Accurate assessment of mass, models and resolution by small-angle scattering

Robert P. Rambo & John A. Tainer

Modern small-angle scattering (SAS) experiments with X-rays or neutrons provide a comprehensive, resolution-limited observation of the thermodynamic state. However, methods for evaluating mass and validating SAS-based models and resolution have been inadequate. Here we define the volume of correlation, \( V_c \), a SAS invariant derived from the scattered intensities that is specific to the structural state of the particle, but independent of concentration and the requirements of a compact, folded particle. We show that \( V_c \) defines a ratio, \( Q_S \), that determines the molecular mass of proteins or RNA ranging from 10 to 1,000 kilodaltons. Furthermore, we propose a statistically robust method for assessing model-data agreements (\( \chi^2_{\text{tree}} \)) akin to cross-validation. Our approach prevents over-fitting of the SAS data and can be used with a newly defined metric, \( R_{\text{SAS}} \), for quantitative evaluation of resolution. Together, these metrics (\( V_c \), \( Q_S \), \( \chi^2_{\text{tree}} \) and \( R_{\text{SAS}} \)) provide analytical tools for unbiased and accurate macromolecular structural characterizations in solution.

Achieving reliable, high-throughput structural characterizations of biological macromolecular complexes is a key challenge in the modern structural-genomics era. In principle, SAS with X-rays (SAXS) or neutrons (SANS) can meet this challenge by efficiently providing information that fully describes the structural state of a macromolecule in solution. SAS can determine a scattering particle’s radius of gyration (\( R_g \)), volume (\( V_P \)), surface-to-volume ratio and correlation length (\( \ell_c \)), with the latter three physical parameters dependent on the Porod invariant (\( Q \)), an empirical SAS value defined for compact folded particles. \( Q \) is unique to a scattering experiment and requires convergence of the SAS data at high scattering vectors (\( q \), Å\(^{-1} \)) versus \( q \) (Kratky) plot. Convergence defines an enclosed area where the degree of convergence reflects the compact (bounded area), flexible or unfolded (unbounded area) solution states (Fig. 1a). Consequently, non-convergence leaves \( Q \) undefined and paradoxically implies that \( V_P \) and \( \ell_c \) are undefined for flexible particles (Supplementary Fig. 1 and Supplementary Notes). This observation leaves \( R_g \) as the only structural parameter that can be reliably derived from SAS data on flexible systems.

Defining the \( V_c \)

SAS is uniquely capable of providing structural information on all particle types, including flexible systems such as intrinsically unstructured proteins. Here, we overcome current limitations of SAS analyses by deriving a SAS invariant, \( V_c \). \( V_c \) is defined as the ratio of the particle’s zero angle scattering intensity, \( I(0) \), to its total scattered intensity (Supplementary Notes). The total scattered intensity is the integrated area of the SAS data transformed as \( q I(q) \) versus \( q \). Unlike the Kratky plot, we observe that the integral of \( q I(q) \) versus \( q \) converges for both folded-compact and unfolded-flexible particles (Fig. 1b). The aforementioned ratio at particle concentration (\( c \)) and contrast (\( p \)) given by

\[
V_c = \frac{\int q I(q) dq}{2\pi l_c} = \frac{V_P}{V_P} \left( \frac{c \left( \rho_0 - \rho \right)^2}{2\pi l_c^2} \right) = \frac{V_P}{2\pi l_c^2}
\]

reduces to the particle’s volume (\( V_P \)) per self-correlation length (\( \ell_c \)) with units of Å\(^2\).

This derivation asserts that \( V_c \), like \( R_g \), can be calculated from a single SAS curve and is concentration independent. We validated concentration independence using well-characterized macromolecules of differing composition and mass. Specifically, for the 173-kDa protein glucose isomerase and the 51-kDa P4–P6 RNA domain from the Tetrahymena group I intron, SAS data collected at seven concentrations ranging from 0.2 to 3 mg ml\(^{-1} \) showed concentration independence: 86% of the variance was contained within 4% of the mean. Further analysis of seven additional protein and RNA samples confirmed the concentration independence (Fig. 1c): 65% of the variance was contained within 2% of the mean, suggesting that \( V_c \) is constant across the concentration ranges for all macromolecular shapes and compositions tested.

\( V_c \) is defined by the particle’s \( l_c \) and implies that a change in conformation should change \( V_c \) (Fig. 1d). We observed this prediction for both the bacterial \( S \)-adenosylmethionine (SAM)–I riboswitch and Pyr1, a plant hormone-binding protein. For these macromolecules, ligand binding decreased both \( R_g \) and \( V_c \), consistent with reported compaction upon binding [11-13]. Furthermore, we examined Mg\(^{2+} \)-dependent structured RNAs for folding by SAXS. Measurements of both the bacterial SAM-I riboswitch and turnip yellow mosaic virus (TYMV) transfer-RNA-like structure (TLS) without Mg\(^{2+} \) displayed the classic hyperbolic feature of a monodisperse multiconformation Gaussian ensemble in the Kratky plot (Supplementary Fig. 1). As predicted, flexibility in the absence of Mg\(^{2+} \) increased the experimentally determined \( V_c \) values (by 14.5% for TYMV TLS and 21% for SAM-I RNA) compared to their compact Mg\(^{2+} \)-folded states (Table 1). Collectively, the observed ligand-dependent changes in \( V_c \) for both Pyr1 and SAM-I RNA or Mg\(^{2+} \)-dependent changes in \( V_c \) for TYMV TLS and SAM-I RNA suggest that \( V_c \) is an informative descriptor of the macromolecular state.

Particle mass determination by \( Q_R \)

Accurate determination of molecular mass has been one of the main difficulties in SAS analysis. Existing methods require an accurate particle concentration, the assumption of a compact near-spherical shape, or SAS measurements on an absolute scale [15-18]. As these
requirements hinder both accuracy and throughput of mass estimates by SAS, we sought to establish a SAS-based statistic suitable for determining the molecular mass of proteins, nucleic acids or mixed complexes in solution without concentration or shape assumptions. We calculated \( R_g \) and \( V_c \) from simulated SAXS profiles for 9,446 protein structures from the Protein Data Bank (PDB)\(^{19} \), ranging in molecular weight from 8 to 400 kDa. We discovered that a parameter, \( Q_{Rg} \), defined as the ratio of the square of \( V_c \) to \( R_g \) with units of \( \AA^3 \), is linear versus molecular mass in a log–log plot (Figs 2, 3 and Supplementary Fig. 2). The linear relationship is a power-law relationship given by:

\[
mass = \left( \frac{Q_{Rg}}{c} \right)^{1/k}
\]

which determines the empirical mass of the scattering biological particle allowing for the direct assessment of oligomeric state and sample quality. Parameters \( k \) and \( c \) are empirically determined and specific to the class of macromolecular particle (Supplementary Fig. 3), with \( c \) as Euler’s number. \( V_c \) and \( R_g \) are both contrast and concentration independent, thus the determination of molecular mass using \( Q_{Rg} \) can be made from SAXS data collected under diverse buffer conditions and concentrations, albeit free of inter-particle interference. In fact, this linear relationship produced an average mass error <4% for the 9,446 proteins in the in vacuo-simulated data set (Fig. 2).

Calculations of \( Q_{Rg} \) from simulated and experimental (Supplementary Tables 1 and 2) buffer-subtracted SAXS data of proteins, mixed protein–nucleic acid complexes or RNA alone (Fig. 3a, b) further verified the power-law relationship between \( Q_{Rg} \) and mass. The mass errors for protein and RNA gel-filtration-purified SAXS samples were

Table 1 | Condition-dependent changes in SAXS invariants

| Macromolecule          | \( V_c \) (kDa) | \( R_g \) (Å) | \( V_c^+ \) (kDa) | SASS mass (kDa) |
|------------------------|----------------|--------------|-----------------|---------------|
| SAM-I (bound): mixture† | 460 (± 2)      | 34.4 (± 0.3) | 80,000          | 50.3          |
| SAM-I (free): mixture  | 407 (± 2)      | 31.0 (± 0.2) | 76,000          | 44.9          |
| SAM-I (bound)          | 280 (± 4)      | 22.8 (± 0.4) | 40,000          | 31.4          |
| SAM-I (free)           | 295 (± 4)      | 24.7 (± 0.7) | 48,000          | 32.0          |
| SAM-I – Mg\(^{2+} \)   | 339 (± 12)     | 31.6 (± 1.0) | ND              | 32.8          |
| P4–P6 RNA domain: mixture | 478 (± 1.0) | 31.0 (± 0.1) | 105,000         | 58.2          |
| P4–P6 RNA domain       | 414 (± 5)      | 29.4 (± 0.2) | 73,000          | 50.8          |
| PYR1 (bound)           | 319 (± 0.5)    | 20.6 (± 0.9) | 59,000          | 41.9          |
| PYR1 (free)            | 343 (± 8)      | 23.2 (± 0.8) | 74,000          | 40.2          |
| TyMV + Mg\(^{2+} \)    | 324 (± 2)      | 25.9 (± 0.1) | 49,000          | 35.9          |
| TyMV – Mg\(^{2+} \)    | 371 (± 1)      | 29.9 (± 0.1) | ND              | 39.8          |

Uncertainties are the standard deviation of 4–8 independent SAXS data sets.

Figure 1 | Concentration independence and conformational dependence of \( V_c \). a, b, Experimental SAXS data plotted on a relative scale for glucose isomerase (cyan), 94-nucleotide SAM-I riboswitch in the absence of Mg\(^{2+} \) (orange) and human RAD51AP1, an intrinsically unfolded protein (green). a, Data transformed as the Kratky plot, \( q^2 I(q) \) versus \( q \), reveals the parabolic convergence for a folded particle (blue) and divergence for a flexible (orange) or fully unfolded (green) particle. b, Data plotted as \( q^2 I(q) \) versus \( q \) show convergence for both folded and flexible particles. Inset, convergence for a fully unfolded polymer. c, Concentration independence of \( V_c \) for experimental SAXS data. For each of nine samples, relative difference is calculated as the deviation from the mean normalized to the mean. Concentrations ranged from 0.2 to 3 mg ml\(^{-1} \) for glucose isomerase (cyan), P4–P6 domain (open red and solid green), xylanase (orange), TyMV TLS RNA (UUAG; solid black), poliovirus del8 competitive inhibitor RNA\(^{31} \) (open purple), Agrobacterium tumefaciens RNase P (open black), SAM-I riboswitch with Mg\(^{2+} \) and ligand (large open purple) and SAM-I riboswitch in the absence of Mg\(^{2+} \) (large open yellow). Data was collected to \( q_{\text{max}} = 0.32 \) Å\(^{-1} \) with the exception of solid green \( q_{\text{max}} = 0.52 \) Å\(^{-1} \). x-axis (sample number) refers to the different concentrations for each sample increasing from left to right. d, Correlated changes in \( V_c \) (red) and \( R_g \) (cyan) for conformations of SAM-I riboswitch (PDB code, 2GIS) simulated from molecular dynamics with CNS\(^{29} \). Horizontal lines demonstrate for \( R_g \) or \( V_c \) that a single value can map to multiple conformations. Dual specification of both \( R_g \) and \( V_c \) reduces multiplicity (vertical bars). Relative change represents the difference calculated from the starting model 2GIS. Asterisks denote the time step of the displayed conformation.

Figure 2 | Defining the power-law relationship between \( V_c \), \( R_g \) and protein mass. MW, molecular mass. \( V_c \) and \( R_g \) were determined from theoretical atomic X-ray scattering profiles for 9,446 protein PDB\(^{19} \) structures. For each profile, SAXS data were simulated to a maximum \( q = 0.5 \) Å\(^{-1} \) (~13 Å). Various ratios of \( V_c \) and \( R_g \) against protein mass were examined in a log–log plot. The linear relationship observed for the ratio \( V_c^2/R_g \) (black) suggests that a power-law relationship exists between the ratio and particle mass of the form \( V_c^2/R_g = c \) (mass)\(^k \). The ratio, \( V_c^2/R_g \), is defined by units of \( \AA^3 \) with mass in Daltons. Additional ratios examined (green, cyan, grey and red) displayed asymmetric nonlinear relationships. In green, the fit included generic \( m \) (0.9246 ± 0.0008) and \( c \) (1.892 ± 0.0005) parameters in a nonlinear surface optimization resulting in an average mass error of 4.9 ± 4.3%. Fitting the linear power-law relationship (black) produces an average mass error of 4.0 ± 3.6%. Truncation of the data to \( q = 0.3 \) Å\(^{-1} \) (~21 Å resolution) increases the mass error by 0.6% (Supplementary Fig. 2).
Resisting over-fitting with $\chi^2_{\text{free}}$

We tested $\chi^2_{\text{free}}$ on SAXS data for xylanase at pH 7.2 (Fig. 4a). Based on the fit to the crystallographic structure (PDB code, 1REF; $\chi^2 = 3.9$), SAXS data imply an alternate conformation in solution. Using 1REF as a reference structure, 1,600 conformations were generated and used in a conventional all-data $\chi^2$ determination. Approximately 7% of the sampling procedure for estimating $\chi^2$ that partitions a SAS data set into $n_s$ equal bins for a given $d_{\text{max}}$. A randomly sampled data point is taken from each bin creating a $n_s$-length data vector that is used in $\chi^2$. To minimize outlier influence, $\chi^2$ is taken as the median over $k$ sampling rounds (typically $k = 1,001$) yielding a statistic we call $\chi^2_{\text{free}}$. Analogous to $R_{\text{free}}$, $\chi^2_{\text{free}}$ uses a cross-validation scheme that excludes data from each bin during a round. This technique is akin to the robust least-trimmed squares method$^{25}$ and provides resistance to outliers, preventing over-fitting and the misidentification of models$^{26,27}$.

Cross-validating SAXS model-data agreements

Atomistic modelling of SAS data relies on the reduced chi-square ($\chi^2$) error-weighted scoring function$^{21,22}$ that can be unreliable with moderately noisy data sets or over-estimated degrees of freedom (Supplementary Figs 4 and 5). This can lead to over-fitting and model misidentification. In crystallographic and NMR analyses, cross-validation statistical methods mitigate over-fitting and increase confidence in selected model(s)$^{23,24}$. Here, we present an analogous robust statistical method based on the Nyquist–Shannon sampling and the noisy-channel coding theorems (Supplementary Notes) for evaluating structural models against SAS data.

For a given maximum dimension ($d_{\text{max}}$), the sampling theorem$^9$ determines that the number of unique, evenly distributed observations, $n_d$, required to represent a particle to a maximum scattering vector ($d_{\text{max}}$) is given by ($d_{\text{max}}q_{\text{max}})^{-1}$. For example, SAS data to a $q_{\text{max}}$ of 0.3 Å$^{-1}$ determines for xylanase ($d_{\text{max}}$ 44 Å) or for 30S ribosomal particle ($d_{\text{max}}$ 240 Å) that the minimum number of observations is 4 and 23, respectively. This represents a ~20- to 125-fold over-sampling of a SAS curve composed of 500 observations. The Nyquist–Shannon limit ($n_s$) is the set of maximally independent observations from the band-limited SAS curve (Supplementary Fig. 7). We reasoned that calculating $\chi^2$ from a data set reduced to $n_s$ should more accurately assess the model-data agreement by restricting $\chi^2$ evaluations to the set of independent random variables (Supplementary Notes).

Owing to over-sampling and the uncertainties in $q$, $R(q)$ and $d_{\text{max}}$ determining the exact set of Nyquist–Shannon points will be difficult. Nevertheless, application of the noisy-channel coding theorem guarantees noise-free recovery of the SAS signal (Supplementary Notes and Supplementary Figs 8, 9); therefore, we propose the following

Cross-validating SAXS model-data agreements

Atomistic modelling of SAS data relies on the reduced chi-square ($\chi^2$) error-weighted scoring function$^{21,22}$ that can be unreliable with moderately noisy data sets or over-estimated degrees of freedom (Supplementary Figs 4 and 5). This can lead to over-fitting and model misidentification. In crystallographic and NMR analyses, cross-validation statistical methods mitigate over-fitting and increase confidence in selected model(s)$^{23,24}$. Here, we present an analogous robust statistical method based on the Nyquist–Shannon sampling and the noisy-channel coding theorems (Supplementary Notes) for evaluating structural models against SAS data.

For a given maximum dimension ($d_{\text{max}}$), the sampling theorem$^9$ determines that the number of unique, evenly distributed observations, $n_d$, required to represent a particle to a maximum scattering vector ($d_{\text{max}}q_{\text{max}}$) is given by ($d_{\text{max}}q_{\text{max}})^{-1}$. For example, SAS data to a $q_{\text{max}}$ of 0.3 Å$^{-1}$ determines for xylanase ($d_{\text{max}}$ 44 Å) or for 30S ribosomal particle ($d_{\text{max}}$ 240 Å) that the minimum number of observations is 4 and 23, respectively. This represents a ~20- to 125-fold over-sampling of a SAS curve composed of 500 observations. The Nyquist–Shannon limit ($n_s$) is the set of maximally independent observations from the band-limited SAS curve (Supplementary Fig. 7). We reasoned that calculating $\chi^2$ from a data set reduced to $n_s$ should more accurately assess the model-data agreement by restricting $\chi^2$ evaluations to the set of independent random variables (Supplementary Notes).

Owing to over-sampling and the uncertainties in $q$, $R(q)$ and $d_{\text{max}}$ determining the exact set of Nyquist–Shannon points will be difficult. Nevertheless, application of the noisy-channel coding theorem guarantees noise-free recovery of the SAS signal (Supplementary Notes and Supplementary Figs 8, 9); therefore, we propose the following

Cross-validating SAXS model-data agreements

Atomistic modelling of SAS data relies on the reduced chi-square ($\chi^2$) error-weighted scoring function$^{21,22}$ that can be unreliable with moderately noisy data sets or over-estimated degrees of freedom (Supplementary Figs 4 and 5). This can lead to over-fitting and model misidentification. In crystallographic and NMR analyses, cross-validation statistical methods mitigate over-fitting and increase confidence in selected model(s)$^{23,24}$. Here, we present an analogous robust statistical method based on the Nyquist–Shannon sampling and the noisy-channel coding theorems (Supplementary Notes) for evaluating structural models against SAS data.

For a given maximum dimension ($d_{\text{max}}$), the sampling theorem$^9$ determines that the number of unique, evenly distributed observations, $n_d$, required to represent a particle to a maximum scattering vector ($d_{\text{max}}q_{\text{max}}$) is given by ($d_{\text{max}}q_{\text{max}})^{-1}$. For example, SAS data to a $q_{\text{max}}$ of 0.3 Å$^{-1}$ determines for xylanase ($d_{\text{max}}$ 44 Å) or for 30S ribosomal particle ($d_{\text{max}}$ 240 Å) that the minimum number of observations is 4 and 23, respectively. This represents a ~20- to 125-fold over-sampling of a SAS curve composed of 500 observations. The Nyquist–Shannon limit ($n_s$) is the set of maximally independent observations from the band-limited SAS curve (Supplementary Fig. 7). We reasoned that calculating $\chi^2$ from a data set reduced to $n_s$ should more accurately assess the model-data agreement by restricting $\chi^2$ evaluations to the set of independent random variables (Supplementary Notes).

Owing to over-sampling and the uncertainties in $q$, $R(q)$ and $d_{\text{max}}$ determining the exact set of Nyquist–Shannon points will be difficult. Nevertheless, application of the noisy-channel coding theorem guarantees noise-free recovery of the SAS signal (Supplementary Notes and Supplementary Figs 8, 9); therefore, we propose the following

Cross-validating SAXS model-data agreements

Atomistic modelling of SAS data relies on the reduced chi-square ($\chi^2$) error-weighted scoring function$^{21,22}$ that can be unreliable with moderately noisy data sets or over-estimated degrees of freedom (Supplementary Figs 4 and 5). This can lead to over-fitting and model misidentification. In crystallographic and NMR analyses, cross-validation statistical methods mitigate over-fitting and increase confidence in selected model(s)$^{23,24}$. Here, we present an analogous robust statistical method based on the Nyquist–Shannon sampling and the noisy-channel coding theorems (Supplementary Notes) for evaluating structural models against SAS data.

For a given maximum dimension ($d_{\text{max}}$), the sampling theorem$^9$ determines that the number of unique, evenly distributed observations, $n_d$, required to represent a particle to a maximum scattering vector ($d_{\text{max}}q_{\text{max}}$) is given by ($d_{\text{max}}q_{\text{max}})^{-1}$. For example, SAS data to a $q_{\text{max}}$ of 0.3 Å$^{-1}$ determines for xylanase ($d_{\text{max}}$ 44 Å) or for 30S ribosomal particle ($d_{\text{max}}$ 240 Å) that the minimum number of observations is 4 and 23, respectively. This represents a ~20- to 125-fold over-sampling of a SAS curve composed of 500 observations. The Nyquist–Shannon limit ($n_s$) is the set of maximally independent observations from the band-limited SAS curve (Supplementary Fig. 7). We reasoned that calculating $\chi^2$ from a data set reduced to $n_s$ should more accurately assess the model-data agreement by restricting $\chi^2$ evaluations to the set of independent random variables (Supplementary Notes).

Owing to over-sampling and the uncertainties in $q$, $R(q)$ and $d_{\text{max}}$ determining the exact set of Nyquist–Shannon points will be difficult. Nevertheless, application of the noisy-channel coding theorem guarantees noise-free recovery of the SAS signal (Supplementary Notes and Supplementary Figs 8, 9); therefore, we propose the following
Validating model-data resolution limits

Determining resolution limits of model-data agreements cannot be achieved by $\chi^2$ alone and requires a metric we define as $R_{\text{SAS}}$, incorporating residuals between modelled and experimental values for both $R_g$ and $V_c$ given by

$$R_{\text{SAS}} = \frac{(R_g^2 - R_{g\text{model}}^2)^2}{(R_g^2)^2} + \frac{(V_c^2 - V_{c\text{model}}^2)^2}{(V_c^2)^2}$$

$R_{\text{SAS}}$ is a difference distance metric derived from the set of Q-independent SAS invariants. Calculation of $R_{\text{SAS}}$ at varying resolutions provides an objective basis to determine appropriate resolution limits for data-model agreements. For dilute xylanase (Supplementary Fig. 4a, b), data were collected to a maximum $q = 0.5 \text{ Å}^{-1}$ ($\sim 13 \text{ Å}$ resolution) and fit to PDB 1REF with a $\chi^2$ of 1.3, suggesting an acceptable data-model agreement. However, inspection of $R_{\text{SAS}}$ and $\chi^2_{\text{free}}$ (20.3 and 1.8, respectively) reveal low agreement. Truncating the SAS data shows a significant decrease in $R_{\text{SAS}}$ with $\chi^2_{\text{free}}$ increasing initially then decreasing as the data-model agreement improves (Supplementary Fig. 4b). Convergence of $R_{\text{SAS}}$ towards zero with a $\chi^2_{\text{free}} \leq 1.5$ implies the limit of the data-model agreement to be $q \approx 0.2 \text{ Å}^{-1}$ or a resolution of 31 Å. The combination of $R_{\text{SAS}}$ and $\chi^2_{\text{free}}$ for a given model provides a quantitative and graphical approach for determining the acceptable resolution between the data and model (Supplementary Figs 4b and 5). As SAS data are often used to filter a large set of conformationally distinct models, the models themselves may not be capable of describing the SAS data to high resolution; therefore, application of $R_{\text{SAS}}$ and $\chi^2_{\text{free}}$ may provide the useful resolution of the data-model agreement. Nevertheless, as done recently for crystallography\(^{27}\), a functional definition of resolution can come from the noisy-channel coding theorem. Here, the useful resolution of the data will be assessed by the highest Nyquist–Shannon point supported by the data.

Perspective

The SAS invariant $V_c$ extends analysis to flexible biopolymers in solution. The volume-per-correlation length, like $R_g$, faithfully informs on the conformational state of the particle and can be calculated for models determined by other structural techniques including electron microscopy, X-ray crystallography, NMR and SANS. $V_c$ provides a unique descriptor of the scattering experiment that is broadly applicable. We expect that $V_c$ may further characterize voids in materials such as bone, polymeric beads or nanomaterials. As the ratio of the square of $V_c$ to $R_g$ defines a mass parameter, $Q_0$, SAS experiments can now inform on particle mass without requiring compactness and instrument calibration. Furthermore, $\chi^2_{\text{free}}$ is a robust statistical metric that we envision will enable cross-validated determination of flexible ensembles against observed SASX data. We anticipate that $Q_0$, $R_g$, $\chi^2_{\text{free}}$ and $R_{\text{SAS}}$ will efficiently and objectively aid characterization of flexible macromolecules, check sample quality, determine mass and assembly states, detect concentration-dependent scattering, reduce model misidentification and over-fitting, and assess resolution for model-to-data agreement.

METHODS SUMMARY

SAXS data were simulated with FoXS\(^{22}\) and CRYSOL\(^{21}\). For each SASX data set, linear fits to the Guinier region were calculated to determine $R_g$ and $l(0)$. The Guinier parameters were used to calculate an extrapolated scattering data set to zero angle.

On the basis of an extrapolated data set, $V_c$ was calculated by dividing the Guinier $l(0)$ by the integrated area of $qI(q)$ versus $q$ calculated using the trapezoid rule. For simulated atomic SAXS profiles, extrapolation was unnecessary. Simulated atomic SAXS profiles by FoXS calculates scattering profiles at specified scattering-angle increments consistent with experimental measurements, whereas CRYSOL (without an input SAXS data set) can only calculate a maximum of 256 scattering intensities at a specified maximum scattering angle. At beamline 12.3.1 (Advanced Light Source Berkeley), typical data sets collected to a maximum $q$ of 0.2 Å\(^{-1}\) produce $\sim 500$ data points with the beamstop centred in the middle of the detector. Visual comparison of atomic SAXS profiles from FoXS with CRYSOL did not indicate any systematic differences.

Full Methods and any associated references are available in the online version of this paper.

Received 11 April 2011; accepted 13 March 2013.

1. Harrison, S. C. Comments on the NIGMS PSI. Structure 15, 1344–1346 (2007).
2. Hura, G. L. et al. Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS), Nature Methods 6, 606–612 (2009).
3. Rambo, R. P. & Tainer, J. A. Bridging the solution divide: comprehensive structural analyses of dynamic RNA, DNA, and protein assemblies by small-angle X-ray scattering. Curr. Opin. Struct. Biol. 20, 128–137 (2010).
4. Sosnick, T. R. & Woodson, S. A. New era of molecular structure and dynamics from solution scattering experiments. Biopolymers 95, 503–504 (2011).
5. Glatter, O. & Kratky, O. Small Angle X-Ray Scattering (Academic, 1982).
6. Putnam, C. D., Hammel, M., Hura, G. L. & Tainer, J. A. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures and conformations and assemblies in solution. Q. Rev. Biophys. 40, 191–285 (2007).
7. Jacques, D. A. & Trewella, J. Small-angle scattering for structural biology—expanding the frontier while avoiding the pitfalls. Protein Sci. 19, 642–657 (2010).
8. Bai, Y., Das, R., Miliett, S. I., Herschlag, D. & Doniach, S. Probing counterion modulation repulsion and attraction between nucleic acid duplexes in solution. Proc. Natl Acad. Sci. USA 102, 1035–1040 (2005).
9. Moore, P. Small-angle scattering. Information content and error analysis. J. Appl. Crystallogr. 13, 168–175 (1980).
10. Rambo, R. P. & Tainer, J. A. Improving small-angle X-ray scattering data for structural analyses of the RNA world. RNA 16, 638–646 (2010).
11. Nishimura, N. et al. Structural mechanism of asbsic acid binding and signaling by dimeric PY1. Science 326, 1373–1379 (2009).
12. Santiago, J. et al. Modulation of drought resistance by the abscisic acid receptor PYLS through inhibition of clade A PP2Cs. Plant J. 60, 575–588 (2009).
13. Stoddard, C. D. et al. Free state conformational sampling of the SAM-I riboswitch aptamer domain. Structure 18, 787–797 (2010).
14. Hammond, J. A., Rambo, R. P. & Kieft, J. S. Multi-domain packing in the Adramia sp. cox1b1–coxa1b2 domain of the cox1b1–coxa1b2 apicomplexan PfEMP1. Nat. Struct. Mol. Biol. 399, 450–463 (2010).
15. Orthaber, D., Bergmann, A. & Glatter, O. SAXS experiments on absolute scale with Kratky systems using water as a secondary standard. J. Appl. Crystallogr. 33, 218–225 (2000).
16. Mylonas, E. & Svergun, D. I. Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering. J. Appl. Crystallogr. 40, s245–s249 (2007).
17. Fischer, H., de Oliveira Neto, M., Napolitano, H. B., Polikarpov, I. & Craievich, A. F. Determination of the molecular weight of proteins in solution from a single small-angle X-ray scattering measurement on a relative scale. J. Appl. Crystallogr. 43, 101–109 (2010).
18. Rambo, R. P. & Tainer, J. A. Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. Biopolymers 95, 559–571 (2011).
19. Berman, H. M. et al. The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000).
20. Wyatt, P. J. Light scattering and the absolute characterization of macromolecules. Anal. Chim. Acta 272, 1–40 (1993).
21. Svergun, D., Barberato, C. & Koch, M. H. J. CRYSOI – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J. Appl. Crystallogr. 28, 768–773 (1995).
22. Schneidman-Duhovny, D., Hammel, M. & Sali, A. FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic Acids Res. 38, W540–W544 (2010).
23. Brünger, A. T. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature 355, 472–475 (1992).
24. Brünger, A. T., Clore, G. M., Gronenborn, A. M., Saffrich, R. & Nilges, M. Assessing the quality of solution nuclear magnetic resonance structures by complete cross-validation. Science 261, 328–331 (1993).
25. Rousseau, P. J. & Leroy, A. M. Robust Regression and Outlier Detection (Wiley, 1987).
26. Jie, Y., Qi, T., Amores, J. & Sebe, N. Toward Robust Distance Metric Analysis for Similarity Estimation (IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006).
27. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. Science 336, 1030–1033 (2012).
28. Brunger, A. T. et al. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr. D 54, 905–921 (1998).
29. de Groot, B. L. et al. Prediction of protein conformational freedom from distance constraints. Proteins 29, 240–251 (1997).
30. Rambo, R. P. & Doudna, J. A. Assembly of an active group II intron–maturase complex by protein dimerization. Biochemistry 43, 6486–6497 (2004).
31. Keel, A. Y., Jha, B. K. & Kieft, J. S. Structural architecture of an RNA that competitively inhibits RNase L. RNA 18, 88–89 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank G. L. Hura, M. Hammel, R. T. Batey, J. Tanamachi and the staff of SIBYLS Beamline 12.3.1 at the Advanced Light Source for discussions and P. Adams for suggestions regarding simulations with CNS. We thank E. Rambo, G. Williams and E. D. Getzoff for manuscript comments. This work is supported in part by funding to foster collaboration with Bruker and the Berkeley Laboratory Directed Research and Development (LDRD) program provided by the Director, Office of Science, US Department of Energy on Novel Technology for Structural Biology. The SIBYLS Beamline 12.3.1 facility and team at the Advanced Light Source is supported by United States Department of Energy program Integrated Diffraction Analysis Technologies (DEAC02-05CH11231) and by National Institutes of Health grant R01GM105404.

Author Contributions R.P.R. developed the theory and computational algorithms with input from J.A.T. Both J.A.T. and R.P.R. designed the experiments and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.P.R. (rprambo@lbl.gov) or J.A.T. (jat@scripps.edu).
METHODS

χ<sub>free</sub> calculation. For a given d<sub>free</sub>, the SAXS/SANS data collected between d<sub>max</sub> and d<sub>min</sub> can be divided into n<sub>s</sub> equal bins, in which n<sub>s</sub> is determined by the Nyquist–Shannon sampling theorem. Here, d<sub>max</sub> is measured from the atomistic model; however, d<sub>max</sub> can be directly inferred using an indirect Fourier transform method such as GNOM. In the case of 500 data points, and n<sub>s</sub> = 10, each bin will contain 50 data points such that a single randomly selected data point will represent that Nyquist–Shannon point. As a selected data point may be biased by inter-particle interference or uncertainties in q or I(q), the selection of the representative data point from the Nyquist–Shannon bin must occur through several selection rounds (k). During each round, the set of randomly selected points comprises the test set for calculating χ<sub>free</sub> against the model. The accepted value is taken as the median over k rounds. The number of rounds, k, will vary with the average noise level of the SAXS/SANS data set. The probability of selecting an erroneous data point from a bin scales directly with the noise. We have found that for high-quality data (<10% noise) k can be as small as a few hundred, whereas for high-noise data, k should be 2,000 to a maximum of 3,000.

Sample preparation. Protein and RNA samples were derived from a variety of sources. For glucose isomerase and xylanase, protein samples were obtained as suspended crystals (Hampton Research). Each protein was further purified by gel-filtration chromatography immediately before SAXS data collection in buffer consisting of 0.1 M Tris–HCl, 5 mM CaCl<sub>2</sub>, 100 mM KCl and 2 mM tris(2-carboxyethyl)phosphine (TCEP), buffer B (40 mM 2-ethansulolonic acid, MES, pH 6.8, 8 mM MgCl<sub>2</sub> and 100 mM KCl) or buffer C (40 mM Na-citrate, pH 5.0, 75 mM KCl and 1% glycerol). Proteins were re-suspended by a 50-fold dilution of the crystals in buffers A or B for glucose isomerase and buffers A, B or C for xylanase. Diluted crystals were incubated at 37 °C on a nutator for 1 h, concentrated to 10 mg ml<sup>-1</sup> and injected on a pre-equilibrated Superdex 200 PC 32 column (GE Healthcare) for glucose isomerase and Superdex 75 PC 32.3 column (GE Healthcare) for xylanase. Fractions corresponding to peak elution were taken for SAXS and quantitated by absorbance at 280 nm.

Polymerase was re-combinantly expressed and purified from Escherichia coli using cells transformed with a PET vector conferring ampicillin resistance. Cells were grown at 37 °C and induced for 4 h with isopropl-β-D-thiogalactoside at D<sub>PET</sub>=0.6 mm<sup>-1</sup> before collection. Cells were lysed as described<sup>16</sup>. Lysate was clarified by low-speed spin in 50-mL Falcon tubes and incubated at 65 °C for 20 min. Lysate was further clarified by high-speed centrifugation at 20,000g for 40 min at 4 °C. Bound nucleic acids were removed by polyethyleneimine treatment and ammonium sulphate precipitation. Protein was re-suspended in buffer B and further purified to homogeneity using Superdex 200 HR 10/30 (GE Healthcare) for SAXS analysis.

Conformational simulation. SAM-I riboswitch molecular dynamics simulations were performed with CNS as described<sup>19</sup>. In brief, the SAM crystal structure (PDB code, 2GIS) was analysed with FIRST and FRODA<sup>20</sup> at several energy cutoffs to determine possible rigid and flexible conformational space of the SAM-I riboswitch. A series of confidence-weighted averages were used to ascribe constraints within the structure for molecular dynamic simulations with CNS using anneal.ipc. The CNS input file was modified to remove the electrical potential from the energy function and calculations were performed as torsional angle dynamics only. For each simulation, 2,000 steps were recorded in the trajectory file and each step was written to file as a PDB.

Simulating noisy SAXS data sets. SAXS intensities over a single exposure will range over a few hundred orders of magnitude, and consequently the noise levels will vary throughout the measured q region. Therefore, we used intensity uncertainties from previously collected SAXS experiments as a source of realistic noise for the simulated SAXS data sets. The noise level of the empirical SAXS curve is reported as the average relative noise in the last third of the observed q-range (Fig. 4).

For a selected q<sub>0</sub>, the simulated I(q<sub>0</sub>) was randomly displaced based on a random draw using the Box–Muller transform of a standard Gaussian distribution parameterized by the empirical intensity, I(q<sub>0</sub>)<sub>obs</sub>, and uncertainty, error(q<sub>0</sub>)<sub>obs</sub>. The Box–Muller transform returns two possible values and a random binary selection was used to provide a final single value for the displacement of the simulated I(q<sub>0</sub>), I<sub>displaced</sub><sub>q</sub><sub>0</sub>. The simulated error(q<sub>0</sub>) was reported as I<sub>displaced</sub><sub>q</sub><sub>0</sub> * error(q<sub>0</sub>)<sub>obs</sub>.

32. Fulle, S. & Gohlke, H. Analyzing the flexibility of RNA structures by constraint counting. Biophys. J. 94, 4202–4219 (2008).