Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

APRV – a program for automated data processing, refinement and visualization

Markus Kroemer,[a]* Matthias K. Dreyer[b]* and K. Ulrich Wendt[b]*

[a]Adalbert-Stifter-Strasse 8, 79102 Freiburg, Germany, and [b]Aventis Pharma Deutschland GmbH, Structural Biology, Building G865, Industrial Park Hoechst, 65926 Frankfurt, Germany

Correspondence e-mail: matthias.dreyer@aventis.com, markus.kroemer@aprv.org, ulrich.wedt@aventis.com

Received 7 March 2004
Accepted 22 June 2004

APRV (Automatic Processing, Refinement and Visualization) is a new program that enables high-throughput batch processing of crystallographic data. The program combines processing of raw diffraction images, initial structure refinement and visual inspection of resulting electron density into a seamless one-step procedure, during which all relevant parameters are refined automatically. It is controlled by a user-friendly graphical interface, facilitating operation by non-experts.

1. Introduction

Detailed studies of molecular recognition and molecular properties often require the elucidation of significant numbers of protein–ligand or mutant structures that only exhibit structural differences around a specific site in a macromolecule. In the area of drug discovery, the need for rapid determination of protein–ligand structures is further underscored by the use of crystallography as a screening tool for low-affinity ligands (Nienaber *et al.*, 2000; Carr & Jhoti, 2002). These procedures require the collection, processing and interpretation of data sets from several dozens to hundreds of soaked crystals, many of which do not contain the desired ligand. To enable the throughput needed for such applications, it is critical to automate routine procedures (Kuhn *et al.*, 2002). The determination of protein–ligand structures based on known protein structures involves the following steps: (i) data collection, (ii) data processing, (iii) initial model refinement and calculation of electron-density maps, (iv) visual evaluation of electron densities and (v) ligand placement and final refinement of the complex structures.

Recent software developments provide solutions for beamline automation (Leslie *et al.*, 2002), automated structure determination starting with processed data (Adams *et al.*, 2002) or partially refined structures (Oldfield, 2001; Brunzelle *et al.*, 2003) or provide an environment for *de novo* structure determination using experimental phases and starting from raw data (Holton & Alber, 2004). Our program *APRV*, on the other hand, addresses the need for a simple tool to easily process large numbers of raw data sets in order to obtain difference electron-density maps that are key decision points for further analysis of the data (steps ii–iv). The subsequent step of ligand placement (step v) regularly requires intervention by an experienced crystallographer.

APRV makes use of well established crystallographic software, using the program packages *XDS*/*XSCALE* (Kabsch, 1988) for indexing, integration, scaling and reduction of diffraction data, *CNS*/*CNX* (Brünger *et al.*, 1998; Accelrys Inc., 2002) for the initial refinement of structures and *O* (Jones *et al.*, 1991) for visualization of refined structures and associated electron-density maps. Once set up, *APRV* generates all necessary scripts and sequentially executes these programs. Decisions that are usually based on user input are made automatically according to project-specific default settings.

APRV is particularly useful to quickly inspect large numbers of soaking or co-crystallization experiments for meaningful difference electron densities and can be operated without in-depth crystallographic knowledge.

2. Design

APRV has a web-based graphical user interface (GUI) for the input of parameters, the control of computational processes, the inspection of data-processing and structure-refinement log files and the visual inspection of electron-density maps and refined models. The program streamlines crystallographic data processing through (i) elimination of repetitive input requirements, (ii) automation of iterative optimization cycles and (iii) making the readout of soaking experiments accessible to non-expert staff. To enable this, user input is grouped into three categories (Fig. 1*a*).

The category ‘detector’ contains the settings related to a specific radiation source and the laboratory settings as required by *XDS* (*e.g.* detector type, number and size of detector pixels and coordinate system of the experi-

© 2004 International Union of Crystallography
Printed in Denmark – all rights reserved

mental setup). This category has to be specified only once per laboratory source or synchrotron beamline used by a given crystallographic group.

The second input category is referred to as 'major project'. It specifies the parameters related to a specific crystal form, including the space group, unit-cell parameters and (if known) atomic coordinates of a reference model. Optional 'major project' parameters are, for example, a test set for the calculation of free R factors, a reference data set for scaling and reindexing or non-crystallographic symmetry restraints. In addition to the parameters that are required to steer the underlying crystallographic programs, *APRV*-specific parameters can be set by the user to define the criteria for automated decisions such as the determination of resolution cutoffs based on completeness, R factor, $R_{\rm mrgd\text{-}F}$ (Diederichs & Karplus, 1997) and $I/\sigma(I)$ (reasonable defaults are provided).

The third category of input parameters is referred to as 'sub project'. This group comprises parameters that are specific for a given data set, four of which have to be provided for each run, namely the generic filename of images, the detector distance, the oscillation range and the wavelength (default set to Cu $K\alpha$ value).

Once a 'major project' is set up the user input is limited to the 'sub project' category. All other parameters needed for processing are inherited from the related 'detector' and 'major project' categories.

This setup corresponds well to the routine procedures in the crystallographic laboratory, where many isomorphous data sets have to be processed. Without *APRV*, quite a number of parameters must be specified by the user for each of the various programs and some parameters must be interactively refined. In contrast, with *APRV* all necessary parameters are passed between the programs and important parameters are refined in iterative cycles (Fig. 1*a*). For example, when indexing problems occur *APRV* will try all reasonable index origins, refine the detector origin and search for suitable intensity cutoffs to optimize the number of spots used for the indexing step. During profile-fitting and integration the 'beam divergence' and the 'reflecting range' are optimized to convergence by iteratively reapplying the final suggested values of each integration cycle into another profile-fitting step. For space groups with indexing ambiguities a reference data set can be provided and *APRV* will automatically select the setting with the highest correlation for further processing and refinement.

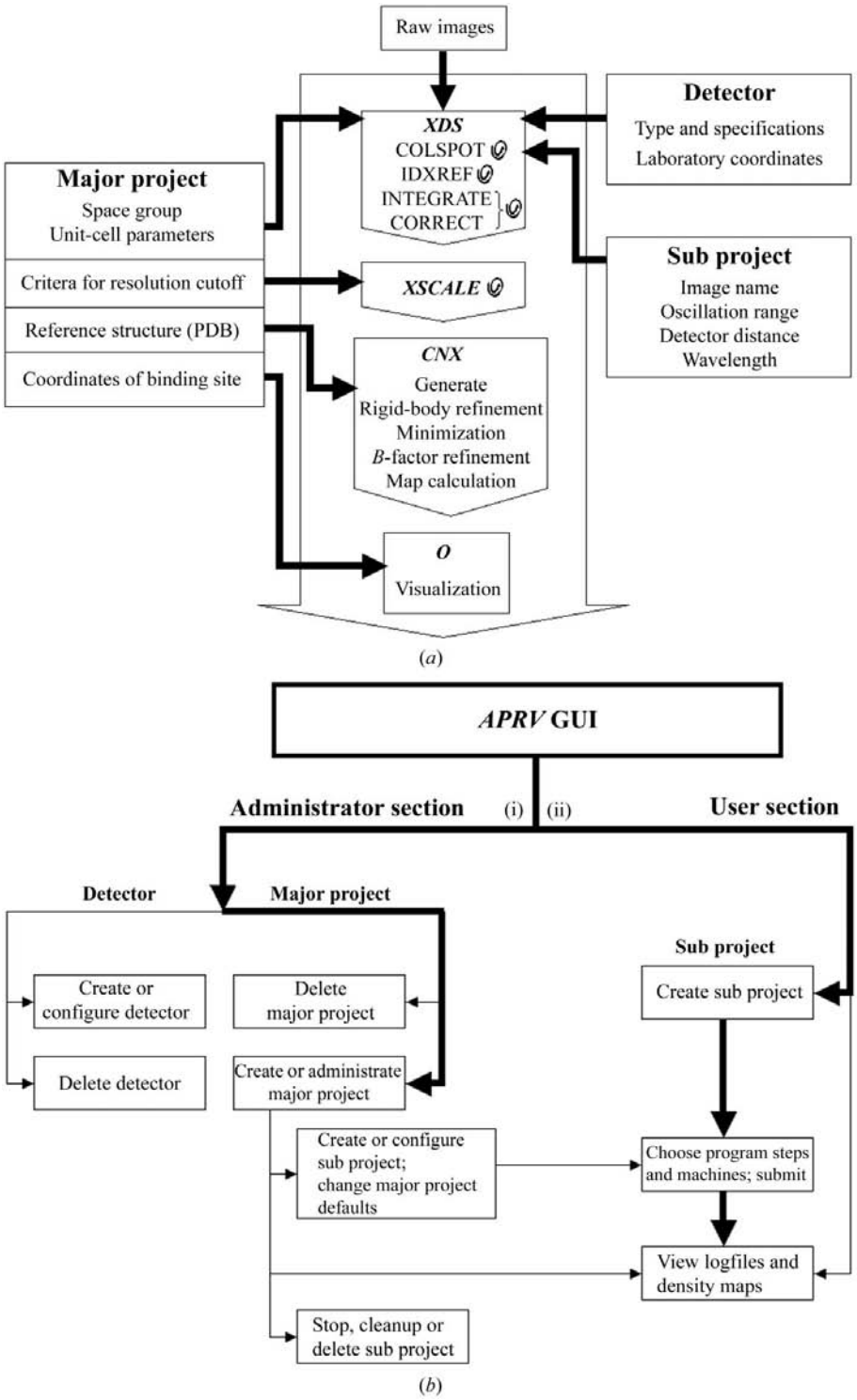

Figure 1
(*a*) Program, data and parameter flow within *APRV*. Parameter input is divided into the three categories: 'detector', 'major project' and 'sub project'. The snail arrows indicate iterative parameter refinement and optimization cycles. (*b*) Organization of the *APRV* GUI. The typical workflow is indicated with bold lines: (i) one-time setting up of a 'major project' in the administrator section; (ii) setting up a 'sub project' for every data set by giving the required input parameters, choosing the desired program steps and selecting a machine for execution. Thin lines indicate additional workflow options.

3. Use

The GUI of *APRV* is built up hierarchically and divided into an 'administrator section' for experienced users and a 'user section' for standard use (Figs. 1*b* and 2). Within the 'administrator section', one can define new

'detectors' and 'major projects' and edit existing ones. Input options in the 'user section' are limited to the 'sub project' category. In routine use on a given crystallographic project non-experts can work in the 'user section' to (i) generate 'sub projects' (usually one for each data set) by setting a few simple parameters, (ii) process data sets and (iii) inspect log files and electron-density maps. Data processed with *APRV* usually exhibit identical or superior quality when compared with manually processed data, while the 'hands-on time' is dramatically reduced (Table 1).

Table 1
Comparison of data quality between manually and automatically processed data.

Manual data processing was performed by an experienced crystallographer in a straightforward manner, *i.e.* without iterative refinement of data-processing parameters. Single program steps were only re-run if they failed. The focus during manual data processing was to obtain data quickly, whereas no special care was taken to obtain the highest possible data quality. Criteria for resolution cutoff were $R_{\rm mrgd\text{-}\mathit{F}} < 35\%$ in the highest resolution shell and $I/\sigma(I) > 3.0$ for at least 33% of the reflections. The quality of data processed with *APRV* is comparable to or better than the manually processed data. Most important, however, is the dramatic decrease in 'hands-on time'. A sample view of the resulting density is shown in Fig. 3.

Data set	Mode	Resolution (Å)	$R_{\rm mrgd\text{-}\mathit{F}}$† (%)	$I/\sigma(I)$†	$R_{\rm cryst}/R_{\rm free}$‡ (%)	$\rho_{\rm max}/\sigma(\rho)$§	Hands-on time¶ (min)
CDK2††	Manual	1.49	8.5	11.4	27.6/28.8	7.0	50
	APRV	1.49	4.0	17.9	24.5/27.1	7.6	3
ProtX††	Manual	2.36	7.9	13.7	27.6/31.2	5.8	50
	APRV	2.36	7.5	14.3	26.1/30.8	6.4	3
CK2††	Manual	2.90	10.3	9.8	20.8/27.3	4.7	35
	APRV	2.98	12.9	14.3	18.7/27.9	4.8	3

† Data taken from *XSCALE* (Kabsch, 1988). ‡ Data taken from coordinate-file header after refinement with *CNX* (Brünger *et al.*, 1998; Accelrys Inc., 2002). § Data for difference-density map as given by *MAPMAN* (Jones & Thirup, 1986). In the cases shown, the maximum peak of the difference-density map corresponds to a bound ligand. ¶ The 'hands-on time' is the real time spent for creating directories, copying and editing input files and running the programs, except for the run time of *XDS*. †† CDK2, cyclin-dependent kinase 2; ProtX is a 65 kDa protein; CK2, casein kinase 2.

3.1. 'Major project' definition

The name of a 'major project' is freely assignable and may describe the protein of interest, *e.g.* 'PROTEIN KINASE A'. The setup of a 'major project' is divided into the configuration of the programs *XDS*, *XSCALE*, *CNS*/*CNX* and *O* (Fig. 2). For data processing, initial refinement and density calculation, a minimum of three 'major project' parameters must be supplied by the user, namely the space group of the crystal form, the unit cell and the name of the reference PDB file (*i.e.* apo or native structure).

3.2. 'Sub project' definition

The name of a 'sub project' is freely assignable and may, for example, describe the inhibitor used for soaking, *e.g.* 'STAUROSPORINE'. Processing of a given data set is started by creating a new 'sub project' based on the higher-level 'major project'. As described above, mandatory parameter values must be supplied and the 'detector' must be selected out of a pop-up list. Finally, the user must select the appropriate program steps to be run by *APRV* from a list. A typical run consists of the sequence *XDS*–*XSCALE*–*CNS*(generate)–*CNS*(rigid)–*CNS*(minimize)–*CNS*(*B* individual)–*CNS*($2F_oF_c$)–*CNS*(F_oF_c).

The progress of running jobs can be followed by viewing the corresponding status page, which provides results summaries from the crystallographic programs used in the process as well as links to their detailed log files. The results are stored in the '*APRV* home directory', which is hierarchically organized into 'major project' and subordinate 'sub project' directories. As soon as refinement has been completed, the refined structure and electron-density maps can be visualized by clicking on a button which opens an *O* session. If the user has supplied a specific coordinate centre for *O* in the 'major project' setup (*e.g.* the expected inhibitor-binding site), the view is centred to

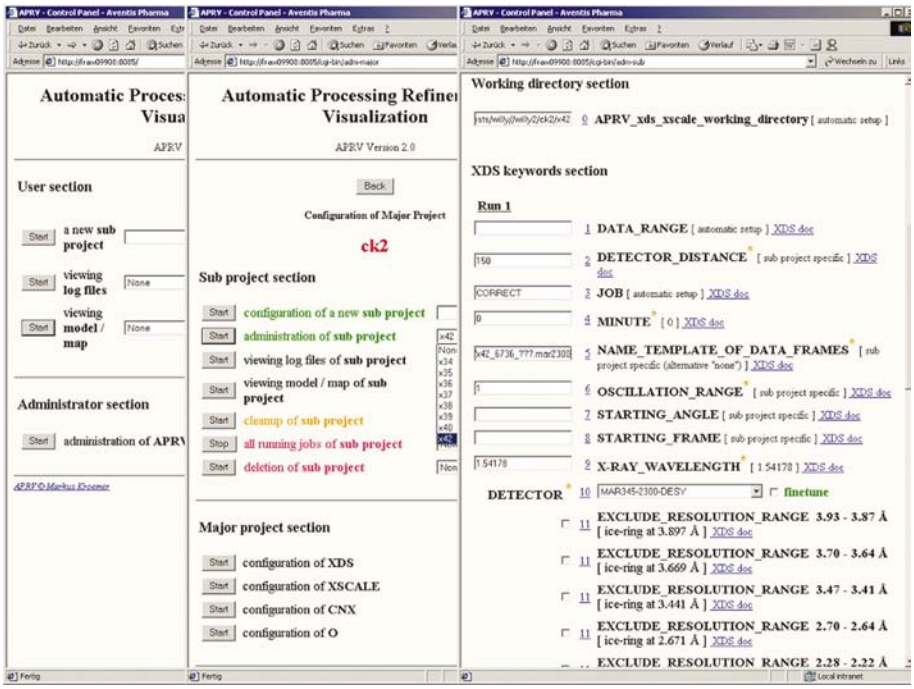

Figure 2
Example view of *APRV* GUI pages. Left, *APRV* start page. Pressing the 'administrator' start button opens the administrator section (centre), from where various actions can be started, *e.g.* the administration of an existing 'sub project' (right). The 'sub project' administration allows the changing of parameters in order to re-run certain program steps, *e.g.* CORRECT within *XDS*, exclusion of ice-ring resolution ranges *etc.*

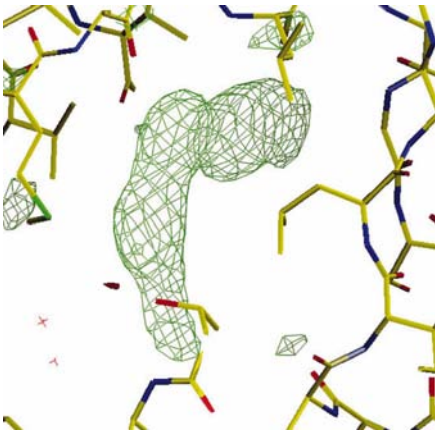

Figure 3
Sample view of resulting F_oF_c electron density. The difference density shown (contoured at 3σ) corresponds to a small-molecule ligand bound to ProtX.

this position and allows a quick inspection of whether a compound soak was successful.

Batch processing of an unlimited number of data sets can be performed and the number of jobs running in parallel is controlled by machine-specific parameters.

4. Implementation

The web-based GUI of *APRV* is based on cgi scripts written in Perl-5. The GUI is provided by an Apache webserver v.1.3.23 (The Apache Software Foundation, 2002) running on a local computer (the *APRV* GUI server) and is accessible from all local computers running a web browser (PCs or workstations). *APRV* does not need to be compiled and is easy to install by running shell scripts that set up the local environment of the *APRV* GUI server. From within the GUI, processing of collected data, refinement of structures and visualization of results can be run either on the *APRV* GUI server itself or on remote computers using remote shell commands. Remote computers have to be set up once by running a shell script for each computer.

The current version of *APRV* runs on Silicon Graphics IRIX64, Digital OSF1 and True64 and various Linux distributions (RedHat, SuSE, Debian). The program is freely available to academic and non-profit users at http://www.aprv.org/. Licences for underlying crystallographic programs used in *APRV* must be requested separately.

We thank Herman Schreuder, Christian Engel, Daniel Kloer and Michael Claus for helpful discussions.

References

Accelrys Inc. (2002). *CNX User Guide*. San Diego, USA.

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Read, R. J., Sacchettini, J. C., Sauter, N. K. & Terwilliger, T. C. (2002). *Acta Cryst.* D**58**, 1948–1954.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, N., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Brunzelle, J. S., Shafaee, P., Yang, X., Weigand, S., Ren, Z. & Anderson, W. F. (2003). *Acta Cryst.* D**59**, 1138–1144.

Carr, R. & Jhoti, H. (2002). *Drug. Discov. Today*, **7**, 522–527.

Diederichs, K. & Karplus, P. A. (1997). *Nature Struct. Biol.* **4**, 269–275.

Holton, J. & Alber, T. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 1537–1542.

Jones, T. A. & Thirup, S. (1986). *EMBO J.* **5**, 819–822.

Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 67–71.

Kuhn, P., Wilson, K., Patch, M. G. & Stevens, R. C. (2002). *Curr. Opin. Chem. Biol.* **6**, 704–710.

Leslie, A. G. W., Powell, H. R., Winter, G., Svensson, O., Spruce, D., McSweeney, S., Love, D., Kinder, S., Duke, E. & Nave, C. (2002). *Acta Cryst.* D**58**, 1924–1928.

Nienaber, V. N., Richardson, P. L., Klighofer, V., Bouska, J. J., Giranda, V. L. & Greer, J. (2000). *Nature Biotechnol.* **18**, 1105–1109.

Oldfield, T. J. (2001). *Acta Cryst.* D**57**, 696–705.

The Apache Software Foundation (2002). *The Apache Software Foundation*, http://www.apache.org/.