J|A|C|S
ARTICLES
Published on Web 04/06/2007

Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering

Pau Bernadó,*[,†,‡] Efstratios Mylonas,[†] Maxim V. Petoukhov,[†,§] Martin Blackledge,[||] and Dmitri I. Svergun*[,†,§]

Contribution from the European Molecular Biology Laboratory, Hamburg Outstation, 22603 Hamburg, Germany, Institut de Recerca Biomèdica, Parc Científic de Barcelona, Josep Samitier 1-5, 08028 Barcelona, Spain, Institute of Crystallography, Russian Academy of Sciences, Leninsky pr. 59, 117333 Moscow, Russia, and Institut de Biologie Structurale Jean-Pierre Ebel, CEA-CNRS-UJF, 41 Rue Jules Horowitz, 38027 Grenoble, France

Received December 20, 2006; E-mail: pbernado@pcb.ub.es; svergun@embl-hamburg.de

Abstract: Structural analysis of flexible macromolecular systems such as intrinsically disordered or multidomain proteins with flexible linkers is a difficult task as high-resolution techniques are barely applicable. A new approach, ensemble optimization method (EOM), is proposed to quantitatively characterize flexible proteins in solution using small-angle X-ray scattering (SAXS). The flexibility is taken into account by allowing for the coexistence of different conformations of the protein contributing to the experimental scattering pattern. These conformers are selected using a genetic algorithm from a pool containing a large number of randomly generated models covering the protein configurational space. Quantitative criteria are developed to analyze the EOM selected models and to determine the optimum number of conformers in the ensemble. Simultaneous fitting of multiple scattering patterns from deletion mutants, if available, provides yet more detailed local information about the structure. The efficiency of EOM is demonstrated in model and practical examples on completely or partially unfolded proteins and on multidomain proteins interconnected by linkers. In the latter case, EOM is able to distinguish between rigid and flexible proteins and to directly assess the interdomain contacts.

1. Introduction

The main paradigm of structural studies in molecular biology is that knowledge of the three-dimensional (3D) structure is a prerequisite of understanding how biological macromolecules function. The advent of the "postgenomic" era, with its large number of available genome sequences, led to a breakthrough in the high-resolution structure determination of individual proteins or domains, largely based on macromolecular X-ray crystallography (MX).[1] Further significant progress is now observed in the studies of large macromolecular complexes. Even in the absence of well-diffracting crystals (required for MX) meaningful models can be constructed from the high-resolution structures of individual subunits by macromolecular docking or rigid body refinement using a combination of MX, Nuclear Magnetic Resonance (NMR), cryo-electron microscopy (cryo-EM), or small-angle X-ray scattering (SAXS).[2,3]

The above progress refers mainly to macromolecules and complexes with well-defined and rather rigid tertiary structures. Much greater difficulties are encountered in the study of flexible objects. Macromolecules possessing inherent flexibility are difficult to crystallize; in addition the conformations of disordered loops or domains cannot be captured by MX, and are difficult to describe uniquely from NMR data. Further, important functions in regulation of gene expression and in the cell cycle are often accomplished by intrinsically unfolded proteins, which do not have a fixed tertiary structure.[4−6] These proteins can obviously not be crystallized, and although NMR, the only high-resolution method applicable, yields valuable information about their local structure,[7−11] there is a clear need for the development of approaches for the quantitative characterization of flexible and intrinsically disordered proteins in solution.

SAXS, a structural method applicable to native particles in solution, is particularly suitable for the study of less structured systems. In a SAXS experiment on a dilute macromolecular solution, the measured scattering intensity emerges from $\sim 10^{16}$ particles in the illuminated sample volume. For monodisperse

[†] European Molecular Biology Laboratory.
[‡] Institut de Recerca Biomèdica.
[§] Russian Academy of Sciences.
[||] Institut de Biologie Structurale Jean-Pierre Ebel.

(1) Gerstein, M.; Edwards, A.; Arrowsmith, C. H.; Montelione, G. T. *Science* **2003**, *299*, 1663.
(2) Sali, A.; Glaeser, R.; Earnest, T.; Baumeister, W. *Nature* **2003**, *422*, 216−25.
(3) Svergun, D. I.; Koch, M. H. J. *Rep. Progr. Phys.* **2003**, *66*, 1735−1782.
(4) Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradovic, Z. *Biochemistry* **2002**, *41*, 6573−6582.
(5) Uversky, V. N. *Protein Sci.* **2002**, *11*, 739−756.
(6) Dyson, H. J.; Wright, P. E. *Nature Rev.* **2005**, *6*, 197−208.
(7) Shortle, D.; Ackerman, M. S. *Science* **2001**, *293*, 487−489.
(8) Louhivouri, M.; Pääkkonen, K.; Fredriksson, K.; Permi, P.; Lounila, J.; Annila, A. *J. Am. Chem. Soc.* **2003**, *125*, 15647−15650.
(9) Mohana-Borges, R.; Goto, N. K.; Kroon, G. J. A.; Dyson, H. J.; Wright, P. E. *J. Mol. Biol.* **2004**, *34*, 1131−1142.
(10) Jha, A. K.; Colubri, A.; Freed, K. F.; Sosnick, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13099−13104.
(11) Bernadó, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W. H.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17002−17007.

10.1021/ja069124n CCC: $37.00

systems containing randomly oriented particles with identical structures, the isotropic SAXS intensity is proportional to the single particle scattering averaged over all orientations. Remarkable recent progress in data analysis methods over the past decade[3] has significantly broadened the possibilities of the technique, enabling one to reconstruct 3D models from these one-dimensional intensities at low resolution. In particular, methods have been developed to generate probable configurations of missing fragments in the incomplete high-resolution structures based on the solution SAXS patterns from the full length proteins,[12] and they were successfully applied in practice.[13−15]

SAXS is often used to study equilibrium mixtures and nonequilibrium processes like protein folding/unfolding or (dis)-assembly pathways, but data interpretation for such mixtures is typically restricted to overall parameters such as the radius of gyration (R_g). This analysis yields valuable albeit limited information about the kinetics of these processes, i.e., folding intermediates.[16,17] More detailed analysis of mixtures and also of flexible systems is hampered by the necessity of taking not only orientational but also conformational averaging into account.

In the present paper, a new approach is proposed for the analysis of flexible systems, in particular unfolded and multidomain proteins with flexible linkers, using 3D models. It is assumed that the experimental data can be represented by an average of a (generally unknown) number of conformers. A pool containing a large number of possible conformations is randomly generated to cover the configurational space, and a genetic algorithm (GA) is used to select appropriate subsets of configurations fitting the experimental data. Quantitative criteria are developed to analyze the solutions and the virtue of the approach, but also its limitations are demonstrated in model and practical examples.

2. Theory and Methods

2.1. Principle of the Ensemble Optimization Method (EOM). Standard analytical methods for studying static structures search for a single model/conformation which satisfies physical and/or biochemical constraints and yields the scattering pattern fitting the experimental data. For flexible structures, we represent the object by an ensemble containing a number N of different conformations. The scattering from such an ensemble is readily computed by averaging the individual scattering patterns from the conformers. To select an appropriate ensemble, a representative pool containing a large number $M \gg N$ of random conformers is first generated, which should cover all the configurational space. A Monte Carlo based search is then employed to select the subsets of configurations fitting the experimental data. Without loss of generality we can assume that the subsets are uniformly populated, so that the intensity of a subset $I(s)$ containing N conformers is

$$I(s) = \frac{1}{N}\sum_{n=1}^{N} I_n(s) \qquad (1)$$

Here, $I_n(s)$ is the scattering from the n-th conformer and the momentum transfer $s = 4\pi \sin\theta/\lambda$ where 2θ is the scattering angle and λ is the wavelength. Nonuniform weights of conformers can be easily accounted for by inclusion of different numbers of conformers with similar shapes. The computation of the scattering intensity $I_n(s)$ from each conformation, especially for large proteins, is a rather time-consuming procedure. To speed up the calculations, the scattering curves from all the structures in the pool are first precomputed and the subsequent selection operators are performed using these patterns and not the structures. The final model should best fit the experimental curve $I_{exptl}(s)$ minimizing the discrepancy between the experimental and calculated curves

$$\chi^2 = \frac{1}{K-1}\sum_{j=1}^{K}\left[\frac{\mu I(s_j) - I_{exptl}(s_j)}{\sigma(s_j)}\right]^2 \qquad (2)$$

where K is the number of experimental points, $\sigma(s)$ are standard deviations, and μ is a scaling factor.

2.2. Generation of Protein Conformations. For unstructured proteins, the pools of possible chain conformations were generated by the program flexible-Meccano,[11] which builds consecutively a polypeptide chain assuming that peptide planes are rigid entities connected through C_α atoms. The specific amino acid conformations are randomly selected from a library derived from a database of coil conformations found in high-resolution X-ray structures.[18] A simple exclusion term is applied to avoid collapse within the chain.[19] Details of this polypeptide building program can be found in the original work (ref 11). Note that, although in this reference only generation of small and middle size proteins is considered, the properties of the algorithm are not chain-length dependent allowing it to be used for arbitrary proteins. The conformer pools for multidomain proteins were constructed using the program Pre_bunch originally developed to generate starting approximations for the program BUNCH.[20] Pre_bunch treats domains as rigid bodies and connects them by self-avoiding linkers, where the dihedral angles of the linkers in the $C_\alpha-C_\alpha$ space are selected randomly but biased to comply with the quasi-Ramachandran plot,[21] and the model generated is free from steric clashes. It should be noted that the most important requirement of the pool generation for EOM, which is to sample the space with feasible initial models, is adequately fulfilled by generation using both flexible-Meccano and Pre_bunch.

For both unstructured and multidomain proteins, $M = 10\,000$ structures are sufficient to cover the configurational space and provide representative pools of conformers. In all cases the scattering patterns from the conformers in the pools were calculated with the program CRYSOL[22] using default parameters in the range of scattering vectors from 0.0 to 0.5 Å^{-1}.

2.3. Ensemble Optimization and Analysis of the Structural Properties. A genetic algorithm (GA) was employed for the subset selection.[23] Each subset (chromosome) is composed of N (typically $N = 50$) scattering profiles (genes) corresponding to different conformers. For each chromosome, the average (eq 1) of its individual SAXS profiles is compared with the experimental scattering yielding the fitness function (eq 2). In the first generation, K (typically, $K = 50$) chromosomes are randomly selected from the pool. In each generation

(12) Petoukhov, M. V.; Eady, N. A.; Brown, K. A.; Svergun, D. I. *Biophys. J.* **2002**, *83*, 3113−3125.
(13) Garcia, P.; Ucurum, Z.; Bucher, R.; Svergun, D. I.; Huber, T.; Lustig, A.; Konarev, P. V.; Marino, M.; Mayans, O. *FASEB J.* **2006**, *20*, 1142−51.
(14) Durand, D.; Cannella, D.; Dubosclard, V.; Pebay-Peyroula, E.; Vachette, P.; Fieschi, F. *Biochemistry* **2006**, *45*, 7185−93.
(15) Petoukhov, M. V.; Monie, T. P.; Allain, F. H.; Matthews, S.; Curry, S.; Svergun, D. I. *Structure* **2006**, *14*, 1021−7.
(16) Chen, L.; Widegger, G.; Kiefhaber, T.; Hodgson, K. O.; Doniach, S. *J. Mol. Biol.* **1998**, *276*, 225−237.
(17) Kimura, T.; Uzawa, T.; Ishimori, K.; Morishima, I.; Takahashi, S.; Konno, T.; Akiyama, S.; Fujisawa, T. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2748−2753.
(18) Lovell, S. C.; Davis, I. W.; Arendall, W. B., III; de Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. *Proteins* **2003**, *50*, 437−450.
(19) Levitt, M. *J. Mol. Biol.* **1974**, *104*, 59−107.
(20) Petoukhov, M. V.; Svergun, D. I. *Biophys. J.* **2005**, *89*, 1237−1250.
(21) Kleywegt, G. J. *J. Mol. Biol.* **1997**, *273*, 371−376.
(22) Svergun, D. I.; Barberato, C.; Koch, M. H. J. *J. Appl. Crystallogr.* **1995**, *28*, 768−773.
(23) Jones, G. *Genetic and Evolutionary Algorithms. Encyclopedia of computational chemistry*; Wiley; Chichester, U.K., 1998.

these *K* chromosomes are submitted to two genetic operators: *random mutation* and *crossing*. In *random mutation* up to 20% of the genes of each chromosome were exchanged for others, 50% of which are randomly selected from the pool and 50% from the other *K*-1 chromosomes of the same generation. In the *crossing* operator two chromosomes were selected randomly and their genes were exchanged, with a minimum of two genes transferred to the offspring. At the end of these two genetic operations, the population is composed of 3*K* chromosomes, the fitness function for each average curve was evaluated, and the best *K* chromosomes were selected for further evolution, typically for up to 5000 generations.

The number of conformers *N* in the subset reflects the flexibility of the system, and its influence on the results was comprehensively analyzed below. The other tunable GA parameters were found to be not critical for the performance of the algorithm.

In all calculations, 200 independent GA runs were performed and the structural characteristics of the resulting 10 000 conformations were analyzed. The three characteristic parameters of the solutions were the radius of gyration (R_g), the maximum distance between two C_α atoms within a conformer (D_{max}), and the anisometry (P). The latter parameter was calculated using the eigenvalues (d1, d2, and d3) obtained from the diagonalization of the radius of gyration tensor as

$$P = 2*\mathrm{d1}/(\mathrm{d2} + \mathrm{d3}) \quad (3)$$

where d1 is the most unique dimension, and d2 and d3 are the most similar dimensions (if $P < 1$ the particle is oblate; if $P > 1$ it is prolate).

A normalized spatial discrepancy (NSD) was used to quantify the diversity between structures within the ensemble derived from the optimization.[24] The NSD between two structure sets S_1 and S_2 containing 3D coordinates of points (e.g., atoms, residues, or beads) is defined as follows. For every point s_{1i} from the set $S_1 = \{s_{1i}, i = 1,..N_1\}$, the minimum value among the distances between s_{1i} and all points in the set $S_2 = \{s_{2i}, i = 1,..N_2\}$ is denoted as $\rho(s_{1i}, S_2)$. The NSD is a normalized average

$$\mathrm{NSD}(S_1, S_2) = \left[\frac{1}{2}\left(\frac{1}{N_1 d_2^2}\sum_{i=1}^{N_1}\rho^2(s_{1i}, S_2) + \frac{1}{N_2 d_1^2}\sum_{i=1}^{N_2}\rho^2(s_{2i}, S_1)\right)\right]^{1/2} \quad (4)$$

where N_i is the number of points in S_i and the fineness d_i equals the average distance between the neighboring points in S_i. For ideally superimposed objects NSD tends to be 0; when it significantly exceeds 1 the objects systematically differ one from the other. When comparing the structures from an ensemble, the NSD between all pairs of conformers (C_α coordinates independently of the residue they represent) are calculated with the program SUPCOMB,[24] and the average value is reported.

2.4. Multiple Curve Fitting. The approach can be extended to simultaneous fitting of multiple scattering curves, e.g., when data from partial constructs or deletion mutants are also available. Indeed, given the *n*-th conformation from the pool the scattering from a subconformation n_{i-j} containing only residues *i* to *j* is readily computed. For the *pM* conformers in the pool, where *p* is the number of fragments measured, the scattering profiles were calculated and used as input files for the EOM. For each chromosome, *p* different averaged theoretical scattering curves are calculated and compared with the *p* experimental curves, and the overall fitness (the sum of the individual χ_i^2 values) guides the GA selection process.

2.5. Generation of the Test Cases. 2.5.1. Synthetic Scattering Profiles. The synthetic intensity profiles calculated applying different long-range and/or conformational restrictions to form specific molecular sizes and shape distributions are listed in Table 1. A 100 amino acid long polyalanine chain was used for testing the performance of the EOM for unstructured proteins. In the first test ensemble, at least one C_β from the A11−A20 fragment was placed within 15 Å from another C_β from the A81−A90 fragment (*compact*). The second ensemble was generated forcing the presence of an extended conformation, $\phi = -112.6°$ and $\psi = 123.0°$, in the A46-A55 fragment of the polypeptide chain (*extended*). The third ensemble was created without restrictions (*control*).

Table 1. Synthetic Test Cases Simulated and Computed Using EOM

test case[a]	structural properties[b]	curves fitted
	Unfolded Proteins (Polyalanine Chain)	
control	no structural restrictions	1
extended	$\phi = -112.6°$, $\psi = 123.0°$ in the A46−A55 fragment	1
compact	15 Å contact between A11−A20 and A81−A90 fragments	1
extended_mf	$\phi = -112.6°$, $\psi = 123.0°$ in the A80−A89 fragment	3
compact_mf	15 Å contact between A50−A60 and A90−A100 fragments	3
	Multidomain Proteins	
symmetric_extended	domains 1 and 4 farther than 130 Å	1
symmetric_compact	domains 1 and 3 closer than 30 Å	1
asymmetric compact	domains 1 and 3 closer than 40 Å	1 and 2

[a] A name tag assigned throughout the text to the test case. [b] Structural properties of the ensemble used to generate the synthetic curve.

As a test for multidomain proteins a polyubiquitin chain was generated composed by four identical ubiquitin domains (1ubq, 72 aa, the last 4 residues were deleted because they are highly flexible) connected through 20 residue long polyalanine linkers, with additional 10 amino acid long tails at both termini. Two ensembles were calculated for this symmetric case forcing the center of masses of the first and the third ubiquitin domains to be closer than 30 Å (*symmetric compact*) and the first and the fourth further than 130 Å (*symmetric extended*). In another test, an asymmetric chain was identical to the previously described but with four different structured domains: SH3 (1awx, 55 aa), ubiquitin (1ubq, 72 aa), MAD2A (1s2h, 205 aa), and MBP (1jvx, 371 aa). In the ensemble created for the asymmetric model, the SH3 (first) and the MAD2A (third) domains were closer than 40 Å (*asymmetric compact*).

2.5.2. Multiple Scattering Curves Fitting. Two test cases were studied for the polyalanine chain: (i) an extended section was forced encompassing residues A80−A89 (*extended_mf*), and (ii) a long-range contact was imposed between two C_β atoms of the A50−A60 and the A90−A100 fragments (*compact_mf*). Two additional constructs of the chain were also generated: (i) a 1−75 fragment, which in both cases will behave as a completely unfolded system, and (ii) a 50−100 fragment containing the structured sections of the chain.

For the *asymmetric compact* chain an additional ensemble was calculated for a construct encompassing the interacting first and the third domains.

For each of these systems, 1000 conformations were generated with the appropriate software, and then their individual scattering profiles were calculated with CRYSOL[22] and averaged to obtain the synthetic curves that were fitted with the EOM. A Gaussian error of 2% was added to the synthetic data prior to the fitting.

2.6. Analysis of the Experimental Scattering Profiles. Synchrotron X-ray scattering data from solutions of lysozyme in 40 mM acetic acid, 50 mM NaCl, 8 M urea, pH 4.0 were collected at the X33 beamline of the EMBL (DESY, Hamburg)[25] using an MAR345 image plate detector. The scattering patterns were obtained with a 3 min exposure time for a solute concentration of around 10 mg/mL at temperature 363 K. A small (10 mM) or large amount (100 mM) of DTT was added to the

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

(25) Koch, M. H. J.; Bordas, J. *Nucl. Instrum. Methods* **1983**, *208*, 461−469.

solutions prior to the measurements. The sample−detector distance was 2.7 m, and the range of momentum transfer covered was $0.008 < s < 0.5$ Å^{-1}. Experimental details about the sample preparations and the SAXS measurement on the Bruton's Tyrosyne Kinase (BTK) constructs have been previously reported.[26]

In both cases, unfolded lysozyme and the BTK, a pool of 10 000 structures, was calculated. The side chains of the lysozyme were added with the program sccomp.[27] The conformations of BTK were calculated using the 3D structures of the four folded domains as described in ref 26 connected with linkers of the appropriate length. The pool of conformers of the three-domain PH−SH3−SH2 construct was calculated as described in the multiple fitting section.

3. Results and Discussion

3.1. Optimal Ensemble Size in Unfolded Proteins. The use of a limited number N of structures (curves) in the chromosome to describe the astronomical amount of conformers coexisting in solution is an unavoidable simplification. The effect of the chromosome size during the EOM was studied in the *control* polyalanine ensemble. A systematic improvement of the merit function was observed when going from a single conformer description, $N = 1$ (the best conformer from the pool), to a chromosome with $N = 100$, $\chi^2_i = 9.57$ and 0.80, respectively. The fitness improved dramatically with the number of conformers in the chromosome for small N to reach a plateau around $N = 10$. The R_g distributions derived from the optimized ensemble can be compared with those originating from the *control* test case. Even for small N, broad distributions are obtained resembling the test case, although spikes and artifacts are observed at the distribution tails. Smoother and more accurate distributions were obtained for $N = 50$ and $N = 100$. These results suggested that $N = 50$ is a good compromise for unfolded proteins between the accuracy of the ensemble and computational resources required (see Supporting Information). Note that similar optimal ensemble sizes have been used in order to describe Paramagnetic Relaxation Experiments in chemically unfolded proteins.[28] In contrast, much larger ensembles are necessary to average Residual Dipolar Couplings (RDCs) in unstructured proteins.[10,11] The difference in the optimum ensemble size may reflect differences in the structural information encoded in the experimental data from these techniques.

3.2. Size/Shape Descriptor Distributions for Unfolded Proteins. SAXS is often used to assess the overall sizes of unfolded and partially folded proteins by monitoring the R_g value.[29] The experimental R_g is an overall size estimate being a z-average over all conformations in solution. The R_g distribution provided by the EOM yields significantly more information about the ensemble.

Figure 1 displays the R_g, D_{max}, and P distributions from the optimized ensembles compared with those derived from the test cases and those corresponding to the initial pool of 10 000 conformers. Very good agreement between the test-case distributions and the EOM-derived ensembles was observed even for situations with broad dispersions such as the *extended* ensemble. Interestingly, EOM is able to enrich the final ensembles with conformations whose size/shape properties are not largely populated in the initial pool. The anisometry parameter P accounts for higher resolution information about the overall shape, and it is worth noting that EOM is able to discern between prolate ($P > 1$) and oblate ($P < 1$) molecules whereby their populations neatly reproduce those of the simulated test cases.

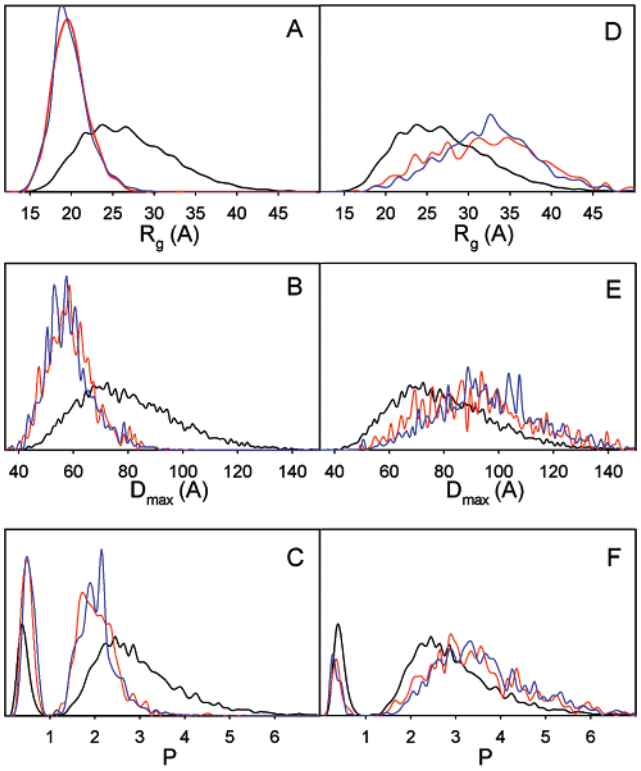

Figure 1. Size and shape descriptor distributions: R_g (A and D), D_{max} (B and E), and P (C and F) for the *compact* (A−C) and *extended* (D−F) polyalanine test cases. The distributions correspond to the pool of 10 000 structures (black), the 1000 structures of the test case (red), and the optimized ensemble (blue) calculated using chromosomes with $N = 50$ structures.

Overall, EOM provides accurate size and shape descriptor distributions of disordered proteins from the SAXS profiles. On the other hand, care must be taken when interpreting the EOM selected ensembles in terms of individual models. SAXS is a low-resolution method, and the structural information that can be reliably extracted without the danger of overinterpretation is the size and shape distribution functions, even though the algorithm employs high-resolution models to describe the conformers. In our simulations, different EOM runs selected different combinations of conformers from the pool, but their size and shape descriptor distributions were very similar to those of the test case.

3.3. Multiple Curve Fitting: Extracting Local Structural Information. It has been shown that both the presence of secondary structural elements[30] and long-range contacts in intrinsically unfolded proteins[31] can play pivotal roles in the biological activity. In addition, protein folding intermediates present local and long-range interactions that delineate the folding funnel.[32,33]

(26) Márquez, J. A.; Smith, C. I. E.; Petoukhov, M. V.; Lo Surdo, P.; Mattson, P. T.; Knekt, M.; Westlund, A.; Scheffzek, K.; Saraste, M.; Svergun, D. I. *EMBO J.* **2003**, *22*, 4616−4624.
(27) Eyal, E.; Najmanovich, R.; McConkey, B. J.; Edelman, M.; Sobolev, V. *J. Comput. Chem.* **2004**, *25*, 712−724.
(28) Lindorff-Larsen, K.; Kristjansdottir, S.; Fieber, W.; Dobson, C. M.; Poulsen, F. M.; Vendruscolo, M. *J. Am. Chem. Soc.* **2004**, *126*, 3291−3299.
(29) Doniach, S. *Chem. Rev.* **2001**, *101*, 1763−1778.
(30) Fuxreiter, M.; Simon, I.; Friedrich, P.; Tompa, P. *J. Mol. Biol.* **2004**, *338*, 1015−1026.
(31) Bertoncini, C. W.; Jung, Y.-S.; Fernandez, C. O.; Hoyer, W.; Griesinger, C.; Jovin, T. M.; Zweckstetter, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 1430−1435.

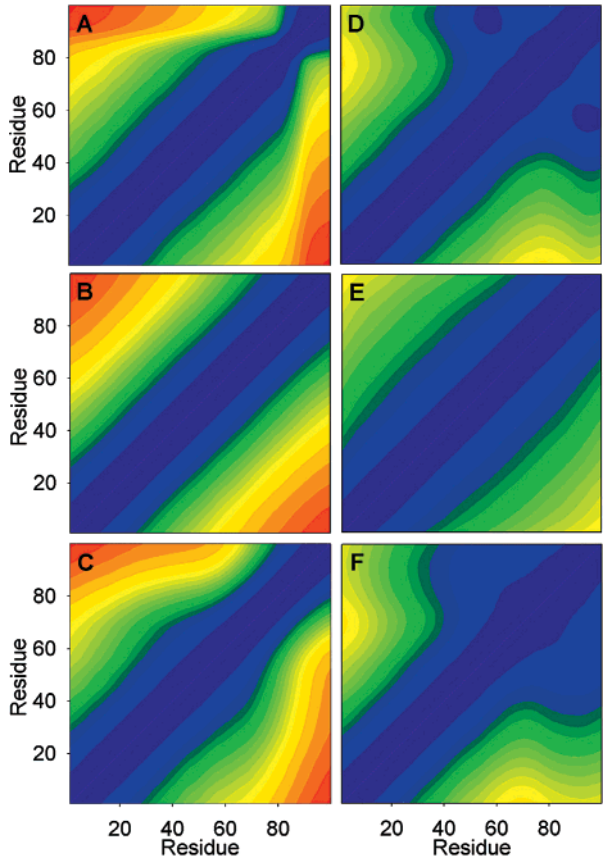

Figure 2. Detection of structuring effects in unfolded proteins. (A) Average $C_\alpha-C_\alpha$ distance matrix for the ensemble of 1000 structures with extended conformation along the A80−A89 fragment and (D) with a long-range contact between A50−A60 and A90−A100 sections of the polyalanine chain. Contact matrices derived from the single (full-length) curve fitting ensemble, for the extended (B) and compact ensembles (E), and those derived from the multiple curve fitting for the (C) extended (*extended_mf*) and (F) compact (*compact_mf*) test cases. The color scale goes from less than 10 Å (purple) to 70 Å (red).

Two simulations were performed to explore the ability of the EOM to characterize these effects (see Theory and Methods). The $C_\alpha-C_\alpha$ average contact matrices for both test cases, *extended_mf* and *compact_mf*, and for their single-curve optimized ensembles are shown in Figure 2. The optimized ensembles from the single-curve fitting clearly detects whether the SAXS curves originate from more elongated (*extended_mf*) or more compact (*compact_mf*) chains compared to the pool of eligible structures (Figure S2 in the Supportiong Information). However, localization of the section responsible for the structuring remains elusive due to the low resolution of SAXS. This lack of information already discussed in the previous section is evident from the symmetry of the contact matrices displayed in Figure 2B,E.

Figure 2C,F display the averaged $C_\alpha-C_\alpha$ contact matrices calculated from the ensembles derived by the simultaneous fitting of three scattering profiles: the full-length and the A1-A75 and A50-A100 synthetic curves (see Theory and Methods). For *extended_mf* the multiple curve analysis unambiguously establishes that the highly extended fragment is located at the C-terminal part of the chain. Similarly, the C-terminal part is identified as the origin of the compactness for the multiple fit of the *compact_mf* test case, and the average $C_\alpha-C_\alpha$ contact matrix becomes very similar to the target one. Note that subject to the low resolution of SAXS the interacting fragments at the residue level cannot be defined.

The multiple curve fitting clearly allows one to extract more information about the unstructured proteins compared to the single curve analysis in the previous section such that secondary structure formation or long-range interaction networks derived from mutational or biological data on unfolded proteins can be detected or validated. A limitation of the multiple curve analysis is that the deletion mutants are assumed to keep the same conformation as the corresponding portions in the full length protein. This condition may not be fulfilled, e.g., when deleting highly charged terminal parts and independent analysis of electrostatic interactions (based on the primary structure or on experimental methods like NMR).

3.4. Detection of Dynamics in Multidomain Proteins. Interdomain motions in multidomain proteins are normally linked to their biological activity. Some data about these large amplitude motions are provided by NMR relaxation experiments[34] or the measurement of RDC in partially aligned proteins.[35] The use of EOM yields complementary quantitative information about these motions from the SAXS patterns.

Two different synthetic SAXS profiles from a polyubiquitin chain under different dynamic regimes were simulated and fitted using EOM. The model objects were (i) a completely flexible polyubiquitin chain and (ii) a rigid polyubiquitin chain with $R_g = 23.6$ Å. The EOM was applied with increasing chromosome size, N, and two structural properties of the derived ensembles have been monitored: NSD, reflecting the structural spread of the conformers in the ensemble, and the distribution of R_g. Figure 3A displays the NSD variation for the two cases compared with the averaged variability of 100 conformers randomly selected from the pool (NSD = 2.06). In the flexible polyubiquitin case the structural diversity of the selected ensembles, independently of the value of N, runs very close to that of the pool, whereas for the static structure the EOM ensemble displays a much lower variability (NSD = 1.3). Further, the polyubiquitin under a dynamic regime presents a broad R_g distribution equivalent to that of the pool, whereas the static regime displays narrow R_g distributions around the R_g of the test conformation as observed in Figure 3B.

An intermediate dynamic regime was also tested where the three first ubiquitin domains and linkers were fixed as in the compact structure, and the fourth domain and the C-terminal tail were free to adopt all sterically allowed conformations. The optimized ensembles present a large value of NSD, around 1.85, slightly lower than the pool, and a broad R_g distribution

In all our tests, the scattering profiles from flexible proteins submitted to EOM yielded highly heterogeneous ensembles with an NSD close to the pool one and broad, more than 20 Å, distributions of R_g. On the other hand rigid proteins are readily detected from the small structural variability (NSD < 1.5) and narrow R_g distributions of a few angstroms around the average R_g. The EOM models for rigid proteins are similar to those

(32) Vendruscolo, M.; Paci, E.; Dobson, C. M. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 14817−14821.
(33) Korzhnev, D. M.; Salvatella, X.; Vendruscolo, M.; Di Nardo, A. A.; Davidson, A. R.; Dobson, C. M.; Kay, L. E. *Nature* **2004**, *430*, 586−590.
(34) Brüschweiler, R.; Liao, X.; Wright, P. E. *Science* **1995**, *268*, 886−889.
(35) Fischer, M. W. F.; Losonczi, J. A.; Weaver, J. L.; Prestegard, J. H. *Biochemistry* **1999**, *38*, 9013−9022.

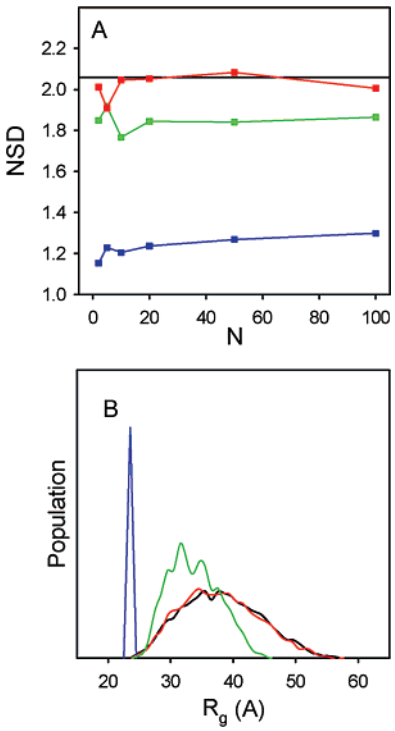

Figure 3. Detection of interdomain motions. (A) Variation of the value of NSD as a function of the number of conformers, *N*, in the optimized ensemble, for a free polyubiquitin chain (red), for compact static conformation (blue), and for an intermediate dynamic regime (green). Black line represents the structural variability in the pool of structures (NSD = 2.06) from 100 randomly selected structures. (B) R_g distributions calculated from the optimized ensembles for the above cases obtained with ensembles of $N = 50$ (the same color code than panel A).

obtained with rigid body refinement programs SASREF and BUNCH.[20] The NSD and the R_g distribution in the selected structures are thus representative parameters to detect the interdomain dynamics in multidomain proteins.

3.5. Structural Information in Multidomain Proteins. The relative orientations of domains in multidomain proteins in solution can be determined using NMR,[34,36] SAXS,[20] or their combination.[37] However, quantitative analysis of large scale movements in flexible multidomain proteins is difficult.[38−40] We have tested the capabilities of EOM to derive domain motions from SAXS. Synthetic scattering data were created for two ensembles of the polyubiquitin chain with the four ubiquitin domains arranged in (i) compact (*symmetric compact*) and (ii) extended conformations (*symmetric extended*) (see Theory and Methods). The structural properties of the EOM-derived ensembles were analyzed. Similar to the above results for disordered proteins, the distributions of the overall parameters R_g, D_{max}, and *P* neatly agree with those of the test cases (Figure S3 in the Supporting Information).

For multidomain proteins, analysis of relative interdomain distances yields the most relevant information about the structural organization. Indeed, the presence of interdomain contacts can modulate protein activity, and their modifications in response to external conditions outline their regulation processes.[41] We have analyzed the *symmetric compact* and *asymmetric* multidomain test cases (see Theory and Methods) using the EOM. The interdomain distance distributions obtained from the *symmetric compact* case (Figure 4) reveal that the model is highly compact. However, the interaction fragment responsible for the compactness remains elusive, resulting in highly similar distance distributions between the terminal domains, 1−2 and 3−4. This observation is similar to the symmetric $C_\alpha-C_\alpha$ averaged distance matrix obtained earlier for polyalanine chain test cases (cf. Figure 2B,E). The ambiguity is therefore caused by the symmetry of the polyubiquitin chain, and it vanishes when the domains have different sizes, as shown in Figure 3. In the asymmetric protein the 1−2 and 3−4 distance distributions become different demonstrating that the N-terminal part is compact. This analysis, however, still shows uncertainty regarding the specific domain−domain interactions. When the scattering curve from a short construct encompassing domains 1−3 is simultaneously fitted with the full length curve this uncertainty is also resolved, indicating that the relevant interdomain contact is between the first and the third domains, whereas the third and the fourth domains behave independently.

The EOM-based analysis cannot uniquely reconstruct distribution of interdomain distances for symmetric multidomain proteins from the optimized ensembles, but this limitation is partially or totally removed for the case of asymmetric proteins and when multiple curves for different constructs are available. Of course, as in the case of the unstructured proteins, the multiple curve fitting implies that the structures of the partial constructs remain largely the same as in the full-length protein.

3.6. Application of EOM to Experimental Systems. 3.6.1. Denatured Lysozyme. The denatured lysozyme sample in the presence of 8 M urea and 10 mM DTT yields $R_g = 26.3$ Å indicating a highly unstructured system (note that the R_g of native lysozyme is 15 Å). The EOM neatly fitted the data with $\chi^2 = 1.64$, but surprisingly the obtained R_g and *P* distributions were biased toward compact structures showing an overpopulation of oblate conformers when compared to the pool (Figure 5). This indicates that the protein was not completely unfolded, presumably due to remaining disulfide bridges. To disrupt these bridges the concentration of the reducing agent, DTT, was increased to 100 mM yielding $R_g = 30.0$ Å. An excellent fit to the experimental scattering profile was obtained using EOM ($\chi^2 = 1.37$) as shown in Figure 5, whereby the R_g and *P* distributions derived from the optimized ensemble are virtually equivalent to those for the pool of structures calculated with flexible-meccano. This further confirms that the latter program generates models adequately describing unfolded proteins.

Our results suggest the presence of one or more disulfide bridges in the low DTT sample, although not all four.[42] The presence of a reduced folding intermediate with a complex network of noncovalent interactions cannot be excluded. However, under high denaturing agent concentration lysozyme lacks permanent long-range contacts. Note that the distributions in this sample arise from a structural averaging and are,

(36) Chou, J. J.; Li, S.; Klee, C. B.; Bax, A. *Nat. Struct. Biol.* **2001**, *8*, 990−997.
(37) Grishaev, A.; Wu, J.; Trewhella, J.; Bax, A. *J. Am. Chem. Soc.* **2005**, *127*, 16621−16628.
(38) Baber, J.; Szabo, A.; Tjandra, N. *J. Am. Chem. Soc.* **2001**, *123*, 3953−3959.
(39) Bertini, I.; del Bianco, C.; Gelis, I.; Katsaros, N.; Luchinat, C.; Parigi, G.; Peana, M.; Provenzani, A.; Zoroddu, A. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6841.
(40) Bernadó, P.; Fernandes, M. X.; Jacobs, D. M.; Fiebig, K. M.; García de la Torre, J.; Pons, M. *J. Biomol. NMR* **2004**, *29*, 21−35.
(41) Jacobs, D. M.; Saxena, K.; Vogtherr, M.; Bernadó, P.; Pons, M.; Fiebig, K. M. *J. Biol. Chem.* **2003**, *28*, 26174−26182.
(42) Hoshino, M.; Hagihara, Y.; Hamada, D.; Kataoka, M.; Goto, Y. *FEBS Lett.* **1997**, *416*, 72−76.

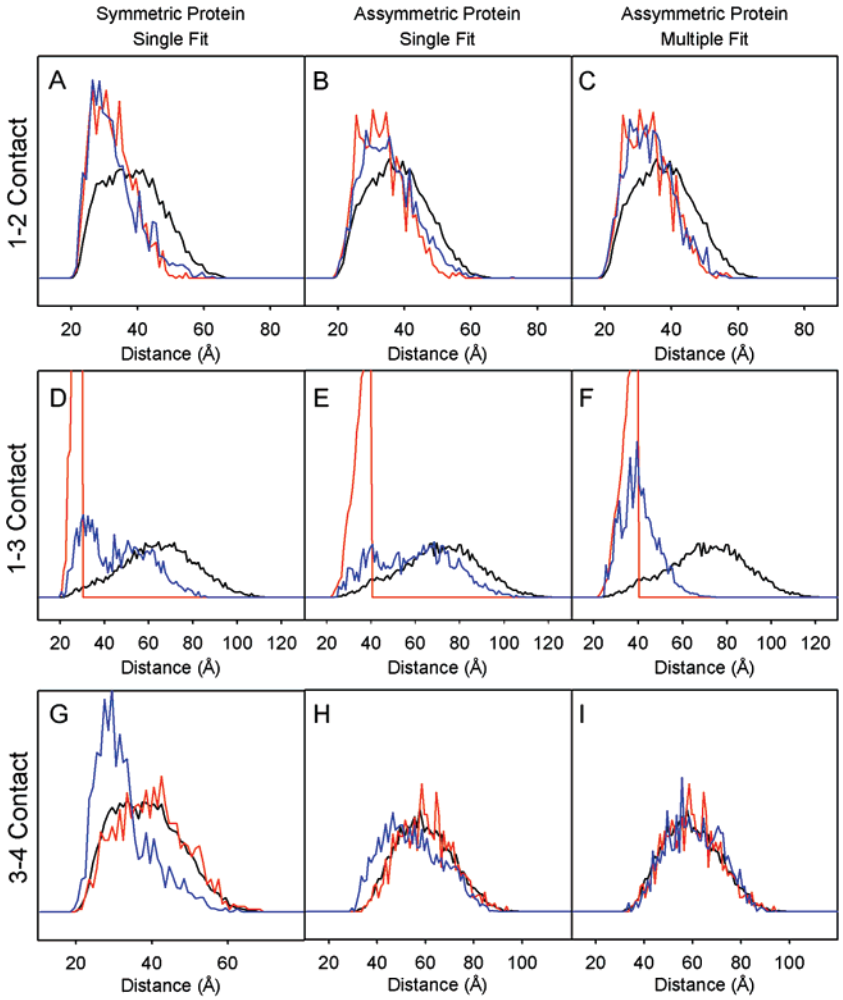

Figure 4. Center-to-center distance distributions of the domains in the optimized ensembles obtained from the fitting of synthetic data of multidomain proteins with long-range contacts imposed. Distributions are shown between domains 1 and 2 (A−C), 1 and 3 (D−F), 3 and 4 (G−I), in the *symmetric compact* chain (A, D, and G), in the *asymmetric compact* chain using the single curve fit (B, E, and H), and in the *asymmetric compact* using the multiple curve fit (C, F, and I). Distributions from the pool of structures, the test cases, and the optimized ensembles are displayed in black, red, and blue, respectively.

therefore, compatible with the transient hydrophobic contacts that have been suggested in a similar lysozyme preparation by measuring the increase in the NMR transverse relaxation rates[43] that arise from a more complex time-averaging phenomena.

3.6.2. Bruton's Protein Tyrosine Kinase (BTK): A Multidomain Protein. Bruton's protein tyrosine kinase (BTK) is a member of the Tec family of cytoplasmatic (or nonreceptor) protein tyrosine kinases, and it is crucial for human and murine B cell development.[44,45] BTK is a multidomain protein formed by four different folded domains: (1) Pleckstrin homology (PH, 169 aa), (2) SH3 (55 aa), (3) SH2 (96 aa), and (4) kinase (Kin, 258 aa) (see Figure 4B).

Scattering profiles from two BTK constructs,[26] the full-length (FL) and a fragment encompassing the PH, SH3, and SH2 domains (PH-SH2), were simultaneously fitted using the EOM. The agreement of the ensemble model with the experimental scattering profiles is excellent, $\chi^2 = 0.41$ and 0.50 for the FL and the PH−SH2, respectively (Supporting Information, Figure S4A). The average NSD value calculated from the ensemble of 50 structures with the lowest χ^2 is 1.973. This value is only marginally smaller than the average NSD = 2.005 characterizing 100 randomly selected structures from the pool, suggesting significant flexibility of the molecule. However, the size and shape descriptor distributions of the optimized ensemble (Supporting Information, Figure S5B) are clearly biased toward extended conformations compared with the distributions obtained from the pool of structures. The distance distributions between domains (Figure 5C) further allow one to assess the degree of extension of specific linkers. Two linkers, connecting the PH−SH3 and SH3−SH2, are clearly more extended than the pool, whereas the SH2−Kin linker does not appear different from the pool.

The fragment of the PH−SH3 linker close to the SH3 domain has a proline-rich region with a duplicate sequence KPLPPE/TP characteristic of the Tec kinases.[45] It is well-known that proline rich stretches present a strong tendency to form polyproline II (PPII) structures,[46] in agreement with the enhanced extension detected with the EOM. Interestingly, all the distance distributions involving PH domain are bimodal (Figure 5C) indicating that the PH−SH3 linker domain may switch between two conformations in solution. This is also

(43) Klein-Seetharaman, J.; Oikawa, M.; Grimshaw, S. B.; Wirmer, J.; Duchart, E.; Ueda, T.; Imoto, T.; Smith, L. J.; Dobson, C. M.; Schwalbe, H. *Science* **2002**, *295*, 1719−1722.
(44) Neet, K.; Hunter, T. *Genes Cells* **1996**, *1*, 147−169.
(45) Smith, C. I. E.; Islam, T. C.; Mattsson, P. T.; Mohamed, A.; Nore, B. F.; Vihinen, M. *BioEssays* **2001**, *23*, 436−446.
(46) Kelly, M. A.; Chellgren, B. W.; Rucker, A. L.; Troutman, J. M.; Fried, M. G.; Miller, A.-M.; Creamer, T. P. *Biochemistry* **2001**, *40*, 14376−40383.

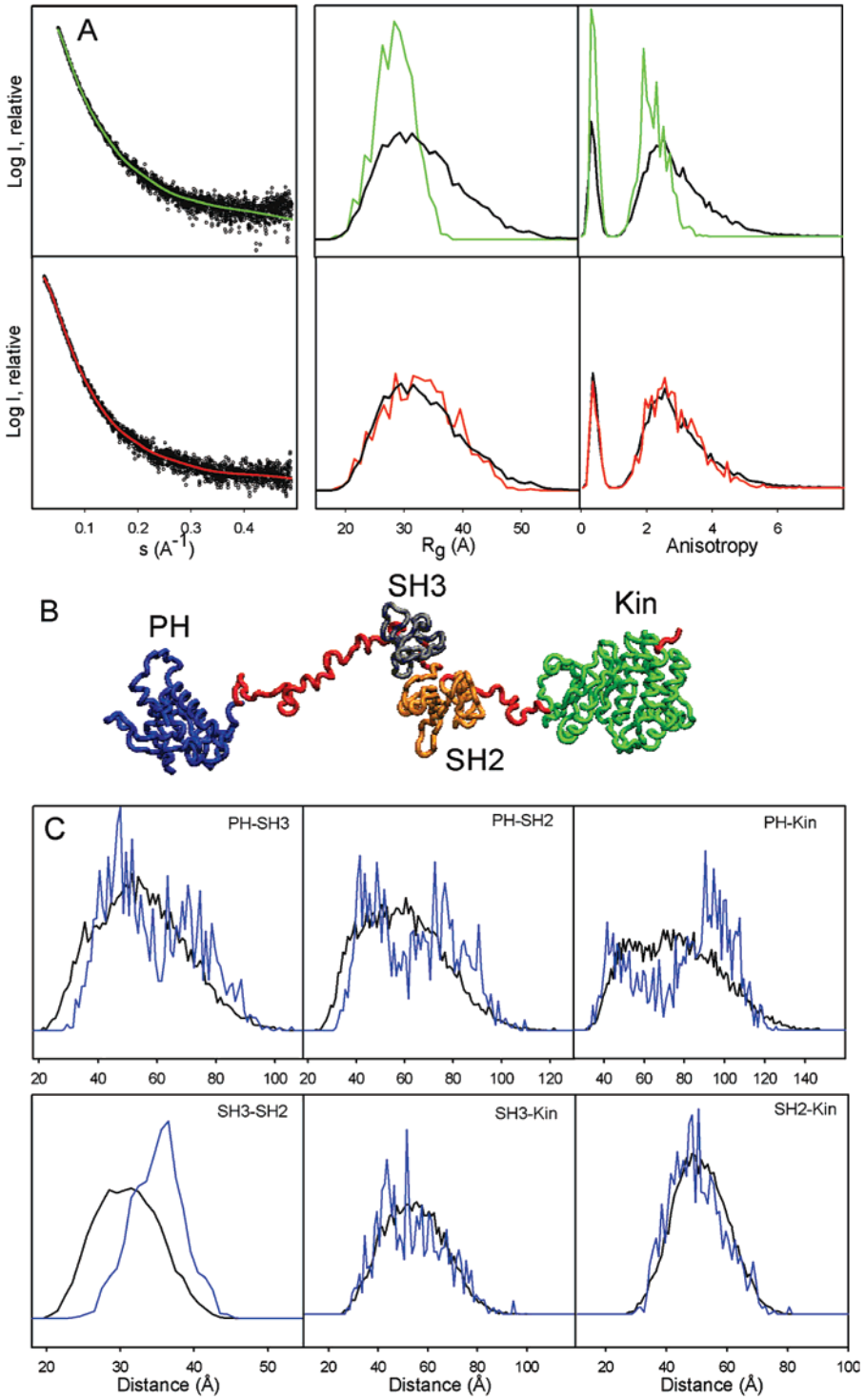

Figure 5. Application of the EOM to experimental data from denatured lysozyme and BTK. (A) EOM fits to the partially reduced (green) and fully reduced (red) lysozyme samples, and the R_g and P distributions from the pool (black) and optimized ensembles for both lysozyme samples. (B) Domain structure of BTK (see text). (C) Interdomain distance distributions for all domain pairs in BTK from the pool of 10 000 structures (black) and the EOM optimized ensemble (blue).

reflected in the bimodal shape of the R_g distribution (Supporting Information, Figure S4B) and may largely contribute to the observed flexibility of the entire BTK. The SH3 and SH2 domains are on average farther apart than expected from the pool simulation pointing to an extended conformation and rigidity of the short (8 residues) linker between them. This result is in agreement with NMR studies done in Abl and Fyn SH3−SH2 isolated constructions using spin relaxation and RDC experiments[47,48] where it was found that SH3 and SH2 domains are significantly coupled, although they still exhibit some interdomain flexibility. In addition, molecular dynamic simulations and mutational studies on c-Src-kinase showed the regulating role of the rigidity in the SH3−SH2 linker.[49] The

(47) Fushman, D.; Xu, R.; Cowburn, D. *Biochemistry* **1999**, *38*, 10225−10230.
(48) Ulmer, T. S.; Werner, J. M.; Campbell, I. D. *Structure* **2002**, *10*, 901−911.

extended conformation of the four domains and overall open conformation of BTK was also proposed in a previous paper based on the rigid body modeling.[26] A similar extended conformation has been described for the activated form of the c-Abl tyrosine kinase based on SAXS measurements for a construct where three anchoring points of the inactive form were removed.[50] Further investigations would be necessary to conclude whether the flexible scenario observed for the BTK is general for other tyrosine kinase proteins. In particular, the EOM application to the data from a construct encompassing the SH2−Kin domains in active and inactivated forms would allow one to analyze the alternative stable states of the kinase, the role of the linkers, and the interactions between SH2 and the N-lobe of the kinase.[47]

4. Conclusions

A method is proposed to quantitatively analyze flexible macromolecules in solution by SAXS. This approach is applicable to highly dynamic systems such as unstructured and multidomain proteins. The optimization method based on a genetic algorithm selects an ensemble of conformers describing the scattering properties of inherently dynamic macromolecules. Using synthetic data we have shown the capabilities of the approach to properly describe the distribution of overall parameters for these systems. EOM can be used to localize short and long-range structuring effects along the chain for partially unfolded proteins. The method is able to detect interdomain motions in multidomain proteins, which are known to play important roles in biomolecular function but remain hardly detectable by other techniques. EOM was further used on experimental data from an unstructured and a multidomain protein. Residual disulfide bridges were observed for a chemically unfolded lysozyme, which could be disrupted by increasing the concentration of the reducing agent. The presence of interdomain dynamics was detected for a multidomain BTK, and flexibility of the individual linkers was characterized.

The method provides 3D models, which can be further validated and refined against other experimental methods such as circular dichroism (CD), Fluorescence Resonance Energy Transfer (FRET), or hydrodynamic measurements. Especially NMR providing complementary local information about the chain properties (as opposed to the global information contained in SAXS data) appears extremely useful for the joint use with EOM. Indeed, the presence of conformational restrictions in the polypeptide chains detected by EOM can be confirmed using RDC measurements. Interdomain coupling, such as that observed between the SH3 and SH2 domains in the manuscript, can also be detected using spin relaxation experiments or RDCs. The next step in the further development of the EOM principle may be incorporation of the relevant data from other methods into the ensemble generation algorithm.

The EOM enlarges the capabilities of SAXS to study highly dynamic biomolecular systems, which are extremely difficult or impossible to characterize using high-resolution techniques. We expect that its application will shed light onto the biological roles played by unstructured and multidomain proteins, as well as the description of folding intermediates, and the delineation of the protein or RNA folding pathways. The executable code of the program for EOM analysis is available from the authors upon request. The proposed principle of the ensemble fitting of SAXS data is of general character and can be readily applied to nonbiological systems like polymers, catalysts, dendrimers, nanocomposites, etc.

Acknowledgment. The authors acknowledge funds from the EU design study SAXIER (contract RIDS No 011934), ANR NT05-4_42781 and Ramón y Cajal program to D.S., M.B., and P.B., respectively. P.B. acknowledges EMBO for the short-term fellowship.

Supporting Information Available: Figures displaying the fitted R_g distributions with increasing values of N; the $C_\alpha-C_\alpha$ average distance matrix distribution for the *control* polyalanine chain test case; the optimized R_g, D_{max}, and P distributions for the polyubiquitin chains; and the fittings to the experimental curves and the size and shape descriptor distributions for the BTK data analysis obtained from the EOM fitting. This material is available free of charge via the Internet at http://pubs.acs.org.

JA069124N

(49) Young, M. A.; Gonfloni, S.; Superti-Furga, G.; Roux, B.; Kuriyan, J. *Cell* **2001**, *105*, 115−126.
(50) Nagar, B.; Hantschel, O.; Seeliger, M.; Davies, J. M.; Weis, W. I.; Superti-Furga, G.; Kuriyan, J. *Mol. Cell* **2006**, *21*, 787−798.