Carbonara: Predicting protein tertiary structure from BioSAXs data.

Arron Bale

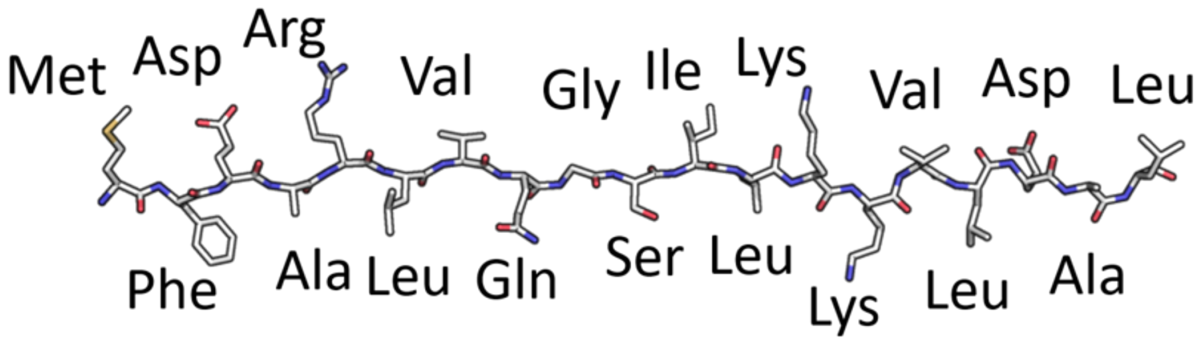
May 24, 2024

Durham University (MoSMed CDT)

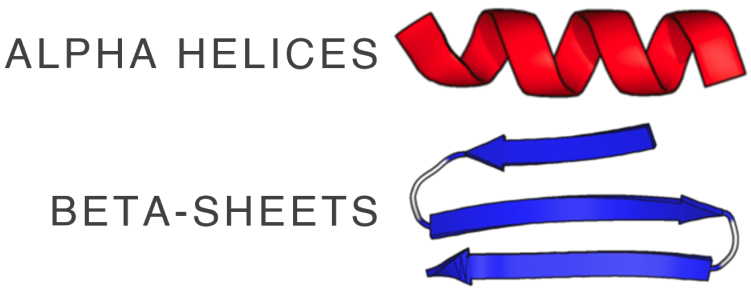
Introduction

Crash course in protein structure.

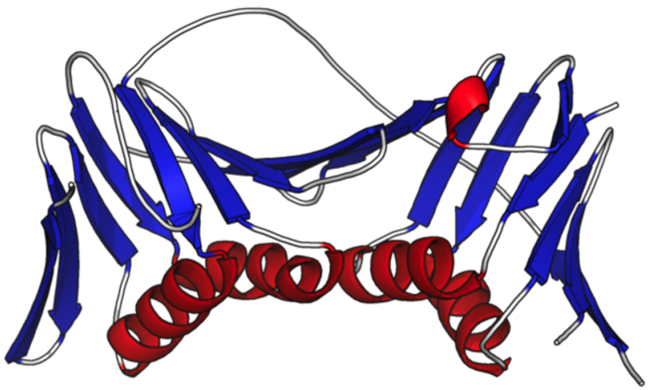
Primary Structure



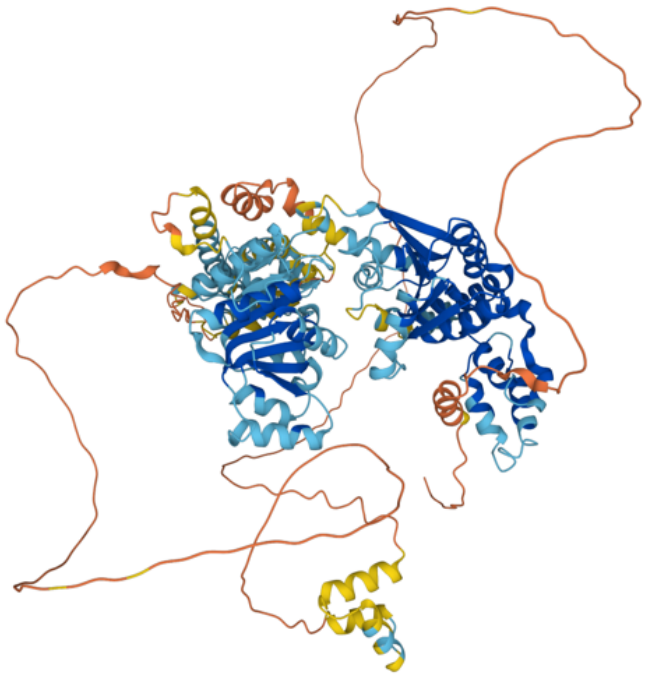
Secondary Structure



Tertiary Structure



A word on the enemy - AlphaFold

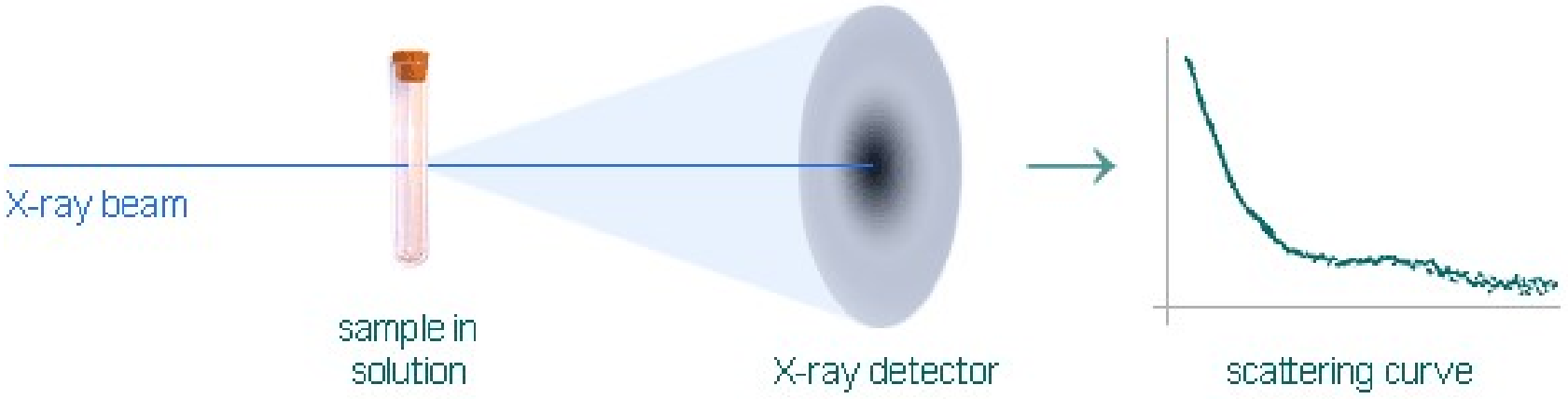


Low accuracy predictions
expected for regions:

- 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 in solution are **flexible!**

The briefest of introductions to SAXS



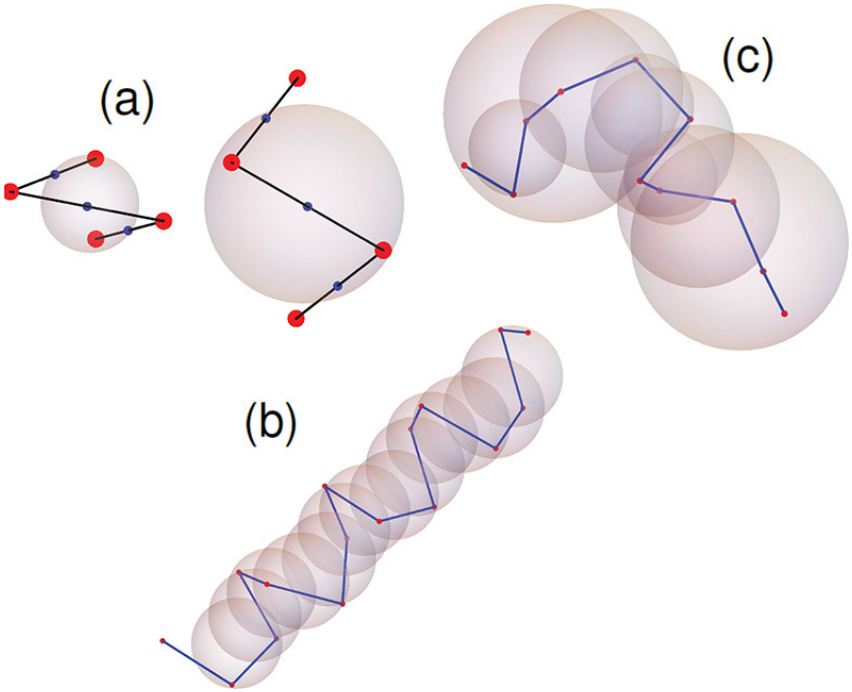
- The incident X-ray beam is scattered by the protein onto the detector.
- The random motion of molecules in solution leads to an isotropic scattering pattern.
- x-axis = the momentum transfer (the sine of the scattering angle divided by the x-ray wavelength)
- y-axis = logarithm of the observed scattering intensity.

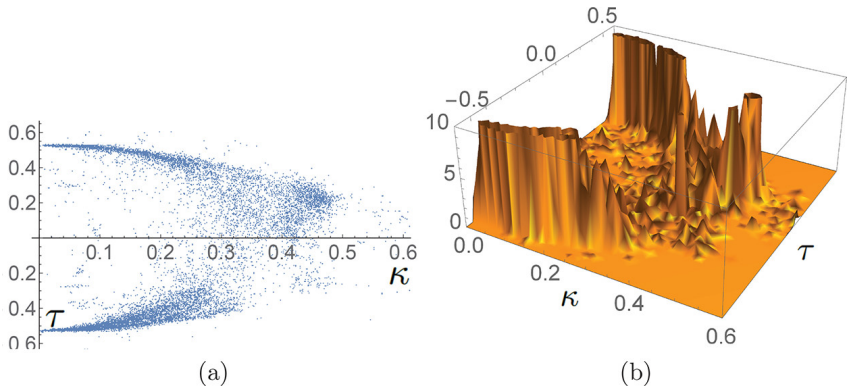
What we do

Backbone Model

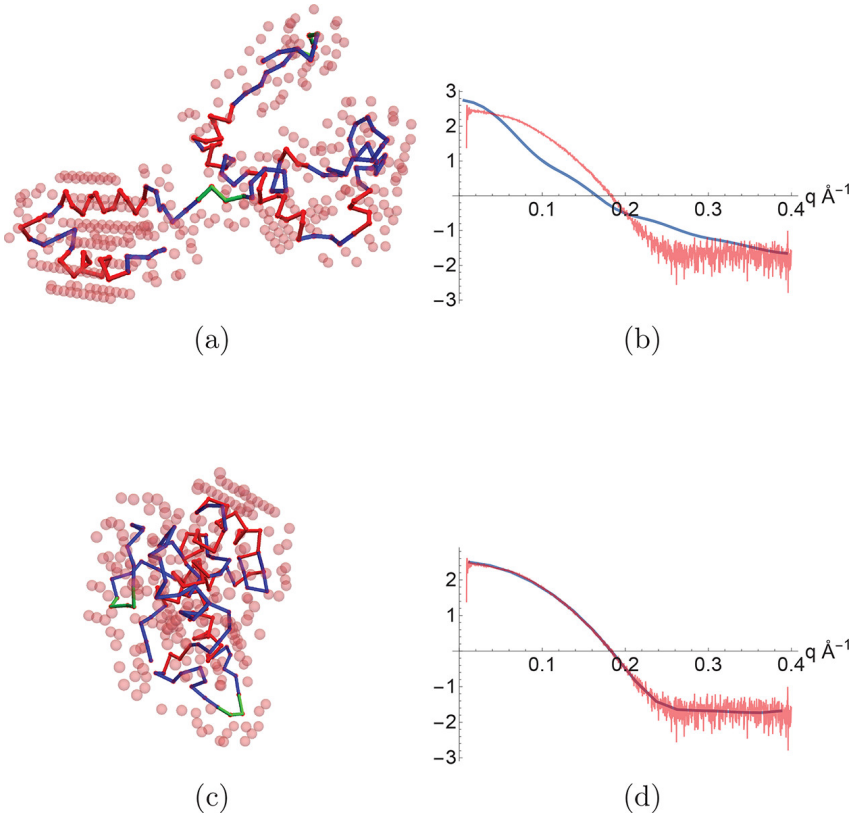
We describe the geometry of the backbone by two quantities:

- Curvature (κ) - Inverse of the radius of the circumscribed sphere defined four points.
- Torsion (τ) - Angle between the two plane normals defined by four points.



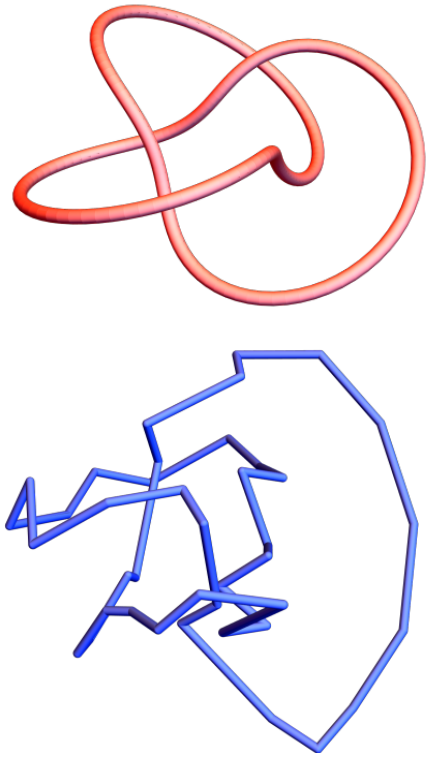


Fitting to SAXs data.

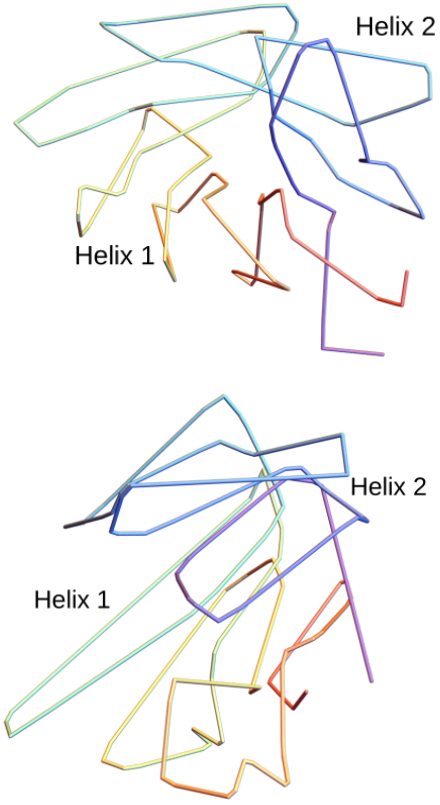


Problem: Many structures can fit the same experimental scattering curve.

We need a topological notion of similarity



A An ideal and distorted trefoil.



B A Rossmann fold and a TIM barrel

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:

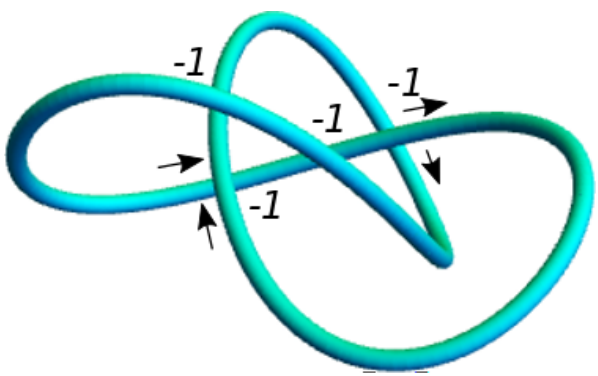
$$Wr = 2 \sum_{l=i+1}^{n} \sum_{m=l+1}^{n} \frac{\Omega_{lm}}{4\pi},$$

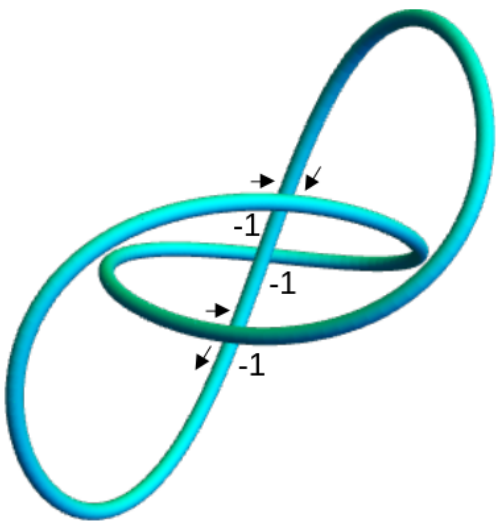
where Ω_{lm} is a signed spherical area representing the contribution to (1) from the crossing of edges connecting the pairs of points $(l-1, l)$ and $(m, m-1)$.

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.

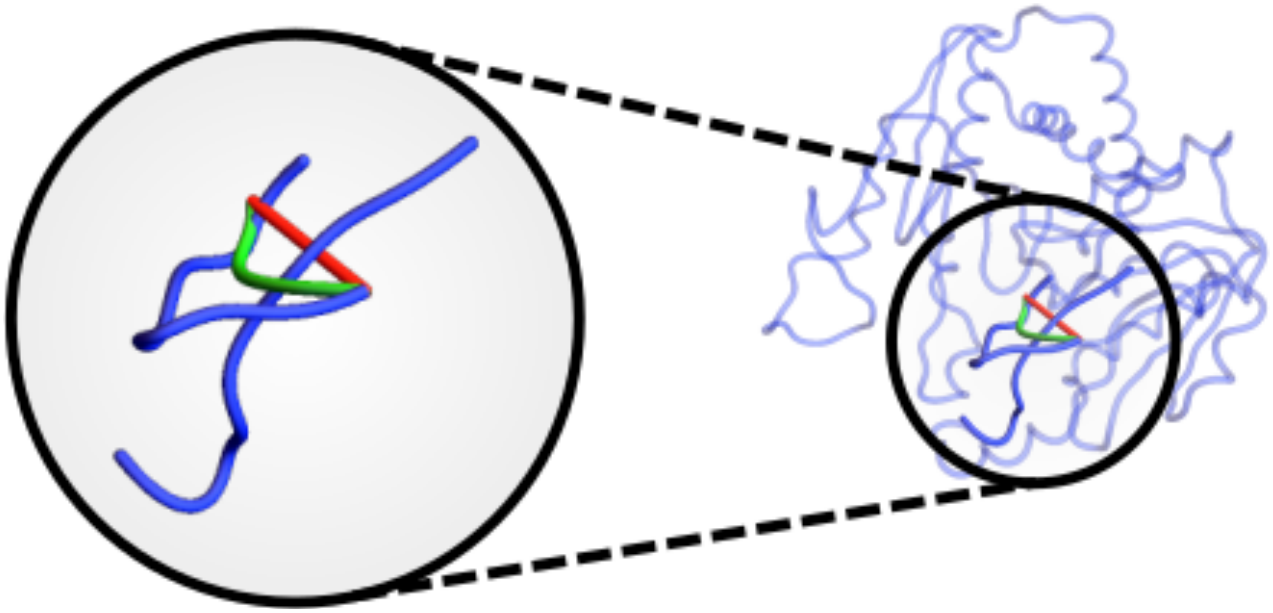
Should we compute this for the full backbone?

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.

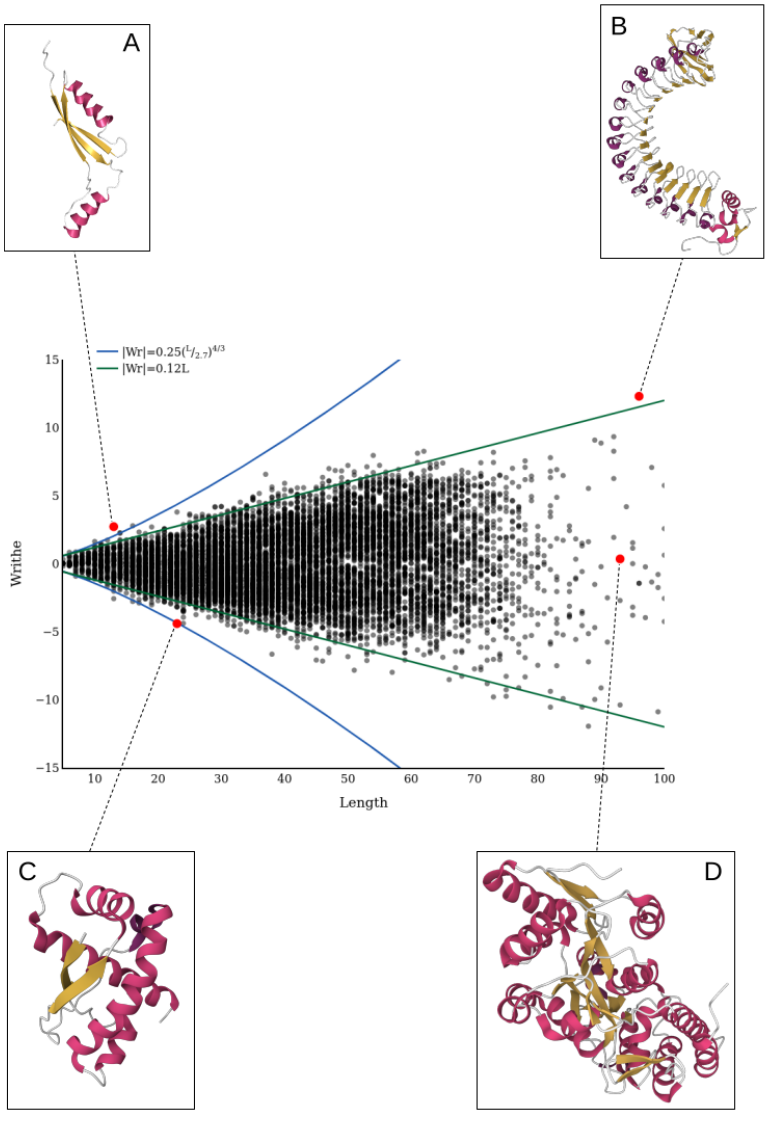
We propose the SKMT algorithm, which adapts the KMT reduction algorithm [*Koniaris and Muthukumar 1991, Taylor 2000*] to act locally on secondary structures.

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.

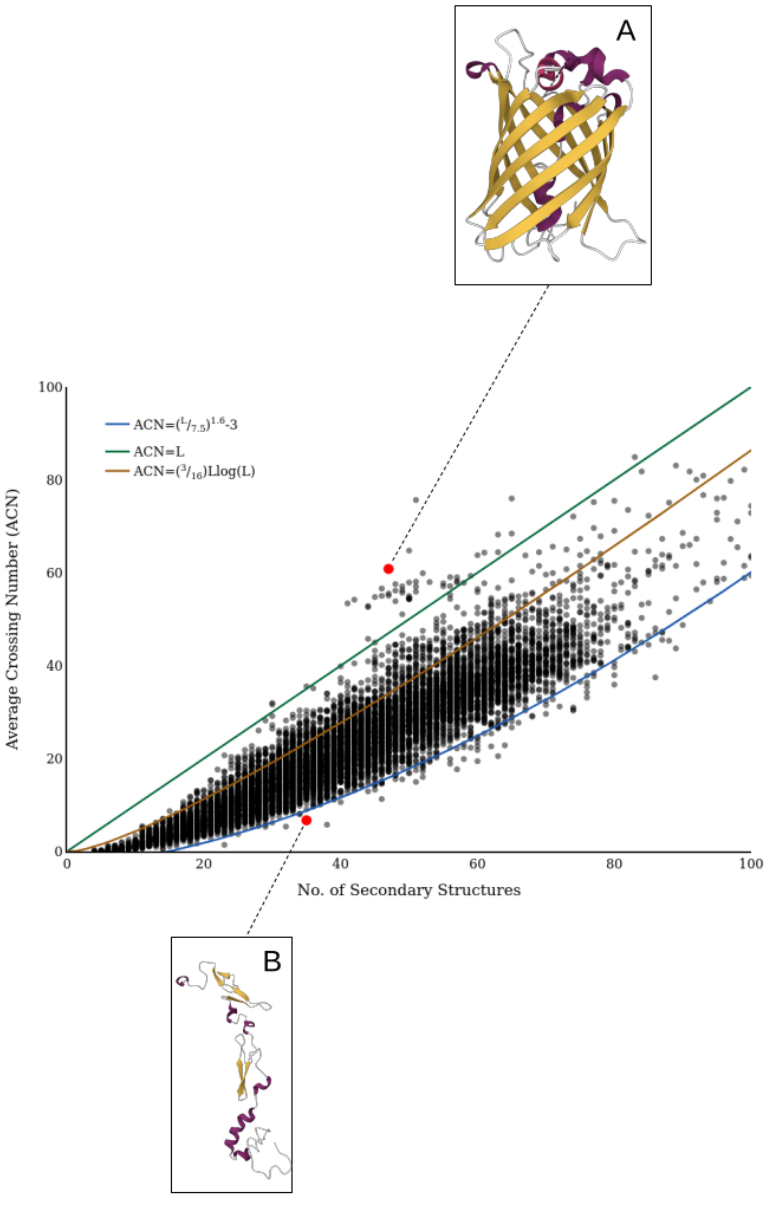
Length constraints on the writhe of proteins



- 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 $0.12L$.

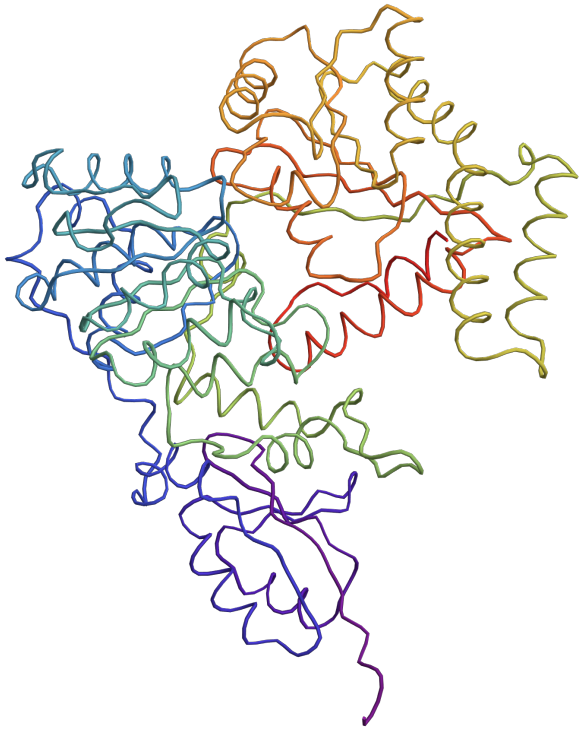
Using the acn to improve structural predictions

Length constraints on the *acn* of proteins

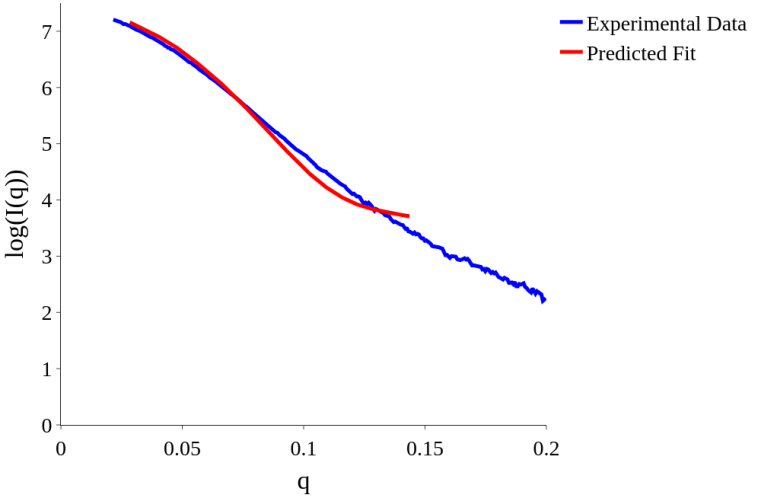


- 99.5% bound from above by a linear growth.
- 98.9% have an *acn* measure above the curve $(L/6)^{1.6} - 3$, a fit obtained by sight.
- A curve $(3/16)L\log(L)$ also acts as a reasonable upper bound, 91.4% of the data falling below this curve.

Human SMARCAL1

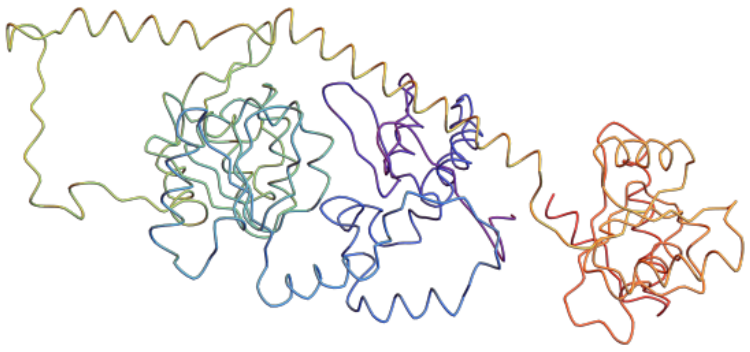


A The AlphaFold predicted structure

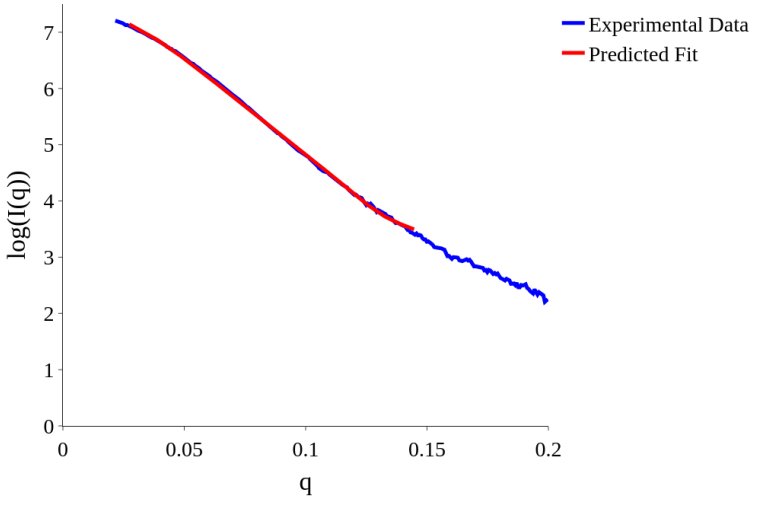


B The experimental scattering in blue. The scattering fit for the predicted structure in red. This shape suggests the structure needs to open out.

Human SMARCAL1

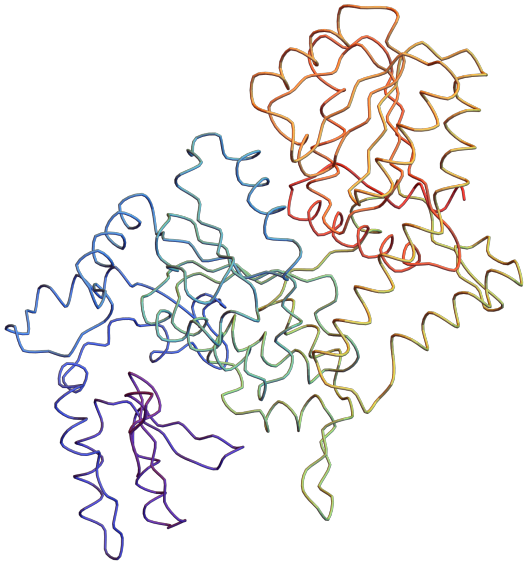


A A predicted structure which fits the scattering data, but whose *acn* falls below the empirically determined bound.

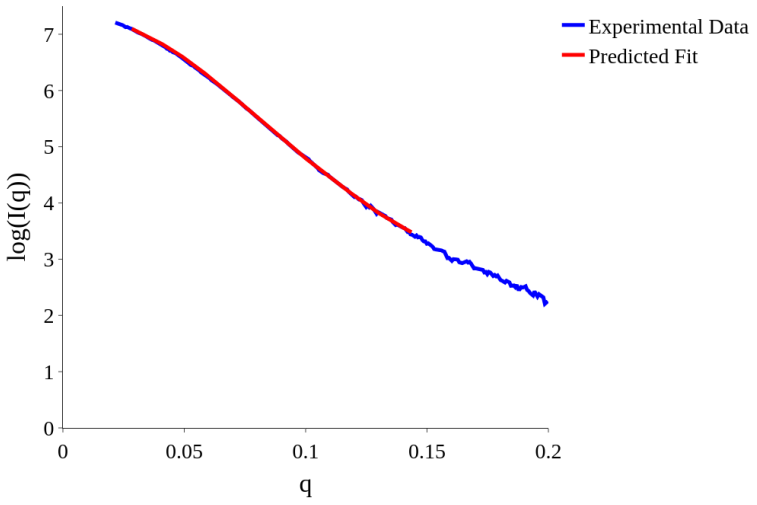


B The experimental scattering data in blue and the scattering fit of the predicted structure in red.

Human SMARCAL1



A A predicted structure which fits the scattering data, whose *acn* falls well within 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)*

Future Developments

- Optimisation Routine: Currently a naive monte carlo sampler, definitely lots of room to be smart here.
- Multimeric Predictions: This is possible, but is currently quite clunky.
- SAXs Profiling: Can we predict which sections we should change in order to affect a specific region of the scattering profile.