Discrete Nonlinear Schrödinger Equation and Polygonal Solitons with Applications to Collapsed Proteins

Nora Molkenthin,[1,2] Shuangwei Hu,[1,2] and Antti J. Niemi[1,2]

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

(Received 7 October 2010; published 16 February 2011)

We introduce a novel generalization of the discrete nonlinear Schrödinger equation. It supports solitons that we utilize to model chiral polymers in the collapsed phase and, in particular, proteins in their native state. As an example we consider the villin headpiece HP35, an archetypal protein for testing both experimental and theoretical approaches to protein folding. We use its backbone as a template to explicitly construct a two-soliton configuration. Each of the two solitons describe well over 7.000 supersecondary structures of folded proteins in the Protein Data Bank with sub-angstrom accuracy suggesting that these solitons are common in nature.

DOI: 10.1103/PhysRevLett.106.078102 PACS numbers: 87.15.Cc, 05.45.Yv, 36.20.Ey

The discrete nonlinear Schrödinger equation [1] is a prime example of a universal equation. It originally appeared in the connection of polarons in molecular crystals [2] but has since had numerous applications from fiber optics and nonlinear acoustics to quantum condensates and ocean waves. The equation supports both stationary and time dependent solitons that were first introduced to describe Davydov solitons in proteins [3], then found in applications to the crystalline state of acetanilide [4], and subsequently emerged in Bose-Einstein condensates [5]. Today the discrete nonlinear Schrödinger equation together with its generalizations (GDNLS) form a very actively studied family of nonlinear equations, widely utilized to describe a multitude of phenomena in disparate physical, chemical, and biological scenarios [1–6].

In this Letter we argue that solitons of GDNLS equation are also common in polymers, they may even be pivotal in describing the collapsed phase: In general, a polymer such as protein displays three different *nontrivial* phases. These are in the universality class of self-avoiding random walk, in the universality class of Brownian motion, and in the universality class of a collapsed polymer [7]. The first two phases are theoretically quite well understood, and several models have been presented to describe them [8]. But the collapsed phase is much more difficult to describe and tractable models are hard to come by. Here we introduce a novel GDNLS equation that relates to an energy function that has been shown to characterize the collapsed phase [9,10]. We propose that the presence of solitons is essential for describing collapsed (chiral) polymers. While the model we consider is applicable for a large class of (chiral) polymers as a concrete example we address the problem of proteins in their native state, in particular, since there is a large amount of data available for comparisons [11].

We describe a polymer by the coordinates $\mathbf{r}_i$ of the N backbone carbons ($i = 1, \ldots, N$), in the case of proteins these coordinates can be downloaded from the Protein Data Bank (PDB) [12]. We compute the tangent vectors

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

The binormal and normal vectors are given by

$$\mathbf{b}_i = \frac{\mathbf{t}_{i-1} \times \mathbf{t}_i}{|\mathbf{t}_{i-1} - \mathbf{t}_i|} \quad \text{and} \quad \mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i.$$

These vectors are subject to the discrete Frenet equation

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

where T^2 and T^3 are two of the standard generators of three-dimensional rotations, explicitly in terms of the permutation tensor we have $(T^i)_{jk} = \epsilon^i_{jk}$.

From (1) and (2) we compute the bond angles κ_i and the torsion angles τ_i in terms of the PDB data for $\mathbf{r}_i$. Alternatively, if κ_i and τ_i are given we can compute the coordinates $\mathbf{r}_i$. The common convention is to select κ_i to be non-negative, the zeros of its continuum version (the curvature) correspond to the inflection points of the ensuing curve.

We determine κ_i and τ_i by locating the critical points of the following energy function [9,10],

$$E = -\sum_{i=1}^{N-1} 2\kappa_{i+1}\kappa_i + \sum_{i=1}^{N} \{2\kappa_i^2 + c(\kappa_i^2 - m^2)^2\} + \sum_{i=1}^{N} \{b\kappa_i^2\tau_i^2 + d\tau_i + e\tau_i^2 + q\kappa_i^2\tau_i\}. \tag{3}$$

We select κ_i to be periodic, $\kappa_i \in [-\pi, \pi] \bmod (2\pi)$. It is subject to both local and nearest-neighbor interactions. The variable $\tau_i \in [-\pi, \pi] \bmod (2\pi)$ is only subject to local interactions. Finally, (b, c, d, e, m, q) are *global* parameters that in applications to folded proteins are specific to a given supersecondary structure, but are quite independent of the detailed monomer structure.

The energy function (3) is a discretized version of the standard Abelian Higgs Model; see [9] for details. The third term is a symmetry breaking potential. The closely related second term is a remnant of the method we have used to discretize second order derivatives and the fourth term has its origin in the familiar Higgs effect. The fifth term is a one-dimensional version of the Chern-Simons functional; its presence provides a very simple explanation of homochirality with a positive (negative) parameter e giving rise to right-handed (left-handed) chirality. The sixth term is a Proca mass, and the last term is a regulator; if this term is removed, the energy function (3) is exactly the Hamiltonian of a discrete Abelian Higgs Model with Chern-Simons term and Proca mass, in supercurrent variables that are commonly introduced in applications to superconductivity [9].

We note that if we delete all but the first term in the second sum, we arrive at the (discrete) Kratky-Porod model [13] of semiflexible polymers. It cannot describe the collapsed phase of polymers and, in particular, it does not support solitons.

In [10] it has been proposed that the critical points of (3) yield solitons, and approximative methods were introduced to describe them as models of supersecondary helix-loop-helix structures. We now show that (3) relates directly to the GDNLS equation. This equation emerges as follows: We first eliminate the auxiliary variable by varying the energy functional with respect to τ_i. This gives us an equation of motion to resolve for τ_i in terms of κ_i,

$$\frac{\partial E}{\partial \tau_i} = 2b\kappa_i^2\tau_i + 2e\tau_i + d + q\kappa_i^2 = 0 \Rightarrow \tau_i[\kappa_i] = -\frac{1}{2}\frac{d + q\kappa_i^2}{e + b\kappa_i^2}. \tag{4}$$

We then perform a variation of the energy functional with respect to κ_i, and substitute $\tau_i[\kappa_i]$ from (4) into the ensuing equation of motion to arrive at our GDNLS equation

$$\kappa_{i+1} - 2\kappa_i + \kappa_{i-1} = U'[\kappa_i]\kappa_i \equiv \frac{dU[\kappa]}{d\kappa_i^2}\kappa_i \quad (i = 1, \dots, N) \tag{5}$$

(with $\kappa_0 = \kappa_{N+1} = 0$). This equation determines the stationary points of the following GDNLS Hamiltonian

$$H = -2\sum_{i=1}^{N-1}\kappa_{i+1}\kappa_i + \sum_{i=1}^{N}\{2\kappa_i^2 + U[\kappa_i]\}$$

where

$$U[\kappa] = -\left(\frac{bd - eq}{2b}\right)^2\frac{1}{e + b\kappa^2} - \left(\frac{q^2 + 8bcm^2}{4b}\right)\kappa^2 + c\kappa^4.$$

Here the second and the third term are familiar in the context of the nonlinear Schrödinger equation [1–6]. If only the third term is present the Hamiltonian relates to the Hasimoto representation of space curves [14]. Finally, the first term is a generalization of the Vinetskii-Kukhtarev potential [15] of nonlinear waveguides. But none of these truncations, even when they describe solitons, yield a model that relates to proteins in their native state.

If we choose the parameters in (3) so that the potential $U[\kappa]$ has two separate local minima, the results in [16] ensure the existence of a dark soliton solution that interpolates between these two minima. Such a qualitative form of $U[\kappa]$ typically follows if away from the vicinity of $\kappa = 0$ the potential becomes dominated by the second contribution to E in (3). This is the familiar double-well potential term, with minima at $\kappa = \pm m$. A dark soliton is then a configuration that interpolates from the ground state in the vicinity of $\kappa_1 \approx \pm m$ to the ground state in the vicinity of $\kappa_N \approx \mp m$ as we traverse the backbone. When we compute κ_i from (5) and τ_i from (4) and integrate the ensuing discrete Frenet equation we obtain a N-vertex polygonal chain such that a ground state with $\kappa \approx \pm m$ and τ given by (4) is a helix, with the dark soliton describing a loop that connects two helices.

We follow [16] to solve (5) iteratively by locating a fixed point of

$$\kappa_i^{(n+1)} = \kappa_i^{(n)} - \epsilon\{\kappa_i^{(n)}U'[\kappa_i^{(n)}] - (\kappa_{i+1}^{(n)} - 2\kappa_i^{(n)} + \kappa_{i-1}^{(n)})\}. \tag{6}$$

Here $\{\kappa_i^{(n)}\}_{i\in N}$ denotes the nth iteration of an initial configuration $\{\kappa_i^{(0)}\}_{i\in N}$ and ϵ is some sufficiently small but otherwise arbitrary numerical constant, for example, we can choose $\epsilon = 0.01$. It is obvious that a fixed point of (6) satisfies the GDNLS equation (5). As an initial configuration we utilize a step function, chosen to have the same overall topology as the desired dark multisoliton solution. Notice that as it stands, the energy functional (3) has the $\kappa \leftrightarrow -\kappa$ reflection symmetry that may not be exactly realized in applications to folded proteins, for example, there are proteins where a loop connects an α helix with a β sheet. Thus we explicitly break this symmetry using the parameter m: We set $m \rightarrow m_a$ for $N_{a-1} \leq i \leq N_a$ along the chain. Typical values for m_a are $m_a \approx \pm\pi/2$ for the α helix, and $m_a \approx \pm 1$ for the β strand.

We have performed extensive numerical investigations of the dark soliton solutions to (6). We have found that for proper values of the parameters solitons indeed exist and can be combined into multisolitons that together with (4) give a *very* high accuracy approximation of various folded protein structures that are stored in the PDB [12].

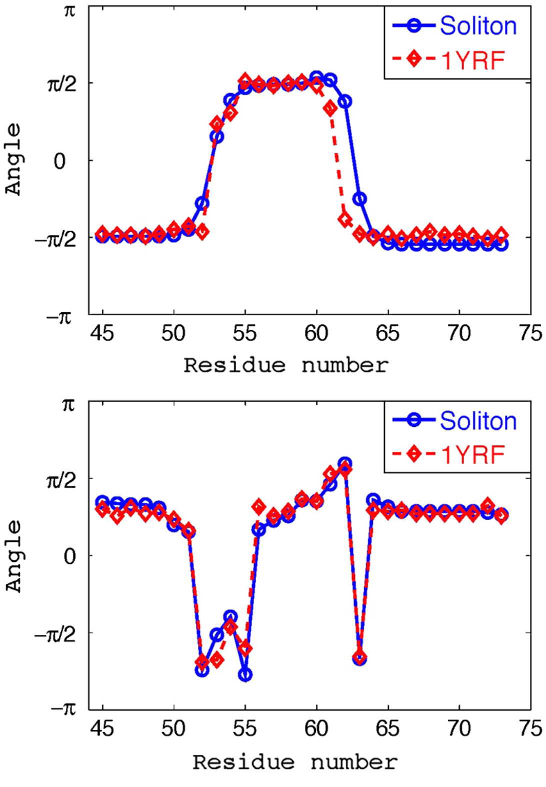

FIG. 1 (color online). (Top): The bond angles κ_i of 1YFR (red) for the sites 3–33 (45–78 in the PDB indexing convention) and their approximation by a soliton solution to Eq. (5) (blue). (Bottom): The torsion angles τ_i of 1YRF (red) for the sites 3–33 (45–78 in the PDB indexing convention) and their approximation by a soliton solution to Eq. (4) (blue).

As an example we construct two dark solitons using as our template the chicken villin headpiece subdomain HP35 (PDB code 1YRF) which is a naturally existing 35-residue protein. It has three α helices separated from each other by two loops. Together with the engineered version (2F4K in PDB) and the very similar HP36 (1VII in PDB), the HP35 has become the subject of very extensive studies both experimentally [17–20] and theoretically [21–24]. Using classical molecular dynamics, the authors of [21–24] report on the construction of native and near-native folds. The native fold in [23] deviates in average around 1.63 Å in C_α RMSD from the x-ray data [19] for the sites 2–34 (counting from the N terminus), and Ref. [24] reports very similar results with a proposed native fold average C_α RMSD around 1.54–1.65 Å for the sites 2–34. The overall resolution in the experimental x-ray data is 1.07 Å in RMSD [19]. We have selected this protein with the hope that by constructing it as a two-soliton solution with loops identified as the solitons, we can provide a new and beneficial perspective for molecular dynamics simulations to become even more effective.

In order to construct a two-soliton solution that describes the HP35 fold in PDB, we first convert the PDB coordinates for the C_α carbons to the bond and torsion angles using (2).

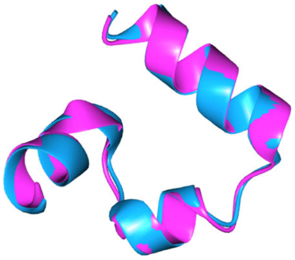

FIG. 2 (color online). Comparison between 1YRF backbone (red [dark gray]) and a soliton solution of (3) (blue [light gray]). The RMSD distance is 0.74 Å.

The result is shown in Fig. 1. The reason we do not consider the entire chain is that in order to compute these angles from the three-dimensional space coordinates we need to know the coordinates of three adjacent C_α carbons. From the κ_i profile we conclude that the C_α backbone of 1YRF consists of two dark solitons. These correspond to the two loops of 1YRF and are located around the sites 49–53 (PDB indexing) and 58–62 in Fig. 1, respectively. These solitons interpolate between ground states that correspond to the three α helices of 1YRF. The first helix is located between the sites 42–49, the second between the loops around sites 53–58, and the third occupies the remaining sites starting from 62 in Fig. 1. While the two-soliton profiles $\{\kappa_i\}$ are clearly identifiable, the profile of $\{\tau_i\}$ is substantially less regular and *a priori* one may expect that the strong irregularity in $\{\tau_i\}$ reflects the amino acid differences in the side chains. *Quite unexpectedly* we have found that this is not the case. The $\{\tau_i\}$ profile can be computed *very* accurately from (4) in terms of the soliton profile κ_i, as the apparent irregularity reflects *solely* the mod(2π) multivalued character of a periodic variable.

To construct the soliton profile for the entire chain, we introduce for each of the two would-be solitons the global parameters (b, c, d, e, m_1, m_2, q): There is one set of parameters for the sites $i = 3$–13 (counting from N terminus) and another set of parameters for the remaining sites. We construct the ensuing soliton solution of (5) by iterating (6) to a fixed point, and compute its RMSD to 1YRF. We then change the parameters randomly and compute the new soliton profile, always starting from the same initial profile for the κ_i. We compare its RMSD to 1YRF with that obtained for the first set of initial parameters using the standard Metropolis algorithm devised to minimize RMSD. By repeating these steps in combination with simulated annealing we eventually produce our final soliton solution.

Note that even though we have seven parameters for each soliton, four of these are determined by the curvature and torsion on each side of the loop and thus *only three* parameters are needed for each of the loops.

Figure 2 compares our minimal RMSD two-soliton configuration with the 1YRF backbone constructed from the x-ray data, for the sites $i = 3$–33. The RMSD between the two configurations is 0.72 Å, well below the overall

TABLE I. Parameter values for a two-soliton solution that describe the entire 1YRF protein with accuracy 0.72 Å. We also present parameters for soliton-1 that describes the first loop (sites 2–13) with accuracy 0.75 Å, and t corresponding values for soliton-2 that describes the second loop (sites 14–33) with accuracy 0.28 Å.

parameter	b	c	d	e	q	m_1	m_2
1st set	$-3.070816340e-04$	$4.461893869e-01$	$1.142581922e-02$	$7.675000601e-04$	$-3.704049149e-03$	1.423206983	1.616099122
2nd set	$-1.095208557e-04$	1.172495797	$5.811514400e-04$	$2.013501270e-04$	$-2.880826898e-04$	1.520126333	1.540139296
soliton-1	$1.800314201e-04$	0.4222887366	$7.02765265e-03$	$4.663610215e-04$	$-2.190120515e-03$	1.444455611	1.565166201
soliton-2	$-2.22159366e-04$	1.088046084	$1.308858509e-03$	$3.94423507e-04$	$-6.4844084e-04$	1.518466566	1.543914339

resolution of the experimental x-ray data (which is 1.07 Å). Indeed, our dark two-soliton pair describes the native 1YRF backbone with an accuracy comparable to that of the radius of a carbon atom. In Table I we provide the parameter values for this configuration. We also present the parameter values for the best individual solitons that we have independently constructed for the two loops.

Since the solitons we have constructed employ the specific profile of 1YRF as a template, one might think that the parameter values in Table I are specific to this particular protein, reflecting its unique amino acid structure. However, this is *not* the case. For example, for the second soliton in Table I we find that there are presently a total of 7.736 unique supersecondary structures in the PDB with RMSD deviation less than 1.0 Å.

In conclusion, we have presented a novel generalized discrete nonlinear Schrödinger equation that supports solitons that describe chiral polymers such as proteins in their collapsed phase. The equation involves only *global* parameters, in particular, the fold is determined by a *single* function. With the 1YRF backbone as a template, we have constructed a soliton configuration that describes the backbone with an atomary level accuracy less than the radius of a carbon atom. Furthermore, we have found that *thousands* of supersecondary structures in the PDB are described with sub-angstrom accuracy by our solitons. Among the future challenges is the enumeration and modeling of the different supersecondary structures in the PDB and developing a relation between genome and a soliton basis of the PDB data.

Our research is supported by a grant from the Swedish Research Council (VR). We thank F. Sha and A.Khorotkin for help in comparing our solitons with the PDB data, M. Lundgren for discussions, and B. Malomed for pointing out a relation to the Vinetskii-Kukhtarev equation. N. M. thanks M. Herrmann for communications.

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

[2] T. Holstein, Ann. Phys. (N.Y.) **8**, 325 (1959).

[3] A. C. Scott, Phys. Rep. **217**, 1 (1992).

[4] J. C. Eilbeck, P. S. Lomdahl, and A. C. Scott, Phys. Rev. B **30**, 4703 (1984).

[5] J. C. Eilbeck and M. Johansson, *The Discrete Nonlinear Schrödinger Equation-20 years on, in Localization and Energy Transfer in Nonlinear Systems*, edited by L. Vázquez, R. S. MacKay, and M. Paz Zorzano (World Scientific, Singapore, 2003).

[6] A. C. Scott, *Nonlinear Science: Emergence and Dynamics of Coherent Structures* (Oxford University Press, Oxford, 2003), 2nd ed..

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

[8] J. F. Marko and E. D. Siggia, Phys. Rev. E **52**, 2912 (1995).

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

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

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

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

[13] O. Kratky and G. Porod, J. Colloid Sci. **4**, 35 (1949).

[14] H. Hasimoto, J. Fluid Mech. **51**, 477 (2006).

[15] V. O. Vinetskii and N. V. Kukhtarev, Sov. Phys. Solid State **16**, 2414 (1975).

[16] M. Herrmann, Applicable Analysis **89**, 1591 (2010).

[17] C. J. McKnight, P. T. Matsudaira, and P. S. Kim, Nat. Struct. Biol. **4**, 180 (1997).

[18] J. Meng, D. Vardar, Y. Wang, H. C. Guo, J. F. Head, and C. J. McKnight, Biochemistry **44**, 11 963 (2005).

[19] T. K. Chiu, J. Kubelka, R. Herbst-Irmer, W. A. Eaton, J. Hofrichter, and D. R. Davies, Proc. Natl. Acad. Sci. U.S.A. **102**, 7517 (2005).

[20] L. Wickstrom, Y. Bi, V. Hornak, D.P Raleigh, and C. Simmerling, Biochemistry **46**, 3624 (2007).

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

[22] D. L. Ensign, P. M. Kasson, and V. S. Pande, J. Mol. Biol. **374**, 806 (2007).

[23] H. Lei and Y. Duan, J. Mol. Biol. **370**, 196 (2007).

[24] P. L. Freddolino and K. Schulten, Biophys. J. **97**, 2338 (2009).