TOPICAL REVIEW — Modeling and simulations for the structures and functions of proteins and nucleic acids

Application of topological soliton in modeling protein folding: Recent progress and perspective

Xu-Biao Peng(彭绪彪)[1,†], Jiao-Jiao Liu(刘娇娇)[1], Jin Dai(戴劲)[1,2], Antti J Niemi[1,2,‡], and Jian-Feng He(何建锋)[1,§]

[1] School of Physics, Beijing Institute of Technology, Beijing 100081, China
[2] Nordita, Stockholm University, Roslagstullsbacken 23, SE-106 91 Stockholm, Sweden

(Received 30 June 2020; revised manuscript received 8 August 2020; accepted manuscript online 13 August 2020)

Proteins are important biological molecules whose structures are closely related to their specific functions. Understanding how the protein folds under physical principles, known as the protein folding problem, is one of the main tasks in modern biophysics. Coarse-grained methods play an increasingly important role in the simulation of protein folding, especially for large proteins. In recent years, we proposed a novel coarse-grained method derived from the topological soliton model, in terms of the backbone C_α chain. In this review, we will first systematically address the theoretical method of topological soliton. Then some successful applications will be displayed, including the thermodynamics simulation of protein folding, the property analysis of dynamic conformations, and the multi-scale simulation scheme. Finally, we will give a perspective on the development and application of topological soliton.

Keywords: protein folding, coarse-grained method, Landau free energy function, topological soliton

PACS: 87.15.Cc, 87.14.E–, 87.15.A–, 87.15.hm **DOI:** 10.1088/1674-1056/abaed9

1. Introduction

Proteins are important functional molecules that participate in most of metabolic processes in living organisms. It is well known that the biological function of a protein depends critically on its three-dimensional native structure that is encoded in its amino acid sequence. This leads us to think about the protein folding problem: how a protein folds into its biologically active structure under physical principles.[1,2] Failing to fold gives rise to loss of function or dysfunction for a specific protein, which may be harmful or dangerous to the living organism. A broad range of human diseases, such as Alzheimer's, Parkinson's, type-II diabetes, and some types of cancers, arise from protein misfolding.[3–5] At the same time, the increasing spread of antibiotic-resistant pathogenic bacteria renders the lose efficacy of current antibiotics at a rapid rate and the lack of new antibacterial drugs.[6–8] And no effective treatments are found to prevent or control virulent viral infections due to the mutability of viral interactions with host cells, such as HIV, Ebola, SARS, and MERS.[9,10] The study of protein folding not only provides insights into the biological mechanisms of proteins but also is highly beneficial to the discovery of new antibacterial and antiviral drugs at the molecular level.

In recent decades, the protein folding problem has attracted considerable attention of researchers.[2] Some successful algorithms/methods have been devised to predict a protein's native structure from its amino acid sequence with increasing accuracy.[11–14] An excellent example is Alpha-Fold that achieves the impressive results in predicting the protein structure by deep learning.[14] Advanced experimental methods, such as optical tweezers and magnetic tweezers, have been developed to survey the kinetic and thermodynamic features at different stages of protein folding.[15–18] Computer-aided computations have distinct advantages in completely depicting the folding process and exploring the key information about the free energy landscape, the folding pathway, the folding rate and the folding mechanism, specifically the intermediate states with short lifetimes. All-atom molecular dynamics (MD) is popular in the simulation of protein folding, however its huge computational cost and limited conformational space sampling are still intractable issues. The success of all-atom MD simulation is limited to the fast folding of small simple proteins, even with most powerful computing technologies.[19,20] Enhanced sampling techniques, such as the replica exchange MD, accelerated MD, umbrella sampling, and meta-dynamics, have been proposed to increase the sampling efficiency in the conformational space.[21–24] On the other hand, some coarse-grained models with different degrees of simplification like UNRES and Gō model have been established to quicken the simulation of protein folding.[25–29] In recent years, coarse-grained methods are playing an increasingly important role in the folding study of large proteins.[30–32]

We develop an original coarse-grained method from the topological soliton for the protein folding research.[33–36]

†E-mail: xubiaopeng@bit.edu.cn
‡E-mail: Antti.Niemi@physics.uu.se
§Corresponding author. E-mail: hjf@bit.edu.cn

From the perspective of theoretical physics, a protein molecule is one of string-like objects. We define the backbone C_α chain of a protein that is regarded as a discrete string. The geometry of backbone C_α chain is analyzed by building local Frenet frames. The Landau free energy function is constructed to model the backbone C_α chain with the bond and torsion angles as variables, on account of its invariance of frame rotation. The discrete nonlinear Schrödinger equation derived from the Landau free energy function supports the dark soliton solution that is used to describe the backbone C_α chain. In recent years, we systematically proposed the topological soliton-based method to the protein folding problem. Compared to other reduced models, the soliton model is an effective theory with a simple Landau free energy and hence gains great advantages in the simulation performance. In this paper, we will first give a comprehensive review on the theoretical method of topological soliton. Then some examples of successful applications will be presented, including the folding simulations for both ordered and disordered proteins, the multi-scale simulations by combining with the all-atom MD, the characteristic analysis of dynamic conformations.[35–38] In the last section, we will bring forward a perspective on the theory and possible applications of topological soliton in protein researches.

2. Theoretical method of topological soliton

2.1. The backbone C_α chain

We simplify a protein molecule to a backbone C_α chain which physically derives from the piecewise linear discrete string.[33] Let $\boldsymbol{r}_i$ with $i = 1,2,\ldots,N$ denote the coordinates of C_α atoms. As shown in Fig. 1, we introduce the unit tangent vector $\boldsymbol{t}_i$, binormal vector $\boldsymbol{b}_i$ and normal vector $\boldsymbol{n}_i$ at each C_α atom

$$\boldsymbol{t}_i = \frac{\boldsymbol{r}_{i+1} - \boldsymbol{r}_i}{|\boldsymbol{r}_{i+1} - \boldsymbol{r}_i|},$$
$$\boldsymbol{b}_i = \frac{\boldsymbol{t}_{i-1} \times \boldsymbol{t}_i}{|\boldsymbol{t}_{i-1} \times \boldsymbol{t}_i|}, \quad \boldsymbol{n}_i = \boldsymbol{b}_i \times \boldsymbol{t}_i. \tag{1}$$

The right-handed orthogonal triplet ($\boldsymbol{t}_i$, $\boldsymbol{n}_i$, $\boldsymbol{b}_i$) constructs the discrete Frenet frame at the i-th C_α atom of backbone chain. The bond and torsion angles formed by the continuous C_α atoms, shown in Fig. 2, are computed by the Frenet frame vectors

$$\kappa_i \equiv \kappa_{i+1,i} = \arccos(\boldsymbol{t}_{i+1} \cdot \boldsymbol{t}_i), \tag{2}$$
$$\tau_i \equiv \tau_{i+1,i} = \mathrm{sgn}((\boldsymbol{b}_i \times \boldsymbol{b}_{i+1}) \cdot \boldsymbol{t}_i) \arccos(\boldsymbol{b}_{i+1} \cdot \boldsymbol{b}_i). \tag{3}$$

From the coordinates of C_α atoms, all of the Frenet frames and bond and torsion angles can be easily obtained by Eqs. (1)–(3). Conversely, if all of bond and torsion angles are given, the Frenet frame at the $(i+1)$-th C_α atom can be educed from the one at the i-th C_α atom by the discrete Frenet equation

$$\begin{pmatrix} \boldsymbol{n}_{i+1} \\ \boldsymbol{b}_{i+1} \\ \boldsymbol{t}_{i+1} \end{pmatrix} = \begin{pmatrix} \cos\kappa\cos\tau & \cos\kappa\sin\tau & -\sin\kappa \\ -\sin\tau & \cos\tau & 0 \\ \sin\kappa\cos\tau & \sin\kappa\sin\tau & \cos\kappa \end{pmatrix}_{i+1,i} \times \begin{pmatrix} \boldsymbol{n}_i \\ \boldsymbol{b}_i \\ \boldsymbol{t}_i \end{pmatrix}. \tag{4}$$

After setting up the Frenet frames along the backbone C_α chain, the coordinates of C_α atoms $\boldsymbol{r}_k$ with $k = 1,2,\ldots,N$ are determined by

$$\boldsymbol{r}_k = \sum_{i=0}^{k-1} |\boldsymbol{r}_{i+1} - \boldsymbol{r}_i| \cdot \boldsymbol{t}_i = \sum_{i=0}^{k-1} \triangle \boldsymbol{r}_{i,i+1} \cdot \boldsymbol{t}_i. \tag{5}$$

Here $\boldsymbol{r}_0 = 0$, $\boldsymbol{t}_0$ points to the direction of positive z axis and $\boldsymbol{t}_1$ locates on the y–z plane. And the distance between two contiguous C_α atoms $\triangle \boldsymbol{r}_{i,i+1}$ takes the constant value of 3.80 Å.[39]

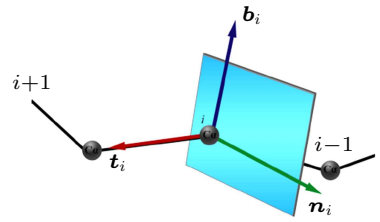

Fig. 1. The Frenet frame vectors ($\boldsymbol{t}_i$,$\boldsymbol{n}_i$, $\boldsymbol{b}_i$) at the i-th C_α atom.

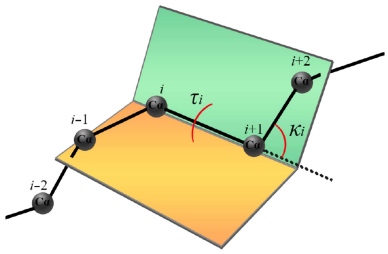

Fig. 2. The virtual bond and torsion angles (κ_i, τ_i) along the backbone C_α chain.

Since equation (5) does not explicitly contain the vectors $\boldsymbol{n}_i$ and $\boldsymbol{b}_i$, the backbone C_α chain remains unchanged under the frames of ($\boldsymbol{n}_i$, $\boldsymbol{b}_i$) are rotated by any angle θ_i around $\boldsymbol{t}_i$. That is, the backbone C_α chain has the invariance for the local SO(2) rotation of Frenet frame. Accordingly, the rotation acts on the bond and torsion angles as follows:

$$\kappa_i \rightarrow \mathrm{e}^{\mathrm{i}\theta_i}\kappa_i, \quad \tau_i \rightarrow \tau_i + \theta_{i-1} - \theta_i. \tag{6}$$

The transformation of (κ_i, τ_i) does not affect on the geometry of the backbone C_α chain. From the definition in Eqs. (2) and (3), the fundamental ranges of bond and torsion angles are $\kappa_i \in [0,\pi]$ and $\tau_i \in [-\pi,\pi)$, which can be regarded as the latitude and longitude angles of the unit sphere. If the two angles are extended into $[-\pi,\pi]\mathrm{mod}(2\pi)$, they can be considered as on torus.[35,39] The two-fold covering of the unit sphere is compensated by the $\mathbb{Z}_2$ gauge transformation

$$\kappa_l \rightarrow -\kappa_l, \quad \tau_i \rightarrow \tau_i - \pi, \ \forall\ l \geq i. \tag{7}$$

This is a special case of Eq. (6) with $\theta_l = \pi$ ($l \geq i+1$) and $\theta_l = 0$ ($l < i+1$). The $\mathbb{Z}_2$ symmetry motivates us to establish the energy function for the backbone C_α chain from the string theory.

2.2. Universal energy function and soliton

According to Anfinsen's dogma,[40] the protein in the native state locates at the minimum of Helmholtz free energy

$$E = U - TS, \tag{8}$$

where U is the internal energy, S is the entropy and T is the temperature T. The angles (κ_i, τ_i) are adopted as the structural variables of the free energy because of their intrinsical connection with the coordinates of C_α atoms. Considering the $\mathbb{Z}_2$ gauge symmetry shown in Eq. (7), the Landau free energy is naturally introduced to describe the backbone C_α chain of protein.[41,42] When the deformations of protein around the energy minimum keep slow and small, the free energy can be expanded as follows:

$$\begin{aligned} E \ = & -\sum_{i=1}^{N-1} 2\kappa_{i+1}\kappa_i + \sum_{i=1}^{N}\{2\kappa_i^2 + c(\kappa_i^2 - m^2)^2\} \\ & + \sum_{i=1}^{N}\{b\kappa_i^2\tau_i^2 + d\tau_i + e\tau_i^2 + q\kappa_i^2\tau_i\} + \cdots . \end{aligned} \tag{9}$$

Here c, m, b, d, e, and q depend on the physicochemical properties and the microstructure of protein. And the parameters are evaluated during training the minimal energy conformation of protein.[35,36]

From the extremum condition of energy function (9), an expression of the torsion angles is deduced in terms of the bond angles

$$\tau_i[\kappa_i] = -\frac{d + q\kappa_i^2}{2e + 2b\kappa_i^2}. \tag{10}$$

To take the derivative of energy function (9) with respect to κ_i and insert Eq. (10) into the ensuing equation, the following equation for the bond angles κ_i is obtained

$$\kappa_{i+1} - 2\kappa_i + \kappa_{i-1} = \frac{\mathrm{d}U[\kappa_i]}{\mathrm{d}\kappa_i^2}\kappa_i \quad (i = 1, 2, \ldots, N). \tag{11}$$

Here $\kappa_0 = \kappa_{N+1} = 0$, and $U[\kappa_i]$ is the effective potential

$$U[\kappa_i] = -\left(\frac{bd - eq}{2b}\right)^2 \frac{1}{e + b\kappa_i^2} - \frac{q^2}{4b}\kappa_i^2 - 2cm^2\kappa_i^2 + c\kappa_i^4. \tag{12}$$

In the case of folded protein, the values of the first two terms in Eq. (12) are much less than the latter two

$$U[\kappa_i] \approx -2cm^2\kappa_i^2 + c\kappa_i^4. \tag{13}$$

It is the familiar double-well potential with two separate minima $\kappa_i \approx \pm m$, in which the positive and negative values of κ_i are related by the $\mathbb{Z}_2$ transformation (7).

Equation (11) is a generalized discrete nonlinear Schrödinger (DNLS) equation that supports a soliton solution. Though the exact analytical solution is unknown, equation (11) can be solved numerically by the iterative equation

$$\kappa_i^{n+1} = \kappa_i^n - \varepsilon\{\kappa_i^n U'[\kappa_i^n] - (\kappa_{i+1}^n - 2\kappa_i^n + \kappa_{i-1}^n)\}. \tag{14}$$

Here $\{\kappa_i^n\}_{i\in N}$ denotes the n-th iteration of an initial conformation $\{\kappa_0^n\}_{i\in N}$ and ε is some sufficiently small but arbitrary constant. The soliton solution is independent of the value of ε which is often chosen to be 0.01 in the actual calculation. Furthermore, we have found a good approximative analytical solution

$$\kappa_i \approx \frac{m_1 e^{c_1(i-s)} - m_2 e^{-c_2(i-s)}}{e^{c_1(i-s)} + e^{-c_2(i-s)}}. \tag{15}$$

The m_1 and m_2 specify the asymptotic values of κ_i which are adjacent to the soliton. The parameter s defines the center of the soliton, and the c_1 and c_2 depict the shape of the variable region for the soliton. All of these parameters are determined by the profile of bond angles for the backbone C_α chain.

The statistics of crystal structures in PDB verifies that the secondary structures of proteins have the local distributions of bond and torsion angles.[39] For examples, $\kappa_i \approx \pi/2$ and $\tau_i \approx 1$ for the right-handed α-helix, $\kappa_i \approx 1$ and $\tau_i \approx \pi$ for the β-strand. Other regular structures, such as 3_{10} helix and left-handed helix, have also the constant values of (κ_i, τ_i). The values of bond and torsion angles are variable for the loops in proteins. Thus, the soliton solution of Eq. (11) can exactly model the profile of bond angles for any super-secondary structure of proteins. For instance, in the case of α-helix-loop-β-strand, the soliton describes the profile that the variable values of κ_i in the loop link the constant $\kappa_i \approx \pm\pi/2$ of the α-helix and the constant $\kappa_i \approx \pm 1$ of the β-strand. Since a protein is made up of a few structural modules like α-helix-loop-β-strand, the backbone C_α chain can be constructed by assembling all of the solitons that model these modules.

2.3. Thermal nonequilibrium dynamics of proteins

Protein folding is a kind of thermal non-equilibrium process. The folding near a thermal equilibrium proceeds in line with the Arrhenius law for simple proteins.[43] On the other hand, the Glauber algorithm manages the dynamics of a simple spin chain system coming close to the thermal equilibrium, which also obeys the Arrhenius law.[44,45] Since proteins can be regarded as spin chains, it is natural to simulate the protein folding in terms of Glauber algorithm. The transition probability $p\ (a \to b)$ between two states a and b is evaluated by

$$p\ (a \to b) = \frac{1}{1 + \exp\left(\dfrac{\Delta E_{ba}}{kT}\right)}. \tag{16}$$

Here $\Delta E_{ba} = E_b - E_a$ is the activation energy which is computed by the energy function (9) during the Monte Carlo (MC)

sampling. In addition, the steric constraint needs to be taken into account: the distance between two non-adjacent C_α atoms is always larger than 3.8 Å

$$|r_i - r_k| > 3.8 \text{ Å}, \quad |i-k| \geq 2. \tag{17}$$

This is considered as a necessary condition to accept a state during the MC simulation.

The parameters of Landau free energy function in Eq. (9) are not associated with the temperature explicitly. As a result, the temperature factor kT in Eq. (16) is not the physical temperature, but it can be related to the physical temperature by the renormalization. In the transition probability, the nearest neighbor coupling in the energy function is normalized to

$$-\frac{2}{kT}\sum_{i=1}^{N-1}\kappa_{i+1}\kappa_i. \tag{18}$$

Thus the temperature factor kT is related to the physical temperature t in such a way

$$\frac{2}{kT} \to \frac{J(t)}{k_B t}, \tag{19}$$

where k_B is the Boltzmann constant. The coefficient $J(t)$ complies with the renormalization group equation

$$t\frac{dJ}{dt} = \beta_J(J;c,m,b,d,e,q). \tag{20}$$

Assuming that the leading order of β_J only depends on $J(t)$, and the expansion of $J(t)$ at the low temperature limit is as follows:

$$J(t) \approx J_0 - J_1 t^\alpha + \cdots. \tag{21}$$

Equation (20) can be turned into

$$\alpha(J(t) - J_0) = \beta_J(J(t)). \tag{22}$$

From Eq. (19), the coefficient $J(t)$ can be expressed by the temperature factor kT and the physical temperature

$$J(t) \approx \frac{2k_B t}{kT}. \tag{23}$$

Combining Eq. (20) with Eq. (23), the renormalization group equation translates into

$$t\frac{d}{dt}\left(\frac{1}{kT}\right) = -\frac{1}{kT} + \frac{1}{2k_B t}\beta_J\left(\frac{2k_B t}{kT}\right). \tag{24}$$

By solving Eq. (24), we find the relation between the temperature factor and the physical temperature

$$kT \approx \frac{2}{J_0}k_B t \cdot \exp\left(\frac{J_1}{J_0}t^\alpha\right). \tag{25}$$

As long as the proper experimental temperatures are known, such as the transition temperature between the folded and completely unfolded states, the temperature factor kT can be converted into the corresponding physical temperature in the practical MC simulation.

3. Applications of the soliton model in protein folding

The soliton model and its related techniques on geometry analysis provide us a series of powerful tools in many fields of the protein structural dynamics, such as the thermodynamics simulation, the dynamical property analysis and the multi-scale simulations. In this section, we will review these applications.

3.1. On the thermodynamics simulations

The soliton model is an effective coarse-grained model based on the backbone C_α chain of the protein native structure. The parameters in the free energy function (9) and the soliton model of a protein can be obtained by fitting the ground state of the free energy into its native structure. Table 1 is a list of proteins that are modelled using topological soliton by far. The RMSD between the soliton model and the experimental structure indicates that the modelling accuracy is as high as sub-Angstrom. Once the parameters in the free energy function (9) are determined, the thermodynamical properties can be investigated by a MC sampling process as described in Subsection 2.3.

Table 1. The proteins that have been fitted using topological soliton model.

Proteins	PDB ID	Length of sequence	RMSD/Å
Villin headpiece 35	1YRF	29 aa	0.38
Myoglobin	1ABS	154 aa	0.78
HIV-1 reverse transcriptase protein	3DLK	18 aa	1.13
λ-repressor	1LMB	84 aa	0.51
Human islet amyloid polypeptide	2L86	37 aa	1.17
Myc proto-oncogene protein	1NKP	88 aa	0.98
Amyloid intra-cellular domain	3DXC	28 aa	0.46
Engrailed homeodomain	2JWT	61 aa	0.67
Parvalbumin-β	2PVB	57 aa	1.28

Depending on the properties of the conformational ensemble at low temperature, the protein can be classified into ordered protein and disordered protein. For an ordered protein, it is assumed to fold into an essentially unique native structure when cooled down to a low temperature. In other words, the native conformational ensemble is highly localized and the free energy landscape nearby the native state is funnel-like. In contrast, when an intrinsically disordered protein is cooled down to a low temperature, the conformational ensemble is structurally disparate but energetically comparable. For intrinsically disordered protein, the energy landscape at the native state is basin-like, *i.e.*, the different substates are separated with each other only by some relatively low energy barriers. The soliton model can deal with both kinds of proteins. In comparison with the Gō-like model which is also native structure based,[25] the number of the parameters in the soliton model is much fewer than the degree of freedom of the

protein structure. As a result, the soliton model can give predictions beyond the pre-known native structure, which makes it possible to give the landscape of the intrinsically disordered protein.

3.1.1. On the ordered protein

We take myoglobin as an example, to show the application of the soliton model in the thermodynamics simulation on the ordered protein.[46] Myoglobin is a kind of globular protein composed of 154 amino acids, and plays an important role in transporting oxygen. The native structure of myoglobin can be measured through the x-ray crystallographical method. Here we take the PDB entry 1ABS with resolution 1.5 Å for setting up the soliton model on myoglobin. The backbone C_α chain of the myoglobin is modelled by 10 solitons, whose parameters are shown in Table 1 in Ref. [46], giving the modelling accuracy about 0.8 Å. From the soliton model, the MC simulation with a heating-cooling cycle is performed to make sure the myoglobin can correctly unfold and refold, and the conformational ensembles are collected at different temperatures in the heating process. The above process is repeated thousands of times until the conformational ensembles get converged. For the detailed settings of the simulation, we refer to the literature.[46] In the end, we analyze the radius of gyration as a function of simulation temperature as shown in Fig. 3.

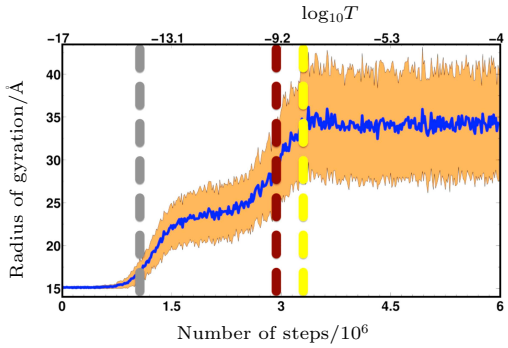

Fig. 3. The radius of gyration evolution with temperature increasing. The gray, red, yellow dashed lines are corresponding to the real temperatures of 25 °C, 75 °C, and 90 °C, respectively. Reproduced with permission from Ref. [46].

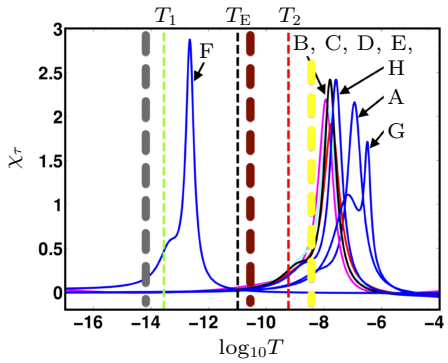

Fig. 4. The susceptibility of helical denucleation. Three transition temperatures are labeled as T_1, T_2, T_E, representing the two transition temperatures for the radius of gyration and for the energy, respectively. The colored thick dash lines are the same as in Fig. 3. Reproduced with permission from Ref. [46].

We can see that there is an metastable intermediate state between the native state and completely unfolded state in the simulation. The radius of gyration of intermediate state is 24 Å, which is very close to the experimental measurement (23.6 Å) on the molten globular state of the apomyoglobin. In addition, α-helical denucleation are also studied. For each helix in the native state, the susceptibility of denucleation is calculated as shown in Fig. 4, where we can see that the order of denucleation is $F \rightarrow BCDE \rightarrow A, H \rightarrow G$.

3.1.2. On the intrinsically disordered protein

For the intrinsically disordered protein, we take two proteins as examples: one is the human islet amyloid polypeptide (hIAPP), and the other is the amyloid precursor protein intra-cellular domain (AICD).[35,36] Both proteins are amyloid related and intrinsically disordered. In PDB, the native structures of hIAPP (PDB code: 2L86) and AICD complex with the C-terminal phosphotyrosine-binding (PTB) domain of Fe65 (PDB code: 3DXC) are measured by NMR and x-ray crystallographic methods, respectively. The superimpositions of the native structures and their corresponding soliton models are shown in Fig. 5, and their corresponding fitting accuracies are 1.17 Å and 0.59 Å, respectively. Then the heating-cooling MC simulations are performed, so that the intrinsically disordered protein can overcome the energy barriers and reach other possible conformations from the initial conformation. In the end, the conformational ensembles in the low temperature are generated and analyzed.

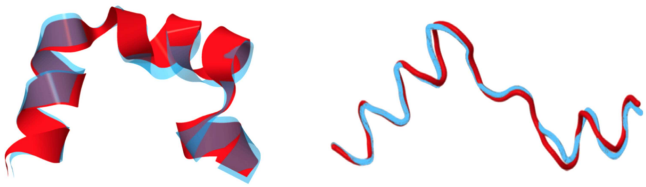

Fig. 5. The superimposition of the soliton model and PDB structures. Left panel is for 2L86 and right panel is for 3DXC. The light blue is from PDB structure and the red is from soliton model. Reproduced with permission from Refs. [35,36].

For hIAPP, after clustering analysis, six different conformational clusters are obtained as shown in Fig. 6. By further analyzing the conformational locality and stability of each cluster, we identify that the cluster 1 composed of two antiparallel helices is a good candidate to trigger the hIAPP aggregation and amyloidosis. This conclusion is consistent with many known observations in the literatures.

For AICD, the conformational landscape at different energy levels are systematically investigated nearby the energy minimum, as shown in Fig. 7. It is found that the native state of the isolated AICD can be a superposition of a family of degenerate conformations, which relates to the mobility of the soliton configuration as shown in Fig. 8. These results should provide a basis for analyzing the isolated AICD structures in the future NMR experiments.

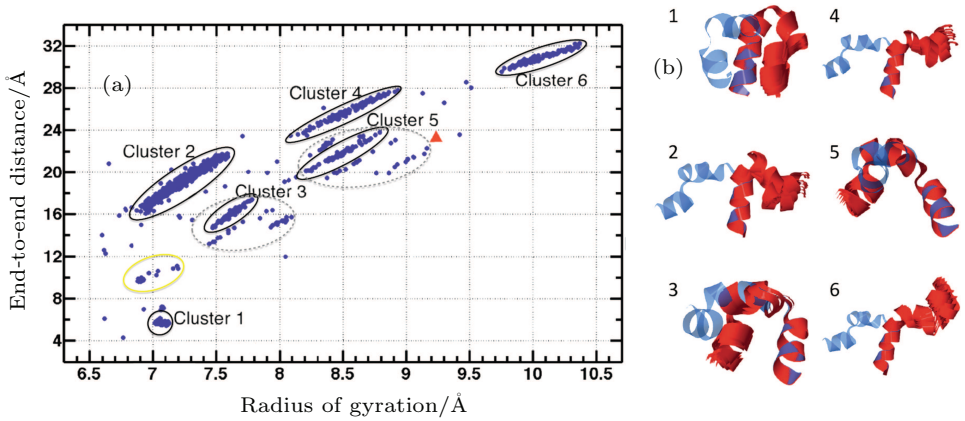

Fig. 6. The conformational clusters for 2L86 at low temperature. Panel (a) is the conformational landscape, and panel (b) is the representative structures. Reproduced with permission from Ref. [35].

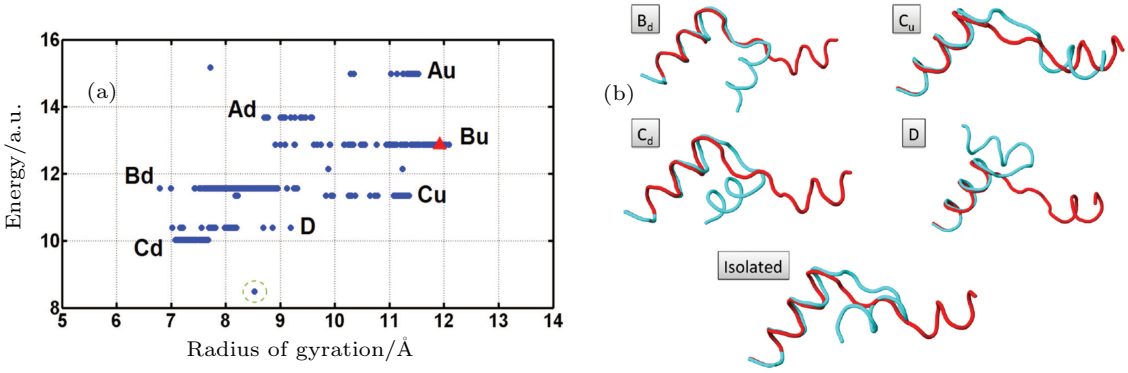

Fig. 7. The conformational clusters for 3DXC at low temperature. Panel (a) is the energy landscape of the conformational ensemble, where the red triangle denotes the initial structure in PDB. Panel (b) is the representative structures whose energies are lower than the initial structures. Reproduced with permission from Ref. [36].

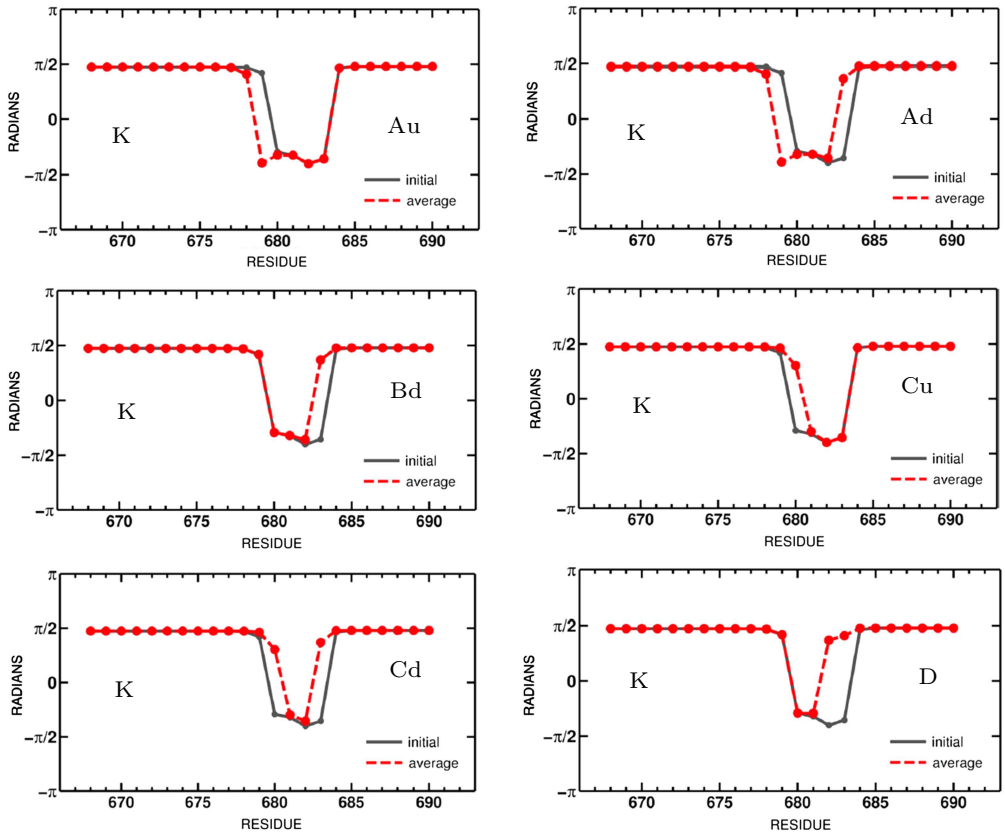

Fig. 8. The corresponding soliton mobility of the clusters in 3DXC. Reproduced with permission from Ref. [36].

3.2. On combination with all-atom MD simulations

Another important application of the soliton model is to combine with all-atom MD simulations. There are two ways to cooperate with all-atom MD simulations: (i) The multi-scale MD simulations, *i.e.*, to perform the MD simulation with specific initial conformation picked up from the thermodynamics simulation of the soliton model. (ii) To analyze the collective dynamical properties in the all-atom MD simulation from the viewpoint of soliton model.

3.2.1. The multi-scale MD simulations

For the application in the multi-scale MD simulations, we study the Myc proto-oncogene protein (PBD code: 1NKP) in the monomeric state,[38] which is intrinsically disordered but highly important in biomedicine. We first perform the thermodynamics simulation from soliton model, to generate the possible structural ensembles in the native state. For Myc protein, we get five clusters as shown in Fig. 9. For each cluster, the all-atom MD simulations at relatively low temperature (290 K) are performed with randomly selected conformation from each cluster as the initial structure. We find that the only simulation with the initial structure in cluster 1 can quickly converge to a stable ensemble. As an example, a comparison of the RMSD evolution on clusters 1, 4, and 5 (the cluster containing the native structure in PDB) is shown in Fig. 10, from which we can infer that the structures in cluster 1 is the most favorable conformation for the isolated Myc protein. Such information could be helpful for the future study on drug design.

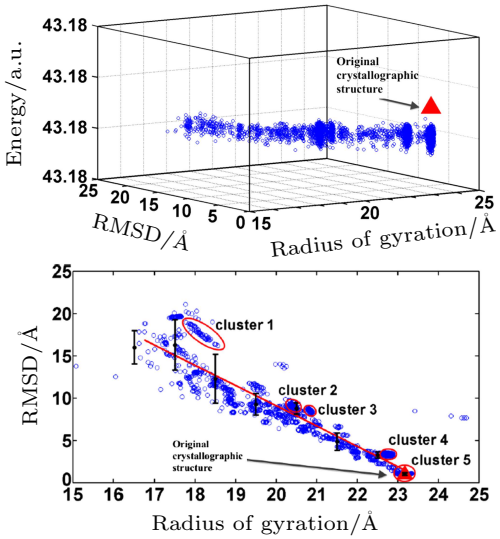

Fig. 9. The conformational clusters for 1NKP at low temperature. Top panel is the energy landscape of the conformational ensemble, where the red triangle denotes the initial structure in PDB. The bottom panel is the corresponding conformational landscape projected from top panel. Reproduced with permission from Ref. [38].

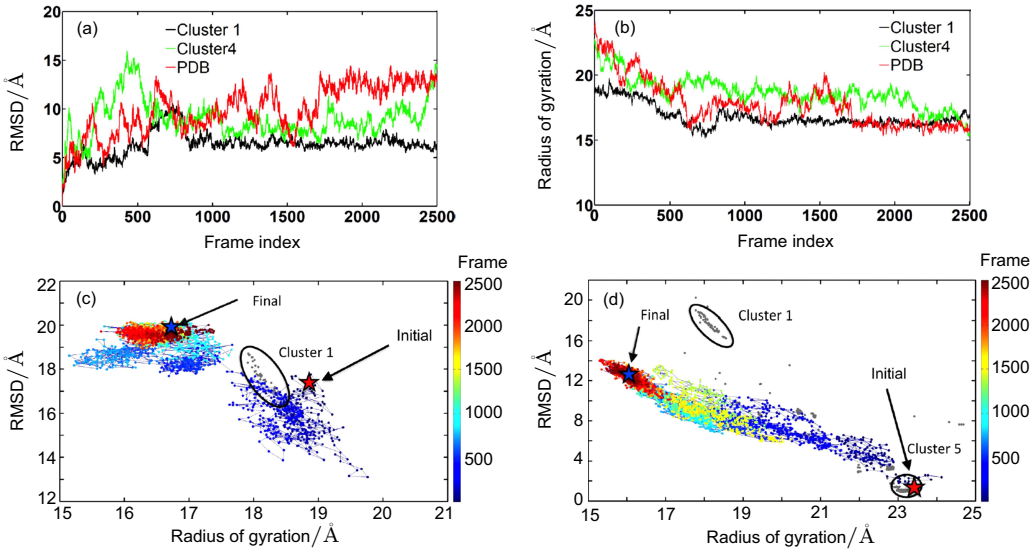

Fig. 10. The stability comparison among clusters in 1NKP, (a) comparison of the RMSD evolutions in MD simulations with initial conformations in clusters 1, 4, and 5 (denoted as PDB in the legend), (b) a comparison of the radius of gyration evolutions in MD simulations with initial conformations in clusters 1, 4, and 5 (denoted as PDB in the legend), (c) the conformational landscape evolution for MD simulations with initial conformation from cluster 1, (d) the conformational landscape evolution for MD simulations with initial conformation from cluster 5. Reproduced with permission from Ref. [38].

3.2.2. The dynamical properties analysis

For the application in the dynamical properties analysis, we study the folding properties of the α-helix subunit in HIV envelope glycoprotein gp41 (PDB code: 1AIK).[37] There are six α-helical subunits in the biological assembly. We isolate a subunit from it and run all-atom MD simulations in explicit solvent. In isolation, the subunit is unstable and starts folding and the trajectory shows the folding process in atomic resolution. For analyzing the folding properties, we define several collective variables based on the geometrical properties from the concept of soliton model, including the spin chain variable and folding index for backbone and sidechain, respectively. In Fig. 11, we can see that during the folding procedure the protein twists itself in different levels (the folding index fluctuates

between −2 and 4) and finally stabilizes itself at folding index being 1. The clear soliton motion is found in the N-terminal sidechain spin variable analysis as shown in Fig. 12. Hence, by analyzing such collective variables in the all-atom MD trajectory, the Bloch spin wave structure and its motion are observed in the folding process.

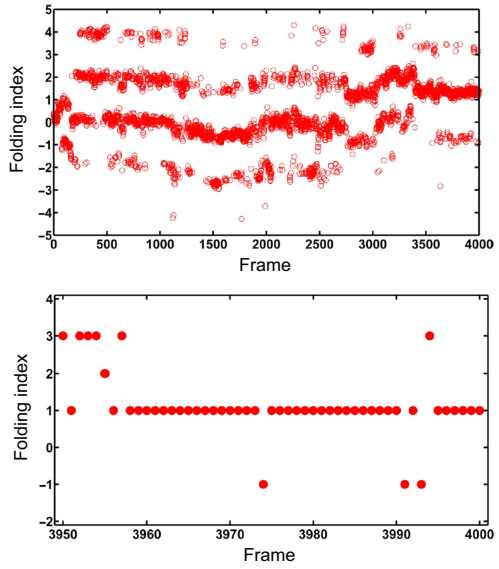

Fig. 11. The folding index evolution in MD simulation. The top panel is the folding index in the entire MD simulation process, and bottom panel is a zoom in of top panel in frame 3950–4000. Reproduced with permission from Ref. [37].

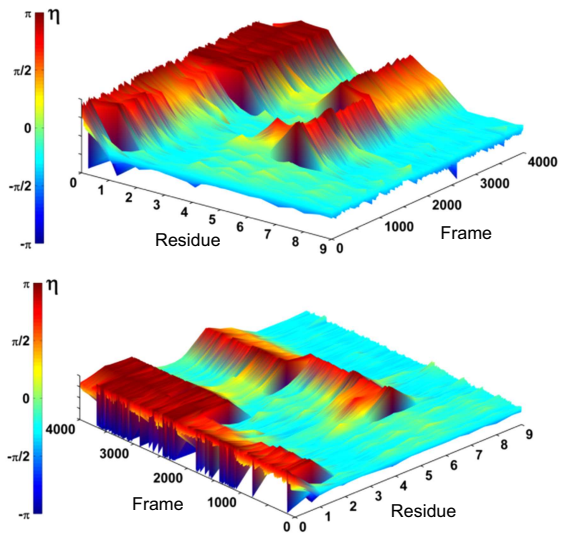

Fig. 12. The sidechain soliton motion in the N-terminal of the protein during the MD simulation. The two panels show the same data, but from different perspectives, for the first ten residues. Reproduced with permission from Ref. [37].

4. Conclusion and perspective

The protein folding is an outstanding conundrum in life science and always attracts plenty of attention. Researches have shown that the soliton model is universal and can cover most of the structures in PDB.[47] Hence, the concept of the soliton model can be widely used in protein research. In this paper, we reviewed the theory and possible applications of the soliton model and its related techniques in several different fields of the protein researches. As a new coarse-grained model derived from the gauge invariant principle of a string, the soliton model can not only reduce the complexity in protein thermodynamics simulations, but also can successfully cooperate with the traditional MD simulations for the multi-scale strategy and the collective dynamical property analysis. In addition, the discrete Frenet frame as well as its similar frames also provide a useful tool for the protein structure visualization, validation, refinement and reconstruction.

However, we note that more efforts are still needed for the further development of the topological soliton model. Firstly, the parameters in soliton are determined by a fitting/training process using the native structure currently, which requires that the native structure must be experimentally measured as accurately as possible. To break this limit, the parameters should be obtained from only the information of sequence and environment. Secondly, the current applications of the soliton model are limited to the monomeric protein. To extend the model to the multi-chain case, the protein–protein interaction such as the effective Lennard–Jones interaction between the chains need to be implemented, which is the future direction of the soliton model. Last but not least, the relation between the realistic force field and soliton model, *i.e.*, how the effective free energy of the topological soliton model is derived from the atomic interactions, needs to be further investigated.

References

[1] Dill K, Ozkan S B, Weikl T R, Chodera J D and Voelz V A 2007 *Curr. Opin. Struc. Biol.* **17** 342
[2] Dill K A and MacCallum J K 2012 *Science* **338** 1042
[3] Chiti F and Dobson C M 2006 *Ann. Rev. Biochem.* **75** 333
[4] Dobson C M 2003 *Nature* **426** 884
[5] Bullock A N and Fersht A R 2001 *Nat. Rev. Cancer* **1** 68
[6] Davies J 1996 *Nature* **383** 219
[7] Walsh C T 2000 *Nature* **406** 775
[8] Fischbach M A and Walsh C T 2009 *Science* **325** 1089
[9] Clercq E D and Li G 2016 *Clin. Microbiol. Rev.* **29** 695
[10] Domingo E 2016 *Virus as Populations* (London: Elsevier) p. 299
[11] Roy A, Kucukural A and Zhang Y 2010 *Nat. Protoc.* **5** 725
[12] Shi Y Z, Wu Y Y, Wang F H and Tan Z J 2014 *Chin. Phys. B* **23** 078701
[13] Wang J, Mao K, Zhao Y, Zeng C, Xiang J, Zhang Y and Xiao Y 2017 *Nucleic Acids Res.* **45** 6299
[14] Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S and Baker D 2020 *Proc. Natl. Acad. Sci. USA* **117** 1496
[15] Sun B, Wei K J, Zhang B, Zhang X H, Dou S X, Li M and Xi X G 2008 *EMBO J.* **27** 3279
[16] Chen H, Yuan G, Winardhi R S, Yao M, Popa I, Fernandez J M and Yan J 2015 *J. Am. Chem. Soc.* **137** 3540
[17] Hong L, Jain N, Cheng X, Bernal A, Tyagi M and Smith J C 2016 *Sci. Adv.* **2** e1600886
[18] Gou L, Jin T, Chen S, Li N, Hao D, Zhang S and Zhang L 2018 *Chin. Phys. B* **27** 028708
[19] Lindorff-Larsen K, Piana S, Dror R O and Shaw D E 2011 *Science* **334** 517
[20] Piana S, Klepeis J L and Shaw D E 2014 *Curr. Opin. Struct. Biol.* **24** 98
[21] Sugita Y, Kitao A and Okamoto Y 2000 *J. Chem. Phys.* **113** 6042
[22] Jing Z F and Sun H 2015 *J. Chem. Theory Comput.* **11** 2395
[23] Torrie G M and Valleau J P 1977 *J. Comput. Phys.* **23** 187
[24] Liao A and Parrinello M 2002 *Proc. Natl. Acad. Sci.* **99** 12562
[25] Gō N 1983 *Ann. Rev. Biophys. Bioeng.* **12** 183

[26] Khalili M, Liwo A, Jagielska A and Scheraha H 2005 *J. Phys. Chem. B* **109** 13798
[27] Zuo G H, Zhang J, Wang J and Wang W 2005 *Chin. Phys. Lett.* **22** 1809
[28] Zuo G H, Hu J and Fang H P 2007 *Chin. Phys. Lett.* **24** 2426
[29] Lu Y K, Zhou X and OuYang Z C 2017 *Chin. Phys. B* **26** 050202
[30] Su J G, Han X M and Zhao S X 2016 *J. Mol. Model.* **22** 1
[31] Li W, Wang W and Takada S 2014 *Proc. Natl. Acad. Sci. USA* **111** 10550
[32] Wang Y, Chu X, Suo Z, Wang E and Wang J 2012 *J. Am. Chem. Soc.* **134** 13755
[33] Niemi A J 2003 *Phys. Rev. D* **67** 106004
[34] Hu S, Jiang Y and Niemi A J 2013 *Phys. Rev. D* **87** 105011
[35] He J F, Dai J, Li J, Peng X B and Niemi A J 2015 *J. Chem. Phys.* **142** 045102
[36] Dai J, Niemi A J and He J F 2016 *J. Chem. Phys.* **145** 045103
[37] Dai J, Niemi A J, He J F, Sieradzan A and Ilieva N 2016 *Phys. Rev. E* **93** 032409
[38] Liu J J, Dai J, He J F, Niemi A J and Ilieva N 2017 *Phys. Rev. E* **95** 032406
[39] Peng X, Chenani A, Hu S, Zhou Y and Niemi A J 2014 *BMC Struct. Biol.* **14** 27
[40] Anfinsen C B and Scheraga H A 1975 *Adv. Protein Chem.* **29** 205
[41] Landau L D and Lifshitz E M 2013 *Statistical Physics* Vol. 5 (London: Elsevier) p. 429
[42] Chernodub M, Hu S and Niemi A J 2010 *Phys. Rev. E* **82** 011916
[43] Scalley M L and Baker D 1997 *Proc. Natl. Acad. Sci.* **94** 10636
[44] Glauber R J 1963 *J. Math. Phys.* **4** 294
[45] Bortz A B, Kalos M H and Lebowitz J L 1975 *J. Comput. Phys.* **17** 10
[46] Peng X, Sieradzan A K and Niemi A J 2016 *Phys. Rev. E* **94** 062405
[47] Krokhotin A, Niemi A J and Peng X 2012 *Phys. Rev. E* **85** 031906