A Pattern-Recognition Procedure for Scanning Oscillation Films

BY W. KABSCH

Max-Planck-Institut für medizinische Forschung, Jahnstrasse 29, 6900 Heidelberg, Federal Republic of Germany

(Received 4 March 1977; accepted 29 April 1977)

A pattern-recognition procedure has been developed which directly locates and integrates all diffraction spots on a given oscillation film. Indexing of all identified spots is carried out subsequently, together with the refinement of the relevant parameters.

Introduction

X-ray diffraction studies of crystalline biological macromolecules with unit-cell dimensions greater than 100 Å are carried out most efficiently with the oscillation camera (Arndt, 1968; Arndt, Champness, Philzackerly & Wonacott, 1973). The processing of the oscillation photographs can be performed by essentially two different methods (Milch & Minor, 1974).

The first method predicts, from the initial crystal orientation and the starting and finishing angles of the oscillation, which spots will appear on the film and calculates their film coordinates. Then the optical density within a reasonably sized box around each calculated spot position is measured and the observed density within a reasonably sized box around each calculated spot position is measured and the observed positions of the centroids are used to refine the relevant parameters. However, the oscillation limiting angles cannot be refined by this method. They are usually obtained by specifying the indices for a number of partially recorded reflexions on the film.

The second method directly locates and integrates all spots on the film without knowing their diffraction indices. The refinement of the crystal setting and indexing are carried out subsequently. As an additional advantage, a complete list of recorded spots on the film offers a simple way for the refinement of the oscillation limiting angles. This paper describes an efficient implementation of this approach.

Spot-finding procedure

A rotating-drum scanner* was used for digitizing the optical densities of the oscillation films in discrete steps of 25, 50, or 100 μm. During the scan the X axis on film is defined to be parallel to the drum axis and optical-density values are transferred to the computer along scan lines parallel to the Y axis. Fig. 1(a) shows the digitized optical densities of a 0.65 x 1.2 mm film area sampled at steps of 50 μm.

The basic pattern which has to be recognized on the film is specified by a spot-definition box (Ford, 1974). Fig. 1(b) shows a typical layout. Negative values indicate a point belonging to the background area, values of zero mean that the measured optical density of that point is to be ignored, while a positive value specifies a point belonging to the spot area. The upper, lower, right and left boundaries of the spot area are marked by values of 1, 2, 3, or 4, respectively.

According to the spot-definition box, the integrated optical density is determined by summing the optical density values falling within the spot area and subtracting the background, which is estimated by averaging the values within the background area.

To find all spots, the centre of the spot-definition box is successively moved to all points on the film and the integrated optical density is evaluated. The integrated optical densities can also be determined to a good approximation from the density readings at the boundary of the spot area alone, thereby reducing

* Optronics International Inc., 7 Stuart Road, Chelmsford, MA, USA.
the number of arithmetic operations considerably. From Fig. 1, moving the spot-definition box by 1 step along Y on the film, the sum of the optical-density readings in the new spot area can be obtained from the old sum just by adding the readings at the lower boundary and subtracting the readings at the upper boundary in the original position.

A good approximation for the local background is obtained at almost no extra effort by averaging the optical density readings at the boundary of the current spot area. With this approximation, a point on the film is regarded as the potential centre of a spot if its associated integrated optical density is above some threshold value which is, in addition, a local maximum within a range of ±2 steps along both X and Y. In this case the centre of the spot-definition box is moved to each of these 25 neighbouring points and the centroid of the background-subtracted optical-density distribution within the spot area is evaluated in each case. If the spot-definition box is optimally placed the distance between the centre of the defined spot area in the box and the centroid of the observed density distribution is a minimum. For a proper diffraction spot this minimum value must not be greater than 1/2/2 steps.

The positions of the spots, their integrated optical density, variance and background are recorded on a scratch file. At the end of the scan all spot positions are checked to determine whether any spot on the film has been recorded more than once. This does happen occasionally because of the spot centring procedure. Any two spots located by the spot-finding algorithm are considered equivalent, if they differ by no more than three scan steps in both X and Y directions. The spot with the highest integrated optical density is selected from this group of equivalent spots. During the scan the number of scan lines that have to be present in the core memory of the computer is given by the number of steps along X of the spot definition box plus four to allow for a centring capability.

By way of illustrating the spot-finding procedure described above, Fig. 1(c) shows the integrated optical densities obtained by moving the spot-definition box across the film area of Fig. 1(a). The background is estimated as the mean value of the optical density readings at the boundary of the spot. The two local maxima and the corresponding positions of the spot-definition box are indicated.

**Indexing and refinement**

In accordance with Milch & Minor (1974) a right-handed Cartesian coordinate system \{e₁, e₂, e₃\} is defined in the laboratory frame: e₃ is parallel to the X-ray beam and points from the crystal to the film, e₂ is parallel to the spindle axis, while e₁ points upwards. The origin is at the intersection of the spindle axis with the X-ray beam which coincides with the centre of the crystal. The angular position of the crystal about the e₂ axis is denoted by ϕ, which increases for a counterclockwise rotation.

The film, at a distance F from the crystal, is assumed to be parallel to the e₁e₂ plane. A spot on the film located at the scanner coordinates X, Y has the coordinates

\[
x = (X - X₀) \cos ω + (Y - Y₀) \sin ω, \\
y = (Y - Y₀) \cos ω - (X - X₀) \sin ω, \\
z = F
\]

in the laboratory system. X₀, Y₀ are the scanner coordinates of the film origin. The tilting angle ω describes the orientation of the film on the scanner. Approximate values for X₀, Y₀ and ω are obtained from fiducial marks on the film, whose coordinates in the laboratory system as well as on the scanner are known.

It is useful to separate the crystal-orientation parameters from the metrical parameters of the unit cell. Therefore, another orthonormal right-handed coordinate system \{e'₁, e'₂, e'₃\} is defined which is fixed in the crystal: e'₃ points along the e*' axis, e'₂ points along the b axis, while e'₁ completes the system. The origin coincides with the origin of the laboratory system. In this crystal system the reciprocal unit-cell vectors a*, b*, e* can be represented as columns of the matrix

\[
A' = \begin{pmatrix}
a* \sin β* & b* (\cos γ* - \cos α* \cos β*)/\sin β* \\
0 & b* [\sin^2 α* - (\cos γ* - \cos α* \cos β*)^2/\sin^2 β*]^{1/2} \\
a* \cos β* & b* \cos α*
\end{pmatrix}
\]

(2)

The unit-cell vectors a, b, c are then represented as rows of the matrix B', which is also the inverse of A'.

\[
B' = \begin{pmatrix}
a \sin γ & a \cos γ & 0 \\
0 & b & 0 \\
c(\cos γ \cos α - \cos β)/\sin γ & c \cos α & c [\sin^2 α - (\cos γ \cos α - \cos β)^2/\sin^2 γ]^{1/2}
\end{pmatrix}
\]

(3)

The connexion between the crystal coordinate system and the laboratory system is then given by an orthogonal orientation matrix U which depends upon the spindle position. The matrices describing the reciprocal cell and the unit cell of the crystal in the
laboratory coordinate system are
\[ A = UA', \quad B = B'U^{-1} = A^{-1}. \quad (4) \]

Hence, a reflexion \( hkl \) has the coordinates
\[ \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} = A \begin{pmatrix} h \\ k \\ l \end{pmatrix} \quad (5) \]
in the laboratory system.

Diffraction occurs if the crystal is rotated by some angle \( \tau \) from the centre of the oscillation range such that the reflexion \( hkl \) intersects the stationary Ewald sphere. At diffraction the coordinates are
\[ \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \cos \tau & 0 & \sin \tau \\ 0 & 1 & 0 \\ -\sin \tau & 0 & \cos \tau \end{pmatrix} \begin{pmatrix} x_0 \\ y_0 \\ z_0 \end{pmatrix} \quad (6) \]
and if \( \lambda \) is the wavelength being used
\[ x = \text{sign} \left( x_0 \right)(x_0^2 + z_0^2 - z^2)^{1/2}, \]
\[ y = y_0, \]
\[ z = -(x_0^2 + y_0^2 + z_0^2)\lambda/2 = -(ha^* + kb^* + lc^*)^2 \lambda/2 \]
\[ \tau \approx \sin \tau = (x_2 - x_0)/(x_2^2 + z_0^2). \quad (7) \]

The diffracted X-rays produce a spot on the film at the scanner coordinates
\[ X = X_0 + (x \cos \omega - y \sin \omega)F/(z + 1/\lambda), \]
\[ Y = Y_0 + (x \sin \omega + y \cos \omega)F/(z + 1/\lambda). \quad (8) \]

It is possible also to derive the unknown reflexion indices \( hkl \) of a spot located at the scanner coordinates \( X, Y \) (Milch & Minor, 1974). At diffraction the coordinates in the laboratory system of the corresponding reflexions are
\[ x = \left[ (X - X_0) \cos \omega + (Y - Y_0) \sin \omega \right]^{1/2} \]
\[ y = \left[ (Y - Y_0) \cos \omega - (X - X_0) \sin \omega \right]^{1/2} \]
\[ z = -1/\lambda + F/\lambda[F^2 + (X - X_0)^2 + (Y - Y_0)^2]^{1/2}. \quad (9) \]

The reflexion indices are determined from (4), (5), (6) and (9) as
\[ \begin{pmatrix} h \\ k \\ l \end{pmatrix} = B \begin{pmatrix} \cos \tau & 0 & -\sin \tau \\ 0 & 1 & 0 \\ \sin \tau & 0 & \cos \tau \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}. \quad (10) \]

Fortunately the unknown angle \( \tau \) is small (< 2.5°) and therefore the three nonintegral numbers \( h_0k_0l_0 \) calculated from (10) by setting \( \tau = 0 \) will be close to the correct reflexion indices \( hkl \). The eight integer triplets nearest to \( h_0k_0l_0 \) are determined ignoring those which are systematically absent according to the lattice type of the crystal. The associated \( \tau \) angles for the remaining reflexions are calculated with (5), (6), and (7) and only those are kept which have a \( \tau \) angle within the oscillation range of the film plus the mosaic spread of the crystal. Finally, their scanner coordinates are determined from (8) and compared with the observed spot position. The spot is considered correctly indexed if there is just one calculated scanner position closer than a specified maximum discrepancy \( D_{\text{max}} \) and all the others are farther away than a specified minimum distance \( D_{\text{min}} \). If none of the calculated spot positions are closer than \( D_{\text{max}} \), the spot located at \( X, Y \) by the spot-finding algorithm is considered as an artefact. If more than one of the calculated spot positions is closer than \( D_{\text{min}} \), a spot-overlap situation has been encountered and no indices are assigned to the observed spot.

The indexing of the spots located by the spot-finding algorithm is repeated several times together with the refinement of the relevant parameters. For each cycle a ring area about the film origin is defined from which the spots are taken for indexing and values for \( D_{\text{min}} \) and \( D_{\text{max}} \) are specified. To allow large initial errors for the parameters only spots close to the film origin are indexed in the first cycle and both \( D_{\text{min}} \) and \( D_{\text{max}} \) are set equal to about half the shortest spot distance in the ring. As the parameters rapidly approach their correct values the ring area on the film is extended and the criteria \( D_{\text{min}} \) and \( D_{\text{max}} \) for properly indexed spots are made more stringent. Scanning in steps of 50 μm, final values of about 2 steps for \( D_{\text{max}} \) and slightly less than the shortest spot distance for \( D_{\text{min}} \) are usually taken.

Two groups of parameters have to be refined; the first consists of the film origin \( X_0, Y_0 \), the crystal to film distance \( F \), and the independent metrical parameters of the unit cell. It is useful to proceed in two steps: keeping the unit-cell dimensions constant, \( X_0, Y_0 \) and \( F \) are first refined with all fully recorded reflexions within the allowed film area. This is done by minimizing the sum of the squares of the deviations of the distance \( R_{\text{obs}} \) between the observed spot position and the film origin and a calculated distance \( R_{\text{calc}} \) from (7)
\[ R_{\text{obs}} = [(X - X_0)^2 + (Y - Y_0)^2]^{1/2} \]
\[ R_{\text{calc}} = F(z^2 - 2z/\lambda)^{1/2} = F(x^2 + y^2)^{1/2}. \quad (11) \]

Normally, the unit-cell parameters are not refined initially because they are known much more accurately than \( X_0, Y_0, \) and \( F \). At later stages, when most of the relevant parameters are already correct, the independent unit-cell parameters are refined such that the \( z \) coordinates from (7) best match the observed \( z \) values from (9).

The second group of parameters that have to be refined consists of the film tilting angle \( \omega \) and the orthogonal matrix \( U \). An initial value for \( U \) is easily obtained by specifying the indices of any two noncollinear vectors in real or reciprocal space and their direction cosines in the laboratory coordinate system at a given reference spindle position, \( \varphi_{\text{ref}} \). The components in the crystal coordinate system are calculated
from the indices of the two vectors with the matrix $A'$ for reciprocal-space vectors and the matrix $B'$ for real-space vectors [see equations (2) and (3)]. The components of these two vectors are normalized to unit length in both coordinate systems, then the best orthogonal matrix $U_{\text{ref}}$ relating the crystal system with the laboratory system can be determined (Kabsch, 1976). The orthogonal matrix $U$ at the centre of the oscillation range of the film is then computed from $U_{\text{ref}}$. For the refinement, $\omega$ and $U$ are replaced by

$$
\omega_1 = \omega + \Delta \omega
$$

and the sum of the squared deviations between the calculated and observed spot position for all reflexions on the film is minimized as a function of the small corrections $\Delta \omega$, $\epsilon_x$, $\epsilon_y$. Note that $\epsilon_y$ - representing a small rotation about the spindle axis - cannot be obtained by this procedure. It is rather determined as the mean value of the $\tau$ angles of all identified reflexions on the film, since the distribution of these angles must be symmetrical about the centre of the oscillation range for the correct $U$.

**Discussion**

The algorithms for locating and indexing of spots as described above are coded in Fortran IV for a Xerox (XDS-530) 16-bit computer and are available from the author upon request. Typically, the computing time for processing a film area of $8 \times 8$ cm in steps of 50 $\mu$m is about 50 min, with approximately 40 K 16-bit words of core memory used (ca 0.9 $\mu$s cycle time). For a computer that can address bytes considerably less memory is required. On a big computer (IBM 360/91), the same job is executed in around 4 min, thus prohibiting any large-scale film processing unless computing is free.

In practice it is found that the proper identification of spots on the film does not critically depend upon the defined spot shape. Therefore, no attempt has been made to specify distinct spot-definition boxes for different areas on the film.

I would like to thank Dr G. C. Ford very much for numerous discussions on the subject.

**References**


