SYMMETRY-CONSTRAINED 3D INTERPOLATION FOR VIRUS X-RAY CRYSTALLOGRAPHY

Yibin Zheng and Peter C. Doerschuk

School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907–1285 USA
doerschu@ecn.purdue.edu

John E. Johnson

Department of Molecular Biology
The Scripps Research Institute
La Jolla, CA

ABSTRACT

An interpolation problem that is important in viral x-ray crystallography is considered. The problem requires new methods because (1) the function is known to have icosahedral symmetry, (2) the data is corrupted by experimental errors and therefore lacks the symmetry, (3) the problem is 3D, (4) the measurements are irregularly spaced, and (5) the number of measurements is large (10**4). A least-squares approach is taken using two sets of basis functions: the functions implied by a minimum-energy band-limited exact interpolation problem and a complete orthonormal set of band-limited functions. A numerical example on Cowpea Mosaic Virus is described.

1. INTRODUCTION

Viruses, like molecules, can sometimes be crystallized and their 3D structure can then be determined by x-ray crystallography. This paper considers an interpolation problem that commonly arises during structure determination for the so-called spherical viruses. Spherical viruses are viruses with a shell of protein (the so-called "capsid") surrounding an inner core of nucleic acid. The capsid is "crystalline" in the sense that it is constructed from many repetitions of the same polypeptides and the entire capsid is invariant under the rotational symmetries of the icosahedron. The icosahedron, as shown in Figure 1, is constructed from 20 equilateral triangles and has 60 rotational symmetries: a 5-fold axis where 5 triangles meet, a 3-fold axis through the center of each triangle, and a 2-fold axis at the midpoint of each edge between two triangles. A typical outer radius of the capsid is in the range 10^2–10^3 Å.

For experimental reasons, the three-dimensional crystal x-ray diffraction data sets used to refine a virus structure are incomplete. However, because of the icosahedral symmetry of the viral particle in real space, the data set in reciprocal space also displays icosahedral symmetry. This redundancy, which can be up to 60-fold, should allow the structure to be solved even though the data set is incomplete (e.g., only 20% of the data was measured). However, in order to use standard refinement algorithms and programs

This work was supported by NSF grants BIR-9513594 and DBI-9630497. Y. Zheng is now at GE Corporate R&D, Schenectady, NY.

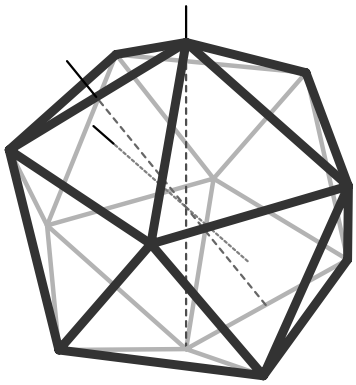

Figure 1: An Icosahedron. One symmetry axis of each type—2-, 3-, and 5-fold—is shown.

(e.g., the electron density modification algorithm [9]), an essentially complete data set is required. Therefore, based on the 60-fold redundancy, the incomplete data set is interpolated to generate a complete data set. However, the current methods do not exploit the icosahedral symmetry of the reciprocal-space data and are therefore inaccurate. In this paper we describe more sophisticated interpolation algorithms based on a least-squares point of view which exploit the icosahedral symmetry.

2. CRYSTALLOGRAPHY AND INTERPOLATION

We first recall the basic framework. For notational simplicity, we consider only the simplest case, which is a cubic P Bravais lattice in which the unit cell measures a units on a side. Let $\rho_u(\vec{x})$, with Fourier transform $P_u(\vec{k})$, be the electron density within a unit cell of the crystal. By definition, $\rho_u(\vec{x})$ is zero outside the unit cell. Let $\vec{n} = (n_1, n_2, n_3)$ be a vector in which each component takes only integer values. The electron density in the entire crystal, denoted by $\rho(\vec{x})$, can then be written $\rho(\vec{x}) = \sum_{\vec{n}} \rho_u(\vec{x} - \vec{n}a)$ and has Fourier transform

$$P(\vec{k}) \quad = \quad P_u(\vec{k}) \left(\frac{2\pi}{a}\right)^3 \sum_{\vec{n}} \prod_{i=1}^{3} \delta(k_i - \frac{2\pi}{a} n_i)$$

($\delta(x)$ is the Dirac delta function). Equivalently, since $\rho(\vec{x})$ is periodic, one can use Fourier series:

$$P_{\vec{n}} = \frac{1}{a^3}\int \rho_u(\vec{x})\exp(-i\frac{2\pi}{a}\vec{n}\cdot\vec{x})\mathrm{d}\vec{x} = \frac{1}{a^3}P_u(\frac{2\pi}{a}\vec{n})$$

and $\rho(\vec{x}) = \sum_{\vec{n}} P_{\vec{n}}\exp(i(2\pi/a)\vec{n}\cdot\vec{x})$. Let $f_u(\vec{x}) = \int \rho_u(\vec{\lambda})\rho_u^*(\vec{\lambda}-\vec{x})\mathrm{d}\vec{\lambda}$, with Fourier transform $F_u(\vec{k}) = |P_u(\vec{k})|^2$, be the autocorrelation function of $\rho_u(\vec{x})$. Define $g(\vec{x}) = \sum_{\vec{n}} f_u(\vec{x}-\vec{n}a)$ which is periodic with period a in each coordinate direction (though $f_u(\cdot)$ is non-zero over a region of size $2a\times 2a\times 2a$), has Fourier transform

$$G(\vec{k}) \quad = \quad |P_u(\vec{k})|^2\left(\frac{2\pi}{a}\right)^3\sum_{\vec{n}}\prod_{i=1}^{3}\delta(k_1-\frac{2\pi}{a}n_i),$$

and has Fourier series $F_{\vec{n}} = a^3|P_{\vec{n}}|^2 = (1/a^3)|P_u((2\pi/a)\vec{n})|^2$. Since only $F_{\vec{n}}$ is measured, the only function that can be directly reconstructed by inverse Fourier series is $g(\vec{x})$, the so-called Patterson function.

We now add the icosahedral symmetry of the virus particle. For notational simplicity, we again consider only the simplest case, which is a single virus particle in each unit cell positioned in a way that none of the 60 icosahedral symmetries are related to the space group symmetries of the crystal. The first key fact is that the Fourier transform of an icosahedrally symmetric object has icosahedral symmetry. Therefore, the icosahedral symmetry of $\rho_u(\vec{x})$ implies that $P_u(\vec{k})$ has icosahedral symmetry and therefore the measurements $F_{\vec{n}} = (1/a^3)|P_u((2\pi/a)\vec{n})|^2$ are samples on the reciprocal lattice of an icosahedrally symmetric function. Let $\vec{k}'$ be one of the 59 vectors that are symmetry related to $\vec{k}$. Therefore, $P_u(\vec{k}) = P_u(\vec{k}')$. The second key fact, due to the assumption that none of the 60 icosahedral symmetries are related by the space group symmetries of the crystal, is that $\vec{k}'$ cannot be written in the form $\vec{k}-(2\pi/a)\vec{n}$ when the components of $\vec{n}$ are restricted to integer values. Therefore, if $\vec{k}$ is on the reciprocal lattice (and therefore $|P_u(\vec{k})|^2$ is measured) then $\vec{k}'$ is not on the reciprocal lattice (and therefore $|P_u(\vec{k}')|^2$ is not measured). Therefore, the icosahedral symmetry provides 59 additional measurements of $|P_u(\vec{k})|^2$ for each measurement that is actually made. (If some of the 60 icosahedral symmetries are also in the crystal space group then, rather than 59 additional values, there is a proportionally smaller number such as 29). The third key fact is that the data is samples of $F_u(\vec{k})$ which has (inverse) Fourier transform $f_u(\vec{x})$ which is space limited to a region of size $2a\times 2a\times 2a$.

Now consider the missing data problem. In order to achieve good performance with iterative phase refinement methods it is necessary to have measured $|P_u(\vec{k})|^2$ for all reciprocal lattice points $(2\pi/a)\vec{n}$. However, as described above, this is often not possible experimentally. Therefore, we want to interpolate the missing values from the measured values. The resulting interpolated values will likely be good enough to refine the phases because we have 60-fold redundant data. The interpolation problem is difficult for the reasons described in the Abstract. Note, however, that the interpolation only needs to be done once, at the beginning of the refinement process, so substantial computation can be done. Tobacco Ring Spot Virus (J.E. Johnson, personal communication, 1995) provides a typical example: In the resolution range of 50–4Å, there were 66594 measured unique reflections which is roughly 22% of the total unique reflections in that resolution range.

3. THEORY OF INTERPOLATION

The general problem is interpolation of a band-limited function from samples of the function. This problem has a long history [6] and, if an infinite number of samples are taken on a rectangular lattice, a well-known answer in terms of "$\sin(x)/x$". Very few exact results giving explicit formulas for the interpolating function are known when an infinite number of samples are taken on a non-rectangular lattice (e.g., Refs. [5, 10]). A finite number of samples never leads to a unique interpolation without an additional constraint and numerous constraints have been studied: polynomial interpolators [1, 7], non-linear warpings of the independent variable [3], linear band-limited interpolators which minimize the energy in the interpolated function [2, 8], and interpolation motivated by the two-dimensional problems that arise in computed tomography (see Ref. [4] and the references cited therein). For the missing data problem the most natural approach appears to be linear band-limited interpolators which minimize the energy in the interpolated function.

Our approach is to consider interpolation as a least squares problem where all of the a priori information (the support and the icosahedral symmetry constraints on $\rho(\vec{x})$) are built into the basis functions and the least squares optimization problem is a natural method for dealing with the inconsistent data. In work not described here we show that a Bayesian estimation problem with Gaussian a priori model and Gaussian measurement noise is a special case of the least squares problem and that, if the data is consistent, minimum-energy band-limited exact interpolation (MEBLEI) is also a special case. We also show how to reduce the computation by the number number of symmetry elements, which is 60 in the complete icosahedral group.

There are two natural sets of basis functions. Both the MEBLEI and the Bayesian formulations suggest the basis functions $b_j(\vec{x}) = 1_{|\vec{x}|\le 2R_+}(\vec{x})(1/60)\sum_{\beta=0}^{59}\exp(iR_\beta\vec{k}_j\cdot\vec{x})$ where 1_S is the indicator function for the set S, $R_\beta\in\mathcal{R}^3$ are the orthonormal matrices that describe the symmetry operations (all rotations) of the icosahedral group, and $\vec{k}_j$ is the $\vec{k}$ location for the jth measurement. The second set of basis functions is $b_{l,n,p}(\vec{x}) = T_{l,n}(\theta,\phi)H_{l,p}(r)$ where $\vec{x} = (r,\theta,\phi)$ in spherical coordinates, $T_{l,n}(\theta,\phi)$ are icosahedral harmonics [12, 11], and $H_{l,p}(r)$ are certain linear combinations of lth order spherical Bessel functions that can be derived by examining a Sturm-Liouville problem. The two key properties of the $T_{l,n}$ functions are that every weighted sum of $T_{l,n}$ functions is a function that has icosahedral symmetry and every smooth icosahedrally symmetric function can be expanded as a weighted sum of $T_{l,n}$ functions. It is standard to require the additional properties that the $T_{l,n}$ functions are real valued and orthonormal. The three key properties of the $H_{l,p}$ functions are that for fixed l they form a complete orthonormal set, they have the correct support ($H_{l,p}(r) = 0$

for $r > R_+$), and the spherical Hankel transform of $H_{l,p}$ can be computed analytically (this transform is the radial component of the 3D Fourier transform of $b_{l,n,p}(\vec{x})$). Note that the $b_j(\vec{x})$ can be used in a least squares problem even if the data is inconsistent. The advantage of the $b_{l,n,p}(\vec{x})$ is that the resolution of the interpolating function and the size of the least squares problem can be controlled by limiting the maximum l and p that are considered while, in contrast, the number of $b_j(\vec{x})$ functions is always equal to the number of measurements.

4. NUMERICAL RESULTS

We consider interpolation problems for Cowpea Mosaic Virus (CpMV) for which the 3D atomic structure is known. The calculations are based on simulated data so that truth is known: Data for $0 < |\vec{k}| \leq k_{\max}$Å are computed, 80% of the computed measurements are deleted by a Bernoulli random process, estimates of the deleted measurements are computed by interpolation based on the remaining 20% of the computed measurements, and performance measures are computed based on the difference between the computed and interpolated measurements for the 80% of the computed measurements that were deleted. (A more realistic deletion process based on cones of retained measurements has also been considered and gives similar results). Three performance measures are considered:

$$\begin{aligned} R_2 &= \left\| G(\vec{k}) - \hat{G}(\vec{k}) \right\|_2^2 / \left\| G(\vec{k}) \right\|_2^2 \\ R_1 &= \left\| G^{1/2}(\vec{k}) - \hat{G}^{1/2}(\vec{k}) \right\|_1 / \left\| G^{1/2}(\vec{k}) \right\|_1 \\ C &= \left\| \left[G(\vec{k}) \hat{G}(\vec{k}) \right]^{1/2} \right\|_1 / \left\{ \left\| G(\vec{k}) \right\|_1 \left\| \hat{G}(\vec{k}) \right\|_1 \right\}^{1/2} \end{aligned}$$

where $G(\vec{k})$ is the data, $\hat{G}(\vec{k})$ is the interpolator in Fourier space, and the sums in the l_p norms $\| \cdot \|_p$ are over those measurements that were deleted.

Only the $b_{l,n,p}(\vec{x})$ basis functions are considered here. The parameters are $k_{\max} = 1/20\text{Å}^{-1}$, $R_+ = 160$Å, the p sum is truncated to $\lfloor k_{\max} 2R_+ \rfloor$, and the l sum is truncated to $L_{\max}$ for which a variety of values are considered. The least squares problem is solved using the singular value decomposition.

Representative results are shown in Table 1. In this example, the interpolator is estimated from 1399 retained measurements and tested on 5674 deleted measurements. In terms of R_1, which is the performance measure that crystallographers traditionally focus on, performance is doubled with $L_{\max} = 30$ which represents an interpolator with 312 parameters (there are no icosahedral harmonics for $l \in \{1$–$5, 7$–$9, 11, 13, 14, 17, 19, 23, 29\}$, one harmonic for the remaining $l < 30$, and two harmonics for $l = 30$; it is not necessary to include even l because $g(\vec{x}) = g(-\vec{x})$).

We are currently working on problems with $k_{\max} = 1/10\text{Å}^{-1}$, which have many more retained measurements; problems with more complicated relationships between the crystal's space group and the virus' icosahedral group, and problems with experimental data.

	Current	$L_{\max}$		
		10	20	30
Self R_2	N/A	0.00194	0.000198	2.72e-05
R_1	0.25	0.322	0.227	0.146
R_2	N/A	0.0756	0.00806	0.000485
C	0.8	0.969	0.988	0.995

Table 1: Performance of the interpolation based on $b_{l,n,p}(\vec{x})$ basis functions. The column "Current" represents current practice. "Self R_2" refers to the R_2 performance measure computed not over the 80% of deleted points but rather over the 20% of retained points. Since there are fewer basis functions than data points, the least squares problem does not generate an exact interpolator.

5. REFERENCES

[1] F. J. Beutler. Recovery of randomly sampled signals by simple interpolators. Info. Contr., 26:313–340, 1974.

[2] D. S. Chen and J. P. Allebach. Analysis of error in reconstruction of two-dimensional signals from irregularly spaced samples. IEEE Trans. ASSP, 35(2):173–180, Feb. 1987.

[3] J. J. Clark, M. R. Palmer, and P. D. Lawrence. A transformation method for the reconstruction of functions from nonuniformly spaced samples. IEEE Trans. ASSP, 33(4):1151–1165, Oct. 1985.

[4] H. Fan and J. L. C. Sanz. Comments on "Direct Fourier reconstruction in computer tomography. IEEE Trans. ASSP, 33(2):446–449, Apr. 1985.

[5] J. R. Higgins. A sampling theorem for irregularly spaced sample points. IEEE Trans. Info. Theory, pages 621–622, Sept. 1976.

[6] A. J. Jerri. The Shannon sampling theorem—Its various extensions and applications: A tutorial review. Proc. IEEE, 65(11):1565–1596, Nov. 1977.

[7] J. Jimenez and J. C. Agui. Approximate reconstruction of randomly sampled signals. Signal Processing, 12:153–168, 1987.

[8] L. Levi. Fitting a bandlimited signal to given points. IEEE Trans. Info. Theory, pages 372–376, July 1965.

[9] B.-C. Wang. Methods in Enzymology vol. 115, chapter Resolution of Phase Ambiguity in Macromolecular Crystallography, pages 90–112. Academic Press, New York, 1985.

[10] J. L. Yen. On nonuniform sampling of bandwidth-limited signals. IRE Transactions on Circuit Theory, pages 251–257, Dec. 1956.

[11] Y. Zheng and P. C. Doerschuk. 3D reconstruction of partially known viral structures from solution x-ray scattering data. In Proceedings of the IEEE 1996 International Conference on Acoustics, Speech, and Signal Processing, volume 4, pages 2076–2079, Atlanta, Georgia, May 7–10 1996.

[12] Y. Zheng and P. C. Doerschuk. Explicit orthonormal fixed bases for spaces of functions that are totally symmetric under the rotational symmetries of a Platonic solid. Acta Cryst., A52:221–235, 1996.