Algebraic Reconstruction Techniques (ART) for Three-dimensional Electron Microscopy and X-ray Photography

RICHARD GORDON, ROBERT BENDER AND GABOR T. HERMAN

Center for Theoretical Biology
and
Department of Computer Science,
State University of New York at Buffalo, Amherst, N.Y. 14226, U.S.A.

(Received 12 August 1970)

We give a new method for direct reconstruction of three-dimensional objects from a few electron micrographs taken at angles which need not exceed a range of 60 degrees. The method works for totally asymmetric objects, and requires little computer time or storage. It is also applicable to X-ray photography, and may greatly reduce the exposure compared to current methods of body-section radiography.

There are a number of biological structures which have not yet been crystallized, and whose structure therefore cannot be determined by X-ray crystallography. The most outstanding example is the ribosome, whose mechanochemical functioning in the translation of messenger RNA into protein is a basic unsolved problem. Other examples are chromosomes and numerous uncrystallizable enzymes. DeRosier & Klug (1968) have given a Fourier method for the reconstruction of three-dimensional objects from electron micrographs. Unfortunately, there are limitations on their method, which make it practical only for highly symmetrical objects. They estimate that in order to obtain a 30 Å reconstruction of a 250 Å ribosome, electron micrographs would have to be taken at approximately 30 different angles, on a stage capable of tilting ±90°. This number of pictures, if taken by ordinary electron microscopy, would destroy the ribosome and cover it with a thick layer of dirt from the microscope chamber.

We will present an entirely new, direct method, an Algebraic Reconstruction Technique (ART), which has the following advantages over the Fourier method:
(1) the ART method works readily for completely asymmetric objects;
(2) it produces considerable detail of such objects with only 5 to 10 views:
(3) ordinary tilting stages may be used, since the views may be taken over
a relatively small range of angles (±30°);
(4) computing time is approximately 30 seconds per section on a Control
Data 6400:
(5) small computers may be used, since little storage is required:
(6) ART is directly applicable to macroscopic X-ray photography, and
should require considerably less radiation than present methods of
body-section radiography (Kane, 1953).

The computing requirements are emphasized, since Crowther, Amos, Finch, DeRosier & Klug (1970), studying highly symmetrical spherical
viruses by the Fourier method, needed 1·5 hours on an IBM 360/44 computer,
and apparently most of its core storage.

Using a tilt stage, we may rotate the object around a single axis. Then each
plane through the object, perpendicular to the axis, projects into a line on the
electron micrograph. We may reconstruct each plane in turn from the
densities of its corresponding lines in the different views, and stack the planes
to get the three-dimensional reconstruction. (The separation of the planes
should be no greater than the resolution of the microscope.) Thus the
problem is reduced from three dimensions to two.

Consider a square around the object in a given plane, and suppose that we
are satisfied with reconstructing this plane at \( n \times n \) points within the square.
Thus we must find the optical density \( \rho_{ij} \) at each point \((i, j), i = 1, \ldots, n;\n\)
\( j = 1, \ldots, n. \)

We define a ray of a projection or view at angle \( \theta \) as a band of width \( w \)
across the plane at that angle. (\( w \) ought to be less than or equal to the resolu-
tion of the microscope.) This ray will intersect a subset of the discrete points
defined above. Thus the total optical density of the ray across a plane of the
real object somehow has to be distributed amongst these discrete points in the
reconstruction.

Let \( R_k \) be the known total optical density of ray \( k \) of the projection at
angle \( \theta \), as measured from the appropriate photograph. Each ray yields a
linear equation

\[
R_k = \sum_{(i, j) \in \text{ray}(k, \theta)} \rho_{ij} \quad k = 1, \ldots, r_{\theta}.
\]

For instance, if \( \theta = 0° \), then \( k = j \) and

\[
R_{j, \theta} = \sum_{i=1}^{n} \rho_{ij} \quad j = 1, \ldots, n
\]

provided \( w = \) one unit of the grid. The number of rays in a given projection
at angle \( \theta, r_{\theta}, \) is of the order of \( n/w \) (see Appendix A). If we take \( p \) photo-
graphs, we have approximately \( pn \) linear equations of type (1) in the \( n^2 \) unknowns \( \rho_{ij} \).

Suppose we are using an electron microscope with 5 Å resolution to ascertain the internal structure of a 250 Å wide ribosome. In order to use the microscope to maximum advantage, we ought to choose \( n \approx 50 \) points across. If we take five photographs of the ribosome at different angles, we will have only, say, 250 equations in 2500 unknowns. The number of solutions is infinite. Yet we would like to believe that we have obtained a certain amount of information about the three-dimensional structure. The problem is how to make this visible.

The real, but unknown structure of the ribosome is, of course, one of the numerous solutions to our undetermined equations. If we had some reason to choose one solution over another, we might get reasonably close to the real structure.

We have previously presented three Monte Carlo algorithms (Gordon & Herman, 1970) for finding solutions to equations (1). In these, the optical densities \( \rho_{ij} \) were quantized, and single density bits were added or subtracted to an originally blank array (\( \rho_{ij} = 0 \)), until the equations were satisfied. In order to test the algorithms, we used a digitized picture of a little girl, Judy, to represent one plane of the three-dimensional object (Plate I). The results of the Monte Carlo algorithms were reconstructions which looked something like the original, but were highly “peppered”. Thus it is apparent that not all solutions are “smooth”.

One way we could get a smooth reconstruction was to average a number of individual peppered reconstructions. This procedure is valid, because the average of two or more solutions to equations (1) is also a solution. Thus it is clear that smooth solutions exist. This is presumably a desirable feature of reconstructions of real objects. Moreover, the average of a number of peppered reconstructions was closer to the original picture \( \{ \rho_{ij} \} \) in terms of the Euclidean “distance”

\[
\delta = \left[ \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\rho_{ij} - \bar{\rho}_{ij})^2 \right]^{1/2}.
\]  

We also found that the “entropy”

\[
S = \frac{-1}{\ln n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left( \frac{\rho_{ij}}{T} \right) \ln \left( \frac{\rho_{ij}}{T} \right)
\]  

was nearly maximized by averaging, which bolstered out intuition that the best solution would be the one which maximized entropy, or alternatively, gives us no more “information” than we put in; i.e. the solution maximizing equation (4) would seem to be the “least biased”. (\( T \) is the total density of the picture.) These notions open up some unanswered questions about the
information content of pictures, which we have discussed in detail (Gordon & Herman, 1970).

While trying to discover mathematically why the Monte Carlo algorithms worked at all, we argued in terms of the mean result of the random addition or subtraction of bits (Gordon & Herman, 1970). We have since realized that this mean could be used directly to give a highly simplified algorithm.

Let the total density of a plane be

$$T = \sum_{k=1}^{ra} R_{\theta \theta} = \sum_{i=1}^{n} \sum_{j=1}^{n} \rho_{ij}$$

(5)

for any \( \theta \). Our new procedure will be iterative, starting with an initial "guess" for the \( \rho_{ij} \):

$$\rho_{ij}^{0} = \frac{T}{n^2} \quad i, j = 1, \ldots, n$$

(6)

and refining this guess by one of the following alternative convergent procedures. Let \( R_{\theta \theta} \) be the optical density of ray \( k \) of projection \( \theta \) at iteration \( q \). Let \( N_{\theta \theta} \) be the number of grid points included in ray \( (k, \theta) \). In symbols

$$R_{\theta \theta}^{q} = \sum_{\text{the } N_{\theta \theta} \text{ points } (i, j) \text{ in ray } (k, \theta)} \rho_{ij}^{q} \quad k = 1, \ldots, r_{\theta}$$

(7)

where \( \rho_{ij}^{q} \) is the \( q \)th iterate of (or approximation to), \( \rho_{ij} \). By the direct multiplicative method the density of a point \( (i, j) \) is changed according to the formula

$$\rho_{ij}^{q+1} = \left( \frac{R_{\theta \theta}^{q}}{R_{\theta \theta}^{q}} \right) \rho_{ij}^{q}$$

(8)

for \( (i, j) \) in ray \( (k, \theta) \). By the direct additive method

$$\rho_{ij}^{q+1} = \max \left[ \rho_{ij}^{q} \left( R_{\theta \theta}^{q} - R_{\theta \theta}^{q} \right) / N_{\theta \theta}, 0 \right].$$

(9)

These formulae are applied to all rays of a given projection at angle \( \theta \), and then used for the next projection, etc.

The multiplicative method temporarily restores each ray to its exact value, since

$$\sum_{(i, j) \in \text{ray } (k, \theta)} \rho_{ij}^{q+1} = R_{\theta \theta} \quad k = 1, \ldots, r_{\theta}.$$  

(10)

However, when subsequent angles are considered, equations (10) will no longer hold for the previous angles. Thus formula (8) must be applied iteratively until the \( \rho \)'s have converged to stable values.

The additive method does not exactly satisfy equations (10) when some of the densities, which would be negative, are set to 0. This "error" can be made up in subsequent iterations. [The additive method is equivalent to the mean of our Monte Carlo Algorithm 1 (Gordon & Herman, 1970).]
In practice, we found that for a $49 \times 49$ grid for Judy's picture, both methods are convergent within 20 or so iterations per projection, using either the Euclidean distance $\delta$ or the entropy $S$ as a criterion (Figs 1, 2, Plate II). [Of these only the latter may be used for unknown objects. Note also that the entropy decreases with the number of iterations. This is exactly the opposite of what we have observed in the Monte Carlo methods of Gordon & Herman (1970). The reason for this is that in the previous method every iteration satisfied all the projections and successive iterations just gave us smoother pictures, while here we start with a perfectly smooth picture and successive iterations force us to satisfy the original rays more and more precisely, and thus reduce the smoothness of the picture. Of course, the degree of match of the $R_{\phi_0}$ to the original rays, $R_{\phi_0}$, could also be used to measure the progress of the calculation.] The computing time was 0·30 sec per iteration per projection on a CDC 6400 for the multiplicative method, and 0·32 sec for the

![Graph](image-url)

**Fig. 1.** The deviation $\delta$ [equation (3)] vs. iteration for the pictures of Judy in Plate II, (top curves), and $\delta - 1$ for the crater (bottom set of curves), both using the direct additive method. The number of evenly spaced angles (projections) between $+30^\circ$ and $-30^\circ$ is indicated for each curve.
additive method, using FORTRAN. The working storage required is only \( \approx n^2 + pn \), or less than 3000 words.

Plate I also shows reconstructions of a lunar crater redigitized by hand from Tzannes, Spencer & Kaplan (1970). It is apparently simpler in structure, since the reconstructions converge sooner than Judy's (Fig. 1).

Both methods give smooth reconstructions. It is difficult to choose between them. The multiplicative method gives a higher entropy, and therefore would intuitively seem to be "less biased". Indeed the multiplicative method does maximize the entropy in the case of only two projections (Gordon & Herman, 1970). On the other hand, for the pictures of Judy and the crater, at least, the additive algorithm gives a lower value of \( \delta \) (Fig. 3), and the reconstruction also looks a little more like the original (Plates I and II).

Some of our recent studies indicate that while the multiplicative method seems to maximize the entropy, the additive method seems to minimize the function

\[
V = \sum_{i=1}^{n} \sum_{j=1}^{n} (\rho_{ij} - T/n^2)^2. \tag{11}
\]

This function is the variance of the gray levels in the picture, and may be considered a measure of the non-uniformity of the picture. If we consider \( 1/V \) to be a measure of the uniformity of the picture, it is easy to prove that the pictures which maximize uniformity and entropy respectively will in general be near to each other. The question whether optimization of uniformity

![Figure 2](image-url)
PLATE I. Under each picture is the number of equally spaced angles between $-30^\circ$ and $-30^\circ$ used for the reconstruction. The $0^\circ$ direction is parallel to the rows.

PLATE I. Row A: original picture of the lunar crater and reconstructions by the direct multiplicative method. The crater is digitized at 50 x 50 points.
Row B: reconstructions of the crater by the direct additive method and the original picture of Judy.
Row C: multiplicative and additive reconstructions of Judy by the fast arithmetic reconstruction technique on the left and actual moiré patterns on the right, done by a Calcomp plotter.
Plate II. A series of reconstructions of Judy done by the direct multiplicative method (top two rows) and the direct additive method (bottom two rows).
Fig. 3. The deviation $\delta$ [equation (3)] vs. the angular increment of the evenly spaced angles between $+30^\circ$ and $-30^\circ$ for pictures of Judy. The top line is the additive fast arithmetic reconstruction technique, equation (C1). The middle and lower lines are the direct multiplicative and additive methods, respectively, at the 20th iteration. The extrapolation to $\Delta \theta = 0$ is speculative. The number of angles is indicated.

rather than optimization of entropy should be the aim of picture reconstruction algorithms, and a discussion of some further algorithms based on the uniformity function, will be the subject matter of future work (Gaarder, Gordon & Herman, manuscript in preparation). Two especially simple algorithms are mentioned in Appendix C.

Finally, it should be noted that both methods work in a small range of angles (Plates I and II). In particular, as the number of projections approaches $n$, the original object is almost completely restored (Fig. 3, Plates I and II). This is reasonable, since with enough linearly independent equations, they ought to be determined, and the original object ought to be completely reconstructed. However, for practical purposes five projections appear to be sufficient for a $50 \times 50$ picture.

It is interesting that the accuracy of these reconstructions cannot be discussed in terms of point to point resolution, since an object consisting of two points a distance $w$ apart will almost always be perfectly reconstructed.
[In fact, a higher resolution than that of a single electron micrograph is obtainable (Gaarder, Gordon & Herman, manuscript in preparation).]

DeRosier & Klug (1968) estimate that 30 views over a 180° span would be necessary for 30 Å resolution of a 250 Å ribosome, which would correspond to solving for the $\rho_{ij}$'s on an 8 x 8 grid.† It is clear that we are doing considerably better than this with only five views over a 60° span (Plate II). Reconstructions of individual ribosomes by the ART method will be presented in the following paper (Bender, Bellman & Gordon, 1970).

In body-section radiography (Kane, 1953) the X-ray source and the film are moved in a coordinated fashion so that only one plane in the patient in between does not blur out. If our methods were used instead, the X-rays need only go across the plane of interest. The tissues above and below need not be exposed. By photometric reading of a fluorescent screen, the intensities could be passed directly to a small computer, and the reconstructed section displayed on a television screen within a minute or so. In effect, our methods provide rapid cross-sectioning of an object, without cutting.

We would like to thank Professor Cyrus Levinthal for suggesting the problem, Karen Gordon for helping digitize the picture of Judy by hand, Judith Carmichael for providing the photograph, and the SUNY/B Computer Center for computing time. We thank Dr Thomas Gaarder for pointing out the relation between the additive method and the uniformity function. We also thank Mr Arthur R. Axelrod and Dr Pierre A. Lavallee of the Xerox Computer Science Laboratory in Rochester, New York for helping us use a Xerox LDX output device for producing the halftones. Supported in part by NASA grant NGR 33–015–016 to the Center for Theoretical Biology, the State University of New York through the Einstein chair budget of Dr C. H. Waddington, and N.S.F. Grant GJ596. Dedicated to Diana Gordon, artist.

REFERENCES


† The Fourier method may be applicable to tiny, naturally occurring ribosome crystals; see Morgan (1968), but cf. DeRosier & Klug (1969).
APPENDIX A

Definition of Rays

The formulae for a family of straight lines at angle $-\theta$ from the $x$ axis, spaced a distance $w$ apart, the first of which goes through the point $(1, 1)$, is

$$y = x \tan \theta + 1 - \tan \theta + kw \cos \theta \quad k = 0, 1, \ldots, r_\theta \quad (A1)$$

The $k$th ray is defined as the band between lines $k-1$ and $k$. The number of rays which intersect a square of $n \times n$ points $\{(1, 1), (1, n), (n, 1), (n, n)\}$ is

$$r_\theta = \left\lfloor \frac{(n-1)(\sin |\theta| + \cos \theta)}{w} \right\rfloor + 1 \quad (A2)$$

where the brackets denote truncation to the next lowest integer. For an angle $+\theta$, a similar family of lines is considered, the first of which goes through the point $(n, 1)$.

In order to find the discrete points intersected by a given ray, we first calculate $i_{\min}$ and $i_{\max}$, the lowest and highest $y$ values at which the ray intersects the square. $j$ is then varied from $j_{\min}$ to $j_{\max}$ in steps of 1. For each value of $j$, we calculate anew $i_{\min}$ and $i_{\max}$, the lowest and highest integer $x$ values which lie within the ray on a horizontal line $y = j$, by solving for $x$ in equations (A1) with parameters $k-1$ and $k$. $i$ is varied from $i_{\min}$ to $i_{\max}$ in steps of 1, and the coordinates $(i, j)$ added to a list. Points lying exactly on line $k-1$ [equation (A1)] are counted as being in ray $k$.

This method of computing the list of points in a ray is so rapid that we did not bother to store the coordinates, but rather generated them as needed. The previous method (Gordon & Herman, 1970) was exceedingly slow. The new method is easily generalized to nonparallel rays, which may occur in X-ray photography.

APPENDIX B

A Generalized Algorithm

A general iterative step may be written

$$\rho_{ij}^{t+1} = \max \left[ 0, \sum_{m=0}^{\infty} A_{k\theta m} (\rho_{ij}^t)^m \right]$$

for $(i, j)$ in ray $(k, \theta)$, with the constraint of equation (10). Ignoring the truncation when $\rho_{ij}^{t+1}$ is set to zero, by summation we find

$$R_{k\theta} = \sum_{m=2}^{\infty} A_{k\theta m} \sum_{(i, j) \in \text{ray} (k, \theta)} (\rho_{ij}^t)^m + A_{k\theta 1} R_{k\theta} + A_{k\theta 0} N_{k\theta}$$

(B2)

Only one of the $A$'s is determined. Any $\{A_{k\theta m}\}$ satisfying (B2), which defines a continuous transformation in (B1), will presumably, under iteration, make
the $\rho$'s converge to a smooth solution of equations (1). The multiplicative method is the special case

$$A_{k\theta} = 0, \quad A_{k\theta_1} = R_{k\theta}/R_{k\theta_1}^\alpha, \quad A_{k\theta m} = 0$$

for $m \geq 2$, and the additive method corresponds to

$$A_{k\theta} = (R_{k\theta} - R_{k\theta_1}^\alpha)/N_{k\theta_1}, \quad A_{k\theta_1} = 1, \quad A_{k\theta m} = 0$$

for $m \geq 2$, so that both are linear transformations.

If any transformation is used with $A_{k\theta m} \neq 0$ for $m \geq 2$, then additional computing time will be necessary to calculate $\sum_{i,j} (\rho_{ij})^n$. The visual properties of some such nonlinear iterations are currently under investigation. Except for the intuitive notions that one might want a solution which maximizes either the "entropy" or the "uniformity", we do not yet have criteria for determining the "most objective" transformation which should be used in all cases. The formal mathematical justification for why our method is successful must await a full analysis of the properties of solutions of undetermined simultaneous linear equations, careful study of the "information content" of pictures (Gordon & Herman, 1970), and analysis of what characterizes those matrices $P_{ij}$ which are likely to arise from real pictures (Gordon & Herman, 1970; Gaarder, Gordon & Herman, manuscript in preparation).

### APPENDIX C

**The Fast Arithmetic Reconstruction Techniques: Moiré Methods**

The reconstructions by the third Monte Carlo algorithm (Gordon & Herman, 1970) and by all of the algorithms for small numbers of projections exhibit streaking in the directions of the projections. This inspired us to devise a simple rough algorithm, which can be carried out with a straight edge and a bit of arithmetic, and is thus designated the fast arithmetic reconstruction technique: for each ray, within its corresponding band across a square, we draw a set of evenly spaced lines. The number of lines is proportional to the value of the ray.

Where the lines from different projections cross, higher densities will be obtained. This method thus reconstructs the object as a moiré pattern. Such reconstructions are shown in Plate I. Some of the major features of the originals are reasonably accurately located.

Plate I also includes a halftone portrayal of the approximately equivalent algorithm

$$P_{ij} \propto \sum_{\theta} R_{k\theta}$$

where $k$ is the ray of projection $\theta$ containing the point $(i, j)$. The $\rho$'s are
normalized so that their total density equals that of the original. These algorithms do not satisfy equations (1). They also show little improvement beyond 5 to 10 projections (Fig. 3).

The contrast of the above additive moiré algorithms is quite low. More detail is obtained in parts of the picture by a multiplicative "moiré" algorithm

\[ \rho_{ij} \propto \prod_{\theta} R_{k\theta} \quad (C2) \]

although other parts have exceedingly high densities (Plate I).

These methods seem to reconstruct the geometry of the original picture, but with erroneous densities. We are investigating contrast-enhancing techniques and other moiré algorithms which may make this approach useful. It may be shown that even the above additive moiré methods are as accurate as current methods of body section radiography (Gaarder & Herman, manuscript in preparation).