Journal of
Applied Crystallography

ISSN 0021-8898

Received 16 August 2006
Accepted 17 January 2007

Accuracy of molecular mass determination of proteins in solution by small-angle X-ray scattering

Efstratios Mylonas[a] and Dmitri I. Svergun[a,b]*

[a]European Molecular Biology Laboratory, Hamburg Outstation, 22603 Hamburg, Germany, and [b]Institute of Crystallography, Moscow, Russia. Correspondence e-mail: dmitri@embl-hamburg.de

One of the most important overall parameters, which can be derived from small-angle X-ray scattering (SAXS) experiments on macromolecular solutions is the molecular mass (MM) of the solute. In particular, for a monodisperse protein solution, MM of the solute is calculated from the extrapolated scattering intensity at zero angle $I(0)$. Assessing MM by SAXS provides valuable information about the oligomeric state and absence of unspecific aggregation in solution. The value of MM can either be estimated by comparison with a protein standard with a known MM or by determining the absolute scattering intensity using, *e.g.*, water scattering. In both cases, knowledge about the solute concentration and about the partial specific volume of the protein is required. By measuring 13 well characterized globular proteins with MMs ranging from 13.7 to 669 kDa we analyze the sources of possible systematic deviations and assess the accuracy of MM determination using SAXS. The data indicate that all these proteins have approximately the same 'effective' value of the partial specific volume of about 0.7425 $cm^3\ g^{-1}$. It is shown that both inter-protein and water calibration can be used for molecular mass determination by SAXS and in most cases the errors do not exceed 10%.

1. Introduction

One of the most straightforward parameters to derive from small-angle scattering data on macromolecular solutions is the molecular mass (MM) of the solute. Although small-angle X-ray scattering (SAXS) is less accurate than, *e.g.* mass spectroscopy, in determining the MM, the former method allows measurements in solution, closer to the native state. One of the most common applications of SAXS is the determination of the oligomeric state of the biomolecule (*e.g.* a protein or a macromolecular complex) or monitoring of aggregation or degradation processes, which can be readily done by assessing the MM value.

For a monodisperse protein solution, the characteristic parameter directly associated with MM is the intensity at zero angle $I(0)$ which can be calculated easily using the Guinier approximation (Guinier, 1939) or an indirect transformation program (Glatter, 1977; Svergun, 1992). As it is not possible to measure the absolute intensity of the protein directly (Russell, 1983), one has to resort to secondary standards. In SAXS, standard proteins with known molecular masses are often used such as lysozyme [*e.g.* Hammel *et al.* (2002)], bovine serum albumin [*e.g.* Petoukhov *et al.* (2003)] or glucose isomerase (Kozak, 2005). Alternatively, scattering from secondary standards like Lupolen (Kratky, 1964), or water (Orthaber *et al.*, 2000) can be used to obtain the scattering from the solute on the absolute scale and then to calculate the MM.

For the calibration, knowledge about the solute concentration (c) and partial specific volume of the protein ($\bar{\nu}$) is crucial. When using the standard proteins it is typically assumed that the $\bar{\nu}$ values of the standard and actually measured protein are identical. With this assumption, the ratio of the molecular masses of the two proteins is identical to the ratio of their $I(0)$s normalized against the concentrations. When using water, the $\bar{\nu}$ value explicitly enters the equation to compute the MM, and in fact, the latter value is rather sensitive to the changes in $\bar{\nu}$ (Feigin & Svergun, 1987). In practice, inaccuracies of concentration and partial specific volume often become larger sources of errors in MM determination than the precision of calculating $I(0)$ itself.

In the present paper, a systematic study is performed to assess the accuracy of MM determination in SAXS. Solutions of well characterized and commercially available proteins covering a wide range of MMs from 13.7 to 669 kDa were measured to determine the $I(0)$ values. Possible sources of systematic errors are analyzed and repetitive measurements are employed to minimize the errors in the measured solute concentrations and in extrapolated $I(0)$ values. Based on the results, an 'optimum' value of partial specific volume for globular proteins in solution is proposed.

2. Materials and methods

2.1. Protein preparation and concentration determination

Proteins ribonuclease A, chymotrypsinogen A, ovalbumin, aldolase, catalase and thyroglobulin were part of the LMW (low-molecular weight) and HMW (high-molecular weight) gel-filtration calibration kits from GE Healthcare (product codes 17-0442-01 and 17-0441-01, respectively). Carbonic anhydrase, alcohol dehydrogenase, β-amylase and apoferritin were part of the kit for molecular weights 29000–700000 from Sigma (product code MWGF 1000). Lysozyme from chicken egg white was from Fluka (product code 62971), BSA Type H2 from Gerbu Biotechnik (product code 1064) and glucose isomerase from Hampton Research (product code HR7-100). All proteins except apoferritin and glucose isomerase were in

powder form and were either dissolved in low-salt buffers or dialyzed overnight after dissolving. Apoferritin and glucose isomerase were dialyzed overnight into the appropriate buffers. The buffers used were 100 m*M* Tris 100 m*M* NaCl, pH 7.5 for all proteins except lysozyme (40 m*M* acetic acid 50 m*M* NaCl pH 4.0), BSA (50 m*M* HEPES pH 7.5) and glucose isomerase (100 m*M* Tris 1 m*M* $MgCl_2$ pH 8.0). The solute concentrations were determined by the absorption of the protein solutions at 280 nm using either an Eppendorf spectrophotometer (in 6 *M* guanidinium chloride buffer) or a Nanodrop ND-1000 spectrophotometer (in normal dialysis buffer).

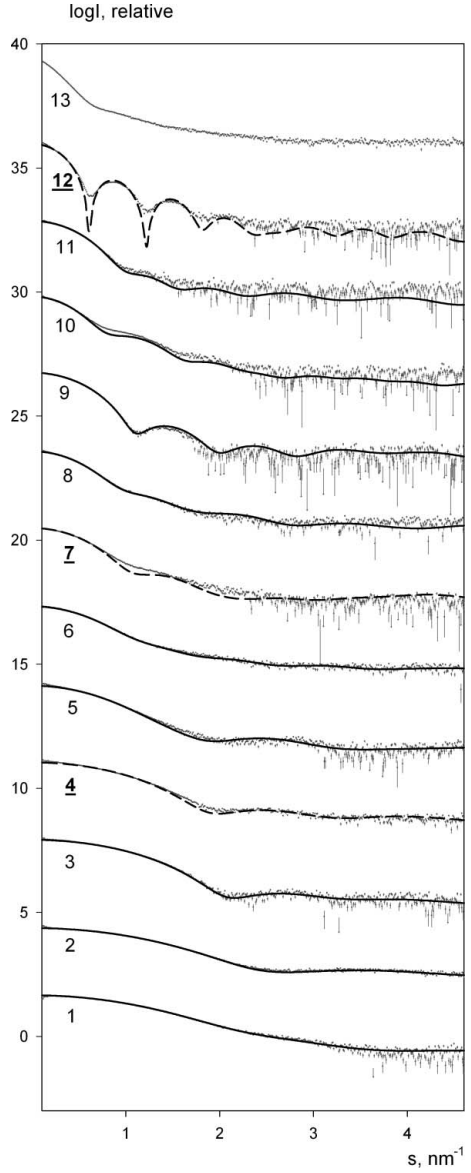

Figure 1
Experimental scattering and the fits. The experimental data are displayed as dots with error bars, the fits from the crystallographic models computed by *CRYSOL* (Svergun *et al.*, 1995) are displayed as solid lines. The fits for the outliers which are shown with dashed lines and their curves are numbered in bold. The logarithm of the scattering intensity (I) is plotted as a function of the modulus of the scattering vector s; the fits are appropriately displaced along the logarithmic axis for better visualization. (1) ribonuclease A, (2) lysozyme, (3) chymotrypsinogen A, (4) carbonic anhydrase, (5) ovalbumin, (6) bovine serum albumin (BSA), (7) alcohol dehydrogenase, (8) aldolase, (9) glucose isomerase, (10) β-amylase, (11) catalase, (12) apoferritin and (13) thyroglobulin (no high resolution model is available for the latter). The outliers (carbonic anhydrase, alcohol dehydrogenase, apoferritin) are underlined.

Repetitive measurements of the absorption were performed to improve the accuracy; no significant differences were observed between the measurements of the native and the denatured states of the protein. The extinction coefficients of the proteins were calculated using the online tool *ProtParam* (Gasteiger *et al.*, 2005).

2.2. Partial specific volume determination

The protein solutions with concentrations 5 and 10 mg ml^{-1} and the appropriate buffers were prepared and measured on an Anton Paar DMA 5000 densitometer. $\bar{\nu}$ can be calculated using the formula

$$\bar{\nu} = \frac{1}{d_{\rm buff}}\left(1 - \frac{d_{\rm p} - d_{\rm buff}}{c}\right)$$

where $d_{\rm buff}$ is the density of the buffer, $d_{\rm p}$ is the density of the protein solution and c is the protein concentration. A freeware program, *SEDNTERP* (Philo *et al.*, 1995–2006) was used to calculate the partial specific volumes based on the amino acid composition of the proteins. This program predicts $\bar{\nu}$ for 298 K but these values were adjusted to the actual temperature of the measurements (288 K), and the correction did not exceed 1%. The experimental values reported by different authors (Durchschlag, 1986; Perkins, 1986; Harpaz *et al.*, 1994) were also used.

2.3. SAXS data collection and processing

Synchrotron X-ray scattering data from solutions of the aforementioned proteins were collected at the X33 beamline of the EMBL (DESY, Hamburg) (Koch & Bordas, 1983) using a MAR345 image-plate detector. The scattering patterns were measured with exposure times ranging 2–5 min at 288 K. The concentration of the solutes was about 2–5 mg ml^{-1} for proteins with MM > 50 kDa and about 5–12 mg ml^{-1} for proteins with MM < 50 kDa. The sample-to-detector distance was 2.7 m and the range of the modulus of the scattering vector covered was $0.09 < s < 5$ nm^{-1} [$s = 4\pi \sin(\theta)/\lambda$, where 2θ is the scattering angle and $\lambda = 0.15$ nm is the X-ray wavelength]. The constant water scattering was determined by subtracting the scattering of the empty cuvette from that filled with distilled water at 293 K. The measurements of the proteins and of water scattering were repeated in six separate experimental sessions and the results were averaged.

The data were processed using standard procedures and normalized against concentration using the program package *PRIMUS* (Konarev *et al.*, 2003). The forward scattering $I(0)$ and the radii of gyration $R_{\rm g}$ were evaluated using the Guinier approximation (Guinier, 1939) assuming that at very small angles ($s < 1.3R_{\rm g}$) the intensity is represented as $I(s) = I(0)\exp[-(sR_g)^2/3]$. Additionally, $I(0)$s and $R_{\rm g}$s as well as the maximum dimensions $D_{\rm max}$ and the interatomic distance distribution functions $p(r)$ were computed using the indirect transform package *GNOM* (Svergun, 1992). Comparison with known high-resolution models was made using the program *CRYSOL* (Svergun *et al.*, 1995), which fits the experimental intensity by adjusting the excluded volume of the particle and the contrast of the hydration layer.

2.4. Molecular mass calculation

2.4.1. Protein calibration. When using a standard protein for calibration the simple formula

$$\mathrm{MM_p} = I(0)_{\rm p}/c_{\rm p}\,\frac{\mathrm{MM_{st}}}{I(0)_{\rm st}/c_{\rm st}}$$

is used, where $I(0)_{\rm p}$ and $I(0)_{\rm st}$ are the scattering intensities at zero angle of the studied and the standard protein, respectively, $\mathrm{MM_p}$ and

Table 1
Experimental and computed structural parameters of the studied proteins.

Protein	MM (kDa)	PDB used	Abs $I(0)/c$ (10^{-2} cm^2 mg^{-1})	R_g (nm)	MM_1 (kDa)	MM_2 (kDa)	MM_3 (kDa)	Δ_1 (%)	Δ_2 (%)	Δ_3 (%)
Ribonuclease A	13.7	1FS3	1.00 ± 0.05	1.58 ± 0.04	10.3	13.9	10.0	−24.6	1.4	−27.0
Lysozyme	14.3	1LYZ	1.03 ± 0.06	1.43 ± 0.04	11.2	14.3	11.1	−21.9	0.0	−22.7
Chymotrypsinogen A	25.0	2CGA	1.85 ± 0.03	1.85 ± 0.01	22.9	25.7	23.4	−8.3	2.6	−6.4
Carbonic anhydrase	29.0	1V9E	2.39 ± 0.15	2.08 ± 0.03	30.8	33.1	29.4	6.4	14.2	1.5
Ovalbumin	45.0	1OVA	3.21 ± 0.08	2.66 ± 0.04	41.9	44.6	46.0	−6.8	−1.0	2.2
BSA	66.0	1N5U	4.84 ± 0.13	2.99 ± 0.08	60.4	67.1	62.2	−8.5	1.7	−5.7
Alcohol dehydrogenase	150	1JVB	6.24 ± 0.26	3.27 ± 0.03	85.3	86.4	92.6	−43.1	−42.4	−38.3
Aldolase	158	1ZAH	11.16 ± 0.70	3.51 ± 0.13	144	155	150	−8.7	−2.2	−5.2
Glucose isomerase	173	1OAD	11.58 ± 0.24	3.25 ± 0.07	137	160	–	−20.6	−7.2	–
β-Amylase	200	1FA2	13.53 ± 1.32	4.22 ± 0.04	174	187	–	−12.9	−6.3	–
Catalase	232	4BLC	15.85 ± 0.40	3.84 ± 0.13	187	220	197	−19.5	−5.4	−15.2
Apoferritin	440	1IER	25.91 ± 2.56	7.05 ± 0.21	324	359	345	−26.3	−18.4	−21.6
Thyroglobulin	669	–	53.01 ± 1.88	7.56 ± 0.19	622	734	–	−7.0	9.7	–
Average								16.5	8.7	14.6

Notation: MM_1 and Δ_1 were computed using $\bar{\nu}$ values predicted from the sequence (Table 2), MM_2 and Δ_2 were computed using $\bar{\nu} = 0.7425$ cm^3 g^{-1} and MM_3 and Δ_3 were computed using reported experimental values for $\bar{\nu}$ (see Table 2). The average in the bottom row is calculated from the absolute values of the deviations.

MM_{st} are the corresponding molecular masses and c_p and c_{st} are the concentrations. Here, the ratio of the expected molecular mass to the $I(0)$ normalized against concentration, was calculated for each protein. Then, the average of these ratios was determined, excluding the outliers (*i.e.* the proteins deviating too much from the rest). Subsequently, the multiplication of this value with the $I(0)$s gives us the MMs by intercalibration between the proteins. Essentially this procedure calculates the MM of each protein considering all other proteins as calibrants, since we know their actual MMs.

2.4.2. Absolute *I*(0) determination using water. To calculate the forward scattering $I(0)$ in the absolute scale, the known scattering of water 1.632×10^{-2} cm^{-1} at 288 K was used (Orthaber *et al.*, 2000). By dividing the relative $I(0)$s of the proteins with the experimental constant scattering of water and then multiplying by the absolute scattering of water one obtains the $I(0)$s of the proteins in absolute scale. To calculate the molecular mass (MM) in kDa we used the formula (Feigin & Svergun, 1987; Orthaber *et al.*, 2000) $MM = [N_A I(0)/c]/\Delta\rho_M^2$, where $I(0)/c$ is the forward scattering normalized against concentration, $\Delta\rho_M = [\rho_{M,prot} - (\rho_{solv}\bar{\nu})]r_o$ is the scattering contrast per mass, $N_A = 6.023 \times 10^{23}$ mol^{-1} is the Avogadro number, $\rho_{M,prot} = 3.22 \times 10^{23}$ e g^{-1} is the number of electrons per mass of dry protein, $\rho_{solv} = 3.34 \times 10^{23}$ e cm^{-3} is the number of electrons per volume of the aqueous solvent, $\bar{\nu}$ is the partial specific volume of the protein and $r_o = 2.8179 \times 10^{-13}$ cm is the scattering length of an electron.

3. Results

Fig. 1 shows representative scattering curves of the proteins and the theoretical patterns computed from the available crystallographic models of the same or of highly homologous proteins taken from the Protein Data Bank (PDB; Bernstein *et al.*, 1977). The PDB codes of the crystallographic models are presented in Table 1 (no homologous structure is available for thyroglobulin). The fits to the curves calculated from the crystallographic models are rather good in most cases but there are also some outliers (fits displayed in dashed lines), which indicate that the crystal structure or oligomeric composition of the protein in the crystal differs from that in solution. Table 1 summarizes the radii of gyration (R_g) and the $I(0)$s (in absolute scale after normalization against concentration and water scattering) of the proteins using the Guinier extrapolation (the results represent average values from six independent experimental sessions). The $I(0)$ and R_g values calculated by *GNOM* (not shown) are very similar. A good agreement is observed with the previously reported values for other proteins, *e.g.* glucose isomerase (Kozak, 2005) and lysozyme (Orthaber *et al.*, 2000).

Table 2
Partial specific volumes ($\bar{\nu}$) (cm^3 g^{-1}) of proteins.

Protein	Experimental this work	Calculated *SEDNTERP*	Experimental previously reported
Ribonuclease A	0.73 ± 0.01	0.7072	0.703
Lysozyme	–	0.7133	0.712
Chymotrypsinogen A	0.76 ± 0.01	0.7296	0.732
Carbonic anhydrase	–	0.7345	0.729
Ovalbumin	0.75 ± 0.01	0.7357	0.746
BSA	0.74 ± 0.02	0.7305	0.734
Alcohol dehydrogenase	–	0.7411	0.750
Aldolase	–	0.7347	0.739
Glucose isomerase	–	0.7246	–
β-Amylase	0.75 ± 0.03	0.7343	–
Catalase	–	0.7239	0.730
Apoferritin	–	0.7309	0.738
Thyroglobulin	–	0.7234	–

Initially, we assessed the overall consistency of the results by intercalibration. Three proteins (carbonic anhydrase, alcohol dehydrogenase and apoferritin) display significant deviations between the calculated and expected MMs while the remaining ten proteins were self-consistent and showed good agreement with an average ratio of expected MM to $I(0)$. As seen from Fig. 1, the scattering patterns computed from the crystal structures of the three former proteins yield the curves deviating significantly from the experimental data. Given these deviations, oligomeric states of the three proteins may be different from that expected from the crystal structure (*e.g.* partial dissociation may take place), but also unspecific aggregation and impurities cannot be excluded. The three proteins were thus omitted from the further analysis, and after disregarding the three outliers the average deviation between the expected and calculated MMs was 8.7%.

In the third column of Table 2 the partial specific volumes are given calculated from the amino acid composition (Durchschlag, 1986; Perkins, 1986; Harpaz *et al.*, 1994). The MMs computed using these values are given in Fig. 2 and Table 1. It can be seen that for most of the proteins the MMs are significantly underestimated. Indeed, for five of the proteins the deviations exceed 20% and not a single MM is within 5% from the expected value (the overall discrepancy between the calculated and expected values is 16.5%). When using the

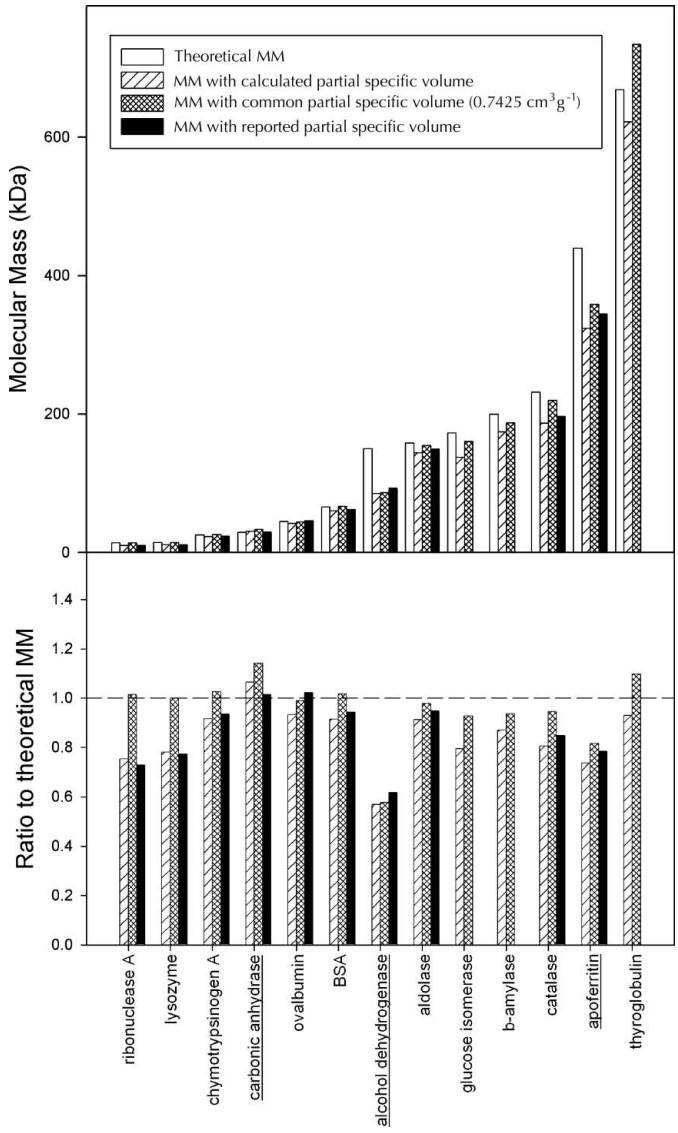

Figure 2
Comparison of the MMs for the different proteins examined. The lower panel displays the ratios of the calculated values to the expected MMs. The outliers (carbonic anhydrase, alcohol dehydrogenase, apoferritin) are underlined.

consensus experimental values published by Durchschlag (1986), Perkins (1986) and Harpaz *et al.* (1994) (Table 2, fourth column; for three proteins no such values are available), similar results with the overall discrepancy of 14.6% were obtained (Fig. 2). We also measured the values of $\bar{\nu}$ for all proteins experimentally on an Anton Paar densitometer as described in *Materials and methods*, but a reliable experimental determination proved to be difficult as a large amount (2 ml) of sample was required for each measurement. Significant deviations were observed between the individual density measurements, and for many samples the value of $\bar{\nu}$ could not thus be reliably assessed. It is interesting to note that in all the cases when reproducible values were obtained, they significantly exceeded the values predicted from the amino acid sequence (see Table 2).

It is a common practice to assume that most of globular proteins have a similar value for partial specific volume [around 0.74 cm^3 g^{-1}; *e.g.* Feigin & Svergun (1987)]. This value is considerably higher than the calculated one for all proteins (except for alcohol dehydrogenase), and the calculated MMs would consequently be much closer to the theoretical values. We have thus optimized the common value of $\bar{\nu}$ to have the smallest deviations between the calculated and expected MMs for all ten proteins, which were consistent with each other while using the intercalibration procedure. The average obtained was 0.7425 cm^3 g^{-1}, very close to the empirical common value above for the globular proteins. The calculated MMs given in Table 1 agree much better with the expected MMs than the values computed using the predicted $\bar{\nu}$, with the average deviation going down to 8.7% (not unexpectedly, to the deviation obtained with the intercalibration procedure).

4. Conclusions and discussion

The results above demonstrate that SAXS is able to provide MM estimates within an error of about 10% provided the solute concentration is measured with an accuracy of 5–10%, usually achievable in spectrophotometric experiments. This range of precision is sufficient for a reliable determination of the oligomeric state of proteins (*e.g.* monomers *vs* dimers). The use of standard proteins and water calibration give the same level of accuracy and the two approaches are easily interchangeable. In particular, lysozyme and bovine serum albumin, the proteins most often used for calibration, display deviations of 0 and 1.7%, respectively, from the expected values, and can be safely used as standards. Both intercalibration and water calibration require additional measurements (standard protein measurement in the first case and blank sample compartment in the second case) so it is a matter of convenience to employ one or the other options.

Another interesting, although perhaps not unexpected, result is that globular proteins in solution appear to have a common 'effective' partial specific volume of about 0.7425 cm^3 g^{-1}, which is significantly larger than most of the amino acid-derived (Philo *et al.*, 1995–2006) or earlier reported experimental (Durchschlag, 1986; Perkins, 1986; Harpaz *et al.*, 1994) partial specific volumes. Indeed, the use of the (smaller) amino acid-derived volumes yields the MM estimates incompatible with the values expected from the sequence, and also our measurements of $\bar{\nu}$ suggest that they should be higher than the predicted values. Moreover, the calculated values based on the amino acid data are valid for proteins in pure water and the cosolvents in the buffer may also influence the partial specific volume. This means that, ideally, they should be determined in the very buffer used for a SAXS experiment. The major difficulty in measuring $\bar{\nu}$ is that reliable experimental determination using a densitometer requires dozens of mg of protein. Usually, the yield of protein expression and purification does not permit one to have a sufficient amount of protein for the densitometric measurements and it is thus difficult to expect that the experimental values will be available in most of practical studies. We suggest therefore the 'effective' value of 0.7425 cm^3 g^{-1}, which should provide sufficiently good accuracy (on average, within 10%) for most of the globular proteins.

We would like to thank Jan Skov Pedersen and the staff of the Physical Chemistry of Soft Condensed Matter group (Department of Chemistry, Aarhus University) for giving us the opportunity to perform the density measurements for the partial specific volume determination.

References

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.

Durchschlag, H. (1986). *Thermodynamic Data for Biochemistry and Biotechnology*, edited by H.-J. Hinz, p. 45. Berlin: Springer-Verlag.
Feigin, L. A. & Svergun, D. I. (1987). *Structure Analysis by Small-Angle X-ray and Neutron Scattering*. New York: Plenum Press.
Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D. & Bairoch A. (2005). *Protein Identification and Analysis Tools on the ExPASy Server* in *The Proteomics Protocols Handbook*, edited by J. M. Walker, pp. 571–607. Totowa, USA: Humana Press.
Glatter, O. (1977). *J. Appl. Cryst.* **10**, 415–421.
Guinier, A. (1939). *Ann. Phys.* (*Paris*), **12**, 161–237.
Hammel, M., Kriechbaum, M., Gries, A., Kostner, G. M., Laggner, P. & Prassl, R. (2002). *J. Mol. Biol.* **321**, 85–97.
Harpaz, Y., Gerstein, M. & Chothia, C. (1994). *Structure*, **2**, 641–649.
Koch, M. H. J. & Bordas, J. (1983). *Nucl. Instrum. Methods*, **208**, 461–469.
Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J. & Svergun, D. I. (2003). *J. Appl. Cryst.* **36**, 1277–1282.
Kozak, M. (2005). *J. Appl. Cryst.* **38**, 555–558.
Kratky, O. (1964). *Fresenius J. Anal. Chem.* **201**, 161–194.
Orthaber, D., Bergmann, A. & Glatter, O. (2000). *J. Appl. Cryst.* **33**, 218–225.
Perkins, S. J. (1986). *Eur. J. Biochem.* **157**, 169–180.
Petoukhov, M. V., Svergun, D. I., Konarev, P. V., Ravasio, S., van den Heuvel, R. H., Curti, B. & Vanoni, M. A. (2003). *J. Biol. Chem.* **278**, 29933–29939.
Philo, J., Hayes, D. B. & Laue, T. (1995–2006). *SEDNTERP*, http://www.jphilo.mailway.com/.
Russell, T. (1983). *J. Appl. Cryst.* **16**, 473–478.
Svergun, D. I. (1992). *J. Appl. Cryst.* **25**, 495–503.
Svergun, D. I., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.