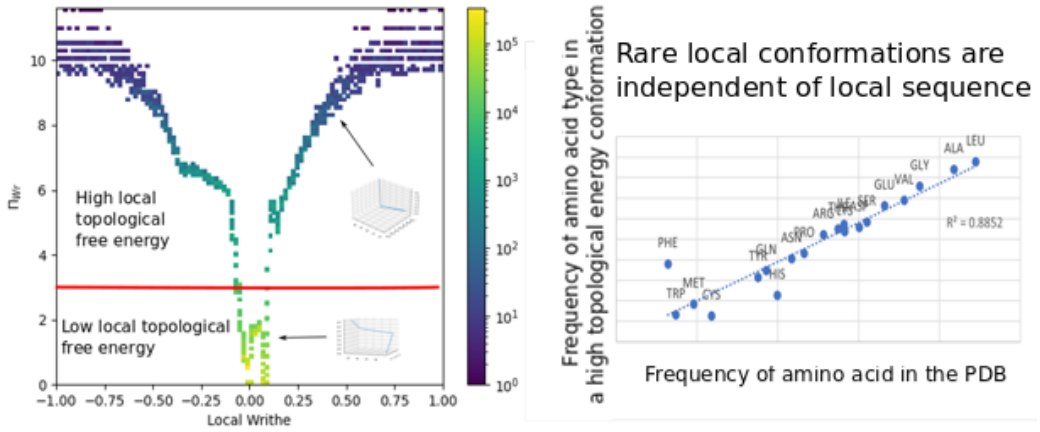
Graphical Abstract

The local topological free energy of proteins

Quenisha Baldwin, Eleni Panagiotou



Highlights

The local topological free energy of proteins

Quenisha Baldwin, Eleni Panagiotou

- The Writhe and Torsion are used to introduce a new local topological/geometrical free energy that can be associated to 4 consecutive amino acids along the protein backbone.
- High local topological free energy conformations are independent of sequence.
- High local topological free energy conformations are independent of secondary structure.
- High local topological free energy conformations may be involved in the rate limiting step in protein folding.

The local topological free energy of proteins

Quenisha Baldwin[a], Eleni Panagiotou[b]

[a]*Department of Biology, Tuskegee University, 1200 W Montgomery Rd, Tuskegee, 36088, AL, USA*
[b]*Department of Mathematics and SimCenter, University of Tennessee at Chattanooga, 615 McCallie Ave, Chattanooga, 37403, TN, USA*

Abstract

Protein folding, the process by which proteins attain a 3-dimensional conformation necessary for their function, remains an important unsolved problem in biology. A major gap in our understanding is how local properties of proteins relate to their global properties. In this manuscript, we use the Writhe and Torsion to introduce a new local topological/geometrical free energy that can be associated to 4 consecutive amino acids along the protein backbone. By analyzing a culled protein dataset from the PDB, our results show that high local topological free energy conformations are independent of sequence and may be involved in the rate limiting step in protein folding. By analyzing a set of 2-state single domain proteins, we find that the total local topological free energy of these proteins correlates with the experimentally observed folding rates reported in [1].

Keywords: Protein structure, protein folding, topology, Writhe, Torsion
PACS: 0210, 8714, 8715
2000 MSC: 57M25, 60-08

1. Introduction

Protein folding is the process by which a protein attains a unique three-dimensional conformation necessary for its function [2]. Many different models of protein folding have been proposed, all of which aim to understand the free energy barrier associated with the transition from the unfolded configuration to the native state of the protein [3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]. To describe this process, which involves multiple lengthscales (from the length scale of an amino acid to that of the

entire protein backbone), it is necessary to have a meaningful characterization of the 3-dimensional configuration of proteins across length scales. In this manuscript we introduce characterizations of protein conformations using tools from mathematics, related to knot theory, that apply to all protein length scales. In particular, we focus on characterizing the local conformations of proteins (those of 4 consecutive α carbon atoms, denoted CA) and exploring the relation to the global configuration of the protein and protein kinetics.

Folded proteins are defined by their primary, secondary, tertiary and quaternary structure [2]. The primary structure refers to the protein amino acid sequence. The secondary structure refers to a sequence of 3-dimensional building blocks the protein attains (beta sheets, alpha helices, coils). The tertiary structure refers to the 3-dimensional conformation of the entire polypeptide chain. The quaternary structure of a protein comprises of 2 or more polypeptide chains. More refined methods to characterize protein conformations than these classifications are also used. For example, at the level of amino acids, the Ramachandran plot, is a traditional way to capture the geometrical signatures of amino acids in terms of their dihedral angles in 3-space. At the length scale of the entire protein, the number of sequence-distant contacts is a way to describe the conformation of the protein, which has shown a remarkable correlation with experimentally observed folding rates [21, 22, 23, 24].

In the last decades, measures from knot theory have been applied to biopolymers [25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40] and in particular to proteins to classify their conformations [41, 42, 43, 44, 45, 46, 47, 48, 49, 50]. One of the simplest measures of conformational complexity of proteins that does not require an approximation of the protein by a knot dates back to Gauss; the Writhe of a curve. Studies have applied the Gauss linking integral to measure the entanglement of the protein backbone by taking the entire backbone of the protein or by looking at linking between parts of the protein. Both approaches have found a correlation between folding rates and these measures of conformational complexity [42, 43, 44, 45]. However, this does not answer how local properties of proteins relate to its tertiary structure. The protein backbone, represented by its CA atoms, can attain interesting conformations even with as few as 4 amino acids. To our knowledge no study has focused at exploring the local topology/geometry of the proteins at the length scale of 4 amino acids using the Writhe. In this manuscript, we use the Writhe to define a novel topological/geometrical free

energy that can be assigned locally to the protein. We do the same also using the Torsion. We use this free energy to identify conformations with high local topological free energy. These high local topological free energy conformations are not favored in folded proteins and we hypothesize they may be important in protein folding. Our results show that the high local topological free energy conformations are independent of the local sequence and of secondary structure and that they may be involved in the rate limiting step in protein folding. By analyzing a set of single domain proteins that have been reported to fold in a concerted, all-or-none, two-state fashion, we show that the previously reported experimental folding rates in [1] correlate with the total local topological free energy of the proteins, with slower folding rates associated to higher total local topological/geometrical free energy.

The paper is organized as follows: Section 2 describes the topological and geometrical functions for characterizing 3-dimensional conformations used in this paper. Section 3 describes our results. Finally, in Section 4, we summarize the findings of our analysis.

2. The local topological/geometrical free energy of proteins

In Section 2.1 we give the definition of the mathematical tools that we will use in this manuscript. In Section 2.2 we introduce the novel definition of a local topological free energy of the protein backbone.

2.1. Measures of topological/geometrical complexity

We represent proteins by their CA atoms, as linear polygonal curves in space. A measure of conformational complexity of curves in 3-space is the Gauss linking integral. When applied to one curve, this integral is called the Writhe of a curve:

Definition 2.1. *(Writhe). For a curve l with arc-length parameterization $\gamma(t)$, the Writhe, Wr, is the double integral over l [51]:*

$$Wr(l) = \frac{1}{2\pi} \int_{[0,1]^*} \int_{[0,1]^*} \frac{(\gamma'(t) \times \gamma'(s)) \cdot (\gamma(t) - \gamma(s))}{||\gamma(t) - \gamma(s)||^3} dtds. \qquad (1)$$

where γ' denotes the derivative of γ and where the integrals run over all $s, t \in [0, 1]$, such that $s \neq t$.

The Writhe measures the average algebraic sum of crossings of the projection of the curve with itself over all possible projection directions. It is a measure of the number of times a chain winds around itself and can have both positive and negative values.

The total Torsion of the chain, describes how much the chain deviates from being planar and is defined as:

Definition 2.2. *The Torsion of a curve l with arc-length parameterization $\gamma(t)$ is the integral over l:*

$$T(l) = \frac{1}{2\pi} \int_{[0,1]} \frac{(\gamma'(t) \times \gamma''(t)) \cdot \gamma'''(t)}{||\gamma'(t) \times \gamma''(t)||^2} dt. \tag{2}$$

where $\gamma', \gamma'', \gamma'''$ denote the first, second and third derivatives of γ.

The Writhe and the Torsion have successfully been applied to study entanglement in biopolymers and proteins in particular [52, 45, 42, 43, 44, 53].

For a polygonal curve l of n edges, the Writhe can be expressed as

$$Wr(l) = \sum_{1 \leq i < j \leq n} \frac{1}{2\pi} \int_{[0,1]} \int_{[0,1]} \frac{(\gamma_i'(t), \gamma_j'(s), \gamma_i(t) - \gamma_j(s))}{||\gamma_i(t) - \gamma_j(s)||^3} dt ds \tag{3}$$

where γ_i (resp. γ_j) denotes the arc-length parametrization of the ith edge (resp. jth edge) and where the sum is taken over all pairs of edges i, j such that $|i - j| > 1$. Each summand in Eq. 3 is then twice the Gauss linking integral of two straight segments. The Gauss linking integral of two straight segments has a closed form that avoids numerical integration and can be computed with no approximation as in [54].

The Torsion of a polygonal curve also has a finite form, as explained in [54]. Namely, for a polygonal curve, the Torsion is equal to the normalized sum of dihedral angles:

$$T(l) = \frac{1}{2\pi} \sum_{2 \leq i \leq n-1} \phi(i) \tag{4}$$

where $\phi(i)$ denotes the dihedral angle centered at the ith edge.

An important property of the Gauss linking integral and the Torsion which makes them useful in practice is that they can be applied to polygonal curves of any length to characterize 3-dimensional conformations at different length scales. In this work, we use the Writhe and the Torsion to characterize

local 3-dimensional conformations of a protein at the length scale of 4 amino acids, we call this the *local Writhe* and the *local Torsion*, respectively.

Definition 2.3. *We define the local Writhe of an amino acid (resp. local Torsion), represented by the CA atom i to be the Writhe (resp. Torsion) of the protein backbone connecting the CA atoms $i, i+1, i+2, i+3$.*

Figure 1 shows examples of the Writhe and Torsion values when applied globally to the entire protein or locally, to 4 consecutive CA atoms of the protein.

We note that in the following we will interchange the name "CA atom" with "amino acid", since there is a unique CA atom in each amino acid.

The local Writhe of an amino acid is thus the Writhe of a polygonal curve of 3 edges. The local Writhe of 3 edges is in practice the Gauss linking integral between the first and third edge (because consecutive edges have zero linking number). The latter is equal to the geometric probability that the two straight segments cross in any projection direction (divided by 2) [54]. The local Writhe thus, can take values $-1 \leq Wr \leq 1$. The local Writhe is a measure of the local orientation of a polygonal curve and a measure of its compactness. For example a tight right-handed turn (resp. left-handed) may have a positive (resp. negative) Writhe value close to 1 (resp. -1), while a relatively straight segment will have a value close to 0. The Torsion of 3 segments is the signed dihedral angle of the 3 segments divided by 2π and thus takes the values $-0.5 \leq T \leq 0.5$. The Torsion is 0 for a planar segment and increases to ± 0.5 as the segment deviates from being planar. We note that for a fixed dihedral angle, we have an infinite possibility of positions of the third edge relative to the first edge, which can contribute a different local Writhe. Thus, for the same value of local Torsion, we can have infinitely many different conformations with different values of local Writhe. For example, it is possible to have low Writhe and high Torsion, and vice versa (see Figure 2).

2.2. Topological/Geometrical free energy

To assign a local topological/geometrical free energy along a protein backbone, we use a method inspired by the framework used in [55] for identifying exotic geometries of hydrogen bonds derived through Density Functional Theory (DFT) calculations. We first derive the distributions of the local Writhe and local Torsion in the ensemble of folded proteins. Then for each

local conformation of a given protein we compare its local Writhe (resp. Torsion) value to those of the ensemble and a free energy is assigned to it based on the population of that value in the ensemble. We can do the same for the global topology/geometry of the entire protein.

We compute the distribution of a topological parameter in the folded state ensemble. To do this in practice, we use a subset of the structures provided in the PDB which are related to one another by no more than some fixed percentage sequence identity (culled subset). Namely, we use the dataset of unbiased, high-quality 3-dimensional structures with less than 60% homology identity from [56].

Definition 2.4. *Let d_{Wr} denote the density (ie. the number of occurrences) of Wr in the folded ensemble. Let m_{Wr} denote the maximum occurrence value for Wr. To any value p of Wr, we associate a normalized quantity, which we will call, topological free energy:*

$$\Pi(p) = \ln[d(m)/d(p)] \tag{5}$$

The same definition can be applied for T. We denote Π_{Wr}, Π_T the topological free energy in Writhe and Torsion, respectively.

We note that the above definition of topological free energy can be applied to different lengths of the protein, by measuring Wr for n consecutive amino acids at a time. In this manuscript, we will focus at $n = 4$ which is the smallest possible n that can be used to define Writhe and Torsion.

Definition 2.5. *We will call the topological free energy of conformations of 4 consecutive amino acids, the local topological free energy.*

Definition 2.6. *We will say that an amino acid has a high local topological free energy in Wr (or is rare in Wr) if it is the first amino acid in a local conformation with value $Wr = p$ is such that $\Pi(p) \geq c$, where c is a threshold corresponding to the 95^{th} percentile of Π-values across the set of folded proteins.*

The same definition applies for the Torsion. We stress that we call rare an amino acid that is at the beginning of a rare conformation. However, a rare conformation is composed by 4 consecutive amino acids. We will say that an amino acid belongs to a rare conformation when it is one of these four amino acids.

We note that the free energy that we defined captures topological effects in long chains. For short sequences of 4 amino acids, the topology is trivial, but the geometry is not. Nevertheless, since we use the Writhe, defined through the Gauss linking integral, a conventional tool in topology, we will use the term topological free energy at all length scales, short as 4 amino acids.

Using the local topological free energy at each local conformation of a protein, we can assign a local topological free energy to the entire protein as follows:

Definition 2.7. *The total local topological free energy of a protein is the sum of all the local topological free energies of each local conformation in the protein.*

The calculation of the total local topological free energy of a protein follows a sliding window approach, where the local topological free energy of each local conformation of 4 consecutive CA atoms is added.

3. Results

In Section 3.1 we present our results on the local topology/geometry of the culled PDB ensemble dataset of unbiased, high-quality 3-dimensional structures with less than 60% homology identity from [56] (a total of 13,192 proteins). In Section 3.2 we examine local conformations of high topological/geometrical free energy. In Section 3.3 we analyze a set of 2-state singe domain proteins to examine the relation between the total local topological free energy along the protein backbone and the experimentally observed folding rate of the protein.

3.1. Local topology in the PDB

In this section we present the analysis of the local topology/geometry of the sample of the PDB proteins at 4 consecutive amino acids at a time along the entire backbone.

Figures 3A and B show the local Writhe and local Torsion distributions in the PDB culled ensemble, respectively. Note that the Writhe and the Torsion of a random polygonal curve of 3 edges follow normal distributions centered at zero [57]. The distributions of the local Writhe and local Torsion of proteins are clearly not those of random polygonal curves. The local

Writhe and local Torsion show one local maximum at positive values and one at negative. This suggests a well defined pattern in the local conformation of folded proteins, which may be due to the secondary structure elements and other characteristics of the amino acids. A peak at a positive Writhe or Torsion value suggests presence of right-handed local conformations. A peak at a negative Writhe or Torsion value suggests presence of left-handed local conformations. This could be a manifestation of the secondary structure of the proteins analyzed: Namely, 98% of the proteins in our sample contain at least one helix which contribute positive Writhe values and 91% contain at least one beta sheet, and β-strands may contribute small negative Writhe values [45]. The local Writhe values are concentrated between -0.2 and 0.2. The local Writhe maxima occur at approximately -0.01 and 0.8. Interestingly, the local Torsion distribution seems to be almost entirely concentrated in two values, a positive (0.25) and a negative (-0.25). Examples of local conformations at the peaks of the distribution of local Torsion are shown in Figure 4. This strong pattern in the dihedral angles between CA atoms may be expected due to the strong patterns observed in the ϕ, ψ and ω dihedral angles in proteins [2]. We see that conformations with low values of Writhe in absolute value can have high values of Torsion and vice-versa (see Figure 2 and also Figure 14 in Appendix 8). Both distributions of local Writhe and Torsion have more pronounced peaks than the global Writhe and Torsion distributions discussed in Appendix 7.

Figures 3C and D show the Π_{Wr}-values and Π_T-values as a function of Wr and T, respectively. We see that most Π values are less than 2, indicative of low topological free energy. Note that conformations with low Π_{Wr} values can have high Π_T values and vice-versa (see Figure 5).

3.2. High local topological free energy conformations in proteins

In this Section we will focus on those conformations outside the 95^{th} percentile of the distributions, which correspond to high Π values (values greater than 95% of the distribution of Π), and that we associate with high local topological/geometrical free energy. For simplicity, we will also refer to them as rare conformations. The complement of the 95^{th} percentile of the Π_{Wr}-value distribution in the PDB corresponds approximately to absolute Writhe values greater than 0.1. The complement of the 95^{th} percentile of the Π_T-value distribution in the PDB corresponds approximately to absolute Torsion values greater than 0.3 or smaller than 0.1. Thus, rare local Writhe values correspond to high Writhe in absolute value, while rare local Torsion

values may be values close to 0 in absolute value. This suggests that rare conformations in Writhe may not necessarily be rare in Torsion and vice-versa (see Figure 5). In general, the high local Writhe values could correspond to tight right-handed turns and the low local Torsion values could correspond to almost planar conformations.

In Section 3.2.1 and Section 3.2.2, respectively, we examine if high local topological free energy conformations are related to secondary structure elements and/or specific amino acid types.

3.2.1. High local topological free energy conformations and secondary structure in the PDB

To better understand the meaning of values of local Writhe and Torsion at the complement of the 95^{th} percentile of the Π-value distribution, we examine the correlation between these values and secondary structure elements.

Figure 6 shows the distribution of secondary structure elements in the protein sample and the distribution of the first, second, third and fourth amino acid in rare local conformations in Writhe in secondary structure elements. The distribution of the rare local conformations in Torsion is shown in Appendix 8. We see that 42% of the rare conformations in both local Writhe are in helices, 37% in coils and 21% in β sheets, which are very similar with the percentages of helices, coils and β sheets in the protein sample. Similar results are found for Torsion (see Appendix 8). Therefore, high local topological energy conformations are independent of secondary structure. We stress that even if the locations of rare conformations in Writhe and Torsion are similarly distributed across secondary structures, they are not pointing to the same amino acids (see Figure 4)

3.2.2. High local topological free energy conformations and amino acid type in the PDB

Amino acids have preferred dihedral angle distributions, specific sizes and other amino acid type dependent physical properties. It is natural therefore to examine whether there is a correlation between an amino acid being part of a rare conformation and its amino acid type.

Figure 7 shows the frequency of each amino acid in the PDB culled dataset versus the frequency each amino acid appears as part of a high local topological free energy conformation in Writhe in the 1st, 2nd, 3rd or 4th position. The same is shown for Torsion in Appendix 8. Overall, we see that the frequency by which an amino acid occurs in a rare conformation is the same

as the frequency by which it appears in the PDB culled dataset. This suggests that high local topological free energy configurations are independent of their sequence. Exceptions might be Phenylalanine and Histidine in both local Writhe and Torsion. Phenylalanine appears to be favoring rare local conformations while Histidine is not favored in rare local conformations. To quantify this we examined the absolute difference between the presence of an amino acid in a rare conformation and the frequency of the amino acid in the culled data set in general in Figure 8. This shows that Phenylalanine appears to be favoring rare local conformations by 3% while Histidine is not favored in rare local conformations by 2%, independently of the location within the local conformation.

We next examine the handedness of the high local topological free energy configurations led by each amino acid. Note that the distribution of Writhe in Figure 3A points to a higher number of positive local Writhe conformations. However, it does not exclude the possibility that some amino acid types are involved in conformations that create negative Writhe values disproportionally. As a proxy for handedness we simply use the sign of the Writhe, where positive sign indicates right-handed, while negative sign indicates left-handed. We find that the percentage of positive Writhe values for each amino acid fall within the range of 56-65% (with the exception of Methionine which is 69% positive in local Writhe and cysteine which is 49% positive in local Writhe). Similar results hold for Torsion, ie. positive Torsion values for each amino acid fall within the range of 51-68%. These results suggest a small but consistent preference for positive local Writhe and Torsion values for high local topological free energy conformations, representative of right-handed conformations.

We also examine the average absolute local Writhe and Torsion for each amino acid in a high local topological free energy conformation, shown in Figure 9 (Left). The local Writhe varies between 0.009762 and 0.030401158 and the local Torsion varies from 0.027668472 to 0.06821071. The outlier, Cysteine, has a Torsion value between this range, but has a Writhe of 0.009, indicative of more extended conformations.

Figure 9 (Right) shows the average local topological free energy for local Writhe and Torsion for each amino acid when involved in a high topological free energy configuration. Our results show that Asparagine, Glutamic acid, Lysine, Histidine and Methionine have on average low Π_T values.

3.3. Local topological free energy and protein folding rates

In the previous paragraph we found that the frequency of a specific amino acid type in a high local topological free energy conformation was on average the same as its frequency in the protein sample, suggesting no significant amino acid preference for being in a high local topological free energy conformation, with the possible exception of phyenylalanine. We may thus infer that the rare local conformations are not related to the local protein sequence. In this section we will examine how the local conformation of proteins may be related to protein folding kinetics. Our hypothesis is that unusual local topological/goemetrical properties in the PDB structures indicate rare local topologies/geometries in the unfolded state ensemble of proteins. This suggests that proteins need to search more in the unfolded state ensemble for such rare conformations and overcome energy barriers.

We analyze the native states of a set of simple, single domain, non-disulfide-bonded proteins that have been reported to fold in a concerted, all-or-none, two-state fashion, whose experimental folding rates in water were obtained in [1]. In [45] it was shown that the logarithm of the experimental folding rate decreases with decreasing global Writhe and Torsion of the protein backbone. We point out that other parameters in the literature are known to correlate with folding rates. The number of sequence-distant contacts is the simplest parameter that shows the best correlation to date [23]. This parameter may be a proxy to some other more physical aspect related to the 3-dimensional conformation of proteins that could provide understanding to mechanisms of protein folding. Several parameters based on the global topology of proteins have been used and have shown a strong correlation with protein folding rates [58, 42, 43, 44]. All these efforts focus on either the entire protein or on large parts of proteins (concatenated loops). In this Section we focus at the smallest lengthscale possible for analysis of Writhe and Torsion and show that topological/geometrical parameters at the length scale of 4 consecutive CA atoms also correlate with protein folding rates.

Figure 10 shows the logarithm of the experimental folding rate versus the normalized total local topological free energy in Torsion (total sum of Π_T values along the backbone divided by the length of the protein). Our results show that the folding rate decreases with increasing total local topological free energy in Torsion along the entire backbone (with $R^2 = 0.38$).

We also examine the logarithm of the experimental folding rate versus the number of rare (high local topological free energy) amino acids a protein has in Figure 11. We find that the folding rate decreases weakly with increasing

number of high local topological free energy in Writhe local conformations, with Spearman coefficient $\tau = -0.274$ and Kendall coefficient $\tau = -0.17$. This correlation supports the hypothesis that conformations of proteins that are rare in the sample of native states may also be rare in the unfolded state ensemble of proteins and thus related to global energy barriers in the protein.

3.4. Local topological free energy and ϕ values

The effect of amino acid sequence on the tertiary structure of a protein is studied experimentally though chemical scannings [59, 60]. Chemical scanning consists in substituting amino acids along a protein backbone (for a protein known to fold) and exploring its folding after the substitution to its native state. This comparison is done using a quantity called ϕ value, which reflects how much the mutated amino acid is involved in the key contacts established during the folding process. A ϕ value that is equal to 0, suggests that the mutation has no effect on the structure and that the region surrounding the mutation is unfolded in the transition state. A ϕ value that is equal to 1 means that the local structure around the mutation closely resembles the structure of the native state. Therefore, the ϕ value represents how a specific amino acid in the sequence has an effect in the global structure of the protein, with $\phi = 1$ suggesting it is not important at the rate limiting step. Here we focus on exploring whether there is a relation between Π-values and experimentally observed ϕ values for a set of well studied proteins: barnase, FK506 binding protein (FKBP12), chrymotypsin inhibitor (CI2) and src SH3 domain (SH3).

We calculate the Π_{Wr} values along the backbone of the proteins and compare them to the experimentally reported ϕ values along their backbone [59, 60]. Our results, shown in Figure 12, display an overall decrease in ϕ as Π_{Wr} increases. The Kendall coefficient is $\tau = -0.1286$ and the Spearman coefficient is $\tau = -0.1968$. This weak decreasing trend is also supported by the trend of the moving average. This further supports our findings showing that high Π_{Wr} values are associated with high free energy conformations which are sensitive in the rate-limiting step of the folding process.

4. Discussion

We used the local topology/geometry of protein structures alone to associate a novel local topological/geometrical free energy to the protein backbone amino acids. By using a culled protein data set from the PDB we

derived the distributions of the local Writhe and local Torsion values. Using these, we computed a local topological free energy for each protein. For the data set we studied in this manuscript (a data set with less than 60% homology identity), our results showed that high local topological free energy conformations are independent of secondary structure and sequence. Interestingly, our results suggest that these high local topological free energy conformations are related to the global conformations of the proteins. Namely, by focusing on a well studied set of 2-state proteins, we found that the logarithm of experimental folding rates decreases with the total local topological free energy in Torsion along the protein backbone. We also found a weak decrease of the logarithm of the folding rate with the number of high local topological free energy conformations in Writhe per protein. Our results also showed that ϕ values decrease with increasing local topological free energy in Writhe. These results point to the fact that the local topological free energy in Torsion and in Writhe capture different information about proteins. Namely, the total local topological free energy in Torsion better captures a free energy for the entire protein backbone, while the local topological free energy in Writhe can be better used to identify local conformations which capture important features of the entire protein conformation.

Our results suggest that the local topological free energy in Writhe and Torsion captures characteristics of the 3-dimensional conformation of proteins that can be helpful in understanding protein folding and function. For example, the local topological free energy of proteins can be applied to families of proteins to detect similarities and possible selection mechanisms based on tertiary structure. The local topological free energy of proteins can also be applied to compare how mutations in proteins known to stabilize or destabilize protein structure, affect their 3-dimensional conformation. Moreover, the local topological free energy in Writhe and Torsion can be easily calculated for protein trajectories obtained through simulations to study the topological landscape of unfolded proteins and to study protein folding. Eventually, the local topological free energy could be a useful parameter in the search of a predictive model of protein folding.

5. Funding

We thank the support of NSF REU 1852042 and internal support of the University of Tennessee at Chattanooga. We thank the support of NSF DMS 1913180.

6. Acknowledgments

We thank Dr. Bobby Sumpter and Dr. Cristian Micheletti for very helpful discussions.

7. The global topology of proteins in the PDB

In this section we analyze the three-dimensional configuration of the entire protein backbone of the proteins in PDB culled protein dataset as a whole and not at the local length scale.

The normalized values of the Writhe by the length of the proteins is shown in Figure 13(A). We see a bimodal distribution with a peak at positive values of Writhe and a smaller one at negative Writhe values. The distribution is overall skewed to the positive values of Writhe, ranging from -0.2 to 0.6. The skewness of the distribution to positive Writhe values indicates a preference for right-handed conformations in the proteins. This is in agreement with results in [58] where it was shown that the logarithm of the experimental folding rate decreases when the Writhe of the native state becomes negative. Note that the Writhe and Torsion of random polygonal curves follows a normal distribution centered at the origin [57, 61]. Clearly, proteins cannot be modeled by random polygonal curves. The skewness of the distribution could be a manifestation of the secondary structure of the proteins analyzed or of the local Writhe values, discussed in Section 3.1. However, we note that even though the secondary structure element Writhe values add to the global Writhe of the protein, there can be helical proteins with negative Writhe and proteins with no helices that have positive Writhe.

The normalized Torsion values are shown in Figure 13(B). We point out that the distributions of Writhe and Torsion are apparently different, which may be expected, since the two parameters capture different characteristics of the 3-dimensional conformation of the proteins. Similarly to the Writhe, the Torsion values may be affected by the secondary structure elements.

8. Supplementary Information

8.1. The local Writhe and local Torsion in the PDB

Figure 14 shows the values of local Writhe and local Torsion for each local conformation in the protein sample. Comparing to the results shown

in Figure 3, we see that conformations of local Torsion at the peaks of the distribution can have values of Writhe that are away from the peaks of the distribution of local Writhe.

8.2. Local topological free energy in Torsion

Figure 15 shows the distribution of high local topological free energy in Torsion conformations in the culled PDB sample. Our results show a similar distribution as that of the distribution of secondary structures in the PDB, suggesting that high local topological free energy conformations are independent of secondary structure. We notice that the first and second amino acid in a rare conformation are 42% in helices, 20% in sheets and 38% in coils, while the third and fourth amino acid in a rare conformation are 41% in helices, 21% in sheets and 38% in coils. This may suggest that some of the rare conformations in Torsion may occur in the end of helices or at the beginning of sheets.

Figure 16 shows the frequency of each of each amino acid type in a rare conformation in Torsion versus its frequency in the PDB culled sample. We see an agreement between the two for all amino acid types except Phenylalanine, which seems to occur more often in a rare conformation in Torsion that predicted by its occurrence in the PDB sample. This is similar to what was observed for the rare conformations in Writhe in Section 3.2.2. To better quantify this result, Figure 17 shows the deviation of the frequency of an amino acid type in a rare conformation in Torsion from its frequency in the PDB culled sample. The results show that Phenylalanine is by 3% more frequent to appear in a rare conformation than its frequency in the PDB, while Histidine appears less frequently in a rare conformation than its frequency in the PDB sample by 2%.

References

[1] K. W. Plaxco, K. T. Simons, I. Ruczinski, D. Baker, Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics, Biochemistry 37 (2000) 11177–11183.

[2] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Molecular Biology of the Cell, New York: Garland Science, 2002.

[3] D. L. Stein, Protein states and proteinquakes, PNAS 82 (1985) 3670–72.

[4] C. Anfinsen, E. Haber, M. Sela, F. J. White, The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain, PNAS 47 (1961) 1309.

[5] C. Anfinsen, Principles that govern the folding of protein chains, Science 181 (1973) 223–30.

[6] K. Lindorff-Larsen, S. Piana, R. Dror, D. E. Shaw, How fast-folding proteins fold, Science 334 (2011) 517–20.

[7] A. N. Adhikari, K. F. Freed, T. R. Sosnick, Simplified protein models can rival all atom simulations in predicting folding pathways and structure, Phys. Rev. Lett. 111 (2013) 028103.

[8] C. Levinthal, Are there pathways for protein folding?, J. Chem. Phys. 65 (1968) 1968.

[9] J. Shea, C. L. Brooks, From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding, Annu. Rev. Phys. Chem. 52 (2001) 499–535.

[10] J. N. Onuchic, Z. Luthey-Schulten, P. G. Wolynes, Theory of protein folding: the energy landscape perspetive, Annu. Rev. Phys. Chem. 48 (1997) 545–600.

[11] M. Oliveberg, P. G. Wolynes, The experimental survey of protein-folding energy landscapes, Q Rev. Biophys. 38 (2005) 245–288.

[12] D. N. Ivankov, A. V. Finkelstein, Prediction of protein folding rates from the amino acid sequence-predicted secondary structure, PNAS 101 (2004) 8942–8944.

[13] W. Englander, L. Mayne, The case for defined protein folding pathways, PNAS 114 (2017) 8253–8258.

[14] W. Englander, L. Mayne, The nature of protein folding pathways, PNAS 111 (2014) 15873–15880.

[15] W. Englander, L. Mayne, Z. Y. Kan, W. Hu, Protein folding-how and why: By hydrogen exchange, fragment separation, and mass spectrometry, Annu Rev Biophys 45 (2016) 135–152.

[16] W. Hu, Z. Y. Kan, L. Mayne, S. W. Englander, Cytochrome c folds through foldon-dependent native-like intermediates in an ordered pathway, PNAS 113 (2016) 3809–3814.

[17] Z. Guo, D. Thirumalai, Kinetics of protein folding: Nucleation mechanism, time scales, and pathways, Biopolymers 36 (1995) 745–57.

[18] J. N. Onuchic, N. D. Socci, Z. Luthey-Schulten, P. G. Wolynes, Protein folding funnels: the nature of the transition state ensemble, Fold. Des. 1 (1996) 441–50.

[19] D. U. Ferreiro, E. A. Komives, P. G. Wolynes, Frustration in biomolecules, Q. Rev. Biophys. 47 (2014) 285–363.

[20] E. Shakhnovich, V. Abkevich, O. Ptitsyn, Conserved residues and the mechanism of protein folding, Nature 379 (1996) 96–98.

[21] D. E. Makarov, K. W. Plaxco, The topomer search model: asimple, quantitative theory of two-state protein folding kinetics, Protein Science 12 (2003) 17–26.

[22] D. E. Makarov, C. A. Keller, K. W. Plaxco, H. Metiu, How the folding rate constant of simple-single domain proteins depends on number of native contacts, PNAS 99 (2002) 3535–3539.

[23] K. W. Plaxco, K. T. Simons, D. Baker, Contact order, transiton state placement and the refolding rates of single domain proteins, J. Mol. Biol. 277 (1998) 985–994.

[24] K. W. Plaxco, R. I. Larson, S., D. S. Riddle, E. C. Thayer, B. Buchwitz, A. R. Davidson, D. Baker, Evolutionary conservation in protein folding kinetics., J. Mol. Biol. 298 (2000) 303–312.

[25] M. Pouokam, B. Cruz, S. Burgess, M. Segal, M. Vazquez, J. Arsuaga, The Rabl configuration limits topological entanglement of chromosomes in budding yeast, Scientific Reports 9 (2019) 6795.

[26] E. J. Rawdon, J. C. Kern, M. Piatek, P. Plunkett, A. Stasiak, K. C. Millett, Effect of knotting on the shape of polymers, Macromolecules 41 (2008) 8281–8287.

[27] S. Trigueros, J. Arsuaga, M. E. Vazquez, D. W. Sumners, J. Roca, Novel display of knotted DNA molecules by two-dimensional gel electrophoresis, Nucleic Acids Research 29 (2001) e67.

[28] J. Arsuaga, M. Vazquez, S. Trigueros, D. W. Sumners, J. Roca, Knotting probability of DNA molecules confined in restricted volumes: DNA knotting in phage capsids, Proc. Natl. Acad. Sci. USA 99 (2002) 5373–5377.

[29] J. Arsuaga, M. Vazquez, P. McGuirk, S. Trigueros, D. W. Sumners, J. Roca, DNA knots reveal a chiral organization of DNA in phage capsids, Proc. Natl. Acad. Sci. (USA) 102 (2005) 9165–9169.

[30] J. Arsuaga, Y. Diao, T. Kaplan, M. Vazquez, The effects of density on the topological structure of the mitochondrial DNA from trypanosomes, Journal of Mathematical Biology 64 (2012) 1087–1108.

[31] X. Hua, B. Raghavan, D. Nguyen, J. Arsuaga, M. Vazquez, Random state transitions of knots: a first step towards modeling unknotting by type II topoisomerases, Topology and its applications 157 (2007) 1381–1397.

[32] R. Stolz, M. Yoshida, R. Brasher, M. Flanner, K. Ishihara, D. J. Sheratt, K. Shimokawa, M. Vazquez, Pathways of DNA unlinking: a story of stepwise simplification, Sci. Reports 7 (2017) 12420.

[33] D. W. Sumners, S. G. Whittington, Untangling DNA, Math Intelligencer 12 (1990) 71–80.

[34] D. Marenduzo, E. Orlandini, A. Stasiak, D. W. Sumners, L. Tubiana, C. Micheletti, DNA-DNA interactions in bacteriophage capsids are responsible for the observed DNA knotting, PNAS 106 (2009) 22269–22274.

[35] C. Micheletti, D. Marenduzzo, E. Orlandini, D. W. Sumners, Knotting of random ring polymers in confined spaces, J. Chem. Phys. 124 (2006) 64903.1–10.

[36] C. Micheletti, H. Orland, Efficient sampling of knotting-unknotting pathways for semiflexible gaussian chains, Polymers 9 (2017) 196.

[37] D. Buck, F. E., A topological characterization of knots and links arising from site-specific recombination, J. Phys. A.: Math. Theor. 40 (2007) 12377–12395.

[38] D. Buck, F. E., Predicting knot or catenane type of site-specific recombination products, J. Mol Biol. 374 (2007) 1186–1199.

[39] E. Flapan, A. He, H. Wong, Topological descriptions of protein folding, PNAS 116 (2019) 9360–9369.

[40] I. Darcy, J. Luecke, M. Vazquez, Tangle analysis of difference topology experiments: applications to a mu protein-DNA complex, Algebraic and Geometric Topology 9 (2009) 2247–2309.

[41] J. I. Sulkowska, E. J. Rawdon, K. C. Millett, J. N. Onuchic, A. Stasiak, Conservation of complex knotting and slpiknotting in patterns in proterins, PNAS 109 (2012) E1715.

[42] M. Baiesi, E. Orlandini, A. Trovato, F. Seno, Linking in domain-swapped protein dimers, Scientific Reports 6 (2016) 1–11.

[43] M. Baiesi, E. Orlandini, F. Seno, A. Trovato, Exploring the correlation between the folding rates of proteins and the entanglement of their native state, J. Phys. A: Math. Theor. 50 (2017) 504001.

[44] M. Baiesi, E. Orlandini, F. Seno, A. Trovato, Sequence and structural patterns detected in entangled proteins reveal the importance of co-translational folding, Scientific Reports 9 (2019) 1–12.

[45] E. Panagiotou, K. W. Plaxco, A topological study of protein folding kinetics, Topology and Geometry of Biopolymers, AMS Contemporary Mathematics Series 746 (2020) 223–233.

[46] K. Shimokawa, K. Ishihara, I. Grainge, D. Sherratt, M. Vazquez, Ftsk-dependent XerCD-dif recombination unlinks replication catenanes in a stepwise manner, PNAS 110 (2013) 20906–20911.

[47] W. Niemyska1, D. Dabrowski-Tumanski, M. Kadlof, E. Haglund, P. Sułkowski, J. I. Sulkowska, Complex lasso: new entangled motifs in proteins, Scientific Reports 6 (2016) 36895.

[48] M. Jamroz, W. Niemyska, E. J. Rawdon, A. Stasiak, K. C. Millett, P. Sulkowski, J. Sulkowska, Knotprot: a database of proteins with knots and slipknots, Nucleic Acids Res. 43 (2015) D306–14.

[49] P. Dabrowski-Tumanski, M. Piejko, S. Niewieczerzal, A. Stasiak, J. I. Sulkowska, Protein knotting by active threading of nascent polypeptide chain exiting from the ribosome exit channel, J. Phys. Chem. B. 122 (2018) 11616–11625.

[50] D. Goundaroulis, N. Gügümcu, S. Lambropoulou, J. Dorier, A. Stasiak, L. H. Kauffman, Topological methods for open-knotted protein chans using the concepts of knotoids and bonded knotoids, Polymers 9 (2017) 444.

[51] K. F. Gauss, Werke, Kgl. Gesellsch. Wiss. Göttingen, 1877.

[52] P. Rogen, B. Fain, Automatic classification of protein structure by using gauss integrals, Proc. Natl Acad. Sci 100 (2003) 119–24.

[53] F. Norbiato, F. Seno, A. trovato, M. Baiesi, Folding rate optimization promotes frustrated interactions in entangled protein structures, International Journal of Molecular Sciences 1 (2020) 213.

[54] T. Banchoff, Self-linking numbers of space polygons, Indiana Univ. Math. J. 25 (1976) 1171–1188.

[55] R. C. Penner, Backbone free energy estimator applied to viral glycoproteins, Journal of Computational Biology 27 (2020) 1–14.

[56] G. Wang, R. L. J. Dunbrack, Pisces: a protein sequence culling server, Bioinformatics 19 (2003) 1589–1591.

[57] E. Panagiotou, K. C. Millett, S. Lambropoulou, The linking number and the writhe of uniform random walks and polygons in confined space, J. Phys. A 43 (2010) 045208–30.

[58] E. Panagiotou, L. Kauffman, Knot polynomials of open and closed curves, (submitted) (2020).

[59] M. M. Gromiha, S. Selvaraj, Important amino acid properties for determining the transition state structures of two-state protein mutants, FEBS Letters 526 (2002) 129–134.

[60] V. Daggett, A. Li, A. R. Fersht, Combined molecular dynamics and -value analysis of structure-reactivity relationships in the transition state and unfolding pathway of Barnase: Structural basis of hammond and anti-hammond effects, J. Am. Chem. Soc. 120 (1998) 12740–12754.

[61] Y. Diao, C. Ernst, K. Hinson, U. Ziegler, The mean-squared writhe of alternating random knot diagrams, J. Phys. A: Math. Theor. 43 (2010) 495202.

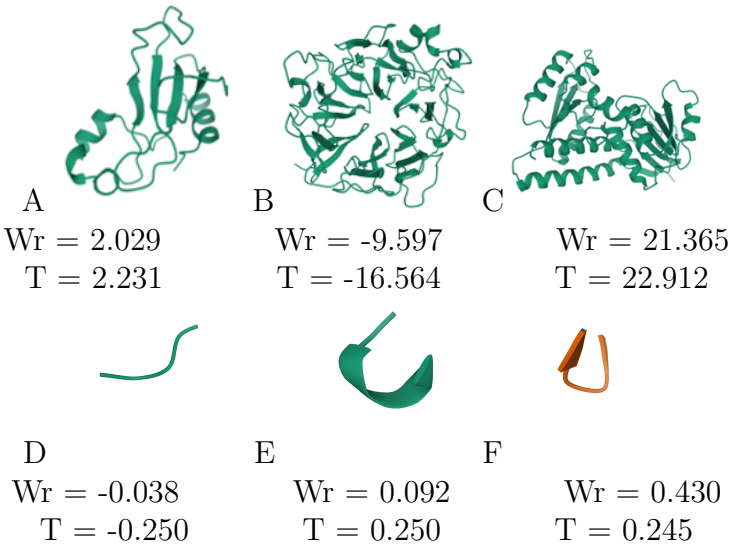
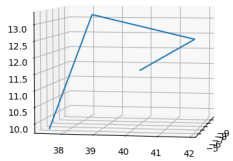
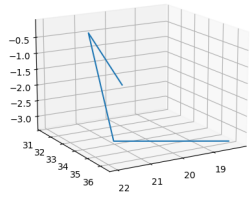


Figure 1: Examples of global and local Writhe and Torsion. (A) Global Writhe (Wr) and Torsion (T) of PDB: 1A2P (resp. (B) 1A12, (C) 1A4I). In these exmaples, Wr and T increase in absolute value as length and complexity of protein increases. (D) Local Writhe and Torsion values of PDB: 16PK amino acids 1-4 (E) 16PK amino acids 4-8 and (F) 1GK9 amino acids 92-96 shown.



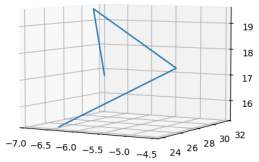
A

Wr = 0.1963, T= 0.2529



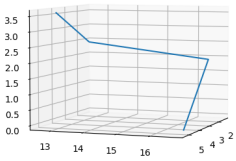
B

Wr = 0.0551, T= 0.2501



C

Wr = -0.0490, T=-0.2499



D

Wr = -0.0466, T=-0.1591

Figure 2: Examples of local conformations in the protein sample and their corresponding local Writhe and local Torsion values. A, B: Similar Torsion, different Writhe. C, D: Similar Writhe, different Torsion. We see that we can have same local Writhe and different local Torsion values and vice-versa.

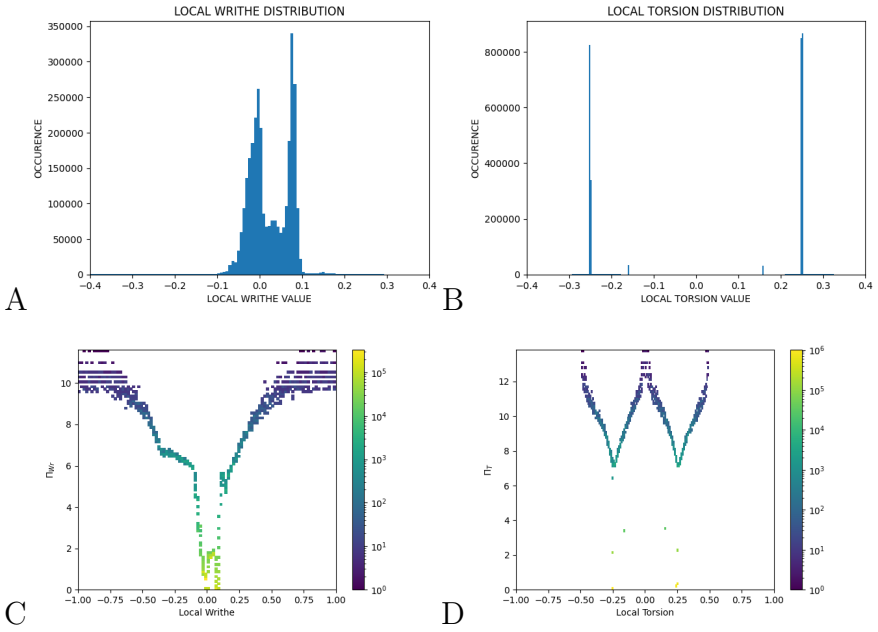


Figure 3: Distribution of the local topology in the PDB. (A) The local Writhe. (B) The local Torsion. Both distributions are bimodal but very different from each other, and from the global Writhe and Torsion distributions shown in the . (C, D) The local Π_{Wr} and Π_T values as a function of Wr and T, respectively. High density is indicated by yellow. Local conformations with $\Pi_{Wr} > 2.6$ correspond to the complement of the 95$^{\text{th}}$ percentile of the Writhe distribution (meaning that they are higher than 95% of the Π_{Wr} values) and local conformations with $\Pi_T > 3.5$ correspond to the complement of the 95$^{\text{th}}$ percentile of the Torsion distribution (meaning that they are higher than 95% of the Π_T values).

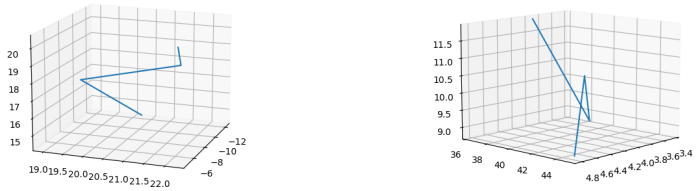
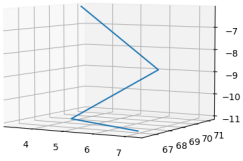
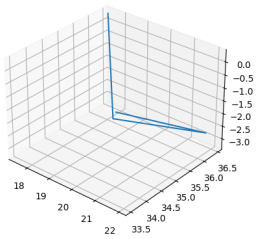


Figure 4: Examples of local conformations with Torsion values at the two peaks of the distribution shown in Figure 3B.



A

Wr =-0.0017, T=0.1591



B

Wr = -0.2726, T=-0.2523

Figure 5: Examples of local conformations in the protein sample with (A) high Π_T, low Π_{Wr}, (B) high Π_{Wr}, low Π_T.

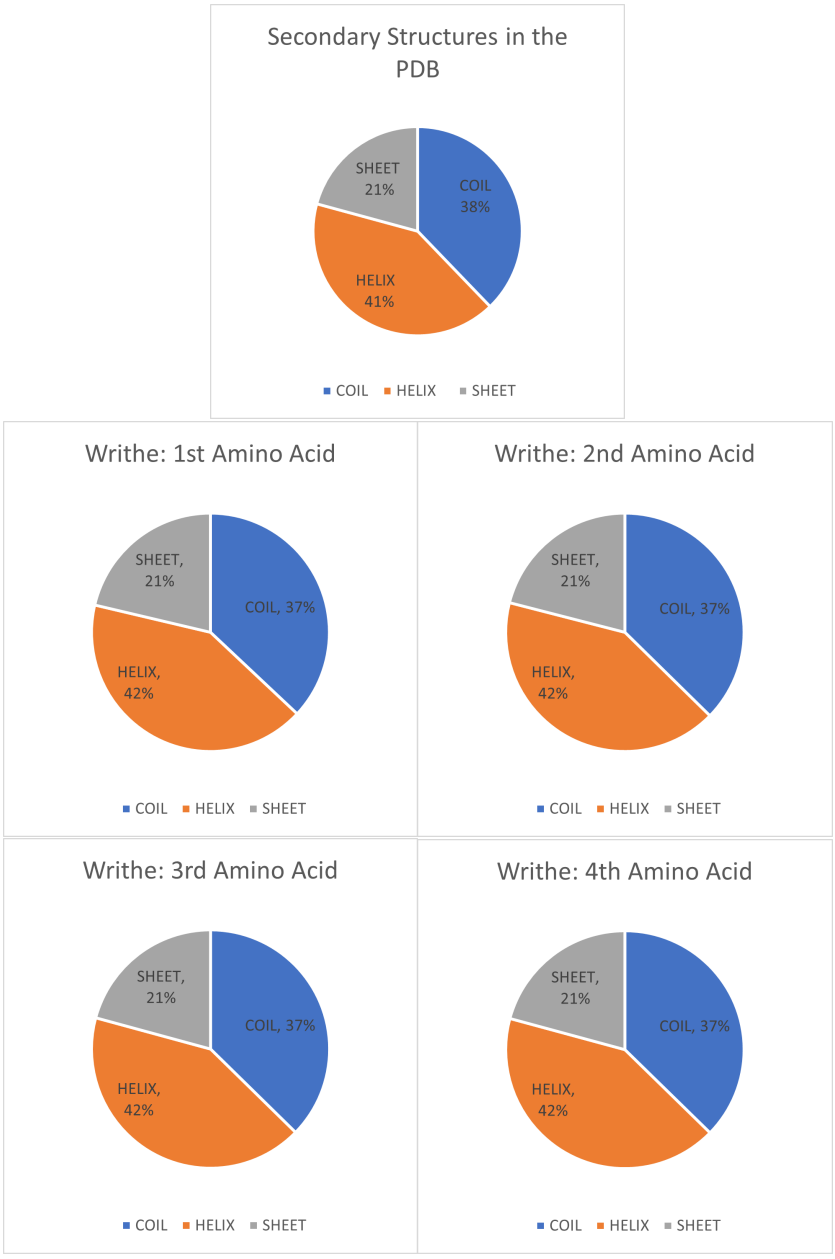


Figure 6: (Top) The distribution of secondary structures in the protein sample. (Bottom 4 Figures) The distribution in secondary structure elements of the first, second, third and fourth amino acid in high local topological/geometrical free energy configurations in Writhe in the PDB culled data set. We notice that the distributions of rare conformations in secondary structure elements are similar to the distribution of secondary structure elements in the PDB sample, indicating that the high local topological free energy conformations are independent of secondary structure. The Torsion distribution is similar (shown in Appendix 8).

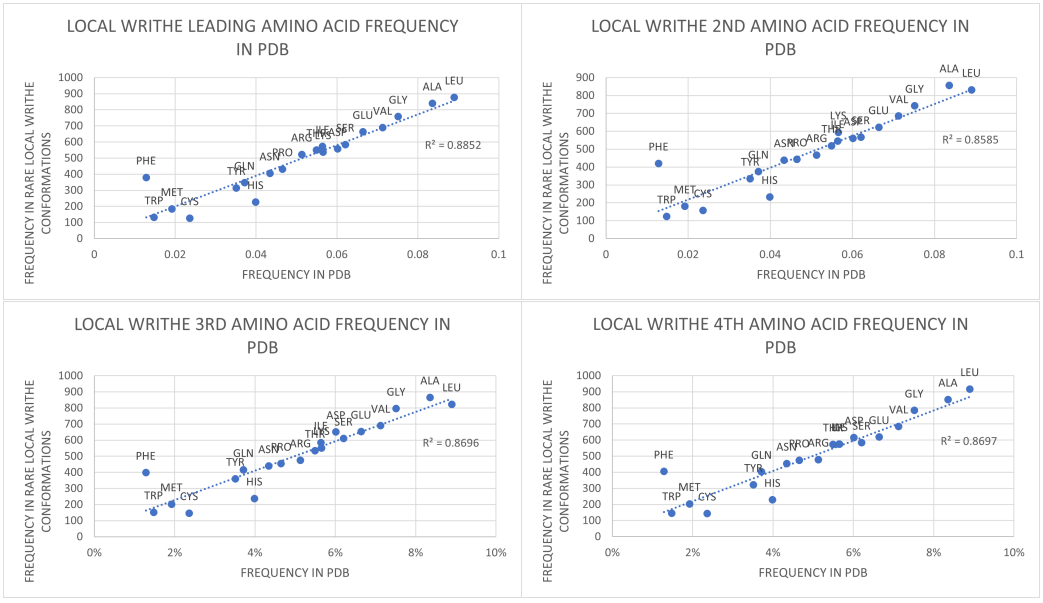


Figure 7: Frequency of amino acid types in the first, second, third and fourth amino acid, respectively, in high local topological free energy configurations in Writhe as a function of the frequency of amino acid types in the PDB in general. We find a linear fit with $R^2 = 0.8852$ for the first amino acid ($R^2 = 0.8585$, $R^2 = 0.8696$, $R^2 = 0.8697$ for the second, third and fourth, respectively) suggesting that high local topological free energy configurations in Writhe are not related to local amino acid sequence.

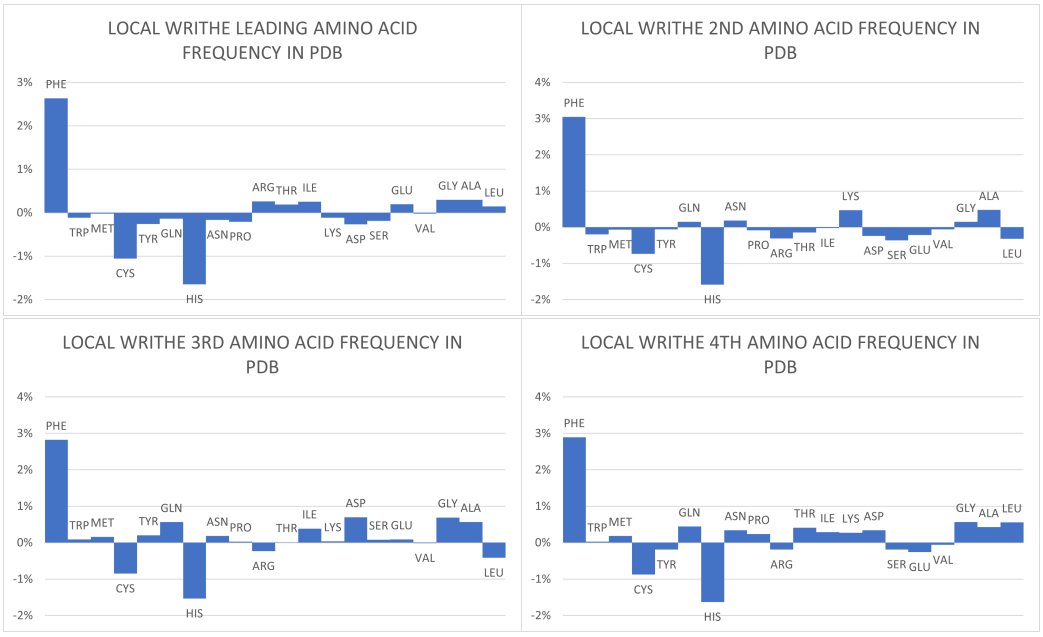


Figure 8: The difference of the frequency of an amino acid in a rare local conformation versus its frequency in the PDB as a percentage. This shows that Phenylalanine appears to be favoring rare local conformations by 3% while Histidine is not favored in rare local conformations by 2%, independently of the location withing the local conformation.

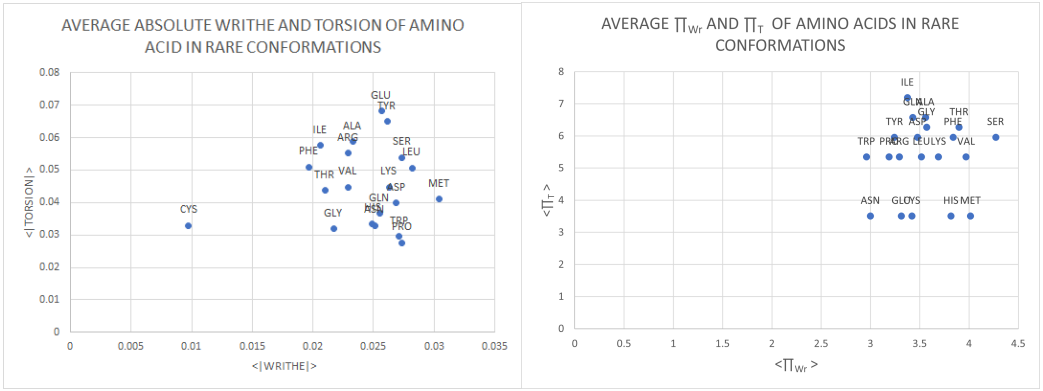


Figure 9: Left: The Average absolute Writhe values and Average absolute Torsion values of high local topological free energy conformations containing each amino acid type. Right: The Average Π_{Wr} values and Π_T values of high local topological free energy conformations containing each amino acid type.

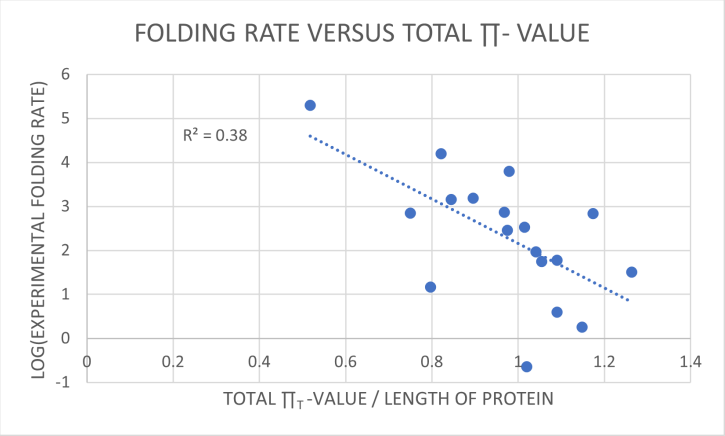


Figure 10: The logarithm of the experimentally observed folding rate of a set of 2-state proteins as a function of the normalized sum of local topological free energy in Torsion (the sum of Π_T-values along the protein backbone).

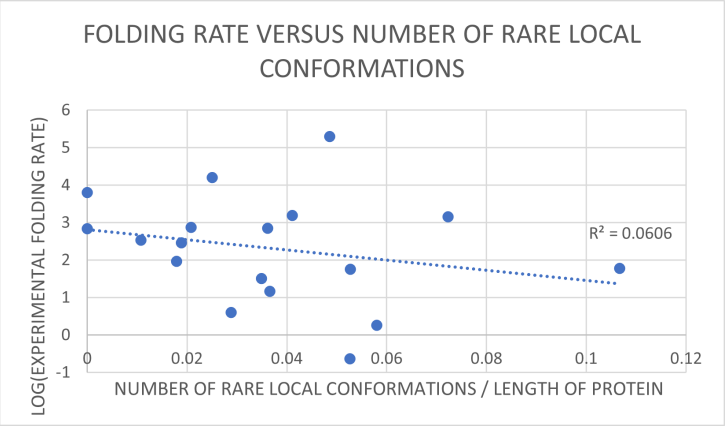


Figure 11: The logarithm of the experimentally observed folding rate of a set of 2-state proteins as a function of the normalized number of high local topological free energy conformations in Writhe, with Spearman coefficient $\tau = -0.274$ and Kendall coefficient $\tau = -0.17$.

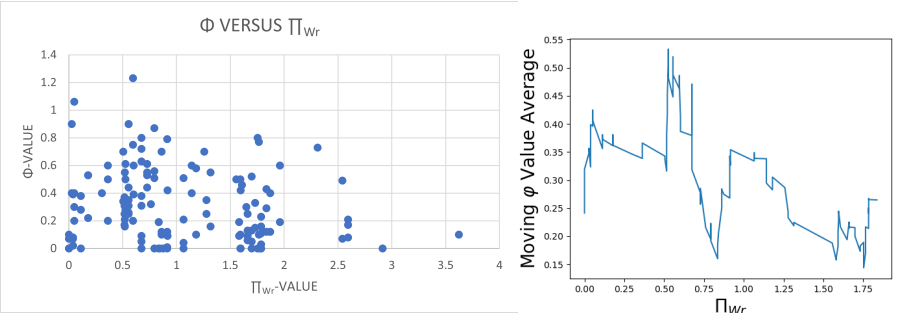


Figure 12: Left: ϕ-values versus the Π-values of proteins barnase (PDB ID: 1BRS), FKBP12 (PDB ID: 1FJK), CI2 (2CI2) and SH3 (PBD ID: 1SRL). Right: Moving average of ϕ values as a function of Π_{Wr} values.

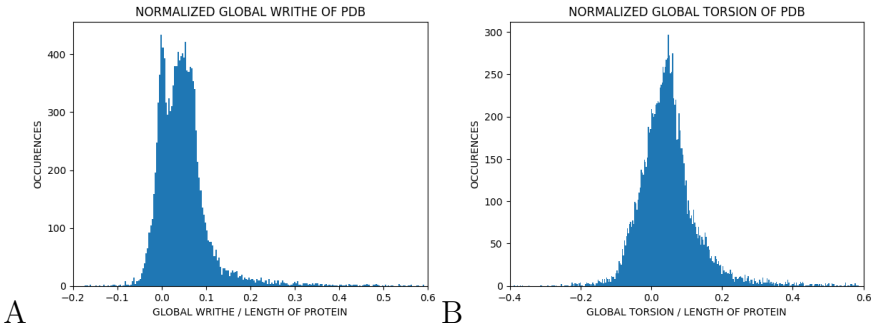


Figure 13: Distribution of the global topology/geometry of the PDB culled dataset.(A) Writhe of a protein normalized by the length of the protein. (B) Torsion of a protein normalized by the length of the protein. Both the normalized Writhe and normalized Torsion of the proteins in the PDB ensemble show a bimodal distribution with a peak at a positive and a negative value, skewed to the right. However, the two distributions are different, indicating that the two parameters capture different aspects of the protein conformation.

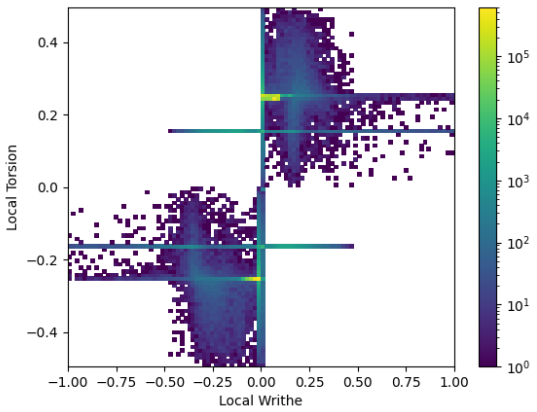


Figure 14: The distribution of local Writhe and local Torsion of each amino acid. There exist local conformations with high Writhe and low Torsion and vice-versa.

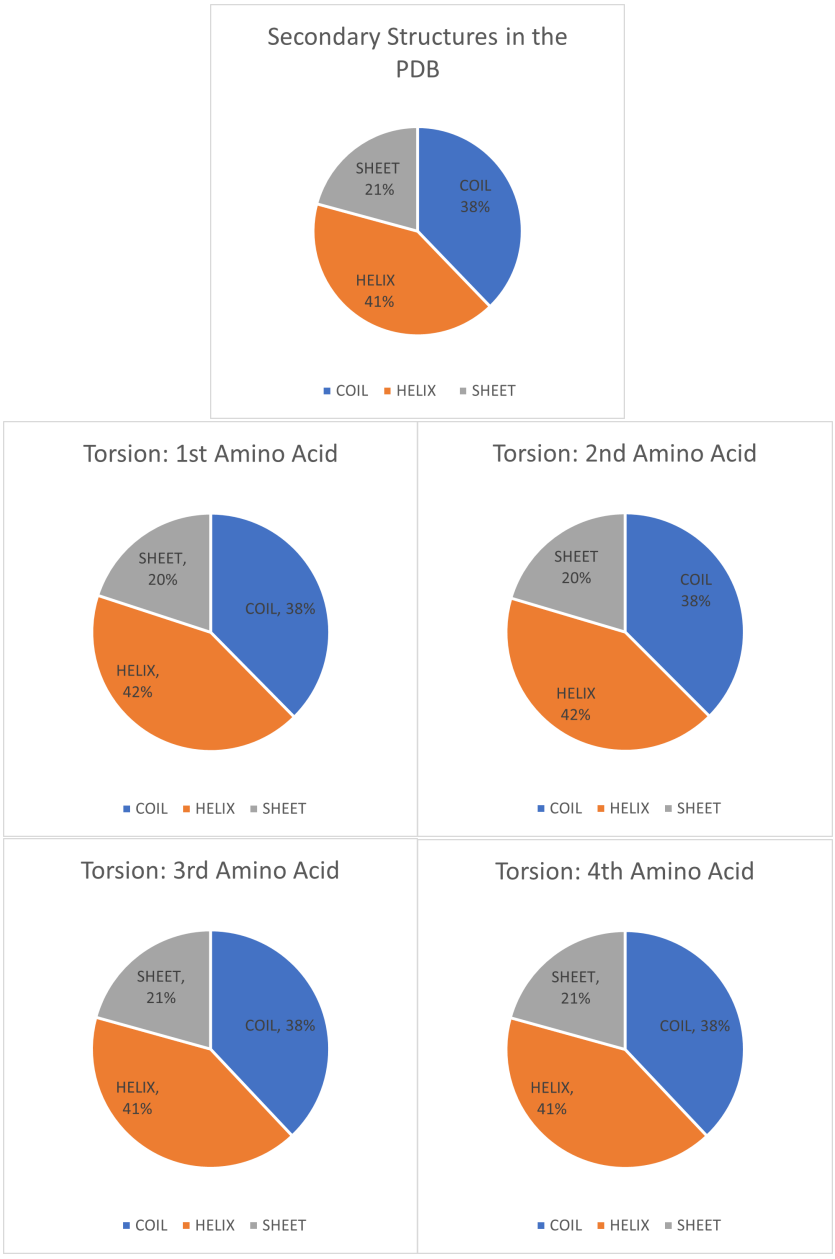


Figure 15: (Top) The distribution of secondary structures in the culled PDB ensemble. (Bottom 4 Figures) The distribution in secondary structure elements of the first, second, third and fourth amino acid in high local topological/geometrical free energy configurations in Torsion in the PDB culled data set. We notice that the distributions are similar, indicating that the high local topological free energy conformations are independent of secondary structure.

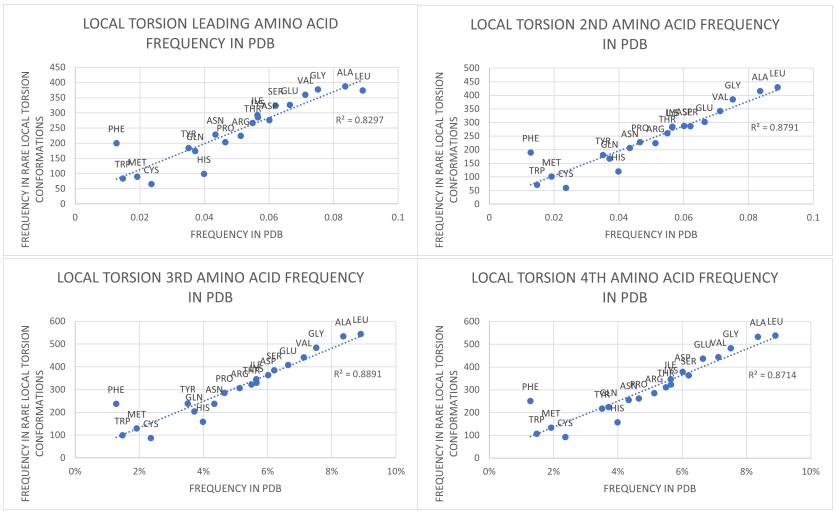


Figure 16: Frequency of amino acid types in the first, second, third and fourth amino acid, respectively, in high local topological free energy configurations in Torsion as a function of the frequency of amino acid types in the PDB in general. We find a linear fit with $R^2 = 0.8297$ (resp. $R^2 = 0.8791$, $R^2 = 0.8891$ and $R^2 = 0.8714$.) suggesting that high local topological free energy configurations in Torsion are not related to local amino acid sequence.

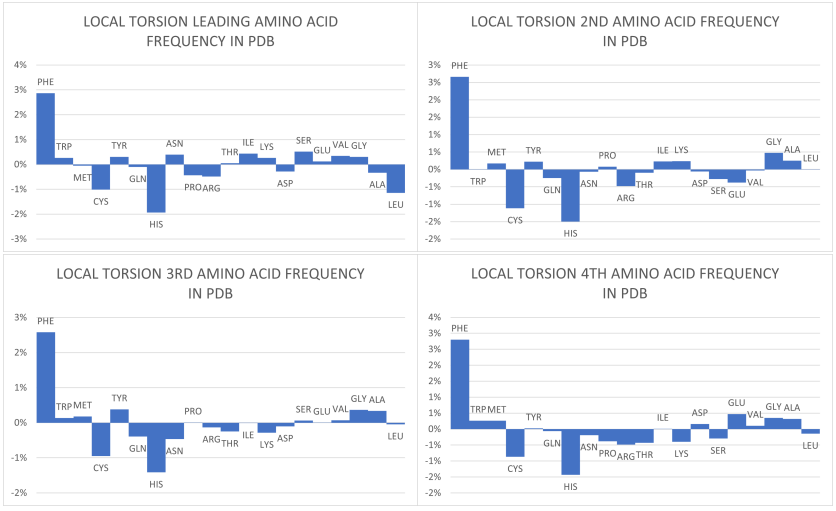


Figure 17: To quantify the degree to which the frequency of an amino acid in a rare conformation deviates from its frequency of appearance in the PDB in general. This shows that Phenylalanine appears to be favoring rare local conformations in Torsion by 3% while Histidine is not favored in rare local conformations by 2%, independently of the location withing the local conformation.