

New methods for domain structure determination of proteins from solution scattering data

Maxim V. Petoukhov and Dmitri I. Svergun

Copyright © International Union of Crystallography

Author(s) of this paper may load this reprint on their own web site provided that this cover page is retained. Republication of this article or its storage in electronic databases or the like is not permitted without prior permission in writing from the IUCr.

New methods for domain structure determination of proteins from solution scattering data

Maxim V. Petoukhov^{ab} and Dmitri I. Svergun^{ab*}

^aEuropean Molecular Biology Laboratory, Hamburg Outstation, EMBL c/o DESY, Notkestraße 85, D-22603 Hamburg, Germany, and ^bInstitute of Crystallography, Russian Academy of Sciences, Leninsky pr. 59, 117333 Moscow, Russia. E-mail: Svergun@EMBL-Hamburg.DE

New approaches for domain structure determination of macromolecules from X-ray solution scattering are presented. An *ab initio* method for building structural models of proteins from the scattering data implemented in the computer program GASBOR employs simulated annealing to find a chain-like spatial distribution of dummy residues (DR), which fits the experimental scattering pattern up to a resolution of 0.5 nm. This method substantially improves the resolution and reliability of models derived from the scattering data. A modification of GASBOR fitting distance distribution data in real space substantially (by a factor of 3 to 7) speeds up the calculation of DR models. The DR approach is further extended to reconstruct missing domains in multisubunit proteins and fusion proteins and to find probable configurations of missing loops in crystallographic models. Four computer programs have been written to provide appropriate tools for different situations in which a high or low resolution structural model lacks a loop or domain. These methods permit solution scattering analysis to usefully complement the results obtained by high-resolution methods like X-ray crystallography and nuclear magnetic resonance spectroscopy. The efficiency of the presented approaches is illustrated by their application for structure analysis of several proteins, with known and unknown crystal structure, from experimental scattering data.

Keywords: small angle scattering; protein fold; domain structure; modelling; simulated annealing

1. Introduction

Recent breakthrough in genome sequencing prompted structural genomics initiatives aiming at determining thousands of new protein structures using X-ray crystallography and nuclear magnetic resonance spectroscopy (NMR) (Burley, 2000; Edwards *et al.*, 2000). Obviously, X-ray crystallography requires crystals of good diffraction quality, while the application of NMR to structure determination is limited to small proteins. It is expected that not more than 30% of genome proteins can be solved at high resolution and development of complementary approaches is important to widen the range of structural targets. Small-angle scattering (SAS) of X-rays and neutrons is a fundamental tool for low resolution structure studies of biological macromolecules in a wide range of sizes. The scattering data can be obtained directly from solutions in nearly physiological conditions. The experimental procedures are relatively fast and simple compared to crystallography and NMR and they do not require special sample preparation. These advantages become crucial in the investigation of large conformational

transitions that occur upon binding of effectors or changes in physicochemical conditions and assembly or (un)folding processes.

The scattering intensity $I(s)$ from a dilute monodisperse 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 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

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

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 non-zero contrast for the shell $\delta\rho_b = \rho_b - \rho_s$ (Svergun *et al.*, 1995).

SAS has attracted a renewed interest after recent development of new analysis methods to reconstruct three dimensional structure of macromolecules from the scattering patterns (Svergun *et al.*, 1996; Chacon *et al.*, 1998; Svergun, 1999; Walther *et al.*, 1999). The most recent approach representing a protein by an assembly of dummy residues (DR) (Svergun *et al.*, 2001) overcomes some of the limitations of the homogeneous approximation used so far in the analysis of SAS data and yields substantially more detailed and reliable models. In the present paper, further developments of the DR technique are described.

2. Dummy residues approach

The DR approach is based on the fact that proteins typically consist of folded polypeptide chains composed of amino acid residues separated by approximately 0.38 nm between adjacent C_{α} atoms in the primary sequence. Up to a resolution of about 0.5 nm, the protein structure can be considered as an assembly of DRs centred at the C_{α} positions. A three-dimensional model of the protein may be constructed from solution scattering data by finding a chain-like spatial arrangement of the DRs that fits the experimental scattering pattern including small and medium angles. The use of the C_{α} positions permits 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 the nearest neighbours. The requirement of chain compatibility is included as a penalty term in the function to be minimised in the form:

$$E(\mathbf{r}) = \chi^2 + \alpha \text{Pen}(\mathbf{r}), \quad (2)$$

where χ^2 is the discrepancy between the experimental $I_{\text{exp}}(s)$ and calculated $I_{\text{DR}}(s)$ scattering curves, $\text{Pen}(\mathbf{r})$ is the penalty to ensure chain-like distribution of neighbours, $\alpha > 0$ its weight.

The scattering intensity from an ensemble of K points ($K = N$ dummy residues + M dummy solvent atoms belonging to a border layer) with coordinates \mathbf{r}_i is calculated using the Debye formula (Debye, 1915):

$$I_{\text{DR}}(s) = \sum_{i=1}^K \sum_{j=1}^K g_i(s) g_j(s) \frac{\sin sr_{ij}}{sr_{ij}}, \quad (3)$$

where $g_i(s)$ is the formfactor of the i -th point and $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ is the distance between the i -th and j -th point.

Given the large quantity of free parameters (coordinates of DRs), minimisation of $E(\mathbf{r})$ is performed using simulated annealing (Kirkpatrick *et al.*, 1983). This *ab initio* method for building structural models of proteins from X-ray solution scattering data is implemented in a computer program GASBOR (Svergun *et al.*,

2001). Its efficiency was illustrated by application for structure analysis of numerous proteins, with known and unknown crystal structure, from the experimental scattering data. It was demonstrated that the DR method accounting for the scattering data up to 0.5 nm resolution substantially improves the reliability of models compared to the shape determination methods, which could only fit low resolution scattering data. The improved detaility of the DR models makes solution scattering a useful complementary technique in large-scale structural characterisation of proteins (e.g. if the crystal is not available).

3. Real space version of GASBOR

Rapid generation of models is a necessary prerequisite for large-scale analysis. It is clear that dealing with thousands primary sequences, the time spent per structure is must be of order of days, not more. A solution scattering experiment using third generation source is a matter of minutes. The generation of a DR model of a protein by the original version of GASBOR is computationally proportional to $N^{3/2}$, where N is the number of residues and for large proteins may require several days of CPU on an average workstation. The most time-consuming operation in the simulated annealing DR procedure is the calculation of scattering intensity $I(s)$ from an ensemble of residues using the Debye formula.

In order to improve the speed of the program, a real-space fitting procedure was implemented. As the original, reciprocal space version, the program starts from a random distribution of DRs in a search volume (a sphere with the maximum diameter equal to that of the particle). A single modification of the model involves moving a randomly selected residue to a point at a distance 0.38 nm from another residue within the search volume, and the rule to accept this modification is defined by the simulated annealing protocol. In contrast to the original GASBOR the discrepancy χ^2 is calculated between 'experimental' distance distribution function $P(r)$ and that of the model $P_{DR}(r)$:

$$\chi^2 = \frac{1}{n-1} \sum_{j=1}^n \left[\frac{cP_{DR}(r_j) - P(r_j)}{\sigma(r_j)} \right]^2. \quad (4)$$

Here, the function $P(r)$ and the correspondent standard deviation $\sigma(r)$ are pre-calculated from the experimental scattering data by the indirect Fourier transform using program GNOM (Svergun, 1992; Svergun *et al.*, 1988), and c is a scaling coefficient. The distance distribution function of the given DR ensemble is calculated as follows. A sampling grid on r with step 0.1 nm is introduced both for $P(r)$ (Fig. 1, curve 1) and $P_{DR}(r)$. Each interresidual distance contributes to $P_{DR}(r)$ with the weight $g_i(0)*g_j(0)$, where $g_i(0)$ and $g_j(0)$ are formfactors of i -th and j -th residues taken at $s=0$. However, computation of $P_{DR}(r)$ from a C_α -only model is not trivial. If one assumes that a given interresidue distance contributes only to a single bin where it directly falls, an oscillating function is obtained, whereas the experimental $P(r)$ functions of proteins are smooth. Thus, even using correct positions of the C_α atoms, one would not reproduce the experimental $P(r)$ (Fig. 1, curve 2). To overcome this difficulty, each interresidue distance is supposed to contribute to several bins which smoothes the $P_{DR}(r)$. The given distance contributes to the bin into which it falls and to l neighbouring bins from both sides, and the contributions are proportional to the areas covered in the given bin by the isosceles triangle with base coinciding with the r axis and is centred at the r matching the given distance (Fig. 2). The base length is equal to $2l$ nm and the full triangle area corresponds to the weight of the distance $g_i(0)*g_j(0)$. Taken into account that an average effective residue size is about 0.7 nm, $l=3$ seems a reasonable choice. Indeed, curves 3 and 4 in Fig 1.

presenting distance distribution functions calculated from C_α -only model with $l=2$ and 3, respectively, indicate that curve 3 still deviates from the GNOM $P(r)$ but curve 4 has a good agreement with the actual distance distribution function. Further increase of l does not lead to a significant improvement of $P_{DR}(r)$ but rather increases the calculation time. Note that, as in the previous GASBOR version, only initial $P_{DR}(r)$ is fully calculated, whereas all the subsequent computations are made by updating this function and the table of distances r_{ij} within the model.

The above way of computation of $P_{DR}(r)$ permits one to account for the finite size of the residue. However, its internal structure is not taken into account and therefore the initial part of the distance distribution function containing the information on the inner structure of a residue should be entirely ignored. As a minimum distance between residues is about 0.38 nm the discrepancy between $P_{DR}(r)$ and $P(r)$ is calculated starting from the fourth bin, i.e. from the distance of 0.35 nm. The Fourier transform of such $P_{DR}(r)$ would differ from the scattering curve calculated from the given DR model at higher angles by approximately a constant term equal to the sum of squared formfactors $g_i(0)$ taken at $s=0$ over all residues. Thus the real space fitting relies less on the high-angle scattering data then the reciprocal space fitting did, which makes the DR method less susceptible to systematic errors in solvent scattering subtraction at higher angles.

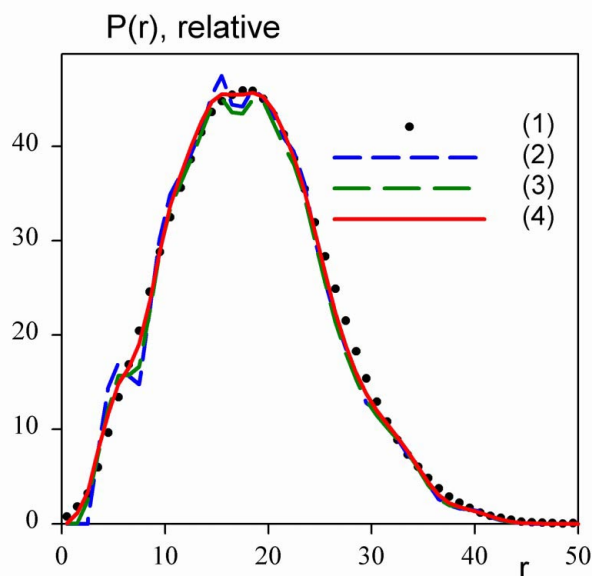


Figure 1

Distance distribution functions of lysozyme. (1) - The $P(r)$ function back transformed from the scattering pattern computed using all-atoms protein representation (Svergun *et al.*, 1995), (2)-(4) - $P(r)$ calculated from C_α -only model, interresidual distances contribute to 1, 5 and 7 bins respectively.

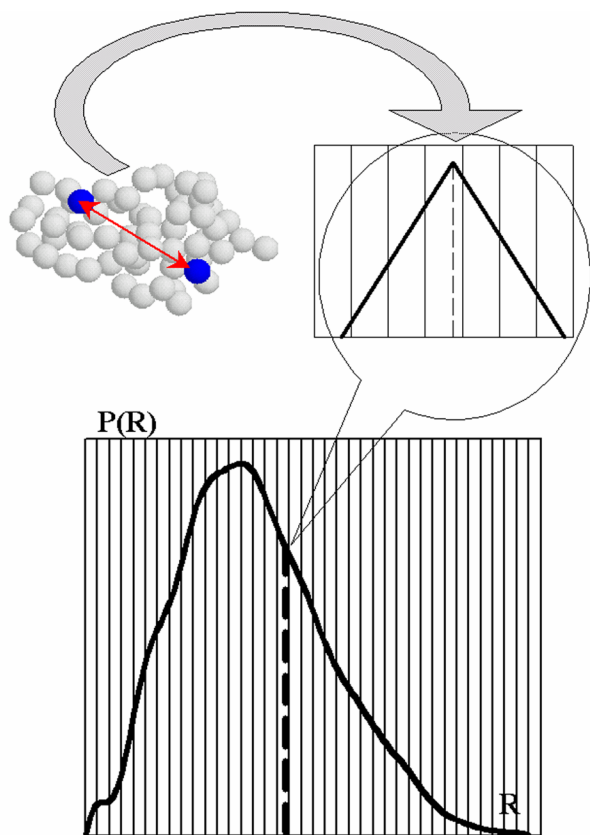


Figure 2

Calculation of $P(r)$ from C_{α} -only model. Each interresidual distance contributes to $2l+1$ bins accordingly to the areas covered by the triangle in the given bin.

4. Addition of missing loops and domains to protein models

Protein function is not only related 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 be absent. 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 using genetic or proteolytic methods. In an attempt of providing a complementary tool for the analysis of incomplete models, Petoukhov *et al.* (2002) have further extended the DR approach to reconstruct missing domains in multisubunit proteins and fusion proteins and to find probable configurations of disordered loops in crystallographic models.

The main idea consists in fixing the known part of the structure (either high or low resolution model) and modelling the missing portions, such as a disordered loop or domain, to fit the experimental scattering data obtained from the entire particle. 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. In all the methods below, the known part of the structure is converted into a set of C_{α} coordinates and represented by an ensemble of DRs or individual residues fixed at these positions. The unknown portion to be generated is also

represented by a set of residues, and the excluded volume penalty prevents its overlap with the fixed part of the model.

Four different algorithms implemented in computer programs CREDO, CHADD, GLOOPY and CHARGE provide an appropriate tool for various situations in which a structure lacks a loop or domain. The choice of the method depends on the information available about 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.*, 2001), 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 residue(s). Further, GLOOPY tries to construct native-like folds by accounting for excluded volumes of side chains (Aszodi *et al.*, 1995), hydrophobic interactions (Huang *et al.*, 1995) knowledge-based potentials (Sippl, 1990; Miyazawa & Jernigan, 1999; Thomas & Dill, 1996) and the distribution of C_{α} bond and dihedral angles (Kleywegt, 1997). If the secondary structure of the missing portion is known, the program CHARGE allows to further constrain the model by incorporating 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 Fig. 3 (see Petoukhov *et al.* (2002) for more detailed description).

The programs GASBOR (both reciprocal and real space versions), CREDO, CHADD, GLOOPY and CHARGE run on IBM PC compatible machines under Windows 9x/NT/2000/XP and Linux as well as on major Unix platforms. 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 through P6 and P222 through P62 are supported). The executable codes of the programs are available from the Web site of the EMBL, Hamburg Outstation: www.embl-hamburg.de/ExternalInfo/Research/Sax.

5. Applications

The above methods were applied to experimental scattering data to construct DR models of a number of proteins with known and unknown crystal structure and to add missing fragments to partial protein models. The experimental wide-angle X-ray scattering pattern of urate oxidase from *Aspergillus Flavus* (UOX) in Fig. 4 was recorded at the D24 beamline at LURE (Orsay, France). The distance distribution function of the particle computed by GNOM is presented in Fig. 4, insert. UOX is a tetramer possessing a P222 point symmetry with MM = 130 kDa and a total of 1180 residues, PDB entry 1uox; (Colloc'h *et al.*, 1997). The models of the tetrameric UOX restored assuming P222 symmetry from the scattering pattern and from the distance distribution function by reciprocal and real space versions of GASBOR are presented in Fig. 5. The evaluation of the model took 111 h of CPU time for the reciprocal space fitting and only 28 hours for the real space fitting. The DR models were automatically aligned with the C_{α} coordinates in the crystal structure using the program SUPCOMB (Kozin & Svergun, 2001) minimizing a dissimilarity measure between two models as a normalized spatial discrepancy (NSD). After alignment both models yield a good agreement with the atomic model in the crystal (NSD = 1.15 and 1.11 for the reciprocal and the real space fitting, respectively).

Another example illustrates the reconstruction of a missing loop using information about secondary structure. The scattering profile from the homodimeric protein R2 of ribonucleotide reductase from *E. coli* (PDB entry 1xik (Logan *et al.*, 1996), MM = 79 kDa) was recorded at the X33 beamline of the EMBL at DESY (Hamburg, Germany). The crystallographic model was solved to 1.7-Å resolution containing 341 residues per monomer. The C-terminal 35 residues are missing 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; Kuprin *et al.*, personal communication).

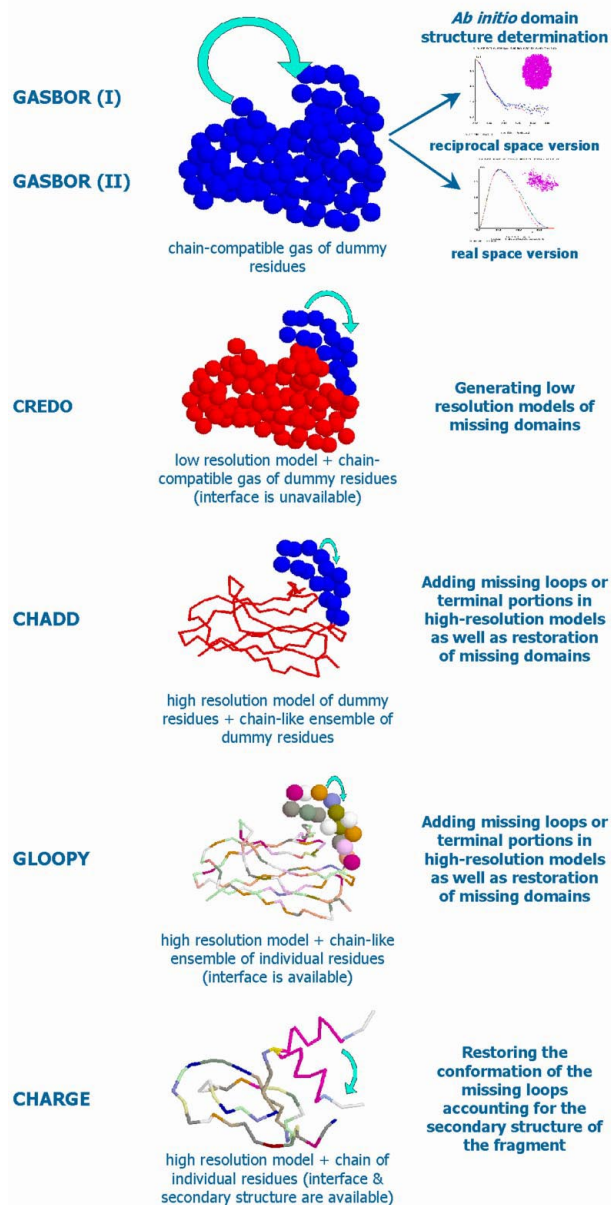


Figure 3
Schematic illustration of the model types, programs and their applications.

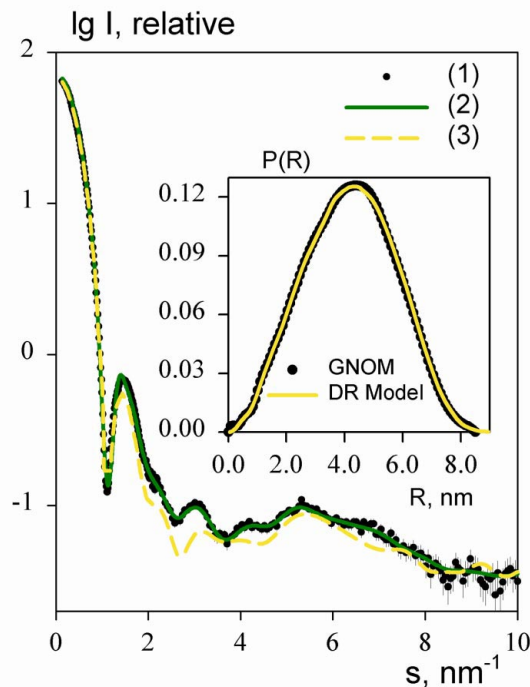


Figure 4
X-ray scattering from urate oxidase (1) and scattering from the DR models: reciprocal space fitting model (2), model obtained by fitting in real space (3). The insert displays the experimental distance distribution function of UOX computed by GNOM (dots) and that of the real space constructed model (full line).

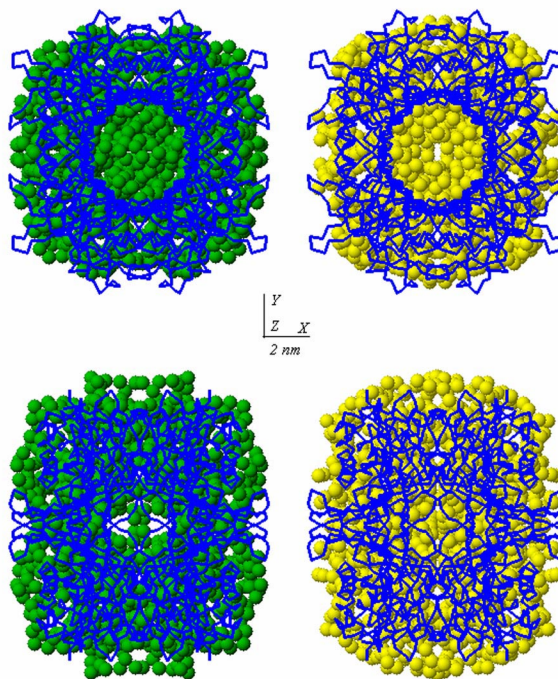


Figure 5
Atomic model of urate oxidase (blue C_{α} -chain) superimposed with the *ab initio* DR models displayed as spheres. Green and yellow models are obtained by fitting in reciprocal and real space, respectively. The bottom view is rotated by 90° counterclockwise around X axis.

According to secondary structure prediction programs (Cuff & Barton, 1999; Cuff & Barton, 2000; Cuff *et al.*, 1998), a major portion of the missing fragment (residues 345 to 373) is predicted to form an α -helix. Fig. 7 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).

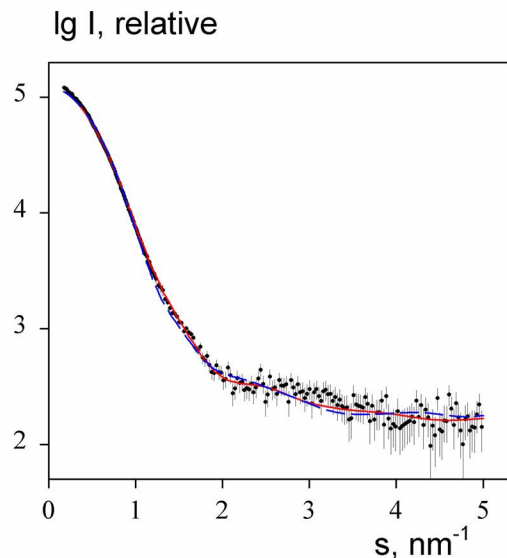


Figure 6

X-ray scattering patterns from the protein R2 (dots with error bars), scattering of the crystallographic model where the missing fragments are absent (dashed line) and scattering from the reconstructed model (full line).

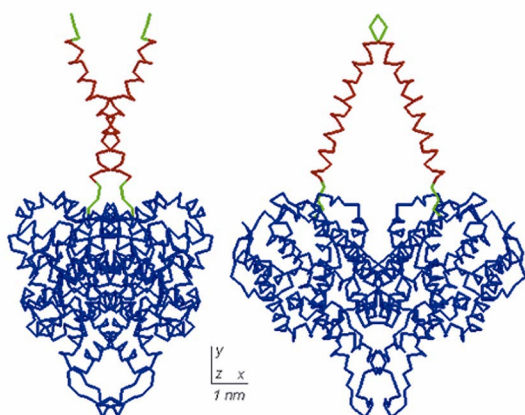


Figure 7

Reconstruction of missing loops in the R2 protein. Crystallographic model is displayed in blue, the reconstructed C-terminal loops in dimeric R2 obtained by CHARGE are shown in green and red (the α -helical portion of the variable domain). The right view is rotated by 90° counterclockwise around twofold symmetry axis (Y).

6. Conclusion

A new *ab initio* method (condensation of a gas of residues) to analyze domain structure of proteins accounting for small and wide angle solution scattering pattern yields substantially more reliable and higher resolution models than previous shape determination methods. The extension of this method further permits to reconstruct missing loops or domains in incomplete high or low resolution models of proteins from the scattering data. The new modelling techniques make solution scattering a useful complementary tool for a large-scale structural characterization of proteins.

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, YSF 00-50 and YSC 02-4364. The authors are indebted to M.H.J. Koch, P. Vachette and S. Kuprin for providing the experimental data and to F. Bonneté for the gift of urate oxydase.

References

- Aszodi, A., Gradwell, M.J., & Taylor, W.R., (1995). *J. Mol. Biol.*, **251**(2): p. 308-326.
- Burley, S. K. (2000). *Nat Struct Biol* **7 Suppl**, 932-934.
- Chacon, P., Moran, F., Diaz, J. F., Pantos, E. & Andreu, J. M. (1998). *Biophys J* **74**, 2760-75.
- Colloc'h, N., el Hajji, M., Bachet, B., L'Hermite, G., Schiltz, M., Prange, T., Castro, B. & Mornon, J. P. (1997). *Nat Struct Biol* **4**, 947-52.
- Cuff, J. A. & Barton, G. J. (1999). *Proteins* **34**, 508-19.
- Cuff, J. A. & Barton, G. J. (2000). *Proteins* **40**, 502-11.
- Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M. & Barton, G. J. (1998). *Bioinformatics* **14**, 892-3.
- Debye, P., (1915). *Ann. Physik* **46**: p. 809-823
- Edwards, A. M., Arrowsmith, C. H., Christendat, D., Dharamsi, A., Friesen, J. D., Greenblatt, J. F. & Vedadi, M. (2000). *Nat Struct Biol* **7 Suppl**, 970-2.
- Huang, E.S., Subbiah, S. & Levitt, M., (1995). *J Mol Biol*, **252**(5): p. 709-20.
- Kirkpatrick, S., Gelatt, C.D., & Vecchi, M.P. (1983) *Science*. **220**: p. 671-680
- Kleywegt, G.J., (1997). *J Mol Biol*, **273**(2): p. 371-376
- Kozin, M. B. & Svergun, D. I. (2001). *J. Appl. Crystallogr.* **34**, 33-41.
- Logan, D. T., Su, X. D., Aberg, A., Regnstrom, K., Hajdu, J., Eklund, H. & Nordlund, P. (1996). *Structure* **4**, 1053-64.
- Miyazawa, S. & Jernigan, R.L., (1999). *PROTEINS: Structure, Function, and Genetics*, **34**: p. 49-68.
- Petoukhov, M. V., Eady, N. A. J., Brown, K. A. & Svergun, D. I. (2002). *Biophys. J.* (in press).
- Sippl, M.J., (1990). *J Mol Biol*, **213**(4): p. 859-83.
- Svergun, D. I. (1992). *J. Appl. Crystallogr.* **25**, 495-503.
- Svergun, D. I. (1999). *Biophys J* **76**, 2879-86.
- Svergun, D. I., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Crystallogr.* **28**, 768-773.
- Svergun, D. I., Petoukhov, M. V. & Koch, M. H. J. (2001). *Biophys J* **80**, 2946-53.
- Svergun, D. I., Semenyuk, A. V. & Feigin, L. A. (1988). *Acta Crystallogr.* **A44**, 244-250.
- Svergun, D. I., Volkov, V. V., Kozin, M. B. & Stuhmann, H. B. (1996). *Acta Crystallogr.*, **A52**: p. 419-426
- Thomas, P.D. & Dill, K.A., (1996). *Proc Natl Acad Sci U S A*, **93**(21): p. 11628-33.
- Walther, D., Cohen, F. E. & Doniach, S. (2000). *J. Appl. Crystallogr.* **33**, 350-363.