Determination of Domain Structure of Proteins from X-Ray Solution Scattering

Dmitri I. Svergun,*† Maxim V. Petoukhov,†‡ and Michel H. J. Koch*
*European Molecular Biology Laboratory, Hamburg Outstation, D-22603 Hamburg, Germany; †Institute of Crystallography, Russian Academy of Sciences, Leninsky pr. 59, 117333 Moscow; and ‡Physics Department, Moscow State University, Moscow 117234, Russia

ABSTRACT An *ab initio* method for building structural models of proteins from x-ray solution scattering data is presented. Simulated annealing is employed to find a chain-compatible spatial distribution of dummy residues which fits the experimental scattering pattern up to a resolution of 0.5 nm. The efficiency of the method is illustrated by the *ab initio* reconstruction of models of several proteins, with known and unknown crystal structure, from experimental scattering data. The new method substantially improves the resolution and reliability of models derived from scattering data and makes solution scattering a useful technique in large-scale structural characterization of proteins.

INTRODUCTION

Structural studies in molecular biology face the challenge of the post-genomic era, with vast numbers of genome sequences becoming available. There is presently an interest in large-scale expression and purification of proteins for subsequent structure determination using x-ray crystallography and nuclear magnetic resonance (NMR) (Edwards et al., 2000). Obviously, x-ray crystallography requires crystals of good diffraction quality, whereas the application of NMR to structure determination is limited to small proteins. It is therefore clear that a significant fraction of proteins could not be analyzed using the two methods.

X-ray scattering from protein solutions is a structural method applicable to a broad range of conditions and sizes of macromolecules (Feigin and Svergun, 1987). In a scattering experiment, a dilute protein solution is exposed to x-rays, and the scattered intensity (I) is recorded as a function of the scattering angle. Because of the random positions and orientations of particles, the intensity is isotropic and proportional to the scattering from a single particle averaged over all orientations. The measured intensity after subtraction of solvent scattering is (Feigin and Svergun, 1987)

$$I(s) = \langle |A_a(\mathbf{s}) - \rho_s A_s(\mathbf{s}) + \delta\rho_b A_b(\mathbf{s})|^2 \rangle_\Omega \qquad (1)$$

where $A_a(\mathbf{s})$, $A_s(\mathbf{s})$, and $A_b(\mathbf{s})$ are, respectively, the scattering amplitudes from the particle in vacuo, from the excluded volume, and from the hydration shell. The electron density of the bulk solvent, ρ_s, may differ from that of the hydration shell, ρ_b, yielding a non-zero contrast of the shell $\delta\rho_b = (\rho_b - \rho_s)$. The scattering vector is $\mathbf{s} = (s, \Omega)$, where $s = 4\pi \sin\theta/\lambda$ denotes the momentum transfer, 2θ is the scattering angle, λ the wavelength of the radiation, and $\langle\ \rangle_\Omega$ stands for the average over the solid angle Ω in reciprocal space.

The main advantage of solution scattering is its ability to study the structure of native particles in nearly physiological conditions and to analyze structural changes in response to variations in external parameters. The price to pay is a dramatic loss of information caused by the spherical averaging in the scattering pattern. This has also led to the widespread opinion that solution scattering provides information only about overall size and anisometry of the solute particles.

Traditionally, only limited portions of the scattering patterns were used to construct three-dimensional models. At low angles (resolution of 2 to 3 nm), x-rays are insensitive to the internal structure and the scattering is essentially determined by the particle shape. Low resolution shape models were thus constructed on a trial-and-error basis using a priori information from other methods. More recently, *ab initio* modeling approaches have been developed representing the shape either by an angular envelope function (Svergun and Stuhrmann, 1991; Svergun et al., 1996), or as an ensemble of densely packed beads (Chacon et al., 1998; Svergun, 1999; Walther et al., 1999), and employing nonlinear minimization to fit the scattering data. In spite of the limited resolution and the restricted range of application of these methods, a number of successful studies (Bada et al., 2000; Chacon et al., 2000; Svergun et al., 2000a, b) have shown that solution scattering curves provide sufficient information to restore particle shapes *ab initio*. In contrast, higher resolution x-ray scattering patterns from proteins are even rarely measured, partly for experimental reasons but mainly because of lack of adequate methods to interpret the data in terms of structural models.

The level of structural detail provided by x-ray solution scattering can be qualitatively assessed by considering Fig. 1 presenting theoretical scattering curves from twenty-five proteins with different folds and molecular masses (MM) ranging from 10 to 300 kDa. The high resolution structures of the proteins were taken from the Protein Data Bank (PDB) (Bernstein et al., 1977) and the curves were com-

Received for publication 17 January 2001 and in final form 21 March 2001.

Address reprint requests to Dmitri Svergun, EMBL c/o DESY; Notkestrasse 85, D-22603 Hamburg, Germany. Tel: 49-40-89902-125; Fax: 40-40-89902-149, E-mail: svergun@embl-hamburg.de.

© 2001 by the Biophysical Society

0006-3495/01/06/2946/08 $2.00

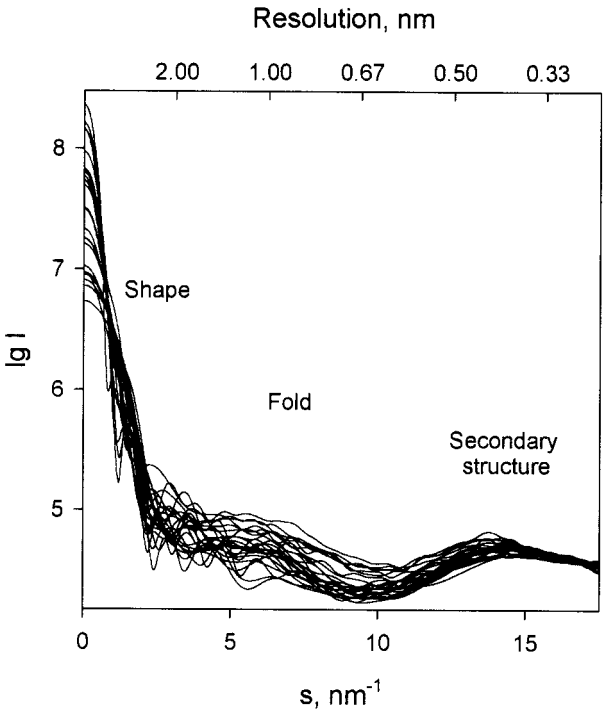

FIGURE 1 Theoretical x-ray solution scattering curves computed from atomic models of 25 different proteins. The upper axis displays the spatial resolution $\Delta = 2\pi/s$ and the text labels indicate the levels of structure organization characteristic of this resolution.

puted using the program CRYSOL (Svergun et al., 1995) assuming the bound solvent to be 10% denser than the bulk ($\delta\rho_b = 30\ e/\text{nm}^3$; Svergun et al., 1995, 1998). The patterns differ significantly up to $s \approx 12\ \text{nm}^{-1}$ but appear to converge at higher resolution. This suggests that solution scattering is not suitable for determination of secondary structure elements but also that the patterns contain information about the shape and fold of proteins up to a resolution of ~0.5 nm. In the present paper, a method is proposed to build structural models of proteins accounting for the entire scattering curve and not just for its initial portion corresponding to the shape scattering. The advantages of the new technique are illustrated by its application to several proteins, with both known and unknown crystal structures.

MATERIALS AND METHODS

Chain-compatible dummy residues model

Proteins typically consist of folded polypeptide chains composed of amino acid residues separated by ~0.38 nm between adjacent C_α atoms in the primary sequence. At a resolution of 0.5 nm, a protein structure can be considered as an assembly of dummy residues (DR) centered at the C_α positions. A three-dimensional model of the protein may therefore be constructed from solution scattering data by finding a chain-compatible spatial arrangement of the DRs that fits the experimental scattering pattern. That such a model adequately describes scattering patterns of proteins was verified by simulations. A protein with a known atomic structure containing N residues was represented by its C_α coordinates. Solvent-corrected spherically averaged scattering amplitudes from the amino acid residues were computed and weighted according to their abundance to yield the averaged residue form factor $f(s)$ in Fig. 2. To represent the bound solvent, the protein was surrounded by a hydration layer of thickness $\Delta r = 0.3$ nm as follows. A quasi-uniform grid of $M \approx N$ angular directions based on Fibonacci numbers was generated (Svergun, 1994). For each direction, the most distant residue was found and a dummy solvent atom was placed 0.5 nm outside the protein. The scattering intensity from the entire assembly of $K = N + M$ centers with coordinates $\mathbf{r}_i$ was calculated using the Debye formula (Debye, 1915)

$$I_{DR}(s) = \sum_{i=1}^{K} \sum_{j=1}^{K} g_i(s) g_j(s) \frac{\sin sr_{ij}}{sr_{ij}} \qquad (2)$$

where $g_i(s) = f(s)$ for a DR and $g_i(s) = (4\pi r_i^2/M)\ \Delta r\ \delta\rho_b$ for a solvent atom, and $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the distance between the ith and jth point. At higher angles, the intensities ($I_{DR}(s)$) were systematically lower than the theoretical ones ($I(s)$) because the internal structure of the residues is neglected in the DR model. The ratio $c(s) = I(s)/I_{DR}(s)$ computed for the 25 proteins in Fig. 1 was used to evaluate the average correction factor $\langle c(s)\rangle$ (Fig. 2) such that the product $\langle c(s)\rangle I_{DR}(s)$ yielded a good agreement with the theoretical scattering patterns up to $s = 15\ \text{nm}^{-1}$ for numerous proteins tested.

The use of the C_α positions allows to impose restrictions on the spatial arrangement of the DRs. In addition to the 0.38-nm separation along the chain, excluded volume effects and local interactions lead to a characteristic distribution of nearest neighbors. A histogram of the average number of C_α atoms in a 0.1-nm thick spherical shell surrounding a given C_α atom as a function of the shell radius $\langle N(R_k)\rangle$ for $0 < R_k < 1$ nm is presented in Fig. 3. It is clear that the histogram $N_{DR}(R_k)$ for a plausible chain-compatible DR model should be similar to $\langle N(R_k)\rangle$.

Minimization algorithm

A DR model of the protein structure can be retrieved from the scattering data as follows. Given the number of residues N (usually known from the

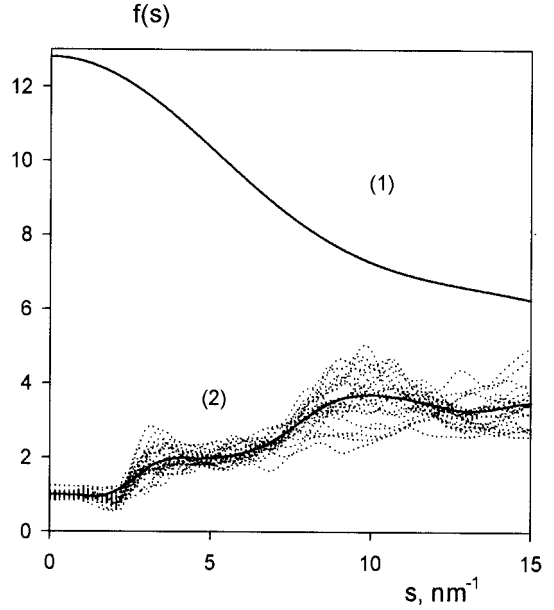

FIGURE 2 Averaged form factor of a residue (*1*) and the average correction factor (*2*). Dotted curves represent individual correction functions for the proteins in Fig. 1.

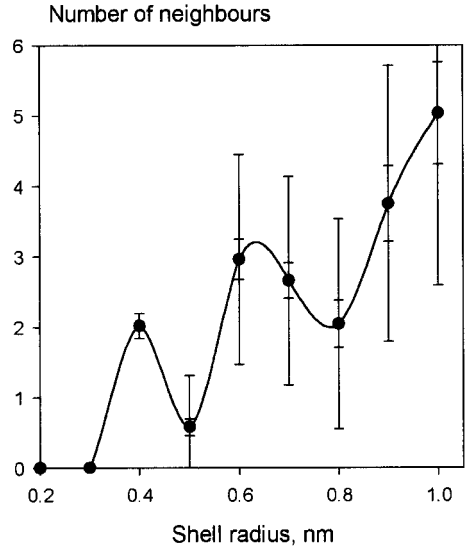

FIGURE 3 Histogram of an average number of C_α atoms in 0.1 nm thick spherical shells around a given C_α atom. Smaller error bars: variation of the averaged values over all proteins; larger error bars: averaged variation within one protein.

protein or translated DNA sequence), coordinates of DRs $\mathbf{r}_i$ are found which minimize a goal function $E(\mathbf{r}) = \chi^2 + \alpha P(\mathbf{r})$. Here, χ^2 is the discrepancy such that

$$\chi^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left[\frac{\langle c(s)\rangle I_{DR}(s_j) - cI_{exp}(s_j)}{\sigma(s_j)}\right]^2 \qquad (3)$$

where $I_{exp}(s)$ is the experimental intensity specified at n points s_j, $j = 1, \ldots n$, $\sigma(s_j)$ is the correspondent standard deviation, and c is a scaling coefficient. The penalty $P(\mathbf{r})$ has the form

$$P(\mathbf{r}) = \sum_k [W(R_k)(N_{DR}(R_k) - \langle N(R_k)\rangle)]^2 + G(\mathbf{r}) + [\max\{0, (|\mathbf{r}_c| - r_0)\}]^2 \qquad (4)$$

The first term in Eq. 4 imposes a protein-like nearest-neighbor distribution (the weights $W(R_k)$ are inversely proportional to the variations of $\langle N(R_k)\rangle$ in Fig. 3). The second term $G(\mathbf{r})$ ensures that the model is interconnected, i.e., each DR has at least one neighbor at a distance of 0.38 nm. All connected fragments (graphs) in the current DR model are found and $G(\mathbf{r})$ is computed as $\ln(N_G/N) \geq 0$, where N_G is the length of the longest graph. The third term keeps the center of mass of the DR model $\mathbf{r}_c$ close to the origin ($r_0 \approx 0.1\ D_{max}$ is the radius of a penalty-free zone and D_{max} is the maximum size of the protein). The penalty weight $\alpha > 0$ is selected such that $P(\mathbf{r})$ yields a significant contribution (~10 to 50%) to $E(\mathbf{r})$ at the end of the minimization.

The easiest way to construct a DR model would be to generate a random-walk C_α chain and let it fold to minimize $E(\mathbf{r})$ (this option will be discussed later). However, a better convergence was obtained for a restrained condensation of a gas of DRs within a spherical search volume (which resembles to some extent an ab initio phasing approach (Subbiah, 1991)). The simulated annealing (Kirkpatrick et al., 1983) condensation protocol is simple: (1) N coordinates $\mathbf{r}_i$ of the DRs are randomly generated within a sphere of radius $R = D_{max}/2 + r_0$, and M dummy solvent atoms are added as described. A value of the goal function $E(\mathbf{r})$ is computed and a high starting temperature (T_0) is selected. (2) A DR taken at random is relocated to an arbitrary point at a distance of 0.38 nm from another randomly selected DR (move from $\mathbf{r}$ *to* $\mathbf{r}'$*).* If the second location falls outside the search volume, repeat (2). (3) Positions of the solvent atoms are updated if necessary and a difference $\{\Delta E = E(\mathbf{r}') - E(\mathbf{r})\}$ is computed. If $\Delta E < 0$, the move is accepted; if $\Delta E \geq 0$, it is accepted with a probability $\exp(-\Delta E/T)$. (4) Steps (2 and 3) are repeated a sufficient number of times (N_T) to equilibrate the system, after which the temperature is lowered ($T' = 0.9T$). The system is cooled until no improvement in $E(\mathbf{r})$ is observed.

Computer program and testing

The above algorithm was implemented in a computer program GASBOR and computations were performed on simulated examples to select values of the parameters ($T_0 \approx 10^{-3}$, $N_T \approx 10^2 K$, $\alpha \approx 10^{-2}$) ensuring its convergence. As usual with simulated annealing, millions of function evaluations are required and it takes a prohibitively long time to fully re-compute $E(\mathbf{r})$ each time. Fortunately, both Debye's formula (Eq. 2) and the penalty (Eq. 4) are computed from the distances r_{ij}. A table of off-diagonal distances $\{r_{ij}, i > j\}$ is evaluated and saved at step (1) and later updated during steps (2 and 3). Moving one DR at a time greatly speeds up the computations (for example, on a 500 MHz Pentium III PC, the CPU time required to build the model of lysozyme below is reduced from ~100 h to ~1 h). The program is also able to take into account particle symmetry by generating symmetry mates for the DRs in the asymmetric unit (point groups P2 to P6 and P222 to P62 are currently supported).

In all tests on proteins presented below, the value of D_{max} was determined directly from the experimental data using the orthogonal expansion program ORTOGNOM (Svergun, 1993), and the DR models were restored without any a priori information except for the number of residues. The results of ~10 independent annealing runs were compared with each other and/or with the crystallographic model if the latter was available. Because the DR models had an arbitrary orientation and handedness, they were automatically aligned with the C_α coordinates in the crystal structures using the program SUPCOMB (Kozin and Svergun, 2001). This program minimizes a dissimilarity measure between two models as a normalized spatial discrepancy (NSD). For every point in the first model (DR or C_α), the minimum value among the distances between the point and all points in the second model is found, and the same is done for the points in the second model. These distances are added and normalized against the average distances between the neighboring points for the two models. In the context of the work described here, the physical meaning of NSD for the best superposition is that, on average, for any C_α atom in the atomic structure there is a DR at a distance of about NSD × 0.38 nm (and vice versa).

Scattering experiments and data treatment

The synchrotron radiation x-ray scattering data from hexokinase, yeast pyruvate decarboxylase and chitin-binding protein were collected following standard procedures using the × 33 camera (Boulin et al., 1986, 1988) of the European Molecular Biology Laboratory at Deutsches Elektronen Synchrotron, (Hamburg) and multiwire proportional chambers with delay line readout (Gabriel and Dauvergne, 1982). The data were recorded at different protein concentrations (2 to 25 mg/ml) for two or three sample-detector distances (3.9, 2.5, and 1.4 m) and the scattering patterns were merged to yield the final composite curves. A shortened camera setup (sample-detector distance, 0.5 m) was designed for the additional wide-angle measurements on lysozyme and bovine serum albumine. Details of the experimental procedures are given elsewhere (Koenig et al., 1993; Svergun et al., 2000a). The data processing (normalization, buffer subtraction, etc) involved statistical error propagation using the program SAPOKO (Svergun and Koch, unpublished).

RESULTS

After validation on simulated examples, the method was used to construct DR models of a number of proteins with

known and unknown crystal structure from experimental scattering data. For test purposes, wide-angle synchrotron x-ray scattering patterns were recorded from two readily available proteins, hen egg-white lysozyme (Merck Eurolab GmbH, Darmstadt, Germany) and bovine serum albumin (BSA; Sigma Chemical Corp., St. Louis, MO). These extensively characterized proteins are often used as model objects for structural and functional studies. The data were collected at protein concentrations ranging from 10 to 150 mg/ml and extrapolated to zero concentration of the solute protein for both small and wide angles to yield accurate scattering curves in the range up to $s = 13$ nm^{-1} as illustrated in Fig. 4.

The reconstructed models of lysozyme (MM = 14 kDa, 129 residues) superimposed with its atomic structure in the crystal (PDB entry 6lyz; Diamond, 1974) are presented in Fig. 5 *A*. The middle column displays the best (NSD = 0.75), the right column the worst (NSD = 0.85) solution out of 10 independent reconstructions. For comparison, the left column presents the low resolution shape of lysozyme restored ab initio using the program DAMMIN (Svergun, 1999). The latter model only fits the low angle portion of the scattering pattern (Fig. 4) whereas the new method neatly fits the entire curve and provides a significantly more detailed model of the protein structure.

An interesting result was obtained for BSA (MW = 67 kDa, 583 residues). Initial attempts at modeling the structure failed to fit the experimental scattering pattern in Fig. 4. To further understand this discrepancy, theoretical scattering curves were computed from the coordinates of two

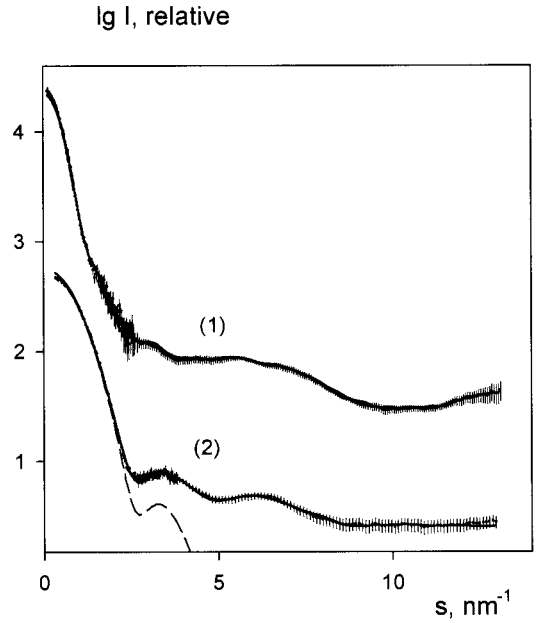

FIGURE 4 X-ray scattering from lysozyme (*1*) and BSA (*2*) (*dots with error bars*) and scattering from the DR models (*full lines*). For lysozyme, scattering from the low resolution shape model is also displayed (*dashed line*).

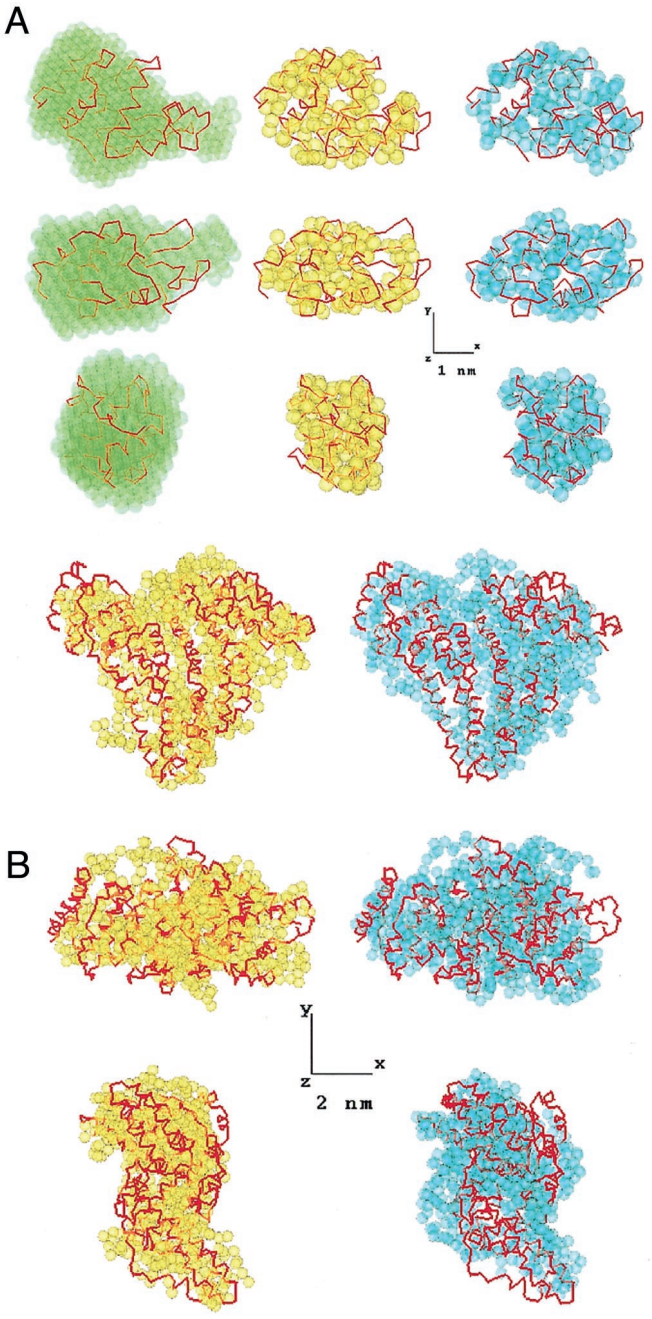

FIGURE 5 Atomic models of lysozyme (*A*) and ligand-bound HSA (*B*) superimposed with DR models from solution scattering. The atomic models are displayed as C_α chains, DR models as semi-transparent spheres (yellow and cyan models yield best and worst agreement with the atomic model, respectively). For lysozyme, a low resolution shape obtained by DAMMIN (Svergun; 1999) is displayed in green (*left column*). The center and bottom rows are rotated counterclockwise by 90^0 around X and Y, respectively. All three-dimensional models were displayed using the program ASSA (Kozin et al., 1997).

available crystallographic models of the human homologue, human serum albumin (HSA), which shares 90% sequence homology with the BSA. One crystal structure is that of the unliganded form of HSA (PDB entry 1ao6; Sugio et al.,

1999) while the other contains five fatty acids bound to the protein (entry 1bke; Curry et al., 1998) and reveals substantial conformational changes with respect to the former structure (r.m.s. displacement of C_α atoms, 0.46 nm; NSD = 0.78). The scattering pattern computed using CRYSOL with coordinates from the ligand-bound form of HSA yielded a better fit to the experimental data compared with the ligand-free form. In both cases, however, a negative contrast of the hydration shell $\delta\rho_b = -30$ *e*/nm^3 must be assumed to fit the data. As a result, the density of the solvation layer surrounding the protein in solution was, somewhat unexpectedly, 10% below that of the bulk rather than 10% larger as is generally observed for globular proteins. Given that the fatty acids are 30% lighter than water, this result suggests that bound hydrophobic fatty acids may be exposed on the surface of this protein (which is a major fatty-acid transport protein in the circulatory system), leading to an apparently less dense hydration shell. In a dozen reconstructions with a negative contrast of the dummy solvent atoms $\delta\rho_b = -30$ *e*/nm, the entire BSA scattering pattern can be neatly fitted by that of a DR model (Fig. 4). The best and worst reconstructions (NSD = 1.22 and 1.36, respectively) are superimposed in Fig. 5 *B* with the atomic model of the ligand-bound HSA (Curry et al., 1998). Note that the DR models yielded somewhat worse agreement with the crystal structure of unliganded HSA (Sugio et al., 1999) (the NSD was, on average, 3 to 5% higher) indicating that the method is able to distinguish between the two conformations.

In the following examples, DR models were constructed from scattering patterns recorded as part of ongoing research projects at the European Molecular Biology Laboratory, Hamburg, Germany. The data collection conditions were not optimized to yield high quality curves at wide angles, and the resolution was not better than 1 nm (Fig. 6). It was thus interesting to monitor the performance of the method against these data sets.

The biologically active subunit of yeast hexokinase is a homodimer (a monomer with MW = 53.8 kDa has 485 residues). The scattering patterns from solutions of monomeric and dimeric hexokinase are presented in Fig. 6, curves 1 and 2, respectively. A crystal structure of the monomer is available (PDB entry 1hkg; Bennett and Steitz, 1980) but the quaternary structure of the dimeric enzyme in solution is uncertain. Theoretical curves computed from the crystallographic dimers, both symmetric and asymmetric (Bennett and Steitz, 1980; Steitz et al., 1976), failed to fit the experimental scattering of dimeric hexokinase. The scattering patterns in Fig. 6 were used to independently build DR models of the monomer and of the dimer (assuming P2 symmetry for the latter). Several reconstructions of the monomer yielded a good agreement with the atomic model in the crystal (NSD between 1.04 and 1.14). A typical superposition with NSD = 1.08 is presented in Fig. 7 *A*, *left panel*. The symmetric model of the dimeric enzyme obtained without any assumption about the structure of the

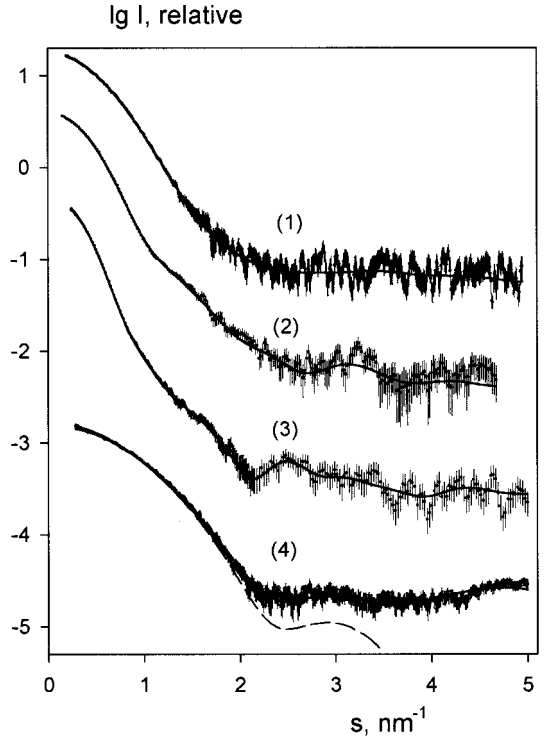

FIGURE 6 x-ray scattering patterns from monomeric (*1*) and dimeric (*2*) hexokinase, PDC (*3*) and CHB1 (*4*) (*dots with error bars*) and scattering from the reconstructed DR models (*full lines*). For chitin-binding protein, the scattering from a low resolution shape model is displayed (*dashed line*). The scattering patterns are displaced by one logarithmic unit for better visualization.

monomer displays two distinct monomers and suggests a clear way of forming the dimer in solution (Fig. 7 *A*, *right panel*).

Yeast pyruvate decarboxylase (PDC) provides another example of the use of symmetry restrictions. At low pH, PDC forms catalytically active tetramers with MM = 236 kDa and a total of 2148 residues (Koenig et al., 1993). The model of the tetrameric enzyme restored from the scattering pattern in Fig. 6 (curve 3) assuming P222 symmetry is presented in Fig. 7 *B* along with the crystal structure (PDB entry 1pvd; Arjunan et al., 1996). The comparison suggests that PDC in solution is more compact than in the crystal, partly as a result of an altered association between the dimers (corresponding to a relative tilt of about 10° with respect to the *x* axis in top orientation). This result is in an agreement with the model of tetrameric PDC in solution obtained earlier by rigid body refinement (Svergun et al., 2000c).

The final example deals with a protein of unknown atomic structure, a chitin-binding protein CHB1 from *Streptomyces* (MM = 18.7 kDa, 201 residues). The low resolution shape of the protein recently obtained (Svergun et al., 2000a) from the scattering pattern in Fig. 6 (*curve 4*) is displayed in Fig. 7 *C*, *left column*. Ten independent DR models were generated, and the two solutions that differ

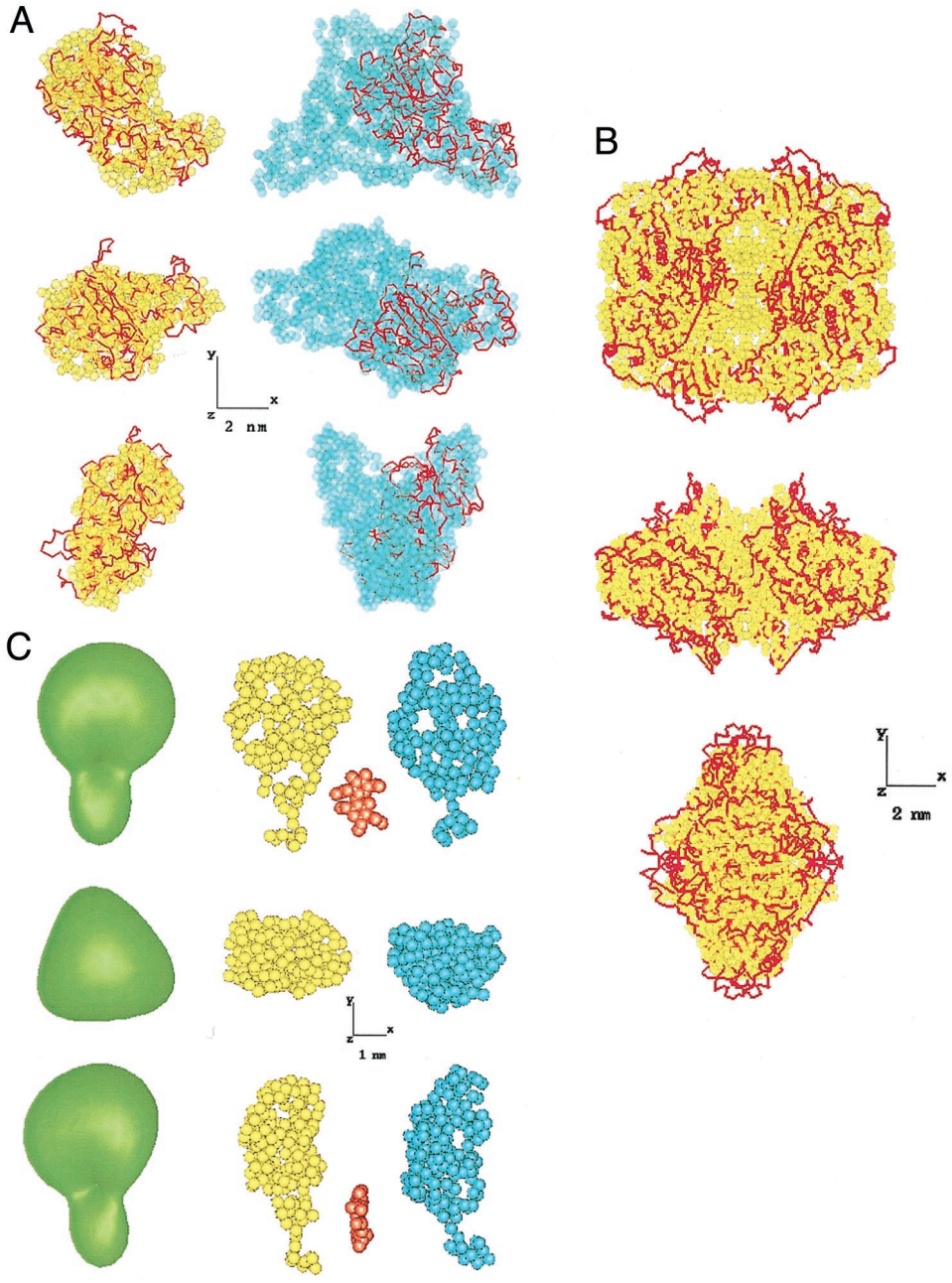

FIGURE 7 (*A*) Comparison of the atomic model of monomeric hexokinase and the DR models of the monomeric (*left*) and dimeric (*right*) hexokinase. (*B*) Atomic model of tetrameric PDC superimposed with the DR model. (*C*) Ab initio models of CHB1. *Left column*, low resolution envelope obtained by the program SASHA (Svergun et al., 1996); *center* and *right columns*, two most different DR models. The atomic models are displayed as C_α chains, DR models as spheres. The center and bottom rows are rotated as in Fig. 5. A space-filling representation of α-chitin is presented in (*C*) (*red*) near the foot-like protuberance of CHB1.

most from each other (NSD = 0.95) are presented in Fig. 7 *C*, *middle* and *right columns*. Even these two most different models are fairly similar and, in particular, display a characteristic foot-like protuberance. Comparison with a space-filling representation of the atomic structure of α-chitin (taken from the Cambridge Crystallographic Database, http://www.cmbi.kun.nl/) suggests this protuberance as a plausible site for binding the chitin molecule.

DISCUSSION

It is clear that the method does not yield a unique solution (spatial distribution of DRs), but rather provides a manifold of configurations corresponding to virtually the same scattering pattern (e.g., yellow and cyan DR models in Fig. 5 *A* and *B*, and Fig. 7 *C*). Calculations on simulated and experimental scattering patterns indicate that the differences be-

tween the DR models are substantially smaller than those observed for the low resolution shape determination using densely packed beads (Svergun, 1999). This is not surprising given that the DR models use fewer independent parameters to fit much larger portions of the scattering pattern than do the bead modeling programs (Chacon et al., 1998; Svergun, 1999; Walther et al., 1999). The variations between the DR models preserve the domain structure of the protein, and, in analogy to NMR studies, an average (most probable) model can be generated by merging the results of independent simulated annealing runs.

Could the method yield more than a detailed shape and domain structure of a protein? Although the DR models do not directly provide the tertiary structure, the map of approximate C_α positions can be incorporated as a constraint in protein folding prediction methods. We have also attempted to directly restore the protein fold from an x-ray scattering pattern. A C_α chain was folded accounting for the primary and secondary structure, backbone angles distribution (Kleywegt, 1997), hydrophobicity (Huang et al., 1995), and interaction potentials between residues (Miyazawa and Jernigan, 1999; Thomas and Dill, 1996). Following a random-walk simulated annealing protocol, various native-like folds could be generated fitting the data and satisfying all constraints (i.e., having a free energy much lower than that of the native protein). These results suggest, not unexpectedly, that there is insufficient information to reconstruct a protein fold described by C_α coordinates alone using a single x-ray pattern. Currently, the folding algorithm is extended to account for the centers of the side chains (Guo et al., 1995) and to incorporate additional information about the internal structure provided by contrast variation using isotopic H/D exchange in neutron scattering experiments. It should also be noted that the principle of DR modeling can be used to characterize the structure of unfolded or partially folded proteins and to construct low resolution ab initio models in protein crystallography.

The DR modeling accounting for the complete scattering pattern already yields substantially more reliable and higher resolution models than previous methods, and the present approach has potential for future development. The new modeling technique makes x-ray scattering, which is free from major limitations of crystallography and NMR, a useful option for a large-scale structural characterization of proteins in solution. .

The authors thank K. Brown for valuable comments on the manuscript, M. Malfois for assistance with measurements, V. Renkwitz for modifying the camera, H. Bartunik, S. Koenig and G. Grueber for providing the protein samples.

The work was supported by International Association for the Promotion of Cooperation with Scientists from the Independent States of the Former Soviet Union, Grants 00-243 and YSF 00-50.

The executable codes of the program GASBOR for IBM-PC and major UNIX platforms are available from http://www.embl-hamburg.de/ExternalInfo/Research/Sax.

REFERENCES

Arjunan, P., T. Umland, F. Dyda, S. Swaminathan, W. Furey, M. Sax, B. Farrenkopf, Y. Gao, D. Zhang, and F. Jordan. 1996. Crystal structure of the thiamin diphosphate-dependent enzyme pyruvate decarboxylase from the yeast Saccharomyces cerevisiae at 2.3 A resolution. *J. Mol. Biol.* 256:590–600.

Bada, M., D. Walther, B. Arcangioli, S. Doniach, and M. Delarue. 2000. Solution structural studies and low-resolution model of the Schizosaccharomyces pombe sap1 protein. *J. Mol. Biol.* 300:563–574.

Bennett, W. S., Jr., and T. A. Steitz. 1980. Structure of a complex between yeast hexokinase A and glucose. II. Detailed comparisons of conformation and active site configuration with the native hexokinase B monomer and dimer. *J. Mol. Biol.* 140:211–230.

Bernstein, F. C., T. F. Koetzle, G. J. B. Williams, E. F. Meyer, Jr., M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542.

Boulin, C. J., R. Kempf, A. Gabriel, and M. H. J. Koch. 1988. Data acquisition systems for linear and area X-ray detectors using delay line readout. *Nucl. Instrum. Meth. A* 269:312–320.

Boulin, C., R. Kempf, M. H. J. Koch, and S. M. McLaughlin. 1986. Data appraisal, evaluation and display for synchrotron radiation experiments: hardware and software. *Nucl. Instrum. Meth. A* 249:399–407.

Chacon, P., J. F. Diaz, F. Moran, and J. M. Andreu. 2000. Reconstruction of protein form with X-ray solution scattering and a genetic algorithm. *J. Mol. Biol.* 299:1289–1302.

Chacon, P., F. Moran, J. F. Diaz, E. Pantos, and J. M. Andreu. 1998. Low-resolution structures of proteins in solution retrieved from X-ray scattering with a genetic algorithm. *Biophys. J.* 74:2760–2775.

Curry, S., H. Mandelkow, P. Brick, and N. Franks. 1998. Crystal structure of human serum albumin complexed with fatty acid reveals an asymmetric distribution of binding sites. *Nat. Struct. Biol.* 5:827–835.

Debye, P. 1915. Zerstreuung von Roentgenstrahlen. *Ann. Physik* 46: 809–823.

Diamond, R. 1974. Real-space refinement of the structure of hen egg-white lysozyme. *J. Mol. Biol.* 82:371–391.

Edwards, A. M., C. H. Arrowsmith, D. Christendat, A. Dharamsi, J. D. Friesen, J. F. Greenblatt, and M. Vedadi. 2000. Protein production: feeding the crystallographers and NMR spectroscopists. *Nat. Struct. Biol.* 7(Suppl):970–972.

Feigin, L. A., and D. I. Svergun. 1987. Structure Analysis by Small-Angle X-Ray and Neutron Scattering. New York: Plenum Press. pp 335

Gabriel, A., and F. Dauvergne. . 1982. The localization method used at EMBL. *Nucl. Instrum. Meth.* 201:223–224.

Guo, D. Y., G. D. Smith, J. F. Griffin, and D. A. Langs. 1995. Use of globic scattering factors for protein structures at low resolution. *Acta Cryst.* A51:945–947.

Huang, E. S., S. Sibbiah, and M. Levitt. 1995. Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.* 252: 709–720.

Kirkpatrick, S., C. D. Gelatt, Jr., and M. P. Vecci. 1983. Optimization by simulated annealing. *Science* 220:671–680.

Kleywegt, G. J. 1997. Validation of protein models from Calpha coordinates alone. *J. Mol. Biol.* 273:371–376.

Koenig, S., D. Svergun, M. H. Koch, G. Hubner, and A. Schellenberger. 1993. The influence of the effectors of yeast pyruvate decarboxylase (PDC) on the conformation of the dimers and tetramers and their pH-dependent equilibrium. *Eur. Biophys. J.* 22:185–194.

Kozin, M. B., and D. I. Svergun. 2001. Automated matching of high- and low-resolution structural models. *J. Appl. Crystallogr.* 34:33–41.

Kozin, M. B., V. V. Volkov, and D. I. Svergun. 1997. ASSA: a program for three-dimensional rendering in solution scattering from biopolymers. *J. Appl. Crystallogr.* 30:811–815.

Miyazawa, S., and R. L. Jernigan. 1999. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins.* 34:49–68.

Steitz, T. A., R. J. Fletterick, W. F. Anderson, and C. M. Anderson. 1976. High resolution x-ray structure of yeast hexokinase, an allosteric protein exhibiting a non-symmetric arrangement of subunits. *J. Mol. Biol.* 104: 197–122.

Subbiah, S. 1991. Low-resolution real-space envelopes: an approach to the ab initio macromolecular phase problem. *Science.* 252:128–133.

Sugio, S., A. Kashima, S. Mochizuki, M. Noda, and K. Kobayashi. 1999. Crystal structure of human serum albumin at 2.5 A resolution. *Protein Eng.* 12:439–446.

Svergun, D. I. 1993. A direct indirect method of small-angle scattering data treatment. *J. Appl. Crystallogr.* 26:258–267.

Svergun, D. I. 1994. Solution scattering from biopolymers: advanced contrast variation data analysis. *Acta Crystallogr.* A50:391–402.

Svergun, D. I. 1999. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.* 76:2879–2886.

Svergun, D. I., C. Barberato, and M. H. J. Koch. 1995. CRYSOL - a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* 28:768–773.

Svergun, D. I., A. Becirevic, H. Schrempf, M. H. Koch, and G. Grueber. 2000a. Solution structure and conformational changes of the Streptomyces chitin-binding protein (CHB1). *Biochemistry.* 39:10677–10683.

Svergun, D. I., M. Malfois, M. H. Koch, S. R. Wigneshweraraj, and M. Buck. 2000b. Low resolution structure of the sigma54 transcription factor revealed by X-ray solution scattering. *J. Biol. Chem.* 275: 4210–4214.

Svergun, D. I., M. V. Petoukhov, M. H. Koch, and S. Konig. 2000c. Crystal versus solution structures of thiamine diphosphate-dependent enzymes. *J. Biol. Chem.* 275:297–302.

Svergun, D. I., S. Richard, M. H. Koch, Z. Sayers, S. Kuprin, and G. Zaccai. 1998. Protein hydration in solution: experimental observation by x-ray and neutron scattering. *Proc. Natl. Acad. Sci. U. S. A.* 95: 2267–2272.

Svergun, D. I., and H. B. Stuhrmann. 1991. New developments in direct shape determination from small-angle scattering 1. Theory and model calculations. *Acta Crystallogr.* A47:736–744.

Svergun, D. I., V. V. Volkov, M. B. Kozin, and H. B. Stuhrmann. 1996. New developments in direct shape determination from small-angle scattering 2. Uniqueness. *Acta Crystallogr.* A52:419–426.

Thomas, P. D., and K. A. Dill. 1996. An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl. Acad. Sci. U. S. A.* 93:11628–11633.

Walther, D., F. E. Cohen, and S. Doniach. 1999. Reconstruction of low-resolution three-dimensional density maps from one-dimensional small-angle X-ray solution scattering data for biomolecules. *J. Appl. Crystallogr.* 33:350–363.