Journal of
Applied
Crystallography

ISSN 0021-8898

Received 11 March 2013
Accepted 15 June 2013

Rapid and accurate calculation of small-angle scattering profiles using the golden ratio

Max C. Watson* and Joseph E. Curtis

NIST Center for Neutron Research, National Institute of Standards and Technology, 100 Bureau Drive, Mail Stop 6102, Gaithersburg, MD 20899-6102, USA. Correspondence e-mail: max.watson@nist.gov

Calculating the scattering intensity of an N-atom system is a numerically exhausting $O(N^2)$ task. A simple approximation technique that scales linearly with the number of atoms is presented. Using an exact expression for the scattering intensity $I(\mathbf{q})$ at a given wavevector $\mathbf{q}$, the rotationally averaged intensity $I(q)$ is computed by evaluating $I(\mathbf{q})$ in several scattering directions. The orientations of the $\mathbf{q}$ vectors are taken from a quasi-uniform spherical grid generated by the golden ratio. Using various biomolecules as examples, this technique is compared with an established multipole expansion method. For a given level of speed, the technique is more accurate than the multipole expansion for anisotropically shaped molecules, while comparable in accuracy for globular shapes. The processing time scales sub-linearly in N when the atoms are identical and lie on a lattice. The procedure is easily implemented and should accelerate the analysis of small-angle scattering data.

1. Introduction

Small-angle scattering is an invaluable tool for probing the nanoscale structure of matter. Small-angle X-ray and neutron scattering have become versatile techniques used by structural biologists, physical chemists and materials scientists to study the shapes of molecules and the interactions between them. As opposed to crystallography, small-angle scattering provides the means to analyze the native structure of biomolecules in solution under near physiological conditions. While methods for determining the degree of folding (Kratky & Porod, 1949) and overall size of particles (Guinier, 1939) have long been established, the determination of three-dimensional molecular shape from the one-dimensional scattering profile remains one of the field's greatest challenges.

The first of these attempts involved the construction of low-resolution shapes described by spherical harmonics (Stuhrmann, 1970*b*; Svergun & Stuhrmann, 1991; Svergun *et al.*, 1997, 1996). With the advent of modern computing, molecules can be modeled using arrangements of coarse-grained beads. The scattering intensity of a given bead configuration is explicitly calculated and compared with experimental data (Chacón *et al.*, 1998, 2000; Walther *et al.*, 2000; Franke & Svergun, 2009; Volkov & Svergun, 2003; Svergun, 1999; Svergun *et al.*, 2001). Using this '*ab initio*' bead-modeling approach, the shape of a molecule can (in principle) be determined by calculating the scattering profile for a multitude of candidate structures until a configuration that agrees with the data is found. Combined with high-throughput analysis pipelines, this has allowed for a turnover rate of 20 proteins per week (Hura *et al.*, 2009). When the crystal structures of individual domains are known, the structure of multidomain proteins may be reconstructed by finding the optimal interdomain arrangement which gives the best agreement between theoretical calculations and the measured scattering intensity (Putnam *et al.*, 2007; Wall *et al.*, 2000; Petoukhov & Svergun, 2005, 2006; Förster *et al.*, 2008; Konarev *et al.*, 2001). In addition to molecular shape, simulations of atomically detailed systems can provide insights into the structure of crowded solutions (McGuffee & Elcock, 2006; Chaudhri *et al.*, 2012; Mereghetti *et al.*, 2010; Curtis *et al.*, 2012) and the dynamics of proteins (Lindorff-Larsen *et al.*, 2012; Monticelli *et al.*, 2008). In all of these cases, the computational cost of calculating the scattering intensity from atomic positions imposes a serious bottleneck on the interpretation of experimental measurements.

2. Background

Determining the scattering profile of an N-atom system is analytically straightforward. We use the term 'atom' loosely, since it can also refer to coarse-grained beads. Using the notation of Roe (2000), the scattering intensity $I(\mathbf{q})$ for a particular scattering orientation is given by

$$I(\mathbf{q}) = |A(\mathbf{q})|^2 \tag{1}$$

with amplitude

$$A(\mathbf{q}) = \sum_{j}^{N} b_j \exp(-i\mathbf{q}\cdot\mathbf{r}_j), \tag{2}$$

where b_j is the generalized scattering length, $A(\mathbf{q})$ represents the normalized scattering amplitude, $\mathbf{r}_j$ refers to atomic posi-

doi:10.1107/S002188981301666X

tions, N is the number of atoms and $\mathbf{q}$ is the scattering vector. The magnitude of the scattering vector $q = |\mathbf{q}|$ is given by $q = 4\pi\sin(\theta)/\lambda$, where 2θ is the scattering angle and λ is the wavelength of the incident radiation. Written in this way, equations (1) and (2) can refer to both neutron and X-ray scattering (where b_j implicitly depends on q in the latter case). When the solution is dilute and consists of one species, intermolecular interactions are negligible and the sum in equation (2) runs over all atoms within a single molecule. However, in general the summation includes all atoms. When there is no preferred molecular orientation, the observed scattering intensity $I(q)$ is the rotational average over all orientations. Debye (1915) showed that integration over all orientations gives

$$I(q) = \sum_j^N \sum_k^N b_j b_k \frac{\sin\left(q|\mathbf{r}_j - \mathbf{r}_k|\right)}{q|\mathbf{r}_j - \mathbf{r}_k|} \tag{3}$$

[for a derivation see Warren (1990)]. Owing to the double summation over atom pairs, the $O(N^2)$ calculation can be very time consuming. For calculations based on fully atomistic data, N can be as high as $\sim10^6$. Though molecules are represented by only a few hundred beads in current *ab initio* bead-modeling studies, potentially millions of possible configurations must be tested (Chacón *et al.*, 1998, 2000; Walther *et al.*, 2000; Franke & Svergun, 2009; Volkov & Svergun, 2003; Svergun, 1999; Svergun *et al.*, 2001). For either of these applications, evaluating Debye's formula is a formidable task, even for large computer clusters (Gelisio *et al.*, 2010). Approximations are often employed for greater efficiency.

The most popular approach (Svergun & Stuhrmann, 1991; Svergun *et al.*, 1995; Stuhrmann, 1970*a*; Merzel & Smith, 2002) employs a multipole expansion, truncating an infinite series at L terms. Because of the mathematical properties of the expansion, the rotationally averaged intensity can then be easily calculated. The $O(L^2N)$ computational effort of this technique grows linearly with the number of atoms, offering considerable time savings. However, the processing time grows quadratically with the number of harmonics L. The method can be inaccurate for anisotropically shaped (*i.e.* nonspherical) molecules using small ($L \lesssim 20$) values of L [see §4 and Gumerov *et al.* (2012)]. The multipole expansion technique may also be inappropriate for systems of multiple interacting proteins, which exhibit even greater anisotropy and whose shape cannot be described by a single bounding surface.

A second approach involves constructing a histogram of atomic separations with a specified bin size (Pantos & Bordas, 1994; Pantos *et al.*, 1996). The Debye formula then becomes a sum over bins and scattering lengths for each magnitude q of the scattering vector. The technique is fast when the number of unique scattering lengths $b_j(q)$ is small compared to the number of q values. This condition holds for neutron scattering, where the scattering length does not depend on the value of q. However, it is not the case for X-ray scattering. The approximation is also limited by the initial $O(N^2)$ construction of the histogram.

In this paper, we present a simple approximation method for calculating the scattering profile of any system that can be expressed as a collection of atoms or coarse-grained spheres. Part of our technique is similar to a method used for analyzing simulations of interacting ellipsoids (Sjöberg, 1999). We have extended that approach to study systems of biological complexity. Our technique relies on an exact $O(N)$ expression for the intensity $I(\mathbf{q})$ at a given scattering vector $\mathbf{q}$. We then approximate the rotationally averaged scattering intensity by evaluating $I(\mathbf{q})$ in several scattering directions. The $\mathbf{q}$ vectors describing each direction are generated using the golden ratio and are isotropically distributed. At a given level of speed, our method is more accurate than the multipole expansion technique (Svergun *et al.*, 1995) for anisotropically shaped molecules, while equally accurate for globular shapes. This makes our approach especially valuable for irregularly shaped and intrinsically disordered proteins, whose structures cannot be determined using crystallography. Though the multipole expansion can be useful for modeling globular proteins, often their overall shapes have already been determined in crystallography studies. Our method can be further expedited when the atoms are identical and lie on a lattice, which is frequently the case with *ab initio* bead modeling and powder diffraction studies. These advances mark an important practical step toward greater accuracy in structure determination and the enhanced ability to compare atomistic descriptions of molecules with small-angle scattering data.

3. Method

Our technique is based on the exact calculation of $I(\mathbf{q})$ for a specific scattering vector $\mathbf{q}$. To obtain $I(\mathbf{q})$, first $A(\mathbf{q})$ is numerically computed and then its complex modulus is taken: $I(\mathbf{q}) = \{\mathrm{Re}[A(\mathbf{q})]\}^2 + \{\mathrm{Im}[A(\mathbf{q})]\}^2$. Using Euler's formula, this can be written as

$$I(\mathbf{q}) = \Big[\sum_j^N b_j \cos\big(\mathbf{q}\cdot\mathbf{r}_j\big)\Big]^2 + \Big[\sum_j^N b_j \sin\big(\mathbf{q}\cdot\mathbf{r}_j\big)\Big]^2. \tag{4}$$

The calculation is $O(N)$ as long as the quantities inside the square brackets are numerically evaluated before being squared. This result is exact and mathematically equivalent to equations (1) and (2). The above formula is very convenient for *ab initio* bead modeling since atoms can be relocated, removed or added without recalculating the entire sums. These operations are of $O(N)$ complexity or greater when using a histogram of atomic separations (Walther *et al.*, 2000). To our knowledge, equation (4) was first used by Sjöberg (1999) for efficiently analyzing many-body simulations. The structure factor of hard-sphere systems has also been calculated by taking the fast Fourier transform of the particle positions (Frenkel *et al.*, 1986). However, the discretization of the positions may lead to numerical artefacts at moderate to large values of q (Cannavacciuolo *et al.*, 2002).

When the molecules adopt all possible orientations, the rotationally averaged intensity $I(q)$ may be obtained by averaging $I(\mathbf{q})$ over all scattering directions. While the scattering vector does not change in the actual experiment, rotating a molecule for a fixed scattering vector is equivalent to changing the scattering vector for a fixed molecular orientation. We approximate $I(q)$ by evaluating equation (4) for n scattering vectors with magnitude q, and simply take the mean result. The directions of the vectors are drawn from a quasi-uniform lattice on a sphere (González, 2010). A similar lattice was first used within the context of proteins by Svergun (1994), where the number of points belongs to the Fibonacci sequence $\{\ldots 1, 2, 3, 5, 8, 13, 21, 34, \ldots\}$. In contrast to the Fibonacci method (Grishaev *et al.*, 2010), our chosen procedure for constructing the lattice allows the number of grid points n to be more finely adjusted, where n can be any odd integer. The scattering vectors are given by

$$\begin{aligned} q_x^{(k)} &= q\cos\left[\sin^{-1}(2k/n)\right]\cos(2\pi k/\Phi), \\ q_y^{(k)} &= q\cos\left[\sin^{-1}(2k/n)\right]\sin(2\pi k/\Phi), \\ q_z^{(k)} &= 2kq/n, \end{aligned} \tag{5}$$

where k runs over $\{-(n-1)/2, \ldots, 0, \ldots, (n-1)/2\}$ and $\Phi = (1+5^{1/2})/2$ is the golden ratio. The orientations of the vectors $\mathbf{q}^{(k)}$ for various values of n are shown in Fig. 1. Note that since q can be factored out of equation (5) the lattice only needs to be generated once. The rotationally averaged intensity is then approximated as

$$I(q) = \frac{1}{n}\left\{\sum_{k=(1-n)/2}^{(n-1)/2} I[\mathbf{q}^{(k)}]\right\}. \tag{6}$$

The total computational effort of this procedure scales as $O(nN)$. Repeating the calculation after adding, removing or relocating an atom is an $O(n)$ task. While the rotational average may be performed using more sophisticated numerical integration techniques (Poitevin *et al.*, 2011; Bardhan *et al.*, 2009), these algorithms require more time to implement, and the number of grid points is not as flexible. Equations (4)–(6) are the main result of this paper, establishing a straightforward technique for calculating scattering profiles at a fraction of the computational cost. An implementation of the algorithm for biological molecules can be downloaded at http://www.smallangles.net/sassie/.

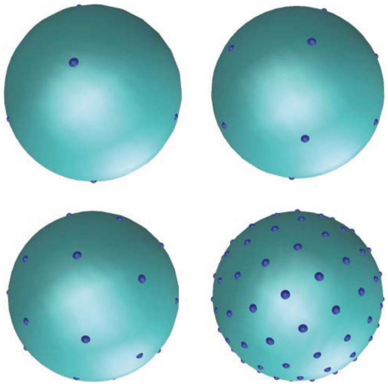

Figure 1
The orientations of the scattering vectors $\mathbf{q}^{(k)}$ generated in equation (5) for $n = \{15, 29, 51, 201\}$. Each $\mathbf{q}^{(k)}$ is represented by a vector pointing from the center of the unit sphere (cyan) to a point on its surface (blue).

The accuracy of the 'golden vector method' depends on the number of vectors n and the geometry of the molecule. As n increases, the average over scattering directions becomes more isotropic, leading to closer agreement with the Debye formula [equation (3)]. For anisotropically shaped molecules, the intensity $I(\mathbf{q})$ is sensitive to the orientation of the scattering vector $\mathbf{q}$. To ensure accuracy, a larger value of n is therefore generally required, compared with that required for more spherical shapes. For a fixed range in q and a given shape, more orientations are also needed for larger molecules. This size effect vanishes for a perfectly spherical shape but becomes relevant for anisotropic molecules. As an example, consider a series of cubes of different sizes. Using the same number of vectors n, the scattering profile of each cube will be accurate up to a q value that *decreases* as the cube size *increases*. If comparable accuracy for all profiles is desired up to the same q value, n must be increased for the larger cubes. However, for a series of spheres of different radii, $I(q)$ does not depend on n, and the size effect just described does not exist. Geometric effects aside, the appropriate value of n can be determined by comparing the corresponding scattering profile with Debye's formula (§4) or with $I(q)$ using a larger n (§5).

4. Application to biological molecules

The golden vector (GV) method was compared with Debye's formula and a multipole expansion (ME) technique (Svergun *et al.*, 1995) for four molecules: lysozyme [Protein Data Bank (PDB) code 6lyz; Diamond, 1974], MCM (mini chromosome maintenance complex) (PDB code 1ltl; Fletcher *et al.*, 2003), ferritin (PDB code 1fha; Lawson *et al.*, 1991) and a 200 Å-long fragment of double-stranded DNA (B form). Additional residues were added to the unstructured portion of MCM (Krueger *et al.*, 2011). The DNA molecule consists of 60 random base pairs arranged along a straight line (Dijk & Bonvin, 2009). Neutron scattering lengths were used for all molecules (for X-ray scattering lengths visit http://www.reflectometry.org/danse/docs/elements/). For simplicity, the calculations were based on the positions of the Cα atoms only, while their scattering lengths were dictated by their respective amino acid (Jacrot & Zaccai, 1981). The resulting scattering profiles were very close to the all-atom intensities for $0 \le q \le 0.5$ Å^{-1}. Since there is no simple analog to Cα atoms for nucleotides, we used the positions of all atoms for DNA. For each molecule, the number of atoms as well as the length along the longest dimension (D) are listed in Table 1. The largest dimension was determined by looping over all atomic separations.

To quantitatively compare the GV and ME methods, we determined the number of vectors/harmonics needed for each approximation to agree with Debye's formula at various degrees of accuracy. Using m values of q equally spaced in the range $0 \le q_i \le 0.5$ Å^{-1}, n and L were increased until the

Table 1
The molecules studied in this paper.

The label (Cα) denotes the use of Cα atoms only. The overall size of each molecule is described by D, its largest linear dimension.

Molecule	Number of atoms (N)	D (Å)
Lysozyme	129 (Cα)	44
Ferritin	1764 (Cα)	94
DNA	2460	205
MCM	3432 (Cα)	373

average difference between the GV/ME approximations and the Debye formula obeyed

$$\frac{1}{m}\sum_{q_i}\frac{|I_{\rm GV/ME}(q_i)-I_{\rm exact}(q_i)|}{I_{\rm exact}(q_i)} \le \Delta \qquad (7)$$

for $\Delta = \{10\%, 7.5\%, 5\%, 2.5\%\}$ and $m = 20$. Since the deviation is the average over points, the exact value of m does not affect the outcome. The χ^2 test was not used since it requires error measurements from experimental data. To calculate $I_{\rm ME}(q)$, we used equations (5) and (11) of Svergun *et al.* (1995) and the same scattering lengths as above. Our method of comparison is relatively strict, since many small-angle scattering measurements have larger error bars in the high-q region and do not always extend to $q = 0.5$ Å^{-1}. If the range were $0 \le q_i \le 0.3$ Å^{-1}, for example, fewer vectors/harmonics would be necessary to yield the same level of agreement with the Debye formula. The results of the analysis are listed in Table 2 and are partially shown in Fig. 2.

For the GV technique, the number of vectors needed for a given level of accuracy generally increases with the molecule's size and anisotropy. At 2.5% agreement with the Debye formula, MCM has the highest n value. While ferritin and lysozyme require roughly similar n values, DNA actually requires the least. This can be understood by noting that $I(\mathbf{q})$ for a rod is highly peaked when $\mathbf{q}$ is perpendicular to the rod's axis (Roe, 2000). This situation occurs multiple times when $\mathbf{q}$ is taken from an isotropic distribution.

For the ME method, the required number of harmonics corresponds to the molecule's size and shape in a more straightforward way. Lysozyme and ferritin are globular and

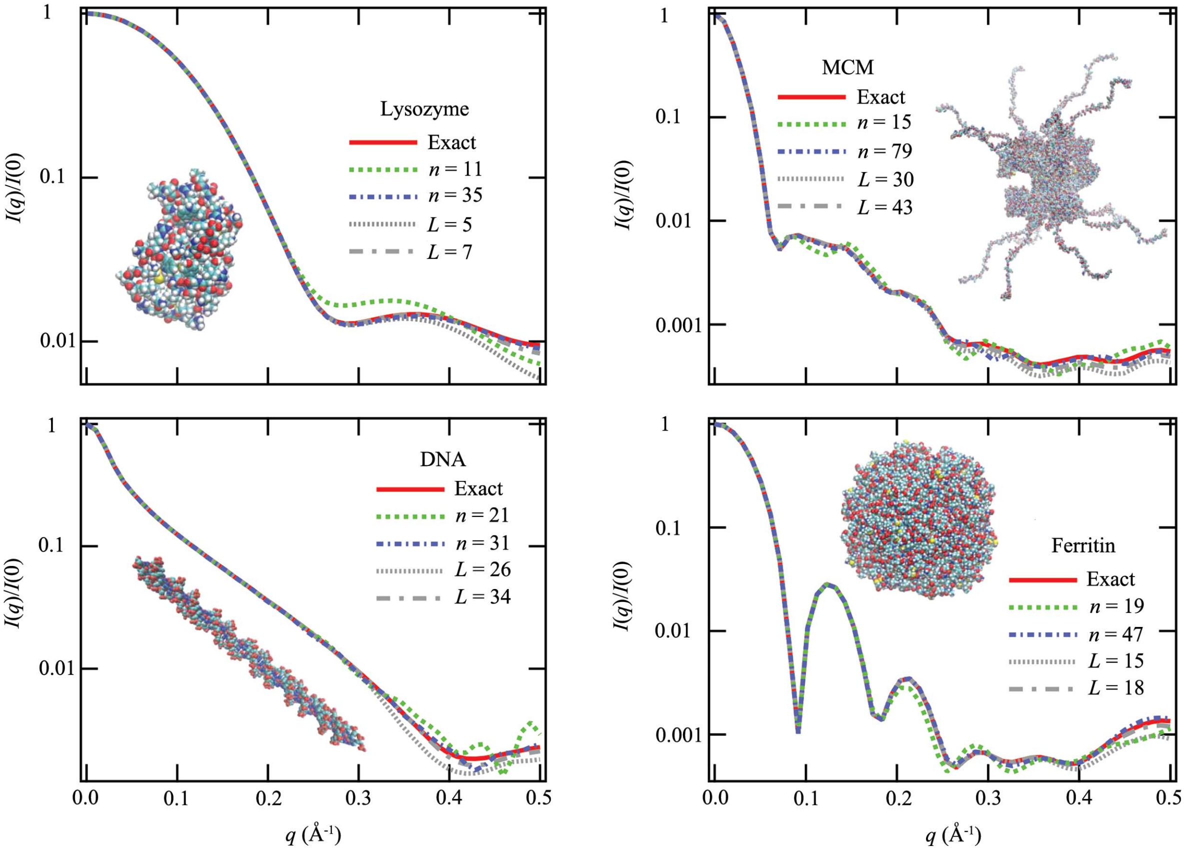

Figure 2
The scattering profiles of lysozyme, MCM, ferritin and a fragment of DNA. All curves are normalized by their zero angle value $I(0)$. The images of each molecule are not on the same scale. The red curves correspond to Debye's formula [equation (3)]. The blue and green curves were obtained using the GV approximation with different numbers of vectors n. The gray curves were calculated using an ME approach (Svergun *et al.*, 1995) with L harmonics. Both approximations improve as n and L increase, and become more accurate on shorter length scales. As listed in Table 2, the n/L values shown correspond to specific levels of accuracy. For both approximation methods, the calculations corresponding to the dotted curves differ from Debye's formula by an average of 10%. For lysozyme, ferritin and DNA, the dash–dotted curves differ by 2.5%. For MCM, the dash–dotted lines are within 5%. The processing times for each approximation are listed in Table 2. Note that, for a given level of accuracy Δ as measured by equation (7), the disagreement between the ME approximation and the Debye formula grows monotonically with increasing q. It grows non-monotonically for the GV method.

Table 2
The speed of the GV and ME methods at different levels of accuracy.

For each molecule, n/L correspond to the minimum number of vectors/harmonics required for the average deviation between the Debye formula and the GV/ME approximations to be less than 10, 7.5, 5 and 2.5%. The corresponding processing times for the GV method, the ME package *CRYSON* (Svergun *et al.*, 1998) and the Debye formula are listed in milliseconds within square brackets. Since the maximum value of L cannot exceed 50, the *CRYSON* timing for MCM at Δ = 2.5% is not available.

	10% match		7.5% match		5% match		2.5% match		
Molecule	n	L	n	L	n	L	n	L	Exact
Lysozyme	11 [1.2]	5 [3.5]	25 [2.4]	5 [5.9]	29 [3.0]	6 [6.3]	35 [3.7]	7 [8.1]	[6.5]
Ferritin	19 [24]	15 [59]	25 [33]	15 [54]	25 [33]	16 [60]	47 [62]	18 [72]	[1300]
DNA	21 [38]	26 [170]	21 [38]	28 [190]	25 [46]	30 [210]	31 [59]	34 [270]	[2400]
MCM	15 [38]	30 [300]	17 [45]	35 [370]	79 [200]	43 [550]	149 [390]	>50 [–]	[5010]

can be modeled with a small set of harmonics. However, L must be substantially increased for MCM and DNA. Note that, in addition to its quasi-spherical shape, lysozyme requires fewer harmonics because of its small size. This can be understood with the same reasoning as in the previous section. For each molecule, the required number of harmonics can be estimated using information theory (Moore, 1980): $L = q_{\rm max}D/\pi$. Using $q_{\rm max} = 0.5$ Å^{-1} and the values of D from Table 1, this equation gives $L = \{7, 15, 33, 59\}$ for lysozyme, ferritin, DNA and MCM, respectively. These predictions are in good overall agreement with the number of harmonics listed in Table 2. However, that equation should only serve as an estimate and does not take the desired level of accuracy or the molecule's geometry into account. For example, a perfect sphere of any size can be exactly described by $L = 1$.

The computer processing times for the GV and ME techniques are listed in Table 2. Both the GV method and Debye's formula were programmed in Fortran. The ME approximations were calculated using *CRYSON* (Svergun *et al.*, 1998). The size of the grid used for defining the molecular envelope within *CRYSON* was set to the minimum value. In order to minimize contributions from peripheral tasks (*e.g.* reading coordinate files, setting up data structures *etc.*) we subtracted the wall clock time required for calculating $I_{\rm GV/ME}(q)$ with $n = 1$ and $L = 1$, respectively. Note that, while the GV technique calculates the *in vacuo* scattering due to the atoms within each molecule, *CRYSON* also calculates the approximate scattering intensity due to the displaced solvent and hydration layer (Svergun *et al.*, 1995, 1998). Thus for computing the *in vacuo* scattering profile, the ME timings may be reduced by ~30% compared to those in Table 1 (an exact comparison is not possible since the source code is unavailable). While the lysozyme and ferritin timings are comparable between the two techniques, the GV method is about four times faster than the ME approximation for DNA and eight times faster for MCM at $\Delta = 10$ and 7.5%. For MCM, the GV method is twice as fast as the ME technique for $\Delta = 5\%$. For the ME approximation to be within 2.5% agreement with the Debye formula, the number of harmonics exceeds 50 and cannot be calculated within *CRYSON*. Judging from these four molecules, we generally expect that the two approximation methods are roughly equivalent for quasi-spherical molecules, while the GV technique is more accurate and efficient for anisotropic shapes.

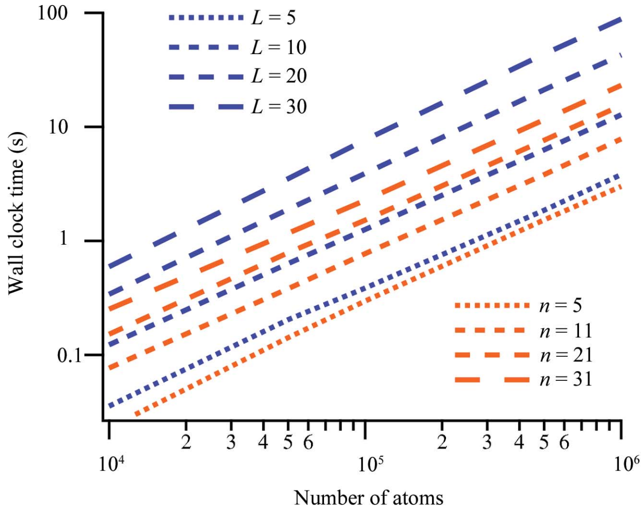

Figure 3
The processing time required to calculate $I(q)$ *versus* the number of atoms using the GV method and an ME technique (Svergun *et al.*, 1998). Timings for the GV method are shown for $n = \{5, 11, 21, 31\}$ vectors (orange). The ME timings for $L = \{5, 10, 20, 30\}$ harmonics are displayed (violet).

In addition to specific molecules, we also examined how the ME and GV methods scale with the number of atoms, n and L. We generated configurations of $N = 10^4$–10^6 atoms whose positions were randomly assigned inside a sphere of average number density $\rho = 0.02$ Å^{-3}. The radius was chosen to correspond to the volume V for which $N/V = \rho$. Each 'molecule' was written to a PDB file, with each atom designated as a Cα ('CA'). The wall clock times are shown in Fig. 3, where $I(q)$ was calculated for 20 values of q. Just as in Table 1, the timings for $n = 1$ and $L = 1$ calculations were subtracted. The timings for both the GV and ME methods are proportional to the number of atoms. Owing to the $O(L^2N)$ effort of the ME method, the timings rapidly increase for larger L. Though not implemented here, the GV method may run even faster by taking advantage of trigonometric identities (Svaneborg & Pedersen, 2000) in equation (4). For example, $\sin(2x)$ does not need to be explicitly evaluated if $\sin(x)$ and $\cos(x)$ are already known. Because of its simple $O(nN)$ linear scaling, the GV method can easily be distributed on graphical processing units (GPUs) as well.

5. Future extensions

While we focused on examples involving a single type of biomolecule in a dilute solution, the GV approach can be applied to a host of other situations. Ensembles of noninteracting molecules contain simple additive contributions to the total scattering intensity (Blanchet & Svergun, 2013). The Debye formula has become increasingly used to interpret neutron and X-ray powder diffraction data for nanocrystalline systems (Cademartiri *et al.*, 2006; Chiche *et al.*, 2008; Thomas, 2009; Cervellino *et al.*, 2003, 2010; Oddershede *et al.*, 2008; Beyerlein *et al.*, 2011, 2010). As with small-angle scattering, the analysis is limited by the severe $O(N^2)$ computational overhead (Gelisio *et al.*, 2010; Beyerlein *et al.*, 2011). Since the GV approximation technique converges to the Debye formula, it could be applicable to powder diffraction as well.

In this paper, we determined the number of vectors required by comparing the GV approximation with the Debye formula. However, the Debye formula is computationally expensive to calculate. In order to determine the sufficient number of vectors without comparing to the Debye formula or experimental data, we suggest the following procedure. Compute the scattering profile using n and $n+2$ vectors (since n must be odd), which we denote by $I_{\rm GV}^{(n)}$ and $I_{\rm GV}^{(n+2)}$, respectively. If the difference between the curves is sufficiently small, $I_{\rm GV}^{(n+2)}$ is a good approximation. If the difference is not small, calculate $I_{\rm GV}^{(n+4)}$, compare it with $I_{\rm GV}^{(n+2)}$, and repeat the process. The 'difference' between curves may be quantified on the basis of the user's favorite metric: the relative deviation described in §4, root-mean square deviation *etc*.

Though not considered in this paper, the bulk solvent displaced by the macromolecules and the surrounding hydration layer also contribute to the scattering intensity (Svergun *et al.*, 1995, 1998). These effects may be directly incorporated within the GV method using an atomic representation of the solvent. The displaced solvent has long been described by dummy atoms located at the position of each solute atom (Svergun *et al.*, 1995; Fraser *et al.*, 1978; Poitevin *et al.*, 2011; Schneidman-Duhovny *et al.*, 2010). Explicit representation of the excluded solvent is also possible (Grishaev *et al.*, 2010). The hydration layer can be constructed by including denser water molecules at the solute/solvent interface (Svergun *et al.*, 2001; Perkins, 2001; Yang *et al.*, 2009; Grishaev *et al.*, 2010; Merzel & Smith, 2002). For these cases, the sum in equation (4) would run over solute atoms and explicit solvent atoms. Alternatively, the scattering length of each solute atom can be renormalized according to its solvent-accessible surface area (Schneidman-Duhovny *et al.*, 2010). Just as in previous studies, coefficients of scattering lengths/amplitudes may be included within our approximation in order to fit experimental data. In *FoXS* (Schneidman-Duhovny *et al.*, 2010) for example, the effective scattering length of each solute atom is modeled as $b_j^{\rm eff} = b_j^{\rm v} - c_1 b_j^{\rm s} + c_2 s_j b_j^{\rm w}$, where $b_j^{\rm v}$, $b_j^{\rm s}$ and $b_j^{\rm w}$ correspond to the scattering lengths of the atom *in vacuo*, the dummy atom and the water in the neighboring hydration layer, respectively. s_j is the solvent-accessible surface area of the atom, while c_1 and c_2 are fit parameters used to adjust the total excluded volume of the atoms and the density of the hydration layer, respectively [analogous to r_0 and $\delta\rho$ in *CRYSOL* (Svergun *et al.*, 1995)]. In our treatment, $b_j^{\rm eff}$ would replace b_j in equation (4). Regardless of the chosen solvent description, the GV method provides a fast and accurate engine for scattering calculations, upon which any atomic model can be built.

The processing time of the GV method can be made to scale sub-linearly with N when the atoms are identical and lie on a lattice. This is the case for several powder diffraction studies (Cervellino *et al.*, 2003, 2010; Chiche *et al.*, 2008; Thomas, 2009; Oddershede *et al.*, 2008; Beyerlein *et al.*, 2011, 2010) and most *ab initio* bead-modeling techniques. On a lattice, the position of each atom may be written as a linear combination of primitive vectors $\mathbf{a}_{1,2,3}$. In any one of these directions $\mathbf{a}_j$, the positions of the atoms along a single row are given by $\mathbf{r}_p = \mathbf{r}_0 + p\mathbf{a}_j$, where p runs from 0 to $N_{\rm r} - 1$. The scattering amplitude of that row is a simple geometric series:

$$\begin{aligned} A_{\rm row}(\mathbf{q}) &= b\sum_{p=0}^{N_{\rm r}-1} \exp\big[-i(\mathbf{q}\cdot\mathbf{r}_0 + p\mathbf{q}\cdot\mathbf{a}_j)\big] \\ &= b\,\frac{\sin\big(N_{\rm r}\mathbf{q}\cdot\mathbf{a}_j/2\big)}{\sin\big(\mathbf{q}\cdot\mathbf{a}_j/2\big)}\exp\left\{-i\left[\mathbf{q}\cdot\mathbf{r}_0 + \frac{(N_{\rm r}-1)\mathbf{q}\cdot\mathbf{a}_j}{2}\right]\right\}, \end{aligned} \tag{8}$$

where $N_{\rm r}$ is the number of atoms in the specific row and b is the scattering length. Equation (2) would then reduce to sums over rows, effectively lowering the dimension of the calculation. The resulting computational overhead scales sub-linearly with the number of atoms: $O(nN^{\alpha})$, where $\alpha < 1$. The value of α depends on the shape of the molecule. For example, applying this technique to a cube made on an N-atom lattice would be an $O(nN^{2/3})$ task.

The GV method will also be useful for rapidly analyzing atomically detailed simulations of dense protein solutions (McGuffee & Elcock, 2006; Chaudhri *et al.*, 2012; Mereghetti *et al.*, 2010). Though the scattering intensity is normally calculated under the assumption that orientational correlations between molecules are negligible (Roe, 2000), these correlations may be significant in highly concentrated protein solutions. For such systems, $I(q)$ must be calculated directly from its definition [equations (1) and (2)]. Using the GV approximation, the sum in equation (4) would run over the atoms of all proteins in the simulation and be averaged over the entire trajectory. The scattering profile could then be directly compared with experimental data.

6. Conclusion

We have presented an efficient and mathematically transparent technique for calculating the scattering profiles of systems described on an atomic level. For a given level of accuracy, the GV method is comparable in speed to the ME approach for quasi-spherical molecules but faster for anisotropic shapes. Conversely, for a given level of speed, the GV approximation is more accurate for irregular shapes. In addition to interpreting small-angle scattering experiments, this approach also has potential applications in powder diffraction and protein simulations. Given its utility, the GV technique should be a valuable extension to any scattering toolbox.

MCW acknowledges the support of the National Research Council. This material is based upon work supported by the National Science Foundation under CHE-1265817.

References

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

Beyerlein, K. R., Snyder, R. L. & Scardi, P. (2011). *J. Appl. Cryst.* **44**, 945–953.

Beyerlein, K., Solla-Gullón, J., Herrero, E., Garnier, E., Pailloux, F., Leoni, M., Scardi, P., Snyder, R., Aldaz, A. & Feliu, J. (2010). *Mater. Sci. Eng. A*, **528**, 83–90.

Blanchet, C. E. & Svergun, D. I. (2013). *Annu. Rev. Phys. Chem.* **64**, 37–54.

Cademartiri, L., Montanari, E., Calestani, G., Migliori, A., Guagliardi, A. & Ozin, G. (2006). *J. Am. Chem. Soc.* **128**, 10337–10346.

Cannavacciuolo, L., Pedersen, J. & Schurtenberger, P. (2002). *Langmuir*, **18**, 2922–2932.

Cervellino, A., Giannini, C. & Guagliardi, A. (2003). *J. Appl. Cryst.* **36**, 1148–1158.

Cervellino, A., Giannini, C. & Guagliardi, A. (2010). *J. Appl. Cryst.* **43**, 1543–1547.

Chacón, P., Díaz, J. F., Morán, F. & Andreu, J. M. (2000). *J. Mol. Biol.* **299**, 1289–1302.

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

Chaudhri, A., Zarraga, I. E., Kamerzell, T. J., Brandt, J. P., Patapoff, T. W., Shire, S. J. & Voth, G. A. (2012). *J. Phys. Chem. B*, **116**, 8045–8057.

Chiche, D., Digne, M., Revel, R., Chanéac, C. & Jolivet, J. (2008). *J. Phys. Chem. C*, **112**, 8524–8533.

Curtis, J. E., Nanda, H., Khodadadi, S., Cicerone, M., Lee, H. J., McAuley, A. & Krueger, S. (2012). *J. Phys. Chem. B*, **116**, 9653–9667.

Debye, P. (1915). *Ann. Phys.* (*Leipzig*), **46**, 809–823.

Diamond, R. (1974). *J. Mol. Biol.* **82**, 371–391.

Dijk, M. V. & Bonvin, A. (2009). *Nucleic Acids Res.* **37**(Suppl. 2), W235–W239.

Fletcher, R., Bishop, B., Leon, R., Sclafani, R., Ogata, C. & Chen, X. (2003). *Nat. Struct. Mol. Biol.* **10**, 160–167.

Förster, F., Webb, B., Krukenberg, K. A., Tsuruta, H., Agard, D. A. & Sali, A. (2008). *J. Mol. Biol.* **382**, 1089–1106.

Franke, D. & Svergun, D. I. (2009). *J. Appl. Cryst.* **42**, 342–346.

Fraser, R. D. B., MacRae, T. P. & Suzuki, E. (1978). *J. Appl. Cryst.* **11**, 693–694.

Frenkel, D., Vos, R., Kruif, C. D. & Vrij, A. (1986). *J. Chem. Phys.* **84**, 4625–4630.

Gelisio, L., Azanza Ricardo, C. L., Leoni, M. & Scardi, P. (2010). *J. Appl. Cryst.* **43**, 647–653.

González, Á. (2010). *Math. Geosci.* **42**, 49–64.

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

Guinier, A. (1939). *Ann. Phys.* (*Paris*), **12**, 161–237.

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

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

Jacrot, B. & Zaccai, G. (1981). *Biopolymers*, **20**, 2413–2426.

Konarev, P. V., Petoukhov, M. V. & Svergun, D. I. (2001). *J. Appl. Cryst.* **34**, 527–532.

Kratky, O. & Porod, G. (1949). *Recl Trav. Chim. Pays-Bas*, **69**, 1106–1122.

Krueger, S., Shin, J. H., Raghunandan, S., Curtis, J. E. & Kelman, Z. (2011). *Biophys. J.* **101**, 2999–3007.

Lawson, D. M., Artymiuk, P. J., Yewdall, S. J., Smith, J. M. A., Livingstone, J. C., Treffry, A., Luzzago, A., Levi, S., Arosio, P., Cesareniparallel, G., Thomas, C. D., Shaw, W. V. & Harrison, P. M. (1991). *Nature*, **349**, 541–544.

Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S. & Shaw, D. (2012). *J. Am. Chem. Soc.* **134**, 3787–3791.

McGuffee, S. & Elcock, A. (2006). *J. Am. Chem. Soc.* **128**, 12098–12110.

Mereghetti, P., Gabdoulline, R. R. & Wade, R. C. (2010). *Biophys. J.* **99**, 3782–3791.

Merzel, F. & Smith, J. C. (2002). *Acta Cryst.* D**58**, 242–249.

Monticelli, L., Sorin, E. J., Tieleman, D. P., Pande, V. S. & Colombo, G. (2008). *J. Comput. Chem.* **29**, 1740–1752.

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

Oddershede, J., Christiansen, T. L. & Ståhl, K. (2008). *J. Appl. Cryst.* **41**, 537–543.

Pantos, E. & Bordas, J. (1994). *Pure Appl. Chem.* **66**, 77.

Pantos, E., van Garderen, H., Hilbers, P., Beelen, T. & van Santen, R. (1996). *J. Mol. Struct.* **383**, 303–308.

Perkins, S. J. (2001). *Biophys. Chem.* **93**, 129–139.

Petoukhov, M. V. & Svergun, D. I. (2005). *Biophys. J.* **89**, 1237–1250.

Petoukhov, M. V. & Svergun, D. I. (2006). *Eur. Biophys. J.* **35**, 567–576.

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

Putnam, C. D., Hammel, M., Hura, G. L. & Tainer, J. A. (2007). *Q. Rev. Biophys.* **40**, 191–285.

Roe, R. (2000). *Methods of X-ray and Neutron Scattering in Polymer Science.* Topics in Polymer Science. New York, Oxford: Oxford University Press.

Schneidman-Duhovny, D., Hammel, M. & Sali, A. (2010). *Nucleic Acids Res.* **38**(Suppl. 2), W540–W544.

Sjöberg, B. (1999). *J. Appl. Cryst.* **32**, 917–923.

Stuhrmann, H. B. (1970*a*). *Acta Cryst.* A**26**, 297–306.

Stuhrmann, H. (1970*b*). *Z. Phys. Chem.* **72**, 177–184.

Svaneborg, C. & Pedersen, J. (2000). *J. Chem. Phys.* **112**, 9661–9670.

Svergun, D. I. (1994). *Acta Cryst.* A**50**, 391–402.

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

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

Svergun, D. I., Petoukhov, M. V. & Koch, M. H. (2001). *Biophys. J.* **80**, 2946–2953.

Svergun, D. I., Richard, S., Koch, M. H., Sayers, Z., Kuprin, S. & Zaccai, G. (1998). *Proc. Natl Acad. Sci.* **95**, 2267–2272.

Svergun, D. I. & Stuhrmann, H. B. (1991). *Acta Cryst.* A**47**, 736–744.

Svergun, D. I., Volkov, V. V., Kozin, M. B. & Stuhrmann, H. B. (1996). *Acta Cryst.* A**52**, 419–426.

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

Thomas, N. W. (2010). *Acta Cryst.* A**66**, 64–77.

Volkov, V. V. & Svergun, D. I. (2003). *J. Appl. Cryst.* **36**, 860–864.

Wall, M. E., Gallagher, S. C. & Trewhella, J. (2000). *Annu. Rev. Phys. Chem.* **51**, 355–380.

Walther, D., Cohen, F. E. & Doniach, S. (2000). *J. Appl. Cryst.* **33**, 350–363.

Warren, B. (1990). *X-ray Diffraction.* New York: Dover Publications.

Yang, S., Park, S., Makowski, L. & Roux, B. (2009). *Biophys. J.* **96**, 4449–4463.