RECONSTRUCTION OF MACROMOLECULAR ENVELOPES FROM CRYSTAL X-RAY DIFFRACTION AMPLITUDES

V. L. Lo, R. P. Millane

Computational Imaging Group
Dept. Electrical & Computer Eng.
University of Canterbury
Private Bag 4800
Christchurch, New Zealand

R. L. Kingston

School of Biological Sciences
University of Auckland
Private Bag 92019
Auckland, New Zealand

ABSTRACT

The envelope problem in macromolecular x-ray crystallography involves determining the boundary of a molecule from measurements of amplitudes of x-rays diffracted from a crystalline specimen. This represents a highly underdetermined image reconstruction problem with a large number of degrees of freedom. We regularize the problem by applying binary and connectivity constraints to the image, and seek the solution using the method of iterated projections. However, since the constraints are highly non-convex, the usual methods of generalized projections are not effective. We use the difference map projection algorithm and show that this is effective with simulated diffraction data from a protein envelope.

Index Terms— X-ray crystallography, phase retrieval, iterative projection algorithms, image reconstruction

1. INTRODUCTION

X-ray crystallography is a technique for determining the atomic structures of molecules, i.e. the positions of the constituent atoms, from measurements of the amplitudes of x-rays diffracted from a crystalline sample [1]. This is achieved by reconstructing the three-dimensional electron density function of the molecule, from which the atomic positions can be inferred. The complex amplitude of the diffracted x-rays is the Fourier Transform (FT) of the electron density function so that, in principle, the former can be calculated from the latter by inverse Fourier Transformation (IFT). However, there are two practical difficulties. Firstly, only the magnitude, but not the phase, of the diffracted x-rays can be measured, thus the IFT cannot be calculated directly. This is an example of what is commonly referred to as a *phase problem* [2]. Secondly, since the specimen is crystalline, i.e. periodic, the Fourier transform is sampled, and its magnitude is undersampled, relative to the Nyquist rate, by a factor of two in each of the three dimensions [2]. The problem is therefore highly underdetermined. For large biological molecules, referred to as macromolecules, a variety of methods are used to solve this problem, most of which entail the collection of additional experimental data. Examples include the methods of multiple isomorphous replacement and molecular replacement [1]. An important intermediate step in structure determination is locating the region within the unit cell (one period of the crystal) that is occupied by the molecule, a process known as *envelope determination*. The molecular envelope is usually determined from preliminary electron density functions, calculated using experimentally-derived phases. However an alternate method for envelope determination exists, termed contrast variation, which may have advantages in certain circumstances (e.g. in the determination of large and complex macromolecular structures).

The protein molecules in a crystal are surrounded by solvent molecules that vary in position from unit cell to unit cell and so behave as a uniform electron density. It is possible, by the addition of salt to a crystal, to modify the electron density of the solvent. By making diffraction measurements from crystals with at least three different solvent electron densities, it is possible to calculate the amplitudes that would be diffracted by the molecular envelope, i.e. a function that is constant in the region occupied by the protein molecule and zero elsewhere [3]. However no general method exists for phasing these amplitudes, and recovering the molecular envelope. This is the problem that is addressed in this paper.

Our objective is to use *a priori* information, or constraints, in image space, i.e. the crystal, to compensate for the lack of information, due to undersampling and the loss of phase, in Fourier space. There are two main constraints. The first results from the envelope being a two-valued function, i.e. a voxel is either inside, or outside, the envelope. We refer to this as a *binary constraint*. The second constraint results from the molecule, and thence the envelope, being a single, connected domain. We refer to this as a *connectivity constraint*. Our approach to solving the reconstruction problem is the use of iterative projection algorithms (IPAs), which are algorithms for finding functions that are subject to a number

of disparate constraints [4]. In particular, we use the difference map (DM) algorithm [5], a variant of IPAs that is resistant to being trapped in limit cycles, a common problem if any of the constraints are non-convex (as they are in the problem considered here).

IPAs are described in the next section. Constraints and projections for the molecular envelope problem are described in the next two sections. Results from simulations are described in the next section, and concluding remarks made in the final section.

2. ITERATIVE PROJECTION ALGORITHMS

Image reconstruction problems with incomplete data can often be solved by combining the data with constraints on the image [4, 6]. However, in many practical problems the image has many degrees of freedom, and finding the solution can be extremely difficult. An effective approach for solving such problems is the method of *iterative projections* [4, 5].

It is convenient to represent the image as a point $\mathbf{x}$ in an N-dimensional Euclidean vector space R^N where N is the number of degrees of freedom (number of pixels) in the image. The full data, denoted by the vector $\mathbf{y}$, is related to the image $\mathbf{x}$ by

$$\mathbf{y} = K\mathbf{x}, \tag{1}$$

where K is some "forward" operator (which may be linear or non-linear). In general, the data $\mathbf{y}$ are not sufficient to uniquely determine $\mathbf{x}$.

If we have *a priori* information on valid images $\mathbf{x}$, then the set of valid images, denoted A, is a subspace of R^N, i.e. $A \subset R^N$. Furthermore, we can define another subspace of R^N, denoted B, as the set of all images that are consistent with the data $\mathbf{y}$, i.e.

$$B = \{\mathbf{x} : \mathbf{y} = K\mathbf{x}\}. \tag{2}$$

The image reconstruction problem then is to find an image $\mathbf{x}^*$ in the intersection of A and B, i.e. $\mathbf{x}^* \in A \cap B$.

An IPA attempts to find a point $\mathbf{x}^*$ in $A \cap B$ as follows. An initial estimate of the image $\mathbf{x}_0$ is adjusted to conform to the image space constraints and then adjusted to satisfy the data, forming the next iterate $\mathbf{x}_1$, and the process repeated. The minimum change to $\mathbf{x}_n$ is made at each stage by minimizing the Euclidean distance. The adjustments are referred to as a *projection* onto the relevant constraint set, which we denote by P_A and P_B for the sets A and B, respectively. The projection operators are then

$$P_A\mathbf{x} = \underset{\mathbf{x}' \in A}{\operatorname{argmin}} \|\mathbf{x}' - \mathbf{x}\| \tag{3}$$

and similarly for P_B.

The simplest IPA, that alluded to above, takes the form

$$\mathbf{x}_{n+1} = P_B P_A \mathbf{x}_n. \tag{4}$$

This is variously referred to as the error reduction (ER), projection onto convex sets (if the constraint sets are all convex), or generalized projection (if at least one of the constraint sets is non-convex), algorithm. The ER algorithm converges to a point in $A \cap B$ if both A and B are convex. Unfortunately, the constraint sets in most image reconstruction problems are not convex and the ER algorithm often fails to converge to a point in the intersection, which is often referred to as *stagnation*. A number of different projection algorithms have been developed that can help avoid stagnation [4]. A particularly effective and versatile algorithm is the DM algorithm, defined by the iteration [5]

$$\begin{aligned}\mathbf{x}_{n+1} = \mathbf{x}_n + \beta[&P_A((1+1/\beta)P_B\mathbf{x}_n - (1/\beta)\mathbf{x}_n) \\ &- P_B((1-1/\beta)P_A\mathbf{x}_n + (1/\beta)\mathbf{x}_n)],\end{aligned} \tag{5}$$

where $\beta \neq 0$ is a parameter. The final solution is obtained by projecting the iterate $\mathbf{x}_n$ onto A or B. This is the algorithm that we use.

3. CONSTRAINTS

There are three constraints used in the problem at hand: (1) the Fourier magnitude constraint, (2) the binary constraint, and (3) the connectivity constraint. Note that the ER and DM algorithms are based on two constraint sets. Therefore, although we discuss the binary and connectivity constraints separately, we combine these as a single constraint when they are implemented in the IPAs. The nature of these constraints is described in the following subsections.

3.1. Fourier magnitude constraint

The Fourier magnitude constraint represents satisfaction of the x-ray diffraction amplitude data. The magnitude data are sampled and so exist on a lattice in 3D Fourier space which is referred to in crystallography as the reciprocal lattice. For computational purposes, the electron density is represented on a 3D grid in image space and the reciprocal lattice coincides with the corresponding grid generated by taking the 3D Discrete Fourier Transform (DFT) of the sampled electron density. A practical consideration is that the diffracted x-ray amplitudes cannot be measured at low resolution close to the undiffracted beam, i.e. close to the origin of Fourier space. Also, diffraction data out to a maximum isotropic resolution are usually measured. Therefore, although the reciprocal lattice may occupy a cuboid domain, the Fourier magnitude constraint is applied within a spherical shell, which we denote by Q.

3.2. Binary and fill fraction constraint

Since the x-ray magnitude data have been collected from a solvent contrast series and appropriately processed, they represent the FT of the molecular envelope, i.e. a two-valued

function. With appropriate scaling of the magnitude data the electron density can be reduced to a binary function equal to 1 within the envelope region and 0 in the solvent region. We refer to this as the binary constraint and it significantly restricts the solution space. The fraction of the unit cell that is occupied by the protein molecule, or the envelope, denoted f, is rather easily determined experimentally and therefore offers an additional constraint. We refer to this as the *fill fraction constraint*.

3.3. Connectivity constraint

Protein molecules are globular structures that are held together by chemical bonds and non-bonding interactions. The individual molecules, and therefore also their envelopes, then form a single connected domain. Furthermore, the integrity of the crystal itself is supported by intermolecular contacts so that connectivity also exists between the molecules within the crystal. We refer to this as the connectivity constraint which is satisfied if all voxels of the envelope form a single connected domain, with connectivity defined in terms of an appropriately defined neighbourhood of each voxel. Connectivity is a well-defined but rather weak constraint. Protein molecules are also reasonably compact in the sense that they do not form highly tenuous connected structures. Compactness is not easily defined but some degree of compactness is incorporated with the connectivity constraint as described below.

4. PROJECTIONS

The DM algorithm requires two constraint sets. Here we develop two projection operators; one for the Fourier magnitude constraint, denoted P_M, and one for the image space constraint, denoted P_I. The projection P_I is based on two projections, denoted P_B and P_C, that incorporate the binary and connectivity constraints, respectively. The three projection operators are described in the following subsections.

4.1. Fourier magnitude projection

The image (molecular envelope) is denoted by $\mathbf{x} = (x_1, x_2, ..., x_n)$ and the DFT of $\mathbf{x}$ by $\mathbf{X} = (X_1, X_2, ..., X_n) = F[\mathbf{x}]$ where $F[\cdot]$ denotes the 3D DFT. The Fourier magnitude data are denoted by $\{M_i : i \in Q\}$. The Fourier magnitude projection in the Fourier vector space, denoted by $\tilde{P}_M$, is given by

$$\tilde{P}_M X_i = \begin{cases} M_i \exp(j\phi(X_i)) & i \in Q \\ X_i & i \notin Q, \end{cases} \tag{6}$$

where $\phi(\cdot)$ denotes the phase. The Fourier magnitude projection P_M is then given by

$$P_M \mathbf{x} = F^{-1}[\tilde{P}_M F[\mathbf{x}]]. \tag{7}$$

4.2. Binary projection

The binary and fill-fraction constraints are combined and the resulting projection is easily seen to be

$$P_B x_i = \begin{cases} 0 & x_i \notin S(f) \\ 1 & x_i \in S(f), \end{cases} \tag{8}$$

where $S(f)$ is the set of the fN largest values of $\mathbf{x}$. The binary constraint is a set of points in the image vector space and maps to a set of points in the Fourier vector space. As a result of Hermitian symmetry, the Fourier magnitude constraint is an $N/2$-dimensional hypersphere in the Fourier vector space. Therefore, the binary constraint is a relatively strong constraint, with good noise tolerance.

4.3. Connectivity projection

For a pure connectivity constraint applied to a binary image, the projection involves adding a network of "filaments" to connect any disconnected regions whose distances apart are smaller than the volume of the smallest region, and removing regions that are smaller than their distance from any other region. Rigorously calculating this projection would be computationally expensive. Furthermore, it would tend to lead to tenuous structures as opposed to compact globular structures and so is not appropriate for the problem at hand. An alternative projection is to remove all regions except the largest. The result satisfies connectivity and will tend not to be tenuous since no filaments are introduced. This is the projection that was used and is denoted P_C.

The full image space projection P_I is applied by applying P_B followed by P_C, i.e. $P_I \mathbf{x} = P_C P_B \mathbf{x}$. Note that applying P_C tends to reduce the fill fraction. However, it was observed that after a small number of iterations the image tended to consist of one large domain and a number of much smaller domains. Therefore, the change to the fill fraction after applying P_C is rather small.

5. RESULTS

The ER and DM algorithms were implemented using the projections P_M and P_I as described above. The algorithms were tested by simulation on a molecular envelope derived from a solved protein structure taken from the Protein Data Bank (PDB). The protein used was the Alkaline protease from *P. aeruginosa* [7]. The crystal lattice is orthogonal (orthorhombic) with unit cell dimensions $77.2 \times 176.7 \times 51.1$Å. There are four molecules in each unit cell related by crystallographic symmetry (space group $P2_12_12_1$). The molecular envelope was derived using standard methods [8] with an averaging radius of 8Å. The fill fraction is $f \simeq 0.35$. The envelope was represented on an $18 \times 40 \times 12$ sampling grid which gives a grid spacing of 4.3Å and approximately 9×10^3 degrees of freedom. The envelope within the unit cell is shown

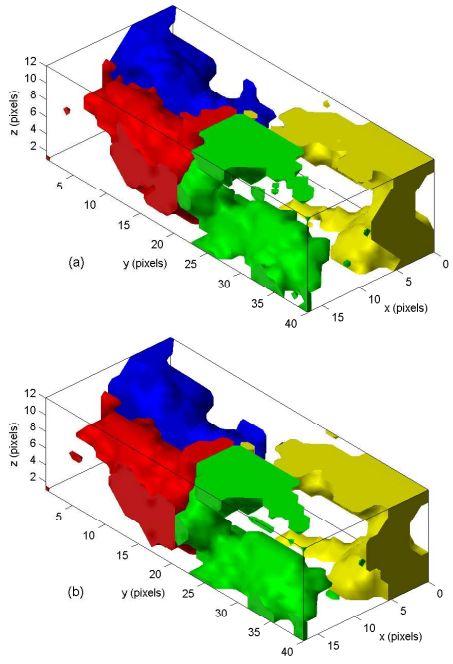

Fig. 1. Original (a) and reconstructed (b) protein envelopes. The symmetry equivalent regions are represented by different grey levels for clarity.

in Fig. 1(a). The Fourier magnitudes were calculated by the DFT, a scale factor applied, 2% Gaussian noise added, and the magnitude data retained within a resolution shell between 40 and 7Å. A 6-neighborhood was used to define connectivity. The algorithms were started with a random binary image with the correct fill fraction.

The ER and DM (with $\beta = 0.9$) algorithms were run for 5×10^5 iterations, which takes about 9 hours on one core of an Intel Q6600 CPU. The normalized squared error between the measured M_j and reconstructed $|X_j|$ Fourier magnitudes was calculated at each iteration to monitor convergence. In no case did the ER algorithm make any progress towards the correct solution. The DM algorithm made good progress towards the solution and then fluctuated around the solution. The solution with the smallest rms error was used to calculate the solution by applying the constraint P_I. The Fourier space normalized errors were approximately 5%, and image space errors were approximately 7%. The reconstructed envelope is shown in Fig. 1(b) and is seen to be a quite faithful representation of the true envelope. The Fourier magnitude error versus iteration is shown in Fig. 2.

6. CONCLUSIONS

The envelope determination problem in macromolecular x-ray crystallography is highly underdetermined but the data deficiency can be compensated for by the binary and connected nature of the image. However, both the image and data constraints are non-convex, making it difficult to find the solution. The DM IPA appears to be an effective method for obtaining the solution. Future work will involve applying this method to solvent contrast data.

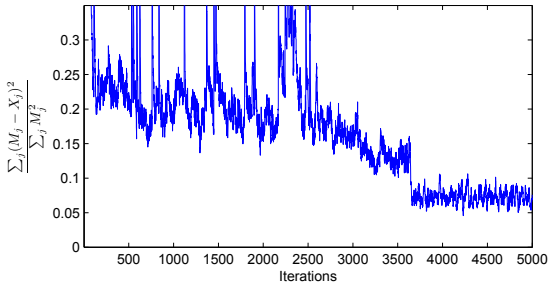

Fig. 2. Normalized Fourier transform error vs Iterations.

7. ACKNOWLEDGEMENTS

VLL is the recipient of a University of Canterbury Doctoral Scholarship.

8. REFERENCES

[1] J. Drenth, *Principles of X-ray Crystallography*. Springer-Verlag, 1994.

[2] R. P. Millane, "Phase retrieval in crystallography and optics," *J. Opt. Soc. Am. A*, vol. 7, pp. 394–411, 1990.

[3] C. W. Carter, K. V. Crumley, D. E. Coleman, F. Hage, and G. Bricogne, "Direct phase determination for the molecular envelope of tryptophanyl-trna synthetase from bacillus stearothermophilus by x-ray contrast variation," *Acta Cryst.*, vol. A46, pp. 57–68, 1990.

[4] R. P. Millane, "Iterative projection algorithms for solving inverse problems," *Proc. Oceans 2003*, pp. 2714–2719, CD-ROM, IEEE, 2003.

[5] V. Elser, "Phase retrieval by iterated projections," *J. Opt. Soc. Am. A*, vol. 20, pp. 40–55, 2003.

[6] M. Betero and P. Boccacci, *Introduction to inverse problems in imaging*. Institute of Physics Publishing, London, 1998.

[7] H. Miyatake, Y. Hata, T. Fujii, K. Hamada, K. Morihara, and Y. Katsube, "Crystal-structure of the unliganded alkaline protease from pseudomonas-aeruginosa IFO3080 and its conformational-changes on ligand binding," *J. Biochemistry*, vol. 118 (3), pp. 474–479, 1995.

[8] B. C. Wang, "Resolution of phase ambiguity in macromolecular crystallography," *Methods in Enzymology*, vol. 115, pp. 90–112 Part B, 1985.