CHARACTERIZING THE SPACE OF INTERATOMIC DISTANCE DISTRIBUTION FUNCTIONS CONSISTENT WITH SOLUTION SCATTERING DATA

Paritosh A. Kavathekar
Department of Computer Science, Dartmouth College, Hanover, NH 03755

Bruce A. Craig
Department of Statistics, Purdue University, West Lafayette, Indiana 47907, USA

Alan M. Friedman
Department of Biological Sciences, Purdue Cancer Center and Bindley Bioscience Center, Purdue University, West Lafayette, Indiana 47907, USA

Chris Bailey-Kellogg and Devin J. Balkcom*
Department of Computer Science, Dartmouth College, Hanover, NH 0375
**Email: devin@cs.dartmouth.edu*

Abstract: Scattering of neutrons and x-rays from molecules in solution offers alternative approaches to the studying of a wide range of macromolecular structures in their solution state without the need of crystallization. In this paper, we study one part of the problem of elucidating three-dimensional structure from solution scattering data, determining the distribution of interatomic distances, P(r). This problem is known to be ill-conditioned; for a single observed diffraction pattern, there may be many consistent distance distribution functions. Due to the ill conditioning, there is a risk of overfitting the observed scattering data. We propose a new approach to avoiding this problem, accepting the validity of multiple alternative P(r) curves rather than seeking a single "best".

We show that there are linear constraints that ensure that a computed P(r) is consistent with the experimental data. The constraints enforce smoothness in the P(r) curve, ensure that the P(r) curve is a probability distribution, and allow for experimental error. We use these constraints to precisely describe the space of all consistent P(r) curves as a polytope of histogram values or Fourier coefficients. This description can then be used to sample the space of potential alternative P(r) curves. We use this description to develop a linear programming approach to sampling the space of consistent, realistic P(r) curves. In tests on both experimental and simulated scattering data, our approach efficiently generates ensembles of such curves that display substantial diversity. In particular, we show that the ensemble of P(r) curves generated for a given protein includes members that are more different from a reference curve for that protein than are reference curves for proteins of other structural topologies. Thus subsequent reconstruction steps must properly account for this P(r) diversity in optimizing structural models.

1. INTRODUCTION

There is currently no single best experimental technique for studying the structure of macromolecules. Crystallizing proteins for x-ray crystallography is often difficult, while NMR is limited by the size of the protein or complex. In this paper, we examine solution scattering, often called small angle x-ray scattering (SAXS) when x-rays are employed and only data to small diffraction angle is collected. Solution scattering is a relatively simple and inexpensive experimental technique that can be applied to a large range of molecular sizes, from 10-1000 Å, without the need for crystallization [1]. SAXS has found widespread applications for diverse problems such as low resolution

*Corresponding author.

structure prediction of proteins and complexes [2–5], protein folding studies with time-resolved data [6, 7], protein dynamics [8], and determining the association model for protein complexes [9].

In a SAXS experiment, a narrow beam of x-rays is directed towards a dilute solution of macromolecules. The electrons in the macromolecule scatter the beam, and a detector measures the intensity of the scattered beams in different directions. The resultant intensity at any point depends on the relative position of the electrons with respect to each other, yielding a curve I(q) that expresses the scattered intensity (I) at different values of the momentum transfer vector (q) .

The I(q) curve is a function of the protein shape, and although spherical averaging and experimental noise cause a loss of information in going from a three-dimensional structure to a one-dimensional scattering pattern, it has proven practically possible to use SAXS to reconstruct the low-resolution structure of a macromolecule [2, 10, 3]. We study an intermediate problem in the reconstruction process: computing a function P(r) that describes the distribution (P) of interatomic distances (r) in a molecule. The P(r) is a more intuitive function of the molecular shape than the I(q) curve, and is a useful intermediate in the reconstruction. For example, the real-space version of the program GASBOR [11, 12] produces models by matching a given P(r) curve.

The P(r) curve is related to the I(q) curve by the following Fourier tranform [13]:

$$P(r) = \frac{2r}{\pi} \int_0^\infty q \cdot I(q) \sin(qr) dq, \qquad (1)$$

where $q = 4\pi \sin\theta/\lambda$, with 2θ the scattering angle and λ the x-ray wavelength.

Figure 1 summarizes the relationship among structure, P(r), and I(q). Although the P(r) curve is related to the scattering curve with the well-established integral relation of equation 1, obtaining the P(r) curve for a protein given its scattering profile is not trivial. This problem is ill-conditioned [14] because experimentally the data is available only for a finite q range, whereas the integral in equation 1 extends between 0 and infinity. Applying the direct transform to the limited data produces non-physical P(r) curves.

The ill-conditioned nature of this problem creates significant potential for overfitting the data. All previous approaches try to produce a single best P(r) curve and avoid overfitting by various mechanisms. Most previous approaches try to produce P(r) curves that reduce the discrepancy, χ^2, between the experimental scattering curve and the one predicted using the Fourier inverse of equation 1. Additional constraints are employed to reduce the potential for overfitting.

Early approaches modeled P(r) curves as the summation of continuous basis functions. For example, Moore [15] used sine functions over a restricted interval and employed a Shannon information content criterion to avoid overfitting by limiting the number of sine functions. Glatter [16] used B-splines to model P(r) curves and avoided overfitting by added a damping term to the χ^2 target function in order to obtain a smooth solution. Steenstrup and Hansen [17, 18] modeled P(r) as a discrete distribution function defined at a fixed number of points. They avoided overfitting by maximizing the entropy of the P(r) distribution function subject to the constraint $\chi^2 \approx 1$, expressed as a Lagrange multiplier.

One of the most widely used programs for constructing P(r) curves is GNOM [14, 19], which uses Tikhonov regularization [20, 21] in order to avoid overfitting. GNOM defines a set of perceptual criteria that the user desires in the P(r) and sets the regularization and smoothing parameters so as to best achieve those properties.

Our alternative approach to avoiding overfitting is not to seek a single "best" P(r) solution, but rather to accept the validity of multiple feasible solutions consistent with a given I(q) curve. Key to determining such an ensemble is a representation for P(r) curves that are not just consistent with a scattering profile, but also have properties such as continuity and smoothness that make them protein-like. The main contributions of this paper are as follows:

(1) We provide a complete characterization, as a convex polytope in an appropriate representation space, of those P(r) curves that are consistent with a given scattering curve and display realistic P(r) properties.
(2) We provide a linear-programming based method to quickly generate consistent, realistic P(r) curves for a given scattering curve.

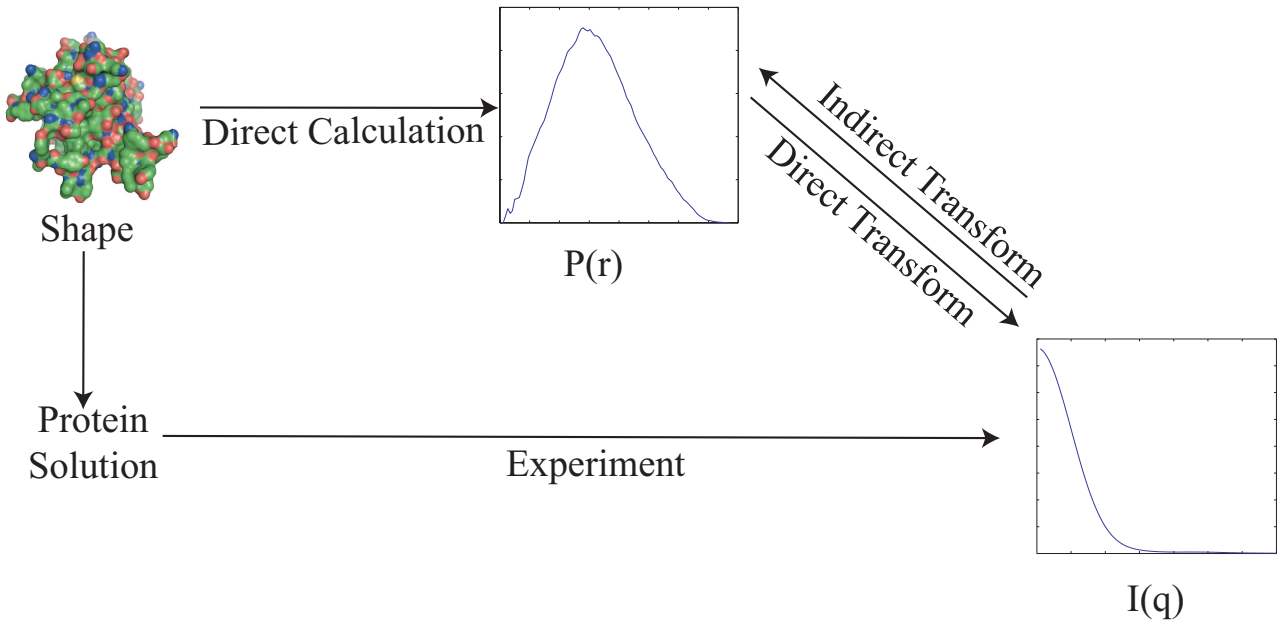


Fig. 1. A schematic diagram showing the correspondence between protein shapes, their P(r) curves, and scattering curves. The labels on the arrows show the techniques used to get one representation from the other. In practice, different shapes might have close P(r) curves, and different I(q) curves may map to similar P(r) curves.

(3) In tests with both experimental and simulated data, we demonstrate that consistent and realistic P(r) curves are significantly diverse, such that ensembles for proteins with different folds can overlap, limiting the direct identification of protein fold from scattering data.

2. METHODS

Equation 1 describes the relationship between the scattering curve and the interatomic distance distribution function. Our aims are to characterize the space of all physical P(r) curves that are consistent with a given I(q) curve, and to quickly generate an ensemble of such curves. We first overview the approach, before providing details in the subsections.

One common representation of a P(r) curve is as a histogram of bins centered at discrete values. A P(r) curve is thus represented by a vector of length n, where n is the number of bins. Conversely, we can say that every point in an n-dimensional space corresponds to a P(r) curve. We call this space the P(r) *space*. We can characterize points in P(r) space as *consistent* (they represent curves that satisfy the given I(q) curve within a predefined error tolerance), *realistic* (they represent curves that are smooth and protein-like), both, or neither. In Section 2.1 we mathematically define consistent and realistic P(r) curves. Under that definition, we show that all such curves lie inside a convex polytope in P(r) space. Mathematically, we can describe the set of all solutions with an equation

$$C \cdot P(r) \leq b \ , \tag{2}$$

where C is a matrix that defines this polytope, P(r) a vector representing the histogram, and b a vector of constants representing the constraints.

Another common representation of a P(r) curve is as a continuous curve, a linear combination of basis functions. Practically, the number of basis functions is finite, say k. Thus a P(r) curve is represented as a set of k coefficients, defining the linear combination. As before points in the *coefficient space* correspond to P(r) curves. In Section 2.2, we extend the results for the histogram approach to coefficient space; if α represents a point in coefficient space, then all consistent and realistic curves lie inside a convex polytope in coefficient space given by

$$C'\alpha \leq b' \ . \tag{3}$$

The geometric characterization of these spaces, in which all consistent curves lie in a contiguous,

well-defined region, has advantages both for understanding the properties of the curves, as well as for producing ensembles using linear combinations. This leads to our algorithms in Section 2.3 for generating consistent and realistic P(r) curves.

2.1. Consistent and realistic P(r) curves in the histogram representation

Both P(r) and I(q) are continuous functions. Experimentally, the scattering intensities are measured at some discrete set of q values. Similarly, P(r) curves are also represented as histograms with a known bin width. Our method treats both P(r) and I(q) as continuous curves sampled at discrete points. Although we use discrete approximations of continuous curves, we do not place any condition on the number, width or uniformity of bins.

Under discrete approximation we can write the Fourier inverse of equation 1 as follows:

$$I(q) = \sum_{i=1}^{n} h_i \frac{\sin(qr_i)}{qr_i} P(r_i) \ , \tag{4}$$

where h_i is the width of the bin. In the interest of clarity we will ignore the bin-width parameter h_i in some of the equations; this omission does not affect our framework.

We scale I(q) curves by dividing by $I(0)$, the "0" angle scattering intensity. Experimentally, $I(0)$ values are not available, and are estimated using the Guinier plot [22, 13].

For a set of discrete q values, we can write equation 4 in a matrix form:

$$\begin{pmatrix} I(q_1) \\ \vdots \\ I(q_m) \end{pmatrix} = \begin{pmatrix} \frac{\sin(q_1 r_1)}{q_1 r_1} & \cdots & \frac{\sin(q_1 r_n)}{q_1 r_n} \\ \vdots & \ddots & \vdots \\ \frac{\sin(q_m r_1)}{q_m r_1} & \cdots & \frac{\sin(q_m r_n)}{q_m r_n} \end{pmatrix} \begin{pmatrix} P(r_1) \\ \vdots \\ P(r_n) \end{pmatrix} \tag{5}$$

$$I(q) = A \cdot P(r) \ . \tag{6}$$

where A is the transform matrix. We use P(r) to represent both the interatomic-distance distribution function and the vector defined in equation 5; the correct interpretation should be clear from the context. $P(r_i)$ refers to the $P(r)$ value at the ith sampled point, r_i. We assume that there are n such points and m sampled points in the q space.

We now mathematically define consistent and realistic P(r) curves in terms of linear constraints on the P(r) histograms.

P(r) distribution. P(r) is a probability distribution, so it must sum to 1. (This is a normalizing constraint; $P(r)$ scales are relative. This normalization implies that we can treat P(r) curves as probability distributions.) Mathematically,

$$\sum_i P(r_i) = 1 \ .$$

Scattering intensity. The P(r) curve must give rise to the observed scattering curve. Ideally the curve should satisfy equation 5, but practically because of experimental limitations and noise we can only constrain the predicted I(q) to a certain interval:

$$I(q) - \sigma(q) < A \cdot P(r) < I(q) + \sigma(q) \ .$$

Here $\sigma(q)$ is a column vector that specifies the allowed error at each q value. In our implementation we use σ values that depend on the standard deviation of the measured intensities.

Non-negativity. Since $P(r_i)$ are probability values, they must all be non-negative:

$$0 \leq P(r_i) \leq 1, \qquad \forall i \ .$$

While these constraints are loose, for some (or all) r values we can enforce stricter constraints that force the probability values to lie in a particular interval. For example, for r close to 0 or $D_{\max}$, the maximum distance which can be obtained from the I(q) curve, we can limit the probability values to the range $[0, \epsilon]$ where ϵ is suitably close to 0.

Continuity. One way to ensure that P(r) curves are smooth is to restrict the amount of variation in the probability values between adjacent bins. Though such a constraint does not eliminate local maxima or minima, it ensures that these extremal points are not sharp. Let β be the maximum permissible difference between adjacent probability values. We can write the continuity constraint as follows:

$$|P(r_{i+1}) - P(r_i)| \leq \beta, \quad \forall i < n \ .$$

Criteria that influence the selection of β include the number of bins, width of the bins, and level of smoothness desired. In practice, a uniform β works well, but β need not be constant over the entire

curve. In some regions, such as near $r = 0$, sharp spikes reflecting atomic packing (with high resolution data) may be acceptable, while in others they may not be desirable.

Smoothness. Continuity constraints restrict abrupt changes in P(r) values, but portions of the curve can still have a saw-tooth pattern without violating continuity. This pattern is characterized by alternating small local maxima and minima. We address this problem by enforcing second-order constraints on consecutive triples of the P(r) curve:

$$|P(r_{i-1}) - 2P(r_i) + P(r_{i+1})| \leq \gamma \quad , \forall i \in [2, n-1] \; .$$

These conditions bound the derivative at each point in the discrete approximation of the P(r) curve. Here, γ is a user-defined parameter that need not be constant over the entire curve; similar to the continuity constraints, we can have different curvature bounds for different portions of the curve.

All constraints used to describe consistent P(r) curves are linear, so we can combine them to produce equation 2, with C as a matrix containing the constraints and b a vector containing the corresponding constants. The normalization constraint is an equality constraint, but may be written as an equivalent pair of inequality constraints. Equation 2 represents a convex polytope in P(r) space that characterizes the space of desired solutions; all consistent P(r) curves lie inside this high-dimensional convex polytope, and conversely any point inside this polytope corresponds to a consistent and realistic P(r) curve.

2.2. Consistent and realistic P(r) curves in the basis function representation

Let us express P(r) curves in a functional basis such as a Fourier basis [15]. Then for any value r_j we can write

$$P(r_j) = \sum_{i=1}^{k} \alpha_i f_i(r_j) \quad ,$$

where f_i are the basis functions, α_i the corresponding coefficients, and k the number of basis functions suitable to represent all P(r) curves to the required resolution.

Let 1_k be a row vector of all ones, with length k. Similarly, define 0_k as a row vector of length k with all zeros. Using 1_k and 0_k we define two matrices, M and D, as follows:

$$M = \begin{pmatrix} 1_k & 0_k & \dots & 0_k \\ 0_k & 1_k & \dots & 0_k \\ \vdots & \vdots & \ddots & \vdots \\ 0_k & 0_k & \dots & 1_k \end{pmatrix}$$

and

$$D = \begin{pmatrix} f_1(r_1) & 0 & \dots & 0 \\ 0 & f_2(r_1) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & f_k(r_1) \\ \vdots & \vdots & \vdots & \vdots \\ f_1(r_n) & 0 & \dots & 0 \\ 0 & f_2(r_n) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & f_k(r_n) \end{pmatrix} \quad ,$$

where M is an $n \times kn$ block-diagonal matrix, D is a $kn \times k$ matrix with the basis functions. If α is a vector of coefficients for the basis functions, then we can express the vector P(r) in terms of M and D as,

$$\begin{pmatrix} P(r_1) \\ P(r_2) \\ \vdots \\ P(r_n) \end{pmatrix} = MD \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{pmatrix} \; . \tag{7}$$

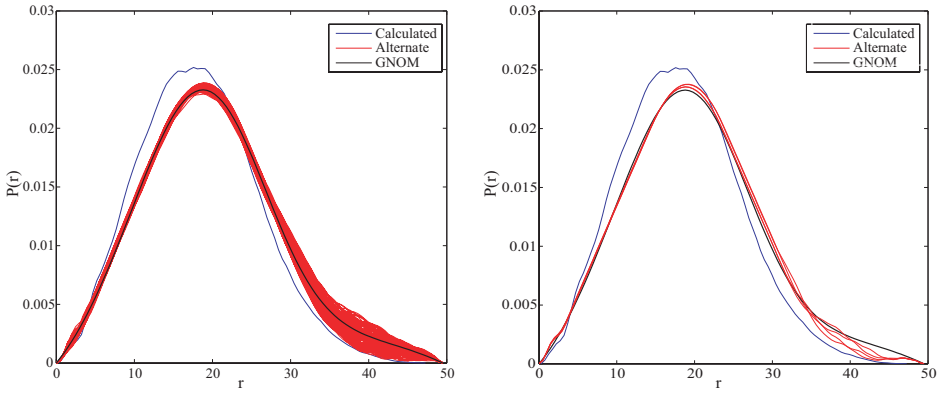
Substituting equation 7 in equation 2 (the constraints from the previous section) gives

$$CMD\alpha \leq b$$

By defining $C' = CMD$, we get equation 3, the equation of a polytope in the coefficient space. We must note that although the constraints are applied at discrete points, the underlying curves are continuous.

2.3. Generating P(r) ensembles

Intuitively, the most diverse curves lie on vertices of our polytope. While there exist vertex enumeration algorithms (*e.g.*, [23]), they become impractical in very high dimensions (of the order of the number of points in a P(r) curve). Furthermore, the number of vertices increases exponentially with dimensionality. Thus we instead seek a diverse subset of the vertices.



(a) 100 alternate curves

(b) Some example curves

Fig. 2. Alternate P(r) curves (red) generated by our method for experimental data to $q_{\max} = 0.5$ for hen egg white lysozyme, along with the curve calculated from the x-ray structure (blue) and the reconstruction by GNOM (black).

We formulate the problem of generating the vertices of the polytope as a linear program in coefficient space. Let c be a point in the coefficient space; then we solve the following linear program:

$$\text{Maximize} \quad c \cdot \alpha \ ,$$
$$\text{subject to} \quad C'\alpha \leq b \ .$$

By maximizing the dot product of c with the coefficient vector α, subject to constraints defined in section 2.1, we obtain a vertex of the polytope (or a point on the facet, in case of interior point methods) in the coefficient space.

To generate a number of vertices, we simply solve the optimization problem for many random vectors c. For a reasonable parameter choice and using Fourier basis functions our program takes about one second to generate a candidate P(r) curve on a 2.4GHz Pentium machine using the Matlab solver. Therefore, it is easy to rapidly generate a large ensemble (and possibly select the most interesting curves from it).

To some extent, these vertices can be described as "maximally diverse," since the linear program picks a direction in coefficient space and tries to find the maximum possible variation in that direction without violating any of the constraints.

While the vertices capture the "envelope" of P(r) curves, one might also want to obtain a more complete ensemble of the satisfying P(r) curves. Since simple generate-and-test algorithms are extremely inefficient in high dimensions, we instead generate additional satisfying curves by repeatedly taking convex combinations of previously identified satisfying curves. Such curves are guaranteed to lie inside the polytope due to its convex nature.

3. RESULTS

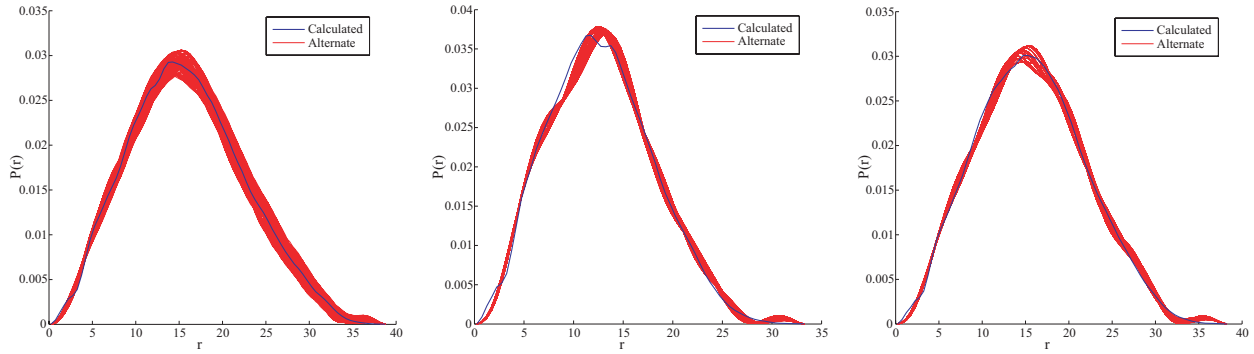
We demonstrate the effectiveness and significance of our approach in characterizing the diversity in realistic P(r) curves consistent with scattering data.

3.1. Ensemble Diversity

We first applied our approach to experimental scattering data up to $q_{\max} = 0.5$ for the protein hen egg-white lysozyme, as distributed with the program GNOM [14]. Figure 2(a) shows an ensemble of 100 curves generated using our method, and Figure 2(b) shows a few examples chosen from the ensemble. For comparison the black curve shows the output from the program GNOM [14], using default parameters, while the blue curve is the distance-distribution calculated from an x-ray crystal structure.

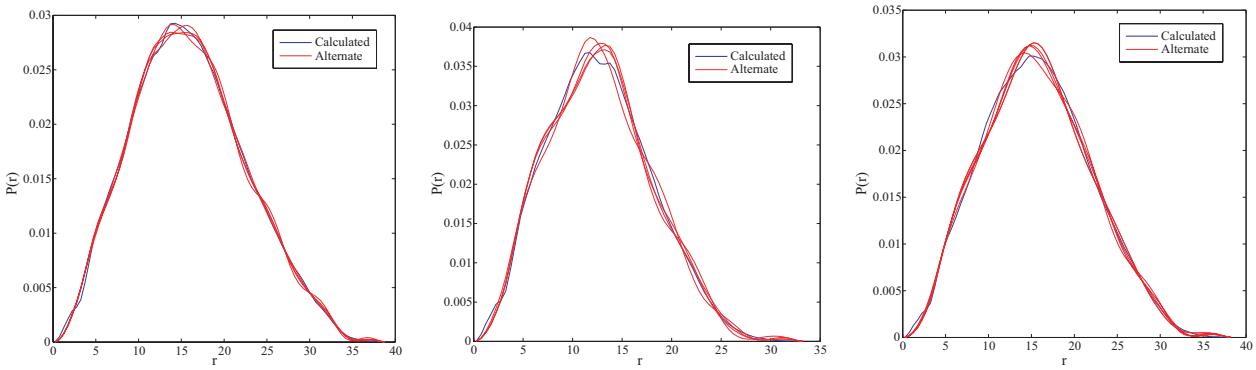
There is significant diversity in the consistent, realistic P(r) curves (see Table 1 for quantification). In addition to evaluating the "global" diversity, we may also evaluate the uncertainty at a given point r by the height of the red band.

We note that both our ensemble and the curve generated using GNOM are shifted relative to the curve from the PDB file. This is due to the contribution to the solution scattering intensity from the bulk solvent that the protein replaces and from a hard water shell around the protein [24, 25]. Since it is difficult to model these solvent interactions, in or-



(a) 1i27A00 (b) 1hp8A02 (c) 1mj4A00

Fig. 3. Ensemble of 100 alternate P(r) curves (red) generated by our method for simulated scattering data to $q_{\max} = 0.7$ for the three reference domains, along with the curve calculated from the x-ray structure (blue).



(a) 1i27A00 (b) 1hp8A02 (c) 1mj4A00

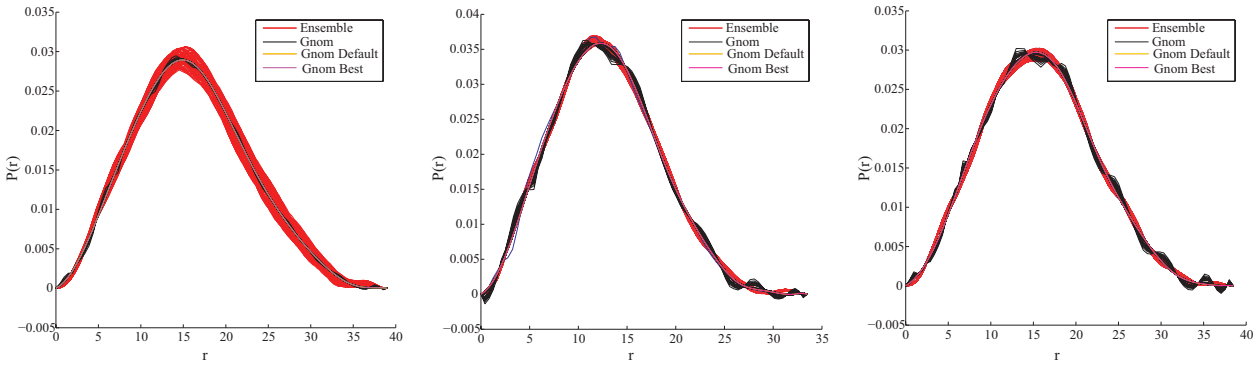
Fig. 4. Some sample alternate P(r) curves generated for the three reference domains. These curves are taken from the set of curves shown in figure 3(a)–3(c). This figure shows that the alternate curves are smooth and protein-like.

der to better evaluate our method with a "ground truth", we next turn to simulation results for which these scattering contributions are not included.

For simulation studies we have selected three very different domains, one each of alpha, beta and alpha plus beta proteins under CATH classification [26, 27]. The CATH ids are 1i27A00 (*Arc Repressor Mutant, subunit A topology*), 1hp8A02 (*Seminal Fluid Protein PDC-109 (DomainB) topology*) and 1mj4A00 (*ubiquitin-like (UB roll)* topology). For simulated data, we use the scattering intensity of the protein in vacuum (an output from CRYSOL [24]) to $q_{\max} = 0.7$. We compare against the corresponding P(r) curve computed from the atomic coordinate file.

Figures 3(a), 3(b), and 3(c) show 100-member P(r) ensembles calculated from simulated data, while Figures 4(a), 4(b), and 4(c) show selected examples. The alternate curves are smooth, validating our approach of modeling smoothness as a constraint on the curves as opposed to a criterion to be optimized.

We have tested the robustness of our method by adding random Gaussian noise to the simulated scattering curves. At each q value, we added a random noise $c \cdot k \cdot \sigma(q)$, where c is a constant, k is a random variable that follows a standard normal distribution and $\sigma(q)$ is the standard deviation in the intensity value. With a small adjustment to the error bounds for scattering-intensity constraint, our method robustly handled noise as we increased c. We tested for c in the range $[0, 5]$ in increments of 1. We found that adding a noise of c standard deviations at each q value required a corresponding increase in error tolerance.



(a) 1i27A00 (b) 1hp8A02 (c) 1mj4A00

Fig. 5. Ensembles generated using our method (red) and by varying the regularization parameter in GNOM[14] (black).

Table 1. Diversity in our ensemble vs. one generated by varying the regularization parameter in GNOM[14]. d(GNOM, Polytope) is the minimum distance of a curve in GNOM to one in our ensemble, and similarly for d(Polytope, GNOM).

Protein	Average d(GNOM, Polytope)	Average d(Polytope, GNOM)	d(GNOM Default, Polytope)	d(GNOM Best, Polytope)
1i27A00	0.0298 ± 0.0011	0.0513 ± 0.0099	0.0292	0.0295
1hp8A02	0.0227 ± 0.0058	0.028 ± 0.0073	0.0225	0.0225
1mj4A00	0.0244 ± 0.0062	0.032 ± 0.0061	0.02732	0.02732

3.2. Comparison with GNOM

Our main approach is to describe completely the set of realistic P(r) curves using linear constraints. Once the set has been described, it can be sampled to find an ensemble of consistent and realistic P(r) curves. There are two advantages to this approach: all generated samples satisfy the constraints, and we can develop algorithms to get a diverse sampling within the polytope of consistent curves.

It is possible to generate an ensemble of P(r) curves using the existing GNOM software, by varying the regularization parameter [14], although this is not the typical use of the regularization parameter and generating ensembles is not the intended use of GNOM. To generate a GNOM derived ensemble we sampled the regularization parameter uniformly in log space. Of the curves so generated, we consider only those that GNOM classified as *GOOD*. We also generated two additional curves using GNOM: *GNOM Default*, the output of GNOM when run with default parameters and *GNOM Best*, the output GNOM produces when it is provided the correct OSCILL and VALCEN criteria [14] determined from the calculated curve.

Table 1 compares the diversity of the ensembles generated by our method and by GNOM. For every curve in the ensemble from GNOM we calculated the distance of the closest curve in our ensemble, and *vice versa*. The second and third columns in Table 1 show the average over these minimum distances, while the final two columns show the distances to the default and best curve from the curves in our ensemble. The average minimum distance of a curve from GNOM's ensemble to that from our polytope ensemble is shorter than *vice versa*. Thus, there are curves in our ensemble that do not have a correspondingly close curve in GNOM's ensemble. We attribute this greater diversity to the nature of our sampling. Our sampling method samples points from the vertices of the convex polytope in the coefficient space. These points, by definition, correspond to curves that are most diverse without violating the constraints in section 2.1.

Figures 5(a)–5(c) show the ensembles generated using our approach and those generated by GNOM. The GNOM curves are closely banded when compared to the curves generated from our ensemble. This graphically validates the results of Table 1; most curves in the GNOM band have a close counterpart in our ensemble, but curves in our ensemble do not

always have a close counterpart to those in GNOM. The oscillations and non-negative P(r) values in the GNOM curves reflect our forcing the regularization parameter to unusual values. Our method generates a diverse set without these problems because it explicitly enforces the smoothness and non-negativity constraints at each r value.

3.3. Structural implications of P(r) diversity

In order to investigate the diversity of our P(r) ensemble we compare it with a diverse set of representative protein structures. We define a set of 1084 diverse structures, which we call *CATHRep*, by selecting the CATH representative structure for each different topology. We generated both the particle scattering curves (using CRYSOL) and interatomic-distance distribution curves (from the pdb file) for all CATHRep proteins. Then we calculated the I(q) distance between each CATHRep protein and each of our example proteins, using RFactor [28] with uniform weights. Similarly, we calculated the distance between the P(r) curves for the three examples and all CATHRep proteins, in this case measuring distance as the area between the two curves. We repeated the procedure for each member of our generated P(r) ensemble, using equation 4 to determine corresponding I(q) curves.

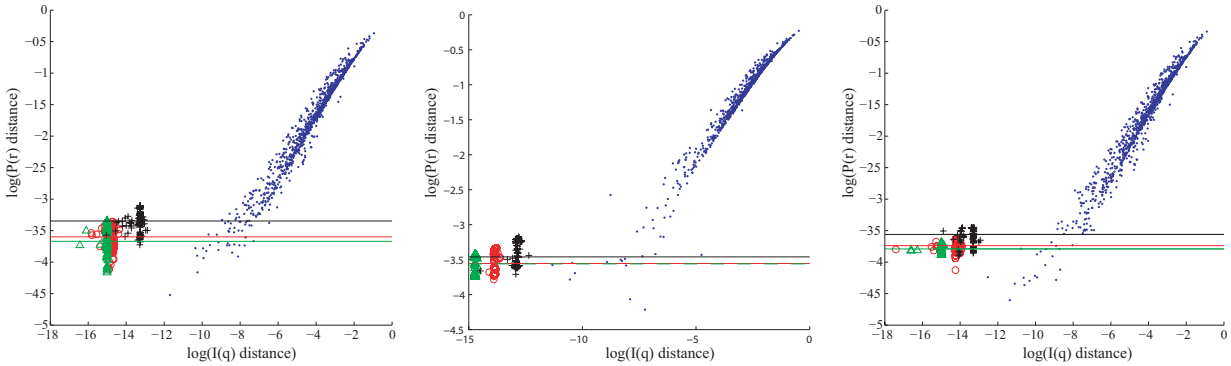
Figures 6(a)–6(c) plot the log of the I(q) distance vs. the log of the P(r) distance, where the blue dots correspond to CATHRep proteins and the other markers correspond to our P(r) ensemble at different values for the intensity constraints. It is clear that the scattering curves for the alternate curves are much closer than those for proteins in CATHRep, a direct consequence of the scattering-intensity constraint. However, there are structures in CATHRep whose P(r) curves are closer to the P(r) curves of the three examples than the alternate curves our method generates (as illustrated by the points below horizontal lines). This shows that the although the alternate P(r) curves give rise to scattering curves that are almost identical to those from the actual structure, they are significantly different from the P(r) curves for the actual structures when compared to the variability seen in CATHRep. Although all three of the examples show considerable variability in candidate P(r) curves, table 2 shows that this variability in P(r) curves is not uniform; 1hp8A02 has far fewer structures in CATHRep with closer P(r) curves than does 1i27A00. The third example 1mj4A00 falls somewhere between these two. An interesting extension of our work might be to evaluate different I(q) curves for their potential to generate P(r) curves with different diversity.

Table 2 summarizes the P(r) variability with respect to CATHRep. For every ensemble we calculated the maximum and median distance of the P(r) curves in the ensemble to that of the reference protein's P(r) curve. Table 2 shows the number of structures in CATHRep that have a P(r) distance to the reference structure below these thresholds. These results show that the diversity of P(r) in our ensemble is large enough to overlap with a substantial number of representative structures. We note that most structures in CATHRep have large differences in both P(r) and I(q) space because of the nature of CATHRep, which is supposed to represent a diverse set of protein structures. Therefore, for the examples we have considered the overlap is limited to 68 out of the 1084 structures in CATHRep.

Our method explicitly bounds the error at each q value for the scattering curve predicted using a P(r) curve. It it natural to ask if such constraints encourage curves that differ by the maximal possible deviation at each point. In practice, this does not appear to be the case for the ensembles our method generates. Typically the χ^2 distance between the scattering curves from the ensemble and those for the reference proteins was small. When a deviation of σ was allowed at every q value, the maximum χ^2 distance over ensembles for all three structures was 0.3057. When a deviation of $\sigma/2$ was permitted this value was 0.1035, and for ensembles with 2σ deviations permitted it was 0.4978. These χ^2 values fall much below the permitted deviation at each point; even if a deviation of one σ is permitted at every point, not every point on the scattering curve varies by that amount.

4. DISCUSSION AND CONCLUSIONS

We have described the first method for characterizing an ensemble of P(r) curves that are consistent with a given scattering profile. At the heart of our approach



(a) 1i27A00 (b) 1hp8A02 (c) 1mj4A00

Fig. 6. Similarity between reference proteins and proteins in the CATH representative database, in terms of I(q) (x-axis) and P(r) (y-axis). Ensemble members are plotted for the example proteins, permitting an error of $\sigma/2$ (green triangles), σ (red circles), or 2σ (black plus signs). A single blue dot is plotted for each CATH representative. Green, red, and black horizontal lines correspond to the median values for P(r) distances for the three sets of alternative curves.

Table 2. Number of members of CATHRep with P(r) distance below the maximum or median distance in the generated

Domain	Max. Error	#CATHRep structures	
		max	median
1i27A00	$\sigma/2$	40	18
	σ	38	20
	2σ	68	39
1hp8A02	$\sigma/2$	13	7
	σ	14	7
	2σ	13	12
1mj4A00	$\sigma/2$	19	16
	σ	20	19
	2σ	32	19

is the idea of representing the desirable properties of the P(r) curves as constraints on the set of solutions. Such a formulation allows us to describe all realistic and consistent P(r) curves as occupying a convex polytope in a high-dimensional space. We use linear programming to sample this space and rapidly generate a diverse ensemble of curves. In this section we discuss practical issues in implementing this approach, limitations, and possible future directions.

Any approach with a few user-defined parameters faces the problem of appropriately selecting those parameters. If the smoothness and continuity parameters are too small, the linear program becomes infeasible; too large, and the P(r) curves become jagged. Practically, we found that a binary search for these parameters worked reasonably well. Another parameter is the number of basis functions used to represent the P(r) curves. We obtained the smoothest solutions when the number of basis functions was the minimum required to satisfy the linear program. The minimum number of basis functions also represents the smallest ensemble that satisfies the linear program.

One can imagine constraints describing other desirable properties of P(r) curves beyond those we considered. So long as those constraints are linear, they may be easily added to the existing framework. Higher-order constraints might be added while still maintaining convexity and contiguity of the set of feasible P(r) curves.

There are limitations in the sampling technique we applied. Sampling uniformly in high dimensions is inherently hard, and we cannot claim to have a uniform coverage of the consistent P(r) space. This formulation does not indicate how big (or small) the polytope is, which depends on the user-defined parameters for various constraints as well as the particular properties of the transform matrix defined in

equation 5. Perhaps most importantly, we need to be aware that changes in some directions in P(r) space may affect the corresponding I(q) curves more than changes in other directions.

We now discuss the results of our approach in the wider context of *ab initio* structure prediction using SAXS. The mapping from structure to P(r) to I(q) is not one-to-one, and every node in CATHRep represents a unique topology. Therefore, combining these two facts, our results in section 3 show that such alternate P(r) curves might correspond to proteins with significant structural differences. This observation should be used both as a caution and as an opportunity for structure elucidation from solution scattering data. As a caution, it reminds us that diverse alternate structures can give rise to similar scattering curves. As an opportunity, our results can be viewed as a first step in producing a complete ensemble of structures compatible with a given scattering curve.

5. ACKNOWLEDGEMENT

This work was supported in part by a grant from NSF SEIII (IIS-0502801) to CBK, AMF, and BAC.

References

1. M. E. Wall, S. C. Gallagher, J. Trewhella.: Large-Scale Shape Changes in Proteins and Macromolecular Complexes. *Annual Review of Physical Chemistry*, 2000, Oct; 51:350–380.
2. D. I. Svergun, H. B. Stuhrmann.: New developments in direct shape determination from small-angle scattering. 1. Theory and model calculations.*Acta Crystallographica Section A*, 1991, Nov; 47:736–744.
3. D. Walther, F. E. Cohen, S. Doniach: Reconstruction of low-resolution three-dimensional density maps from one-dimensional small-angle X-ray solution scattering data for biomolecules. *Journal of Applied Crystallography*, 2000, Apr; 33:350–363.
4. P. Chacón, F. Morán, J. F. Díaz, E. Pantos, J. M. Andreu: Low-Resolution Structures of Proteins in Solution Retrieved from X-Ray Scattering with a Genetic Algorithm. *Biophysical Journal*; 1998, 74, 6:2760–2775.
5. M. V. Petoukhov, D. I. Svergun: Global Rigid Body Modeling of Macromolecular Complexes against Small-Angle Scattering Data. *Biophysical Journal*;2005, May, 89, 2:1237–1250.
6. T. R. Sosnick, J. Trewhella: Denatured states of ribonuclease A have compact dimensions and residual secondary structure. *Biochemistry*; 1992, Sep, 31, 35: 8329–8335.
7. D. J. Segel, A. Bachmann, J. Hofrichter, K. O. Hodgson, S. Doniach, T. Kiefhaber: Characterization of transient intermediates in lysozyme folding with time-resolved small-angle X-ray scattering. *Journal of Molecular Biology*; 1999, May, 288 3:489–499.
8. G. Olah, R. D. Mitchell, T. R. Sosnick, D. A. Walsh, J. Trewhella: Solution structure of the cAMP-dependent protein kinase catalytic subunit and its contraction upon binding the protein kinase inhibitor peptide. *Biochemistry*; 1993, Apr, 32, 14: 3649–3657.
9. T. E. Williamson, B. A. Craig, E. Kondrashkina, C. Bailey-Kellogg, A. M. Friedman: Analysis of self-associating proteins by singular value decomposition of solution scattering data. *Biophysical Journal*; 2008, 94: 4906–4923.
10. D. I. Svergun, V. V. Volkov, M. B. Kozin, H. B. Stuhrmann: New Developments in Direct Shape Determination from Small-Angle Scattering. 2. Uniqueness. *Acta Crystallographica Section A*; 1996, May, 52, 3: 419–426.
11. D. I. Svergun, M. V. Petoukhov, M. B. Kozin: Determination of Domain Structure of Proteins from X-Ray Solution Scattering. *Biohysical Journal*; 2001, Jun, 80, 6: 2956–2953.
12. M. V. Petoukhov, D. I. Svergun: New methods for domain structure determination of proteins from solution scattering data. *Journal of Applied Crystallography*; 2003, Jun, 36, 3 Part 1:540–544.
13. A. Guinier, G. Fournet: Small-Angle Scattering of X-rays. Wiley, 1955.
14. D. I. Svergun: Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *Journal of Applied Crystallography*; 1992, Aug, 25, 4: 495–503.
15. P. B. Moore: Small-angle scattering. Information content and error analysis. *Journal of Applied Crystallography*; 1980, Apr, 13, 2:168–175.
16. O. Glatter: A new method for the evaluation of small-angle scattering data. *Journal of Applied Crystallography*; 1977, Oct, 10, 5: 415–421.
17. S. Steenstrup: Deconvolution in the presence of noise using maximum entropy principle. *Australian Journal of Physics*; 1985, 38: 319–327.
18. S. Steenstrup, S. Hansen: The maximum-entropy method without the positivity constraint – applications to the determination of the distance-distribution function in small-angle scattering. *Journal of Applied Crystallography*, 1994, Aug, 27, 4: 574–580.
19. D. I. Svergun, A. V. Semenyuk, L. A. Feigin: Small-angle-scattering-data treatment by the regularization method. *Acta Crystallographica Section A*; 1988, May, 44, 3: 244–250.
20. A. N. Tikhonov: On the stability of inverse problems. *Doklady Akademii Nauk SSSR*; 1943, 39, 5: 195–198.
21. A. N. Tikhonov, V. A. Arsenin: Solution of Ill-posed

Problems. Wiley,1977.
22. A. Guinier: La diffraction des rayons X aux trés petits angles, application á létude de phénoménes ultramicroscopiques. *Ann. Physique*; 1939, 12: 161–237.
23. D. Avis and K. Fukuda: A Pivoting Algorithm for Convex Hulls and Vertex Enumeration of Arrangements and Polyhedra. *Discrete & Computational Geometry*; 1992, 8: 295-313.
24. D. Svergun, C. Barberato, M. H. J. Koch: *CRYSOL* – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *Journal of Applied Crystallography*, 1995, Dec, 28, 6: 768–773.
25. D. I. Svergun, M. H. Koch: Small-angle scattering studies of biological macromolecules in solution.*Reports on Progress in Physics*; 2003, 66, 10: 1735–1782.
26. C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, J.M. Thornton: CATH- A Hierarchic Classification of Protein Domain Structures. *Structure*, 1997, Aug, 5, 8: 1093-1108.
27. F. M. G. Pearl, C. F. Bennett, J. E. Bray, A. P. Harrison, N. Martin, A. Shepherd, I. Sillitoe, J. Thornton, C. A. Orengo: The CATH database: an extended protein family resource for structural and functional genomics. *Nucl. Acids Res.*; 2003, 31, 1:452-455.
28. A. V. Sokolova, V. V Volkov, D. I. Svergun: Prototype of a database for rapid protein classification based on solution scattering data. *Journal of Applied Crystallography*; 2003, Jun, 36, 3 Part 1: 865–868.