SCIENTIFIC REPORTS

natureresearch

OPEN

Topological Indices of Proteins

Dmitry Melnikov[1,2], Antti J. Niemi[3,4,5,6] & Ara Sedrakyan[1,7]

Protein molecules can be approximated by discrete polygonal chains of amino acids. Standard topological tools can be applied to the smoothening of the polygons to introduce a topological classification of folded states of proteins, for example, using the self-linking number of the corresponding framed curves. In this paper we extend this classification to the discrete version, taking advantage of the "randomness" of such curves. Known definitions of the self-linking number apply to non-singular framings: for example, the Frenet framing cannot be used if the curve has inflection points. However, in the discrete proteins the special points are naturally resolved. Consequently, a separate integer topological characteristics can be introduced, which takes into account the intrinsic features of the special points. This works well for the proteins in our analysis, for which we compute integer topological indices associated with the singularities of the Frenet framing. We show how a version of the Calugareanu's theorem is satisfied for the associated self-linking number of a discrete curve. Since the singularities of the Frenet framing correspond to the structural motifs of proteins, we propose topological indices as a technical tool for the description of the folding dynamics of proteins.

Up to space translations and rigid rotations, curves in three dimensions can be defined in terms of a pair of scalar functions of a single scalar parameter. One possible choice is curvature $\kappa(s)$ and torsion $\tau(s)$ (here selected to be functions of the arc-length parameter s), which respectively provide a local measure of the given curve failing to be straight and planar. Curvature and torsion characterize the local rotation of a right triple of vectors (Frenet frame) $\{\mathbf{n}(s), \mathbf{b}(s), \mathbf{t}(s)\}$ along the curve. Here $\mathbf{t}$ is the tangent vector, $\mathbf{n}$ and $\mathbf{b}$ are the normal and binormal vectors respectively. Given functions $\kappa(s)$ and $\tau(s)$ one can recover the frame at each point and the parameterization $\mathbf{x}(s)$, up to space isometries, using for example, Frenet equations[1].

Given curvature and torsion, Frenet equations define a *framed* curve, that is a ribbon defined by tangent vector $\mathbf{t}$ and a vector $\mathbf{u}$ transverse to $\mathbf{t}$ at every point on the curve. The Frenet frame corresponds to a particular choice $\mathbf{u} = \mathbf{b}$ (or $\mathbf{u} = \mathbf{n}$). It is not always a convenient choice, since at *inflection points*, where $\kappa = 0$, the direction of $\mathbf{b}$ and $\mathbf{n}$ is not defined. These vectors experience a discrete jump by angle π across the inflection point (see the left panel of Fig. 1). Nevertheless, for a smooth curve, one can always introduce a different framing, well-defined at inflection points.

Framed curves can be endowed with a topological characteristic called *self-linking number Lk*, introduced by the Gauss integral formula

$$Lk = \frac{1}{4\pi} \oint_\gamma ds \oint_{\gamma_\mathbf{u}} ds' \frac{[\dot{\mathbf{x}}(s) \times \dot{\mathbf{x}}(s')] \cdot (\mathbf{x}(s) - \mathbf{x}(s'))}{|\mathbf{x}(s) - \mathbf{x}(s')|^3},$$

(1)

where γ is a closed curve and $\gamma_\mathbf{u}$ is its framing, *i.e.* a curve parameterized by $\mathbf{x}(s) + \varepsilon\mathbf{u}(s)$, with some small $\varepsilon > 0$. The integral computes the Gauss' linking number of γ and $\gamma_\mathbf{u}$. It is an integer number invariant under smooth generic deformations of the curve, or its framing. Provided the above mentioned properties of the Frenet framing, one should conclude that it is not a good choice for the calculation of the self-linking number of a curve with inflection points[2]. Nevertheless, the Frenet framing can be useful in detecting such points, providing an

[1]International Institute of Physics, Federal University of Rio Grande do Norte, Campus Universitário, Lagoa Nova, Natal-RN, 59078-970, Brazil. [2]Institute for Theoretical and Experimental Physics, B. Cheremushkinskaya 25, Moscow, 117218, Russia. [3]Present address: Laboratoire de Mathematiques et Physique Theorique CNRS UMR 6083, Fédération Denis Poisson, Université de Tours, Parc de Grandmont, F37200, Tours, France. [4]Nordita, Stockholm University, Roslagstullsbacken 23, SE-106 91, Stockholm, Sweden. [5]Laboratory of Physics of Living Matter, Far Eastern Federal University, Sukhanova 8, Vladivostok, 690950, Russia. [6]Department of Physics, Beijing Institute of Technology, Haidian District, Beijing, 100081, P. R. China. [7]National Laboratory after Alikhanyan, Yerevan Physics Institute, Br.Alikhanyan 2, Yerevan, 0036, Armenia. Correspondence and requests for materials should be addressed to D.M. (email: dmitry@iip.ufrn.br)
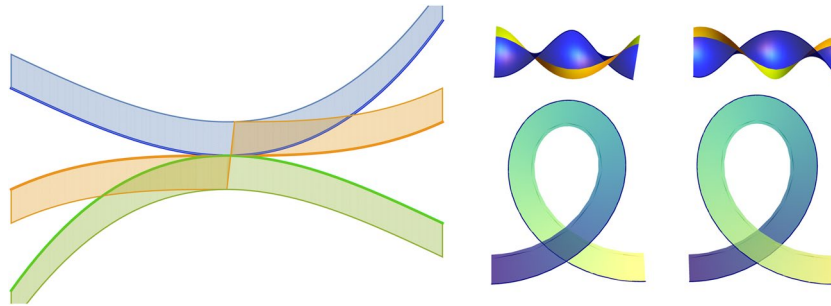
Figure 1. (Left) Three members of a one-parametric family of planar curves corresponding to positive curvature (blue), negative curvature (green) and curvature that changes sign in the middle point (orange). The Frenet framing of the orange curve has a discontinuous jump of the direction of the normal vector across the inflection point. (Right) Segments of a framed 3D curve with $tw = -1$ (top left), $tw = 1$ (top right), $wr = -1$ (bottom left) and $wr = 1$ (bottom right).

additional information about the curve. As non-generic points of the three-dimensional embedding of a segment (or a circle), inflection points can have a physical significance.

The purpose of this paper is to show how one can use the Frenet framing and some of its extensions to detect and classify special points of the discretized versions of curves, the polygons. The primary motivation of this exercise is to connect the special points of framed polygons with the secondary structure motifs in protein molecules and understand their role in the folding process as well as the biological function of the proteins.

We start from a theorem of Calugareanu[3] about basic topology of closed curves. Self-linking number can be calculated from a two-dimensional diagram of a framed curve, obtained by its projection on a selected plane. Calugareanu's theorem (Calugareanu-White-Fuller[4,5]) states that the self-linking number is provided by the sum,

$$Lk = wr + tw, \qquad (2)$$

of two quantities known as writhe and twist. The writhe is defined as a difference of positive and negative self-intersections of the projection $P(\gamma)$ of the curve γ on the plane (bottom right of Fig. 1), while the twist is the half-difference of positive and negative intersections of $P(\gamma)$ and $P(\gamma_{\mathbf{u}})$ (top right of the same figure). Under the smooth variations (isotopies) of the three-dimensional curve, or under the change of the projection plane, writhe and twist can only change in such a way that their sum Lk is preserved. (See[6,7] for three-dimensional definitions of tw and wr).

Calugareanu's formula can be applied to examples of extended quasi-one-dimensional objects in biophysics and soft matter, for example in the DNA supercoiling and folding of polymers and protein molecules. See[5,8–15] for references. A common point of those studies can be put as follows: topology, to a certain extent, controls dynamics of such biological systems.

In this paper we discuss a special version of the self-linking number applicable to protein molecules. This number arises as a result of a "resolution" of the ill-defined linking number associated with the Frenet framing. For proteins, which are not closed curves, the total self-linking number can be defined as a sum of a regular, but not integer, piece, computed with respect to a non-singular framing, and an integer piece, characterizing the singular points of the Frenet framing,

$$Lk = Lk_0 + \frac{\theta}{2}, \ \ \theta \in \mathbb{Z}. \qquad (3)$$

We argue that in the Frenet framing index θ has a meaning of the twist, counting large local rotations of the normal vector. One can locally undo the large rotations by a transformation, which do not change the curve, but introduces a different framing (gauge transformation). In the new framing the total self-linking number can be written as

$$Lk = Lk_0 + \frac{\omega}{2}, \ \ \omega \in \mathbb{Z}, \qquad (4)$$

where ω has the meaning of writhe.

Similarly to the Calugareanu's theorem, one would like to claim that for a generic framing there is an invariant given by the sum $\omega + \theta$. There is however a complication related to the fact that the Frenet framing is singular at the inflection points. In particular, the signs of the discrete contributions to ω and θ are not uniquely defined. We propose that the correct choice for ω and θ is the one, which simply counts the number of singularities of the Frenet framing. This number is a topological invariant, satisfying a version of the Calugareanu's theorem.

The paper is organized as follows. Sections 2 and 3 contain preparatory material necessary for the discussion of the main results in sections 4 and 5. In section 2 we give a definition of the self-linking number in terms of the matrices rotating frames along curves. The version of this definition applicable to discrete curves is discussed in section 3. In section 4 we introduce and compute two different sets of indices for discrete curves, which are shown to take almost integer values for protein molecules. We generalize previous works by showing how different types

of indices can be introduced by considering different choices of discrete framing. In section 5 we explain how the integer indices are related to the self-linking number, as defined in section 3. We conclude that the two indices considered in section 4 are similar to the twist and writhe. Specifically, Frenet framing is a special framing with a pure twist index. We observe that the two indices, and the corresponding twist and writhe, are not equal for the framing choices naturally provided by the proteins. We explain how the definition of the twist and writhe must be rectified in order be compatible with the Calugareanu's theorem. In section 6 we briefly discuss the continuous presentation of curves and the associated gauge symmetry. Loci of the indices are solitons of this continuous model. We conclude in section 7.

Matrix Presentation

It is convenient to use a different way to compute the self-linking number, which stems from the gauge theory formulation of topological invariants of knots and links[16]. This definition is obtained as follows. The frame $\mathbf{e}_a \equiv \{\mathbf{n}(s), \mathbf{b}(s), \mathbf{t}(s)\}$ can be conveniently presented as a 2×2 matrix using Pauli matrices $\hat{\sigma}^i, i = 1, 2, 3$. In other words, we would like to work with spinor representation of the frame vectors. In the topological context such presentation of strings was introduced in[17,18]. The spinor representation of Frenet equation was also considered in[19].

Matrix $\hat{e}$ representing one of the three frame vectors can be introduced as a contraction

$$\hat{e}_a \equiv \{\hat{n}, \hat{b}, \hat{t}\} = \sum_i e_a^i \hat{\sigma}^i. \tag{5}$$

Consequently, three-dimensional rotations will be acting on $\hat{e}$ by $SU(2)$ matrices, such that rotation about axis $\mathbf{u} = \{u^i\}$ by angle α is realized by a matrix

$$S = \exp\left(i\,\frac{\alpha}{2}\hat{u}\right), \quad \hat{u} = \sum_i u^i \hat{\sigma}^i. \tag{6}$$

Note that alternatively one can define vector matrices $\hat{\mathbf{e}} = \mathbf{e}_a \cdot \hat{\sigma}^a$. In such notations, the frames will be rotated by matrices acting in a different space, on indices of a different type. However, the two approaches are equivalent, the difference between them is such that, in the first case, one rotates the frame vectors with respect to a fixed choice of the basis vectors in space, while in the second case, the basis vectors are rotated with respect to a fixed-position frame.

The frame at point s is defined as a $SU(2)$ rotation $S(s)$ of the frame at the origin:

$$\hat{e}_a(s) = S(s)\hat{e}_a(0)S^{-1}(s), \quad a = 1, 2, 3. \tag{7}$$

We note that Frenet equations are just the infinitesimal form of the above formula, while $\kappa(s)ds$ and $\tau(s)ds$ are local infinitesimal rotation angles. Given $\kappa(s)$ and $\tau(s)$ one can recover the rotation matrix at any point s in terms of a path ordered exponential

$$S(s) = P\exp\left(-\int_{\gamma_s} SdS^{-1}\right), \quad SdS^{-1} = \frac{i}{2}\begin{pmatrix} \tau & -i\kappa \\ i\kappa & -\tau \end{pmatrix}ds, \tag{8}$$

where γ_s is a part of the curve parameterized by the arc-length parameter on the interval $[0, s]$.

If the curve is closed the frame should return to itself after completing a closed path along it. In this case $S(0)S^{-1}(L) = \pm I$, where L is the length of the curve. We note that in Eq. (8) the integrand is a flat $SU(2)$ connection, characterized by an integer number of possible singular points of the non-Abelian field strength of the gauge field. Consequently the integral over a closed curve is proportional to an integer number times π. This integer number characterizing the total rotation of the frame along the curve is the self-linking Lk, $S = \exp(\pi i \cdot Lk)$. It can also be related to homotopy classes $\pi_3(S^2)$ of the Hopf map $S^3 \rightarrow S^2$ [17,18].

For open curves number Lk, computed by integral (8), is not an integer. In general, it describes the rotation between the initial and the final orientation of the frame. Since the space of such rotations is not simply connected, the latter split into non-trivial equivalence classes, distinguished by integer numbers, which are equivalent to Lk for closed curves.

Geometry of Discrete Chains

Now we would like to discuss how the above definitions work in the case of discretized curves, *i.e.* polygonal chains. Moreover, we will focus on a nature-given set of chains, the proteins. Protein molecules are quasi-one-dimensional sequences of amino acids. To visualize their geometric structure and study the folding dynamics, a coarse grained description of the molecule is commonly used. One choice is to represent the chain by the positions of the C_α carbon atoms of the amino acids. For our purposes C_α chain will be a discretized version of a smooth three-dimensional curve. Note that proteins provide a specific class of polygonal chains: chemical bonds and steric (excluded volume) constraints introduce a large degree of regularity to the protein molecules, as will be reviewed below. In particular, this allows to describe proteins in terms of a spin-chain-like model, *cf.*[20]. We will discuss the consequences of the constraints for topology.

First, we introduce the discretized Frenet framing (left panel of Fig. 2). The nodes of the polygon $\mathbf{x}_n$ are labeled by index n running from one to N. Tangent vectors are defined as normalized differences of the positions of the
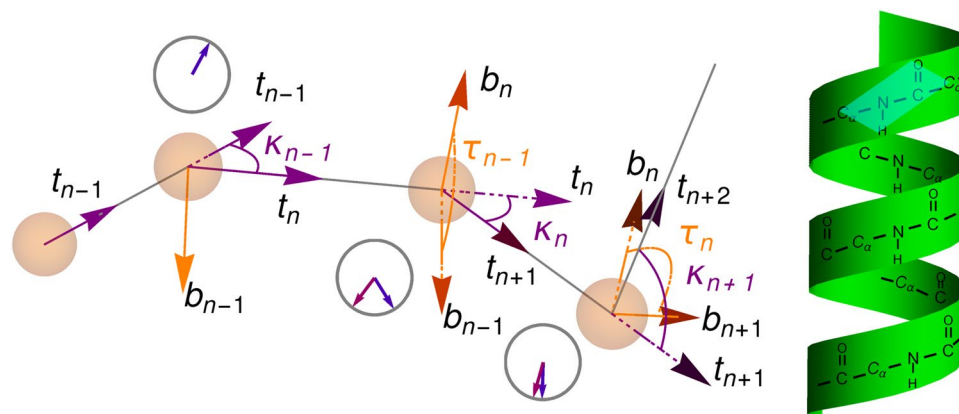
Figure 2. (Left) Definition of the geometric parameters. Curvature angles κ_n define rotation of vectors $\mathbf{t}_n$ towards adjacent $\mathbf{t}_{n+1}$ around $\mathbf{b}_n$. Torsion angles τ_n define rotations of binormal vectors $\mathbf{b}_n$ towards the subsequent vector $\mathbf{b}_{n+1}$ around $\mathbf{t}_{n+1}$. Clocks show the magnitude of the angles for either curvature (blue) or torsion (purple). (Right) Illustration showing peptide framing of the protein chain. In particular, the bonds connecting C_α and side chains are not shown, because they do not belong to the peptide planes. The parallelogram illustrates a single peptide plane.

C_α atoms, $\mathbf{t}_n \propto \mathbf{x}_{n+1} - \mathbf{x}_n$. Binormal vectors are normal to the plane of the local rotation of the tangent vectors, $\mathbf{b}_n \propto \mathbf{t}_n \times \mathbf{t}_{n+1}$, while $\mathbf{n}_n$ complete the right triples (see[21,22] for more details).

Using the discrete Frenet frames one can define curvature and torsion angles, κ_n and τ_n, which serve as discrete versions of the continuous curvature $\kappa(s)$ and torsion $\tau(s)$. As can be deduced from the infinitesimal form of the rotation matrix in Eq. (8) κ_n is the infinitesimal rotation (bond) angle around the binormal vector $\mathbf{b}_n$, while τ_n is the infinitesimal rotation (twist) angle around the corresponding tangent vector. From an example on Fig. 2 (left) one can calculate the discrete rotation angles as follows.

$$\kappa_n = \arccos(\mathbf{t}_n \cdot \mathbf{t}_{n+1}), \;\; n = 1, \dots, N - 2, \tag{9}$$

$$\tau_n = s_n \arccos(\mathbf{b}_n \cdot \mathbf{b}_{n+1}), \;\; n = 1, \dots, N - 3. \tag{10}$$

Here $0 \leq \kappa_n < \pi$, while τ_n is defined with the sign $s_n = \pm 1$, which determines the direction of the rotation of the frame around the tangent vector. This direction can be determined using two consecutive binormal vectors $\mathbf{b}_n$ and $\mathbf{b}_{n+1}$ connected by the bond parallel to $\mathbf{t}_{n+1}$ using equation

$$s_n = \mathrm{sign}([\mathbf{b}_{n+1} \times \mathbf{b}_n] \cdot \mathbf{t}_{n+1}). \tag{11}$$

Consequently, $-\pi \leq \tau_n < \pi$. Note that torsion angles are defined with respect to two consecutive frames, which are subject to the condition that binormal vectors of both frames are both perpendicular to the bond connecting the frames and to the respective tangent vector. The curvature angle is only defined locally, for a single frame. The direction of the rotation around the binormal vectors is fixed and the angle takes values in $[0, \pi]$.

An alternative "natural" framing is provided by the peptide planes of the protein chain. The peptide planes can be defined by the subsequent $C_\alpha C$ and CN bonds along the backbone chain. A non-trivial fact is that also the oxygen of the CO bond and the hydrogen of the NH bond, as well the next C_α connected to N, all lie approximately within the same plane (there are six atoms contained in one plane). At the level of secondary structures, the peptide chain is traditionally visualized as a ribbon, which can be thought as a smoothening of the sequence of the connected peptide planes (approximately as on the right of Fig. 2). The ribbon version of the chain introduces a *peptide framing* of the protein molecule.

Instead of the natural peptide framing, in this paper we will focus on the study of the discrete Frenet framing. We will show that this framing has special points and that those points correspond to secondary structure motifs. To that end we will need to compute torsion and curvature angles and the self-linking numbers of a series of proteins. Following the above discussion the self-linking number is computed by the discrete version of path ordered exponential (8), which is simply a product of discrete rotation matrices ordered along the chain:

$$S(N) = S_{N,N-1} \cdot S_{N-1,N-2} \cdots S_{3,2} \cdot S_{2,1}. \tag{12}$$

According to Fig. 2 (left) there is a natural parametrization of the rotation matrices at each step: first, one rotates the frame around vector $\mathbf{b}_n$ by angle κ_n, and next, rotates around $\mathbf{t}_{n+1}$ by angle τ_n,

$$S_{n+1,n} = e^{i\frac{\tau_n}{2}\hat{t}_{n+1}} \cdot e^{i\frac{\kappa_n}{2}\hat{b}_n}. \tag{13}$$
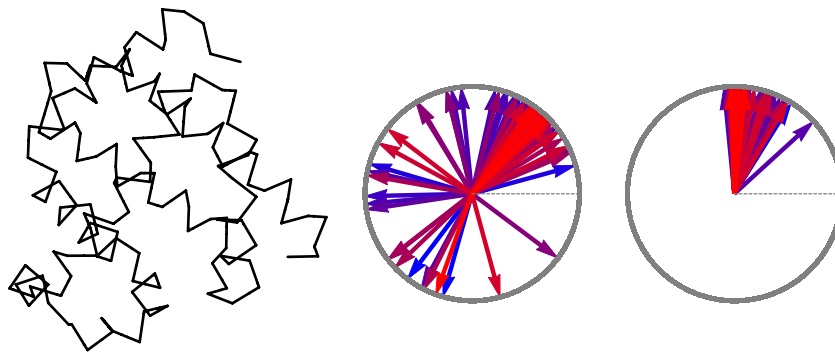
Figure 3. (Left) Chain of the C_α atoms in myoglobin molecule, *1a6m*. (Right) Clocks representing torsion angles τ_n and curvature angles κ_n. Color coding shows the order of rotations, with more blue ones occurring earlier than more red ones.

For a closed curve, the result should just be positive or negative identity matrix, $S = \pm I = \exp(\pi i L k)$, where an integer self-linking number appears as an ambiguously defined phase.

We also note an ambiguity in the definition of the sign of the torsion angle in Eq. (11) if $\mathbf{b}_{n+1} = \pm \mathbf{b}_n$. The case $\mathbf{b}_{n+1} = \mathbf{b}_n$ is a flattening point, torsion angle vanishes across the bond, but the curvature is well defined so there is no actually an ambiguity. The case $\mathbf{b}_{n+1} = -\mathbf{b}_n$ is analogous to an inflection, which is accompanied by a π flip of the direction of the normal and binormal vectors.

Our main interest will be in the latter special points, at which the binormal vector changes sign. If the flip is an exact reflection by π, then one cannot say, whether this gives a positive, or a negative contribution to the self-linking number. On the other hand, in a nature-given chain like protein, the rotation is never exactly by π, so modulo possible experimental error, or even larger rotations by multiples of 2π, the direction of the rotation can be defined unambiguously.

Indices of Proteins

We will now analyze the topological data of a set of proteins, whose structure was obtained with higher resolution and passed additional consistency checks. Technically, these are PDB structures with homology equivalence less than 30%, which are obtained using diffraction data with the resolution better than 1.0 Å and verified to only include structures, which do not have a unit cell containing more than one peptide chain, which do not have a missing heavy atom in the backbone, or do not have alternate positions for heavy atoms, and those, whose chains do not have non-contiguous residue numbers[23]. (We thank K. Hinsen for sharing with us the results of the checks.) For this set of 212 selected proteins we will construct the Frenet framing, determine the local rotation angles and compute topological indices. We will demonstrate the statistics of the indices for all the studied proteins and consider a couple of examples in more detail.

We start with an example of the myoglobin, code *1a6m* in the Protein Data Bank (PDB). Figure 3 shows the C_α chain of this protein and the set of all curvature and torsion angles. The angles are mapped to the phases of unit vectors on a complex plane.

In the case of torsion, 124 out of the total 148 plotted vectors on the clocks of Fig. 3 point to a direction between zero and $\pi/2$ with the average value of $\overline{\tau} \simeq 50°$. The remaining 24 vectors reflect few possible deviations of the rotation from this standard angle. Behavior of the curvature angles is even more regular. Most of the rotations are concentrated close to 90° with the average value of $\overline{\kappa} \simeq 86°$.

Similar behavior of the curvature and torsion angles can be observed for the remaining proteins. Therefore, we would like to work with the following simple model. We will consider a protein as a set of regular periodic structures connected by "kinks" – some irregular connections[24]. Regular structures are characterized by a uniform curvature and torsion angles $\overline{\kappa}$ and $\overline{\tau}$, while angles, considerably deviating from those, correspond to the kinks. This correlates with the common secondary structure classification of proteins, with regular structures corresponding to helices, and kinks – to structural motifs connecting them. We would like to think of regular structures as of a "ground state" of the polygon chain associated with the protein molecule, and of the kinks – as of walls separating domains of the ground state.

Torsion and curvature indices. Basing on the simple view of proteins given above, we would like to introduce a topological classification of the irregular parts of a protein, that is a characteristic that will not care that much about the absolute positions of the kinks, but rather about their sequence and their intrinsic features. One such classification can be produced by a number of full rotations that the clock vectors representing the torsion angle (*e.g.* on Fig. 3) make around the center of the circle, as one follows the polygon chain.
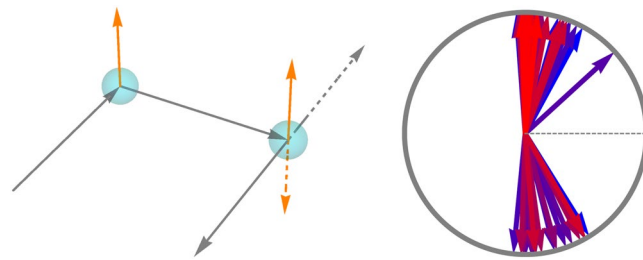
Given the set of torsion angles τ_i one computes the following quantity

$$\vartheta = \frac{1}{\pi} \sum_{i=1}^{N-4} \Delta\tau_i.$$

(14)
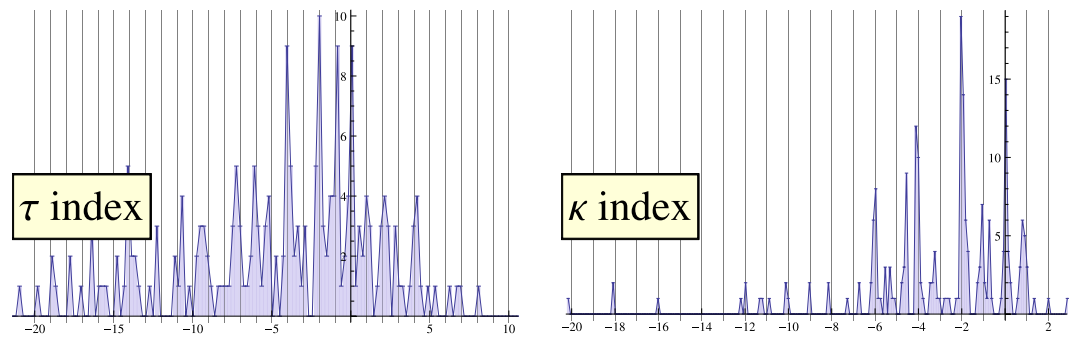
Figure 4. (Left) Distribution of the torsion index ϑ for the set of analyzed proteins (bin size $\simeq 0.2$). (Right) Distribution of the curvature index ϖ in the same set (bin size $\simeq 0.1$). The histograms demonstrate the propensity of the indices towards integer values.

Figure 5. (Left) chain without and with inflection (dashed vectors). (Right) Distribution of κ_n curvature angles in the myoglobin taken with a sign.

This index, first introduced in[24], computes the sum of planar rotations of τ_n vectors on Fig. 3. The rotations are taken in the direction of the smaller angle between τ_n and τ_{n+1}. Using the coordinate presentation of vectors, $\vec{\tau_i} = (\cos\tau_i,\ \sin\tau_i,\ 0)$, the change of angle is defined through

$$\Delta\tau_i = \mathrm{sign}(\hat{\mathbf{z}} \cdot (\vec{\tau_i} \times \vec{\tau_{i+1}}))\arccos(\vec{\tau_i} \cdot \vec{\tau_{i+1}}), \tag{15}$$

where $\hat{\mathbf{z}} = (0,\ 0,\ 1)$ denotes a unit vector perpendicular to the plane.

The analysis of 212 selected proteins is summarized by the histogram on Fig. 4 (left), which shows the distribution of the quantity ϑ. As can be seen from the figure, it has a high propensity towards integer values. This quantization was first discussed in[24], where ϑ was called *folding index*. The statistical distribution of the index around integer values gives an idea of the robustness of this quantization.

Note that instead of summing $\Delta\tau_i$ we could instead sum the angles τ_i themselves. Such an index would be an analog of the integral of the torsion in the continuous case. The latter is the twist *tw*, which appears in the Calugareanu's theorem (2) and is not in general an integer number. Index ϑ computed here, is a discrete version of the integral of the derivative of the torsion, which has loci on the irregular pieces of the chain. It is integer in units of π.

Do the curvature angles carry any similar topological information? A calculation using definition (14), but with $\Delta\kappa_i$ instead of $\Delta\tau_i$ would produce a trivial answer. Few reasons for that is that κ_n are restricted to take values in $[0,\ \pi]$ and all fluctuate close to the same value $\pi/2$. In particular, κ_n do not distinguish two situations shown on Fig. 5 (left), where the chain has two alternative opposite directions. The alternative, shown by dashed vectors, is an inflection, characterized by a flip of the orientation of the binormal vector (orange). In order for inflections to be taken into account by κ_n we decorate the angles with an additional sign, which is determined as follows.

For the first frame we assign κ_1 to be positive. At any position $n + 1$ the sign is defined with respect to the relative orientation of the consecutive **b** vectors: $\mathrm{sign}(\kappa_{n+1}) = \mathrm{sign}(\kappa_n)\,\mathrm{sign}(\mathbf{b}_n \cdot \mathbf{b}_{n+1})$. That is the sign at $n + 1$ remains the same as the sign at n, if the angle between vector $\mathbf{b}_n$ and $\mathbf{b}_{n+1}$ is less than 90° and vice versa. Note that modified κ_n is no longer as local as the original one: it requires at least two consecutive binormal vectors. There is a corresponding global Z_2 symmetry that distinguishes two choices of the sign. This symmetry is spontaneously broken forming domains of different signs of curvature angles. In the myoglobin molecule, the distribution of the curvature angles calculated with the sign is shown on Fig. 5 (right). In this diagram 21 out of 148 curvature angles have negative sign.

Given the set of oriented κ_i one can now calculate index
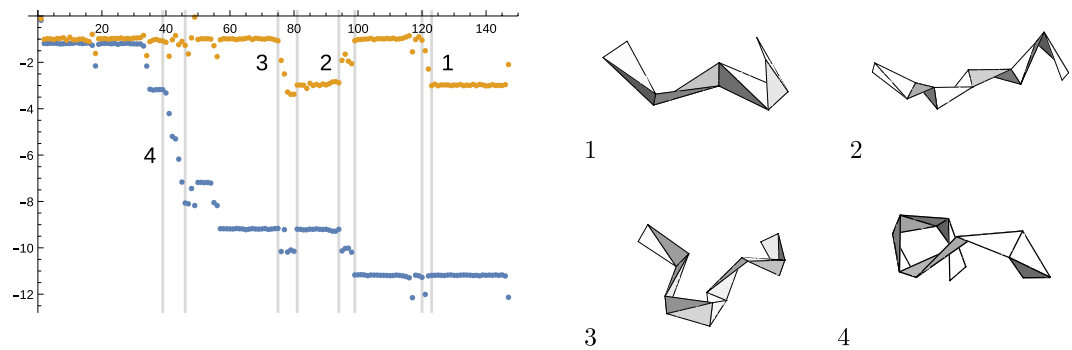
Figure 6. Accumulation of ϑ (yellow) and ϖ (blue) indices along the chain of myoglobin (*1a6m*). Horizontal axis labels the amino acid sequence. Insets show the shapes of framed polygons at the labeled positions.
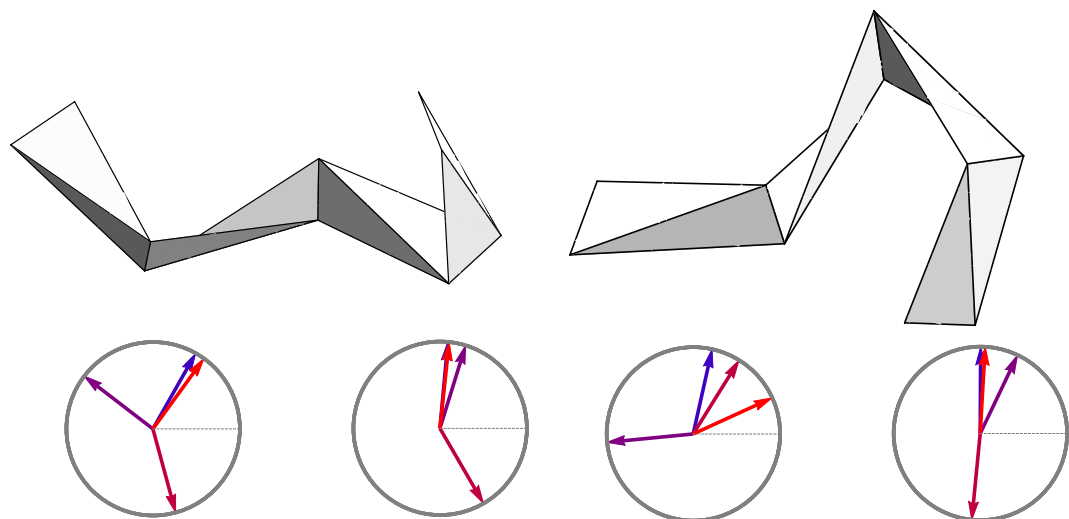
Figure 7. Jump in ϑ (left) and jump in ϖ (right). In each case clock diagrams show the sequence of τ_n (left) and κ_n (right).

$$\varpi = \frac{1}{\pi} \sum_{i=1}^{N-4} \Delta \kappa_i, \tag{16}$$

with $\Delta\kappa_i$ defined in a similar way to $\Delta\tau_i$, *cf*. Eq. (15). Analyzing 212 proteins we found the distribution of index ϖ shown on the right panel of Fig. 4. Again, the index appears to be "quantized" in units of π.

Above we have compared index ϑ with twist *tw*. One can observe that the second index is similar to writhe. Indeed, it is clear that for a straight line, the tangent vector is parallel to itself, the normal vectors rotate in a perpendicular plane and a full 2π rotation adds or removes a unit of twist. Similarly, for a closed planar curve the binormal vector is always parallel to itself and a full 2π rotation gives a shift of the writhe by one unit. Similarly to ϑ, index ϖ computes the relative contribution of large inhomogeneous rotations of the κ vector around the origin.

Evolution of indices and kink structures. We can look a little closer at the protein data to see how indices ϑ and ϖ are built. We consider two characteristic examples. The first one, again, is the myoglobin (*1a6m*), which is a helical protein. The regularity of its structure can be observed from the plot of the index accumulation on Fig. 6. The plot shows the regions of constant index, which correspond to helical structures, and irregular connections through which the index jumps.

One might be interested in how the loci of index accumulation may look like. This is shown on the insets on Fig. 6, which visualize framed kinks at the selected locations along the chain. More specifically, from the left panel of Fig. 7 one can see how the unit of index ϑ is obtained, while the right panel of the same figure shows the locus of the ϖ index.

The second characteristic example is shown on Fig. 8 (left) with the xylanase protein (*1i1w*). This is a longer protein with a richer structure. In contrast to *1a6m*, there is a small variation of ϖ index, while ϑ index exhibits cascades of significant changes.
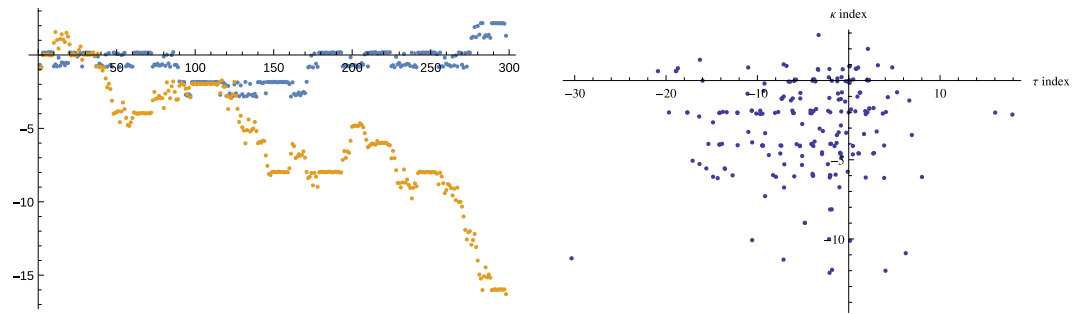
Figure 8. (Left) Evolution of index ϑ (yellow) along the chain of xylanase (*1i1w*). The same for the ϖ index (blue). (Right) Statistics of pairs (ϑ, ϖ) for 212 analyzed proteins. This plot shows the absense of obvious correlations between the two indices.

One can notice already from the two examples considered here that the two indices are not quite correlated. This can also be observed in the plot of (ϑ, ϖ) pairs for many proteins on Fig. 8 (right). In the next section we will discuss a more precise connection of the indices with the self-linking number discussed before.

Self-Linking Number

Twist. Consider the following topological model of proteins. The protein backbone can be mapped to a chain (12) of elementary rotations (13). At regular positions, corresponding to helices, one inserts a rotation matrix $S_{n+1,n}$ with uniform rotation angles $\overline{\kappa}$ and $\overline{\tau}$, as in the previous section. For a small number of positions non-uniform rotations are inserted. In a simple model, we will assume that the non-uniform rotations introduce an additional shift of the τ angle by $\pm\pi$. In other words, we will assume the following form of the rotation matrix at every position

$$\hat{S}_{n+1,n} = e^{i\theta_n \pi \hat{t}_{n+1}/2} \cdot S_{n+1,n}, \ \ \pi\theta_n = \tau_n - \overline{\tau}. \tag{17}$$

In the present model θ_n can take values 0, 1 and -1, so that at every position there is either a uniform rotation, or a rotation with an additional shift.

It follows from relation (7) between frames and rotation matrices that

$$S_{n+1,n} \cdot \hat{e}_a(n) = \hat{e}_a(n+1) \cdot S_{n+1,n}. \tag{18}$$

In other words, rotation matrix $S_{n+1,n}$ transports the frame at position n to the frame at position $n + 1$. We can view the extra rotation around $\mathbf{t}_{n+1}$ in Eq. (17) as a component of a frame following

$$\exp\left(\pm i\frac{\pi}{2}\hat{t}_n\right) = e^{\pm i\pi/2}\hat{t}_n, \ \ \exp\left(\pm i\frac{\pi}{2}\hat{b}_n\right) = e^{\pm i\pi/2}\hat{b}_n. \tag{19}$$

Hence we can apply Eq. (18) replacing $\hat{e}$ by $\hat{t}$: matrices $S_{n+1,n}$ will then transport $\mathbf{t}_n$ at position n to $\mathbf{t}_{n-1}$ at position $N - 1$ and vice versa.

In the string of discrete rotations (12), one can commute the large π rotations through the chain of regular rotations and collect them in the beginning of the chain, arriving at a string

$$S(N) = e^{i\frac{\pi}{2}\theta} \cdot S_{N-2,N-3}\cdots S_{3,2} \cdot S_{2,1} \cdot \hat{t}_1^{\theta}, \tag{20}$$

where

$$\theta = \theta_2 + \cdots + \theta_{N-2}. \tag{21}$$

In Eq. (20) the full rotation matrix splits into three factors. The last factor is an overall $\theta\pi$ rotation of the first frame $\hat{e}_a(1)$ around the tangent vector. It comes from commuting all the extra rotation matrices through the uniform rotations $S_{n+1,n}$. Since $\hat{t}_1^2 = 1$, it is enough to consider θ modulo two in this factor.

The rotation of $\hat{e}_a(1)$ comes with a phase – the first factor in Eq. (20). We understand the phase θ as a (half of the) discrete self-linking number associated with the large π rotations of the Frenet frame. It is an integer number and being associated with rotations around a tangent vector it is a discrete contribution to "twist" *tw*. While a π rotation of a frame as vectors is ambiguous, the spinor representation distinguishes the direction of such a rotation.

Finally, there is a factor corresponding to a chain of uniform rotations that computes the self-linking number of a helix,

$$S_{N-2,N-3}\cdots S_{3,2} \cdot S_{2,1} = S_{N-2,1} = e^{\pi i L k_0 \hat{t}'}, \tag{22}$$

where $\mathbf{t}'$ is the "instantaneous" axis of rotation of the first frame to the last frame.

Let us consider the case of a closed chain. For example, we can assume identifying initial and final points $N \sim 1$. In this case there will be $N - 1$ tangent vectors and $N - 1$ frames. An additional factor $\hat{S}_{1,N-1}\hat{S}_{N-1,N-2}$ will be added to the chain of rotations (12), for example, by left multiplication. Periodic boundary conditions will then require that the total product in Eq. (20) combines into a trivial rotation, that is a plus or minus identity in the spinor representation. The two possibilities are that either rotations around $\mathbf{t}_1$ and around $\mathbf{t}'$ are both trivial, producing phases $e^{i\pi Lk_0}$ and $e^{i\pi/2\theta}$, for even θ, or they are both π rotations around $\mathbf{t}_1$ with θ being odd and Lk_0 half-integer. In either case

$$S_{\text{closed}} = \exp(i\pi Lk) = e^{i\frac{\pi}{2}(\theta + 2Lk_0)}. \tag{23}$$

In other words, we find that the full self-linking number splits in the sum of the regular self-linking number counting the uniform (small) rotations and the large π jumps:
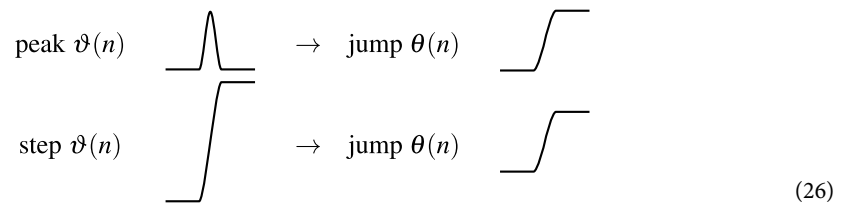
$$Lk = Lk_0 + \frac{\theta}{2}. \tag{24}$$

Lk_0 is a half-integer if the number of π jumps is odd. In most general open chain case Lk_0 is a real number, while θ is an integer. An appropriate generalization of Eq. (24) would be

$$Lk = [Lk_0] + \begin{cases} \dfrac{\theta}{2}, & \theta \text{ even} \\ \dfrac{\theta - 1}{2}, & \theta \text{ odd} \end{cases}, \tag{25}$$

where $[Lk_0]$ stands for the integer part of Lk_0.

We should yet connect the index θ discussed here to the index ϑ computed in section 4. We noted before that the sum of torsion angles, or equivalently, the integral of torsion, is not a topological invariant – in particular, it is not the full self-linking number. From the point of view of our discussion this happens, because by non-commutativity of three-dimensional rotations, we cannot simply add torsion angles to obtain Lk. However, this can be done with the discrete π rotations, which "commute" with the rotation matrices. Consequently, θ is counting the sum of the individual jumps accounting for their direction.

Index ϑ is closely related to θ, but is not precisely the same. It computes local difference between τ_n and in continuous case would correspond to integrating a derivative of τ. A key subtlety is that torsion angles are defined modulo 2π, to which θ is less sensitive. To explain this point we first notice that there are two kind of structures in the behavior of $\vartheta(n)$ on Fig. 7 (left): the peaks and the steps. Both of these structures indicate the same jump θ_n as in the indicated on the following diagram corresponding to a positive jump

$$\text{peak } \vartheta(n) \quad \underline{\wedge} \quad \rightarrow \quad \text{jump } \theta(n) \quad \underline{\int}$$

$$\text{step } \vartheta(n) \quad \underline{\int} \quad \rightarrow \quad \text{jump } \theta(n) \quad \underline{\int} \tag{26}$$

Note that the peak has a magnitude π, while the height of the step is 2π, and both of the contribute $+\pi$ to θ (similarly for the negative peaks, steps and jumps). If $\vartheta(n)$ only consisted of steps than one would find $\vartheta = 2\theta$, but due to additional winding information contained in ϑ, values of ϑ and θ are independent.

Writhe. Finally, we should explain the meaning of index ϖ. To find a non-trivial index in terms of the curvature angles we have extended their domain to negative values. As was mentioned, a positive κ_n rotation around vector $\mathbf{b}_n$ is equivalent to a negative rotation around vector $\mathbf{u}_n = -\mathbf{b}_n$. So a non-trivial index ϖ corresponds to a different, non-Frenet framing of the polygon. This new framing can be introduced by additional local rotations of the frame

$$S(N) = \cdots \hat{S}_{n+1,n}\hat{S}_{n,n+1}\cdots = \cdots \hat{S}_{n+1,n}U_n^{-1}U_n\hat{S}_{n,n+1}\cdots = \cdots \hat{S}'_{n+1,n}\hat{S}'_{n,n+1}\cdots, \tag{27}$$

where U_n rotates $\mathbf{b}$ and $\mathbf{n}$ vectors by π at positions n. The tangent vectors remain the same, so only the framing is changed. In the new framing we will use labels $\mathbf{u}_n$ for the counterparts of the binormal vectors. This rearrangement is equivalent to a gauge transformation in the definition of the self-linking number (12), so it does not change the topology.

In the $\mathbf{u}$-framing we cast the rotation matrices in the form

$$\hat{S}'_{n+1,n} = S_{n+1,n} \cdot e^{i\omega_n \pi \hat{u}_n/2}, \tag{28}$$

where now ω_n, which can be either 0, 1, or -1, represent additional shifts of rotations relative to the uniform curvature angles. In this way, we can still assume $0 \le \kappa_n < \pi$, while the "negative" rotations around $\mathbf{u}_n$ would correspond to $\omega_n = \pm 1$. This change is reflected in the transition from the curvature angle diagram on Fig. 3

(right) to the one on Fig. 5 (right). We note that in this case there are no extra shifts in the torsion angles, since those shifts are "undone" by transformations U_n.

We can play the same game defining an index

$$\omega = \omega_1 + \cdots + \omega_{N-3},\tag{29}$$

which corresponds to collecting all the extra rotations localized at the initial frame.

$$S(N) = e^{i\frac{\pi}{2}\omega} \cdot S_{N-2,N-3}\cdots S_{3,2} \cdot \hat{b}_1^{\,\omega},\tag{30}$$

where, by definition, $\mathbf{u}_1 = \mathbf{b}_1$.

Once again, imagining a closed polygon, the product of uniform rotations and $\omega\pi$ rotation around $\mathbf{b}_1$ should conspire to produce a phase $\exp(i\pi \cdot Lk)$ with an integer Lk. As rotations around a $\mathbf{b}$ vector, ω_n will contribute to the "writhe" part of the self-linking number.

As in the case of θ, one can compare index ω with index ϖ computed in section 4. The relation between ϖ and ω is the same as between ϑ and θ. Consequently, one can recover ω from the evolution of ϖ, as in the examples of Figs 6 (left) and 8 (left), using the rules outlined by Eq. (26).

An obvious question is whether one should expect θ equal to ω, since after all, both indices count the points with a large flip of the $\mathbf{b}$ vector. It is clear from Fig. 8 (right), that $\vartheta \neq \varpi$. Moreover, one can also find examples in Figs 6 (left) and 8 (left), where the direction of the change of ϑ is different from the direction of the change of ϖ. Therefore, in general, $\theta \neq \omega$. The discrepancy has the following explanation.

Indices θ and ω identify and quantify inflection points, where $\mathbf{b}$ and $\mathbf{n}$ have discontinuous π jumps. The direction of the jump, or the sign of π is not defined. Meanwhile, in the discrete polygon, like the protein molecule, there are no well-defined inflection points. What one has is some discretized resolution of the framing across inflections. The resolution depends, among other things, on independent "random fluctuations" of the vectors κ_n and τ_n around their mean values. For example, the jump in θ is defined as $\pi\theta_n = \tau_n - \tau_{n-1}$, where τ_{n-1} is supposed to be close to $\overline{\tau}$. The difference is never exactly π, which allows to determine the sign of the rotation, but the value of the sign depends on particular values of the angles at position $\mathbf{n}$. Consequently, the $\mathbf{b}$ and $\mathbf{u}$ framings represent two different topological classes characterized by two self-linking numbers, whose values differ by $|\theta - \omega|$.

Nevertheless some "invariant" information can be obtained. Since by construction, and by observation of Fig. 6 (left), large jumps occur simultaneously in curvature and torsion channels, one at least expects that

$$\theta = \omega \bmod 2.\tag{31}$$

We checked that for 80% of the analyzed proteins, rounded values of ϑ and ϖ are indeed equal modulo 2. Moreover, one can count the total number of jumps, which should agree between two channels,

$$\sum_n |\omega_n| = \sum_m |\theta_m|.\tag{32}$$

This is a topological invariant since any local large rotation of the frame removing θ_n will create a contribution to ω_n.

More generally, one can define the set of vectors $\mathbf{u}_n$ in terms of two-dimensional planes V in the three-dimensional space. In the case of a generic plane all vectors $\mathbf{b}$ will either have a positive or a negative scalar product with one of the two normal vectors to the plane. We will flip the direction of those $\mathbf{b}_n \to \mathbf{u}_n$ whose scalar product sign would be opposite to that of vector $\mathbf{b}_1$. If $\mathbf{b}_n$ and $\mathbf{b}_{n+1}$ had opposite orientations (not within V) one of them would always become a $\mathbf{u}$-vector giving a non-zero ω_{n+1}. Since large rotations of $\mathbf{b}$ vectors do not always happen through exactly a π angle, there will also be non-zero θ contributions. The topological invariant of the polygon will be the sum

$$\sum_n |\omega_n| + \sum_m |\theta_m|,\tag{33}$$

where only the values ω_n and θ_m are V-dependent, but not the sum of their absolute values. This is a special version of the Calugareanu's theorem.

Gauge Theory Description

In the discrete approach three-dimensional curves (polygons) can be described using set of angles κ_n and τ_n. One might have noticed that curvature angles are associated to the nodes of the polygons, while torsion angles correspond to the bonds connecting the nodes. In the language of lattice gauge theory κ_n can be understood as vertex (matter) degrees of freedom, while τ_n are "connections" (gauge fields). Indeed this point can be further elaborated in the continuous presentation (see[25]).

In the Frenet picture curvature $\kappa(s)$ and torsion $\tau(s)$ characterize the rotation of the Frenet frame around the curve. However, the particular choice of the Frenet framing is not physical. There is an infinite number of possible choices of framing for a smooth generic curve. For example, a different local choice can be realized by introducing a local rotation of the Frenet frame around the tangent vector. It is not hard to see, that from the point of view of the Frenet description, such a local rotation induces transformations of the curvature and torsion[25],

$$\kappa(s) \to e^{i\alpha(s)}\kappa(s), \ \ \tau(s) \to \tau(s) + \alpha'(s),\tag{34}$$

where it is convenient to generalize curvature to a complex quantity. Indeed, this means that curvature behaves like a complex scalar field, while τ is a one-dimensional analog of a gauge fields.

At the same time, we have seen from Eq. (8) that one-dimensional gauge field τ and complex field κ can be viewed as components of a non-Abelian $SU(2)$ gauge connection. Moreover the non-Abelian one-dimensional connection is a projection of a three-dimensional connection on the given curve, see *e.g.*[26]. The three-dimensional connection in turn, defines framing at any point in space. It is well known, that framed three-dimensional manifolds are classified by an integer framing number. Consequently, one can think of the framing as coming from an embedding of the curve in a framed manifold. The choice of the framed manifold will define the framing and the associated integer number – the self-linking. Projection of the three-dimensional connection on a selected curve breaks the $SU(2)$ symmetry down to the $U(1)$ subgroup (34) of frame rotations around the tangent vector.

Based on the standard symmetry approach to construction of effective actions in field theory one can ask, what is the effective description of curves defined by functions $\kappa(s)$ and $\tau(s)$, subject to gauge transformations (34). A simplest effective model of such degrees of freedom is provided by the Abelian Higgs model[25]. In order for the theory to reproduce solutions with non-zero torsion the minimal extension is to add a one-dimensional Chern-Simons term (see[26,27] for the general approach to the construction of the effective model), as in the following functional, *cf.*[25],

$$F = \int ds \left(\frac{1}{2} (\nabla - i\hat{\tau}) \hat{\kappa}^* (\nabla + i\hat{\tau}) \hat{\kappa} - \frac{\mu^2}{2} |\hat{\kappa}|^2 + \frac{\lambda}{2} |\hat{\kappa}|^4 \right) - F \int ds \, \hat{\tau},$$

(35)

Here hats upon torsion and curvature indicates that they are gauge dependent quantities transforming according to Eq. (34). In the Frenet gauge curvature is a real scalar $\hat{\kappa} = \kappa$ and $\hat{\tau} = \tau$. Constants μ, λ and F are phenomenological parameters of the curves, which are obtained by comparing the curves and basic geometric features of the protein molecules. ∇ is the one-dimensional gradient. The last term is an analog of the Chern-Simons action for gauge fields in three dimensions. To make the theory regular it might also be useful to add a gauge invariant Proca mass term for the gauge field proportional to τ^2.

Functional (35) can serve as an effective free energy functional, whose minimum energy configurations are curves with constant curvature and torsion, *i.e.* helices. Apart from the lowest energy translationally invariant solutions (ground states) such theories can also have solitons, *i.e.* solutions interpolating between the same, or different ground states[21]. The solitons can be either kinks or lumps (dents) in curvature and torsion. Clearly plots of the topological indices on Figs 3 and 8 are examples of such solitons. In other words, topological indices θ and ω count the total soliton numbers of the proteins. We do not intend to describe the corresponding solitons here, though many details about their discrete version, in direct application to proteins, can be found in works[28–31].

Discussion

In this paper we considered effects of framing for nature given discretizations of smooth curves, such as protein molecules. Protein backbones equipped with the Frenet framing underline basic features of the secondary structure of proteins, such as existence of regular helical pieces connected by structural kinks (motifs). Studying the structure at the topological level we have introduced a simple spin-chain-like model of proteins, in which regular pieces corresponded to parallel oriented vectors (ground state), while kinks corresponded to short sequences of vectors that deviated from the standard orientation.

We have demonstrated that the kinks-motifs can be counted by topological indices taking discrete values. This owes to the fact that the kinks correspond to inflection points of smooth versions of the discrete curves, where the Frenet frame experiences large rotations. The discrete indices can be related to the self-linking number Lk of a framed curve, which is a topological invariant. Although the self-linking number is defined for smooth framed curves, we showed how the notion of the self-linking number can be extended to the case of singular framings, such as the Frenet framing with inflection points. In this case the invariant splits into a sum of the regular self-linking number of a regular framing and an integer index associated to the inflection points.

We have shown that the Frenet framing corresponds to a special type of the singular self-linking number – a pure twist. To have also a writhe-type contributions to Lk one has to flip at least a part of the normal vectors in the Frenet framing. Such flips are local (gauge) transformations that convert units of twist to units of writhe in accordance with the Calugareanu's theorem. We showed that in terms of the indices defined here, the theorem is satisfied only if we compute all the indices modulo sign. Otherwise our indices are defined with respect to incompatible framings and cannot be compared. This is, of course, a result of the singularity of the Frenet framing.

Spin chains can be extended to effective field theory models. In such models curvature is promoted to a complex scalar field $\kappa(s)$, while torsion is a one-dimensional scalar field. Indices computed here can be related to topological charges of field theory solitons.

As an open direction, it is interesting to study how the topological indices defined here behave in the protein's dynamics. Formation of the structural motifs is one of the basic steps in the protein folding, so understanding the behavior of the indices, e.g. during a protein folding process, could be a good way to understand better how proteins fold. Moreover, the evolution of the indices during the time evolution of a dynamical protein could tell us how proteins move. At the moment, we are not aware of any systematic experimental data, that could be used for such a study. The only data that one could conceivably use, may come from molecular dynamics simulations with quite long proteins. This is an ambitious program for the future. As an alternative, one could try to make simulations with coarse grained force fields such as UNRES[32]. We hope to look at this, in a future publication.

References

1. Frenet, F. Sur les courbes double courbure. *J. Math. Pures Appl.* **17**, 437–447 (1852).
2. Aicardi, F. Self-linking of spatial curves without inflections and its applications. *Funct. Anal. Its and Appl.* **34**, 79–85, https://doi.org/10.1007/BF02482420 (2000).
3. Călugăreanu, G. L'intégrale de gauss et l'analyse des noeuds tridimensionnels. *Rev. Math. Pures Appl.* **4**, 5–20 (1959).
4. White, J. H. Self-linking and the gauss integral in higher dimensions. *Am. J. Math.* **91**, 693–728 (1969).
5. Fuller, F. B. The writhing number of a space curve. *Proc. Natl. Acad. Sci. USA* **68**, 815–819, https://doi.org/10.1073/pnas.68.4.815 (1971).
6. Kauffman, L. H. *Knots and Physics*. 3rd edn. (World Scientific, Singapore, 2001).
7. Dennis, M. R. & Hannay, J. H. Geometry of calugareanu's theorem. *Proc. R. Soc. Lond. A* **461**, 3245–3254, https://doi.org/10.1098/rspa.2005.1527 (2005).
8. Fuller, F. B. Decomposition of the linking number of a closed ribbon: a problem from molecular biology. *Proc. Natl. Acad. Sci. USA* **75**, 3557–3561, https://doi.org/10.1073/pnas.75.8.3557 (1978).
9. Pohl, W. F. Dna and differential geometry. *Math. Intell.* **3**, 20–27, https://doi.org/10.1007/BF03023391 (1980).
10. Bauer, W. R., Crick, F. H. & White, J. H. Supercoiled dna. *Sci. Am.* **243**, 100–113 (1980).
11. Frank-Kamenetski, M. D. & Vologodski, A. V. Topological aspects of the physics of polymers: The theory and its biophysical applications. *Soviet Physics Uspekhi* **24**(8), 679–696, https://doi.org/10.1070/pu1981v024n08abeh004835 (1981).
12. Millett, K. C. & Sumners, D. W. L. *Random Knotting and Linking*. K & E series on knots and everything (World Scientific Publishing Company Pte Limited, 1994).
13. Klapper, I. Biological applications of the dynamics of twisted elastic rods. *Journal of Computational Physics* **125**, 325–337, https://doi.org/10.1006/jcph.1996.0097 (1996).
14. Katrich, V. & Vologodskii, A. The effect of intrinsic curvature on conformational properties of circular dna. *Biophys. J.* **42**, 1070–1079, https://doi.org/10.1016/S0006-3495(97)78757-5 (1997).
15. Hoffman, K. A., Manning, R. S. & Maddocks, J. H. Link, twist, energy, and the stability of dna minicircles. *Biopolymers* **70**, 145–157, https://doi.org/10.1002/bip.10430 (2003).
16. Polyakov, A. M. Fermi-bose transmutations induced by gauge fields. *Mod. Phys. Lett.* **A3**, 325, https://doi.org/10.1142/S0217732388000398 (1988).
17. Kavalov, A. R. & Sedrakian, A. G. Fermion representation of the three-dimensional ising model. *Nucl. Phys. B* **285**, 264–278, https://doi.org/10.1016/0550-3213(87)90338-5 (1987).
18. Kavalov, A. R. & Sedrakyan, A. G. The sign factor of the three-dimensional ising model and the quantum fermionic string. *Phys. Lett. B* **173**, 449–452, https://doi.org/10.1016/0370-2693(86)90413-2 (1986).
19. Ioannidou, T., Jiang, Y. & Niemi, A. J. Spinors, strings, integrable models, and decomposed yang-mills theory. *Phys. Rev. D* **90**, 025012, https://doi.org/10.1103/PhysRevD.90.025012 (2014).
20. Bryngelson, D. & Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA* **84**, 7524–7528, https://doi.org/10.1073/pnas.84.21.7524 (1987).
21. Chernodub, M., Hu, S. & Niemi, A. J. Topological solitons and folded proteins. *Phys. Rev. E* **82**, 011916, https://doi.org/10.1103/PhysRevE.82.011916 (2010).
22. Hu, S., Lundgren, M. & Niemi, A. J. The discrete frenet frame, inflection point solitons and curve visualization with applications to folded proteins. *Phys. Rev. A* **83**, 061908, https://doi.org/10.1103/PhysRevE.83.061908 (2011).
23. Hinsen, K., Hu, S., Kneller, G. R. & Niemi, A. J. A comparison of reduced coordinate sets for describing protein structure. *J. Chem. Phys.* **139**, 124115, https://doi.org/10.1063/1.4821598 (2013).
24. Lundgren, M., Krokhotin, A. & Niemi, A. J. Topology and structural self-organization in folded proteins. *Phys. Rev. E* **88**, 042709, https://doi.org/10.1103/PhysRevE.88.042709 (2013).
25. Danielsson, U. H., Lundgren, M. & Niemi, A. J. Gauge field theory of chirally folded homopolymers with applications to folded proteins. *Phys. Rev. E* **82**, 021910, https://doi.org/10.1103/PhysRevE.82.021910 (2010).
26. Gordeli, I., Melnikov, D., Niemi, A. J. & Sedrakyan, A. Chern-simons improved hamiltonians for strings in three space dimensions. *Phys. Rev. D* **94**, 021701, https://doi.org/10.1103/PhysRevD.94.021701 (2016).
27. Hu, S., Jiang, Y. & Niemi, A. J. Energy functions for stringlike continuous curves, discrete chains, and space-filling one dimensional structures. *Phys. Rev. D* **87**, 105011, https://doi.org/10.1103/PhysRevD.87.105011 (2013).
28. Molkenthin, N., Hu, S. & Niemi, A. J. Discrete nonlinear schrödinger equation and polygonal solitons with applications to collapsed proteins. *Phys. Rev. Lett.* **106**, 078102, https://doi.org/10.1103/PhysRevLett.106.078102 (2011).
29. Hu, S., Krokhotin, A., Niemi, A. J. & Peng, X. Towards quantitative classification of folded proteins in terms of elementary functions. *Phys. Rev. E* **83**, 041907, https://doi.org/10.1103/PhysRevE.83.041907 (2011).
30. Krokhotin, A., Niemi, A. J. & Peng, X. Soliton concepts and protein structure. *Phys. Rev. E* **85**, 031906, https://doi.org/10.1103/PhysRevE.85.031906 (2012).
31. Krokhotin, A., Lundgren, M. & Niemi, A. J. Solitons and collapse in the λ–repressor protein. *Phys. Rev. E* **86**, 021923, https://doi.org/10.1103/PhysRevE.86.021923 (2012).
32. UNRES. A package to carry out coarse-grained simulations of protein structure and dynamics, https://unres.pl.

Acknowledgements

Author Contributions

D.M., A.J.N. and A.S. equally contributed to conceiving and developing the theory and methodology of this work. D.M. performed the analysis of the PDB data. All authors reviewed the manuscript.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.