Using writhe to provide limits on entanglement for protein backbones and identify shared helical domains

Arron Bale

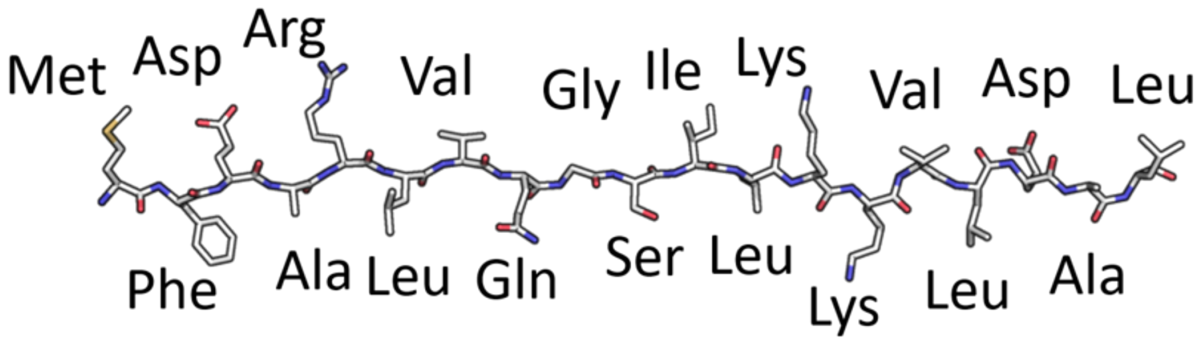
January 8, 2024

Durham University (MoSMed CDT)

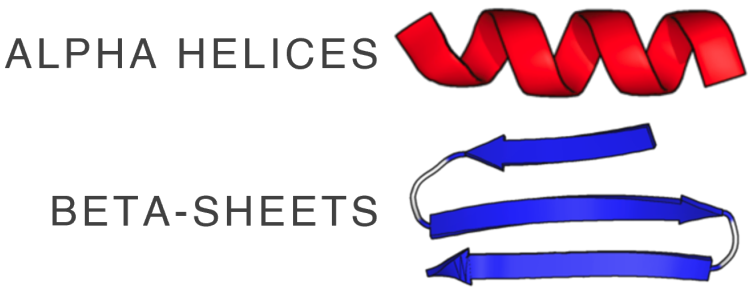
Introduction

It’s late, let’s start easy

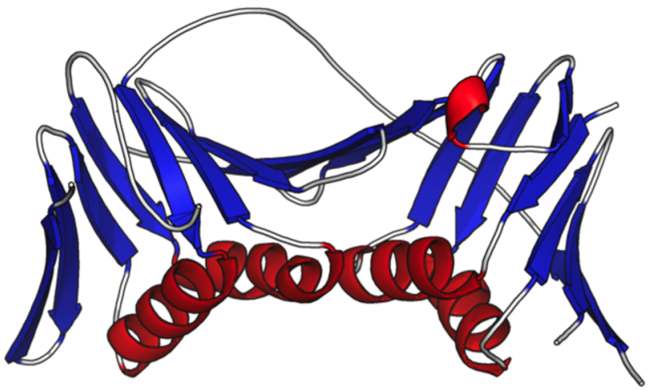
Primary Structure



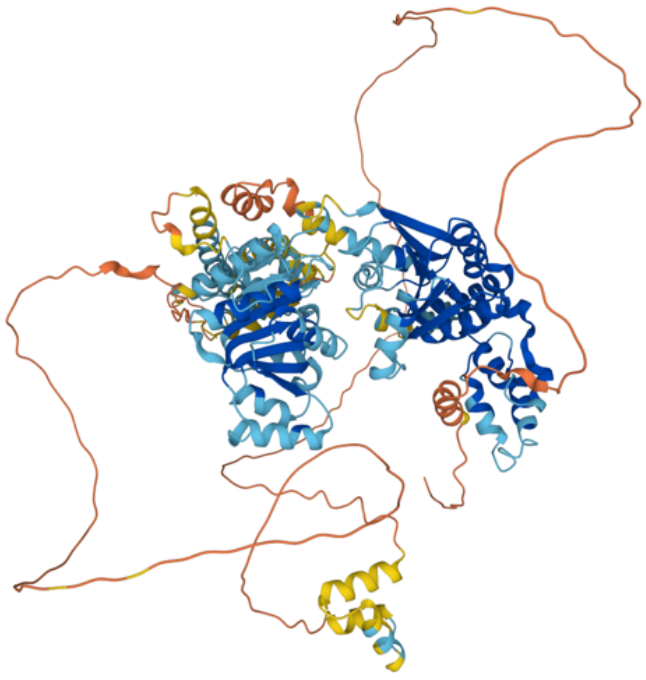
Secondary Structure



Tertiary Structure



What to do about AlphaFold?

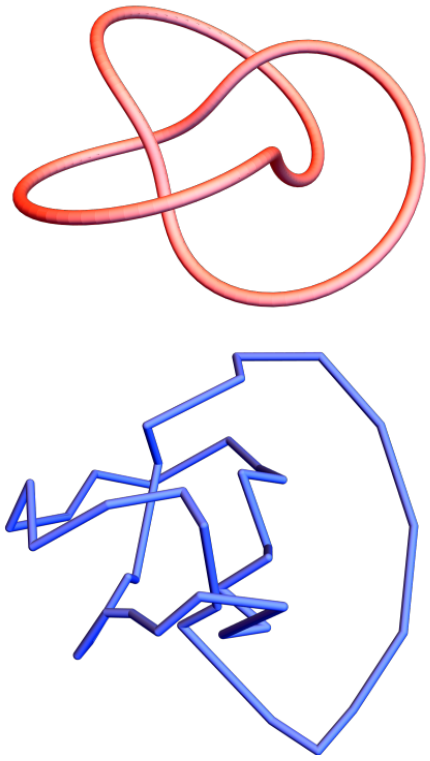


Low accuracy predictions expected for regions:

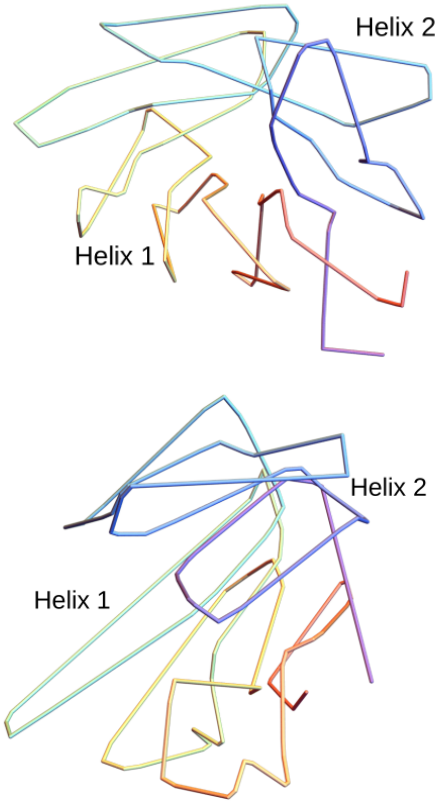
- without homologous sequences in the Protein Data Bank (PDB)
- with sequences having widely different structures.

We are also seeing that AlphaFold predictions do not fit our experimental data.
Crucially, proteins are **flexible!**

What do we mean by similar?



A An ideal and distorted trefoil.



B A Rossmann fold and a TIM barrel

There may even be an evolutionary link between these structures. [*Figueroa Yévenes et al. 2016*]

Methods/Results

Writhe

The **writhe** of a three-dimensional curve $\mathbf{x}$ with tangent vector $\mathbf{T}$ is given by the Gauss linking integral:

$$Wr = \frac{1}{4\pi} \int_{\mathbf{x}} \int_{\mathbf{x}} \mathbf{T}(s) \times \mathbf{T}(t) \cdot \frac{\mathbf{x}(s) - \mathbf{x}(t)}{\|\mathbf{x}(s) - \mathbf{x}(t)\|^3} \, \mathrm{d}s \, \mathrm{d}t. \tag{1}$$

We use the discrete analogue of (1) given by:

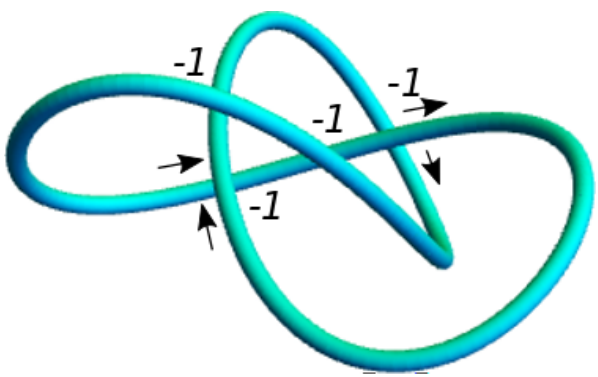
$$Wr = 2 \sum_{l=i+1}^{n} \sum_{m=l+1}^{n} \frac{\Omega_{lm}}{4\pi},$$

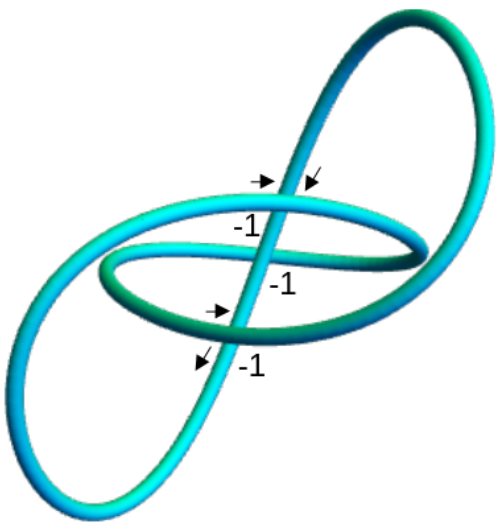
where Ω_{lm} is a signed spherical area representing the contribution to (1) from the crossing of edges connecting the pairs of points $(l-1, l)$ and $(m, m-1)$.

Writhe

The **writhe** of a three-dimensional curve $\mathbf{x}$ with tangent vector $\mathbf{T}$ is given by the Gauss linking integral:

$$Wr = \frac{1}{4\pi} \int_{\mathbf{x}} \int_{\mathbf{x}} \mathbf{T}(s) \times \mathbf{T}(t) \cdot \frac{\mathbf{x}(s) - \mathbf{x}(t)}{\|\mathbf{x}(s) - \mathbf{x}(t)\|^3} \, \mathrm{d}s \, \mathrm{d}t.$$





Average Crossing Number

The **average crossing number** (*acn*) of a three-dimensional curve $\mathbf{x}$ with tangent vector $\mathbf{T}$ is given by the Gauss linking integral:

$$Wr = \frac{1}{4\pi} \int_{\mathbf{x}} \int_{\mathbf{x}} |\mathbf{T}(s) \times \mathbf{T}(t)| \cdot \frac{\|\mathbf{x}(s) - \mathbf{x}(t)\|}{\|\mathbf{x}(s) - \mathbf{x}(t)\|^3} \, \mathrm{d}s \, \mathrm{d}t.$$

For bounds on entanglement in particular it is useful to count the number of crossings without sign as a positive definite measure of complexity of the fold.

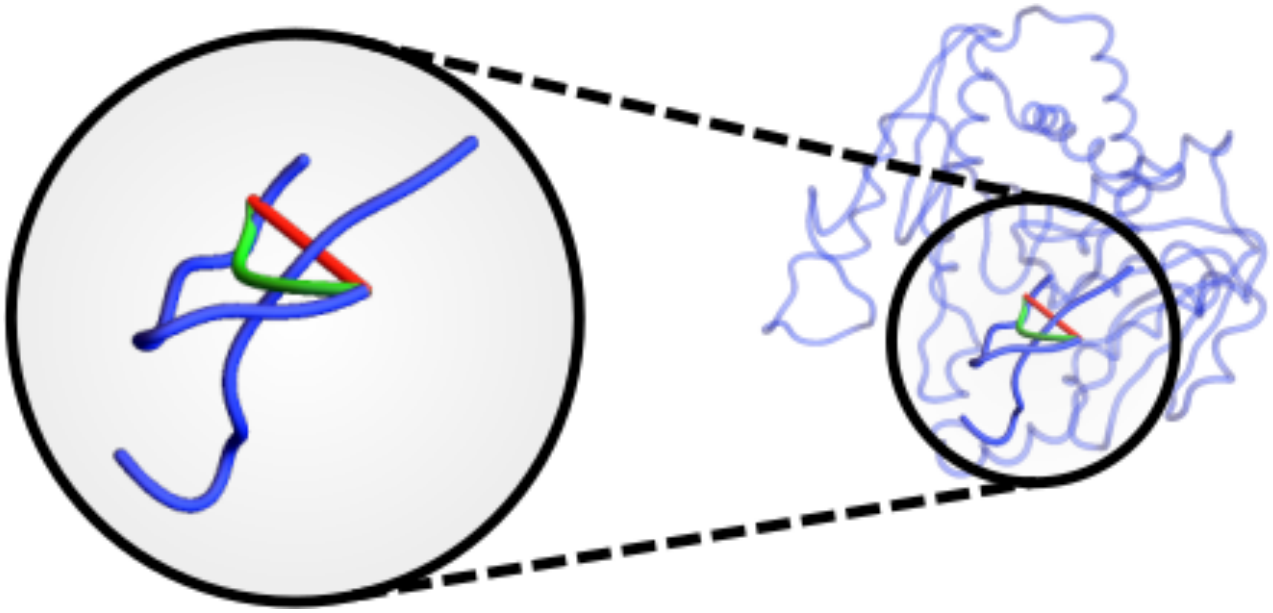
Should we compute this for the full backbone?

The backbone curve is locally helical, leading to a big build of writhe. Therefore, we would like to smooth the backbone such that:

- we preserve any essential entanglement (think knotting, slipknotting, etc.)
- we capture the rigid nature of secondary structures on the global scale.
- we have a sensible length scaling of entanglement.

We propose the SKMT algorithm, which adapts the KMT reduction algorithm [*Koniaris and Muthukumar 1991, Taylor 2000*] to act locally on secondary structures.

How the SKMT smoothing works



The full backbone of PDB 3KZK in blue. In red we see the effect of directly replacing the highlighted linker by an edge connecting endpoints. In green, we see the subsection output by the SKMT algorithm, which maintains the threading of the C-terminus strand, and so+ the knotting.

SKMT Length

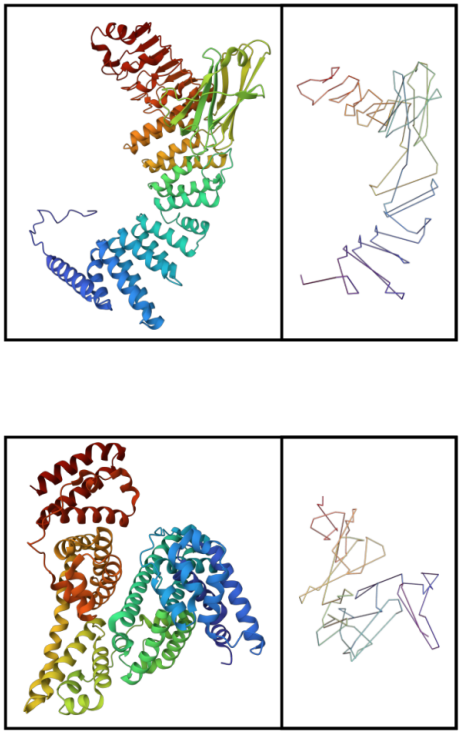


Figure 3: The backbone curves of PDB entries 1DCE and 3V03 alongside their SKMT representations.

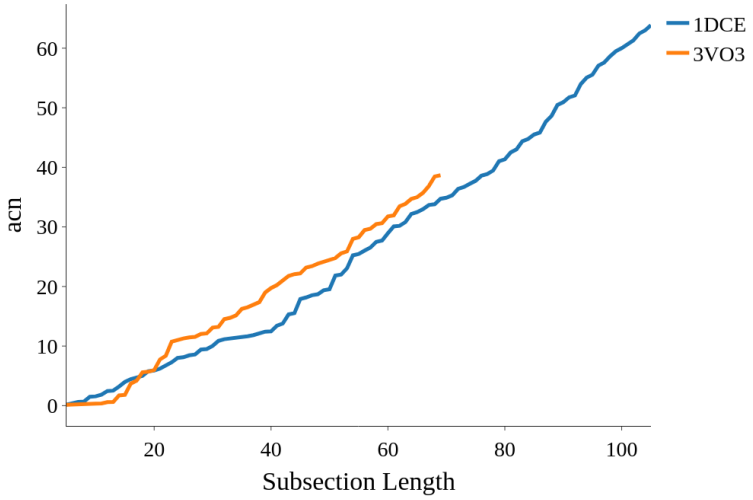
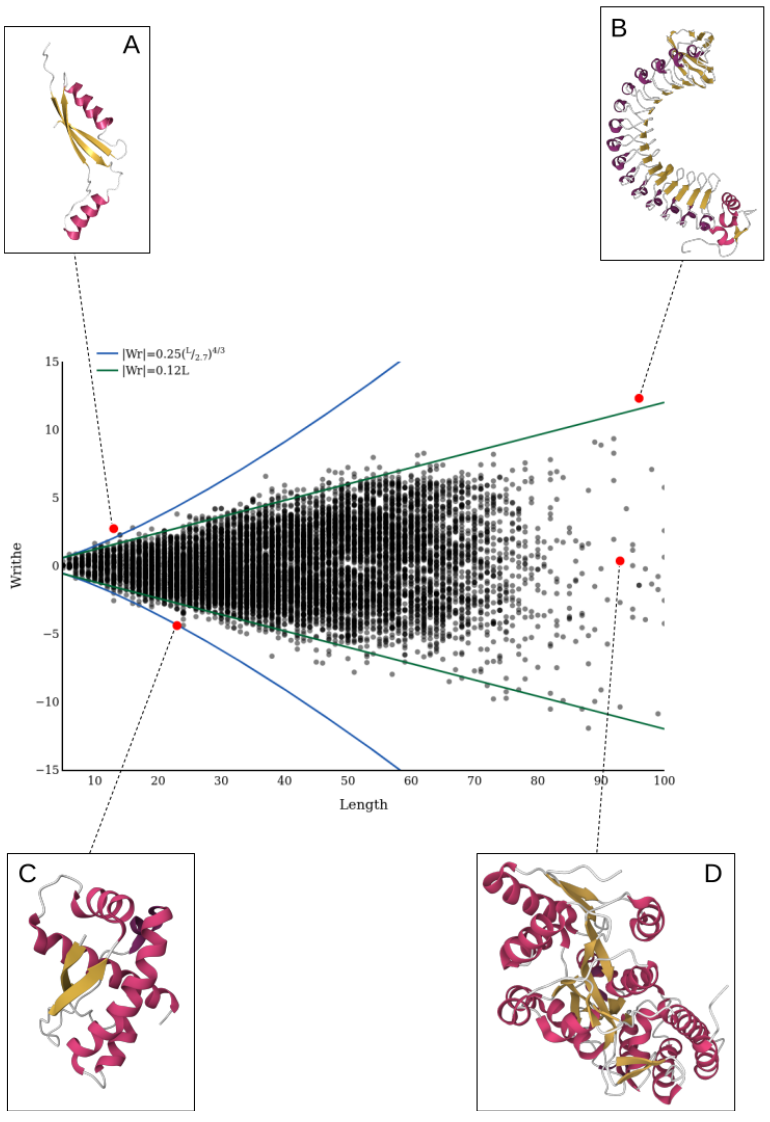


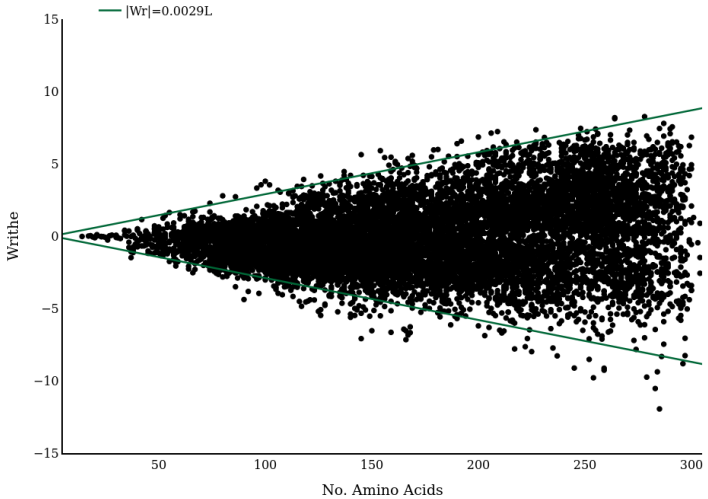
Figure 4: The *acn* profile as a function of the subsection length *n* of the SKMT smoothed backbones of 1DCE and 3V03 respectively.

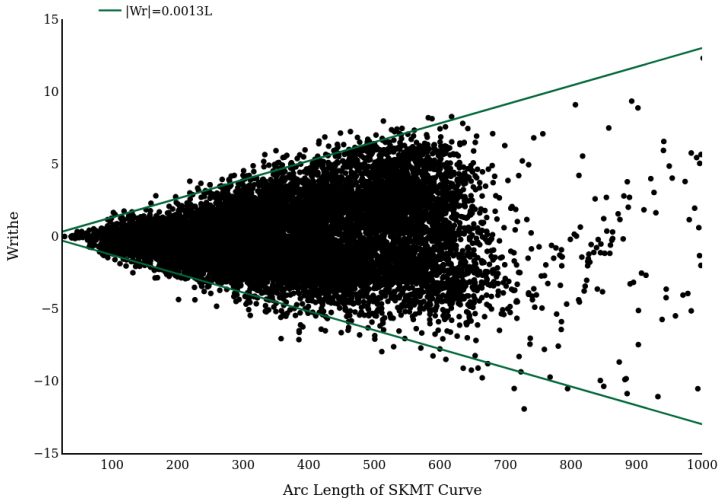
Length constraints on the writhe of proteins



- 99.9% of all values lie within the theoretical knot bound [*Cantarella, DeTurck, and Gluck 2001*].
- 97.9% of the structures fit within a linear bound $0.12L$.
- A significant number of proteins have a low or even close to zero writhe, consistent across all scales.

Other definitions of length





The distribution of writhe for some alternative definitions of length. Choosing a linear bounding curve such that 97.9% of the data lies within as before, we struggle to get as tight a fit across all length scales.

Using this distribution to build a writhe comparison metric

Comparing structures using $W(C_{ij})$

We define the similarity of subsections $\mathcal{C}^1_{ij}, \mathcal{C}^2_{lk}$ of SKMT smoothed curves $\mathcal{C}^1, \mathcal{C}^2$ (with $j - i = k - l$) as:

$$S(\mathcal{C}^1_{ij}, \mathcal{C}^2_{lk}) = \frac{1}{j - i - 4} \sum_{m=4}^{j-i} \frac{1}{0.24m} |Wr(\mathcal{C}^1_{i,i+m}) - Wr(\mathcal{C}^2_{l,l+m})|. \quad (2)$$

This measures the mean absolute difference in writhe relative to the typical maximal linear growth just discussed. To find similar subsections, we look for the largest disjoint subsections such that $S(\mathcal{C}^1_{ij}, \mathcal{C}^2_{lk})$ is less than some tolerance s_0.

Comparing structures using $W(C_{ij})$

We define the similarity of subsections $\mathcal{C}^1_{ij}, \mathcal{C}^2_{lk}$ of SKMT smoothed curves $\mathcal{C}^1, \mathcal{C}^2$, with $j - i = k - l$:

$$S(\mathcal{C}^1_{ij}, \mathcal{C}^2_{lk}) = \frac{1}{j - i - 4} \sum_{m=4}^{j-i} \frac{1}{0.24m} |Wr(\mathcal{C}^1_{i,i+m}) - Wr(\mathcal{C}^2_{l,l+m})|. \quad (3)$$

- We only consider m ≥ 4 as the value of writhe for smaller sections is not meaningful.
- Accounts for the maximal observed difference of subsections of opposing sign writhe as a function of their length.

What does this look like?

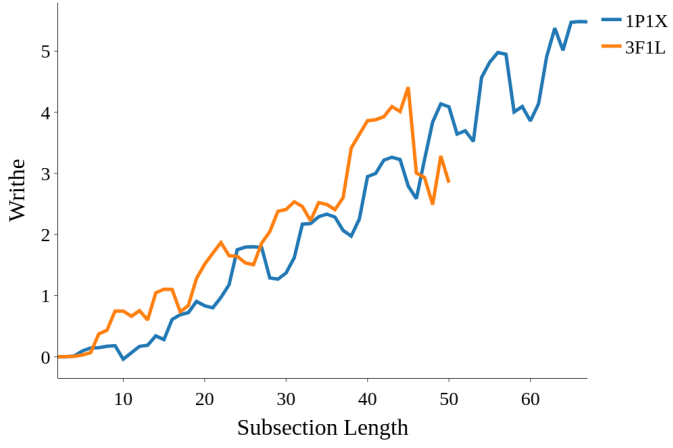
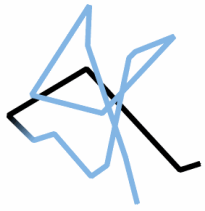


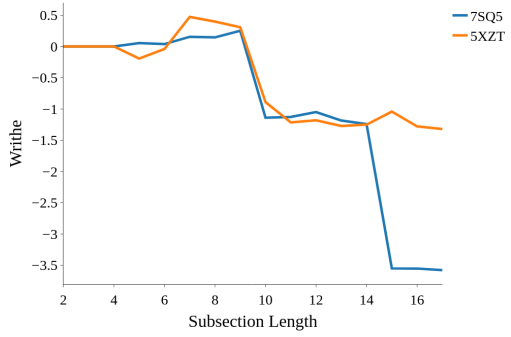
Figure 6: Visualisations of the similarity metric $S(\mathcal{C}^1_{ij}, \mathcal{C}^2_{lk})$ for the example Rossmann Fold (3F1L) and TIM Barrel (1P1X) domains.

Identifying ”knot-quites”

A PDB 7SQ5

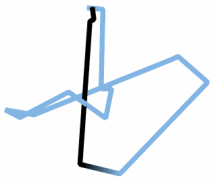


B PDB 5XZT

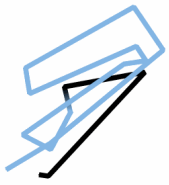


Comparison to our database for a designed trefoil knot protein (PDB 7SQ5) gives 9 matches at 80% coverage. One such example is PDB entry 5XZT, which is not knotted. The similarity metric identifies a similar profile up to the threading of the C terminus.

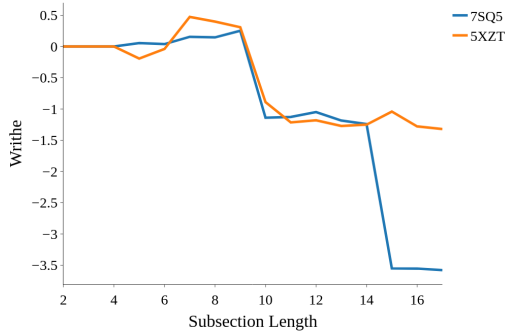
Identifying ”knot-quites”



A PDB 7SQ5



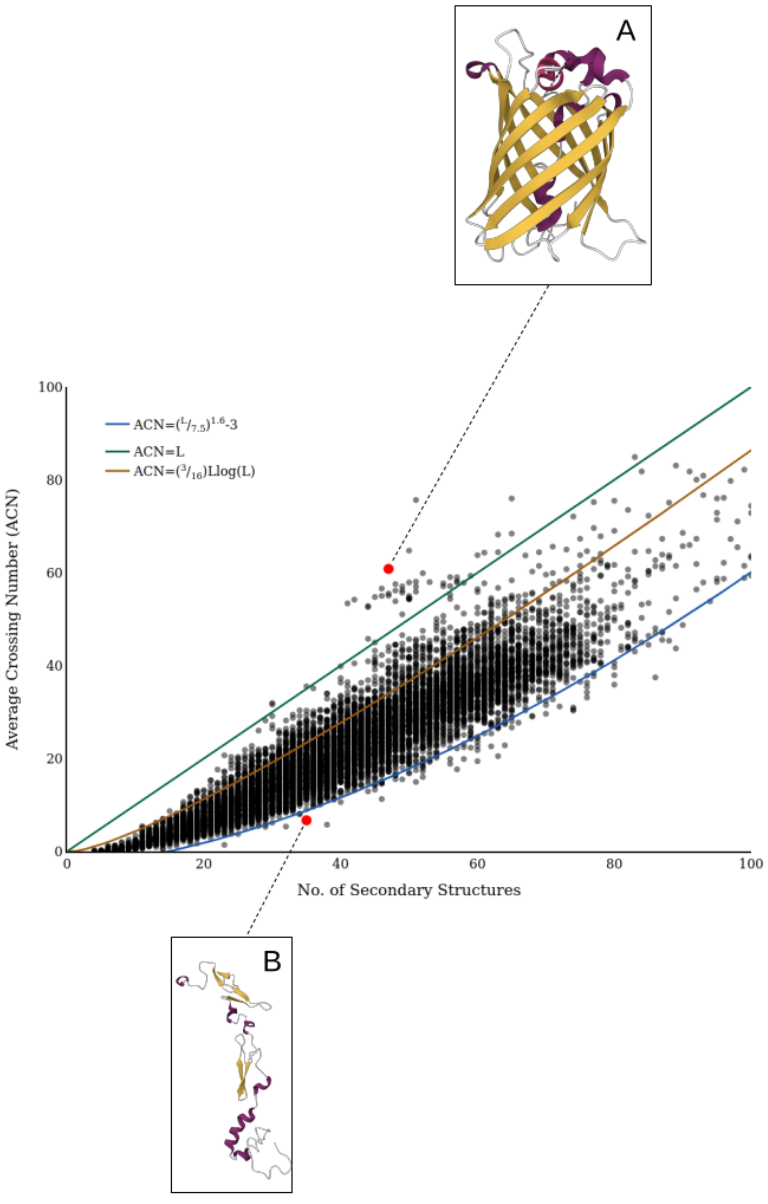
B PDB 5XZT



Comparison to our database for a designed trefoil knot protein (PDB 7SQ5) gives 9 matches at 80% coverage. One such example is PDB entry 5XZT, which is not knotted. The similarity metric identifies a similar profile up to the threading of the C terminus.

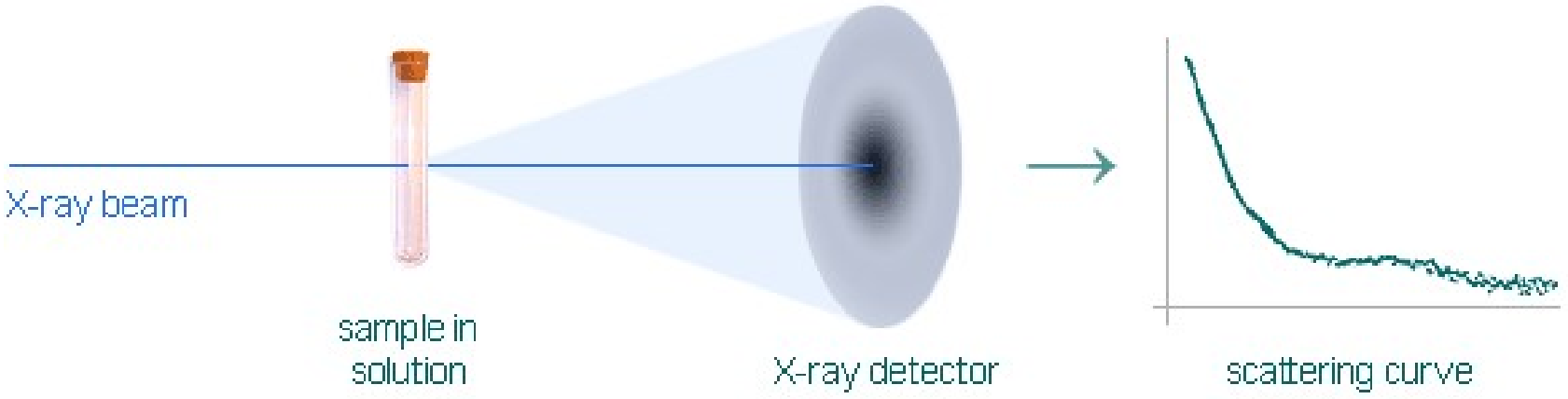
Using the acn to improve structural predictions

Length constraints on the *acn* of proteins



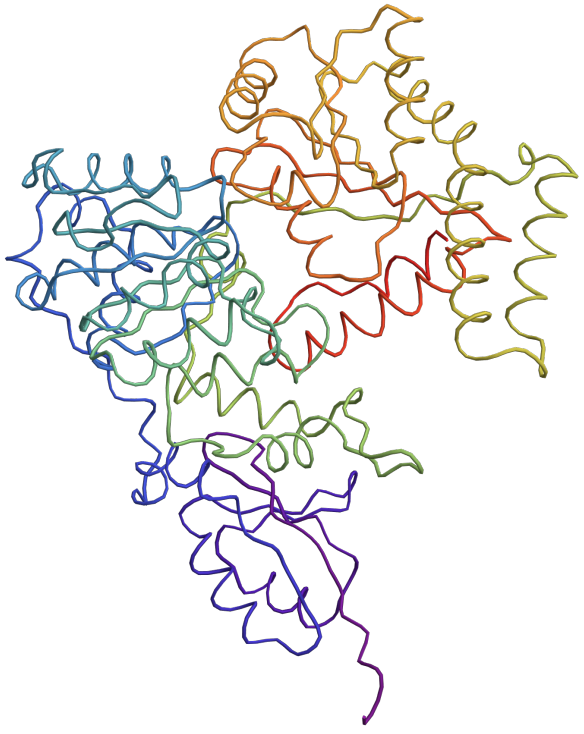
- 99.5% bound from above by a linear growth.
- 98.9% have an *acn* measure above the curve $(L/6)^{1.6} - 3$, a fit obtained by sight.
- A curve $(3/16)L\log(L)$ also acts as a reasonable upper bound, 91.4% of the data falling below this curve.

The briefest of introductions to SAXS

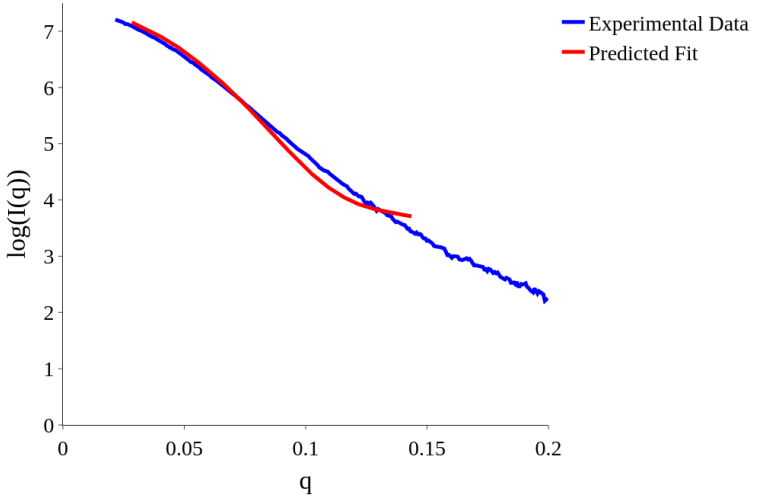


- Small Angle X-Ray Scattering studies proteins in solution.
- The incident beam is scattered by the protein onto the detector.
- The random motion of molecules in solution leads to an isotropic scattering pattern.
- The x-axis measures the momentum transfer, the y-axis shows the logarithm of the observed scattering intensity.

Human SMARCAL1

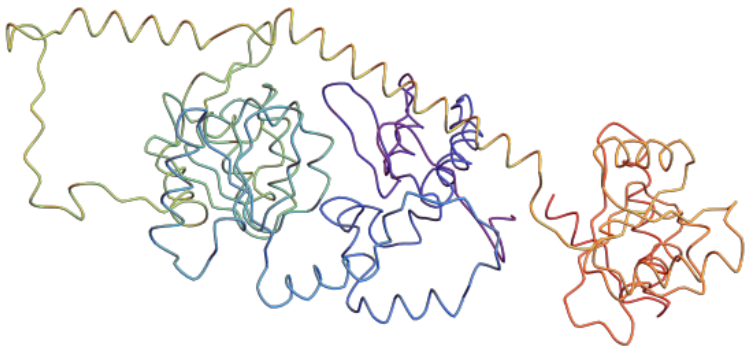


A The AlphaFold predicted structure

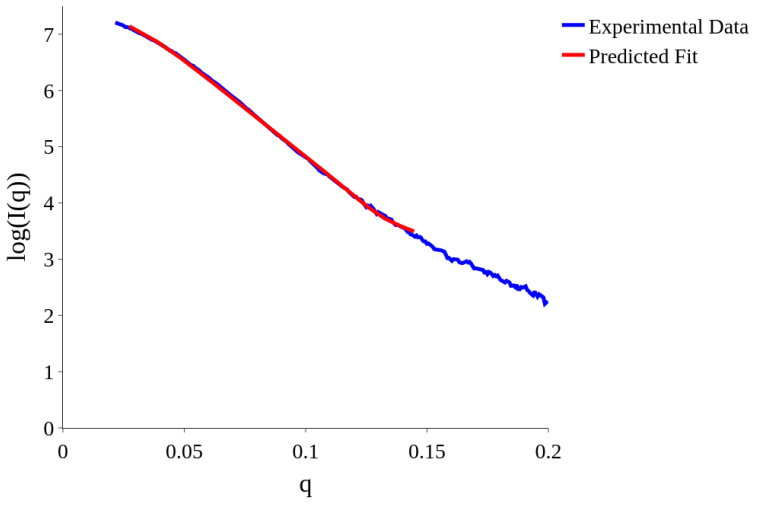


B The experimental scattering in blue. The scattering fit for the predicted structure in red. This shape suggests the structure needs to open out.

Human SMARCAL1

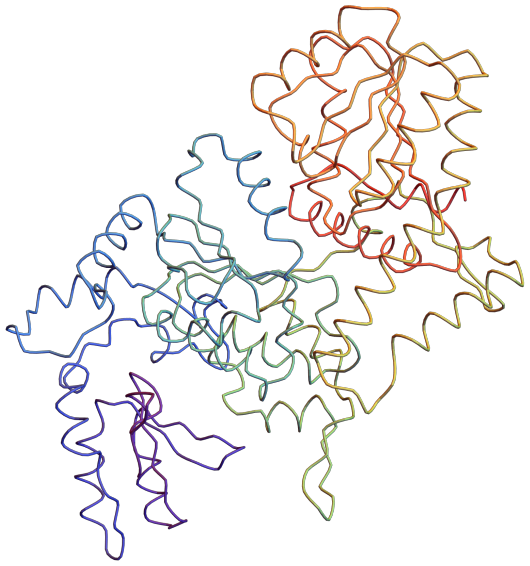


A A predicted structure which fits the scattering data, but whose *acn* falls below the empirically determined bound.

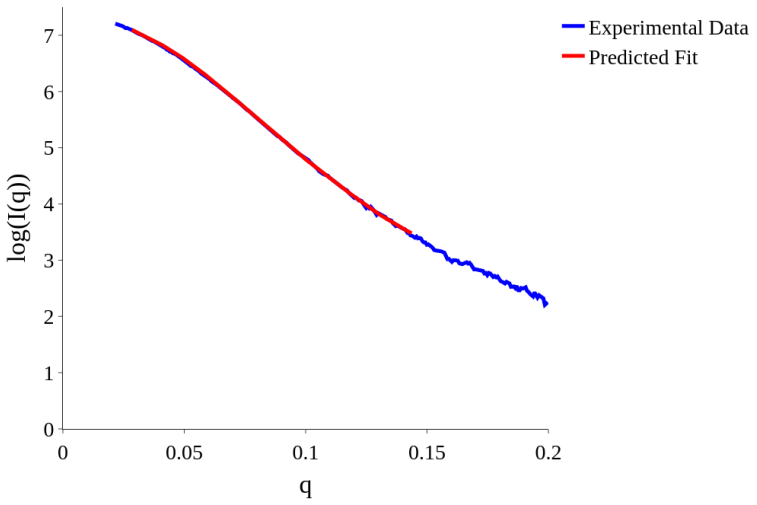


B The experimental scattering data in blue and the scattering fit of the predicted structure in red.

Human SMARCAL1



A A predicted structure which fits the scattering data, but whose *acn* falls below the empirically determined bound.



B The experimental scattering data in blue and the scattering fit of the predicted structure in red.

Conclusions

Conclusions

- We developed a novel method of smoothing protein backbone curves to uncover their global entanglement.
- In particular this method of smoothing highlights the relationship between secondary structure and global entanglement better than existing smoothing techniques.
- We derived an empirical bound on the writhe of these smoothed backbone curves, which is used to define a structural similarity measure.
- An empirical bound on the absolute complexity of entanglement is used to improve stuctural predictions to BioSAXS data.

Acknowledgements

This work was made possible by:

- Chris Prior
- Rob Rambo and Diamond Light Source Ltd.
- Ehmke Pohl and the MoSMed CDT (with funding from EPSRC)

And thanks to you for listening. Any questions?

Based on *Bale A, Rambo R, Prior C (2023) The SKMT Algorithm: A method for assessing and comparing underlying protein entanglement. PLOS Computational Biology 19(11)*

Conclusions

- We developed a novel method of smoothing protein backbone curves to uncover their global entanglement.
- In particular this method of smoothing highlights the relationship between secondary structure and global entanglement better than existing smoothing techniques.
- We derived an empirical bound on the writhe of these smoothed backbone curves, which is used to define a structural similarity measure.
- An empirical bound on the absolute complexity of entanglement is used to improve stuctural predictions to BioSAXS data.