Journal of
Applied Crystallography
ISSN 0021-8898

Advanced solution scattering data analysis methods and their applications

D. I. Svergun

J. Appl. Cryst. (2000). **33**, 530–534

Advanced solution scattering data analysis methods and their applications†

† Advanced data analysis methods

Dmitri I. Svergun [ab*]

[a]*EMBL, Hamburg Outstation, Notkestraße 85, D-22603 Hamburg, Germany*
[b]*Institute of Crystallography, Russian Academy of Sciences, Leninsky pr. 59, 117333 Moscow, Russia*
Email:Svergun@EMBL-hamburg.de

A method for *ab initio* low-resolution shape and internal structure retrieval from contrast variation solution scattering data is described. The method uses a multiphase model of a particle build from densely packed dummy atoms and employs simulated annealing to find a compact interconnected configuration of phases that fits the available experimental data. In the particular case of a single phase particle (shape determination) the method is compared to another *ab initio* method using low resolution envelope functions. Examples of the shape determination of several proteins from experimental X-ray scattering data are presented.

1. Introduction

Investigation of quaternary structure of biological macromolecules in solution remains one of the most important fields of application of the small angle scattering (SAS) technique. SAS permits to analyze individual macromolecules and their complexes in nearly physiological conditions and to directly study structural responses to the changes in physical and chemical environment. Although yielding low-resolution information only, SAS is applicable in a broad range of conditions and sizes of macromolecules (Feigin & Svergun, 1987). The method is an important complementary tool to the high-resolution techniques (X-ray crystallography and nuclear magnetic resonance (NMR)) as the latter are usually applied to individual macromolecules in rather specific conditions.

The SAS intensity $I(s)$ from a dilute monodisperse solution of macromolecules is proportional to the spherically averaged intensity from a single particle (s denotes the modulus of the scattering vector, $s=(4\pi/\lambda)\sin\theta$, λ is the wavelength, 2θ is the scattering angle). The Shannon sampling theorem (Shannon & Weaver, 1949, Moore, 1980) states that the number of parameters (Shannon channels) required to represent $I(s)$ on an interval $[s_{min}, s_{max}]$ is equal to $N_s = D_{max}(s_{max} - s_{min})/\pi$ where D_{max} is the maximum particle size. In practice, solution scattering curves decay rapidly with s and the typical number of the Shannon channels does not exceed 10 to 15.

Given the low information content, unambiguous interpretation of the scattering data using three-dimensional (3D) models is only possible if the variety of models is appropriately restrained. In particular, biologically meaningful models of macromplecular complexes can be constructed by rigid body refinement if high resolution atomic structures of the individual domains are available from X-ray crystallography or NMR (*e.g.* Krueger *et al.*, 1997; Ashton *et al.*, 1997; Svergun *et al.*, 1998). A computer system for automated and interactive rigid body refinement is described in this issue (Kozin & Svergun, submitted). For *ab initio* methods, models described by $M\approx N_s$ parameters or those with $M>>N_s$ parameters out of which only $M'\approx N_s$ are independent, should be used. Here, M' can be reduced by incorporating *a priori* information (*e.g.* on particle symmetry or that taken from electron microscopy), whereas N_s can be increased by performing contrast variation experiments.

An *ab initio* shape determination method developed by Svergun & Stuhrmann (1991) and Svergun *et al.* (1997) represents the particle shape using an angular envelope function

$$F(\omega) = \sum_{l=0}^{L} \sum_{m=-l}^{l} f_{lm} Y_{lm}(\omega), \qquad (1)$$

where (r,ω) are polar coordinates, $Y_{lm}(\omega)$ are spherical harmonics and the multipole coefficients f_{lm} are complex numbers (Stuhrmann, 1970). Using this parameterization, 3D bodies are at low resolution represented by a few independent parameters only (in the general case, $M=(L+1)^2-6$). It was demonstrated by Svergun *et al.* (1996) that a unique shape reconstruction is achieved if M does not exceed $1.5N_s$. In the present paper, a new method for low-resolution shape and internal structure retrieval is presented and the shape determination using the two *ab initio* methods is compared using experimental scattering data.

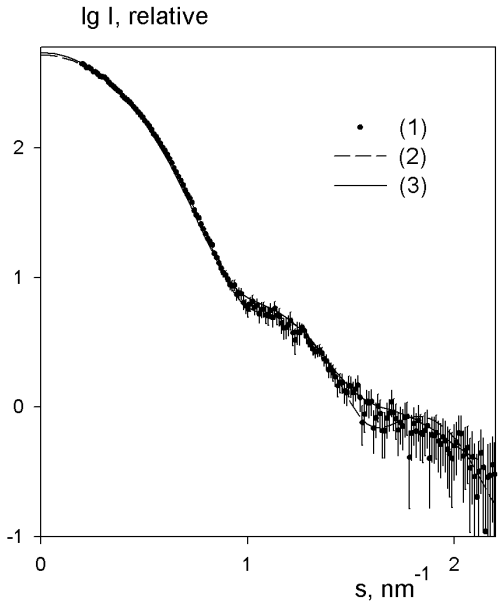
2. Structure analysis using dummy atoms model

Below, the new technique is briefly described (for more details, see Svergun, 1999). A model of a K-phase particle ($K\geq 1$) is constructed as follows. A volume enclosing the particle (*e.g.* an overall particle shape provided by electron microscopy or simply a sphere of radius $R=D_{max}/2$) is filled by densely packed spheres of radius $r_0 << D_{max}$ (dummy atoms). Each dummy atom is assigned a number k_j indicating to which phase it belongs [k_j ranges from *0* (=solvent) to K]. The particle shape and structure are completely described by a phase assignment (configuration) vector X_j with $M\approx(D_{max}/2r_0)^3>>1$ components.

Assuming that the atoms of the k-th phase have contrast $\Delta\rho_k$, the scattering intensity from such a dummy atoms model (DAM) is expressed as

$$I(s)=2\pi^2 \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \left(\sum_{k=1}^{K} \left[\Delta\rho_k A_{lm}^{(k)}(s)\right]^2 + 2\sum_{n>k} \Delta\rho_k A_{lm}^{(k)}(s) \Delta\rho_n \left[A_{lm}^{(n)}(s)\right]^* \right) \qquad (2)$$

Here, the partial amplitudes from the volume occupied by the k-th phase are



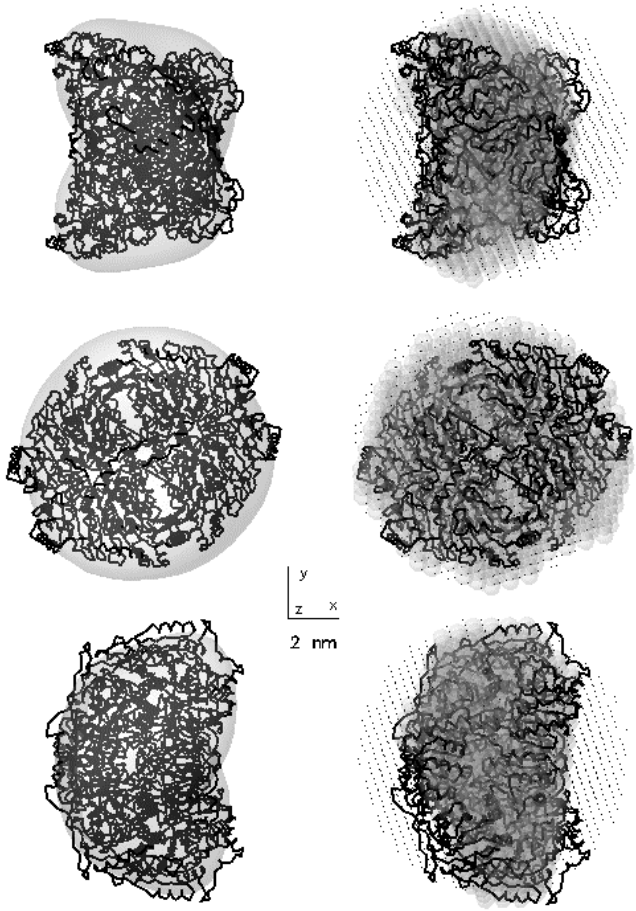


Figure 1
Shape determination of the pyruvate decarboxylase from *Zymmomonas mobilis*. (a): experimental solution scattering curve (1), calculated scattering from the envelope model (2), processed data after subtraction of a constant term and scattering from the SA model (3). (b): comparison of the atomic model (C_α-trace) with the envelope function (left panel) and with the restored DAM (right panel; the dummy atoms belonging to the particle are shown as semi-transparent spheres, those belonging to the solvent as dots). Middle and bottom row are rotated counterclockwise by 90^0 around the Y and X axes, respectively. The models were displayed on a SUN Workstation using the program *ASSA* (Kozin, Volkov & Svergun, 1997).

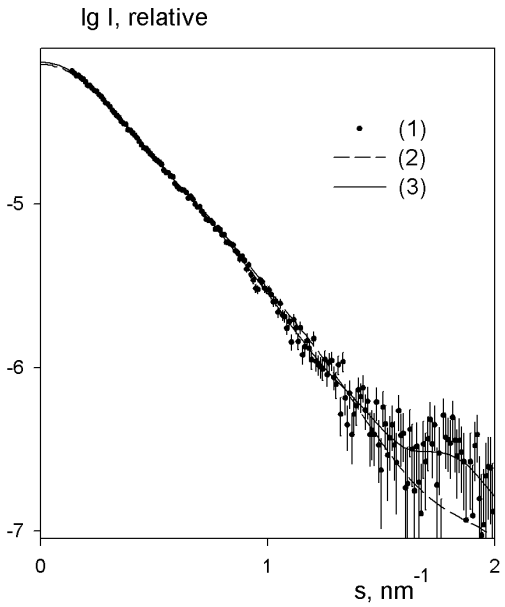
$$A_{lm}^{(k)}(s) = i^l \sqrt{2/\pi} \sum_{j=1} j_l(sr_j) Y_{lm}^*(\omega_j) \qquad (3)$$

where the sum runs over the dummy atoms belonging to the *k*-th phase and $j_l(x)$ denotes the spherical Bessel function. Equations (2-3) allow to rapidly compute the scattering curves from a multi phase DAM for an arbitrary configuration *X* and arbitrary contrasts of the components.

As the number of dummy atoms *M* significantly exceeds N_s, constraints must be imposed on possible configurations to reduce the number of independent parameters *M*'. First, the model should have low resolution with respect to r_0, *i.e.* the volumes occupied by the phases cannot contain just a single atom or a few atoms only. Having defined for each dummy atom a list of contacts in the first coordination sphere (*i.e.* list of neighboring atoms at an offset $2r_0$), looseness (degree of isolation) of a non-solvent atom is calculated as $P(N_e) = exp(-0.5N_e) - exp(-0.5N_c)$. Here, N_e is the number of contacts belonging to the same phase and $N_c=12$ is the coordination number for hexagonal packing. Looseness of the configuration *X* (*i.e.* its non-compactness) is characterized by the average value $P(X) = \langle P(N_e) \rangle$.

Another constraint requires interconnectivity of phases. A phase is interconnected if two arbitrarily selected atoms belonging to the phase can be connected by a trajectory lying within this phase (a trajectory consists of steps made within the first coordination sphere). For a given configuration, all interconnected fragments (graphs) in a phase are found, and the value $G_k(X)=ln(N_k/M_k)\geq 0$ is computed, where N_k and M_k are the numbers of dummy atoms in the entire *k*-th phase and in the longest graph, respectively.

The task of retrieving a low resolution DAM from the scattering data is formulated as follows: find the vector $\{X\}$ for *M* dummy atoms minimizing the function $f(X) = \chi^2 + \alpha[P(X)+\Sigma G_k(X)]$. Here, χ denotes the overall discrepancy between the experimental $I_{exp}(s)$ and calculated $I_{calc}(s)$ intensity sets at all contrasts available (possibly more then one), $\alpha>0$ is the weight of the looseness penalty, and the sum runs over the phases that must be interconnected. The global minimization of $f(X)$ is performed using simulated annealing (SA). The idea of this method is to randomly modify the vector *X* while always moving to the configurations that decrease $f(X)$ but sometimes moving also to those increasing $f(X)$. The probability of accepting the latter moves decreases in the course of the minimization (Kirkpatrick *et al.*, 1983). Each move involves a random change of the phase assignment of one atom only, and it is sufficient to update a single term in sum (3) to compute the partial amplitudes, which speeds up the evaluation of $f(X)$ significantly.



3. Shape determination

The virtue of the above method was illustrated (Svergun, 1999) by an *ab initio* structure restoration of a ribosome-like two-phase model. Its further use for the analysis of contrast variation experiments on hybrid 70S *E.coli* ribosomes permitted to simultaneously fit 42 neutron and X-ray scattering curves and yielded a 3D map of the protein-RNA distribution revealing the likely positions of individual proteins in the ribosomal subunits (Svergun & Nierhaus, submitted). Below, the practically important case of a single-phase particle (*K=1*) is considered, in which the method performs *ab initio* shape determination from a single ("shape") scattering curve. The search volume is a sphere of radius $D_{max}/2$ and the vector X contains values 0 or 1, similar to the bead modeling of Chacón *et al*. (1998). The shape determination using the DAM is illustrated in practical examples (X-ray scattering data from proteins, some of them with known structure in the crystal) and compared to the multipole expansion method of Svergun *et al*. (1996, 1997).

In all computations below, a constant was subtracted from the experimental data to diminish the scattering contribution due to internal particle structure and to ensure that the intensity decays as s^{-4} following Porod's (1982) law. The values of D_{max} were determined using the program ORTOGNOM (Svergun, 1993) and the r_0 was selected to have $M\approx10^3$ atoms in the DAM. The normalized discrepancy χ was taken

$$\chi^2 = \frac{\sum_{k=1}^{N} \left\{ W(s_k)[I_{calc}(s_k) - I_{\exp}(s_k)] \right\}^2}{\sum_{k=1}^{N} \left[W(s_k) I_{\exp}(s_k) \right]^2} \qquad (4)$$

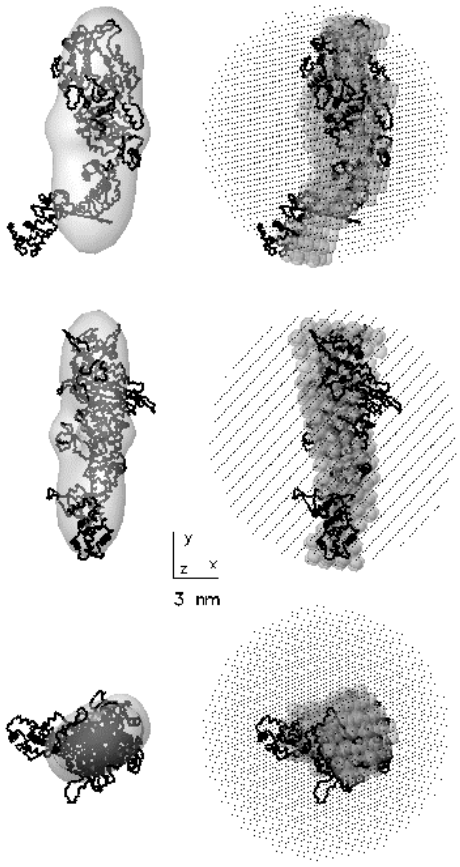
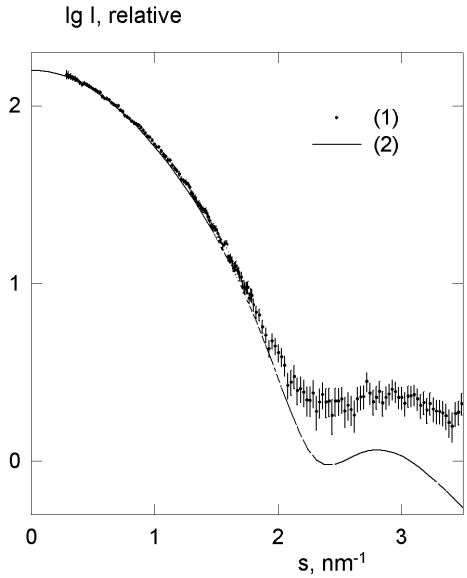


Figure 2
Shape determination of the myosin subfragment S1. Notations are as in Figure 1.

where N is the number of knots in the reciprocal space. To speed up the computations, the experimental data were processed using the program GNOM (Svergun, 1992). The processed curves were recomputed on N=51 knots and the weighting function $W(s)=s^2$ was used. The processed intensities were always neatly fitted and the default penalty weight α=0.01 ensured that the penalty term yielded major contribution to $f(X)$ at the end of the minimisation (typically, $\chi^2\approx10^{-5}$ and $\alpha P(X)\approx10^{-4}$). The synchrotron radiation scattering curves from the pyruvate decarboxylase (PDC, König *et al*., 1998) and chitin binding protein CHB1 (Svergun *et al*., submitted) were recorded as part of ongoing projects at the EMBL, Hamburg Outstation. The scattering data from the myosin subfragment 1 (S1) were measured at the Stanford Synchrotron Radiation Laboratory (Mendelson *et al*., 1996). A typical minimization run required about 2-3 hours on a 400 MHz Pentium-II machine. As both shape determination methods yielded models at an arbitrary orientation and handedness, they were appropriately rotated for the comparisons.

Fig. 1 presents the *ab initio* shape determination of PDC from *Zymmomonas mobilis*, a tetrameric protein of molecular weight (MW) 244 kDa (König *et al.*, 1998). The maximum diameter of the particle is D_{max}=11 nm and the scattering curve in Fig. 1a contains N_s=*8.3* Shannon channels. The envelope in Fig. 1b, left column, was restored by the program SASHA (Svergun *et al*., 1997) using



spherical harmonics up to L=6 and assuming the point 222 symmetry (13 free parameters). The DAM (Fig. 1b, right column) was obtaned at r_0=0.4 nm (total number of dummy atoms M=1939, that in the final model M_1=1031). As seen from comparison with the PDC structure from the Brookhaven Protein Data Bank (PDB, Bernstein *et al*, 1977), entry 1zpd (Dobritzsch *et al.*, 1998), both *ab initio* methods yield a fair low resolution shape restoration. The envelope function represents the structural details somewhat better, which is partially due to use of the symmetry restrictions.

The shape restoration of the S1, a very elongated protein of (MW= 125 kDa, D_{max}=17 nm) is illustrated in Fig. 2. The scattering curve in Fig. 2a contains N_s=12.1 Shannon channels and the envelope reconstruction was performed at L=4 without symmetry restrictions (19 free parameters). The crystallographic model in Fig. 2b is based on the PDB entry 2mys (Rayment *et al*., 1993) and contains additional loops missing in the crystal structure generated using secondary structure prediction methods (Mendelson *et al*., 1996). As seen from Fig. 2b, left panel, the envelope function is only able to represent general asymmetry of the protein. In contrast, the DAM in Fig. 2b, right column (r_0=0.6 nm, M=2103, M_1=214) provides fairly adequate description of the particle shape.

The two examples indicate that the restoration of the envelope function might be preferable for globular particles. For highly anisometric particles, angular envelope function (1) may not adequately represent details of the particle shape and the use of the DAM should be superior. In any case, comparison of the two reconstructions enables one to assess the reliability of the model.

Fig. 3 presents the shape restoration of the CHB1 from *Streptomyces olivaceoviridis* illustrating the influence of the DAM packing radius and of the penalty term. CHB1 has MW=18 kDa, D_{max}= 6 nm and the

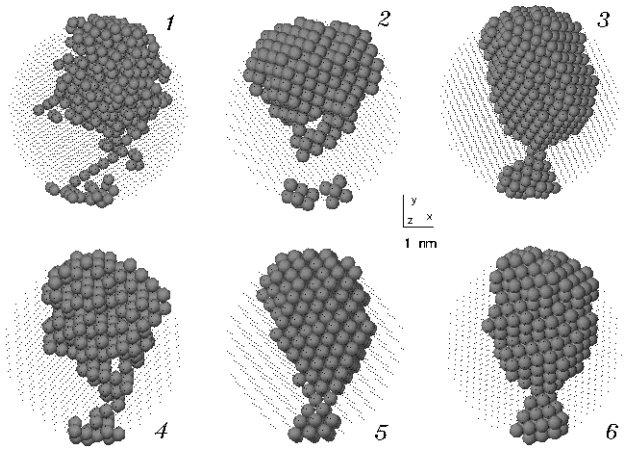


Figure 3
Shape determination of the chitin binding protein. (a): experimental solution scattering curve (1) and processed data after subtraction of a constant term and scattering from the DAMs (2). (b): the DAMs restored at different conditions (see text).

portion of the scattering curve useful for the shape analysis (Fig. 3a) contains N_s= *6.6* Shannon channels. The models 1-3 in Fig. 3b were obtained for r_0=0.2 nm (M=2503) at default α=0.01, but, for models 1 and 2, looseness and disconnectivity contributions, respectively, were omitted in the penalty term leading to obviously unacceptable results. Models 4-6 correspond to r_0=0.25 nm (M=1272) with increasing penalty weights (α=0.001, 0.005 and 0.02, respectively). Scattering from all these models in Fig. 3a is indistinguishable from each other and also from the processed data.

4. Discussion

On one hand, the comparison in Fig. 3 demonstrates that an unconstrained shape restoration using the DAM with $M>>N_s$ parameters is ambiguous. On the other hand, requirements of compactness and connectivity seem to be sufficient for unambiguous reconstruction. Moreover, low-resolution features of the restored models are stable to variations of r_0 and α (cf. models 3, 5 and 6 in Fig. 3b). In fact, at later annealing stages the penalty is decreased rather than χ, *i.e.* the program searches for a compact interconnected solution constrained by the fit. The resolution of the final model thus depends on the information content in the data rather than on the total number of dummy atoms *M*. The more information is provided by the data, the more stringent is the constraint, *i.e.* the more detail should be kept in the model.

The looseness and disconnectivity penalty apparently force the method to select the level of detail required for uniqueness; another question is, whether this unique solution represents adequately the actual shape of the particle. Calculations on geometrical bodies indicate that globular and elongated models are uniquely restored from the simulated scattering curves. For extremely flat bodies with the anisometry factor exceeding 1:5, the method reproducibly finds configurations that are more compact than the initial model. (and yield virtually the same scattering curve in the given range of the

momentum transfer). Further analysis of the uniqueness problem is now in progress.

The results obtained for various proteins (see Figs 1b and 2b and examples presented by Svergun, 1999) suggest that in practice the simulated annealing does provide an adequate restoration of the low resolution shape. Caution is, however, required, especially when studying low MW macromolecules, where the contribution from the internal structure limits significantly the angular range suitable for the shape determination.

The computer program DAMMIN implementing the shape determination method runs on an IBM-PC and on the major UNIX platforms and permits to include additional information about the particle shape (*e.g.* point symmetry restrictions, expected anisometry). The program executables and user instructions are available from the URL
http://www.embl-hamburg.de/ExternalInfo/Research/Sax/index.html.

The author thanks M. H. J. Koch, S. König, R. A. Mendelson and G. Grüber for providing the experimental scattering data. The work was supported by the EU Grant BIO4-CT97-2143.

References

Ashton, A. W., Boehm, M. K., Gallimore, J. R., Pepys M. B. & Perkins, S. J. (1997) *J. Mol. Biol.* **272**, 408-422.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol*. **112**, 535-542.

Chacón, P., Morán, F., Díaz, J. F., Pantos, E. & Andreu, J. M. (1998) *Biophys. J.* **74**, 2760-2775.

Dobritzsch, D., König, S., Schneider, G. & Lu, G. (1998) *J. Biol. Chem.* **273**, 20196-20204

Feigin, L. A. & Svergun, D. I. (1987) *Structure analysis by small-angle X-ray and neutron scattering.* New York: Plenum Press.

König, S., Svergun, D.I., Volkov, V.V., Feigin, L.A. & Koch, M. H. J. (1998) *Biochemistry*, **37**, 5329-5334

Kirkpatrick, S., Gelatt, C. D., Jr. & Vecci, M.P. (1983) Science **220,** 671-680.

Kozin, M. B., Volkov, V. V. & Svergun, D. I. (1997) *J. Appl. Cryst.* **30**, 811-815.

Krueger, J. K., Olah, G. A., Rokop, S. E., Zhi, G., Stull, J. T., and Trewhella, J. (1997). *Biochemistry* **36**, 6017-6023.

Mendelson, R.A., Schneider, D.K. & Stone, D.B. (1996) *J. Mol. Biol.* **256**, 1-7.

Moore, P. B. (1980) *J. Appl. Cryst*. **13**, 168-175.

Porod, G. (1982) General theory. In *Small-angle X-ray scattering*, O.Glatter and O.Kratky, editors. London: Academic Press, pp. 17-51.

Rayment, I., Rypniewski, W.R., Schmidt-Base K., Smith, R., Tomchick, D.R., Benning, M.M., Winkelmann, D.A., Wesenberg, G. & Holden H.M. (1993) *Science* **261,** 50-58.

Shannon, C. E. & Weaver, W. (1949) *The mathematical theory of communication*. Urbana: University of Illinois Press.

Stuhrmann, H. B. (1970) *Zeitschr. Physik. Chem. Neue Folge* **72**, 177-198.

Svergun, D. I. (1992) *J. Appl. Cryst*. **25**, 495-503.

Svergun, D. I. (1993) *J. Appl. Cryst*. **26**, 258-267.

Svergun, D. I. (1999) *Biophys. J.* **76,** 2879-2886.

Svergun, D. I. & Stuhrmann, H. B. (1991) *Acta Cryst*. **A47**, 736-744.

Svergun, D. I., Volkov, V. V., Kozin, M. B. & Stuhrmann, H. B. (1996) *Acta Cryst*. **A52**, 419-426.

Svergun, D.I., Volkov, V.V., Kozin, M.B., Stuhrmann, H.B., Barberato, C. & Koch, M.H.J. (1997) *J. Appl. Cryst.* **30**, 798-802.

Svergun, D.I., Aldag, I., Sieck, T., Altendorf, K.-H., Koch, M.H.J., Kane, D.J., Kozin, M.B. & Grüber, G. (1998) *Biophys. J.* **75,** 2212-2219.

 Received 12 June 1999 • Accepted 24 January 2000