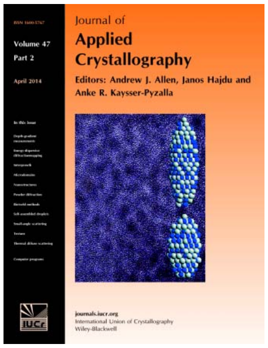
Journal of
Applied
Crystallography
ISSN 1600-5767

Hierarchical $O(N)$ computation of small-angle scattering profiles and their associated derivatives

Konstantin Berlin, Nail A. Gumerov, David Fushman and Ramani Duraiswami

J. Appl. Cryst. (2014). **47**, 755–761



Journal of
Applied Crystallography
ISSN 1600-5767

Received 9 October 2013
Accepted 28 February 2014

Hierarchical $O(N)$ computation of small-angle scattering profiles and their associated derivatives

Konstantin Berlin,[a,b]‡ Nail A. Gumerov,[b,c]‡ David Fushman[a,b] and Ramani Duraiswami[b,c]*

[a]Department of Chemistry and Biochemistry, Center for Biomolecular Structure and Organization, University of Maryland, College Park, MD, USA, [b]Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA, and [c]Fantalgo LLC, Elkridge, MD, USA. Correspondence e-mail: ramani@umiacs.umd.edu

The need for fast approximate algorithms for Debye summation arises in computations performed in crystallography, small/wide-angle X-ray scattering and small-angle neutron scattering. When integrated into structure refinement protocols these algorithms can provide significant speed up over direct all-atom-to-all-atom computation. However, these protocols often employ an iterative gradient-based optimization procedure, which then requires derivatives of the profile with respect to atomic coordinates. This article presents an accurate, $O(N)$ cost algorithm for the computation of scattering profile derivatives. The results reported here show orders of magnitude improvement in computational efficiency, while maintaining the prescribed accuracy. This opens the possibility to efficiently integrate small-angle scattering data into the structure determination and refinement of macromolecular systems.

1. Introduction

Accurate characterization of biomolecular structures in solution is required for understanding their biological function and for therapeutic applications. Small-angle scattering (SAS) of X-rays and neutrons can indirectly measure the distribution of interatomic distances of a molecule in solution, providing a set of structural restraints (Koch *et al.*, 2003). As a result, solution SAS studies have become increasingly popular in structural biology, with a broad range of applications including structure refinement of biological macromolecules and their complexes (Grishaev, Tugarinov *et al.*, 2008; Grishaev, Ying *et al.*, 2008; Grishaev *et al.*, 2005; Pons *et al.*, 2010), analysis of conformational ensembles and flexibility in solution (Bernadó *et al.*, 2007, 2010; Datta *et al.*, 2009; Jehle *et al.*, 2011), and high-throughput structural studies (Hura *et al.*, 2009; Grant *et al.*, 2011).

In order to use SAS experimental data as an atomic level structural restraint, the SAS profile needs to be predicted *ab initio* from the assumed molecular atomic structure, which requires computing all-pairs interactions of the atoms in the molecule (also referred to as an N-body problem). In addition, this computation must be performed numerous times in an iterative structure refinement algorithm (Grishaev, Tugarinov *et al.*, 2008; Grishaev *et al.*, 2010) or for a high-throughput structural analysis.

Several approximation methods have been proposed to speed up this computation (Svergun *et al.*, 1995; Bardhan *et al.*, 2009; Grishaev *et al.*, 2010; Poitevin *et al.*, 2011). However, depending on the size of the molecule, the approximations can introduce significant errors (Gumerov *et al.*, 2012). Recently we developed and demonstrated a hierarchical harmonic expansion method, based on ideas from the fast multipole method, which has superior asymptotical performance to previously proposed approximation methods, while maintaining any prescribed accuracy (Gumerov *et al.*, 2012). This large speedup in computation has the potential to significantly improve integration of SAS into structure refinement protocols, like *Xplor-NIH* (Schwieters *et al.*, 2003) and *HADDOCK* (Dominguez *et al.*, 2003), where the scattering profile needs to be computed thousands of times during iterative structure refinement.

A potential problem with integrating SAS data into a structure refinement protocol is the need to accurately and quickly compute the gradient ('force') of the SAS profile, in addition to the actual SAS profile (Brünger *et al.*, 1998). In order to save on computation time it was suggested by Gabel *et al.* (2006) to approximate the low-frequency part of the SAS profile using Taylor expansion. The drawback of this approach is that significant structural information contained in the high-frequency region of the scattering profile is ignored. Furthermore, the low-frequency profile region becomes smaller as the molecule's size increases, since this approximation depends on the magnitude of the product of the wavenumber and molecule's size. An alternative approach was suggested by Grishaev, Ying *et al.* (2008), where the gradient was approximated by taking the derivative of the scattering amplitude explicitly and then averaging over a large number of orientations. However, there are no accuracy guarantees for

‡ These authors contributed equally to the manuscript.

doi:10.1107/S1600576714004671

such an approximation, which might result in an incorrect gradient direction.

In addition to the computational complexity of the derivative computation, macromolecules can be combined with a solvation layer, to improve the accuracy of the SAS profile prediction (Grishaev *et al.*, 2010). These additional water molecules, especially the large number of atoms that result from discretizing a continuous water model, can significantly increase the computation time needed for estimation of the SAS profile and its derivatives (Stumpe *et al.*, 2010; Poitevin *et al.*, 2011; Virtanen *et al.*, 2011), which might make profile and/or derivative calculation computationally intractable, even for a properly error-bounded approximation approach.

Currently, we are not aware of any previously published method for the computation of arbitrarily accurate SAS profile derivatives, other than direct differentiation followed by all-to-all summation of the scattering equation, or any method that can tractably compute them for very large and/or dense objects. While initially it would seem that accurate computation of derivatives is not critical for structural refinement protocols, in practice, inaccurate derivatives have significant performance implications for any optimization algorithm, since using the derivatives to compute an inaccurate descent direction can significantly increase the number of iterations, and by extension SAS profile computations, required for an optimization algorithm to converge to a local minimum.

Here we propose a method for computing SAS profile derivatives based on analytical differentiation of the spherical harmonic expansion of the scattering equation, which can be quickly and accurately computed using a hierarchical approach developed previously (Gumerov *et al.*, 2012). This differentiation method completely avoids computationally intractable all-to-all-atom summations, and as we demonstrate, can compute arbitrarily accurate derivatives, while only slightly adding to the overall computation time required to calculate the scattering profile without derivatives (Gumerov *et al.*, 2012). Furthermore, leveraging the computational advantages of the hierarchical approach, we demonstrate that the profile and/or derivatives for a solvated macromolecule (using explicit or dense grid water models) can be obtained many orders of magnitude faster than by alternative methods.

2. Method

The SAS profile value of an N-atom system such as a macromolecule and associated solvent molecules can be computed as

$$I(\mathbf{r}_i, \ldots, \mathbf{r}_N; q) = \left\langle \left| A(\mathbf{q}) \right|^2 \right\rangle_\Omega, \tag{1}$$

for $q = q_1 \ldots q_Q$, where q_Q is also the largest value of q,

$$A(\mathbf{q}) = \sum_{j=1}^{N} f_j(q) \exp\left(i\mathbf{q} \cdot \mathbf{r}_j\right) \tag{2}$$

is the scattering amplitude from the particle *in vacuo*, $q = |\mathbf{q}|$ is the scattering vector magnitude [$q = (4\pi/\lambda)\sin\vartheta$, 2ϑ is the scattering angle and λ the wavelength], $\mathbf{r}_j = [x_j, y_j, z_j]$ is the position of the jth atom, $f_j(q)$ is the form factor of the jth atom and $\langle\cdot\rangle_\Omega$ indicates averaging of $\mathbf{q}$ over a sphere, to reflect averaging over all possible molecular orientations in solution or powder.

Equation (1) can be analytically averaged, such that

$$\left\langle \left| A(\mathbf{q}) \right|^2 \right\rangle_\Omega = \sum_{i=1}^{N} f_i(q) \sum_{j=1}^{N} f_j(q) s(qr_{ij}) = \sum_{i=1}^{N} f_i(q) \Psi_i, \tag{3}$$

where s is the sinc function,

$$s(qr_{ij}) = \frac{\sin(q\|\mathbf{r}_i - \mathbf{r}_j\|_2)}{q\|\mathbf{r}_i - \mathbf{r}_j\|_2}, \tag{4}$$

Ψ_i is defined as in equation (3) and $\|\cdot\|_2$ is the Euclidean norm (Debye, 1915; Koch *et al.*, 2003).

In order to refine the structure of a macromolecule, which is uniquely defined by its atomic positions $[\mathbf{r}_i, \ldots, \mathbf{r}_N]$, against experimentally collected scattering data, the discrepancy between the experimental scattering, $I^{\exp}(q)$, and the predicted scattering, $I(\mathbf{r}_i, \ldots, \mathbf{r}_N; q)$, for $q = q_1, \ldots, q_Q$, is minimized. To resolve the structural degeneracy of SAS, additional energy terms coming from NMR, X-ray crystallography, other structural methods and force-field restraints may also be used. This refinement is typically formulated as an energy minimization problem,

$$\min_{\mathbf{r}_i, \ldots, \mathbf{r}_N} \chi^2(\mathbf{r}_i, \ldots, \mathbf{r}_N) + \Lambda(\mathbf{r}_i, \ldots, \mathbf{r}_N), \tag{5}$$

where

$$\begin{aligned}\chi^2(\mathbf{r}_i, \ldots, \mathbf{r}_N) &= \sum_{k=1}^{Q} \|I(\mathbf{r}_i, \ldots, \mathbf{r}_N; q_k) - I^{\exp}(q_k)\|_2^2 \\ &= \|\mathbf{d}\|_2^2 = \sum_{k=1}^{Q} \mathrm{d}_k^2,\end{aligned} \tag{6}$$

is the discrepancy between the predicted and the experimental scattering profiles and $\Lambda(\mathbf{r}_i, \ldots, \mathbf{r}_N)$ is the contribution from all the other structural restraints, as well as potential regularization terms. In what follows we abbreviate $I(\mathbf{r}_i, \ldots, \mathbf{r}_N; q_k)$ as $I(q_k)$.

The computational complexity of evaluating the profile value $I(q)$ in equation (1) is directly related to the computational complexity of the double summation in equation (3) and so has $O(N^2)$ computational complexity if the double summation is evaluated directly. The widely used computer program *CRYSOL* (Svergun *et al.*, 1995) uses a constant size truncated harmonic expansion to speed up these computations, suggesting a computational complexity $O(N)$ for evaluation of $I(q)$. It has been shown (Gumerov *et al.*, 2012) that without an associated error bound the harmonic expansion methods are likely to give incorrect results, especially at large q and/or large molecule sizes. As an alternative, an arbitrarily accurate method based on the hierarchical decomposition of the spatial domain, local spherical expansion and translation was presented. We will use that method as a basis for our proposed differentiation algorithm, which can, for a slight increase in computational cost relative to the original algorithm, not only fully evaluate all the profile values $I(q)$ but also compute the derivatives of the profile with respect to the Cartesian atomic coordinates.

2.1. Spherical expansion

To efficiently compute equation (3), we first expand equation (4) using a spherical harmonic expansion (Gumerov *et al.*, 2012), such that

$$s(qr_{ij}) = 4\pi \sum_{n=0}^{\infty} \sum_{m=-n}^{n} R_n^{-m}(\mathbf{r}_i) R_n^m(\mathbf{r}_j), \tag{7}$$

for all i and j, where R_n^m is a regular spherical basis function:

$$R_n^m(\mathbf{r}) = j_n(qr) Y_n^m(\mathbf{s}) = j_n(qr) Y_n^m(\theta, \varphi), \quad \mathbf{r} = r\mathbf{s}. \tag{8}$$

The functions $R_n^m(\mathbf{r})$ here are given in spherical coordinates (r, θ, φ), $\mathbf{r} = r(\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)$, $j_n(qr)$ are spherical Bessel functions, while $Y_n^m(\theta, \varphi)$ are the normalized spherical harmonics of degree n and order m,

$$Y_n^m(\theta, \varphi) = (-1)^m \left[\frac{2n+1}{4\pi} \frac{(n-|m|)!}{(n+|m|)!} \right]^{1/2} P_n^{|m|}(\cos\theta) \exp(im\varphi),$$
$$n = 0, 1, 2, \ldots, \qquad m = -n, \ldots, n. \tag{9}$$

$P_n^{|m|}(\mu)$ are the associated Legendre functions consistent with those given by Abramowitz & Stegun (1970) or the Rodrigues formulae,

$$P_n^m(\mu) = (-1)^m \left(1 - \mu^2\right)^{m/2} \frac{\mathrm{d}^m}{\mathrm{d}\mu^m} P_n(\mu), \quad n \geq 0, \quad m \geq 0,$$
$$P_n(\mu) = \frac{1}{2^n n!} \frac{\mathrm{d}^n}{\mathrm{d}\mu^n} \left(\mu^2 - 1\right)^n, \quad n \geq 0, \tag{10}$$

where $P_n(\mu)$ are the Legendre polynomials.

Truncating the infinite spherical harmonic expansion in equation (7) to finite length p and aggregating N^2 such expansions into the $N\Psi_i$ functions yields a spherical harmonic approximation of Ψ_i in terms of the spherical basis functions $R_n^m(\mathbf{r})$:

$$\Psi_i = \sum_{n=0}^{p-1} \sum_{m=-n}^{m=n} R_n^m(\mathbf{r}_i) B_n^m(\mathbf{r}_j) + \varepsilon_p, \tag{11}$$

where

$$B_n^m(\mathbf{r}_j) = 4\pi \sum_{j=1}^{N} f_j(q) R_n^{-m}(\mathbf{r}_j), \tag{12}$$

and p and ε_p are the cutoff degree and associated error, respectively. The appropriate value for cutoff degree p that will guarantee a relative error smaller than ε_p has been derived by Gumerov *et al.* (2012).

Even though the direct computation of Ψ_i, for all $i = 1, \ldots, N$, would take $O(p^2N^2)$ time, we can speed up the computation to $O(p^2N)$ by changing the order of summation, such that we first precompute $B_n^m(\mathbf{r}_j)$ in $O(N)$ time (since it is the same for all Ψ_i) and then evaluate $N\Psi_i$ values, each in $O(p^2)$ time. Thus computing all the values of Ψ_i, as well as their combined weighted sum $I(q)$ [equation (3)], takes $O(p^2N)$ total time. We will refer to this method for computing all Ψ_i and $I(q)$ as the 'middleman' approach.

Recently, we have demonstrated that it is actually possible to compute these expansions significantly faster, using our hierarchical algorithm, in $O[p^3\log(p) + N\log(N)]$ time (Gumerov *et al.*, 2012). This allows fast computation of these expansions, as well as the actual SAS profile, for large molecular systems.

In the derivation of $I(q)$ derivatives below, we assume that we have already computed the spherical harmonic expansions of Ψ_i, using the middleman, hierarchical or some other method.

2.2. Computation of derivatives

We now derive an ε-accurate algorithm for computing the gradient of $I(q)$, sometimes referred to as the 'force' of $I(q)$, which is required for a local convex optimization of equation (5) using a Newton-type minimization, such as done in *Xplor-NIH* (as well as other programs). Computation of the Hessian is typically avoided, either by an iterative approximation of the Hessian using a quasi-Newton method or, in the case of least squares, by an approximation based on the Jacobian (Boyd & Vandenberghe, 2004).

Assuming the derivatives of Λ are already computed, based on the linearity of differentiation, in order to compute the gradient of equation (5) we only need to compute the Jacobian of $I(q)$,

$$\mathbf{J} = \begin{pmatrix} \frac{\partial I_1}{\partial x_1} & \frac{\partial I_1}{\partial y_1} & \frac{\partial I_1}{\partial z_1} & \cdots & \frac{\partial I_1}{\partial x_N} & \frac{\partial I_1}{\partial y_N} & \frac{\partial I_1}{\partial z_N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial I_Q}{\partial x_1} & \frac{\partial I_Q}{\partial y_1} & \frac{\partial I_Q}{\partial z_1} & \cdots & \frac{\partial I_Q}{\partial x_N} & \frac{\partial I_Q}{\partial y_N} & \frac{\partial I_Q}{\partial z_N} \end{pmatrix}, \tag{13}$$

where $\mathbf{J}$ is a $Q \times 3N$ matrix. From the Jacobian, the gradient can be computed as

$$\nabla\chi^2 = 2\mathbf{J}^{\mathrm{T}}\mathbf{d}, \tag{14}$$

and the Hessian can be approximated as

$$\mathbf{H} \simeq 2\mathbf{J}^{\mathrm{T}}\mathbf{J} + 2\lambda\mathbf{I}, \tag{15}$$

where $\mathbf{I}$ is the identity matrix, for some $\lambda \in [0, 1]$, in a nonlinear least-squares minimization. In the case of an iterative quasi-Newton minimization of equation (5), only equation (14) is needed, while the Hessian is iteratively updated, and thus the full Jacobian $\mathbf{J}$ does not need to be explicitly formed.

Without loss of generality for $\partial I_k/\partial y_i$ and $\partial I_k/\partial z_i$,

$$\frac{\partial I_k}{\partial x_i} = 2f_i(q_k)\Psi_i' = 2f_i(q_k) \sum_{j=1}^{N} f_j(q_k) s'(q_k r_{ij}), \tag{16}$$

where $s'(q_k r_{ij})$ and Ψ_i' are the partial derivatives with respect to x_i. The derivative $s'(q_k r_{ij})$ can be analytically computed as

$$s'(q_k r_{ij}) = (x_i - x_j) \left[\frac{\cos(q_k r_{ij})}{r_{ij}^2} - \frac{\sin(q_k r_{ij})}{q_k r_{ij}^3} \right], \tag{17}$$

for $r_{ij} \neq 0$. Therefore, direct computation of $\mathbf{J}$ (and $\nabla\chi^2$) has complexity of $O(N^2Q)$.

Alternatively, once the spherical harmonic expansion of Ψ_i is computed, taking the derivative of equation (11) yields

$$\Psi_i' = \sum_{n=0}^{p-1}\sum_{m=-n}^{m=n} R_n'^m(\mathbf{r}_i)B_n^m(\mathbf{r}_j) + \hat{\varepsilon}_p, \qquad (18)$$

where $\hat{\varepsilon}_p$ is the error in the derivative associated with cutoff degree p. The derivatives of the spherical function, $R_n'^m$, have been derived by Gumerov & Duraiswami (2004):

$$\frac{1}{q}\frac{\partial R_n^m(\mathbf{r})}{\partial x} = \frac{1}{2}\Big[b_{n+1}^{-m-1}R_{n+1}^{m+1}(\mathbf{r}) - b_n^m R_{n-1}^{m+1}(\mathbf{r}) + b_{n+1}^{m-1}R_{n+1}^{m-1}(\mathbf{r}) - b_n^{-m}R_{n-1}^{m-1}(\mathbf{r})\Big], \qquad (19)$$

$$\frac{1}{q}\frac{\partial R_n^m(\mathbf{r})}{\partial y} = \frac{-i}{2}\Big[b_{n+1}^{-m-1}R_{n+1}^{m+1}(\mathbf{r}) - b_n^m R_{n-1}^{m+1}(\mathbf{r}) - b_{n+1}^{m-1}R_{n+1}^{m-1}(\mathbf{r}) + b_n^{-m}R_{n-1}^{m-1}(\mathbf{r})\Big], \qquad (20)$$

$$\frac{1}{q}\frac{\partial R_n^m(\mathbf{r})}{\partial z} = a_{n-1}^m R_{n-1}^m(\mathbf{r}) - a_n^m R_{n+1}^m(\mathbf{r}), \qquad (21)$$

where

$$b_n^m = \begin{cases} \left[\dfrac{(n-m-1)(n-m)}{(2n-1)(2n+1)}\right]^{1/2} & \text{for } 0 \le m \le n, \\ -\left[\dfrac{(n-m-1)(n-m)}{(2n-1)(2n+1)}\right]^{1/2} & \text{for } -n \le m < 0, \\ 0 & \text{for } n < |m|, \end{cases} \qquad (22)$$

$$a_n^m = \begin{cases} \left[\dfrac{(n+1+m)(n+1-m)}{(2n+1)(2n+3)}\right]^{1/2} & \text{for } n \ge |m|, \\ 0 & \text{for } n < |m|, \end{cases} \qquad (23)$$

and $i^2 = -1$.

This derivative can be written as a weighted linear combination of two to four adjacent spherical basis functions (depending on the particular spatial derivative needed), where the weights are independent of $\mathbf{r}_i$. Thus, computation of the derivative can be expressed as a sparse matrix product of the transformation matrix multiplied by the spherical expansion coefficients of Ψ_i. Therefore,

$$\Psi_i' = \sum_{n=0}^{p-1}\sum_{m=-n}^{m=n} R_n^m(\mathbf{r}_i)\,\alpha_{(j)n}^m + \hat{\varepsilon}_p, \qquad (24)$$

where $\alpha_{(j)n}^m$ can be quickly computed by an $O(p^2)$ matrix multiplication and is the same for all Ψ_i'. As noted by Gumerov & Duraiswami (2004), the error bound for derivative computations requires slightly higher truncation numbers, which are one or two higher than those for the expansion of Ψ.

Thus, the overall complexity of computing $\mathbf{J}$, given that the harmonic expansion has already been computed for all Ψ_i for all q, is $O(p^2NQ)$. The appropriate value of p for a specific value of q is proportional to qD, where D is the diameter of the molecule (largest distance between any two atoms) (Gumerov *et al.*, 2012). Therefore, the overall complexity of the gradient computation is $O(q_Q^2D^2NQ)$, for an arbitrary error bound. In practice, this is typically faster than any of the methods for the computation of the spherical harmonic expansion of equation (3) and hence does not significantly change the runtime of the algorithm.

3. Results

Above we showed that, theoretically, the hierarchical algorithm has asymptotically better computational complexity than previous approaches. However, this advantage depends on the scattering vector magnitude q and the diameter and atom density of the molecule, since that dictates if the number of terms in the harmonic expansions, p^2, is small enough compared to N. Therefore, we directly compare the runtime for computing a theoretical scattering profile and its Jacobian using our algorithm ('hierarchical') with that for the direct all-to-all computation of equations (3) and (16) ('direct') as well as a spherical harmonic expansion method like *CRYSOL* (Svergun *et al.*, 1995) with a proper spherical harmonic truncation value ('middleman') (Gumerov *et al.*, 2012), for various-sized random and actual macromolecules solvated using different water grid densities. This directly demonstrates the computational advantage of our hierarchical approach for the various biomolecular small-angle scattering problems that are currently studied.

3.1. Derivatives

We ran the benchmarks on a 2.80 GHz 64 bit Linux machine with 24 GB ECC DDR3 SDRAM and a Dual Quad-

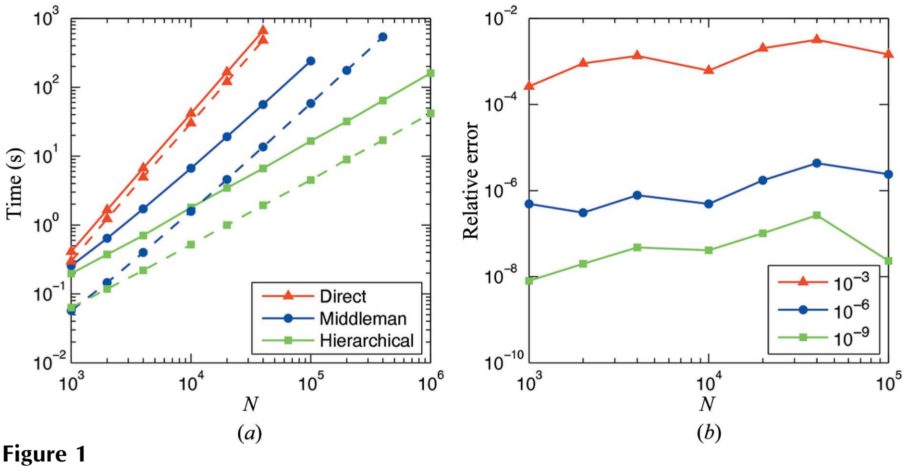


Figure 1
Timing results for randomly generated proteins, using uniformly distributed 50-point scattering profiles, $0 < q \le 0.5$ Å^{-1}. (*a*) Computation time for just the SAS profile (dashed) and the SAS profile with Jacobian (solid). (*b*) Maximum relative error in the Jacobian computed using the hierarchical approach, as measured by $\max_i \|\mathbf{J}_{i,*} - \hat{\mathbf{J}}_{i,*}\|_2/\|\hat{\mathbf{J}}_{i,*}\|_2$, where $\hat{\mathbf{J}}$ is the true Jacobian and $\mathbf{J}_{i,*}$ is the ith row of $\mathbf{J}$. For smaller molecules, $\hat{\mathbf{J}}$ was taken to be the Jacobian computed using the direct method, while for very large molecules, $\hat{\mathbf{J}}$ was estimated using the middleman approach with $\varepsilon_p = 10^{-12}$ cutoff.

Core Intel Xeon X5560 CPU. All programs were compiled using the Intel 11.1 compiler using the '-parallel -O3 -ipo' flags. We measured the computational time for generating a SAS profile composed of 50 uniformly spaced evaluations of the profile in the range $0.01 \le q \le 0.5$ Å^{-1}, for a randomly generated molecule with an atomic density of $d = 0.02$ Å^{-3}. The timing and accuracy results for the randomly generated molecules are presented in Fig. 1. The accuracy of the Jacobian computation is assessed by comparing the Jacobian of the truncated harmonic expansion series, computed by our hierarchical algorithm, with the true Jacobian, computed directly using equation (16). In the case when the molecule is too large for direct N^2 computation of the true Jacobian, as a proxy, the Jacobian is computed from the harmonic expansion using the middleman approach and a very high truncation number ($\varepsilon_p = 10^{-12}$).

Fig. 1(*a*) shows that computation of the derivatives does not change the asymptotic properties of any of the algorithms. For the hierarchical method, the computation of the Jacobian increases the computation time by about 3.5-fold. This increase is partly because of the addition of a downward pass in our hierarchical method. The hierarchical method, with or without Jacobian computation, is faster than the *CRYSOL*-like middleman algorithm, as well as direct computation. Fig. 1(*b*) shows that the accuracy of the Jacobian computation is within an order of magnitude of the prescribed accuracy for $I(q)$.

3.2. Water layer

We have demonstrated that computing the scattering profile and its Jacobian is significantly faster using our hierarchical algorithm than the two known algorithms (direct and middleman). In addition, one of the principal advantages of our hierarchical method is that we decouple the number of atoms from the molecule's diameter during the computation. This gives us an even larger computational advantage over previous approaches when computing SAS profiles for solvated molecules, since water models tend to significantly increase the number of atoms in the computation, without significantly increasing D.

The computational advantage of the hierarchical algorithm depends on the specific water model used, and various water models have been suggested for SAS prediction (Stumpe *et al.*, 2010; Poitevin *et al.*, 2011; Virtanen *et al.*, 2011). Assessing the accuracy of these models against actual experimental data is outside the scope of this article. However, since only the atomic density and diameter of the macromolecular system impact the runtime of the three tested algorithms, while the specific positions of the water atoms (or pseudo-atoms) do not, for simplicity we simulate the effect of various water models on the runtime by generating a uniform water grid layer at various atomic densities. The grid size (the distance between adjacent pseudo-water atoms) ranges from infinity (no water atoms) to 0.5 Å (a suggested value for various discretized continuous water models). A grid size of 3.1 Å is approximately equivalent to an explicit water model (assuming one pseudo-atom representing each water molecule).

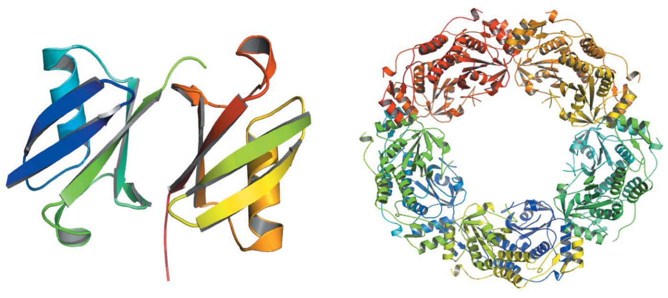
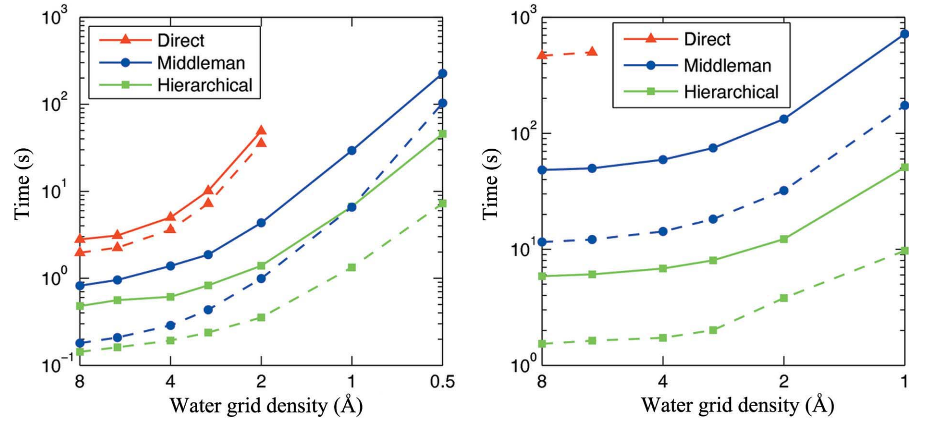

(*a*) (*b*)

Figure 2
Cartoon representations of (*a*) 1aar and (*b*) 2e2g.

We chose two proteins, archaeal peroxiredoxin [Protein Data Bank (PDB) code 2e2g; Nakamura *et al.*, 2008] and diubiquitin (PDB code 1aar; Cook *et al.*, 1992), that have very different diameters and interatomic distance distributions, to illustrate the speedup the algorithm provides. Archaeal peroxiredoxin has a large empty cavity in the center (see Fig. 2). The number of atoms for 1aar increases from 2575 atoms with no water layer to several orders of magnitude larger (548 462 atoms) for a 0.5 Å water layer grid. For 2e2g the number of atoms increases from 39 609 to 543 832 atoms for a slightly coarser 1.0 Å water layer grid. The speedup of our algorithm compared with previously published algorithms is shown in Fig. 3.

Fig. 3 demonstrates that, not only is our hierarchical approach significantly faster, the speedup gets progressively larger for higher water grid density.



(*a*) (*b*)

Figure 3
Timing results for computation of the profile (computation of just the SAS profile is shown as a dashed line, and computation of the SAS profile with the Jacobian is shown as a solid line), with an 8 Å water layer, at different water grid densities (interatomic distances). (*a*) 1aar and (*b*) 2e2g.

4. Discussion

The computational advantage of the hierarchical algorithm, especially for molecules with a large diameter and/or solvated using various water models, can be understood in terms of how it relates to the three main alternatives for computing $I(q)$: (i) direct all-to-all summation of equation (3); (ii) numerical integration or averaging of equation (1), as suggested by Grishaev, Tugarinov *et al.* (2008), Bardhan *et al.* (2009) and Poitevin *et al.* (2011); and (iii) analytical expansion of equation (2) (Svergun *et al.*, 1995; Liu *et al.*, 2012). As discussed by Gumerov *et al.* (2012), approach (i) has $O(N^2)$ complexity, making it intractable for larger macromolecules, while approaches (ii) and (iii) are bounded by the bandwidth of equation (2), requiring either $O(q^2D^2)$ function evaluation over a sphere or basis functions for any type of series expansion in order to accurately approximate the scattering profile, resulting in $O(q^2D^2N)$ computational complexity. Additionally, a spherical harmonic series expansion of $A(\mathbf{q})$ is not sparse (meaning most, if not all, coefficients of the harmonic expansion are nonzero), so it is not possible to accurately approximate equation (3) by using a non-uniform subsampling of the sphere.

When considering large molecules and/or dense water grid models, where D and N are large, the above approaches become computationally intractable. However, a large speedup is still possible if we can compute an approximate series expansion of equation (2) faster than in $O(q^2D^2N)$ time. That is the main idea behind our hierarchical method, where a small expansion is performed locally, where D is small, and then translated hierarchically onto progressively larger domains. This grouping of adjacent atoms together is somewhat similar to the intuitive idea of grouping of atoms by their secondary structure as suggested by Liu *et al.* (2012). However, because grouping is done at multiple hierarchical levels and these groups are combined together using proven theoretical bounds, this paradigm allows much faster computation, can be made arbitrarily accurate and can better accommodate various macromolecular shapes (including cavities).

As a result of the hierarchical grouping of atoms, our algorithm has $O[(qD)^3\log(qD) + N\log(N)]$ complexity for each $I(q)$ evaluation, which is a significant theoretical improvement for large molecules and various water models, because it separates the size of the domain, D, from the number of atoms in the macromolecule, N. In practice, this allows us to compute $I(q)$ accurately for extremely large systems made of millions of atoms or even when the water grid density is high enough to accurately approximate a continuous water model. As a by-product of this computation, we also compute the $O(q^2D^2)$-sized series expansion of equation (3). Once this expansion is computed, the derivatives of the minimization function can be quickly computed by taking the derivative of each basis function.

Even though the hierarchical method is faster than basic spherical harmonic expansion of $I(q)$, it is not required for computing the Jacobian, since only the final series expansion is needed to compute the derivative. This means that programs like *CRYSOL*, which are based on direct series expansion of equation (2), can be easily adapted to compute derivatives using equation (24), as long as the series expansion is properly truncated according to an appropriate error bound.

The hierarchical approach is a general approach for computation of Debye summations and places no restriction on how atoms are distributed in the coordinate space or how those placements are derived. Given the atomic coordinates from any water model, the form factors and the desired precision, the hierarchical approach can efficiently compute the scattering profile and its derivative without any manual tuning of parameters (like size of the expansion series or number of points sampled on a sphere). The approach is also not limited to any specific spatial subdivision of macromolecules and can be adapted to use various other groupings, including secondary structures, or other rigid components, further speeding up the computation.

For certain applications the derivatives for the water molecules might not be required, or only the cross scattering between several molecular domains is needed. Our hierarchical approach allows us to take advantage of these cases by computing a downward pass only on the required atomic domain, if necessary. We note that this algorithm is a proof of principle, and thus the performance can be significantly improved, either by improving the translation operators or by removing complex number multiplications in several instances.

5. Conclusion

We have developed and demonstrated a fast method for computation of the SAS profile and its associated derivatives with respect to the atomic coordinates of a molecule. The method is based on our previously developed hierarchical expansion method and is able to compute the SAS profile and associated derivatives to within arbitrary ε accuracy, while being an order of magnitude faster than existing methods. The computational advantage of the algorithm is even larger for solvated molecules. This opens up the possibility to efficiently integrate small-angle scattering data into the existing protocols for structure determination, molecular dynamics and refinement of macromolecular systems.

This study has been partially supported by the New Research Frontiers Award of the Institute of Advanced Computer Studies of the University of Maryland, by Fantalgo LLC and by NIH grant GM065334 to DF.

References

Abramowitz, M. & Stegun, I. A. (1970). *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables.* National Bureau of Standards Applied Mathematics Series, No. 55. Washington, DC: US Government Printing Office.

Bardhan, J., Park, S. & Makowski, L. (2009). *J. Appl. Cryst.* **42**, 932–943.

Bernadó, P., Modig, K., Grela, P., Svergun, D. I., Tchorzewski, M., Pons, M. & Akke, M. (2010). *Biophys. J.* **98**, 2374–2382.

Bernadó, P., Mylonas, E., Petoukhov, M. V., Blackledge, M. & Svergun, D. I. (2007). *J. Am. Chem. Soc.* **129**, 5656–5664.

Boyd, S. & Vandenberghe, L. (2004). *Convex Optimization.* Cambridge University Press.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Cook, W. J., Jeffrey, L. C., Carson, M., Chen, Z. & Pickart, C. M. (1992). *J. Biol. Chem.* **267**, 16467–16471.

Datta, A., Hura, G. & Wolberger, C. (2009). *J. Mol. Biol.* **392**, 1117–1124.

Debye, P. (1915). *Ann. Phys.* **351**, 809–823.

Dominguez, C., Boelens, R. & Bonvin, A. (2003). *J. Am. Chem. Soc.* **125**, 1731–1737.

Gabel, F., Simon, B. & Sattler, M. (2006). *Eur. Biophys. J.* **35**, 313–327.

Grant, T. D., Luft, J. R., Wolfley, J. R., Tsuruta, H., Martel, A., Montelione, G. T. & Snell, E. H. (2011). *Biopolymers*, **95**, 517–530.

Grishaev, A., Guo, L., Irving, T. & Bax, A. (2010). *J. Am. Chem. Soc.* **132**, 15484–15486.

Grishaev, A., Tugarinov, V., Kay, L. E., Trewhella, J. & Bax, A. (2008). *J. Biomol. NMR*, **40**, 95–106.

Grishaev, A., Wu, J., Trewhella, J. & Bax, A. (2005). *J. Am. Chem. Soc.* **127**, 16621–16628.

Grishaev, A., Ying, J., Canny, M., Pardi, A. & Bax, A. (2008). *J. Biomol. NMR*, **42**, 99–109.

Gumerov, N. A., Berlin, K., Fushman, D. & Duraiswami, R. (2012). *J. Comput. Chem.* **33**, 1981–1996.

Gumerov, N. A. & Duraiswami, R. (2004). *Fast Multipole Methods for the Helmholtz Equation in Three Dimensions.* San Diego: Elsevier Science.

Hura, G. L., Menon, A. L., Hammel, M., Rambo, R. P., Poole, F. L. II, Tsutakawa, S. E., Jenney, F. E. Jr, Classen, S., Frankel, K. A., Hopkins, R. C., Yang, S., Scott, J. W., Dillard, B. D., Adams, M. W. W. & Tainer, J. A. (2009). *Nat. Methods*, **6**, 606–612.

Jehle, S., Vollmar, B. S., Bardiaux, B., Dove, K. K., Rajagopal, P., Gonen, T., Oschkinat, H. & Klevit, R. E. (2011). *Proc. Natl Acad. Sci.* **108**, 6409–6414.

Koch, M. H. J., Vachette, P. & Svergun, D. I. (2003). *Q. Rev. Biophys.* **36**, 147–227.

Liu, H., Morris, R. J., Hexemer, A., Grandison, S. & Zwart, P. H. (2012). *Acta Cryst.* A**68**, 278–285.

Nakamura, T., Yamamoto, T., Abe, M., Matsumura, H., Hagihara, Y., Goto, T., Yamaguchi, T. & Inoue, T. (2008). *Proc. Natl Acad. Sci. USA*, **105**, 6238–6242.

Poitevin, F., Orland, H., Doniach, S., Koehl, P. & Delarue, M. (2011). *Nucleic Acids Res.* **39**(Suppl. 2), W184–W189.

Pons, C., D'Abramo, M., Svergun, D. I., Orozco, M., Bernadó, P. & Fernández-Recio, J. (2010). *J. Mol. Biol.* **403**, 217–230.

Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Marius Clore, G. (2003). *J. Magn. Res.* **160**, 65–73.

Stumpe, M. C., Blinov, N., Wishart, D., Kovalenko, A. & Pande, V. S. (2010). *J. Phys. Chem. B*, **115**, 319–328.

Svergun, D., Barberato, C. & Koch, M. H. J. (1995). *J. Appl. Cryst.* **28**, 768–773.

Virtanen, J. J., Makowski, L., Sosnick, T. R. & Freed, K. F. (2011). *Biophys. J.* **101**, 2061–2069.