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

# addenda and errata

## Parameter-space screening: a powerful tool for high-throughput crystal structure determination. Corrigendum

**Zhi-Jie Liu,* Dawei Lin, Wolfram Tempel, Jeremy L. Praissman, John P. Rose and Bi-Cheng Wang***

Southeast Collaboratory for Structural Genomics, Department of Biochemistry and Molecular Biology, The University of Georgia, Athens, Georgia, USA. Correspondence e-mail: liu@secsg.uga.edu, wang@bcl1.bmb.uga.edu

Fig. 4 in the article by Liu *et al.* [(2005), *Acta Cryst.* D**61**, 520–527] was labelled incorrectly. A corrected version of the figure is given here. Also in §3.1.3 of the original article the Cr $K\alpha$ wavelength was given incorrectly. It should be 2.29 Å.

### References

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

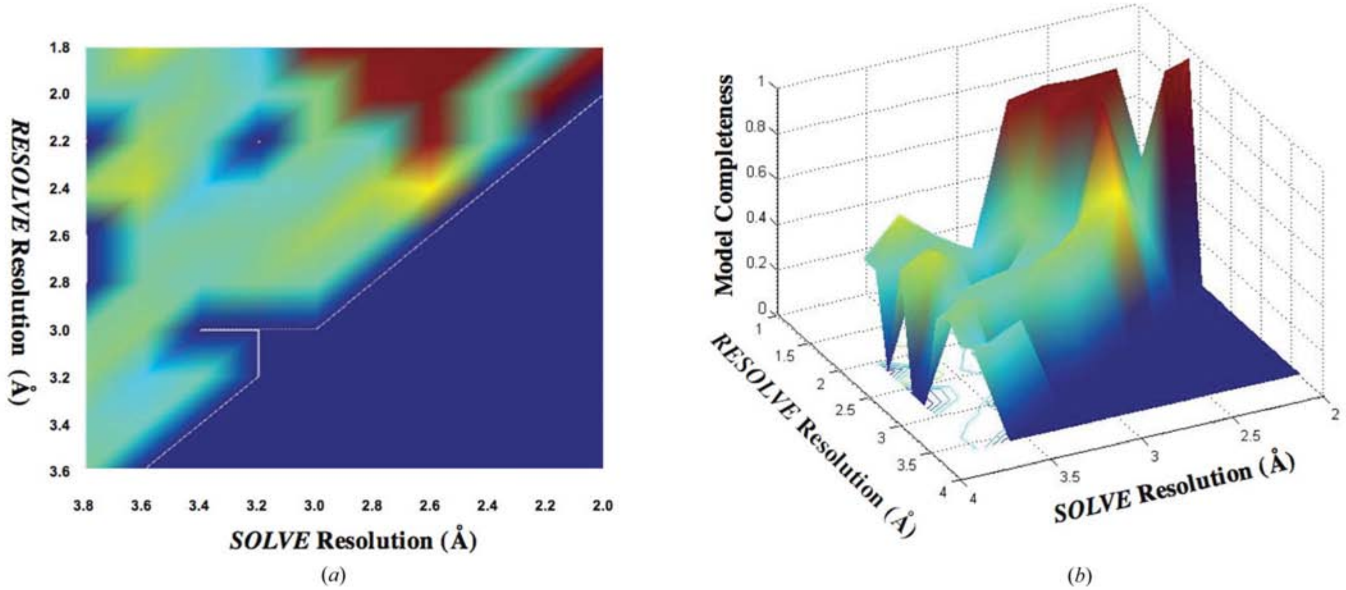


**Figure 4**
A graphical representation of pipeline success space for the PA-L1 example. A total of 55 *SOLVE*/*RESOLVE* phase sets were used as input to *ARP*/*wARP*. The colour scheme used represents success (number of residues fitted), with red indicating a near-complete model and blue/cyan representing cases where model building failed. An interesting and unexpected feature is that success space is not continuous with regions of low success sandwiched between regions of high success.

doi:10.1107/S0907444905024996

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

# Parameter-space screening: a powerful tool for high-throughput crystal structure determination

**Zhi-Jie Liu,* Dawei Lin, Wolfram Tempel, Jeremy L. Praissman, John P. Rose and Bi-Cheng Wang***

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

Correspondence e-mail: liu@secsg.uga.edu, wang@bcl1.bmb.uga.edu

Received 4 October 2004
Accepted 28 January 2005

The determination of protein structures on a genomic scale requires both computing capacity and efficiency increases at many stages along the complex process. By combining bioinformatics workflow-management techniques, cluster-based computing and popular crystallographic structure-determination software packages, an efficient and powerful new tool for structural biology/genomics has been developed. Using the workflow manager and a simple web interface, the researcher can, in a few easy steps, set up hundreds of structure-determination jobs, each using a slightly different set of program input parameters, thus efficiently screening parameter space for the optimal input-parameter combination, *i.e.* a set of parameters that leads to a successful structure determination. Upon completion, results from the programs are harvested, analyzed, sorted based on success and presented to the user *via* the web interface. This approach has been applied with success in more than 30 cases. Examples of successful structure determinations based on single-wavelength scattering (SAS) are described and include cases where the 'rational' crystallographer-based selection of input parameters values had failed.

## 1. Introduction

Owing to the continued improvements in hardware, software and experimental techniques over the past decade, X-ray diffraction experiments produce data of higher quality and resolution than ever before. However, crystal structure determination, in the case of macromolecules, continues to be a complicated multi-step process that typically includes identification and refinement of the phasing substructure (heavy atoms or anomalous scatterers), generation of protein phases, density modification, tracing the peptide chain, building and refining the protein model, validation and publication. Because of the complexity of the protein crystal structure determination, many bottlenecks and decision points remain that slow down the process. Automation of some aspects of protein structure determination has advanced considerably. Program packages such as *SOLVE*/*RESOLVE* (Terwilliger, 2002), *ARP*/*wARP* (Perrakis *et al.*, 1999) and *CCP*4 (Collaborative Computational Project, Number 4, 1994) have partially automated protein structure determination, but the crystallographer's attention is still required in order to answer the following questions. (i) Is the data set of sufficient quality to permit solution of the structure? (ii) Among several alternative strategies, methods and computer programs, all with specific strengths and weaknesses, which one(s) is(are) most appropriate for the given problem? (iii) What are the appropriate values for the input parameters for each program? (iv)

If more than one source of data (native, derivatives *etc.*) is available, which data set should be used? Or, what is the best way to combine them, if appropriate? (v) At each step, do the results/output indicate that one can reasonably proceed to the next step? If not, should more and/or better data be collected? The crystallographer generally addresses these questions in a trial-and-error process based on his/her experience by adjusting the parameters based on the previous results and repeating the computation. This process is not only very inefficient owing to the limitations of a manual operation, but it also often results in missing a solution even if the data could provide one. Increased throughput requirements of the structural genomics era aggravate this shortcoming.

Continued growth in computational power and maturing computer cluster technology gives today's crystallographer computer resources unheard of a decade ago and, together with improved algorithms and new approaches, has significantly reduced the average time of the structure-determination process (from data collection to Protein Data Bank submission) from a number of months to a matter of days. The Southeast Collaboratory for Structural Genomics (SECSG; Adams *et al.*, 2003), like other structural genomics centers (Norvell, 2000), is pursuing the integration of different crystallographic programs into a structure-determination pipeline. Examples of existing pipelines include a combination of *SHARP* with *ARP*/*wARP*, *ACrS* (Brunzelle *et al.*, 2003) and *ELVES* (Holton & Alber, 2004). The availability of a 128-processor computer cluster at SECSG and a custom dictionary-driven workflow-management system (Praissman *et al.*, 2003) allows multiple structure-determination jobs to be run in parallel, with each job run with a slightly different set of program input parameters. This approach increases the success rate of structure solution by (i) exploring a significantly larger fraction of program parameter space and (ii) by sampling program parameter space in finer increments than is feasible with manual job submission. Using this approach, we have found structure solutions in a number of cases where conventional 'crystallographer-directed' screening of program parameter space had failed.

The SECSG *SCA2Structure* pipeline described here was designed and implemented using the *BioPERL* pipeline platform (Stajich *et al.*, 2002) with the aim of producing a partially refined structure from a set of scaled single-wavelength anomalous scattering (SAS) data. The current version integrates *SOLVE*/*RESOLVE*, *ISAS* (Wang, 1985), *DM* (Cowtan & Zhang, 1999), part of the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994), *ARP*/*wARP* and *REFMAC* (Murshudov *et al.*, 1997), also part of *CCP*4, into a pipeline that is capable of spawning hundreds of jobs in parallel on a Linux cluster using various combinations of programs and/or input-parameter values. Our results have shown that the pipeline dramatically increases the efficiency and success rate of the structure-determination process. This pipeline approach not only increases the speed of determining a crystal structure, it also increases the likelihood of success owing to finer sampling of program parameter space.

*SCA2Structure* has been used to solve over 30 structures (Protein Data Bank; Berman *et al.*, 2000) with codes 1l7l, 1nnh, 1nnq, 1nnw, 1pry, 1ups, 1ryq, 1s36, 1sen, 1sgw, 1she, 1vjk, 1vk1, 1vka, 1vkc, 1xe1, 1xg9, 1xg7, 1xhc, 1xho, 1xi3, 1xi9, 1xk8, 1xkc, 1xma, 1xqu, 1xrg, 1xx7, 1y82, 1y81, 1yb3, 1ybx, 1yby, 1ybz, 1ycy, 1yd7). Of these, the following seven structures will be used to demonstrate the capabilities of the pipeline: 1l7l, 1nnq, 1sl8, 1nnh, 1vjk and 1ryq. Included in these examples are two cases where experienced crystallographers failed to solve the structure using 'rational' values for program input parameters. The total time necessary to complete and refine the structure ranged from 4 h to one week depending on the resolution of the data.

## 2. Materials and methods

### 2.1. Pipeline architecture

The pipeline is composed of three major components (Fig. 1): (i) a dictionary-driven web-based user interface, (ii) a *BioPERL*-based workflow-management system and (iii) a set of analytical tools for harvesting and visualizing data from the resultant log files. The web interface is used to authenticate users, upload experimental data and to input values for the various parameters (or parameter range) that will be used in the calculations. This networked client–server model has several advantages. Apart from the platform-independence of the client side, the crystallographic computing environment is administered centrally, relieving the user from tasks such as software installation and updates. It also allows the authentication procedure, project management, basic interface layout, session tracking and monitoring functions to be shared among different pipelines. Thus, once users become familiar with the usage of one pipeline, they can easily use other pipelines. The web form (Fig. 2) used to collect the input parameters required to set up a structure-determination run is generated by a dictionary-driven form generator that shares architectural features with the PDB structure-deposition tool *AutoDep* (Lin *et al.*, 2000) currently running at the European Bioinformatics

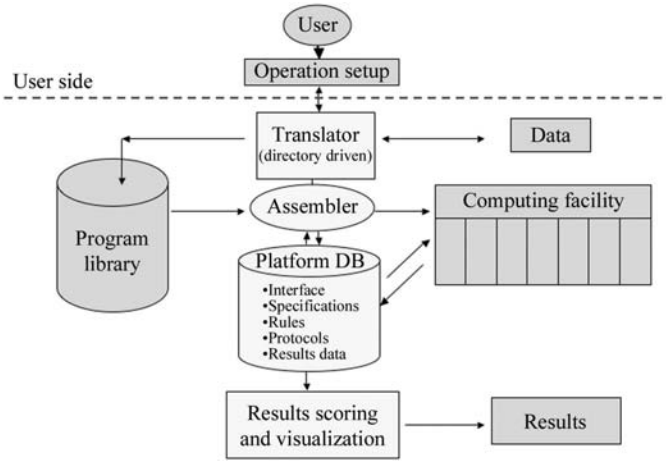
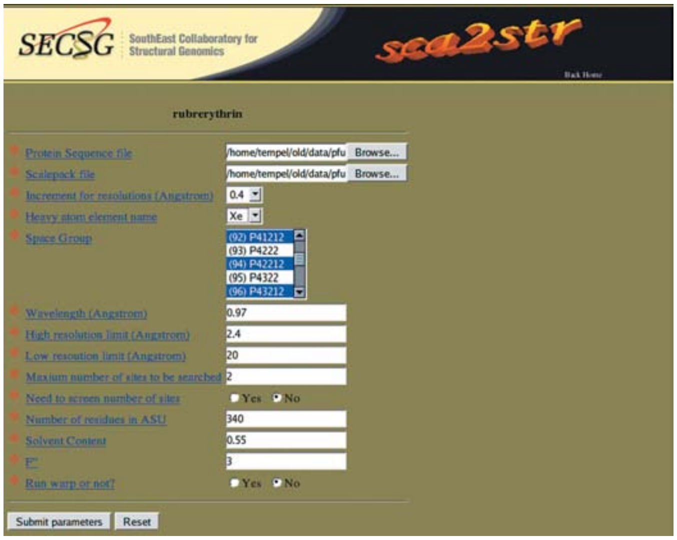


**Figure 1**
Diagram outlining the various components of the pipeline.

Institute PDB mirror (http://autodep.ebi.ac.uk). All information related to user input such as parameter name, parameter description, validation rules and HTML representation information are specified in a dictionary. With this dictionary-based approach, the programming of interfaces for new pipelines is greatly facilitated since it only requires the addition of the appropriate entries to the dictionary.

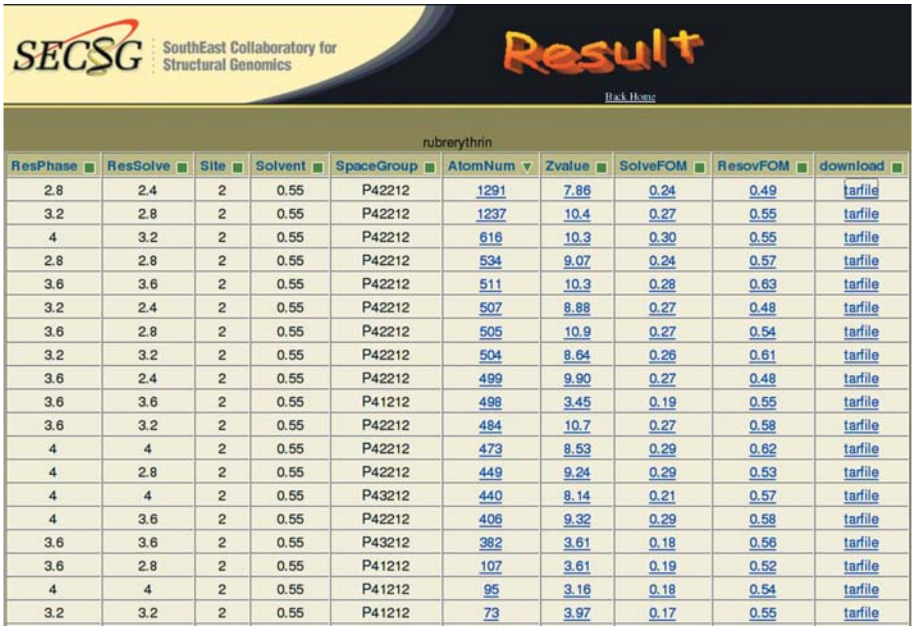


rubrerythrin

| | |
|---|---|
| Protein Sequence file | /home/tempel/old/data/pfu Browse... |
| Scalepack file | /home/tempel/old/data/pfu Browse... |
| Increment for resolutions (Angstrom) | 0.4 |
| Heavy atom element name | Xe |
| Space Group | (92) P41212<br>(93) P4222<br>(94) P42212<br>(95) P4322<br>(96) P43212 |
| Wavelength (Angstrom) | 0.97 |
| High resolution limit (Angstrom) | 2.4 |
| Low resoution limit (Angstrom) | 20 |
| Maxium number of sites to be searched | 2 |
| Need to screen number of sites | ○ Yes ● No |
| Number of residues in ASU | 340 |
| Solvent Content | 0.55 |
| F" | 3 |
| Run warp or not? | ○ Yes ● No |

Submit parameters Reset

**Figure 2**
A view of the *SCA2Structure* job-submission web page. The page is used to upload sequence and structure-factor files and to define input parameters and their ranges used to generate the structure determination jobs multiple space groups (highlighted in blue) can be selected by selecting the desired space groups while holding down the Ctrl key. Help files and examples for the various input parameters can be obtained by clicking on the appropriate web link (shown at left in blue).



rubrerythrin

| ResPhase | ResSolve | Site | Solvent | SpaceGroup | AtomNum | Zvalue | SolveFOM | ResovFOM | download |
|---|---|---|---|---|---|---|---|---|---|
| 2.8 | 2.4 | 2 | 0.55 | P42212 | 1291 | 7.86 | 0.24 | 0.49 | tarfile |
| 3.2 | 2.8 | 2 | 0.55 | P42212 | 1237 | 10.4 | 0.27 | 0.55 | tarfile |
| 4 | 3.2 | 2 | 0.55 | P42212 | 616 | 10.3 | 0.30 | 0.55 | tarfile |
| 2.8 | 2.8 | 2 | 0.55 | P42212 | 534 | 9.07 | 0.24 | 0.57 | tarfile |
| 3.6 | 3.6 | 2 | 0.55 | P42212 | 511 | 10.3 | 0.28 | 0.63 | tarfile |
| 3.2 | 2.4 | 2 | 0.55 | P42212 | 507 | 8.88 | 0.27 | 0.48 | tarfile |
| 3.6 | 2.8 | 2 | 0.55 | P42212 | 505 | 10.9 | 0.27 | 0.54 | tarfile |
| 3.2 | 3.2 | 2 | 0.55 | P42212 | 504 | 8.64 | 0.26 | 0.61 | tarfile |
| 3.6 | 2.4 | 2 | 0.55 | P42212 | 499 | 9.90 | 0.27 | 0.48 | tarfile |
| 3.6 | 3.6 | 2 | 0.55 | P41212 | 498 | 3.45 | 0.19 | 0.55 | tarfile |
| 3.6 | 3.2 | 2 | 0.55 | P42212 | 484 | 10.7 | 0.27 | 0.58 | tarfile |
| 4 | 4 | 2 | 0.55 | P42212 | 473 | 8.53 | 0.29 | 0.62 | tarfile |
| 4 | 2.8 | 2 | 0.55 | P42212 | 449 | 9.24 | 0.29 | 0.53 | tarfile |
| 4 | 4 | 2 | 0.55 | P43212 | 440 | 8.14 | 0.21 | 0.57 | tarfile |
| 4 | 3.6 | 2 | 0.55 | P42212 | 406 | 9.32 | 0.29 | 0.58 | tarfile |
| 3.6 | 3.6 | 2 | 0.55 | P43212 | 382 | 3.61 | 0.18 | 0.56 | tarfile |
| 3.6 | 2.8 | 2 | 0.55 | P41212 | 107 | 3.61 | 0.19 | 0.52 | tarfile |
| 4 | 4 | 2 | 0.55 | P41212 | 95 | 3.16 | 0.18 | 0.54 | tarfile |
| 3.2 | 3.2 | 2 | 0.55 | P41212 | 73 | 3.97 | 0.17 | 0.55 | tarfile |

**Figure 3**
A view of the *SCA2Structure* web report page. By default, the results are sorted based on the number of atoms fitted by *RESOLVE*. In this case, the top solution is most likely to be correct since it has the highest number of atoms fitted (AtomNum), the *RESOLVE* resolution (ResSolve) is high (for this data set) and it has a high *SOLVE Z* score (Zvalue). The solution is confirmed by downloading the files associated with the solution (tarfile) and manually inspecting the fit of the *RESOLVE* model to the electron-density map.

All information collected by the web interface is then transferred to the second layer of the pipeline platform, where workflow technology is used to manage the interaction of the different software tools. Based on this concept, the various crystallography software tools in the program library are converted to modules using the appropriate wrapper and a configuration file is used to specify how the various modules are connected and what input is to be used for each module. New pipelines are then assembled by the addition or rearrangement of these modules within the configuration file. The configuration file is processed by a *BioPERL*-based (http://www.bioperl.org) pipeline workflow manager that submits jobs to the cluster in the order specified by a given pipeline configuration file. The *BioPERL* workflow manager also ensures that computing resources are used as efficiently as possible. Details of the workflow-management system used by the pipeline will be published elsewhere.

Upon completion of the structure-determination run, a collection of analysis and visualization tools are used to harvest pertinent data from the numerous (typically between 500 to 1000) program log files generated by the structure-determination run. The tools parse out key data items relevant to the crystallographer, which are formatted into web-based tables (the *SCA2Structure* report web page) that can be easily sorted or filtered by the user (see Fig. 3). Currently, the following items are provided: (i) resolution values for phasing and phase extension/heavy-atom refinement, (ii) number of sites used in the search, (iii) solvent content used in the calculations, (iv) space group, (v) number of atoms traced by *RESOLVE*, (vi) *SOLVE Z* score, (vii) *SOLVE* figure of merit (FOM), (viii) *RESOLVE* FOM and (ix) a link to a tar file containing all output related to a given solution.

A relational database is used to archive the job process history, input and output data and pipeline input form dictionaries. Using this approach, a job can be rerun if necessary based on archived data. In addition, the relational database format facilitates the mining of archived data.

Based on the pipeline workflow platform described above, the *SCA2Structure* high-throughput crystal structure determination pipeline was constructed. In its original implementation and for the purpose of actual structure solutions discussed herein, the *SOLVE* and *RESOLVE* programs provide the core crystallographic functionality. Its capabilities have been extended over time with the integration of the programs *ISAS*, *DM*, *ARP/*

**Table 1**
Crystallization results.

| Protein | Crystallization condition | Derivatization |
|---|---|---|
| PA-L1 | Karaveg *et al.* (2003) | |
| Pfu-1210814 | Tempel, Liu, Schubot *et al.* (2005) | |
| Ca-aequorin | 100 m*M* sodium acetate pH 4.6, 30%(*v*/*v*) 2-methyl-2,4-pentanediol and 0.02 *M* calcium chloride, incubated at 277 K | |
| Pfu-1801964 | 100 m*M* sodium citrate pH 5.9, 10%(*w*/*v*) PEG 3000 and 500 m*M* magnesium chloride, incubated at 291 K | Addition of 0.2 µl of a 9:1 mixture of the precipitant solution and a 0.1 m*M* aqueous solution of potassium tetrachloroplatinate(II) to the crystallization drop 1 h prior to harvesting |
| Endo-$\beta$-Gal$_{\rm GnGa}$ | Deng *et al.* (2004) | Tempel, Liu, Horanyi *et al.* (2005) |
| Pfu-562899 | Native crystal: 100 m*M* sodium cacodylate pH 6.5, 30%(*w*/*v*) PEG 8000 and 200 m*M* ammonium sulfate, incubated at 291 K<br>Derivative crystal: 100 m*M* Tris pH 7.0, 2 *M* ammonium sulfate and 200 m*M* lithium sulfate, incubated at 291 K | Addition of small grain of potassium iodide to the drop and soaking for 4 h |
| Pfu-263306 | 100 m*M* sodium citrate pH 6.6 and 25%(*w*/*v*) PEG 3000, incubated at 291 K | Addition of small grain of potassium iodide to the drop and soaking for 4 h |

**Table 2**
Data-collection and data-processing results.

Values in parentheses were observed in the high-resolution shell.

| Protein | PA-L1 | Pfu-1210814 | Ca-aequorin | Pfu-1801964 | Endo-$\beta$-Gal$_{\rm GnGa}$ | Pfu-562899 | Pfu-263306 |
|---|---|---|---|---|---|---|---|
| Molecular weight (kDa) | 12.9 | 19.5 | 22.5 | 34.0 | 49.4 | 10.3 | 6.95 |
| X-ray source | Cu anode | SER-CAT† | Cr anode | Cu anode | SER-CAT† | SER-CAT† | Cu anode |
| Wavelength (Å) | 1.54 | 0.97 | 2.29 | 1.54 | 1.70 | 2.00 | 1.54 |
| Detector | R-AXIS IV | MAR CCD165 | R-AXIS IV | Smart 6000 | MAR CCD165 | MAR CCD225 | Smart 6000 |
| Exposure (s) | 300 | 20 | 420 | 60 | 5 | 3 | 60 |
| Oscillation range (°) | 360 × 0.5 | 2 × 200 × 0.5 | 202 × 1.0 | 2 × 350 × 0.3 | 6 × 160 × 0.5 | 360 × 1.0 | 2 × 400 × 0.3 |
| Distance (mm) | 150 | 170 | 126.2 | 90 | 110 | 80 | 60 |
| Space group | $I222$ | $P4_22_12$ | $P4_32_12$ | $I4_1$ | $P6_3$ | $P6_522$ | $P3_221$ |
| Unit-cell parameters | | | | | | | |
| $a$ (Å) | 40.25 | 105.92 | 54.34 | 68.36 | 159.53 | 81.35 | 45.66 |
| $b$ (Å) | 72.30 | 105.92 | 54.34 | 68.36 | 159.53 | 81.35 | 45.66 |
| $c$ (Å) | 133.82 | 81.00 | 135.06 | 151.64 | 85.85 | 63.55 | 50.78 |
| High-resolution shell (Å) | 1.86–1.80 | 2.43–2.35 | 2.59–2.50 | 2.29–2.10 | 2.78–2.68 | 2.38–2.30 | 1.96–1.80 |
| Completeness (%) | 93.7 (57.9) | 99.9 (100.0) | 99.5 (95.4) | 83.3 (44.9) | 98.0 (80.0) | 99.9 (100.0) | 87.5 (50.5) |
| $R_{\rm sym}$ (%) | 3.5 (9.1) | 7.1 (36.9) | 7.0 (12.9) | 3.8 (9.3) | 6.9 (33.3) | 8.9 (12.1) | 3.6 (6.7) |
| $I/\sigma(I)$ | 44.8 (11.1) | 36.7 (6.3) | 34.1 (11.0) | 14.5 (5.0) | 55.7 (7.2) | 78.9 (49.7) | 16.4 (4.9) |

† SER-CAT: Southeast Regional Collaborative Access Team, Sector 22, Advanced Photon Source, Argonne National Laboratory.

*wARP* and *REFMAC*. Based on its core components, the pipeline only requires the scaled reflection data (*SCALE-SCALEPACK* or MTZ format), the polypeptide sequence and the expected solvent content of the crystal to produce a partial model of the peptide under investigation. With the addition of the *ARP*/*wARP* module, the pipeline has produced, in the case of PA-L1, a nearly complete refined model of the protein. The *SCA2Structure* pipeline user interface provides reasonable default parameters (or screening range) including step size for inexperienced users. It has been our experience that the default parameters work very well in most cases. In its current implementation, *SCA2Structure* permits screening of the following.

(i) The number of expected heavy-atom (or anomalous scatterer) sites (*SOLVE*).

(ii) Space groups.

(iii) High-resolution data cutoff for the heavy-atom search (*SOLVE*).

(iv) High-resolution data cutoff for initial phasing (*SOLVE*/*ISAS*).

(v) High-resolution data cutoff for phase improvement/extension (*RESOLVE*).

(vi) Phasing programs (*SOLVE*/*ISAS*).

### 2.2. Protein samples

The protein samples used in the analyses were expressed and purified according to published procedures. The *Pseudomonas aeruginosa* lectin-1 (PA-L1) sample was prepared according to the procedure of Karaveg *et al.* (2003). The *P. furiosus* samples (Pfu-263306, Pfu-562899, Pfu-1210814 and Pfu-1801964) were prepared by the SECSG *P. furiosus* Protein Production Core following a general procedure (Adams *et al.*, 2003) and using the genes encoding the respective proteins (Robb *et al.*, 2001). The *Aequorea victoria* aequorin (Ca-aequorin) sample was prepared according to Deng *et al.* (manuscript in preparation). The *Clostridium perfringens* GlcNAc$\alpha$1-4Gal-releasing endo-$\beta$-galactosidase (Endo-$\beta$-Gal) sample was prepared according to the method of Ashida *et al.* (2002).

**Table 3**
Parameters screened by *SCA2Structure* pipeline.

| Protein | PA-L1 | Pfu-1210814 | Ca-aequorin | Pfu-1801964 | Endo-$\beta$-Gal$_{\rm GnGa}$ | Pfu-562899 | Pfu-263306 |
|---|---|---|---|---|---|---|---|
| Resolution range screened | 3.4–2.0 | 4.0–2.4 | 3.8–2.6 | 3.8–2.2 | 3.8–2.9 | 4–2.4 | 3.6–2.0 |
| Increment (Å) | 0.2 | 0.4 | 0.3 | 0.2 | 0.3 | 0.2 | 0.2 |
| Optimal resolution for initial phasing (Å) | 2.0 | 2.8 | 2.6 | 2.4 | 3.2 | 2.6 | 2.0 |
| Optimal resolution for phase extension (Å) | 1.8 | 2.4 | 2.5 | 2.2 | 2.7 | 2.5 | 2.0 |
| No. of heavy-atoms found/sought | 4/4 | 2/2 | 9/10 | 2/2 | 17/20 | 4/4 | 1/4 |
| Solvent content used in phase extension | 0.65 | 0.55 | 0.46 | 0.52 | 0.44/0.62 | 0.55 | 0.43 |
| Space groups screened | $I222$ | $P4_22_12$, $P4_12_12$, $P4_32_12$ | $P4_32_12$, $P4_12_12$, $P4_22_12$ | $I4_1$ | $P6_3$ | $P6_522$, $P6_122$ | $P3_221$, $P3_121$ |
| Total No. of jobs | 55 | 180 | 180 | 72 | 80 | 56 | 90 |
| Total time | 1 h | 5 h | 2 h | 2 h | 5 h | 33 min | 32 min |

### 2.3. Crystallization

With the exception of PA-L1, all crystals were obtained by the microbatch-under-oil method (D'Arcy *et al.*, 2003) using a modified Douglas Instruments ORYX 6 robot (Shah *et al.*, 2005) and 72-well Nunc plates. The crystallization drops contain 0.5 µl protein solution mixed with 0.5 µl precipitate solution. The drops were covered with a 7:3(*v*:*v*) layer of paraffin and silicon oils. The crystallization experiments are summarized in Table 1.

### 2.4. X-ray diffraction and data reduction

For data collection, the crystals were mounted in nylon loops (Teng, 1990), flash-cooled in liquid nitrogen (Hope, 1988), mounted on the goniometer and maintained at 100 K in a nitrogen-gas cryostream. The data collection and processing was optimized for single-wavelength anomalous scattering phasing. Details of the data collection for the various samples are given in Table 2. Data were indexed, integrated and scaled using the *HKL* (*DENZO*/*SCALEPACK*) suite (Otwinowski & Minor, 1997) in all cases with the exception of Pfu-263306 and Pfu-1801964, where the *PROTEUM* package (Bruker AXS) was used for data processing (see Table 2).

### 2.5. Computing hardware and software

Calculations were carried out on a 64-node cluster of two-way servers (International Business Machines) based on the x86 architecture. Resource management and job scheduling were handled by a combination of the *OpenPBS* (http://www.openpbs.org) and *MAUI* (http://www.supercluster.org/maui) packages. Job preparation and tracking was based on the *BioPERL* (http://www.bioperl.org) suite. Web content for job submission and result retrieval was served by the Apache (http://httpd.apache.org) HTTP server.

### 2.6. Structure solution, phase improvement, chain tracing and refinement

The anomalous substructure and initial phases were determined using *SOLVE* (Terwilliger & Berendzen, 1999) in SAS mode. Phase refinement was carried out using *RESOLVE* (Terwilliger, 1999). Resolution cutoffs for initial phasing and phase extension were screened within the limits shown in Table 3. As shown, in some instances calculations were performed for several candidate space groups and oligomeric states. For all structures described here, calculations involving *SOLVE* and *RESOLVE* were performed on the high-throughput pipeline platform. In the case of PA-L1, the pipeline was extended to also run *ARP*/*wARP*. With the exception of PA-L1, all structures were manually refitted and refined prior to Protein Data Bank (PDB; Berman *et al.*, 2000; Bernstein *et al.*, 1977) deposition.

## 3. Results and discussion

### 3.1. Examples

Table 3 lists the parameters screened by the *SCA2Structure* pipeline for the seven examples given below, including the number of jobs spawned by the pipeline and the amount of time it took to produce a structure using the pipeline.

**3.1.1. PA-L1 (a galactophilic lectin from *Pseudomonas aeruginosa*)**. The protein contains a calcium ion and three ordered sulfur-containing amino-acid residues (two cysteinyl residues and one ordered methioninyl residue). Initial attempts to solve the structure using SAS data collected in-house and analysed with *SOLVE* did not produce an interpretable electron-density map. The first model of this protein (PDB code 1l7l) was instead based on synchrotron data. The in-house data set was revisited during the initial tests of the *SCA2Structure* pipeline. The pipeline was able to solve the PA-L1 structure using in-house data giving a complete (98%) *ARP*/*wARP* trace. However, as one would expect, not all parameter combinations generated by the pipeline led to a successful structure determination, as illustrated in Fig. 4. One surprising outcome of this study was that success using two values for a resolution cutoff did not guarantee success with an intermediate value. This is illustrated in Fig. 4, which shows that when a high-resolution cutoff of 2.0 Å was used for both *SOLVE* and *RESOLVE*, the pipeline failed to produce a structure. However, when the *RESOLVE* resolution cutoff is either 1.8 or 2.2 Å a solution is obtained. Analysis of the anomalous scattering substructures for these three cases reveals that for the unsuccessful case *SOLVE* produced the enantiomer of the correct anomalous substructure, resulting in an uninterpretable electron-density map. This finding is in accordance with a suggestion by the author of *SOLVE* (T. Terwilliger, personal communication). Phase extension and model building were automatically performed with the program *ARP*/*wARP* within the pipeline platform.

**3.1.2. Pfu-1210814 (rubrerythrin)**. This was the first 'unknown' pipeline SAS test structure. *P. furiosus* rubrerythrin, similar to its known homologues, contains iron-binding motifs and an experiment was designed to exploit the iron anomalous scattering signal by recording phasing data using 1.74 Å X-rays. A second set of data was recorded to higher resolution using 0.97 Å X-rays for refinement purposes. Both data sets were processed keeping Bijvoet-related reflections separate. When the pipeline failed to produce a structure using the phasing data, the high-resolution data set was subjected to the same analysis. To our surprise, the pipeline produced a structure (80% complete *RESOLVE*) from this data set. The *RESOLVE* phases and initial trace were then used to manually complete the model with *XFIT* (McRee, 1999). The final refinement was carried out with *REFMAC* within *CCP*4 and the MOLPROBITY web service (Lovell *et al.*, 2003). The coordinates have been deposited in the PDB (entry 1nnq). The structure revealed that zinc has replaced iron in the iron-binding site, which explains why the phasing data failed to produce a structure. By chance, the high-resolution data were collected using a wavelength where the zinc anomalous scattering signal, although not optimal, was sufficient to solve the structure. In addition, since the space group could only be unambiguously assigned once the structure had been solved, the pipeline setup included the screening of several candidate space groups. The fully refined model has been described elsewhere (Tempel, Liu, Schubot *et al.*, 2005).

**3.1.3. Ca-aequorin (a calcium-sensitive photoprotein from *Aequorea aequorea*)**. A single data set was collected on a crystal of calcium-loaded Ca-aequorin using Cr *Kα* X-rays (λ = 2.909). The pipeline produced a structure based on three calcium ions and eight S atoms determined by *SOLVE*. Again, three space groups were analyzed by the pipeline during the structure determination. The *RESOLVE* phases and initial trace were then used to manually complete the model with *XFIT*. The final refinement was carried out with *REFMAC* in *CCP*4. The coordinates have been deposited in the PDB (entry 1sl8).

**3.1.4. Pfu-1801964 (a putative asparaginyl-tRNA synthetase from *P. furiosus*)**. The structure was solved by the pipeline from a single set of SAS data collected on a $K_2PtCl_4$ derivative of the protein using Cu *Kα* X-rays. The *RESOLVE* peptide trace produced by the pipeline was manually completed using *XFIT*. The resulting model was then used to generate molecular-replacement (*EPMR*; Kissinger *et al.*, 1999) phases for a higher resolution data set followed by automated rebuilding (*ARP*/*wARP*). The final refinement was carried out with *REFMAC* in *CCP*4. The coordinates have been deposited in the PDB (entry 1nnh).

**3.1.5. Endo-*β*-Gal (GlcNAc*α*1-4Gal-releasing endo-*β*-galactosidase from *Clostridium perfringens*)**. The structure was solved using data collected on an iodide derivative (iodide quick soak; Dauter *et al.*, 2000) of the protein recorded using 1.74 Å X-rays. The phases from *RESOLVE* and the *RESOLVE* trace were used to manually complete the model with *XFIT*. The final refinement was carried with *REFMAC* in *CCP*4 and the MOLPROBITY web service (Lovell *et al.*, 2003). The coordinates have been deposited in the PDB (entry 1ups). The details for this model have been published elsewhere (Tempel, Liu, Horanyi *et al.*, 2005).

**3.1.6. Pfu-562899 (molybdopterin-converting factor subunit 1 from *P. furiosus*)**. The structure was solved using the pipeline from data collected from a halide derivative (Dauter *et al.*, 2000) crystal. Phases from *RESOLVE* were used for automated model building in *ARP*/*wARP*. The model was refined [*REFMAC* and the MOLPROBITY web service (Lovell *et al.*, 2003)] against an isomorphous higher resolution data set. The coordinates have been deposited in the PDB (entry 1vjk).

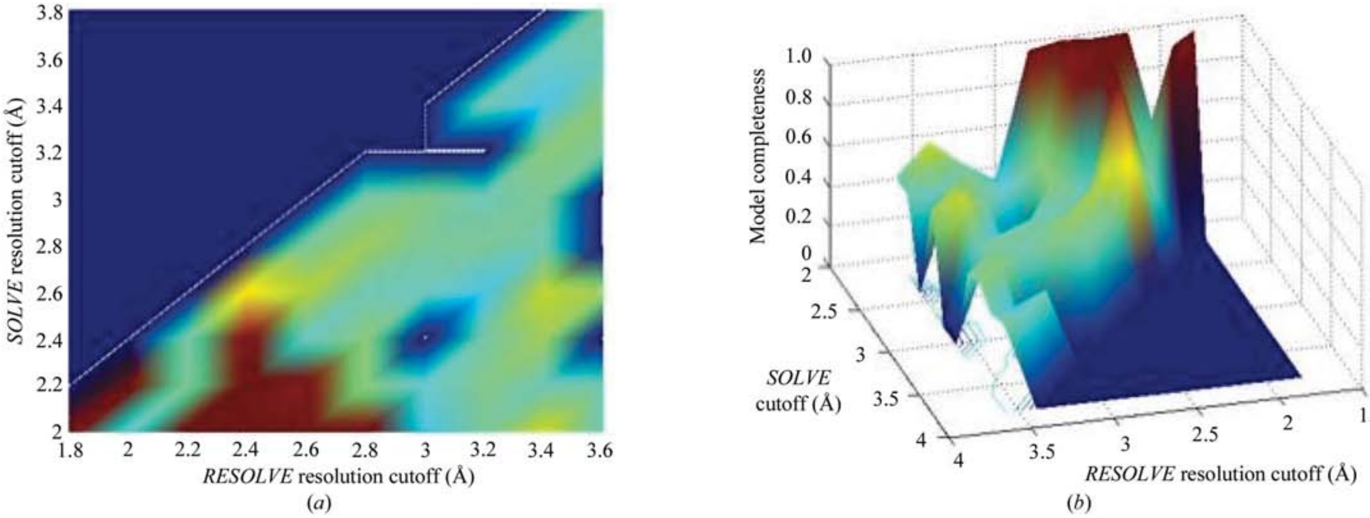


**Figure 4**
A graphical representation of pipeline success space for the PA-L1 example. A total of 55 *SOLVE*/*RESOLVE* phase sets were used as input to *ARP*/*wARP*. The colour scheme used represents success (number of residues fitted), with red indicating a near-complete model and blue/cyan representing cases where model building failed. An interesting and unexpected feature is that success space is not continuous with regions of low success sandwiched between regions of high success.

#### 3.1.7. Pfu-263306 (a putative DNA-directed RNA polymerase subunit $\varepsilon''$).

In this case, the enantiomorphic space-group ambiguity ($P3_121$ or $P3_221$) had to be resolved as part of the structure-determination process. The structure was solved using data collected from an iodide derivative. Phases from *RESOLVE* were used for automated model building in *ARP/wARP*. The model was refined [*REFMAC* and the MOLPROBITY web service (Lovell *et al.*, 2003)] against a set of isomorphous higher resolution data. The resulting model revealed a zinc–sulfur site involving four cysteinyl residues. Interestingly, electron density also defined some residues of the N-terminal histidine purification tag. The coordinates have been deposited in the PDB (entry 1ryq).

### 3.2. Interpreting results

Because of the distributed nature of the pipeline, hundreds of log files can be generated in a typical structure-determination run. Analyzing this vast amount of data manually would be a formidable task, so a set of analytical tools for extracting and visualizing the results in an organized manner *via* the pipeline Result web page (see Fig. 3) has been developed. In our experience, sorting the data based on the number of residues fitted by *RESOLVE* gives the best indication of a successful structure determination. Generally, a solution will have the greatest number of atoms fit by *RESOLVE*, providing that the resolution used for the phasing (*SOLVE*) and phase-extension (*RESOLVE*) calculations are high. Additionally, a correct solution will have a high *SOLVE Z* score and *SOLVE* FOM. Using the above criteria, tar files for the potential solutions are downloaded from the pipeline Result page to the client computer where the experimental electron-density maps can be inspected manually to confirm the solution.

Experience has shown that screening for the number of heavy-atom sites to be found did not produce better results than when a single slightly overestimated value for this parameter was used in the calculations since *SOLVE* automatically rejects doubtful heavy-atom sites. However, as noted above in the case of PA-L1, a variation of the resolution cutoffs produced interesting and non-intuitive results (Fig. 4). In this case, the resolution ranges that resulted in successful automated model building were not continuous and imply that structure-determination failure may be the result of the inopportune choice of resolution cutoffs for *SOLVE* and *RESOLVE* even when the data were capable of providing a solution using different resolution cutoffs. The addition of higher resolution data into calculations also does not always guarantee success. This is because although the observation-to-parameter ratio is increased owing to the added data, the anomalous differences, which are increasingly weak at high resolution, decrease the signal-to-noise ratio in the data.

As indicated in the above examples, the *ARP/wARP* module of the pipeline is typically not used in the structure-determination process but is run independently after a successful solution is found. This is because SECSG crystallographers usually separate data collection for phasing purposes from data collection to be used for high-resolution refinement. This approach allows for the optimization of the anomalous signal in the phasing data and for the optimization the intensities of the weak reflections at high resolution in the refinement data. In addition, running *ARP/wARP* in every case would waste considerable CPU time since only a few of the pipeline jobs submitted will produce useful phases. Instead, the results from the *SOLVE/RESOLVE* runs are first analyzed and if a successful solution is found and the resolution of the data permits *ARP/wARP* is run (usually on a higher resolution data set or a data set collected at a higher energy which should have lower absorption effects).

## 4. Conclusions

The *SCA2structure* pipeline has become the primary method of *de novo* structure solution at SECSG and has been instrumental in the determination of over 40 structures. The simple job-submission web page coupled with a fine sampling of program parameter space can help to answer several of the questions posed in §1. These include (i) are the data of sufficient quality to permit solution of the structure? (ii) Are more data needed? and (iii) what are the optimal values for the input parameters for the programs used in the structure-determination process? SECSG crystallographers have used the pipeline for on-site structure determination at the beamline to answer these questions. Typically, once data are collected and processed at the beamline, a structure-determination run is submitted to the UGA cluster *via* the pipeline job-submission web page. The crystallographer is then free to begin data collection on the next target on the list. Once the structure-determination jobs have finished (usually between 1 and 2 h), the results can be quickly analyzed (using the pipeline Web Report page) to determine whether more data are needed to solve the structure. This approach has proven to be quite efficient and in one recent case five structures were solved in a 23 h period by SECSG crystallographers on-site at SER-CAT (supporting data to be published elsewhere).

The success of the pipeline is based on several factors: (i) by parameter-space screening, the *SCA2Structure* pipeline dramatically increases the structure-solution success rate. This in turn decreases the number of trials required and thus reduces the time needed for structure determination. (ii) The web-based user interface allows easy job submission and result retrieval since it can be accessed from any location including the synchrotron beamline. (iii) Its ease of use and SECSG's 128-processor cluster make the pipeline an almost real-time tool for the analysis of data quality (with the capability to produce an SAS structure) and structure production. (iv) The dictionary-driven design and facile extensibility of the platform permit easy adoption of new pipeline modules and/or alternative computational protocols while maintaining a consistent user-interface layout.

The power of the pipeline comes from the parameter-space screening. This innovation overcomes the peculiarities associated with a given data set arising from crystal quality, experimental errors and other factors that make each data set a unique case; optimal values can be quickly found that are best suited for a given data set. Since the user only needs to supply the parameter range and sampling step to be used, the process becomes very efficient. The current pipeline runs on the SECSG 128-processor Linux cluster. However, the *BioPerl* job-management system allows easy configuration to any system including a single processor. The efficiency of the process is however dramatically reduced as the number of processors decreases.

The *SCA2Structure* pipeline provides a powerful tool for SAS phasing problems. It has fundamentally changed the way in which structure determination is carried out at SECSG. At the same time, we are continuously working on making the pipeline more versatile and intelligent; for example, future versions of the pipeline will be able to handle MAD, SIRAS and MIR phasing automatically. We are also in the process of mining pipeline results from different structure-determination calculations at SECSG in an attempt to find a general set of indicators which lead to the best solution. The discovery of trends will not only influence the development of future pipelines but also general aspects of crystallographic structure determination. For example, pipelines such as the SAS pipeline discussed here provide a convenient quality-assessment tool for diffraction data. Because the pipeline produces structure solutions for some data sets while it fails to find answers with others, it is possible to identify characteristics that may act as predictors for success or failure. Because some of these characteristics are directly affected by experimental procedures, the identification of decisive factors will allow rational adjustment of data-collection design and/or parameters to increase the probability of success.

Work was supported in part by funds from the National Institutes of Health (GM62407), The Georgia Research Alliance and The University of Georgia Research Foundation. Use of the Advanced Photon Source was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. W-31-109-ENG-38. Data were collected at Southeast Regional Collaborative Access Team (SER-CAT) 22-ID beamline at the Advanced Photon Source, Argonne National Laboratory. Supporting institutions may be found at http://www.ser-cat.org/members.html. Use of the IMCA-CAT beamline 17-ID at the Advanced Photon Source was supported by the companies of the Industrial Macromolecular Crystallography Association through a contract with Illinois Institute of Technology. International Business Machines is acknowledged for a Shared University Research Award.

## References

Adams, M. W., Dailey, H. A., DeLucas, L. J., Luo, M., Prestegard, J. H., Rose, J. P. & Wang, B.-C. (2003). *Acc. Chem. Res.* **36**, 191–198.

Ashida, H., Maskos, K., Li, S. C. & Li, Y. T. (2002). *Biochemistry*, **41**, 2388–2395.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissing, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **26**, 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rogers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

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

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

Cowtan, K. D. & Zhang, K. Y. J. (1999). *Prog. Biophys. Mol. Biol.* **72**, 245–270.

D'Arcy, A., Mac Sweeney, A., Stihle, M. & Haber, A. (2003). *Acta Cryst.* D**59**, 396–399.

Dauter, Z., Dauter, M. & Rajashankar, K. R. (2000). *Acta Cryst.* D**56**, 232–237.

Deng, L., Liu, Z. J., Ashida, H., Li, S. C., Li, Y. T., Horanyi, P., Tempel, W., Rose, J. & Wang, B.-C. (2004). *Acta Cryst.* D**60**, 537–538.

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

Hope, H. (1988). *Acta Cryst.* B**44**, 22–26.

Karaveg, K., Liu, Z. J., Tempel, W., Doyle, R. J., Rose, J. P. & Wang, B.-C. (2003). *Acta Cryst.* D**59**, 1241–1242.

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

Lin, D., Manning, N. O., Jiang, J., Abola, E. E., Stampf, D., Prilusky, J. & Sussman, J. L. (2000). *Acta Cryst.* D**56**, 828–841.

Lovell, S. C., Davis, I. W., Adrendall, W. B. III, de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). *Proteins*, **50**, 437–450.

McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Norvell, J. C. & Machalek, A. Z. (2000). *Nature Struct. Biol.* **7**, Suppl. 931.

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

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

Praissman, J., Lin, D., Liu, Z. J., Tempel, W., Rose, J. & Wang, B.-C. (2003). *Am. Cryst. Assoc. Abstr. Papers (Annu. Meet.)*, **30**, 43.

Robb, F. T., Maeder, D. L., Brown, J. R., DiRuggiero, J., Stump, M. D., Yeh, R. K., Weiss, R. B. & Dunn, D. M. (2001). *Methods Enzymol.* **330**, 134–157.

Shah, A. K., Liu. Z. J., Stewart, P. D., Schubot, F. D., Rose, J. P., Newton, M. G. & Wang, B.-C. (2005). *Acta Cryst.* D**61**, 123–129.

Stajich, J. E. *et al.* (2002). *Genome Res.* **12**, 1611–1618.

Tempel, W., Liu, Z. J., Horanyi, P. S., Deng, L., Lee, D., Newton, M. G., Rose, J. P., Ashida, H., Li, S. C., Li, Y. T. & Wang, B.-C. (2005). In the press.

Tempel, W., Liu, Z. J., Schubot, F. D., Shah, A., Weinberg, M. V., Jenney, F. E. Jr, Arendall, W. B. 3rd, Adams, M. W., Richardson, J. S., Richardson, D. C., Rose, J. P. & Wang, B.-C. (2005). *Proteins*, **57**, 878–882.

Teng, T.-Y. (1990). *J. Appl. Cryst.* **23**, 387–391.

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

Terwilliger, T. C. (2002). *Acta Cryst.* D**58**, 1937–1940.

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

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