Low-Resolution Structures of Transient Protein-Protein Complexes using Small-Angle X-Ray Scattering

Jascha Blobel[1], Pau Bernadó[1], Dmitri I. Svergun[2,3], Romà Tauler[4] and Miquel Pons[1,5,*]

1. Laboratory of Biomolecular NMR, Institute for Research in Biomedicine. Parc Científic de Barcelona, Baldiri Reixac, 10, 08028 Barcelona, Spain.

2. European Molecular Biology Laboratory, Hamburg Outstation, Notkestrasse 85, 22603 Hamburg, Germany.

3. Institute of Crystallography, Russian Academy of Sciences, Leninsky pr. 59, 117333 Moscow, Russia.

4. Department of Environmental Chemistry, IIQAB-CSIC, Jordi Girona 18, Barcelona 08034, Spain.

5. Departament de Química Orgànica, Universitat de Barcelona, Martí i Franquès, 1-11, 08028 Barcelona, Spain.

* Corresponding author: Miquel Pons (mpons@ub.edu)

ABSTRACT:

The determination of the three-dimensional structure of a weak protein-protein complex in solution using Small-Angle X-ray Scattering requires the deconvolution of its contribution from those of other components coexisting in equilibrium.
Using the oligomerization equilibrium of low molecular weight phosphatase (lmwPTP) as a model system, we show computationally and experimentally that the individual low-resolution structures of monomeric and dimeric lmwPTP can be determined from a small number of SAXS curves using the Multivariate Curve Resolution with Alternating Least Squares (MCR-ALS) algorithm. The dimeric complex represents no more than 15% of the macromolecules in the most concentrated sample. The derived structures are in good agreement with the crystallographic ones and the dissociation constant match the one measured by NMR. These results demonstrate the power of SAXS, in combination with MCR-ALS, to study transient biomolecular complexes. The limits of the method were explored using a three species model that describes the oligomerization of lmwPTP at higher concentrations.

INTRODUCTION

Weak protein-protein interactions are inherent components of the functional interactome, i.e. the ensemble of macromolecular interactions responsible for regulatory processes.[1,2] In contrast to the long lived constitutive interactions that involve large contact areas and substantial interaction energies, regulatory interactions involve small binding energies, comparable to k•T, resulting in a dynamic situation in which bound and free forms are easily inter-converted. The corresponding equilibrium is readily displaced in response to minor environmental changes and the dynamic nature of the complexes depicts a major experimental challenge for their structural characterization. Solution techniques such as NMR and Small Angle Scattering of X-rays (SAXS) or neutrons (SANS) are best suited for the study of transient protein complexes. NMR and scattering techniques can be measured under very similar experimental conditions and are highly complementary: NMR can provide information about the contact interface (through chemical shift perturbation or NOEs) and the relative orientation (from residual dipolar couplings or spin relaxation rates) of the interacting partners. A limited long range NMR information can be provided by paramagnetic induced relaxation or pseudocontact shifts.[3-5] On the other hand, SAXS can provide low-resolution structures of pure species in solution, which define the overall shape of the complex.[6-9]

When different species are present, the SAXS curves are the weighted average of all the coexisting species in solution and pure scattering curves of all participating species are required to interpret SAXS data.[10-12] Individual scattering curves can be obtained

experimentally from isolated species or computed from 3D models. As shown recently, transient protein complexes can be isolated through cross-linking experiments and their pure scattering curves recorded.[13]

An elegant strategy based on Singular Value Decomposition (SVD) has been employed to study protein folding by SAXS using a series of scattering curves measured in different conditions. Exploiting the significantly different features of the scattering curves of folded and unfolded proteins, the pure profiles for the coexisting species along the folding pathway were obtained.[14-17]

Weak oligomerization processes involving the interaction between identical proteins are specially challenging for both NMR and SAXS. Previous studies from our group showed that ^{15}N NMR relaxation data at different protein concentrations could be used to characterize the formation equilibrium of different oligomeric species of a low molecular weight phosphatase (lmwPTP).[18,19] The analysis of the NMR relaxation data required the previous knowledge of the three-dimensional structure of the major species present (monomer and dimer), in this case derived from crystals.[20,21] These studies were complemented with ^{129}Xe NMR experiments in which the noble gas atoms acted as spies of the formation of specific protein oligomeric species.[19]

Recent work using synthetic SAXS data has suggested that the SVD based approach could also be applicable for the study of protein oligomerization.[22]

Here we use lmwPTP as a test system to show that SAXS data obtained at different protein concentrations can be used to simultaneously characterize oligomerization equilibria and to obtain the *ab initio* low-resolution structure of a dimer representing only a minor portion (less than 20%) of the total protein in the sample. At higher lmwPTP

concentrations, variable concentration SAXS also provides evidence about the presence of at least one additional oligomer and about its stoichiometry, although the information obtained is not sufficient to derive a reliable structure.

The layout of the paper is the following: In the first part we show a global analysis of variable concentration SAXS data, the evidence of the formation of oligomers at the higher concentrations and the derivation of the number of species that contribute significantly to the observed data. In the second part we briefly present the Multivariate Curve Resolution method using an Alternating Least Square (MCR-ALS) algorithm to extract the SAXS contributions from individual species. In the following sections we use this method to study the low concentration part of the experiment, where only two species are present and we show, using simulated and real data, that low-resolution structures of lmwPTP monomer and dimer can be obtained without *a priori* information. Finally, we analyze the complete concentration range in which higher oligomers are formed to explore the limits of the approach.

RESULTS

Concentration dependence of global parameters derived from SAXS data

The eight scattering curves of lmwPTP recorded at total protein concentrations of 0.056, 0.17, 0.25, 0.34, 0.41, 0.48, 0.55, and 0.60 mM (1.01 to 10.8 mg/ml) are displayed in Fig. 1a. Different parameters reporting on size and shape characteristics of each sample can be used to compare the different curves. The forward intensity, $I(0)$, obtained through the extrapolation of the scattering curve to zero angle using Guinier's approach, is proportional to the total protein concentration and the molecular mass of the solute (Fig. 1b). For a pure species, the forward scattering increases linearly with the concentration. However, the observed increase is not linear with lmwPTP concentration, indicating that several species are present and that their relative populations depend on the total protein concentration, as expected for an oligomerization process. The dashed lines represent the expected values for pure monomer (extrapolated from $I(0)$ at the lowest concentration) and pure dimer. The theoretical values are based on the crystal structures and take into account, in addition to the fact that the concentration of dimers is half that of monomers for the same amount of protein, the differences in the solvent shell in the monomer and the dimer. The comparison of the calculated scattering profiles from the crystallographic monomeric and dimeric structures of the lmwPTP shows that $I(0)_D = 3.45 \cdot I(0)_M$.

The radius of gyration (R_g) characterizes the average size of the dissolved particles and can therefore provide insight into the oligomerization processes. The apparent R_gs, estimated using Guinier's approximation for each scattering curve, are plotted versus the

corresponding total protein concentration in Fig. 1c. The increase in apparent R_g is also consistent with an oligomerization process. Even at the lowest concentration studied, which is very close to the sensitivity limit of SAXS, the experimental R_g is larger than expected for pure lmwPTP monomer. Contrary to the $I(0)$ value yielding a weight average of the molecular mass, R_g is a so-called z-average which is more sensitive to the presence of larger particles i.e. higher oligomers. The R_g versus concentration plot is approximately linear up to 0.34 mM but significant deviations from linearity are observed beyond 0.41 mM.

The diameter of the largest particle in each sample (D_{max}) is obtained from the distance distribution function calculated with the program GNOM.[23] As in the case of $I(0)$ and R_g, a systematic increase with lmwPTP concentration is observed in D_{max}. The concentration dependence of D_{max} is plotted in Fig. 1d. A plateau with an average D_{max} of 7.46 nm is reached between 0.34 to 0.48 mM lmwPTP. This dimension corresponds closely to the largest dimension of the crystallographic lmwPTP dimer (7.5 nm).[20] At higher concentrations, increasing values of D_{max} point to the presence of larger oligomers.

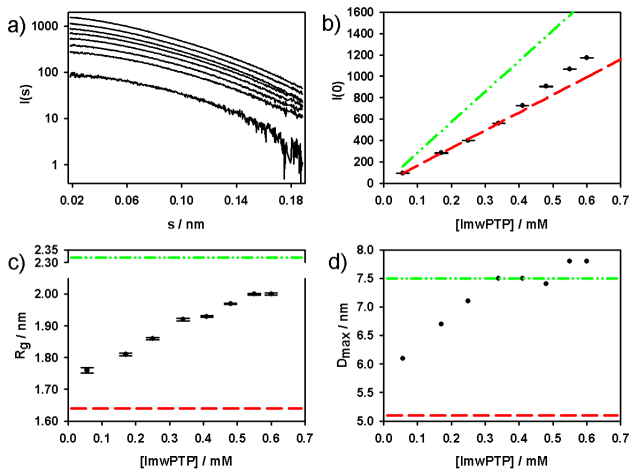


Figure 1. a) Scattering intensities measured for lmwPTP at total concentrations of (from bottom to top) 0.056, 0.17, 0.25, 0.34, 0.41, 0.48, 0.55 and 0.60 mM as a function of the momentum transfer $s = 4\pi \sin(\theta) / \lambda$. b-d) lmwPTP concentration dependency of different parameters derived from SAXS. b) Forward intensity $I(0)$ at different lmwPTP concentrations. c) Apparent R_g. d) D_{max}. The dashed lines are the values calculated for pure monomer (red long dash) and dimer (green dash dot dot) solutions.

Determination of the number of species by Principal Component Analysis (PCA)

SAXS data of complex mixtures are concentration weighted linear combinations of the scattering curves arising from the different species present. Principal Component Analysis (PCA)[24] of a series of experiments, in which the same species are present in different proportions, can be used to determine the minimum number of coexisting species required to account for the data. Following the analysis of R_g and D_{max}, we analyzed the complete set of curves and a subset containing the curves up to 0.41 mM. Figure 2 shows the first four PCA eigenvectors, together with the corresponding autocorrelation functions C(r), for the low concentration subset and the complete set. Clearly in both cases the first two eigenvectors contain significant features while the fourth eigenvector shows only random fluctuations around zero. The third eigenvector of the complete set show small but significant systematic deviations from zero, most clearly

seen in the autocorrelation function. In contrast, the third eigenvector of the low concentration subset is already at the noise level. The significance of the different PCA eigenvectors can be quantified by the relative values of the associated eigenvalues (Supplementary material, figure S1). When the analysis is restricted to the low concentration samples the third eigenvalue is already at the noise level. In contrast, when the complete concentration range is analyzed, the third eigenvalue is significantly higher than the successive ones. The PCA in combination with the analysis of the size descriptors derived from SAXS data suggests the presence of only two significant species up to a lmwPTP concentration of 0.41 mM, while at least one additional species is needed to account for the SAXS data obtained at higher concentrations.

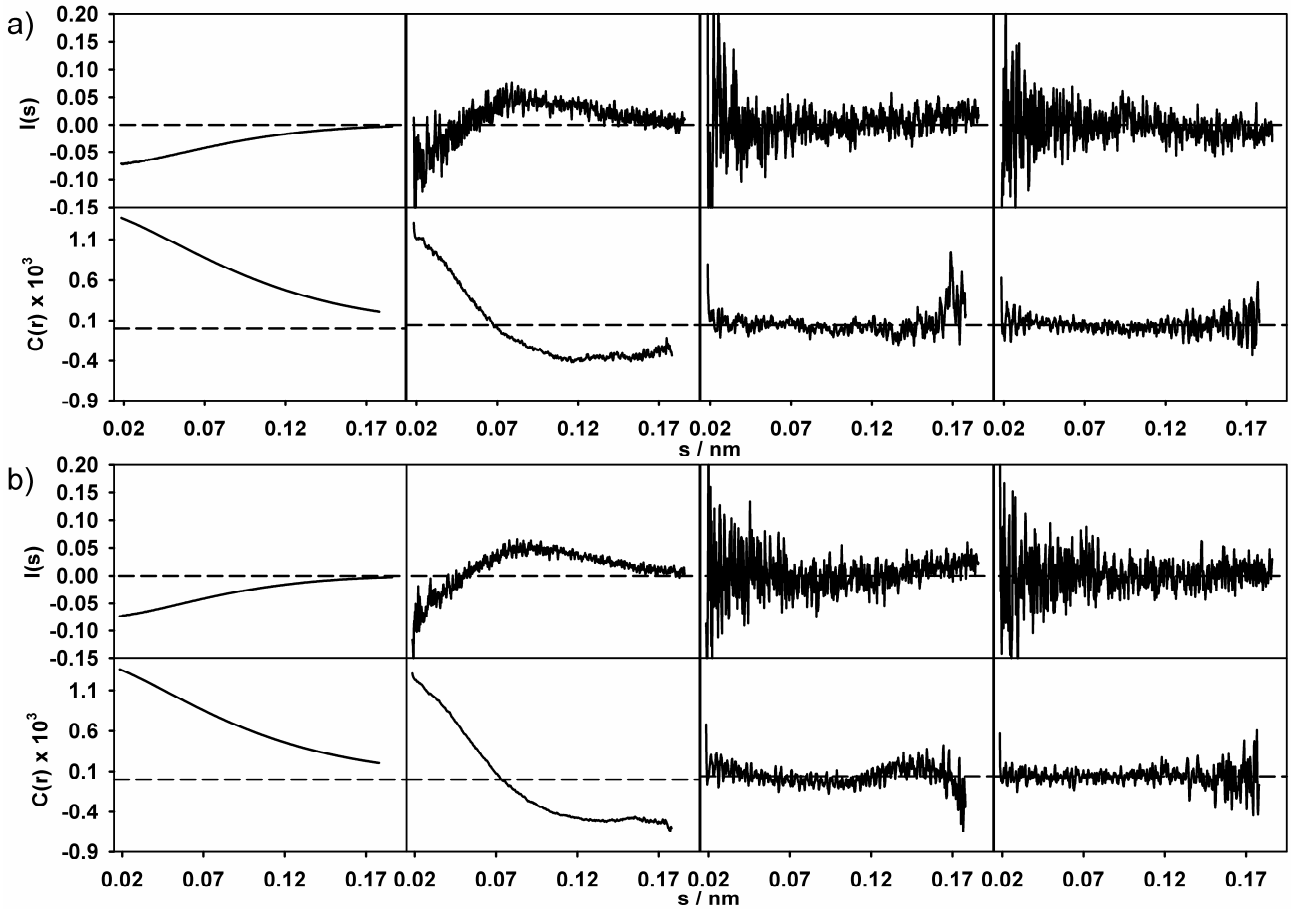


Figure 2. First four eigenvectors resulting from PCA of concentration dependent SAXS data of lmwPTP in the concentration ranges 0.056 – 0.41 mM (a) and 0.056 – 0.60 mM (b). In each set, the bottom row corresponds to the vectors' autocorrelation functions.

Low concentration experiments: monomer-dimer model

The crystallographic structures of the monomeric (1pnt)[21] and dimeric (1c0e)[20] species were used to characterize the two-component equilibrium of lmwPTP through fitting the SAXS curves using OLIGOMER.[10] Each scattering curve was fitted individually to linear combinations of the theoretical scattering curves of monomer and dimer yielding their relative populations. The estimated populations of monomer and dimer at each lmwPTP concentration and the fitting error χ^2 are shown in Fig. 3. The fitting error to a two-state model decreases with decreasing protein concentration down to the detection limit,

consistent with the PCA results. A dissociation constant (K_d) of 0.93 mM is obtained using data up to 0.41 mM.

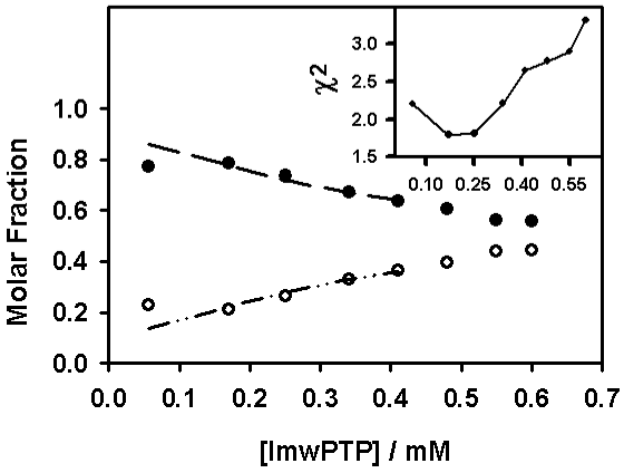


Figure 3. Molar fractions of monomer (filled circles) and dimer (open circles) calculated with OLIGOMER using the crystal structures of monomer and dimer lmwPTP. The dashed (monomer) and dash dot dot (dimer) lines correspond to the molar fractions calculated using a K_d = 0.93 mM. The inset shows the OLIGOMER fitting errors at different lmwPTP concentrations.

Multivariate Curve Resolution

In the previous section we used known structures to analyze the SAXS data. We wanted to investigate if, in the absence of prior structural information, it would be possible to derive simultaneously both the dissociation constant and the low-resolution structures of the monomer and dimer using only SAXS data taken at different protein concentrations. For this, we employed Multivariate Curve Resolution (MCR), a powerful chemometrics method used to analyze sets of multivariate (multichannel, multiwavelength, etc) signals arising from weighted linear combinations of pure components. Details about MCR and its applications in different fields are given elsewhere.[25-29] An extended description of the method is provided as supplementary information.

In the present application, the initial data set consisted of a matrix of scattering curves obtained at different protein concentrations from which MCR analysis can provide estimates of two factor matrices, one related with the unmixed (pure) SAXS curves of the components in the mixtures (the spectral matrix, SM) and another one related with their relative concentration profiles (the concentration matrix, CM). The MCR factor decomposition method imposes chemically relevant constraints, like non-negativity or unimodality, and uses an Alternating Least Squares (MCR-ALS) algorithm to estimate the best combination of concentration and spectral matrices.

Singular Value Decomposition (SVD) has been used in other studies with a similar purpose of decomposing a matrix of spectra obtained in different conditions.[14-17,22] The problem using SVD lays in its interpretability. Even though SVD provides a mathematically consistent decomposition of the data and identifies the number of components, the extraction of the curves of the pure components is not warranted.

MCR-ALS of SAXS curves using a monomer-dimer equilibrium: synthetic data

First, the MCR-ALS strategy to derive pure scattering curves of different species from SAXS mixture curves was investigated using synthetic data. Using the curves calculated from the available X-ray structures and a K_d = 1 mM, 101 synthetic scattering curves of lmwPTP samples ranging from 0.05 mM (0.9 mg/ml, the experimental limit of detection) up to 1.55 mM (27.9 mg/ml) in steps of 0.015 mM were generated. At the lowest concentration, the molar fraction of monomer is more than 91% whereas the dimer represents up to 40% of the macromolecules at 1.55 mM lmwPTP. Noise was included as explained in the Materials and Methods section. The 101 curves are shown in Fig. 4a.

MCR-ALS of the synthetic data assuming a monomer-dimer equilibrium but no structural information about either of the two species provided an estimated K_d = 0.992 mM (Fig. 4b inset), in good agreement with the expected K_d = 1 mM. A fit using the software CRYSOL[30] of the estimated pure SAXS curves with the original structures gave χ_i^2 of 1.260 and 1.266 for the monomer and dimer, respectively (Fig. 4b). The calculated ratio $I(0)_D/I(0)_M$ is 3.45, equivalent to the ratio calculated from the 3D structures of the dimer and the monomer with CRYSOL.

The MCR-ALS application to SAXS data was further tested using an experimentally realistic number of scattering curves in the concentration range of the available experimental lmwPTP SAXS data. Synthetic SAXS curves were simulated at the five concentrations of 0.085, 0.17, 0.25, 0.34 and 0.41 mM (Fig. 4c). With K_d = 1 mM, the maximum percentage of dimer is 21.1% at the highest protein concentration. The best solutions found using the MCR-ALS strategy were very similar to the ones obtained from the set of 101 SAXS curves, although the minimum was less well defined (Fig. 4d). The average K_d for the five best solutions was 0.898 ± 0.096 mM. The scattering curves extracted from the best solution are in very good agreement with the theoretical curves, having χ_i^2 of 1.292 and 1.283 for monomer and dimer, respectively. The ten best results have $\chi_i^2 < 1.300$. The $I(0)_D/I(0)_M$ ratio for the best solution is 3.44, in perfect agreement with the theoretical value of 3.45.

The *ab initio* low-resolution structures obtained using DAMMIN[31] show a very good superposition with the input structures (see Fig S1 in Supplementary Information). These results show that MCR-ALS is a robust method that can be successfully applied to a

small number of curves involving mixtures in which one of the components is below 20%.

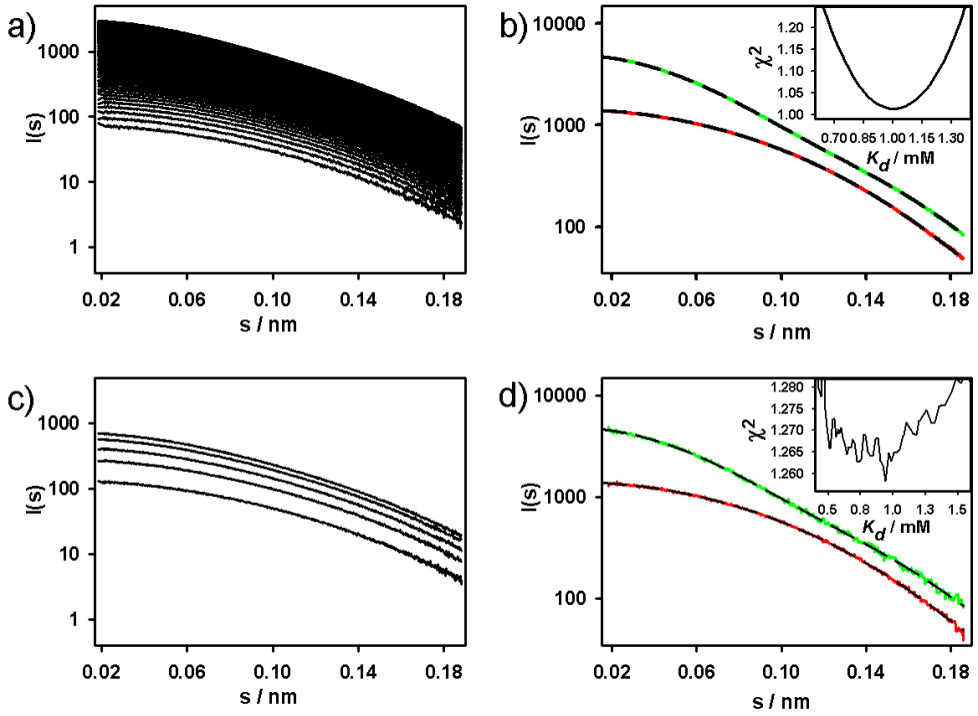


Figure 4. MCR-ALS of synthetic SAXS curves generated for the monomer-dimer equilibrium of lmwPTP. a) Simulated scattering curves at 101 different concentrations between 0.05 and 1.55 mM lmwPTP. b) Pure scattering curves of monomer (red) and dimer (green) fitted with CRYSOL. The theoretical curves are superimposed (black dashed lines). c) Reduced set corresponding to lmwPTP concentrations of 0.085, 0.17, 0.25, 0.34, and 0.41 mM. d) Pure scattering curves obtained from the reduced set. The fitting errors of MCR-ALS are shown in b) and d) as a function of K_d in the insets.

MCR-ALS of experimental SAXS curves from low concentration lmwPTP solutions

The MCR-ALS algorithm was used to analyze the five experimental scattering curves obtained from lmwPTP samples at protein concentrations up to 0.41 mM, which can be described as a monomer-dimer equilibrium. A well-defined minimum was found at a K_d = 1.72 mM with a χ^2 = 3.993 (Fig. 5a inset). The average of the dissociation constants for the five best results having χ^2 < 3.997 gives K_d = 1.62 ± 0.12 mM, in good agreement

with the 1.5 ± 0.1 mM value determined by ^{1}H-NMR chemical shift changes assuming a two component equilibrium.[32] Fig. 5a shows the SAXS curves of the pure components extracted by the MCR-ALS algorithm in color, superimposed on the by CRYSOL fitted crystal structure scattering curves of lmwPTP monomer and dimer. A very good fit to the crystallographic structures of the monomer and dimer are obtained, although visually the fitting of the monomer is slightly better. A quantitative estimate of the goodness of fit depends on the estimate of the noise level of the pure curves (see materials and methods). The average $I(0)_D/I(0)_M$ ratio of the five best solutions is 3.80 ± 0.05, slightly exceeding the theoretical value of 3.45, probably being caused by the presence of small quantities of higher oligomers. The R_g values derived from the scattering curves using Guinier's approximation are in very good agreement with the theoretical values for the monomer (16.37 ± 0.01 Å versus 16.45 Å) and the dimer (23.90 ± 0.1 Å versus 23.32 Å). *Ab initio* shape reconstructions of the scattering curves extracted by the MCR-ALS method are calculated by DAMMIN and show to be in very good agreement with the crystal structures of the monomeric and dimeric lmwPTP (Fig. 5b).

The low-resolution structure correctly represents the oblate shape of the monomer and the characteristic dumbbell shape of the dimer. Although optimizations were performed without symmetry constraints, the final shape of the dimer shows a nearly symmetric object.

These results confirm that the combination of MCR-ALS and SAXS provides correct *ab initio* low-resolution structures of two individual species in equilibrium, of which the dimer represents only 3.0 to 15.1% of the total macromolecular concentration. For the fitting protocol, no *a priori* structural assumptions had to be made.

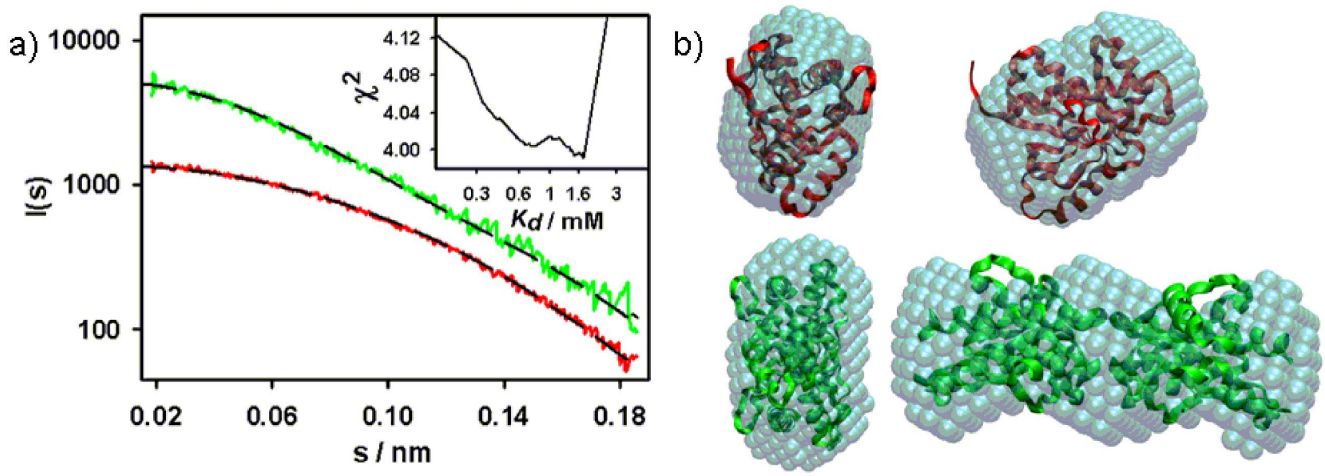


Figure 5. a) Scattering curves of the pure components superimposed on the theoretical scattering curves of monomeric (red) and dimeric (green) species (black dashed lines) as calculated by CRYSOL. b) A*b initio* low-resolution reconstructions derived from the pure curves shown in a), superimposed with the corresponding crystal structures of either lmwPTP monomer (1pnt, red) or dimer (1c0e, green).

Exploring the limits of MCR-ALS analysis of SAXS data

Previous results using NMR relaxation data showed that lmwPTP participates in a more complex oligomerization equilibrium involving at least one species in addition to the monomer and dimer.[18,19] This is also observed by SAXS at concentrations above 0.41 mM. The feasibility of studying this rather complex equilibrium by MCR-ALS was first tested by using eight synthetic scattering curves simulated using the dissociation constants previously determined by NMR (K_d = 6.00 mM & K_{tet} = 0.07 mM).[19] As the three pure scattering curves, the one derived from the X-ray structures of the monomer, the dimer, and a putative compact tetramer with a R_g of 26.86 Å, compatible with NMR relaxation experiments, were used. The proportions of dimer and tetramer at the highest protein concentration under the simulated conditions are 6.9% and 3.4%.

Attempts to decompose the SAXS curves without any prior structural information provided only the scattering curve of the major species, which was consistent with the monomer structure. The other scattering curves were combinations of dimer and tetramer. Using the calculated curves from the known structures of the monomer and dimer as additional input, the correct dissociation constants were found (K_d = 6.017 ± 0.098 mM & K_{tet} = 0.072 ± 0.003 mM, average of the five best solutions) and the extracted scattering curve for the tetramer was in very good agreement (χ_i^2 = 2.823 for the CRYSOL fit, Fig. 6c) with the theoretical structure used to generate the synthetic data. The agreement between the extracted and theoretical values of R_g (27.70 Å and 26.86 Å, respectively) and $I(0)_T/I(0)_M$ (13.1 and 12.9, respectively) confirmed that the putative tetramer structure could be derived from the simulated data if the structure of monomer and dimer were known beforehand.

The MCR-ALS analysis of the eight experimental SAXS curves obtained up to 0.60 mM lmwPTP, assuming a monomer-dimer-tetramer equilibrium and the knowledge of the structure of the monomeric and dimeric species, failed to provide a scattering curve from which a low-resolution structure of the additional species could be extracted. Restricting the analysis to the forward scattering $I(0)_T$ derived from the extracted curve and comparing it with the value from the monomeric species ($I(0)_T/I(0)_M$), we find a value of 12.847 ± 2.00 for the five best solutions, well within the range of values computed from 5,000 tetramers generated by random docking of two dimers ($12.55 < I(0)_T/I(0)_M < 13.23$), see Materials and Methods. The results using experimental SAXS data are in agreement with previous NMR results concerning the stoichiometry and dissociation constants (K_d & K_{tet}), although it was not possible to extract a low-resolution structure

from a third minor species, representing less than 4% of the protein in solution. Although at this concentration the tetramer structure used to generate synthetic data could still be retrieved, the failure using experimental data may reflect the effect of deviations from the ideal three-species model or additional noise sources not included in the synthetic data. These results illustrate the current limitations of the MCR-ALS analysis of complex equilibria using SAXS.

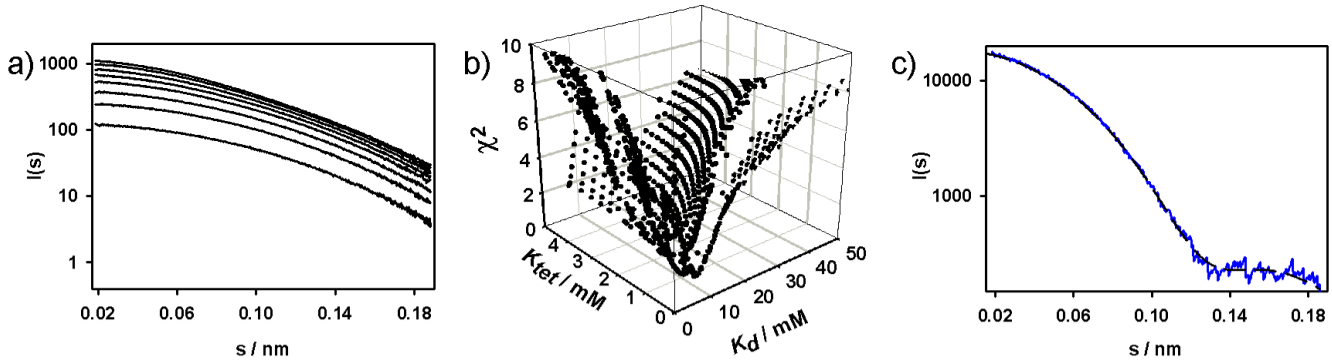


Figure 6. a) Simulated scattering curves at (from bottom to top) 0.085, 0.17, 0.25, 0.34, 0.41, 0.48, 0.55, and 0.60 mM of lmwPTP assuming a monomer-dimer-tetramer oligomerization model governed by a K_d = 6.0 mM and a K_{tet} = 0.07 mM. b) Two dimensional grid search showing a minimum around the expected K_d and K_{tet} values. c) MCR-ALS derived scattering curve of the model tetramer (blue), superimposed on the theoretical curve (black dashed line) calculated from the tetrameric model with CRYSOL.

DISCUSSION

The functionally important weak protein-protein complexes can usually not be isolated in their pure form as they coexist with their individual components in a dynamic equilibrium in solution. Therefore, their structure determination presents the challenge to extract the relevant structural information unique to each component of the mixture. In this work we have used the oligomerization of lmwPTP as a model system to study weak protein-protein interactions.[33] This system is well defined, the crystallographic structures

of monomer and dimer are known,[20,21] and the equilibrium has been extensively characterized using ultracentrifugation,[20] chemical shift perturbation,[32] ^{15}N NMR relaxation,[18] and ^{129}Xe NMR.[19]

SAXS data were obtained under identical conditions at different concentrations from a single protein batch. It should be noted that the complete data set was measured consecutively ensuring a constant scaling factor for all the curves. This is essential for the simultaneous analysis of all the data, which lies in the heart of the MCR-ALS method. The data have been analyzed at different levels to explore the limits of the information that can be derived from SAXS depending on the prior knowledge about the system.

The first level of analysis is based on simple shape parameters obtained from the analysis of each scattering curve, such as the forward scattering $I(0)$, the apparent radius of gyration R_g or the distance distribution function and, in particular, its maximum value D_{max}. Shape analysis of variable concentration SAXS data obtained at increasing lmwPTP concentrations provided a simple diagnostic for the formation of different oligomers.

At a second level, Principal Component Analysis (PCA) provides information on the number of species present and the concentration ranges at which a given model applies. For lmwPTP, PCA confirmed the almost exclusive existence of monomer and dimer at concentrations up to 0.41 mM but the presence of a more complex equilibrium at higher concentrations.

Finally, the MCR-ALS method extracts the individual scattering curves of the complex and its components present in equilibrium. Simultaneously, the dissociation constant governing the relative concentration of the different species is obtained. The separated

scattering curves of lmwPTP monomer and dimer, the later representing only 3.0 to 15.1% of the protein in the mixtures, could be extracted without using any prior structural knowledge from the analysis of only five experimental SAXS curves obtained at concentrations up to 0.41 mM. From the extracted curves, low-resolution structures in good agreement with the crystallographic ones could be obtained. The dissociation constant also matched the one derived from NMR.[32]

Deconvolution of concentration dependent SAXS curves of oligomeric mixtures had previously been shown using SVD and simulated data, although no analysis with experimental data was presented. In addition, the simulated conditions assumed rather low dissociation constants (μM to low mM) such that even at protein concentrations below the usual experimental detection limit (0.25 mg/ml) a large variation in the concentration of all protein species was ensured and, in some cases, the complex was the dominant species.[22]

The limits of MCR-ALS were explored by studying the more complex equilibrium that exists at higher lmwPTP concentrations where additional very low concentration species are present. Simulations assuming the monomer-dimer-tetramer equilibrium previously suggested by NMR data suggested that MCR-ALS could provide the structure of the tetramer if the structures of the other species were known. Experimentally, we could only extract an estimate of the molecular mass of the tetramer from the forward scattering value but we could not obtain a pure scattering curve for this species, which is estimated to represent only around 3.4% in the most concentrated sample. At the corresponding minimum, the MCR-ALS determined dissociation constants of both dimer and tetramer did match the ones determined by NMR though.[19]

Our results further underline the usefulness of SAXS in the large-scale studies of biomolecular complexes. Although we have applied the approach to an oligomerization process, other experimental data sets in which the relative concentrations of two or more partners are changed in a controlled way could eventually allow the study of transient hetero-complexes, provided that the presence of additional species (e.g. homodimers as well as heterodimers) could be ruled out or accounted for using previous information.[34] The scattering curves extracted using MCR-ALS represent a direct insight into the structure of the contributing partners. For all these reasons we envision a prominent role of SAXS in combination with intelligent data analysis, such as MCR-ALS, in the large-scale studies of macromolecular assemblies.

MATERIALS AND METHODS

Protein expression and purification

Unlabeled lmwPTP was expressed as described before.[19,35] Three days before SAXS measurements, the solution was re-purified over a size exclusion column (Superdex 75) and the buffer exchanged to 200 mM potassium phosphate, 3 mM sodium azide and 10 mM TCEP*HCl at a pH = 6.00. The final protein concentration was determined by UV absorption as described before.[32]

SAXS measurements and overall parameter determination

Synchrotron radiation X-ray scattering data was collected following standard procedures on the X33 camera at the European Molecular Biology Laboratory (EMBL) on the storage ring DORIS III of the Deutsches Elektronen Synchrotron (DESY).[36] Scattering curves were recorded on a MAR345 image plate detector for the eight different lmwPTP concentrations of 0.60, 0.55, 0.48, 0.41, 0.34, 0.25, 0.17 and 0.056 mM. (10.8 to 1.01 mg/ml). The samples were prepared by successive dilution of a 1 mM lmwPTP sample with the corresponding buffer. Scattering curves of the buffer were collected before and after each acquisition of a protein sample to avoid systematic error. SAXS measurements were performed at 37°C after letting the sample equilibrate for two minutes. The scattering curves covered the range of momentum transfer of $0.0956 < s < 5.0455$ nm^{-1}. The scattering due to buffer was subtracted by averaging the buffer measurements

enclosing the actual protein measurement. All data manipulations were performed with the program PRIMUS.[10]

The forward scattering $I(0)$ and the radius of gyration R_g were evaluated with the Guinier approximation assuming that, at very small angles ($s < 1.3/R_g$), the intensity is represented as $I(s)=I(0)\exp(-(sR_g)^2/3)$.[37] The actual concentration of the lowest lmwPTP concentration sample (0.056 mM) was determined from its $I(0)$ value by extrapolation of the $I(0)$ obtained at 0.17 mM lmwPTP.

Calculation of synthetic SAXS data

Synthetic scattering curves were generated with CRYSOL[30] using the crystal structures of the monomer (1pnt.pdb),[21] dimer (1c0e.pdb),[20] or a hypothetical compact model of the tetramer. Scattering curves from mixtures were generated as the concentration weighted sum of all species present. The concentrations of all species at various total protein concentrations were calculated from the dissociation constants of dimer (K_d) and tetramer (K_{tet}) defined as:

$$K_d = [M]^2/[D] \quad (1)$$

$$K_{tet} = [D]^2/[T] \quad (2)$$

Where [M], [D], and [T] correspond to the molar concentrations of the monomer, dimer, and tetramer, respectively.

To each synthetic SAXS curve, an error based on the experimental data obtained from the measurement of the sample at 0.60 mM lmwPTP was added as described before.[22] Briefly, the relative noise from the experimental data set was calculated by dividing the experimental noise ($\sigma_{exp}(s)$) by the intensity ($I_{exp}(s)$) observed:

$$k_{\exp}(s) = \sigma_{\exp}(s) / I_{\exp}(s) \quad (3)$$

The resulting factor $k_{\exp}(s)$ was related to the relevant lmwPTP concentration $[lmwPTP]_{sim}$ through its ratio with $[lmwPTP]_{exp}$ = 0.60 mM and multiplied by the intensity of the simulated scattering curve ($I_{\text{sim}}(s)$) giving the noise $\sigma_{\text{sim}}(s)$:

$$\sigma_{sim}(s) = k_{\exp}(s) \cdot \sqrt{([lmwPTP]_{\exp} / [lmwPTP]_{sim})} \cdot I_{sim}(s) \quad (4)$$

For the inclusion of noise in the simulated scattering curve, $\sigma_{\text{sim}}(s)$ was multiplied by a Gaussian distribution centered around one and added to the simulated scattering curve.

Principal Component Analysis

Principal Component Analysis (PCA) was conducted by using the Singular Value Decomposition (SVD) algorithm implemented in Matlab© for determining the singular values and vectors of the data matrix containing the set of SAXS spectral curves.[38] As the noise level increases significantly for consecutive eigenvectors, we calculated the autocorrelation function C(r) of each eigenvector as:

$$C(r) = \frac{1}{N-r} \sum_{i=1}^{N-r} I(i) I(i+r) \quad (5)$$

where N is the total number of points. The autocorrelation functions have a higher signal-to-noise level and reveal systematic deviations from zero in the third eigenvector resulting from the analysis of the low concentration subset.

Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) analysis of the SAXS curves

SAXS curves ranging from s = 0.185 – 1.86 nm^{-1} were used, consisting of 734 points. The maximum *s*-value was chosen due to the existence of only positive intensities for all scattering curves up to 1.86 nm^{-1}. The large divergences in the intensities up to 0.185 nm^{-1} of the lowest concentration curve (0.056 mM) restricted the analysis to low angle measures.

Initial estimations of the pure scattering curves best describing the set of SAXS curves were obtained using the same approach as in the SIMPLISMA procedure, looking for the 'purest' curves amongst the experimental ones.[29] The purest curves are those spectra channels that depend only (or mostly) on one of the components of the mixture. They served as the starting point for the MCR-ALS determination of the optimized pure SAXS scattering curves of the macromolecules present in solution and of their corresponding concentrations. Scattering curves number 20 (0.3 mM) and 101 (1.55 mM) were selected as starting search points in the first analysis conducted, consisting of 101 synthetic SAXS data. For the experimental data set the curves corresponding to the extreme concentrations are the obvious choice. When the signals of the individual components are highly overlapped, as is the case of SAXS curves, the pure spectra depend basically from the applied constraints used during the ALS optimization and the choice of the initial estimates of the curves is not crucial.

Different parameters had to be selected throughout the ALS optimization, including the number of maximum iterations = 50 and the convergence criteria, which was set to a very small value of 0.0005 for the monomer-dimer equilibrium and 0.01 for the monomer-

dimer-tetramer one, as small differences are expected around a determined set of dissociation constants. In case of the equilibrium bearing a third species, the scattering curves of monomer and dimer were fixed through the option “vclos1” during the ALS optimization. In this case, scattering curves of these two components were constrained to be the theoretical scattering curves obtained from the crystal structures of monomer and dimer, which were also in good agreement with those obtained by ALS in the reduced concentration range where only these two species were present. Before the ALS optimization, the raw input SAXS experimental data is filtered by PCA for a selected number of components and the PCA filtered data matrix is used instead of the original raw SAXS curves data matrix. As a consequence of this, a more stable ALS optimization is in general achieved. ALS optimization is performed under constraints like non-negativity and unimodality (a single profile maximum), which are applied to the concentration profiles as well as to the pure SAXS curves. Details about the implementation of these constraints are given elsewhere.[25-29] Furthermore, in this work an additional constraint had been implemented to force concentration profiles to follow the mass-action oligomerization equilibria law. The way to implement such a constraint is similar to previous implementations for acid-base multi-equilibria systems and for multicomponent kinetic systems (with concentration profiles fulfilling rate laws).[39-41] Here, the analyzed monomer-dimer equilibrium is explained by a single equilibrium constant referred to as K_d (equation (1)). In order to ensure that the global minimum was found, the algorithm was initiated from different values of possible dissociation constants in a grid search manner.

Hereby, 401 different starting K_ds equally spaced on a logarithmic scale in the range of 0.01 to 100 mM were used in case of a monomer-dimer equilibrium. For the monomer-dimer-tetramer equilibrium, an additional dissociation constant (K_{tet}) is necessary. In this case the grid search included 61 different K_d and 61 different K_{tet} values equally spaced on logarithmic basis. Starting K_d values ranged from 0.2 mM to 200 mM (-0.7 – 2.3 in steps of 0.05), and K_{tet} values ranged from 0.001 mM to 1 mM (-3.0 – 0.0 in steps of 0.05) and 0.0002 to 6.3 mM (-3.7 – 0.8 in 61 steps of 0.075) for the theoretical data and experimental data, respectively.

After the adjustment of the theoretical dissociation constants to the ALS resolved concentration profile matrix (CM), a new concentration matrix is obtained. As a result, an increase in the overall error occurs, which is reduced by estimating a new scattering curve matrix (SM) by ALS optimization with previously mentioned constraints. Other than in case of the typical MCR-ALS approach, the scoring function driving the grid-search minimization is based on the real experimental data set $(I_{\exp})_{nsp,ns}$ and the corresponding error $(\sigma_{\exp})_{nsp,ns}$:

$$\chi^2 = \sqrt{\{\sum_{nsp}\sum_{ns}((I_{\exp} - I_{calc})/\sigma_{\exp})^2_{nsp,ns}\}/\{nsp \cdot ns\}} + \varepsilon \cdot \sum_{n}\{([conc]_n - [conc\_kdis]_n)^2/[conc]_n\} \quad (6)$$

Further variables are $(I_{\text{calc}})_{nsp,ns}$, which are the calculated scattering curves, *nsp* and *ns* are the number of scattering curves and the number of points *s*, respectively. $[conc]_n$ is the concentration matrix calculated by ALS for each species *n* and $[conc\_kdis]_n$ is the concentration calculated for each species from the dissociation constants. ε is a factor accounting for the differences in magnitude of the error in the spectra and the error in the concentration. It was chosen to weight the overall error χ^2 by roughly 5% for the

monomer-dimer and by 20% for a three species equilibrium at the best minimum found, resulting in $\varepsilon = 50$ in case of monomer-dimer data and $\varepsilon = 1$ and 10 for the theoretical and experimental data of a three species equilibrium. The error is mainly determined by the pure scattering curves of each species (SM), whereas the reason for the incorporation of an error due to the concentration matrix is to help guiding MCR-ALS in the right direction. Furthermore, it is of importance to remark again that the final results are based on the matrix of the scattering curves (SM) obtained from PCA filtering of the experimental data set.

Visually, the fitting of the pure scattering curves with the ones predicted from the crystallographic structures of monomer and dimer is very good. In order to calculate χ^2 values one has to estimate the real noise level of the curve extracted by MCR-ALS. Clearly, the curve for the minor species shows a higher statistical noise. Using the experimental error of the lowest concentration SAXS curve as an estimate of the experimental error of the pure dimer curve we obtained an upper limit of $\chi^2 = 1.9$. The error in the monomer curve fit should be lower. Using a constant relative noise of 3% (the CRYSOL default value) as an estimate of the noise of the monomer curve gives $\chi^2 = 2.0$.

The forward intensities $I(0)$ and the R_gs were calculated using the program autoRg[42] from the scattering curves generated by MCR-ALS.

The forward scattering $I(0)$ depends on the molecular mass. In the absence of hydration contribution $I(0)_D/I(0)_M$ equals 4. However the volume of the hydration layer for a dimer is smaller than the sum of two monomers because of the buried surface upon binding. Theoretical $I(0)_D/I(0)_M$ values were derived from the crystallographic structures using

CRYSOL. To estimate the range of possible $I(0)_T/I(0)_M$ ratios, 5,000 different tetramers of lmwPTP were obtained with the FT-Dock[43] software using the structure of the dimer as a starting template.

Low-resolution Models of monomeric and dimeric species

Low-resolution structures from pure scattering curves were built with DAMMIN.[31] This program represents the protein shape as an ensemble of *M* densely packed beads inside of the search volume, a sphere of diameter D_{max}. Maximum dimensions, D_{max}, from the monomer and dimer pure curves were obtained from the scattering curves with the indirect Fourier transform package GNOM.[23] Ten independent DAMMIN reconstructions were performed for the optimized scattering curves of the monomer and the dimer. The optimized shapes were averaged to yield the most probable models of both proteins with the software package SUPCOMB.[44]

ACKNOWLEDGEMENTS

Authors are indebted with Juan Fernández-Recio (Barcelona Supercomputing Center) for the calculation of the lmwPTP tetramers. The work was partially supported by funds from the Spanish Ministry of Education (BIO2007-63458 to MP). We acknowledge the support of the European Community-Research Infrastructure Action (under FP6 "Structuring the European Research Area Program contract no. RII/2004/ 5060008") to the EMBL-Hamburg Outstation for covering the travel and accommodation expenses at EMBL-Hamburg. DS acknowledges support from the HFSP grant RGP0055/2006-C. J.B. is a recipient of a predoctoral fellowship from the Spanish Ministerio de Education y Ciencia. P.B. holds a Ramón y Cajal contract that is partially financed by the Spanish Ministry of Education and by funds provided to the IRB by the Generalitat de Catalunya.

Supporting Information available. This material is available free of charge via the Internet at http://pubs.acs.org

REFERENCES

(1) Giot, L.; Bader, J.S.; Brouwer, C.; Chaudhuri, A.; Luang, B.; Li, Y. *Science* **2003**, *302*, 1727-1736.

(2) Robinson, C. V.; Sali, A.; Baumeister, W. *Nature* **2007,** *450*, 973-982.

(3) Zuiderweg, E. R. *Biochemistry* **2002**, *41*, 1–7.

(4) Bonvin, A. M.; Boelens, R.; Kaptein, R. *Curr. Opin. Chem. Biol.* **2005**, *9*, 501–508.

(5) Tang, C.; Ghirlando, R.; Clore, G. M. *J. Am. Chem. Soc.* **2008**, *130*, 4048-4056

(6) Koch, M. H. J.; Vachette, P.; Svergun, D. I. *Q. Rev. Biophys.* **2003**, *36*, 147-227.

(7) Svergun, D. I.; Koch, M. H. J. *Rep. Prog. Phys.* **2003**, *66*, 1735-1782.

(8) Svergun, D. I.; Koch, M. H. J. *Curr. Op. Struct. Biol.* **2002**, *12*, 654-660.

(9) Petoukhov, M. V.; Svergun, D. I. *Curr. Op. Struct. Biol.* **2007**, *17*, 562-571.

(10) Konarev, P. V.; Volkov, V. V.; Sokolova, A. V.; Koch, M.H. J.; Svergun, D.I. *J. Appl. Crystallogr.* **32003**, *36*, 1277–1282.

(11) Vestergaard, B.; Groenning, M.; Roessle, M.; Kastrup, J. S.; van de Weert, M.; Flink, J. M.; Frokjaer, S.; Gajhede, M.; Svergun, D. I. *PLoS Biology* **2007**, *5*, 1089-1097.

(12) Bernadó, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. *J. Am. Chem. Soc.* **2007**, *129*, 5656-5674.

(13) Xu, X.; Reinle, W.; Hannemann, F.; Konarev, P. V.; Svergun, D. I.; Bernhardt, R.; Ubbink, M. *J. Am. Chem. Soc.* **2008**, *130*, 6395-6403.

(14) Chen, L.; Hodgson, K. O.; Doniach, S. *J. Mol. Biol.* **1996**, *261*, 658-671.

(15) Segel, D. J.; Fink, A. L.; Hodgson, K. O.; Doniach, S. *Biochemistry* **1998**, *37*, 12443-12451.

(16) Doniach, S. *Chem. Rev.* **2001**, *101*, 1763-1778.

(17) Akiyama, S.; Takahashi, S.; Kimura, T.; Ishimori, K.; Morishima, I.; Nishikawa, Y.; Fujisawa, T. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1329-1334.

(18) Bernadó, P.; Åkerud, T.; de la Torre, J. G.; Akke, M.; Pons, M. *J. Am. Chem. Soc.* **2003**, *125*, 916-923.

(19) Blobel, J.; Schmidl, S.; Vidal, D.; Nisius, L.; Bernadó, P.; Millet, O.; Brunner, E.; Pons, M. *J. Am. Chem. Soc.* **2007**, *129*, 5946-5953.

(20) Tabernero, L.; Evans, B. N.; Tishmack, P. A.; Van Etten, R. L.; Stauffacher, C. V. *Biochemistry* **1999**, *38*, 11651-11658.

(21) Zhang, M.; Van Etten, R. L.; Stauffacher, C. V. *Biochemistry* **1994**, *33*, 11097-11105.

(22) Williamson, T. E.; Craig, B. A.; Kondrashkina, E.; Bailey-Kellogg, C.; Friedman, A. M. *Biophys. J.* **2008**, 4906-4923.

(23) Svergun, D. I. *J. Appl. Crystallogr.* **1992**, *25*, 495–503.

(24) Joliffe, I. T. *Principal Component Analysis*; Springer: New York, 2002.

(25) Tauler, R.; Smilde, A.K.; Kowalski, B. R. *J. Chemometr.* **1995**, *9*, 31-58.

(26) Tauler, R. *Chemometrics and Intelligent Laboratory Systems* **1995**, *30*,133-146.

(27) Jaumot, J.; Gargallo, R.; de Juan, A.; Tauler, R. *Chemometrics and Intelligent Laboratory Systems* **2005**, *76*, 101-110.

(28) de Juan, A.; Tauler, R. *Analytica Chimica Acta* **2003**, *500*, 195-210.

(29) de Juan, A.; Tauler, R. *Crit. Rev. Anal. Chem.* **2006**, *36*, 163-176.

(30) Svergun, D. I.; Barberato, C.; Koch, M. H. J. *J. Appl. Crystallogr.* **1995**, *28*, 768–773.

(31) Svergun, D. I. *Biophys. J.* **1999**, *76*, 2879–2886.

(32) Åkerud, T.; Thulin, E.; Van Etten, R. L.; Akke, M. *J. Mol. Biol.* **2002**, *322*, 137-152.

(33) Tabernero, L.; Aricescu, A. R.; Jones, E. Y.; Szedlacsek, S. E. *FEBS* **2008**, *275*, 867-882.

(34) Aloy, P.; Russell, R.B. *Nature Biotechnol.* **2004**, *22*, 1317-1321.

(35) Wo, Y. Y.; Zhou, M. M.; Stevis, P.; Davis, J. P.; Zhang, Z. Y.; van Etten, R. L. *Biochemistry* **1992**, *31*, 1712-1721.

(36) Roessle, M. W.; Klaering, R.; Ristau, U.; Robrahn, B.; Jahn, D.; Gehrmann, T. *J. Appl. Crystallogr.* **2007**, *40*, 190–194.

(37) Guinier, A. *Ann. Phys. (Paris)* **1939**, *12*, 161–237.

(38) Golub, G. H.; Van Loan, Ch. F. *Matrix Computations, 2nd Ed.*; John Hopkins Univ.Press: London, 1989.

(39) Diework, J.; de Juan, A.; Marcel, M.; Tauler, R.; Lendl, B. *Analyt. Chem.* **2003**, *76*, 641-647.

(40) Diework, J.; de Juan, A.; Tauler, R.; Lendl, B. *Appl. Spectr.* **2002**, *56*, 40-50.

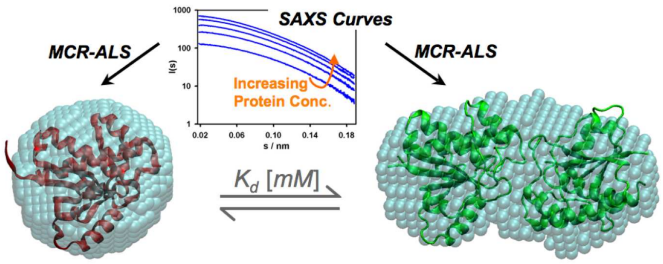
(41) de Juan, A.; Maeder, M.; Martínez, M.; Tauler, R. *Chemometr. Intell. Lab. Syst.* **2000**, *54*, 123-141.

(42) Konarev, P. V.; Petoukhov, M. V.; Volkov, V. V.; Svergun, D. I. *J. Appl. Crystallogr.* **2006**, *39*, 277-286.

(43) Gabb, H. A.; Jackson, R. M.; Sternberg, M. J. E. *J. Mol. Biol.* **1997**, *272*, 106-120.

(44) Kozin, M. B.; Svergun, D. I. *J. Appl. Crystallogr.* **2001**, *34*, 33–41.

TOC Graphic


SAXS Curves
MCR-ALS
MCR-ALS
Increasing
Protein Conc.
I(s)
s / nm
K_d [mM]