Acta Crystallographica Section D
**Biological
Crystallography**
ISSN 0907-4449

Auto-Rickshaw: an automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment

Santosh Panjikar,* Venkataraman Parthasarathy, Victor S. Lamzin, Manfred S. Weiss and Paul A. Tucker

EMBL Hamburg Outstation, c/o DESY,
Notkestrasse 85, D-22603 Hamburg, Germany

Correspondence e-mail:
panjikar@embl-hamburg.de

Received 16 November 2004
Accepted 13 January 2005

The *EMBL-Hamburg Automated Crystal Structure Determination Platform* is a system that combines a number of existing macromolecular crystallographic computer programs and several decision-makers into a software pipeline for automated and efficient crystal structure determination. The pipeline can be invoked as soon as X-ray data from derivatized protein crystals have been collected and processed. It is controlled by a web-based graphical user interface for data and parameter input, and for monitoring the progress of structure determination. A large number of possible structure-solution paths are encoded in the system and the optimal path is selected by the decision-makers as the structure solution evolves. The processes have been optimized for speed so that the pipeline can be used effectively for validating the X-ray experiment at a synchrotron beamline.

1. Introduction

The use of anomalous scattering coupled with the ready availability of tuneable synchrotron beamlines has revolutionized the determination of initial phases in macromolecular crystallographic structure solution. Many protein structures have now been solved using either single or multiple isomorphous replacement with anomalous scattering (SIRAS or MIRAS). In particular, the number of structures determined using information derived solely from anomalous scattering, be it by the single wavelength (SAD) or multiple wavelength anomalous diffraction (MAD) methods (Hendrickson, 1991), is rapidly increasing. The advantage of the SAD/MAD techniques is that only one crystal is required for data collection, although in the case of MAD the number of wavelengths at which the data need to be acquired may vary (Nagem *et al.*, 2001; González, 2003; Burla *et al.*, 2004). The methods themselves, however, are well established in macromolecular crystallography (MX).

The recent launch of various structural genomics projects worldwide has brought about another revolution in the field of MX. These initiatives have provided an enormous drive for the development of high-throughput methods, which has resulted in the automation of the many different steps in structure determination and the construction of so-called pipelines. Examples include the developments towards the automatic set-up of crystallization experiments (Soriano & Fontecilla-Camps, 1993; Sadaoui *et al.*, 1994; Chayen *et al.*, 1990, 1994), crystal detection in crystallization drops (Luft *et al.*, 2001; Wilson, 2002; Spraggon *et al.*, 2002; Rupp, 2003; Cumbaa *et al.*, 2003; Bern *et al.*, 2004), crystal mounting and centring (Andrey *et al.*, 2004) and further developments in X-ray data collection and processing (Leslie *et al.*, 2002) as

doi:10.1107/S0907444905001307

well as crystal structure determination, which will be discussed below.

Crystal structure determination both by isomorphous replacement and by anomalous scattering techniques is a multi-step process in which each step, from substructure determination to model building and validation, requires certain decisions to be made. These decisions comprise the choice of the crystallographic computer programs that are most suitable to perform the specific tasks and the optimal input parameters for each of these programs. The important parameters include the space group of the crystal, the number of molecules in the asymmetric unit, the type of heavy-atom derivative, the extent of derivatization, the diffraction limit of both the native and the derivatized crystal and the quality of the collected diffraction data. After the collection of the X-ray data (of native or derivative crystals or both), existing crystallographic computer programs for X-ray data processing and scaling, for solving and validating the substructure, for the refinement of the substructure atom parameters, phase calculation, density modification, phase extension and non-crystallographic symmetry (NCS) averaging (if more than one molecule is present in the asymmetric unit) are normally relied upon in order to achieve an interpretable electron-density map. The interpretability of the map depends to a large extent on the success of the preceding steps and is generally limited by the resolution of the data and the quality of the phase information. Traditionally, each of the steps described was carried out by an experienced crystallographer, whose skill manifested itself in finding the optimum, or at least a successful, path towards the completion of the structure determination.

In the recent past, several structure-determination software packages have been assembled by various authors with different goals and degrees of built-in automation, *e.g.* *SOLVE*/*RESOLVE* (Terwilliger, 1999, 2000), *SHARP* (de La Fortelle & Bricogne, 1997), *BnP* (Weeks *et al.*, 2001) and *HKL2MAP* (Pape & Schneider, 2004). More recently, automated systems for structure determination have been developed that combine different crystallographic computer programs to build a crystal structure determination pipeline. Examples include *CHART* (Emsley, 1999), *AUTOSHARP* (C. Vonrhein, E. Blanc, P. Roversi & G. Bricogne, unpublished work), *ACrS* (Brunzelle *et al.*, 2003), *ADSP* (http://asdp.bnl.gov/), *ELVES* (Holton & Alber, 2004), *APRV* (Kroemer *et al.*, 2004) and *CRANK* (Ness *et al.*, 2004). Most of these systems require sufficiently high-resolution X-ray data and reasonable phase information, mainly because the model-building step is based on either the program package *ARP*/*wARP* (Perrakis *et al.*, 1999; Morris *et al.*, 2004) or *RESOLVE* (Terwilliger, 2000).

Like most of the other systems, the *EMBL-Hamburg Automated Crystal Structure Determination Platform* (colloquially termed *Auto-Rickshaw*) is entirely based on common crystallographic methodologies. Various available crystallographic computer programs are combined with several decision-makers. These decision-makers are coded in an attempt to mimic the approach an experienced crystallographer would take. Their role is to choose the appropriate crystallographic computer programs and the required input parameters at each step of the structure determination. The system uses a web-based graphical user interface to sequester the minimal initial data needed to determine a macromolecular structure using the SAD, SIRAS, two-wavelength or three-wavelength MAD (2W-MAD or 3W-MAD) approaches.

2. Objectives

The primary aim of the system presented here is to obtain an interpretable electron-density map and a partial structure in order to confirm the success of the X-ray experiment at the synchrotron while the crystal is still at or near the beamline. In practice, as soon as the first data set is collected and processed, the structure-determination pipeline can be invoked. While the computations are running, X-ray diffraction data collection may be continued at other wavelengths in the case of a MAD experiment or for other candidate derivatives in the case of isomorphous replacement. If the data collected in the first experiment can be successfully interpreted, further data collection can be halted (Dauter, 2002). The philosophy of the present system is to provide the user with an as-simple-as-possible and easy-to-use interface with the option for different phasing protocols. The number of required input parameters should be minimal and the time required for structure solution should be as short as possible. Such a setup should ensure a more efficient use of synchrotron beamtime.

The ultimate aim of the platform is to achieve a model which is correct and as complete as possible. At present, this is only possible with relatively high-resolution X-ray diffraction data (higher than 2.6 Å) when reasonable phase information is coupled with the automated model-building program *ARP*/*wARP* (version 6.1, available since July 2004; http://www.arp-warp.org). A medium-term goal is to reduce the resolution requirement to about 3.2 Å.

3. Design of the platform

A C-shell script has been written to combine different macromolecular crystallographic software. The script includes several decision-makers, which ensure that crystal structure determination is performed automatically as soon as X-ray diffraction data from a derivatized protein crystal are collected and processed. A graphical user interface (GUI) has been designed using cgi, Perl, Java and HTML applications. The required input parameters are the space group, the number of amino-acid residues per subunit, the expected number of heavy atoms bound to each subunit, the number of subunits in the asymmetric unit and the name of the project. The desired phasing protocol (SAD, SIRAS, 2W-MAD, or 3W-MAD) must also be selected by the user (Fig. 1). The GUI allows the user to follow the progress of the structure determination and it provides visualization of the initial model together with the electron density. The output is either a partial α-helical model or an almost complete model (see

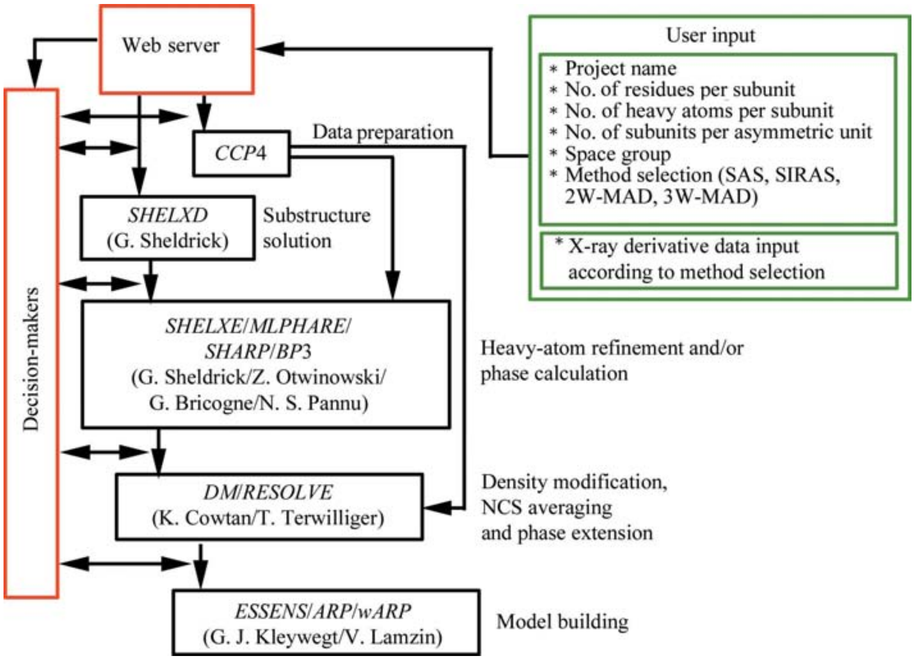
Web server
Decision-makers
CCP4
Data preparation
SHELXD
(G. Sheldrick)
Substructure solution
SHELXE/MLPHARE/
SHARP/BP3
(G. Sheldrick/Z. Otwinowski/
G. Bricogne/N. S. Pannu)
Heavy-atom refinement and/or
phase calculation
DM/RESOLVE
(K. Cowtan/T. Terwilliger)
Density modification,
NCS averaging
and phase extension
ESSENS/ARP/wARP
(G. J. Kleywegt/V. Lamzin)
Model building
User input
* Project name
* No. of residues per subunit
* No. of heavy atoms per subunit
* No. of subunits per asymmetric unit
* Space group
* Method selection (SAS, SIRAS,
2W-MAD, 3W-MAD)
* X-ray derivative data input
according to method selection

Figure 1
Overview of the architecture of the *EMBL-Hamburg Automated Crystal Structure Determination Platform*. The existing crystallographic computer programs in the pipeline are shown in black, the web server and decision-makers in red and the user input in green boxes. Steps from data reduction through initial model building are addressed by the platform and run without user intervention. Data collection, processing, manual model completion, refinement and structure validation are not included.

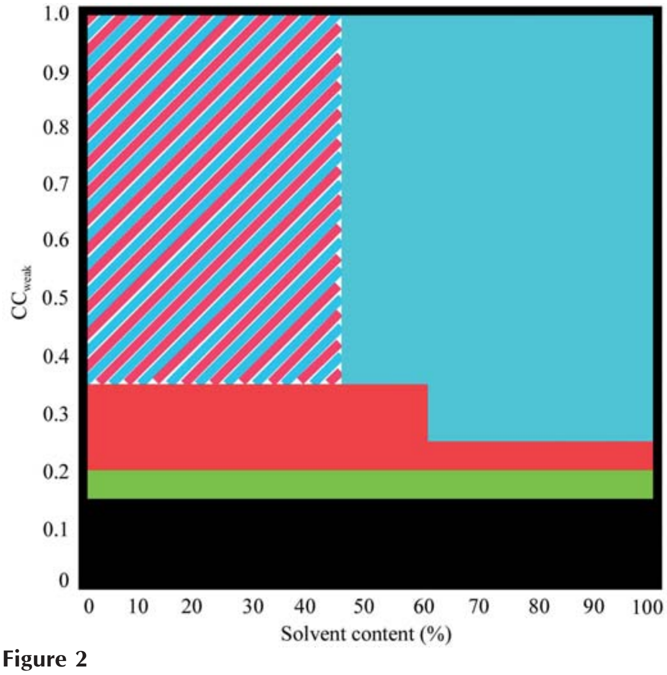

Figure 2
Two-dimensional plot depicting the areas in which the respective heavy-atom parameter refinement and/or phase-calculation program is chosen dependent on CC_{weak} and the solvent content. Colour code for areas: green, *BP*3 (for SAD and SIRAS) or *SHARP* (for 2W-MAD or 3W-MAD); blue, *SHELXE*; red, *MLPHARE*. The black area denotes the regime where further calculations are halted owing to too weak a signal or too poor a heavy-atom model. The maximum resolution of the data as a third dimension is indicated in the red and blue area in the top left corner. Here, either *MLPHARE* ($d_{min} > 2.0$ Å) or *SHELXE* ($d_{min} \leq 2.0$ Å) is chosen. It is important to mention that the boundaries between the areas are subject to change as the pipeline evolves.

below). Each of the phasing protocols performs the following tasks: (i) reading in the necessary X-ray data, (ii) preparation of the X-ray data for the subsequent steps, (iii) scaling of the data (in cases when more than one data set is available), (iv) substructure determination, (v) substructure site selection and enantiomorph determination, (vi) heavy-atom refinement and phase calculation, (vii) density-modification and phase extension, (viii) non-crystallographic symmetry (NCS) search as well as NCS averaging (if applicable), (ix) partial/initial model building and (x) interpretation of electron-density maps.

(i) *Input data*. The processing of the X-ray data is not part of the platform. The following formats of processed data are recognized automatically by the system and converted using *CCP*4 programs (Collaborative Computational Project, Number 4, 1994): *DENZO*/*SCALEPACK* (Otwinowski & Minor, 1997), *MOSFLM*/*SCALA* (Collaborative Computational Project, Number 4, 1994), *DENZO*/*SCALA*, *XDS*/*XSCALE* (Kabsch, 1988) or *d**TREK* (Pflugrath, 1999). It is assumed that the maximum resolution of the data has been sensibly defined by the user, therefore no resolution cutoff is applied at this step. Unless stated otherwise, all data are used in the subsequent steps.

(ii) *Preparation of X-ray data*. Various programs from the *CCP*4 suite are used to prepare the X-ray data for experimental phasing. *SCALEPACK2MTZ* and *TRUNCATE* (French & Wilson, 1978) convert the intensities to structure-factor amplitudes and prepare the MTZ-type files. Based on the unit-cell parameters and the number of amino-acid residues given, the solvent content is estimated. The unit-cell parameters and the presumed number of molecules in the asymmetric unit are then stored for later use.

(iii) *Scaling*. If the experiment type is either SIRAS, 2W-MAD or 3W-MAD then the data sets are scaled to each other using the *CCP*4 program *SCALEIT*. At this stage, a decision is made on the basis of the scaling *R* factor between two data sets as to whether the data are non-isomorphous or whether they are possibly mis-indexed. For SIR, the *R*-factor cutoff is set to 30% and for MAD to 14%. Higher scaling *R* factors cause the structure solution to be halted and the platform suggests continuing without the outlying data set. If the scaling step fulfils the criteria, the scaled data are then passed on to the next step. It should also be mentioned that the *R*-factor cutoffs given have not yet been tested very thoroughly. One of the reasons is the lack of a sufficient number of SIR test cases. Mis-indexing is identified by examining the signed correlation coefficient (SCC) between the anomalous differences for the MAD case as computed in *SHELXC*. According to our experience, a negative value of the SCC in each resolution shell is a good indicator of misindexing.

(iv) *Substructure solution*. The first crucial step of structure determination is the successful location of the heavy atoms. This is undertaken with the program *SHELXD* (Usón & Sheldrick, 1999), which is based on dual-space recycling and the combination of direct methods and Patterson methods. The program *SHELXC* (Sheldrick *et al.*, 2001) writes the instruction and *hkl* files for *SHELXD*. The default number of cycles is set to 1000, since a weak signal or a large substructure may require many trials in order to successfully locate all atoms of the substructure. However, once the correlation coefficient between observed and calculated weak E values $\mathrm{CC}_{\mathrm{weak}}(E_{\mathrm{obs}}, E_{\mathrm{calc}})$ as computed in *SHELXD* matches or exceeds a preset value ($\mathrm{CC}_{\mathrm{weak}} > 20\%$ for SAD and SIRAS and $\mathrm{CC}_{\mathrm{weak}} > 35\%$ for 2W-MAD and 3W-MAD), thereby indicating that a solution has been found, the *SHELXD* job is terminated automatically. If this criterion is not matched *SHELXD* continues to run with the default number of cycles. The maximum resolution used for the substructure-determination step is chosen based on the significance of the anomalous signal. The decision as to whether the anomalous signal is significant is based on the values of $\langle\mathrm{DANO}\rangle/\langle\sigma(I)\rangle$ for the SAD and SIRAS cases and of the SCC between the anomalous differences for the MAD case as computed in *SHELXC*. Values of 1.3 or higher for the former and 30% or higher for the latter are used for the significance check.

(v) *Site selection and enantiomorph determination*. The site selection is based on the peak-height list produced by *SHELXD*. If the substructure is found in the first trial, the decision-makers use the peak heights as the basis to select the sites. Initially, the top $(n + 1)$ sites from the peak list, with n being the number of sites requested by the user, are selected. This list is then examined and all sites that are above a threshold identified by a drop in the peak height of more than 40% between successive sites are selected. If no such drop can be identified in the selected peak list, the remainder of the peak list is searched. When *SHELXD* requires more than one trial to find the substructure, the site selection is based on the common sites between the two best trials. For this purpose, the program *NANTMRF* (Smith, 2002) is used. Once the substructure is solved, the handedness is inspected. As *SHELXD* is based on direct methods, which operate with the normalized structure factors, both enantiomorphs can be expected to be present among the solutions (Schneider & Sheldrick, 2002). Four levels of hand determination are encoded in the script. At first, the program *ABS* (Hao, 2004) is employed. The maximum resolution for the *ABS* run is set to 3.0 Å or to the maximum resolution of the data if it is less than 3.0 Å. If the absolute value of the parameter C (for the definition of C see Hao, 2004) is less than 2%, the quality of the substructure is considered poor and the second level of hand determination is invoked. Here, *SHELXE* (Sheldrick, 2002) is used. The program is run for a single cycle in both substructure hands to check the absolute value of the contrast. The higher value of the contrast should indicate the correct hand. If the results at these two levels are contradictory, the third or fourth level is invoked. In the third level (only applicable to MAD and SIRAS cases) heavy-atom parameter refinement is performed in both hands and the resulting FOM is examined. If the difference between the resulting values for FOM is smaller than 0.005, the decision is deferred to the fourth level. At this level, the correct hand is decided upon after density modification and map skeletonization by checking the connectivity of the 'bones'. The 'bones' are produced by skeletonizing the resulting electron density (Greer, 1974) using the program *MAPMAN* (Jones & Thirup, 1986) with default parameters. For the correct and more easily interpreted map, the connectivity of the bones should be significantly higher than for the map of the incorrect hand. Thus, the larger maximum fragment size of the bones identifies the map belonging to the correct hand. When the anomalous signal is strong (see below), the enantiomorph is usually already determined in the first or first two levels. If the anomalous signal is weak or the substructure model is poor, a more detailed check is made at the first three levels. However, ultimately it is validated at the fourth level by considering the interpretability of the resulting electron density.

(vi) *Heavy-atom refinement and phase calculation*. The platform can invoke three heavy-atom-parameter refinement and phase-calculation programs: *MLPHARE* (Collaborative Computational Project, Number 4, 1994), *SHARP* (de La Fortelle & Bricogne, 1997) and *BP*3 (Pannu *et al.*, 2003; Pannu & Read, 2004). In addition, it also uses *SHELXE* for phase calculation. The platform's decision-makers choose the respective program depending upon the strength of the anomalous signal, the solvent content and the resolution limit of the data. The strength of the anomalous signal is quantified by the correlation coefficient between observed and calculated E values for weak reflections $\mathrm{CC}_{\mathrm{weak}}(E_{\mathrm{obs}}, E_{\mathrm{calc}})$ computed in *SHELXD*. A two-dimensional plot describing the choice of the programs as a function of $\mathrm{CC}_{\mathrm{weak}}$ and solvent content is shown in Fig. 2. The choice between *BP*3 and *SHARP* is made according to the protocol chosen: *BP*3 is chosen for SAD and SIRAS cases and *SHARP* is chosen for the 2W-MAD and 3W-MAD cases.

(vii) *Density modification*. The program *DM* (Cowtan, 1994) is used for density modification and phase extension to the maximum resolution limit. Density modification is required to improve the initial phases calculated from either the SAD, SIRAS, 2W-MAD or 3W-MAD protocols. The density-modification protocol involves solvent flattening, histogram matching and multi-resolution modification with the use of perturbation γ correction for bias reduction. All data are used. The number of cycles is determined by the automatic mode of the program.

(viii) *Non-crystallographic symmetry (NCS) search and NCS-averaging*. The use of NCS depends upon the input given by the user regarding the number of chemically identical molecules in the asymmetric unit. If there are two subunits in the asymmetric unit, the *CCP*4 program *PROFESSS* is launched after the density-modification step in order to search for the twofold operator using the heavy-atom coordinates of the correct hand. At the end of the NCS averaging in *DM* (Cowtan, 1994), the correlation coefficients (CC) are calculated between related areas of electron density. If the calculated initial and final CC are good enough, then subsequent programs use the NCS-averaged map. If the NCS operator has not been found, for example owing to a poor heavy-atom model, the calculations are continued on the *DM* map of the correct hand. In such a case, the NCS symmetry will be ignored. If an NCS operator is found, but after NCS averaging the initial CC is less than 25% and the final CC is not higher than 75%, the calculations are continued on the non-averaged *DM* map. When there are more than two molecules in the asymmetric unit, density modification and NCS averaging is performed using the program *RESOLVE* (Terwilliger, 2000), which is based on statistical density-modification algorithms.

(ix) *Partial/initial model building*. The choice of partial/initial model building depends on the resolution of the X-ray data. If the maximum resolution is lower than 2.6 Å, the script continues with the program *ESSENS* (Kleywegt & Jones, 1996, 1997) in fast mode and looks for ten-residue-long α-helices in the electron-density map. At the end of the search, *ESSENS* creates two maps (the score and display maps) and rotation files that contain rotation information for the best fit of the template. The *SOLEX* program, which is part of *ESSENS*, uses all these files to extract the best solutions and to combine them in order to autotrace α-helices. Such interpretation is possible and useful for proteins whose secondary structure is mostly α-helical or mixed α/β. If less than ten residues are built, then it is concluded that the map either contains predominantly β-sheet structure or that it is not interpretable (see below). If the ratio of the fragment size of the 'bones' generated from the electron density of both hands is larger than 1.02 and the ratio of the *DM* free *R* factor as determined in the density-modification step is smaller than 0.98, the map is considered interpretable. Furthermore, the complete history of the structure determination is examined. If the substructure has been determined successfully (see above) and NCS (if present) has been detected, then it is suggested that the map is likely to contain features corresponding to β-strands and that a visual check on the graphics is needed before halting the data collection at the beamline. If the value for the approximate resolution for 50% solvent content d_{50} (according to the formula $d_{50} = d_{\min}[\mathrm{sc}^{-1} - 1]^{1/3}$, where $d_{\min}$ is the maximum resolution of the X-ray data and sc is the solvent fraction of the crystal) is higher than 2.6 Å, the initial model building is carried out with *ARP/wARP* v.6.1. The number of building cycles is dependent on the map quality, which is assessed from the number of residues built in the first building cycle. If $d_{\min}$ is less than 2.0 Å and in the first building cycle more than 70% of the model is built, the total number of building cycles is set to five, whereas in all other cases the default number of ten building cycles is used.

(x) *Interpretability of electron-density maps*. A good indicator for the interpretability of an electron-density map is the length of the 'bones' fragments. These are typically significantly longer in an interpretable map than in a map that is uninterpretable. If the ratio of the fragment size of the 'bones' generated from the electron density of both hands is close to 1.0 (0.98–1.02) and in addition the ratio of the *DM* free *R* factor as determined in the density-modification step is also close to 1.0 (1.02–0.98), then the map is considered to be not interpretable. At this stage, one possibility is that the structure determination has been attempted in the wrong space group. It is obvious that in such a case the electron density will not be interpretable, neither in the original nor in the reverse hand. It is then suggested to try an alternative space group.

4. Functionality of the design

Currently, there are two versions of *Auto-Rickshaw*: a 'Beamline Version' and an 'Advanced Version' (Fig. 3). Both versions use a similar GUI for input, but the Advanced Version requires the sequence information for the protein target in order to build the side chains during the model-building step. There are 24 unique paths for each coded phasing protocol to solve the crystal structure. The decision-makers select a single path dependent on the input parameters and the evolution of the structure-determination process. The process is started after a button 'SUBMIT' is pressed and its progress can be viewed in the web browser. The Beamline Version is explicitly geared towards its use for validation of the X-ray experiment at the synchrotron as soon as the data have been collected and processed. Its main feature is that it is fast. It can also be useful for finding the correct space group in cases when the space-group ambiguity cannot be resolved at the data-processing and scaling stage. The Beamline Version provides a means to display the electron density together with an initial α-helical model or, if α-helices are not present in the structure (see previous section), together with the 'bones'.

Once the X-ray experiment is validated, the Advanced Version can be used for more complete model building if the resolution of the data permits. The Advanced Version of the platform uses *ARP/wARP* for automated model building.

All decision-making is currently handled within the script. The system avoids proceeding further towards structure solution, if it finds, for example, that the quality of the data is insufficient and/or the anomalous signal is too low to solve the substructure using the embedded protocols. The system then suggests checking the space group and data quality or trying the structure solution manually. If a difficulty is met at a later stage, the decision-makers suggest inspection of the X-ray data and the input. At present, the decision-makers cannot deal adequately with problems arising from non-isomorphism, radiation damage or twinning.

5. Experimental

All calculations presented in this manuscript were performed on a Toshiba Laptop (Satellite 5200-701) with a 1.9 GHz Mobile Intel Pentium 4 processor and 1024 MB RAM.

6. Results and discussion

Tables 1 and 2 depict the list of 14 test and 11 real cases that were used to test and validate the platform, sorted by the maximum resolution of the data. All of the 11 real cases R1–R11 are new structures solved using the platform from data collected at the EMBL Hamburg beamlines by either beamline users or in-house staff. The examples span a resolution range from 1.3 to 3.2 Å, are distributed among many crystal classes and their asymmetric unit content ranges from 85 to 2630 amino-acid residues. The types of heavy atoms are limited to S, Xe, Se and Br, with a dominance of SeMet-containing proteins. In each case, the process began with the input of minimal data through the web interface. The decision-makers selected the appropriate path in order to obtain interpretable electron density and to deliver an initial model in the shortest possible time. Many of the test cases have of course been tried with various protocols, but Tables 1 and 2 depict only some representative examples. It should be mentioned that for all the cases tried so far, at least an interpretable electron-density map has been obtained. Table 3 gives an overview of the results obtained for most of the cases. In the following, a few of the examples will be discussed in more detail.

6.1. Real case R3

R3 represents a typical case for a low-resolution MAD data collection at a synchrotron site. The content of the asymmetric unit is average, as is the solvent content. An interpretable map and partial model could be achieved using the SAD protocol with either the peak-wavelength or the remote-wavelength data within 6 min of CPU time, although the X-ray data had been collected at all three wavelengths. The heavy-atom sites were refined in *MLPHARE*. A twofold NCS operator was found and NCS averaging was performed using *DM*. Electron-density interpretation was carried out using *ESSENS*. More than half of the total CPU time was consumed by building the partial α-helical model containing 149 out of 430 residues, which was sufficient to demonstrate the viability of the phases. When either the 2W-MAD or the 3W-MAD protocol was used, a similar pathway was selected by the decision-makers. In this case, 7 min CPU time was required to obtain an interpretable map and a partial model containing 204 residues.

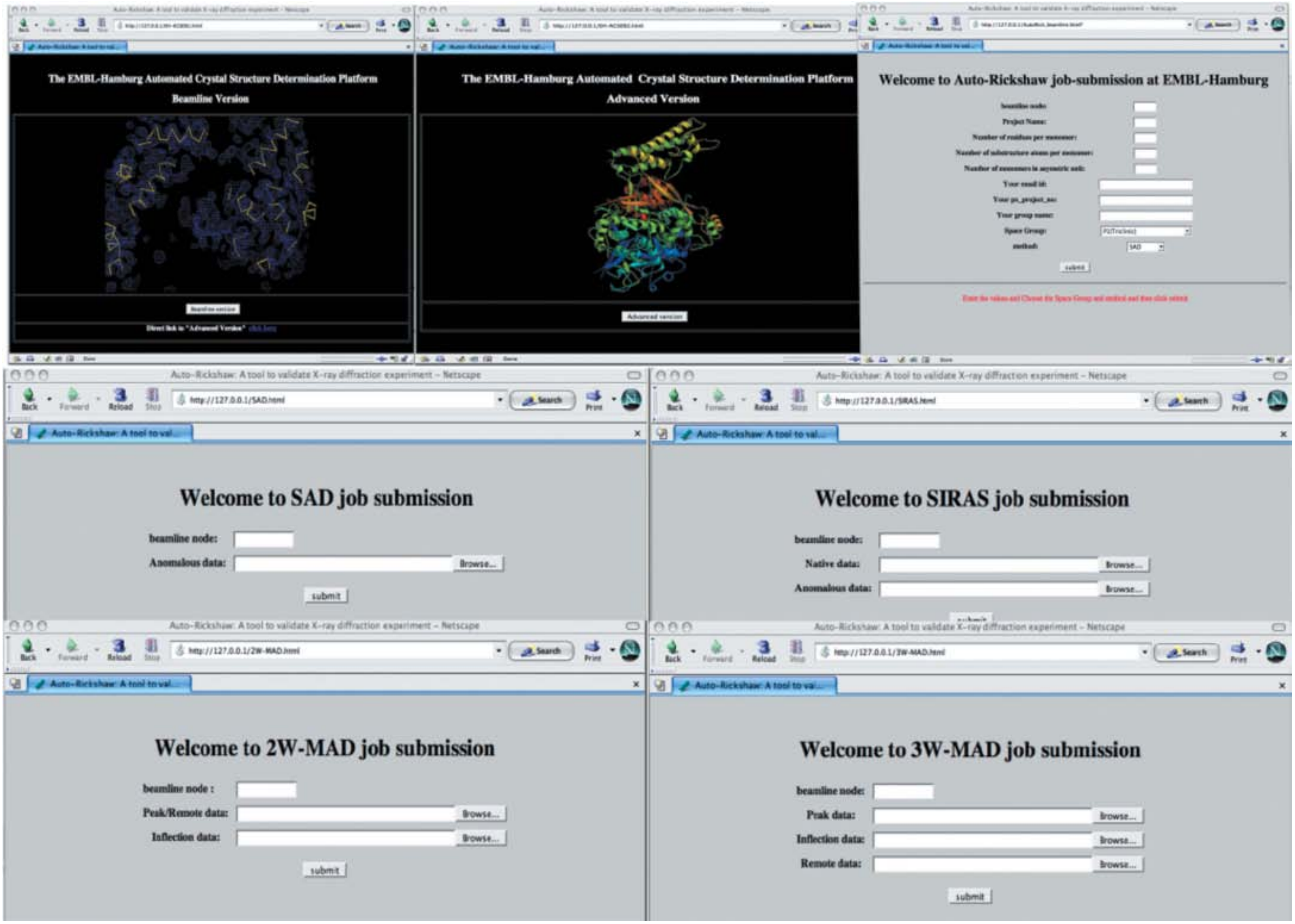
The EMBL-Hamburg Automated Crystal Structure Determination Platform
Beamline Version
Beamline version
The EMBL-Hamburg Automated Crystal Structure Determination Platform
Advanced Version
Advanced version
Welcome to Auto-Rickshaw job-submission at EMBL-Hamburg
Project Name:
Number of residues per monomer:
Number of substructure atoms per monomer:
Number of monomers in asymmetric unit:
Your email id:
Your group name:
Space Group:
method:
SAD
submit
Auto-Rickshaw: A tool to validate X-ray diffraction experiment – Netscape
http://127.0.0.1/SAD.html
Welcome to SAD job submission
beamline node:
Anomalous data:
Browse...
submit
http://127.0.0.1/SIRAS.html
Welcome to SIRAS job submission
beamline node:
Native data:
Anomalous data:
Browse...
http://127.0.0.1/2W-MAD.html
Welcome to 2W-MAD job submission
beamline node :
Peak/Remote data:
Inflection data:
Browse...
submit
http://127.0.0.1/3W-MAD.html
Welcome to 3W-MAD job submission
beamline node:
Peak data:
Inflection data:
Remote data:
Browse...
submit

Figure 3
Graphical user interface for the *EMBL Hamburg Automated Crystal Structure Determination Platform*. The Beamline Version entry window (shown at the top left) appears on the first page of the server. It is linked to the Advanced Version (top middle). The data input interface as well as the interfaces to the SAD, SIRAS, 2W-MAD or 3W-MAD protocols are also shown.

Table 1
Description of the test cases T1–T14.

	Protein	Residues in AU	Sites in AU and element	Subunits in AU	Space group	$d_{\rm min}$ (Å)	Solvent content (%)	Method
T1	NS2	366	15 Se	2	$P6_5$	3.00	45	3W-MAD
T2	Aldolase	2630	40 Se	10	$P2_1$	2.90	40	SAD
T3	Gere†	444	12 Se	6	$C2$	2.75	47	3W-MAD
T4	ACT-II†	370	8 Se	2	$C222_1$	2.50	67	2W-MAD
T5	SS†	400	16 Se	2	$P4_122$	2.45	51	3W-MAD
T6	Cyanase†	1560	40 Se	10	$P1$	2.45	52	3W-MAD
T7	Adaptin	120	2 Xe	1	$P2_12_12_1$	2.00	47	SIRAS
T8	Aldolase	2630	40 Se	10	$P2_1$	1.90	40	SIRAS
T9	Thaumatin	210	17 S	1	$P4_12_12$	1.90	53	SAD
T10	APT-1†	408	20 Br	2	$P2_1$	1.80	42	SAD
T11	PSCP†	372	10 Br	1	$P6_2$	1.80	55	3W-MAD
T12	P9	147	3 Se	1	$I4$	1.70	61	3W-MAD
T13	Elastase	240	30 Br	1	$P2_12_12_1$	1.50	35	SAD
T14	Adaptin	120	2 Xe	1	$P2_12_12_1$	1.30	47	SAD

† Taken from http://www.ccp4.ac.uk/autostruct/testdata.

Table 2
Description of the real cases R1–R11.

Protein	Residues in AU	Sites in AU and element	Subunits in AU	Space group	$d_{\rm min}$ (Å)	Solvent content (%)	Method
R1	800	20 Se	2	$P2_12_12_1$	3.20	56	3W-MAD
R2	636	16 Se	2	$R3$	3.20	58	3W-MAD
R3	430	10 Se	2	$P2_12_12$	3.00	48	3W-MAD
R4	388	12 Se	4	$C222_1$	3.00	46	3W-MAD
R5	200	4 Se	2	$P4_12_12$	2.60	44	SAD
R6	1500	30 Se	6	$P3_212$	2.60	70	2W-MAD
R7	593	9 Se	1	$P6_422$	2.55	78	SAD
R8	468	20 Br	3	$P6_5$	2.50	45	SAD
R9	85	10 Br	1	$P6_322$	2.40	62	2W-MAD
R10	645	11 Se	1	$C2$	2.20	57	SAD
R11	842	24 Se	2	$P2_1$	1.70	42	SAD

It becomes clear that about 2/3 of the beam time could have been saved if the platform had been tried after the first data set.

6.2. Real case R2

R2 is larger than R3 and diffracted less well, but its solvent content was higher. Because of the lower resolution (3.2 Å) and the larger substructure, *SHELXD* required significantly more time to find the sites. Therefore, the platform needed 30 min CPU time to produce an interpretable map. This example demonstrates that the platform is in principle capable of handling low-resolution data, although a partial α-helical model could not be built in this case, since R2 is a predominantly β-sheet structure.

6.3. Real case R7

The resolution limit of R7 was 2.55 Å and 87% of the model was built in 4.5 h CPU time using the Advanced Version of the platform. As this crystal contained 78% solvent, after substructure solution the decision-makers selected the program *SHELXE* for phase calculation and continued with *ARP*/*wARP* for automated model building. The data collection at the second wavelength of a MAD experiment was actually halted by the user when it had become clear that an interpretable map and a partial α-helical model with 363 residues out of 593 built had been obtained by the Beamline Version in less than 1 h CPU time. At this stage, the space-group ambiguity had already been resolved and the solvent content correctly estimated.

Table 3
Results obtained with test and real cases.

n.d., not determined; —, resolution and/or solvent content too low for Advanced Version.

			Beamline Version		Advanced Version	
Protein	Method	$d_{\rm min}$ (Å)	Residues built (%)	Time (min)	Residues built (%)	Time (min)
T1	3W-MAD	3.00	17	12	—	—
T2	SAD	2.90	29	160	—	—
T3	3W-MAD	2.75	55	25	—	—
T4	2W-MAD	2.50	52	22	72	335
T5	3W-MAD	2.45	30	14	70	120
T6	3W-MAD	2.45	n.d.	n.d.	77	835
T7	SIRAS	2.00	0	3	90	88
T8†	SIRAS	1.90	37	180	81	980
T9	SAD	1.90	5	8	96	81
T10	SAD	1.80	18	14	n.d.	n.d.
T11	3W-MAD	1.80	59	40	n.d.	n.d.
T12	3W-MAD	1.70	0	7	90	180
T13	SAD	1.50	n.d.	n.d.	99	112
T14†	SAD	1.30	0	3	85	85
R1	3W-MAD	3.20	16	18	—	—
R2†	3W-MAD	3.20	0	30	—	—
R3†	3W-MAD	3.00	44	7	—	—
R4	3W-MAD	3.00	49	8	—	—
R5	SAD	2.60	26	5	—	—
R6	2W-MAD	2.60	35	111	60	630
R7†	SAD	2.55	69	51	87	270
R8	SAD	2.50	28	22	65	200
R9	2W-MAD	2.40	30	35	75	45
R10	SAD	2.30	28	21	72	230
R11	SAD	1.70	31	13	95	300

† Discussed in greater detail in the text.

6.4. Test case T8

T8 is the largest structure solved so far using the platform. There are ten subunits in the asymmetric unit and each subunit contains 263 residues. The Beamline Version of the platform required 3 h to arrive at an interpretable map and a partial α-helical model containing 976 out of 2630 residues. Using the Advanced Version, the total time required to build 2130 residues (81%) was approximately 16 h. It is clear that for such large structures the platform is not sufficiently fast to be used in real time with data collection. However, given the steady development of crystallographic software as well as computing power, it is conceivable that even such large structures will in the future be determined in minutes, rather than in hours. Another approach to further minimize the computing time is parallelization, which will also be exploited in future versions of the platform.

6.5. Test case T14

T4 is the smallest structure of the examples. It contains 120 residues in the asymmetric unit. The xenon derivative of the crystal diffracted to 1.3 Å. An interpretable map was achieved in 3 min of CPU time using the Beamline Version; however, no partial model was built owing to the predominantly β-sheet structure. The Advanced Version produced an 85% complete model in 85 min CPU time.

7. Relation to other automatic structure determination pipelines

As outlined in §1, a number of other automatic structure determination pipelines have recently been developed elsewhere. Each of these pipelines distinguishes itself from the others by its primary aims and degree of automation. Some rely mainly on one software package; others are more comprehensive. The *EMBL-Hamburg Automated Crystal Structure Determination Platform* follows the latter philosophy, by including modules from many different software packages. Its strength lies in its flexibility and the ability to decide on the path to be taken dependent on the outcome of a previous step. A good representative of a multi-option pipeline is *ELVES* (Holton & Alber, 2004), which is aimed at automating the complete structure determination including the data processing. In contrast, *APRV* (Kroemer *et al.*, 2004) concentrates on the identification of ligand binding. The main philosophy behind the platform presented here is to validate the X-ray diffraction experiment in the minimal time. Results of the computations should become available sufficiently early that decisions with respect to further data collection can be made. In addition, the design is targeted towards the use of pre-coded sequences of computational steps and the means of identifying the most optimum path for structure determination for a given project.

8. Availability of the *EMBL-Hamburg Automated Crystal Structure Determination Platform*

As mentioned above, all the examples presented here were carried out on a laptop computer. The platform has now been implemented on a server at the EMBL Hamburg Outstation connected to a 16-processor Linux cluster. It can be anticipated that this will speed up the process considerably. Users of the EMBL Hamburg beamlines will have access to the server. This server could be made available for other academic institutions in the future.

9. Conclusions

In summary, the *EMBL-Hamburg Automated Crystal Structure Determination Platform* presented here combines several decision-makers with various macromolecular crystallographic computer programs in order to allow fast structure solution from different types of crystallographic experiments automatically without or with minimum user intervention.

Currently, the platform is restricted to SAD, SIRAS, 2W-MAD or 3W-MAD phase determination. So far, many possible paths for structure solution have been explored using the available test data, although only a limited number of the possible paths available for structure solution are explicitly referred to in the tables. No test example has been encountered which met the data-analysis criteria and failed to achieve an interpretable map.

The platform is currently installed on EMBL Hamburg beamline computers and allows users and EMBL staff to validate their X-ray diffraction experiments in the shortest possible time. This ensures an efficient use of the beam time available. The platform is undergoing continuous development. This includes the incorporation of new functionalities as well as continuous software upgrades. Another important aspect is the evolution and improvement of the decision-making. As more data become available, the structure-determination paths can be scrutinized thoroughly in order to increase the efficiency of the coded decision making.

10. Future perspectives

The ultimate aim of the platform is to build a model which is correct and as complete as possible both at low and high resolution. A number of additional tasks can in the future be incorporated in the platform. These include a twinning test, brute-force space-group determination assuming the correct point group, the automatic determination of the number of molecules in the asymmetric unit, different density-modification protocols and ultimately the linkage to automatic data-collection software such as *DNA* (Leslie *et al.*, 2002). Other model-building programs can be incorporated into the platform with the possibility of parallel execution. In the future, further refinement of the constructed model can also be considered. Another plan is to advance the platform by including automated molecular-replacement programs.

The number of input parameters will be further reduced. It is envisioned that only the amino-acid sequence will be required in addition to the diffraction data. Hopefully, the system will evolve into a 'self-learning' system, where a concrete path will be chosen not only based on the input but also on the accumulated history.

We would like to thank the developers of the various computer programs for their kind permission to use their software in this pipeline. We also gratefully acknowledge the generous supply of X-ray data by A. B. Datta, P. Chakrabarti, Y. Liang, X. Su, J. P. Declercq, M. G. Jeppesen, X. Ma, J. Stoeckigt, W. Liu, R. Ladenstein, A. Romero, E. Nowak, C. Butan, S. Müller, B. Kauffmann and E. Lorentzen. Part of the work was supported by the EC-funded projects BIOXHIT (contract No. LHSG-CT-2003-503420) and SPINE (contract No. QLG2-CT-2002-00988).

References

Andrey, P., Lavault, B., Florent, C. & Maurin, Y. (2004). *J. Appl. Cryst.* **37**, 265–269.
Bern, M., Goldberg, D., Stevens, R. C. & Kuhn, P. (2004). *J. Appl. Cryst.* **37**, 279–287.
Brunzelle, J. S., Shafaee, P., Yang, X., Weigand, S., Ren, Z. & Anderson, W. F. (2003). *Acta Cryst.* D**59**, 1138–1144.

Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Moustiakimov, M. C., Polidori, G. & Siliqi, D. (2004). *Acta Cryst.* D**60**, 1683–1686.

Chayen, N. E., Shaw Stewart, P. D. & Baldock, P. (1994). *Acta Cryst.* D**50**, 456–458.

Chayen, N. E., Shaw Stewart, P. D., Maeder, D. L. & Blow, D. M. (1990). *J. Appl. Cryst.* **23**, 297–302.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Cowtan, K. (1994). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **31**, 34–38.

Cumbaa, C. A., Lauricella, A., Fehrmann, N., Veatch, C., Collins, R., Luft, J. R., DeTitta, G. & Jurisica, I. (2003). *Acta Cryst.* D**59**, 1619–1627.

Dauter, Z. (2002). *Acta Cryst.* D**58**, 1958–1967.

Emsley, P. (1999). *CCP4 Newsl. Protein Crystallogr.* **36**. http://www.ccp4.ac.uk/newsletter36/08_chart.html.

French, G. S. & Wilson, K. S. (1978). *Acta Cryst.* A**34**, 517–525.

González, A. (2003). *Acta Cryst.* D**59**, 315–322.

Greer, J. (1974). *J. Mol. Biol.* **82**, 279–301.

Hao, Q. (2004). *J. Appl. Cryst.* **37**, 498–499.

Hendrickson, W. A. (1991). *Science*, **254**, 51–58.

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

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

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

Kleywegt, G. J. & Jones, T. A. (1996). *Acta Cryst.* D**52**, 826–828.

Kleywegt, G. J. & Jones, T. A. (1997). *Acta Cryst.* D**53**, 179–185.

Kroemer, M., Dreyer, M. K. & Wendt, K. U. (2004). *Acta Cryst.* D**60**, 1679–1682.

La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.

Luft, J., Wolfley, J., Jurisica, I., Glasgow, J., Fortier, S. & DeTitta, G. (2001). *J. Cryst. Growth*, **232**, 591–595.

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.

Morris, R. J., Zwart, P. H., Cohen, S., Fernandez, F. J., Kakaris, M., Kirillova, O., Vonrhein, C., Perrakis, A. & Lamzin, V. S. (2004). *J. Synchrotron Rad.* **11**, 56–59.

Nagem, R. A., Dauter, Z. & Polikarpov, I. (2001). *Acta Cryst.* D**57**, 996–1002.

Ness, S. R., de Graaff, R. A., Abrahams, J. P. & Pannu, N. S. (2004). *Structure*, **12**, 1753–1761.

Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.

Pape, T. & Schneider, T. R. (2004). *J. Appl. Cryst.* **37**, 843–844.

Pannu, N. S., McCoy, A. J. & Read, R. J. (2003). *Acta Cryst.* D**59**, 1801–1808.

Pannu, N. S. & Read, R. J. (2004). *Acta Cryst.* D**60**, 22–27.

Perrakis, A., Morris, R. J. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.

Pflugrath, J. W. (1999). *Acta Cryst.* D**55**, 1718–1725.

Rupp, B. (2003). *Acc. Chem. Res.* **36**, 173–181.

Sadaoui, N., Janin, J. & Lewit-Bentley, A. (1994). *J. Appl. Cryst.* **27**, 622–626.

Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.

Sheldrick, G. M. (2002). *Z. Kristallogr.* **217**, 644–650.

Sheldrick, G. M., Hauptman, H. A., Weeks, C. M., Miller, R. & Usón, I. (2001). *International Tables for Macromolecular Crystallography*, Vol. *F*, edited by M. G. Rossmann & E. Arnold, ch. 16, pp. 333–345. Dordrecht: Kluwer Academic Publishers.

Smith, G. D. (2002). *J. Appl. Cryst.* **35**, 368–370.

Soriano, T. M. B. & Fontecilla-Camps, J. C. (1993). *J. Appl. Cryst.* **26**, 558–562.

Spraggon, G., Lesley, S. A., Kreusch, A. & Priestle, J. P. (2002). *Acta Cryst.* D**58**, 1915–1923.

Terwilliger, T. C. (1999). *Acta Cryst.* D**55**, 1863–1871.

Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.

Usón, I. & Sheldrick, G. M. (1999). *Curr. Opin. Struct. Biol.* **9**, 643–648.

Weeks, C. M., Rappeleye, J., Furey, W., Miller, R., Potter, S. A., Smith, G. D. & Xu, H. (2001). Annu. Meet. Am. Crystallogr. Assoc., Abstract W0264.

Wilson, J. (2002). *Acta Cryst.* D**58**, 1907–1914.