

Addition of Missing Loops and Domains to Protein Models by X-Ray Solution Scattering

Maxim V. Petoukhov,^{*†} Nigel A. J. Eady,[‡] Katherine A. Brown,[‡] and Dmitri I. Svergun^{*§}

^{*}European Molecular Biology Laboratory, Hamburg Outstation, D-22603 Hamburg, Germany; [†]Physics Department, Moscow State University, 117234 Moscow, Russia; [‡]Department of Biological Sciences, Centre for Molecular Microbiology and Infection, Imperial College of Science, Technology and Medicine, London SW7 2AY, United Kingdom; and [§]Institute of Crystallography, Russian Academy of Sciences, 117333 Moscow, Russia

ABSTRACT Inherent flexibility and conformational heterogeneity in proteins can often result in the absence of loops and even entire domains in structures determined by x-ray crystallographic or NMR methods. X-ray solution scattering offers the possibility of obtaining complementary information regarding the structures of these disordered protein regions. Methods are presented for adding missing loops or domains by fixing a known structure and building the unknown regions to fit the experimental scattering data obtained from the entire particle. Simulated annealing was used to minimize a scoring function containing the discrepancy between the experimental and calculated patterns and the relevant penalty terms. In low-resolution models where interface location between known and unknown parts is not available, a gas of dummy residues represents the missing domain. In high-resolution models where the interface is known, loops or domains are represented as interconnected chains (or ensembles of residues with spring forces between the C_α atoms), attached to known position(s) in the available structure. Native-like folds of missing fragments can be obtained by imposing residue-specific constraints. After validation in simulated examples, the methods have been applied to add missing loops or domains to several proteins where partial structures were available.

INTRODUCTION

Protein function is related not only to the three-dimensional arrangement of polypeptide chains but also to their intrinsic mobility. Techniques such as x-ray crystallography and NMR can yield high-resolution information regarding the positions of individual atomic groups within a macromolecule, but flexible or disordered regions may appear to be absent. Such regions may be of significant functional importance and can include, for example, a loop in an enzyme active site, a receptor-binding motif, or an antigenic epitope. In large multi-domain proteins, inherent flexibility between domains can prevent successful crystallization, and in these cases crystallographic or NMR data may be limited to studies of individual domains produced by genetic or proteolytic methods. However, it is apparent that complementary approaches are required to analyze the structure of intact multi-domain proteins and assemblies, especially in view of recent initiatives aimed at large-scale expression and purification of proteins for subsequent structure determination (e.g., Edwards et al., 2000).

One such approach is small-angle x-ray scattering (SAXS) (Feigin and Svergun, 1987). This technique can yield structural information about macromolecules in solution with proteins from as small as 6 kDa (e.g., Sayers et al., 1999) to large macromolecular complexes such as the ribosome (Svergun and Nierhaus, 2000). SAXS patterns result

from an average of the scattering from the entire ensemble of randomly oriented particles in the sample, and this lowers the resolution of the method. Nevertheless, in contrast to x-ray crystallographic analysis where flexible regions of a structure may result in poorly interpretable electron density, solution scattering patterns are sensitive to these disordered regions, yielding information about their average conformation. The SAXS method can thus provide (at low resolution) information complementary to that of crystallography and NMR. Solution scattering also permits one to construct models of multi-domain proteins and macromolecular complexes from high-resolution structures of individual domains or subunits. Rigid body modeling, successfully used by different groups (Ashton et al., 1997; Krueger et al., 1997; Svergun et al., 1997, 1998a, 2000), is an effective way to characterize complex structures. The methods to compute solution scattering patterns accurately from atomic models and rapidly evaluate scattering from complex particles are now well established (Svergun, 1994, 1995, 1998b). These methods coupled with three-dimensional display and manipulation programs allow interactive or automated searches of positional parameters to fit the experimental scattering from the complex (Konarev et al., 2001; Kozin and Svergun, 2000).

In cases where the portions of a macromolecule or complex lack a three-dimensional structure description, alternative methods are required (beyond rigid-body refinement) to generate a model. For example, the known part of the structure (either high- or low-resolution model) can be fixed, and missing portions, such as the disordered loops or domains, can be then modeled to fit the experimental scattering data obtained from the intact particle. In the present paper, a recently proposed dummy-residues model

Submitted June 7, 2002, and accepted for publication July 17, 2002.

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

© 2002 by the Biophysical Society

0006-3495/02/12/3113/13 \$2.00

(Svergun et al., 2001a) is further developed to construct the algorithms for complementing high- and low-resolution partial models of protein structures. Simulated annealing is used to minimize a scoring function containing the discrepancy between the experimental and calculated patterns and relevant penalty terms. Where applicable, information about the primary and secondary structure is used to restrain the model and to provide native-like conformations of the missing structural fragments.

After validation in simulated examples, the potential of this approach has been explored using three model systems. These methods have first been applied to develop models for small contiguous loops (~30–35 residues), which are absent in the crystal structures of a *Drosophila* motor protein (Kozielski et al., 1999) and the R2 protein of *Escherichia coli* ribonucleotide reductase (Logan et al., 1996). Second, reconstruction of an entire missing domain has been attempted using experimentally observed scattering data from a fusion protein. This fusion consists of *Schistosoma japonicum* glutathione *S*-transferase (GST) and *E. coli* dihydrofolate reductase (DHFR). Although a few examples exist of crystal structures of GST fused with relatively small fusion fragments (Lim et al., 1994; Ware et al., 1999; Zhang et al., 1998), proteins of interest are often isolated by proteolytic digestion of the linker region (Nagai and Thogersen, 1984) before structural analysis. Therefore, little is known about the conformation of the linker region or the structure of a globular protein fused to GST. Using SAXS and the reconstruction methods, new information regarding domain and linker orientations in this popular fusion system is presented. In combination, analysis of these model systems provides an insight into the possible scope of these reconstruction techniques, from small loops to multi-domain assemblies.

MATERIALS AND METHODS

Dummy-residues approach

The scattering intensity $I(s)$ from a dilute monodispersed solution of macromolecules is an isotropic function depending on the modulus of the scattering vector $s = (s, \Omega)$, where Ω is the solid angle in reciprocal space, $s = (4\pi/\lambda)\sin\theta$, λ is the wavelength, and 2θ is the scattering angle. The x-ray scattering intensity is proportional to the scattering from a single particle averaged over all orientations and can be expressed as:

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

where $A_a(s)$, $A_s(s)$, and $A_b(s)$ 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 nonzero contrast for the shell $\delta\rho_b = \rho_b - \rho_s$ (Svergun et al., 1995).

In the method (Svergun et al., 2001a), the protein structure is represented by an ensemble of dummy residues (DRs) centered at the positions of virtual C_α atoms. A simulated annealing (SA) procedure (Kirkpatrick et al., 1983) is used to find DR positions by fitting the experimental data and simultaneously providing a chain-compatible structure. This is achieved by minimizing a scoring function $E(r) = \chi^2 + \sum \alpha_i P_i(r)$ where χ^2 is the

discrepancy between the experimental and calculated scattering patterns and the penalties $P_i(r)$ restrain the solution to ensure a chain-compatible arrangement of the DRs. The weights α_i are selected in such a way that the total penalty $\sum \alpha_i P_i(r)$ yields a significant contribution (~10–50%) to $E(r)$ at the end of the minimization. It has been demonstrated that the DR representation adequately represents solution scattering patterns up to a resolution of 0.5 nm and that the method allows an ab initio restoration of domain structures of proteins (Svergun et al., 2001a). In the present paper, the DR approach is extended to build missing domains or loops around a known part of the protein structure. Depending on the information available, four models are considered, differing by the representation of the missing portion of the structure and by the set of constraints.

Computation of the scattering intensity

The scattering intensity $I_{\text{mod}}(s)$ from a protein model consisting of N DRs positioned at \mathbf{r}_i is calculated as described (Svergun et al., 2001a) using Debye's formula (Debye, 1915):

$$I_{\text{mod}}(s) = \sum_{i=1}^K \sum_{j=1}^K g_i(s) g_j(s) \frac{\sin sr_{ij}}{sr_{ij}}, \quad (2)$$

where $K = N + M$ and M is the number of dummy solvent atoms in the hydration shell of the particle, $g_i(s)$ is the form factor of i th residue or solvent atom, and $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the distance between the i th and j th point. To generate the hydration shell of thickness $\Delta r = 0.3$ nm, the most distant residue is found along each direction of a quasi-uniform angular grid of $M \approx N$ vectors, and a solvent atom with the form factor $g_i(s) = (4\pi r_i^2 / M) \Delta r \delta\rho_b$ is placed 0.5 nm outside the protein. Following previously published methods (Svergun et al., 1995, 1998b), the contrast of the hydration shell is taken to be 30 e/nm³. Solvent-corrected spherically averaged scattering intensities from the amino acid residues are weighted according to their abundance in proteins, yielding an average residue form factor $\langle g(s) \rangle$ (Fig. 1 a). The DR-form factor $g_i(s) = \langle g(s) \rangle$ is taken when using models that do not account for the primary structure of the protein. To account for the internal residue structure, a correction factor $\langle c(s) \rangle$ is introduced. More than 100 proteins with known structures were taken from the Protein Data Bank (PDB) (Bernstein et al., 1977), and the scattering intensities of the full-atom representations of $I_{\text{full}}(s)$ were computed by the program CRY SOL (Svergun et al., 1995) where the average ratio $\langle c(s) \rangle = \langle I_{\text{full}}(s) / I_{\text{mod}}(s) \rangle$ is evaluated over the ensemble (Fig. 1 b). As demonstrated in Svergun et al. (2001a), the function $\langle c(s) \rangle I_{\text{mod}}(s)$ yields an adequate representation of the scattering pattern of a protein up to a resolution of 0.5 nm.

For the DR models accounting for the primary structure, form factors of the individual residues (Fig. 1 a) are computed by averaging the form factors of different conformations in the PDB files. The correction factor calculated as described above for the case of DRs is presented in Fig. 1 b. Simulations performed on proteins with known structures demonstrated, not unexpectedly, that the use of individual residues yields an even better accuracy than the dummy residues.

The discrepancy χ^2 between the calculated curve and experimental data $I_{\text{exp}}(s)$ measured at n points s_j , $j = 1, \dots, n$ is computed as:

$$\chi^2 = \frac{1}{n-1} \sum_{j=1}^n \left[\frac{\mu \langle c(s) \rangle I_{\text{mod}}(s_j) - I_{\text{exp}}(s_j)}{\sigma(s_j)} \right]^2, \quad (3)$$

where $\sigma(s_j)$ are the experimental errors and μ is an overall scaling coefficient.

Simulated annealing protocol

For all the models described here, SA (Kirkpatrick et al., 1983) is used for global minimization of the scoring function. The main aim of this method

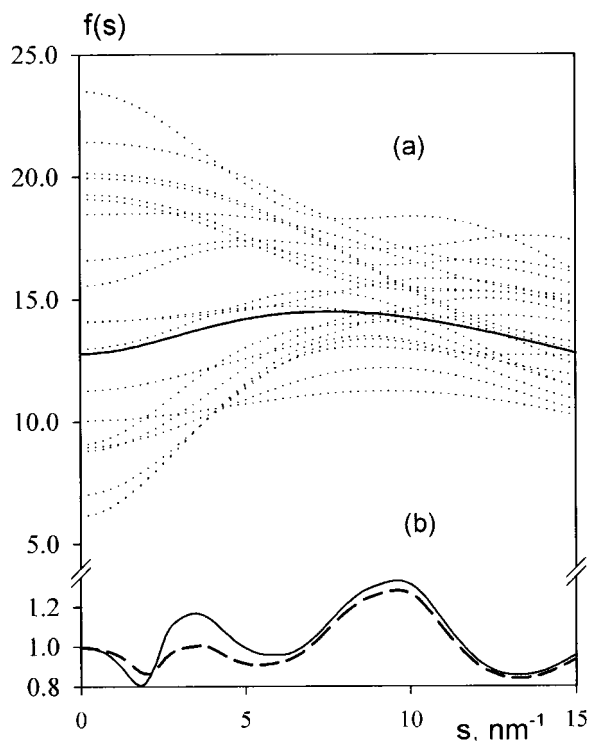


FIGURE 1 (a) Form factors of the 20 amino acid residues (\cdots) and an average form factor (—); (b) Average correction factor $\langle c(s) \rangle$ for the intensity computation using individual (— — —) and dummy residues (—).

is to perform random modifications of the system (i.e., of the current residue arrangement) by always moving to configurations that decrease the scoring function $E(\mathbf{r})$ but to also occasionally move to configurations that increase $E(\mathbf{r})$. The probability of accepting the latter moves decreases in the course of the minimization (the system is cooled). At the beginning, the temperature is high and the changes are almost random, whereas at the end a configuration with (nearly) minimum $E(\mathbf{r})$ is reached. The algorithm is implemented in its faster simulated quenching (Ingber, 1993; Press et al., 1992) version as follows. 1) The known part of the structure is loaded and moved to the origin and remains fixed during minimization. The rest of the structure is then generated depending on the model used. A value of the goal function $E(\mathbf{r})$ is computed and a high starting temperature T_0 is selected. 2) A random modification (move from \mathbf{r} to \mathbf{r}') of the system is performed (specific ways of generation and modification of the system are considered below). 3) Positions of the solvent atoms accounting for the border solvent layer 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 $\Delta E \geq 0$, the move 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, and the temperature is lowered ($T' = \eta T$, $\eta < 1$) afterwards. The system is cooled until no improvement in $E(\mathbf{r})$ is observed.

Types of dummy-residue models

Free dummy-residues model

This model is an extension of the original DR model (Svergun et al., 2001a) and can be used when the location of the interface between the known and missing portions of the structure is unknown. This usually takes place when a low-resolution model represents the known portion, although

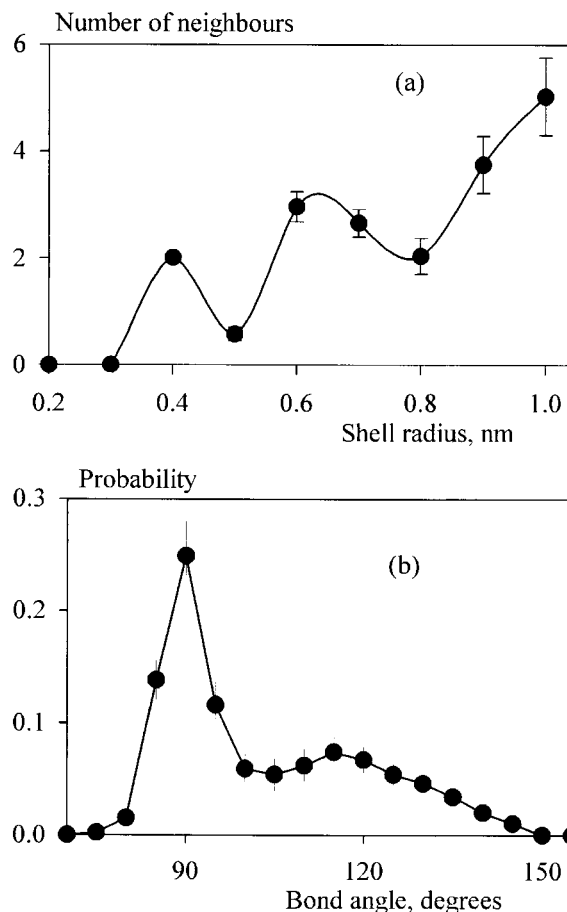


FIGURE 2 Histograms of the average distributions of nearest neighbors (a) and of the C_α - C_α - C_α bond angles (b) in a polypeptide chain.

high-resolution models can also be used. The known part of the structure is fixed and the unknown part is represented as a gas of free DRs within a search volume (the latter is a sphere with a diameter equal to the maximum size D_{\max} of the entire particle). The numbers of residues in the fixed and variable parts (N_0 and N_D , respectively) are assumed to be available a priori, whereas the value of D_{\max} can be determined from the solution scattering pattern of the particle. The scoring function is:

$$E(\mathbf{r}) = RF^2 + \alpha_{\text{dst}} P_{\text{dst}} + \alpha_{\text{con}}^1 P_{\text{con}}^1 + \alpha_{\text{con}}^2 P_{\text{con}}^2 + \alpha_{\text{gyr}} P_{\text{gyr}}$$

$$RF^2 = (n - 1) \chi^2 \left\{ \sum_{j=1}^n \left[\frac{I_{\text{exp}}(s_j)}{\sigma(s_j)} \right]^2 \right\}^{-1} \quad (4)$$

Here and below, a normalized R -factor RF will enter the scoring function instead of the discrepancy to facilitate the choice of the SA parameters and the penalty weights.

The first penalty ensuring a protein-like distribution of the nearest neighbors in the model has the form introduced in Svergun et al. (2001a):

$$P_{\text{dst}} = \sum_k [W(R_k)(N_{\text{mod}}(R_k) - \langle N(R_k) \rangle)]^2, \quad (5)$$

where $\langle N(R_k) \rangle$ is 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 for $0 < R_k < 1$ nm observed for real proteins. $N_{\text{mod}}(R_k)$ is such a histogram for the model, and the weights $W(R_k)$ are inversely proportional to the variations of $\langle N(R_k) \rangle$ (Fig. 2 a). The summation in Eq.

5 is performed over the DRs in the variable portion of the model.

The second penalty requires the model to be interconnected so that each residue has at least one neighbor at a distance of 0.38 nm:

$$P_{\text{con}} = \ln(N/N_1), \quad (6)$$

where N_1 is the length of the longest interconnected fragment of the model. This penalty is applied twice: once for the entire structure and separately for the variable part.

The third penalty restrains the space occupied by the variable part whose radius of gyration R_g can approximately be estimated as $R_g^{\text{est}} \sim 3\sqrt[3]{N_D}$. The penalty has the form:

$$P_{\text{gyr}} = ((R_g^{\text{mod}} - R_g^{\text{est}})/R_g^{\text{est}})^2, \quad (7)$$

where R_g^{mod} is the radius of gyration of the variable portion.

Initially, the missing DRs are randomly positioned inside the search volume but outside the fixed portion of the model. A single SA step involves relocation of a randomly selected residue to a point at a distance of 0.38 nm from another randomly selected residue in the variable portion. The penalties force the variable portion to condense to a compact chain-compatible model, and the procedure (implemented in the program CREDO) is best suited for generating low-resolution models of missing domains without using information about primary and secondary structures.

Dummy-residues model with spring forces between neighbors

This approach builds chains of DRs attached to given point(s) or residue(s) in the known part of the structure. In contrast to the previous model, it is explicitly required that the i th DR be separated by 0.38 nm from the $(i + 1)$ th one. The scoring function is:

$$E(\mathbf{r}) = RF^2 + \alpha_{\text{dst}}P_{\text{dst}} + \alpha_{\text{spr}}P_{\text{spr}} + \alpha_{\text{gyr}}P_{\text{gyr}}, \quad (8)$$

where P_{dst} and P_{gyr} are the same penalties as in Eq. 7, but instead of the disconnectivity penalty (P_{con}), spring potentials P_{spr} between the neighboring DRs are used:

$$P_{\text{spr}}(\mathbf{r}) = \frac{1}{ND_{\text{max}}^2} \sum_{i=1}^{N-1} (|\mathbf{r}(i+1) - \mathbf{r}(i)| - 0.38)^2 \quad (9)$$

The first (and, if appropriate, the last) DR(s) in the variable part are required to contact the interface point(s) between the known and variable parts of the structure. The initial approximation of the variable part is randomly generated inside a sphere with radius of gyration $R_g^{\text{est}} \sim 3\sqrt[3]{N_D}$ centered at the interface point (or between the two interface points). The SA step involves moving a randomly selected residue to an arbitrary point at a distance of 0.38 nm from one of two adjacent residues. The variable part converges to a quasi- C_α chain attached to the given point(s) in the known structure. This algorithm (program CHADD) is useful for adding missing loops or terminal portions to high-resolution models but can also be used for missing domain restoration.

Individual-residues model with spring forces between neighbors

This model is similar to the previous one but accounts for the primary structure of the protein. Not only is the scattering intensity computed using the individual form factors, but also residue-specific information is formulated as additional penalties to further restrain the solution and to generate

native-like folds of the missing loop/domain. The scoring function has the form:

$$E(\mathbf{r}) = RF^2 + \alpha_{\text{dst}}P_{\text{dst}} + \alpha_{\text{spr}}P_{\text{spr}} + \alpha_{\text{hyd}}P_{\text{hyd}} + \alpha_{\text{bur}}P_{\text{bur}} + \alpha_{\text{eng}}P_{\text{eng}} + \alpha_{\text{vol}}P_{\text{vol}} + \alpha_{\text{ang}}P_{\text{ang}} + \alpha_{\text{dih}}P_{\text{dih}}, \quad (10)$$

where the penalties P_{dst} and P_{spr} are as discussed above and the additional terms contain the residue-specific information.

The two penalties accounting for the hydrophobicity of the residues are from Huang et al. (1995):

$$P_{\text{hyd}} = -\frac{1}{n} \sum_j (H_j - C_j h_j / N_j) \quad (11)$$

$$P_{\text{bur}} = -\frac{1}{n} \sum_j B_j \quad (12)$$

The sums (Eqs. 11 and 12) run over the hydrophobic residues in the entire model. The penalty of Eq. 11 promotes contacts between the hydrophobic residues. Here, n is the total number of hydrophobic residues; C_j and H_j are the numbers of all contacts and nonhydrophilic contacts of the j th residue, respectively (contacting distance is assumed to be 0.73 nm); and N_j and h_j are the total number and the number of nonhydrophilic residues, respectively, except for the $(j - 1)$ th, j th, and $(j + 1)$ th residues. The penalty of Eq. 12 forces the hydrophobic residues to be buried in the interior of the protein. Here, B_j is the number of all neighbors of j th residue except for the $(j - 2)$ th, $(j - 1)$ th, $(j + 1)$ th, and $(j + 2)$ th residues (the neighboring distance equals 1 nm).

The penalty P_{eng} uses knowledge-based potentials to minimize the empirical free energy of the model. The interaction potentials between residues in proteins can be computed from the analysis of the PDB structures (Miyazawa and Jernigan, 1999; Sippl, 1990; Thomas and Dill, 1996). The total energy of the model is calculated as the sum over all inter-residue contacts, and the penalty has the form:

$$P_{\text{eng}} = \frac{1}{N} \sum_i \sum_{j < i-1} U_{ij}, \quad (13)$$

where the summation is performed over the residues separated by less than 0.73 nm and the potentials U_{ij} are tabulated in Miyazawa and Jernigan (1999) and (Thomas and Dill, 1996).

In keeping with the low resolution of the solution scattering data, the DR model describes the C_α backbone only, and the excluded volume effects between the backbone atoms due to the penalties P_{dst} and P_{spr} do not account for the side chains. To compensate for this, pseudo- C_β atoms representing the side chains are introduced following the lolly-loop model (Aszodi et al., 1995). The direction of the i th C_α - C_β vector depends on the positions of the $(i - 1)$ th, i th, and $(i + 1)$ th C_α atoms. The C_α - C_β distance and the van der Waals radius $r_{\beta i}$ of the pseudo- C_β atom depend on the type of the i th residue. The additional excluded volume effect is taken into account by minimizing the averaged cross-volume of all spheres representing the C_α (van der Waals radius $r_\alpha = 0.19$ nm) and pseudo- C_β atoms:

$$P_{\text{vol}} = \frac{1}{V} \sum_i \sum_j (V_{ij}^{\alpha\beta} + 0.5V_{ij}^{\beta\beta}), \quad (14)$$

where V is the total excluded volume of the protein, and $V_{ij}^{\alpha\beta}$ and $V_{ij}^{\beta\beta}$ are the cross-volumes of the C_α or C_β atom belonging to i th residue with the C_β atom belonging to j th residue, respectively.

The two other penalties impose restrictions on the distribution of bond and dihedral angles of the model chain. It is well known (Irbaeck et al., 1997; Levitt, 1976) that the C_α - C_α - C_α bond angles in a protein backbone have a specific distribution. Fig. 2 *b* presents a histogram of the distribution of C_α - C_α - C_α angles $\langle F(\gamma_k) \rangle$ averaged over more than 100 protein structures

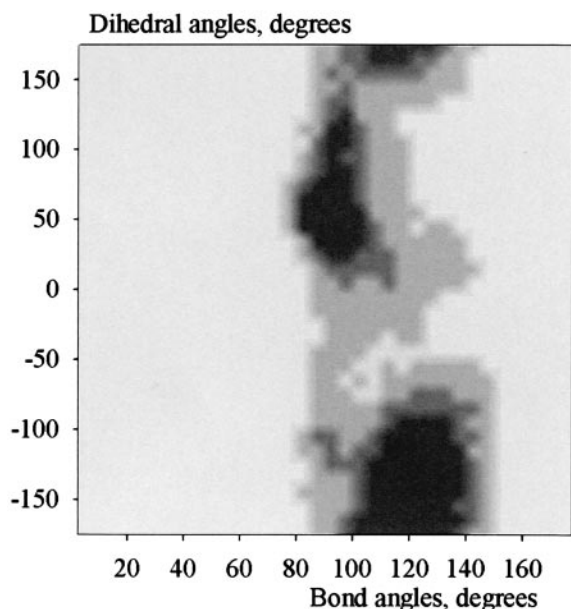


FIGURE 3 Distribution of the C_{α} backbone angles and dihedrals. The core area is shown in black, the additionally allowed regions in dark gray, the generously allowed in light gray, and the disallowed region in white. Sampling rates of the bond angles and of the dihedrals equal 5° and 10° , respectively.

deposited in the PDB. Similar to the neighbors penalty, P_{dst} (Eq. 5), the bond angle penalty is computed as:

$$P_{\text{ang}} = \sum_k \left[\frac{(F_{\text{mod}}(\gamma_k) - \langle F(\gamma_k) \rangle)}{0.1 \max(\langle F(\gamma_k) \rangle, 0.02)} \right]^2, \quad (15)$$

where $F_{\text{mod}}(\gamma_k)$ is the histogram of the current model (bin step equals 5°).

Fig. 3 displays a histogram of the distribution of C_{α} - C_{α} - C_{α} - C_{α} dihedral angles versus C_{α} - C_{α} - C_{α} angles (quasi-Ramachandran plot) computed by averaging the distributions for the above PDB models. Following Kleywegt (1997), the histogram can be split into four areas: core (index = 1), additionally allowed (2), generously allowed (3), and disallowed (4). A plausible model should display bond angles and dihedrals concentrated in the core and additionally allowed regions. Each pair of C_{α} - C_{α} - C_{α} - C_{α} dihedral angles versus C_{α} - C_{α} - C_{α} bond angles in the model is attributed to a cell in the quasi-Ramachandran plot, and the sum:

$$P_{\text{dih}} = \frac{1}{N} \sum_{i=2}^{N-2} (\text{index}(i) - 1)^2, \quad (16)$$

gives the penalty for improper dihedrals.

The generation and modification of the model during SA are the same as in the previous section. The algorithm, implemented in the program GLOOPY, allows native-like configurations to be attributed to the missing loops or domains.

Folding of a model chain composed of individual residues

The most straightforward protein model consisting of C_{α} atoms is an interconnected polypeptide chain. This model does not require a connectivity constraint, and the secondary structure elements, if known, can be easily introduced. The chain model is less flexible than the gas of residues, and this increases the chances of being trapped in an incorrect conformation during minimization. As indicated in Svergun et al. (2001a), attempts

at ab initio fitting of x-ray solution scattering data starting from a random-walk C_{α} chain led to a manifold of native-like models with different fold topologies. The chain model is, however, very useful as a means of restoring the conformation of shorter fragments such as missing loops.

The missing loop(s) are attached to the appropriate residue(s) in the known part of the structure, initially as random-walk chain(s) with a step of 0.38 nm between joints. If a specific portion of the loop is known to form an α -helix or β -sheet (e.g., from secondary structure prediction), an idealized secondary structure template of the appropriate length is inserted. The scoring function is the same as in Eq. 10 but without the spring potentials P_{spr} . Two types of moves, local and global, are used to modify the variable part of the model maintaining the distances between the adjacent residues and preserving the secondary structure. In both cases, a residue is selected at random among those not belonging to the secondary structure elements. A local move involves random rotation of a residue around the axis drawn through its two neighbors. For a global move, made after each N_D local moves, the second residue in the variable domain is selected, which does not belong to the secondary structure elements. The part of the chain between the selected residues is rotated by an arbitrary angle around the axis drawn through these two residues.

The algorithm, implemented in the program CHARGE, is aimed at restoring the conformation of the missing loop(s) and is most useful if information about their secondary structure is available.

Materials

Oligonucleotides were synthesized by Sigma-Genosys (Pampisford, UK). Media reagents were from Merck (Lutterworth, UK). Isopropyl β -D-thiogalactoside was from Genesys (London, UK). *Taq* polymerase chain reaction (PCR) Ready-To-Go beads, precast native and SDS polyacrylamide gels, protein molecular weight standards, Coomassie Brilliant Blue, glutathione Sepharose 4B, and Factor Xa were from Amersham Pharmacia Biotech (St. Alban's, UK). DNA molecular weight standards were from Gibco BRL (Paisley, UK). DNA Qiaquick gel extraction and Miniprep kits were from Qiagen (Crawley, UK). Restriction enzymes and T4 DNA ligase were from New England Biolabs (Hitchin, UK). Protein Microcon, Centricon, and Centriprep devices were from Millipore (Watford, UK). Bradford assay reagent and disposable plastic columns were from Biorad (Hemel Hempstead, UK). All other chemicals were from Sigma-Aldrich (Poole, UK).

Construction of plasmid

Plasmid pGEX-DHFR is a pGEX-5X-1 derivative (Amersham). The plasmid encodes *Schistosoma japonicum* GST, a 10-residue C-terminal linker peptide (containing a protease cleavage site) and the *Escherichia coli folA*-encoded DHFR. The *folA* gene was PCR amplified from a genomic DNA preparation of *E. coli* K-12 cells using primer 1 (5'-GAGTGGATC-CCTATCAGTCTGATTGCGGCG-3'), which contains a *Bam*HI restriction site upstream of the second codon (ATC), and primer 2 (5'-CTATCTCGAGTTACCGCCGCTCCAGAAT-3'), which incorporates a unique *Xho*I restriction site downstream of the TAA stop codon. One cycle of 96°C for 5 min followed by 35 cycles of 96°C for 1 min, 50°C for 1 min, and 72°C for 1.5 min, linked to a final cycle of 72°C for 10 min, generated a 500-bp PCR fragment encoding the *E. coli* K-12 *folA* gene. This fragment was gel purified, digested with *Bam*HI and *Xho*I, and ligated into the *Bam*HI-*Xho*I restriction sites of the pGEX-5X-1 vector to produce plasmid pGEX-DHFR. Initial clones were obtained by heat-shock transformation into *E. coli* strain BL21-CODONPLUS(DE3)-RIL. The presence of the *folA* gene was confirmed by restriction digestion of the transformed construct and by DNA sequencing with an ABI/Perkin-Elmer 377 Automated Sequencer (Perkin-Elmer Applied Biosystems, Norwalk, CT) using the dideoxy method with BigDye Terminator Ready Reaction Kits (Perkin-Elmer).

Protein purification

GST and GST-DHFR proteins were purified as follows. One liter of 2XYT (1.6% (w/v) tryptone, 1.0% (w/v) yeast extract, and 0.5% (w/v) NaCl in distilled water) containing 100 $\mu\text{g/ml}$ ampicillin was inoculated with 10 ml of an overnight culture of *E. coli* BL21-CODONPLUS(DE3)-RIL transformed with pGEX-DHFR. Cells were grown at 37°C with shaking until a cell density corresponding to an OD_{600} of 0.6 was reached. Isopropyl β -D-thiogalactoside was then added to a final concentration of 1 mM, and growth was allowed to continue for another 4 h. Centrifugation of the culture at $8000 \times g$ for 20 min yielded a cell pellet that was resuspended in PBS (0.01 M $\text{KH}_2\text{PO}_4/\text{K}_2\text{HPO}_4$ buffer, 0.0027 M KCl, 0.137 M NaCl, pH 7.4; Sigma-Aldrich). Cells were lysed by sonication with three 30-s bursts at full power, and insoluble material was removed by centrifugation at $12,000 \times g$ for 45 min.

Six milliliters of 50% (w/v) glutathione Sepharose 4B (Amersham) in PBS was added to the cell lysate supernatant (typically, 15 ml), which was then incubated at 4°C for 1.5 h with rotation. The material was transferred into a plastic column (Biorad) and washed seven times with 10 ml of PBS. GST, produced from pGEX-5X-1, was eluted by resuspending the glutathione Sepharose in 5 ml of 10 mM glutathione in 50 mM Tris-HCl, pH 8.0, and collecting the flow through from the column after incubation for 10 min at room temperature. This was repeated twice more to retrieve all the GST protein. GST-DHFR, produced from pGEX-DHFR, was eluted intact from the glutathione Sepharose as above for GST. Alternatively, cleavage of the linker between DHFR and GST was attempted by resuspending the glutathione Sepharose in 6 ml of PBS and incubating with 200 μl of 100 $\mu\text{g/ml}$ Factor Xa (Amersham) at 4°C for 24 h with rotation. Protein cleaved from the bound GST was initially collected as flow-through from the column. Subsequent washing of the column with 6 ml of PBS retrieved any additional protein. Proteins eluted from the column with glutathione and those cleaved with Factor Xa were analyzed by SDS and native polyacrylamide gel electrophoresis.

Sample preparation

All GST proteins were buffer exchanged into PBS using HiTrap Desalt columns (Amersham). Pooled samples of GST-DHFR were concentrated in Centriprep and Centricon YM-30 concentrators (Millipore) whereas pooled GST samples required Centriprep and Centricon YM-10 concentrators (Millipore). Protein concentrations were determined by Bradford assay (Biorad). For R2, SAXS measurements were performed at 2.5, 5, 10, and 20 mg/ml; GST measurements were performed at 3, 6, 7.8, and 24 mg/ml; GST-DHFR measurements were performed at 3.7, 5.4, 8.1, and 14.4 mg/ml; nonclaret disjunctional (ncd) measurements were as described (Svergun et al., 2001b).

Scattering experiments, data processing, and analysis

The experimental x-ray scattering data from protein solutions were collected following standard procedures using the X33 camera (Boulin et al., 1986, 1988; Koch and Bordas, 1983) of the European Molecular Biology Laboratory on the storage ring DORIS III of the Deutsches Elektronen Synchrotron with multiwire proportional chambers with delay line readout (Gabriel and Dauvergne, 1982). The data processing (normalization, buffer subtraction, etc.) involved statistical error propagation using the program SAPOKO (D. I. Svergun and M. H. J. Koch, unpublished data). The scattering patterns from R2, GST and GST-DHFR were recorded at sample-detector distances of 3.2 m and 1.4 m, and the wavelength $\lambda = 0.15$. The scattering patterns recorded at the two sample-detector distances were merged to yield the final composite curves to cover the range of momentum transfer $0.1 \text{ nm}^{-1} < s < 5.2 \text{ nm}^{-1}$. Additional details of the experimental procedures and the ncd data collection are described elsewhere

(Svergun et al., 2001b). The value of D_{max} was determined from the scattering patterns using the orthogonal expansion program ORTOGNOM (Svergun, 1993). The x-ray scattering patterns for simulated examples and those from the incomplete atomic models of proteins were computed from the structures taken from the PDB using the program CRY SOL (Svergun et al., 1995). The models without a one-to-one residue correspondence were superimposed using the program SUPCOMB (Kozin and Svergun, 2001), and those with such correspondence were computed with the algorithm (Kabsch, 1978).

RESULTS AND DISCUSSION

Computer programs and testing

The programs CREDO, CHADD, GLOOPY, and CHARGE all run on IBM PC-compatible machines under Windows 9x/NT/2000/XP and Linux as well as on major Unix platforms. To reduce the time required for computations, the model scattering intensity and the penalties are not recomputed after each modification of the structure but rather updated as previously described (Svergun et al., 2001a). All the programs are able to take into account particle symmetry by generating symmetry mates for the residues in the asymmetric unit (point groups P2 to P6 and P222 to P62 are supported). The programs were tested on simulated examples to adjust the parameters of the SA procedures. The values $T_0 = 10^{-3}$, $N_T = 5000\sqrt{N_D}$ and $\eta = 0.9$ were found to ensure convergence. The default values of the penalty weights for different algorithms are summarized in Table 1.

Method validation using simulated examples

To validate reconstruction procedures, a simulated fusion protein was constructed using the crystallographic coordinates of hen egg-white lysozyme (129 residues, PDB file 6lyz) (Diamond, 1974) as the N-terminal domain, with bovine pancreas trypsin inhibitor (BPTI; 58 residues, PDB entry 4pti) (Marquart et al., 1983) fused to the C-terminus of the former protein (Fig. 4, *a* and *b*). The theoretical scattering curve of the fusion protein was computed using CRY SOL (Fig. 5, *curve 1*) and was then used to reconstruct the structure of the BPTI domain assuming that the lysozyme structure is known. The program CREDO was used to fit DR models to the simulated data yielding a reasonable representation of the overall shape of the BPTI. However, in some cases the BPTI domain was not oriented next to the C-terminus of lysozyme as in the original simulated fusion protein (see typical example in Fig. 4 *a*). This is not surprising given that information about the interface between the proteins is missing in the CREDO reconstruction. It is interesting to note that even though the location of the interface is incorrect, the overall low-resolution structure of the restored models after appropriate rotation and translation agrees well with that of the simulated fusion protein (Fig. 4 *a*). To improve relative domain orientation, we used the program CHADD, which explicitly uses information

TABLE 1 Types of the models, default penalty weights, and typical applications

Program	Model	Application	Weight										
			P_{dst}	P_{con}	P_{gyr}	P_{spr}	P_{hyd}	P_{bur}	P_{eng}	P_{vol}	P_{ang}	P_{dih}	
CREDO	Free dummy residues	Generating low-resolution models of missing domains	0.001	0.01	0.001								
CHADD	Dummy residues with spring forces between neighbors	Adding missing loops or terminals in high-resolution models as well as missing domain restoration	0.001		0.1 [†]	0.01							
GLOOPY	Individual residues with spring forces between neighbors	Search for native-like configuration of the missing loops or domains	0.0001*			0.01	0.0001	0.0001	0.001	0.1	0.00001*	0.0001*	
CHARGE	Chain composed of individual residues	Restoring the conformation of the missing loops accounting for the secondary structure of the fragment	0.001				0.0001	0.0001	0.001	0.1	0.001	0.001	

*To ensure interconnectivity of the solution, the weights of penalties P_{dst} , P_{ang} , and P_{dih} are lowered by an order of magnitude in GLOOPY as compared with CHARGE.

[†]The weight of penalty P_{gyr} in CHADD is multiplied by η^2 after each temperature cycle.

about the location of the interface. In the example presented here, the C-terminus of lysozyme was identified as the fusion point. The shapes of the resulting added domains obtained in independent runs of CHADD were consistent with the crystal structure of BPTI, although their positions varied by 0.3–0.5 nm. Fig. 4 *b* presents an averaged result of 10 independent runs, which predicts the correct position and shape of the BPTI domain fairly well.

To validate the loop reconstruction procedures, several lysozyme models were made containing deletions in the following regions: 1) residues 120–129 located at the C-terminus, 2) residues 1–15 containing an α -helix located at the N-terminus, and 3) residues 40–55 containing a β -sheet located on the surface of the structure. First, the theoretical scattering pattern of the intact protein was calculated using CRY SOL. Using this scattering pattern and the coordinates of each deletion model, missing loop regions were reconstructed using the program GLOOPY. In all cases, theoretical scattering curves of the reconstructed proteins, obtained after addition of the missing loop regions, gave good fits to the simulated scattering pattern of the intact protein (Fig. 5, *curves 2–4*). When compared with C_α coordinates of the crystal structure of lysozyme, typical restored models (Fig. 4, *c–e*) have an overall RMSD equal to 0.17, 0.24, and 0.25 nm for deletions 1, 2, and 3, respectively. For comparison, generation of the missing fragments as random-walk self-avoiding chains yields the average RMSD values of 0.37, 0.53, and 0.51, respectively. Use of the program CHARGE for deletion 2 forces residues 5–15 to form an α -helix, thus further reducing the RMSD (to \sim 0.15–0.2 nm; results not shown).

Conformational mobility in small loops/domains

The failure to observe structural elements in electron density maps arising from protein crystal structures is often due to conformational mobility or heterogeneity. The application of reconstruction methods offers the possibility of constructing a model for the missing loops or domains both in terms of their structure and their position in three-dimensional space. Two examples are presented below that illustrate these concepts.

In the first example, a truncated form of the *Drosophila* motor protein ncd was studied using SAXS (Svergun et al., 2001b). The native ncd protein is 700 residues in length. A construct named MC6 was made that expresses the C-terminal 368 residues (M333-K700) of ncd. This construct appears to be monomeric in solution as it lacks an N-terminal coiled-coil region (residues 196–347) that would otherwise mediate dimerization. Using crystallographic coordinates of a ncd variant (PDB entry 1cz7) (Kozielski et al., 1999), a partial three-dimensional model of MC6 was produced (Svergun et al., 2001b). This model lacked the 33 C-terminal residues absent in the crystal structure. The scattering curve computed from the MC6 model fails to fit the scattering pattern of the protein in solution (Fig. 6, *curve I*; $\chi = 1.98$). Addition of the missing loop using the programs GLOOPY and CHARGE in these studies significantly improved the fit ($\chi = 0.89$). The loop conformations yielded by the programs in different independent reconstructions are similar to each other, suggesting a fan-like manifold of orientations (Fig. 7 *a*). In an earlier study using trial secondary structure motifs, this region was modeled as

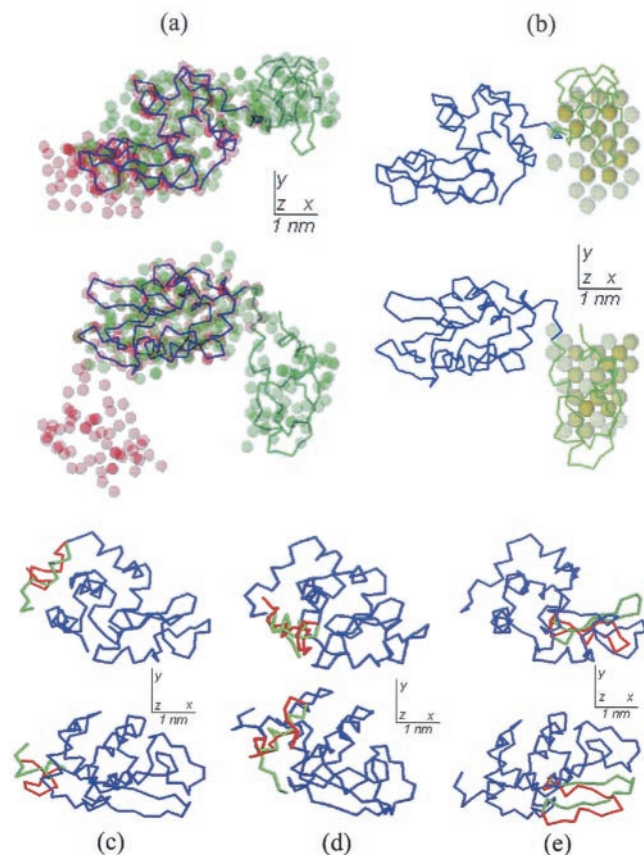


FIGURE 4 Reconstruction of the missing domain in a fictitious fusion protein. A molecule of BPTI (*green*) is attached to the C-terminal of hen egg-white lysozyme (*blue*). The two molecules are displayed as C_{α} traces and the reconstructed models as semitransparent spheres. (a) Typical reconstruction by CREDO. The orientation with the lysozyme molecule overlapped is shown in red; the orientation yielding the best overlap with the entire complex is shown in green. (b) Average of five independent reconstructions by CHADD (probable shape and position of the BPTI domain is displayed in *yellow*). Comparison of the atomic structure of lysozyme with the models reconstructed by the program GLOOPY: green, correct fold of the missing loop; red, typical restored fold; blue, the rest of the structure. (c) Missing C-terminal tail; (d) missing N-terminal tail; (e) missing loop in the middle of the sequence. On all panels, the bottom view is rotated by 90° counterclockwise around the x axis.

an antiparallel two-stranded β -sheet (Svergun et al., 2001b). The conformation of this tentative model fits within the plane of the fan and is also of a similar length as the conformations provided by GLOOPY and CHARGE. These results suggest that the loop is flexible in solution, moving predominantly in the plane of the fan.

The second example illustrates the use of information about secondary structure for the reconstruction of a missing loop. Specifically, the crystallographic model of a homodimeric protein R2 of ribonucleotide reductase from *E. coli* (PDB entry 1xik; molecular mass = 79 kDa) (Logan et al., 1996) was solved to 1.7-Å resolution containing 341 residues per monomer. The C-terminal 35 residues are miss-

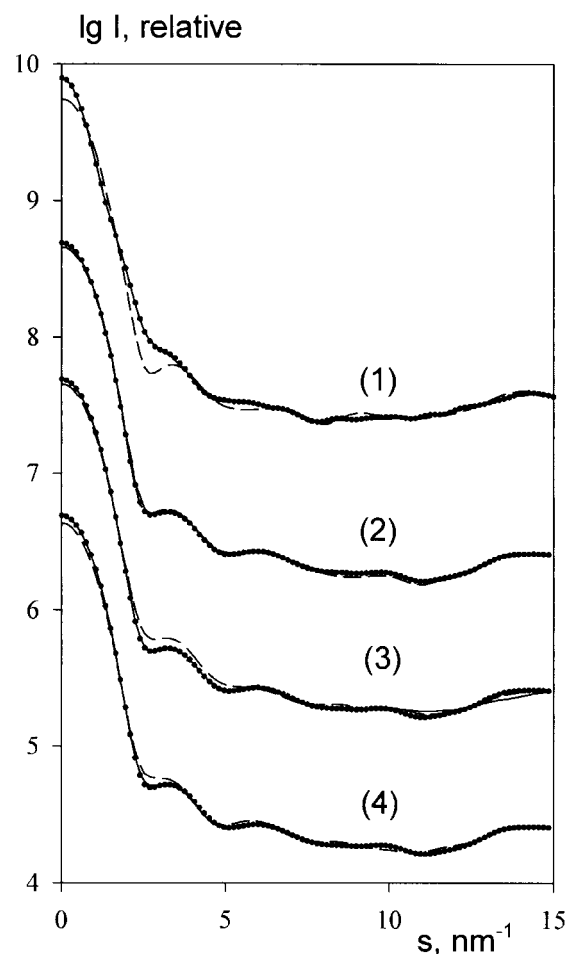


FIGURE 5 Scattering patterns from the model lysozyme structures. (1) Complex with BPTI; (2) missing 10 residues at the C-terminal; (3) missing 15 residues at the N-terminal; (4) missing 15 residues in the middle. \cdots , scattering from full-length structures; $-\ -$, scattering from the models without the missing fragments; $—$, scattering from the restored models.

ing in the crystal structure, and the scattering curve computed from the crystallographic model displays small but significant systematic deviations from the experimental data ($\chi = 1.30$; Fig. 6, *curve 2* and *inset*; S. Kuprin, Karolinska Institute, Stockholm, Sweden, personal communication, 1998). According to secondary structure prediction programs (Cuff and Barton, 1999, 2000; Cuff et al., 1998), a major portion of the missing fragment (residues 345–373) is predicted to form an α -helix. Fig. 7 *b* shows the position of a typical reconstruction of the fragment using the program CHARGE, which gives a significant improvement in the fit to the experimental data ($\chi = 1.07$). The result suggests that the α -helix from each monomer subunit extends away from the core structure of the protein to produce a biantennary structure in the dimer. This structure is likely to occupy a number of conformations, which is consistent with the lack of interpretable electron density in the original crystal structure (Logan et al., 1996).

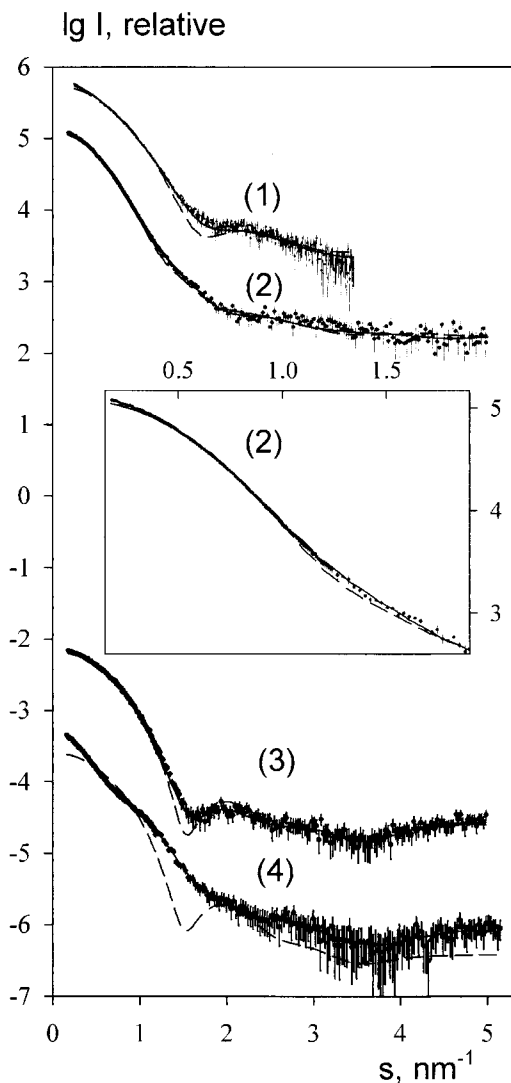


FIGURE 6 X-ray scattering patterns from MC6 construct (1), the protein R2 (2), GST (3), and its fusion with DHFR (4) ($\cdot \cdot \cdot$ with error bars); scattering of the crystallographic models where the missing fragments are absent ($- - -$); and scattering from the reconstructed models ($—$). The scattering patterns are appropriately displaced in the logarithmic scale, and the inner part of the R2 pattern is shown in the inset for better visualization.

GST-fusion protein domains

The pGEX series of vectors (Amersham) (Smith and Johnson, 1988) are designed to enable inducible, high-level intracellular expression of genes as fusions with the *Schistosoma japonicum* GST, a 26-kDa protein forming homodimers in solution. Crystal structures are available for GSTs from a number of sources (Ji et al., 1992; Parker et al., 1990), including recombinant *S. japonicum* GST purified from pGEX-3X (Amersham) (McTigue et al., 1995). This recombinant *S. japonicum* GST contains an extra 13 residues at the C-terminus compared with the native *S. japonicum* GST, but this linker peptide is absent in the PDB entry

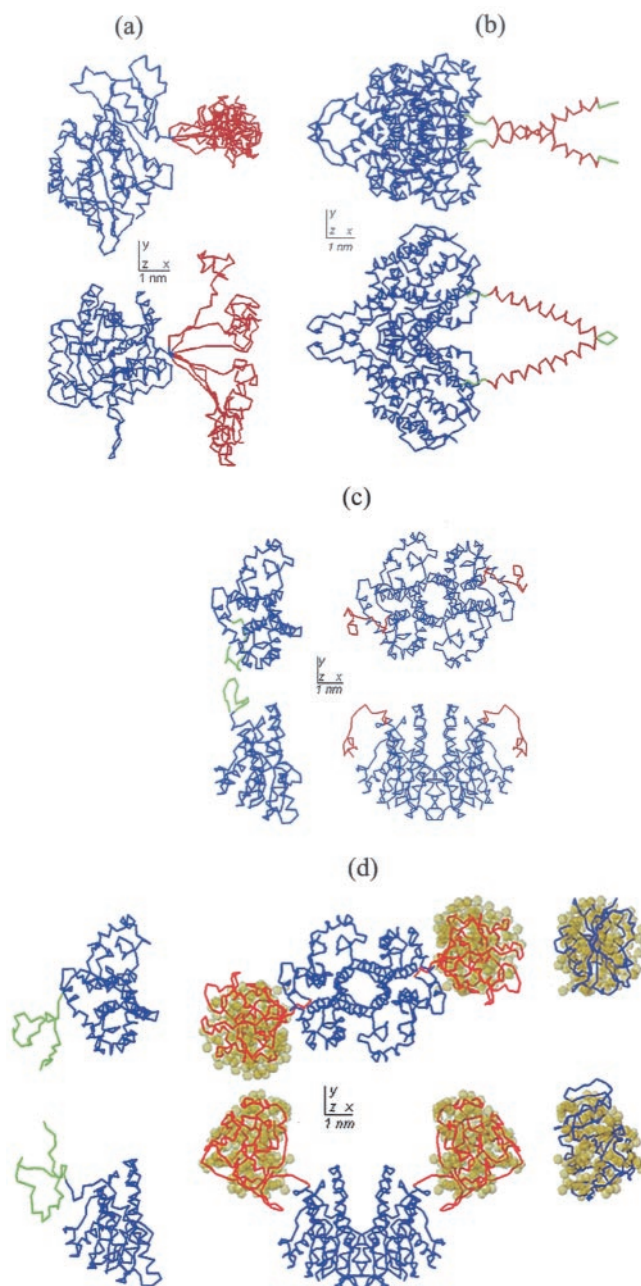


FIGURE 7 Reconstruction of missing loops in proteins. Crystallographic models are displayed in blue, with the reconstructed fragments as red C_{α} -traces and yellow semitransparent spheres. (a) Missing loop in the MC6 construct obtained by GLOOPY and CHARGE (several independent reconstructions are displayed); (b) Missing C-terminal loop in dimeric R2 obtained by CHARGE (the α -helical portion of the variable domain and the portion for which no secondary structure was assumed are displayed in red and green, respectively); (c) A linker at the C-terminal in dimeric GST reconstructed by GLOOPY (right panel) and the structure of the GST monomer fused with a conserved neutralizing epitope GP41 (15 residues displayed in green, PDB entry 1g9e); (d) A DHFR domain in a fusion GST+DHFR protein reconstructed by CREDO (spheres) and CHADD (C_{α} -traces). Solution by CREDO is superimposed with the atomic model of DHFR on the right panel. The left panel presents the crystallographic structure of the GST fusion with α -Na,K-ATPase (36 residues displayed in green, PDB entry 1bg5). On all panels, the bottom view is rotated by 90° counterclockwise around the x axis.

(1gta) (McTigue et al., 1995). GST was expressed and purified for SAXS data collection and analysis from similar plasmid in the pGEX series, pGEX-5X-1 (Amersham). Compared with the native *S. japonicum* GST, this GST has an extra 22 residues at the C-terminus. The x-ray solution scattering pattern from the latter protein is presented in Fig. 6. The scattering curve computed from a homodimer built from the crystallographic structure of GST lacking the C-terminal residues yields a poor fit to the experimental data (Fig. 6, *curve 3*; $\chi = 1.30$). GLOOPY was used to model the missing linker, assuming P2 symmetry for the entire structure. Several independent runs produced similar extended conformations of the modeled linker (a typical result is presented in Fig. 7 *c*). The theoretical scattering of the GST crystal structure combined with this modeled linker gave a significant improvement in the fit to the experimental data with $\chi = 0.81$.

SAXS analysis was also performed on a GST fusion protein. The *folA* gene, encoding dihydrofolate reductase from *E. coli* K-12 was cloned into the same vector, pGEX-5X-1, from which GST was expressed. This enabled production of a fusion protein, GST-DHFR, consisting of the 218 residues of the *S. japonicum* GST followed by a 10-residue linker containing the Factor Xa cleavage site and 158 residues of *E. coli* K-12 DHFR. Fig. 6 (*curve 4*) shows the experimental scattering pattern of GST-DHFR. Models of the dimeric fusion protein were built using the programs CREDO and CHADD by fixing the structure of dimeric GST and then adding the linker and the DHFR domain as a variable part and assuming P2 symmetry for the entire complex. Both programs gave good fits to the experimental data with $\chi = 1.02$ (Fig. 6, *curve 4*). The shape and position of the missing domain for each model reconstructed by the two methods were consistent with each other and also with the crystallographic model of DHFR (PDB entry 1ra9) (Sawaya and Kraut, 1997), as illustrated in Fig. 7 *d*. It would appear from comparison of the linker regions of GST and GST-DHFR that the configuration of the linker is different in each case (cf. Fig. 7 *c*). Attempts to cleave the linker region of GST-DHFR using Factor Xa were unsuccessful, even when the wt % was increased above 1% and the incubation time was >16 h. Even in the presence of 0.05% SDS, the percentage of protein cleaved was minimal. Resistance to proteolytic cleavage may well be due to a more compact conformation of the linker region in GST-DHFR and/or steric hindrance causing the Factor Xa cleavage site to be inaccessible. Although there are no crystal structures in the PDB of proteins >40 residues fused to GST, comparison of the structures of two GST fusions (Fig. 8, *a* and *b*), PDB entries 1gne (Lim et al., 1994) and 1bg5 (Zhang et al., 1998), with the model of GST-DHFR shows that there appears to be a similar orientation of GST and its fusion partner in all three cases.

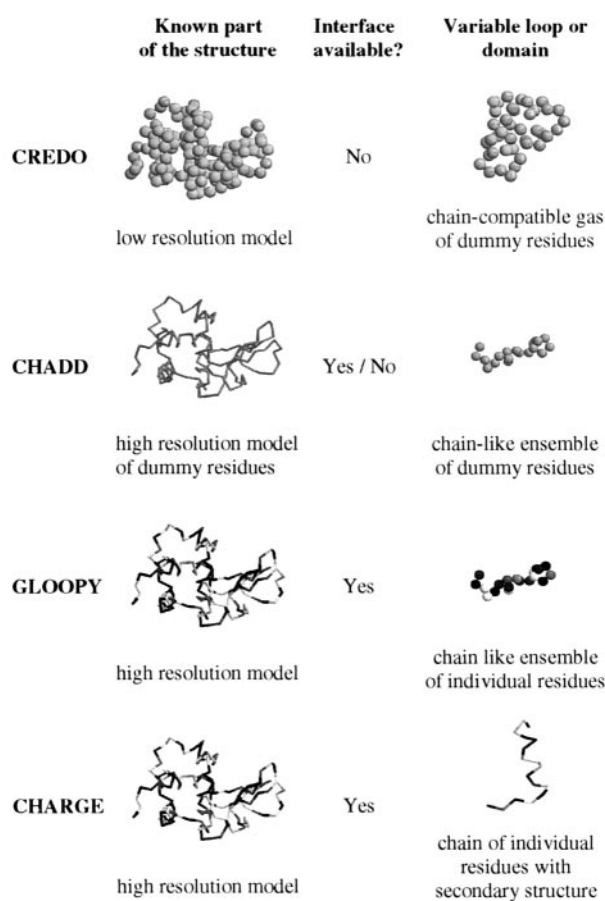


FIGURE 8 Schematic illustration of the model types and additional information used by different programs.

CONCLUSION

To summarize, four algorithms have been written to provide an appropriate tool for each of the various situations in which a structure lacks a loop or domain. The choice of method depends on the information available regarding the known part of the model, the missing fragment, and the interface. If a low-resolution model of the known part is available (e.g., from electron microscopy or from SAXS by ab initio methods (Svergun, 1999; Svergun et al., 2001a)), the location of the interface is usually unknown and the missing fragment can be added using the program CREDO. In this case, the result is a low-resolution model of the domain structure of the complex. For high-resolution models, the programs CHADD and GLOOPY can build missing loops and domains attached to specific residues(s). Furthermore, GLOOPY tries to construct native-like folds by accounting for excluded volumes of side chains, hydrophobic interactions, knowledge-based potentials, and the C_{α} bond and dihedral angles. If the secondary structure of the missing portion is known, the program CHARGE allows additional constraints to be applied to the model by incorporat-

ing α -helices and/or β -sheets in the variable fragment. As the model of an interconnected C_α chain used by CHARGE is less flexible than a free gas of residues implemented in the other programs, CHARGE is better suited to reconstruct missing loops rather than missing domains. The main features and possible applications of the four algorithms are summarized in Table 1 and Fig. 8.

Even though the programs CHADD, GLOOPY, and CHARGE yield the missing fragments in the form of folded C_α chains, these should be considered as approximate models only. Solution scattering, being a low-resolution method, does not provide an exact fold but rather a probable configuration of the volume occupied by the missing portion. In all of the above algorithms, scattering from the model is computed using Eqs. 1–3, which do not explicitly take averaging over possible different conformations of flexible loops or terminal fragments into account. Such an average would not significantly influence the results given the low resolution of the scattering data but would take much longer computation times. The methods are trying to obtain a single equivalent conformation of the missing domain, and it is also useful to analyze the results of several independent SA runs to generate averaged probability maps. This analysis allows refinement of the shape and position of missing domains (see Fig. 4, *a* and *b*) and better visualization of regions occupied by the missing loops (Fig. 7 *a*). When using CHARGE, care must be taken not to restrict the model too much based on secondary structure predictions (which generally are no better than 70% accurate). In the above example for R2, all major techniques predicted an α -helix with high probability, which made it possible to use a long helical fragment for constructing the model in Fig. 7 *b*.

Missing loop residues can be added to known high-resolution structures using homology modeling (Mendelson and Morris, 1997; Perera et al., 2000). In general, the reliability of structures produced using homology modeling is high for short loops but decreases as the length of the fragment to be added is increased. Using solution scattering, the situation is precisely the opposite: the larger the missing fragment, the more significant its contribution to the entire scattering pattern, and the missing residues can be modeled more reliably. In practical terms, one can expect the methods presented here to be useful for missing fragments consisting of ~5–10% of the entire structure (20–40 residues for a 50-kDa protein) and higher. For shorter loops, homology modeling may be sufficient; however, solution scattering data can be used as an additional restraint (in particular, for rigid body refinement of the orientation of the fragment to be added) and can also be used for validation of the final model (see, e.g., Zheng and Doniach, 2002). Moreover, it should be stressed that the methods presented are not limited to amending crystallographic models with disordered loops but are also applicable to the addition of missing domains to low-resolution models and to fusion proteins, especially when no crystals are available.

In the model systems presented here, experimental SAXS data has allowed the reconstruction of both missing loops and domains, providing a structural description of disordered regions. The reconstructions are based on the experimental data of ~1.2-nm resolution, but the actual resolution of the models may be higher because of the additional information used. In particular, histogram and angular penalties (Figs. 2 and 3) ensure adequate behavior of the model scattering curves at higher momentum transfers. In the case of the *Drosophila* motor and R2 ribonuclease reductase proteins, modeling predicts extended structures from the surface of the globular core. Such structures could indeed show large flexibility in solution, which may explain why the regions could not be modeled from the crystallographic electron density maps. Modeling studies of GST expressed from the pGEX system give a description of the linker region, which was not visible in the original crystal structure (McTigue et al., 1995). The model of GST-DHFR also provides a visualization of how such fusions appear in solution. In particular, the linker region appears to adequately separate both globular domains, suggesting that GST does not, per se, directly influence folding of its partner in protein-protein interactions. In addition, the model shows how the fused protein (i.e., DHFR) can occlude the linker, resulting in resistance to protease digestion in this case. Taken together, these examples demonstrate how such reconstruction methods using SAXS data have the potential to add missing fragments to available high- or low-resolution protein models. Indeed, as three-dimensional structural information from larger multi-protein complexes emerges, the true potential of these techniques may be realized for modeling both domains and interfaces responsible for macromolecular assembly where inherent flexibility and conformational heterogeneity limit high-resolution visualization. Modeling using the protein structure representation as an ensemble of residues could also become useful for the interpretation of low-resolution crystallographic maps (Guo et al., 1999).

The executable codes of the programs CREDO, CHADD, GLOOPY, and CHARGE are available as Wintel β -releases from the EMBL-Hamburg website (<http://www.embl-hamburg.de/ExternalInfo/Research/Sax>). The executables for Linux and major UNIX platforms can be obtained from the authors upon request.

We are indebted to M. H. J. Koch for stimulating discussions and help with the x-ray data collection and to F. Kozielski and S. Kuprin for providing the experimental data on MC6 and R2.

The work was supported by the 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 U.K. Biotechnology and Biological Sciences Research Council (studentship to N.A.J.E.) and the European Community Access to Research Infrastructure Action of the Improving Human Potential Program to the EMBL Hamburg Outstation, contract HPRI-CT-1999–00017.

REFERENCES

- Ashton, A. W., M. K. Boehm, J. R. Gallimore, M. B. Pepys, and S. J. Perkins. 1997. Pentameric and decameric structures in solution of serum amyloid P component by X-ray and neutron scattering and molecular modelling analyses. *J. Mol. Biol.* 272:408–422.
- Aszodi, A., M. J. Gradwell, and W. R. Taylor. 1995. Global fold determination from a small number of distance restraints. *J. Mol. Biol.* 251:308–326.
- 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. *Nuclear Instrum. Methods 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. *Nuclear Instrum. Methods A.* 249:399–407.
- Cuff, J. A., and G. J. Barton. 1999. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins.* 34:508–519.
- Cuff, J. A., and G. J. Barton. 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins.* 40:502–511.
- Cuff, J. A., M. E. Clamp, A. S. Siddiqui, M. Finlay, and G. J. Barton. 1998. JPred: a consensus secondary structure prediction server. *Bioinformatics.* 14:892–893.
- Debye, P. 1915. Zerstreung von Roentgenstrahlen. *Ann. Phys.* 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.* 7:970–972.
- Feigin, L. A., and D. I. Svergun. 1987. *Structure Analysis by Small-Angle X-Ray and Neutron Scattering*. Plenum Press, New York.
- Gabriel, A., and F. Dauvergne. 1982. The localization method used at EMBL. *Nuclear Instrum. Methods.* 201:223–224.
- Guo, D. Y., R. H. Blessing, D. A. Langs, and G. D. Smith. 1999. On 'globbicity' of low-resolution protein structures. *Acta Crystallogr. D.* 55:230–237.
- Huang, E. S., S. Subbiah, and M. Levitt. 1995. Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.* 252:709–720.
- Ingber, L. 1993. Simulated annealing: practice versus theory. *Math. Comput. Model.* 18:29–57.
- Irbæck, A., C. Peterson, F. Potthast, and O. Sommelius. 1997. Local interactions and protein foldig: a three-dimensional off-lattice approach. *J. Chem. Phys.* 107:273–282.
- Ji, X., P. Zhang, R. N. Armstrong, and G. L. Gilliland. 1992. The three-dimensional structure of a glutathione S-transferase from the mu gene class: structural analysis of the binary complex of isoenzyme 3–3 and glutathione at 2.2-Å resolution. *Biochemistry.* 31:10169–10184.
- Kabsch, W. A. 1978. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A.* 34:827–828.
- Kirkpatrick, S., C. D. Gelatt, Jr., and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science.* 220:671–680.
- Kleywegt, G. J. 1997. Validation of protein models from C α coordinates alone. *J. Mol. Biol.* 273:371–376.
- Koch, M. H. J., and J. Bordas. 1983. X-ray diffraction and scattering on disordered systems using synchrotron radiation. *Nuclear Instrum. Methods.* 208:461–469.
- Konarev, P. V., M. V. Petoukhov, and D. I. Svergun. 2001. MASSHA: a graphic system for rigid body modelling of macromolecular complexes against solution scattering data. *J. Appl. Crystallogr.* 34:527–532.
- Kozielski, F., S. De Bonis, W. P. Burmeister, C. Cohen-Addad, and R. H. Wade. 1999. The crystal structure of the minus-end-directed microtubule motor protein ncd reveals variable dimer conformations. *Struct. Fold Des.* 7:1407–1416.
- Kozin, M. B., and D. I. Svergun. 2000. A software system for automated and interactive rigid body modeling of solution scattering data. *J. Appl. Crystallogr.* 33:775–777.
- Kozin, M. B., and D. I. Svergun. 2001. Automated matching of high- and low-resolution structural models. *J. Appl. Crystallogr.* 34:33–41.
- Krueger, J. K., G. A. Olah, S. E. Rokop, G. Zhi, J. T. Stull, and J. Trehwella. 1997. Structures of calmodulin and a functional myosin light chain kinase in the activated complex: a neutron scattering study. *Biochemistry.* 36:6017–6023.
- Levitt, M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* 104:59–107.
- Lim, K., J. X. Ho, K. Keeling, G. L. Gilliland, X. Ji, F. Ruker, and D. C. Carter. 1994. Three-dimensional structure of *Schistosoma japonicum* glutathione S-transferase fused with a six-amino acid conserved neutralizing epitope of gp41 from HIV. *Protein Sci.* 3:2233–2244.
- Logan, D. T., X. D. Su, A. Aberg, K. Regnstrom, J. Hajdu, H. Eklund, and P. Nordlund. 1996. Crystal structure of reduced protein R2 of ribonucleotide reductase: the structural basis for oxygen activation at a dinuclear iron site. *Structure.* 4:1053–1064.
- Marquart, M., J. Walter, J. Deisenhofer, W. Bode, and R. Huber. 1983. The geometry of the reactive site and of the peptide groups in trypsin, trypsinogen and its complex with inhibitors. *Acta Crystallogr. B.* 39:480–490.
- McTigue, M. A., D. R. Williams, and J. A. Tainer. 1995. Crystal structures of a schistosomal drug and vaccine target: glutathione S-transferase from *Schistosoma japonica* and its complex with the leading antischistosomal drug praziquantel. *J. Mol. Biol.* 246:21–27.
- Mendelson, R., and E. P. Morris. 1997. The structure of the acto-myosin subfragment 1 complex: results of searches using data from electron microscopy and x-ray crystallography. *Proc. Natl. Acad. Sci. U.S.A.* 94:8533–8538.
- 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 Struct. Funct. Genet.* 34:49–68.
- Nagai, K., and H. C. Thogersen. 1984. Generation of β -globin by sequence-specific proteolysis of a hybrid protein produced in *Escherichia coli*. *Nature.* 309:810–812.
- Parker, M. W., M. Lo Bello, and G. Federici. 1990. Crystallization of glutathione S-transferase from human placenta. *J. Mol. Biol.* 213:221–222.
- Perera, L., C. Foley, T. A. Darden, D. Stafford, T. Mather, C. T. Esmon, and L. G. Pedersen. 2000. Modeling zymogen protein C. *Biophys. J.* 79:2925–2943.
- Press, W. H., S. A. Teukolsky, W. T. Wetterling, and B. P. Flannery. 1992. *Numerical Recipes*. University Press, Cambridge, UK.
- Sawaya, M. R., and J. Kraut. 1997. Loop and subdomain movements in the mechanism of *Escherichia coli* dihydrofolate reductase: crystallographic evidence. *Biochemistry.* 36:586–603.
- Sayers, Z., P. Brouillon, D. I. Svergun, P. Zielenkiewicz, and M. H. J. Koch. 1999. Biochemical and structural characterization of recombinant copper-metallothionein from *Saccharomyces cerevisiae*. *Eur. J. Biochem.* 262:858–865.
- Sippl, M. J. 1990. Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* 213:859–883.
- Smith, D. B., and K. S. Johnson. 1988. Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. *Gene.* 67:31–40.
- 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. A.* 50:391–402.

- 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., I. Aldag, T. Sieck, K. Altendorf, M. H. J. Koch, D. J. Kane, M. B. Kozin, and G. Grueber. 1998a. A model of the quaternary structure of the *Escherichia coli* F1 ATPase from x-ray solution scattering and evidence for structural changes in the delta subunit during ATP hydrolysis. *Biophys. J.* 75:2212–2219.
- Svergun, D. I., C. Barberato, and M. H. J. Koch. 1995. CRY SOL: a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* 28:768–773.
- Svergun, D. I., C. Barberato, M. H. J. Koch, L. Fetler, and P. Vachette. 1997. Large differences are observed between the crystal and solution quaternary structures of allosteric aspartate transcarbamylase in the R state. *Proteins.* 27:110–117.
- Svergun, D. I., and K. H. Nierhaus. 2000. A map of protein-rRNA distribution in the 70 S *Escherichia coli* ribosome. *J. Biol. Chem.* 275:14432–14439.
- Svergun, D. I., M. V. Petoukhov, and M. H. J. Koch. 2001a. Determination of domain structure of proteins from x-ray solution scattering. *Biophys. J.* 80:2946–2953.
- Svergun, D. I., M. V. Petoukhov, M. H. J. Koch, and S. Koenig. 2000. Crystal versus solution structures of thiamine diphosphate-dependent enzymes. *J. Biol. Chem.* 275:297–302.
- Svergun, D. I., S. Richard, M. H. J. Koch, Z. Sayers, S. Kuprin, and G. Zaccai. 1998b. 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., G. Zaccai, M. Malfois, R. H. Wade, M. H. J. Koch, and F. Kozielski. 2001b. Conformation of the *Drosophila* motor protein non-claret disjunctional in solution from x-ray and neutron scattering. *J. Biol. Chem.* 276:24826–24832.
- 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.
- Ware, S., J. P. Donahue, J. Hawiger, and W. F. Anderson. 1999. Structure of the fibrinogen γ -chain integrin binding and factor XIIIa cross-linking sites obtained through carrier protein driven crystallization. *Protein Sci.* 8:2663–2671.
- Zhang, Z., P. Devarajan, A. L. Dorfman, and J. S. Morrow. 1998. Structure of the ankyrin-binding domain of α -Na,K-ATPase. *J. Biol. Chem.* 273:18681–18684.
- Zheng, W., and S. Doniach. 2002. Protein structure prediction constrained by solution x-ray scattering data and structural homology identification. *J. Mol. Biol.* 316:173–187.