Pictorial Pattern Recognition and the Phase Problem of X-ray Crystallography

Arthur M. Lesk
Princeton University*

The availability of interactive, three-dimensional, computer graphics systems coupled to powerful digital computers encourages the development of algorithms adapted to this environment. Pictorial pattern recognition techniques make possible a number of approaches to X-ray structure determination based on molecular model building, i.e. the use of chemical information to frame "structural hypotheses" which can computationally be tested and refined by reference to the experimental data.

Application of standard pattern recognition algorithms is hindered by the fact that the cross-correlation between a model and the correct structure cannot be computed because of a fundamental incompleteness in the measured data. However, it is possible to compute an upper bound to such a cross-correlation. A simple example demonstrates that this information can be the basis of a technique for structure determination that can make effective use of an interactive graphics system.

Model building by cross-correlations has intrinsic advantages over usual crystallographic techniques based on the autocorrelation or Patterson function, especially for large structures. This is significant, for crystallography of biological macromolecules has been and will continue to be a field of intense interest.

Key Words and Phrases: pictorial pattern recognition, phase problem, X-ray crystallography, interactive graphics

CR Categories: 3.13, 3.17, 3.63

The Phase Problem of X-ray Crystallography

Powerful algorithms are essential in the determination of chemical structures by X-ray crystallography for several reasons. First, the measurements are *indirect*, since the data contain information about the Fourier transform of the electron density distribution in a molecule. Second, the measurements are *incomplete* in a fundamental sense, in that they fix the absolute magnitudes of the Fourier coefficients but in general give no information about their arguments.

This ambiguity leads to "the phase problem," of determining the arguments (or phase angles) of the Fourier coefficients. Once phase angles are known, a straightforward Fourier transformation produces the electron density distribution itself. A variety of techniques for phase determination, both experimental and computational, have led to successful structure determinations [3].

Often, a crystallographer knows a good deal about a compound before attempting to solve its structure by X-ray diffraction. If he can construct even a rough model for his structure, the determination of the location and orientation of the model that best fits the data is a useful step toward a solution, since phases taken from the Fourier transform of a properly positioned model may be applied to the experimental data. Many crystallographic structure determinations utilize such a com-

* Department of Biochemical Sciences, Princeton University, Princeton, NJ 08540. This is the second in a series of papers on the phase problem of X-ray crystallography. This work was supported by National Institutes of Health, U.S. Public Health Service research grant GM-16539, and U.S. National Science Foundation research grant GE-7857. This work made use of computer facilities supported in part by National Science Foundation grants NSF-GJ-34 and NSF-GU-3157.

putation of an approximate electron density map, derived by combining a set of approximated phases for the Fourier coefficients with measured magnitudes. The crystallographer may then select for further analysis those features he believes to be correct and reject those he discredits.

Pepinsky, Vand, and co-workers explored the application of interactive computer displays to the phase problem approximately 20 years ago, using an analog computer called X-RAC [7–9]. Although newer digital computers superseded that equipment in size and speed, the power of interactive displays was recognized and demonstrated then.

These considerations suggest that pictorial pattern recognition techniques might effectively be applied to X-ray structure determination using modern interactive computer graphics systems [4, 17].

Patterson (Autocorrelation) Techniques

The Patterson or autocorrelation function has been an extremely powerful and widely used tool for introducing chemical information into a crystallographic structure determination [1, 2, 6, 12]. The Patterson function is computable directly from measurable quantities, requiring no information about phases. It has a meaningful structural interpretation: a peak in the autocorrelation function corresponds to an interatomic vector somewhere within the structure. If there are N atoms in a structure, there are N^2 peaks in the Patterson function (N of which pile up at the origin, corresponding to the vectors from each atom to itself—these may be subtracted). Thus the complexity of a Patterson function increases very rapidly with the size of the structure.

Techniques based on cross-correlations, in which a structural model serves directly as a pattern, have intrinsic advantages over the autocorrelation approach, especially for macromolecules. Although the cross-correlation between an unknown structure and a pattern is not computable without knowing phases for the structure, an *upper bound* to the cross-correlation can be computed by transferring phases from the model to the structure. In at least some cases this information is sufficient for accurate positioning of a structural model. Such an approach is closely related to that of Stout, et al. [15, 16] and to rigid body refinement techniques [13].

Pattern Recognition by Maximal Normalized Cross-Correlation

The normalized cross-correlation between two functions f and g is a convenient, scale independent measure of their agreement; it is analogous to the cosine of the angle between two vectors [11]:

$$\cos\theta(f, g) = \int f(\mathbf{x})\cdot g(\mathbf{x})\,d\mathbf{x}/[\int |f|^2\,d\mathbf{x}\cdot\int |g|^2\,d\mathbf{x}]^{1/2}.$$

The functions considered here are $\rho(\mathbf{x})$, the electron density distribution in a unit cell of the crystal, and $\rho_m(\mathbf{x})$, a charge distribution corresponding to some model. The magnitudes of the Fourier coefficients of the structure are denoted by F_o, and those of the pattern by F_c.

To locate the model in the unit cell so as best to approximate the structure, a translation vector $\mathbf{t}$ and a rotation matrix $\mathbf{R}$ are sought to maximize $\cos\theta(\rho(\mathbf{x}), \rho_m(\mathbf{R}\cdot\mathbf{x} + \mathbf{t}))$. This can be done equivalently in reciprocal space:

$$\cos\theta(\rho, \rho_m(\mathbf{R}, \mathbf{t})) = \Sigma F_o\cdot F_c(\mathbf{R}, \mathbf{t})/[\Sigma |F_o|^2\cdot\Sigma |F_c|^2]^{1/2}$$

in which $F_c(\mathbf{R}, \mathbf{t})$ are the Fourier coefficients of $\rho(\mathbf{R}\cdot\mathbf{x} + \mathbf{t})$.

$\cos\theta(\rho, \rho_m(\mathbf{R}, \mathbf{t}))$ cannot be evaluated without knowing phases for the Fourier coefficients of ρ. An upper bound is available by assigning to the Fourier coefficients of ρ the phases of those of the model

$$\cos\theta(\rho, \rho_m(\mathbf{R}, \mathbf{T})) \leq \Sigma |F_o|\cdot|F_c(\mathbf{R}, \mathbf{t})|/[\Sigma |F_o|^2\cdot\Sigma |F_c|^2]^{1/2}.$$

The quantity on the right-hand side is larger than or equal to the normalized cross-correlation of the model with *any* function consistent with the measured Fourier coefficient magnitudes. If for some choice of $\mathbf{R}$ and $\mathbf{t}$ the bound is *small*, then a good match between the structure and the model placed with that position and orientation is ruled out; if *large*, then for some choice of phases the model can be well matched.

Thus a peak in the maximal normalized cross-correlation as a function of positional or orientational parameters suggests the proper placement of a pattern in the unit cell.

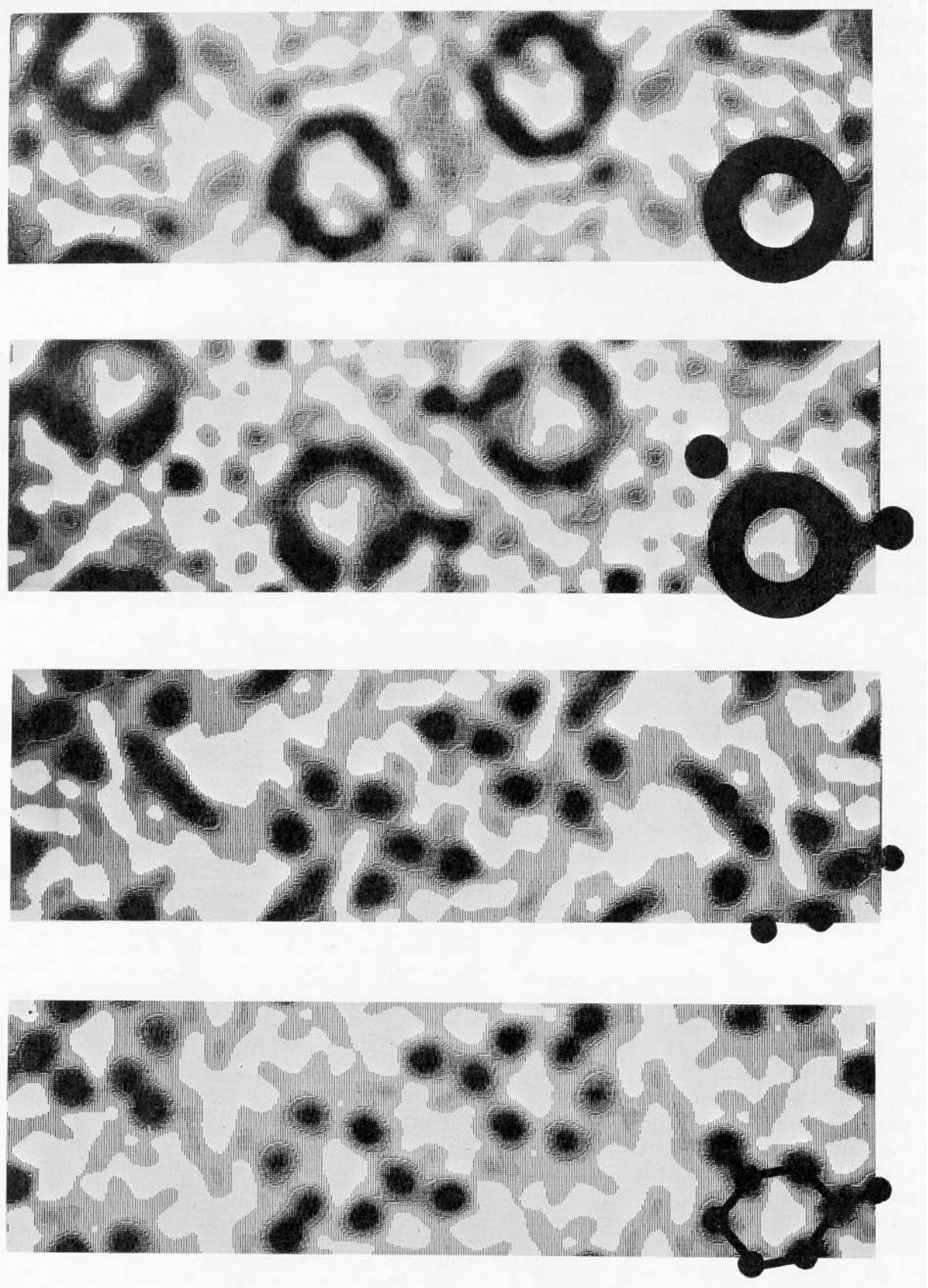
An Example: $0k\ell$ Projection of Dihydrouracil

Dihydrouracil, a structure solved recently by other methods, is particularly favorable for the application of pattern recognition techniques since all atoms are visible in projection [10]. The cell dimensions are $a = 4.201$ Å, $b = 5.816$ Å, $c = 19.777$ Å, $\beta = 95.15°$, the $0k\ell$ projection has symmetry *pgg*, and there are four molecules per unit cell.

One possible route to the solution of this structure proceeded in four steps, using progressively more refined patterns. In Figure 1, the pattern used in each step is shown set into the structure obtained by combining

Fig. 1. Solution of $0k\ell$ projection of dihydrouracil by recognition of progressively refined patterns. Patterns shown over structures computed by Fourier transformation from structure factor intensities measured by X-ray diffraction and phases determined from the patterns. The computer program used to produce the half-tone drawings differed only in minor details from one coded by I.D.G. MacLeod [5].

The final frame is in effect a portrait of the electron density distribution in dihydrouracil at approximately 100 million times magnification.



measured Fourier coefficient magnitudes with model phases.

The first pattern is an annulus of inner radius 0.85 Å and outer radius 1.7 Å (plus three other identical annuli generated by the unit cell symmetry). Its symmetry reduces the problem to a two-variable search for its position. There are 16 peaks in the maximal normalized overlap, which form four sets, each corresponding to four molecules related by symmetry. Each set corresponds to a different possible choice of origin in the unit cell. All are physically equivalent.

With the ring in a position corresponding to one choice of origin, two point atoms are placed outside the ring, 120° apart, and rotated in tandem through 360° to locate the angular positions of the oxygen atoms. This model consisted of the annulus plus two atoms.

The next step is the replacement of the annulus by a regular hexagon of atoms. Appropriate form factors were assigned to carbon, nitrogen, and oxygen. In this model, all bond lengths and angles are equal, but those in the structure computed from it are not. This stage already satisfies a criterion stated by Stout and Jensen [14]: "... a structure not containing heavy atoms can generally be completed without excessive difficulty if 50 to 75 percent of the electron density is located within an average error of about 0.3 Å."

The fourth section of the photograph shows the structure partially refined. The maximum deviation of any atomic coordinate from those reported is 0.16 Å.

Discussion and Conclusions

An approach to the phase problem of X-ray crystallography employs a cross-correlation technique to test and refine structural models. This has certain intrinsic advantages over some standard crystallographic techniques based on the Patterson or autocorrelation function: it avoids the quadratic dependence in complexity of the Patterson function upon the size of the structure; it does not require the identification of individual interatomic vector peaks in the Patterson; and models for fragments of a complex structure may be combined linearly.

A test of the method in a simple two-dimensional case was successful, demonstrating that the technique can solve a real structure using real data. The computer time required for this example was trivial—the cost of evaluating all cross-correlations was less than $10, at about $500 per hour—which suggests that more complex cases will also be tractable. The generalization from two-dimension to three-dimension requires finding the maximum of a function of at most six variables, a task within the reach of current algorithms.

As pointed out in the test example, computation can be simplified in two ways: by the choice of symmetrical patterns, at least in the early stages of analysis; and by the building up of a complex structure by "synthesis" of models of simpler fragments. Choice of the route for such a synthesis may advantageously be left to the discretion of a skilled crystallographer, and it could be carried out particularly effectively using an interactive computer graphics system.

It is hoped that this approach will be helpful in solving structures of biological macromolecules, for which models are generally available but for which the Patterson functions are quite complicated to analyze.

Acknowledgments. I am grateful to Miss J. Kaufman for expert technical assistance, Professor R. Langridge and Drs. E. and P. Huber and J. Madden for criticism of the manuscript, and R. Mathews, J. Thornberry, and Dr. J. Haga for preparation of the photograph.

Received February 1971, revised April 1971

References

1. Ahmed, F.R., Hall, S.R., and Huber, C. (Eds.) *Crystallographic Computing*. Munksgaard, Copenhagen, 1970, Topic B, pp. 81–123.
2. Harrison, H.R. Application of Patterson methods to the solution of molecular crystal structures containing a known rigid group. *Acta Cryst. A26* (Nov. 1970), 692–694.
3. Karle, J. The phase problem in structure analysis. *Adv. Chem. Phys. 16* (1969), 131–222.
4. Levinthal, C. Molecular model-building by computer. *Sci. Amer. 214* (June 1966), 42–52.
5. MacLeod, I.D.G. Pictorial output with a line printer. *IEEE Trans. Comput. C-19* (Feb. 1970), 160–162.
6. Nordman, C.E., and Nakatsu, K. Interpretation of the Patterson function of crystals containing a known molecular fragment. The structure of an *Alstonia* alkaloid. *J. Am. Chem. Soc. 85* (Feb. 1963), 353–354.
7. Pepinsky, R. An electronic computer for X-ray crystal structure analyses. *J. Appl. Phys. 18* (July 1947), 601–604.
8. Pepinsky, R., Van den Hende, J., and Vand, V. X-RAC and digital computing methods. In *Computing Methods and the Phase Problem in X-ray Crystal Analysis*, R. Pepinsky, J.M. Robertson, and J.C. Speakman (Eds.) Pergamon Press, New York, 1961, pp. 154–160.
9. Robertson, J.M. *Organic Crystals and Molecules*, Cornell U. Press, Ithaca, N.Y., 1953, pp. 105–108.
10. Rohrer, D.C., and Sundaralingam, M. Stereochemistry of nucleic acids and their constituents. VI. The crystal structure and conformation of dihydrouracil: a minor base of transferribonucleic acid. *Acta Cryst. B26* (May 1970), 546–553.
11. Rosenfeld, A. *Picture Processing by Computer*. Academic Press, New York, 1969.
12. Rossman, M.G., and Blow, D.M. The detection of sub-units within the crystallographic asymmetric unit. *Acta Cryst. 15* (Jan. 1962), 24–31.
13. Scheringer, C. Least-squares refinement with the minimum number of parameters for structures containing rigid-body groups of atoms. *Acta Cryst. 16* (June 1963), 546–550.
14. Stout, G.H., and Jensen, L.H. *X-ray Structure Determination*. Macmillan, New York, 1968, p. 353.
15. Stout, G.H., Malofsky, B.M., and Stout, V.F. Phytolaccagenin: a light atom X-ray structure proof using chemical information. *J. Am. Chem. Soc. 86* (Mar. 1964), 957–958.
16. Stout, G.H., Stout, V.F., and Welsh, M.J. Celebixanthone/ a combined chemical and crystallographic structure proof. *Tetrahedron 19* (Apr. 1963), 667–676.
17. Sutherland, I.E. Computer displays. *Sci. Amer. 222* (June 1970), 56–81.