doi:10.1016/j.jmb.2008.07.074 J. Mol. Biol. (2008) 382, 1089–1106

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Integration of Small-Angle X-Ray Scattering Data into Structural Modeling of Proteins and Their Assemblies

Friedrich Förster[1,2,3]*, Benjamin Webb[1,2,3], Kristin A. Krukenberg[4], Hiro Tsuruta[5], David A. Agard[6,7]* and Andrej Sali[1,2,3]*

[1]*Department of Bioengineering and Therapeutic Sciences, University of California at San Francisco, San Francisco, CA, USA*

[2]*Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, CA, USA*

[3]*California Institute for Quantitative Biosciences (QB3), University of California at San Francisco, San Francisco, CA, USA*

[4]*Graduate Program in Chemistry and Chemical Biology, University of California at San Francisco, San Francisco, CA, USA*

[5]*Stanford Synchrotron Radiation Laboratory, Stanford Linear Accelerator Center, 2575 Sand Hill Road, MS 69, Menlo Park, CA 94025-7015, USA*

[6]*Howard Hughes Medical Institute, University of California at San Francisco, San Francisco, CA, USA*

[7]*Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, CA, USA*

A major challenge in structural biology is to determine the configuration of domains and proteins in multidomain proteins and assemblies, respectively. All available data should be considered to maximize the accuracy and precision of these models. Small-angle X-ray scattering (SAXS) efficiently provides low-resolution experimental data about the shapes of proteins and their assemblies. Thus, we integrated SAXS profiles into our software for modeling proteins and their assemblies by satisfaction of spatial restraints. Specifically, we modeled the quaternary structures of multidomain proteins with structurally defined rigid domains as well as quaternary structures of binary complexes of structurally defined rigid proteins. In addition to SAXS profiles and the component structures, we used stereochemical restraints and an atomic distance-dependent statistical potential. The scoring function is optimized by a biased Monte Carlo protocol, including quasi-Newton and simulated annealing schemes. The final prediction corresponds to the best scoring solution in the largest cluster of many independently calculated solutions. To quantify how well the quaternary structures are determined based on their SAXS profiles, we used a benchmark of 12 simulated examples as well as an experimental SAXS profile of the homotetramer D-xylose isomerase. Optimization of the SAXS-dependent scoring function generally results in accurate models if sufficiently precise approximations for the constituent rigid bodies are available; otherwise, the best scoring models can have significant errors. Thus, SAXS profiles can play a useful role in the structural characterization of proteins and assemblies if they are combined with additional data and used judiciously. Our integration of a SAXS profile into modeling by satisfaction of spatial restraints will facilitate further integration of different kinds of data for structure determination of proteins and their assemblies.

*Corresponding authors. F. Förster is to be contacted at University of California at San Francisco, UCSF MC 2552, Byers Hall at Mission Bay, Suite 503B, 1700 4th Street, San Francisco, CA 94158, USA. D. A. Agard, Howard Hughes Medical Institute and the Department of Biochemistry & Biophysics, University of California at San Francisco, 600 16th St., San Frnacisco, CA 94158, USA. A. Sali, University of California at San Francisco, UCSF MC 2552, Byers Hall at Mission Bay, Suite 503B, 1700 4th Street, San Francisco, CA 94158, USA. E-mail addresses: frido@salilab.org; agard@msg.ucsf.edu; sali@salilab.org.

Abbreviations used: SAXS, small-angle X-ray scattering; EM, electron microscopy; IMP, Integrative Modeling Platform; DOPE, discrete optimized protein energy; XI, xylose isomerase; RBNC, rigid body in the native configuration.

Received 19 April 2008;
received in revised form
20 July 2008;
accepted 25 July 2008
Available online
31 July 2008

Edited by M. Levitt

Keywords: small-angle X-ray scattering; quaternary structure; macromolecular assembly modeling; statistical potentials; protein structure prediction

Introduction

A comprehensive structural description of proteins, nucleic acids, and their assemblies will help us discover the principles that underlie cellular processes and bridge the gaps between genome sequencing, functional genomics, proteomics, and systems biology.[1,2] While X-ray crystallography and NMR spectroscopy can provide accurate high-resolution structures, these methods are limited by the difficulties in protein purification, stability of large complexes, crystallization (X-ray), and size (NMR). Single-particle cryo-electron microscopy (EM) generally does not provide atomic-resolution structures and currently cannot be applied to systems smaller than approximately 250 kDa. While efficient, computational protein structure prediction methods are limited by their accuracy. These difficulties may be overcome by computational methods that effectively combine experimental, theoretical, and statistical information.[2,3]

Small-angle X-ray scattering (SAXS) can rapidly provide low-resolution information about the shape of a macromolecule or a complex in solution.[4–7] A SAXS measurement determines the molecule's rotationally averaged scattering intensity as a function of spatial frequency, $I(q)$, typically at 1- to 3-nm resolution.[5,7] This profile can be readily transformed into an electron pair distance distribution function, $P(r)$, which is essentially a histogram of all pairwise distances, r, of the electrons in the sample. Due to the rotational averaging, the information content of a SAXS profile is dramatically reduced compared with an X-ray crystallographic diffraction pattern or even a density map from cryo-EM. However, one of the advantages of SAXS is its applicability to a wide range of assemblies; its applications range from DNA fragments[8] to whole virions.[9] Moreover, data collection and processing are very rapid (typically, from seconds to minutes), allowing high-throughput analyses of a large number of samples at many conditions. These advantages are in stark contrast to crystallography, cryo-EM, and NMR spectroscopy. The ease of altering solution conditions makes SAXS ideal for mapping structural differences between varied conformational states of a macromolecular system; if the structure of one conformational state is known, even the relatively sparse information content of a SAXS profile can be sufficient to determine structural rearrangements, such as hinge motions in proteins.[10–12]

Information from SAXS can in principle be incorporated into the modeling process in two ways. First, a SAXS profile can be used to assess different models that were produced based on other considerations. For example, experimental SAXS profiles have been used to choose one of the many different quaternary-structure arrangements produced by molecular docking of the son-of-sevenless domains.[13] Similarly, simulations indicated that SAXS profiles can be used to choose a close-to-native solution from a large set of alternative homology models of a given protein.[14]

Second, a SAXS profile can be used during the model-building stage itself. The first such calculation of a model based on a SAXS profile relied on representing a macromolecular surface using spherical harmonics.[15] However, this representation has a relatively low resolution, thus leading to the development of alternative methods. Due to the limited information content of SAXS profiles, virtually all subsequently developed methods have aimed to integrate additional information into structure determination to reduce the manifold of solutions consistent with a given SAXS profile to a usefully small number.

Depending on the nature and resolution of additional information, different representations have been proposed. Early approaches modeled the molecule's envelope enforcing compactness.[16] Early coarse-grained approaches represented the macromolecule as an assembly of unconnected beads on a grid.[17–19] This representation enforces an overall mass by using a required number of beads; geometric symmetry may also be incorporated through symmetrical sampling. In addition, compactness of the models is ensured by restricting the sampling to the vicinity of a compact initial model[17,18] or by including appropriate terms into the scoring function.[19,20] Other modeling approaches represent a protein as a chain of beads rather than a set of disconnected beads on a grid.[20] In a recent application, atomic models were fitted into 6-fold symmetrical bead reconstructions to gain insights into domain rearrangements of the AAA-ATPase p97 during its functional cycle.[10] Higher-resolution *a priori* structural information about some parts of the protein can also be integrated into the modeling process by focusing the conformational sampling on the undefined segments, such as loops, only[21] or on the configuration of structurally defined domains and their flexible linkers.[22] For example, rigid-body

modeling was applied to give qualitative insights into the conformation of the polypyrimidine binding protein.[23] Recently, SAXS profiles have been incorporated into the modeling of protein structures based on NMR-derived restraints, which significantly increased the accuracy of models for multidomain proteins compared with models based on NMR profiles alone.[24] In addition, the inclusion of simulated SAXS profiles into folding simulations led to models of small helical proteins.[25]

Here, we describe the newly developed SAXS module in *Integrative Modeling Platform* (IMP)†, our software platform for modeling macromolecular assemblies by satisfaction of spatial restraints.[3,26] This integration in turn allowed us to combine SAXS profiles with various other types of data already used by IMP. Specifically, we present a protocol for modeling multidomain proteins and complexes. In particular, the protocol calculates the quaternary structures of multidomain proteins with structurally defined rigid domains as well as quaternary structures for complexes of structurally defined rigid proteins. In addition to a SAXS profile and rigid-body constraints, we use stereochemical restraints derived from a molecular mechanics force field,[27] a simple nonbonded atom pair term,[28] the atomistic distance-dependent statistical potential discrete optimized protein energy (DOPE),[29] and an optional symmetry-enforcing term.[26] We quantify the performance of the protocol using a benchmark of simulated examples. Finally, to test the method in a realistic setting, we also model the quaternary structure of the homotetramer D-xylose isomerase (XI) based on an experimental SAXS profile. The method already revealed large domain rearrangements between the nucleotide-free and the nucleotide-bound forms of *Escherichia coli* Hsp90.[12]

Theory and Methods

We developed a method to calculate atomic models of proteins and their assemblies that are consistent with experimental SAXS profiles as well as other spatial restraints. The solution is found by optimizing a scoring function that quantifies how consistent a model is with the SAXS profile and the other restraints. Next, the method is described by specifying its three components: (i) the representation of the modeled system; (ii) the terms that contribute to the scoring function; and (iii) the optimization protocol. We also describe how to evaluate the ensemble of models obtained from independent optimizations of the scoring function.

Representation of a protein or a complex

The system is represented by its N_{at} nonhydrogen atoms. A major problem that needs to be overcome is the large size of the search space compared with the amount of input data. The model is partitioned into L rigid bodies, b_l, $l=1,2,\ldots,L \leq N_{at}$ to reduce the number of degrees of freedom. Each rigid body $b_l(\vec{r_l}, \vec{\theta_l})$ is characterized by its center of mass $\vec{r_l}$ and its orientation $\vec{\theta_l}$. The orientation is measured by three rotations $(\theta_x, \theta_y, \theta_z)$ around the x-, y-, and z-axes of the coordinate system with its origin at the center of mass. A rigid body $b_l(\vec{r_l}, \vec{\theta_l})$ can be any set of atoms, including a single atom, a secondary-structure element, a domain, or the whole protein. Each atom of the whole macromolecule is a member of exactly one rigid body. The two extreme cases correspond to either the entire structure or each atom being a rigid body. Here, in our applications to proteins, we model user-defined domains as rigid bodies, while the connecting linkers are flexible. In our application to assemblies of proteins, we model the whole proteins as rigid bodies.

Scoring function

The scoring function is defined as follows:

$$S = S_{\text{stereo}} + S_{\text{overlap}} + S_{\text{DOPE}} + S_{\text{sym}} + S_{\text{SAXS}} \quad (1)$$

S_{stereo} accounts for the inter-rigid body stereochemical features of the atoms in the protein (i.e., chemical bond lengths, bond angles, etc.) according to the CHARMM molecular mechanics force field.[27] The term S_{overlap} penalizes steric clashes in the model;[28] a harmonic potential penalizes distances between two atoms if they are closer than the sum of their van der Waals radii. The term S_{DOPE} is an atomic distance-dependent statistical potential.[29] Additionally, a symmetry restraint S_{sym} can be imposed on the arrangement of the rigid bodies:[26] S_{sym} is the sum of the RMSs of the differences between equivalent distances in two symmetry units. The term S_{SAXS} accounts for the SAXS profile and is described in detail next.

SAXS experiments measure the rotationally averaged X-ray scattering of the macromolecule under scrutiny. The measured quantity is a one-dimensional curve that gives the scattered intensity I_{exp} as a function of the momentum transfer $q=(4\pi/\lambda)\sin(\theta)$, where λ is the wavelength of the incident X-ray beam and 2θ is the scattering angle.

Similar to previous approaches to modeling a macromolecule based on its SAXS profile,[17,18,20,24] we score a model based on the deviation between the calculated (I_{m}) and experimental (I_{exp}) SAXS profiles:

$$\chi^2 = \frac{1}{Q}\sum_{k=1}^{Q} \frac{1}{\sigma_{\text{exp}}^2(q_k)} \left(I_{\text{exp}}(q_k) - cI_{\text{m}}(q_k)\right)^2 \quad (2)$$

where k denotes the index of the measured frequency q, Q is the total number of frequencies, and σ_{exp} is the experimental error. The relative scaling of I_{m} with respect to I_{exp} cannot be determined precisely because the protein concentrations generally cannot be measured with sufficient accuracy. Thus, the profile I_{m} is scaled by a constant c, which is chosen by minimizing χ^2 (Supplementary Theory and Methods).

† http://salilab.org/imp

Moreover, we require S_{SAXS} to be comparable in size with the other four types of terms in S. We used the following rationale: $S_{overlap}$, S_{stereo}, S_{DOPE}, and S_{sym} are extensive functions in that they increase proportionally to the number of atoms, N_{at}, in the model. For the native structure, the value of χ^2 from Eq. (2) is on the order of 1. We choose the SAXS penalty of the native structure to be of comparable magnitude with the internal potential energy, defining the scaling as follows:

$$S_{SAXS} = N_{at} k_B T \chi^2 \quad (3)$$

where k_B is the Boltzmann constant and $T = 100$ K.

The experimental SAXS profile of a macromolecule in solution is the difference between the scattering of the solution with and that without the protein. Thus, the profile is approximately equal to the difference between the scattering intensities of the macromolecule and the solvent in the same volume. This model neglects the approximately 3-Å thin hydration layer around the macromolecule, which has a slightly higher density than bulk water. We neglected the hydration layer because the density difference between the hydration layer and bulk water is relatively small (<0.060 e$^-$Å^{-3}) compared with the densities of water (0.334 e$^-$Å^{-3}) and protein (approximately 0.440 e$^-$Å^{-3}) and because the volume of the hydration layer is small compared with that of the macromolecule. We assess the error due to neglecting the hydration layer in Results by comparing our profiles with those using the program CRYSOL, which does account for the hydration layer.

We use the Debye formula[30] to calculate the SAXS profile I_m of a given atomic model of a protein or an assembly:

$$I_m(q) = \sum_{j=1}^{N_A} \sum_{j=1}^{N_A} f_i(q) f_j(q) \frac{\sin(q d_{ij})}{q d_{ij}} \quad (4)$$

where $f_i(q)$ and $f_j(q)$ are the isotropic atomic form factors of the atoms i and j, respectively, and d_{ij} is the Euclidean distance between these atoms. In solution, the scattering of the protein is reduced by the contribution of the solvent in the protein excluded volume; we followed a previous formulation[31] to account for this effect (Supplementary Theory and Methods). To accelerate computation, we did not include hydrogen atoms explicitly; their vacuum scattering and their excluded volumes were added to those of their bound atoms.

For optimization of S, we need the gradient of χ^2. After some algebra, we obtain the partial derivative of χ^2 (Eq. (3)) with respect to the x-coordinate of atom i using Eq. (4):

$$\frac{\partial}{\partial x_i}\chi^2 = 4c \sum_{j\neq 1}^{N_A} \frac{x_i - x_j}{d_{ij}^2} \sum_{k=1}^{Q} \frac{I_{exp}(q_k) - c I_m(q_k)}{\sigma_{exp}^2(q_k)} f_i(q_k) f_j(q_k) \left(\frac{\sin(q_k d_{ij})}{q_k d_{ij}} - \cos(q_k d_{ij}) \right) \quad (5)$$

The derivatives with respect to y and z are equivalent to those for x.

The calculations of χ^2 and its derivatives are computationally demanding. However, we can substantially shorten the computation time for both quantities by approximating the modeled system with atoms of different scattering masses but equal shape (Supplementary Theory and Methods). The computation time is reduced by a factor equal to the number of frequencies at which $I(q)$ is sampled, typically 100, using this approximation. Even at the resolution of $q_{max} = 1$ Å^{-1}, the loss in accuracy due to this approximation is marginal: The deviation between a profile calculated with and that calculated without the approximation is of the order of $\chi^2 = 10^{-3}$, which is at least 2 orders of magnitude below typical optimal χ^2 scores for our models (Supplementary Theory and Methods).

The rigidity constraints imply that atoms belonging to the same rigid body b_l are not allowed to move with respect to one another. Thus, all forces within b_l must be zero, which is achieved for S_{SAXS} by excluding all atoms that are part of the same rigid body as atom i from the summation in Eq. (5).

Modeling with multiple SAXS profiles

In IMP, we provide the option to add more than one SAXS profile term to S (Supplementary Theory and Methods). In addition to the profile of the macromolecule under scrutiny, SAXS profiles of the gold-labeled macromolecule or profiles of subsets of the original system can be acquired. When the corresponding structures are conserved in these constructs, these profiles provide additional information about the macromolecule.

Optimization

We optimize the model with respect to S to obtain an ensemble of good scoring solutions (Fig. 1). Many independent optimizations are carried out starting from different random initial configurations to sample different minima of S. We differentiate between the global search mode, in which we sample 1000 initial configurations, and the local search mode, in which we sample 100 initial configurations in the vicinity of a specific configuration. In both cases, the resulting models are clustered to make the final prediction.

Starting from an initial model of a monomeric protein or a complex (e.g., a crystal structure or a homology model), the domains and the connecting linker residues (monomeric protein) or the proteins (complex) are individually rotated and translated by random values $\Delta\Phi$ and ΔT, respectively. Depending on the magnitude of these parameters, the protocol explores the intermediate neighborhood of the initial model (local search mode) or the whole space of solutions (global search mode). We chose $\Delta T = 40$ Å and $\Delta\Phi = 180°$ for global sampling and $\Delta T = 1$ Å and $\Delta\Phi = 2°$ for local sampling. In the case of monomers, the domains are initially brought into relative vicinity by optimization with respect to S_{stereo} using the method of conjugate gradients,[32]

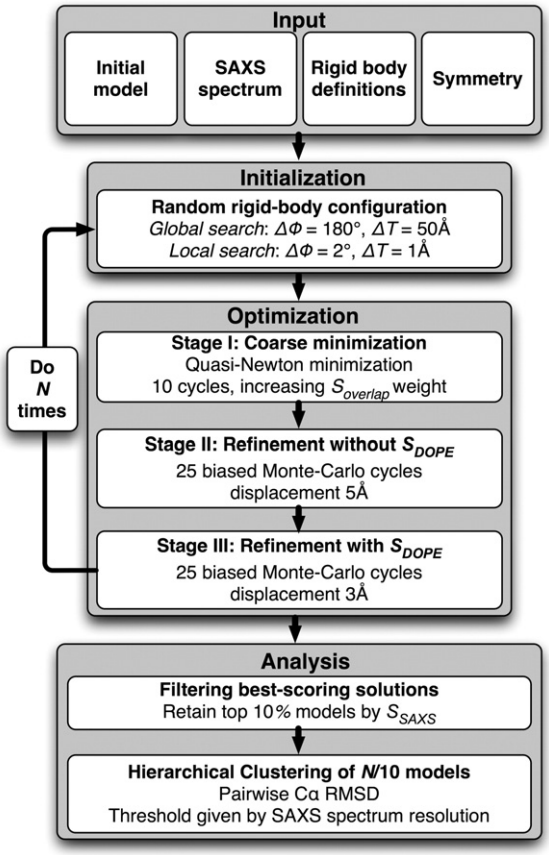

Fig. 1. Flowchart for the modeling protocol. As input, we need a SAXS profile, an initial model, the definition of rigid bodies, and, optionally, the symmetry. Initialization yields N random configurations, which are subsequently optimized in three stages. For the final analysis, only 10% of the N models with the best S_{SAXS} scores are retained and clustered.

which reduces the sampling problem at a negligible computational cost.

Next, in stage I of optimization, a coarse relaxation of the model is carried out with 10 cycles of a quasi-Newton method,[33] relying on the scoring function consisting only of S_{stereo}, S_{overlap}, S_{sym}, and S_{SAXS}; the weight of S_{overlap} is gradually increased from 0.005 to 1. The low initial weight of S_{overlap} allows large variations of the model during the initial iterations.

In stage II, we further optimize the model with respect to the sum of S_{stereo}, S_{overlap}, S_{sym}, and S_{SAXS} using a biased Monte Carlo algorithm: all atoms of the model are randomly displaced with a standard deviation of 5 Å, maintaining the rigid-body constraints. This perturbed model is then relaxed using one cycle of quasi-Newton optimization. Twenty-five Monte Carlo steps are carried out, decreasing the temperature exponentially. At each step, the resulting model is accepted or rejected according to the Metropolis criterion.

In stage III, 25 additional Monte Carlo steps are finally carried out to optimize the models with respect to the complete S, including S_{DOPE}.

On average, more than 200 models are generated for each of the 1000 initial independent optimizations, resulting in a total of ~200,000 evaluated models. On a Linux-based cluster of 100 processors, the computations take from a few hours to 2 days, depending on the size of the modeled system.

Analysis of optimized models

Because each optimization of a random starting configuration samples only a fraction of the configuration space, it generally ends in a local minimum. Thus, we rank all solutions by S_{SAXS} and retain only the top 10% for further analysis. These solutions are hierarchically clustered[34] with MATLAB (The Mathworks Inc.), relying on C^{α} RMSDs for all pairs of structures. Each C^{α} RMSD is obtained by superposing the largest domains and then computing the deviation between the whole structures. As a threshold level, t, for clustering, we use the resolution of the experimental SAXS profile, $t = 2\pi/q_{\max}$.

Benchmark of multidomain proteins and binary complexes

To assess the method with statistical rigor, we applied it to a benchmark set of 12 known structures with calculated SAXS profiles. The benchmark includes 9 multidomain proteins (Table 1 and Supplementary Table S1); their domains are treated as rigid bodies in our calculations. The domain definitions for the native structures were taken from the CATH database. Domains for 5 of the 9 proteins were represented by experimentally determined structures with identical sequences in a different state. Domains for the other 4 proteins were modeled by comparative modeling based on related structures. The alignments for comparative modeling were obtained from our comprehensive database of structural alignments, DBAli.[35] The models were built with the 'automodel' class in MODELLER 9.0. Except for 1o0vA, the models cover the whole sequence. The benchmark also includes 3 protein complexes, each consisting of 2 proteins (Table 2), that were obtained from 'Docking Benchmark 2.0.'[36] Here, the rigid bodies for each protein corresponded to crystallographic structures of the same sequence in a different state.[36]

We aimed to map the accuracy of the predicted configurations as a function of rigid-body accuracy. We modeled the configurations of rigid bodies as described above using the global and local search modes. For comparison, we also created an initial model with all rigid bodies superposed onto their counterparts in the native structure (RBNCs, rigid bodies in the native configuration) using Chimera[37] and optimized the linker segments with respect to S_{stereo} and S_{overlap}. Starting from this initial model, we performed 10 independent local optimizations ("RBNC refined" in Supplementary Table S1). Similarly, we also performed 10 local optimizations starting with the native state ("native refined" in Supplementary Table S1).

Table 1. Benchmark of multidomain proteins

Target[a]		Template[b]		Optimization protocol[c]		Most accurate cluster[d]										
								Best model by S			Best model by S_{SAXS}			Best model by S_{DOPE}		
PDB	Rigid bodies (residues)	PDB	Sequence identity (%)	Initial model	Sampling	Size (%)	Min RMSD (Å)	RMSD (Å)	Δr (Å)	$\Delta\alpha$ (°)	RMSD (Å)	Δr (Å)	$\Delta\alpha$ (°)	RMSD (Å)	Δr (Å)	$\Delta\alpha$ (°)
1ha0	1–316	2viuA	100	Near-template	Global	52	2.2	2.9	0.5	0.8	2.7	0.7	1.2	2.9	0.5	0.8
	326–494	2viuB	100	Near-template	Local	100	2.4	2.4	0.6	0.9	2.4	0.6	0.9	2.4	0.6	0.9
				Near-template	None	NA	2.8	2.8	0.1	0.1	2.8	0.1	0.1	2.8	0.1	0.1
				RBNC	None	NA	2.4	2.4	0.2	0.2	2.4	0.2	0.2	2.4	0.2	0.2
1gv2	89–136	1h88C	100	Near-template	Global	37	1.2	2.8	5.0	24.6	2.6	4.3	25.0	5.0	3.8	45.2
	149–190	1h88C		Near-template	Local	100	3.0	3.0	2.2	36.9	10.7	10.9	112.4	3.0	2.2	36.9
				Near-template	None	NA	9.2	9.2	7.8	101.9	9.2	7.8	101.9	9.2	7.8	101.9
				RBNC	None	NA	0.7	0.7	0.2	1.2	0.7	0.2	1.2	0.7	0.2	1.2
1mdtA	1–375	1ddtA	100	Near-template	Global	17	1.4	1.5	1.5	6.7	1.9	2.0	6.8	1.5	1.5	6.7
	389–535	1ddtA		Near-template	Local	100	10.4	11.8	8.9	123.6	13.3	3.8	141.7	11.7	7.8	128.4
				Near-template	None	NA	20.3	20.3	31.0	177.4	20.3	31.0	177.4	20.3	31.0	177.4
				RBNC	None	NA	1.0	1.0	1.0	4.0	1.0	1.0	4.0	1.0	1.0	4.0
1a62	1–45	1pv0A	100	Near-template	Global	53	1.3	1.6	3.2	19.2	1.5	2.7	19.2	1.4	1.2	12.2
	49–125	1pv0A		Near-template	Local	100	10	1.1	1.1	0.7	6.6	1.1	0.7	6.6	1.1	0.7
				Near-template	None	NA	1.0	1.0	1.0	4.0	1.0	1.0	4.0	1.0	1.0	4.0
				RBNC	None	NA	1.0	1.0	1.0	4.0	1.0	1.0	4.0	1.0	1.0	4.0
1iknA	19–184	1svcP	46	Near-template	Global	46	8.3	12.6	14.1	166.3	12.7	15.0	160.1	12.6	14.1	171.8
	196–291	1svcP		Near-template	Local	100	12.0	12.8	14.2	167.0	12.2	13.3	178.0	12.8	14.2	167.0
				Near-template	None	NA	13.3	13.3	16.0	133.4	13.3	16.0	133.4	13.3	16.0	133.4
				RBNC	None	NA	2.1	2.1	2.4	5.2	2.1	2.4	5.2	2.1	2.4	5.2
1ko9A	12–100	1fn7A	100	Near-template	Global	27	1.8	2.1	1.7	10.8	2.4	2.6	15.3	1.8	1.0	8.7
	104–126, 265–323	1fn7A		Near-template	Local	100	1.7	1.9	1.8	7.3	1.9	1.8	7.3	1.8	1.8	7.3
	135–261	1fn7A		Near-template	None	NA	1.5	1.5	0.9	5.1	1.5	0.9	5.1	1.5	0.9	5.1
				RBNC	None	NA	1.5	1.5	0.9	2.0	1.5	0.9	2.0	1.5	0.9	2.0
1o0vA	1–102	1a6tB	33	RBNC	Global	17	27.2	50.6	58.2	172.5	46.9	55.2	164.2	50.6	58.2	172.5
	115–204	1l6xA	31	RBNC	None	NA	5.6	5.6	4.6	21.4	5.6	4.6	21.4	5.6	4.6	21.4
	215–313	1l6xA	31													
1exmA	7–211	1f60A	35	Near-template	Global	24	25.4	29.8	37.8	157.6	29.6	37.6	149.6	25.4	31.8	117.8
	217–308	1f60A		Near-template	Local	100	11.6	18.2	20.7	61.3	17.7	21.8	56.1	18.2	20.7	61.3
	312–405	1f60A		Near-template	None	NA	9.1	9.1	7.1	33.5	9.1	7.1	33.5	9.1	7.1	33.5
				RBNC	None	NA	6.4	0.3	0.5	1.6	0.3	0.5	1.6	0.3	0.5	1.6
1cb6	1–87, 252–330	1h76A	63	Near-template	Global	27	4.5	4.5	3.6	11.5	23.9	27.8	98.3	8.7	5.4	18.8
	89–249	1h76A		Near-template	Local	100	3.4	3.5	4.6	11.2	3.9	4.2	12.6	3.5	4.6	11.2
	333–428, 596–687	1h76A		Near-template	None	NA	8.2	8.2	5.0	24.0	8.2	5.0	24.0	8.2	5.0	24.0
	433–592	1h76A		RBNC	None	NA	2.1	2.1	0.2	1.3	2.1	0.2	1.3	2.1	0.2	1.3

[a] We modeled nine proteins (targets) using rigid bodies for the specified amino acid residue ranges connected by flexible linkers.
[b] The rigid bodies corresponded to comparative models built using the specified template structures with different sequence identities, except for 100% sequence identity cases for which the specified template structure was used.
[c] We predicted domain configurations using protocols that differed by the initial configuration (near-template or RBNC) and the optimization (local or global).
[d] The models from each protocol were then clustered. The most accurate cluster was identified as the cluster with the model that was closest to the native configuration in terms of C^{α} RMSD. The size of the cluster is the percentage of all models in it. Min RMSD is the C^{α} RMSD of the most accurate model in the most accurate cluster. We also list C^{α} RMSD as well as Δr and $\Delta\alpha$ (the translation and the rotation of the rigid bodies for the optimal superposition on their native counterparts, respectively) for the models with the best S, S_{SAXS}, and S_{DOPE} in the most accurate cluster.

Table 2. Benchmark of binary complexes

Target		Template	Optimization protocol			Most accurate cluster								
								Best model by S			Best model by S_{SAXS}			Best model by S_{DOPE}		
PDB	Rigid bodies	PDB	Initial model	Sampling	Size (%)	Min RMSD (Å)	RMSD (Å)	Δr (Å)	$\Delta\alpha$ (°)	RMSD (Å)	Δr (Å)	$\Delta\alpha$ (°)	RMSD (Å)	Δr (Å)	$\Delta\alpha$ (°)
1avxAB	16–245:A	1qquA	Principal axes[a]	Global	45	9.3	10.5	5.1	81.6	10.5	5.1	81.6	14.6	8.8	116.1
	501–677:B	1ba7B	RBNC	None	NA	2.6	2.6	0.2	1.0	2.6	0.2	1.0	2.6	0.2	1.0
1ibr	9–177:A	1qg4A	Principal axes[a]	Global	51	12.6	13.0	18.5	66.5	13.2	18.9	66.6	13.2	19.0	66.4
	2–440:B	1f59A	RBNC	None	NA	3.4	3.4	0.3	1.2	3.4	0.3	1.2	3.4	0.3	1.2
1fq1	23–200:A	1fpzF	Principal axes[a]	Global	28	46.8	50.1	67.7	94.4	50.1	67.7	94.4	46.9	68.5	101.8
	1–296:B	1b39A	RBNC	None	NA	4.0	4.0	1.6	1.2	4.0	1.6	1.2	4.0	1.6	1.2

[a] The category "principal axes" indicates that the initial model was obtained by centering the respective rigid protein structures on their centers of mass and aligning their moments of inertia. See Table 1 for definitions of the other terms.

Assessment of similarity to the native state

The models were compared with the native state using two measures: (i) C^{α} RMSD with respect to the native structure and (ii) rigid-body translation Δr and rotation $\Delta\alpha$ of the rigid bodies relative to their positions in the native state. The C^{α} RMSD was calculated using the 'superpose' command of MODELLER 9.0. For calculation of Δr and $\Delta\alpha$, the reference frame was defined by first superposing the largest rigid body from the model on its equivalent fragment in the native structure. Next, each of the remaining rigid bodies was rotated around its center of mass and subsequently translated such that it superposed with the equivalent part of the native structure. The corresponding rotation and translation define $\Delta\alpha$ and Δr, respectively. If the model consisted of more than two rigid bodies, we computed the mean values of $\Delta\alpha$ and Δr to characterize the similarity between two configurations of rigid bodies, always using the largest rigid body to define the reference frame.

Acquisition of experimental XI profiles

We prepared XI solutions with concentrations of 0.55, 1.1, 2.7, and approximately 20 mg/ml from XI crystals (Hampton Research, Aliso Viejo, CA) and a buffer solution containing 10 mM Hepes, pH 7.4, and 150 mM NaCl. SAXS profiles were recorded at beam line 4-2 at the Stanford Synchrotron Radiation Laboratory:[38] Each solution was placed in a cuvette, which was maintained at 20 °C and located 2.5 m from a MarCCD165 detector (MarUSA, Evanston, IL). Twenty 15-s exposures (X-ray wavelength $\lambda = 1.381$ Å) were acquired in series for each concentration. For a range of concentrations, we obtained approximately constant radii of gyration R_G in the $R_G \cdot q_{min}$–$R_G \cdot q_{max}$ range of 0.37–1.27, indicating no protein aggregation or change in the homotetramer quaternary structure: R_G of 32.7±0.4 $Å^{-1}$ at 0.55 mg/ml, 32.9±0.2 $Å^{-1}$ at 1.1 mg/ml, and 32.8±0.1 $Å^{-1}$ at 2.7 mg/ml; these R_G values are similar to those reported previously.[39] SAXS profiles were computed from the scattering images using MarParse,[38] and profiles recorded at 2.7-mg/ml ($0.01\ Å^{-1} < q < 0.10\ Å^{-1}$) and 20-mg/ml ($0.055\ Å^{-1} < q < 0.27\ Å^{-1}$) concentrations were scaled and merged using Primus.[40] The complete profiles ranged from $q = 0.01\ Å^{-1}$ to $q = 0.27\ Å^{-1}$. We finally resampled the profile on a uniform grid of 100 mesh points using linear interpolation.

Results

Accuracy of calculated SAXS profiles

We determined that it is not necessary to include the hydration shell of a protein in the Debye model for calculating SAXS profiles because the inclusion of the hydration shell has a much smaller impact on

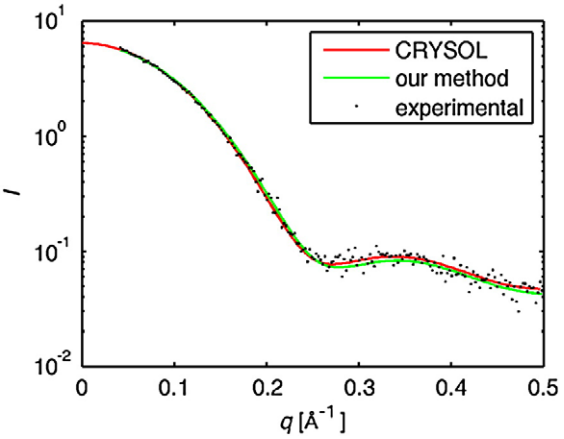

Fig. 2. Accuracy of the Debye model for calculating SAXS profiles. We compare an experimental profile of lysozyme (obtained from http://www.embl-hamburg.de/ExternalInfo/Research/Sax/crysol.html; shown in black) with the profiles calculated using our methodology (green) and the program CRYSOL[41] (red).

χ^2 than the errors in an experimentally measured SAXS profile. Specifically, we compared the experimentally measured lysozyme SAXS profile with the profiles computed by our method and CRYSOL[41] [Fig. 2; Protein Data Bank (PDB) code 6lyz]. The profiles calculated by our program and by CRYSOL agree with the experimental profile within its error, although the CRYSOL profile is a slightly better fit ($\chi^2 = 0.20$ *versus* 0.26). We also determined that using a single consensus atomic shape has no notable effect on the accuracy of calculated SAXS profiles; that is, the χ^2 difference between profiles calculated with the consensus and specific atomic shapes is approximately 0.001 (Supplementary Theory and Methods).

Modeling diphtheria toxin using a simulated SAXS profile

We first illustrate the method by its application to monomeric diphtheria toxin (Fig. 3a; PDB code 1mdt). We calculated the SAXS profile of the crystallographic structure: The q values ranged from $q_{min} = 0.02$ Å^{-1} to $q_{max} = 0.5$ Å^{-1}, and the $I(q)$ values were sampled on 100 equidistant grid points. Experimental error was simulated by adding white Gaussian noise with a standard deviation of 0.3 times $I(q_{max})$. The two diphtheria toxin domains were treated as rigid bodies; they were taken from the dimeric form of the protein, which possesses a dramatically different domain arrangement (Fig. 3b; PDB code 1ddtA).

In our approach, the predicted quaternary structure depends primarily on the experimental S_{SAXS} term and the modeling terms $S_{overlap}$ and S_{DOPE}. The main role of the additional term S_{stereo} is to keep

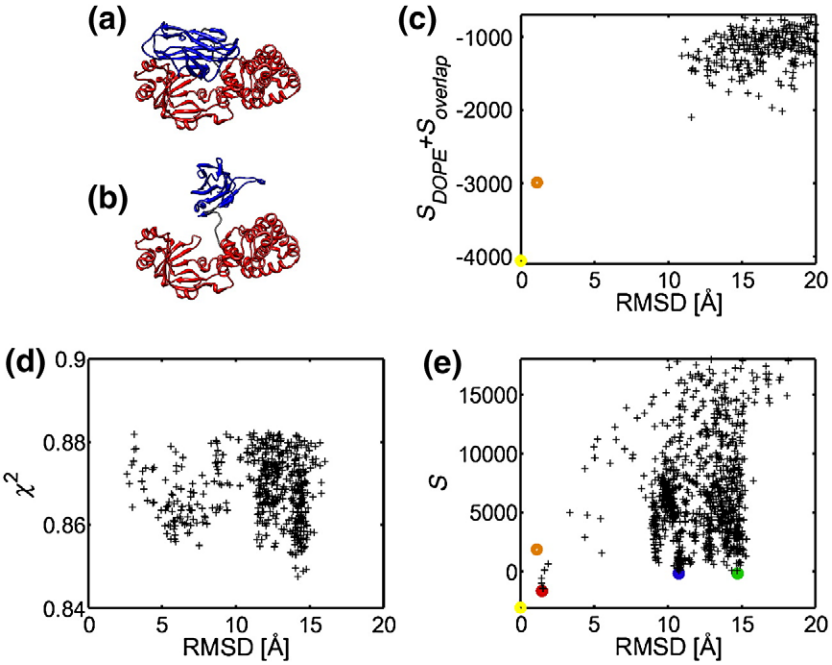

Fig. 3. Modeling diphtheria toxin using a simulated SAXS profile. (a) Native monomeric diphtheria toxin (PDB code 1mdtA) has two domains (blue and red). (b) We approximated these domains by their structures in the dimeric form (PDB code 2ddtA). (c) The sum of S_{DOPE} and $S_{overlap}$ of all models optimized with respect to $S_{DOPE} + S_{overlap} + S_{stereo}$, plotted against their C^α RMSD relative to the crystallographic structure. The yellow and orange circles depict the native structure and the RBNC, respectively. (d) The χ^2 of all models optimized with respect to $S_{SAXS} + S_{stereo}$, plotted against C^α RMSD. The χ^2 values for the native structure and the RBNC are outside the plotted range ($\chi^2 = 1.21$ and 1.57, respectively). (e) Overall S score of all models optimized with respect to the complete S, plotted against C^α RMSD. The best scoring models from the three clusters in Fig. 4 are highlighted in red, blue, and green.

domains in relative proximity and thus restrict the number of possible conformations. To assess the effect of combining experimental information (S_{SAXS}) and modeling information ($S_{overlap}$ and S_{DOPE}) on the predicted quaternary structures, we calculated models with different combined scoring functions consisting of (i) only S_{stereo}, $S_{overlap}$, and S_{DOPE}; (ii) only S_{stereo} and S_{SAXS}; and (iii) the complete S. One thousand independent optimizations were carried out for each scoring function variant in the global sampling mode. For each scoring function, we evaluated the score distribution and the C^{α} RMSD values of the models with respect to the native state (Fig. 3c–e). The first scoring function, which lacks S_{SAXS}, has minima for configurations that differ substantially from the native state (Fig. 3c); none of the models has a C^{α} RMSD below 10 Å. For the scoring function including only S_{SAXS} and S_{stereo}, a set of different configurations is consistent with the SAXS profile ($\chi^2<1$; Fig. 3d); the lowest scoring configurations are not near-native (their C^{α} RMSD is between 10 and 15 Å). In contrast, optimization of the complete S successfully results in near-native models (Fig. 3e). Thus, the combination of the SAXS and modeling terms results in a global minimum close to the native state, while the individual terms do not. This synergy justifies the integration of SAXS and protein structure modeling.

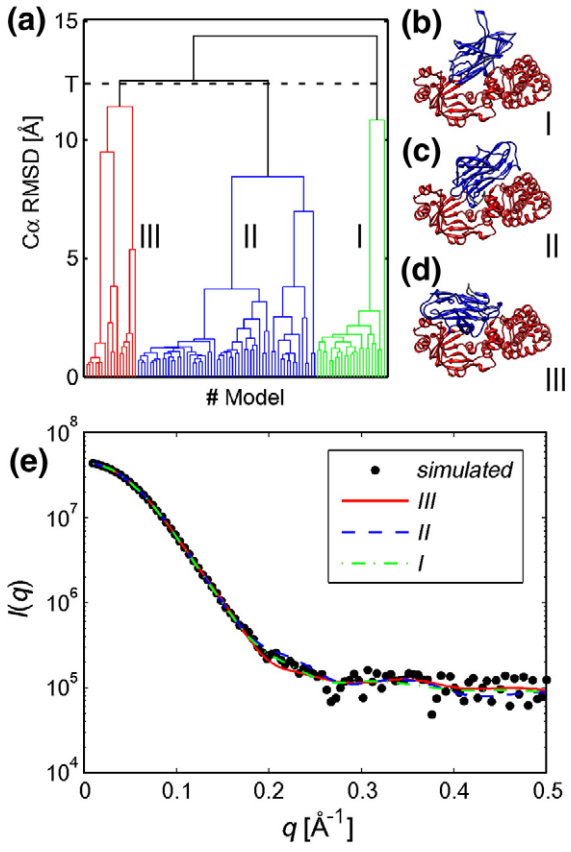

Fig. 4. Clustering of diphtheria toxin models. (a) Hierarchical tree of the 10% best scoring models according to their χ^2 values. The tree is cut at a pairwise C^{α} RMSD of 12.5 Å to cluster the models. Panels (b), (c), and (d) are models with the best S within clusters I, II, and III, respectively. (e) The $I(q)$ profiles of the best scoring models from clusters I, II, and III compared with the calculated SAXS profiles. The χ^2 values are 1.4, 1.5, and 1.3, respectively.

We further analyzed the models obtained from the optimization of S. Using C^{α} RMSD, we hierarchically clustered 10% of the models with the best χ^2 (Fig. 4a). The models clearly cluster into three groups, separated by more than 12.5-Å C^{α} RMSD from one another. Each cluster is represented by the model with the best S (Fig. 4b–d). Albeit all models correspond to distinct configurations, their shapes and calculated profiles are similar, as expected from the nature of the SAXS restraint. The model from cluster III (Fig. 4d) has the lowest S and the lowest SAXS penalty ($\chi^2=1.3$ compared with 1.5 and 1.4 for the models from clusters I and II, respectively; Fig. 4e) and is closest to the native state in terms of its C^{α} RMSD (1.4 Å compared with 10.4 and 14.0 Å, respectively). Interestingly, cluster III does not constitute the largest cluster, nor does it include the models with the lowest S_{DOPE} (Supplementary Table S1). Thus, in this case, S and χ^2 are indicative of the cluster with near-native models, whereas S_{DOPE} and the cluster size are not.

Benchmark on multidomain proteins

To assess the modeling protocol with statistical rigor, we applied it to our benchmark set of nine multidomain proteins (Theory and Methods). The benchmark comprised proteins with two domains (five cases), three domains (three cases), and four domains (one case; Table 1 and Supplementary Table S1). For five proteins, the domains were represented by experimentally determined structures that are identical in sequence but part of a different assembly. The rigid bodies of the remaining four proteins were modeled by comparative modeling based on related template structures. Thus, the benchmark covered different scenarios in terms of protein size and available structural information.

After optimization and clustering, the cluster with the most accurate model (in terms of C^{α} RMSD) was termed the most accurate cluster. In the global search mode, the best scoring models in the most accurate cluster were close to the native state (C^{α} RMSD<4 Å) for six proteins, of medium accuracy for one protein (C^{α} RMSD=12.8 Å), and of low accuracy in two cases only (C^{α} RMSD>18 Å; Table 1 and Supplementary Table S1). When crystallographic domain structures were used, the resulting configurations were always highly accurate. In contrast, when comparative models were used, high-accuracy configurations and medium-accuracy configurations could only be obtained for 1cb6 and for 1iknA, respectively. In our benchmark, these proteins also possessed the highest sequence identity with their template structures (63% and 46%, respectively). It is well established that the accuracy of a comparative model is correlated with the

sequence identity on which it is based.[42] Therefore, not surprisingly, the accuracy of whole protein models produced by our method is correlated with the sequence identity of the structural template for the modeling of the individual domains.

In real applications, we do not have the modeled native structure. Thus, we do not know which one of the generated models is most accurate but need to select it based on some criterion. Possible criteria include the S and S_{SAXS} (χ^2) scores, as well as the cluster size. In all but one high-accuracy case and one medium-accuracy case (global sampling for 1mdtA; Figs. 3 and 4), the most accurate model was found in the largest cluster. Similarly, in all but one case (global sampling for 1cb6), the most accurate model was also the one with the best S and S_{SAXS} scores. Thus, the selection of the final prediction by any of these three criteria is likely to identify the most accurate of the generated models.

In almost all cases, the local search produced more accurate models than the global search. Only for 1mdtA did the local search result in a less accurate model than the global search, consistent with the largest distortion in the initial structure relative to the native state (C^α RMSD>20 Å) among all benchmark cases. Remarkably, for the four-domain protein 1cb6 (C^α RMSD of the initial model is 8.2 Å), local sampling yielded more accurate results than global sampling, albeit global sampling results in models with lower scores. Thus, in this case, the near-native configurations are within the radius of convergence for the local sampling, whereas the global minimum, which is not a near-native configuration, is not. Based on the examples of 1mdtA and 1cb6, we conclude that the radius of convergence for local sampling is on the order of 10 Å.

Benchmark on binary protein complexes

We also applied our protocol to three binary protein complexes (Theory and Methods). The accuracy of binary complex models is generally lower than that for the individual two-domain proteins (Table 2). The highest accuracy was achieved for 1avxAB, whose best scoring models had a C^α RMSD of approximately 10 Å. This target was classified as 'easy' in Docking Benchmark 2.0.[36] The high C^α RMSD is due largely to inaccurate relative orientation of the two proteins ($\Delta\alpha = 81.6°$); in comparison, the relative position is quite accurate ($\Delta r = 5.1$ Å). Interestingly, local sampling in the vicinity of the RBNC state resulted in a model similar to the configuration using global sampling, which also scored better than the RBNC state according to S and χ^2 but not S_{DOPE} (Supplementary Table S2).

The results for target 1ibr are similar to those for 1avx; the best scoring models were different from the native state and scored better than the RBNC state. The target 1ibr is classified as 'difficult' in Docking Benchmark 2.0.[36]

For the third example, 1fq1 ('difficult' in Docking Benchmark 2.0[36]'), the largest three clusters produced by the global sampling were of similar size and different from the native state (C^α RMSD>30 Å). Moreover, the individual S, S_{SAXS}, and S_{DOPE} scores for the RBNC configuration were worse than those for optimized models.

If we assemble the native protein structures instead of their models, the local sampling around the native state always finds a configuration close to the native state (C^α RMSD<1.2 Å); moreover, the optimized S, S_{SAXS}, and S_{DOPE} are significantly better than those for the assembly of nonnative rigid bodies (Supplementary Table S2).

Quaternary structure of XI from an experimental SAXS profile

To test our methodology on experimental rather than simulated profiles, we modeled the quaternary structure of XI from *Streptomyces rubiginosus*. The model was calculated using three approximations for the monomers: (i) the native subunit structure in the complex; (ii) a comparative model based on 4xim as a template (67% sequence identity); and (iii) a comparative model derived from 1a0d (28% sequence identity). The best possible superpositions of four copies of the monomer comparative models onto the native structure of the XI complex (i.e., the C^α RMSD values of the RBNCs) were 2.7 and 5.1 Å for 4xim and 1a0d, respectively.

To reproduce the 222 symmetry of the XI tetramer consisting of identical subunits A, B, C, and D, we added a symmetry term S_{sym} to S. Specifically, we restrained equivalent intrasubunit C^α atom distances between equivalent chain pairs to be similar to each other.[26,43]

We acquired an experimental XI SAXS profile ranging from $q_{min} = 0.01$ Å^{-1} to $q_{max} = 0.27$ Å^{-1} (Theory and Methods). For each of the three assembly calculations, we built 500 configurations using the global sampling mode and clustered the 10% best scoring (χ^2) configurations (Table 3). Because XI is a homotetramer, the labels on the four subunits are irrelevant when comparing two configurations (i.e., swapping monomers has absolutely no impact on the structure). Therefore, we calculated C^α RMSDs between a model and the native state for all subunit permutations in the model and chose the minimum value as the final C^α RMSD.

For the native monomer, the largest cluster contains the models with the lowest S score as well as the most native-like models (C^α RMSD = 3.5 Å), indicating an accurate prediction. However, if we only consider S_{SAXS}, the best scoring models are not closest to the native state (Fig. 5a–d), illustrating the positive role played by the other terms in the combined scoring function S (Fig. 5e and f). The agreement between the model SAXS profiles and the experimental profile is generally lower than that in the benchmark simulations ($\chi^2_{min} \approx 8$ *versus* $\chi^2_{min} \approx 1$). We also checked our χ^2 against that implemented in CRYSOL,[41] confirming that the CRYSOL χ^2 likewise favors nonnative solutions and correlates well with our χ^2 (Supplementary Fig.

Table 3. Modeling of the XI tetramer using monomer models based on a range of sequence identity with the templates

Template		Cluster[a]		Best S in cluster[b]		Best χ^2 in cluster[b]		Best S_{DOPE} in cluster[b]		Best $S_{overlap}$ in cluster[b]	
PDB	Sequence identity (%)	Size (%)	Min RMSD (Å)	S	RMSD (Å)	χ^2	RMSD (Å)	S_{DOPE}	RMSD (Å)	$S_{overlap}$	RMSD (Å)
1xib	100	8	51.9	202,190	53.5	8.4	51.9	−437	51.9	174,417	53.5
		50	2.0	5828	2.3	9.6	2.3	−20,171	2.1	1965	2.3
		16	49.0	196,400	50.1	7.5	50.1	5292	49.7	172,413	50.1
		20	27.3	380,590	32.9	7.9	29.3	−513	27.3	351,926	32.9
4xim	67	48	3.4	53,800	3.5	12.8	3.6	−9242	3.4	31,709	3.5
		14	50.6	191,787	51.1	9.5	50.6	7396	50.8	155,453	51.1
		8	28.0	526,269	29.4	10.8	28.0	2799	30.4	494,592	29.4
1a0d	28	20	43.2	118,494	46.7	7.0	46.6	−1031	46.0	101,594	46.7
		22	47.8	92,027	50.2	6.9	50.1	−1938	50.1	75,363	50.2
		16	10.9	86,644	14.5	8.0	14.5	−2715	15.3	69,575	14.5
		10	47.2	394,099	50.5	8.8	50.5	3591	47.2	368,160	50.5
		8	51.7	136,284	51.8	8.3	51.7	3316	51.7	112,201	51.8
		18	26.6	116,515	29.9	7.1	29.8	−1148	30.9	99,543	29.9

[a] Clusters of all models are characterized by their relative sizes and the minimum C^α RMSDs.
[b] From each cluster, the best scoring model according to S, χ^2, S_{DOPE}, and $S_{overlap}$ is chosen, and its respective score and C^α RMSD are shown.

S1). Comparison of the SAXS profiles of the best scoring models in the largest cluster with the experimental and simulated SAXS profiles of the native state reveals that deviations from the experimental profile are most significant in the range of $0.11<q<0.12$ Å^{-1} (Fig. 5a–c). Similar observations were reported previously for XI.[39] In contrast to S_{SAXS}, S_{DOPE} is substantially lower for models in the cluster close to the native state than for the models in the other clusters.

The calculation with the 4xim-based monomer models follows the same trends as those with the native subunit (Fig. 5e and Table 3). Therefore, in a realistic setting, our modeling protocol would have correctly determined the quaternary arrangement of four subunits of the XI monomer using its experimental SAXS profile and a high-accuracy subunit model based on 67% sequence identity with the template structure.

In contrast, the results for the comparative model based on the remotely related 1a0d are qualitatively different. In this case, the largest cluster does not contain the most native-like models (C^α RMSD= 49.1 Å). If we select the cluster according to S or S_{DOPE}, the model is closer to the native state, albeit far from it in terms of C^α RMSD (15.3 Å). Nevertheless, this model still correctly predicts many of the residues at the subunit interfaces: 14% of the native contacts (the number of correctly predicted residue–residue contacts in the model divided by the number of contacts in the native structure) are identified correctly, which is considered 'acceptable' in blind assessments of protein docking methods at CAPRI meetings.[44]

Scoring *versus* sampling

We now analyze the extent to which model accuracy is limited by sampling and scoring; sampling is limiting when configurations close to the native state are not generated during optimization, and scoring is limiting when the most native-like configurations do not correspond to the global minimum of S.

For three of our benchmark cases (1ha0, 1ibr, and 1ko9), we plotted the best score (S_{min}), the corresponding C^α RMSD [RMSD(S_{min})], and the minimum C^α RMSD ($RMSD_{min}$) for a set of structures resulting from global sampling against the number of independent optimizations (Fig. 6). For the two-domain protein 1ha0, S_{min}, RMSD(S_{min}), and $RMSD_{min}$ do not improve substantially beyond 100 independent optimizations. Moreover, global sampling performs as well as local sampling starting from the RBNC (Fig. 6a; Supplementary Table S1). For the protein complex 1ibr, S_{min} and RMSD(S_{min}) reach a plateau at approximately 100 optimizations. RMSD(S_{min}) asymptotically reaches the value for local sampling around the RBNC, which is well above 10 Å, showing that a nonnative configuration scores better than the RBNC (Fig. 6b; Supplementary Table S2). The values for $RMSD_{min}$ asymptotically approach a value of approximately 3 Å (C^α RMSD of RBNC is 2.3 Å), indicating that near-native configurations are sampled but do not score well. In summary, we observed for all two-component systems (two-domain proteins and binary complexes) that the best scoring models scored approximately the same as the refined RBNC configurations; therefore, sampling does not appear to be limiting the accuracy of our predictions in these cases.

For the three-domain protein 1ko9, S_{min}, RMSD (S_{min}), and $RMSD_{min}$ improve slowly with an increase in the number of independent optimizations, beginning to reach a plateau at 400 independent optimizations. S_{min} from global optimization exceeds S_{min} of the models obtained by refining the RBNC, indicating that in this case the accuracy of the predicted quaternary structures is limited by sampling. Similar results are obtained for the other three- and four-domain proteins; the global mini-

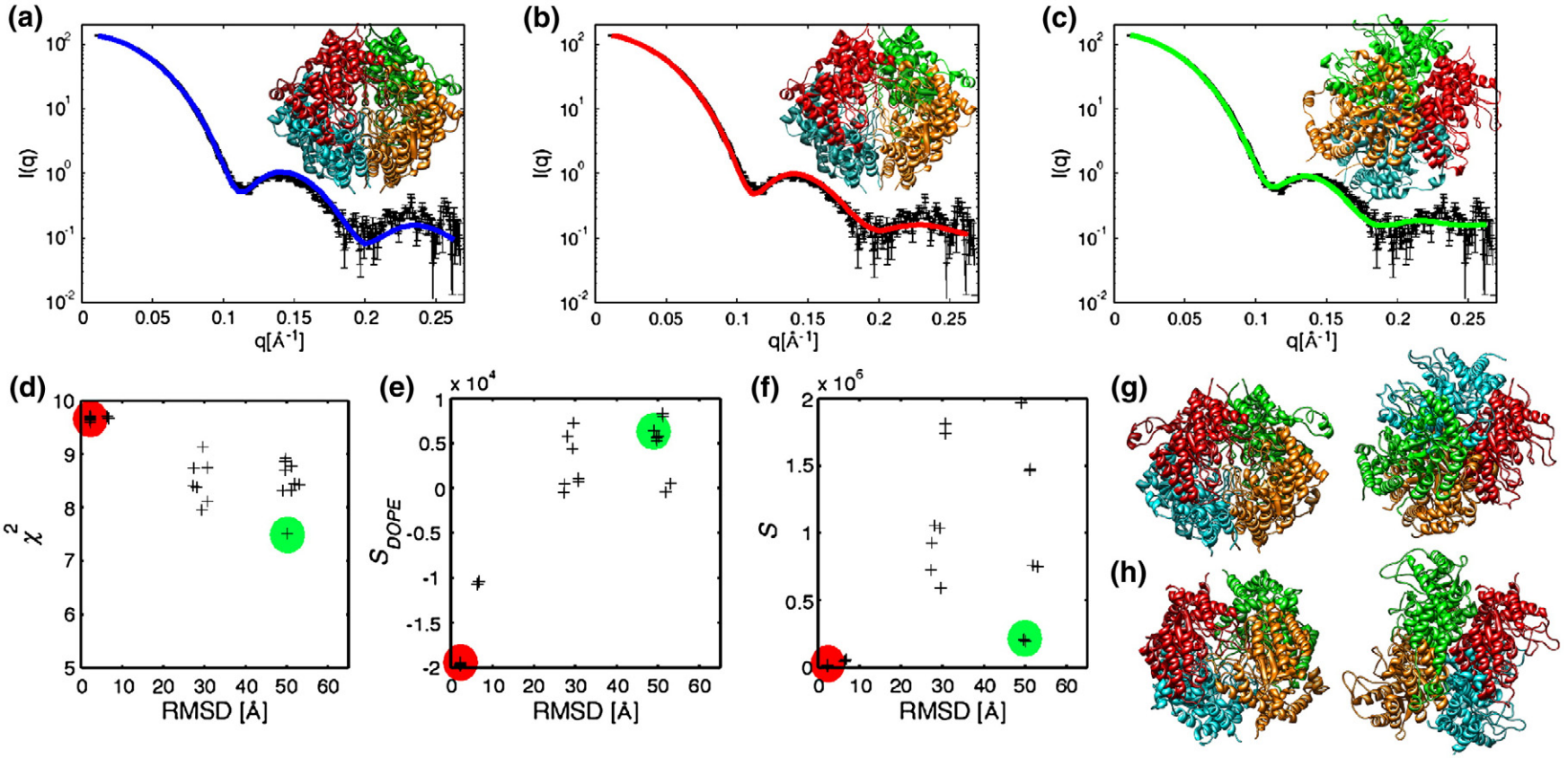

Fig. 5. Models of XI and their calculated SAXS profiles compared with the experimental profile. All chain A's (red) across the models are oriented identically to facilitate visual comparison of the different models. (a) Native XI ($\chi^2=17.7$; PDB code 1xib). (b) The XI model, based on the native subunit, with the best S_{DOPE} ($\chi^2=9.6$). (c) The XI model, based on the native subunit, with the best S_{SAXS} ($\chi^2=7.5$). (d) The χ^2 of the top 10% models (i.e., clustered models) plotted against their C^α RMSD with respect to the crystallographic structure. The models with the best S_{DOPE} and S_{SAXS} are indicated in red and green, respectively. (e) S_{DOPE} of the top 10% models plotted against their C^α RMSD with respect to the crystallographic structure. (f) Combined S score of the top 10% models plotted against their C^α RMSD. (g) The XI configuration of subunit comparative models based on 4xim (67% sequence identity); left, best S_{DOPE}; right, best S_{SAXS}. (h) The XI configuration of subunit comparative models based on 1a0d (27% sequence identity); left, best S_{DOPE}; right, best S_{SAXS}.

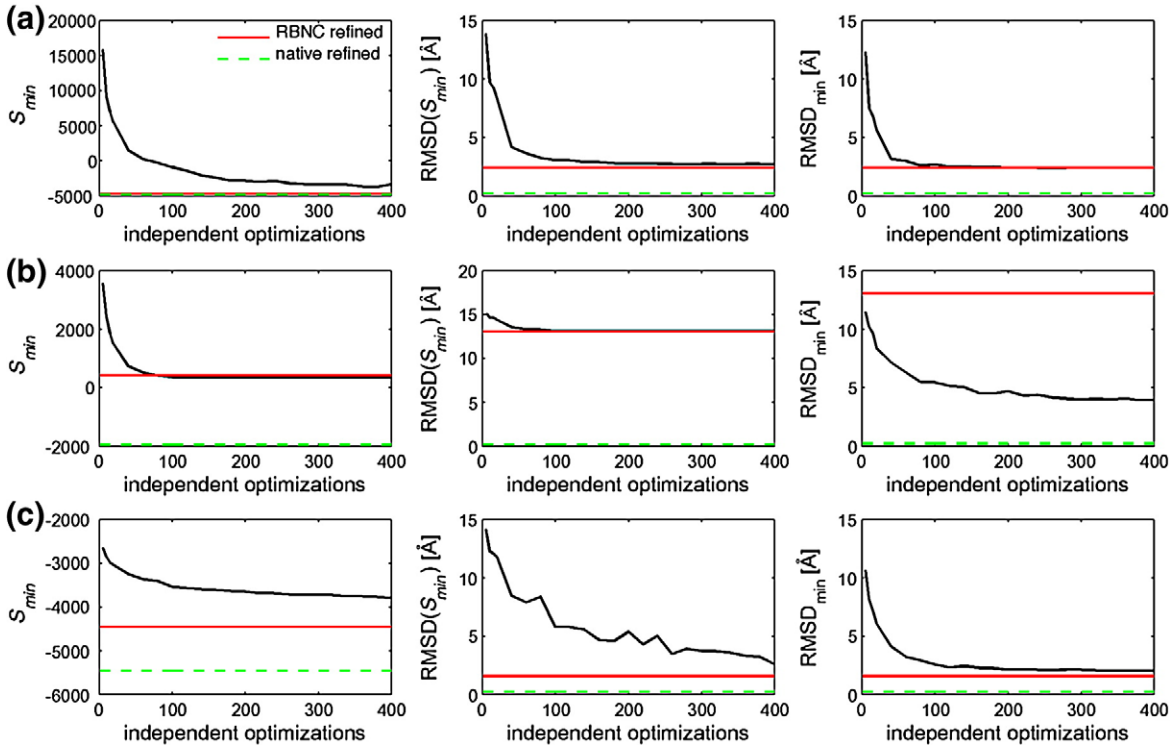

Fig. 6. Scoring *versus* sampling. (a) The minimum $S_{\min}$ score (left), the corresponding C^{α} RMSD (middle), and the minimum C^{α} RMSD ($RMSD_{min}$; right) for all models are plotted as a function of the number of independent optimizations for the benchmark case 1ha0. For comparison, the values for the best scores achieved for local sampling in the vicinity of the RBNC state (RBNC refined) and for local optimization of the native bodies in the vicinity of the native configuration are also shown. (b) Dimeric complex 1ibr. (c) Three-domain protein 1ko9.

mum of *S* could only be reached using global sampling if we performed at least 1000 independent optimizations. However, when a sufficiently accurate starting configuration is available (e.g., from a similar template protein), highly accurate configuration models can be obtained using only 100 independent optimizations (C^{α} RMSD<2 Å; Supplementary Table S1).

Discussion

We incorporated information from a SAXS profile into protein structure modeling by satisfaction of spatial restraints implemented in our IMP.[3,26] We then benchmarked IMP by modeling quaternary structures of multidomain proteins and protein assemblies using rigid domains and proteins, respectively. We discuss here (i) the relationship between our method and the methods of others; (ii) the benefits of integrating protein structure modeling and SAXS fitting; (iii) the limitations arising from inaccurate scoring, imperfect sampling, and errors in rigid bodies; and (iv) the scope for integration of additional information into the modeling process.

Comparison with other methods

Recently, experimental SAXS profiles have also been used to calculate atomic models of proteins by BUNCH[22] and CNS, which rely on NMR-derived restraints as well as a SAXS profile.[24] We now outline similarities and differences between these two approaches and our method.

Our SAXS penalty is similar to the score implemented previously in CNS.[24] Identically, both scores use the Debye formula to treat the excluded solvent and rely on χ^2 as the SAXS penalty. Moreover, they both calculate the partial derivatives, allowing them to use gradient-based optimization techniques. However, the computation of the SAXS penalty and its derivative in our approach is significantly faster compared with the original implementation because we use the electron pair distribution function $P(r)$ for the calculation of χ^2 and its derivative (Supplementary Theory and Methods). The gain in efficiency depends on the granularity of $I(q)$ sampling. For example, we reduce the computation time by 2 orders of magnitude for a data set with 100 data points and more for finer-sampled profiles. This gain in computational speed does not reduce the precision of the calculated SAXS profiles by more than the typical precision of experimental SAXS profiles. Moreover, we can gain additional efficiency relative to CNS through the use of rigid bodies, which required changes in the calculation of the partial derivatives of S_{SAXS} (Supplementary Theory and Methods). The gains in computational efficiency are important because they allowed us to sample the space of possible solutions more exhaustively.

Our SAXS penalty is different from the penalty in BUNCH,[22] which is calculated by CRYSOL[41] and upweighs high-frequency components in χ^2. While we also tested such a scoring function, it did not result in a significant improvement for our benchmark. For example, we modeled the diphtheria toxin using a χ^2 that weights frequencies according to q^2. The corresponding S_{SAXS} term did not result in a more accurate model if used only in conjunction with S_{stereo} (Supplementary Fig. S1). It also did not produce a higher yield of near-native configurations when using the full score S, compared with the uniform weighting (Supplementary Fig. S1). Also, in our experimental test case XI, the IMP and BUNCH scores are highly correlated (Supplementary Fig. S2). Therefore, for parsimony, we used the scoring function with uniform weighting of the SAXS profile at different frequencies. Moreover, unlike our score, the BUNCH SAXS penalty includes two additional fitting parameters that modify I_m[41] corresponding to the average displaced volume per atomic group and the density of the hydration layer. The slightly better fit between the CRYSOL and experimental profiles of lysozyme (Fig. 2) is likely due to these additional parameters, but their effect is negligibly small ($\Delta\chi^2 = 0.05$). Thus, the additional fitting parameters in CRYSOL cannot explain the errors in the calculated profile of XI for $0.11\ \text{Å}^{-1} < q < 0.12\ \text{Å}^{-1}$ (Fig. 5a). Likewise, the program solX[45] calculates SAXS profiles of known atomic structures that fit experimental data as well as CRYSOL's profiles[6] without using fitting parameters other than the protein concentration, similarly to our approach. It is unlikely that errors in our experimental data collection are responsible for this discrepancy because independent SAXS measurements of XI yielded comparable differences between experimental profiles and profiles computed from the crystallographic XI structure.[39] Moreover, qualitatively similar discrepancies could be observed for *E. coli* aspartate transcarbamylase, a 303-kDa homododecamer, in the original CRYSOL publication.[41] The calculated aspartate transcarbamylase profile has a pronounced local minimum at $q \approx 0.08\ \text{Å}^{-1}$, while the corresponding minimum in the experimental SAXS profile is significantly less prominent. In addition to the imperfect calculation of the SAXS profile for a given structure, these discrepancies might also reflect structural differences between the crystallographic and solution structures.

Our optimization protocol consists of independent minimizations of the scoring function from many random starting configurations, with the aim to sample the entire configuration space (for the global optimization mode; Fig. 1). For each minimization, we use a simulated annealing biased Monte Carlo algorithm; each Monte Carlo step is followed by a local quasi-Newton relaxation, for which the first derivatives of the scoring function are needed. Thus, the biased Monte Carlo process samples only local minima. In contrast, BUNCH[22] uses a conventional simulated annealing Monte Carlo protocol, in which the sampling is not restricted to local minima. However, in many applications, such as X-ray crystallography,[46] NMR spectroscopy,[47] comparative protein structure modeling,[28] and *ab initio* structure prediction of proteins[48] and assemblies,[49] optimization methods that use the first derivatives are known to be significantly more efficient than Monte Carlo schemes. Thus, we efficiently implemented and used the first derivatives in our optimization.

A perennial problem in structure modeling is whether or not an optimization scheme finds all the good scoring solutions. To at least partly address this problem, we run many minimizations in parallel and independently on a large computer cluster (i.e., hundreds of nodes). The resulting large sample of solutions is then clustered, to present them more parsimoniously for subsequent analysis. By construction, the structures in one cluster are generally distinct from the structures in another cluster; they involve dissimilar interfaces and have C^α RMSD values worse than 12 Å (Figs. 4 and 5). An analysis of the sampling shows that our protocol usually finds the global minimum for up to four rigid bodies (below); thus, most good scoring local minima are also expected to be sampled in these cases. A large computer cluster with at least 100 processors is currently needed for efficient use of our method, perhaps limiting its practical utility. Nevertheless, such computing clusters are becoming increasingly available to many users. In addition, our software is being adapted to run on graphics processing units, such as NVIDIA's Tesla with 240 processors‡, which might enable efficient application on a single desktop computer.

Integration of SAXS and protein structure modeling

Integrative computational methods can exploit various kinds of spatial information to determine the assembly structures at higher accuracy and precision than are possible based on each individual type of data;[3,26] in conjunction, pieces of data that are relatively uninformative by themselves can still result in accurate and precise models of proteins and assemblies.

Here, we combined a SAXS profile with information about protein structures that can be calculated only from their sequence. Specifically, we supplemented the SAXS term (S_{SAXS}) by the penalties for steric clashes ($S_{overlap}$), an atomic distance-dependent statistical potential (S_{DOPE}), stereochemistry restraints from a molecular mechanics force field (S_{stereo}), and symmetry restraints (S_{sym}). Importantly, an integration of these different types of information can significantly improve modeling accuracy relative to relying on a SAXS profile or protein structure modeling alone (e.g., Fig. 3; 1cb6 in Table 1; XI in Fig. 5). Moreover, the sum of the different terms was not observed to be less accurate than any of the individual terms; in other words,

‡ http://www.nvidia.com/object/tesla_c1060.html

when the final configurations were inaccurate, no optimization of the individual terms, including S_{SAXS}, produced an accurate configuration.

Our method does not necessarily predict unique best scoring solutions; due to the low information content of the input restraints, models from different clusters can have comparable scores. For example, S_{SAXS} is completely invariant to rotations of a spherical rigid body. For the benchmark case 1cb6, models from the near-native cluster and a nonnative cluster have similar scores (Supplementary Table S1). Nevertheless, due to the combination of different types of information, the number of distinct configurations compatible with the input data is generally much smaller compared with that using only a single type of information (e.g., Fig. 3).

Influence of scoring, sampling, and rigid bodies on model accuracy

Our benchmark allows us to assess the limitations of the protocol and highlight opportunities for future research. Modeling by optimization depends on two conditions: (i) the scoring function needs to have a global minimum at the native or near-native state and (ii) the sampling needs to be sufficiently thorough to find the global or near-global minimum. Therefore, we tested the degree to which our method is limited by the accuracy of the scoring function and the thoroughness of sampling. We also asked how accurate the rigid bodies need to be so that the scoring function still has the global minimum at the native state. The assessment of sampling allowed us to judge when a global search without reliance on a suitable initial structure can be successful or, conversely, when we need a sufficiently accurate initial model so that at least a near-native state can be found by local sampling alone. We also analyzed the accuracy of the method as a function of the number of rigid bodies, which allows us to further qualify sampling and scoring limitations.

For systems of two rigid bodies (i.e., two-domain proteins and binary protein complexes), even the global sampling produced numerous configurations close to the global minimum of the scoring function S (Fig. 6a and b; Supplementary Tables S1 and S2). Thus, the accuracy of these models is largely determined by the accuracy of S. The global minimum of S corresponded to the native or near-native state only if the rigid bodies were not too distorted (i.e., C^{α} RMSD of less than 3 Å). Therefore, as expected, the accuracy of the rigid bodies crucially influences the landscape of S.

Among the individual terms of S, it is not surprising that S_{DOPE} sometimes favors nonnative configurations if sufficiently distorted rigid bodies are used. Errors in the positions of exposed atoms interfere with their packing, which is the aspect scored by S_{DOPE}. Moreover, distorted rigid bodies can cause steric clashes and thus increase $S_{overlap}$. In contrast to S_{DOPE}, a strong dependence of S_{SAXS} on the rigid-body accuracy is somewhat surprising; in all cases, the rigid domains and proteins possessed the correct fold, but S_{SAXS} did not always favor near-native configurations. If the C^{α} RMSD errors of the rigid bodies were above approximately 3 Å, configurations with minimum S_{SAXS} were significantly different from the native state (i.e., higher than 18 Å for C^{α} RMSD). This result can be explained by the ambiguity of the SAXS restraint; many different quaternary structures have a similar envelope and therefore similar SAXS profiles. Subtle structural differences can then make the S_{SAXS} of near-native configurations worse than that for some nonnative configurations. The deteriorating effect of errors in the rigid bodies on the accuracy of the predicted configurations turns out to be most pronounced for complexes (Table 2; Supplementary Table S2). For multidomain proteins, the connecting linker limits the number of possible configurations. Thus, the number of possible false minima of S is generally larger for protein complexes than for monomeric proteins. We anticipate that the precise degree of tolerable structural differences will also depend on the shape of the rigid body: If a rigid body does not have a distinctive shape (e.g., it is a sphere or a cylinder), small distortions can be sufficient to alter the position of the global minimum; for more distinct shapes, the sensitivity of the prediction accuracy to the rigid-body errors is expected to be smaller.

For three or more rigid bodies, the typical number of independent optimizations we used (1000) was insufficient to reliably find the global minimum of S in a global search (Fig. 6c). For four domains (benchmark case 1cb6), we increased the number of initial configurations to 5000, requiring 3 days on 200 central processing units. Such an exhaustive sampling is impossible without using a large computer cluster. However, configurations close to the global minimum of S could be found at dramatically reduced computational cost using local sampling if the template (initial) configuration was sufficiently close to the native state (C^{α} RMSD$<$10 Å). In such a case, the local search is more efficient than the global search because no computing time is wasted on searching far from the native configuration. The development of computationally more expensive and sophisticated sampling strategies may allow sampling the configurations of five or more domains with a larger radius of convergence than that of our present optimization.

To overcome the limitations on prediction accuracy arising from rigid-body errors, we probably need to abandon the rigid-body approximation. In the field of protein–protein docking, simultaneous sampling of different component conformations and their configuration has been described.[49] Similar approaches could be used for fitting a configuration to a SAXS profile, requiring both more sophisticated scoring functions and sampling algorithms than those described here. The use of high-angle scattering data might allow us to compute atomic structures more accurately, but this goal would require flexible modeling of the atomic structures.

Structural changes at the domain level result in signals at frequencies beyond $q=0.5$ Å^{-1}, which we did not consider in our calculations because we represented domains and proteins as rigid bodies.

Our protocol currently relies on the sample protein or complex existing in a single state. However, proteins and complexes can exist in equilibrium among different states, corresponding to varied packing between secondary-structure segments, domains, and proteins, as well as variations in unstructured regions, such as long domain linkers. An approach to score a given ensemble of models has been proposed recently.[50] Our protocol could also be extended to optimize an ensemble of models using a similar score. However, the increased computational effort as well as the limited information in SAXS profiles will limit such approaches.

Integration of SAXS profiles with additional information

Given the relatively low information content of a SAXS profile and the limitations in our protein structure modeling terms, incorporating additional information into the scoring function S is desirable. Using supplementary information is further justified by the sensitivity of S to rigid-body errors and possible systematic experimental errors in a SAXS profile (e.g., due to aggregation of macromolecules in solution). Further integration is facilitated by our implementation of SAXS fitting in IMP, which can already produce models by simultaneously satisfying a large variety of other spatial restraints. In addition to the SAXS and modeling terms used here, IMP can incorporate (i) restraints implied by an alignment between the modeled sequence and related structures;[28] (ii) restraints implied by an alignment of the modeled sequence and many short segments of known structure; (iii) bioinformatics analysis of protein interaction modes;[51] (iv) protein–protein docking that is restrained by the composition of interacting surfaces determined by NMR spectroscopy (M. F. Kim, A. Sali, V. Dotsch, and M. Rees, unpublished results); (v) symmetry and density from a cryo-EM map of the assembly;[52] and (vi) proximity of subunits inferred from immunoaffinity purifications, yeast two-hybrid system, footprinting, and chemical cross-linking.[26] The global shape information from SAXS is especially complementary to local restraints, such as the atomic distance restraints derived from chemical cross-linking detected by mass spectrometry.[53] Another attractive application is the integration of SAXS for structural characterization of component structures of a large assembly whose overall density is determined by cryo-EM.

Conclusions

An accurate quaternary-structure model of a protein or a protein assembly can be obtained using only a SAXS profile, stereochemistry restraints from a molecular mechanics force field, and an atomic distance-dependent statistical potential provided that sufficiently accurate approximations for the constituent domain and protein structures are available. Otherwise, the predictions are generally ambiguous and have large errors. Our integration of a SAXS profile into modeling by satisfaction of spatial restraints will facilitate further integration of different kinds of data for structure determination of proteins and their assemblies.

Acknowledgements

F.F. is grateful for a long-term fellowship from the Human Frontier Science Project Organization. K.A. K. was supported by a National Defense Science and Engineering Graduate fellowship. The Stanford Synchrotron Radiation Laboratory is funded by the Department of Energy Basic Energy Sciences Program, and the Stanford Synchrotron Radiation Laboratory Structural Molecular Biology Program is supported by the Department of Energy Office of Biological and Environmental Research and the National Institutes of Health National Center for Research Resources Biomedical Technology Program through grant P41 RR001209. D.A.A. has been supported by the Howard Hughes Medical Institute; D.A.A. and A.S. have been supported by a University of California Discovery Grant (bio03-10401/Agard). A.S. has also been supported by the Sandler Family Supporting Foundation, the National Institutes of Health (R01 GM54762, R01 GM083960, U54 RR022220, and PN2 EY016525), the National Science Foundation (EIA-032645 and IIS-0705196), Hewlett–Packard, NetApps, IBM, and Intel. We thank Maya Topf, Narayanan Eswar, Frank Alber, Fred Davis, Min-Yi Shen, and Marc Marti-Renom for fruitful discussions.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2008.07.074

References

1. Sali, A., Glaeser, R., Earnest, T. & Baumeister, W. (2003). From words to literature in structural proteomics. *Nature*, **422**, 216–225.
2. Robinson, C. V., Sali, A. & Baumeister, W. (2007). The molecular sociology of the cell. *Nature*, **450**, 973–982.
3. Alber, F., Förster, F., Korkin, K., Topf, M., & Sali, A. (in press). Integrating diverse data for structure determination of macromolecular assemblies. *Annu. Rev. Biochem.* **77**, 443–477.
4. Doniach, S. (2001). Changes in biomolecular conformation seen by small angle X-ray scattering. *Chem. Rev.* **101**, 1763–1778.

5. Koch, M. H., Vachette, P. & Svergun, D. I. (2003). Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Q. Rev. Biophys.* **36**, 147–227.
6. Putnam, C. D., Hammel, M., Hura, G. L. & Tainer, J. A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution.*Q. Rev. Biophys.* **40**, 191–285; Cambridge Journals Online.
7. Svergun, D. & Koch, M. (2003). Small-angle scattering studies of biological macromolecules in solution. *Rep. Prog. Phys.* **66**, 1735–1782.
8. Das, R., Kwok, L. W., Millett, I. S., Bai, Y., Mills, T. T., Jacob, J. *et al.* (2003). The fastest global events in RNA folding: electrostatic relaxation and tertiary collapse of the *Tetrahymena* ribozyme. *J. Mol. Biol.* **332**, 311–319.
9. Canady, M. A., Tsuruta, H. & Johnson, J. E. (2001). Analysis of rapid, large-scale protein quaternary structural changes: time-resolved X-ray solution scattering of *Nudaurelia capensis* omega virus (NomegaV) maturation. *J. Mol. Biol.* **311**, 803–814.
10. Davies, J. M., Tsuruta, H., May, A. P. & Weis, W. I. (2005). Conformational changes of p97 during nucleotide hydrolysis determined by small-angle X-ray scattering. *Structure*, **13**, 183–195.
11. Yamagata, A. & Tainer, J. A. (2007). Hexameric structures of the archaeal secretion ATPase GspE and implications for a universal secretion mechanism. *EMBO J.* **26**, 878–890; Epub: January 25, 2007.
12. Krukenberg, K. A., Förster, F., Rice, L., Sali, A., & Agard, D. A. (in press). A novel conformation of *E. coli*Hsp90 in solution: insights into the conformational dynamics of Hsp90. *Structure*, **16**, 755–765.
13. Sondermann, H., Nagar, B., Bar-Sagi, D. & Kuriyan, J. (2005). Computational docking and solution X-ray scattering predict a membrane-interacting role for the histone domain of the Ras activator son of sevenless. *Proc. Natl. Acad. Sci. USA*, **102**, 16632–16637.
14. Zheng, W. & Doniach, S. (2005). Fold recognition aided by constraints from small angle X-ray scattering data.*Protein Eng. Des. Sel.* **18**, 209–219; Epub: April 21, 2005.
15. Stuhrmann, H. (1970). 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., Sect. A*, **26**, 297–306.
16. Svergun, D. & Stuhrmann, H. (1991). New developments in direct shape determination from small-angle scattering: 1. Theory and model calculations. *Acta Crystallogr., Sect. A*, **47**, 736–744.
17. Chacon, P., Moran, F., Diaz, J. F., Pantos, E. & Andreu, J. M. (1998). Low-resolution structures of proteins in solution retrieved from X-ray scattering with a genetic algorithm. *Biophys. J.* **74**, 2760–2775.
18. Walther, D., Cohen, F. E. & Doniach, S. (2000). Reconstruction of low-resolution three-dimensional density maps from one-dimensional small-angle X-ray solution scattering data for biomolecules. *J. Appl. Crystallogr.* **33**, 350–363.
19. Svergun, D. I. (1999). Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.* **76**, 2879–2886.
20. Svergun, D. I., Petoukhov, M. V. & Koch, M. H. (2001). Determination of domain structure of proteins from X-ray solution scattering. *Biophys. J.* **80**, 2946–2953.
21. Petoukhov, M. V., Eady, N. A., Brown, K. A. & Svergun, D. I. (2002). Addition of missing loops and domains to protein models by X-ray solution scattering. *Biophys. J.* **83**, 3113–3125.
22. Petoukhov, M. V. & Svergun, D. I. (2005). Global rigid body modeling of macromolecular complexes against small-angle scattering data. *Biophys. J.* **89**, 1237–1250.
23. Petoukhov, M. V., Monie, T. P., Allain, F. H., Matthews, S., Curry, S. & Svergun, D. I. (2006). Conformation of polypyrimidine tract binding protein in solution. *Structure*, **14**, 1021–1027.
24. Grishaev, A., Wu, J., Trewhella, J. & Bax, A. (2005). Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data. *J. Am. Chem. Soc.* **127**, 16621–16628.
25. Wu, Y., Tian, X., Lu, M., Chen, M., Wang, Q. & Ma, J. (2005). Folding of small helical proteins assisted by small-angle X-ray scattering profiles. *Structure*, **13**, 1587–1597.
26. Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D. *et al.* (2007). Determining the architectures of macromolecular assemblies. *Nature*, **450**, 683–694.
27. MacKerell, A. D., Jr, Bashford, D., Bellott, M., Dunbrack, R. L., Jr, Evanseck, J., Field, M. J. *et al.* (1998). All-atom empirical potential for molecular modeling and dynamics studies of protein. *J. Phys. Chem. B*, **102**, 3586–3616.
28. Sali, A. & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815.
29. Shen, M. Y. & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–2524.
30. Debye, P. (1915). Zerstreuung von Roentgenstrahlen. *Ann. Phys.* **46**, 809–823.
31. Fraser, R., MacRae, T. & Suzuki, E. (1978). An improved method for calculating the contribution of solvent to the X-ray diffraction pattern of biological molecules. *J. Appl. Crystallogr.* **11**, 693–694.
32. Shanno, D. & Phua, K. (1980). Remark on algorithm 500. *ACM Trans. Math. Software*, **6**, 618–622.
33. Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes in FORTRAN: The Art of Scientific Computing*, 2nd edit. Cambridge University Press, Cambridge, UK.
34. Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, **2**, 241–254.
35. Marti-Renom, M. A., Ilyin, V. A. & Sali, A. (2001). DBAli: a database of protein structure alignments. *Bioinformatics*, **17**, 746–747.
36. Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S. *et al.* (2003). CAPRI: a Critical Assessment of PRedicted Interactions. *Proteins*, **52**, 2–9.
37. Goddard, T. D., Huang, C. C. & Ferrin, T. E. (2007). Visualizing density maps with UCSF Chimera. *J. Struct. Biol.* **157**, 281–287.
38. Smolsky, I. L., Liu, P., Niebuhr, M., Ito, L., Weiss, T. M. & Tsuruta, H. (2007). Biological small-angle X-ray scattering facility at the Stanford Synchrotron Radiation Laboratory. *J. Appl. Crystallogr.* **40**, s453–s458.
39. Kozak, M. (2005). Direct comparison of the crystal and solution structure of glucose/xylose isomerase from *Streptomyces rubiginosus*. *Protein Pept. Lett.* **12**, 547–550.
40. Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J. & Svergun, D. I. (2003). Primus: a Windows

PC-based system for small angle scattering data analysis. *J. Appl. Crystallogr.* **36**, 1277–1282.
41. Svergun, D., Barberato, C. & Koch, M. (1995). CRYSOL—a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* **28**, 768–773.
42. Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F. & Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325.
43. Alber, F., Kim, M. F. & Sali, A. (2005). Structural characterization of assemblies from overall shape and subcomplex compositions. *Structure*, **13**, 435–445.
44. Mendez, R., Leplae, R., De Maria, L. & Wodak, S. J. (2003). Assessment of blind predictions of protein–protein interactions: current status of docking methods. *Proteins*, **52**, 51–67.
45. Zuo, X. & Tiede, D. M. (2005). Resolving conflicting crystallographic and NMR models for solution-state DNA with solution X-ray diffraction. *J. Am. Chem. Soc.* **127**, 16–17.
46. Brunger, A. T., Kuriyan, J. & Karplus, M. (1987). Crystallographic *R* factor refinement by molecular dynamics. *Science*, **235**, 458–460.
47. Brunger, A. T., Clore, G. M., Gronenborn, A. M. & Karplus, M. (1986). Three-dimensional structure of proteins determined by molecular dynamics with interproton distance restraints: application to crambin. *Proc. Natl. Acad. Sci. USA*, **83**, 3801–3805.
48. Bradley, P., Misura, K. M. & Baker, D. (2005). Toward high-resolution *de novo* structure prediction for small proteins. *Science*, **309**, 1868–1871.
49. Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A. & Baker, D. (2003). Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* **331**, 281–299.
50. Bernado, P., Mylonas, E., Petoukhov, M. V., Blackledge, M. & Svergun, D. I. (2007). Structural characterization of flexible proteins using small-angle X-ray scattering.*J. Am. Chem. Soc.* **129**, 5656–5664; Epub: April 6, 2007.
51. Korkin, D., Davis, F. P., Alber, F., Luong, T., Shen, M. Y., Lucic, V. *et al.* (2006). Structural modeling of protein interactions by analogy: application to PSD-95. *PLoS Comput. Biol.* **2**, e153.
52. Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W. & Sali, A. (2008). Protein structure fitting and refinement guided by cryo-EM density. *Structure*, **16**, 295–307.
53. Seebacher, J., Mallick, P., Zhang, N., Eddes, J. S., Aebersold, R. & Gelb, M. H. (2006). Protein cross-linking analysis using mass spectrometry, isotope-coded cross-linkers, and integrated computational data processing. *J. Proteome Res.* **5**, 2270–2282.