PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

BCL::SAXS: GPU accelerated Debye method for computation of small angle X-ray scattering profiles

Daniel K. Putnam,[1] Brian E. Weiner,[2] Nils Woetzel,[2] Edward W. Lowe Jr.,[2] and Jens Meiler[1,2]

[1] Department of Biomedical Informatics, Vanderbilt University, Nashville, Tennessee 37235

[2] Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235

ABSTRACT

Small angle X-ray scattering (SAXS) is an experimental technique used for structural characterization of macromolecules in solution. Here, we introduce BCL::SAXS—an algorithm designed to replicate SAXS profiles from rigid protein models at different levels of detail. We first show our derivation of BCL::SAXS and compare our results with the experimental scattering profile of hen egg white lysozyme. Using this protein we show how to generate SAXS profiles representing: (1) complete models, (2) models with approximated side chain coordinates, and (3) models with approximated side chain and loop region coordinates. We evaluated the ability of SAXS profiles to identify a correct protein topology from a non-redundant benchmark set of proteins. We find that complete SAXS profiles can be used to identify the correct protein by receiver operating characteristic (ROC) analysis with an area under the curve (AUC) > 99%. We show how our approximation of loop coordinates between secondary structure elements improves protein recognition by SAχS for protein models without loop regions and side chains. Agreement with SAXS data is a necessary but not sufficient condition for structure determination. We conclude that experimental SAXS data can be used as a filter to exclude protein models with large structural differences from the native.

Proteins 2015; 83:1500–1512.

Key words: proteins; SAXS; Debye formula; GPU acceleration.

INTRODUCTION

Protein structure determination remains a major challenge in the field of structural biology.[1] While X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy can provide high resolution structures, these techniques can be limited by size,[2] high flexibility,[3] and membrane environment.[3] Computational de novo protein structure prediction methods have been developed, but are limited by the vast conformational search space that needs to be searched when no template structure is available.[4] To overcome these experimental and computational limitations, hybrid methods—that is, the combination of multiple techniques—can be utilized to gain structural insights of proteins.[5–7]

SAXS offers an alternative to traditional structure determination techniques

Small angle X-ray scattering (SAXS) is an experimental structural characterization method for rapid analysis of biological macromolecules in solution.[8–12] During data acquisition in SAXS, macromolecules move freely in solution while a beam of X-rays with constant wavelength λ irradiate the sample. At the point of interaction between X-rays and electrons in the sample, both elastic and inelastic scattering occur. This work considers the case of elastic scattering by electrons. The intensity of the scattered X-rays captured on the detector is proportional to the Fourier Transform of a pairwise distance function $\rho(r)$ that gives the probability of finding two atoms a certain distance apart. This distance function is weighted

Grant sponsor: NLM (to D.K.P.); Grant number: 5T15LM007450-09; Grant sponsor: NIH; Grant number: R01 GM080403, R01 MH090192, and R01 GM099842; Grant sponsor: NSF; Grant number: Career 0742762 and OIA 0959454.
*Correspondence to: Jens Meiler, Vanderbilt University 7330 Stevenson Center, Station B 351822 Nashville, TN 37235. E-mail:jens@meilerlab.org
Received 13 March 2015; Revised 8 May 2015; Accepted 19 May 2015
Published online 27 May 2015 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24838

by the excess scattering density of the respective scattering volume compared to the solvent.

For a more comprehensive review of SAXS theory we recommend several reviews.[8,9,13–16] SAXS profiles are reported by intensity (I) as a function of momentum transfer vector (q). Large interatomic distances contribute to the intensity profile at small q, while short interatomic distances contribute to the intensity at large q. Several parameters can be extracted directly from the scattering profile including: the molecular mass (MM), radius of gyration (R_g), hydrated particle volume (V_p), and maximum particle diameter (D_{max}). The state of the protein (folded vs. unfolded) can be observed from the Kratky representation of the scattering data plotting q vs. $q^2I(q)$. The scattering profile can be transformed into the pairwise distance density function which is a histogram of distances between pairs of points in a particle. This shape information has been used for the validation of structural models.[17,18]

Use of SAXS experimental data in computation

The experimental SAXS profile has been used to filter a set of proposed models by comparing the computed SAXS profile of each model with the experimental data.[5,19] Furthermore, the experimental profile has been incorporated into an energy function for protein folding to obtain a model consistent with experimental data.[20] More recently SAXS has been used to identify and model protein flexibility from an ensemble set of conformers.[21] In this approach a large library of initial conformers are given as input. After a sufficient library of conformers has been found, the experimental SAXS data are used to ascertain which combination of conformers optimally fit the data. In this case, the scattering intensity (I) is represented by a linear combination of the selected conformers. The crucial step in this analysis is computation of a SAXS profile from a proposed protein model.

Protein structure prediction

De novo protein structure prediction methods have two major components—a sampling algorithm and a scoring function. During the sampling phase, the protein model is perturbed. The protein is then scored, using a scoring function designed to identify native-like topologies. This process is iterated in order to minimize the scoring function. The challenge in this process is sampling the large conformational space of a protein densely enough so that one model approaches the native conformation. To be time-efficient, the protein model is often simplified to remove conformational degrees of freedom (coarse grained sampling) and the scoring function is therefore rapid but inaccurate. Sampling for larger proteins is further complicated by nonlocal contacts, amino acids in contact in Euclidean space (<8 Å), that are far apart in sequence (>12 residues). As the number of non-local contacts increase, the accuracy of de novo protein structure prediction methods drastically decreases.[22] Atomic detail is added in a later stage of the protocol and the model is rescored/optimized with a higher accuracy scoring function. The accuracy necessary to identify the correct topology by its superior energy at this stage is a RMSD value of approximately 2 Å when compared with the native structure.

BCL::Fold is designed to address the sampling bottleneck

BCL::Fold is a protein structure prediction method that rapidly assembles secondary structure elements (SSEs) into topologies.[23,24] This approach provides a means to focus sampling on long range contacts between amino acid pairs. To begin, a pool of predicted SSEs is generated from an input FASTA sequence of amino acids. SSEs are randomly selected from the pool and assembled using a Monte Carlo Metropolis (MCM) assembly protocol to produce a coarse grained representation of the protein without side chain atoms and loop region residues. During assembly the model is evaluated using a consensus knowledge-based scoring function. This process is repeated 10,000 to 100,000 times. The underlying hypothesis of BCL::Fold is that the interactions between SSEs determine the majority of the protein core and give rise to its thermodynamic stability. Once the models have been generated, they are clustered by RMSD100 into N cluster centers. The medoid from each cluster center is selected for loop construction and side chain addition using Rosetta.[25] This produces a set of proposed conformations for a given protein sequence in the absence of experimental data.

BCL::SAXS is a GPU accelerated Debye implementation for profile reconstruction

The use of experimental SAXS profiles during the construction of protein models with BCL::Fold would provide additional constraints on the sampling space of a given protein sequence. To incorporate experimental SAXS restraints into BCL::Fold, we must first develop a method to compare experimental SAXS profiles with profiles generated from protein models produced by BCL::Fold, i.e. missing loop region and side chain residues.

Here, we describe our newly developed algorithm BCL::SAXS. It computes complete SAXS scattering profiles for complete protein models and an approximate scattering profile for protein models that consist of secondary structure elements only as used in BCL::Fold.[23,24,26,27] The main methods to calculate a SAXS scattering profile from atomic coordinates are

spherical harmonics with multipole expansion, Monte Carlo methods, and the Debye formula.[28–31] Multipole expansion methods have been shown to be highly accurate, but difficult to modify for incomplete protein models. The Debye formula is easy to modify, but comes with a high computational cost. Ultimately we want to compare SAXS Profiles generated from BCL::Fold models[23,24]—i.e. protein structure that lack loops and side chains—with experimental SAXS profiles. To facilitate this, we chose to use the Debye formula, implement approximations for missing loops and side chain atoms, and address the computational cost with graphical processing unit (GPU) acceleration.

Overall approach

In BCL::SAXS interatomic pairwise distances are computed explicitly for each heavy atom using the Debye formula for atomic scatterers.[32] It models the hydration layer based on the solvent accessible surface area of each atom. To maximize the fit to experimental data BCL::SAXS optimizes the hydration layer density and the excluded volume of the protein. We accelerate the algorithm performance by using GPU parallel threads. We demonstrate the discriminatory power of SAXS at three different abstraction levels consistent with the BCL::Fold folding protocol:[23] (1) complete protein models, (2) protein models with approximated side chain coordinates, (3) protein models with approximated side chain coordinates and approximated loop regions. We quantify the performance of the protocol from a set of 455 proteins with SAXS profiles computed *in silico* and experimental data from hen egg white lysozyme. Furthermore, our work introduces a new approximation of the coordinates of residues in loop regions for crude protein models missing these residues. BCL::SAXS is available to the scientific community via the BCL::Commons user interface (www.meilerb.org). It is free for academic use.

MATERIALS AND METHODS

To accurately determine the SAXS profile from the atomic coordinates of full atom protein models we utilized several key equations—the Debye formula for atomic scatterers and three equations to calculate the form factors.[28,29,32–35] The form factors are continuous functions of the magnitude of the momentum transfer vector $\vec{q}$. Using the Euclidean atomic coordinates from structures stored in the protein data bank (PDB),[36] scattering profiles are reconstructed. The following equations, starting with the Debye formula, depict the method:

$$I(q) = \sum_{i=1}^{M}\sum_{j=1}^{M} F_i(q)F_j(q)\frac{\sin\left(qr_{ij}\right)}{qr_{ij}} \qquad (1)$$

where the intensity, $I(q)$ is a function of the magnitude of the momentum transfer vector $\vec{q}$. It is given by $|\vec{q}| = (4\pi\sin\theta)/\lambda$, where θ is given by a scattering angle of 2θ, and λ is the wavelength of the incident beam. $F_i(q)$ and $F_j(q)$ are the atomic form factors and r_{ij} is the pairwise Euclidean distance between atom i and atom j. M is the number of atoms in the protein and the summations run over all atoms. To calculate the form factors, we subtracted the displaced solvent contribution from the form factor in vacuo and added the contribution of the hydration layer:

$$F_i(q) = f_{v,i}(q) - c_1 f_{s,i}(q) + c_2 S_i f_{w,i}(q) \qquad (2)$$

where $f_{v,i}(q)$ is the atomic form factor in vacuo, $f_{s,i}(q)$ is the form factor of the hypothetical atom that represents the displaced solvent,[30] and $f_{w,i}(q)$ is the contribution from the hydration layer. S_i is the solvent accessible surface area of the given atom. C_1 is used to modify the total excluded volume of the atoms and C_2 is used to modify the water density in the hydration shell. The atomic form factor in vacuo approximation is based on the combination of relativistic Dirac-Slater wave functions and numerical Hartree-Fock wave functions.[33,34,37,38] These Hartree-Fock scattering factors were previously computed from $q = 0$ to $q = 1.5$ at intervals of 0.01Å^{-1}.[39] For convenience, these scattering factors were previously fit to the 5-gaussian (Cromer-Mann) analytic function:

$$f_{v,i}(q) = \sum_{i=1}^{4} a_i \cdot e^{-b_i\left(\frac{q}{4\pi}\right)^2} + c \qquad (3)$$

where a, b, and c are the constants for each atom, and q is the momentum transfer vector. This approximation is only valid with a q range from 0 to 2.0 Å[33,34,37] which is sufficient for SAXS scattering experiments where the valid scattering angle range is from 0 to ≈ 0.33 Å.[8,9] For larger scattering angles, a 6-gaussian approximation must be used which is valid from 0 to ≈ 6.0 Å.[38] The displaced solvent scattering $f_{s,i}(q)$ was approximated by V_i,[30] the excluded solvent volume V displaced by atom i:

$$f_{s,i}(q) = q_s V_i e^{\frac{-q^2 V_i^{2/3}}{4\pi}} \qquad (4)$$

where q_s is the solvent density of 0.334e Å^{-3}.[35] The combination of these equations yields a SAXS scattering profile from rigid body data stored in a pdb file.

GPU parallel processing to accelerate algorithm

The pairwise nature of the Debye formula has a computational cost of $O(N^2)$ for each value of q evaluated, where N represents the number of atoms contained in

the protein. This high computational cost and time requirement has precluded the use of the direct calculation of SAXS profiles using the Debye formula during folding simulations. To circumvent this computational limitation, alternative approaches for this calculation including multipole expansion methods for spherical harmonics[30] and approximation of the individual form factors have been developed.[29] In contrast, to directly compute the SAXS profile using the Debye formula we leverage here the parallel architecture of graphical processing unit (GPU) threads using OpenCL and computed SAXS profiles directly.

GPU implementations of the Debye formula for SAXS profile reconstruction

In 2013, Antonov *et al.* showed how to use GPU acceleration to evaluate SAXS profiles in a Markov Chain Monte Carlo framework.[40] From a protein structure created *in silico*, they reconstructed the SAXS profile using the Debye formula and GPU Acceleration. To address the $O(N^2)$ complexity of the Debye formula they created a coarse grain representation of the protein model with a one or two-body "dummy atom" approximation for each residue. The two-body representation required the development of 21 form factors to represent each new atom type—one for Alanine, one for Glycine, one for the Backbone, and 18 for the remaining side chains. These form factors were derived using a Monte Carlo simulation of a set of 297 high resolution crystal structures from the Protein Data Bank (PDB).[29] This algorithm was benchmarked on problem sizes ranging from 64 to 8192 scattering bodies. The speed up ranges from 16× to 394×. A protein represented by 1888 bodies with 51 discrete q values took 2408 ms on a central processing unit (CPU) and 9 ms with GPU acceleration.

BCL::SAXS GPU implementations of the Debye formula for SAXS profile reconstruction

To build upon the previous work we parameterize the excluded volume and hydration shell in the form factor calculation and operate on individual atoms. For full atom representations of proteins we can account for deviations in electron density and hydration shell thickness. The Debye formula can be visualized as an $N \times N$ square matrix of N-atom rows by N-atom columns where N is the number of atoms in the protein. The pairwise Euclidean distances are calculated for each entry in the matrix with the diagonal represented by zeros. Pairwise distance calculations in a matrix form are an ideal calculation type for GPU acceleration because each GPU thread can calculate a single Euclidean distance with the only limitation being memory. To address memory requirements, the algorithm was restructured to have each thread calculate a Debye partial sum for a current atom i:

$$I_{\text{partial}} = F_i(q) \sum_{j=1}^{M} F_j(q) \frac{\sin\left(qr_{ij}\right)}{qr_{ij}} \tag{5}$$

This technique enables the application of this accelerated algorithm to very large multimeric systems in excess of 90,000 atoms with the current GPU memory constraints while leveraging device shared memory in a tiling technique. The result of this partial sum is a matrix of q rows by N-atom columns where q is the momentum transfer vector and N is the total number of atoms. These partial sums are then summed across each column to completion for each q using a GPU reduction sum kernel to arrive at the desired q number of sums.

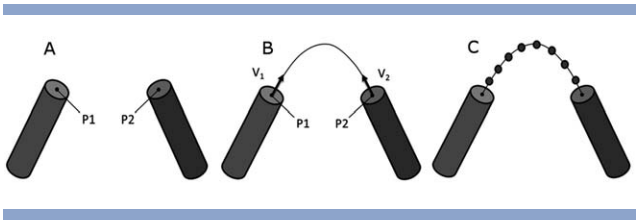

Figure 1

Construction of curvilinear path and placement of residues in region between two SSEs. (**A**) Protein model with two α-helical structures, p_1 and p_2. (**B**) Approximated path with unit vectors v_1 and v_2 pointing in the helical direction of SSE_1 and the helical direction of SSE_2. (**C**) Residues placed equidistant along the curvilinear path between SSEs.

Generation of SAXS scattering profile from atomic coordinates with CRYSOL

To measure the time the algorithm takes on different types of GPUs, experimental scattering curves were approximated from high resolution protein structures in the PDB using the program CRYSOL.[30] This program computes the scattering profile using spherical harmonics and multipole expansion for fast calculation of the spherically averaged scattering profile.

Approximate SAXS scattering profiles for protein models without side chain and loop regions

To approximate the side chain regions of a given amino acid, the form factors for the atoms with missing side chain coordinates were added to the C_β position of the respective amino acid. This approach is analogous to how the form factors for hydrogen are folded into their respective heavy atom in CRYSOL.[30] The loop regions were approximated by removing atomic coordinate data between secondary structure elements (SSEs) and computing a path from the c-terminus of the first SSE to the n-terminus of the second SSE. The amino acid residues in the loop regions were placed at points along the path

(Fig. 1). While crude, this approach is much more rapid than actual construction of loops.

Vector calculations to approximate the loop path between two secondary structure elements

P_1 represents the C_β position vector of the last residue in the N-terminal SSE, while P_2 represents the C_β position vector of the first residue in the C-terminal SSE.

$$\overrightarrow{P_{1,n}} = \{x_1, y_1, z_1\} \quad (6)$$
$$\overrightarrow{P_{2,c}} = \{x_1, y_1, z_1\} \quad (7)$$

CP_1 represents the center position vector of the last residue in the N-terminal SSE, while CP_2 represents the center position vector of the first residue on the C-terminal SSE.

$$\overrightarrow{CP_{1,n}} = \{x_2, y_2, z_2\} \quad (8)$$
$$\overrightarrow{CP_{2,c}} = \{x_2, y_2, z_2\} \quad (9)$$

We computed a vector pointing in the same orientation of the SSE by subtracting the C_β position of the center of the SSE from P_1 and P_2.

$$\overrightarrow{V_n} = \overrightarrow{P_n} - \overrightarrow{CP_n} \quad (10)$$

where n is the index of the point. The direction of the vectors V_1 and V_2 were computed by dividing them by their magnitude.

$$\overrightarrow{D_n} = \frac{\overrightarrow{V_n}}{\sqrt{V_{nx}^2 + V_{ny}^2 + V_{nz}^2}} \quad (11)$$

The scalar distance (D_{sse}) between two SSEs was computed by subtracting P_2 from P_1 and then taking the norm of the resulting vector. The percentage to move from P_1 toward P_2 at each step (L) along path (S) was computed by dividing one by one more than the number of amino acids in the loop region.

$$L = \frac{1}{N_{aa}+1} \quad (12)$$

The predicted Euclidean loop length (P) was computed by multiplying the number of amino acids by the C_α–C_α spacing of 3.2 Å. The 3.2 Å term is the average distance between amino acids in the coil region of a protein. It was computed by averaging the C_α distance between residues in the engrailed homeodomain (PDB ID: 1ENH).[41]

$$P = N_{aa} \times 3.2 \text{ Å} \quad (13)$$

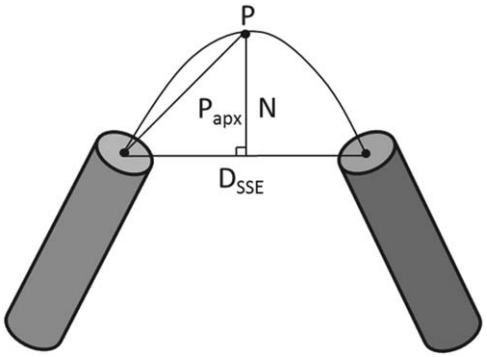

Figure 2

Depiction of the parabolic height approximation method. D_{sse} is the Euclidean distance between SSEs, P_{apx} is the estimated length of the hypotenuse side of a right triangle. N is the normalization factor and controls the height of the parabola.

Pathway calculations for loop approximation

The path length (S) between two SSEs was approximated as a curve starting in the direction of SSE_1 and ending in the direction of SSE_2. The curve calculation consists of a linear, parabolic, and a directional component. The linear component is given by:

$$\overrightarrow{l(L)} = (1-L)\overrightarrow{P}_1 + L\overrightarrow{P}_2 \quad (14)$$

where L is the percentage between [0, 1]. When $L = 0$, the equation reduces to the Euclidean vector coordinates of the starting point. When $L = 1$, the equation reduces to the Euclidean vector coordinates of the end point. The parabolic component is given by:

$$p(L) = N \times L(1-L) \quad (15)$$

where N is a normalization factor to size the height of the parabola and control parabolic path length. The directional component is given by:

$$\overrightarrow{d(L)} = \left[(1-L)\overrightarrow{d_1} + L\overrightarrow{d_2}\right] \quad (16)$$

where d_1 and d_2 are unit directional vectors pointing in the direction of SSE_1 and SSE_2, respectively. The complete parabolic approximation function is:

$$\overrightarrow{P(L)} = (1-L)\overrightarrow{P_1} + L\overrightarrow{P_2} + NL(1-L) \times \left[(1-L)\overrightarrow{d_1} + L\overrightarrow{d_2}\right] \quad (17)$$

Normalization factor and path length calculations

The normalization factor (N) controls the height of the curve and corresponding path length. To calculate N for a given loop region we divided the curve in half and

approximated the arc to be the hypotenuse of a right triangle. The base of the triangle was the Euclidean distance between the SSEs divided by two (Fig. 2). With these approximations, the normalization factor (N) is given by the Pythagorean Theorem:

$$N = \frac{1}{2}\sqrt{P^2 - D_{sse}^2} \tag{18}$$

where N is the normalization factor, P is the predicted loop length, and D_{sse} is the Euclidean distance between P_1 and P_2.

Model quality was assessed by the χ agreement between the calculated and experimental SAXS curves

To compare the scattering profiles, we first normalized the experimental and calculated scattering intensities to be between (0, 1). To magnify the effects of small distances, (higher q values), the scattering intensities (I) for both data sets were converted to a $\log_{10}$ scale. To account for concentration differences in experimental data, the calculated curve was multiplied by a scaling weight (c) that minimizes the χ score.[28,30]

$$c = \left[\sum_{k=1}^{Q} \frac{I_{cal}(q_k) \cdot I_{exp}(q_k)}{\sigma_{exp}^2(q_k)}\right]\left[\sum_{k=1}^{Q} \frac{I_{cal}^2(q_k)}{\sigma_{exp}^2(q_k)}\right]^{-1} \tag{19}$$

where I_{cal} is the intensity of the calculated curve, I_{exp} is the intensity of the experimental curve, σ is the experimental error, and q is the momentum transfer vector. Using cubic splines, the derivative of the intensities for both data sets were computed. Similar to other approaches to modeling proteins from a SAXS scattering profile,[11,42,43] we score a model based on the χ score between the experimental profile and the profile computed by our algorithm BCL::SAXS.

$$\chi = \sqrt{\frac{1}{Q}\sum_{i=1}^{q}\left(\frac{I_{exp}(q_i) - cI_{cal}(q_i)}{\sigma(q_i)}\right)^2} \tag{20}$$

where Q is the number of entries in the data set and σ is the experimental error of the measured profile. In cases where no experimental error is provided it is simulated. We compute the χ score from different states of the experimental and calculated scattering profiles. The first state on the absolute scale is to compute the χ score right after the initial profile reconstruction with the Debye formula and scaling. The second state is to compute the χ score after converting the both experimental and computed data to the $\log_{10}$ scale. The third state is to compute the χ score after taking the derivative of the $\log_{10}$ representation of the experimental and calculated curves.

Table I
SSE Definitions for Hen Egg White Lysozyme

Type	SSE number	Start residue	Sequence location	End residue	Sequence location
Helix	1	ARG	5	HIS	15
Helix	2	LEU	25	SER	36
Helix	3	CYS	80	LEU	84
Helix	4	ILE	88	ASP	101
Helix	5	VAL	109	CYS	115
Helix	6	ASP	119	ARG	125
Strand	1	LYS	1	PHE	3
Strand	2	PHE	38	THR	40
Strand	3	ALA	42	ASN	46
Strand	4	SER	50	GLY	54
Strand	5	GLN	57	SER	60

For complete models, we identify the optimal χ values by optimizing combinations of the excluded volume parameter, C_1 and the hydration layer parameter, C_2 inside a boundary ($0.8 \leq C_1 \leq 1.2$ and $0 \leq C_2 \leq 4.0$). Using these parameters we compute the scaling parameter c that minimizes χ for each C_1, C_2 combination.

RESULTS

To illustrate the use of BCL::SAXS, we show the results using hen egg white lysozyme (PDB ID: 6LYZ, molecular weight 14 kDa). The X-ray scattering results for this protein were obtained from an open access database, BIOISIS, containing experimental SAXS data for hen egg white lysozyme (BIOSIS ID: LYSOZP). The SAXS profile for this protein was collected at the SIBYLS Beamline ASL BL12.3.1 and the experimental setup has been previously described.[44] To account for uncertainty in the PDB definitions of secondary structure of 6LYZ, we added additional SSEs by taking the consensus prediction of the secondary structure server 2Struc.[45] This meta server runs secondary structure prediction using the Dictionary of Secondary Structure of Proteins (DSSP),[46] DSSPcont,[47] Stride,[48] P-SEA,[49] PALSSE,[50] STICK,[51] KAKSI,[52] and TM-Align.[53] The final SSE definitions used for analysis are shown in Table I. The final model with loop approximations is shown in Figure 3.

The SAXS comparison derivative χ score

When comparing SAXS profiles between two distinct proteins, the common method is to use the χ formula previously shown.[28,30,54] However, when computing a SAXS profile for models with approximate the side chain atoms and loop regions, we observe a systematic upward shift from the original $I(q)$ profile [Fig. 4(A)]. This shift between the experimental and approximated profiles increases the rate of false positive identification by SAXS scores (Fig. 5). We observe also that minima and

Figure 3

Depiction of hen egg white lysozyme PDB ID:6lyz. (**A**) The crystal structure of lysozyme with the n-terminal region colored blue and the c terminal region colored red. (**B**) Depiction of the native structure with the loop regions removed and approximated by pseudo atoms along the curvilinear path between SSEs. (**C**) Overlay of the native and approximated version of hen egg white lysozyme. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

maxima of the $I(q)$ profile are less affected. Therefore, by comparing the derivative of the profiles, we take the shape of the SAXS profile into account which decreases the rate of false positive identification by SAXS score.

For this derivative comparison, a curve was fit through the experimental data points using locally weighted scatterplot smoothing (LOESS)[55,56] using a span of 0.2 and a polynomial degree of 1 in *R*. The span variable determines how much of the data is used to fit each local polynomial. A large span produces the smoothest function while the smaller the span, the closer the regression will conform to the data. Splines were used to numerically differentiate the fit profile. The derivative results and scores are shown in Figure 4 and Table II. To measure the similarity between an experimental SAXS profile and complete protein models, we use the standard χ score. By using this score, we can easily compare our method with other established methods in the field such as CRYSOL. The user can specify what metric to use during analysis.

Nonredundant dataset for protein discrimination benchmark

To determine how well the SAXS score can distinguish protein folds from each other, we evaluated a representative subset of 455 proteins with a 20% identify cutoff, 1.6 Å resolution cutoff, and 0.25 *R*-factor cutoff from the PICES databank.[57,58] These proteins can be formed into a 455 × 455 matrix (207,025 pairings) where the diagonal represents a protein paired with itself (a true positive) and the off diagonal elements represents a protein paired with a different protein. Using scattering profiles generated through CRYSOL, we computed the difference between the native protein and the test protein for each pairing. If the minimum SAXS score for a given protein was on the diagonal for the *i*th row and *j*th column, then we correctly identified the protein from all other candidate proteins and classified that as a true positive. If the minimum SAXS score was not on the diagonal, we classified it as a false positive. Using receiver operating characteristic (ROC) curves, we plotted the false positive rate on the *x*-axis and the true positive rate on the *y*-axis.

The area under the curve (AUC) for complete protein models is >99%. When side chains are removed, the AUC remains >99%. The AUC for proteins without side chains and loop regions is 76%. When loop regions are approximated, the AUC is 84%. The derivative score improves the

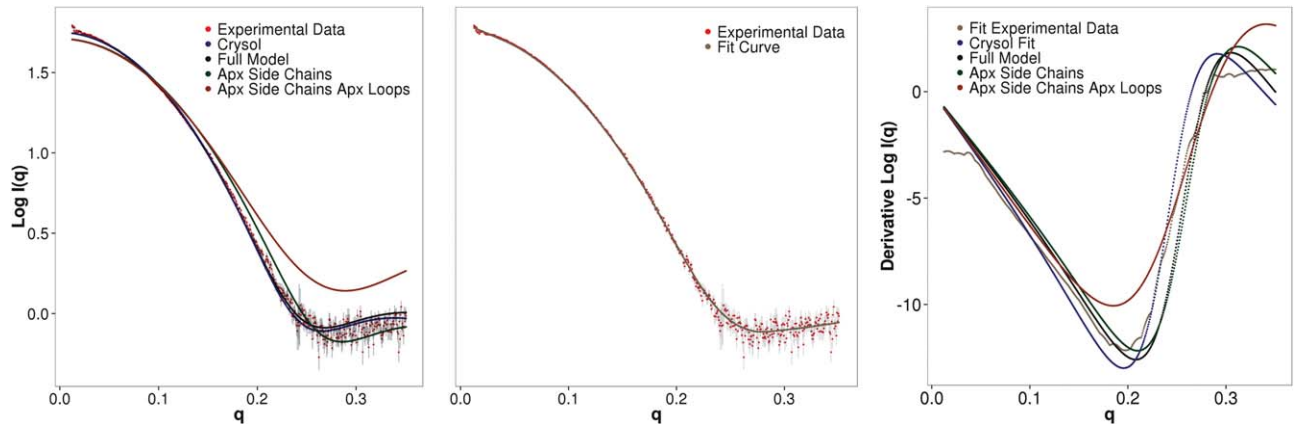

Figure 4

Depiction of the Experimental SAXS profile for Hen Egg White Lysozyme and SAXS profiles computed with BCL::SAXS for different protein states. **A**, Represents the fit on a log10 scale with experimental data being the SAXS profile of Hen Egg White Lysozyme. Crysol is the curve generated through Crysol from 6lyz and fit to the experimental data. Full Model is the curve generated through BCL::SAXS from 6lyz. Apx Side Chains is the curve generated through BCL::SAXS using backbone atoms only and summing the form factors for all side chain atoms at the Cβ coordinate of the residue. Apx side chains Apx loops is the curve generated through BCL::SAXS using loop approximation and side chain approximation. **B**, Locally weighted scatterplot smoothing (LOESS) of the experimental SAXS data points. **C**, Fit of previous data types from panel A using the derivative of the log10 profiles.

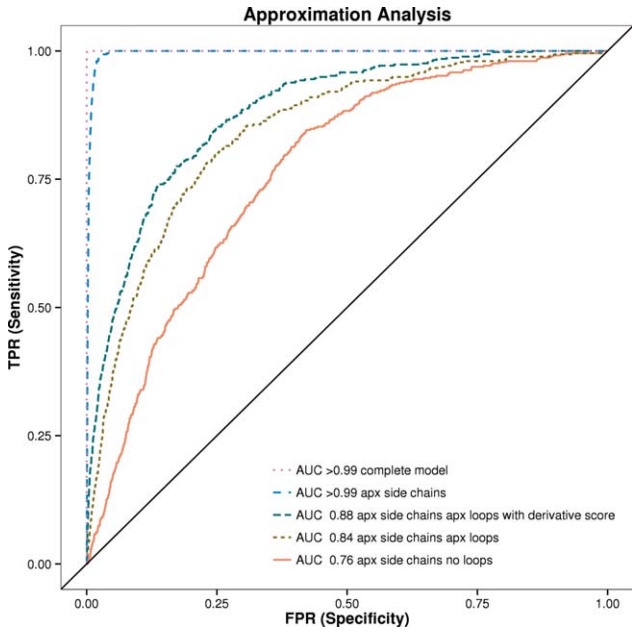

Figure 5

ROC analysis of 455 proteins from Pisces dataset in different states. The area under the curve (AUC) is shown with BCL::SAXS profiles generated for complete protein models (purple), models with approximated side chains (sky blue), approximated side chains and with loop approximation method (gold), approximated side chains without loop approximation method (orange), and the derivative of the approximated side chains with the loop approximation method (teal). The standard χ score was used to compare the profiles for all plots except for teal, where the derivative χ score was used. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

AUC to 88% (see Fig. 5). There were 207,025 total pairing evaluated in this experiment. In all but three cases the lowest SAXS score was the native protein when using complete protein models for analysis. For proteins 1YOZA and 3I31A the native was ranked second, while for protein 3L42A the native was ranked third.

Structural similarity of proteins with similar SAXS scores

To determine if protein models with similar SAXS scores were similar in protein structure, MAMMOTH[59] was used to score the structural similarity of a pair of proteins. The diagonal represents self-paired proteins. The higher the *Z*-score, the more similar the two structures are. A *Z*-score below four indicates that two proteins are structurally different. In the SAXS analysis, a lower SAXS score indicates the scattering profiles of two proteins are very similar. In this analysis, a high *Z*-Score and a low SAXS score indicate that proteins identified by SAXS as similar are structurally similar. Figure 6(A) depicts 3H5L chain A (molecular weight 44.92 kDA) paired with a copy of itself. As expected the SAXS similarity score is very low and the *Z*-score is high. Interestingly, panel B depicts 1N1F chain A (molecular weight 18.35 kDA) paired with 2GPE chain A (molecular weight 5.95 kDA). Although there is a difference of 12.4 kDA, the SAXS score indicates that the proteins are similar. Figure 6 shows that structurally similar proteins (high Mammoth *Z*-score) always have a low SAXS score (bottom left corner). However, while structurally dissimilar proteins (low Mammoth Z-score) tend to have increased SAXS scores, the observed range of SAXS scores widens. As expected, structurally different proteins can appear similar in a SAXS experiment if their overall shapes are similar.

Table II

χ^2 Scores Comparing Experimental SAXS Data for Hen Egg White Lysozyme with Profiles Generated from the Crystal Structure (6LYZ) for CRYSOL and BCL::SAXS

Type	$\text{Log}_{10}\ \chi$	Derivative χ
Crysol	2.81	0.96
BCL::SAXS Full Model	2.32	1.01
BCL::SAXS Apx Side Chains	9.16	1.17
BCL::SAXS Apx Side Chains and Loops	19.83	1.25

SAXS degeneracy in the scattering profile

During elastic scattering, energy is conserved between incident X-rays that scatter by interactions with electrons in the target sample. The magnitude of the wave vector $\vec{k}$ for both the incident and scattered wave is given by $2\pi/\lambda$. The change in wave-vector is only in direction and the difference between $\vec{k_i}$ and $\vec{k_f}$ is given by $\vec{q}$-the momentum transfer vector. The X-ray scattering amplitude at $\vec{q}$ by a particle at position $\vec{r_j}$ is given by:

$$A_j(\vec{q}) = f(q)e^{\left(i\vec{q}\cdot\vec{r}_j\right)} \tag{21}$$

where f is the form factor for the atom j at a magnitude for q given by $4\pi\sin\theta/\lambda$. The form factor decreases from a maximum at $q = 0$. At this q value, the form factor is equivalent to the atomic number Z of the atom. Hence, atoms with higher Z are stronger scatterers. The amplitude for an ensemble of particles is a summation of the amplitudes of all particles:

$$A(\vec{q}) = \sum_{j=1}^{n} f(q)e^{\left(i\vec{q}\cdot\vec{r}_j\right)} \tag{22}$$

The scattering intensity is given by the amplitude multiplied by its complex conjugate $A(\vec{q})^*$:

$$I(\vec{q}) = A(\vec{q})A(\vec{q})^* \tag{23}$$

The observed scattering pattern is not the complex amplitude function. It is the modulus squared of the amplitude function. Most of the structural information obtained from X-ray scattering experiments reside in the phase of the wave-function. This phase information is

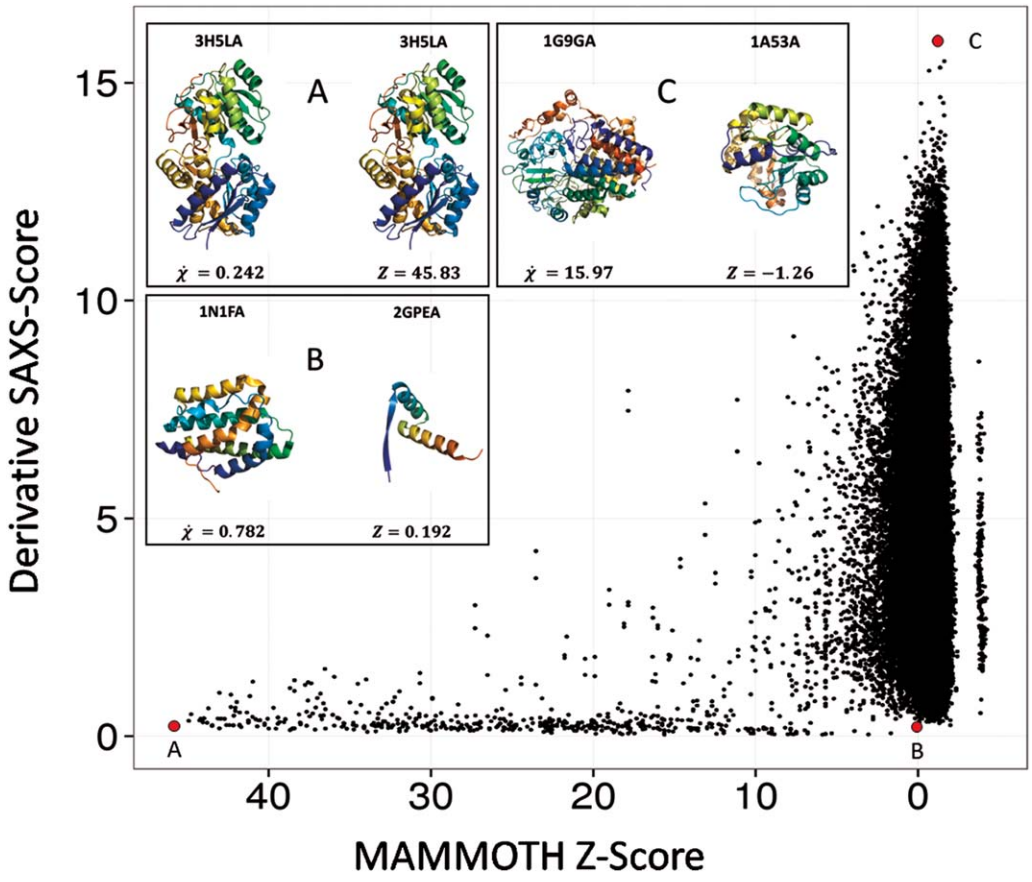

Figure 6

Structural MAMMOTH *Z*-score versus SAXS profile similarity score of 455 proteins from Pisces dataset. All 455 proteins were scored by structural similarity to each other with self-pairing receiving the highest *z*-score (*x*-axis). SAXS profiles for all 455 proteins were generated and the χ score between all scores was computed (*y*-axis). **A–C** correlate with their respective red dot. Panel A depicts 3H5LA paired with itself. Panel B depicts 1N1FA paired with 2GPEA. Panel C depicts 1G9GA paired with 1A53A. The derivative χ score was used to compare the 455 SAXS profiles. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

stored in the imaginary part of the amplitude function and is lost when multiplied by the complex conjugate. This loss of phase information results in a loss of structural uniqueness. Furthermore, the effect is compounded because during a SAXS experiment samples are free to rotate. The observed $I(q)$ function is therefore also an average over possible orientations. The loss of orientation and phase information results in the degeneracy in the scatting profile (multiple structures yielding similar SAXS profiles) as observed in Figure 6.

To show the relation between the molecular weight of the compared proteins and the similarity of the SAXS profiles, we calculated molecular weights for all 455 proteins in the PISCES data set used in the MAMMTOH analysis. We then combined the molecular weight difference with the derivative SAXS score to generate a density plot (Fig. 7). As expected, we observe that for proteins of similar molecular weight a range of SAXS similarity scores χ are possible from very similar to dissimilar determined solely by the similarity in overall shape. As the difference in molecular weight increases, the minimum SAXS similarity scores χ increases also, that is, structures with large molecular weight differences do not have similar SAXS profiles, even if the overall shape is similar.

Scoring BCL::models with SAXS

BCL::Fold was run to generate 10,000 protein models of 3FRR. These models were only comprised of secondary structure elements. Using the side chain and loop region approximations, BCL::SAXS was used to construct SAXS profiles for all 10,000 models generated by BCL::Fold (Fig. 8). From this figure, we observe that the correct topology has a very low SAXS score. We note that model C has a lower SAXS score (1.43) than model B (1.71) although model B has a much lower RMSD100 score (7.72) than model C (16.29). This behavior is expected because SAXS cannot distinguish topologies that fit inside the overall SAXS envelope. Agreement with

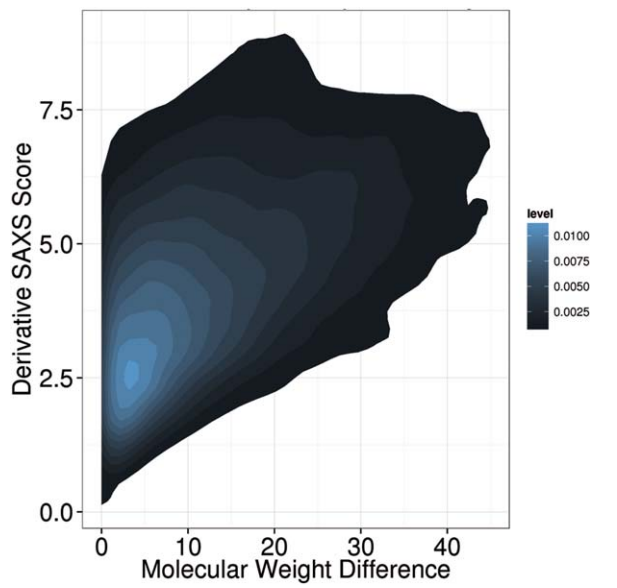

Figure 7

The SAXS similarity scores χ in relation to molecular weight difference. Molecular weights for all 455 proteins from the PISCES data set were calculated. The absolute value of the difference in weight between two proteins was computed for all pairs. The density plot depicts the difference in molecular weight on the *x*-axis and the derivative SAXS similarity score χ on the *y*-axis. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

by SAXS score is a necessary condition for correct protein identification, but not sufficient to uniquely identify the correct model. However, because of this, the SAXS score can be used as a filter to remove models that score above a threshold.

GPU algorithm yields orders of magnitude speed improvements

The GPU accelerated Debye calculation was benchmarked on several protein systems from the PDB with sizes ranging from 1800 atoms to 92,000 atoms. The benchmark was performed on several devices ranging from low-end workstation class GPUs (Quadro 600) to high-end consumer grade GPUs (GTX680) (see Table III). The speed was determined by measuring the time in seconds from the start of the Debye formula to the SAXS profile return from the Debye formula. The Maximum Speed up is the maximum of the ratio of the CPU time in seconds divided by the GPU time in seconds.

DISCUSSION

We have demonstrated how to compute SAXS profiles from atomic coordinates. In our approach for complete protein models we did not make approximations to the Debye formula, rather we used GPU acceleration to handle the double summation of all atoms and used the Hartree-Fock scattering factors directly. For proteins of sizes ranging from 1832 atoms to 91,846 atoms we find, as expected, that without GPU acceleration, the $O(N^2)$ computational cost of the Debye formula results in a significant slow-down when compared to the $O(q^2D^2N)$ algorithm implemented in CRYSOL (Table III). The magnitude of the momentum transfer vector is given by q and D is the max dimension of the macromolecule. With GPU acceleration computation times are comparable. The GPU card that gave the best performance was GTX680.

In order to compare experimental scattering profiles with approximated profiles we computed the first derivative of the profiles and then computed the similarity score ($\dot{\chi}$), between the derivatives of the SAXS profiles. This enabled us to reduce the amount of false positives obtained during our analysis and improve the accuracy in structure identification using SAXS profiles from 84% to 88%. BCL::SAXS was >99% accurate in picking the native protein from a set of other proteins when using complete proteins from the PDB and using the standard χ comparison score. With the side chains approximated, BCL::SAXS remained >99% accurate in picking the native protein from a set of other proteins. With the loop regions removed, the accuracy dropped from >99% to 76%. This result shows that loop regions play an important role in defining overall protein shape. Using our loop approximation algorithm and the derivative of the χ score, the accuracy increased to 88%. This result shows that having an approximate estimate of a protein location can have significant impact on the accuracy of SAXS scattering profiles generated from rigid bodies.

The MAMMOTH analysis shows that proteins with very similar z-scores (structurally similar proteins) also have a low SAXS $\dot{\chi}$ score. Importantly, the analysis shows that very similar structures do not have high SAXS scores. In the middle range of the analysis, we observe that SAXS scores are degenerate. Different structures can have similar SAXS scores. This degeneracy is inherently due to the spherical averaging of atoms in the SAXS data collection process. Because of this degeneracy SAXS cannot be used exclusively to predict protein structure.

CONCLUSION

We explored the idea of approximating the SAXS score for protein models without side chain and loop coordinates by placing dummy atoms along a path between secondary structure elements. The SAXS profile can be used to distinguish different proteins from each other, but cannot be used exclusively to distinguish different permutations of the same topology. However, the SAXS profile can be used as a filter to exclude protein models that are very different from the native from further analysis as a filter.

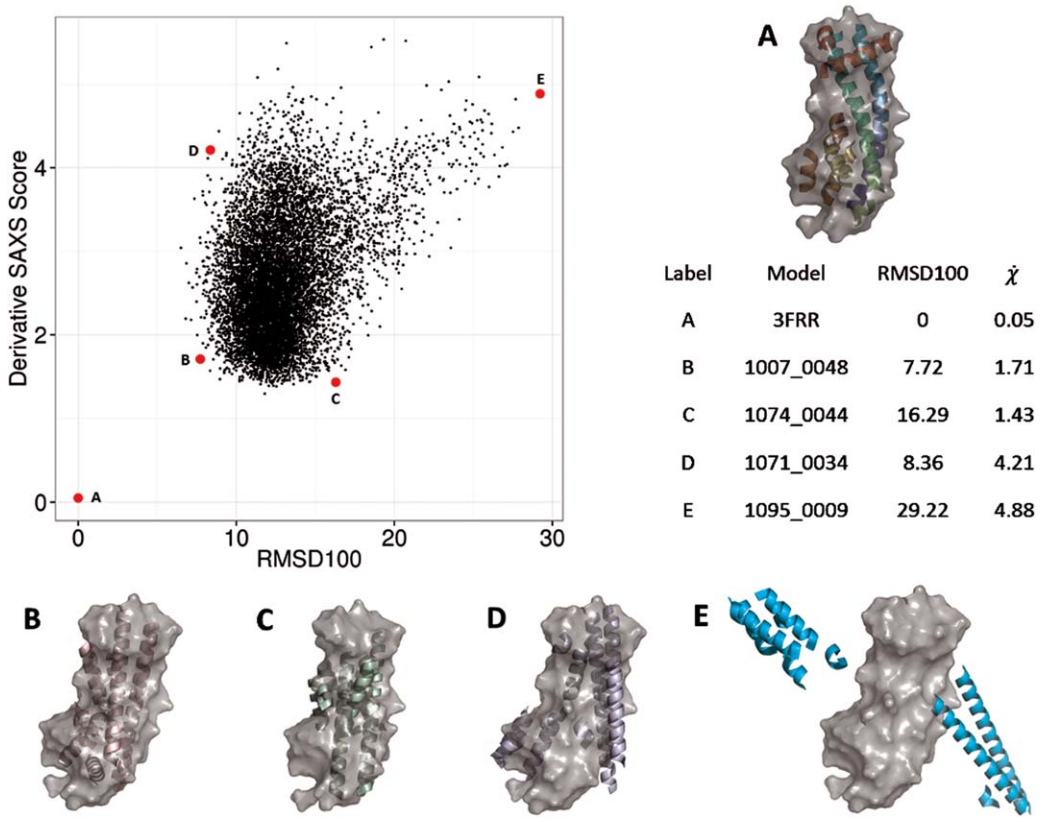

Figure 8

Filtering models produced by BCL::Fold by SAXS score. BCL::SAXS was used to score 10,000 protein confirmations of 3FRRA generated by BCL::Fold. In each case the surface of the native confirmations is shown in gray. Each black dot represents one model. The red dots labeled with **A–E** show examples of different conformations sampled by BLC::Fold and their respective scores. The derivative χ score was used to compare the 10,000 BCL models with the native structure. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table III
Timing Results of GPU Versus CPU Benchmarks

PDB	Atoms	BCL::SAXS CPU $1300	Crysol	Quadro 600 $225	GTX470 $325	GTX480 $325	GTX580 $550	GTX680 $1000	C1060 $1300	Maximum speed up
1O26_A	1832	3.6	–	0.1	0.07	0.07	0.07	0.07	0.09	51×
1WA5_C	7543	65	–	1	0.31	0.28	0.27	0.20	0.37	325×
1NR1	23,217	624	2	9.3	2	1.9	1.8	1.2	2.7	520×
1ZUM	43,243	2300	5	30	4.9	4.1	3.9	2.4	6.5	958×
1VSZ	91,846	15365	10	132	19.8	16.9	15.8	9.0	26.3	1707×

All timings are reported in seconds. Crysol reported timings to the nearest second. The first two measurements were not accurate and have been omitted.

ACKNOWLEDGMENTS

The authors thank Mariusz Butkiewicz, and Jeff Mendenhall for their insight and assistance throughout the development of BCL::SAXS. The authors acknowledge Oak Ridge National Laboratory supercomputing resources on TITAN. The authors thank Oanh Vu for her assistance in manuscript preparation.

REFERENCES

1. Karplus M. The Levinthal paradox: yesterday and today. Fold Des 1997;2:S69–S75.
2. Skrisovska L, Schubert M, Allain FH. Recent advances in segmental isotope labeling of proteins: NMR applications to large proteins and glycoproteins. J Biomol NMR 2010;46:51–65.
3. Bill RM, Henderson PJ, Iwata S, Kunji ER, Michel H, Neutze R, Newstead S, Poolman B, Tate CG, Vogel H. Overcoming barriers to membrane protein structure determination. Nat Biotechnol 2011;29: 335–340.
4. J.T. Ngo JM, M. Karplus. Computational complexity, protein structure prediction, and the Levinthal paradox. In: Merz KJLGS, editor. The protein folding problem and tertiary structure prediction. Boston, MA: Birkhauser; 1994. pp 435–508.

5. Alber F, Forster F, Korkin D, Topf M, Sali A. Integrating diverse data for structure determination of macromolecular assemblies. Annu Rev Biochem 2008;77:443–477.
6. Lindert S, Staritzbichler R, Wotzel N, Karakas M, Stewart PL, Meiler J. EM-fold: de novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps. Structure 2009;17:990–1003.
7. Robinson CV, Sali A, Baumeister W. The molecular sociology of the cell. Nature 2007;450:973–982.
8. Koch MHJ, Vachette P, Svergun DI. Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. Q Rev Biophys 2003;36:147–227.
9. Putnam CD, Hammel M, Hura GL, Tainer JA. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Q Rev Biophys 2007;40:191–285.
10. Svergun DI, Koch MHJ. Small-angle scattering studies of biological macromolecules in solution. Rep Prog Phys 2003;66:1735–1782.
11. Svergun DI, Petoukhov MV, Koch MH. Determination of domain structure of proteins from X-ray solution scattering. Biophys J 2001; 80:2946–2953.
12. Tsuruta H, Irving TC. Experimental approaches for solution X-ray scattering and fiber diffraction. Curr Opin Struct Biol 2008;18:601–608.
13. Lars N, Mark H, Jan P. Fast n-body simulation with CUDA, Simulation 3. In: Nguyen H, editor. GPU Gems 3. Boston MA: Pearson Education, Inc; 2008. pp 677–697.
14. Feigin LA, Svergun DI. Structure analysis by small-angle X-ray and neutron scattering. New York: Plenum Press; 1987.
15. Glatter O, Kratky O. Small angle X-ray scattering. New York: Academic Press Inc; 1982. pp 515.
16. Putnam DK, Lowe EW, Meiler J. Reconstruction of SAXS profiles from protein structures. Comput Struct Biotechnol J 2013;8:1–12.
17. Mertens HD, Svergun DI. Structural characterization of proteins and complexes using small-angle X-ray solution scattering. J Struct Biol 2010;172:128–141.
18. Forster F, Webb B, Krukenberg KA, Tsuruta H, Agard DA, Sali A. Integration of small-angle X-ray scattering data into structural modeling of proteins and their assemblies. J Mol Biol 2008;382:1089–1106.
19. Zheng W, Doniach S. Fold recognition aided by constraints from small angle X-ray scattering data. Protein Eng Des Sel 2005;18:209–219.
20. Stuhrmann HB. Interpretation of small-angle scattering functions of dilute solutions and gases. A representation of the structures related to a one-particle scattering function. Acta Crystallogr A 1970;26: 297–306.
21. Pelikan M, Hura GL, Hammel M. Structure and flexibility within proteins as identified through small angle X-ray scattering. Gen Physiol Biophys 2009;28:174–189.
22. Bonneau R, Ruczinski I, Tsai J, Baker D. Contact order and ab initio protein structure prediction. Protein Sci 2002;11:1937–1944.
23. Karakas M, Woetzel N, Staritzbichler R, Alexander N, Weiner BE, Meiler J. BCL::Fold - de novo prediction of complex and large protein topologies by assembly of secondary structure elements. PLoS One 2012;7:e49240.
24. Woetzel N, Karakas M, Staritzbichler R, Muller R, Weiner BE, Meiler J. BCL::Score-knowledge based energy potentials for ranking protein models represented by idealized secondary structure elements. PLoS One 2012;7:e49242.
25. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, Kaufman K, Renfrew PD, Smith CA, Sheffler W, Davis IW, Cooper S, Treuille A, Mandell DJ, Richter F, Ban YE, Fleishman SJ, Corn JE, Kim DE, Lyskov S, Berrondo M, Mentzer S, Popović Z, Havranek JJ, Karanicolas J, Das R, Meiler J, Kortemme T, Gray JJ, Kuhlman B, Baker D, Bradley P. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol 2011;487:545–74.
26. Heinze S, Putnam DK, Fischer AW, Kohlmann T, Weiner BE, Meiler J. CASP10-BCL::Fold efficiently samples topologies of large proteins. Proteins 2015;83(3):547–63.
27. Weiner BE, Woetzel N, Karakas M, Alexander N, Meiler J. BCL::MP-Fold: folding membrane proteins through assembly of transmembrane helices. Structure 2013;21:1107–1117.
28. Schneidman-Duhovny D, Hammel M, Sali A. FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic Acids Res 2010;38(Web Server issue):W540–W544.
29. Stovgaard K, Andreetta C, Ferkinghoff-Borg J, Hamelryck T. Calculation of accurate small angle X-ray scattering curves from coarse-grained protein models. BMC Bioinform 2010;11:429.
30. Svergun D, Barberato C, Koch MHJ. CRYSOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. J Appl Crystallogr 1995;28:768–773.
31. Tjioe E, Heller WT. ORNL_SAS: software for calculation of small-angle scattering intensities of proteins and protein complexes. J Appl Crystallogr 2007;40:782–785.
32. Debye P. Zerstreuung von Röntgenstrahlen. Annalen der Physik 1915;351:809–823.
33. Cromer DT, Mann JB. X-ray scattering factors computed from numerical Hartree-Fock Wave Functions. Los Alamos: University of California; 1967.
34. Cromer DT, Waber JT. Scattering factors computed from relativistic Dirac-Slater wave functions. Acta Crystallogr 1965;18:104.
35. Fraser RDB, MacRae TP, Suzuki E. An improved method for calculating the contribution of solvent to the X-ray diffraction pattern of biological molecules. J Appl Crystallogr 1978;11:693–694.
36. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein data bank: a computer-based archival file for macromolecular structures. Arch Biochem Biophys 1978;185:584–591.
37. Doyle PA, Turner PS. Relativistic Hartree–Fock X-ray and electron scattering factors. Acta Crystallogr A 1968;24:390–397.
38. Fox AG, Okeefe MA, Tabbernor MA. Relativistic Hartree-Fock X-ray and electron atomic scattering factors at high angles. Acta Crystallogr A 1989;45:786–793.
39. Brown PJR, Maslen AG, O'Keefe EN, Willis B.T.M. MA. Intensity of diffracted intensities. In: Prince E, editor. International Tables for Crystallography, Vol. C. John Wiley and Sons; 2006. pp 554–595.
40. Antonov LD, Andretta C, Hamelryck T. Parallel GPGPU Evaluation of Small Angle X-Ray Scattering Profiles in a Markov Chain Monte Carlo Framework. In: Gabriel J, Schier J, VanHuffel S, Conchon E, Correia C, Fred A, Gamboa H, editors. Biomedical Engineering Systems and Technologies, Vol. 357. Communications in Computer and Information Science: Springer Berlin Heidelberg; 2013. pp 222–235.
41. Clarke ND, Kissinger CR, Desjarlais J, Gilliland GL, Pabo CO. Structural studies of the engrailed homeodomain. Protein Sci 1994; 3:1779–1787.
42. Grishaev A, Wu J, Trewhella J, Bax A. Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data. J Am Chem Soc 2005;127:16621–16628.
43. Walther D, Cohen FE, Doniach S. Reconstruction of low-resolution three-dimensional density maps from one-dimensional small-angle X-ray solution scattering data for biomolecules. J Appl Crystallogr 2000;33:350–363.
44. Hura GL, Menon AL, Hammel M, Rambo RP, Poole FL, 2nd, Tsutakawa SE, Jenney FE, Jr., Classen S, Frankel KA, Hopkins RC, et al. Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). Nat Methods 2009;6:606–612.
45. Klose DP, Wallace BA, Janes RW. 2Struc: the secondary structure server. Bioinformatics 2010;26:2624–2625.

46. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
47. Andersen CA, Palmer AG, Brunak S, Rost B. Continuum secondary structure captures protein flexibility. Structure 2002;10:175–184.
48. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23:566–579.
49. Labesse G, Colloc'h N, Pothier J, Mornon JP. P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. Comput Appl Biosci 1997;13:291–295.
50. Majumdar I, Krishna SS, Grishin NV. PALSSE: a program to delineate linear secondary structural elements from protein structures. BMC Bioinform 2005;6:202.
51. Taylor WR. Defining linear segments in protein structure. J Mol Biol 2001;310:1135–1150.
52. Martin J, Letellier G, Marin A, Taly JF, de Brevern AG, Gibrat JF. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. BMC Struct Biol 2005;5:17.
53. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–2309.
54. dos Reis MA, Aparicio R, Zhang Y. Improving protein template recognition by using small-angle X-ray scattering profiles. Biophys J 2011;101:2770–2781.
55. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc 1979;74:829–836.
56. Cleveland WS, Devlin SJ. Locally weighted regression - an approach to regression-analysis by local fitting. J Am Stat Assoc 1988;83:596–610.
57. Wang G, Dunbrack RL, Jr. PISCES: recent improvements to a PDB sequence culling server. Nucleic Acids Res 2005;33(Web Server issue):W94–W98.
58. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–1591.
59. Ortiz AR, Strauss CE, Olmea O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. Protein Sci 2002;11:2606–2621.