

Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS)

Greg L Hura^{1,6}, Angeli L Menon^{2,6}, Michal Hammel^{1,6}, Robert P Rambo¹, Farris L Poole II², Susan E Tsutakawa³, Francis E Jenney Jr^{2,4}, Scott Classen¹, Kenneth A Frankel¹, Robert C Hopkins², Sung-jae Yang², Joseph W Scott², Bret D Dillard², Michael W W Adams² & John A Tainer^{3,5}

We present an efficient pipeline enabling high-throughput analysis of protein structure in solution with small angle X-ray scattering (SAXS). Our SAXS pipeline combines automated sample handling of microliter volumes, temperature and anaerobic control, rapid data collection and data analysis, and couples structural analysis with automated archiving. We subjected 50 representative proteins, mostly from *Pyrococcus furiosus*, to this pipeline and found that 30 were multimeric structures in solution. SAXS analysis allowed us to distinguish aggregated and unfolded proteins, define global structural parameters and oligomeric states for most samples, identify shapes and similar structures for 25 unknown structures, and determine envelopes for 41 proteins. We believe that high-throughput SAXS is an enabling technology that may change the way that structural genomics research is done.

Visualizing macromolecular shapes and assemblies that principally determine function is a central challenge for structural molecular biology¹. Addressing this challenge requires the capacity to characterize the many complexes and conformations that underlie biological outcomes. Yet growing metagenomics, proteomics and bioinformatics contributions are outpacing classical structural biology approaches, creating an increasing structural knowledge gap^{2,3}.

X-ray diffraction and scattering are powerful methods for unraveling structural details and molecular shapes⁴. Macromolecular X-ray crystallography has been the cornerstone of the structural genomics initiatives⁵; both crystallography and NMR spectroscopy have provided a deep and broad survey of macromolecular structural properties at high resolution^{6–8}. Yet the stochastic nature of crystallization and the molecular size and time constraints of NMR limit the throughput of these technologies. The application of X-ray scattering in solution, known as small angle X-ray scattering (SAXS), to structural biology has lagged behind crystallography despite its strength in other fields⁹. However, SAXS use has sharply increased with advances in synchrotron X-ray sources and detectors that improve data quality and reduce

the amount of sample required. New algorithms have been developed that can identify accurate shapes and assemblies based on the scattering data^{4,10,11}. Notably, SAXS analyses can build on and be combined with other results to test experimental hypotheses and computational models⁴.

Though lower in spatial resolution than crystallography or NMR spectroscopy, SAXS offers fundamental advantages for high-throughput structural analyses: structural measurements are carried out in solution, sample preparation is simple, quality global parameters can be obtained for most samples, and SAXS is compatible with and complementary to other biophysical techniques. With samples containing monodispersed and homogeneous species, SAXS data can be used to define envelopes (three-dimensional surfaces that define the molecular shape of the macromolecule in solution) to better than 15 Å resolution. This resolution is often sufficient to address key biological questions, and several high-impact SAXS results have recently been described^{12–15}. Because sample preparation is minimal and data can be rapidly collected and analyzed, SAXS is potentially the highest-throughput structural determination and analysis technique. As most macromolecular structures are amenable to SAXS analysis, for example, the structural analysis of all complexes of a metabolic pathway can be considered. In the US alone, the National Institutes of Health will spend \$80 million this year on the Protein Structure Initiative¹⁶, which provides structures for 3–15% of its targets¹⁷, so a cost-effective and efficient means to improve the fraction of protein samples yielding structural information would be very valuable.

Here we report the development of an efficient pipeline enabling robust, broadly applicable and largely automated SAXS-based structural analyses. Though alternative collection approaches have been reported^{18,19}, we obtained high-quality data from small volumes (12 µl) and protein concentrations (~1 mg ml⁻¹), with temperature and anaerobic control for sample stability in a modular 96-well format. We subjected 50 proteins, mostly from *Pyrococcus furiosus*, to our pipeline. Our high-throughput SAXS

¹Physical Bioscience Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA. ²Department of Biochemistry and Molecular Biology, University of Georgia, Athens, Georgia, USA. ³Life Science Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA. ⁴Georgia Campus–Philadelphia College of Osteopathic Medicine, Suwanee, Georgia, USA. ⁵Department of Molecular Biology and The Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, California, USA. ⁶These authors contributed equally to this work. Correspondence should be addressed to J.A.T. (jat@scripps.edu) or M.W.W.A. (adams@bmb.uga.edu).

RECEIVED 5 JANUARY; ACCEPTED 9 JUNE; PUBLISHED ONLINE 20 JULY 2009; DOI:10.1038/NMETH.1353

pipeline provided global information (Table 1) for most samples as well as folding, assembly and three-dimensional envelope information on monodisperse samples. Such information can be used to judge the amenability of proteins for crystallographic studies and can even be used to infer protein function. Our results demonstrate how automated, high-throughput SAXS can provide a critical enabling technology for producing unique, comprehensive and complementary solution structural information.

RESULTS

High-throughput SAXS data collection platform

To achieve sufficient X-ray flux for informative scattering with low protein concentration and small volumes, we designed the SIBYLS beamline at the Advanced Light Source. We used a light path generated by a super-bend²⁰ magnet to provide a 10^{12} photons s^{-1} flux (1 Å wavelength). The tunable incident wavelength enables rapid adjustment of the momentum transfer (q) appropriate for the experiment without changing the sample-to-detector configuration ($q = 4\pi\sin(\theta/2)/\lambda$ where θ is the scattering angle and λ is the

wavelength). Scattering is measured on a MAR165 area detector (Rayonix) coaxial with the incident beam and 1.5 m from the sample allowing a q range from a minimum of 0.007 Å^{-1} to a maximum of 4.2 Å^{-1} (Fig. 1a).

To transfer 96-well plate samples to the SAXS sample cell, we implemented a Hamilton pipetting robot. Both sample cell and the 96-well plate were temperature-controlled with the sample plate sealed by a pierceable aluminum sheet. The robot needle transferred samples to the helium-filled sample holder (Fig. 1b), providing an anaerobic environment with low X-ray scattering cross-section, reducing background.

Protocol for high-throughput SAXS data analysis

For efficient analysis of data quality and information, we developed a SAXS analysis tree (Fig. 1c). We automated the program data flow with Perl scripts (Supplementary Software) for the ATSAS²¹ program suite similar to those recently reported²². We standardized the output for automated incorporation into our database and automated job scheduling on computer clusters. Data analysis

Table 1 | SAXS characterizations for 34 *P. furiosus* samples including structural information

Class	Identifier ^a	Ortholog ^b	R_G (Å)	D_{max} (Å)	Assemblies	Envelope
Aggregates	PF0418	ATPase			Separable	No
	PF1733	Conserved hypothetical ^c			Inseparable	No
	PF1951	Aspartate-ammonia ligase			Inseparable	No
Mixtures of oligomers of unknown structure	PF0230	ArsR transcription regulator			Mostly 2-mer	Yes
	PF0259	Conserved hypothetical	26.0	84	Mostly 8-mer	Yes
	PF0741	Thioredoxin-related	20	> 80	Mostly 1-mer	Yes
	PF1548	Conserved hypothetical			Rings	Yes
	PF1605	Molybdopterine synthase	21	> 90	Mostly 1-mer	Yes
Mixtures of oligomers of known structure	PF0094 ^d	Glutaredoxin-like	28	110	92% 2-mer/8% 4-mer	Yes
	PF0380 ^d	Conserved hypothetical	21	125	68% 2-mer/32% 1-mer	Yes
	PF0939	Isopropylmalate dehydratase	23.1	82	73% 2-mer/27% 1-mer	Yes
	PF1909 ^d	Ferredoxin	13.0	38	40% 2-mer/60% 1-mer	Yes
Matching PDB model	PF0863 ^d	Adenylyl cyclase CyaB	27.4	87	Matching 1-mer	Yes
	PF1061 ^e	Ferredoxin β -grasp fold	17.7	78	Matching 1-mer	Yes
	PF1281 ^d	Superoxide reductase	22.1	80	Matching 4-mer	Yes
	PF1282 ^e	Rubredoxin	11.0	29	Matching 1-mer	Yes
Crystal structure of homolog	PF0619	Conserved hypothetical	23.1	73	Matching 3-mer	Yes
	PF1026	Malic enzyme, NAD-binding	31.8	110	Matching 1-mer	Yes
	PF1033	Thioredoxin-like fold	51.2	150	Matching 10-mer	Yes
	PF1528	SNO glutamine amidotransferase	19.9	80	Matching 1-mer	Yes
	PF1674	Tyrosine/serine phosphatase	16.7	58	Matching 1-mer	Yes
	PF1787	Acetyl-CoA synthetase	33.9	98	New 3-mer	Yes
	PF0014/0015 ^f	Conserved hypothetical	55.0	165	> 8-mer	Yes
	PF0553	Tyrosine phosphatase	19.2	110	1-mer	Yes
	PF706.1	Zinc finger	18.6	80	1-mer	Unfolded
Proteins of unknown structure	PF0699	Conserved hypothetical	23.7	74	2-mer	Yes
	PF0715	NADH oxidase	23.1	96	2-mer	Yes
	PF0965/0966/0967/0971 ^g	Pyruvate ferredoxin oxidoreductase	36.9	120	244 kDa	Yes
	PF1282/1205 ^h	Nucleotide-binding protein	24.3	95	1-mer	Unfolded
	PF1291	Phosphoesterase	35.6	110	4-mer	Yes
	PF1372	Conserved hypothetical	23	75	4-mer	Yes
	PF1911	Ferredoxin NADP reductase	30.9	101	2-mer	Yes
	PF1950	Phosphoribosyl transferase	25.3	100	2-mer	Yes
	PF2047.1	Conserved hypothetical	29.7	155	1-mer	Unfolded

^a*P. furiosus* gene product as labeled by gene number³⁰. ^bProtein names (orthologs) were based on bioinformatics analyses (<http://www.ebi.ac.uk/interpro/>). ^cConserved hypothetical proteins, which have unknown function. ^dProteins that have been crystallized. ^eProteins with models determined by NMR spectroscopy. ^fPF0014 and PF0015 were tandemly expressed in *E. coli* and form a complex. ^gPF0965, PF0966, PF0967 and PF0971 form pyruvate ferredoxin oxidoreductase and were purified from native biomass. ^hThe PF1282-PF1205 recombinant fusion protein has the rubredoxin (PF1282) of *P. furiosus* as the 'tag' to an unrelated putative nucleotide binding protein (PF1205). Although the PF1282 portion was folded as evident by its red color (iron), PF1205 was unfolded.

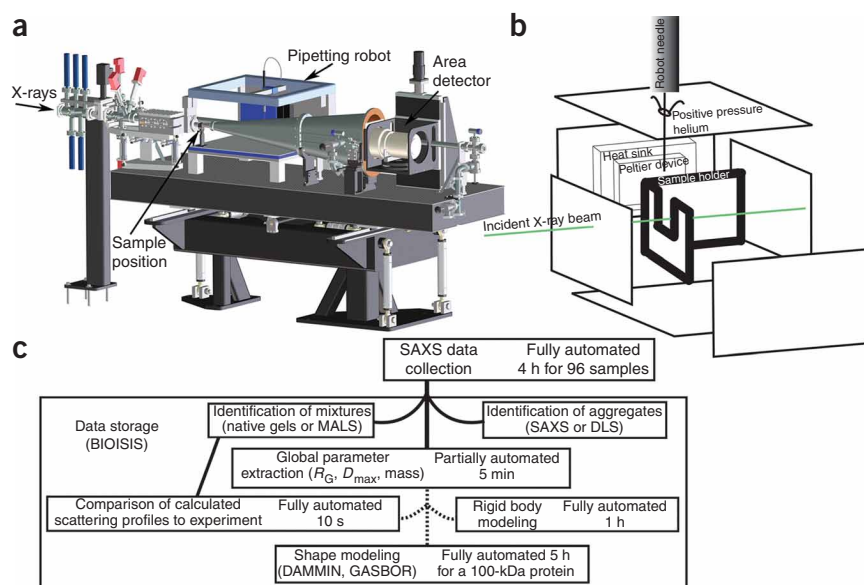


Figure 1 | High-throughput SAXS pipeline.

(a) Configuration of the SAXS endstation.

(b) Schematic of the sample area showing how the sample is loaded by the robot into a temperature-controlled cell. Positive helium pressure reduces air scatter and oxidative damage. (c) SAXS data are collected at the SIBYLS beamline. Subsequent analysis depends on whether the macromolecules are categorized as aggregated (using either the SAXS curve itself or dynamic light scattering (DLS) analysis), as mixtures (based on native gel electrophoresis (native gel) or multiangle light scattering (MALS)) analyses, or monodisperse samples. For monodisperse samples, SAXS data next define global solution structural parameters radius of gyration, maximum dimension and calculated mass. Sequence-based homology search discovers existing structures that can be used to analyze both mixtures and monodisperse samples. Approximate time scales are noted in each step. Perl scripts are used to collect information and begin processes for dashed paths. Both primary data and derived shapes are stored at the BIOISIS web-accessible utility.

begins by defining global sample parameters and comparing experimental and calculated scattering curves where prior structural information exists. To test the scattering information, we used two different molecular envelope determination programs, DAMMIN²³ and GASBOR¹⁰. DAMMIN searches compact configurations of adjustable diameter spheres to create a shape which best matches the scattering profile. The sphere size depends on the maximum dimension of the molecule, and thus computation time is independent of molecule size. GASBOR uses spheres adjusted in size and scattering power to match an amino acid, and a penalty is enforced to promote connectivity. Both algorithms use a simulated annealing search algorithm.

Ten independent DAMMIN runs are spawned by default once data enter the system. Mass is estimated using half the Porod volume⁹ calculated from $q < 0.25 \text{ \AA}^{-1}$. For most samples, we found this estimate sufficient to identify oligomeric state. When ambiguous, we estimated the mass by the extrapolated intensity at zero scattering angle^{9,24}. The time required to traverse the analysis tree was size-dependent: 40 min for a 20 kDa protein to 1.5 d for a 500 kDa complex run in parallel with other proteins. With current computational resources, our throughput exceeds 20 proteins per week for a full analysis; > 1,000 macromolecules could be analyzed per year.

Automated data storage and quality control

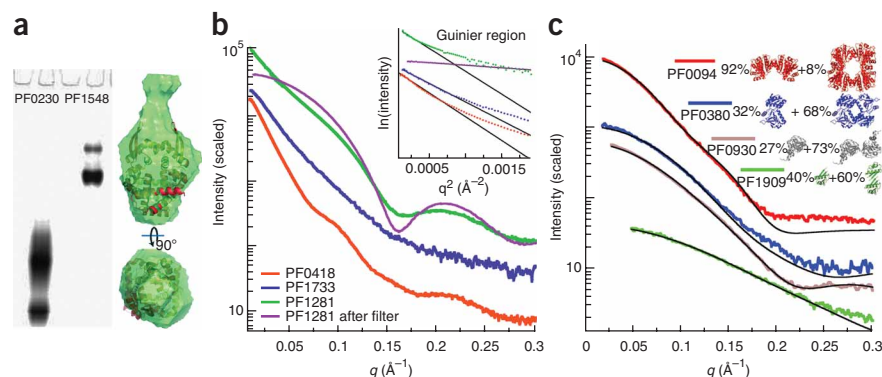
To aid communication of our results as well as to promote objective quality assessment, test newly available atomic resolution models and to aid in the development of new SAXS algorithms, we created the web-accessible database biologically integrated structures in solution (BIOISIS; <http://www.bioisis.net/>). A powerful aspect of SAXS data collection is the ability to characterize macromolecules in many solution conditions. In the BIOISIS database all experimental details are saved and associated with each sample. Database functionality was enhanced for proteins obtained from *P. furiosus*, *Sulfolobus solfataricus* and *Halobacterium salinarum*, including gene annotations and a search engine for gene number or a keyword in the annotation.

Testing prototypical sample sets

To test whether SAXS can provide proteomic-scale information, we analyzed protein targets from two sources: 34 recombinantly expressed *P. furiosus* samples with a 9-amino-acid His tag (Table 1) plus 16 Joint Center for Structural Genomics (JCSG) targets with 19-amino-acid His tags (Supplementary Table 1). We focus here on the results from *P. furiosus* samples in which 29 of the 34 proteins had not crystallized despite systematic efforts. These are

Figure 2 | SAXS analysis provides feedback on 'challenging' samples that are polydisperse or inhomogeneous.

(a) Native gel electrophoresis analysis of samples PF0230 and PF1548 (left). Overlaying the SAXS-predicted PF0230 envelope with a close homolog (PDB 2CWE) revealed consistency to the homolog dimer with additional density indicating a larger species (right). (b) SAXS results directly discerned aggregation based on low-angle Guinier regions (insert) for the indicated protein samples. Features (oscillations) in the SAXS scattering curve for PF0418 and PF1281 suggest that small adjustments in sample preparation may yield workable data, for example, PF1281 data were markedly improved after passing through a filter. (c) Probable multimers may be identified when atomic resolution results are available of the protein or a homolog. Here multimers in crystal lattices (PF0094 homolog PDB 1J08, PF0380 PDB 1VK1, PF0930 homolog PDB 1V7L and PF1090 PDB 1S31) are used to identify a best fit to the SAXS data.



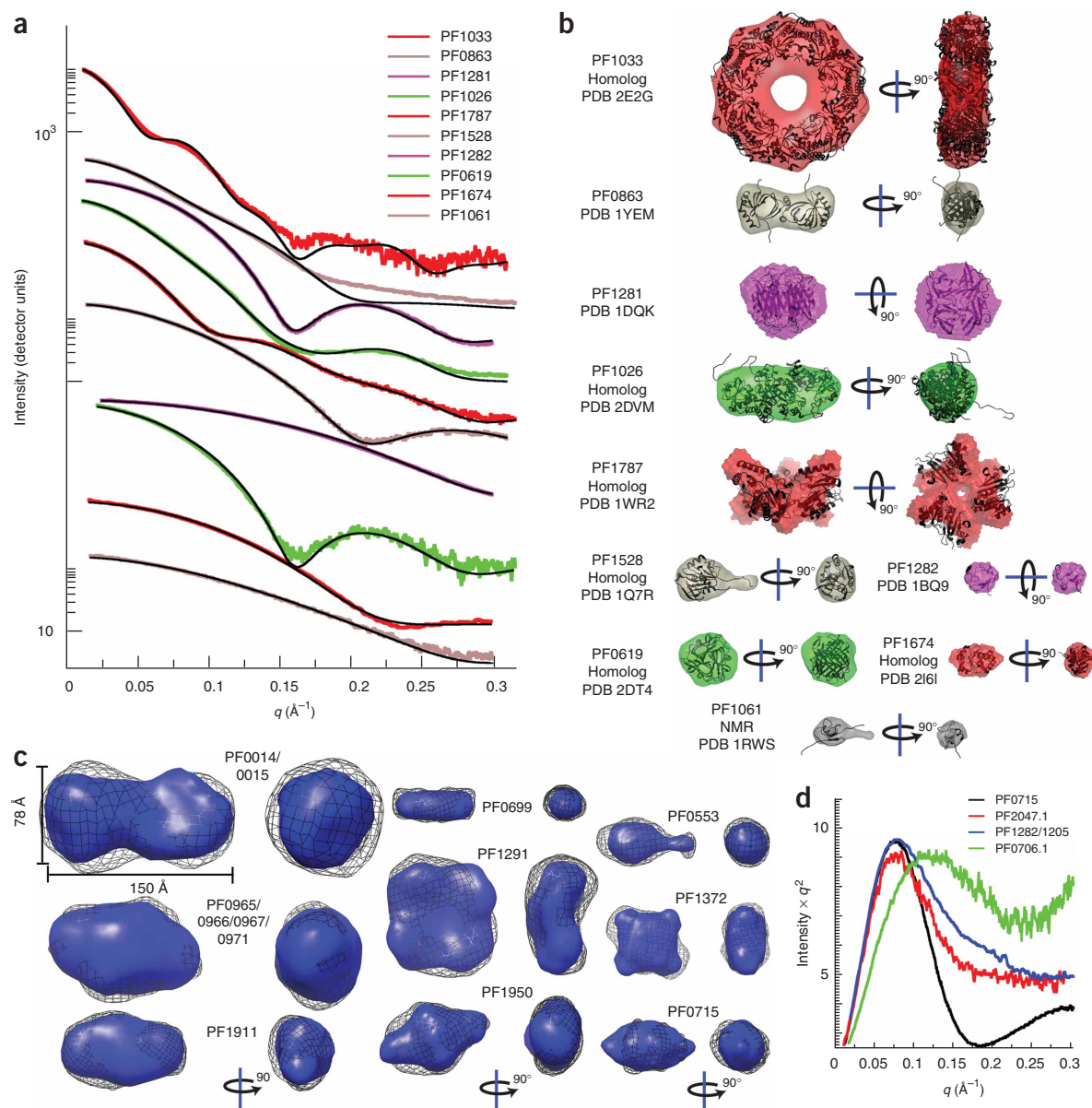


Figure 3 | SAXS provides accurate shape and assembly in solution for most samples. **(a)** For the ten proteins with structural homologs or existing structures, the experimental scattering data (colors) were compared with the scattering curve calculated for the matching structure (black). **(b)** For monodisperse samples, the envelope determinations (colored as in **a**) were overlaid with the existing structures (ribbons). All monomeric units had a 9-amino-acid His tag attached. **(c)** For the 9 proteins with no available structural information, envelope predictions from two independent programs were compared and generally agreed. The DAMMIN results (black mesh) were generated without symmetry. The GASBOR results used twofold symmetry for the PF0014-PF0015 protein complex, PF0965-PF0966-PF0967-pF0971 protein complex, PF1911 (dimer), PF00716 (dimer), PF0699 (dimer) and PF1950 (dimer). Fourfold symmetry was imposed on tetrameric PF1291 and PF1372. **(d)** Plotting the SAXS data as $I \times q^2$ versus q (Kratky plot) highlights proteins with large unfolded regions. The Kratky plot of PF0715 is shown for comparison of a folded protein and shows characteristic parabolic behavior at wide angles. In contrast PF0706.1, PF2047.1 and PF1282-PF1205 fusion protein have SAXS data consistent with unfolded regions as reflected in the nonparabolic wide-angle properties.

referred to by the 'PF number' of the gene that encodes them and are prototypic of gene products providing sequences for current structural genomics efforts.

To aid analysis, we divided samples into three general classes (**Table 1**): nonideal proteins (light scattering or other techniques suggest aggregated or mixed assembly states), proteins with existing structural information (either directly or from a sequence homolog) and proteins with unknown structures. We first characterized the samples by nondenaturing gel electrophoresis and

light scattering. Nonideal samples exhibited mixtures of states or aggregation, which restricts SAXS analyses (**Fig. 2**). Proteins with existing structural information (for themselves or sequence homologs) allowed higher-resolution analyses. Proteins of unknown structure were monodispersed with no or incomplete structural homology. For these new protein structures, SAXS not only provided shape and assembly information but also identified similar known structures based on direct comparisons of experimental scattering with that calculated from known structures.

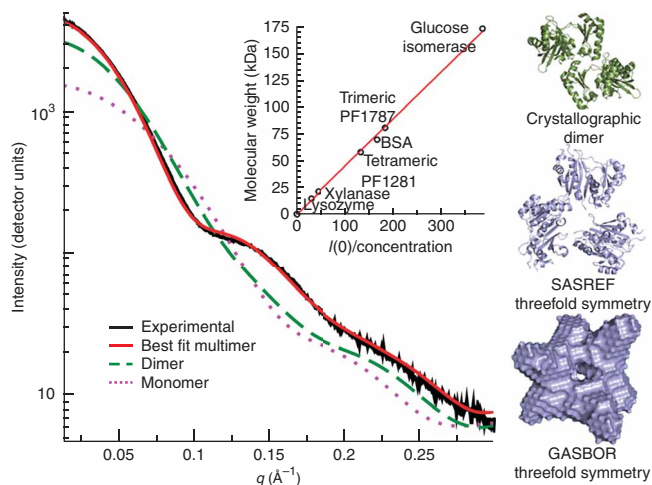


Figure 4 | SAXS determines accurate assembly state in solution, as shown for acetyl-CoA synthetase subunit (PF1787). The experimental scattering curve for PF1787 is shown with calculated scattering curves for monomeric and dimeric atomic resolution structures of homologs. The best fit (best fit multimer) to the experimental SAXS data are calculated from a threefold symmetric trimer derived from a monomeric homolog (PDB 1WR2). The trimeric form of PF1787 was confirmed using the extrapolated intensity at 0 scattering angle ($I(0)$), normalized for concentration (inset). The indicated protein standards were used to place the data on a relative scale. Relevant structures from analysis of PF1787 are shown on the right. The crystallographic dimer is a flexibly linked two-domain protein. Models with threefold symmetry enforced match the SAXS results.

Nonideal protein samples can guide sample improvements

Native gel electrophoresis results showed that PF0230 and PF1548 were mixed oligomeric species (Fig. 2a). After purification, 'mixtures' are the result of proteins that form various multimers in dynamic equilibrium with one another. Results for all SAXS-derived parameters on mixtures are electron number-weighted and population-weighted averages of parameters determined from each component individually. Algorithms for envelope determination assume homogeneous solutions, so interpretations must take any mixed state into account. The PF0230 envelope (Fig. 2a), for example, is overlaid on the proposed biological unit from a homolog crystal structure. The lower portion fits the dimer; yet the extension is probably an average of dimers mixed with larger oligomers. For PF1548, gel filtration analysis and forward scattering indicate large multimeric assemblies. Reconstructed envelopes suggest rings with a propensity to stack.

Aggregate samples produce scattering curves dominated at the smallest angles by the largest particles, which can confound subsequent analysis. However, SAXS probes structural details at and below 15 Å, so such samples may generate useful information if interpreted cautiously as SAXS is additive. We identified three proteins, PF0418, PF1281 and PF1733 as aggregated based on SAXS profiles with nonlinearity on a Guinier plot (natural log of intensity versus q^2) for $q \times R_G < 1.4$. (Fig. 2b). Given our measured q range, this metric identifies particles with $R_G > 75$ Å and a longest dimension across the molecule (D_{\max}) of at least 340 Å (larger than a ribosome). Scattering curve oscillations beyond $q > 0.1$ Å⁻¹ with

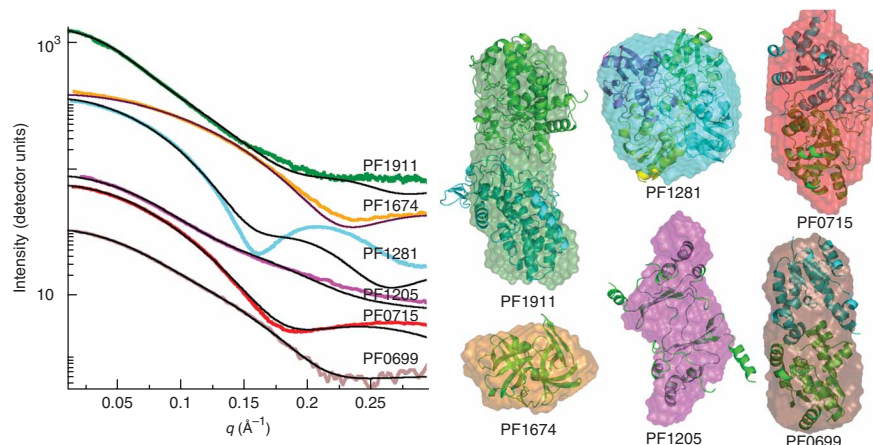
such a large R_G indicate ordered and population-wide correlations on a much smaller length scale (for example, data from PF0418 and PF1281). The absence of such oscillations (for example, PF1733) in aggregated samples indicates the lack of population-wide correlations on small length scales and typifies aggregates composed of irreversibly misfolded protein.

We also examined proteins provided by the JCSG crystallographic pipelines. These samples were concentrated to on average 23 mg ml⁻¹. High concentrations are common starting points for crystallization trials. However, throughout this study we found that proteins concentrated to greater than 5 mg ml⁻¹ after purification often contained artificial multimerization states or aggregates. High concentration may increase the chances of crystal nucleation but also increase heterogeneity, adversely affecting SAXS and other analyses. Aggregated samples whose scattering shows oscillations are often salvageable by removing aggregates. For example, passing samples through a 100 kDa filter yielded scattering characteristic of a monodisperse solution for PF1281 (Fig. 2b). Given the scattering features observable for PF0418, interpretable SAXS results would likely be obtained with additional sample preparation, such as filtration. We obtained data for six JCSG samples in this manner (Supplementary Table 1).

Homologous structures improve resolution

To take full advantage of scattering information, it is important to identify and use additional information when available⁴. An initial step in our analysis tree was the application of sequence analysis to identify any known detailed structures for samples. Atomic models were available for seven *P. furiosus* samples. Seven others had sequence homology to proteins of comparable size with an existing structure in the Protein Data Bank (PDB) (Table 1).

Figure 5 | SAXS defines accurate shape and assembly in solution for unknown structures and can uncover unsuspected structural similarity. Experimental scattering curves for proteins with no known structural homolog (left, color) were compared with calculated scattering (black) from PDB structures identified by DARA²⁶, a database of scattering curves calculated from the PDB database. Results from the shape reconstruction program GASBOR (colored envelopes) are overlaid onto the structures identified by DARA (ribbon models, right). In addition, PF1674 and PF1281 with known structures show a limitation in the DARA search and the need for better comparative algorithms.



Existing detailed structures allowed comparisons of measured scattering curves with those calculated from atomic models.

In four cases (**Table 1** and **Fig. 2c**), nondenaturing gel electrophoresis analysis showed multimeric forms and the data could be fit as a mixture of assemblies found in crystallographic lattices. SAXS data can identify relevant multimers even when mixtures are present. For example PF0094 fit multimers found in a homolog better than those found in its own determined crystal lattice.

SAXS analyses provided shape and assembly in solution for most samples including nine proteins with no preexisting structural information and three proteins with unfolded regions (**Fig. 3**). Six samples had SAXS curves that matched those calculated from single multimeric states suggested by PDB structures (**Fig. 3a,b**). We identified PF1281 as aggregated based on the SAXS results, but we obtained a homogenous solution after spin column filtration just before data collection (**Fig. 2b**).

PF1674 matched the scattering profile calculated from the monomeric state of a distant homolog (**Table 1**). In contrast, PF1787 did not fit the monomer scattering profile, nor any multimer in the crystal structure of a homolog with 55% sequence identity (PDB 1WR2). We applied rigid body modeling of three subunits and found a best fit to the experimental data, supporting the reconstructed envelope with threefold symmetry (**Fig. 3b** and **4**).

Visualizing novel assemblies and envelopes

Nine of the *P. furiosus* proteins (and three of the JCSG targets) that we analyzed by SAXS were new with no known atomic models of sequence homologs of similar length. We assumed monodispersity based on our observation of single bands by native gel electrophoresis analysis. We determined shape and assembly from scattering curves (**Fig. 3d**). To test envelope consistency, we compared models from both DAMMIN without enforced symmetry and GASBOR with appropriate symmetry. The shapes generated by these independent approaches are consistent with one another. GASBOR yielded contour shapes with greater detail. For three proteins, a Kratky plot²⁵ indicated considerable unfolded regions (**Fig. 3d**). Envelopes generated for PF2047.1 and the PF1205-PF1282 fusion reveal a compact region. However, conformationally heterogeneous samples yielded envelopes representing the average shape.

DISCUSSION

Macromolecular information is encoded in shape and assembly, so methods that bridge the growing gap between structural information and highly productive genomic and proteomic advances are needed. Structural genomics efforts have greatly increased the throughput of protein structure analysis (<http://sg.pdb.org/>), but even with the best efforts, up to ~85% to ~97% of samples cannot be easily characterized by crystallography¹⁷. In contrast, our SAXS pipeline yielded solution structural information for 31 of 34 *P. furiosus* samples and 10 of 16 JCSG targets, a success rate of 82%, whereas crystallography efforts only characterized 7 of 34 *P. furiosus* targets (21%), typical of structural genomics efforts. Furthermore, SAXS provides superior accuracy for solution conformation and assembly, compared to complementary, higher-resolution methods such as crystallography and NMR spectroscopy.

SAXS data can have direct implications for determining biological functions as well as for guiding crystallization and biochemical

characterizations. For example, we created and purified a soluble construct of PF1205 fused to rubredoxin (PF1282) to aid PF1205 purification, but as indicated by SAXS analyses the fusion protein (PF1282/1205) was unstructured and would be unlikely to crystallize. The SAXS profile alone can provide insight for comparative studies, and datasets can be generated at a rate of 96 samples in 4 h. For crystallographic purposes, the observation of reversible aggregation as function of concentration is a metric for identifying likely crystallization targets. Comparison of SAXS data to those calculated from known structures may guide molecular replacement efforts and identify new folds (**Fig. 5**). The scattering curve we determine for PF0699 matched remarkably well to a scattering profile calculated from a solved structure in the PDB (see the database for rapid search of structural neighbors based upon their SAXS patterns (DARA)²⁶). PF0699 is a conserved hypothetical protein that we matched to *Escherichia coli* shikimate kinase I (PDB 1KAG²⁷), which acts in the chorismate biosynthesis pathway. *P. furiosus* has this pathway involving a known shikimate kinase (PF1694), so the presence of an analogous protein (PF0699) is intriguing. We also identified promising functional leads for PF0715 and PF1911. We expect to see improvements in identifying structural homologs using calculated profiles from existing structures. For example, the solved crystal structure of superoxide reductase PF1281 (1DQE) is DARA's second ranked tetramer with a higher score given to PDB 1JTK. Yet comparison of the scattering curves over a wider q range immediately highlighted the superior fit of the correct structure (**Figs. 3a** and **5**). Similarly, for conserved hypothetical protein PF1674, its homolog was the 25th ranked structure, whereas structures with poorer overall fit to the data were ranked higher. This is likely the result of overweighting low-resolution features. An additional limitation is the small number of solved crystal structures for very small and very large proteins.

Symmetry provides powerful constraints on SAXS reconstructions, so our observation that 60% of samples formed multimers bodes well for accurate reconstructions. Our SAXS results indicated a trimeric assembly for PF1787 (**Fig. 4**): a flexibly linked, two-domain protein, which is one of two acetyl CoA synthetase (ACS) subunits. ACS generates ATP, CoASH and acetate, and was purified from *P. furiosus* biomass as a heteromeric complex (PF1781 and PF1540) with an $\alpha_2\beta_2$ stoichiometry²⁸. How the trimeric solution structure of PF1787 acts in this ACS reaction can now be experimentally investigated.

The JCSG protein set allowed testing of proteins containing a 19-residue His tag. His tags (on average representing 8% of the scattering of the JCSG proteins) increase D_{\max} , and add considerable shape heterogeneity, resulting in lower resolution. The disordered His tags are also asymmetric, making symmetry in envelope calculations less valid although the core is symmetric. Yet tags may be modeled if core atomic models are available (**Supplementary Fig. 1**).

A serious stated challenge to current structural genomics efforts is the absence of a clear path for a more comprehensive characterization of proteins, including their biologically relevant complexes and conformations²⁹. Our high-throughput SAXS pipeline can be used to examine complexes and conformations in solution, can rapidly evaluate many physiological conditions and ligand interactions, can characterize proteins with unstructured regions, and can identify structural similarities without requiring sequence

homology. In general, SAXS can provide solution structural information at resolutions often sufficient for functional insights into how these proteins work in the context of their pathways and networks. Whereas crystallography provides precision of high-resolution structures, it does not guarantee accuracy of conformational and assembly state under physiological conditions as well as SAXS does. We anticipate that high-throughput SAXS may therefore help address bottlenecks in current structural genomics efforts and aid fundamental research in proteomics and systems biology.

METHODS

Methods and any associated references are available in the online version of the paper at <http://www.nature.com/naturemethods/>.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This research is part of the Molecular Assemblies: Genes and Genomes Integrated Efficiently (MAGGIE) project supported by the US Department of Energy (DOE; DE-FG0207ER64326) and benefited from allocation of supercomputer time at the National Energy Research Scientific Computing Center (NERSC). Support for advancement of SAXS technologies at the Lawrence Berkeley National Laboratory SIBYLS beamline of the Advanced Light Source came from the DOE program Integrated Diffraction Analysis Technologies (IDAT) under contract DE-AC02-05CH11231 with the DOE. We thank I. Wilson and M. Knuth (the Scripps Research Institute) for providing protein samples from their Joint Center for Structural Genomics (JCSG).

AUTHOR CONTRIBUTIONS

G.L.H., J.A.T., M.H., S.C. and K.A.F. designed the SIBYLS beamline for high throughput. G.L.H., J.A.T. and M.W.W.A. wrote the manuscript. A.L.M., F.L.P., F.E.J., S.E.T., R.P.R., R.C.H. and G.L.H. prepared samples for data collection. G.L.H. and S.E.T. collected SAXS data. M.H. and G.L.H. wrote code for analysis. R.P.R. designed <http://www.bioisis.net/>. S.-j.Y. prepared PF0380 and PF2047.1. B.D.D. prepared PF0014/0015. J.W.S. prepared PF1787.

Published online at <http://www.nature.com/naturemethods/>.

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>.

- Robinson, C.V., Sali, A. & Baumeister, W. The molecular sociology of the cell. *Nature* **450**, 973–982 (2007).
- Green, B.D. & Keller, M. Capturing the uncultivated majority. *Curr. Opin. Biotechnol.* **17**, 236–240 (2006).
- Wilmes, P. & Bond, P.L. Metaproteomics: studying functional gene expression in microbial ecosystems. *Trends Microbiol.* **14**, 92–97 (2006).
- 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, conformations and assemblies in solution. *Q. Rev. Biophys.* **40**, 191–285 (2007).
- Fox, B.G. *et al.* Structural genomics: from genes to structures with valuable materials and many questions in between. *Nat. Methods* **5**, 129–132 (2008).
- Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
- Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
- Sigrist, C.J. *et al.* PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* **3**, 265–274 (2002).
- Glatter, O. & Kratky, O. *Small Angle X-ray Scattering* (Academic Press, London, 1982).
- Svergun, D.I., Petoukhov, M.V. & Koch, M.H. Determination of domain structure of proteins from X-ray solution scattering. *Biophys. J.* **80**, 2946–2953 (2001).
- Chacon, P., Diaz, J.F., Moran, F. & Andreu, J.M. Reconstruction of protein form with X-ray solution scattering and a genetic algorithm. *J. Mol. Biol.* **299**, 1289–1302 (2000).
- Yamagata, A. & Tainer, J.A. Hexameric structures of the archaeal secretion ATPase GspE and implications for a universal secretion mechanism. *EMBO J.* **26**, 878–890 (2007).
- Krukenberg, K.A. *et al.* Multiple conformations of *E. coli* Hsp90 in solution: insights into the conformational dynamics of Hsp90. *Structure* **16**, 755–765 (2008).
- Pascal, J.M. *et al.* A flexible interface between DNA ligase and PCNA supports conformational switching and efficient ligation of DNA. *Mol. Cell* **24**, 279–291 (2006).
- Chen, B. *et al.* ATP ground- and transition states of bacterial enhancer binding AAA+ ATPases support complex formation with their target protein, sigma54. *Structure* **15**, 429–440 (2007).
- Service, R.F. Structural biology. Protein structure initiative: phase 3 or phase out. *Science* **319**, 1610–1613 (2008).
- Matthews, B.W. Protein Structure Initiative: getting into gear. *Nat. Struct. Mol. Biol.* **14**, 459–460 (2007).
- Round, A.R. *et al.* Automated sample-changing robot for solution scattering experiments at the EMBL Hamburg SAXS station X33. *J. Appl. Crystallogr.* **41**, 913–917 (2008).
- Toft, K.N. *et al.* High-throughput small angle X-ray scattering from proteins in solution using a microfluidic front-end. *Anal. Chem.* **80**, 3648–3654 (2008).
- Robin, D. *et al.* Superbend upgrade on the advanced light source. *Nucl. Instrum. Methods Phys. Res. A* **538**, 65–92 (2005).
- Konarev, P.V. *et al.* PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *J. Appl. Crystallogr.* **36**, 1277–1282 (2003).
- Petoukhov, M.V., Konarev, P.V., Kikhney, A.G. & Svergun, D.I. ATSAS 2.1—towards automated and web-supported small-angle scattering data analysis. *J. Appl. Crystallogr.* **40**, S223–S228 (2007).
- Svergun, D.I. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.* **76**, 2879–2886 (1999).
- 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).
- Perez, J. *et al.* Heat-induced unfolding of neocarzinostatin, a small all-beta protein investigated by small-angle X-ray scattering. *J. Mol. Biol.* **308**, 721–743 (2001).
- Sokolova, A.V., Volkov, V.V. & Svergun, D.I. Prototype of a database for rapid protein classification based on solution scattering data. *J. Appl. Crystallogr.* **36**, 865–868 (2003).
- Romanowski, M.J. & Burley, S.K. Crystal structure of the *Escherichia coli* shikimate kinase I (AroK) that confers sensitivity to mecillinam. *Proteins* **47**, 558–562 (2002).
- Mai, X. & Adams, M.W. Purification and characterization of two reversible and ADP-dependent acetyl coenzyme A synthetases from the hyperthermophilic archaeon *Pyrococcus furiosus*. *J. Bacteriol.* **178**, 5897–5903 (1996).
- Harrison, S.C. Comments on the NIGMS PSI. *Structure* **15**, 1344–1346 (2007).
- Poole, F.L. II *et al.* Defining genes in the genome of the hyperthermophilic archaeon *Pyrococcus furiosus*: implications for all microbial genomes. *J. Bacteriol.* **187**, 7325–7332 (2005).

ONLINE METHODS

SAXS data collection. All SAXS data collection was performed at the SIBYLS beamline, an international user facility. An application for experiments is accessible at <http://www.bl1231.als.lbl.gov/>. The data collection strategy has been designed to minimize errors owing to instrumentation, radiation damage and concentration-dependent phenomena. The strategy applied depended on available stock concentration and size of the protein. SAXS data were collected on 3 serial dilutions of each sample preparation starting at a maximum of 10 and a minimum of 1 mg ml⁻¹. Sample loading for data collection for each protein proceeded in the following order: lowest concentration, middle concentration, highest concentration followed by a final buffer measurement. The sample cell was washed between loading protein solutions using a mild detergent soak for 1 min followed by 3 rinses with buffer solution. The subtraction of data for the buffer collected before the sample was compared to data collected for the buffer after each sample to insure the subtraction process was not subject to instrument variations. Data were collected from two short and one long X-ray exposure for each protein sample. The short exposures were compared against one another to identify whether radiation damage occurred on this time scale. The beam size at the sample was 4 × 1 mm and converged at the detector to a 100 × 100 μm spot. The large beam size at the sample spread radiation over the entire sample, greatly reducing radiation damage. Concentrations were compared to one another to determine whether concentration-dependent structure factors contributed to the data. In two cases minor concentration dependence was observed and corrected by extrapolating behavior to zero concentration⁹ using code developed in-house. A final scattering curve used for analysis was created for each sample (**Supplementary Fig. 2**).

For most samples, only 1 Å X-rays were used. Short exposures were 0.5 s, and long exposures were 5 s. However, for proteins with long dimensions such as PF1548, 1.5 Å wavelength was also used to better define the maximum distances. Short and long exposures were 4 and 40 s, respectively. All data were collected at room temperature (18–21 °C).

SAXS data analysis. For global parameter (**Table 1** and **Supplementary Tables 1,2**) and pair distribution (**Supplementary Fig. 3**) extraction, we used PRIMUS²¹. X-ray scattering curves calculated from atomic models by CRY SOL³¹ were compared to observed. Molecular envelopes were generated by both DAMMIN²³ and GASBOR¹⁰. Mass was estimated from the Porod volume and by the extrapolated intensity at zero q based upon three standards collected in the same experimental settings²⁴. GASBOR requires the number of residues. Mixtures of proteins with known structures were analyzed with OLIGOMER²¹. SASREF³² was used for rigid body docking.

Leveraging the protein structure database. BLAST³³ was used to identify homologs with PDB structures. To test SAXS identification of similar structures, we used the web utility DARA²⁶ (Database for rapid protein characterization) to rank agreement between experimental data and scattering curves ($q < 0.15 \text{ Å}^{-1}$) calculated from PDB structures. Stored scattering profiles calculated from PDB atomic coordinates were scanned to match profiles to experimental data.

Sample preparation: expression clones. The PF0015-PF0014 coexpression pET24d Bam plasmid consisted of a gene encoding His-tagged PF0015 with in-frame TEV site between the His tag and the protein N terminus, followed by nontagged PF0014, and the pET24d Bam expression plasmid for PF1205 included the gene encoding *P. furiosus* rubredoxin fused in frame between the His tag and PF1205. The remaining His-tagged recombinant proteins had previously been prepared by an X-ray crystallographic structural genomics effort, and their genes were cloned in the expression plasmid, pET24d Bam³⁴. The expression clones for superoxide reductase (SOR), rubredoxin (Rd) and ferredoxin (Fd) have been previously described^{35–37} and were used for the production of native (nontagged) recombinant proteins.

Expression in *E. coli* and purification. All the His-tagged proteins were produced in the *E. coli* strain, BL21 Star DE3 pRIL (Stratagene) as the host. The His-tagged recombinant proteins were purified according to the high-throughput protocols established for *P. furiosus* protein production³⁸. In brief, cells from 1 l induced cultures were lysed and heated at 80 °C for 30 min to precipitate *E. coli* proteins, cooled to 4 °C and then clarified by centrifugation (40,000g). The clarified supernatant was applied to a 5 ml His-trap Ni affinity column (5 ml) using an AKTA explorer (GE Healthcare). The column was washed with 5 column volumes (CV) of 20 mM phosphate buffer (pH 7.0) containing 500 mM NaCl, 10 mM imidazole, 5% (vol/vol) glycerol and 2 mM dithiothreitol. The absorbed protein was eluted with a gradient of 0–500 mM imidazole over 20 CV. The major protein peak was collected and concentrated to 10 ml by ultrafiltration (Millipore), diluted 15-fold in 20 mM Tris buffer (pH 8.0) containing 5% (vol/vol) glycerol and 2 mM dithiothreitol, and then applied to a column (5 ml) of Q Sepharose (GE Healthcare). The column was washed with 5 CV of the same buffer, and the bound proteins were eluted with a 0–1 M NaCl gradient over 20 CV. The major protein peak was concentrated to 5 ml and applied to a 16/60 column size exclusion column of Superdex 75 or Superdex 200 (for PF0015-PF0014) (GE Healthcare) equilibrated with the same Tris buffer. The major protein peak from this column was collected and concentrated to a volume of ~1 ml by ultrafiltration. Samples were buffer exchanged into 20 mM Tris (pH 8.0), 300 mM NaCl and 2 mM DTT for SAXS analysis. Recombinant, native (untagged) rubredoxin (PF1282, Rd), superoxide reductase (PF1281, SOR) and ferredoxin (PF1909, Fd) were expressed and purified as described previously^{39–41}.

Analytical procedures. Protein concentrations were estimated using the Biuret protein assay⁴². SDS-PAGE and native-PAGE analysis of protein samples were done using 4–20% gradient gels (Criterion gel system; Biorad) and run according to the manufacturer's instructions.

31. Svergun, D., Barabero, C. & Koch, M.H. CRY SOL—a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* **28**, 768–773 (1995).

32. Petoukhov, M.V. & Svergun, D.I. Global rigid body modeling of macromolecular complexes against small-angle scattering data. *Biophys. J.* **89**, 1237–1250 (2005).

33. Altschul, S.F. *et al.* Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

34. Adams, M.W. The Southeast Collaboratory for Structural Genomics: a high-throughput gene to structure factory. *Acc. Chem. Res.* **36**, 191–193 (2003).

35. Yeh, A.P. *et al.* Structures of the superoxide reductase from *Pyrococcus furiosus* in the oxidized and reduced states. *Biochemistry* **39**, 2499–2508 (2000).

36. Bau, R. *et al.* Crystal structure of rubredoxin from *Pyrococcus furiosus* at 0.95 Å resolution, and the structures of N-terminal methionine and formylmethionine variants of Pf Rd. Contributions of N-terminal interactions to thermostability. *J. Biol. Inorg. Chem.* **3**, 484–493 (1998).
37. Nielsen, M.S., Harris, P., Ooi, B.L. & Christensen, H.E. The 1.5 Å resolution crystal structure of [Fe₃S₄]-ferredoxin from the hyperthermophilic archaeon *Pyrococcus furiosus*. *Biochemistry* **43**, 5188–5194 (2004).
38. Sugar, F.J. *et al.* Comparison of small- and large-scale expression of selected *Pyrococcus furiosus* genes as an aid to high-throughput protein production. *J. Struct. Funct. Genomics* **6**, 149–158 (2005).
39. Jenney, F.E. Jr & Adams, M.W. Rubredoxin from *Pyrococcus furiosus*. *Methods Enzymol.* **334**, 45–55 (2001).
40. Clay, M.D. *et al.* Spectroscopic studies of *Pyrococcus furiosus* superoxide reductase: implications for active-site structures and the catalytic mechanism. *J. Am. Chem. Soc.* **124**, 788–805 (2002).
41. Brereton, P.S., Verhagen, M.F., Zhou, Z.H. & Adams, M.W. Effect of iron-sulfur cluster environment in modulating the thermodynamic properties and biological function of ferredoxin from *Pyrococcus furiosus*. *Biochemistry* **37**, 7351–7362 (1998).
42. Goa, J. A micro biuret method for protein determination; determination of total protein in cerebrospinal fluid. *Scand. J. Clin. Lab. Invest.* **5**, 218–222 (1953).