Simulation-Based Fitting of Protein-Protein Interaction Potentials to SAXS Experiments

Seung Joong Kim,* Charles Dumont,* and Martin Gruebele*†

*Department of Physics, and †Center for Biophysics and Computational Biology and Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois

ABSTRACT We present a new method for computing interaction potentials of solvated proteins directly from small-angle x-ray scattering data. An ensemble of proteins is modeled by Monte Carlo or molecular dynamics simulation. The global x-ray scattering of the whole model ensemble is then computed at each snapshot of the simulation, and averaged to obtain the x-ray scattering intensity. Finally, the interaction potential parameters are adjusted by an optimization algorithm, and the procedure is iterated until the best agreement between simulation and experiment is obtained. This new approach obviates the need for approximations that must be made in simplified analytical models. We apply the method to lambda repressor fragment 6-85 and *fyn*-SH3. With the increased availability of fast computer clusters, Monte Carlo and molecular dynamics analysis using residue-level or even atomistic potentials may soon become feasible.

INTRODUCTION

Small angle x-ray scattering (SAXS) is a convenient tool for determining protein-protein interaction potentials in solution. A major driving force of this work has been the need for determining ideal conditions for protein crystallization. Thus, the focus has been on the effect of the concentration of precipitation agents and cosolvents (1,2).

Two additional areas could benefit greatly from the effective potentials provided by SAXS studies. One is the study of solvation shells around proteins. Neutron scattering, NMR spectroscopy, simulation, and terahertz spectroscopy have shown that solvent shells of substantial thickness exist around proteins (3–6). Dynamical solvation effects studied by terahertz spectroscopy extend to >10 Å from the protein surface (7). Protein-protein interactions are mediated by such solvent shells, and thus contain information about the solvent shells when measured at sufficiently high concentrations. The other area is the study of transient protein aggregation. Very rapidly folding proteins have folding timescales comparable to the lifetime of transient aggregates (8,9). Such transient aggregates can nucleate irreversible aggregation (10,11), a process linked with numerous diseases. Protein-protein interaction potentials play a key role in defining how easily such nuclei form.

Effective interaction potentials are currently extracted from SAXS data with the aid of analytical approximations to speed up the calculation (1). The random phase approximation treats each protein molecule as an independent scatterer characterized by a form factor. The form factor can be obtained approximately by extrapolating SAXS measurements to infinite dilution (12). The observed scattering intensity is then assumed to be a product of the form factor and a scattering factor, an approximation strictly valid over the full range of scattering angles only for dilute particles. From the scattering factor, a radial pair distribution function and corresponding radial effective potential are obtained. Square-well and Yukawa potentials are used because they have simple Fourier transforms (2). The commonly used DLVO form consists of a hard-sphere cutoff, and two Yukawa potentials ($\sim\exp[-(r - r_0)/\delta]/r$) for long-range repulsion and short-range attraction between proteins.

Increases in computing power enable a more direct approach, which we introduce here. Simulation of multiprotein ensemble dynamics is followed by evaluation of the x-ray scattering of the whole ensemble. Iteration can then be used to refine force fields ''on the fly'' without any low-concentration approximations or scattering analysis approximations.

Fig. 1 outlines our approach. We first simulate the dynamics of an ensemble of dozens to hundreds of model proteins that interact via an adjustable interaction potential. Either Monte Carlo or molecular dynamics simulations are used to sample configurations of the ensemble. We then calculate the global x-ray scattering intensity of the entire model ensemble at each configuration, eliminating the need for low-concentration or random phase approximations. The resulting series of scattering intensities is averaged to obtain the steady-state SAXS intensity as a function of scattering angle. An optimization algorithm compares the computed signal with the experimental signal, and modifies the adjustable interaction potential for the next round of simulation. The process repeats iteratively until the experimental data is matched with the smallest least-squares deviation. Any form of potential can thus be fitted exactly for polydisperse model particles at any concentration.

In this first application, we determine isotropic interaction potentials, and hence assume spherical model protein monomers. Aggregates can have any shape made from these monomer building blocks, up to the size of the box used for

Submitted October 8, 2007, and accepted for publication January 30, 2008.

Address reprint requests to Martin Gruebele, Dept. of Chemistry, University of Illinois at Urbana-Champaign, 600 S. Mathews Ave., Urbana, IL 61801. Tel.: 217-333-1624; E-mail: gruebele@scs.uiuc.edu.

Editor: Jill Trewhella.

0006-3495/08/06/4924/08 $2.00

doi: 10.1529/biophysj.107.123240

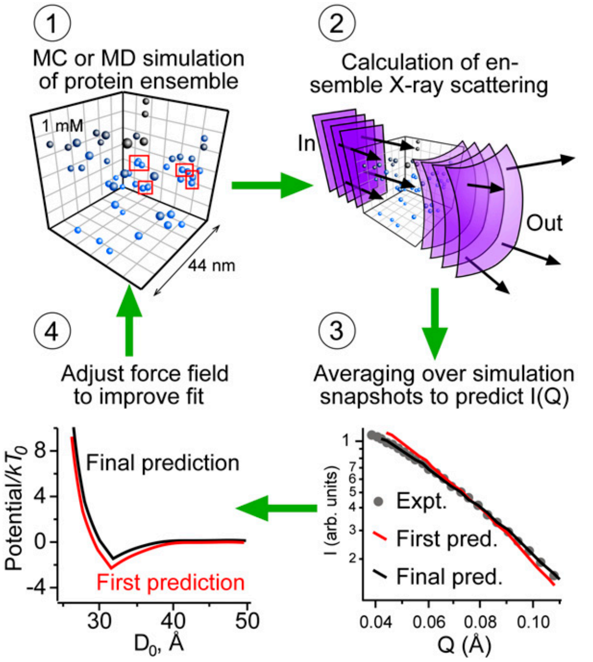
1
MC or MD simulation of protein ensemble
1 mM
44 nm
2
Calculation of ensemble X-ray scattering
In
Out
3
Averaging over simulation snapshots to predict I(Q)
4
Adjust force field to improve fit
Potential/kT_0
8
4
0
-4
Final prediction
First prediction
30
40
50
D_0, Å
I (arb. units)
1
7
6
5
4
3
2
Expt.
First pred.
Final pred.
0.04
0.06
0.08
0.10
Q (Å)

FIGURE 1 Method for extracting the protein-protein potential from SAXS data. In step 1, a protein ensemble of up to 100 molecules is simulated by Monte Carlo or molecular dynamics. In step 2, the exact x-ray scattering for the model ensemble is evaluated at each simulation snapshot. In step 3, the average x-ray scattering curve is obtained and compared with experimental data. In step 4, the interaction potential is adjusted by steepest descent for the next round of simulation.

simulation, typically 20 monomer diameters or more. Thus the analysis must be truncated at large scattering angles, but it does not assume spherical aggregates or low monomer concentration. We illustrate the method by fitting experimental data for the two proteins $\lambda_{6\text{-}85}$ and *fyn*-SH3 to several potential models. The ethylene glycol-water solvent we use is similar to the one used in recent SAXS studies of folding kinetics (13). With the advent of interaction potentials based on sums of amino-acid-pair interactions, the simulation direct fitting approach could yield anisotropic interaction potentials in the near future, revealing potential aggregation sites or local changes in the protein solvation shell.

METHODS

Proteins

$\lambda^*_{6\text{–}85}$ is an 80-residue, five-helix globular protein of molecular mass 9.2 kDa (see Fig. 4, *inset*). The protein we used in SAXS experiments contained the mutations Tyr22Trp, Glu33Tyr, Gly46Ala, and Gly48Ala, engineered by site-directed mutagenesis (Stratagene Quickchange kit, La Jolla, CA) based on a wild-type plasmid donated by Terry Oas (14). *fyn*-SH3 is a predominantly β-sheet protein (molecular mass 9.3 kDa) (see Fig. 4, *inset*) with 78 residues and a tag of six histidine residues. The sequence, donated by Alan Davidson, has mutations Val1Ser, Val5Glu, Ala39Val, and Val55Phe (15).

Genes for the two proteins were inserted into the pET-15b vector, expressed in Rosetta TM (DE3) pLysS cells (Novagen, San Diego, CA), and grown in LB broth at 37°C for 8 h. After induction with isopropyl-β-D-thiogalactopyranoside at 25°C for 12 h, cells were lysed with a French press, and the supernatant was collected after centrifugation. Proteins were selectively bound to a nickel-agarose his-tag binding column (Pharmacia, Peapack, NJ) and eluted with a 250 mM imidazole buffer. The six-histidine tag of $\lambda_{6\text{–}85}$ was cleaved by thrombin (VWR, West Chester, PA), and additional purification was performed with Amicon 3 kDa and 30 kDa membranes (Fisher Scientific, Hampton, NH). *fyn*-SH3 was used with the his tag. The identity of $\lambda_{6\text{–}85}$ and *fyn*-SH3 was confirmed by electrospray ionization mass spectroscopy and their purity by sodium dodecyl sulfate polyacrylamide gel electrophoresis.

Final protein concentrations in buffers used for experiments were determined by near-UV absorption spectroscopy at 280 nm of the tryptophan and tyrosine residues as described by Edelhoch (16). We have found this procedure to yield similar results in aqueous and aqueous-osmolyte buffers. We estimate a relative accuracy of $\pm 1\%$ for dilution series from the same sample, and an absolute accuracy of $\sim\pm 5\%$. Results are rounded to the nearest 10 μM.

SAXS measurements

SAXS measurements were performed at the Biophysics Collaborative Access Team Beamline of the Advanced Photon Source at Argonne National Laboratory (Argonne, IL) (17). An Aviex CCD camera with an active area of $\sim 160 \times 80$ mm^2 (2084×1042 pixels, 78 μm gap between pixels), located 1.9 m from sample, was used to collect data in the scattering angle range of $Q = 4\pi\sin\theta/\lambda = 0.03\text{–}0.12$ Å^{-1}, at a nominal wavelength of 1 Å. Low concentration data for *fyn*-SH3 were also collected with a Pilatus CCD camera. The x-ray beam was collimated to a spot size of 300×130 μm^2 at the sample cell.

To reduce radiation damage, and to enable a direct comparison with our previous SAXS folding study of $\lambda_{6\text{–}85}$, we performed our experiments in a 45:55 vol % ethylene glycol/water buffer. The ionic strength was 50 mM phosphate at pH 7.0. The temperature in all experiments was -28 ± 1°C, cooled by a Neslab ULT-80DD recirculator. Steady-state SAXS data were collected in a UNISOKU sample cell with 80 μl volume and 50 μm thick sapphire windows. The exposure time was 500 ms for $\lambda_{6\text{-}85}$, and 300 ms for *fyn*-SH3 (four frames of 200 ms on the Pilatus detector), based on extensive exposure/concentration tests for protein damage. We measured steady-state SAXS data of $\lambda_{6\text{-}85}$ up to 2.92 mM, and of *fyn*-SH3 up to 1.68 mM, without any visible aggregation at room temperature or at -28°C. Each sample was filtered with a 0.2 μM pore syringe filter (Corning, Toledo, OH) before use. The raw data were angle-averaged with logarithmic weighting in Q, and a reference buffer curve was subtracted.

Interaction potentials

To enable Monte Carlo or molecular dynamics simulations, a protein-protein interaction potential has to be chosen. We tested several pairwise-additive isotropic interaction potentials not easily fitted by analytical methods. At short distance, an r^{12} repulsive term was used instead of the commonplace hard sphere wall:

$$U_{\rm L} = \varepsilon\left[\left(\frac{D_0}{r}\right)^{12} - 2\right] \quad (r < D_0). \tag{1}$$

Past the potential minimum at D_0, exponential, Gaussian and Yukawa forms were used in various combinations to model both attractive and repulsive-attractive potentials:

$$\left.\begin{aligned} U_{\rm E} &= -\varepsilon\exp\left[-\left(\frac{r-D_0}{\delta}\right)\right] \\ U_{\rm G} &= -\varepsilon\exp\left[-\left(\frac{r-D_0}{\delta}\right)^2\right] \\ U_{\rm Y} &= -\varepsilon\left(\tfrac{D_0}{r}\right)\exp\left[-\left(\frac{r-D_0}{\delta}\right)\right] \end{aligned}\right\}(r > D_0), \tag{2}$$

where ε is the potential depth, D_0 is the center-of-mass distance between proteins where the repulsive potential wall begins, and δ is the attractive potential range.

The softer than hard-sphere potential wall, not easily amenable to the analytical treatment, highlights the fact that no reference potential assumptions need to be made. In our first application, we assumed isotropically interacting particles and pairwise additive potentials, although nonspherical particles and *n*-body potentials could be implemented in the future because our approach requires only that the total potential energy for the multiprotein system can be evaluated.

Configurational sampling

To avoid the need for low-concentration approximations, we sample a whole protein ensemble much larger than the typical aggregate size. Protein configurations were sampled by two methods: Metropolis Monte Carlo sampling (MMC), which illustrates computation of the scattering curve from a thermal simulation, and Langevin molecular dynamics (LMD), to illustrate computation of scattering curves from real-time dynamics simulations. In both approaches, we distributed $n = 25$–100 spherically symmetric model protein particles in a spherical or cube-shaped volume, the latter with periodic boundary conditions. The diameter of the simulation volume was determined by the experimental protein concentration. To reduce oscillatory boundary artifacts in the SAXS calculation, the diameter of the volume was varied randomly about the average. Test runs with up to 20,000 protein particles confirmed that full convergence over the desired range of Q could be achieved rapidly with 25 particles for *fyn*-SH3 and with 100 particles for λ_{6-85} over the full experimental concentration range.

For MMC sampling, we started out with a random distribution of particles. Single particles were then chosen at random, and moved by random displacements inside the spherical volume. Each move was accepted or rejected based on the Metropolis criterion by computing the change in total energy, ΔE (18). When the net energy change was negative, the move was accepted, whereas a positive energy change was accepted with a probability of $\exp(-\Delta E/k_BT)$. Equilibration of the total energy to within the statistical noise typically required $50n$ moves for λ_{6-85}. This sampling was repeated until the scattering intensity (see below) was a smooth function of Q. The longest runs provide estimated error bounds for the computed scattering curve.

For molecular dynamics sampling in real time, we used an LMD approach in a cubic volume with periodic boundary conditions. Each protein particle was subject to a vectorial force resulting from the other protein particles, and to a random Brownian force simulating the implicit solvent dynamics. In addition, the Brownian motion was countered by a vectorial damping term. Inertial forces were neglected, resulting in $3n$ equations of motion

$$-\frac{\partial V}{\partial r_{\mathrm{i,m}}} - \gamma\,\frac{dr_{\mathrm{i,m}}}{dt} + \xi_{\mathrm{i,m}}(t) = 0. \quad (3)$$

For nonspherical particles subject to anisotropic interaction potentials, an additional set of $3n$ equations for rotational diffusion would have to be solved, but no additional complications are introduced by our approach. In Eq. 3, V is the interaction potential summed over all protein pairs (Eqs. 1 and 2). Protein particle m is at position $\mathbf{r}_{\mathrm{m}} = (r_{\mathrm{x,m}}, r_{\mathrm{y,m}}, r_{\mathrm{z,m}}) \cdot \gamma = k_BT/D(T,P)$ denotes the velocity relaxation rate, which depends on the diffusion coefficient D, assumed independent of coordinate. $\xi_{\mathrm{i}}(t)$ is Gaussian white noise with zero mean, and a variance set to satisfy the Onsager fluctuation-dissipation theorem that relates ξ and γ (19). The equations of motion were integrated by a standard integrator using finite-difference derivatives (thus, Brownian noise or discontinuities in the potential derivative are not a problem). Derivatives with respect to a single particle, like the energy change ΔE, could be evaluated efficiently. The protein distribution was allowed to evolve to a mean particle deviation of at least 3.4 R_{g} before sampling the next configuration, to ensure that the scattering calculation did not needlessly sample very similar configurations.

Scattering signal

For each multiprotein configuration from the MMC or LMD simulations, we calculated the exact x-ray scattering by evaluating

$$F_{\mathrm{total}}(\mathbf{q}) = \sum_{\mathrm{m=1}}^{\mathrm{n}} F_{\mathrm{m}} e^{-\mathrm{i}\mathbf{q}\cdot\mathbf{r}_{\mathrm{m}}}, \quad (4)$$

where F_{m} is the scattering amplitude for particle m. Because we are determining isotropic interaction potentials here, we approximated each protein particle by a sphere and used the corresponding F_{m} (20,21). The assumption of individual spherical particles sets an upper limit on the Q values that can be fitted. A more realistic electron distribution based on diffraction data would have to be used if anisotropic potentials and large Q values are to be used in fitting. Equation 4 treats the scattering of the model protein ensemble exactly at any concentration and for any aggregation state that is small compared to the size of the simulation box. Thus, no extrapolations to dilute samples or analytical approximations usually needed for polydisperse systems need to be made. The total scattering intensity is obtained from

$$I(\mathbf{q}) = |F_{\mathrm{total}}(\mathbf{q})|^2 \quad (5)$$

and averaged over all configurations sampled by the simulations to yield the average SAXS scattering intensity, $I(Q)$, for direct comparison with experiment.

Approximately 100,000 configurations were averaged for each concentration to obtain a smooth $I(Q)$ for comparison with experiment. To minimize boundary effects and oscillations of the intensity at low Q below the experimental noise level, either a spherical volume was chosen, and its volume was changed randomly about the average value required by each protein concentration (22), or a spherical volume from the center of a periodic boundary condition box was chosen for the x-ray scattering calculation.

Data fitting

We fitted three potential parameters: potential depth, ε; potential range, δ; and potential wall, $D_0 \equiv 2R_0$. An efficient Levenberg-Marquardt optimization algorithm (23,24) was applied to fit the potential parameters to the experimental concentration-dependent scattering data. Minimal evaluation of $I(Q)$ is desirable because each concentration point requires a large number of MMC/LMD simulations to yield a smooth curve.

We also fitted a fourth parameter, the radius of gyration, R_{g}, of the model proteins, to account for the direct effect of particle size on the scattering data. R_0 measures monomer size from the point of view of the interaction potential, whereas R_{g} measures monomer size from the point of view of the scattering intensity. R_{g} is not entirely independent of R_0. For an ideal hard-sphere monomer, $R_{\mathrm{g}}/R_0 = \sqrt{3/5}$. Deviations from spherical shape, and a tapering of the electron density distribution due to solvation or a soft potential wall (as in Eq. 1), are both effectively accounted for by allowing deviations from this ratio. A large deviation would indicate that a better model for the monomeric proteins is needed.

RESULTS

Concentration-dependent SAXS of λ_{6-85}

Fig. 2 shows the concentration dependence of the scattering intensity as a function of Q for the λ_{6-85} Q33Y G46A G48A mutant. A Guinier plot (ln(I) vs. Q^2, not shown) deviates from linearity below $Q^2 = 0.006$ Å^{-2}, indicating some aggregation. Dilution of samples shows that this aggregation is reversible over the concentration range we studied. No deviations were observed at concentrations <100 μM or for Q up to 0.11 Å^{-1}, indicating that the spherical approximation

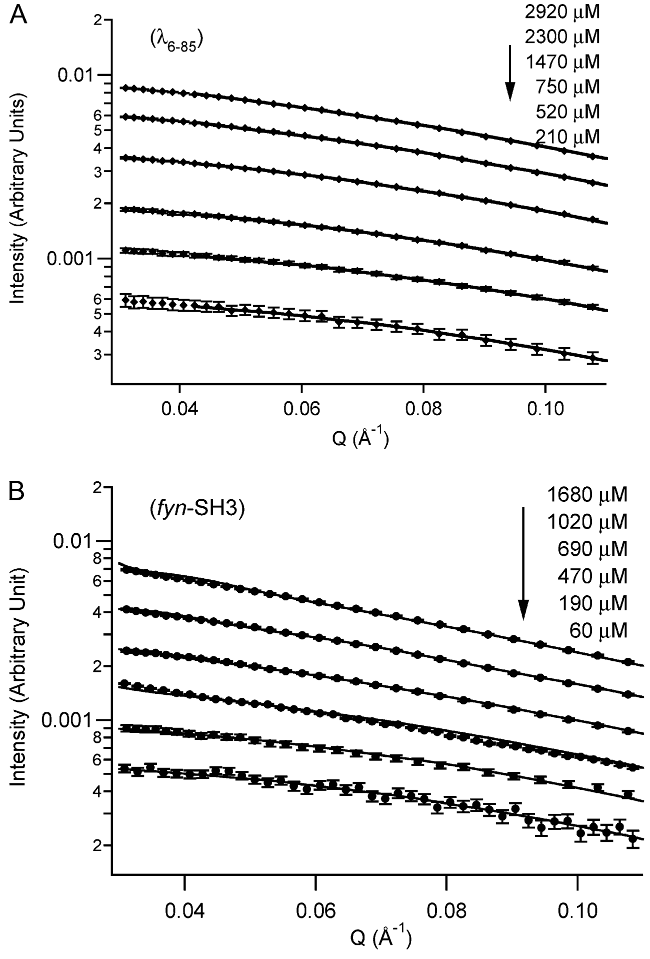

FIGURE 2 Scattering intensity versus magnitude of the scattering vector for λ_{6-85} (*A*) and *fyn*-SH3 (*B*). The lines going through the experimental data points are fits from Monte Carlo simulation.

for protein monomers is good for λ_{6-85} over the range of scattering angles considered here.

MMC fitting results for λ_{6-85}

Simulations were performed by the MMC method. The best fit to experimental data (Fig. 2 *A*) was obtained with a $U_L + U_E$ potential (Lennard-Jones wall, attractive exponential). The calculated radius of gyration is 13.52 Å, the potential depth is 1.5 kT_0, with 3.6 Å of potential range, and the potential wall beginning at $D_0 \approx 31.8$ Å (Table 1 and Fig. 3 *A*; $T_0 = 245$ K). A total of 100 proteins was used in 5000 Metropolis iterations to obtain equilibrated results for each configuration, and 100,000 configurations were sampled. As one might expect, two parameters of this fit are somewhat correlated, the potential range and depth.

LMD simulation for λ_{6-85}

We also performed an LMD simulation with the same potential as MMC at 2920 μM concentration, to confirm consistency of the MMC fitting results with molecular dynamics. We tested a range of different timescales (500 ns, 50 ns, 5 μs, 20 μs) for 25 proteins in a cube having periodic boundary condition. The resulting *I*(*Q*) is shown in Fig. 4 *A*, and agrees with the experimental data within sampling uncertainty. The sampling uncertainty of the molecular dynamics simulations is shown by the error bars. The timescale between successive configurations chosen for scattering calculations was estimated form the diffusion equation $\langle r^2 \rangle = 6Dt$ in 3-D, allowing the protein ensemble to move enough so that successive configurations were independent of one another.

Concentration-dependent SAXS of *fyn*-SH3

Fig. 2 *B* shows the concentration dependence *I*(*Q*) for *fyn*-SH3. The slope of a Guinier plot (not shown) deviates more strongly from linearity at low *Q* than for λ_{6-85}, indicating more extensive aggregation and a stronger interaction potential. As in the case of λ_{6-85}, the spherical monomer approximation works to the largest *Q* values for which data were collected.

Fitting results for *fyn*-SH3

Ensemble configurations were generated by MMC simulation. The best fit was obtained with a $U_L + U_Y + U_E$ potential (Lennard Jones repulsive wall, attractive Yukawa potential well and repulsive exponential potential). Potentials without a repulsive long-range interaction produced significantly worse fits ($\chi^2/\chi^2_{\text{optimal}}$>2). For the three-term potential, the calculated radius of gyration is 14.85 Å and the potential wall size is 42.0 Å. The attractive Yukawa potential depth is 11.2 kT_0 with a 1 Å range, The repulsive exponential potential depth is 7.5 kT_0, with a range of 2.0 Å, which results in a net potential depth of 3.65 kT (Table 2 and Fig. 3 *A*). We used 625,000 Metropolis iterations to obtain converged results, and 50,000 final configurations were sampled. Compared to λ_{6-85}, SH3 consistently produced fits with shorter range but deeper potential wells.

LMD simulation for *fyn*-SH3

We also performed an LMD simulation with the converged MMC potential at 1690 μM concentration, to confirm consistency of the MMC fitting results and the molecular dynamics simulations. Again, we tested a range of different timescales (from 50 ns to 5 μs) for 25 proteins in a cube having periodic boundary condition. The resulting *I*(*Q*) is shown in Fig. 4 *B*, and also agrees with the experimental data within sampling uncertainty. The timescale between successive configurations was chosen by the same criterion as for λ_{6-85}.

DISCUSSION

We have obtained interaction potentials for two proteins under identical buffer conditions by using the four-step procedure in Fig. 1. First, Monte Carlo or molecular dynamics

TABLE 1 Best fit of the λ_{6-85} SAXS data to a $U_L + U_E$ (r^{12} repulsive, exponential attractive) potential

Potential type	R_g (Å)	Potential depth ε (kT_0)	Potential range δ (Å)	Potential wall D_0 (Å)	RMSE
$U_L + U_E$	13.5 ± 0.2	1.5 ± 0.5	3.6 ± 0.5	31.8 ± 3.0	0.0073

Best fit	2920 μM	2300 μM	1470 μM	750 μM	520 μM	210 μM	Weighted average
RMSE	0.0050	0.0049	0.0040	0.0048	0.0072	0.013	0.0073

Also shown in the table are the root mean-square errors (RMSE) for the best fit at individual concentrations. All RMSEs of the fit lie within the experimental error. kT_0 corresponds to 245 K.

simulations of a model protein ensemble compute thermally averaged or time-averaged particle distributions for up to 100 protein particles. Next, x-ray scattering functions, $F_{\text{total}}(\mathbf{q})$, are computed directly for the whole ensemble. These are essentially exact for scattering angles corresponding to the size range from monomer particle to simulation box. In the third step, the resulting scattering intensity is computed without further approximations and then compared to SAXS data. In the last step, a least-squares algorithm refines the potential parameters, so that a new simulation can be started to iterate until the best fit is obtained. The best fits are summarized in Tables 1 and 2.

Although MMC sampling and MD simulations are computationally much more expensive than the analytical approximations commonly used, direct simulation provides a correct description of the scattering amplitude at any concentration, for any monomer size, and for any aggregate shape consistent with the model monomers and up to the size of the simulation box. Any functional form of the potential, rather than a perturbing potential added to a hard-sphere repulsion, can be fitted without additional effort simply by replacing the two-body interaction potential in the simulation.

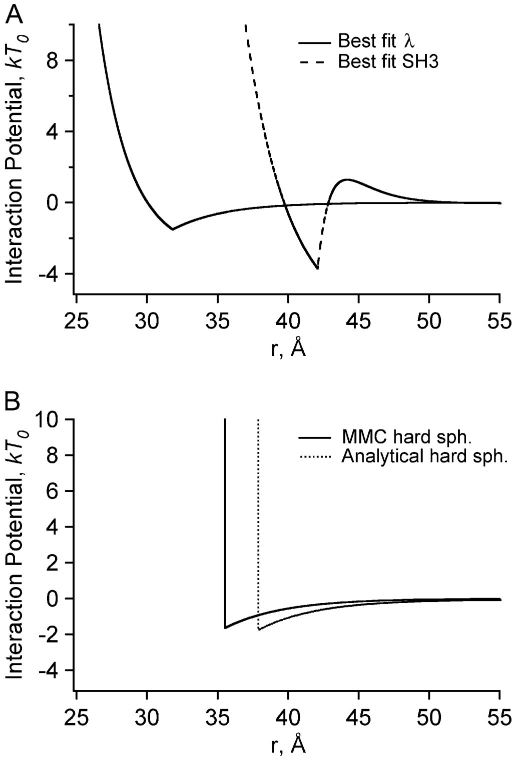

FIGURE 3 (*A*) Best-fit interaction potential for λ_{6-85} and *fyn*-SH3 (in 45% ethylene glycol buffer at −28°C). (*B*) Comparison of the MMC and analytical best-fit hard sphere plus Yukawa potentials for λ_{6-85}. The greatest variation between the three λ_{6-85} shown is in D_0. (MMC parameters: $D_0 =$ 35.5 Å, $\delta = 4.14$ Å, $\varepsilon = 1.65$ kT_0, and $R_g = 13.8$ Å; analytical: $D_0 = 37.8$ Å, $\delta = 4.14$ Å (fixed), $\varepsilon = 1.71$ kT_0, and $R_g = 13.6$ Å).

The simplifying assumptions we retained in the present application are an isotropic interaction potential and hence an isotropic monomer shape, limiting the maximum Q values that could be fitted. The ratio $R_g/R_0 = 2R_g/D_0$ provides a connection between the interaction potential (characterized by D_0) and how the protein scatters (characterized by R_g). Both proteins had a ratio within 9% of the $\sqrt{3/5}$ ratio expected for spherical monomers (Tables 1 and 2). Over the Q-range we examined, neither deviations of protein shapes from a sphere nor electron-density variations are likely to fully account for the difference from the ideal $\sqrt{3/5}$ ratio. More likely, hydration water that interacts strongly with the protein surface could explain the discrepancy between the fitted values of R_g and D_0, because the effective size of the hydrated protein could simply be different for the two different physical processes of x-ray scattering and protein-protein interaction.

Indeed, our fitted radii of gyration in Tables 1 and 2 are larger than the values obtained by taking the bare protein structures from the Protein Data Bank. For example, one would expect $R_g = 11.85$ Å for bare λ_{6-85}, not the 13.1–13.8 Å range obtained from our fits, the best of which has $R_g = 13.5$ Å (Table1). It has been shown previously that the hydration layer around proteins perturbs SAXS such as to increase the effective radius by 1–2 Å. The program CRYSOL takes this effect into account (25,26). Its predicted hydrated radius of gyration is 13.5–13.8 Å, depending on the method used, in excellent agreement with the value we derived from fitting interaction potentials to the SAXS experiment. A similar result is obtained for *fyn*-SH3, although our experimentally fitted radius of gyration is yet another 0.5 Å larger than the one obtained from CRYSOL. This could be due to the histidine tag on our *fyn*-SH3 protein, which was not included in the CRYSOL calculation (no structure is available for the tag).

Extrapolations of the scattering data in Fig. 1 to zero concentration are fitted well by CRYSOL with Protein Data

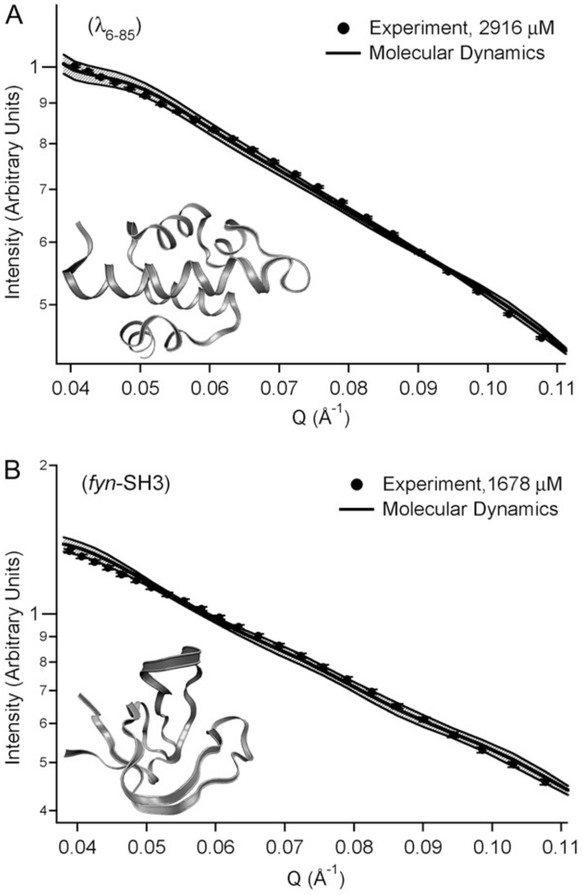

FIGURE 4 Experiment (*circles* with *error bars*) and molecular dynamics simulation (*thick solid line*) of the scattering intensity versus magnitude of the scattering vector for λ_{6-85} (*A*) and *fyn*-SH3 (*B*), confirming the quality of the parameter set obtained by MC modeling. The estimated 1σ sampling error we achieved in the MD simulations is indicated by the envelopes. (*Insets*) Native PDB structures for the protein fragments, as visualized with VMD (29).

Bank structural data as input, showing that the folded monomer shapes remain consistent throughout the concentration range. Our fitting approach clearly does not require a low concentration extrapolation to yield reliable results.

This leads to the question: What range of concentrations is needed to reliably fit the potential parameters, and which parameters remain least reliably determined? The fitting uncertainties are largest for D_0. We confirm in two ways that D_0 is the least well constrained parameter in our fits of λ_{6-85} and *fyn*-SH3. First, we fixed it at the hard-sphere value $2\sqrt{5/3}R_g$. This yielded radii of gyration, R_g, well depths, ε, and potential ranges, δ, that agreed with Tables 1 and 2 within the indicated uncertainties. D_0, on the other hand, shifted by up to 11%, showing that R_g is much more strongly constrained by the SAXS data than is D_0. Still, the χ^2 of the fits did increase by up to 70% when the constraint relating D_0 and R_g was introduced. Thus the differences between D_0 and R_g cannot be explained just by parameter uncertainties.

To investigate how many concentrations are needed to determine parameters, we performed fits with as few as two of the concentration series. For example, 2920 and 520 μM for λ_{6-85} yielded a very similar potential shape ($\varepsilon = -1.6$ kT_0, $\delta = 3.8$ Å for comparison with Table 1), but the parameter D_0 varies greatly (as low as $D_0 = 25$ Å). When more concentrations are added, D_0 approaches values more consistent with R_g. We conclude, at least for λ_{6-85} and *fyn*-SH3, that two concentrations are sufficient to define the shape of the potential, but that D_0 must either be constrained by R_g, or requires at least 5 or 6 concentrations, including high concentrations, to be adequately constrained.

It is worth noting that analytical fitting methods also have problems determining D_0 accurately. For example, two studies of the lysozyme interaction potential had to fix D_0 at values ranging from 28 to 36 Å to fit the other potential parameters (2,27). The value for an ideal sphere is ~37 Å in that case. Our numerical scattering method can be used to validate the analytical approximations usually used to obtain isotropic interaction potentials. To do so, we compared an analytical potential for λ_{6-85} to a simulation-derived potential. To make the comparison feasible within the limitations of the analytical approach, we used a hard-sphere reference potential, coupled with an attractive Yukawa term, to yield a potential similar in shape to our best fit in Table 1. We employed the analytical method described by Winter and co-workers (2), after verifying that our analytical code reproduced their experimental SAXS intensities from their potential parameters. Fig. 3 *B* compares the numerical λ_{6-85} potential with the analytical potential. Either D_0 or the potential range δ was highly correlated with potential depth in the analytical fit, so we had to fix one at the MMC value (δ in Fig. 3 *B*; the result looks even closer with D_0 fixed). With that restriction, reasonable agreement is obtained between the analytical and simulation result. However, as already discussed above, the simulation yields a much more robust fit than the analytical model when more than two concentrations are used; it does

TABLE 2 Best fit of the *fyn*-SH3 SAXS data to a $U_L + U_Y + U_E$ (r^{12} repulsive, exponential attractive, exponential repulsive) potential

		Attractive		Repulsive			Net well depth	
Potential type	R_g (Å)	ε_1 (kT_0)	δ_1 (Å)	ε_2 (kT_0)	δ_2 (Å)	D_0 (Å)	ε (kT_0)	RMSE
$U_L + U_Y + U_E$	14.85 ± 0.2	11.2 ± 0.5	1.0	7.5 ± 0.2	2.0	42.0 ± 4.0	3.7	0.035

Best fit	1680 μM	1020 μM	690 μM	470 μM	190 μM	60 μM	Weighted average
RMSE	0.016	0.0074	0.010	0.034	0.028	0.060	0.035

Also shown in the table are the root mean-square errors (RMSE) for the best overall fit at individual concentrations. kT_0 corresponds to 245 K.

not treat the potential as a small perturbation to a hard-sphere wall. In particular, D_0 can be floated as a free parameter and yields results consistent with R_g (<9% discrepancy) when enough concentrations are fitted. To the best of our knowledge, we did not find any analytical treatments in the literature where adjusting D_0 and R_g independently was possible, let alone yielded consistent results.

We examined a number of isotropic interaction potentials in addition to the best-fit and hard-wall shapes, and found that Gaussian attractive potentials generally performed more poorly than the exponential or Yukawa forms used in the DLVO model. In all fits, the λ_{6-85} potential was longer-range than the *fyn*-SH3, which resembles a ''sticky sphere''. A long-range but weak attractive potential for λ_{6-85} is compatible with recent terahertz measurements of hydration shells around the same mutant (7). These measurements indicated that the dynamics of water molecules are affected by the protein to >10 Å from the protein surface. Such hydration water may significantly mediate protein-protein interactions. It is even possible that the protein-protein interaction potential depends on protein concentration because of concentration-induced changes in the hydration shell. However, our current SAXS data was adequately modeled by a concentration-independent interaction potential.

λ_{6-85} has a significantly lower propensity for aggregation than *fyn*-SH3, but only the latter requires a repulsive potential in the fit to match the data within experimental uncertainty (Fig. 3 *A*). Both proteins were examined in identical buffer solutions of 45:55 vol % ethylene glycol/water, 50 mM phosphate at pH 7.0 and −28°C. As discussed by Winter and co-workers (2), the size of the repulsive potential is very sensitive to the ionic strength and ionic composition of the buffer. Given the isoelectric points of pI = 8.25 (λ_{6-85}) and pI = 4.84 (*fyn*-SH3), it is not surprising that there are differences between λ_{6-85} and *fyn*-SH3 in the screening of the long-range electrostatic repulsion.

As measurements over wide *Q*-ranges become available with new high brightness synchrotron sources, the direct fitting approach will also be useful for determining anisotropic interaction potentials. This requires two additions to our treatment: the potential itself must treat anisotropic interactions, and the scattering calculation can no longer assume spherical monomers. Regarding the potential, Ha-Duong and co-workers have developed residue-residue pair potentials that can be applied to surface residues of interacting proteins (28). To treat arbitrary protein shapes, one adds a rotational diffusion term to Eq. 3 and replaces F_m in Eq. 4 with the orientation-dependent structure factor of the monomeric protein computed using a program such as CRYSOL (25). It remains to be seen how much information might be extracted from scattering data at larger angles using this approach.

In conclusion, direct fitting of SAXS data to interaction potentials via Monte Carlo or molecular dynamics simulation of a model protein ensemble provides a useful alternative to analytical approximations. The form of the potential is unrestricted and no approximations regarding the scattering amplitude of the model protein ensemble need to be made. A range of concentrations still provides the best sampling of protein-protein distances to determine the potential (the potential wall location D_0 in particular), but extrapolations to zero concentration are not necessary. When the potential is restricted to have a hard-sphere wall, our method validates the analytical methods used to date, but actually fits D_0 more consistently with the protein size determined by the scattering amplitude (R_g). With the advent of higher-power computing, the numerical approach demonstrated here can be extended straightforwardly to include coarse-grained anisotropic interaction potentials, and randomly reorienting nonspherical protein shapes.

We thank Dr. Liang Guo at Argonne National Laboratory for helpful assistance in setting up the beam line for experiments, and Mr. Y. Matsumura for assistance during data collection.

This work was supported by National Science Foundation grant MCB 0613643. Time on BioCAT-18, managed by Prof. Thomas Irving, was supported by proposal GUP-6360 of the Advanced Photon Source at the Argonne National Laboratory. M.G. was supported by a Lycan Professorship.

REFERENCES

1. Tardieu, A., A. Le Verge, M. Malfois, F. Bonneté, S. Finet, M. Riés-Kautt, and L. Belloni. 1999. Proteins in solution: from x-ray scattering intensities to interaction potentials. *J. Cryst. Growth.* 196:193–203.
2. Javid, N., K. Vogtt, C. Krywka, M. Tolan, and R. Winter. 2007. Protein-protein interactions in complex cosolvent solutions. *Phys. Chem. Chem. Phys.* 8:679–689.
3. Otting, G., and K. Wüthrich. 1989. Studies of protein hydration in aqueous solution by direct NMR observation of individual protein-bound water molecules. *Biochemistry.* 111:1871–1875.
4. Zanotti, J. M., M. C. Bellissent-Funel, and J. Parello. 1999. Hydration-coupled dynamics in proteins studied by neutron scattering and NMR: the case of the typical EF-hand calcium-binding parvalbumin. *Biophys. J.* 76:2390–2411.
5. Rösgen, J., B. M. Pettitt, and D. W. Bolen. 2005. Protein folding, stability, and solvation structure in osmolyte solutions. *Biophys. J.* 89: 2988–2997.
6. Leitner, D. M., M. Havenith, and M. Gruebele. 2006. Biomolecule large-amplitude motion and solvation dynamics: modeling and probes from THz to X-rays. *Int. Rev. Phys. Chem.* 25:553–582.
7. Ebbinghaus, S., S. J. Kim, M. Heyden, X. Yu, U. Heugen, M. Gruebele, D. Leitner, and M. Havenith. 2007. An extended dynamical solvation shell around proteins. *Proc. Natl. Acad. Sci. USA.* In press.
8. Yang, W. Y., and M. Gruebele. 2003. Folding at the speed limit. *Nature.* 423:193–197.
9. Silow, M., Y. Tan, A. R. Fersht, and M. Oliveberg. 1999. Formation of short-lived protein aggregates directly from the coil in two-state folding. *Biochemistry.* 38:13006–13012.
10. Yang, W., and M. Gruebele. 2006. Binary and ternary aggregation with tethered protein constructs. *Biophys. J.* 90:2930–2937.
11. Otzen, D. E., S. Miron, M. Akke, and M. Oliveberg. 2004. Transient aggregation and stable dimerization induced by introducing an Alzheimer sequence into a water-soluble protein. *Biochemistry.* 43: 12964–12978.
12. Niehbur, M., and M. H. J. Koch. 2005. Effects of urea and trimethylamine-N-oxide (TMAO) on the interactions of lysozyme in solution. *Biophys. J.* 89:1978–1983.

13. Dumont, C., Y. Matsumura, S. J. Kim, J. Li, E. Kondrashkina, H. Kihara, and M. Gruebele. 2006. Solvent-tuning collapse and helix formation time scales of $\lambda_{6\text{-}85}$. *Protein Sci.* 15:2596–2604.

14. Ghaemmaghami, S., J. M. Word, R. E. Burton, J. S. Richardson, and T. G. Oas. 1998. Folding kinetics of a fluorescent variant of monomeric lambda repressor. *Biochemistry.* 37:9179–9185.

15. Larson, S. M., A. A. D. Nardo, and A. R. Davidson. 2000. Analysis of covariation in an SH3 domain sequence alignment: applications in tertiary contact prediction and the design of compensating hydrophobic core substitutions. *J. Mol. Biol.* 303:433–446.

16. Edelhoch, H. 1967. Spectroscopic determination of tryptophan and tyrosine in proteins. *Biochemistry.* 6:1948–1954.

17. Fischetti, R., S. Stepanov, G. Rosenbaum, R. Barrea, E. Black, D. Gore, R. Heurich, E. Kondrashkina, A. J. Kropf, S. Wang, K. Zhang, T. C. Irving, and G. B. Bunke. 2004. The BioCAT undulator beamline 18ID: a facility for biological non-crystalline diffraction and X-ray absorption spectroscopy at the Advanced Photon Source. *J. Synchrotron Radiat.* 11:399–405.

18. Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21:1087–1092.

19. Chandler, D. 1989. Modern Statistical Mechanics. Oxford University Press, Oxford, UK.

20. Rayleigh, L. 1914. On the diffraction of light by spheres of small relative index. *Proc. Roy. Soc. Lond.* A-90:219–225.

21. Rayleigh, L. 1910. The incidence of light upon a transparent sphere of dimension comparable to the wavelength. *Proc. Roy. Soc. Lond.* A-84: 25–46.

22. Suhonen, H. 2005. Simulation of Small-Angle X-Ray Scattering from Collagen Fibrils. Master's thesis. University of Helsinki, Helsinki, Finland.

23. Moré, J. J., B. S. Garbow, and K. E. Hillstrom. 1980. User Guide for MINPACK-1. *In* Argonne National Laboratory Report ANL-80-74. Argonne National Laboratory, Argonne, IL. 1–48.

24. Moré, J. J., D. C. Sorensen, and K. E. Hillstrom. 1984. The MINPACK Project. *In* Sources and Development of Mathematical Software. W. J. Cowell, editor. Prentice-Hall, New York. 88–111.

25. Svergun, D., C. Barberato, and M. H. J. Koch. 1995. CRYSOL: a program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Cryst.* 28:768–773.

26. Svergun, D., S. Richard, M. H. J. Koch, Z. Sayers, S. Kuprin, and G. Zaccai. 1998. Hydration shell in solution: validation by x-ray and neutron scattering. *Proc. Natl. Acad. Sci. USA.* 95:2267–2272.

27. Narayanan, J., and X. Y. Liu. 2003. Protein interactions in undersaturated and supersaturated solutions: a study using light and x-ray scattering. *Biophys. J.* 84:523–532.

28. Bosdevant, N., D. Borgis, and T. Ha-Duong. 2007. A coarse-grained protein-protein potential derived from an all-atom force field. *J. Phys. Chem. B.* 111:9390–9399.

29. Humphrey, W. F., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14:33–38.