Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

SASSIM: a method for calculating small-angle X-ray and neutron scattering and the associated molecular envelope from explicit-atom models of solvated proteins

Franci Merzel[a,b] and Jeremy C. Smith[a]*

[a]IWR – Biocomputing, Universität Heidelberg, INF 368, D-69120 Heidelberg, Germany, and [b]National Institute of Chemistry, Hajdrihova 19, SI-1000 Ljubljana, Slovenia

Correspondence e-mail: biocomputing@iwr.uni-heidelberg.de

Received 26 July 2001
Accepted 15 November 2001

A method is presented to calculate efficiently small-angle neutron and X-ray solution scattering intensities from explicit-atom models of macromolecules and the surrounding solvent. The method is based on a multipole expansion of the scattering amplitude. It is particularly appropriate for extensive configurational averaging, as is required for calculations based on computer-simulation results. In test calculations, excellent agreement with experiment is found between neutron and X-ray scattering profiles calculated from a molecular-dynamics simulation of lysozyme in water. The question of definition of the protein surface is also addressed. For comparison with the continuum model, an analytical envelope around the protein is defined in terms of spherical harmonics and is calculated using a Lebedev grid. The analytical surface thus defined is shown to reproduce well the scattering profile calculated from the explicit-atom model of the protein.

1. Introduction

The purpose of most computational techniques for interpreting small-angle scattering (SAS) is to derive the low-resolution particle structure from the observed scattered intensity. Examples of model-fitting approaches for this are given in Henderson (1996), Zhang *et al.* (1996) and Svergun (1999). However, as small-angle solution X-ray and neutron scattering (SAS) techniques improve, it is becoming of increasing interest to develop methods for rapid evaluation of SAS profiles from explicit-atom coordinates (see, for example, Pickover & Engelman, 1982; Svergun *et al.*, 1995, 1998).

Explicit-atom models of biological macromolecules are available from X-ray crystallography and NMR spectroscopy. Many of these models also contain a partial description of the water of hydration of the system. In addition, molecular-dynamics (MD) simulation is frequently used with explicit solvent to examine detailed macromolecular and solvent properties in atomic detail.

In this paper, we present a method for rapidly evaluating SAS profiles from explicit-atom models. The accompanying program is called *SASSIM* (Small-Angle Scattering SIMulation). The method uses a multipole expansion that efficiently calculates the spherically averaged scattering pattern from the model system. For subsequent analysis, an analytical envelope representing the protein surface is defined.

To test the method, the results of a molecular-dynamics (MD) simulation of hen egg-white lysozyme in explicit water are used and compared with the corresponding experimental profiles obtained from Svergun *et al.* (1998), who performed

X-ray scattering in H_2O and neutron scattering in H_2O and D_2O. The calculated and experimental profiles are in good agreement, tesitifying to the accuracy of the method. The importance of inclusion of the explicit solvent molecules is demonstrated. Finally, it is shown that the continuum model of the protein calculated using the analytical envelope reproduces well the scattering from the explicit-atom model of the protein.

2. Methods

2.1. Small-angle scattering

SAS techniques provide information on low- and medium-resolution structural features of the examined system (see Higgins & Benoit, 1994). The coherent neutron scattering intensity I is a function of the scattering vector $\mathbf{q}$ and is defined as a product of the total scattering amplitude,

$$A(\mathbf{q}) = \sum_j^N b_j \exp(-i\mathbf{q}\mathbf{r}_j), \tag{1}$$

and its complex conjugate A^*,

$$I(\mathbf{q}) = A(\mathbf{q})A^*(\mathbf{q}) = \sum_{ij}^N b_i b_j \exp(-i\mathbf{q}\mathbf{r}_{ij}), \tag{2}$$

where $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ and the summation is over N scatterers, j of scattering length b_j. The scattering vector $\mathbf{q}$ is related to the scattering angle 2θ and the wavelength of the incident radiation λ by the equation $|\mathbf{q}| = (4\pi/\lambda)\sin\theta$. For X-ray scattering, b_j is determined by the electron density $\rho_j(\mathbf{r})$ of atom j, $b_j \rightarrow b_j(\mathbf{q}) = \int \rho_j(\mathbf{r}) \exp(i\mathbf{q}\mathbf{r})\, \mathrm{d}r^3$. The X-ray scattering lengths are q dependent.

As molecules in solution scatter isotropically, (2) must be orientationally averaged, which we denote $\langle\rangle_\Omega$. Moreover, since the system undergoes configurational changes during the experiment/simulation, one also needs to include configurational averaging of the system, denoted by $\langle\rangle_c$,

$$I(q) = \langle\langle I(\mathbf{q})\rangle_\Omega\rangle_c = \sum_{ij} b_i b_j \left\langle \frac{\sin(qr_{ij})}{qr_{ij}} \right\rangle_c. \tag{3}$$

The experimental conditions of solution scattering lead to results which can be interpreted owing to the *excess* scattering from the scattering object, *i.e.* from the protein and the solvent perturbed from the bulk. Therefore, it is convenient to distinguish between two parts of the scattering object. The first part, which gives rise to the excess scattering, consists of the protein plus all solvent that contains perturbation of the time-averaged density from the bulk. The second part consists of bulk solvent which is characterized by the scattering length density b_0 – this provides the reference scattering density and does not contribute to the excess scattering.

We define a model system to be one protein molecule (lysozyme in the present test case) surrounded by water molecules forming a sphere centred at the centre of mass of the protein (Fig. 1). The radius R of the sphere is chosen to be sufficiently large that the time-averaged density of the water in the outer shell, Σ_0, is homogeneous and that of bulk water. In other words, the influence of the protein on the density of the solvent shell in this region is negligible. This assures that there is no excess scattering arising from the finite size of the model scattering system. The value of R in the present calculations was 34.2 Å and Δ was chosen to be 4 Å.

In the present calculations, the excess scattering density is determined from all explicit atoms in the MD simulation, *i.e.* all protein and water atoms $j = 1, \ldots, N$ within radius R. The effect of the solvent outside radius R is modelled by a continuum and invokes Babinet's principle (see Fraser *et al.*, 1978). According to this principle, the scattering lengths b_j are corrected in the following way,

$$b_j \rightarrow b_j - V_j \bar{b}_0 f_j(q), \tag{4}$$

where V_j is the volume displaced (excluded) by the jth explicit atom, $\bar{b}_0$ is the bulk scattering density of the solvent and $f_j(q)$ is the normalized Fourier transform of the shape of the excluded volume associated with atom j.

Including redefinition (4) in (1) together with the time dependence of the explicit-atom coordinates gives the total excess scattering amplitude,

$$A(\mathbf{q}, t) = A_0(\mathbf{q}, t) - B_0(\mathbf{q}, t), \tag{5}$$

where

$$A_0(\mathbf{q}, t) = \sum_j^N b_j \exp[-i\mathbf{q}\mathbf{r}_j(t)] \tag{6}$$

and

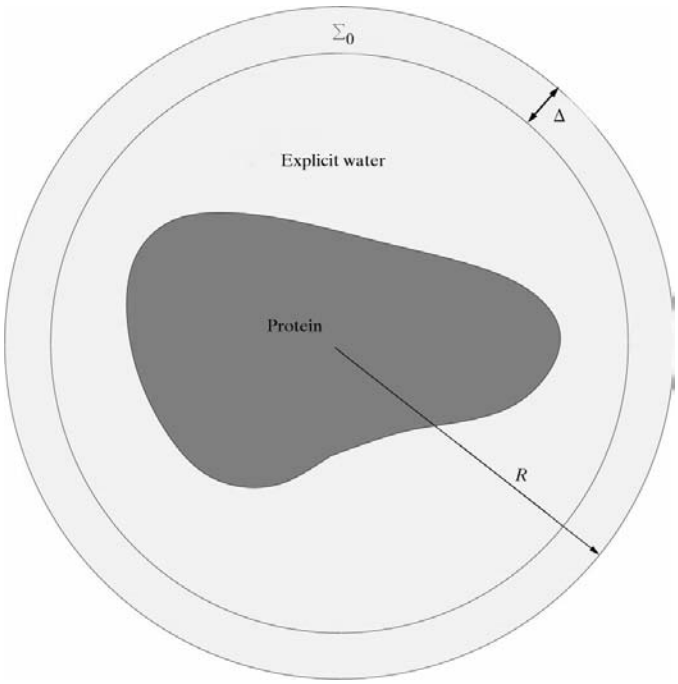

Figure 1
The model system. The average distribution of the explicit water in the outer shell Σ_0 of thickness Δ is used for determination of properties of the bulk solvent.

$$B_0(\mathbf{q}, t) = \bar{b}_0 \sum_j^N V_j f_j(q) \exp[-i\mathbf{q}\mathbf{r}_j(t)]. \quad (7)$$

$A_0(\mathbf{q}, t)$ is the scattering amplitude of the atomic detail system. $\mathbf{r}_j(t)$ denotes the time-dependent radius vector of the jth atom. $B_0(\mathbf{q}, t)$ describes the scattering amplitude of the system volume that is excluded from the background uniform bulk solvent and is modelled using virtual atoms as described later. The bulk scattering density $\bar{b}_0$ can be determined for each type of scattering from the outer shell Σ_0 or, alternatively, from a separate pure solvent simulation. In the former case, the spherically averaged scattering is given by

$$\bar{b}_0(t) = \frac{\sum_{j\in\Sigma_0} b_j \sin[qr_j(t)]/qr_j(t)}{4\pi \int_{R-\Delta}^{R} r^2 \sin(qr)/qr \, \mathrm{d}r}.$$

In the test case, for both methods for determining $\bar{b}_0$ the same results were obtained, confirming that the outer shell water is not perturbed from the bulk. The following $\bar{b}_0$ values resulted: $\bar{b}_0$(X-ray) = 0.3419 Å^{-3}, $\bar{b}_0$(n-H_2O) = -0.5694×10^{10} cm^{-2} and $\bar{b}_0$(n-D_2O) = 6.486×10^{10} cm^{-2}. The values were found to converge. They were obtained averaged over 2000 configurations of the system corresponding to each 0.2 ps frame in the simulated trajectory from 100 to 500 ps. The above values of $\bar{b}_0$ differ slightly from that of pure water, as nine chloride ions were included in the simulation as explained in §2.2. However, we also ran a simulation of the same system without ions and found no significant difference in the scattered intensities. The calculated X-ray $\bar{b}_0$ was then 0.3348 Å^{-3}, which is in excellent agreement with the theoretical value for pure water (0.3342 Å^{-3}).

Finally, the small-angle excess scattering intensity can be expressed as

$$I(q) = \langle\langle |A_0(\mathbf{q}, t) - B_0(\mathbf{q}, t)|^2\rangle_\Omega\rangle_t, \quad (9)$$

where the configurational average $\langle\rangle_c$ is replaced by the MD-simulation time average $\langle\rangle_t$.

It has been shown that one obtains the expected behaviour of the corrected scattering lengths (4) for X-ray scattering if the atomic excluded volumes are represented by Gaussian spheres instead of a uniform volume V_j (Fraser *et al.*, 1978). The excluded volume density, $\mathcal{G}_j(\mathbf{r})$, is then given by

$$\mathcal{G}_j(\mathbf{r}) = \exp\left[-\left(\frac{\mathbf{r}}{\rho_j}\right)^2\right], \qquad \rho_j = \frac{V_j^{1/3}}{\pi^{1/2}}, \quad (10)$$

where the constant ρ_j is determined from the normalization condition $\int \mathcal{G}_j(\mathbf{r})\,\mathrm{d}r^3 = V_j$. Here, the Gaussian spheres are also used for representing the excluded volume in the neutron scattering calculations.

The Fourier transform of a Gaussian sphere is again Gaussian and corresponds to the 'excluded' atomic form factor,

$$f_j(q) = V_j \exp\left(-\frac{q^2 V_j^{2/3}}{4\pi}\right). \quad (11)$$

To determine the excluded volumes, V_j values are taken from *International Tables for X-ray Crystallography* (1974, Vols. III and IV) and scaled. For the water, the scaling factor κ_s is adjusted such that

$$\kappa_s \left\langle \sum_{i\in\Sigma_0} V_i \right\rangle_t = V_{\Sigma_0}, \quad (12)$$

where the summation includes all the atoms in the outer shell Σ_0 and the average is taken over all MD frames. The excluded volumes of the solvent atoms are then $V_i \rightarrow \kappa_s V_i$.

The protein atom excluded volumes are then determined as follows. From the system trajectory the average number of water molecules $\langle N_w\rangle$ within the sphere of radius R is calculated. Owing to the large amount of water used, the perturbation of the average solvent density by the protein is small. In this case the protein volume V_p is $V_p = (4\pi R^3/3) - \langle N_w\rangle\kappa_s(2V_\mathrm{H} + V_\mathrm{O})$, where V_H and V_O are the tabulated excluded volumes of the H and O atoms, respectively. Consequently, the excluded volumes of the protein atoms are rescaled by κ_p, which follows from $\kappa_p \sum_{i\in p} V_i = V_p$, where i runs over all protein atoms. Values of $\kappa_s = 1.49$ and $\kappa_p = 1.02$ were found in the present test case. That κ_p is close to one testifies to the correctness of the tabulated atomic excluded volumes for proteins. The difference in the scaling factors κ_p and κ_s is consistent with the finding that the packing densities of proteins are close to optimal, whereas that of water is not (Harpaz *et al.*, 1994).

The scattering amplitude of the background is then

$$B_0(\mathbf{q}, t) = \sum_j^N \kappa_{s/p} \bar{b}_0 V_j^2 \exp\left(-\frac{q^2 V_j^{2/3}}{4\pi}\right) \exp[-i\mathbf{q}\mathbf{r}_j(t)]. \quad (13)$$

Here, multipole expansions are used for the representation of total scattering amplitudes, A_0 and B_0, permitting fast evaluation of the spherically averaged scattering intensities. This approach was first applied by Stuhrmann (1970*a*) and is also used in the program *CRYSOL* (Svergun *et al.*, 1995).

The simulation system was selected to be sufficiently large that the excess scattering in the outer shell (defined in Fig. 1) is negligible, *i.e.* $I(q)$ in (9) is effectively zero when it is calculated with the index j (6 and 7) running only over the outer shell atoms. Moreover, simulation frames were deleted from the $I(q)$ calculation if the individual bulk scattering density $\bar{b}_0(t)$ in (8) happened to deviate more than σ (the standard deviation of $\{\bar{b}_0(t)\}$) from the average value of $\bar{b}_0$. Thus, we have a smooth transition to bulk water behaviour, all time-averaged fluctuations from the bulk are encompassed within our simulation volume and thus the simulation volume is not a source of error.

However, in contrast to *CRYSOL*, the present method allows analysis of explicit solvent, avoiding the use of a continuum model of the hydration layer. All required parameters, such as the background scattering density and the solvent atom excluded volumes, are determined *a priori* from the simulation data. The present method does not require a model for the protein surface and therefore does not produce results that depend on this model.

To perform the multipole expansions of scattering amplitudes, the following well known relation is used (see Abramowitz & Stegun, 1970),

$$\exp(i\mathbf{qr}) = 4\pi \sum_{l=0}^{\infty} \sum_{m=-l}^{l} i^l j_l(qr) Y^*_{lm}(\omega_{\mathbf{r}}) Y_{lm}(\Omega_{\mathbf{q}}), \qquad (14)$$

where j_l denotes a spherical Bessel function. The notations $\mathbf{r} = (r, \omega_{\mathbf{r}})$ and $\mathbf{q} = (q, \Omega_{\mathbf{q}})$ are also used here.

The multipole expansions of the scattering amplitudes are

$$A_0(\mathbf{q}, t) = \sum_{lm} A^0_{lm}(q, t) Y_{lm}(\Omega_{\mathbf{q}}),$$

$$A^0_{lm}(q, t) = 4\pi i^l \sum_{k}^{N} b_k j_l[qr_k(t)] Y^*_{lm}[\omega_{\mathbf{r}_k}(t)], \qquad (15)$$

$$B_0(\mathbf{q}, t) = \sum_{lm} B^0_{lm}(q, t) Y_{lm}(\Omega_{\mathbf{q}}),$$

$$B^0_{lm}(q, t) = 4\pi i^l \bar{b}_0 \sum_{k}^{N} V_k f_k(q) j_l[qr_k(t)] Y^*_{lm}[\omega_{\mathbf{r}_k}(t)] \qquad (16)$$

and the intensity (9) can then be expressed using only multipole coefficients,

$$I(q) = \left\langle \sum_{lm} |A^0_{lm}(q, t) - B^0_{lm}(q, t)|^2 \right\rangle_t. \qquad (17)$$

(16) and (17) were evaluated for each set of coordinates generated in the simulated trajectory of the system and averaged to obtain the final SAS profiles. The multipole expansion in (16) and (17) turned out to fully converge for $l \leq l_{\max}$, where $l_{\max} = 17$, when evaluating the scattering profile $0 \leq q \leq 0.5$ Å^{-1}.

Finally, in order to compare the calculated neutron scattering results with experiment the wavelength spread of the neutron beam is included and the calculated intensity modified by the resolution function $R(q, q')$ (see Pedersen *et al.*, 1990)

$$I(q) = \int R(q, q') I(q') \, \mathrm{d}q', \qquad (18)$$

$$R(q, q') = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{(q - q')^2}{2\sigma^2}\right], \quad \sigma = \frac{\Delta\lambda}{\lambda} \frac{q}{2(2\ln 2)^{1/2}},$$

where $\Delta\lambda/\lambda$ is in the present case 0.1 (Svergun *et al.*, 1998). For calculation of the neutron scattering profile in the D_2O solution the water and labile protein protons were exchanged for deuterium.

2.2. MD simulation

The simulation was performed in the canonical ensemble using the *CHARMM* (Brooks *et al.*, 1983) program, version 27b1. All protein and solvent atoms were treated explicitly. The 1.33 Å resolution structure of hen egg-white lysozyme was taken from the Protein Data Bank, entry 193l (Vaney *et al.*, 1996) and was embedded in an explicit aqueous environment. The missing H atoms in the crystal structure of the lysozyme were placed using the *HBUILD* method (Brünger & Karplus, 1988). For convergence in the scattering-profile calculations a sufficiently large amount of water was needed. A box of equlibrated water with the form of truncated octahedron originating from a cubic box of side length 84 Å was used. The TIP3P water model was used for the water (Jorgensen *et al.*, 1983). After immersing the protein into the box of water and removing all the water molecules within 2.7 Å of any protein 8577 water molecules remained. In order to electrostatically neutralize the system, nine chloride ions were also included at random positions. This was required so as to enable the use of the Ewald method in representing the electrostatics in the simulation.

The system was simulated with periodic boundary conditions as an isothermal–isobaric (NPT) ensemble at $T = 300$ K and $p = 101$ kPa. The total simulation time was 500 ps, which was found to be sufficiently long for accurate sampling of the relevent system configurations. The average RMS heavy-atom deviation from the crystal structure was to be 1.72 Å indicating that the protein structure was conserved. For analysis the trajectory from 100 ps to 500 ps was used and coordinates saved every 0.2 ps. Fuller details of the simulation will be published elsewhere.

2.3. Definition of the protein surface

The above method for calculating the explicit-atom scattering profiles does not require the definition of a protein surface. However, for subsequent analysis it is useful to have such a definition and indeed a protein surface is a useful concept in many areas of biophysics. It is of particular interest here in the derivation of a continuum model that reproduces the explicit-atom SAS results for the protein atoms. The way in which this surface passes between the two regions cannot be uniquely defined. A variety of possibilities exists; for example, a contour surface of electronic charge density, a molecular van der Waals surface generated by fused atomic spheres, a solvent-accessible surface, a Voronoi surface *etc.* (Lee & Richards, 1971; Connolly, 1985).

Consider the representation of individual atoms as spheres of some suitably defined radii. Various choices have been proposed for atomic radii according to the property that they represent, *e.g.* excluded volume, electron density, electrostatic potential *etc.*, and the molecular surface obtained with these atomic radii indeed depends on this choice. The individual atom representation leads to a model of fused spheres that has a well defined envelope describing the protein surface. However, this envelope forms a very rough surface, owing to the junctions between the spheres. To obtain a smoother surface closer to an analytical envelope around the protein, we approximate each atom by a Gaussian sphere $\mathcal{G}_j$ as in (10), whose volume V_j corresponds to that of a normal sphere of radius r_j.

Since many proteins have an approximately globular structure, we make use of spherical coordinates (r, θ, φ) with their origin at the centre of mass of the protein. A mesh of points is constructed on the spherical surface, each point representing a direction $\omega_k = (\theta_k, \varphi_k)$. Starting at a distance far from the protein, one can move towards the protein centre of mass along any given radial direction ω_k in a sequence of

distances $r_{k1} > r_{k2} > r_{k3} > \ldots$. At each point $\mathbf{r}_{k\mu}$ the contribution from the protein atoms to the volume density function $\mathcal{V}$ is calculated,

$$\mathcal{V}(\mathbf{r}_{k\mu}) = \sum_j \mathcal{G}_j(\mathbf{r}_{k\mu}) = \sum_j \exp\{-[(\mathbf{r}_{k\mu} - \mathbf{R}_j)/\rho_j]^2\}. \quad (19)$$

The same procedure is performed with the solvent atoms but in the reverse direction, *i.e.* sampling the solvent volume density function from the interior to far outside the protein. In both directions monotonically increasing functions are obtained. Their intersection is considered to define locally the protein surface along the direction ω_k, which we denote by $S_P\omega_k$. This definition considers equally both protein and solvent atoms, in the spirit of the Voronoi definition of a molecular surface (Voronoi, 1908).

Projecting the surface points $\mathbf{r}_k$ onto a spherical harmonics expansion allows the protein surface to be represented as an analytical function of the radial angles θ and φ. This formulation was first introduced by Stuhrmann (1970*b*).

We recall that it is always possible to expand an arbitrary single-valued function $F(\mathbf{r})$ over spherical harmonics $Y_{lm}(\omega)$,

$$F(\mathbf{r}) = F(r, \omega) \simeq F_L(r, \omega) = \sum_{l=0}^{L} \sum_{m=-l}^{l} f_{lm}(r) Y_{lm}(\omega), \quad (20)$$

where the expansion coefficients $f_{lm}(r)$ are defined by

$$f_{lm}(r) = \int_0^{2\pi} \mathrm{d}\varphi \int_0^{\pi} \sin(\theta) F(r, \omega) Y_{lm}(\omega)\, \mathrm{d}\theta. \quad (21)$$

The resolution of the expansion is determined by the truncation value L.

The integral in (21) can be computed using the Gaussian-like quadrature scheme of Lebedev (1975),

$$\int_0^{2\pi} \mathrm{d}\varphi \int_0^{\pi} \sin(\theta) F(r, \omega) Y_{lm}(\omega)\, \mathrm{d}\theta = \sum_k^{N_p} w_k \sin(\theta_k) F(r, \omega_k) Y_{lm}(\omega_k). \quad (22)$$

The quadrature is performed by summing up the values of the function $\sin(\theta)F(r, \omega)Y_{lm}(\omega)$ at predefined directions ω_k weighted by constants w_k.

In the Lebedev scheme all directions ω_k correspond to the nodes of the octahedral grids that provide the optimal efficiency, *i.e.* the highest accuracy with the least possible function evaluations. In order to integrate an arbitrary angular function which is assumed to have an exact expansion over spherical harmonics for $l \leq L$, we need to sample this function at the least possible number of points N_p.

Here, the Lebedev grid is applied with $N_p = 434$ points, which is sufficient for accurate integration of functions with resolution $L = 35$, *i.e.* functions having exact spherical harmonic expansion for l, $0 \leq l \leq 35$. The directions of the grid points ω_k and the weights w_k were taken from Treutler & Ahlrichs (1994). To our knowledge, Lebedev grids are available for up to 1202 points, accurate for $L = 59$.

According to (20) we express the protein surface S_P by its multipolar expansion of resolution L_S, S_l, providing the analytical form of the surface,

$$S_P(\omega) \simeq S_l(\omega) = \sum_{l=0}^{L_S} \sum_{m=-l}^{l} \mathcal{R}_{lm} \mathcal{Y}_{lm}(\omega). \quad (23)$$

As the protein surface is a real function, real spherical harmonics $\mathcal{Y}_{lm}$ can be used, defined as

$$\mathcal{Y}_{lm} = \begin{cases} m > 0 & 1/2^{1/2}[Y_{l-m} + (-1)^m Y_{lm}] \\ m = 0 & Y_{l0} \\ m < 0 & i/2^{1/2}[Y_{l-m} - (-1)^m Y_{lm}] \end{cases}. \quad (24)$$

Given $S_P(\omega_k) = r_k$, the coefficients $\mathcal{R}_{lm}$ are obtained by means of projections

$$\mathcal{R}_{lm} = \int_0^{2\pi} \mathrm{d}\varphi \int_0^{\pi} \sin(\theta) S_P(\omega) \mathcal{Y}_{lm}(\omega)\, \mathrm{d}\theta = \sum_k^{N_p} w_k r_k \mathcal{Y}_{lm}(\omega_k). \quad (25)$$

The integrand in (25) consists of two product functions, $S_P(\omega)$ and $\mathcal{Y}_{lm}(\omega)$. The allowed resolution L for exact integration is related to the whole integrand. In other words, the maximum dimension of the integrand in the space spanned by spherical harmonics is L, which is equal to 35 for this type of Lebedev grid.

The dimension L of the function in the space of the spherical harmonics is the direct sum of dimensions of the composing product functions l_1 and l_2, $L = l_1 \oplus l_2$. Inserting (23) into (25) makes it clear that the highest resolution L_S of the surface S_P one can afford is at most $L/2$ (another $L/2$ drops to $\mathcal{Y}$). In the present case, the maximum resolution of the surface is therefore $L_S = 17$. In the following, we will assume $S_P(\omega) = S_l(\omega)$.

As any definition of a molecular surface contains a degree of arbitrariness, it should be tested by computation of measurable physical quantities. In the present case, the appropriate quantity is the SAS intensity. Here, we compute the X-ray SAS profile for the continuum model of the protein that is defined by the surface $S_P(\omega)$ and compare this with that obtained from the full atomic protein structure. For the continuum protein one obtains the following multipole coefficients of the scattering amplitude,

$$C_{lm}(q) = i^l (2/\pi)^{1/2} \int_{\omega} \mathrm{d}\omega Y^*_{lm}(\omega) \int_0^{S_P(\omega)} j_l(qr) r^2\, \mathrm{d}r, \quad (26)$$

while those for the atomic structure are given by (16), with index k running over only the protein atoms. The corresponding continuum model intensities are obtained through summing the squares of the multipole coefficients.

3. Results

The calculated SAS intensities are compared with experiment in Fig. 2. χ values, a measure of the similarity between the calculated and experimental profiles, are given in Table 1. The agreement between the simulation-derived and experimental profiles is found to be excellent and the differences in the profiles for the three types of experiment are also well reproduced.

In the limit $q \rightarrow 0$, the Guinier approximation to the scattering intensity gives,

$$I(q) \simeq I_0 \exp(-q^2 R_g^2/3), \qquad (27)$$

allowing the radius of gyration R_g of the scattering object to be obtained (Guinier, 1939). The radii of gyration from the different types of scattering extracted from the experimental and calculated scattering profiles are listed in the second and third columns of Table 1. These are also found to lie within the experimental error. These R_g values can also be compared with the 'true' value 14.12 ± 0.10 Å obtained from the mass-weighted atomic structure of the protein, $R_g'^2 = 1/M \sum_i^n (m_i r_i^2)$, where $M = \sum_i m_i$. The systematic error ±0.10 Å in the MD-derived geometric radius of gyration arises from the internal protein dynamics. One can see that the true value is about 5% overestimated by the X-ray technique, meaning that the solvent effects result in the protein appearing larger with X-rays. This suggests an increase of the solvent density in the hydration shell, which is consistent with Svergun *et al.* (1998). R_g values turn out to be underestimated 5–10% by neutron scattering. The protein in D_2O solution has a negative contrast with respect to the background scattering length density $\bar{b}_0$, while in H_2O solution the protein is positive with respect to the background. In both cases, the increase of the solvent density in the hydration shell reduces the apparent radius of gyration. The origin of difference in R_g values is therefore in the non-uniform distribution of scattering lengths in the systems and different contrasts provided by different background scattering length densities $\bar{b}_0$.

An important question concerns to what extent the perturbation of the solvent from the behaviour of bulk water influences the calculated SAS profiles. To examine this, SAS profiles were computed and compared in which the solvent perturbation effect is included (by including all protein and solvent atoms in the summations over k in equations 16 and 17) and when they are neglected (including only the protein atoms, with the entire solvent region modelled as bulk continuum). The χ values in Table 1 show that in all three types of scattering experiment the calculated profile is significantly closer to the experimental one when the solvent molecules are explicitly included.

We now examine whether the surface-dependent continuum model of the protein allows the SAS data generated from the explicit-atom protein model to be reproduced. To do this, the $L = 17$ excluded volume protein surface was calculated and the interior region defined as a continuum. It was found that resolution $L = 17$ is sufficient for obtaining convergent results for X-ray SAS intensities at $q < 0.5$ Å^{-1}. (26) and (16) were used to derive the multipole coefficients of the continuum and atomic models, respectively. The corresponding intensities, obtained by summing the squares of the coefficients, are shown in Fig. 3. Good agreement is seen for small q, the region of the profile responsible for the size and shape of the scattering object. The less good agreement for $q > 0.2$ Å^{-1} is as expected and is a consequence of the lack of internal structure in a continuum model. In additional calculations the surface was radially expanded and contracted by varying d and the corresponding profiles were calculated. The RMS deviation of the calculated spectrum from the explicit-atom model [$\chi(d)$] is shown in Fig. 3. The best agreement between the continuum and explicit-atom model is obtained for d close to (slightly less than) zero, which corresponds to the calculated surface. The radius of gyration as a function of d is

Table 1
Comparison between simulation-derived and experimental radii of gyration for different types of scattering.

The quality of the agreement is given by the χ function defined as $\chi^2 = [1/(N-1)]\sum_i^N \{[I^{\rm calc}(q_i) - I^{\rm exp}(q_i)]/\sigma_i\}^2$, where σ_i denotes the standard deviation of the ith experimental point. χ is given for calculations in which the solvent molecules are included explicitly ($P + S$) and when they are represented as an unperturbed continuum (P). The X-ray results are in significantly better agreement with experiment than the neutron profiles, owing to the improved statistical accuracy of the experimental X-ray profile at high q. Radii of gyration were obtained by fitting the profiles to (27) in the range $qR_g < 1$.

	$R_g^{\rm exp}$ (Å)	$R_g^{\rm calc}$ (Å)	$\chi^2(P)$	$\chi^2(P+S)$
X-ray	15.4 ± 0.2	15.25 ± 0.19	0.902	0.614
Neutron in H_2O	13.8 ± 0.2	13.62 ± 0.24	2.883	2.774
Neutron in D_2O	12.4 ± 0.2	12.45 ± 0.22	2.085	1.916

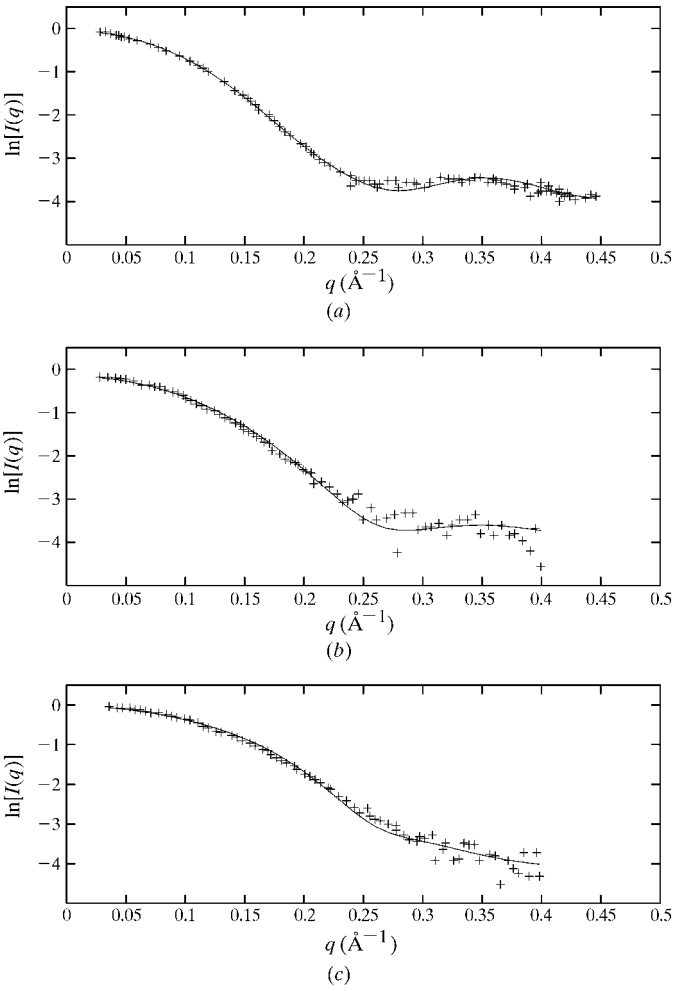

Figure 2
Comparison of calculated (solid lines) with experimental (crosses) SAS profiles of protein in solution for different types of scattering. The experimental data are from Svergun *et al.* (1998). (*a*) X-ray SAS in H_2O. (*b*) Neutron SAS in H_2O. (*c*) Neutron SAS in D_2O.

also shown in Fig. 3 and is also found to coincide with the explicit-atom radius of gyration at $d \simeq 0$. These results concur in indicating that the excluded volume $L = 17$ is a good reference surface for representing the SAS profiles. A three-dimensional plot of the $L = 17$ surface is shown in Fig. 4.

4. Conclusions

The present method efficiently calculates SAS profiles from explicit-atom models of proteins and the surrounding solvent. The use of a multipole expansion allows rapid calculation of the SAS profiles from multiple configurations of systems of large numbers of atoms. Without using a multipole expansion the computational costs scale as N^2, where N is the number of atoms. In the present case, they scale as $N(l_{\max}+1)^2$, where $l_{\max}$ is the multipole expansion truncation value. The speed-up for our system consisting of ∼20 000 atoms is therefore a factor of ∼50 for $l_{\max} = 17$. As such, the method is particularly suitable for the computation of SAS data from computer simulations, such as molecular dynamics or Monte Carlo, where many configurations have to be evaluated. In the present case, the profile was calculated averaged over ∼2000 configurations.

Here, the published experimental lysozyme SAS profiles for X-ray scattering in H_2O and for neutron scattering in H_2O and D_2O were compared with the results of an MD simulation of the same system. Excellent agreement with experiment is seen. This result provides impetus for further work aimed at using the detailed information present in the MD simulation to decompose the contributions to the scattering profile and, in particular, the effect of the protein on the average water structure in the first layer of hydration.

The *SASSIM* program also includes a method for defining the protein surface, again using spherical harmonics. The analytical protein envelope will be of use in many problems in biophysics that require continuum models for the solvent and/or protein. Because of the truncation of multipole expansion in (23) and the finite number of grid points N_P in (25) the surface $S_P(\theta, \varphi)$ is not hermetically closed for protein atoms, meaning it does not adjust optimally to all cavities and ridges on the 'true' protein surface. The proposed definition of the model protein surface also does not represent cavities at directions ω where the 'true' protein surface intersects the corresponding radial line more than once. In this latter case the model surface covers over the pocket. As a consequence, there are a few protein atoms (n'_p) outside the surface and a few water molecules (n'_w) inside. In the case of the model surface expanded over spherical harmonics up to $L = 17$ for lysozyme $n'_p \simeq 3$ and $n'_w \simeq 120$, compared with a total of ∼550 water molecules found in the first 3 Å layer. However, the principle followed here in defining the surface is not to separate all the water from the protein, but rather to try to

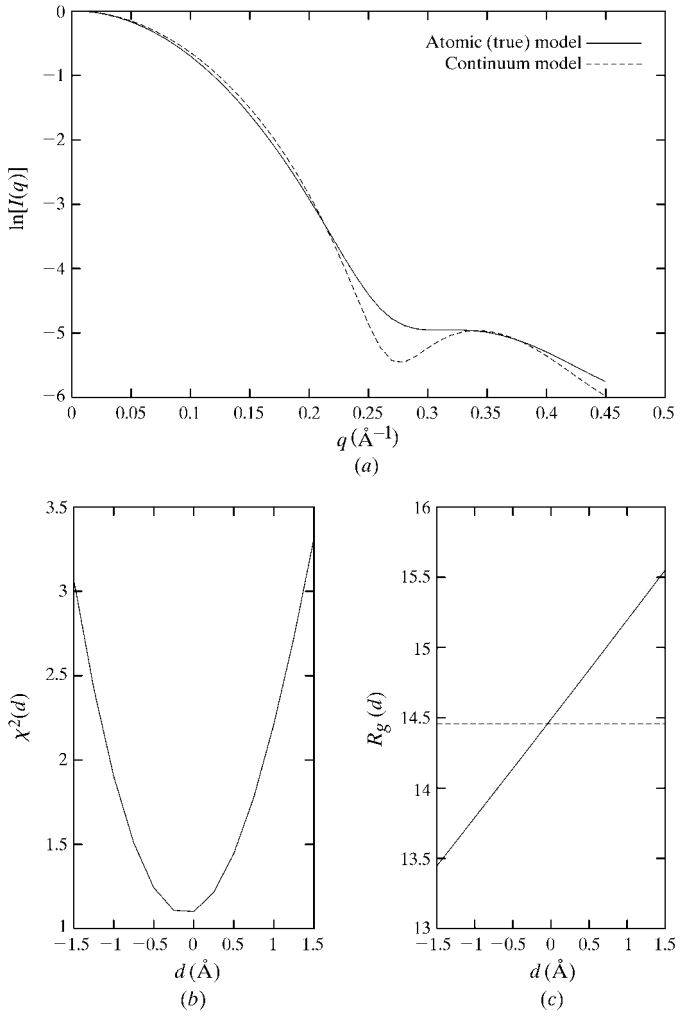

Figure 3
Tests of the protein surface definition. (*a*) Scattering intensity calculated from the full atomic detail and continuum protein models. (*b*) The agreement factor between the true and continuum model as a function of d, $\chi(d)$. (*c*) The dependence of the continnum model radius of gyration on d. The explicit-atom value is the dashed line.

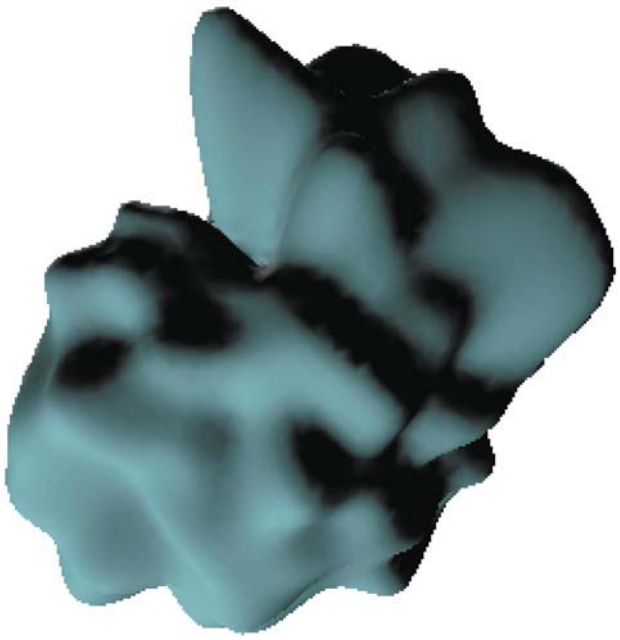

Figure 4
Protein surface expanded over spherical harmonics up to $L_{\max} = 17$.

represent as much as possible of the surface at the smallest possible computational cost. Furthermore, as demonstrated here, the analytical envelope reproduces well the SAXS profile calculated from the explicit protein atoms.

We thank Dr Stefan Fischer for the stimulating discussions. The support to FM from DLR, Germany and the Ministry for Science and Technology of the Republic of Slovenia is gratefully acknowledged. The *SASSIM* program is freely available for use and can be obtained from FM at franc@cmm.ki.si.

References

Abramowitz, M. & Stegun, I. A. (1970). *Handbook of Mathematical Functions.* New York: Dover.
Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). *J. Comput. Chem.* **4**, 187–217.
Brünger, A. T. & Karplus, M. (1988). *Proteins*, **4**, 148–156.
Connolly, M. L. (1985). *J. Am. Chem. Soc.* **107**, 1118–1125.
Fraser, R. D. B., MacRae, T. P. & Suzuki, E. (1978). *J. Appl. Cryst.* **11**, 693–694.
Guinier, A. (1939). *Ann. Phys. (Leipzig)*, **12**, 161–237.
Harpaz, Y., Gerstein, M. & Chothia, C. (1994). *Structure*, **2**, 641–649.
Henderson, S. J. (1996). *Biophys. J.* **70**, 1618–1627.
Higgins, J. S. & Benoit, H. C. (1994). *Polymers and Neutron Scattering.* Oxford: Clarendon Press.
Jorgensen, W. L., Chandrasekhar, J. & Madura, J. D. (1983). *J. Chem. Phys.* **79**, 926–935.
Lebedev, V. I. (1975). *Zh. Vychisl. Mat. Mat. Fiz.* **15**, 48–54.
Lee, B. & Richards, F. M. (1971). *J. Mol. Biol.* **55**, 379–400.
Pedersen, J. S., Posselt, D. & Mortensen, K. (1990). *J. Appl. Cryst.* **23**, 321–333.
Pickover, C. A. & Engelman, D. M. (1982). *Biopolymers*, **21**, 817–831.
Stuhrmann, H. B. (1970*a*). *Acta Cryst.* A**26**, 297–306.
Stuhrmann, H. B. (1970*b*). *Z. Physik. Chem. Neue Folge*, **72**, 177–198.
Svergun, D. (1999). *Biophys. J.* **76**, 2879–2886.
Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.
Svergun, D., Richard, S., Koch, M. H., Sayers, Z., Kuprin, S. & Zaccai, G. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 2267–2272.
Treutler, O. & Ahlrichs, R. (1994). *J. Chem. Phys.* **102**, 346–354.
Vaney, M.C., Maignan, S., Riès-Kautt, M. & Ducruix, A. (1996). *Acta Cryst.* D**52**, 505–517.
Voronoi, G. F. (1908). *Z. Reine Angew. Math.* **134**, 198–287.
Zhang, R., Tristram-Nagle, S., Sun, W., Headrick, R. L., Irving, T. C., Suter, R. M. & Nagle, J. F. (1996). *Biophys. J.* **70**, 349–357.