Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

SGXPro: a parallel workflow engine enabling optimization of program performance and automation of structure determination

Zheng-Qing Fu,* John Rose and Bi-Cheng Wang

Southeast Collaboratory for Structural Genomics, Department of Biochemistry and Molecular Biology, The University Of Georgia, Athens, GA 30602, USA

Correspondence e-mail: fuzq@uga.edu

Received 1 January 2005
Accepted 4 April 2005

SGXPro consists of four components. (i) A parallel workflow engine that was designed to automatically manage communication between the different processes and build systematic searches of algorithm/program/parameter space to generate the best possible result for a given data set. This is performed by offering the user a palette of programs and techniques commonly used in X-ray structure determination in an environment that lets the user choose programs in a mix-and-match manner, without worrying about inter-program communication and file formats, during the structure-determination process. The current *SGXPro* program palette includes *3DSCALE*, *SHELXD*, *ISAS*, *SOLVE*/*RESOLVE*, *DM*, *SOLOMON*, *DMMULTI*, *BLAST*, *AMoRe*, *EPMR*, *XTALVIEW*, *ARP*/*wARP* and *MAID*. (ii) A client/server architecture that allows the user to utilize the best computing facility available. (iii) Plug-in-and-play design, which allows easily integration of new programs into the system. (iv) User-friendly interface.

1. Introduction

Structure determination by X-ray crystallography is a multi-step process consisting of the following key activities: (i) data collection, (ii) generation of the heavy-atom/anomalous substructure or a suitable search model, (iii) phasing of the experimental structure factors (usually involving some form of density modification), (iv) interpretation of the electron-density map (usually performed automatically if the resolution of the data permits) and (v) refinement of the structural model against the observed X-ray structure factors (usually involving manual model adjustments).

The structure-determination process is heavily dependent on the computer program or programs used in carrying out the various steps in the process. Many programs have been developed by various groups over the years to address each of the above five areas. The large ensemble of programs reflects the fact that no single software solution has been adopted by the crystallographic community. This is because some programs work better than others for a given set of data (Fu *et al.*, 2003; Calderone, 2004). Thus, the crystallographer generally begins the structure-determination process using the program (or programs) with which he/she is most familiar. If this approach fails, then either more data are collected or another program (or program package) is explored.

In addition, several attempts have been made to reduce human intervention in the structure-determination process and several integrated program packages are now in general use. These include *AUTO-RICKSHAW* (Panjikar *et al.*, 2005),

doi:10.1107/S0907444905010541

AUTOSHARP (Blanc *et al.*, 2000), *BnP* (Weeks *et al.*, 2001), *CCP4i* (Potterton *et al.*, 2003) and *SOLVE*/*RESOLVE*. With the advent of structural genomics, several automated structure-determination pipelines such as *ANTPHARM* (Brunzelle *et al.*, 2003), *ASDP* (Jiang & Lin, 2005), *ELVES* (Holton & Alber, 2004), *PHENIX* (Adams *et al.*, 2004) and *SCA2STRUCTURE* (Liu *et al.*, 2005) have been developed. Although these automated packages and pipelines reduce user interaction with the structure-determination process, they generally either only utilize a few specific programs or lack a systematic approach for searching both program and parameter space (individual programs and their program input parameters) to identify which sets of programs are optimal for structure determination from a given data set. To address this point, we have investigated which set of popular crystallographic programs and their respective input parameters, when chained together, gives the highest probability of producing a structure from a given set of data.

Our results show that given the high variability in data quality, no optimal program (or set of programs) can be identified that will generate the best results in every case. Moreover, since the choice of program (or programs) used in a given structure determination is usually a matter of familiarity or convenience (different file formats, input strings *etc.*), the implication is that in some cases the lack of a successful structure determination may be a consequence of the improper choice of programs (or sequence of programs) used in the analysis. Thus, the ideal system would be one that allows the user to easily carry out the structure-determination process by automatically and systematically searching program/parameter space using various programs available to the community in an environment that frees the user from dealing with the wide range of file formats and program-specific input.

SGXPro, described here, is an attempt to meet this challenge. *SGXPro* is a crystallographic computational environment that handles inter-process communication between varieties of popular crystallographic programs. Its design allows the user to pick the programs (in a mix-and-match manner) that he/she wants to use in the structure-determination process *via* a user-friendly GUI. Intelligent parameter defaults coupled with the ability to sample the range of the various input parameters in small steps adds to the power of the approach. The *SGXPro* client/server architecture allows the user to utilize the best computing facilities available, while its plug-in-and-play design allows easy integration of new programs into the system.

2. Methods

SGXPro is the software suite developed at the Southeast Collaboratory for Structural Genomics (denoted as SECSG hereafter) based on the parallel workflow engine described below. It was designed to increase the efficiency of macromolecular structure determination by systematically searching both program and parameter space to optimize each step in the process. Unlike the cluster-based *SCA2STRUCTURE* pipeline also developed at SECSG, *SGXPro* does not require a large Linux cluster and offers considerably more flexibility in terms of programs and pathways available to the user. In terms of performance, *SGXPro* also allows a more hands-on approach, using a wide variety of tools, in cases that for one reason or another fail in the *SCA2STRUCTURE* pipeline (*e.g.* low- or average-quality data).

SGXPro offers the user a palette of programs and the ability to seamlessly and dynamically build and execute complex structure-determination pipelines without worrying about file formats or data input. The approach used in developing *SGXPro* was fourfold. The software should be powerful (in functionality), upgradeable (to keep in step with methodology and technology developments), beautiful (provide a user interface that is both functional and pleasing to the eye) and simple (intuitive and easy to use). To achieve these objectives, the following architectural components have been implemented into *SGXPro*: (i) a parallel workflow

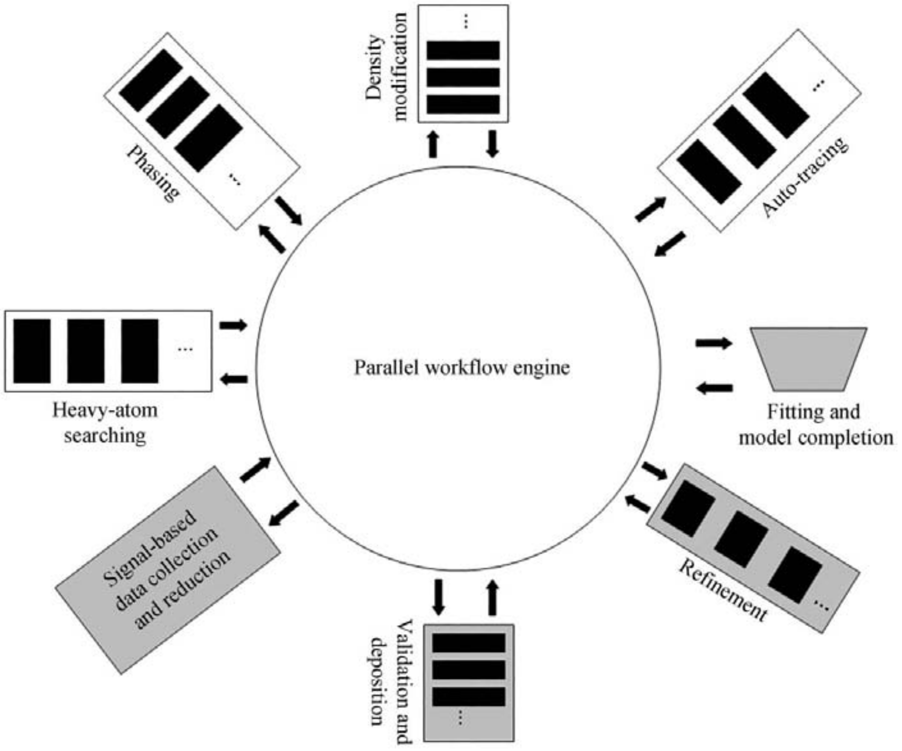

Figure 1
The design of the parallel workflow engine for automation of the structure-solution process. The dark blocks represent parallel tasks dynamically generated from various crystallographic computing programs with different parameter settings. Plug-in interfaces for the programs in the shaded area are under development and are not available in the current version of *SGXPro*.

engine that automatically and systematically searches program and parameter space to arrive at the combination of programs (and their respective input parameters) that will produce a structure for a given data set, (ii) a client/server architecture that allows the user to utilize the best computing hardware available, (iii) a plug-in-and-play design that allows easy integration of new programs into the system and (iv) a user-friendly interface.

2.1. The parallel workflow engine

A novel parallel workflow engine (Fig. 1) has been developed using the C++ computing language. By design, crystallographic programs and utilities are organized into the following categories according to the functions they perform: (i) data acquisition and reduction, (ii) identification of the heavy-atom (or anomalous) substructure, (iii) experimental phasing, (iv) phasing by molecular replacement, (v) density modification and phase improvement, (vi) auto-tracing and model building, (vii) structure refinement, validation and deposition and (viii) general utilities.

The structure-determination job (or process) can be viewed as a cascade of tasks resulting from a common starting point (unit cell, space group and structure factors). The task cascade may involve different programs, with each program executed using different sets of input parameters. Thus, each task can be defined as a computational run of a program with a given set of data and control parameters. For efficiency of task management, a unique class is developed for each computer program that encapsulates the input and output data files and the program control parameters. Each task is internally represented as an object of such a class.

A parallel workflow engine serves as the central control system and manages the whole process using agents (computational modules) for input and job interpretation, task and workflow generation, data flow and control parameter settings, harvesting and analyzing results, task distribution and communication among the various tasks and with the client.

As a job starts, the parallel workflow engine first interprets requests from the client and a parallel workflow of tasks is then dynamically generated according to these requests and the crystallographic computing logic pertaining to the job. For example, if the job is to solve a new protein structure from a set of anomalous diffraction data, the SAS (Hendrickson & Teeter, 1981; Wang, 1985; Dauter *et al.*, 2002; Dodson, 2003) and/or MAD (Phillips & Hodgson, 1980; Karle, 1980; Hendrickson, 1991; Gonzalez *et al.*, 1999) structure-determination workflow will look as follows.

The positions of the anomalous scatterers will be identified by a number of tasks (HA tasks). Each unique HA task is dynamically generated and will employ either a different program, for example *SHELXD* (Schneider & Sheldrick, 2002) or *SOLVE* (Terwilliger & Berendzen, 1999), or a different set of input values (resolution range, number of sites or other adjustable control parameters of the programs). The HA tasks will then be distributed by the workflow engine to the computing server and run in parallel. Upon the completion of the HA tasks, the workflow engine will analyze the outputs from *SHELXD* (including CC all/weak, PATFOM and PSMF values) as suggested by the program manual (Schneider & Sheldrick, 2002) to generate a list of possible solutions. Each solution includes the number of sites and the coordinates of the set of heavy atoms. If the number of sites found is less than that requested, the number of sites found will be used. The solutions from *SOLVE* are sorted by matching the number of sites with those from *SHELXD* if multiple searching of numbers of sites is requested. By doing this, two sorted sets of solutions are generated, one from *SHELXD* and one from *SOLVE*. The user can choose multiple solutions (by default, only the top solution) from each set that will be passed on to the handedness test by *ISAS* (Wang, 1985). If the handedness of a solution is wrong, the workflow engine will make the correction automatically. After the handedness tests, the heavy-atom substructures will be passed on to the phasing step. With the heavy-atom substructure solutions, the workflow engine will proceed to prepare the input data and parameters settings for the next set of tasks for phasing and density modification (PH tasks). Again, each unique PH task is dynamically generated and will employ either a different program (such as *ISAS*, *SOLVE*/*RESOLVE*) or a different set of input values (including resolution range, solvent content and other adjustable parameters of each program). These PH tasks will then be distributed by the workflow to the computing server and run in parallel. As the PH tasks complete, their results will be analyzed and input prepared for subsequent auto-tracing and model-building tasks (MB tasks), if resolution permits. Here again, each unique MB task is dynamically generated and will employ either a different program [such as *ARP*/*wARP* (Perrakis *et al.*, 1999) or *RESOLVE* (Terwilliger, 2000)] or a different set of program input values. These MB tasks, like the others in the structure-determination workflow, will then be distributed to the computing server and run in parallel.

Once the requested job has been completed, results from the various tasks associated with the workflow will be harvested and analyzed. A summary report for the whole process will be generated and presented to the user *via* a built-in text editor and graphical tool on the client side. In the case of a MAD analysis, parallel SAS phasing tasks for peak data will automatically be set up together with the MAD phasing tasks and these results will be included in the summary.

In the case of structure determination by molecular replacement, the workflow engine takes the input sequence file of the target protein and begins with *BLAST* (Altschul *et al.*, 1990), searching the Protein Sequence Databases at NCBI (National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov). The hits are sorted by LIS score and listed (see Table 2) on the *SGXPro* client side. The LIS score is defined as the product of the *BLAST* score and the aligned sequence length. By default, the top structure on the list (the user can also select multiple structures) will be downloaded automatically from the Protein Data Bank (http://www.pdb.org) and passed on to the MR programs as the template (potential sequence homologue) structures. Through the interface, the user is also allowed to customize the workflow engine by choosing different programs and/or different parameter settings for each program. As before, each unique MR task is dynamically generated and will employ either a different program [for example, *AMoRe* (Navaza, 1994) or *EPMR* (Kissinger *et al.*, 1999)] or a different set of parameters (such as template structure and resolution range) and run in parallel on the computing server. Upon completion, the workflow engine will summarize the output and generate a separate list of models from each individual program employed, which is sorted according to CC-F and *R* factor (correlation coefficient and classic *R* factor, respectively, between the observed amplitudes for the crystal and the calculated amplitudes for the model). The CC-F and *R* factor are sensible sorting targets of MR solutions. However, to determine if a model is a real solution or not, further checking of the density map, clashes when put into the crystal cell and refinement of the structure are needed. Fig. 2 illustrates how the parallel workflow proceeds in the above MAD/SAS and MR phasing cases.

2.2. Client/server architecture

Owing to advances in computing technology, a variety of computing platforms are accessible to the structural biology community. In order to support a wide variety of platforms while maintaining a uniform and user-friendly interface, *SGXPro* has been developed using client/server architecture with the parallel workflow engine described above acting as the server and the GUI (graphical user interface) acting as the client. This approach separates the functional modules from the GUI and allows the user to select the best computing platform available, including those at remote sites. Compared with the web-based designs adopted by some of the other packages, the *SGXPro* suite is self-contained and can be installed on any computing platforms, including single-CPU

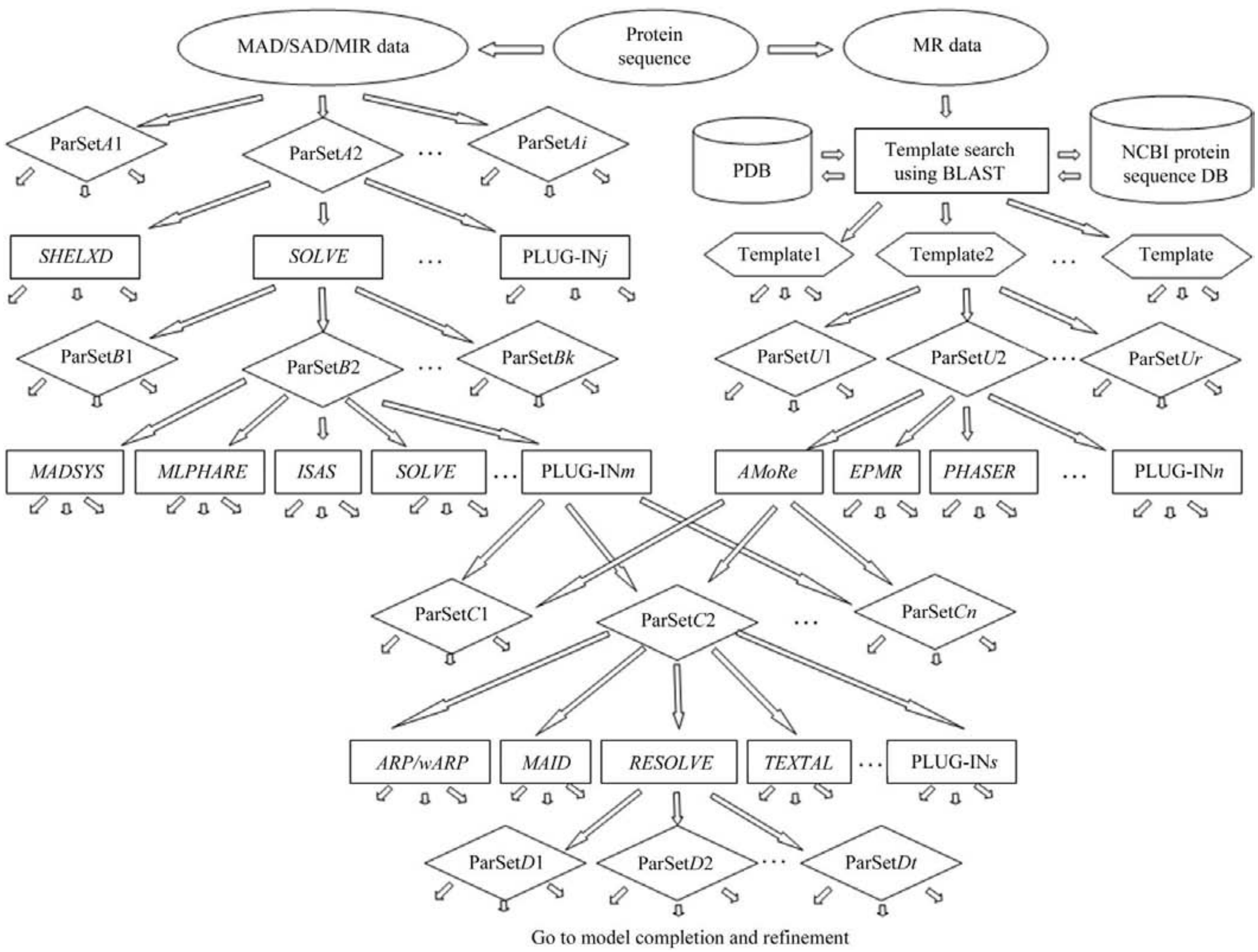

Figure 2
Flow charts of the *SGXPro* jobs to solve structures using MAD/SAS/MIR or MR methods described in §2.1. The workflow and the parallel tasks at each step are dynamically generated based on the user's requests, the input data, the analysis of the results of the previous step and crystallographic computing logic pertaining to the job.

personal computers, multi-CPU workstations and powerful Linux clusters. It also allows the user to choose a private port for transferring data to and from the server, which avoids using the public port 80 that is visible to and shared by all internet traffic.

In addition, *SGXPro* can create a subsystem on a local or remote computing facility, which allows the owner of the *SGXPro* server to set up accounts for other users. This ability is useful in the case of a group of users who want to install and share the *SGXPro* server while maintaining privacy for each member of the group.

2.3. Plug-in-and-play

The plug-in-and-play architecture of *SGXPro* was designed to allow the incorporation of program updates or for the addition of any programs. It also allows users to choose a group of programs to customize their own *SGXPro*. Basically, *SGXPro* provides the workflow engine and the interfaces for all the supported programs. The users do not need to install all the supported programs to run *SGXPro*. They can select to install (plug-in) a group of programs that are available on their systems.

2.4. User-friendly interface

A user-friendly interface, with intelligent defaults, has been designed that includes a variety of helper/expert functions for analyzing, deriving and validating data and parameter settings to simplify input and eliminate errors. In the process of macromolecular structure determination, the user is normally asked to provide a lot of information. Some of this information is readily available from file headers or the log files from previous steps. Other information, although not explicitly stated, can be derived, which may include resolution range, symmetry operators, wavelength, molecular weight, solvent content, atomic scattering factors and anomalous scattering factors. The *SGXPro* interface has been designed to minimize user input with most of the information internally derived by the use of these helper/expert functions. This greatly reduces user effort and eliminates errors in preparing and configuring program input. In a typical case, only a few mouse clicks are needed to set up a complicated structure-determination job that will search program/parameter space to give the best possible solution for the given data.

3. Test and results

The current version of *SGXPro* supports over a dozen popular programs covering most aspects of the structure-determination process, *i.e.* heavy-atom search, phasing, density modification, molecular replacement, sequence analysis, auto-tracing and model building. These programs include *SHELXD*, *SOLVE*, *RESOLVE*, *ISAS*, *BLAST*, *AMoRe*, *EPMR*, *DM* (Cowtan, 1994), *DMMULTI* (Cowtan, 1994), *SOLOMON* (Abrahams & Leslie, 1996), *ARP*/*wARP* and *MAID* (Levitt, 2001). The next version of *SGXPro* will include plug-ins of some other popular programs including those for structure refinement, validation and deposition. The software package is currently under beta test at several institutions, including two synchrotron beamlines. Students also used *SGXPro* at the 2004 American Crystallography Summer Course for Crystallography held at the Illinois Institute of Technology/Advanced Photon Source. Shown below are results from three test cases of SAS/MAD phasing and one case of MR (molecular replacement).

3.1. SAS/MAD phasing

3.1.1. Test data.

(i) *Nigerythrin*. The SAS data (provided by courtesy of Dr Lanzilotta at the University of Georgia) were collected from a single nigerythrin crystal (space group $P2_12_12_1$, unit-cell parameters a = 46.84, b = 72.79, c = 119.69 Å) to 1.85 Å resolution at the Advanced Light Source (beamline 8.2.2) using 1.65 Å X-rays and an ADSC Quantum 315 CCD detector. The data were processed using the *DENZO*/*SCALEPACK* package. The overall completeness of the data set is 96.9%, with an $R_{\rm merge}$ of 4.2% and a redundancy of 8.5 (90.6%, 12.1% and 6.7, respectively, for the highest resolution bin 1.96–1.85 Å). The protein has 202 amino-acid residues and three Fe atoms.

(ii) *C-terminal domain of a corrinoid-binding protein* (denoted hereafter as CBP). The protein contains 125 amino acids with an unknown number of Co sites. The SAS data were collected from one crystal (space group $P2_12_12$, unit-cell parameters a = 55.52, b = 62.65, c = 34.45 Å) to a resolution of 2.30 Å using a Bruker Proteum-R CCD detector mounted on a Rigaku FRD generator using Cu $K\alpha$ X-rays with Rigaku/MSC HiRes^2 optics. A total of 1200 0.3° oscillation images were recorded using an exposure time of 30 s. The intensities were indexed and integrated using the Bruker *PROTEUM* data-reduction package. Experimental error correction and scaling were performed using *PROSCALE* (Fu *et al.*, 2000). The overall completeness of the data set is 99.2%, with an $R_{\rm merge}$ of 4.9% and a redundancy of 7.6 (92.1%, 7.8% and 6.4, respectively, for the highest resolution bin 2.45–2.30 Å).

(iii) *Pfu631545*. The Se-containing protein contains 133 amino acids (including an N-terminal six-His tag) with one Se site. The MAD data were collected (space group $P2_1$, unit-cell parameters a = 36.35, b = 61.09, c = 51.58 Å) on a MAR CCD225 detector at APS SERCAT beamline 22-ID with three wavelengths. The data were integrated using *DENZO* and scaled using *PROSCALE*: 260° (two 130° passes) of data at 0.97828 Å to resolution 1.99 Å (highest resolution bin 2.08–1.99 Å) with $R_{\rm merge}$ = 7.0% (23.8%), a redundancy of 5.2 (5.1) and a completeness of 89.7 (87.6); 180° of data at 0.97941 Å to 2.0 Å resolution (highest resolution bin 2.09–2.00 Å) with $R_{\rm merge}$ = 7.1% (24.4%), a redundancy of 3.5 (3.2) and a completeness of 99.2% (96.1); 360° of data at 0.98086 Å to a resolution of 1.98 Å (highest resolution bin 2.08–1.98 Å) with $R_{\rm merge}$ = 7.8% (25.3%), a redundancy of 8.3 (7.4) and a completeness of 99.5% (95.0). Pfu631545 and CBP data are from projects at SECSG.

3.1.2. Results. Each test started with the following set of minimum required information: amino-acid sequence, name of file containing the reduced data, unit cell and space group. All *SGXPro* structure-determination jobs were set to run on a remote 128-processor IBM Linux Cluster and include identification of anomalous scattering sites, phasing/density modification and autotracing/model building. Programs explored in these analyses were *SHELXD* and *SOLVE* for locating the anomalous scattering sites, *ISAS* and *SOLVE*/*RESOLVE* for phasing and density modification and *ARP*/*wARP* and *RESOLVE* for autotracing and model building. Five different data-resolution cutoffs were used in each step: 1.85, 2.25, 2.65, 3.05 and 3.5 Å for the nigerythrin data, 2.30, 2.60, 2.90, 3.20 and 3.5 Å for the CBP data and 2.00, 2.50, 2.75, 3.00 and 3.50 Å for the Pfu631545 data.

Each job began by using built-in tools to quickly calculate the protein's molecular weight and investigate the possible number of molecules in an asymmetric unit by estimating the solvent content using the Matthews algorithm (Matthews, 1968). From the Matthews analysis, both monomeric and dimeric forms are possible for the nigerythrin crystal (73.3 and 46.7% solvent, respectively), monomers for the CBP crystal (43.5% solvent) and monomers or dimers for Pfu631545 (corresponding to 70.5 and 41.0% solvent, respectively). Based on this, the following number of heavy-atom sites were assigned in searching for the location of the anomalous scatterers: three or six for nigerythrin, four for CBP (since the number of heavy-atom sites in the CBP molecule was not known, four sites were sought) and one and two for Pfu631545.

The results from the initial search for heavy-atom sites clearly suggested that the nigerythrin is a dimer since six sites were found. CBP is a monomer, with only one site found. Two sites were found for Pfu631545, also suggesting a dimer. These intermediate results were used to dynamically set up the parallel workflow engine for the phasing and density modification tasks that follow. As the process proceeded to the end of the workflow, initial models automatically built from the finished phasing tasks were analyzed together with the other output files. The overall process summary includes a list of solutions sorted by descending order of number of amino-acid residues automatically traced and is presented on the client side. There are 77, 66 and 98 finished tasks for the nigerythrin, CBP and Pfu631545 jobs described above. The top 20 solutions on each of these lists are shown in Tables 1(*a*), 1(*b*) and 1(*c*).

Table 1
Results of *SGXPro* jobs in solving the three protein structures by SAS/MAD phasing.

Each row lists a set (or a pipe) of tasks from heavy-atom searching, phasing and density modification to automatic tracing and model building. The results are sorted according to the descending order of number (Naa) or percentage (%) of amino-acid residues automatically traced. Tables 1(*a*), 1(*b*) and 1(*c*) are for the nigerythrin, CBP and Pfu631545 data, which have 77, 66 and 98 finished sets of tasks, respectively. Only the top 20 sets of tasks were listed in the table to save space.

(*a*) Nigerythrin.

Heavy-atom search		Phasing and density modification		Automatic tracing and model building			
Programs	Resolution (Å)	Programs	Resolution (Å)	Programs	Resolution (Å)	Naa	%
SHELXD	3.50	*SOLVE*/*RESOLVE*	1.85	*ARP*/*wARP*	1.85	400	99.0
SHELXD	3.50	*ISAS*	1.85	*ARP*/*wARP*	1.85	399	98.8
SHELXD	3.05	*SOLVE*/*RESOLVE*	1.85	*ARP*/*wARP*	1.85	394	97.5
SHELXD	3.50	*SOLVE*/*RESOLVE*	2.25	*RESOLVE*	2.25	394	97.5
SHELXD	2.65	*ISAS*	1.85	*ARP*/*wARP*	1.85	392	97.0
SOLVE	1.85	*SOLVE*/*RESOLVE*	1.85	*ARP*/*wARP*	1.85	391	96.8
SHELXD	2.65	*SOLVE*/*RESOLVE*	1.85	*ARP*/*wARP*	1.85	390	96.5
SHELXD	2.65	*SOLVE*/*RESOLVE*	2.25	*ARP*/*wARP*	2.25	389	96.3
SHELXD	3.05	*SOLVE*/*RESOLVE*	2.25	*ARP*/*wARP*	2.25	388	96.0
SOLVE	1.85	*SOLVE*/*RESOLVE*	1.85	*RESOLVE*	1.85	367	90.8
SHELXD	2.65	*SOLVE*/*RESOLVE*	1.85	*RESOLVE*	1.85	367	90.8
SHELXD	3.05	*SOLVE*/*RESOLVE*	1.85	*RESOLVE*	1.85	367	90.8
SHELXD	3.05	*ISAS*	1.85	*ARP*/*wARP*	1.85	363	89.9
SHELXD	3.05	*SOLVE*/*RESOLVE*	2.25	*RESOLVE*	2.25	358	88.6
SHELXD	3.50	*SOLVE*/*RESOLVE*	1.85	*RESOLVE*	1.85	357	88.4
SHELXD	3.50	*SOLVE*/*RESOLVE*	2.25	*RESOLVE*	2.25	353	87.4
SHELXD	2.65	*SOLVE*/*RESOLVE*	2.25	*RESOLVE*	2.25	352	87.1
SHELXD	3.05	*ISAS*	1.85	*RESOLVE*	1.85	325	80.4
SHELXD	3.50	*SOLVE*/*RESOLVE*	2.65	*RESOLVE*	2.65	325	80.4
SHELXD	3.05	*SOLVE*/*RESOLVE*	2.65	*RESOLVE*	2.65	324	80.2

(*b*) CBP.

Heavy-atom search		Phasing and density modification		Automatic tracing and model building			
Programs	Resolution (Å)	Programs	Resolution (Å)	Programs	Resolution (Å)	Naa	%
SHELXD	3.50	*SOLVE*/*RESOLVE*	2.60	*RESOLVE*	2.60	102	81.6
SHELXD	3.20	*SOLVE*/*RESOLVE*	2.30	*RESOLVE*	2.30	102	81.6
SHELXD	3.20	*ISAS*	2.60	*RESOLVE*	2.60	100	80.0
SHELXD	2.60	*SOLVE*/*RESOLVE*	2.60	*RESOLVE*	2.60	99	79.2
SHELXD	2.90	*SOLVE*/*RESOLVE*	2.30	*RESOLVE*	2.30	95	76.0
SHELXD	3.20	*SOLVE*/*RESOLVE*	2.60	*RESOLVE*	2.60	94	75.2
SHELXD	3.20	*SOLVE*/*RESOLVE*	3.20	*RESOLVE*	3.20	93	74.4
SHELXD	3.20	*ISAS*	3.20	*RESOLVE*	3.20	91	72.8
SHELXD	2.60	*SOLVE*/*RESOLVE*	2.30	*RESOLVE*	2.30	91	72.8
SHELXD	3.50	*SOLVE*/*RESOLVE*	3.20	*RESOLVE*	3.20	89	71.2
SHELXD	2.90	*ISAS*	2.30	*RESOLVE*	2.30	87	69.6
SHELXD	2.60	*ISAS*	3.20	*RESOLVE*	3.20	87	69.6
SHELXD	3.50	*ISAS*	2.30	*RESOLVE*	2.30	85	68.0
SHELXD	3.20	*ISAS*	2.90	*RESOLVE*	2.90	85	68.0
SHELXD	2.90	*SOLVE*/*RESOLVE*	2.90	*RESOLVE*	2.90	84	67.2
SHELXD	3.50	*ISAS*	2.60	*RESOLVE*	2.60	83	66.4
SHELXD	3.20	*SOLVE*/*RESOLVE*	2.90	*RESOLVE*	2.90	83	66.4
SHELXD	3.20	*ISAS*	2.30	*RESOLVE*	2.30	83	66.4
SHELXD	3.50	*SOLVE*/*RESOLVE*	2.30	*RESOLVE*	2.30	82	65.6
SHELXD	3.50	*SOLVE*/*RESOLVE*	2.90	*RESOLVE*	2.90	80	64.0

Table 1 (continued)

(*c*) Pfu631545. # indicates MAD phasing.

Heavy-atom search		Phasing and density modification		Automatic tracing and model building			
Programs	Resolution (Å)	Programs	Resolution (Å)	Programs	Resolution (Å)	Naa	%
SHELXD	3.50	*SOLVE*/*RESOLVE*	3.00	*RESOLVE*	3.00	160	60.2
#*SHELXD*	2.50	*SOLVE*/*RESOLVE*	3.00	*RESOLVE*	3.00	145	54.5
SHELXD	3.50	*ISAS*	3.00	*RESOLVE*	3.00	143	53.8
#*SHELXD*	3.50	*SOLVE*/*RESOLVE*	2.75	*RESOLVE*	2.75	140	52.6
#*SHELXD*	2.50	*SOLVE*/*RESOLVE*	2.75	*RESOLVE*	2.75	138	51.9
SHELXD	3.50	*SOLVE*/*RESOLVE*	2.50	*RESOLVE*	2.50	135	50.8
SHELXD	2.75	*ISAS*	3.50	*RESOLVE*	3.50	130	48.9
SHELXD	2.50	*ISAS*	3.50	*RESOLVE*	3.50	130	48.9
#*SHELXD*	2.75	*SOLVE*/*RESOLVE*	3.00	*RESOLVE*	3.00	128	48.1
#*SHELXD*	2.00	*SOLVE*/*RESOLVE*	3.00	*RESOLVE*	3.00	125	47.0
SHELXD	3.00	*ISAS*	3.00	*RESOLVE*	3.00	123	46.2
#*SHELXD*	3.50	*SOLVE*/*RESOLVE*	3.00	*RESOLVE*	3.00	123	46.2
#*SHELXD*	3.50	*SOLVE*/*RESOLVE*	3.50	*RESOLVE*	3.50	123	46.2
#*SHELXD*	2.75	*SOLVE*/*RESOLVE*	2.75	*RESOLVE*	2.75	122	45.9
SHELXD	3.50	*SOLVE*/*RESOLVE*	2.75	*RESOLVE*	2.75	120	45.1
SHELXD	2.75	*ISAS*	2.00	*RESOLVE*	2.00	120	45.1
SHELXD	2.50	*SOLVE*/*RESOLVE*	3.00	*RESOLVE*	3.00	120	45.1
#*SHELXD*	2.00	*SOLVE*/*RESOLVE*	3.50	*RESOLVE*	3.50	120	45.1
SHELXD	3.50	*SOLVE*/*RESOLVE*	2.00	*RESOLVE*	2.00	118	44.4
SHELXD	2.50	*ISAS*	3.00	*RESOLVE*	3.00	118	44.4

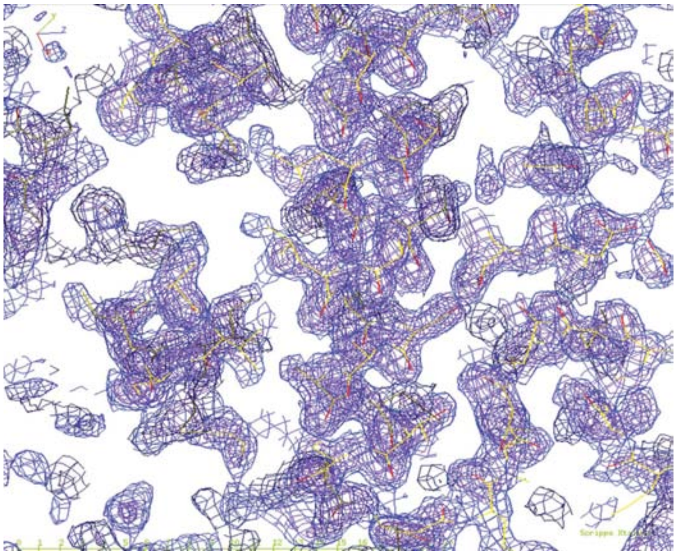

Figure 3
Electron-density map around part of the autotraced model calculated by using phases from the top solution in Table 1(*c*).

The results show a broad range of the percentage of the number of residues automatically traced: ranging from 0.7 to 99.0% for nigerythrin, 3.2 to 81.6% for CBP and 20.7 to 60.2% for Pfu631545. As seen in Table 1, the top results were all produced by tasks that used a combination of different programs and resolutions. For high-quality data in the nigerythrin case, nine of the 77 finished tasks generated near-complete models with over 96% of the total 404 amino-acid residues automatically traced. Using all data (without resolution cutoff), *SOLVE*/*RESOLVE* alone also gave a 90.8% complete model. In the case of CBP, a data set of average quality, the top ten of the 66 finished tasks automatically traced over 70% of the total 125 amino-acid residues with a maximum trace of 81.6%. These top solutions are all from tasks combining different programs and resolutions. Using all data with *SOLVE*/*RESOLVE* alone generated a model with 52.8% (not shown in Table 1*b*) traced amino-acid residues. Although the electron-density maps from all these tasks are traceable, the models of the top solutions provide the best model. The top solution normally has a better electron-density map, which can make the manual fitting much easier to complete and refine the structure. The best solution has been used to complete the structure, which was refined against a set of high-resolution data and deposited in the Protein Data Bank with entry code 1y80. For the Pfu631545 case, a low-quality data set, only the top six (also from tasks combining different programs and resolutions) of the 98 finished tasks automatically traced over 50% of the amino-acid residues. Some efforts are necessary to manually fit the density to complete the model. Fig. 3, drawn using *XTALVIEW* (McRee, 1992), shows the electron-density map around part of the auto-traced model from the top solution in Table 1(*c*). Manual fitting is under way to complete and refine the structure, which will be published separately.

3.2. Molecular replacement

3.2.1. Test data. The native data were collected from one crystal of Humhsp40 C-terminal binding domain (space group $C222_1$, unit-cell parameters a = 97.01, b = 191.13, c = 40.96 Å) to 2.82 Å resolution at the APS SERCAT beamline 22-ID using a MAR CCD225 detector. The data were processed using the *DENZO*/*SCALEPACK* package. The overall completeness of the data set is 93.6%, with an $R_{\rm merge}$ of 6.2% and a redundancy of 6.3 (75.8%, 29.8% and 4.6, respectively, for the highest resolution bin 2.96–2.82 Å). The protein has 170 amino-acid residues. The data are provided courtesy of Dr Sha at University of Alabama at Birmingham.

3.2.2. Results. The workflow engine started with *BLAST* searching, which generated a list of structures (see Table 2). The top two hits, 1c3g (40.0% sequence identity and 94.12% aligned length) and 1hdj (100% sequence identity and 44.12% aligned length), were selected as potential structural homologs for molecular replacement. 1hdj was selected because its sequence is equal to a segment of the target protein Humhsp40 sequence (100% identity), although the structure has only 77 residues, about 44.12% of the sequence length of Humhsp40. All others have both sequence identity and aligned sequence length lower than 50%. In this test, *AMoRe* and *EPMR* were chosen, each with three different resolution cutoffs (3.0, 3.5 and 4.0 Å), to perform the orientation and translation search. With this setup, the workflow engine

Table 2
BLAST search results of Humhsp40.

PDB code is the entry code of the structure in the Protein Data Bank. *E* value and Score are statistics from *BLAST*. LIS is the sorting score of template structures that is defined in §2.1.

PDB code	Chain ID	Identity (%)	Aligned length (%)	Score	LIS	*E* value
1c3g	*A*	40.00	94.12	110.2	103.7	3.9×10^{-25}
1hdj		100.00	44.12	145.2	64.1	9.7×10^{-36}
1bqz		48.00	44.12	78.95	34.8	8.3×10^{-16}
1bq0		48.61	42.35	78.18	33.1	1.3×10^{-15}
1xbl		48.61	42.35	77.41	32.8	2.8×10^{-15}
1nz6	*A*	25.37	39.41	32.34	12.7	0.106
1gh6	*A*	35.29	30.00	29.65	8.9	0.626
1c20	*A*	37.50	14.12	30.42	4.3	0.383
1kqq	*A*	37.50	14.12	30.03	4.2	0.513

automatically generated 12 MR tasks which were then distributed to the Linux cluster. Table 3 lists the top hit from each *AMoRe* or *EPMR* task. In this case, the model from *EPMR* running at 3.0 Å with template 1c3g gave the best solution which generates a high-quality density map. After putting in the sequence of the target protein Humhsp40, the *R* factor and $R_{\rm free}$ are 28.7 and 34.2%, respectively, from rigid-body refinement.

Table 3
The top solution from each of the MR tasks.

PDB code is the entry code of the template structure in the Protein Data Bank. Res. is the high-resolution cutoff used for MR search. CC-F and *R* factor are the correlation coefficient and classic *R* factor, respectively, between the observed amplitudes for the crystal and the calculated amplitudes for the model, which are extracted from the output of *EPMR* or *AMoRe*.

Program	PDB code	Res. (Å)	CC-F	*R* factor
EPMR	1c3g	3.0	0.476	0.583
EPMR	1c3g	3.5	0.407	0.575
EPMR	1c3g	4.0	0.395	0.551
EPMR	1hdj	3.0	0.356	0.587
EPMR	1hdj	3.5	0.285	0.589
EPMR	1hdj	4.0	0.244	0.596
AMoRe	1c3g	3.0	0.341	0.617
AMoRe	1c3g	4.0	0.339	0.616
AMoRe	1c3g	3.5	0.337	0.611
AMoRe	1hdj	3.0	0.228	0.634
AMoRe	1hdj	4.0	0.221	0.639
AMoRe	1hdj	3.5	0.219	0.641

Table 4
Statistics and number of heavy-atom sites found by *SHELXD*.

CC all/weak and PATFOM are the correlation coefficient and figure of merit from *SHELXD*.

Data	Resolution (Å)	No. sites	CC all/weak	PATFOM
Nigerythrin	1.85	6	35.99/20.82	28.99
Nigerythrin	2.25	6	44.18/27.30	39.43
Nigerythrin	2.65	6	52.27/34.18	51.58
Nigerythrin	3.05	6	49.76/32.35	46.95
Nigerythrin	3.50	6	50.87/34.63	41.78
CBP	2.30	1	34.97/22.07	193.43
CBP	2.60	1	36.59/23.10	216.27
CBP	2.90	1	36.71/23.75	173.15
CBP	3.20	1	38.28/23.37	215.80
CBP	3.50	1	40.99/24.37	261.57
Pfu631545	2.00	2	28.48/18.49	58.28
Pfu631545	2.50	2	34.76/25.07	127.92
Pfu631545	2.75	2	35.84/25.23	129.55
Pfu631545	3.00	2	37.88/26.15	132.47
Pfu631545	3.50	2	38.64/26.65	131.01

4. Discussion

In the test examples described above, the search is five-dimensional: (i) programs/algorithms for finding heavy-atom or anomalous sites, (ii) resolution for heavy-atom or anomalous site searching, (iii) programs/algorithms for phasing and electron-density modification, (iv) resolution for phasing and electron-density modification and (v) programs/algorithms for autotracing and model building. The parallel workflow engine is not limited to the five dimensions as used in the test examples. Other plugged-in programs and parameters can also be searched from *SGXPro* if needed.

From the test results, it is interesting to note that the number of amino-acid residues automatically traced (which is the indicator of the quality of solution) is widely distributed in program/parameter space and that different combinations of these variables produced quite different results, which suggests that it is necessary to search not only parameter space but also the program space in order to obtain the best possible initial model from a given data set. In practice, the best results cannot be expected from a fixed combination of programs and parameters. Therefore, the systematic search of these variables is beneficial for finding the best solution, especially in the case of data with marginal quality where in general high-throughput structure-determination pipelines fail.

As indicated in the above test cases, in order to achieve the best possible solution for a given data set one must explore both program and parameter space. A manual trial-and-error approach could be a very tedious and time-consuming process, with the user usually giving up before all combinations are explored. The parallel workflow engine described in this study can greatly reduce the time and human intervention for arriving at the best solution for a given data set, allowing the user to focus on the science of the project and not on the details of the structure determination.

In the case of MAD data, *SGXPro* will try both MAD phasing and SAS phasing with the peak data. It has been shown that SAS phasing may produce a better initial model in some cases owing to the less stringent wavelength requirements (Dauter *et al.*, 2002; Dodson, 2003).

In addition, the *SGXPro* client/server architecture will allow users to utilize the most powerful computing facility available. The plug-in-and-play design provides an open platform that not only allows easy integration of any program into the system but also allows the user to run *SGXPro* with a subset of supported programs that are available. All control parameters of the programs have defaults which are either from the programs, if available, or derived from the data and/or previous input. For example, the default resolution cutoff of *SHELXD* was set to 3.5 Å as suggested by the manual of the program, which works well for most cases, and the resolution cutoff of the phasing programs was set to the highest resolution derived from the input data file. The test cases also show that the statistics from the heavy-atom sites searching do not have obvious correlation with the success of the phasing step. Table 4 lists the statistics from *SHELXD*. From this table it can be seen that *SHELXD* gave reasonably good statistics at different resolutions for all three SAS/MAD cases, while the phasing results are quite different. The resolution search is more sensitive at the phasing step, especially for average- and low-quality data, as shown in Table 1. The use of intelligent defaults and a user-friendly GUI results in minimal user input for the setup of complicated structure process as shown in Fig. 2. All these features make *SGXPro* flexible to fit into different hardware and software environments so that the user can easily set up the program/parameter space searching accordingly. For example, the user can start with the program-space search using all parameter defaults then move on to parameter search by either slightly adjusting around the default or using large steps based on the results from the program-space search. This would limit the number of tasks to only a few at one time, which can be handled by a current common desktop PC.

We thank the many PhD students from the Department of Computer Sciences at University of Georgia, including George Wu, Dongsheng Che, Jizheng Zhao, Haijin Yan, Feng Sun, Joe Zhu and Xiaochuan Yi, all of whom helped with the implementation of the GUI and network computing components. We thank Drs Zhongmin Jin, John Chrzas, Jim Fait at APS SER-CAT beamline, and Andy Howard at Illinois Institute of Technology for suggestions during the beta test. We thank Dr Bingdong Sha at University of Alabama at Birmingham for the use of the Humhsp40 data and Dr William Lanzilotta at University of Georgia for the use of the nigerythrin data. We also thank Dr Gary Newton for help in preparation of the manuscript. This work was supported in part with funds from the National Institutes of Health (GM62407), The Georgia Research Alliance and the University of Georgia Research Foundation.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* **215**, 403–410.

Adams, P. D., Gopal, K., Grosse-Kunstleve, R. W., Hung, L.-W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., Pai, R. K., Read, R. J., Romo, T. D., Sacchettini, J. C., Sauter, N. K., Storoni, L. C. & Terwilliger, T. C. (2004). *J. Synchrotron Rad.* **11**, 53–55.

Blanc, E., Vonrhein, C., Roversi, P. & Bricogne, G. (2000). *Acta Cryst.* A**56**, S107.

Brunzelle, J. S., Shafaee, P., Yang, X., Weigand, S., Ren, Z. & Anderson, W. F. (2003). *Acta Cryst.* D**59**, 1138–1141.

Calderone, V. (2004). *Acta Cryst.* D**60**, 2150–2155.

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

Dauter, Z., Dauter, M. & Dodson, E. (2002). *Acta Cryst.* D**58**, 494–506.

Dodson, E. (2003). *Acta Cryst.* D**59**, 1958–1965.

Fu, Z.-Q., Pressprich, M., Sparks, R., Foundling, S. & Phillips, J. (2000). Abstr. Annu. Meet. Am. Crystallogr. Assoc., Abstract P066.

Fu, Z.-Q., Rose, J. & Wang, B.-C. (2003). *Proceedings of the Fifth International Conference on Molecular Structural Biology, Vienna, Austria, September 3–7*, p. 24.

Gonzalez, A., Pedelacq, J. D., Sola, M., Gomis-Ruth, F.-X., Coll, M., Samama, J. P. & Benini, S. (1999). *Acta Cryst.* D**55**, 1449–1458.

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

Hendrickson, W. A. & Teeter, M. M. (1981). *Nature (London)*, **290**, 107–113.

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

Jiang, J.-S. & Lin, Z. (2005). Submitted.

Karle, J. (1980). *Int. J. Quant. Chem.* **7**, 357–367.

Kissinger, C. R., Gehlhaar, D. K & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.

Levitt, D. G. (2001). *Acta Cryst.* D**57**, 1013–1019.

Liu, Z.-J., Lin, D., Tempel, W., Praissman, J., Rose, J. & Wang, B.-C. (2005). *Acta Cryst.* D**61**, 520–527.

McRee, D. E. (1992). *J. Mol. Graph.* **10**, 44–47.

Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.

Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.

Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. (2005). *Acta Cryst.* D**61**, 449–457.

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

Phillips, J. C. & Hodgson, K. O. (1980). *Acta Cryst.* A**36**, 856–864.

Potterton, E., Briggs, P., Turkenburg, M. & Dodson, E. (2003). *Acta Cryst.* D**59**, 1131–1137.

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

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

Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.

Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.

Weeks, C. M., Rappeleye, J., Furey, W., Miller, R., Potter, S. A., Smith, G. D. & Xu, H. (2001). Abstr. Annu. Meet. Am. Crystallogr. Assoc., Abstract No. W0264.