The analysis of 3DXC establishes that it is composed of two solitons that are both *abundant* in the $\nu \approx 1/3$ cluster of collapsed proteins. In particular, there is *nothing*, either in the regular secondary structures or in the loop regions of the AICD component in the 3DXC complex, that is *a priori* unusual for a collapsed protein in the $\nu \approx 1/3$ phase. Nevertheless, this particular protein chain is located in the cluster $R_g^{(2)}$ that describes linear rod-like $\nu \approx 1$ structures. No such single-chain protein has been found in the PDB above the $\nu = 3/5$ Flory line in Figure 2. It is proposed that the reason for this exceptional structure of AICD is a peculiarity of the phase co-existence identified in Figure 4. The mutual interaction between the AICD and Fe65 in the complex is so strong, that it overcomes the poor solvent effect and prevents AICD from collapsing into the $\nu \approx 1/3$ phase.

According to Davydov,[67] a soliton is a localized carrier of energy and propagates without dissipation along an α-helix. This propagation evokes a deformation in the shape of the protein. If the soliton becomes trapped along the chain, it causes a change in the shape of the protein.[67] The soliton pair along AICD is preceded by a long α-helix, and there is a proline at site 669. Since proline is an effective initiator of a loop in an isolated protein, the propagation of the first soliton along the α-helix until it becomes trapped by Pro(669) is proposed. It should be noted that there is also another proline at site 685, and it initiates the second soliton in its place.

Since the compactness index ν can have only discrete values [Eq. (4)], it can not change when the soliton propagates continuously as a localized deformation along the α-helix. In particular, when the two solitons that are apart from each other are continuously translated along the backbone by shifting the value of s in Eq. (31), a collapsed conformation with $\nu = 1/3$ can never be reached. The AICD will remain in the $\nu \approx 1$ phase [see the discussion under Eq. (4) in Sec. II A].

The explicit profile of the angles defined by Eqs. (11) and (12) is used to investigate what takes place when the first soliton propagates along the backbone towards Pro(669). To propagate a soliton, the value of the parameter s in Eq. (31) is shifted. In Figure 8, it is shown how the radius of gyration R_g of the AICD depends on the position of the first soliton as it propagates towards Pro(669) while the second soliton remains fixed by Pro(685). The shape of the protein changes as the soliton propagates, and, in particular, the α-helix becomes converted into a β-strand by the soliton. The final (θ_i, γ_i) soliton profile is displayed in Figure 9. In Figure 10, the three-dimensional shapes for both the initial PDB conformation and the conformation in which the first soliton has propagated to site 669 are compared.

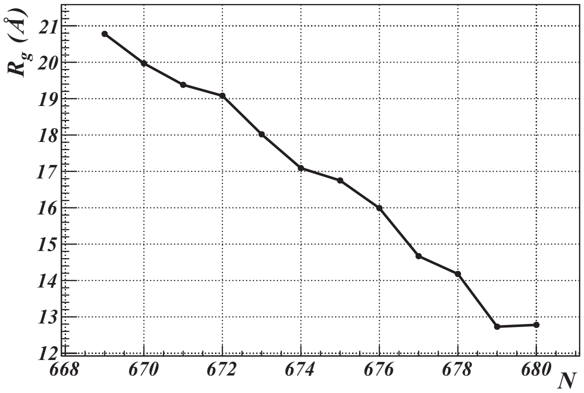

FIG. 8. The evolution of the radius of gyration for AICD in 3DXC (chain B), during the propagation of the first soliton from site 680 towards the proline at site 669; see also Figure 6(b).

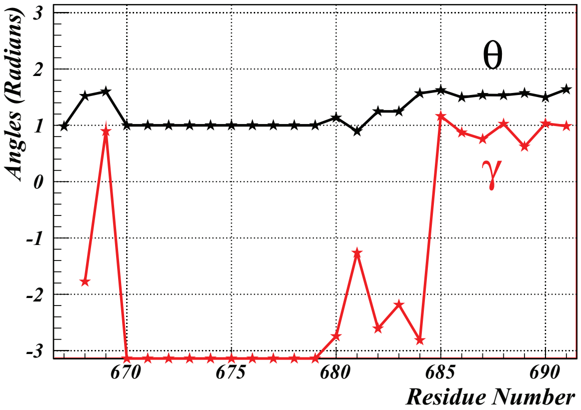

FIG. 9. The (θ_i, γ_i) profile of the 3DXC (chain B) after propagation of the first soliton onto the proline at site 669. See Figure 6 for the initial spectrum. The black lines and symbols correspond to the virtual-bond angles θ and the red lines and symbols correspond to the virtual-bond dihedral angles γ, respectively.

It is found that R_g increases monotonically as the first soliton propagates. When it reaches the position where it becomes locked into Pro(669), the R_g value of the AICD has relocated itself from the $R_g^{(2)}$ cluster to the $R_g^{(3)}$ cluster. Since there are no other known $\nu \approx 1$ clusters in the PDB, it is proposed that these are the only two possible trapped conformations that AICD can have in the phase coexistent complex with Fe65.

It is proposed that the presence of two natural but alternative soliton locations in AICD could give rise to a genetic switch, with potentially interesting biological consequences that deserve to be investigated. This should be particularly interesting, since AICD is a product and Fe65 is a participant in the proteolytic cleavage processing of amyloid precursor protein (APP) into Aβ42 that relates to Alzheimer's disease.[30,34–38] It is interesting how the biological function of the AICD/Fe65 complex could be influenced by the soliton position, when the dimer translocates to the nucleus and participates in gene transcription. It is proposed that the two possible AICD conformations corresponding to the two clusters $R_g^{(2)}$ and $R_g^{(3)}$ might have quite different biological properties.

C. Collapse simulations of isolated AICD and Fe65

It is expected that, in isolation and under physiological conditions, the two possible $\nu \approx 1$ conformations of the AICD in 3DXC are unstable because they are not located in the cluster of $\nu \approx 1/3$ which corresponds to compact monomeric proteins in Figure 2. Indeed, no single-chain protein has been found in the PDB which is located above the $\nu = 3/5$ Flory line in Figure 2. On the other hand, the two solitons that are the modular building blocks of AICD in 3DXC, are both abundant among the collapsed $\nu \approx 1/3$ proteins. Consequently, there are good reasons to expect that, in isolation, the AICD chain becomes subject to a phase transition-like changeover that takes it into the collapsed $\nu \approx 1/3$ cluster. Because there are no monomers in the PDB with compactness index $\nu \approx 1/2$, it can also be expected that Fe65 is similarly unstable in isolation, and becomes subject to a transition to the $\nu \approx 1/3$ cluster.

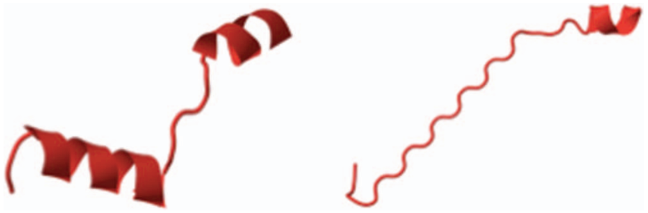

FIG. 10. The cartoon pictures of AICD. On the left, the PDB conformation corresponding to the (θ, γ) spectrum in Figure 6(b) and, on the right, the conformation corresponding to the spectrum in Figure 9.

On general grounds, it can be expected that AICD can collapse along various pathways. For example, there could be a process in which the two solitons first annihilate each other and a new soliton structure is formed to bring about the phase transition from $\nu \approx 1$ to $\nu \approx 1/3$. Alternatively, there could be formation of a new soliton pair near Pro(669), with the ensuing collapse to $\nu \approx 1/3$. Other alternatives also exist; for example, the first soliton could become locked by Pro(669) and the relatively long β-strand could then buckle to form a new soliton pair, separated, e.g., by a β-turn. It is proposed here that detailed experiments should be designed to determine the structural properties of AICD under physiologically relevant conditions.

In order to better understand what takes place both in AICD and in Fe65 when they become removed from each other, canonical Langevin-dynamics simulations of *isolated* AICD and Fe65, respectively, have been performed using the UNRES force field [Eq. (32)], starting from the conformations that AICD and Fe65 have in the 3DXC (heterodimer) experimental structure. The plots of the radius of gyration [computed from Eq. (2)] vs. simulation time are shown in Figure 11 for both AICD and Fe65.

It should be noted that, in a living cell, such a separation between AICD and Fe65 may take place, for example, when the dimer becomes translocated to the nucleus. Because of a relation between the AICD/Fe65 complex and the amyloid precursor protein and the ensuing Alzheimer-disease-causing Aβ42, the separation of the two subchains from each other could somehow relate to Alzheimer's disease which makes the present UNRES investigation particularly interesting.

For both separate AICD and Fe65, a rapid cross-over transition in the value of the radius of gyration, followed by a stabilization in the collapsed form is found. This transition is suggestive of a phase transition, from a non-compact form that is present in the heterodimer to a globular compact form.

For AICD, the mean value of the radius of gyration converges to

$$R_g(AICD) \approx 8.375\ (\text{Å}) \tag{41}$$

and for Fe65

$$R_g(Fe65) \approx 13.33\ (\text{Å}). \tag{42}$$

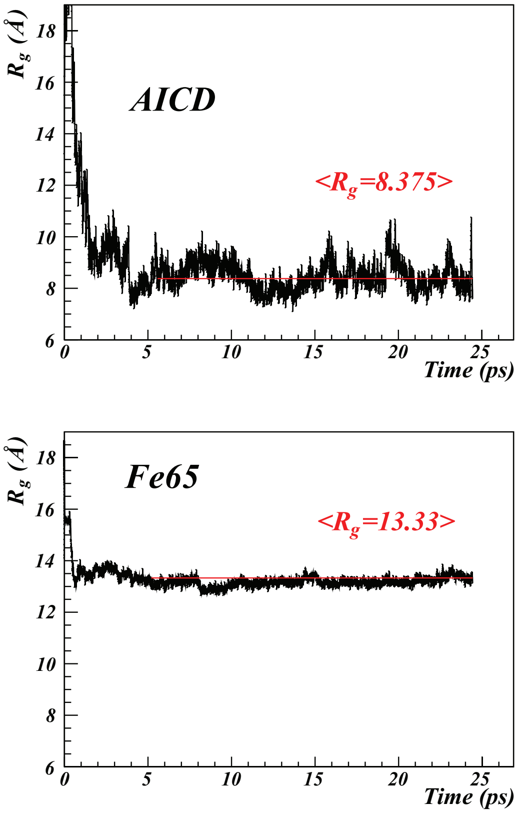

FIG. 11. Plots of the time evolution of the radius of gyration of isolated AICD (top) and isolated Fe65 (bottom) during Langevin molecular dynamics simulation with UNRES.

These are both very close to the collapsed $\nu = 1/3$ values of the mostly-α cluster [Eq. (37)] that are

$$R_g \approx \begin{cases} 8.06\,(\text{Å}) & (AICD), \\ 13.94\,(\text{Å}) & (Fe65). \end{cases} \tag{43}$$

It should be noted that, as stated in the discussion following Eq. (37), the difference between the clusters for mostly-α and other types of collapsed proteins is minuscule.

The remarkable observation in our simulations is that, for both AICD and Fe65, the final values of the radius of gyration fluctuate relatively strongly (Fig. 11). It is proposed here that this reflects the experimental observation that both proteins are unstructured,[3] when they are in isolation and under *in vivo* conditions.

IV. SUMMARY

In summary, by analyzing PDB structures, it has been found that physiologically relevant protein oligomers can display a behavior reminiscent of phase coexistence, in which the various subchains are in different phases. In particular, a family of dimers, for which one of the two subchains is in the same phase with a linear rod-like structure while the other is in the phase of a fully flexible chain, has been found. However, despite the apparent phase coexistence, both chains display the soliton structure that is characteristic of a single chain protein that is in the biologically active collapsed phase. It appears that the mutual interaction between the chains is so strong that it can overcome the effects of a poor solvent.

It has been proposed here that, if the two subchains are detached, they collapse. To confirm this, as an example, a detailed analysis of the potentially Alzheimer-disease related AICD/Fe65 complex has been carried out. It has been found that the AICD subchain can be described in terms of a two-soliton state at a very high precision, that exceeds the accuracy of experimental data as characterized by the B-factors. The soliton propagation along an α-helix of AICD has also been inspected, and it is proposed that this could give rise to a genetic-switch mechanism with potentially interesting biological consequences. Simulations using the UNRES energy function have been performed, and have confirmed that, in isolation, both AICD and Fe65 are indeed unstable and amenable to a transition, akin to a phase collapse, into a folded conformation. From a consideration of the results presented here, it is proposed that the analysis of protein complexes that display phase coexistence should be an interesting and biologically relevant challenge for future theoretical and experimental investigations.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health (Grant No. GM-14312), the National Science Foundation (Grant No. MCB10-19767), the Polish Ministry of Science and Higher Education (Grant No. DS 8372-4-0138-12), and a CNRS PEPS collaboration grant. Computational resources were provided by (a) Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-06CH11357, (b) the National Science Foundation (http://www.nics.tennessee.edu/), and by the National Science Foundation through TeraGrid resources provided by the Pittsburgh Supercomputing Center (c) the Informatics Center of the Metropolitan Academic Network (IC MAN) in Gdańsk, (d) our 624-processor Beowulf cluster at Baker Laboratory of Chemistry, Cornell University, and (e) our 184-processor Beowulf cluster at the Faculty of Chemistry, University of Gdańsk.

[1] A. R. Crofts, Annu. Rev. Physiol. **66**, 689 (2004).
[2] T. Althoff, D. J. Mills, J.-L. Popot, and W. Kuehlbrandt, EMBO J. **30**, 4652 (2011).
[3] H. J. Dyson and P. E. Wright, Nat. Rev. Mol. Cell Biol. **6**, 197 (2005).
[4] V. N. Uversky, Chem. Soc. Rev. **40**, 1623 (2011).
[5] C. L. Maters, G. Simms, N. A. Weinman, G. Multhaup, B. L. McDonald, and K. Beyreuther, Proc. Natl. Acad. Sci. U.S.A. **82**, 4245 (1985).
[6] D. Selkoe, Neuron **6**, 487 (1991).
[7] A. L. Garner, Curr. Topics Med. Chem. **11**, 258 (2011).
[8] S. Jones and J. M. Thornton, Prog. Biophys. Mol. Biol. **63**, 31 (1995).
[9] S. Jones and J. M. Thornton, Proc. Natl. Acad. Sci. U.S.A. **93**, 13 (1996).
[10] H. A. Scheraga, Biophys. Chem. **112**, 117 (2004).
[11] A. Marintchev, D. Frueh, and G. Wagner, Methods Enzymol. **130**, 283 (2007).
[12] Y. Chen, J. Johnson, P. McDonald, B. Wu, and J. D. Mueller, Methods Enzymol. **472**, 345 (2010).
[13] A.-C. Gavin, K. Maeda, and S. Kuehner, Curr. Opin. Struct. Biol. **22**, 42 (2011).
[14] M. Jessulat, S. Pitre, Y. Gui, M. Hooshyar, K. Omidi, B. Samanfar, L. H. Tan, M. Alamgir, J. Green, F. Dehne, and A. Golshani, Expert Opinion on Drug Discovery **6**, 921 (2011).

[15] G. R. Smith and M. J. E. Sternberg, Curr. Opin. Struct. Biol. **12**, 28 (2002).

[16] J. Janin, Mol. Biosys. **6**, 2351 (2010).

[17] A. Rojas, A. Liwo, D. Browne, and H. A. Scheraga, J. Mol. Biol. **537**, 404 (2010).

[18] Y. He, A. Liwo, H. Weinstein, and H. A. Scheraga, J. Mol. Biol. **405**, 298 (2011).

[19] L. Lo Conte, C. Chothia, and J. Janin, J. Mol. Biol. **285**, 2177 (1999).

[20] F. B. Sheinermann, R. Norel, and B. Honig, Curr. Opin. Struct. Biol. **10**, 153 (2000).

[21] P. L. Privalov and N. N. J. Khechinashvili, J. Mol. Biol. **86**, 665 (1974).

[22] M. L. Huggins, J. Chem. Phys. **9**, 440 (1941).

[23] P. J. Flory, J. Chem. Phys. **9**, 660 (1941).

[24] P. G. De Gennes, *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, 1979).

[25] L. Schäfer, *Excluded Volume Effects in Polymer Solutions, as Explained by the Renormalization Group* (Springer-Verlag, Berlin, 1999).

[26] C. B. Anfinsen, Science **181**, 223 (1973).

[27] F. S. Bates, Science **251**, 898 (1991).

[28] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).

[29] J. Hardy and D. J. Selkoe, Õs Disease: Science **297**, 353 (2002).

[30] T. Kristin, K. T. Jacobsen, and Æ. K. Iverfeldt, Cell. Mol. Life Sci. **66**, 2299 (2009).

[31] H. Zheng and E. H. Koo, Molec. Neurodeg. **1**, 1 (2006).

[32] N. K. Robakis, Neurobiol. Aging **32**, 372 (2011).

[33] S. B. Roberts, J. A. Ripellino, K. M. Ingalls, N. K. Robakis, and K. M. Felsenstein, J. Biol. Chem. **269**, 3111 (1994).

[34] T. Müller, H. E. Meyer, R. Egensperger, and K. Marcus, Prog. Neurobiol. **85**, 393 (2008).

[35] S. L. Sabo, L. M. Lanier, A. F. Ikin, O. Khorkova, S. Sahasrabudhe, P. Greengard, and J. D. Buxbaum, J. Biol. Chem. **274**, 7952 (1999).

[36] D. M. McLoughlin, C. J. Christopher, and C. C. J. Miller, J. Neurosci. Res. **86**, 744 (2008).

[37] T. Mueller, C. Loosse, A. Schroetter, A. Schnabel, S. Helling, R. Egensperger, and K. Marcus, Curr. Alz. Res. **8**, 573 (2011).

[38] J. Radzimanowski, B. Simon, M. Sattler, K. Beyreuther, I. Sinning, and K. Wild, EMBO Rep. **9**, 1134 (2008).

[39] A. Liwo, C. Czaplewski, J. Pillardy, and H. A. Scheraga, J. Chem. Phys. **115**, 2323 (2001).

[40] A. Liwo, M. Khalili, C. Czaplewski, S. Kalinowski, S. Ołdziej, K. Wachucik, and H. A. Scheraga, J. Phys. Chem. B **111**, 260 (2007).

[41] A. Liwo, C. Czaplewski, S. Ołdziej, A. V. Rojas, R. Kaźmierkiewicz, M. Makowski, R. K. Murarka, and H. A. Scheraga, in *Coarse-Graining of Condensed Phase and Biomolecular Systems*, edited by G. Voth (CRC, 2008).

[42] A. Liwo, S. Oldziej, C. Czaplewski, D. S. Kleinerman, P. Blood, and H. A. Scheraga, J. Chem. Theory Comput. **6**, 890 (2010).

[43] M. Khalili, A. Liwo, F. Rakowski, P. Grochowski, and H. A. Scheraga, J. Phys. Chem. B **109**, 13758 (2005).

[44] M. Khalili, A. Liwo, A. Jagielska, and H. A. Scheraga, J. Phys. Chem. B **109**, 13798 (2005).

[45] B. Widom, J. Chem. Phys. **43**, 3892 (1965).

[46] L. P. Kadanoff, Phys. **2**, 263 (1966).

[47] K. G. Wilson, Phys. Rev. B **4**, 3174 (1971).

[48] M. E. Fisher, Rev. Mod. Phys. **46**, 597 (1974).

[49] B. Li, N. Madras, and A. Sokal, J. Stat. Phys. **80**, 661 (1995).

[50] T. G. Dewey, J. Chem. Phys. **98**, 2250 (1993).

[51] L. Onsager, Ann. N.Y. Acad. Sci. **51**, 627 (1949).

[52] P. G. De Gennes, Phys. Lett. **38A**, 339 (1972).

[53] J. C. LeGuillou and J. Zinn-Justin, Phys. Rev. B **21**, 3976 (1980).

[54] C. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **90**, 6369 (1993).

[55] D. Thirumalai and D. K. Klimov, Curr. Opin. Struct. Biol. **9**, 197 (1999).

[56] R. I. Dima and D. Thirumalai, J. Phys. Chem. B **108**, 6564 (2004).

[57] L. Hong and J. Lei, J. Polym. Sci. B - Polym Phys. **47**, 207 (2009).

[58] A. Sieradzan, H. A. Scheraga, and A. Liwo, J. Chem. Theory Comput. **8**, 1334 (2012).

[59] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, J. Mol. Biol. **247**, 536 (1995).

[60] L. H. Greene, T. E. Lewis, S. Addou, A. Cuff, T. Dallman, M. Dibley, O. Redfern, F. Pearl, R. Nambudiry, A. Reid, I. Sillitoe, C. Yeats, J. M. Thornton, and C. A. Orengo, Nucl. Acids Res. **35**, D291 (2007).

[61] T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. U.S.A. **101**, 7960 (2004).

[62] Y. Zhang, I. A. Hubner, A. K. Arakaki, E. Shakhnovich, and J. Skolnick, Proc. Natl. Acad. Sci. U.S.A. **103**, 2605 (2006).

[63] L. Holm, C. Ouzounis, C. Sander, G. Tuparev, and G. Vriend, Protein Sci. **1**, 1691 (1992).

[64] A. Krokhotin, A. J. Niemi, and X. Peng, Phys. Rev. E **85**, 031906 (2012).

[65] N. Manton and P. Sutcliffe, *Topological Solitons* (Cambridge University Press, Cambridge, 2004).

[66] T. Dauxois and M. Peyrard, *Physics of Solitons* (Cambridge University Press, Cambridge, 2006).

[67] A. S. Davydov, J. Theor. Biol. **66**, 379 (1977).

[68] L. D. Faddeev and L. A. Takhtajan, *Hamiltonian Methods in the Theory of Solitons* (Springer Verlag, Berlin, 1987).

[69] A. C. Scott, Phys. Reps. **217**, 1 (1992).

[70] P. G. Kevrekidis, *The Discrete Nonlinear Schrödinger Equation: Mathematical Analysis, Numerical Computations and Physical Perspectives* (Springer-Verlag, Berlin, 2009).

[71] N. Molkenthin, S. Hu, and A. J. Niemi, Phys. Rev. Lett. **106**, 078102 (2011).

[72] A. J. Niemi, Phys. Rev. D **67**, 106004 (2003).

[73] M. Chernodub, S. Hu, and A. J. Niemi, Phys. Rev. E **82**, 011916 (2010).

[74] U. Danielsson, M. Lundgren, and A. J. Niemi, Phys. Rev. E **82**, 021910 (2010).

[75] M. Chernodub, M. Lundgren, and A. J. Niemi, Phys. Rev. E **83**, 011126 (2011).

[76] S. Hu, A. Krokhotin, A. J. Niemi, and X. Peng, Phys. Rev. E **83**, 041907 (2011).

[77] S. Hu, M. Lundgren, and A. J. Niemi, Phys. Rev. E **83**, 061908 (2011).

[78] M. Nanias, C. Czaplewski, and H. A. Scheraga, J. Chem. Theor. Comput. **2**, 513 (2006).

[79] C. Czaplewski, S. Kalinowski, A. Liwo, and H. A. Scheraga, J. Chem. Theor. Comput. **5**, 627 (2009).

[80] U. H. E. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).

[81] V. S. Pandé, I. Baker, J. Chapman, S. Elmer, S. Khaliq, S. M. Larson, Y. M. Rhee, M. R. Shirts, C. D. Snow, E. J. Sorin, and B. Zagrovic, Biopolymers **68**, 91 (2003).

[82] This is the universality class of one-dimensional rod-like structures. However, it should be noted that this observation does not imply that the structures are actually rods but means that a protein molecule is much less compact than the same protein when squeezed into a ball. Although the three principal axes of the moment of inertia or the anisotropy parameter[84] characterize the shape of a protein better, R_g is a parameter that can be used to judge this behavior: if the system is viewed as an ellipsoid with the principal dimensions a, b, and c, then $R_g = \sqrt{a^2 + b^2 + c^2}$. Thus, if one of the axes is much longer than the other ones, it will make a large contribution to R_g. For rod-like chains R_g has the strongest dependence on chain length, as follows from Eqs. (3) and (4). From Eqs. (3) and (4) it also follows that polymer (protein) chains can be classified based on the dependence of the radius of gyration on chain length; hence we find that the plot of R_g vs. chain length contains several line-like clusters (Figure 2).

[83] This bears a similarity to the decreasing compactness of folded polypeptides after binding to a mica surface, found in molecular dynamics simulations.[84]

[84] A. Starzyk and P. Cieplak, J. Chem. Phys. **135**, 235103 (2011).

[85] J. L. Smith, W. J. Freebern, I. Collins, A. De Siervi, I. Montano, I. C. M. Haggerty, M. C. McNutt, W. G. Butscher, I. Dzekunova, D. W. Petersen, E. Kawasaki, J. L. Merchant, and K. Gardner, Proc. Natl Acad. Sci. U.S.A. **101**, 11554 (2004).