PHYSICAL REVIEW E **82**, 011916 (2010)

Topological solitons and folded proteins

Maxim Chernodub,[1,2,*] Shuangwei Hu,[1,3,†] and Antti J. Niemi[1,3,‡]

[1] *Laboratoire de Mathematiques et Physique Theorique CNRS UMR 6083, Fédération Denis Poisson, Université de Tours, Parc de Grandmont, F37200 Tours, France*

[2] *Department of Mathematical Physics and Astronomy, Ghent University, Krijgslaan 281, 59, Gent B-9000, Belgium*

[3] *Department of Physics and Astronomy, Uppsala University, P.O. Box 803, S-75108 Uppsala, Sweden*

(Received 1 April 2010; published 21 July 2010)

We argue that protein loops can be described by topological domain-wall solitons that interpolate between ground states which are the α helices and β strands. We present an energy function that realizes loops as soliton solutions to its equation of motion, and apply these solitons to model a number of biologically active proteins including 1VII, 2RB8, and 3EBX (Protein Data Bank codes). In all the examples that we have considered we are able to numerically construct soliton solutions that reproduce secondary structural motifs such as α-helix-loop-α-helix and β-sheet-loop-β-sheet with an overall root-mean-square-distance accuracy of around 1.0 Å or less for the central α-carbons, i.e., close to the limits of current experimental accuracy.

DOI: 10.1103/PhysRevE.82.011916

PACS number(s): 87.15.Cc, 05.45.Yv

Solitons are ubiquitous and widely studied objects that can be materialized in a variety of practical and theoretical scenarios [1,2]. For example solitons can be deployed for data transmission in transoceanic cables, for conducting electricity in organic polymers [1], and they may also transport chemical energy in proteins [3]. Solitons explain the Meissner effect in superconductivity and dislocations in liquid crystals [1]. They also model hadronic particles, cosmic strings, and magnetic monopoles in high energy physics [2] and so on. The first soliton to be identified is the *Wave of Translation* that was observed by John Scott Russell in the Union Canal of Scotland. This wave can be accurately described by an exact soliton solution of the Korteweg-de Vries (KdV) equation [1]. At least in principle it can also be constructed in an atomary level simulation where one accounts for each and every water molecule in the Canal, together with all of their mutual interactions. However, in such a *Gedanken* simulation it would probably become a real challenge to unravel the collective excitations that combine into the *Wave of Translation* without any guidance from the known soliton solution of the KdV equation: Solitons can *not* be constructed simply by adding up small perturbations around some ground state. Instead, a (topological) soliton emerges when non-linear interactions combine elementary constituents into a localized collective excitation that is stable against small perturbations and cannot decay, unwrap or disentangle [1,2].

In this Communication we argue that topological solitons describe proteins in their native folded state [4,6]. We characterize a folded protein by the Cartesian coordinates $\mathbf{r}_i$ of its N central α carbons, with $i=1,\ldots,N$. For many biologically active proteins these coordinates can be downloaded from protein data bank (PDB) [7]. Alternatively, the protein can be described in terms of its bond and torsion angles that can be computed from the PDB data. For this we introduce the tangent vector $\mathbf{t}_i$ and the binormal vector $\mathbf{b}_i$

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

Together with the normal vector $\mathbf{n}_i=\mathbf{b}_i\times\mathbf{t}_i$ we then have three vectors that are subject to the discrete Frenet equation [8].

$$\begin{pmatrix}\mathbf{n}_{i+1}\\ \mathbf{b}_{i+1}\\ \mathbf{t}_{i+1}\end{pmatrix} = \exp\{-\kappa_i\cdot T^2\}\cdot\exp\{-\tau_i\cdot T^3\}\begin{pmatrix}\mathbf{n}_i\\ \mathbf{b}_i\\ \mathbf{t}_i\end{pmatrix}. \tag{2}$$

Here, T^2 and T^3 are two of the standard generators of three dimensional rotations, explicitly in terms of the permutation tensor we have $(T^i)^{jk}=\epsilon^{ijk}$. From Eqs. (1) and (2) we can compute the bond angles κ_i and the torsion angles τ_i using PDB data for $\mathbf{r}_i$. Alternatively, if we know these angles we can compute the coordinates $\mathbf{r}_i$. The common convention is to select the range of these angles so that κ_i is positive. In the continuum limit where Eq. (2) becomes the standard Frenet equation for a continuous curve, $\kappa_i\rightarrow\kappa(x)$ then corresponds to local curvature which is defined to be non-negative.

As a concrete example we now describe the 35 residue villin headpiece protein with PDB code 1VII that has been widely investigated, both theoretically and experimentally [4]. For example in the state-of-art simulation [5] succeeded in producing its fold for a short time with a root mean square distance (RMSD) accuracy of ~2–3 Å.

From the PDB data we compute the values of bond angles κ_i and torsion angles τ_i and the result is displayed in Fig. 1(a). when we use the (standard) convention that the discrete Frenet curvature κ is positive. In 1VII there are three α helices that are separated by two loops. When we use the PDB (NMR) convention for indexing the residues the first, longer, loop is located at sites 49–54 and the second, shorter, between 59–62.

We shall now show that Fig. 1(a) describes two soliton configurations, albeit in an encrypted form. In order to decrypt the data in Fig. 1(a) so that these solitons become unveiled we observe that the Eq. (2) has the following local $\mathbb{Z}_2$ gauge symmetry: At every site we can send

[*] chernodub@lmpt.univ-tours.fr

[†] shuangwei.hu@lmpt.univ-tours.fr

[‡] antti.niemi@physics.uu.se

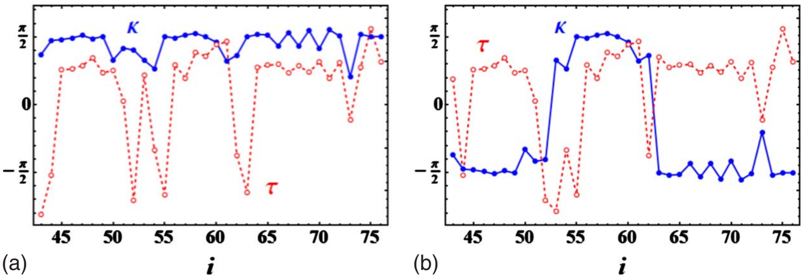

FIG. 1. (Color online) (a) (left): The bond and torsion angles of 1VII, computed with the standard convention that the discrete Frenet curvature κ is positive. (b) (right): The Z_2 gauge transformed bond and torsion angles.

$$\mathbb{Z}_2 : \begin{cases} \kappa_i \to \kappa_i \cdot \cos(\Delta_i) \\ \tau_i \to \tau_i + \Delta_{i-1} - \Delta_i \end{cases} \tag{3}$$

and when we choose at each site $\Delta_i = 0$ or $\Delta_i = \pi$ where $\Delta_i = \pi$ is the nontrivial element of the $\mathbb{Z}_2$ gauge group, the Cartesian coordinates $\mathbf{r}_i$ computed from the discrete Frenet equation remain intact.

The gauge transformation that we introduce is a remnant of the continuum convention to choose the curvature $\kappa(x)$ to be always non-negative. As a consequence it can only be defined in a piecewise manner, between the straight segments and the conjugacy points where the curvature vanishes and the Frenet frame cannot be introduced. For a continuum curve, it can then become an issue how to determine $\kappa(x)$ through these conjugacy points so that its first derivative along the curve remains continuous. Instead of the common convention of a piecewise defined and non-negative curvature, which often leads to a discontinuous first derivative at the conjugacy points, we here use a definition where we allow $\kappa(x)$ to change its sign over points/segments where it vanishes, in such a manner that its first derivative along the curve remains continuous. This introduces a discrete gauge structure that ensures the equivalence of the two alternative descriptions.

For a discrete curve the continuity is not really an issue, if we define the bond angle κ_i to be non-negative the vanishing of κ_i does not pose a similar kind of a problem as in the continuum. But it turns out that by demanding κ_i to be non-negative, we translate the presence of a conjugacy point into sign changes in the torsion angle τ_i between adjacent sites. This allows us to easily locate the potential presence of a loop in the raw PDB data, since a (continuum) topological domain wall necessarily involves the presence of a conjugacy point.

If we implement the $\mathbb{Z}_2$ gauge transformation in the data displayed in Fig. 1(a), at the points where τ_i changes its sign between adjacent points, we arrive at the apparently quite different Fig. 1(b). Unlike in Fig. 1(a), the profile of κ_i in Fig. 1(b) now clearly displays the hallmark profile of a topological soliton-(anti)soliton pair in a double-well potential: The two soliton profiles are located around the sites with indices 49–54 and 59–62 which are the locations of the two loops in 1VII. These profiles interpolate between the two "ground-state" values $\kappa_i \approx \pm\pi/2$ that pinpoint the locations of the α helices in 1VII. Moreover, the two downswings in the value of τ_i from the value $\tau_i \approx 1$ that mark the locations of the α helices, coincide with the locations of the two soliton profiles. The ensuing combined profile of κ_i and τ_i is qualitatively consistent with a double-well potential structure in the (κ,τ) plane that has the form displayed in Fig. 2: When we move from left to right in Fig. 1(b), we follow a trajectory in the (κ,τ) plane that starts by fluctuating around the potential energy minimum at $(\kappa,\tau) \approx (-\pi/2,1)$ in Fig. 2, corresponding to the first α helix. The trajectory then moves through the first loop a.k.a. soliton (red line) to the second potential energy minimum i.e., α helix at $(\kappa,\tau) \approx (+\pi/2,1)$ in Fig. 2, and finally back through the second loop a.k.a. soliton (blue line) to the first potential energy minimum at $(\kappa,\tau) = (-\pi/2,1)$.

We now describe a theoretical model introduced in [9,10] that reproduces the (κ,τ) profile in Fig. 1(b) as a combination of two soliton solutions to its equations of motion, with a very high accuracy for the central α carbons. The model is defined by the energy functional

$$E = \sum_{i=1}^{N-1} (\kappa_{i+1} - \kappa_i)^2 + \sum_{i=1}^{N} c \cdot (\kappa_i^2 - m^2)^2 + \sum_{i=1}^{N} \{ b\kappa_i^2\tau_i^2 + d\tau_i + e\tau_i^2 + q\kappa_i^2\tau_i \}. \tag{4}$$

Here N is the number of central α carbons and (c,m,b,d,e,q) are parameters. We refer to [9,10] for a detailed motivation of Eq. (4): The first sum describes nearest neighbor interactions along the protein. The second sum describes a local self-interaction of the bond angles. The third sum describes local interactions between bond and torsion

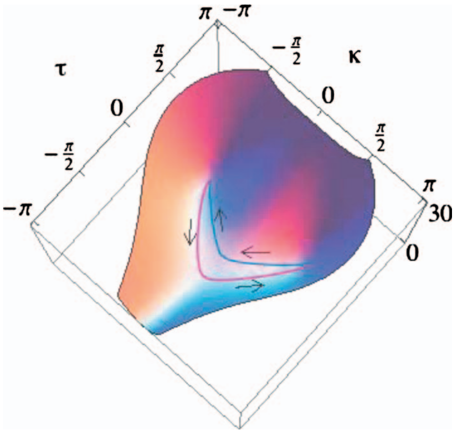

FIG. 2. (Color) The potential energy on (κ,τ) plane that corresponds qualitatively to the data in Fig. 1(b), the soliton between sites 49–54 corresponds to the red trajectory and the soliton between sites 59–62 to the blue trajectory.

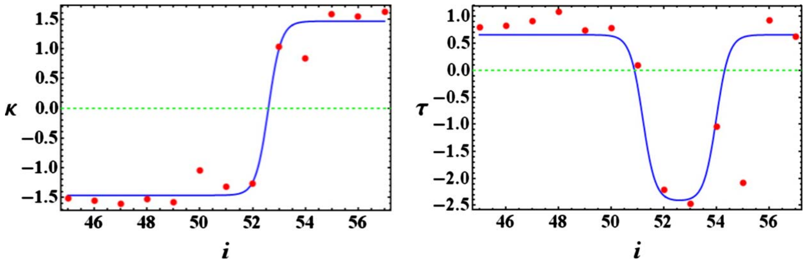

FIG. 3. (Color online) The PDB data for the first α-helix-loop-α-helix motif in 1VII, on the left κ_i and on the right τ_i, together with the least square approximations Eqs. (5) and (6) (solid blue line).

angles, its first term has an origin in a Higgs effect which is due to the potential term in the second sum. The second term in the third sum is the Chern-Simons term, it is responsible for the chirality of the protein chain. The third term is a Proca mass term and the last term can also be related to the Abelian Higgs model, and it is also chiral. As explained in [10] this energy functional is essentially unique, and in particular it can be related to a gauge invariant (supercurrent) version of the energy of 1+1 dimensional lattice Abelian Higgs model. In three space dimensions this model is also known as the Ginzburg-Landau model of conventional superconductivity [2].

We fully appreciate that the detailed fold of a given protein is determined by the specifics of its unique amino acid sequence. The interactions that contribute to the fold include hydrophobic, hydrophilic, long-range Coulomb, van der Waals, saturating hydrogen bonds and so forth interactions [11]. Consequently, *a priori* a given protein should not be approximated by a homopolymer model.

Note that in Eq. (4) there is no reference to the specifics of the interactions that are presumed to drive the folding process. The only explicit long-range force present in Eq. (4) is the nearest-neighbor interaction described by the first term. Moreover, as it stands Eq. (4) depends only on *six* site-independent, homogeneous parameters. There is no direct reference whatsoever to the underlying in general highly inhomogeneous amino acid structure of a protein. We argue that this becomes possible since Eq. (4) supports *solitons* that describe the common secondary structural motifs such as α-helix/β-strand-loop-α-helix/β-strand as solutions to its classical equations of motion. Furthermore, even though the actual numerical values of the parameters are certainly motif dependent and for long loops that constitute bound states of several solitons one might need to introduce more than six parameters, we expect there to be wide *universality* so that a given soliton with its relatively few parameters describes a general class of homologous motifs. Consequently only a relatively small set of parameters is needed to provide soliton templates for structure prediction. In fact, we propose that solitons are the mathematical manifestation of the experimental observation, that the number of different protein folds is surprisingly limited. The presence of solitons could then be the reason for the success of bioinformatics based homology modeling in predicting native folds [4].

In order to quantitatively disclose the soliton solution of Eq. (4) we start by observing that the first two sums in Eq. (4) can be interpreted as a discrete version of the energy of the 1+1 dimensional double well $\lambda\phi^4$ model that is known to support the topological kink soliton. In the continuum limit the kink soliton has the analytic form [1,2],

$$\kappa(x) = m \cdot \tanh(m\sqrt{c} \cdot [x - x_0]).$$

We can try to estimate the parameters m and c for each of the two solitons in the Fig. 1(b) by a least square fitting where we use this continuum soliton to approximate the exact soliton solution of the discrete equations of motion. We consider here explicitly only the first soliton of 1VII, located between (PDB index) sites 49–54. We assume that the discretized kink-soliton describes the profile of κ_i, and using the sites 46–56 we find the following least square fit

$$\kappa(x) \approx 1.4627 \cdot \tanh(2.0816[x - 52.597]). \tag{5}$$

In order to construct $\tau(x)$ we solve for its equation of motion in Eq. (4). Up to the parameters the dependence of τ_i on the kink soliton is then uniquely determined by the model, and the result is

$$\tau(x) \approx -2.4068 \cdot \frac{1 - 0.4689 \cdot \kappa^2(x)}{1 - 0.4619 \cdot \kappa^2(x)}. \tag{6}$$

In Fig. 3 we show how the data in Fig. 1(b) is described by the approximate soliton profile Eqs. (5) and (6). When we construct the ensuing discrete curve in the three dimensional ambient space by solving Eq. (2) with for κ_i and τ_i given by Eqs. (5) and (6), we reproduce the first loop of 1VII with a surprisingly good RMSD accuracy of ~1.4 Å for the PDB indices 46–56. We think that this is quite remarkable, in particular by taking into account the simplicity of our approximation: Our Ansatz depends on *only one single function*, the hyperbolic tangent, that is determined by Eq. (4). In addition, there are the six parameters in Eq. (4). But a minimum of six characteristic parameters are needed to describe any loop configuration, and each of these can be given a very definite interpretation. The parameters are as follows:

(1) The location of the soliton along the protein (in κ_i)

(2) The size of the soliton in number of sites

(3) The asymptotic value of κ_i away from the soliton

(4) The asymptotic value of τ_i away from the soliton

(5) The value of τ_i at the center of the soliton

(6) The relative position of κ_i and τ_i for the center of soliton

For both (3) and (4) there are two possible values, corresponding to α helix and β strand. For (6) we have found that the location of the center of the soliton is slightly different in the variables κ_i and τ_i.

We take the remarkable success of our construction Eqs. (5) and (6) to be a strong argument in support of universality in protein folding. The same set of six parameters should describe corresponding loops in *any* homologically related protein. Obviously this needs to be confirmed, and we are now in the process of constructing the explicit soliton profiles for several homologically related proteins in the PDB.

In order to construct a more accurate description of 1VII, we resort to a numerical construction of a soliton solution to the equations of motion if Eq. (4). We use simulated annealing that involves a Monte Carlo energy minimization of the energy functional

$$F = -\beta_1 \cdot \sum_{i=1}^{N} \left\{ \left(\frac{\partial E}{\partial \kappa_i} \right)^2 + \left(\frac{\partial E}{\partial \tau_i} \right)^2 \right\} - \beta_2 \cdot \sqrt{\frac{1}{N} \sum_{i=1}^{N} |\mathbf{r}_{PDB}(i) - \mathbf{r}_{soliton}(i)|^2}. \qquad (7)$$

with a simultaneous cooling of the two (inverse) temperatures β_1 and β_2. Here, the first sum vanishes when we have a solution to the classical difference equation of motion of Eq. (4), the cooling simulates a gradient flow toward the critical points i.e., classical solutions of Eq. (4). Since Eq. (4) can have several different critical points, we introduce the second term that computes the RMSD distance between the ith α carbon of the solution and the protein we wish to construct. The second term in Eq. (7) then acts like a chemical potential that selects the parameters in Eq. (4) so that we arrive at a soliton solution that corresponds to the actual, given protein fold.

We have numerically constructed the classical solutions of Eq. (4) that describe the secondary structural motifs in proteins with PDB codes 1VII, 2RB8 and 3EBX. The first one has three α helices separated by loops, while the second and third have β-strand-loop-β-strand motifs. Both cases can be described equally by Eq. (4), the only difference is that in the case of β strands the two minima of the (classical) potential in Eq. (4) are located at $(\kappa,\tau) \approx (\pm 1, \pi)$. In each of the proteins that we have studied we have routinely been able to reproduce the secondary structural motifs as classical soliton solutions to the equations of motion for Eq. (4) in terms of only six parameters and with an overall RMSD accuracy of less than 1.0 Å per motif which is essentially the experimental accuracy in x-ray crystallography and NMR; in our simulations the first sum in Eq. (7) decreases typically by around ten orders of magnitude indicating that the final configuration is a solution, essentially within numerical accuracy. Consequently at least in these proteins the secondary structural motifs can be viewed as solitons of the model Eq. (4), within experimental accuracy. Since the motifs that we have considered are quite generic in PDB data, we have very little doubt that our results will continue to persist whenever we have loops that connect α helices and/or β strands. And as long as the loops are not very long and do not describe bound states of several solitons there does not appear to be any need to introduce more than six parameters. Work is now in progress to systematically construct and classify the solitons that describe the secondary structural motifs in a large class of biologically active proteins.

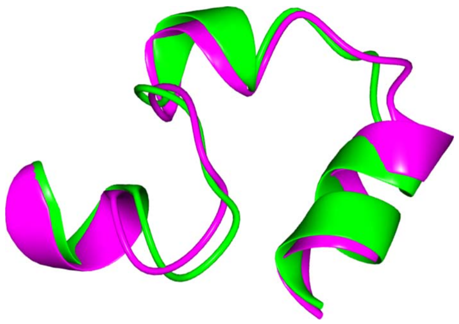

FIG. 4. (Color online) The helix-loop-helix-loop-helix structure of the 1VII protein (light grey, green online) together with its reconstruction in terms of two solitons (dark grey, purple online). The RMSD distance between the two configurations is ≈1.2 Å.

We have also made tentative attempts to use our solitons to reconstruct entire proteins, by *naively* joining the solitons that describe the secondary structural motifs at their ends. In the case of 1VII we have been able to reproduce in this manner the entire protein as a classical soliton with an overall RMSD accuracy of around 1.2 Å and the result is shown in Fig. 4. Even though the accuracy we obtain is very good, the loss of accuracy from ~0.7 to ~1.2 Å when we combine the two solitons in this particular case, suggests that we can still substantially improve the method of assembling an entire folded protein from its solitons. Work is now in progress to develop more efficient methods for assembling entire proteins from their solitons.

In conclusion, we have proposed that the common secondary structural motifs that describe loops connecting α helices and/or β strands can be interpreted as topological solitons, with the α helices and β sheets viewed as ground states that are interpolated by the loops as solitons. Entire proteins can then be assembled simply by combining these solitons together one after another. We have also presented a model that allows us to describe folded proteins in terms of its solitons within experimental accuracy. In its simplest form that we have considered here, the model describes a loop in terms of a single function and six site independent but in general motif dependent parameters, each of which have a direct relation to the overall geometric characteristics of the loop. This observation that all the details and complexities of amino acids and their interactions can be summarized in so simple terms suggests the existence of wide universality in protein folding. It can be viewed as a mathematically precise formulation of the experimental observation that the number of protein conformations is far more limited than the number of different amino acid combinations. Finally, we leave it as a future challenge to expand the model so that it incorporates an order parameter that describes the local orientation of the amino acids along the α carbon backbone.

Our research was supported by grant from the Swedish Research Council (V.R.). We thank Martin Lundgren for discussions.

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

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

[3] A. S. Davydov, J. Theor. Biol. **38**, 559 (1973).

[4] K. A. Dill, O. S. Banu, M. S. Shell, and T. R. Weikl, Ann. Rev. Biophys. **37**, 289 (2008).

[5] G. Jayachandran, V. Vishal, and V. S. Pane, J. Chem. Phys. **124**, 164902 (2006).

[6] K. Huang, *Lectures On Statistical Physics And Protein Folding* (World Scientific, Singapore, 2005).

[7] H. M. Berman, K. Henrick, H. Nakamura, and J. L. Markley, Nucleic Acids Res. **35**, D301 (2007).

[8] P. J. Flory, *Statistical Mechanics of Chain Molecules* (Wiley, New York, 1969).

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

[10] U. H. Danielsson, M. Lundgren, and A. J. Niemi, e-print arXiv:0902.2920.

[11] See for example K. Cahill, Phys. Rev. E **72**, 062901 (2005).