Acta Cryst. (1996). D**52**, 687–692

Data Reduction from Twinned RNA Crystals

SUSAN E. LIETZKE, VASILI E. CARPEROS AND CRAIG E. KUNDROT

Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309-0215, USA.
E-mail: kundrot@colorado.edu

(Received 18 October 1995; accepted 5 January 1996)

Abstract

Methods were developed to process diffraction data from epitaxially twinned crystals. Four programs for data reduction and two display programs were developed to augment the data-reduction program *XDS* [Kabsch (1988). *J. Appl. Cryst.* **21**, 916–924]. The programs can be generalized for use with other data-reduction software that provides the user with a list of the reflections used to determine lattice constants and crystal orientation. *LATTICE_VIEW* generates a PDB file containing 'water molecules' at the reciprocal-space coordinates of the strong spots found in the initial data frames. The PDB file is visualized to identify spots that belong to the same lattice, obtain unit-cell dimensions for a lattice, and assess data quality. *VECTOR_MATCH* is used to find additional spots belonging to a lattice. *ACCOUNT*4 determines which spots have been processed by *XDS*. *COMFORT* discards reflections that are too close to a reflection in another lattice. The display programs provide useful visual information on the quality of the crystal orientations used. Data with an $R_{\rm merge}$ of 7.1% at 2.4 Å resolution were obtained from epitaxially twinned crystals of an RNA dodecamer. The data were of sufficient quality to solve the structure with a combination of molecular replacement and single isomorphous replacement methods.

1. Introduction

Stronger X-ray sources now permit more users to collect diffraction data from crystals much smaller in size than was possible ten to 15 years ago. These hardware improvements have done little to allow users to collect and reduce data from twinned crystals. There has been some success at solving macromolecular structures from twinned data sets by some innovative software methods (Fisher & Sweet, 1980; Goldman, Ollis & Steitz, 1987; Redinbo & Yeates, 1993; Lu, Lindqvist & Schneider, 1995).

The RNA dodecamer GGCGCUUGCGUC forms twinned crystals whose diffraction patterns do not fully overlap (Doudna, Grosshans, Gooding & Kundrot, 1993) (Fig. 1). This type of twinning has been referred to as epitaxial twinning (Redinbo & Yeates, 1993). The twin domains are apparent when viewed between crossed polarized light, but attempts to physically separate the domains were unsuccessful.

Attempts to reduce and index the data with *XDS* (Kabsch, 1988) were not successful. Therefore, programs were developed for: (1) viewing the reciprocal-space coordinates listed in the SPOT.XDS file produced by *XDS*, (2) selecting spots that belong to one lattice, (3) identifying which spots correspond to a particular lattice and (4) merging data from different lattices. The programs used in this analysis allowed the dodecamer structure to be solved by molecular replacement and single isomorphous replacement methods using the data from epitaxially twinned crystals.

The programs developed are general in that they can treat more than two lattices and the lattices can have different cell dimensions. Furthermore, they can be generalized for any data-reduction software that provides the user with a file containing the reflections used to determine lattice constants and crystal orientation.

2. Methods

Four programs have been developed for use with *XDS* (Kabsch, 1988) to enable data reduction from epitaxially twinned crystals (Fig. 2). Following the terminology of *XDS*, 'spots' are putative, strong reflections

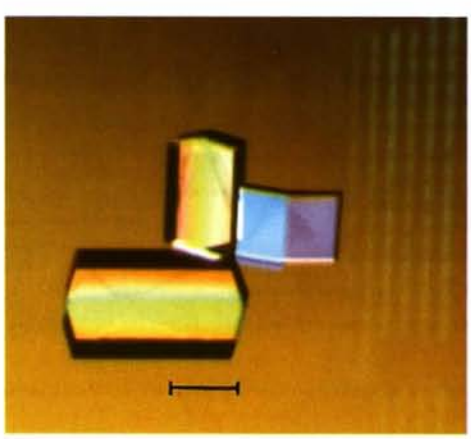


Fig. 1. Crystal of the RNA dodecamer GGCGCUUGCGUC. The bar is 100 μm.

© 1996 International Union of Crystallography
Printed in Great Britain – all rights reserved

identified from the first frames of a data set and 'reflections' are reflections that are reduced to (h, k, l, I, σ) later on in the data-reduction process.

The essence of the approach is to intervene in the data-reduction process between the stages of generating a list of spots and using the spots to determine unit-cell parameters and a crystal orientation. The approach can be used with any data-reduction software that provides the user with a file containing spot positions and the opportunity to intervene at the appropriate stage in the data-reduction process. Unless otherwise indicated, all programs are written in FORTRAN 77. A description of each program follows, along with general remarks about its use. The *Results* section contains more specific results concerning the dodecamer.

2.1. *LATTICE_VIEW*

2.1.1. *Description.* *XDS* reads a number of initial data frames (typically 30), identifies strong reflections and writes the (x, y, φ) coordinates to the file SPOT.XDS. x and y are the coordinates on the detector and φ is the value of the spindle axis.

LATTICE_VIEW reads the SPOT.XDS file and generates a coordinate file [in Protein Data Bank (PDB) format] that contains 'water molecules' at the reciprocal-space coordinates of the spots. Each residue in the file contains one water molecule in the form of an O atom only. The PDB file can be examined with any macromolecular graphics program to visually identify spots that belong to the same lattice (Fig. 3).

There are three chains in the PDB file: *A*, *B* and *C*. Water molecules in the *A* chain have coordinates (0,0,0), (20,0,0), (0,20,0), (0,0,20) in Å, and are useful for visualizing the laboratory coordinate system used by *XDS*.

The *B* chain contains waters at the reciprocal-space coordinates of the spots in the SPOT.XDS file (Fig. 3). The crystal-to-detector distance, direct-beam position and detector swing angle are read in from the XDS.DATA file. The (x, y, φ) coordinates and intensity of each spot are obtained from the SPOT.XDS file. The (x, y, φ) coordinates are transformed into orthogonal reciprocal-space coordinates using a coordinate system in which the origin is located where the direct beam intersects Ewald's sphere, the x axis is perpendicular to the direct beam and the crystal rotation axis, the y axis points from the crystal toward the detector along the direct beam and the z axis is parallel to the crystal rotation axis. For convenience in viewing, the reciprocal-space coordinates are multiplied by a conversion factor of 400 Å^2 prior to writing the PDB file. The coordinates of the spots are written out to the PDB file in the same order that they are read in from the SPOT.XDS file (*i.e.* in order of decreasing intensity).

```
While an unprocessed lattice exists

   Run LATTICE_VIEW (and HKL_VIEW) and identify spots belonging to an
         unprocessed lattice

   While less than 25 spots have been identified

         Pick two spots for reference vector

         Run VECTOR_MATCH to get more spots

   End while

   Run XDS with spots picked by VECTOR_MATCH

   Check quality of lattice orientation with PREDICT2POS

   Run ACCOUNT4 to identify spots not belonging to a lattice

End while

Run COMFORT to eliminate reflections that are too close in reciprocal
   space.

Scale data together with XSCALE.
```

Fig. 2. Pseudocode describing the data-reduction process.

The *C* chain contains coordinates of vectors between the spots (Fig. 4). The vectors between all pairs of spots are calculated and those with a magnitude less than 0.025 Å^{-1} are retained. The vectors are sorted by the product of the intensities of the two spots in decreasing order. For example, residue number one is generated from the two strongest spots. The top 6000 difference vectors are written to the PDB file.

2.1.2. *Use.* The *B* chain shows reciprocal-space coordinates of the spots used by *XDS* to index the reflections. Noise peaks at the perimeter of the detector are readily recognized as not belonging to a regular lattice. Since the residue number of the spots corresponds to their rank order intensity, one can easily determine a residue range that, when displayed, appears free of noise. The SPOT.XDS file can then be edited to retain only these spots. The reciprocal-lattice sampling revealed by the *B* chain is usually too sparse to allow one to measure all six reciprocal unit-cell constants, particularly in the direction perpendicular to the detector face. In preparation for *VECTOR_MATCH*, the residue numbers of spots that appear to belong to the same lattice are recorded.

The *C* chain is useful for obtaining unit-cell dimensions for the lattice(s). An untwinned crystal will produce a cluster of *C*-chain 'water molecules' at each reciprocal-lattice point. If there are many low-intensity noise spots, then many *C*-chain residues with high residue numbers will be scattered through reciprocal space. The *C*-chain residues with low residue numbers can still be used to identify the reciprocal unit-cell dimensions. However, if the crystal slipped during the data collection or if it is twinned, the waters may be too dispersed to obtain reciprocal unit-cell dimensions. If the crystal is twinned, but one twin diffracts much stronger than the other, it is often still possible to obtain reciprocal unit-cell dimensions from the low residue number spots. In our experience, a *C* chain with tight clustering about the reciprocal-lattice positions always indexes well in the *XDS* subroutine *IDXREF*.

2.2. *VECTOR_MATCH*

2.2.1. *Description.* *VECTOR_MATCH* identifies spots in the SPOT.XDS file that belong to the same lattice by identifying pairs of spots that define equivalent vectors. The magnitudes of two vectors are defined as equivalent if they differ by less than a user-specified fraction. The directions of two vectors are regarded as equivalent if the angle between them is less than a user-specified value.

VECTOR_MATCH reads the SPOT.XDS file and the PDB file created by *LATTICE_VIEW*. The user inputs the residue numbers of two spots from the *B* chain created by *LATTICE_VIEW* to define the reference vector, Δ. All possible vectors, $\mathbf{d}_{ij}$, between spots i and j in the SPOT.XDS file are calculated and compared to Δ. If the magnitudes and directions of $\mathbf{d}_{ij}$ and Δ are equivalent, then spots i and j are considered to be part of the same lattice and they are written to the output file. The output file is in SPOT.XDS format.

2.2.2. *Use.* One selects length and angular criteria to get at least 25 spots from *VECTOR_MATCH*. To check the quality of the spots chosen, *LATTICE_VIEW* is run using the new spots from the *VECTOR_MATCH* file as input. If the residues in the resulting file do not cluster tightly, more stringent length and/or angle criteria are needed.

If too few spots are chosen, the procedure can be repeated with a different pair of spots and the output from both runs of *VECTOR_MATCH* combined into one SPOT.XDS file. The spots file generated at this stage is used in *XDS* to process data from one lattice.

2.3. *ACCOUNT4*

2.3.1. *Description.* *ACCOUNT4* is used to determine which spots in the original SPOT.XDS file are accounted for by the lattice(s) processed by *XDS*. The program reads in the original SPOT.XDS file, an XDS.DATA file and an XDS.HKL file for each lattice. The reciprocal-space distance between each spot and the nearest reflection in the XDS.HKL files is determined. If the distance is less than a user-specified value, the spot is regarded as being the same as the reflection and is, therefore, accounted for by a particular orientation of a particular reciprocal lattice. The output is a set of files in SPOT.XDS format. For each XDS.HKL file, there is a file that contains the spots accounted for by the reflections in that file. One additional file contains the spots that were not accounted for by reflections in any of the XDS.HKL files.

2.3.2. *Use.* The objective is to account for all spots in the original SPOT.XDS file. Some spots are never accounted for because *XDS* does not process reflections that have incomplete peak profiles, that are near the spindle axis or that are outside the trusted region of the detector. Some spots may be noise rather than reflections. If a large number of spots remain, the *LATTICE_VIEW/VECTOR_MATCH/ACCOUNT4* sequence can be repeated to find additional lattices.

2.4. *COMFORT*

2.4.1. *Description.* *COMFORT* compares the reciprocal-space coordinates of reflections from two XDS.HKL files and discards those reflections that are 'too close for comfort'. The user specifies the distance criterion for designating two reflections as being too close. *COMFORT* writes the accepted reflections to two XDS.HKL formatted files, one for each of the input files.

2.4.2. *Use.* One uses *COMFORT* in conjunction with the merging program *XSCALE* (Kabsch, 1988) to decide what distance cut off should be used. One seeks to balance a low $R_{\rm merge}$ against a high completeness.

2.5. *PREDICT2POS*

2.5.1. *Description.* *PREDICT2POS* is a program that displays a data frame and shows where reflections are predicted by two different runs of *XDS*, *i.e.* one for each lattice (Fig. 5). *PREDICT2POS* is not necessary for the data reduction, but provides useful visual information on the quality of the crystal orientations used. *PREDICT2POS* is written in IDL (Research Systems, Inc., Boulder, CO). It reads a data frame and the following files from two different *XDS* runs: MODPIX.XDS, XDS.DATA, XPARM.XDS, XYCORR.TABEL. It then calls two FORTRAN subroutines. One subroutine reads a data frame and returns an INTEGER*2 array 512 by 512 pixels containing the intensity of each pixel. The other subroutine is based on the *XDS* subroutine *COLPROF*. Using the data from the *XDS*-generated files, it returns the values of an INTEGER*2 array with dimensions 512 by 512. The array elements are equal to one if the corresponding pixel borders the active region of a reflection and are zero otherwise. *PREDICT2POS* then displays the data frame and the outlines of the active areas.

2.5.2. *Use.* One uses *PREDICT2POS* to determine how well the crystal unit-cell dimensions and orientation are defined. Reflections should only appear in the center of the active region outlines. If the crystal is twinned, one can determine how well each lattice is oriented and how well separated the reflections are.

2.6. *HKL_VIEW*

2.6.1. *Description.* *HKL_VIEW* is a companion program to *LATTICE_VIEW* that helps one determine if unaccounted for spots actually belong to a lattice. The program reads an XDS.HKL file and its corresponding XDS.DATA file and produces a PDB file containing waters at the reciprocal-space coordinates of the

processed reflections. These spots are then viewed along with the spots from a *LATTICE_VIEW* run.

2.6.2. *Use*. Some reflections are not processed by *XDS* because they have incomplete peak profiles, are near the spindle axis or are outside the trusted region of the detector. It is easy to determine if an unaccounted for spot in SPOT.XDS belongs to a lattice if there are many neighboring spots available for comparison. *HKL_VIEW* provides such spots, *i.e.*, the processed reflections. By examining unaccounted for spots this way, one can determine if another lattice exists or not.

3. Results and discussion

The programs described above were used with *XDS* to process and merge data from twinned crystals of the RNA dodecamer GGCGCUUGCGUC. This molecule contains non-canonical U–U base pairs at the center of the duplex. Dodecamer crystals were grown at 311 K from 10% 2-methyl-2,4-pentanediol, 50 m*M* sodium cacodylate pH 7.0, 100–300 m*M* ammonium acetate, and 25 m*M* magnesium chloride. The crystals were twinned (Fig. 1) and broke into irregular fragments when physical manipulation was attempted. Diffraction data to 2.4 Å were collected with a Siemens area detector from crystals at 100 K. The dodecamer crystals contained two lattices; both belonged to space group P_1 and had the same unit-cell dimensions: $a = 29.4$, $b = 28.9$, $c = 46.5$ Å, $\alpha = 98.9$, $\beta = 72.9$ and $\gamma = 96.1°$. The two lattices were oriented in different directions, but shared the same a axis. Prior to implementation of the methods described here, data from the twinned dodecamer crystals were not successfully processed by *XDS*.

The general procedure used to process a dodecamer data set is summarized in Fig. 2. After a SPOT.XDS file was generated by running *XDS*, this original SPOT.XDS file was used as input to *LATTICE_VIEW*. The resulting coordinate file was examined with *InsightII* (Biosym Inc., 1994). For a typical twinned dodecamer data set, portions of the *B* chain appeared to belong to a single lattice while spots in other regions

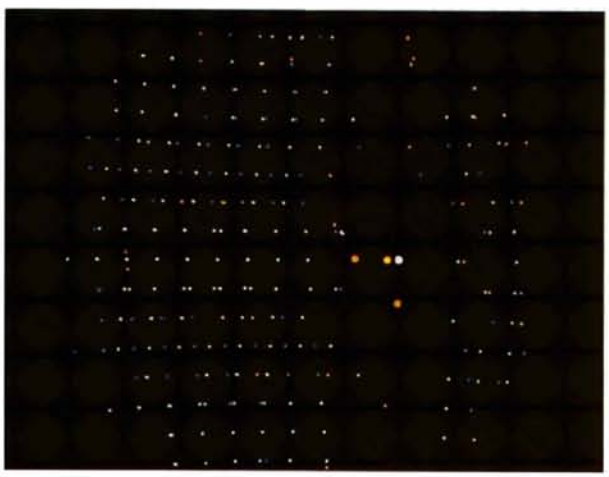
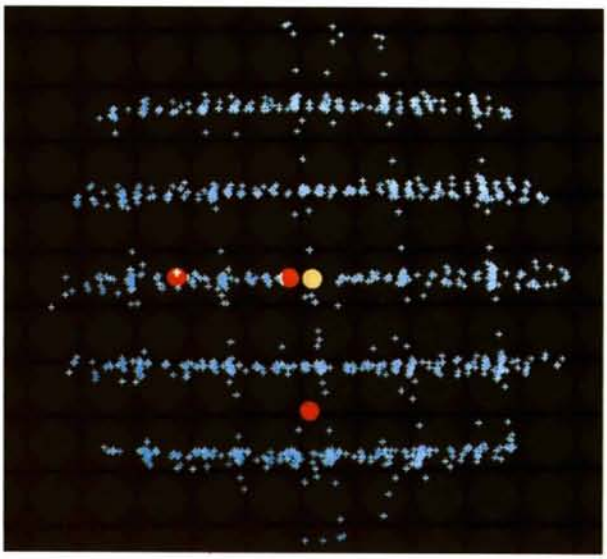
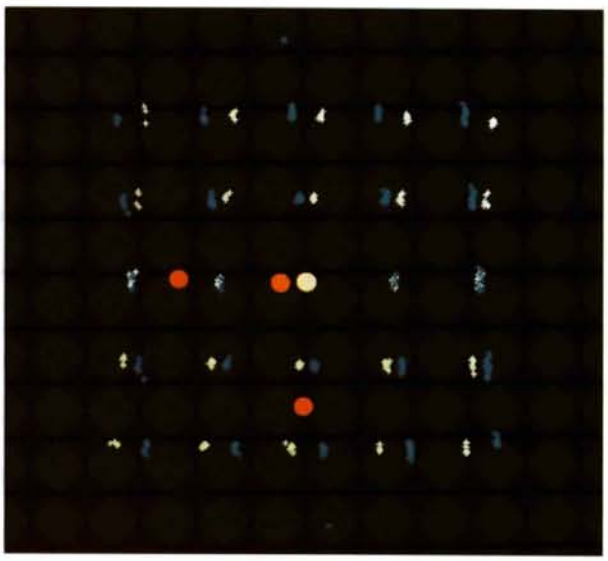


Fig. 3. The *B* chain from *LATTICE_VIEW*. The *B* chain contains spots belonging to lattice 1 (blue), lattice 2 (yellow–green), both lattices (cyan) and neither lattice (orange). The *A* chain is shown in yellow (origin) and orange (x, y and z axes).



(*a*)



(*b*)

Fig. 4. The *C* chain from *LATTICE_VIEW*. (*a*) The difference vectors from the original spots file, (*b*) difference vectors belonging to lattice 1 (blue), lattice 2 (yellow–green) after spots have been assigned to one of the lattices. The *A* chain is shown in yellow (origin) and orange (x, y and z axes).

Table 1. *Statistics on dodecamer native data set*

	Spots accounted for by lattice			Lattice 1				Lattice 2				Merged lattices			
Wedge	1	2	Neither	R_{sym}* (%)	No. of unique reflections	Total No. of observations	% Complete	R_{sym}* (%)	No. of unique reflections	Total. No. of observations	% Complete	R_{merge} (%)	No. of unique reflections	Total. No. of observations	% Complete
a	127	67	38	2.9	4395	5062	77	3.5	4437	5143	78	6.7	5408	10042	95
b	131	76	37	2.9	2366	2398	42	2.8	2369	2415	42	7.2	2893	4409	51
c	132	75	44	—†	442	442	8	—†	451	452	8	5.9	858	890	15
d	92	108	44	2.6	2219	2261	39	2.3	2239	2273	39	5.4	3633	4442	64

* Data from 47 to 2.4 Å. † Did not contain enough observations to calculate R_{sym}.

Table 2. *XSCALE merging statistics for one wedge of native data from 47 to 2.4 Å after running the program COMFORT*

COMFORT cut off (Å)	R_{merge} (%)	No. of unique reflections accepted	Total No. of observations accepted	Completeness (%)
0.000	6.7	5408	10042	95.1
0.003	6.4	5121	9428	90.0
0.005	6.4	5058	9358	88.9
0.008	6.2	4722	8241	83.0
0.011	6.2	3231	5439	56.8

indicated that there was more than one lattice. In the *C* chain, one dimension was fairly well resolved, however the spots in the other two dimensions were scattered.

In order to obtain a SPOT.XDS file that contained spots belonging to a single lattice orientation, the program *VECTOR_MATCH* was used. The fractional length and angular cutoffs used were generally 0.1–0.3 and 1–5°. In the dodecamer case, it was important to avoid choosing spots along the *a* axis, since the twinned domains have this axis in common. The SPOT.XDS file output by *VECTOR_MATCH* was input to *LATTICE_VIEW* and the PDB coordinate file visually examined. This procedure of choosing a pair of spots, running *VECTOR_MATCH*, and looking at the new SPOT.XDS file generally had to be repeated multiple times before a SPOT.XDS file suitable for processing a lattice orientation was obtained.

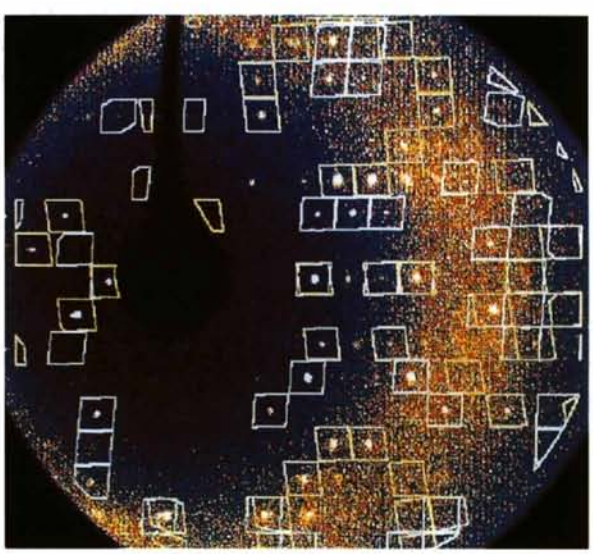


Fig. 5. *PREDICT2POS* output showing a data frame and the predicted positions of reflections from two lattices. The predicted positions are shown in white for one lattice and yellow for the other lattice.

The unit-cell dimensions of the dodecamer crystals were measured from the lattice(s) formed by the *C* chain vectors. Several trials were usually required to obtain a SPOT.XDS file that clearly defined a reciprocal lattice. If the lattices were hard to separate, it was frequently helpful to select a different wedge of data to identify the lattices. Unlike the method of Lu *et al.* (Lu *et al.*, 1995), identification of the two lattices did not require a zone to be aligned in a particular orientation.

Next, the SPOT.XDS file from *VECTOR_MATCH* was substituted for the original SPOT.XDS file and used for processing with *XDS*. After the first lattice orientation, 'lattice 1', was processed, the program *ACCOUNT4* was used to identify spots belonging to 'lattice 2'. *LATTICE_VIEW* and *VECTOR_MATCH* were used to generate a SPOT.XDS file for *XDS* for processing lattice 2. During processing, *PREDICT2-POS* was used to evaluate how well the two lattice orientations were being processed (Fig. 4).

ACCOUNT4 was used to verify that the crystals contained only two lattices. Out of 330 total spots in the SPOT.XDS file from one specimen, 136 belonged to lattice '1' and 105 to lattice '2'. Because the *a* axis of both lattices is the same, 67 spots were accounted for by both lattices. Only 22 spots were unaccounted for by either lattice. These spots were examined and all were not processed by *XDS* because they fell along the spindle axis, were not fully recorded, or were noise spots and not reflections. The partial reflections were readily identified by comparing the spot coordinate to the reflection coordinates produced by *HKL_VIEW*.

Data from a native crystal were collected in four wedges resulting in eight reduced data sets. The two

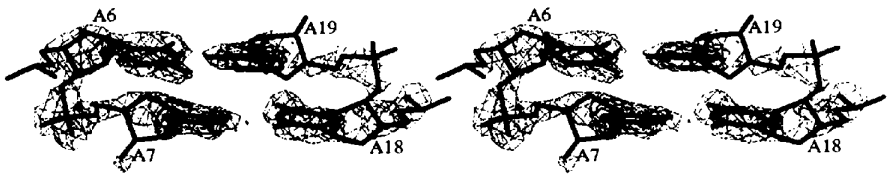


Fig. 6. The current model of the dodecamer. Base pairs U16–U19 and U7–U18 are shown along with $2F_o - F_c$ electron density contoured at 1.0σ.

lattice orientations in each wedge were processed individually, following the procedure above (Table 1). Typically, the two lattices have very comparable data-collection statistics. Data in the highest resolution shell of lattice 1, however, were often more complete than those of lattice 2 because of the stronger diffraction from lattice 1. The R_{merge} values from the two lattice orientations in the native crystal ranged from 5.4 to 7.2% for data between 47 and 2.4 Å.

Prior to merging the data with the program *XSCALE* (Kabsch, 1988), the program *COMFORT* was run in order to remove reflections that were close in reciprocal space. 'Comforting' the two data sets led to a decrease in the R_{merge} value. The first wedge of data was used to test the effect of varying the comfort cut-off level (Table 2). A *COMFORT* cut-off level of 0.003 Å was chosen to balance a low R_{merge} against high completeness. The data from each of the eight data sets, filtered through *COMFORT* with a 0.003 Å cut off, were then merged together. In addition, resolution cut offs were also applied to some of the data because of the weaker diffraction of lattice 2. The final merged native data set had an R_{merge} of 7.1% and was 91% complete. There were 5185 unique reflections and 16679 total observations. Data sets from six candidate heavy-atom derivatives and other native crystals were also successfully processed with the same techniques.

These methods should be applicable to crystals with larger unit cells. The unit cell of the dodecamer is relatively small for a macromolecule and the reciprocal-lattice points are far apart. *LATTICE_VIEW*, however, has provided initial unit-cell dimensions from crystals of other macromolecules with dimensions up to 180 Å. Extrapolating from our experience with the dodecamer suggests that lattice orientations differing by 5–10° can be isolated with *VECTOR_MATCH*. Our experience with *COMFORT* suggests that there is little problem with reflections from different lattices being too close together. It may, however, be useful to modify the background updating procedure in *XDS* for crystals with larger unit cells. The background counts for a pixel are updated when a pixel is far from the nearest predicted reflection. When more than one lattice is diffracting X-rays, a pixel may receive counts from a reflection in a lattice not being processed. One way to solve this problem would be to make sure that a pixel is not unusually strong before it is used to update background.

The data obtained using the methods described in this paper were of sufficient quality to determine the dodecamer structure. Isomorphous-replacement and molecular-replacement techniques were used. The current model includes two duplexes and 78 solvent molecules in the asymmetric unit. It has been refined to a crystallographic R value of 20.3% and R_{free} of 27.0% for data to 2.4 Å (Fig. 6).

4. Conclusions

A method was developed to allow currently available software to reduce data from epitaxially twinned crystals. Programs written to interface with the program *XDS* were created and tested on data from crystals of a RNA dodecamer. The programs developed are general in that they can treat more than two lattices and the lattices can have different cell dimensions. Furthermore, they can be generalized for any data-reduction software that provides the user with a file containing the reflections used to determine lattice constants and crystal orientation. The method is successful; the structure of the dodecamer has been determined and refinement is in progress. The current model has an R value of 20.3% and R_{free} of 27.0%. An analysis of the structure and details of the structure determination will be published elsewhere.

Requests for the programs can be made to CEK at kundrot@colorado.edu. This work was funded by the Colorado RNA Center, W.M. Keck Foundation, and the National Science Foundation (MCB-9221307).

References

Biosym Inc. (1994). *Insight*II. Biosym Inc., San Diego, CA, USA.

Doudna, J. A., Grosshans, C., Gooding, A. & Kundrot, C. E. (1993). *Proc. Natl Acad. Sci. USA*, **90**, 7829–7833.

Fisher, R. G. & Sweet, R. M. (1980). *Acta Cryst.* A**36**, 755–760.

Goldman, A., Ollis, D. L. & Steitz, T. A. (1987). *J. Mol. Biol.* **194**, 143–153.

Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 916–924.

Lu, G., Lindqvist, Y. & Schneider, G. (1995). *Acta Cryst.* D**51**, 13–20.

Redinbo, M. R. & Yeates, T. O. (1993). *Acta Cryst.* D**49**, 375–380.