J|A|C|S
ARTICLES

Published on Web 07/10/2009

Determination of Multicomponent Protein Structures in Solution Using Global Orientation and Shape Restraints

Jinbu Wang,[†] Xiaobing Zuo,[†] Ping Yu,[†,‡] In-Ja L. Byeon,[§] Jinwon Jung,[§] Xiaoxia Wang,[||] Marzena Dyba,[‡,⊥] Soenke Seifert,[#] Charles D. Schwieters,[∇] Jun Qin,[||] Angela M. Gronenborn,[§] and Yun-Xing Wang[*,†]

Protein Nucleic Acid Interaction Section, Biophysics Resource, Structural Biophysics Laboratory, National Cancer Institute at Frederick, National Institutes of Health, Frederick, Maryland 21702, SAIC-Frederick, Inc., National Cancer Institute at Frederick, National Institutes of Health, Frederick, Maryland 21702, Department of Structural Biology, University of Pittsburgh School of Medicine, 1050 BST3, Pittsburgh, Pennsylvania 15261, X-ray Science Division, Advanced Photon Source, Argonne National Laboratory, Argonne, Illinois 60439, Division of Computational Bioscience, Building 12A, Center for Information, Technology, National Institutes of Health, Bethesda, Maryland 20892-5624, and Structural Biology Program, Department of Molecular Cardiology, Lerner Research Institute, NB20, 9500 Euclid Avenue, Cleveland, Ohio 44195

Received April 6, 2009; E-mail: wangyu@ncifcrf.gov

Abstract: Determining architectures of multicomponent proteins or protein complexes in solution is a challenging problem. Here we report a methodology that simultaneously uses residual dipolar couplings (RDC) and the small-angle X-ray scattering (SAXS) restraints to mutually orient subunits and define the global shape of multicomponent proteins and protein complexes. Our methodology is implemented in an efficient algorithm and demonstrated using five examples. First, we demonstrate the general approach with simulated data for the HIV-1 protease, a globular homodimeric protein. Second, we use experimental data to determine the structures of the two-domain proteins L11 and γD-Crystallin, in which the linkers between the domains are relatively rigid. Finally, complexes with K_d values in the high micro- to millimolar range (weakly associating proteins), such as a homodimeric GB1 variant, and with K_d values in the nanomolar range (tightly bound), such as the heterodimeric complex of the ILK ankyrin repeat domain (ARD) and PINCH LIM1 domain, respectively, are evaluated. Furthermore, the proteins or protein complexes that were determined using this method exhibit better solution structures than those obtained by either NMR or X-ray crystallography alone as judged based on the pair-distance distribution functions (PDDF) calculated from experimental SAXS data and back-calculated from the structures.

Introduction

Determining architectures of multicomponent proteins or protein complexes is essential for understanding cellular signaling that involves communication between the domains or subunits. In solution, intermolecular interfaces are determined by nuclear magnetic resonance (NMR) spectroscopy using distance restraints extracted from classical nuclear Overhauser effect (NOE) spectra,[1,2] or from isotope-filtered and edited NOE experiments[3−6] on mixed labeled/nonlabeled samples.[6,7] The latter experiments are less sensitive than their nonfiltered counterparts, with NOEs between domains or subunits difficult to detect and few in numbers, as found for the ARD ILK/PINCH LIM1 complex[8] or not detected at all, as in the protein L11.[9] Furthermore, even in cases with detectable NOEs, their assignment is often challenging and time-consuming. In the best case, even if there are sufficient numbers of NOEs to define interfaces between two components, the global architecture is often underdetermined due to lack of global dimension restraints. The interfaces of multicomponent protein complexes can also be accurately determined using complementary residual dipolar

[†] Protein Nucleic Acid Interaction Section, National Cancer Institute at Frederick.

[‡] SAIC-Frederick, Inc., National Cancer Institute at Frederick.

[§] University of Pittsburgh School of Medicine.

[||] Lerner Research Institute.

[⊥] Structural Biophysics Laboratory, National Cancer Institute at Frederick.

[#] Argonne National Laboratory.

[∇] Division of Computational Bioscience.

[1] Arrowsmith, C. H.; Pachter, R.; Altman, R. B.; Iyer, S. B.; Jardetzky, O. *Biochemistry* **1990**, *29*, 6332–41.

[2] Breg, J. N.; Boelens, R.; George, A. V.; Kaptein, R. *Biochemistry* **1989**, *28*, 9826–33.

[3] Burgering, M. J.; Boelens, R.; Gilbert, D. E.; Breg, J. N.; Knight, K. L.; Sauer, R. T.; Kaptein, R. *Biochemistry* **1994**, *33*, 15036–45.

[4] Clore, G. M.; Appella, E.; Yamada, M.; Matsushima, K.; Gronenborn, A. M. *Biochemistry* **1990**, *29*, 1689–96.

[5] Griffey, R. H.; Redfield, A. G. *Q. Rev. Biophys.* **1987**, *19*, 51–82.

[6] Walters, K. J.; Ferentz, A. E.; Hare, B. J.; Hidalgo, P.; Jasanoff, A.; Matsuo, H.; Wagner, G. *Methods Enzymol.* **2001**, *339*, 238–58.

[7] Venters, R. A.; Huang, C. C.; Farmer, B. T., 2nd; Trolard, R.; Spicer, L. D.; Fierke, C. A. *J. Biomol. NMR* **1995**, *5*, 339–44.

[8] Yang, Y.; Wang, X.; Hawkins, C. A.; Chen, K.; Vaynberg, J.; Mao, X.; Tu, Y.; Zuo, X.; Wang, J.; Wang, Y. X.; Wu, C.; Tjandra, N.; Qin, J. *J. Biol. Chem.* **2008**, *284*, 5836−44.

[9] Lee, D.; Walsh, J. D.; Yu, P.; Markus, M. A.; Choli-Papadopoulou, T.; Schwieters, C. D.; Krueger, S.; Draper, D. E.; Wang, Y. X. *J. Mol. Biol.* **2007**, *367*, 1007–22.

10.1021/ja902528f CCC: $40.75 © 2009 American Chemical Society

coupling (RDC)[10] and intermolecular NOE distance restraints, or chemical shift perturbation aided with computational docking,[11] provided that chemical shift perturbations as a result of protein−protein direct contacts can be distinguished from those that result from structural changes upon bindings. The ambiguity of contact interfaces may be resolved using cross-saturation experiments, as demonstrated in protein−protein complex VDAC-Bcl-x_L.[12]

RDC and SAXS data contain orientation and global shape information about protein structures.[13,14] Both types of data have been used individually to validate relative orientation of domains/subunits in complexes or their global dimension, based on structural models derived from NMR or X-ray crystallography.[15−17] However, for multicomponent proteins or complexes, neither data alone is sufficient to fully validate and determine their structures with respect to unknown relative subunit orientation and unknown architectures, even if coordinates of component structures are known. Indeed, SAXS and small angle neutron scattering (SANS) data together with RDC data have been used successfully to refine known solution NMR structures of single-chain proteins with simulated annealing (SA) protocols.[9,18] In addition, using a MacLaurin series as a target function to simulate the initial part of SAXS data with momentum transfer less than 0.07 $Å^{-1}$ for limited shape discrimination has also been reported.[19] However, presently, we are not aware of any report using experimental SAXS and RDC data and solution NMR-derived component structures for the determination of global architectures of complexes, such as presented here.

For a single set of RDCs and a structural model, four different possible orientations with respect to the alignment tensor are possible. This translates into four unique combinations of relative orientations for a heterodimeric protein or a complex, and three for a homodimeric protein because of restrictions imposed by the C_2 symmetry.[20] A unique combination of relative orientations can be derived via a second set of RDCs in an independent alignment medium.[10] However, the unique relative orientation can also be derived using SAXS data, provided that the subunit structure is sufficiently asymmetric, well-defined and an efficient search algorithm is available. A special case, in which a subunit structure was highly elongated, has been reported for a RNA:RNA complex.[21] Although most protein structures are more globular than nucleic acids, they tend to be mostly asymmetric, which makes it feasible to determine the relative orientation as well as relative position of two components using SAXS. In cases where ambiguity remains due to shape degeneracy, this uncertainty can be resolved with sparse distance information that can readily be obtained by various means. It should be pointed out that without a specific distance restraint(s), the simulated annealing (SA) and molecular dynamics (MD) calculation will not automatically result in the correct architectures of the complex, given a set of discrete orientations and SAXS data restraints. Here, we present an integrated approach that is implemented in the program Global Architecture derived from SAXS and RDC (GASR) and demonstrate the utility of our methodology using five distinct systems.

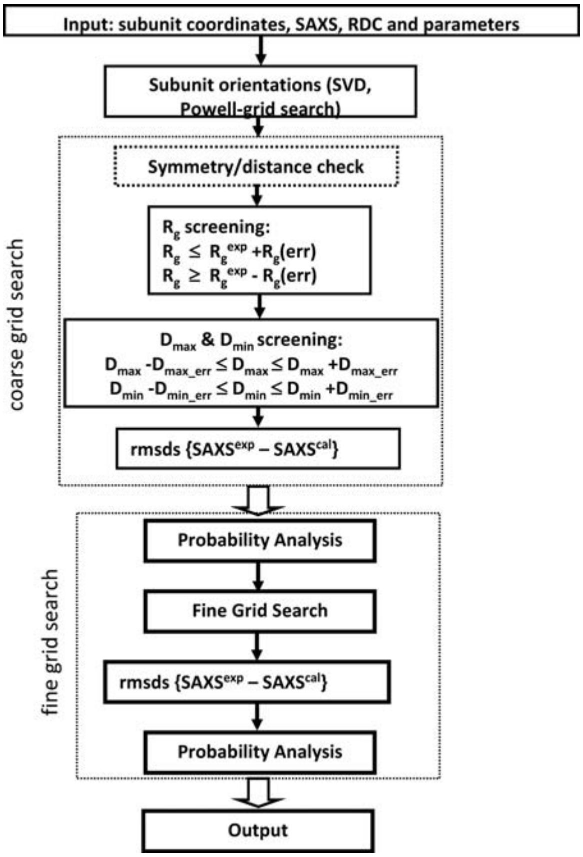
Input: subunit coordinates, SAXS, RDC and parameters
Subunit orientations (SVD, Powell-grid search)
Symmetry/distance check
coarse grid search
R_g screening:
$R_g \le R_g^{exp} + R_g(err)$
$R_g \ge R_g^{exp} - R_g(err)$
D_{max} & D_{min} screening:
$D_{max} - D_{max\_err} \le D_{max} \le D_{max} + D_{max\_err}$
$D_{min} - D_{min\_err} \le D_{min} \le D_{min} + D_{min\_err}$
rmsds {$SAXS^{exp} - SAXS^{cal}$}
fine grid search
Probability Analysis
Fine Grid Search
rmsds {$SAXS^{exp} - SAXS^{cal}$}
Probability Analysis
Output

Figure 1. Flowchart of the GASR program. The individual steps are listed in the boxed regions and the coarse and fine grid searches are enclosed in dotted boxes. R_g(err), D_{max_err} and D_{min_err} are error ranges allowed for R_g, D_{max} and D_{min}, respectively (see the text). $SAXS^{exp}$ and $SAXS^{cal}$ are the experimental and back-calculated SAXS values.

Results

GASR Program. An overview flowchart of the GASR program is shown in Figure 1. The computationally intense core subroutines of the GASR program were written in C and wrapped in Python. A sample input file that was used for calculating the GB1 structure is provided in Supporting Information. The best fit to the scattering data is found using grid searches (Figure 2A) in translational space using spherical polar coordinates (ϕ,θ,r) with components such as subunits or domains being treated as rigid bodies whose relative orientations are fixed, with r as the distance between the centers of masses of the subunits. To increase the computational speed of the search algorithm, the geometric center of one of the subunits (in the case of a heterodimer, the larger subunit) is kept fixed in one of four possible orientations, with the second subunit translated relative to the first on a coarse grid of 1.0 Å for r, and 10° for both ϕ and θ, and later on a fine grid of 1.0 Å and an angular

[10] Fischer, M. W.; Losonczi, J. A.; Weaver, J. L.; Prestegard, J. H. *Biochemistry* **1999**, *38*, 9013–22.

[11] Dominguez, C.; Boelens, R.; Bonvin, A. M. *J. Am. Chem. Soc.* **2003**, *125*, 1731–7.

[12] Malia, T. J.; Wagner, G. *Biochemistry* **2007**, *46*, 514–25.

[13] Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 9279–83.

[14] Koch, M. H.; Vachette, P.; Svergun, D. I. *Q. Rev. Biophys.* **2003**, *36*, 147–227.

[15] Bax, A.; Kontaxis, G.; Tjandra, N. *Methods Enzymol.* **2001**, *339*, 127–74.

[16] Ninio, J.; Luzzati, V.; Yaniv, M. *J. Mol. Biol.* **1972**, *71*, 217–29.

[17] Lattman, E. E. *Proteins* **1989**, *5*, 149–55.

[18] Grishaev, A.; Wu, J.; Trewhella, J.; Bax, A. *J. Am. Chem. Soc.* **2005**, *127*, 16621–8.

[19] Gabel, F.; Simon, B.; Nilges, M.; Petoukhov, M.; Svergun, D.; Sattler, M. *J. Biomol. NMR* **2008**, *41*, 199–208.

[20] Al-Hashimi, H. M.; Bolon, P. J.; Prestegard, J. H. *J. Magn. Reson.* **2000**, *142*, 153–8.

[21] Zuo, X.; Wang, J.; Foster, T. R.; Schwieters, C. D.; Tiede, D. M.; Butcher, S. E.; Wang, Y. X. *J. Am. Chem. Soc.* **2008**, *130*, 3292–3.

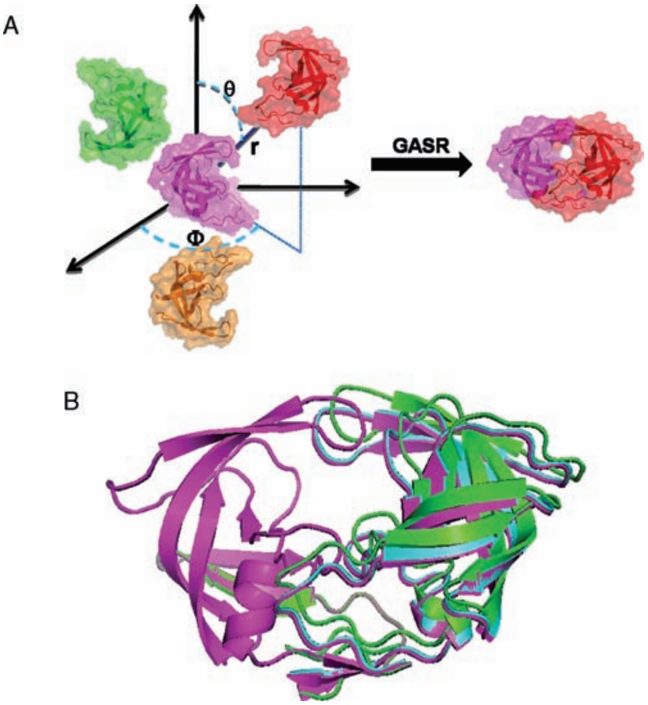

Figure 2. (A) Illustration of spatial search used in the GASR program for a two-subunit protein system in a spherical polar axis system, shown here are the two subunits of the HIV-1 protease. Subunit 1 in magenta is fixed at the origin of axis in space, while subunit 2 in three discrete possible orientations, depicted in red, green and light brown "translates" around subunit 1 without "change in orientation relative to subunit 1" in steps of 1.0 Å for *r*; 10° for both *ϕ* and *θ* in the coarse search, or 1.0 Å and 1.0° in the fine search under restraint of SAXS and other dimensional parameters (see the text). (B) Ribbon diagrams of the HIV-1 protease structures: Superposition of the original NMR structure (PDB code: 1BVE) in magenta, the GASR structure calculated using the simulated RDC and SAXS data without added noise in cyan (pairwise backbone rmsd = 0.22 Å) and with added noise in the SAXS data and RDCs (Gaussian error = 5 Hz) in green (pairwise backbone rmsd = 0.87 Å). The displayed structures were not further refined with the SAXS data. For all three structures the best-fit superposition was carried out for one subunit (left) to clearly convey the differences in relative positioning of the second suunit. All ribbon diagrams in this figure and the following ones were drawn using Pymol (DeLano, W.L. DeLano Scientific, San Carlos, CA,http://www.pymol.org).

one (θ and ϕ) of 1°. The search boundaries are defined by R_g, $D_{\max}$, and $D_{\min}$, the respective radius of gyration, and the maximum and minimum distances between any pair of heavy atoms within the two subunits. $D_{\max}$ is estimated from the pair-distance distribution function, PDDF, derived from experimental SAXS data, as well as the approximate dimension according to the subunit or domain sizes. $D_{\max}$ is usually set with an error of ±10 Å to account for the larger uncertainty in estimating this value based on PDDF and initial component structures. $D_{\min}$ was set to 1.5−3.0 Å, depending on whether the domains are covalently (dual domain proteins) or noncovalently linked. In the first step, the program filters out gross outliers that do not satisfy the boundary conditions at each grid point. Optional restraints such as a C_2-symmetry axis in a homodimer or a heuristic intersubunit distance, if applicable, can also be turned on in the search algorithm to accelerate the search.

The orientation-averaged scattering intensity, $I(q)$, is calculated using the Debye formula:

$$I(q) = \sum_{j}^{N} \sum_{k}^{N} A_j(q) A_k(q) \frac{\sin(qr_{j,k})}{qr_{j,k}} \tag{1}$$

where q is the momentum transfer vector expressed as $(4\pi/\lambda)\sin\theta$; λ is the X-ray wavelength; 2θ is the scattering angle; $I(q)$ is the scattered intensity at q; and $r_{j,k}$ is the distance between *jth* and *kth* atoms; and N is the number of atoms. $A_j(q)$ and $A_k(q)$ are the apparent form factors for the *jth* and *kth* atoms, or the atomic groups, such as CHx, NHx, OHwith considering the effect of the portion of excluded solvent[22] Calculating the scattering profile using eq 1 is a time-consuming step in the grid-search. To speed up the calculation for a two-subunit system in the rigid-body grid search, we rewrote the Debye equation (eq 1) as:

$$\begin{gathered} I(q) = I_a(q) + I_b(q) + I_{ab}(q) \\ I_a(q) = \sum_{i=1}^{m} I_i(q) + 2\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} A_i(q)A_j(q)\frac{\sin qr_{ij}}{qr_{ij}} \\ I_b(q) = \sum_{i=m+1}^{N} I_i(q) + 2\sum_{i=m+1}^{N-1} \sum_{j=i+1}^{N} A_i(q)A_j(q)\frac{\sin qr_{ij}}{qr_{ij}} \\ I_{ab}(q) = 2\sum_{i=1}^{m} \sum_{j=m+1}^{N} A_i(q)A_j(q)\frac{\sin qr_{ij}}{qr_{ij}} \end{gathered} \tag{2}$$

where m is the number of atoms in the first subunits, and $I_a(q)$ and $I_b(q)$ are the scattering intensity contributions from the individual subunits A and B alone, respectively, that remain constant during the search. $I_{ab}(q)$, the contribution from the pair of atoms between two subunits, varies depending on the relative positions of the two subunits; therefore, $I_{ab}(q)$ becomes the only term that needs to be calculated at the each step in the grid search for evaluating the resulting structure. Equation 2 can easily be reformulated for systems containing more than two components.

The root-mean-square deviation (rmsd), defined below (3), is computed between experimental SAXS data, $I_{exp}(q)$, and the back-calculated scattering profile, $I_{\mathrm{cal}}(q)$ to evaluate fit:

$$rmsd = \sqrt{\frac{1}{N-1} \sum_{q} \left[\frac{I_{\exp}(q) - cI_{\mathrm{cal}}(q)}{I_{\exp}(q)}\right]^2} \tag{3}$$

where N is the total number of the scattering data points used in the calculation and c is a scaling factor. The same search process is repeated for the other three (heterodimeric complex) or two (homodimeric proteins) possible orientations for the second subunit. This coarse-grid search generates a list of possible dimeric structures, ranked based on their rmsds in SAXS relative to the experimental data. During the next stage, the top 10% of solutons in the list is submitted to a fine-grid search. The results of the fine-grid search are analyzed using a probability distribution based on the assumption that similar or better solutions are more likely to be found in the neighborhood of good solutions (Supporting Information).

Structure and Interfaces of the HIV-1 Protease Dimer. We benchmarked the program using a simulated set of SAXS and ^{15}N−^{1}H RDC data for the HIV-1 protease homodimer. ^{15}N−^{1}H RDCs are the most readily measurable RDCs in proteins. This allowed us to evaluate program performance with a known structure. The HIV-1 protease is a homodimeric globular protein, with each subunit comprising 99 residues. The experimental interface between the two subunits was determined using 218

[22] Svergun, D.; Barberato, C.; Koch, M. H. J. *J. Appl. Crystallogr.* **1995**, *28*, 768–773.

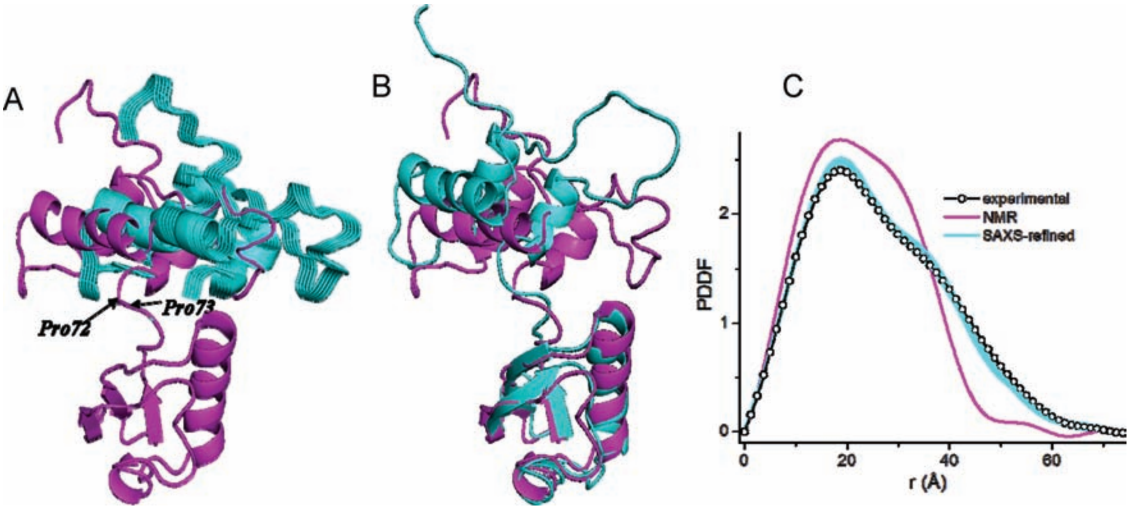

Figure 3. L11 structures and comparison of the back-calculated PDDF with the experimental scattering curve of L11. (A) Ribbon diagram of L11: the top 10% GASR structures calculated without a single interdomain distance restraint are shown in cyan and the non-SAXS refined NMR structure[9] (PDB code: 2e35) in magenta. The positions of Pro72 and Pro73 where the subunits break for the input of GASR are labeled in the NMR structure. (B) Superposition of the NMR+SAXS-refined GASR structure in cyan and the non-SAXS-refined NMR structure in magenta; (C) PDDF curves. These curves are color-coded using the same scheme as in (B) for the calculated curves. The PDDF curve calculated from the experimental SAXS data in black is displayed with small open circles. In both A and B, the structures were best-fitted to the bottom domain in order to emphasize the difference in relative positioning of the other domain.

intersubunit distance restraints.[23] We first simulated SAXS and 70% of amide $^{15}N-^{1}H$ RDCs to derive the architecture and the interfaces between the two subunits using GASR. The backbone rmsd between the original and the GASR derived structure is 0.22 Å, without applying any rigid-body SA refinement. We next added a 5 Hz Gaussian error to each simulated RDC value with 3σ up to 15 Hz. This error range is much larger than the actual experimental error in RDC measurements. The resulting structure exhibits a backbone rmsd relative to the starting one of 0.87 Å. Using hybrid rigid-body SA restrained refinement with SAXS data reduces the rmsd to 0.19 again.

Simulated noisy SAXS data was calculated by averaging SAXS data back-calculated from the NMR ensemble of 28 structures (which exhibited a backbone rmsd relative to the mean of 0.8 Å[23]). It should be noted that the experimental error in SAXS data recorded at the Advanced Photon Sources at Argonne National Laboratory is typically less than 5%, significantly less than the rmsd differences in the SAXS curves that back-calculated from the NMR ensemble (Supporting Information). We nevertheless carried out a calculation using both noisy RDC and SAXS data. The final top 10% of the calculated structures using noisy RDC and SAXS data exhibited backbone rmsd values of 0.84 to 1.2 Å before hybrid rigid-body SA refinement (Figure 2B).

Determination of the Architecture of the Two-Domain Proteins L11 and P23T γD-Crystallin. Using two-domain proteins as test cases serves two purposes: it demonstrates the approach for nonsingle domain proteins and allows testing the applicability of GASR for various initial structure topologies and structure quality. Two-domain proteins are simply special cases of heterodimeric proteins or complexes, with the only difference that they are covalently linked. Thus, the method is also applicable to determine architectures of two domain proteins where few interdomain distance restraints are available to restrain the relative positioning of the two domains. We selected two proteins, L11 and γD-Crystallin that differ significantly in topology and quality, and applied the GASR method to derive architectures of these two proteins.

Protein L11 is a 147-aa ribosomal protein that plays an important role in translocation during peptide synthesis. The three-dimensional structure of L11 has been determined using NOE distance-, RDC- and the small-angle neutron scattering (SANS) constraints.[9] The protein consists of two domains, connected by a five-residue linker that contains tandem prolines (Pro72 and 73). The relative orientation and positioning of two domains in L11 are considered to be important for the coordinated movements that occur during translocation on the ribosome.[24,25] The antibiotic thiostrepton binds between L11's N-terminal domain and the rRNA and inhibits protein synthesis, possibly by interfering with the reorientation and positioning of the domain.[9] During structure determination of L11, few NOEs between the two domains were observed, although the linker is short and fairly rigid, based on analyses of alignment and diffusion tensors.[9]

We used the NMR solution structure that was refined with NOE, dihedral angle and RDC restraints.[9] The NMR solution structure of L11 was treated as a heterodimeric protein by breaking the covalent bond Pro72 and 73 in the linker, with each individual domain used as component inputs in the program GASR, together with RDC and SAXS data. The GASR grid-search was performed without constraining the distance between Pro72 (end of one domain) and Pro73 (beginning of the other domain). The resulting structures (we call them GASR structures thereafter) are displayed in Figure 3A. The top 10% GASR structures with the lowest SAXS rmsd exhibit the correct relative orientation between the two domains, most likely due to highly asymmetrical shapes of the two domains, even though the L11 NMR structure is of relatively low to medium quality in terms of relative positioning of the domains, as evidenced by a large

[23] Yamazaki, T.; Hinck, A. P.; Wang, Y. X.; Nicholson, L. K.; Torchia, D. A.; Wingfield, P.; Stahl, S. J.; Kaufman, J. D.; Chang, C. H.; Domaille, P. J.; Lam, P. Y. *Protein Sci.* **1996**, *5*, 495–506.

[24] Agrawal, R. K.; Linde, J.; Sengupta, J.; Nierhaus, K. H.; Frank, J. *J. Mol. Biol.* **2001**, *311*, 777–87.

[25] Wimberly, B. T.; Guymon, R.; McCutcheon, J. P.; White, S. W.; Ramakrishnan, V. *Cell* **1999**, *97*, 491–502.

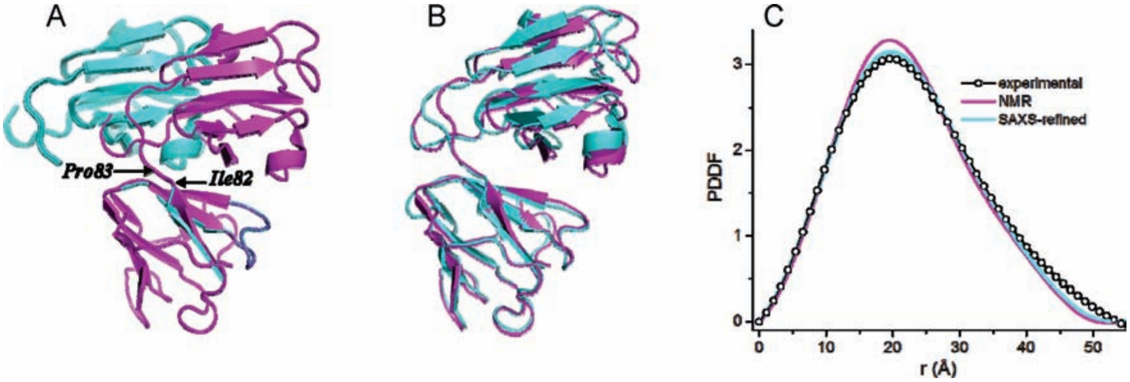

Figure 4. γD-Crystallin P23T mutant structures and comparisons of the back-calculated PDDFs with the experimental curve. (A) Superposition of ribbon diagram representations of the top 10% of the GASR structure (a single, 20 Å distance constraint was used to take into account the covalent linkage between the domains) in cyan and the non-SAXS refined NMR structure (PDB code: 2kfb) in magenta. The positions of Ile82 and Pro83 where the subunits break for the input of GASR are labeled in the NMR structure. (B) Superposition of the NMR+SAXS-refined GASR structure in cyan and the non-SAXS-refined NMR structure in magenta. (C) PDDF comparison. These curves are color-coded using the same scheme as for the ribbon diagrams in (B) with the experimental curve drawn in black with small circles. The overall backbone rmsd between the non-SAXS-refined NMR and the NMR+SAXS-refined structures is about 1.1 Å. In both A and B, best-fit superpositions were carried out for the bottom domain only to highlight the difference in the orientations of the other domain.

rmsd between the experimental SAXS data and the back-calculated one (Figure 3C). Compared to the NMR structure, the two domains in the GASR structures are shifted by about 5 Å, possibly due to the fact that the architectures deviates from the true solution structures. The final relative positioning of the two domains (Figure 3B) was determined using restrained SA with RDC and SAXS data using an SA protocol that regularizes covalent geometry before simulated annealing (Supporting Information). As expected, the original NMR and SAXS-refined NMR structures are locally very similar, but differ in the positioning of the two domains (Figure 3A and B), resulting in a relatively large overall backbone rmsd between the two structures of ~4.9 Å. The relative positioning of the two domains is improved by inclusion of the SAXS restraints, even for this low to medium quality starting structure. A comparison of the back-calculated SAXS curves is provided as Figure S3 in Supporting Information.

The human γD-Crystallin protein also consists of two domains that are linked by a short and nonflexible linker. As both domains are highly globular, this protein constitutes a challenging case for the GASR approach. The NMR solution structure of a congenital cataract forming mutant of human γD-Crystallin, P23T, was recently determined[26] and the set of experimental ^{15}N$-$^{1}H RDCs was used in the present study. GASR was able to generate the correct structure for this protein if D_{max} and R_g values were set to very narrow ranges. However, using loose constraints, GASR generated ambiguous results that contained structures in which one domain was positioned on the wrong side of the other domain. One single distance restraint between the two ends of the break in the chain, the carbonyl carbon of Ile82 and the amide nitrogen of Pro83, allowed to remedy this problem (Figure 4). Interestingly, even a distance as large as 28 Å suffices for this purpose, similar to the case when only RDC based filtering of structural models is performed.[27] However, γD-Crystallin is a single chain, dual domain protein, thus one would naturally set this distance much smaller, given the covalent linkage between the two domains. The GASR structure was regularized and refined using the SA protocol described for L11 (Supporting Information). Comparisons of the SAXS-refined vs non-SAXS-refined NMR structures and their PDDFs are shown in Figure 4B and C.

Homodimeric Complex: The GB1 Side-by-Side Dimer. This complex represents a case for which a well determined monomer structure is available, but the dissociation constant lies in the micromolar range. Such complexes are ideally suited for NMR studies, complemented by SAXS. The structure of the GB1-A34F variant side-by-side dimer was determined by NMR using a 2.2 mM sample and the dimer interface was experimentally determined by 50 intersubunit distance constraints.[28] Both monomer and dimer species coexist in equilibrium in solution and their relative proportions were calculated from the dissociation constant that was measured as 27 ± 4 μM at room temperature.[28] The scattering contribution from dimer was calculated by subtracting the monomer contribution (Supporting Information). D_{min} was set to 3.0 ± 1.0 Å to prevent close contacts between the two subunits, and a 2-fold symmetry restraint was employed during the calculation. The GASR calculation unambiguously yielded the correct structures, as shown in Figure 5A. Further refinement using either rigid-body minimization or SA of the GASR structures improves the fit to the experimental SAXS data even further (Figure 5B and C). It is noteworthy to mention that, even without correction for the monomer, the correct GB1 dimer structures were on the top 30% of the accepted structures generated by GASR. However, in practice, if the final structure is unknown, we cannot consider the top 30% of the accepted structures as unambiguously correct, unless additional information such as a single distance restraint is available. The overall backbone rmsds of the non-SAXS refined NMR structure (PDB ID 2rmm) vs the GASR structure, the rigid-body refined structure and the SAXS-refined GASR structure are ~0.66, 0.89, and 1.53 Å, respectively. Our results for the GB1 dimer represent a relatively easy and best-case scenario for GASR since the monomer structure contains only few mobile regions, the dimerization constant is well determined and the SAXS data was of high quality.

Architecture of the ILK ARD-PINCH LIM1 Complex. The adaptor protein PINCH plays a pivotal role in the assembly of

[26] Jung, J.; Byeon, I. J.; Wang, Y.; King, J.; Gronenborn, A. *Biochemistry* **2009**, *48*, 2597–2609.

[27] Dobrodumov, A.; Gronenborn, A. M. *Proteins* **2003**, *53*, 18–32.

[28] Jee, J.; Byeon, I. J.; Louis, J. M.; Gronenborn, A. M. *Proteins* **2008**, *71*, 1420–31.

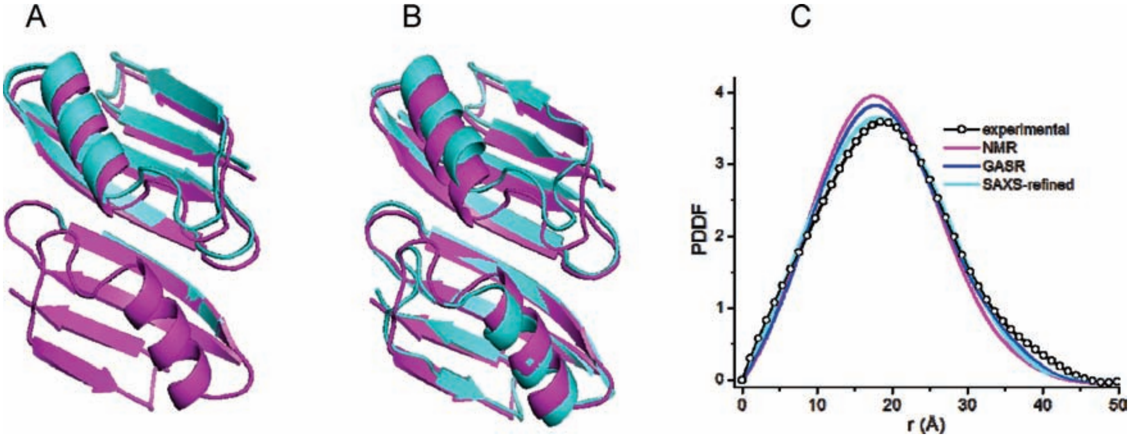

Figure 5. Side-by-side GB1 (A34F mutant) dimer structures and comparisons of the back-calculated PDDFs with the experimental scattering curve. (A) Superpositions of ribbon diagrams of the GASR structure in cyan and the non-SAXS refined NMR structure (PDB code: 2rmm) in magenta.[28] (B) Superposition of the top 10% lowest energy NMR+SAXS-refined GASR average structures in cyan and the original NMR structure[28] in magenta. The best fit superpositions were carried out for one subunit to clearly convey the difference in the positioning of the second subunit in the two structures. (C) PDDF comparison. Back-calculated curves for the original NMR structures are shown in magenta, for the GASR structure in blue, the NMR+SAXS-refined GASR structures in magenta, and the experimental curve is shown in black with small open circles.

focal adhesions (FAs), which are supramolecular complexes that transmit information between the extracellular matrix and the actin cytoskeleton.[29−31] A key step in the PINCH function is its localization to FAs, which depends critically on the tight binding of the LIM1 domain of PINCH to the N-terminal domain of the integrin-linked kinase (ILK).[32,33] The K_d of the complex involving the ILK ankyrin repeat domain (ARD) and the PINCH LIM1, measured using isothermal titration calorimetry, is ~68 nM. The binding is enthalpy-driven and the complex is more stable at low ionic strength, suggesting that the high affinity is primarily due to electrostatic interactions. Despite extensive efforts, only very few unambiguous NOE distance restraints could be measured in a high-sensitivity ^{15}N-edited NOESY spectrum at 900 MHz. The initial architecture of the complex was determined relying heavily on highly ambiguous chemical shift perturbation restraints,[8] rendering its accuracy less than was hoped for. At the same time as the NMR structure became available, an X-ray crystal structure of the ILK ARD/PINCH LIM1 complex was also determined.[34]

The ILK ARD/PINCH LIM1 complex poses a unique challenge for the following reasons: (i) the solution NMR structures of the ILK ARD and the PINCH LIM1 are of a low (LIM1) and medium (ILK) quality;[8] (ii) both proteins contain relatively large numbers of nonstructured regions,[8] complicating the interpretation of the SAXS and RDC data; (iii) relatively a low quality of RDC data, especially for PINCH LIM1, resulted in uncertainty in the determination of the alignment tensor and translated into errors in the four discrete orientations of each subunit in the complex. We tested our GASR approach for determining the architecture of the complex with and without one NOE distance restraint. The ILK structure, the larger component of the complex, was determined to somewhat higher precision than its partner and therefore was used in the Powell grid-search routine[4] in GASR to derive the tensor for the complex, using a scaled D_a value for LIM1. This strategy was adopted since the RDCs for ILK and LIM1 were measured in separate samples with slightly different concentrations of the alignment medium.[8] In the case without the NOE-derived distance constraint, GASR produced the correct relative orientation of the proteins in the complex using the SAXS data, however with a large offset in positioning (Figure S6, Supporting Information). This translational offset most likely is a direct result of the low quality of the NMR structures of both proteins, especially LIM1. This translational offset was removed when the single distance restraint was employed. For the original ILK ARD/PINCH LIM1 structure, seven intermolecular NOE distance restraints were extracted from unambiguous NOEs observed in ^{13}C- and ^{15}N-edit NOESY spectra.[8] Those NOEs cluster around Leu266 and Ala239 in PINCH LIM1 and residues 65−68 in ILK ARD. Here, we tested every one of the seven NOEs as a single loose constraint of 2.8−8.0 Å and used this single constraint in our calculations. All resulting structures of the complex are very similar, exhibiting backbone rmsd values <1 Å. The obtained GASR structure was further refined using the SAXS data and a single intermolecular distance restraint, yielding the structure displayed in Figure 6. The backbone rmsd between the non-SAXS refined (PDB code: 2kbx; magenta, Figure 6A) and SAXS refined (green, Figure 6A) NMR structures for the structured regions (residues 2−154 in ILK ARD and 208−267 in PINCH LIM1) is ~3.9 Å. The backbone rmsd between the SAXS-refined and the X-ray crystal[34] (PDB code: 3f6q) structures is ~4.1 Å. A comparison of all three structures is shown in Figure 6 and the improvement by the experimental SAXS is apparent from Figure 6B and Figure S5 (Supporting Information). Interestingly, despite the similarity between the NMR and crystal structures in terms of global orientation and binding interface,[8,34] the NMR structure appears to better fit to the SAXS data than the crystal structure (Figure 6B and Figure S5, Supporting Information), possibly due to crystal packing artifacts in the latter. In addition, some parts of the polypeptide chains, such as the N-terminal region of PINCH LIM1, are highly flexible in solution, but well-packed in the

[29] Tu, Y.; Li, F.; Goicoechea, S.; Wu, C. *Mol. Cell. Biol.* **1999**, *19*, 2425–34.

[30] Zhang, Y.; Chen, K.; Tu, Y.; Velyvis, A.; Yang, Y.; Qin, J.; Wu, C. *J. Cell Sci.* **2002**, *115*, 4777–86.

[31] Fukuda, T.; Chen, K.; Shi, X.; Wu, C. *J. Biol. Chem.* **2003**, *278*, 51324–33.

[32] Huang, H. C.; Hu, C. H.; Tang, M. C.; Wang, W. S.; Chen, P. M.; Su, Y. *Oncogene* **2007**, *26*, 2781–90.

[33] Hannigan, G.; Troussard, A. A.; Dedhar, S. *Nat. Rev. Cancer* **2005**, *5*, 51–63.

[34] Chiswell, B. P.; Zhang, R.; Murphy, J. W.; Boggon, T. J.; Calderwood, D. A. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 20677–82.

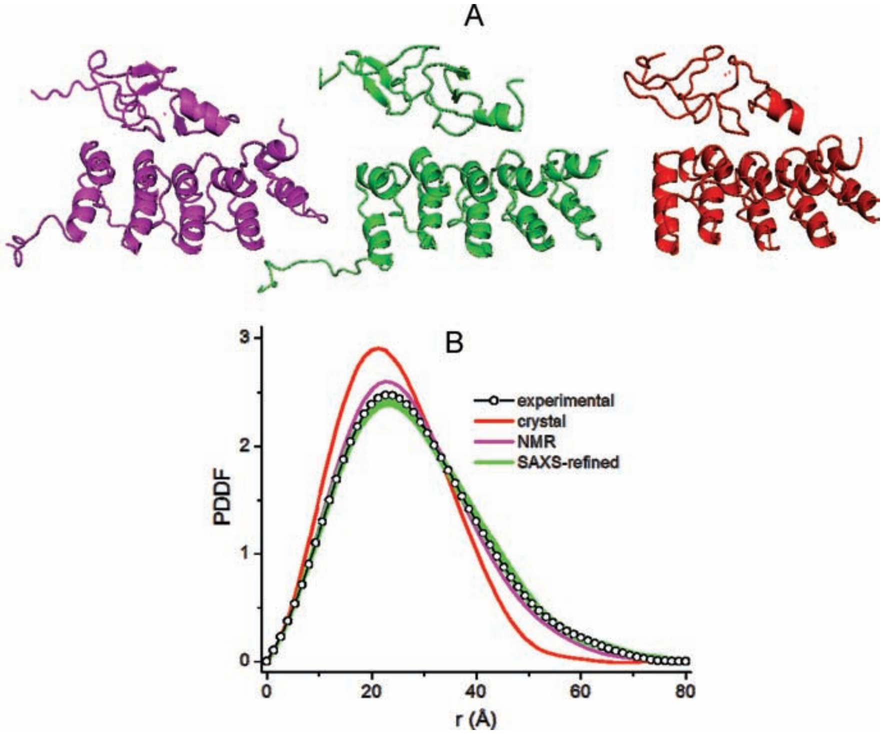

Figure 6. ILK ARD/PINCH LIM1 complex structures and comparisons of the back-calculated PDDFs with the experimental scattering curve. (A) Ribbon diagrams of the ILK ARD/PINCH LIM1 complex structures: NMR structure (magenta) (PDB code: 2kbx), SAXS-refined NMR structure (green) and X-ray crystal structure (red) (PDB code: 6f6q). The NMR+SAXS-refined structure was calculated with the same set of constraints that was used in the previous publication[8] supplemented by the SAXS data. (B) PDDF comparison. These curves are colored using the same color scheme as in (A) for the calculated curves while the experimental curve is shown in black with small open circles. The large difference in PDDF between the back-calculated curve from the X-ray structure and the experimental one suggest that a significant difference between the structure in solution and the crystal is present.

crystal, contributing to this difference. The SAXS refinement statistics for the structures of the ILK ARD/PINCH LIM1 domain complex are provided in Table S2 in the Supporting Information.

Discussion

The GASR approach presented here provides an integrated and efficient way to determine architectures of multicomponent proteins and complexes in solution. SAXS data can be recorded on small amounts of nonisotope-labeled samples and data collection can be completed in a few minutes at a synchrotron or several hours on a home source. Data evaluation using GASR is relatively straightforward and less labor intensive than recording and interpreting various types of heteronuclear multidimensional NMR spectra. The GASR program is fast and efficient: for example, it takes ~10 min on a laptop computer to calculate the structure of a 2 × 14 kD_a homodimeric protein. Therefore, the GASR approach is ideally suited to aid in the structure determination of multicomponent proteins and complexes in solution. If high-quality structures of components are available, derivation of the architecture of a multicomponent protein using GASR is straightforward, even in cases of protein complexes with micromolar K_d vaules. For those cases where only medium quality initial structures are available, GASR is able to calculate the correct structure, provided that the component structures are either highly asymmetrical, such as in L11, or elongated, such as in an RNA:RNA complex.[21] Furthermore, it should be pointed out that the GASR approach yields an overall architecture that reflects the solution properties. It can independently be used to solve architecture of multicomponent proteins, provided that RDCs and high-quality subunit coordinates are available, or, it complements solution NMR methods that exploit intermolecular NOEs to assemble components whose individual structures were determined by either X-ray crystallography or NMR. Here, we only used the GASR approach on two-domain proteins and dimers; the program, however, with little modifications of the program is capable of handling complexes or proteins that have more than two components using a similar strategy.

In addition to being complementary to structural NMR, the GASR approach may also be useful in resolving interface problems encountered in crystallography. For complexes for which the components have low association constants and in which the buried intersubunit surfaces are relatively small or comparable to those across a unit cell, it sometimes is difficult to determine which of the possible interfaces is the biologically relevant. This question is frequently addressed by mutagenesis, altering interface residues and measuring the effect of such changes on the dissociation constant or activity. However, as with all indirect measurements, the results can be difficult to interpret or misleading, as it has been the case for the N-terminus of the dimeric STAT4 complex.[35,36] In such cases, the GASR approach permits one to unambiguously derive architectures of multicomponent proteins or complexes, provided RDC data are available and the component structures are available.

Although very powerful, the GASR approach is not without limitations. Since it relies on topology-based discrimination to determine the correct relative position and orientation, ambigu-

[35] Chen, X.; Bhandari, R.; Vinkemeier, U.; Van Den Akker, F.; Darnell, J. E., Jr.; Kuriyan, J. *Protein Sci.* **2003**, *12*, 361–5.

[36] Vinkemeier, U.; Moarefi, I.; Darnell, J. E., Jr.; Kuriyan, J. *Science* **1998**, *279*, 1048–52.

ous results can ensue when component shapes are highly globular or very symmetrical, or initial structures of components are less well determined, such as in the case of γD-Crystallin, or ILK ARD/PINCH LIM1complex, respectively. One powerful remedy to lift these ambiguities is any information about "proximity". Such distance information can readily be obtained from chemical cross-linking, heuristic biochemical information, compensatory mutagenesis, chemical shift perturbation[37] or paramagnetic relaxation enhancement.[38] Naturally, one can frequently obtain RDC measurements in a second independent alignment medium to eliminate the degeneracy in relative orientation between components.[10] Furthermore, the impact of coexistence of conformers in solution to SAXS data should also be considered (Supporting Information).

In conclusion, the architecture of multicomponent proteins and complexes can be determined using the GASR approach, provided that suitable structures of components are available. This approach is applicable to weekly or strongly associating complexes, of the homo- or hetero multimeric type. It is expected that determining architectures of multicomponent proteins and complexes will greatly benefit from combining NMR and SAXS data with algorithms such as GASR.

Materials and Methods

Sample Preparation, NMR Experiments, and Calculation of Discrete Orientations. Detailed sample preparations for the protein L11, the ILK ARD /PINCH LIM1 complex, the γD-Crystallin P23T mutant and the GB1-A34F mutant were described elsewhere.[9,8,28,26] Monodispersity of samples was examined by dynamic light scattering (DLS) prior to X-ray scattering experiments. DLS studies were performed on a DynaPro Tytan instrument equipped with a Temperature-Controlled MicroSampler (Wyatt Technology Corp., Santa Barbara, CA) at a laser wavelength of 830 nm, scattering angle of 90° in a quartz cuvette at 23 °C. Each measurement consisted of thirty 10 s acquisitions. To obtain the hydrodynamic radii (R_h), the intensity autocorrelation functions were analyzed by *Dynamics 6.7.7.9.* software (Wyatt Technology Corp., Santa Barbara, CA.).

The four possible discrete subunit orientations, all of which are compatible with a single set of RDCs, can be determined using singular value decomposition (SVD),[39] implemented in GASR. If RDCs measured in a second, independent alignment medium are available, the unique orientation of a subunit can also be derived using GASR. In the case that the initial structures are of low or medium-low quality, however, it may be better to derive the discrete orientations using a Powell grid-search algorithm,[4] which was also implemented in the GASR program.

X-ray Scattering Experiments and Data Processing. The solution conditions for recording SAXS data were similar to those for NMR experiments, except for the concentrations of proteins and buffers. The phosphate buffer commonly used in NMR experiments was replaced with either Bis/Tris or MES buffer for the SAXS experiments in order to reduce radiation damage to the proteins. The protein concentrations used for the SAXS experiments in this study were the following: 2.5 mg/mL for L11, 3.6 mg/mL for GB1-A34F (subunit concentration), 3.5 mg/mL for γD-cystallin P23T, and 1.8 mg/mL for ARD ILK/PINCH LIM1. Diluted samples were also run for GB1, L11 and Crystallin samples, but not for ARD ILK/PINCH LIM1, to check for potential interparticle interactions that could distort the SAXS data at low q near zero. No detectable distortions were observed for the higher concentration samples (see radius of gyration, R_g, values in Table S1, Supporting Information).

X-ray scattering measurements were carried out on the BESSRC Sector (12-ID) undulator beamline, a high-flux third generation synchrotron beamline, at the Advanced Photon Source (APS), Argonne National Laboratory. This brilliant beamline offers the advantage of the high sensitivity and reproducibility of experiments. One additional advantage over an in-house scattering device is that much smaller amounts of sample are required, both in terms of concentration and absolute amount. This can be crucially important in cases where concentration dependent aggregation occurs or a protein is too small for scattering on a benchtop scatterer. The experimental setup was as follows: The X-ray wavelength was set at $\lambda = 1.033$ Å. Two setups were used: small- and wide-angle X-ray scattering (SAXS and WAXS, respectively), where the sample to charge-coupled device (CCD) detector (MAR Research, Hamburg) distances were adjusted to achieve scattering q values of 0.006 Å$^{-1} < q < 2.5$ Å^{-1}, where $q = (4\pi/\lambda) \sin \theta$, and 2θ is the scattering angle. Radiation damage was minimized by flowing samples during data acquisition. No degradation of the samples was detected, as confirmed by the absence of systematic signal changes in sequentially collected X-ray scattering images. The accumulated CCD detector image exposure was set to 1−2 s, and data from 20 images were azimuthally averaged after solid-angle correction and normalization with incident primary X-ray beam intensities. The protein scattering data were obtained after subtraction of background solvent scattering from the averaged one-dimensional solution scattering intensity. The WAXS data were used to guide accurate background subtraction for the SAXS data by tuning SAXS background subtraction to coincide with WAXS data in the overlapping q range, approximately between 0.1 and 0.20 Å^{-1}. Although solution X-ray scattering is a high-background measurement, owing to the high brilliancy of the X-ray beam at APS and beamline optimization, the standard deviations (σ) of scattering data were found to be less than 5% of the scattering data throughout the range of 0.006 Å$^{-1} < q < 2.5$ Å^{-1}.

Radii of gyration (R_g) were calculated using the linear Guinier relationship:[40]

$$\ln[I(q)] = \ln[I(0)] - q^2R_g^2/3 \tag{4}$$

The experimental R_g values of the proteins used in this study are listed in Table S1 (Supporting Information). The approximate maximum distance, $D_{\max}$, was estimated from the pair distance distribution function that was obtained using the GNOM program,[41] and dimensions of components based on atomic coordinates. Estimated $D_{\max}$ values are also listed in Table S1 (Supporting Information). $D_{\min}$ was set to 2.5 ± 1.0 Å for dimeric complexes or proteins and 1.5 ± 1.0 Å for the dual-domain proteins. The latter was estimated based on the length of a covalent bond.

When SAXS data contains contributions from both dimeric and monomeric species, a correction can be made for the latter. Such a correction was applied for the GB1-A34F mutant ($K_d \approx 27$ μM). The molar concentrations of the monomeric and dimeric GB1-A34F were calculated as 79.3 and 232.7 μM, respectively, for the 3.6 mg/mL protein solution that was used for the SAXS experiments. The dimer scattering was calculated by subtracting the putative monomer contribution, based on the assumption that no gross conformational change exists between the free monomer structure and the subunit structure in the dimer. At momentum transfer $q = 0$, the scattering intensity [$I(q=0)$] of a sample is proportional to both the concentration and the square of the number of electrons in the molecule. The relative X-ray scattering contribution percentage (p) at $q = 0$ from the monomeric GB1-A34F was estimated as 7.8%, assuming a similar solvation for both the monomeric and the dimeric GB1. After normalizing the experimental SAXS of the

[37] Clore, G. M.; Schwieters, C. D. *J. Am. Chem. Soc.* **2003**, *125*, 2902–12.

[38] Gaponenko, V.; Howarth, J. W.; Columbus, L.; Gasmi-Seabrook, G.; Yuan, J.; Hubbell, W. L.; Rosevear, P. R. *Protein Sci.* **2000**, *9*, 302–9.

[39] Losonczi, J. A.; Andrec, M.; Fischer, M. W.; Prestegard, J. H. *J. Magn. Reson.* **1999**, *138*, 334–42.

[40] Guinier, A. *Acta Metall.* **1955**, *3*, 510–512.

[41] Svergun, D. I. *J. Appl. Crystallogr.* **1992**, *25*, 495–503.

sample [$I(q)^{\text{norm-exp}}$] and the back-calculated SAXS data of GB1-A34F monomer [$I(q)^{\text{norm-mono}}$] relative to the intensity at $q = 0$ as $I(0) = 1$, the relative contribution from dimeric GB1 was calculated by subtracting the contribution of monomer from the normalized experimental SAXS data, that is, $I(q)^{\text{norm-exp}} - p \times I(q)^{\text{norm-mono}}$. For the ILK ARD/PINCH LIM1 complex, the K_d is ~68 nM, therefore no correction was necessary. A figure that displays simulated scattering curves calculated for various dissociation constants is provided in Figure S4 in Supporting Information. The GASR software and Xplor-NIH SA and rigid-body refinement protocols can be downloaded from the author's web page: http://sblweb.ncifcrf.gov/PNAI/files/GASR. Coordinates for all structures were deposited in the Protein Data Bank with PDB codes: 2klj for SAXS- and NMR-refined P23T γD-crystallin, 2klm for refined L11, and 2klk for refined GB1 A34F dimer.

Acknowledgment. We thank Dr. Xinhua Ji for insightful discussions about the crystallography of weakly associated multi-subunit proteins and Dr. David M. Tiede for useful discussion and allowing us using his software for X-ray scattering 2D image processing. This project has been funded in part with federal funds from an intramural grant from the NCI of NIH to Y.X.W., the CIT Intramural Research Program of the NIH to C.D.S., and NIH grants GM082251 and HL58758, GM62823 to A.M.G. and J.Q., respectively. This research was also supported in whole or in part with federal funds from the National Cancer Institute, National Institues of health, under under contract N01-CO-12400. This Research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

Supporting Information Available: A sample input file for GASR, description of simulated annealing refinement, discussions on probability analysis of GASR research and impact of coexistence of conformers in solution, dimensions of protein samples obtained from SAXS data (Table S1), statistics of SA refinements for L11, γD-crystallin, Gb1-A34F dimer, and ILK/PINCH (Tables S2−S5), comparisons of experimental SAXS data and back-calculated curves for L11 (Figure S3) and ILK/PINCH (Figure S5), corrections of SAXS data for GB1-A34F dimer (Figure S4) and comparison of non-SAXS refined and GASR structures for ILK/PINCH (Figure S6). This material is available free of charge via the Internet at http://pubs.acs.org.

JA902528F