Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Influence of multiple well defined conformations on small-angle scattering of proteins in solution

William T. Heller

Condensed Matter Sciences Division and Center for Structural Molecular Biology, Oak Ridge National Laboratory, USA

Correspondence e-mail: hellerwt@ornl.gov

Received 23 March 2004
Accepted 13 October 2004

A common structural motif for many proteins comprises rigid domains connected by a flexible hinge or linker. The flexibility afforded by these domains is important for proper function and such proteins may be able to adopt more than one conformation in solution under equilibrium conditions. Small-angle scattering of proteins in solution samples all conformations that exist in the sampled volume during the time of the measurement, providing an ensemble-averaged intensity. In this paper, the influence of sampling an ensemble of well defined protein structures on the small-angle solution scattering intensity profile is examined through common analysis methods. Two tests were performed using simulated data: one with the extended and collapsed states of the bilobal calcium-binding protein calmodulin and the second with the catalytic subunit of protein kinase A, which has two globular domains connected by a glycine hinge. In addition to analyzing the simulated data for the radii of gyration R_g, distance distribution function $P(r)$ and particle volume, shape restoration was applied to the simulated data. R_g and $P(r)$ of the ensemble profiles could be easily mistaken for a single intermediate state. The particle volumes and models of the ensemble intensity profiles show that some indication of multiple conformations exists in the case of calmodulin, which manifests an enlarged volume and shapes that are clear superpositions of the conformations used. The effect on the structural parameters and models is much more subtle in the case of the catalytic subunit of protein kinase A. Examples of how noise influences the data and analyses are also presented. These examples demonstrate the loss of the indications of multiple conformations in cases where even broad distributions of structures exist. While the tests using calmodulin show that the ensemble states remain discernible from the other ensembles tested or a single partially collapsed state, the tests performed using the simulated catalytic subunit of protein kinase A with noise added demonstrate that it can mask out the ensemble-dependent effects observed for the noiseless profiles.

1. Introduction

Small-angle scattering (SAS) is experiencing an increase in popularity for the study of proteins and protein complexes in solution. Protein complexes are particularly important in the post-genomic era as researchers try to understand the interactions involved in the molecular machines of life. The development of SAS instrumentation for biology at the various synchrotron sources, such as those at the Advanced Photon Source (Irving *et al.*, 2000) and SSRL (Wakatsuki *et al.*, 1992), has opened new opportunities in both high-throughput and time-resolved SAS experiments. The X-ray scattering facilities are complemented by several small-angle neutron

scattering facilities well suited to the study of biological systems, including those at the National Institute of Standards and Technology (Hammouda *et al.*, 1996) and the Institut Max von Laue–Paul Langevin (Ibel, 1976) and the instruments being constructed at the High Flux Isotope Reactor (Lynn *et al.*, 2003) and the Spallation Neutron Source of Oak Ridge National Laboratory. The new instruments available to the user community provide more opportunities for studying proteins and protein complexes by SAS. Advances in the analysis of small-angle scattering data for the shape of the scattering particle (Chacón *et al.*, 1998, 2000; Svergun, 1999; Svergun *et al.*, 1997, 2001; Walther *et al.*, 2000; Heller *et al.*, 2003), which promise higher resolution structural information than is provided by simple Guinier (1939) and distance distribution function [$P(r)$] analyses (Moore, 1980; Svergun *et al.*, 1988), have also generated interest in the technique among members of the structural biology community who do not specialize in scattering techniques.

It is understood how non-ideal behavior of proteins in solution, particularly those arising from aggregation and interparticle interference, influences SAS data. Less well explored is how multiple well defined protein conformations in the sample volume influence the scattering data. Understanding such effects becomes important when the existence of multiple conformations is not known at the beginning of the experiment. As expectations increase for extracting structural information from SAS data, the user must understand how inherent flexibility and multiple conformations can influence the data and analyses, particularly in cases where a protein is expected to have a single well defined conformation.

Flexibility in some portion of a protein's structure is vital to the function of a wide variety of proteins and is capable of producing large-scale domain reorganizations. Some of the more striking examples are the collapse of the calcium-binding proteins calmodulin (Heidorn *et al.*, 1989; Yoshino *et al.*, 1989; Kataoka *et al.*, 1989, 1991; Trewhella *et al.*, 1990) and troponin C (Heidorn *et al.*, 1989; Blechner *et al.*, 1992) about peptides. Other examples include ligand-induced conformational changes in cGMP-dependent protein kinase (Wall *et al.*, 2003), transferrin (Grossmann *et al.*, 1998) and creatine kinase (Forstner *et al.*, 1996). These systems may also display flexibility in the absence of the trigger, such as the binding of a ligand, that causes the conformational change.

Small-angle scattering has been applied to the denaturation and unfolding of proteins (Damaschun *et al.*, 1991; Sosnick & Trewhella, 1992; Chen *et al.*, 1996; Garcia *et al.*, 2001; Choy *et al.*, 2002). Such studies, which also sample multiple conformations of proteins, track changes in the gross structural features of the unfolded protein. An interesting example is the work of Segel *et al.* (1998), in which the data were analyzed in terms of three specific states of cytochrome *c*: native, partially unfolded and fully unfolded. The data were analyzed for the radius of gyration R_g and the decay exponent of the intensity profile at high q. In the case of studying changes in R_g, the parameter represents the average of the structures present in solution and its dependence on the population of structures is understood. The idea that higher order information can be extracted from the data was rejected because of the known conformational disorder of the proteins.

In this paper, SAS data simulated from ensembles of protein structures are analyzed for R_g, $P(r)$, the maximum linear dimension D_{max} and the particle volume. Shape restoration is also applied to the intensity profiles to develop low-resolution models of the structures. The states considered are those in which large segments of the structure remain static but their relative positions vary in a well defined manner. The first test employs ensembles of the extended and collapsed states of calmodulin (CaM), a small protein having two globular domains connected by a flexible helical linker. The high-resolution structures of the extended (Taylor *et al.*, 1991) and collapsed (Ikura *et al.*, 1992) states of the protein have been solved. As expected, the analysis of the simulated data produced for this study of the mixed states of CaM shows a particle that is clearly an intermediate to the fully extended and collapsed states. Shape restoration applied to the intensity profiles shows structures with characteristics of both of the original states. The second test system studied is the catalytic subunit of protein kinase A (cPKA), the crystal structure of which shows a cleft between two globular domains (Knighton *et al.*, 1991). Distributions of cleft openings about a previously identified glycine hinge (Olah *et al.*, 1993) were used to develop averaged SAS intensity profiles. Analysis of the data indicates that even for wide ranges of hinge openings, the effect on the data and low-resolution analyses is subtle. Modeling of the low-resolution structures from the intensity profiles of the ensemble of states also provides little indication that multiple conformations are present. Both sets of results show that careful analysis of protein SAS data can provide some indication of multiple conformations in solution in cases where there are large differences between the structures in the ensemble, but it can be difficult to discern when the structures are similar. The effect of noise on the analysis of data from ensembles of structures was also investigated for both proteins. While the states of CaM are significantly distinct to be discernible even with noise, the addition of noise to the intensity profiles generated for cPKA results in a loss of the subtle distinctions between the static structure and ensemble intensity profiles.

2. Methods

2.1. Theory

The small-angle scattering intensity $I(q)$ of a protein with a single conformation in solution is

$$I(q) = \left| \left\langle \int_V [\rho(\mathbf{r}) - \rho_s] \exp(-i\mathbf{q}\cdot\mathbf{r})\, \mathrm{d}^3 r \right\rangle \right|^2, \qquad (1)$$

where $\rho(\mathbf{r})$ is the scattering length density of the protein, ρ_s is the average scattering length density of the solvent and $\mathbf{q}$ is the momentum transfer, having magnitude $q = (4\pi \sin\theta)/\lambda$, where 2θ is the scattering angle and λ is the wavelength. The integration over the protein volume is both time and rotationally

averaged because the scattering particles are free to diffuse in the solution.

Suppose that the protein being studied can adopt multiple conformations in solution. If each conformational state of the protein is denoted as φ, then we define $f(\varphi)$ to be the frequency of the conformational states of the protein. $f(\varphi)$ is normalized such that $\int f(\varphi)\,\mathrm{d}\varphi = N$, where N is the number of particles in the sampled volume. The scattering-length density of a conformational state φ is written as $\rho(\mathbf{r}, \varphi)$. The total scattered intensity of the ensemble in the sample is then

$$I_{\mathrm{tot}}(q) = \int \left| \left\langle \int_V [\rho(\mathbf{r}, \varphi) - \rho_{\mathrm{s}}] \exp(-i\mathbf{q}\cdot\mathbf{r})\,\mathrm{d}^3 r \right\rangle \right|^2 f(\varphi)\,\mathrm{d}\varphi. \quad (2)$$

Note that the forward scatter, $I_{\mathrm{tot}}(0)$, which depends only on the total excess electron density of the particles in solution, is not influenced by the existence of multiple well defined conformations of the same protein.

2.2. Calculation of the model intensity profiles

The 4cln (Taylor *et al.*, 1991) and 2bbm (Ikura *et al.*, 1992) high-resolution structures of CaM from *Drosophila melanogaster*, representing the extended and collapsed states, respectively, were employed in the first set of tests. The 2bbm structure contains a peptide, which the lobes of CaM wrap around. For this work, the peptide was removed from the structure. The 2cpk crystal structure of the catalytic subunit of protein kinase A (cPKA) in the closed state (Knighton *et al.*, 1991) served as the basis for the model structures for the second set of tests. The program *CRYSOL* (Svergun *et al.*, 1995) was used to calculate the intensity of each model structure used in the testing. The default options were employed for the calculations. Simulated experimental uncertainties were produced for the averaged intensities in proportion to the square root of the ensemble-averaged intensity for use with the fitting routines.

Solution studies of CaM (Heidorn & Trewhella, 1988; Barberato *et al.*, 1992) demonstrated that the conformation of isolated CaM in solution is not as extended as the 4cln crystal structure (Taylor *et al.*, 1991), suggesting that the linker region between the calcium-binding domains is flexible in solution. However, it is assumed here that the protein adopts only two states: either fully extended or fully collapsed. These two conformations are very distinct and were chosen for the sake of simplicity, as performing the simulations necessary to generate the structures intermediate to the end points would be prohibitive. In this case, the ensemble distribution function $f(\varphi)$ is a pair of delta functions representing the two structures. Five different weightings of the states were used for this test: 100% collapsed, 75% collapsed/25% extended, 50% collapsed/50% extended, 25% collapsed/75% extended and 100% extended.

The 2cpk (Knighton *et al.*, 1991) high-resolution structure shows cPKA to be a simple globular protein when the catalytic cleft is closed. The protein contains two globular domains connected by a flexible glycine hinge that opens to expose the catalytic site of the enzyme. To generate the ensemble of structures, the cleft was opened about the glycine hinge identified previously (Olah *et al.*, 1993) in 1° increments from the fully closed state to a maximum opening of 90°. The integral defined in (2) was then approximated as a summation. Again, five angular distributions were tested. The static 45° opened structure served as the basis for the remaining tests. The uniform distribution over all angles was another test performed. Additionally, three Gaussian distributions centered at 45°, with standard deviations equal to 10°, 20° and 30°, were tested. As the wider standard deviations extended significantly beyond the range of angles of interest (0–90°), $f(\varphi)$ was renormalized.

To test the sensitivity of the analysis of the ensemble-averaged intensity profiles to noise, noise was added to the 50% collapsed/50% extended CaM intensity profiles and to the profiles generated from the cPKA ensembles. The noise was added on the basis of the simulated uncertainty $\sigma(q)$ at the given point. A random noise value was picked from the range $[-4\sigma(q), 4\sigma(q)]$. A Gaussian function having a width $2\sigma(q)$ was evaluated using the random noise value. The increased width was used to ensure that sufficient noise was added to the profiles. A random number between 0 and 1 was then generated. The noise value was kept if the random number was less than the value of the Gaussian function evaluated at the random noise value. Random noise values were generated at each point until a value was found that satisfied the above condition.

2.3. Data analysis

Model intensities calculated for the test cases were subjected to analyses that are commonly applied to SAS data of proteins in solution. First, the curves were analyzed for R_{g} (Guinier, 1939) and distance distribution function $P(r)$. Both the algorithm of Moore (1980) and that implemented in *GNOM* (Svergun *et al.*, 1988) were used to determine $P(r)$. The implementation of the Moore algorithm employed uses a $1/q^4$ extrapolation at high q values to provide an estimate of the volume of the scattering particle from the Porod (1951) invariant. This approach is preferable to calculating the volume from the forward scatter because the forward scatter is unaltered by the existence of multiple well defined conformations of a single protein. $P(r)$ analysis by *GNOM* is required for modeling the low-resolution shape of the protein by *DAMMIN* (Svergun, 1999). To provide a secondary test of the shape restoration, the program *GA_STRUCT* (Heller *et al.*, 2003), which determines an average structure from a set of independent modeling runs, was also used.

3. Results

3.1. Calmodulin

The simulated intensity profiles of the five states of CaM are shown in Fig. 1(*a*). The difference between the profiles is pronounced, as expected. Fig. 1(*b*) is a Guinier plot {$\ln[I(q)]$ *versus* q^2} of the low-q region of the intensity profiles. The Guinier regions of the intensity profiles are linear. R_{g} and

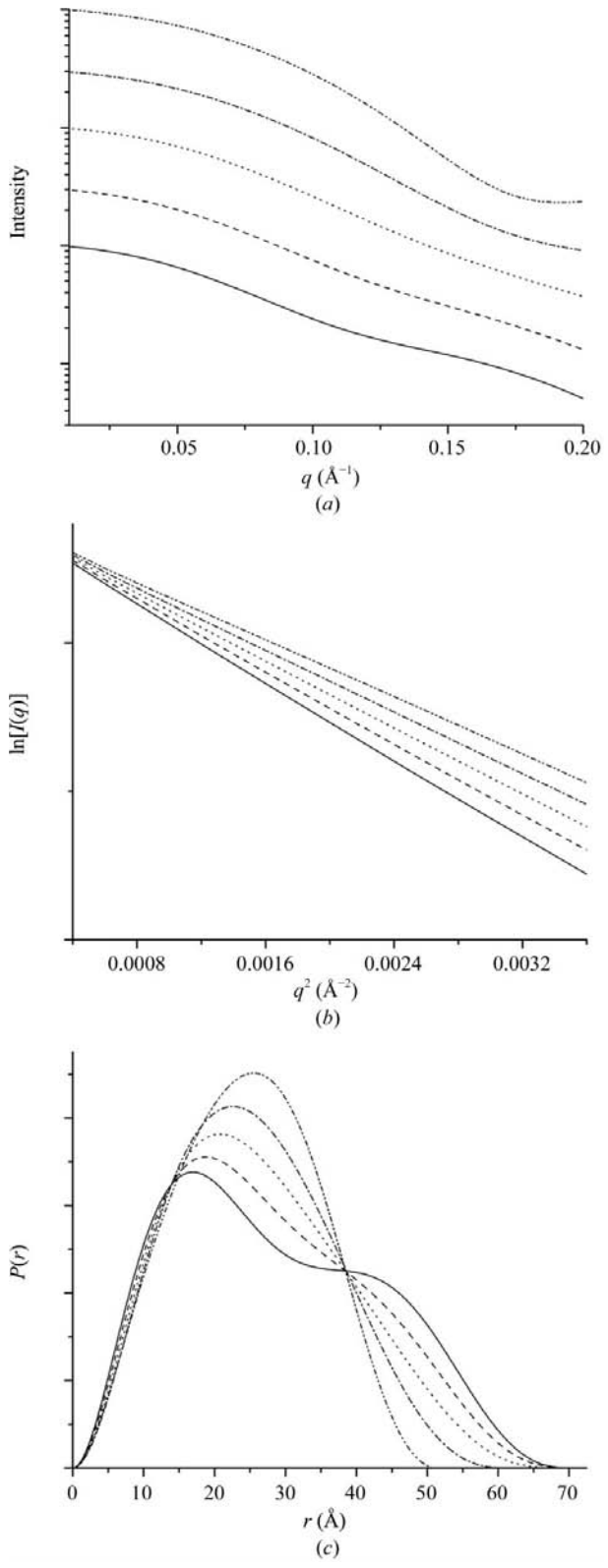

Figure 1
(*a*) Intensity profiles, (*b*) Guinier regions and (*c*) $P(r)$ of the mixed states of extended and collapsed CaM are shown. The line style in each plot corresponds to the same state: fully extended (solid line), 75% extended/25% collapsed (dashed line), 50% extended/50% collapsed (dotted line), 25% extended/75% collapsed (dash/dotted line) and fully collapsed (dash/dot/dotted line). The curves are offset for clarity.

D_{max} determined from the intensity profiles are listed in Table 1 with the volumes determined from the $P(r)$ fitting. The $P(r)$ curves derived from the intensity profiles are plotted in Fig. 1(*c*). The curves change significantly as the system transitions from fully extended to fully collapsed. R_g decreases with increasing content of the intensity profile of the collapsed state of the protein. D_{max} also decreases, but it retains a significant portion of the length of the extended state even in the 50% extended/50% collapsed ensemble. The volume determined from the $P(r)$ fitting of the intensity profile of the 25% extended/75% collapsed ensemble is 32% larger than the fully extended protein. The volume does not peak at the 50% extended/50% collapsed ensemble. The volume estimated from the intensity profile of the fully collapsed state of CaM is also larger than that of the fully extended state. In all cases, the volume determined from the $P(r)$ fitting is larger than the volume expected on the basis of the molecular weight ($\sim$20 000 Å^3).

Low-resolution models of the mixed states of CaM produced by *DAMMIN* and *GA_STRUCT* are shown in Figs. 2 and 3, respectively, docked onto either the extended CaM structure (4cln; Taylor *et al.*, 1991) or the peptide-free collapsed state (2bbm; Ikura *et al.*, 1992). The model resulting from the application of *DAMMIN* to the fully collapsed state agrees very well with the original structure. The fully extended model reproduces the length, aspect ratio and bilobal character of the original structure. Neither model of the collapsed state of CaM reproduces the hole through the center of the structure originally occupied by the peptide that was removed for this study. The intermediate state structures show an increase in bulk in the central region up to the 25% extended/75% collapsed ensemble model, which has a bulge around the linker region of the extended structure. The length of the structure decreases less than was observed for D_{max} in the $P(r)$ fitting. The consensus envelopes generated by *GA_STRUCT* for the fully extended and collapsed states agree with the original structures and the models generated by *DAMMIN*, but are larger, in part as a result of the structural averaging that is an integral part of the modeling process of *GA_STRUCT*. The consensus envelopes produced for the mixed states of CaM clearly show characteristics of both the extended and the collapsed calmodulin, with the middle section of the model growing from a narrow linker to a roughly spherical domain for the 25% extended/75% collapsed ensemble. The extended CaM structure fits inside all of the consensus envelopes except for the model resulting from the fully collapsed state of the protein.

The volumes of the *DAMMIN* models and the *GA_STRUCT* consensus envelopes, also shown in Table 1, increase significantly with increasing content of collapsed state, to a maximum for the fully collapsed state. In all cases, the *DAMMIN* model volumes are larger than the expected volume of CaM and the *GA_STRUCT* consensus envelopes are consistently larger than the *DAMMIN* models. This result can be attributed in part to the structural average that the consensus envelope represents. There is a trend of increasing volume with increasing content of the collapsed state.

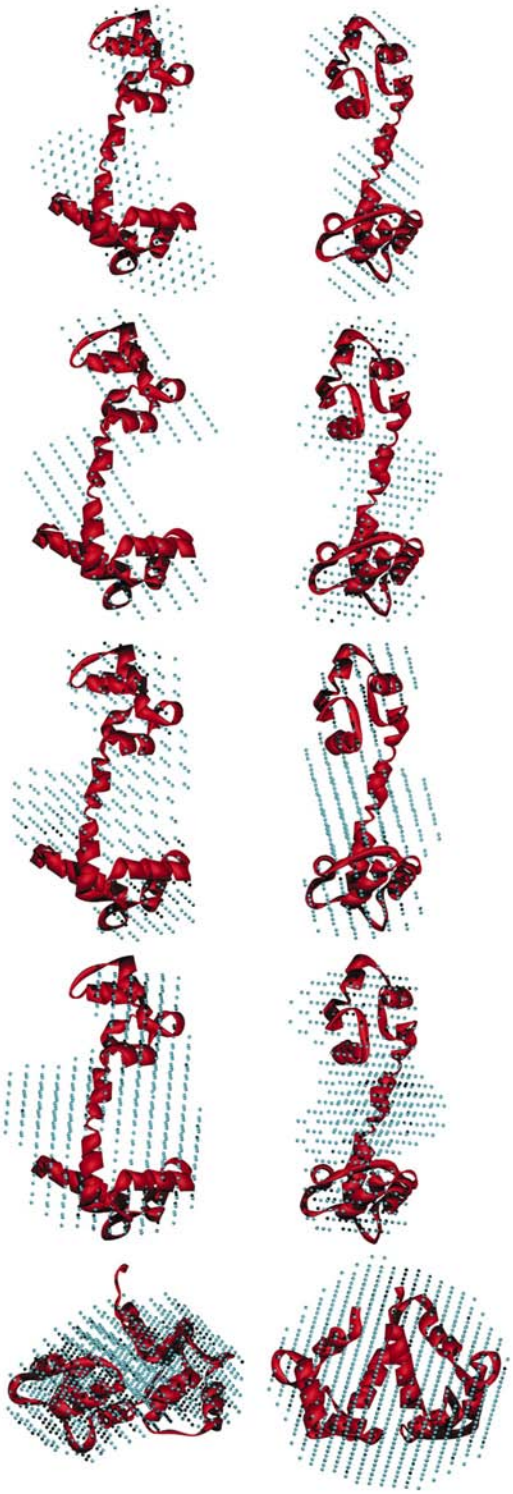

Figure 2
The models produced by *DAMMIN* for CaM for the fully extended, 75% extended/25% collapsed, 50% extended/50% collapsed, 25% extended/75% collapsed and fully collapsed states are shown (shown in order, top to bottom) docked onto the 4cln structure (Taylor *et al.*, 1991), except for the model developed from the fully collapsed state, which is docked onto the 2bbm structure (Ikura *et al.*, 1992). The *DAMMIN* models are shown in cyan and the high-resolution structures are shown in red. Two orthogonal views of each structure are presented (the left and right columns). The images were generated using *ViewerLite* (Accelrys Inc., San Diego, CA, USA) and the *Persistence of Vision Raytracer* (POV-Team; http://www.povray.org/).

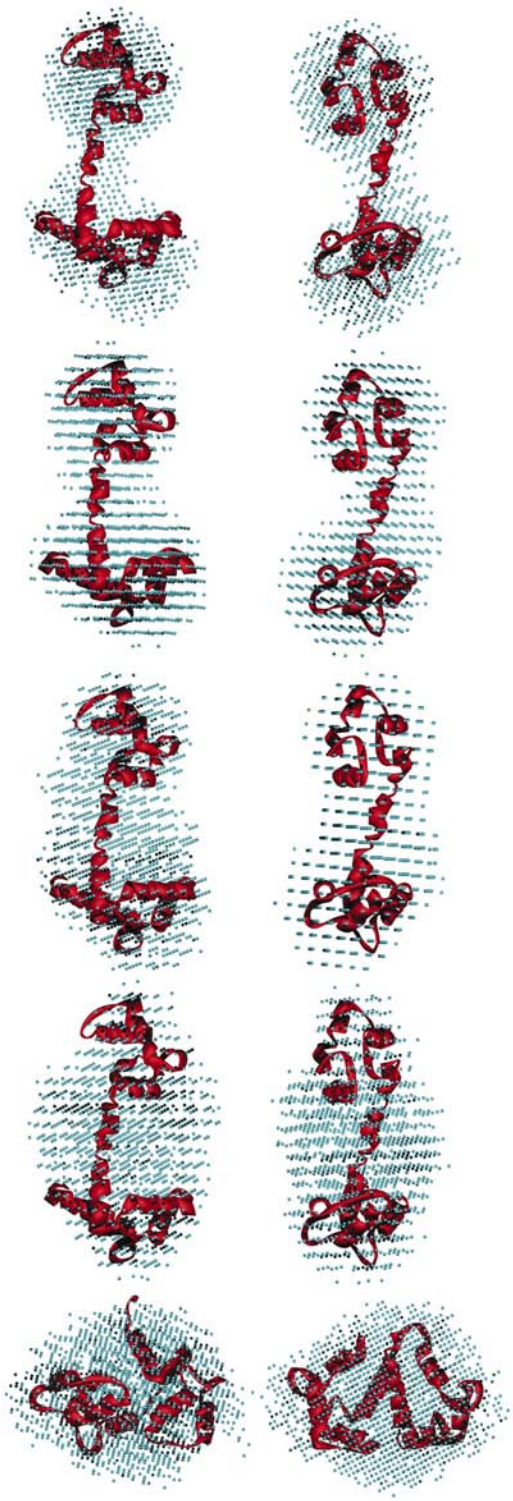

Figure 3
The consensus envelopes produced by *GA_STRUCT* for CaM for the fully extended, 75% extended/25% collapsed, 50% extended/50% collapsed, 25% extended/75% collapsed and fully collapsed states are shown (shown in order, top to bottom) docked onto the 4cln structure (Taylor *et al.*, 1991), except for the model developed from the fully collapsed state, which is docked onto the 2bbm structure (Ikura *et al.*, 1992). The *GA_STRUCT* consensus envelopes are shown in cyan and the high-resolution structures are shown in red. Two orthogonal views of each structure are presented (the left and right columns). The images were generated using *ViewerLite* and the *Persistence of Vision Raytracer*.

Table 1
Structural parameters determined from the intensity profiles simulated for the mixed states of CaM.

R_g was determined according to Guinier (1939). D_{max} and the volume were determined by the Moore (1980) algorithm for determining $P(r)$. The volumes of the models produced by *DAMMIN* and the consensus envelopes produced by *GA_STRUCT* are also provided.

Extended/ collapsed (%)	R_g (Å)	D_{max} (Å)	Volume from $P(r)$ (Å^3)	*DAMMIN* model volume (Å^3)	Consensus envelope volume (Å^3)
100/0	21.2	70	24100	28700	40900
75/25	21.3	69	25800	30300	42600
50/50	20.4	67	27800	33400	49000
25/75	19.4	61	29300	35600	48500
0/100	18.6	51	27200	43200	50400

Table 2
Structural parameters determined from the intensity profiles simulated for the ensembles of cPKA structures.

R_g was determined according to Guinier (1939). D_{max} and the volume were determined by the Moore (1980) algorithm for determining $P(r)$. The volumes of the models produced by *DAMMIN* and the consensus envelopes produced by *GA_STRUCT* are also provided.

Extended/ collapsed (%)	R_g (Å)	D_{max} (Å)	Volume from $P(r)$ (Å^3)	*DAMMIN* model volume (Å^3)	Consensus envelope volume (Å^3)
Static 45 structure	22.50	65	40700	54600	60400
10° std dev.	22.46	66	41000	54100	58500
20° std dev.	22.41	66	41100	54600	58500
30° std dev.	22.37	67	41300	55000	58500
Uniform distribution	22.32	68	41400	53800	60300

Interestingly, the volumes of both sets of models do not follow the trend defined by the volume estimated from the Porod (1951) invariant of intensity profile. The volume estimated in this manner for the fully collapsed state is lower than the two mixed states with the highest collapsed state content. The volumes of the *DAMMIN* models increase monotonically with increasing collapsed state content, while the volumes of the *GA_STRUCT* consensus envelopes roughly level off once the ensemble contains 50% of the collapsed-state intensity profile.

3.2. Catalytic subunit of protein kinase A

The five intensity profiles simulated from the cPKA structure (Knighton *et al.*, 1991) for the various degrees of flexibility about the glycine hinge are shown in Fig. 4(*a*). The differences between the curves are much more subtle than for the mixed states of CaM. The Guinier plots of the low-q regions of the intensity profiles, shown in Fig. 4(*b*), are linear in all cases. Table 2 lists R_g, D_{max} and volume determined from the $P(r)$ fitting with the Moore (1980) algorithm for each test case. The $P(r)$ curves derived from the ensemble intensity profiles are plotted in Fig. 4(*c*). As the width of the distribution of cleft openings increases to the uniform distribution, the value of R_g decreases slightly, while D_{max} increases by 3 Å. The change is more linear than observed in the tests using calmodulin. The volume estimated by the $P(r)$ fitting increases as the breadth of the distribution of cleft openings increases, but in all cases is lower than expected on the basis of the

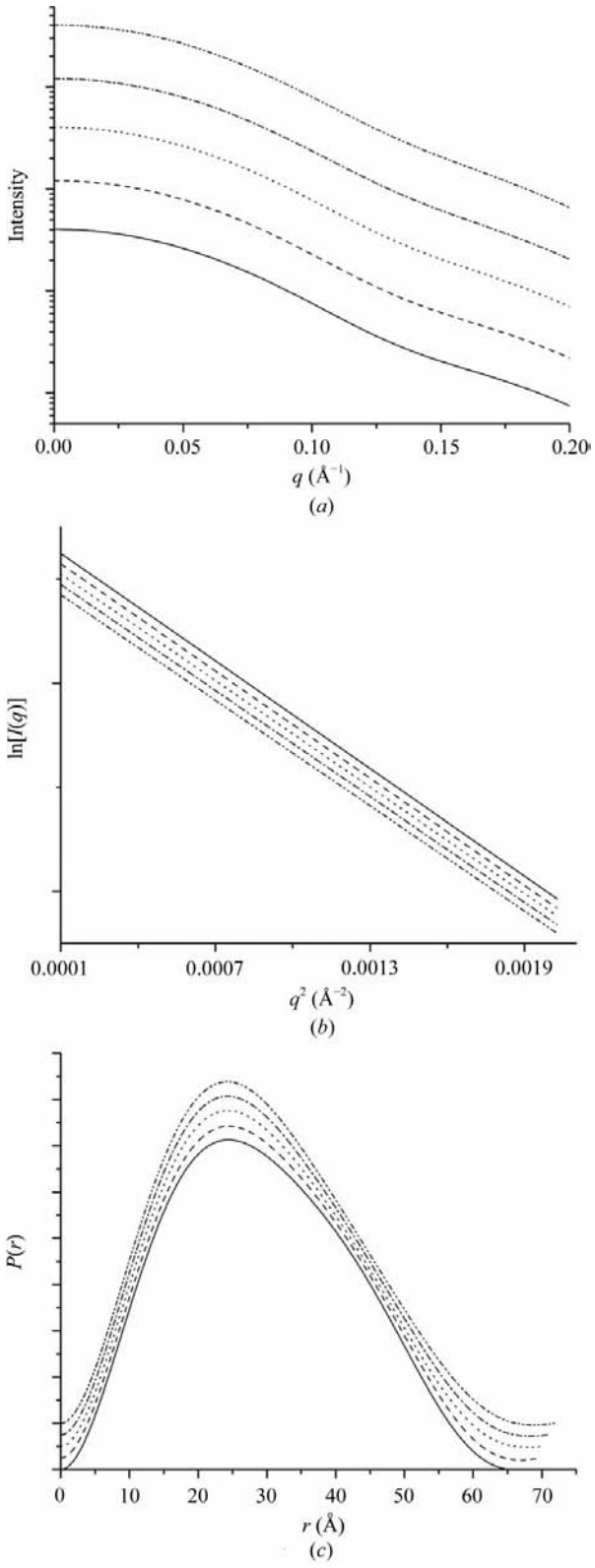

Figure 4
(*a*) Intensity profiles, (*b*) Guinier regions and (*c*) $P(r)$ of the ensembles calculated for the cPKA structure. The line style in each plot corresponds to the same state: static (solid line), 10° standard deviation (dashed line), 20° standard deviation (dotted line), 30° standard deviation (dash/dotted line) and the uniform distribution (dash/dot/dotted line). The curves are offset for clarity.

molecular weight of the protein (~50 200 Å^3); this result may again be a consequence of the extrapolation used to obtain the estimate of the volume. The shape of $P(r)$ changes only slightly, but consistently, with increasing flexibility.

Low-resolution models from the cPKA ensemble intensities were developed using both *DAMMIN* and *GA_STRUCT*. In all tests, the fit to the data of the model intensity profiles generated by the shape-restoration algorithms is excellent (data not shown). The *DAMMIN* models, shown in Fig. 5, do a much better job of reproducing the cleft in the original structure, but consistently do not encompass the helix on the left side of the structures shown in the right column of Fig. 5. In the case of the 20° standard deviation distribution model produced by *DAMMIN*, the structure is somewhat flattened compared with the other structures and this feature was reproduced in independent runs of *DAMMIN*. The volumes of all of the models, shown in Table 2, do not follow the trend from the $P(r)$ analysis. The increase in volume is subtle enough to be lost in the *ab initio* modeling, especially in the case of the consensus envelopes, because of the structural averaging performed.

Much like the models resulting from mixed states of CaM, the *GA_STRUCT* consensus envelopes, shown in Fig. 6, for the different distributions of cleft openings of cPKA change consistently with increasing breadth of the distribution. The models are roughly L-shaped. The helix on the left of the structure, visible in the right column of images, is not encompassed by the models. The models for the broader distributions show a reduction of one leg of the 'L' to a more globular particle. The model produced from the uniform distribution is the clearest example of this effect (Fig. 4, bottom images). The portion of the envelope assigned by the docking to the larger lobe of cPKA is much less pronounced than in the other models. The consensus-envelope volumes, shown in Table 1, do not follow the trend found by the $P(r)$ fitting, which increase only slightly as the breadth of the angular distribution used to calculate the simulated intensity profiles increases.

3.3. Effect of noise

The simulated intensity profiles with noise are plotted in Fig. 7(*a*) with the original noiseless profiles for the 50% extended/50% collapsed CaM and the five profiles generated for cPKA. The $P(r)$ curves determined using the Moore (1980) algorithm are shown in Fig. 7(*b*). The structural parameters determined from the intensity profiles are listed in Table 3. The addition of noise to the 50% extended/50% collapsed CaM intensity profile has not changed R_g, D_{max} or $P(r)$ significantly. In contrast, the addition of noise to the ensemble intensity profiles for cPKA shows that the trend observed in R_g for the noiseless data is washed out by the noise, while that observed in D_{max} is roughly preserved. The $P(r)$ curves no longer follow a clear trend in shape, as found for the noiseless data (Fig. 4*b*). The uncertainties in the R_g values wash out all distinction between the cPKA ensembles. The volumes

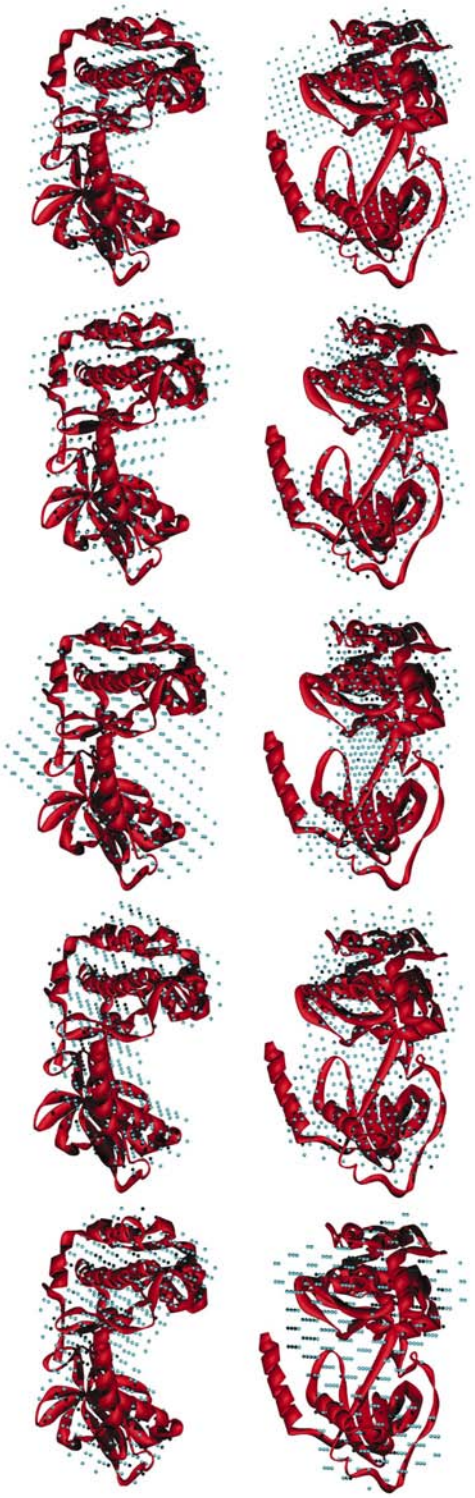

Figure 5
The models produced by *DAMMIN* for cPKA for the static and dynamic structures having Gaussian distributions with 10, 20 and 30° standard deviations and the uniform distribution (in order, top to bottom) docked onto the 2cpk structure. Two orthogonal views of each structure are shown (the left and right columns). The *DAMMIN* models are shown in cyan and the high-resolution structures are shown in red. The structures are not to scale with the images of CaM shown in Fig. 2. The images were generated using *ViewerLite* and the *Persistence of Vision Raytracer*.

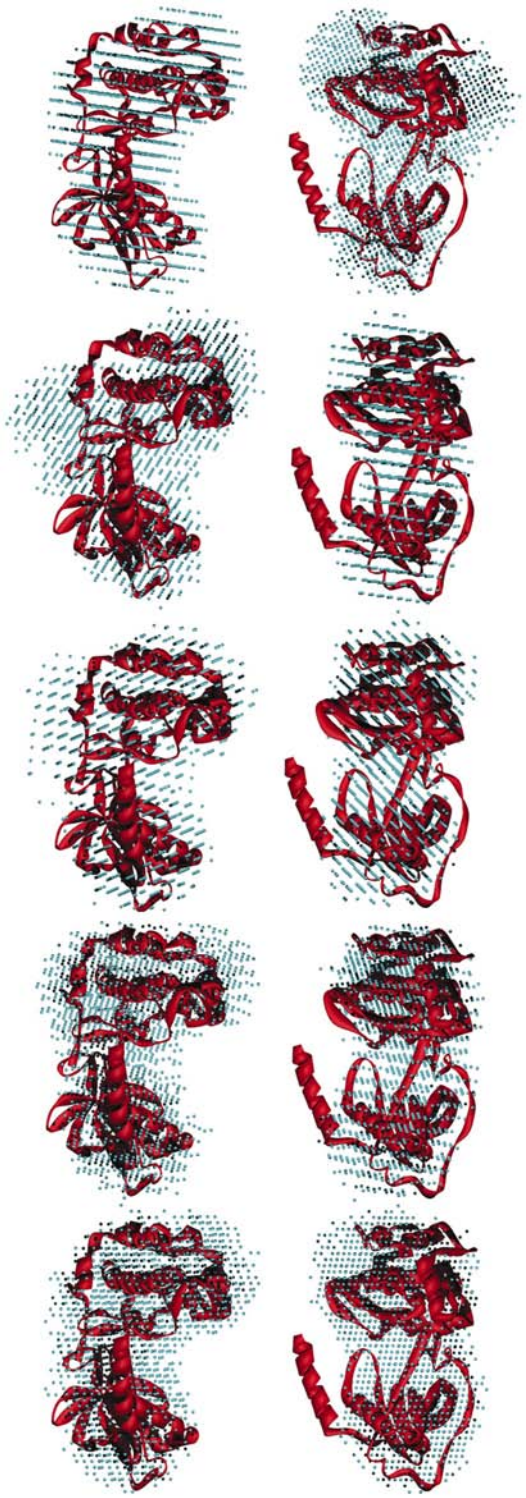

Figure 6
The consensus envelopes produced by *GA_STRUCT* for cPKA for the static and dynamic structures having Gaussian distributions with 10, 20 and 30° standard deviations and the uniform distribution (in order, top to bottom) docked onto the 2cpk structure. Two orthogonal views of each structure are shown (the left and right columns). The *GA_STRUCT* consensus envelopes are shown in cyan and the high-resolution structures are shown in red. The structures are shown to scale with Fig. 5. The images were generated using *ViewerLite* and the *Persistence of Vision Raytracer*.

determined from the $P(r)$ fitting agree with those of the noiseless data to within error.

The models produced for CaM by *DAMMIN* and *GA_STRUCT* are shown in the top and bottom pairs of images of Fig. 8, respectively. The addition of noise has not changed the structure significantly from the models shown in the center pairs of images in Figs. 2 and 3. The *DAMMIN* models (Fig. 9) and *GA_STRUCT* consensus envelopes (Fig. 10) of the cPKA ensembles from the noisy data exhibit more variability than the models generated from the noiseless data, but still do a reasonable job of reproducing the original static cPKA structure. The variability is a manifestation of the slight changes in the shapes of the intensity profiles caused by the noise, which can be most clearly seen in the $P(r)$ plots in Fig. 7(*b*). The consensus envelopes no longer consistently have

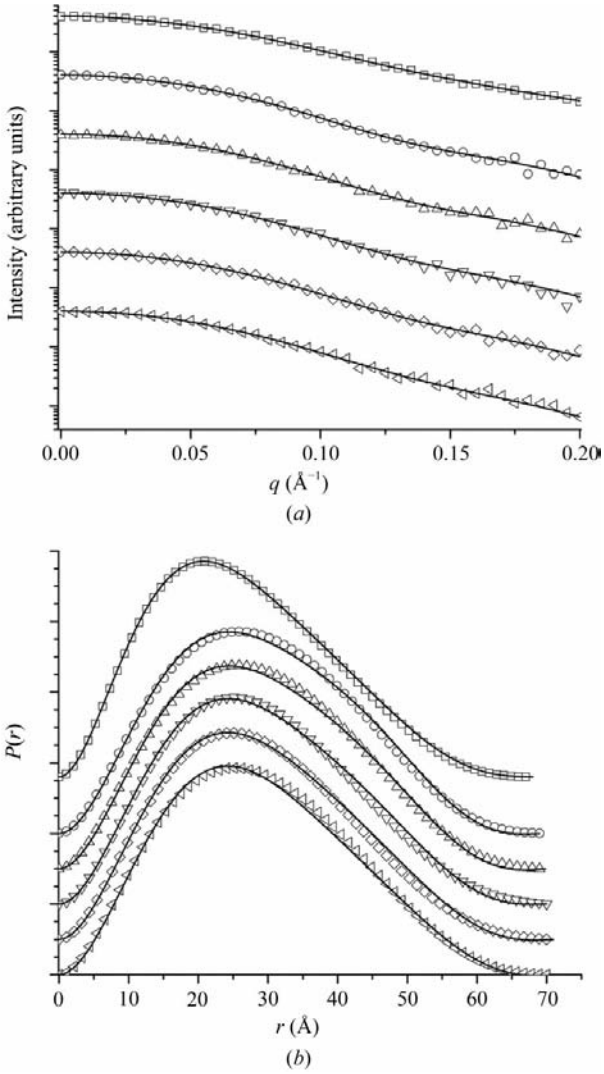

Figure 7
The intensity profiles with added noise (*a*) and associated $P(r)$ (*b*) of the 50% extended/50% collapsed CaM (squares) and the ensembles of cPKA: static (circles), 10° standard deviation (triangles), 20° standard deviation (inverted triangles), 30° standard deviation (diamonds), and the uniform distribution (left-pointing triangles). The solid lines are the noiseless intensity profiles and associated $P(r)$. The curves are offset for clarity.

Table 3
Structural parameters determined from the simulated intensity profiles with noise added.

R_g was determined according to Guinier (1939). D_{max} and the volume were determined by the Moore (1980) algorithm for determining $P(r)$. The volumes of the models produced by *DAMMIN* and the consensus envelopes produced by *GA_STRUCT* are also provided.

Extended/collapsed (%)	R_g (Å)	D_{max} (Å)	Volume from $P(r)$ (Å^3)	*DAMMIN* model volume (Å^3)	Consensus envelope volume (Å^3)
CaM 50% extended/50% collapsed	20.6 ± 0.5	67	27500 ± 700	34300	47100
cPKA static 45 structure	22.30 ± 0.7	69	40600 ± 1700	60600	66700
cPKA 10° std dev.	22.30 ± 0.7	69	40700 ± 1800	59600	58600
cPKA 20° std dev.	22.42 ± 0.7	70	41800 ± 1800	55600	50800
cPKA 30° std dev.	22.51 ± 0.7	70	40700 ± 1700	57700	64900
cPKA uniform distribution	22.28 ± 0.7	70	40100 ± 1600	55500	55100

an L shape. The model volumes, shown in Table 3, roughly agree with those of the noiseless models (Table 1 for CaM and Table 2 for cPKA). In the case of cPKA, there is some variability in the *DAMMIN* model volumes. The volumes of the consensus envelopes vary more than the *DAMMIN* model volumes. There is no longer a trend in the volumes in the $P(r)$ fitting, the *DAMMIN* models or the *GA_STRUCT* consensus envelopes.

4. Conclusions

SAS data resulting from an ensemble of structures can be difficult to identify, depending strongly on the nature of the differences between the states. The extended and collapsed states of CaM are very distinct and this distinction carries over to the intensity profiles. As a result, the mixed states are easy to identify. In contrast to the CaM results, the ensembles of cleft opening of the cPKA structure result in far more subtle differences in the simulated intensity profiles. The structural parameters and $P(r)$ curves extracted from the mixed states do not vary a great deal from the static structure. In both tests performed, the volumes determined from the ensemble intensity profiles increase when multiple conformations are present. The model volumes followed a similar trend in the case of CaM, but the effect is too subtle to be observed in the case of the cPKA models. The addition of noise to the CaM profiles had little effect on the ability to distinguish between the different ensembles. The addition of noise to the cPKA ensemble intensity profiles blurred out the distinguishing characteristics between the various ensembles tested.

While R_g behaves perfectly reasonably for the ensembles of states, D_{max} determined from the $P(r)$ fitting does not behave as expected. One expects that the longest dimension of any particle in solution would be preserved in the data. It can be shown that the $P(r)$ of the ensemble is the population-weighted sum of the individual $P(r)$ curves, making the D_{max} of the ensemble that of the largest particle. This result is not observed in the $P(r)$ fitting by the Moore (1980) algorithm (shown in Table 1) or when *GNOM* is used for the *DAMMIN* modeling (not shown) for either ensemble system studied. In the case of CaM, D_{max} does not decrease significantly until the ensemble intensity profile contains 75% of the collapsed state, at which point D_{max} is at the mid-point of the two extremal states of the protein. The cPKA structure with the cleft opened 90° has a maximum length of 79 Å, yet none of the ensemble $P(r)$ curves extend this far. This effect remains when noise is added to the data. There is some subjectivity in fitting SAS data for $P(r)$, because one does not normally know D_{max} at the beginning of an experiment. As Moore (1980) suggests, the resulting $P(r)$ fitting normally provides a lower bound for D_{max}. In the present study, attempts to fit the mixed state data with the Moore algorithm using D_{max} of the fully extended state resulted in negative amplitudes at the longest vector lengths and this result is interpreted as arising from a D_{max} that is too long for the particle measured. *GNOM* also indicated that the shorter D_{max} values were better solutions. These results suggest that $P(r)$ and D_{max} determined from the experimental data using indirect transform methods may not accurately reproduce the longest dimension present in an

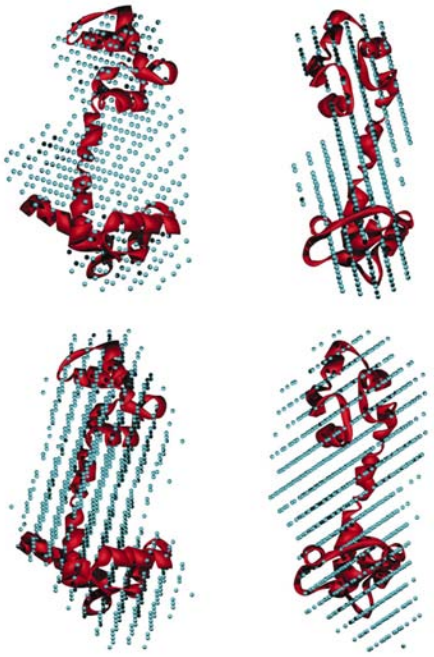

Figure 8
The models produced by *DAMMIN* (top images) and *GA_STRUCT* (bottom images) for CaM for the 50% extended/50% collapsed intensity profile with noise added docked onto the 4cln structure (Taylor *et al.*, 1991). The low-resolution models are shown in cyan and the high-resolution structures are shown in red. Two orthogonal views of each structure are presented (the left and right columns). The images were generated using *ViewerLite* and the *Persistence of Vision Raytracer*.

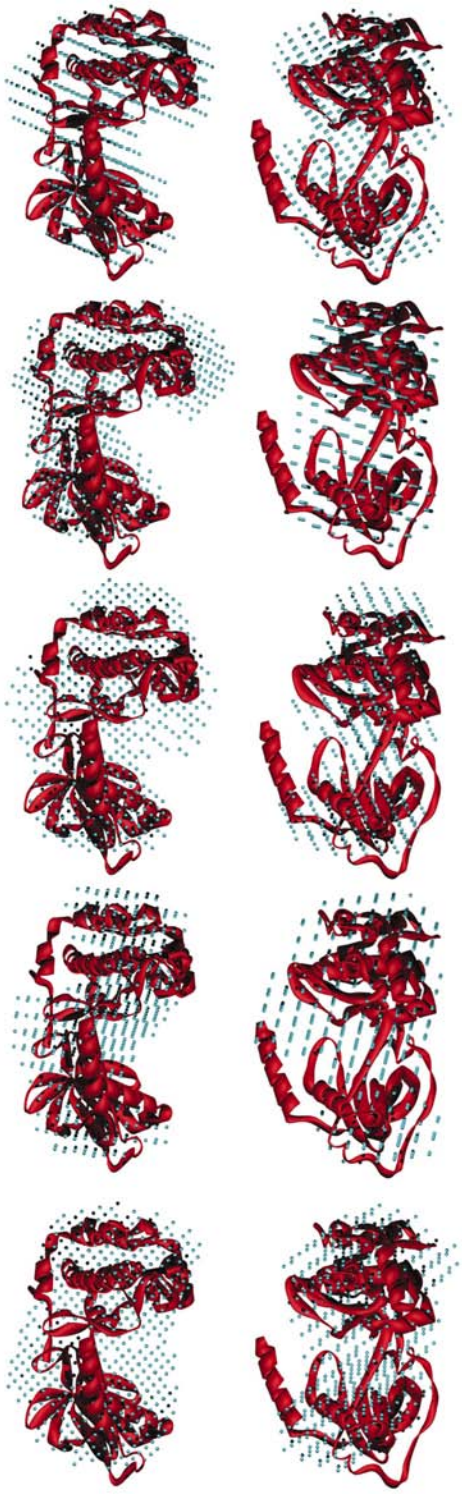

Figure 9
The models produced by *DAMMIN* for the cPKA with added noise for the static and dynamic structures having Gaussian distributions with 10, 20 and 30° standard deviations and the uniform distribution (in order, top to bottom) docked onto the 2cpk structure. Two orthogonal views of each structure are shown (the left and right columns). The *DAMMIN* models are shown in cyan and the high-resolution structures are shown in red. The structures are to scale with Fig. 5. The images were generated using *ViewerLite* and the *Persistence of Vision Raytracer*.

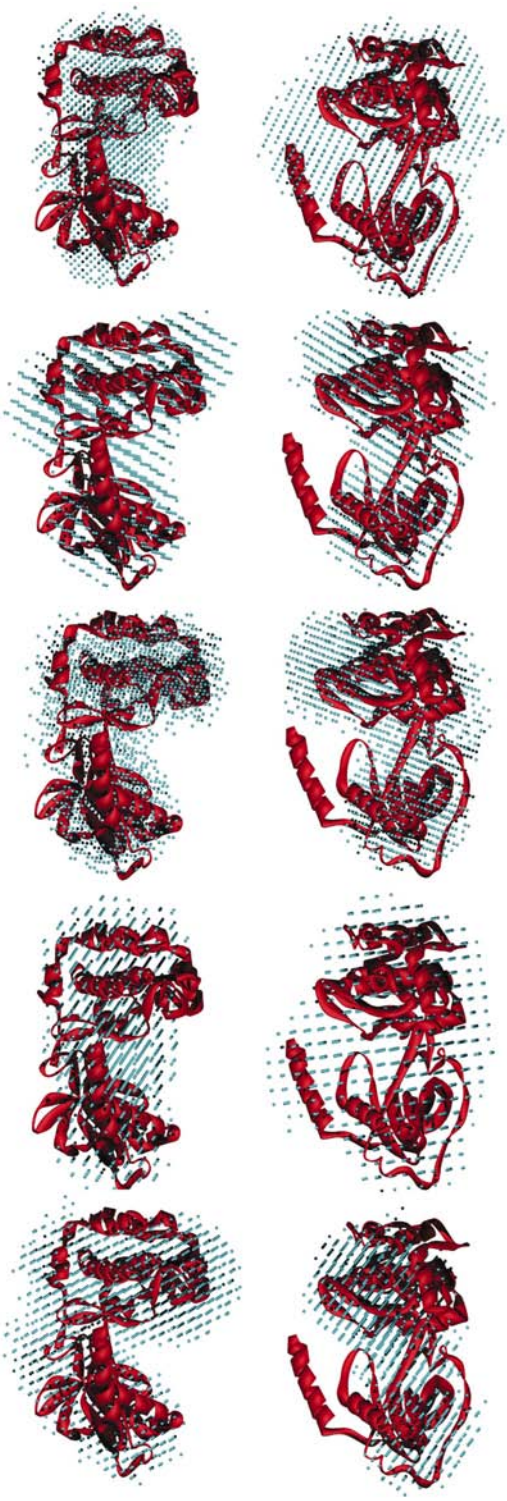

Figure 10
The consensus envelopes produced by *GA_STRUCT* for cPKA with added noise for the static and dynamic structures having Gaussian distributions with 10, 20 and 30° standard deviations and the uniform distribution (in order, top to bottom) docked onto the 2cpk structure. Two orthogonal views of each structure are shown (the left and right columns). The *GA_STRUCT* consensus envelopes are shown in cyan and the high-resolution structures are shown in red. The structures are to scale with those shown in Fig. 5. The images were generated using *ViewerLite* and the *Persistence of Vision Raytracer*.

ensemble of structures. This effect would make it even more difficult to discern that multiple conformations of a protein are present in the sample.

The utility of *ab initio* modeling and volume determination from the $P(r)$ fitting for differentiating static structures from dynamic structures depends strongly on the extent of the variability of the ensemble. The CaM ensembles with or without noise studied here can be easily distinguished from each other and the two structures used to generate the ensembles. The volumes determined from the $P(r)$ fitting are also clearly distinguishable. The extent of the differentiation arises from the relatively large differences between the ensembles. The modeling or the data analysis will not enable the user to differentiate between two ensembles that only differ in extended/collapsed content by a few percent, particularly for experimental data. The ensembles of cPKA with and without noise tested here demonstrate that despite large amounts of flexibility, *ab initio* models are of little use in determining whether the sample contains a static or dynamic structure.

The results demonstrate the importance of additional information when interpreting biological SAS data for specific features, particularly from noisy data. The single partially collapsed state of CaM can be easily distinguished from the ensemble of structures using knowledge about the domain structure of CaM. Heidorn & Trewhella (1988) studied the solution structures of CaM and the related calcium-binding protein troponin C and used knowledge of the domain structure of the proteins to develop models of the proteins in solution. R_g and D_{max} of CaM in EGTA at pH 7.4 are consistent with the 25% extended/75% collapsed ensemble studied here. An examination of the $P(r)$ curves shown here and in Fig. 4 of Heidorn & Trewhella (1988) demonstrates that the experimentally measured solution structure retains its clear bilobal character, with a maximum at 18 Å and a shoulder at ∼38 Å. In stark contrast, $P(r)$ of the mixed ensemble studied here is consistent with a globular particle, having a single maximum near 22 Å. In the end, the experimental data were fit well by a bent dumbbell shape, rather than the globular shape found for the 25% extended/75% collapsed state. It is unlikely that any acceptable amount of noise would make it possible to mistake one state for another. The tests with the ensembles of cPKA suggest that it would not have been possible to distinguish a static structure from an ensemble of structures with the experimental data used to identify the opening of the glycine hinge of cPKA (Olah *et al.*, 1993), although the analysis employed in the study utilized the crystal structure of the protein (Knighton *et al.*, 1991). The noise in the data would have masked out any indication of multiple conformations in solution.

The greatest strength of SAS for the study of proteins in solution is that there is no need to crystallize the system. As a result, this approach is applicable to the large class of proteins with structurally disordered segments, such as the linker regions of the regulatory subunit of cyclic AMP-dependent protein kinase A (Li *et al.*, 2000), which connects the dimerization/docking domain to the cyclic AMP-binding domains and makes contacts with the active site of the catalytic domain in the holoenzyme. Such segments often contain important regulatory sequences, such as phosphorylation or ligand-binding sites, which are critical for biological function. Dynamic fluctuations between the structured domains of the proteins would produce the kind of effects on SAS data demonstrated here and careful analysis of the data could prove useful for understanding the extent of the fluctuations and the role that they play in function.

Recently, there has been increased interest in proteins that are intrinsically unstructured (for a review, see Uversky, 2002). The current work suggests that it may be possible to utilize SAS to gain some insight into the extent of the flexibility of the unstructured system. It would also be very useful for observing decreases in flexibility that may result from the binding of small ligands, such as ions, small molecules or possibly peptides. Such systems might be approached in the same manner as denatured proteins (Damaschun *et al.*, 1991; Sosnick & Trewhella, 1992; Chen *et al.*, 1996; Garcia *et al.*, 2001; Choy *et al.*, 2002). Recent work studying protein denaturation with urea by NMR indicates that denatured states are not entirely random (Shortle & Ackerman, 2001; Ackerman & Shortle, 2002*a*, 2002*b*; Ohnishi *et al.*, 2004). As a result, it may be possible to gain meaningful information about the extent of variation of the unfolded state using SAS, whether it be intrinsic or chemically induced. It is in such cases that shape restoration may play an important role in studying the solution structure of the protein and the extent of the structural variation, particularly in conjunction with any high-resolution information that is available for the system.

The results presented here reinforce the idea that complications can arise when interpreting data from SAS intensity profiles of flexible proteins in solution. Proteins that adopt multiple conformations in solution or display well defined kinds of flexibility are difficult to identify on the basis of the small-angle scattering intensity profile. A larger than expected particle volume may be a sign that some flexibility exists between well defined domains, as both tests demonstrate. In cases where the conformational flexibility does not result in large variations in structure, such as the cleft openings of cPKA, the effect is very subtle and is easily masked by noise. Other artifacts, such as systematic errors and uncorrected resolution effects for a particular instrument, may also mask out such effects. Additional sources of information, such as high-resolution structures of subunits or domains and data pertaining to flexibility, are required in order to truly understand the small-angle scattering intensity profiles of flexible proteins in solution. In such cases it may be possible to understand the nature and extent of the motions of the protein in solution. This work also suggests that improved methods for volume estimation from the scattered intensity profile would be extremely beneficial.

This work was supported by the Office of Biological and Environmental Research project KP1101010 of the US Department of Energy under contract No. DE-AC05-

00OR22725 with Oak Ridge National Laboratory, managed and operated by UT-Batelle, LLC. The author thanks Dr William A. Hamilton for comments on the manuscript.

References

Ackerman, M. S. & Shortle, D. (2002*a*). *Biochemistry*, **41**, 3089–3095.
Ackerman, M. S. & Shortle, D. (2002*b*). *Biochemistry*, **41**, 13791–13797.
Barberato, G., Ikura, M., Kay, L. E., Pastor, R. W. & Bax, A. (1992). *Biochemistry*, **31**, 5269–5278.
Blechner, S. L., Olah, G. A., Strynadka, N. C. J., Hodges, R. S. & Trewhella, J. (1992). *Biochemistry*, **31**, 11326–11334.
Chacón, P., Díaz, J. F., Morán, F. & Andreu, J. M. (2000). *J. Mol. Biol.* **299**, 1289–1302.
Chacón, P., Morán, F., Díaz, J. F., Pantos, E. & Andreu, J. M. (1998). *Biophys. J.* **74**, 2760–2775.
Chen, L., Hodgson, K. O. & Doniach, S. (1996). *J. Mol. Biol.* **261**, 658–671.
Choy, W.-Y., Mulder, F. A. A., Crowhurst, K. A., Muhandiram, D. R., Millett, I. S., Doniach, S., Forman-Kay, J. D. & Kay, L. E. (2002). *J. Mol. Biol.* **316**, 101–112.
Damaschun, G., Damaschun, H., Gast, K., Gernat, C. & Zirwer, D. (1991). *Biochim. Biophys. Acta*, **1078**, 289–295.
Forstner, M., Kriechbaum, M., Laggner, P. & Wallimann, T. (1996). *J. Mol. Struct.* **383**, 217–222.
Garcia, P., Serrano, L., Durand, D., Rico, M. & Bruix, M. (2001). *Protein Sci.* **10**, 1100–1112.
Grossmann, J. G., Crawley, J. B., Strange, R. W., Patel, K. J., Murphy, L. M., Neu, M., Evans, R. W. & Hasnain, S. S. (1998). *J. Mol. Biol.* **279**, 461–472.
Guinier, A. (1939). *Ann. Phys. (Paris)*, **12**, 161–237.
Hammouda, B., Barker, J. G. & Krueger, S. (1996). *Small-Angle Neutron Scattering Manuals.* Gaithersburg, MD, USA: National Institute of Standards and Technology.
Heidorn, D. B., Seeger, P. A., Rokop, S. E., Blumenthal, D. K., Means, A. R., Crespi, H. & Trewhella, J. (1989). *Biochemistry*, **28**, 6757–6764.
Heidorn, D. B. & Trewhella, J. (1988). *Biochemistry*, **27**, 909–915.
Heller, W. T., Krueger, J. K. & Trewhella, J. (2003). *Biochemistry*, **42**, 10579–10588.
Ibel, K. (1976). *J. Appl. Cryst.* **9**, 296–309.
Ikura, M., Clore, G. M., Gronenborn, A. M., Zhu, G., Klee, C. B. & Bax, A. (1992). *Science*, **256**, 632–638.
Irving, T. C., Fischetti, R., Rosenbaum, G. & Bunker, G. B. (2000). *Nucl. Instrum. Methods*, **448**, 250–254.
Kataoka, M., Head, J. F., Seaton, B. A. & Engelman, D. M. (1989). *Proc. Natl Acad. Sci. USA*, **86**, 6944–6948.
Kataoka, M., Head, J. F., Vorherr, T., Krebs, J. & Carafoli, E. (1991). *Biochemistry*, **30**, 6247–6251.
Knighton, D. R., Zheng, J., Ten Eyck, L. F., Xuong, N.-H., Taylor, S. S. & Sowadski, J. M. (1991). *Science*, **253**, 407–414.
Li, F., Gangal, M., Jones, J. M., Deich, J., Lovett, K. E., Taylor, S. S. & Johnson, D. A. (2000). *Biochemistry*, **39**, 15626–15632.
Lynn, G. W., Buchanan, M. V., Butler, P. D., Magid, L. J. & Wignall, G. D. (2003). *J. Appl. Cryst.* **36**, 829–831.
Moore, P. B. (1980). *J. Appl. Cryst.* **13**, 168–175.
Ohnishi, S., Lee, A. L., Edgell, M. H. & Shortle, D. (2004). *Biochemistry*, **43**, 4064–4070.
Olah, G. A., Mitchell, R. D., Sosnick, T. R., Walsh, D. A. & Trewhella, J. (1993). *Biochemistry*, **32**, 3649–3657.
Porod, G. (1951). *Kolloid Z. Z. Polym.* **124**, 83–114.
Segel, D. J., Fink, A. L., Hodgson, K. O. & Doniach, S. (1998). *Biochemistry*, **37**, 12443–12451.
Shortle, D. & Ackerman, M. S. (2001). *Science*, **293**, 487–489.
Sosnick, T. R. & Trewhella, J. (1992). *Biochemistry*, **31**, 8329–8335.
Svergun, D. I. (1999). *Biophys. J.* **76**, 2879–2886.
Svergun, D. I., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.
Svergun, D. I., Petoukhov, M. V. & Koch, M. H. J. (2001). *Biophys. J.* **80**, 2946–2953.
Svergun, D. I., Semenyuk, A. V. & Feigin, L. A. (1988). *Acta Cryst.* A**44**, 244–250.
Svergun, D. I., Volkov, V. V., Kozin, M. B., Stuhrmann, H. B., Barberato, C. & Koch, M. H. J. (1997). *J. Appl. Cryst.* **30**, 798–802.
Taylor, D. A., Sack, J. S., Maune, J. F., Beckingham, K. & Quiocho, F. A. (1991). *J. Biol. Chem.* **266**, 21375–21380.
Trewhella, J., Blumenthal, D. K., Rokop, S. E. & Seeger, P. A. (1990). *Biochemistry*, **29**, 9316–9324.
Uversky, V. N. (2002). *Protein Sci.* **11**, 739–756.
Wakatsuki, S., Hodgson, K. O., Eliezer, D. & Rice, M. (1992). *Rev. Sci. Instrum.* **63**, 1736–1740.
Wall, M. E., Francis, S. H., Corbin, J. D., Grimes, K., Richie-Jannetta, R., Kotera, J., Macdonald, B. A., Gibson, R. R. & Trewhella, J. (2003). *Proc. Natl. Acad. Sci. USA*, **100**, 2380–2385.
Walther, D., Cohen, F. E. & Doniach, S. (2000). *J. Appl. Cryst.* **33**, 350–363.
Yoshino, H., Minari, O., Matsushima, N., Ueki, T., Miyake, Y., Matsuo, T. & Izumi, Y. (1989). *J. Biol. Chem.* **264**, 19706–19709.