ARL-TR-8576 ● Nov 2018

US Army Research Laboratory

Protein Modeling with X-ray Scatter Data

by Michael S Lee, Nick Bedford, Sasha Teymorian,
and Mark H Griep

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-8576 ● Nov 2018

US Army Research Laboratory

Protein Modeling with X-ray Scatter Data

by Michael S Lee
Computational and Information Sciences Directorate, ARL

Nick Bedford
University of New South Wales, Australia

Sasha Teymorian and Mark H Griep
Weapons and Materials Research Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
November 2018	DRI Report	October 2016–October 2017

4. TITLE AND SUBTITLE
Protein Modeling with X-ray Scatter Data

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Michael S Lee, Nick Bedford, Sasha Teymorian, and Mark H Griep

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
US Army Research Laboratory
Computational and Information Sciences Directorate (ATTN: RDRL-CIH-C)
Aberdeen Proving Ground, MD 21005

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-8576

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
In principle, X-ray scattering could provide an alternative means to characterize the structure of proteins that do not form clean periodic crystals. In this work, we compute radial distribution functions from three known X-ray protein crystal structures (green fluorescent protein, bovine serum albumin, and hen egg white lysozyme), and compare them to acquired X-ray scattering data on the target proteins.

15. SUBJECT TERMS
machine learning, pruning, network compression, explainability, image painting

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Michael S Lee
Unclassified	Unclassified	Unclassified	UU	16	19b. TELEPHONE NUMBER (Include area code) (410) 278-5888

Contents

List of Figures

Acknowledgments

We thank Dr J Hyatt for helpful discussions.

1. Introduction

X-ray crystallography is one of the primary means for deducing the 3-D atomic structure of biomolecules and small biomolecular assemblies (Smyth 2000). Tantamount to the success of this approach is crystallization of the desired species. When crystallization is not possible, nuclear magnetic resonance (NMR) and cryo-electron microscopy (Cryo-EM) are viable options. NMR is mainly useful in the smaller protein regime, although medium-sized proteins are beginning to be tackled (Wüthrich 2001). Cryo-EM works well in the protein assembly regime; however, resolution is significantly reduced compared to X-ray crystallography and NMR (Costa 2017).

When other options are not available, X-ray scattering appears to be a possible alternative, whereby the one-dimensional radial distribution function (RDF) of the electron density is obtainable. However, the RDF is significantly less informative than a 3-D electron density derived from X-ray crystallography. Therefore, it is still an unsolved problem if it is possible to derive protein-level models from RDFs. In this work, we test two techniques and three strategies for computing the RDF from known structural data and compare against experimentally observed X-ray RDF data.

2. Theory

2.1 Radial Distribution Function (RDF)

The Fourier inverse of the X-ray scattering function (Keen 2001) is estimated from a model as a pairwise summation,

$$G^x(r) = \sum_{i,j=1}^{n} c_i c_j \frac{K_i K_j}{\left(\sum_{i=1}^{n} c_i K_i\right)^2} \left[g_{ij}(r) - 1\right], \quad (1)$$

where K_i is the effective number of electrons for species *i*, which is roughly the atomic number of *i*; c_i is the fraction of species *i* in the sample; and $g_{ij}(r)$ is the partial radial distribution function of species *j* with respect to species *i*. Explicitly,

$$g_{ij}(r) = \frac{n_{ij}(r)}{4\pi r^2 dr \rho_j}, \quad (2)$$

where $n_{ij}(r)$ is the number of *j*-type entities between (*r, r* + *dr*) from an *i*-type entity. The density is $\rho_j = c_j \rho_0$, where $\rho_0 = N/V$ and *N* is the number of particles and *V* is the sample volume.

2.2 Computing the RDF with a Numerical Grid

Given that the experimentally obtained signal is averaged over many small variations of an ideal system, it might make sense to use a numerical grid with atomic position treated as a small spread. In this work, besides the pairwise strategy (Eq. 1), we also look at "painting" the atomic density onto a regular cubic grid (1400 by 1400 by 1400), where each grid point had a spacing of 0.25 Å. Each atom is treated as a sharp exponentially decaying radial basis function of electron density, $\rho(r) = Ne^{-18r}$, and interpolated onto the grid, where *N* was set to 64 * 0.915, to normalize the computed integral of each atomic function to be roughly the atomic number. The density of the grid outside of the protein atoms, $\rho(r) < 10^{-16}$, was set to be the electron number density of liquid water, 0.264 electrons/Å^3.

From this grid, the radial density function, $G^x(r)$, was computed by quadrature with a spherical grid (Lee 2002). Briefly, a spherical surface grid was composed of equal intervals of latitude, ψ, and varying intervals of longitude, θ as a function of latitude, $N_\theta(\psi) = 1 + \left(2N_\psi - 1\right) * \cos(\psi)$. In this work, N_ψ was set to 40, for a total of 1980 spherical surface sampling points. For simplicity, quadrature weights were all set to the same value of 1.0 / (# points).

3. Modeling Strategies

In this section, we describe and test three strategies for predicting the RDF derived from X-ray scattering. We compare our predictions with actual X-ray scattering data. All proteins used for X-ray scattering experiments were purchased from Sigma-Aldrich Chemicals (St Louis, MO) and used without further purification. Protein samples were dissolved in Millipore deionized-distilled water (MilliporeSigma, Burlington, MA) to the selected concentration as described elsewhere (Teymorian et al. 2018).

3.1 Strategy #1: Full Crystallographic Unit

One straightforward model for the assembly of protein monomers is the crystallographic unit found when performing X-ray crystallography of a given protein. In Fig. 1, for example, green fluorescent protein (GFP) forms a multi-chain complex.

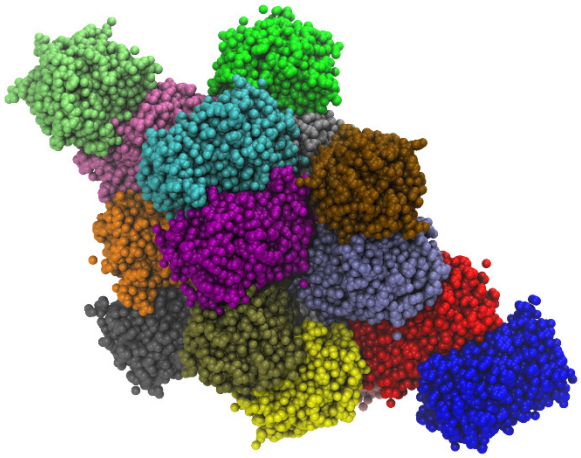


Fig. 1 Full crystallographic assembly of green fluorescent protein (GFP), protein data bank (PDB) code: 1GFL

The advantage of using a crystallographic model as a starting point is that it is firmly based in reality, as it is derived from the 3-D electron density obtained from X-ray reflections of a crystal. However, there is a major drawback when assuming a crystal-like assembly. Crystalline structures are typically very compact because the macroscopic sample is a solid. Thus, random rotations/translations of monomers to fit to experimental RDF data would be difficult. (e.g., using Reverse Monte Carlo [Aoun 2016]). At best, minor model optimizations could be made at the atomic level. In Fig. 2, the computed RDF using this model does not have density oscillations as large as the actual experimental RDF. However, some of the peaks are in the same location.

3.2 Strategy #2: Periodic box

A common method for simulating proteins on a computer is to simply place a monomer in a rectangular box and fill up the void with water molecules (Fig. 3). This roughly corresponds to a semi-dilute protein solution where neighboring monomers are all pointing in the same direction. As can be seen in Fig. 4, there is some resemblance of the predicted RDFs to the experimental ones. The predictions are far too sharp, however, which is consistent with lack of rotational incoherence (random relative orientations) between monomer neighbors and lack of allowance for natural internal vibratory modes.

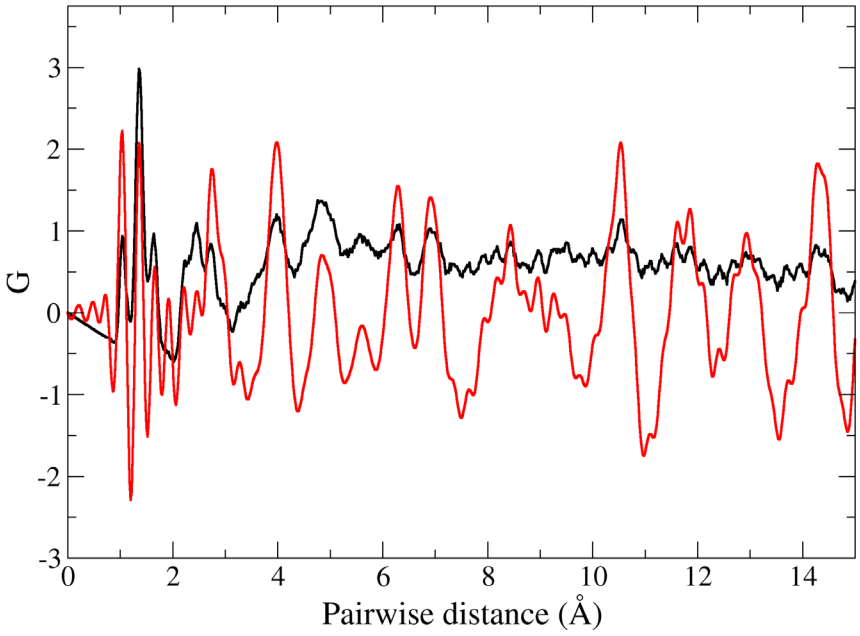


Fig. 2 Comparison of computed (black) and actual (red) RDF for GFP

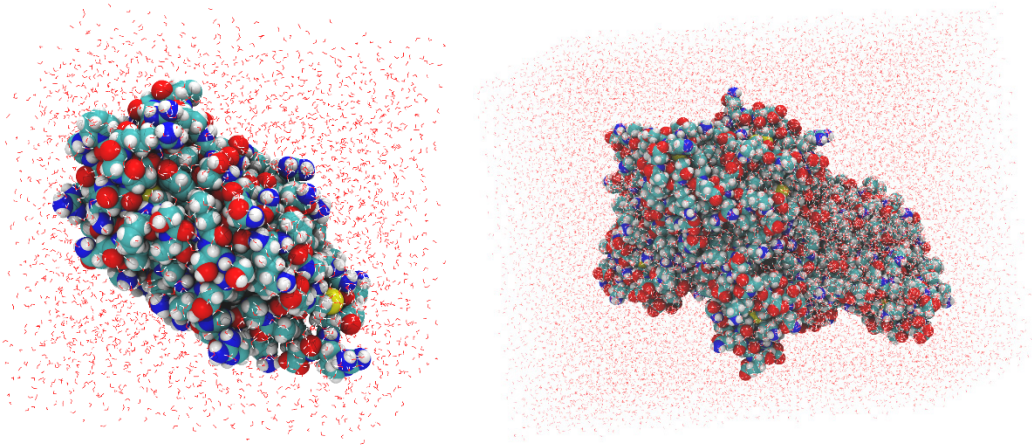
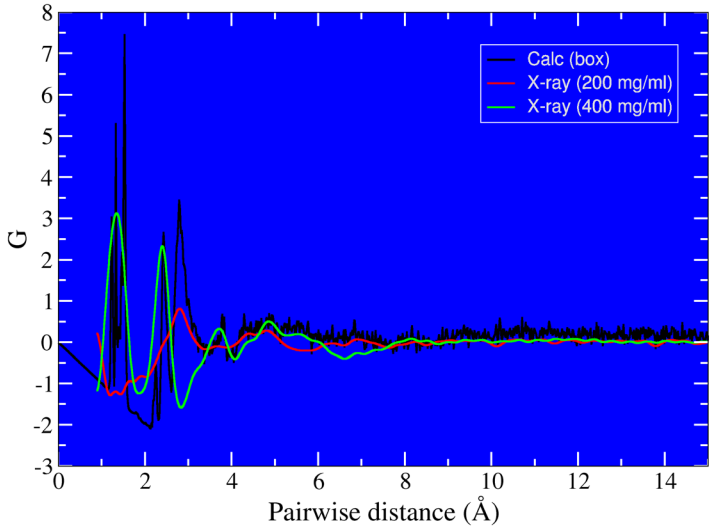
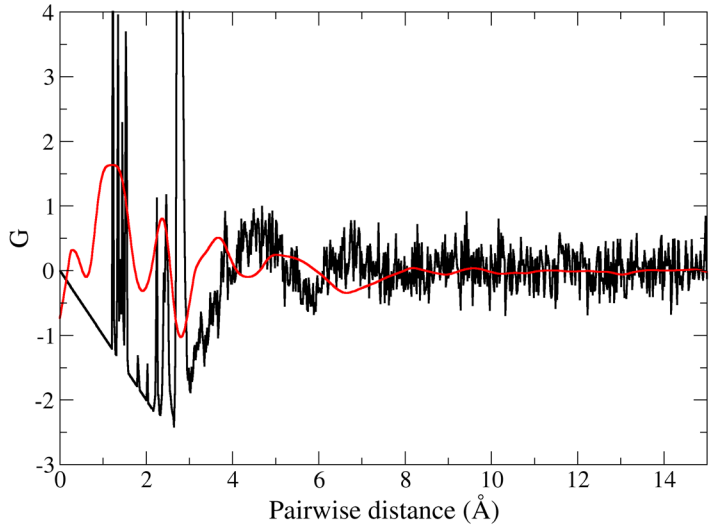


Fig. 3 a) Bovine serum albumin, PDB code: 4F5S (Bujacz 2012) and b) lysozyme, PDB code: 2VB1 (Wang et al. 2007). Each monomer is computationally solvated in a periodic rectangular box.



a)



b)

Fig. 4 Comparison of RDF from water box (black) vs. experimental (red). a) bovine serum albumin (300 mg/ml) and b) lysozyme (300 mg/ml)

3.3 Strategy #3: Protein in Continuum Solvent

Instead of treating the system as a periodic rectangular box with explicit solvent water molecules, we looked at modeling the protein with implicit solvent (i.e., a single continuum solvent density parameter) using the numerical grid approach. The results in Fig. 5 suggest that except for the short-range part of the RDF, which is very similarly predicted for two different proteins, there is not very much agreement with the experimental spectrum.

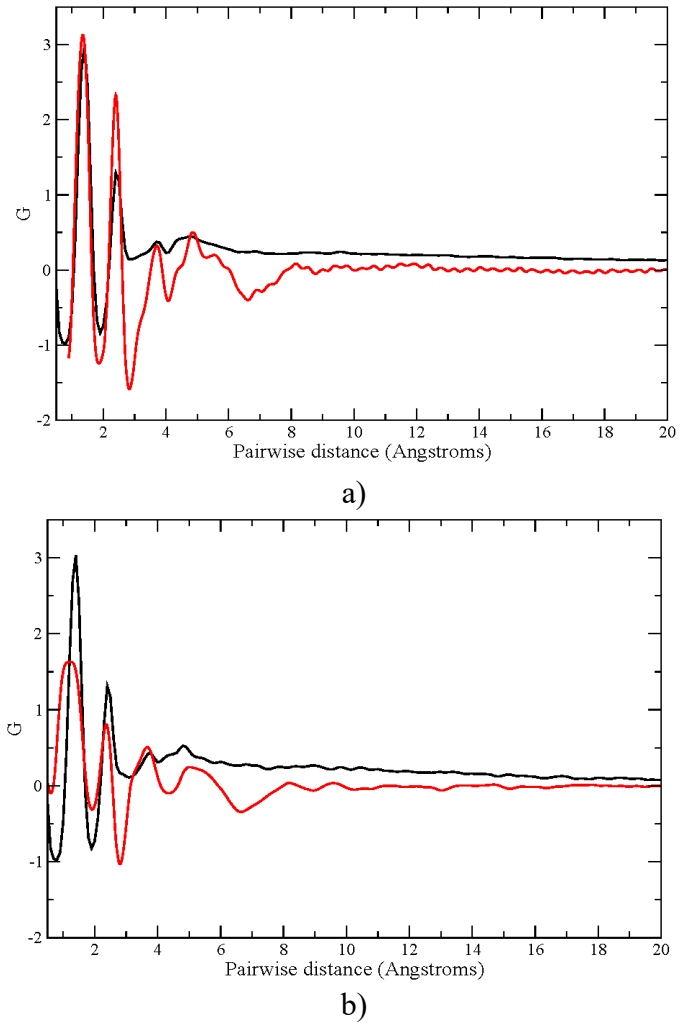


Fig. 5 Comparison of RDF from single monomer (black) vs. experimental (red). a) bovine serum albumin (400 mg/ml) and b) lysozyme (300 mg/ml)

4. Conclusions

In this work, we attempted to predict a protein system that would yield an RDF that matched X-ray scattering. While we were unsuccessful, we note that there is some advantage to using a numerical grid, as it smooths out the curve to look qualitatively like the experimental data. The short range of the RDF was modeled adequately, but the long-range tail could not be properly predicted. Our main suggestion for the future would be to consider modeling an ensemble of various protein conformations and aggregate. This would likely be computationally intensive, but potentially informative in understanding the true nature of the species being studied. Another suggestion would be to use Reverse Monte Carlo (Aoun 2016) on the protein with a simple, implicit water-density model to simplify the stochastic contribution from water molecules.

5. References

Aoun B. Fullrmc, a rigid body Reverse Monte Carlo modeling package enabled with machine learning and artificial intelligence. J Comp Chem. 2016;37(12):1102–1111.

Bujacz A. Structures of bovine, equine and leporine serum albumin. Acta Crystallogr D. 2012;68(10):1278–1289.

Costa TRD, Ignatiou A, Orlova EV. Structural analysis of protein complexes by cryo electron microscopy. Bacterial Protein Secretion Systems. New York (NY): Humana Press; 2017. p. 377–413.

Keen DA. A comparison of various commonly used correlation functions for describing total scattering. J Appl Crystallogr. 2001;34(2):172–177.

Lee MS, Salsbury FR Jr, Brooks CL III. Novel generalized Born methods. J Chem Phys. 2002;116(24):10606–10614.

Smyth MS, Martin JHJ. X-ray crystallography. Mol Path. 2000;53(1):8–14.

Teymorian S, West A, Lee M, Bedford N, Griep M. Site-specific nanocluster synthesis in energy-coupled biomolecular hosts. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2018. Report No.: ARL-TR-8371.

Wang J, Dauter M, Alkire R, Joachimiak A, Dauter Z. Triclinic lysozyme at 0.65 Å resolution. Appl Crystallogr D: Biol Crystallogr. 2007;63(12):1254–1268.

Wüthrich K. The way to NMR structures of proteins. Nat Struct Mol Bio. 2001;8(11):923.

List of Symbols, Abbreviations, and Acronyms

3-D	three-dimensional
ARL	US Army Research Laboratory
Cryo-EM	cryo-electron microscopy
GFP	green fluorescent protein
NMR	nuclear magnetic resonance
PDB	protein data bank
RDF	radial distribution function

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
2 (PDF)	DIR ARL IMAL HRA RECORDS MGMT RDRL DCL TECH LIB
1 (PDF)	GOVT PRINTG OFC A MALHOTRA
4 (PDF)	RDRL CIH S D SHIRES RDRL CIH C E CHIN RDRL WMM G J LENHART RDRL WMM A J SANDS