Atomic-resolution structural information from scattering experiments on macromolecules in solution

Jürgen Köfinger* and Gerhard Hummer†

Laboratory of Chemical Physics, Bldg. 5, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA

(Received 22 October 2012; revised manuscript received 26 March 2013; published 20 May 2013)

The pair-distance distribution function (PDDF) contains all structural information probed in an elastic scattering experiment of macromolecular solutions. However, in small-angle x-ray scattering (SAXS) or small-angle neutron scattering (SANS) experiments only their Fourier transform is measured over a restricted range of scattering angles. We therefore developed a mathematically simple and computationally efficient method to calculate the PDDFs as well as accurate scattering intensities from molecular dynamics simulations. The calculated solution scattering intensities are in excellent agreement with SAXS and wide-angle x-ray scattering (WAXS) experiments for a series of proteins. The corresponding PDDFs are remarkably rich in features reporting on the detailed protein structure. Using an inverse Fourier transform method, most of these features can be recovered if scattering intensities are measured up to a momentum transfer of $q \approx 2$–3 Å^{-1}. Our results establish that high-precision solution scattering experiments utilizing x-ray free-electron lasers and third generation synchrotron sources can resolve subnanometer structural detail, well beyond size, shape, and fold.

DOI: 10.1103/PhysRevE.87.052712 PACS number(s): 87.15.hp, 61.05.cf, 61.05.fg, 87.10.Tf

I. INTRODUCTION

Intense monochromatic x rays at free-electron lasers and third generation synchrotrons, and high-flux neutrons at spallation sources enable high-precision scattering measurements of (bio)macromolecules in solution [1,2]. With these new sources, attention moves beyond the traditional small-angle regime (SAXS and SANS for x-ray and neutron scattering, respectively) probing the size and shape of macromolecules and their assemblies [3,4]. Scattering beyond the SAXS regime reports on protein secondary structure [5] and fold [6]. Time-resolved wide-angle x-ray scattering (WAXS) has already been used to probe protein functional dynamics [7]. However, neither the amount of structural information accessible in WAXS experiments nor the means for their analysis are currently established.

Here, we take a step towards quantifying and extracting the maximum information possible from elastic scattering experiments on macromolecular solutions. All information is contained in the pair-distance distribution function (PDDF) $p(r)$, whose finite range is determined by the correlation length of the macromolecular solution. Scattering experiments measure their Fourier transform $I(q)$, though only over a limited q range. Despite the resulting information loss, scattering data have high discriminatory power and are widely used to assess the validity of structural models. These models have to account not only for scattering contributions of the macromolecules, but also for their excluded volume and solvation shell [8,9]. CRYSOL [10] estimates these solvation contributions using a structureless fluid, which is suitable for SAXS, whereas AXES [11] uses the structure of bulk water, which improves the fitting in the WAXS regime. In both methods, the solvation layer and the excluded volume are modeled to match the shape of the macromolecule, and their respective electron density contrasts to bulk solvent are used as fitting parameters. Nonuniform average electron density near the surface of a protein can be estimated by using the HyPred method [12]. In this method, as in CRYSOL, solvent contributions to the scattering intensity are calculated for an approximate average solvent structure instead of averaging the intensity over an ensemble of structures. Solvent fluctuations are therefore ignored in these methods.

Molecular dynamics (MD) simulations of macromolecules in explicit solvent do not require such approximations and permit, in principle, the direct calculation of scattering intensities, without the need to fit bulk solvent and solvation contributions. However, the calculation of scattering intensities from MD simulations has to account for the finite size of the simulation box, which usually fits the macromolecule tightly. For SAXS and SANS a structureless reference buffer [13] is sufficient, but for WAXS atomistic detail [14,15] is necessary. These methods determine $I(q)$ over a finite q range by using multipole expansions [13,14] or numerical orientational averages [15] and therefore, as in experiment, provide only incomplete real-space information. Here, we instead calculate the complete PDDFs directly to access the maximum amount of structural information provided by solution scattering experiments. In Fig. 1, we validate our real-space approach against accurate scattering intensities measured in the SAXS and WAXS regimes [11].

II. THEORY

A. Method

Solution scattering experiments measure intensities $I(q)$, i.e., the number of photons, as a function of the momentum transfer $q = 4\pi \sin\theta/\lambda$, where 2θ is the scattering angle and λ is the wavelength of the monochromatic incident beam. We calculate the intensity $I(q)$ using

$$I(q) = \sum_{i,j} f_i(q) f_j(q) [\Delta I_{ij}(q) + v I_{ij}(q)], \qquad (1)$$

*juergen.koefinger@nih.gov
†gerhard.hummer@nih.gov

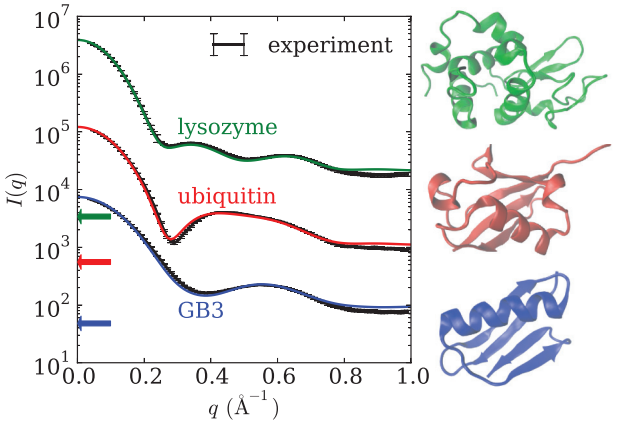

FIG. 1. (Color online) SAXS and WAXS intensities from experiment (black lines, error bars) and all-atom explicit-solvent MD simulations (top to bottom: lysozyme, green; ubiquitin, red; GB3, blue; backbone structures at right). Arrows indicate additive constants correcting intensities for dark currents, possible differences in the electron-density contrast between experiment and MD, and uncertainties in buffer scaling [11]. The latter alone, within the experimental range, account for the magnitude of the constants (see text). Intensities of GB3 and ubiquitin are scaled by 0.01 and 0.1 for clarity.

where the double sum extends over all pairs of particle species i and j with form factors $f_i(q)$ and $f_j(q)$, respectively. The latter are the distinguishing property of the different particle species present in the solvent, the macromolecule, or both. $\Delta I_{ij}(q)$ is the partial intensity difference between macromolecular solution and pure solvent. The excluded-volume (v) term $vI_{ij}(q)$ minimizes bulk solvent contributions by adding bulk solvent scattering intensity oversubtracted in $\Delta I_{ij}(q)$ (see Sec. III C).

$$I_{ij}(q) = \delta_{ij}\rho_i + \rho_i\rho_j 4\pi \int_0^\infty r^2[g_{ij}(r) - \gamma_{ij}]\frac{\sin(qr)}{qr}dr \quad (2)$$

are partial bulk solvent intensities. For large distances, the radial distribution functions $g_{ij}(r)$ of pure solvent approach constants γ_{ij} that can deviate slightly from one for finite simulation systems. ρ_i is the particle number density of species i in the pure solvent and δ_{ij} is the Kronecker symbol. The partial scattering intensity differences satisfy

$$\Delta I_{ij}(q) = \delta_{ij}\Delta N_i + \int_0^{2R} \Delta H_{ij}(r)\frac{\sin(qr)}{qr}dr, \quad (3)$$

with $\Delta N_i = N_i - \rho_i V$ and N_i the average number of particles of species i in an observation sphere of radius R and volume $V = 4\pi R^3/3$ centered on the macromolecule (R sphere). Equation (3) is Debye's formula [16], rewritten here using the difference distribution functions $\Delta H_{ij}(r)$ of interparticle pair distances between the macromolecular solution and pure solvent.

We perform MD simulations of a macromolecule in solution and calculate the partial interatomic distance histograms $H_{ij}(r)dr$ of particles in the R sphere, and the partial histograms $h_i(x)dx$ of their distances x from its center. For the reference buffer, we only have to calculate ρ_i and $g_{ij}(r)$ from a pure solvent simulation. The difference distance distribution functions then satisfy

$$\Delta H_{ij}(r) = H_{ij}(r) - 2\rho_i \int_0^R h_j(x)S_R(x,r)dx - [g_{ij}(r) - 2]\rho_i\rho_j V^2 p_R(r), \quad (4)$$

where $S_R(x,r)$ is the surface area of a sphere of radius r contained within the R sphere at a center distance x and $p_R(r)$ is the PDDF of a structureless R sphere (see Appendix A). Equations (3) and (4) are based on the same scattering theory that also underlies the methods of Oroguchi *et al.* [14,17], and of Park *et al.* for spheres [15], and in this sense the methods for the calculation of $I(q)$ are physically equivalent. In Eq. (1), we additionally include the excluded volume term, which is important in the wide-angle regime (see Sec. III C).

By using distance histograms we can directly calculate the PDDF

$$p(r) = \sum_{i,j}[\Delta p_{ij}(r) + vp_{ij}(r)], \quad (5)$$

where

$$p_{ij}(r) = \delta_{ij}\rho_i c_{ii}(r,0) + \rho_i\rho_j 4\pi \int_0^\infty r'^2[g_{ij}(r') - \gamma_{ij}]c_{ij}(r,r')dr' \quad (6)$$

is the excess bulk-solvent partial PDDF per unit volume that also contains intraparticle contributions, and has the ideal-gas $4\pi r^2$ contributions subtracted.

$$\Delta p_{ij}(r) = \delta_{ij}\Delta N_i c_{ii}(r,0) + \int_0^\infty \Delta H_{ij}(r')c_{ij}(r,r')dr' \quad (7)$$

is the partial PDDF difference between macromolecular solution and pure solvent. The $c_{ij}(r,r') = 2\pi^{-1}\int_0^\infty f_i(q)f_j(q)\sin(qr)\sin(qr')r/r'dq$ are electron PDDFs of two particles of species i and j, whose centers are separated by a distance r'. Atomic radial electron densities $\varrho_i(r)$ determine the x-ray form factors $f_i(q) = 4\pi\int_0^\infty r^2\varrho_i(r)\sin(qr)/(qr)dr$ and in turn $c(r,r')$. For the five-Gaussian approximation [18], analytical expressions for $c_{ij}(r,r')$ simplify the calculation of $p(r)$ (see Appendix B).

B. Derivation

The distance distribution function of the infinite system is given by $H_{ij}^\infty(r) = \langle\sum'_{\alpha\neq\beta}\delta(r - r_{\alpha\beta})\rangle$, where the prime indicates that the sum extends over all distinct pairs of particles α of species i and β of species j, each pair counted exactly once. $r_{\alpha\beta}$ is the particle distance, $\delta(r)$ is Dirac's delta function, and $\langle\cdots\rangle$ indicates an ensemble average. We break up the distribution into contributions coming from inside the R sphere (II), outside the R sphere (OO), and between inside and outside (IO), $H_{ij}^\infty(r) = H_{ij}^{(\mathrm{II})}(r) + H_{ij}^{(\mathrm{OO})}(r) + 2H_{ij}^{(\mathrm{IO})}(r)$. We evaluate these terms first for the bulk solvent and then for the macromolecule in solution. In the difference, the (OO) terms cancel exactly and we obtain Eq. (4), as shown in the following.

For a homogeneous and isotropic system, $H_{ij}^\infty(r)dr$ can be determined by counting particles of species j in a spherical shell of radius r, thickness dr, and surface area $4\pi r^2$ centered at a particle of species i. The expected number of particles within this shell is given by $\rho_j 4\pi r^2 dr g_{ij}(r)$. We divide the

area $4\pi r^2$ into parts within and outside the R sphere, $4\pi r^2 = S_R^{(\mathrm{I})}(x,r) + S_R^{(\mathrm{O})}(x,r)$, where x is the distance of the particle at the shell center from the R-sphere center. $S_R^{(\mathrm{I})}(x,r) \equiv S_R(x,r)$ is given in Appendix A. To calculate the MD simulation ensemble average, we define the distance-from-center distribution functions $h_i^{\infty}(x) = \langle \sum_\alpha \delta(x - x_\alpha)\rangle = h_i^{(\mathrm{I})}(x) + h_i^{(\mathrm{O})}(x)$ for regions I and O. For bulk solvent, $h_i(x) = \rho_i 4\pi x^2$ and $h_i^{(\mathrm{I})}(x) = h_i^{\infty}(x)$ for $x \leqslant R$ and zero otherwise. Consequently, $H_{ij}^{NM}(r) = g_{ij}(r)\rho_i \int_0^\infty h_j^{(\mathrm{N})}(x) S_R^{(\mathrm{M})}(x,r)dx$, where N and M represent regions I or O. For bulk solvent, we obtain $H_{ij}^{(\mathrm{II,s})}(r) = g_{ij}(r)\rho_i\rho_j V^2 p_R(r)$ and $H_{ij}^{(\mathrm{IO,s})}(r) = g_{ij}(r)\rho_i\rho_j[V 4\pi r^2 - V^2 p_R(r)]$.

For the R sphere containing the macromolecule, $H_{ij}^{(\mathrm{II})}(r) = \langle \sum'_{\alpha\neq\beta} \delta(r_{\alpha\beta} - r)\rangle$ and $h_i^{(\mathrm{I})}(x) = \langle \sum_\alpha \delta(x - x_\alpha)\rangle$ are calculated from the simulation trajectory, with $H_{ij}(r) \equiv H_{ij}^{(\mathrm{II})}(r)$ and $h_i(x) \equiv h_i^{(\mathrm{I})}(x)$ in Eq. (4). We obtain $H_{ij}^{(\mathrm{IO})}(r) = g_{ij}(r)\rho_i \int_0^R h_j^{(\mathrm{I})}(x)[4\pi r^2 - S_R(x,r)]dx$ assuming that we can replace the correlations of I and O particles by the bulk pair distribution functions. This assumption is justified if the R sphere contains a sufficiently thick layer of bulk solvent around the macromolecule. Then, particles in the I and O region are either at large distance and thus uncorrelated, or correlated according to $g_{ij}(r)$ at short distance but with contributions that exactly cancel in the difference between macromolecular solution and solvent. Therefore, we can set $g_{ij}(r) = 1$ in the (IO) terms. Terms proportional to r^2 in the IO contributions are neglected because in experiment the corresponding scattering intensities appear at $q \approx 0$, inside the beam stop, due to the large size of the illuminated volumes. The difference $\Delta H_{ij}(r) = H_{ij}^{(\mathrm{II})}(r) + 2H_{ij}^{(\mathrm{IO})}(r) - H_{ij}^{(\mathrm{II,s})}(r) - 2H_{ij}^{(\mathrm{IO,s})}(r)$ then becomes Eq. (4).

C. Efficiency

The calculation of the distance histogram, which is the single most computationally expensive calculation of our method, scales as N^2, where N is the particle number. This scaling might, at first sight, appear inferior to that of other methods. However, the multipole expansion [19], as used for example in CRYSOL [10] and AXES [11], by Oroguchi *et al.* [14], and in various other methods, is efficient in the small-angle regime, but actually becomes cumbersome for wide angles. To reach a given level of accuracy at a maximum scattering angle $q_{\max}$, one finds that spherical harmonics up order $l_{\max} \sim Dq_{\max}$ have to be evaluated for every one of the N atoms, where D is the maximum dimension of the structure. With $\sim N l_{\max}^2$ such terms, and $D \sim N^{1/3}$ for globular structures, the overall computational cost scales as $N^{5/3}$ (and worse for elongated structures) with system size and like $q_{\max}^2$ with the largest q value. Thus, the difference in scaling with system size is only $\sim N^{0.33}$ in the worst case. With such small differences in scaling, the relative efficiencies depend strongly on the prefactor, which is small for the calculation of histograms.

Our method is mathematically simple and allows us to calculate the scattering intensity up to essentially arbitrarily large values of q at no additional computational cost, once the distance histograms have been calculated. This histogram calculation can be easily performed on GPUs with high efficiency [20]. For the intensity calculation over a restricted q range, the Fourier transform of $H_{ij}(r)$ in Eq. (4) can also be calculated directly, e.g., by using massively parallel computers [21].

D. Self-consistent solvent matching

By working in real space in terms of ρ_i and $g_{ij}(r)$, we can control statistical errors by tapering noise in $g_{ij}(r)$. Moreover, we can self-consistently match the densities and density fluctuations of the solvent that differ slightly in pure solvent and macromolecular simulations due to unavoidable finite-size effects. Otherwise, the signal due to the shape of the finite sample volume (i.e., of the R sphere) contaminates the scattering intensity at small angles.

In a first step, we match the solvent electron density $\rho_e = \sum_i \rho_i f_i(q = 0)$ to the electron density in the bulk layer of the macromolecular system by uniformly scaling all ρ_i. As in previous methods [14,15], such a bulk layer is necessary to account for all particle correlations that do not cancel in the difference signal (see Sec. II B). The number of electrons $\mathcal{N}$ inside the R sphere containing the macromolecule satisfies $\langle \mathcal{N}(R)\rangle = \langle \mathcal{N}(R_0)\rangle + [V(R) - V(R_0)]\rho_e$, where $V(R)$ is the sphere volume and $R_0 < R$ is the inner radius of the bulk solvent layer. $\mathcal{N}$ is calculated by summing $f_i(q = 0)$ over all particles within the R sphere. We determine ρ_e by minimizing the least square deviation to the numerical data for $\langle \mathcal{N}(R)\rangle$ for a range of sphere sizes R with fixed R_0.

In a second step, we exploit the fact that $I(0)$ should be independent of the size of the R sphere. From Eq. (1), ignoring the R-independent excluded volume term, we have $I(0) = \langle \mathcal{N}^2\rangle - \langle \mathcal{N}_s^2\rangle - 2\langle \mathcal{N}_s\rangle(\langle \mathcal{N}\rangle - \langle \mathcal{N}_s\rangle)$, where $\langle \mathcal{N}_s\rangle = \rho_e V$. Using that $\langle \mathcal{N}_s^2\rangle - \langle \mathcal{N}_s\rangle = V^2\rho_e^2 \int dr\, g_e(r) p_R(r)$, where $g_e(r) = \sum_{i,j} g_{ij}(r) V^2 \rho_i \rho_j f_i(0) f_j(0)/\langle \mathcal{N}_s\rangle^2$ is the electron-weighted sum of partial radial distance distribution functions, we obtain

$$I(0) = \langle \mathcal{N}^2\rangle - \langle \mathcal{N}_s\rangle^2 \int_0^{2R} g_e(r) p_R(r) dr + \langle \mathcal{N}_s\rangle - 2\langle \mathcal{N}_s\rangle(\langle \mathcal{N}\rangle - \langle \mathcal{N}_s\rangle). \quad (8)$$

If the solvent properties are matched exactly then $I(0)$ will be independent of the radius R. This matching is achieved by scaling $g_e(r)$, and therefore $g_{ij}(r)$, by a factor s. Whereas s usually deviates from one by less than 10^{-4}, the resulting correction effectively suppresses statistical noise and significantly increases the accuracy of the calculated $I(q)$ in the small-angle regime.

E. Simulation details

To investigate the q-range dependence of the structural information probed in experiments, we performed all-atom explicit-solvent MD simulations for a selection of small proteins: GB3 (PDB code 1IGD; 4 N-terminal residues removed as in experiment) [22], ubiquitin (1D3Z) [23], and lysozyme (193L) [24] in 200 mM NaCl solutions of TIP3P water [25]. We also performed simulations using pure solvent. The simulations used the GROMACS suite of programs [26] and the AMBER03w force field [27,28]. We chose these proteins to compare with Grishaev *et al.* [11], who measured scattering intensities up to $q = 1$ Å^{-1} using an aqueous solution of 150 mM

NaCl, 40 mM Na acetate, and small amounts of NaN_3 and DTT. During the 40-ns simulations, we saved 20 000 frames at 2-ps intervals for scattering calculations. We used a rhombic dodecahedron as simulation box, with a nearest image distance under periodic boundary conditions of 90 Å and 16 725, 16 607, and 16 329 water molecules; 63, 61, and 60 Na^+ ions; and 61, 61, and 68 Cl^- ions in the simulations of GB3, ubiquitin, and lysozyme, respectively. All C_α positions were restrained to the experimental structures, except for ubiquitin where we left the last four N-terminal residues free. We confirmed in independent simulations that our conclusions remain valid for completely free proteins.

III. RESULTS AND DISCUSSION

A. Scattering intensities

In Fig. 1, we compare the scattering intensities calculated using Eq. (1) to the experimental results of Ref. [11]. As in experiment, excluded volumes are calculated using a mass specific protein volume $v_m = 7.425 \times 10^{-4}$ l/g resulting in volumes $v = 7525, 10\,470$, and 17 488 Å^3 for GB3, ubiquitin, and lysozyme, respectively. Excluded volumes estimated by subtracting the volumes occupied by solvent from the R-sphere volumes, assuming bulk solvent density, deviate from these estimates by less than 2%.

The agreement between theory and experiment is excellent over the whole q range, despite limitations of the water model and a simpler solvent composition [15]. Different solvent models can lead to contrast differences, which in a first approximation, result in an additive constant to $I(q)$. Such constants also account for residual dark currents and, in the range $q \lesssim 1$ Å^{-1}, for uncertainties in the scaling of intensities in the buffer subtraction. For the experimental data used here, the relative statistical error of the scaling factors was estimated as $\sigma = 0.01$ [11]. To account for these effects, we add constants to the scattering intensities shown in Fig. 1. We could also account for these constants by changing the buffer scaling by $\sim 1.4\sigma$, $\sim 1.2\sigma$, and $\sim 0.4\sigma$ for GB3, ubiquitin, and lysozyme, respectively. In the WAXS regime, the intensities of all three proteins have a peak located at $q \approx 1.4$ Å^{-1} with a shoulder at $q \approx 1.7$ Å^{-1} (Fig. 2). In the SAXS regime (inset), $\ln I(q)/I(0)$ as a function of q^2 shows excellent linear Guinier behavior. From the slopes of the linear fits we obtain radii of gyration of $R_g \approx 11.5 \pm 0.1$ Å (GB3), 13.0 ± 0.1 Å (ubiquitin), and 14.6 ± 0.1 Å (lysozyme), which are all in excellent agreement with the experimental radii $R_g \approx 11.3$, 12.9, and 14.5 Å, respectively, determined from the measured data [11] of Fig. 1.

B. Pair-distance distribution functions

Having established and validated our analysis method, we now study the electron PDDFs of proteins GB3, ubiquitin, and lysozyme, calculated directly using Eq. (5). The PDDFs in Fig. 3 exhibit several distinct peaks, common to the three proteins investigated here, and consistent with the characteristic distances of their secondary structure elements. Similar peak positions were reported for the PDDF of hemoglobin, obtained from high-precision scattering intensity measurements up to $q \approx 3$ Å^{-1} and inverse Fourier transformation [29]. For reference, we include curves for the proteins in vacuum,

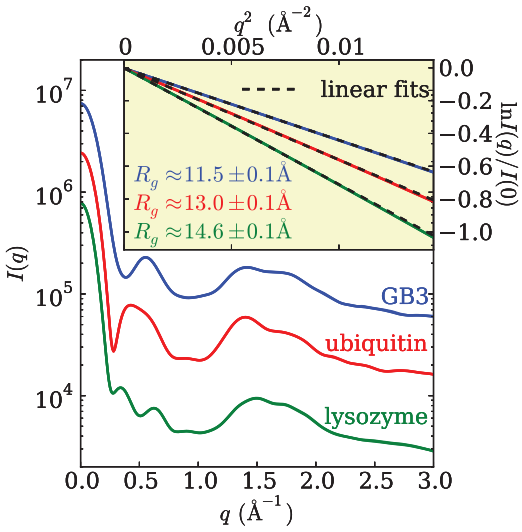

FIG. 2. (Color online) Calculated scattering intensities are independent of the radius R of the observation sphere for proteins GB3 (blue), ubiquitin (red), and lysozyme (green), top to bottom in both main plot and inset. $I(q)$ curves for $R = 30$, 35, and 40 Å are indistinguishable (for clarity, intensities for GB3, ubiquitin, and lysozyme are scaled by factors of 10, 2, and 0.02, respectively). Inset: Guinier plot together with linear fits, $\ln I(q)/I(0) \approx -q^2 R_g^2/3$ (dashed black lines), that give the radii of gyration R_g.

rescaled to have the same normalization $\int p(r)dr = I(0)$ as the corresponding solution $p(r)$. The electron density contrast of proteins against aqueous solution is $\sim 1/4$ of the contrast of proteins against vacuum, leading to a rescaling factor of $\sim 1/4^2$. We find that the vacuum PDDFs appear much smoother with less distinct features. The reason is that in vacuum, $p(r)$ is dominated by the shape contribution due to the large contrast. The fine structure, whose signal strength is less affected by the contrast, effectively sits on top of this shape signal and becomes relatively enhanced in experiments with nearly matched contrast between macromolecule and solvent.

Figure 3 also shows PDDFs obtained by indirect Fourier transforms of the scattering intensities of Fig. 2 over limited q ranges from $q = 0$ up to $q_{\max} = 1, 2$, and 3 Å^{-1} performed with GNOM by using standard procedures [30,31]. The maximum PDDF ranges of 33 (GB3), 35 (ubiquitin), and 45 Å (lysozyme) were determined iteratively by matching the back-transformed intensity with the original data. The agreement improves for larger q ranges, and for $q_{\max} = 3$ Å^{-1} most features are well reproduced.

C. Excluded volume term

The excluded volume term in Eq. (1) is necessary to estimate molecular weight from SAXS and SANS experiments [32] and useful to minimize bulk solvent contributions at wider angles [33]. For this reason, the latter is especially important when comparing experimental and/or theoretical data sets using solvents that are not exactly identical. In Figs. 4(g) and 4(h) we show $p(r)$ and $I(q)$ for ubiquitin with and without the excluded volume term. Without this term, $I(q)$ becomes negative at $q \approx 2$ Å^{-1}, the location of the peak in buffer scattering [Fig. 4(b) and gray vertical line]. Correspondingly,

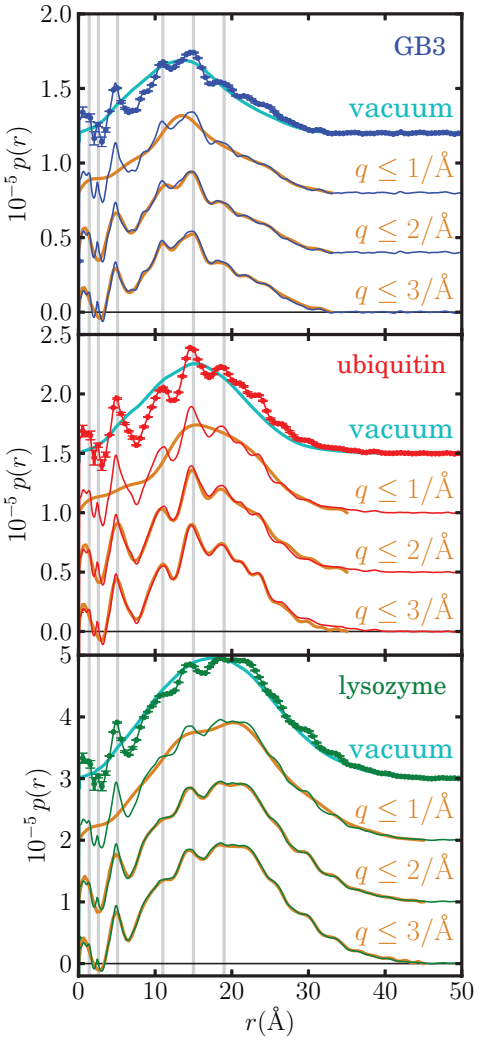

FIG. 3. (Color online) Pair-distance distribution functions for GB3 (blue), ubiquitin (red), and lysozyme (green) in 200 mM NaCl solution calculated directly from MD simulations using Eq. (5) (offset for clarity; standard errors from block averages); calculated indirectly using Fourier transforms (orange lines) of the simulation $I(q)$ in Fig. 1 for truncated q ranges from zero to $q_{\max} = 1,\ 2$, and 3 Å^{-1}; and calculated directly using Eq. (5) for the same simulation ensembles of protein structures, but stripped of solvent (vacuum, cyan; scaled). Peak positions in the PDDFs are consistent for all proteins, as indicated by vertical gray lines. By definition, PDDFs are normalized to $I(0)$.

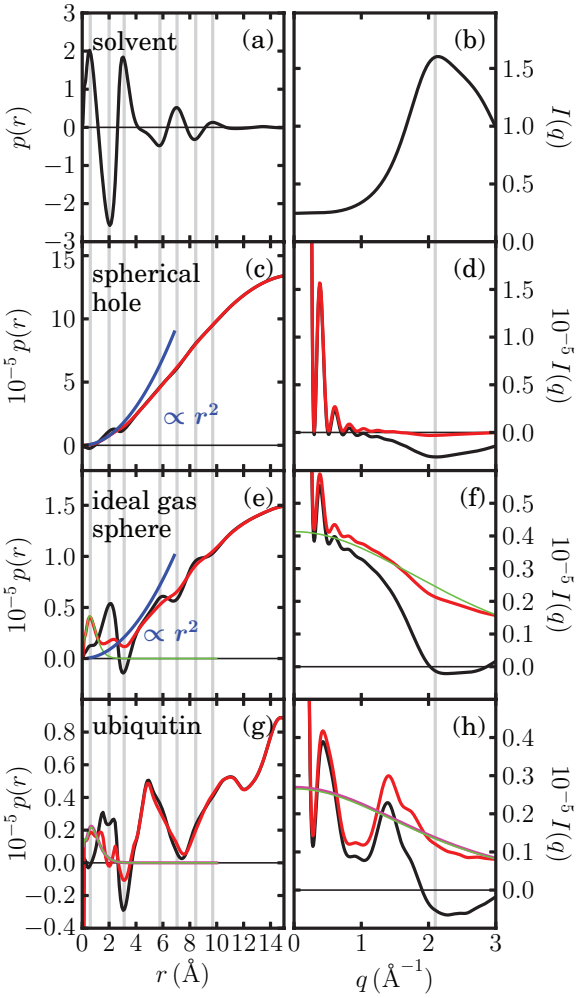

FIG. 4. (Color online) Effect of the excluded volume terms in Eqs. (1) and (5). Vertical gray lines indicate peak positions of the excess bulk solvent $p(r)$ [Eq. (6)] (a) and $I(q)$ [Eq. (2)] (b). $p(r)$ (left) and $I(q)$ (right) for a spherical hole with a radius of 15 Å in bulk solvent [(c), (d)], the same hole filled with an ideal gas of bulk solvent atoms mimicking protein electron density [(e), (f)], and ubiquitin in solution [(g), (h)] with and without the excluded volume term (red and black). For small distances, $p(r)$ grows like r^2 (blue). With the excluded volume term, the intra-atom scattering (green) shows as a positive peak at small distances in $p(r)$ and as a long-wavelength intensity contribution in $I(q)$. For ubiquitin, these contributions agree well with results for ubiquitin in vacuum (magenta).

$p(r)$ is negative without the excluded volume term at $r \approx 3$ Å, close to the peak in the bulk solvent $g(r)$. Such negative values of the intensity stem from oversubtracted buffer signal. This effect is minimized by the excluded volume term.

We illustrate the usefulness of the excluded volume term with the simple example of a spherical hole of volume V cut out of bulk solvent. The partial scattering intensity difference between sample and buffer is $\Delta I_{ij}(q) = -\delta_{ij}\rho_i V + \rho_i\rho_j V^2 \int \{g_{ij}(r)p_R(r) - 8\pi r^2[g_{ij}(r) - \gamma_{ij}]/V\} \sin(qr)/(qr)dr$. The calculated $I(q)$ and $p(r)$ match those for a sphere to a good approximation, as shown in Figs. 4(c) and 4(d). However, without the excluded volume term there are noticeable deviations at $q \approx 2$ Å^{-1}, where $I(q)$ becomes negative after buffer subtraction. Correspondingly, there are significant contributions to $p(r)$ at small r values, where $p_R(r) \approx 4\pi r^2/V$, that carry the signature of the bulk solvent. Adding back bulk solvent intensity [Eq. (2)] proportional to V, we approximately obtain the partial scattering intensity of the hole filled with an ideal gas of solvent atoms, $\rho_i\rho_j V^2 \int p_R(r) \sin(qr)/(qr)dr$. This expression only contains contributions due to interatom scattering, is independent of the bulk solvent structure, and becomes exact for increasing hole size.

To mimic the contrast of a protein in solution, we fill the hole with an ideal gas of solvent atoms with densities $X\rho_i$, where $X = 4/3$ approximates the electron density ratio of protein and solvent. We therefore add to the

difference scattering intensity of the hole contributions due to intra-atom scattering $X\rho_i V$ and interatom scattering $(X^2 - 2X)\rho_i\rho_j V^2 \int p_R(r)\sin(qr)/(qr)dr$ [Figs. 4 (e) and 4(f)]. The excluded volume term minimizes bulk solvent contributions and restores the intra-atom scattering contributions of the ideal-gas sphere. The latter shows up as a peak at small distances $r \lesssim 2$ Å in $p(r)$ and as a long-wavelength scattering intensity causing the offset of $I(q)$ with respect to zero. For ubiquitin, the restored intra-atom scattering contributions agree well with those of ubiquitin in vacuum [Figs. 4(g) and 4(h)].

Even if the excluded volume term is not applied and another convention for buffer subtraction is chosen, bulk solvent $I(q)$ [Eq. (2)] and $p(r)$ [Eq. (6)] can be used to account for uncertainties in the absolute scaling of intensities or in macromolecular concentration. At a protein concentration of $c = 10$ mg/ml, proteins occupy $cv_m \approx 0.7\%$ of the sample volume, where $v_m = 7.425 \times 10^{-4}$ l/g is the mass specific protein volume. Correspondingly, a ∼0.7% error in scaling corresponds to a 100% effective change of the excluded volume or concentration with a big effect on the difference signal, as shown in Fig. 4. Simulations do not suffer from these experimental uncertainties and make strong predictions in the wide-angle regime, which can be used to refine the experimental buffer subtraction.

IV. CONCLUSIONS

The PDDFs contain all structural information probed in elastic x-ray scattering experiments on macromolecules in solution, and are remarkably rich in structural features for proteins. With inverse Fourier transforms most structural features can be recovered from scattering intensities measured up to $q = 3$ Å^{-1}, providing guidance for the design of future scattering experiments. The quantification of the structural information content is especially important for time-resolved scattering experiments [34,35], where the signatures of structural changes tend to be small, about 1% in relative intensities [7]. For macromolecules in nonhydrogenous solution, we expect that subnanometer resolution can also be achieved with neutron scattering measurements probing the wide-angle regime.

The discriminatory power of SAXS and WAXS experiments relies on accurately and precisely measured scattering intensities, on error models that go beyond photon counting statistics, and on structural models making strong predictions about the measured intensities. Then, we can expect to be able to resolve structural differences despite qualitative similarities in $I(q)$ [36].

Our method for the calculation of $p(r)$ and $I(q)$, and previous methods to calculate $I(q)$ [14,15], are not restricted to biological macromolecules in solution, but can be applied to any simulation of a localized inhomogeneity in a fluid where the correlation length is of the order of the box size. Examples are nonbiological macromolecules and nanostructures (e.g., polymers, fullerenes, carbon nanotubes, nanocrystals), crystallization, vaporization, and condensation nuclei, macromolecular assemblies like micelles and polymer clusters, etc. Our method can also be applied to nonspherical observation volumes by effectively embedding them into an ideal-gas R sphere.

ACKNOWLEDGMENTS

We thank Dr. Alexander Grishaev and Dr. Ad Bax for providing the experimental data shown in Fig. 1 of this article. We also thank Dr. Alexander Grishaev and Dr. Philip Anfinrud for many helpful discussions. This work was supported by the Intramural Research Program of the National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, and utilized the biowulf and helix systems at the National Institutes of Health.

APPENDIX A: SPHERE-SPHERE INTERSECTION

$S_R(x,r)$ in Eq. (4) is the area of a sphere of radius r contained within another sphere of radius R, with their centers separated by a distance x. For $x^2 \leqslant R^2 - r^2$, we have

$$S_R(x,r) = \begin{cases} 4\pi r^2 & \text{for} \quad x \leqslant R - r \\ 4\pi r^2 - 2\pi r h_R^+(x,r) & \text{otherwise}; \end{cases} \tag{A1}$$

and for $x^2 > R^2 - r^2$,

$$S_R(x,r) = \begin{cases} 2\pi r h_R^-(x,r) & \text{for} \quad r - R < x < r + R \\ 0 & \text{otherwise}, \end{cases} \tag{A2}$$

where

$$h_R^{\pm}(x,r) = \pm\frac{(x \pm r)^2 - R^2}{2x}. \tag{A3}$$

The PDDF $p_R(r)$ of a structureless sphere of radius R, normalized to one, can then be expressed as

$$\begin{aligned} p_R(r) &= \frac{4\pi}{V^2}\int_0^R x^2 S_R(x,r)dx \\ &= \frac{3r^5}{16R^6} - \frac{9r^3}{4R^4} + \frac{3r^2}{R^3} \quad \text{for} \quad r \leqslant 2R, \end{aligned} \tag{A4}$$

where $V = 4\pi R^3/3$ is the volume of the sphere.

APPENDIX B: ELECTRON PAIR-DISTANCE DISTRIBUTION FUNCTION CALCULATION

The PDDF $p(r)$ can be obtained from the scattering intensity $I(q)$ by a sine transform, i.e.,

$$p(r) = \frac{2}{\pi}\int_0^\infty I(q) j_0(qr)(qr)^2 dq\,, \tag{B1}$$

where $j_0(x) = \sin(x)/x$ is the spherical Bessel function of order zero. We now express $p(r)$ and $I(q)$ using partial distance distribution functions $p_{ij}(r)$ and partial intensities $I_{ij}(q)$, i.e.,

$$p_{ij}(r) = \frac{2}{\pi}\int_0^\infty f_i(q) f_j(q) I_{ij}(q) j_0(qr)(qr)^2 dq, \tag{B2}$$

where $f_i(q)$ is the form factor of particle species i. Using Debye's formula [16] and partial distance histograms $H_{ij}(r)$, we write the partial scattering intensities as

$$I_{ij}(q) = N_i\delta_{ij} + \int_0^\infty H_{ij}(r) j_0(qr)dr. \tag{B3}$$

Inserting this expression for $I_{ij}(q)$ into Eq. (B2), we obtain

$$\Delta p_{ij}(r) = N_i\delta_{ij}c_{ii}(r,r') + \int_0^\infty H_{ij}(r')c_{ij}(r,r')dr', \tag{B4}$$

where the

$$c_{ij}(r,r') = \frac{2}{\pi}\int_0^\infty f_i(q) f_j(q) j_0(qr') j_0(qr)(qr)^2 dq \quad \text{(B5)}$$

are PDDFs of two spherical scatterers of species i and j separated by a distance r'. Using the atomic form factors

$$f_i(q) = 4\pi \int_0^\infty \varrho_i(x) x^2 j_0(qx) dx, \quad \text{(B6)}$$

expressed in terms of radially symmetric electron densities $\varrho_i(r)$, we obtain

$$c_{ij}(r,r') = -\frac{r}{r'}\frac{\pi}{16}\int_0^\infty dy\, \varrho_i(y) y \int_0^\infty dx\, \varrho_j(x) x \times \sum_{s_x,s_y,s'=\pm 1} s_x s_y s' |r + s_x x + s_y y + s' r'|. \quad \text{(B7)}$$

For a single particle, i.e., $i = j$ and in the limit $r' \to 0$, we obtain

$$c_{ii}(r,0) = \frac{(4\pi)^2}{2} r \int_0^\infty dy\, \varrho_i(y) y \int_{|y-r|}^{y+r} dx\, \varrho_i(x) x\,. \quad \text{(B8)}$$

For x-ray scattering, form factors are well approximated by sums of Gaussians [18], i.e.,

$$f_i(q) = \sum_\nu a_{i\nu} e^{-q^2 b_{i\nu}} \quad \text{(B9)}$$

with real valued parameters $a_{i\nu}$ and $b_{i\nu} \geqslant 0$. These form factors appear as products in the expressions for scattering intensities, $f_i(q) f_j(q) = \sum_{\nu,\mu} a_{i\nu} a_{j\mu} e^{-q^2(b_{i\nu}+b_{j\mu})}$. Equation (B5) becomes $c_{ij}(r,r') = \sum_{\nu,\mu} c_{ij}^{\nu\mu}(r,r')$. For $b_{i\nu} + b_{i\mu} \neq 0$ we obtain

$$c_{ii}^{\nu\mu}(r,0) = r^2 \frac{a_{i\nu} a_{i\mu}}{2\sqrt{\pi (b_{i\nu} + b_{i\mu})^3}} \exp\left[-\frac{r^2}{4(b_{i\nu} + b_{i\mu})}\right]$$

and for $b_{i\nu} = b_{i\mu} = 0$ we obtain $c_{ii}^{\nu\mu}(r,0) = a_{i\nu} a_{i\mu} \delta(r)$. For $b_{i\nu} + b_{j\mu} \neq 0$ we obtain

$$c_{ij}^{\nu\mu}(r,r') = \frac{r}{r'}\frac{a_{i\nu} a_{j\mu}}{2\sqrt{\pi (b_{i\nu} + b_{j\mu})}}\left[\exp\left(-\frac{(r-r')^2}{4(b_{i\nu} + b_{j\mu})}\right) - \exp\left(-\frac{(r+r')^2}{4(b_{i\nu} + b_{j\mu})}\right)\right] \quad \text{(B10)}$$

and for $b_{i\nu} = b_{j\mu} = 0$ we obtain $c_{ij}^{\nu\mu}(r,r') = a_{i\nu} a_{j\mu} \delta(r - r')$.

[1] H. N. Chapman, Nat. Mater. **8**, 299 (2009).
[2] S. C. M. Teixeira *et al.*, Chem. Phys. **345**, 133 (2008).
[3] F. Förster, B. Webb, K. A. Krukenberg, H. Tsuruta, D. A. Agard, and A. Sali, J. Mol. Biol. **382**, 1089 (2008).
[4] R. P. Rambo and J. A. Tainer, Curr. Opin. Struct. Biol. **20**, 128 (2010).
[5] L. Makowski, D. J. Rodi, S. Mandava, S. Devarapalli, and R. F. Fischetti, J. Mol. Biol. **383**, 731 (2008).
[6] S. Yang, S. Park, L. Makowski, and B. Roux, Biophys. J. **96**, 4449 (2009).
[7] H. S. Cho, N. Dashdorj, F. Schotte, T. Graber, R. Henning, and P. Anfinrud, Proc. Natl. Acad. Sci. USA **107**, 7281 (2010).
[8] D. I. Svergun, S. Richard, M. H. J. Koch, Z. Sayers, S. Kuprin, and G. Zaccai, Proc. Natl. Acad. Sci. USA **95**, 2267 (1998).
[9] F. Merzel and J. C. Smith, Proc. Natl. Acad. Sci. USA **99**, 5378 (2002).
[10] D. Svergun, C. Barberato, and M. H. J. Koch, J. Appl. Crystallogr. **28**, 768 (1995).
[11] A. Grishaev, L. Guo, T. Irving, and A. Bax, J. Am. Chem. Soc. **132**, 15484 (2010).
[12] J. J. Virtanen, L. Makowski, T. R. Sosnick, and K. F. Freed, Biophys. J. **8**, 2061 (2011).
[13] F. Merzel and J. C. Smith, Acta Crystallogr., Sect. D: Biol. Crystallogr. **58**, 242 (2002).
[14] T. Oroguchi, H. Hashimoto, T. Shimizu, M. Sato, and M. Ikeguchi, Biophys. J. **96**, 2808 (2009).
[15] S. Park, J. P. Bardhan, B. Roux, and L. Makowski, J. Chem. Phys. **130**, 134114 (2009).
[16] P. Debye, J. Phys. Colloid Chem. **51**, 18 (1947).
[17] T. Oroguchi and M. Ikeguchi, J. Chem. Phys. **134**, 025102 (2011).
[18] E. N. Maslen, A. G. Fox, and M. A. O'Keefe, in *International Tables for Crystallography*, edited by A. J. C. Wilson and E. Prince (Kluwer Academic Publishers, Dordrecht/Boston/London, 1999), Vol. C, p. 548.
[19] H. B. Stuhrmann, Acta. Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. **27**, 297 (1970).
[20] B. G. Levine, J. E. Stone, and A. Kohlmeyer, J. Comput. Phys. **230**, 3556 (2011).
[21] B. Lindner and J. C. Smith, Comput. Phys. Commun. **183**, 1491 (2012).
[22] J. P. Derrick and D. B. Wigley, J. Mol. Biol. **243**, 906 (1994).
[23] G. Cornilescu, J. L. Marquardt, M. Ottiger, and A. Bax, J. Am. Chem. Soc. **120**, 6836 (1998).
[24] M. C. Vaney, S. Maignan, M. Riès-Kautt, and A. Ducruix, Acta Crystallogr., Sect. D: Biol. Crystallogr. **52**, 505 (1996).
[25] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, J. Chem. Phys. **79**, 926 (1983).
[26] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, J. Chem. Theory Comput. **4**, 435 (2008).
[27] R. B. Best and J. Mittal, J Phys. Chem. B **114**, 14916 (2010).
[28] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang, and P. Kollman, J. Comput. Chem. **24**, 1999 (2003).
[29] X. Hong and Q. Hao, Appl. Phys. Lett. **94**, 083903 (2009).
[30] O. Glatter, J. Appl. Crystallogr. **10**, 415 (1977).
[31] D. I. Svergun, J. Appl. Crystallogr. **25**, 495 (1992).
[32] O. Kratky, G. Porod, and L. Kahovec, Z. Elektrochem. **55**, 53 (1951).
[33] O. Kratky and O. Glatter, *Small Angle X-Ray Scattering* (Academic Press, London, 1982).
[34] M. Cammarata, M. Levantino, F. Schotte, P. A. Anfinrud, F. Ewald, J. Choi, A. Cupane, M. Wulff, and H. Ihee, Nat. Methods **5**, 881 (2008).
[35] L. Pollack, M. W. Tate, A. C. Finnefrock, C. Kalidas, S. Trotter, N. C. Darnton, L. Lurio, R. H. Austin, C. A. Batt, S. M. Gruner, and S. G. J. Mochrie, Phys. Rev. Lett. **86**, 4962 (2001).
[36] B. Zagrovic and V. S. Pande, J. Am. Chem. Soc. **128**, 11742 (2006).