A tertiary structure protein model for the ab-initio interpretation of small angle X-ray scattering data

Christopher Prior Surname[1*], Owen R Davies Surname[2], Ehmke Pohl [3],

1 Department of Mathematical Sciences, Durham University, Durham DH1 3LE, United Kingdom
2 Institute for Cell and Molecular Bioscience, Medical School, University of Newcastle, Newcastle upon Tyne, NE2 4HH
3 Department of Biosciences & Department of Chemistry, Durham University, Durham DH1 3LE, United Kingdom

* christopher.prior@durham.ac.uk * ehmke.pohl@durham.ac.uk

Abstract

Biological small angle X-ray scattering (SAXS) has become an important tool to investigate the structure of proteins in solution. In this paper we present a novel ab-initio method to represent polypeptide chains as discrete curves that can be used to derive a meaningful three-dimensional model from only secondary structure annotation and experimental SAXS data. High resolution crystal structures were used to generate probability density functions for each of the common secondary structural elements found in proteins. These are used to place realistic restraints on the model curve's geometry. To evaluate the quality of potential models and demonstrate the efficacy of this novel technique we developed a new statistic to compare the entangled geometry of two open curves, based on mathematical techniques from knot theory. The chain model is coupled with an explicit hydration shell model in order derive physically meaningful 3D models by optimizing configurations against experimental SAXS data using a monte-caro based algorithm. We show that the combination of our ab-initio method with spatial restraints based on contact predictions successfully derives a biologically plausible model of the coiled–coil component of the human synaptonemal complex central element protein.

Author summary

CP: Co-conceived the project and wrote the paper. Constructed the model, tested and calibrated it, additionally he developed the knot fingerprint evaluation routine. OD: Provided experimental data. Provided biological background, reviewed the paper. EP: Co-conceived the project and wrote the paper. provided substantial guidance and biological/biochemical insight towards the model's construction as well as obtaining the experimental data.

Introduction

Biological small angle X-ray scattering (BioSAXS) is an increasingly important method for characterising protein structures in solution [1–3]. Its primary advantages over rival techniques such as crystallography and Nuclear Magnetic Resonance (NMR) is its

ability to provide information under native conditions about large protein molecules not accessible by complementary methods. However, there is a price to pay for this advantage; the random motion and orientation of molecules in solution leads to a loss of information due to an effective averaging of the scattering, leaving only information about the protein's intra-molecular distances not their spatial orientations [4]. Furthermore, scattering extends only to low resolution providing limited experimental data. The correct interpretation leading to meaningful biological results remains therefore challenging [5].

Two main methods have been developed to interpret BioSAXS data. The first assumes an accurate 3D model of the protein backbone, usually derived from X-ray crystallography [6–9]. This model is used to calculate the X-ray scattering curve once the excluded solvent volume is taken into account. A major advance, first presented in the CRYSOL algorithm [6], was the inclusion of the solvation layer - the ordered water molecules at the surface of the protein. CRYSOL and FOXS [7] adjust an implicit "shell" of scattering (implicit meaning they do not model individual solvent molecules). Other packages treat the shell explicitly using either molecular dynamics [8] or a geometric filling approach [9]. Allowing for a shell which can have gaps and fill cavities gives a more reliable fit to the data [5]. The second method does not assume an initial structure (ab-initio) but simplifies the protein model as either a volume [10] or a chain [11] of scattering beads without explicit secondary structure. These methods are hence applicable to *de novo* structural prediction, but the lack of secondary structure means interpreting these predictions is a difficult task [5].

Here we propose an alternative ab-initio technique which uses a curve model of the 3D structure of the polypeptide chain. It is parameterised by consecutive discretised descriptions of the four major secondary structural elements, α-helices, β-strands, flexible sections and random coils. The permissible geometry of these curves is restricted by empirically determined constraints, which are akin to Ramachandran constraints [12]. Variations on this curve representation have so far only been used to describe protein structures [13–15]. In order to utilize the curve representation for the interpretation of SAXS data the protein model is combined with a water model for the first hydration shell and an empirically calibrated scattering model. The geometry of the model can then be optimized against the experimental SAXS data. The main advantage of this approach over existing methods is the explicit description of the protein's fold based on realistic secondary structure. Further, this description has a much reduced number of parameters by comparison to all atomistic models. A second advantage is that by accurately characterizing the protein's secondary structure it can reliably incorporate additional structural information in order to improve the results of this technique. In this study contact predictions, based on sequence alignments alone, are used to improve the model predictions.

In order to show the potential of this technique we apply it to BioSAXS data of structural core of the human synaptonemal complex central element protein 1 (SYCE1). This protein represents an essential structural component of the synaptonemal complex (SC) that binds together homologous chromosomes during meiosis and provides the necessary three-dimensional environment for crossover formation [16–18]. The SC is formed of oligomeric α-helical coiled-coil proteins that undergo self-assembly to create a lattice-like assembly [19–21]. In a recent biochemical and biophysical study, human SYCE1 was shown to adopt a homodimeric structure in which its structural core is provided by N-terminal region (amino acids 25-179) forming an extended dimeric α-helical coiled-coil [22]. On the basis of its demonstrated anti-parallel configuration, the structural core was expressed in an engineered construct in which two SYCE1 25-179 sequences were tethered together through a short linker sequence. This construct faithfully reproduced the native structure, and substantially improved protein stability

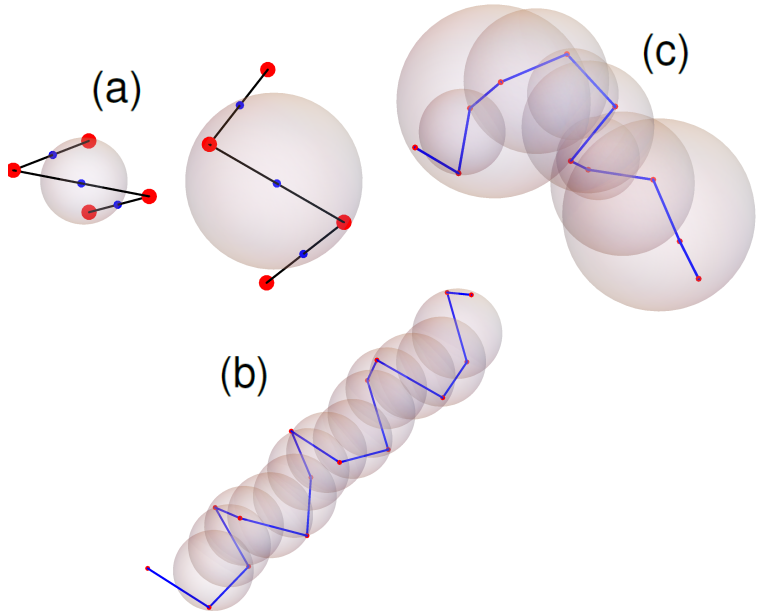


Fig 1. Figures depicting elements of the backbone model. (a) curve subsections $(\mathbf{c}_i, \mathbf{c}_{i+1}, \mathbf{c}_{i+2}, \mathbf{c}_{i+3})$ (red points) and their mid section points $(\mathbf{c}_{m1}, \mathbf{c}_{m2}, \mathbf{c}_{m3})$ (blue), the first example is more tightly wound and has a smaller sphere, hence a higher κ value. The sphere defined by these mid-section points is shown, the inverse of it's radius is the curvature κ. (b) an α-helical section with uniformly similar (κ, τ) values. (c) a flexible (linker) section with varying (κ, τ) values.

in solution and thereby enhanced the quality of BioSAXS data [22]. In this study, using secondary structure predictions and distance restraints purely based on the sequence of protein construct alone, an excellent model of an anti-parallel extended but bent coiled-coil is derived, which is fully consistent with biological data.

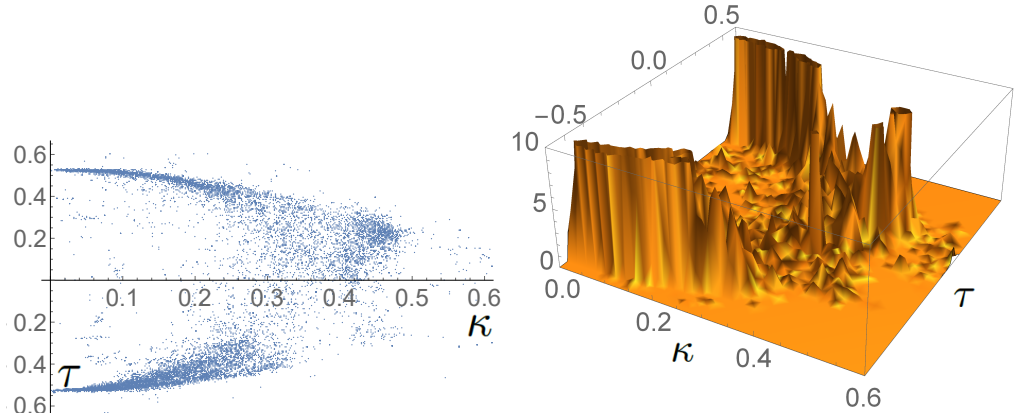
A curve description of the protein molecule

Polypeptide chain

The polypeptide chain is represented as a set of points in 3D space $\{\mathbf{c}_i\}_{i=1}^n$, the positions of the C^αatoms in each amino acid. The geometry of four consecutive points $(\mathbf{c}_i, \mathbf{c}_{i+1}, \mathbf{c}_{i+2}, \mathbf{c}_{i+3})$ can be characterized by two parameters, the curvature κ and torsion τ (see Methods and Materials). κ is defined by the unique sphere made by the centre of the joining edges (see Fig 1(a)), the smaller the sphere the more tightly the curve joining the points fold on themselves, τ measures the chirality of the section, it is positive for right-handed coiling negative if left-handed. An algorithm for generating a curve of length n from $n-3$ pairs of values of (κ_i, τ_i) is defined in Methods and Materials. Examples of an alpha-helical and flexible linker sections (taken from the structure of BSA (PDB=3V03) [23]) are shown in Fig 1(b) and (c).

Secondary structure geometry restraints

C^αcoordinates were extracted from over 60 PDB protein structures taken from the SAS database [24] and the κ and τ values calculated for all sub-sections $(\mathbf{c}_i, \mathbf{c}_{i+1}, \mathbf{c}_{i+2}, \mathbf{c}_{i+3})$. The κ-τ pairs are shown in Fig 2(a). There are three main populations of values



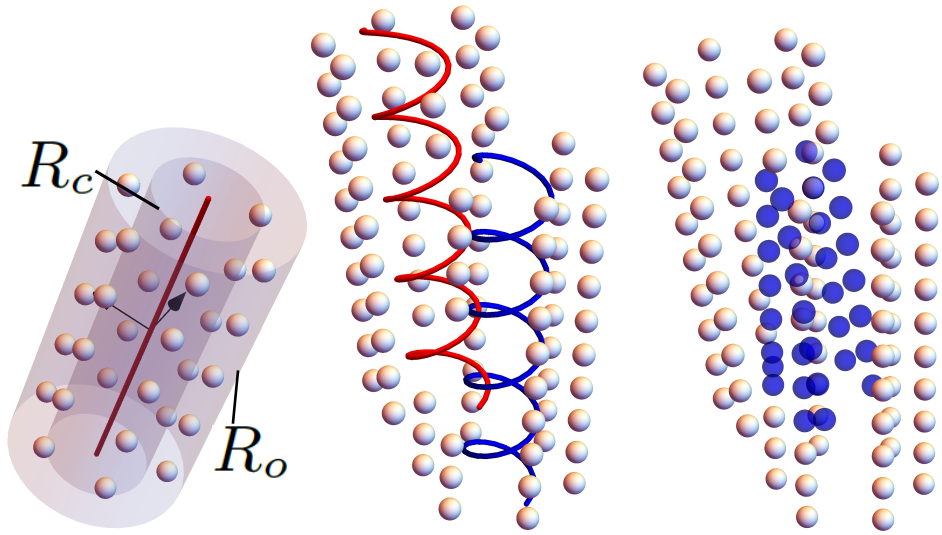
(a) **(b)**

Fig 2. Illustrations of the κ-τ spaces used to impose realistic geometry constraints on the polypeptide chain. (a) (κ, τ) pairs obtained from crystal structures, plotted as points with κ on the horizontal axis and τ the vertical axis. (b) is a P.D.F, created from the data in (a), which correspond to linker sections. There are three distinct domains of high probability corresponding to the preferred corresponding to the preferred secondary structural elements.

(preferential regions). As shown in S1 Appendix :Sec 1 these regions of (κ, τ) space correspond to the three preferential domains of Ramachandran space ([12]). Using the PDB's secondary structure annotation this data was split into categories of β-strands, α-helices and the rest which are not identified (referred to here as linkers). To account for random coils the data were further divided into subsets whose values remained in one preferential domain (as in Fig 1(b)) and those whose κ-τ values belong to multiple domains (like Fig 1(c)). For each set of data a representative probability density function (P.D.F) was calculated using Kernel smoothing techniques [25] (see S1 Appendix :Sec 1), an example is shown in Fig 2(b).

Generating models from secondary structure annotation

In order to generate models based on secondary structure information alone a protein of n amino acids is split into l distinct sub-domains of length m_i ($\sum_{i=1}^{l} m_i = n$). Each section l is classified as α-helical, β-strand or linker, for the purpose of testing and calibration a PDB file's secondary structure assignment was used to perform this task. For each section of length m_i, $m_i - 3$ (κ, τ) pairs are drawn from an appropriate P.D.F and the section is constructed. This process creates the l individual secondary structures. They must then be linked together. Two neighboring sections with specified geometry (for example an α helix and linker) still have a relative rotational degree of freedom. To ensure this remains physically realistic the geometry of the last three and first C^{α}positions of neighboring secondary sections were extracted from the PDB set and further P.D.F's for the set of permissible (κ, τ) pairs of these joining sections were generated for each type of join (*i.e.* α-helix to linker or linker to β strand). So the final step of the process is to obtain all (κ, τ) values for the joint geometry and then construct the whole backbone. A precise mathematical description of this algorithm, *constrained backbone algorithm* (**CB**), is given in S1 Appendix :Sec 1. One example of a structure generated using this algrotihm is shown in Fig 6(b), based on the 1LYZ Lysozyme secondary structure annotation; this particular structure was used as a starting point for an ab-initio structure optimization in this study.



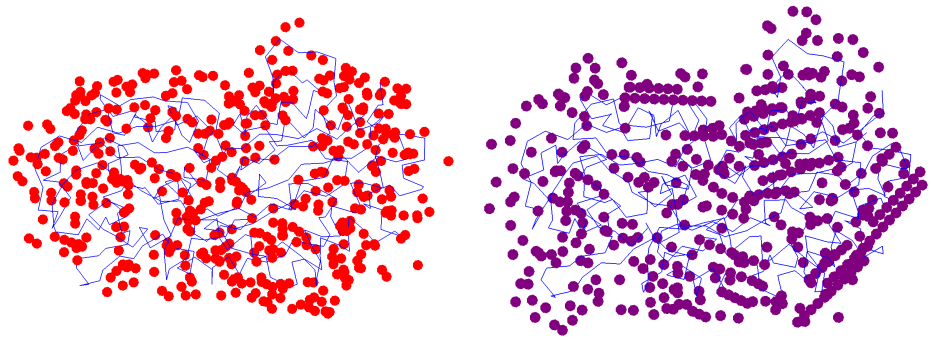
(a) **(b)** **(c)**

Fig 3. Visualizations of the hydration layer model. (a) the initial solvent layer, shown as silver spheres with the core R_c and outer R_o cylinders surrounding the axis of the section (red curve). (b) overlapping sections and solvent layers, (c) shows, in blue, the removed solvent molecules of the pair of sections shown in (b).

The hydration layer

Once the curve representation is obtained it is crucial to to add a model of the hydration layer in order to generate realistic scattering curves. To this aim a novel method was developed whereby solvent molecules are placed in-between a pair of cylindrical surfaces surrounding the axis of a section of the backbone (Fig 3(a)). This layer is then reduced by removing all overlapping solvent molecules. This ensures the shell remains in hollow sections between the fold and on the protein surface, whilst the water molecules are removed form significantly folded regions. This method is illustrated in Fig 3 where the two cylinders of radius R_c (core) and R_o (outer), $R_o > R_c$ are centered on a section i's helical axis (a). Consider a solvent molecule belonging to another section j whose nearest distance from the axis of section i is R_s. If $R_s < R_c$ the solvent is too close to the backbone and removed. If $R_c < R_s < R_o$ the solvent is classed as being shared by the sections i and j and only counted once. This process is applied to all solvent molecules from section i and j on each other, an example of the outcome is shown in Figures 3(b) and (c). Applying this process pairwise to all sections of a C^{α}backbone yields the final hydration layer. Examples of this shell model are shown in Figures 8(a) and (d).

The exact mathematical specification of this hydration layer is detailed in S1 Appendix :Sec 2. The values of the radii (R_c, R_o) and a number of other parameters controlling the solvent density were determined by fitting the model to high resolution crystal structures which contained the first hydration shell. An example model shell, generated with these parameters, is shown in comparison to the model solvent positions from the subatomic resolution structure of a phosphate binding protein from the PDB 4F1V [26] in Fig 4. It is shown the two distributions are statistically similar in S1 Appendix :Sec 2 and hence that the model is realistic.



(a) **(b)**

Fig 4. Comparisons of crystallographic and model solvent positions from the crystal structure of a phosphate binding protein PDB=4F1V, determined at an ultra- high resolution of 0.88 Å [26]. (a) the PDB backbone and the relevant solvent molecules. (b) the model solvent positions (surrounding the same curve as in (a)) obtained with the experimentally determined hydration shell model parameters.

The scattering formula

Once the polypeptide chain and hydration layer models are determined, the Debye formula [27],

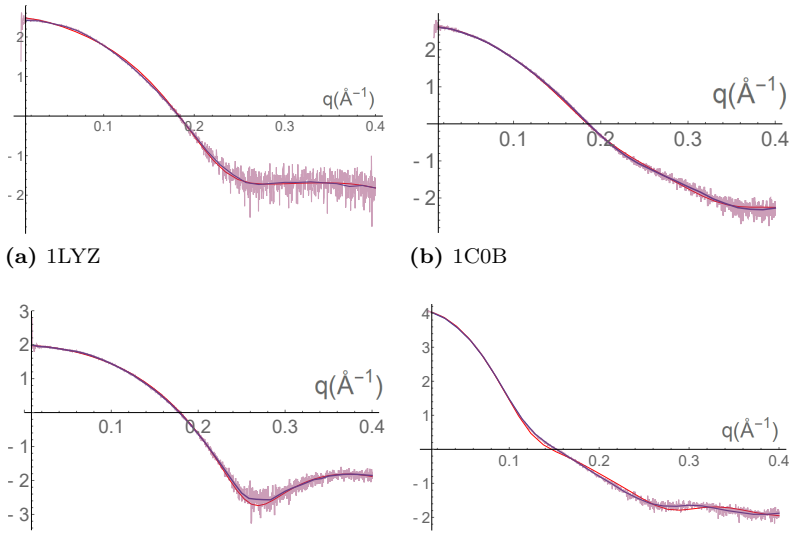
$$I(q) = \sum_{i=1}^{N} \sum_{j=1}^{N} f_i(q) f_j(q) \frac{\sin(q r_{ij})}{q r_{ij}}, \tag{1}$$

is used to calculate the scattered intensity $I(q)$ as a function of momentum transfer $q = \pi \sin(\theta)/\lambda$. Here N is total number of C^{α}'s and solvent molecules and $f_i(q)$ the form factor for residue i. There are two types, one for an amino acid with an excluded volume correction and one for a solvent molecule (see Methods and Materials). The form factors require a number of constants to be specified. To determine these constants the hydration layer model defined the previous section was applied to backbone models obtained from x-ray structures with high quality BioSAXS data available from the SAS database [24]. The parameters were then obtained by a global optimization to the experimental data in which (1) was applied to this composite backbone-hydration shell model (S1 Appendix :Sec 3). The quality of fit obtained, as shown in Fig 5, is a further endorsement of the hydration layer model introduced above.

Evaluating structural similarity.

In the next step the geometry of each model generated by the CB algorithm is optimized by refinement against the scattering data. However, since the problem is under-determined, many models will fit the experimental data so a method is required to compare structures and determine which predictions are "essentially the same" in that they only differ by small local conformational changes (as one should expect in solution).

The standard methods in protein crystallography for comparing similar protein structures are based on root mean squared deviations (RMSD) where two structures are superimposed to minimize the sum of all distances of equivalent C^{α}atoms [28, 29]. This measure and variants on it are known to be overly sensitive to large deviations in single loops (as discussed in [28]). Unlike homologous crystal structures, which will often only differ by the change in a small subsection of the whole structure, the comparison here



(a) 1LYZ **(b)** 1C0B

(c) 1HRC **(d)** 3V03

Fig 5. Fits to scattering data for various molecules using the PDB C^{α}coordinates as a backbone model $\{\mathbf{c}\}_{i=1}^{n}$. The data is shown overlayed by the smoothed data used for fitting (blue curve) and and the model fit (red curve). Variously the χ_s^2 values (see S1 Appendix :Sec 3 for its definition) are (to 3.d.p) (a) 0.001, (b) 0.002 ,(c) 0.004 ,(d) 0.003

will be made between structures generated by a random algorithm, so the significant build up of relatively small individual RMSD errors is likely. In S1 Appendix :Sec 4 a number of additional problems with using the RMSD measure in this context are discussed in detail. To mitigate these problems a novel and more robust approach based on knot theoretic techniques was developed.

Knot fingerprints

Techniques from knot theory have long been used to identify specific (knotted) entanglements in protein structure [30]. To compare two protein structures using knot theory the N and C termini need to be joined [31]. As in [30] the procedure used here is to surround the backbone with a sphere, then choose two random points on the sphere and join the end termini to these points, finally this extended curve is closed with a geodesic arc. The knot is then classified (*e.g.* via Jones polynomials). This procedure is repeated a significant number of times (10000 in this study) and the most common knot (MCK) chosen to indicate the knotting of the curve. To obtain additional information the MCK is calculated for all subsets $\{\mathbf{c}_i | \, i = k, k+1 \ldots j, \, j > k, \, j - k > 3\}$ of the curve. One can then plot this data on a "staircase" diagram with j and k on the axes and each square of the domain colored by its most common knot (*e.g.* [32]) (examples of staircase diagrams are shown in Fig 6(c), (d) and (e)). The fingerprint is found to be preserved across protein families [32], even when there is low sequence identity [33].

Secondary knot fingerprints

Fig 6(c) is the knot fingerprint for one set of Lysozyme coordinates (shown in Fig 6(a)), of the **second** most common knot identified during the random closure process. The secondary fingerprint shown in 6(d) is from a second set of Lysozyme coordinates, (c) and (d) are significantly similar. The secondary fingerprint (e) is derived from a CB generated backbone model, shown in (b), which has the same secondary structure

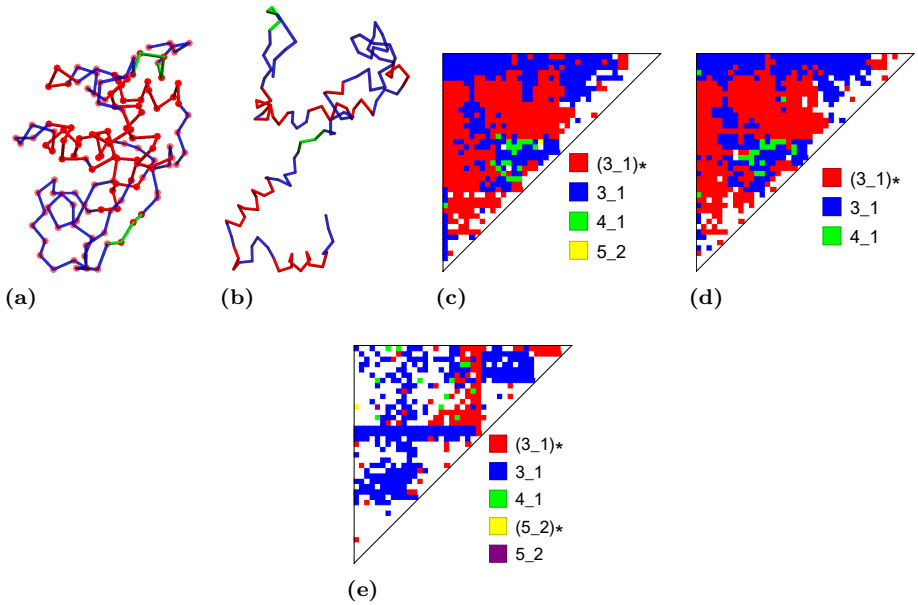
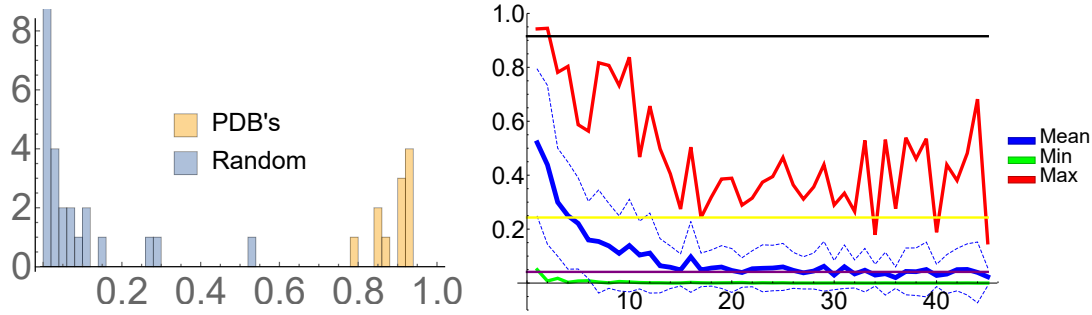


Fig 6. Secondary knot fingerprint analysis of the Lysozyme structure. (a) The C^{α}trace of Lysozyme (PDB 1LYZ [34]). The α-helices are shown in red, β-strand structures green, and linker sections light blue. (b) A structure generated using the CB algorithm which has the same secondary structural elements as Lysozyme. This could be a starting model for the fitting procedure. Panels (c) and (d) are secondary fingerprints of two different crystal structure of Lysozome (1LYZ and the 1AKI respectively). The knot types are indicated (Rolfsen classification [**?**]), white spaces indicate no secondary knots (all knots were of the primary type). (e) Secondary fingerprint for the random structure shown in (b), it differs significantly from (c) and (d) and has a larger range of knots present.

sequence as the 1LYZ PDB. The secondary fingerprint differences between the correct structure (c) and the randomly generated structure (d) is immediately obvious. All primary (MCK) fingerprints in these cases are *identical* and all have the unknot as the MCK. It is clear secondary (and possibly Tertiary) knot fingerprints can differentiate un-knotted folds. A knot fingerprint statistic $\mathcal{K}_l(K_1, K_2)$ is defined in Methods and Materials which quantifies the weighted similarity of knot fingerprints at level l associated with the curves K_1 and K_2 ($l = 2$ for Figures 6(c)-(e)); it yields a value between 0, completely dissimilar, and 1, identically folded.

In S1 Appendix :Sec 4 it is demonstrated that the statistic has the following properties. Firstly it quantifies crystal structures of the same molecule as highly similar $\mathcal{K}_2(K_1, K_2) > 0.77$ and randomly generated structures (with the same secondary structure sequence) as significantly dissimilar, generally $\mathcal{K}_2(K_1, K_2) < 0.1$ (see Fig 7(a)). Secondly it judges crystal monomer structures of similar length as being significantly different (typically $\mathcal{K}_2 < 0.4$), *i.e.* it can differentiate folds. Thirdly it is shown to have excellent properties under deformation. To demonstrate n randomly distributed changes were applied to a crystal structure K_{pdb} using the CB algorithm. For each n 50 such structures K_n were generated and the values of the statistic $\mathcal{K}_2(K_{pdb}, K_n)$ calculated. The results are plotted as a function of n in Fig 7(b) for Lysozyme. The mean value drops off rapidly to the same value as the average of the randomly generated structures (after about 15 changes). The maximum value always remains significantly higher than the mean, it drops below PDB quality after only 2 changes. So a high $\mathcal{K}_2(K_{pdb}, K_n)$



(a) **(b)**

Fig 7. Properties of the (secondary) knot fingerprint statistic $\mathcal{K}_2$ based on variations of the Lysozyme structure. (a) Secondary knot statistics $\mathcal{K}_2(K_{1LYZ}, K)$ of various structures K compared to the curve shown in Fig 6(a). The two distinct sets are Lysozyme PDB's and random structures with secondary structure alignment to Lysozyme (generated using the CB algorithm). (b) Plots of the mean, maximum and minimum value of the 50 secondary knot statistics comparing the 1LYZ structure and the same structure subjected to n random changes in its secondary structure. The dotted lines show 1 standard deviation from the mean. The black line is the average of the PDB structure secondary fingerprint statistics (see (a)) the purple line the Random structure average (crossing the mean at about $n = 15$) and the yellow line the average of secondary fingerprint values for models which fit the experimental data (crossing the mean at about $n = 3$).

value > 0.75 indicates the structure is likely only a few changes away from the original structure.

1 Results

In the previous sections a protein model was developed for the interpretation of BioSAXS data which can derive a realistic tertiary fold for the protein using **only** its secondary structure annotation. This process is summarized in Fig 8 for Lysozyme. First the secondary structure annotation is used to generate a random 3D structure. This is surrounded with an explicit hydration layer (Fig 8(a)) and a model scattering curve is calculated and compared to the experimental data (b). The curve is then changed by using a Monte-Carlo algorithm to generate new secondary structure units (along with a new hydration shell), thus altering the model's fold until it attains a sufficiently good fit to the scattering data Fig 8(e), the structure, shown in Fig 8(d), is now significantly folded like the crystal structure. This improvement is reflected in the fingerprint comparisons, *c.f.* 8(c) and 8(f).

To test the fitting procedure the model was applied to model the protein Lysozyme (PDB:1LYZ). Random initial models were generated using the CB algorithm. The following chi -square statistic χ^2_f was used to to asses the fit quality

$$\chi^2_f = \frac{1}{n_s - 1} \sum_{i=1}^{n_s} \left[\log(I_m(q_i)) - \log(I^s_e(q_i)) - L_d\right]^2 , \tag{2}$$

$$L_d = \frac{1}{n_s} \sum_{i=1}^{n_s} \log(I_m(q_i)) - \log(I^s_e(q_i)).$$

Where n_s is the discrete number of points on the domain $q \in [0, 0.4]$ on which the

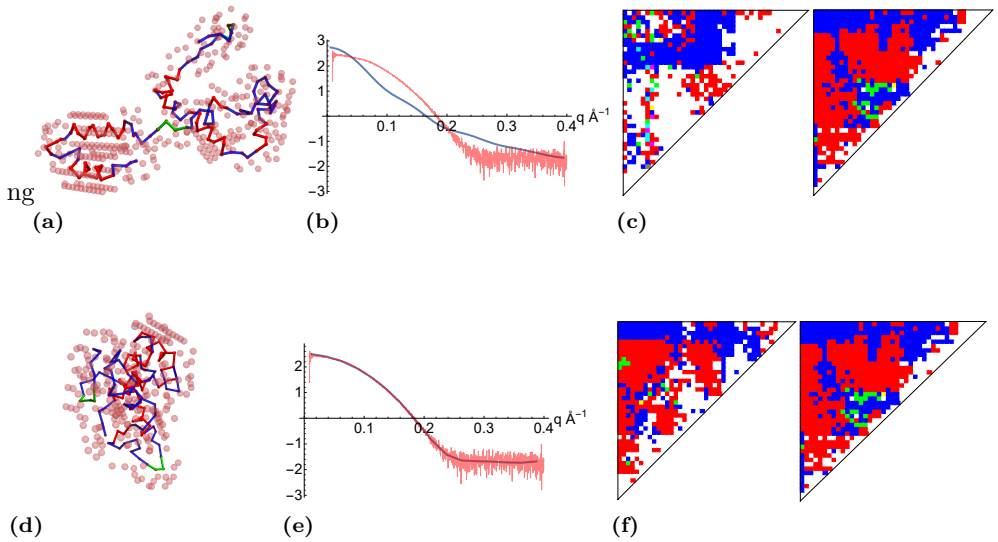


Fig 8. Figures illustrating the fitting process. (a) An initial configuration of the backbone based only on the secondary structure assignment of Lysozyme (PDB:1LYZ). Also shown as spheres are the molecules of the hydration layer. (b) The model scattering curve compared the experimental data. (c) A comparison of the secondary knot fingerprints of the structure in (a) and the crystal structure of Lysozyme (which is the fingerprint on the right). (d) a final structure (and hydration layer) obtained from the fitting process and its model scattering curve (e). A comparison of the secondary knot fingerprints of the structure in (d) and the crystal structure of Lysozyme.

scattering is sampled (a commonly used domain *e.g.* [7]). I_m is the model scattering calculated using the Debye formula (1) and I_e^s, the smoothed experimental data (smoothed using the procedure described in [35]). The factor L_d, which will superimpose identical curves which differ by a translation, is used because the protein concentration can only be measured with relatively low accuracy [6, 7] (when taking a logarithm of the data a scaling factor becomes a vertical translation). In addition, to prevent chemically unreasonable conformations, a penalty is applied if the C^α-C^αdistance of ≤ 3.8Å occurs for any pair of non-adjacent C^αpositions, this quantity is labelled χ_{nl} (see Methods and Materials). The initial randomly generated model is optimized as described above until $\chi_f^2 + \chi_{nl} < 0.008$. Values below this threshold represent an excellent fit, as shown in Fig 8(e).

Knot statistics of optimized Lysozyme models

In order to validate our method optimized models K_f generated from 40 random starting points were generated by optimization against experimental SAXS data (from [24]). They were analyzed using the secondary knot statistic $\mathcal{K}_2(K_{1\mathrm{LYZ}}, K_f)$ (Fig 9). A number of models are significantly structurally similar to the crystal structure $\mathcal{K}_2 > 0.75$, but a significant number perform little better than random. The average $\mathcal{K}_2$ value 0.241, which from Fig 7(b) is the same as that of structures obtained by making 3 changes to sections of the Lysozyme backbone (PDB:1LYZ).

When the protein's structure is unknown the aim of this technique will be to produce a reasonable number of distinct possible tertiary structures. To see if the 40 structures obtained do separate into distinct sub-categories we isolated two sets of structures. Those structurally similar (using $\mathcal{K}_2$) to the model K_f^h with the **highest**

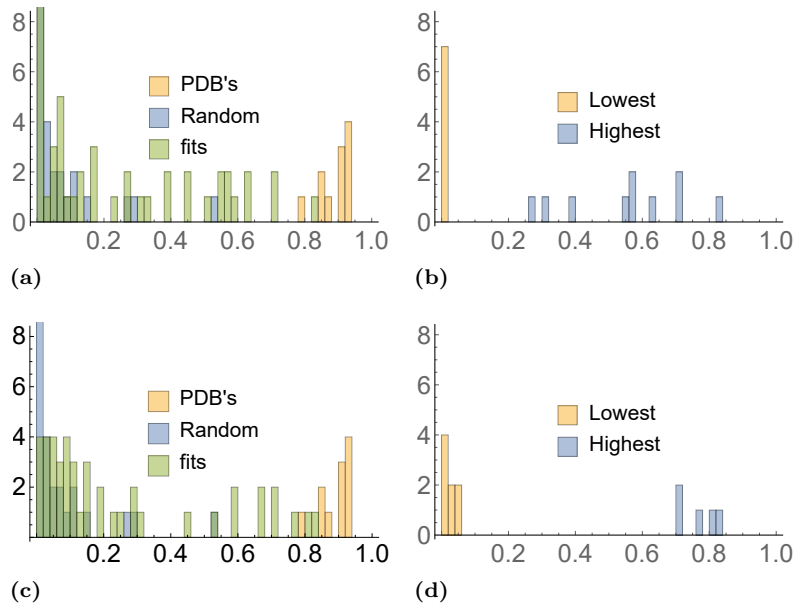


Fig 9. Histograms of $\mathcal{K}_2$ values for ab-initio fits to Lysozyme scattering data. (a) $\mathcal{K}_2(K_{1\mathrm{LYZ}}, K_f)$ values for all K_f obtained by model optimization, plotted in conjunction with the random and PDB data shown in Fig 6. A good number of fits are PDB or close to PDB quality, but there are some which are no better than the random model. (b) histograms of subsets of the data shown in (a) which are classed as structurally similar to the models K_f with the highest and lowest $\mathcal{K}_2(K_{1\mathrm{LYZ}}, K_f)$ values, labelled respectively high and low. (c) as in (a) but for fits obtained with the additional contact predictions. (d) as in (b), shows subsets of structurally similar predictions, the separation of the two sets is significantly more pronounced than in (b).

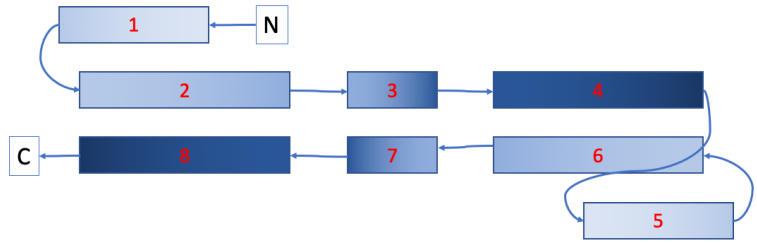


Fig 10. Schematic drawing of the SYCE1 construct with each box corresponding to one predicted alpha helix. The SYCE sequence of approximately 120 amino-acids corresponding to helices 1-4 is duplicated and linked by a tether to a repeat of the same sequence comprising of helices 5-8.

$\mathcal{K}_2(K_{1\mathrm{LYZ}}, K_f)$ value, and those structurally similar to the model K_f^l with the **lowest** $\mathcal{K}_2(K_{1\mathrm{LYZ}}, K_f)$ value. The subsets of the histogram data shown in Fig 9(a) which belong to these two sets are shown in Fig 9(b). They do not overlap, but there is significant variance in the set of $\mathcal{K}_2$ values for those structures similar to K_f^h. It is clear additional structural restraints to the model are required to improve the separation of the algorithm's predictions.

To test the value of additional information contact predictions, based on a large number of homologous sequences, were incorporated into the model restraints. Data from the Raptor X web server [36] for the for the sequence of Lysozyme were obtained. The C^αpairs with the 10 highest correlations were selected. An extra potential χ_{con} was added to the optimization statistic to ensure the distance between these pairs was restricted to be within 5 and 15 Å(see Methods and Materials).

The $\mathcal{K}_2(K_{1\mathrm{LYZ}}, K_f)$ statistic of 40 model structures K_f for which $\chi_f^2 + \chi_{nl} + \chi_{con} < 0.008$ are shown in Fig 9(c). A number of the structures are PDB quality. The folded structure shown in Fig 8(d) is one of these structures. The average is 0.262, higher than in the standard optimization.

In addition the data show better separation properties. Similar to above, subsets of high and low $\mathcal{K}_2(K_{1\mathrm{LYZ}}, K_f)$ structures were isolated. The results show a much clearer separation than found in the absence of contact predictions (*c.f.* Fig 9(b) and (d)).

Application to the structural core of human SYCE1

Based on the success of utilizing contact predictions to constrain potential models we applied the algorithm on the structural core of the human SYCE1 protein, a tethered construct where the sequence is repeated to allow formation of an extended anti-parallel coiled-coils with two short additional helices at each end that could fold back to form a small 3-helix bundle. The secondary structure of the tethered protein construct resulted in eight stretches of alpha-helices where based on the heptad repeats helices 2, 3, and 4, can be aligned to helices 6, 7, and 8 corresponding to the same sequence, respectively in an anti-parallel fashion. This resulted in 14 close contact predictions between helices 2 and 8, and helices 4 and 6, respectively, as shown in Fig 10.

Deriving the models

Based on the sequence and secondary structure predictions (a combination of those of Raptor X [36] and HHPRED [37]) a number of initial configurations were generated using the CB algorithm. An example is shown in Fig 11(a) along with its hydration layer, its scattering curve is compared to the experimental data (from [22]) in Fig 11(b). As shown the fitting is limited to the domain $q \in [0, 0.3]\mathring{A}^{-1}$, which balances the twin

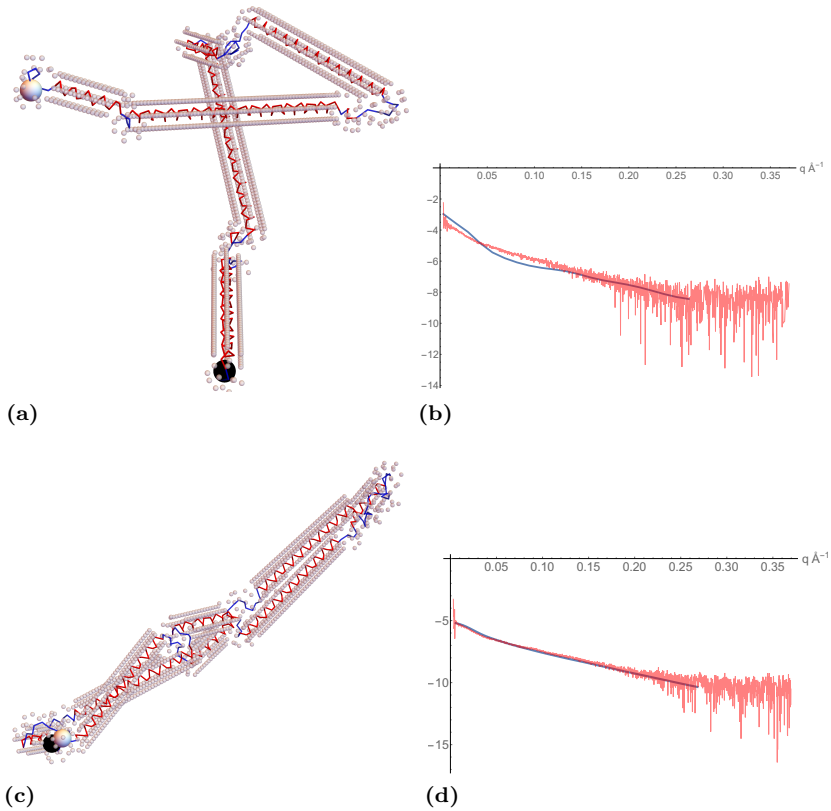


Fig 11. Illustrations of the optimization process used to obtain the model predictions forthe structural core of human SYCE1 . Panel (a) An initial configuration of the backbone based only on the sequence data shown in Fig 10. Also shown as spheres are the molecules of the hydration layer. Large black and white spheres indicate the end termini. (b) the scattering curve of the initial configuration (blue) over-layed on the scattering data (red). (c) the model prediction for which $\chi_f^2 + \chi_{nl} + \chi_{con} < 0.008$, the end termini are next to each other. (d) the final scattering curve compared to the experimental data.

consideration of a sufficient resolution and reliable signal to noise ratio. Using monte-carlo optimization the structure is altered until a reliable fit $\chi_f^2 + \chi_{nl} + \chi_{con} < 0.008$ is obtained, where the potential χ_{con} is based on the contact predictions described above. One such model is shown in Fig 11(c) along with its scattering curve in Fig 11(d). The identical chains of the structure have folded to lie (nearly) parallel with the end termini occupying a local neighbourhood. Three example models for which $\chi_f^2 + \chi_{nl} + \chi_{con} < 0.008$ are shown in Fig 12(a)-(c). Figures 12(d) and (e) indicate one of the coiled coil structures and depict the pairwise distances associated with the contact prediction terms χ_{con}. All models share the elongated bend shape with a anti-parallel coiled-coil arrangement of helix 2-4 to 6-8, respectively. The first helix in each helix (helices 1 and 5, respectively) show different orientations which reflect the expected conformational flexibility of the protein in solution. Importantly, the central coiled coil (made of helices 3 and 7, respectively) is not based on the constraints given a-priori but is entirely based on the optimization against the experimental data. Although a bead model results in a similar overall shape ([22]) our methods is able to derive a more detailed molecular model with distinct structural features such as a coiled coil.

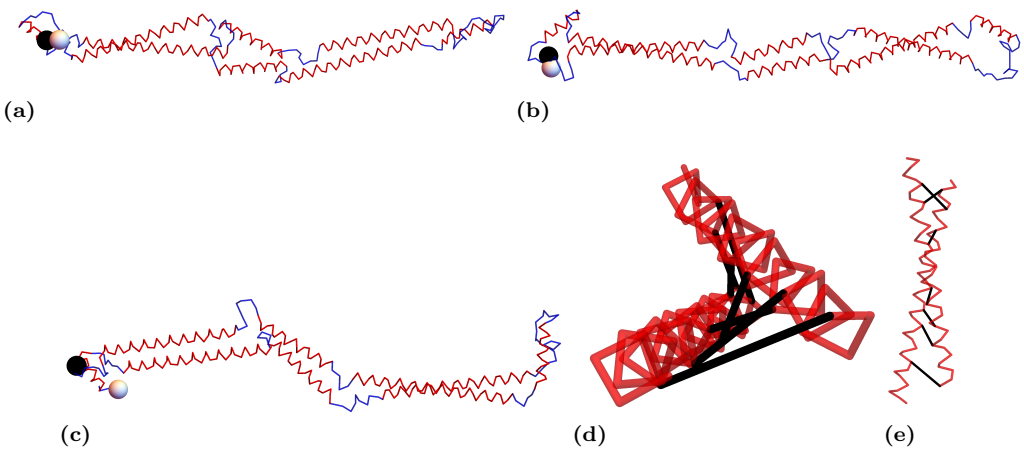


Fig 12. Illustrations of the model predictions. (a)-(c) are all model predictions. (d) One of the coiled coil units of (a) with black tubes representing the contact prediction distances, as see along the axis of the unit. (e) the tilted helical structure of the coiled coil unit.

Discussion

In this paper we have shown that curve representation with hydration shell provides a molecular model for BioSAXS data with fits as good or better than traditional bead and envelope models. Unlike these models our model includes a complete secondary and tertiary model description. Importantly, starting from random models that only take secondary structure information and sequence-dependent distance constraints into account, a physically meaningful 3D model can be obtained by fitting models against the experimental data. That this is possible is due to the fact that the model is described with far fewer parameters compared to even a coarse-grain model that required three coordinates for each amino-acid **combined** with use of geometric constraints for regular secondary structural elements.

In order to show the potential of this ab-initio technique it was applied to a tethered core component of the human SYCE1 protein, for which no high-resolution structural data is available. The model derived was based on sequence information **alone** match those of a model that was previously reported in [22], where the model was based on manual inspection of the sequences coupled with the fitting of ideal helical and coiled coil segments, connected by flexible linkers to experimental scattering data. Importantly, whilst the previously modelled structure includes two coiled coil segments, it was recognised that this was the minimum number of segments required to explain the curved structure and that the true structure could consist of multiple coiled coils interrupted by short linkers. Thus, our novel ab-initio method has successfully generated a highly plausible model from experimental scattering data without the need for any more than minimal manual evaluation. This facility will be crucial for ab-initio structural determination (from biosaxs data) of larger molecules where it would not be practical to generate structures manually.

Further experimental information such as distance information from any other source can easily been added in the form of additional restraints into the optimization algorithm. The model's explicit description of realistic secondary structure means additional information, like contact predictions, radius of gyration, hydrophobicity of the chain and disulfide bonding can be employed as model constraints in the future. This will further enhance the accuracy of all potential models, and in particular help the

end-user to distinguish mathematically correct but physically less likely models from correct solution. The secondary knot fingerprint statistic developed shows significant potential to evaluate structural similarity of models and hence to further automate this vital validation step.

The two future next steps are (i) the application of this method to multimeric structures where each known monomer structure can initially be treated as rigid-body and then refined in order to account for local changes in solution (ii) the application to larger, de-novo structures where the exact 3D structure remains elusive. The application to homo-multimers is straightforward and requires only minor addition to the existing code, we expect this to be the major initial application of our methods. Due to the limited information content of small-angle X-ray scattering data the ab-initio fold determination will depend on the accuracy of secondary structure prediction combined with appropriately weighted distance constraints such as those discussed above.

Methods and Materials

The backbone model

Curvature κ

A section $(\mathbf{c}_i, \mathbf{c}_{i+1}, \mathbf{c}_{i+2}, \mathbf{c}_{i+3})$ defines three edges with midpoints $\mathbf{c}_{ml} = (\mathbf{c}_{i+l-1} + \mathbf{c}_{i+l})/2$, which in turn define the curvature sphere [38] (see Fig 1). The curvature, the inverse of its radius is

$$\kappa(\mathbf{c}_{m1}, \mathbf{c}_{m2}, \mathbf{c}_{m3}) = \frac{2|\sin(\theta_{123})|}{||\mathbf{c}_{m1} - \mathbf{c}_{m2}||} \tag{3}$$

where θ_{123} is the angle between the vectors $\mathbf{c}_{m1} - \mathbf{c}_{m3}$ and $\mathbf{c}_{m2} - \mathbf{c}_{m3}$.

Torsion τ

Three points define a plane (with normal vector $\mathbf{n}$) and the four points $(\mathbf{c}_i, \mathbf{c}_{i+1}, \mathbf{c}_{i+2}, \mathbf{c}_{i+3})$ define two planes through

$$\begin{aligned}
&\mathbf{N}_1 = (\mathbf{c}_{i+1} - \mathbf{c}_i) \times (\mathbf{c}_{i+2} - \mathbf{c}_{i+1}), \\
&\mathbf{N}_2 = (\mathbf{c}_{i+2} - \mathbf{c}_{i+1}) \times (\mathbf{c}_{i+3} - \mathbf{c}_{i+2}), \mathbf{n}_\alpha = \mathbf{N}_\alpha / ||\mathbf{N}_\alpha||, \ \alpha = 1, 2.
\end{aligned}$$

The torsion magnitude is the (length weighted) angle these planes make with each other,

$$\begin{aligned}
&\tau(\mathbf{c}_i, \mathbf{c}_{i+1}, \mathbf{c}_{i+2}, \mathbf{c}_{i+3}) = \frac{2}{l} \sin(\theta_n/2), \\
&l = (||\mathbf{c}_{i+1} - \mathbf{c}_i|| + ||\mathbf{c}_{i+2} - \mathbf{c}_{i+1}|| + ||\mathbf{c}_{i+3} - \mathbf{c}_{i+2}||)/3.
\end{aligned} \tag{4}$$

with θ_n is the angle between $\mathbf{n}_1$ and $\mathbf{n}_2$, see *e.g.* [39]. Its sign is given by the sign of the triple product $\mathbf{n}_1 \times \mathbf{n}_2 \cdot (\mathbf{c}_{i+2} - \mathbf{c}_{i+1})$.

Generating a curve section from (κ, τ) values

Consider a section of curve of length m and $m - 3$ pairs (κ_i, τ_i), whose three initial points $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3$ are randomly chosen (with fixed separation distance $R = 3.8$Å). Since scattering expressions are invariant under an arbitrary translation and rotation ([4]) the exact values of the first two points do not matter (as long as their separation is R). The third point is a structural degree of freedom but it is restricted such that the

C^α-C^αdistance between $\mathbf{c}_1$ and $\mathbf{c}_3$ is greater than R. Once these points are specified the fourth point will be

$$\mathbf{c}_4 = \mathbf{c}_3 + R\left(\sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi), \cos(\theta)\right). \tag{5}$$

with $\theta =\in [0, \pi], \phi \in [0, 2\pi]$. The set $(\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \theta, \phi)$ define four points and hence κ and τ values. Using values of κ_1 and τ_1 equations (3) and (4) are solved for θ and ϕ, this gives $\mathbf{c}_4$. The next point $\mathbf{c}_5$ can similarly be found from the values κ_2 and τ_3, and so on until all $m-3$ (κ_i, τ_i) have been used to yield the m points $\mathbf{c}_i$.

Amino acid form factors

The form factor f_{am} of an amino acid, centered on the C^αatom position, are

$$f_{am}(q) = f_b(q) - \rho_{ex} f_{ex}(q), \tag{6}$$

where f_b is the scattering of the amino acid in a vacuum and f_{ex} is the adjustment due to the excluded volume of solvent. Each amino acid is assigned the same scattering function $f_b(q)$, a five-factor exponential representation

$$f_b = \sum_{i=1}^{5} A_i \mathrm{e}^{-B_i q^2} + C, \tag{7}$$

where $\{A_i, B_i\}_{i=1}^5$ and C are empirically determined constants (a standard form used to fit molecular form factors [40]). The excluded volume effect is captured using an exponential model in the form

$$f_{ex}^a(r_w, q) = v(r_w)\mathrm{e}^{\pi q^2 v(r_w)^{3/2}}, \quad v(r_w) = \frac{4\pi}{3} r_w^3, \tag{8}$$

where r_w is the average atomic radius of the atom [6, 7, 11]. To calculate the excluded volume for amino acids coordinates for all 20 amino acids [41], and values of r_w for Carbon, Nitrogen, Oxygen, Hydrogen and Sulphur (*e.g.* [42]) were used to compute the excluded volume scattering centered at the C^αthrough

$$f_{ex}^{am}(q) = \sum_{i=1}^{N_{am}} f_{ex}^a(r_{wi}, q) \frac{\sin(q r_i^\alpha)}{q r_i^\alpha}, \tag{9}$$

where r_i^α is the distance of atom i from the C^αmolecule and N_{am} the number of atoms in the amino acid. Since f_b does not discriminate individual amino acids this value f_{ex}^{am} was averaged over all 20 amino acids, weighted by their abundance in globular proteins (see [43]). This averaged function, shown in S1 Appendix : Sec3 Fig , gives $f_{ex}(q)$. Finally (6) includes a constant ρ_{ex} which modulates the effect of the excluded volume scatter by comparison to f_b, this value is constrained to lie withing 0.75 and 1.25 (similar constraints are used in [6, 7, 11]). The scattering form for an individual water molecule in the hydration layer is

$$f_h(q) = \rho_h(2 f_{hy}(q) + f_{ox}(q)), \tag{10}$$

where f_{hy} and f_{ox} are the vacuum scattering of Hydrogen and Oxygen respectively [40]. The constant ρ_h was empirically determined (as in [7]). A detailed description of the parameter determination method S1 Appendix: Sec 3.

Knot fingerprint statistic

The knot fingerprint statistic $\mathcal{K}_l$ yields a number between 0 and 1, with 1 implying the two knot fingerprints of interest are identical on a square-by-square basis, and 0 if they are completely dissimilar. It accounts for the number of knots in each square of the lth fingerprint and the similarity in number (per square) between the two (l^{th} level) fingerprints, a value of 1 implies these values are identical for all squares.

Consider two discrete curves $K_1 = \left\{\mathbf{c}^1\right\}_i^n$ and $K_2 = \left\{\mathbf{c}^2\right\}_i^n$. For each subsection $\left\{\mathbf{c}_i^1 |\, i \in [j,k]\right\}$, $\left\{\mathbf{c}_i^2 |\, i \in [j,k]\right\}$ label i_1^{jk}, i_2^{jk} as the knot type of the lth most common knot and n_1^{jk}, n_2^{jk} the number of times that it occurs. Define a "weight function" $w_{jk} \in [0,1]$,

$$w_{jk} = \exp\left[\ln 2 \left(\frac{n_1^{jk} + n_2^{jk}}{S_{max}}\right)^p\right] - 1, \tag{11}$$

$$S_{max} = \max_{j,k \in [1,n],\, k-j>3} n_1^{jk} + n_2^{jk}. \tag{12}$$

An exponentially increasing weight is assigned based on the relative number of knots produced for a given subsection of K_1 and K_2 (relative to the maximum number over all subsections S_{max}). The higher the value of p the greater the relative weight given to larger sums high-resolution crystal structures from the PDB $n_1^{kj} + n_2^{kj}$, all calculations in this note used $p = 5$, a value which was found to give a statistic which had excellent properties in distinguishing folds (see S1 Appendix :Sec 4). Using this weighting the following matching function function c_{jk} is defined

$$c_{jk} = \begin{cases} w_{jk}\left(1 - \left|\frac{n_1^{jk} - n_2^{jk}}{n_1^{jk} + n_2^{jk}}\right|\right) & i_1^{jk} = i_2^{jk}, \\ 0 & i_1^{jk} \neq i_2^{jk}. \end{cases} \tag{13}$$

This assigns a weighting if there is a match in knot type and gives preference to matches which have similar numbers n_1^{kj}, n_2^{kj}. The knot fingerprint statistic $\mathcal{K}_l(K_1, K_2)$ is following real number which takes values on $[0,1]$

$$\mathcal{K}_l(K_1, K_2) = \frac{\sum_{k=4}^{n} \sum_{j=1}^{k-3} c_{jk}}{\sum_{k=4}^{n} \sum_{j=1}^{k-3} w_{jk}}. \tag{14}$$

This can only be one if all i_1^{jk}, i_2^{jk} are the same and have exactly the same individual number of occurrences $n_1^{jk} = n_2^{jk}$.

Overlap potential χ_{nl}

To prevent non-physical conformations, an additional penalty χ_{nl} is applied if a C^α-C^αdistance less than 3.8Å occurs for any pair of non-adjacent C^αpositions, *i.e.*

$$\chi_{nl} = \begin{cases} 1000 & \text{if } ||\mathbf{c}_i - \mathbf{c}_j|| < 3.8, |i-j| \geq 2,\ i,j \in \{1,n\}. \\ 0 & \text{Otherwise.} \end{cases} \tag{15}$$

Contact potential χ_{con}

If $l = 1, \ldots n_c$ labels the n_c pairs of constrained points with mutual distances d_l^c then the quality of contact match χ_{con} is defined as follows:

$$\chi_{con} = \frac{C}{n_c} \sum_{l=1}^{n_c} \left(d_l^c - d_f^c\right)^2, \tag{16}$$

with C a constant and d_c^f a reference distance (7Å was used in this study). The value of C controls the likely variation in the distances d_c^l, a value of $C = 0.001$ in this study was found to give good results.

Biosaxs data for the core component of SYCE1

The details on how this data was obtained can be found in [22].

Supporting information

S1 Appendix. Attached document [online this will be a link to the PDF]

Acknowledgments

The authors would like to thank Daniel Bruce for his help with contact predictions and Christina Law for testing the scattering and hydration models. Financial support from the Biophysical Sciences Institute and the Addison-Wheeler fellowship for CP are gratefully acknowledged. O.R.D is a Sir Henry Dale Fellow jointly funded by the Welcome Trust and Royal Society (grant number 104158/Z/14/Z) and also supported by a Royal society Research Grant (Grant number RG170118). We would also like to thank Dr Mark Miller whose knot identification code was used to calculate knot fingerprints as well as Prof Alain Goriely and Dr Andrew Hausrath for their input and advice in the development of the backbone model. Finally we are grateful to Dr Beth Bromley for her help with the contact predictions within the coiled coil of SYCE1.

References

1. Petoukhov MV, Svergun DI. Applications of small-angle X-ray scattering to biomacromolecular solutions. The international journal of biochemistry & cell biology. 2013;45(2):429–437.

2. Kikhney AG, Svergun DI. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. FEBS letters. 2015;589(19PartA):2570–2577.

3. Mina JG, Thye JK, Alqaisi AQ, Bird LE, Dods RH, Grøftehauge MK, et al. Functional and phylogenetic evidence of a bacterial origin for the first enzyme in sphingolipid biosynthesis in a phylum of eukaryotic protozoan parasites. Journal of Biological Chemistry. 2017;292(29):12208–12219.

4. Svergun DI, Koch MH, Timmins PA, May RP. Small angle X-ray and neutron scattering from solutions of biological macromolecules. vol. 19. Oxford University Press; 2013.

5. Rambo RP, Tainer JA. Bridging the solution divide: comprehensive structural analyses of dynamic RNA, DNA, and protein assemblies by small-angle X-ray scattering. Current opinion in structural biology. 2010;20(1):128–137.

6. Svergun D, Barberato C, Koch MH. CRYSOL–a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. Journal of applied crystallography. 1995;28(6):768–773.

7. Schneidman-Duhovny D, Hammel M, Sali A. FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic acids research. 2010;38(suppl 2):W540–W544.

8. Poitevin F, Orland H, Doniach S, Koehl P, Delarue M. AquaSAXS: a web server for computation and fitting of SAXS profiles with non-uniformally hydrated atomic models. Nucleic acids research. 2011; p. gkr430.

9. Wright DW, Perkins SJ. SCT: a suite of programs for comparing atomistic models with small-angle scattering data. Journal of applied crystallography. 2015;48(3):953–961.

10. Franke D, Svergun DI. DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. Journal of applied crystallography. 2009;42(2):342–346.

11. Svergun DI, Petoukhov MV, Koch MH. Determination of domain structure of proteins from X-ray solution scattering. Biophysical journal. 2001;80(6):2946–2953.

12. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. Journal of molecular biology. 1963;7(1):95–99.

13. Hausrath A, Goriely A. Continuous representations of proteins: Construction of coordinate models from curvature profiles. Journal of structural biology. 2007;158(3):267–281.

14. Lundgren M, Krokhotin A, Niemi AJ. Topology and structural self-organization in folded proteins. Physical Review E. 2013;88(4):042709.

15. Kneller GR, Hinsen K. Protein secondary-structure description with a coarse-grained model. Acta Crystallographica Section D: Biological Crystallography. 2015;71(7):1411–1422.

16. Bolcun-Filas E, Speed R, Taggart M, Grey C, de Massy B, Benavente R, et al. Mutation of the mouse Syce1 gene disrupts synapsis and suggests a link between synaptonemal complex structural components and DNA repair. PLoS genetics. 2009;5(2):e1000393.

17. Costa Y, Speed R, Öllinger R, Alsheimer M, Semple CA, Gautier P, et al. Two novel proteins recruited by synaptonemal complex protein 1 (SYCP1) are at the centre of meiosis. Journal of cell science. 2005;118(12):2755–2762.

18. Park HH. X-ray crystallographic studies of the middle part of the human synaptonemal complex protein 1 coiled-coil domain. Acta Crystallographica Section F: Structural Biology Communications. 2015;71(9):1131–1134.

19. Davies OR, Maman JD, Pellegrini L. Structural analysis of the human SYCE2–TEX12 complex provides molecular insights into synaptonemal complex assembly. Open biology. 2012;2(7):120099.

20. Syrjänen JL, Pellegrini L, Davies OR. A molecular model for the role of SYCP3 in meiotic chromosome organisation. Elife. 2014;3:e02963.

21. Dunce JM, Dunne OM, Ratcliff M, Millán C, Madgwick S, Usón I, et al. Structural basis of meiotic chromosome synapsis through SYCP1 self-assembly. Nature structural & molecular biology. 2018; p. 1.

22. Dunne OM, Davies OR. Molecular structure of human synaptonemal complex protein SYCE1. Chromosoma. 2019; p. 1–14.

23. Majorek KA, Porebski PJ, Dayal A, Zimmerman MD, Jablonska K, Stewart AJ, et al. Structural and immunologic characterization of bovine, horse, and rabbit serum albumins. Molecular immunology. 2012;52(3-4):174–182.

24. Valentini E, Kikhney AG, Previtali G, Jeffries CM, Svergun DI. SASBDB, a repository for biological small-angle scattering data. Nucleic acids research. 2014;43(D1):D357–D363.

25. Wand MP, Jones MC. Kernel smoothing. Crc Press; 1994.

26. Elias M, Wellner A, Goldin-Azulay K, Chabriere E, Vorholt JA, Erb TJ, et al. The molecular basis of phosphate discrimination in arsenate-rich environments. Nature. 2012;491(7422):134.

27. Debye P. Zerstreuung von röntgenstrahlen. Annalen der Physik. 1915;351(6):809–823.

28. Kufareva I, Abagyan R. Methods of protein structure comparison. In: Homology Modeling. Springer; 2011. p. 231–257.

29. Zemla A, Venclovas Č, Moult J, Fidelis K. Processing and evaluation of predictions in CASP4. Proteins: Structure, Function, and Bioinformatics. 2001;45(S5):13–21.

30. Millett KC, Rawdon EJ, Stasiak A, Sułkowska JI. Identifying knots in proteins; 2013.

31. Tubiana L, Orlandini E, Micheletti C. Probing the entanglement and locating knots in ring polymers: a comparative study of different arc closure schemes. Progress of Theoretical Physics Supplement. 2011;191:192–204.

32. Jamroz M, Niemyska W, Rawdon EJ, Stasiak A, Millett KC, Sułkowski P, et al. KnotProt: a database of proteins with knots and slipknots. Nucleic acids research. 2014;43(D1):D306–D314.

33. Sułkowska JI, Rawdon EJ, Millett KC, Onuchic JN, Stasiak A. Conservation of complex knotting and slipknotting patterns in proteins. Proceedings of the National Academy of Sciences. 2012;109(26):E1715–E1723.

34. Diamond R. Real-space refinement of the structure of hen egg-white lysozyme. Journal of molecular biology. 1974;82(3):371–391.

35. Rambo RP, Tainer JA. Accurate assessment of mass, models and resolution by small-angle scattering. Nature. 2013;496(7446):477.

36. Peng J, Xu J. RaptorX: exploiting structure information for protein alignment by statistical inference. Proteins: Structure, Function, and Bioinformatics. 2011;79(S10):161–171.

37. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, et al. A completely Reimplemented MPI bioinformatics toolkit with a new HHpred server at its Core. Journal of molecular biology. 2018;430(15):2237–2243.

38. Gonzalez O, Maddocks JH. Global curvature, thickness, and the ideal shapes of knots. Proceedings of the National Academy of Sciences. 1999;96(9):4769–4773.

39. Carroll D, Hankins E, Kose E, Sterling I. A survey of the differential geometry of discrete curves. The Mathematical Intelligencer. 2014;36(4):28–35.

40. Brown P, Fox A, Maslen E, O'Keefe M, Willis B. Intensity of diffracted intensities. In: International Tables for Crystallography Volume C: Mathematical, physical and chemical tables. Springer; 2006. p. 554–595.

41. Kleywegt GJ, Jones TA. Databases in protein crystallography. Acta Crystallographica Section D: Biological Crystallography. 1998;54(6):1119–1131.

42. Fraser R, MacRae T, Suzuki E. An improved method for calculating the contribution of solvent to the X-ray diffraction pattern of biological molecules. Journal of Applied Crystallography. 1978;11(6):693–694.

43. Schwartz R, Istrail S, King J. Frequencies of amino acid strings in globular protein sequences indicate suppression of blocks of consecutive hydrophobic residues. Protein Science. 2001;10(5):1023–1031.