Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Effective electron-density map improvement and structure validation on a Linux multi-CPU web cluster: The TB Structural Genomics Consortium Bias Removal Web Service

Vinod Reddy,[a] Stanley M. Swanson,[a] Brent Segelke,[b] Katherine A. Kantardjieff,[c] James C. Sacchettini[a] and Bernhard Rupp[a,b]*

[a]Biochemistry and Biophysics Department, Texas A&M University, 2128 TAMU, College Station, TX 77843-2128, USA, [b]Macromolecular Crystallography and Structural Genomics Group, Lawrence Livermore National Laboratory, University of California, Livermore, CA 94551, USA, and [c]W. M. Keck Foundation Center for Molecular Structure, Department of Chemistry and Biochemistry, California State University Fullerton, 800 North State College Boulevard, Fullerton, CA 92834-6866, USA

Correspondence e-mail: br@llnl.gov

Received 18 July 2003
Accepted 12 September 2003

Anticipating a continuing increase in the number of structures solved by molecular replacement in high-throughput crystallography and drug-discovery programs, a user-friendly web service for automated molecular replacement, map improvement, bias removal and real-space correlation structure validation has been implemented. The service is based on an efficient bias-removal protocol, *Shake&wARP*, and implemented using *EPMR* and the *CCP*4 suite of programs, combined with various shell scripts and Fortran90 routines. The service returns improved maps, converted data files and real-space correlation and *B*-factor plots. User data are uploaded through a web interface and the CPU-intensive iteration cycles are executed on a low-cost Linux multi-CPU cluster using the Condor job-queuing package. Examples of map improvement at various resolutions are provided and include model completion and reconstruction of absent parts, sequence correction, and ligand validation in drug-target structures.

1. Introduction

The number of structures obtained by molecular replacement (MR) is expected to grow rapidly in coming years both in academic laboratories and in pharmaceutical structure-based drug-discovery efforts (Blundell *et al.*, 2001). The increasing accessibility of powerful molecular-replacement programs and the increasing availability of search models owing to the discovery of novel folds by public and commercial structural genomics efforts (Norvell & Zapp-Machalek, 2000) are major contributing factors. An estimate of structures solved in a commercial structural genomics effort indicates that about 70% of all structures processed were solved by MR, and in drug-discovery efforts the numbers may be even higher (Kissinger *et al.*, 2001).

Anticipating a corresponding need for map improvement and electron-density-based structure validation, we present an effective easy-to-use web service for map improvement, phase-bias removal and rapid assessment of local model quality, which complements geometry-based structure-validation programs such as *WHAT-IF* (Vriend, 1990) and *PROCHECK* (Laskowski *et al.*, 1993). The protocol, *Shake&wARP*, achieves effective bias removal using a modified *wARP* (Perrakis *et al.*, 1997) procedure and is implemented using the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994) of programs, various shell scripts and Fortran90 applets executable in parallel mode on multiple CPUs. *Shake&wARP* differs in details and choice of parameters from the routines distributed with *wARP*, most notably

© 2003 International Union of Crystallography
Printed in Denmark – all rights reserved

in model perturbation, dummy-atom placement/removal criteria and map averaging. Given a modest model, *Shake&wARP* works efficiently at resolutions as low as 2.6–2.8 Å and yields improved map quality in direct comparison with other bias-reduced map-reconstruction methods. The web service allows fully automated MR solution using *EPMR* (Kissinger *et al.*, 1999), followed by *Shake&wARP* map averaging. Residue-by-residue real-space correlation-coefficient and *B*-factor plots are automatically created in GIF format. Inspection of real-space correlation-coefficient plots provides a quick assessment of local structure quality, ensuring that no details have been overlooked in important areas, while less time is wasted on over-refinement in areas of little interest and which may serve only to create artificially low global quality descriptors.

We present details of the protocol and implementation of a web service, which extends our low-cost approach to high-throughput protein crystallography (Rupp, 2003) to parallel execution and job queuing using Intel/AMD-based hardware and Linux/Condor[1] as a portable operating system platform. A number of general examples demonstrate commonly observed phase-bias and map-interpretation problems and illustrate the importance of effective bias removal and electron-density map improvement.

2. Model phase bias and map improvement

Model phase bias is a major concern in any crystallographic structure determination, in particular when the experimental phases are suboptimal or significantly biased, as in maps derived from MR phases. The effects of insidious model bias can be dramatic and are not easily recognized by commonly used global structure-quality descriptors such as R and $R_{\rm free}$ (for a review, see Kleywegt & Jones, 1997). In severe cases, model bias can introduce artifacts that seriously limit the usefulness of a structure and questionable conclusions affecting the biological significance of results may be drawn (Rupp & Segelke, 2001; Hanson & Stevens, 2000; Hanson *et al.*, 2002).

2.1. Source of model phase bias

Model (or phase) bias results from the fact that reflection phase angles (α_{hkl}), which are required to complete the Fourier transformation of structure amplitudes $|F|_{hkl}$ back to electron density $\rho_{(xyz)}$ (1) are not directly observable quantities. They must be provided by additional phasing experiments: for example, multiple isomorphous replacement methods (Blundell & Johnson, 1976; Islam *et al.*, 1998) or anomalous phasing, such as single-wavelength (Matthews, 2001) or multi-wavelength (Hendrickson & Ogata, 1997) anomalous diffraction. Phases may also originate from initial MR models, where they tend to be marginal and highly biased (Adams *et al.*, 1999).

$$\rho_{(xyz)} = \frac{1}{V}\sum_{-h}^{h}\sum_{-k}^{k}\sum_{-l}^{l}|F|_{hkl}\exp[-2\pi i(hx+ky+lz-\alpha_{hkl})]. \quad (1)$$

2.2. Model-bias reduction

The need for countermeasures against model bias has long been recognized and a variety of bias-reduction methods are presently used and implemented in modern program packages (for a comprehensive discussion of model bias see, for example, Read, 1997). A number of general strategies or combinations thereof are commonly employed to combat model bias: (i) omission of parts of the model, (ii) perturbation ('shaking') of the model coordinates between refinement/rebuilding cycles, (iii) allowance for model errors in the refinement target functions and map coefficients, (iv) repeated cycling of real-space and reciprocal-space refinement (real-space fitting of the model into electron density *versus* refining model coordinates against reciprocal-space structure factors) and (v) map-averaging techniques. During refinement, the implementation of maximum-likelihood (ML) targets, as implemented in *REFMAC* (Murshudov *et al.*, 1997) or *CNS* (Brünger *et al.*, 1998), together with σ_A-weighted map coefficients (Read, 1986) of general form $2m|F_o| - D|F_c|\exp(i\alpha_c)$ accounting for partial or incorrect model, produces maps with significantly reduced model bias. When these strategies are used together with strict $R_{\rm free}$ cross-validation (Brünger, 1992), relatively 'safe' crystallography should be possible. Nevertheless, in many cases a weak part of the structure, a ligand or cofactor may need additional individual confirmation to ensure that its density is not a result of remaining model bias. In these cases, omission of the questionable model part in the phase calculation and perturbation of the remainder of the structure to eliminate 'memory' followed by ML refinement will yield a map of reduced bias. Electron density will either confirm a disordered residue or perhaps obliterate the hope of the presence of a ligand. 'Classical' omit maps (Bhat & Cohen, 1994; Bhat, 1988), σ_A omit maps (Read, 1986, 1990), simulated-annealing omit maps (Hodel *et al.*, 1992) and shake omit maps (Zeng *et al.*, 1997) are commonly used for this purpose. An ML-based reciprocal-space density-modification method (*Prime&Switch*), which can be applied to initial experimental maps or model-phased maps and appears to perform well at low resolution and with marginal models, has been recently implemented in *RESOLVE* (Terwilliger, 1999, 2000).

2.3. Map improvement and bias reduction with *Shake&wARP*

The basic idea behind *Shake&wARP* (*S&W*) is to combine most of the available means for bias reduction into one single protocol. The strategies implemented include omitting parts (random atoms and/or specific parts of the model), perturbation ('shaking') of coordinates, use of ML refinement targets in *REFMAC* (Murshudov *et al.*, 1997), iterating in multiple cycles with real-space dummy-atom placement using the

[1] The Condor Software Program (Condor) was developed by the Condor Team at the Computer Sciences Department of the University of Wisconsin-Madison. All rights, title, and interest in Condor are owned by the Condor Team.

*CCP*4 program *ARP_WATERS* (Lamzin & Wilson, 1993) and finally, probably the most effective contribution to *S&W*, averaging of six maps resulting from differently perturbed starting models (Perrakis *et al.*, 1997).

The original *wARP* procedure by default places atoms into high density peaks (3.5σ) and removes atoms below 1.5σ density, although the *ARP* (now *ARP_WATERS*) program (Lamzin & Wilson, 1993) will add additional atoms below this level if it has not found the number of atoms set in keyword FIND. By comparision, *Shake&wARP* (i) builds into much lower density (1.0σ), (ii) removes atoms below 0.6σ, (iii) begins from six differently and optimally perturbed starting models (detailed in the caption to Fig. 1), where 10% of atoms have been randomly removed, and (iv) perturbs the remaining coordinates by an average of 0.25 Å r.m.s.d. (Fig. 1). Starting with different models and building into relatively low solvent density compared with the original *wARP* procedure followed by weighted map averaging can effectively be viewed as a real-space solvent-flattening procedure which significantly increases map contrast. The density features contained in each map are amplified and the noise density represented by variably placed atoms will effectively be averaged out. The power of map averaging for phase improvement has been well established in cross-crystal form and NCS-averaging techniques (Kleywegt & Read, 1997) and is one of the major reasons for increased clarity and contrast in *S&W* maps compared with those reconstructed by other techniques (Fig. 5 provides a strong example of the drastic improvement obtained by averaging).

The automated model-building program *ARP*/*wARP* (Perrakis *et al.*, 1999), also known as 'warpNtrace', which can build protein models into empty experimentally phased maps and can also improve MR models, does not currently employ map averaging.

2.4. Convergence and model perturbation in dummy-atom refinement

As the only initial information available to *Shake&wARP* originates from the input model phases and the diffraction data, a balance must be maintained between data quality and staring phase quality, beyond which the *ARP*/*REFMAC* cycles will fail to converge or to provide phase improvement. An estimate for the minimum resolution (better than 2.4 Å) for the applicability of *ARP* has been provided (Perrakis *et al.*, 1997) and it was noted that the higher the resolution, the better the method will work (subject to other omni-valid criteria such as data quality and completeness). The need for convergence also affects the amount and mode of permissible coordinate perturbation. To investigate both the effect of various model-perturbation methods as well as to provide an estimate for convergence of *S&W*, we calculated deviations between initial and perturbed models against the final structure, representing the error as the *R* value and phase error in variation with resolution (Fig. 1). As expected, the higher the resolution of the available data, the more robust the protocol becomes when starting from weak initial models; given high-resolution data, the capability of *S&W* to extend phases from marginal models having essential random phases at higher resolution is remarkable.

Structure factors and phase errors up to 1.2 Å for variously perturbed models used in the structure solution of a cytochrome c' dimer (RSCP; PDB code 1gqa) from *Rhodobacter sphaeroides* (Ramirez *et al.*, 2003) were calculated and the results are explained in Fig. 1. In summary, Gaussian (error function, graph B) perturbation alone tends to have little effect at low resolution (compared with deletion alone, graph A), while at the same time it introduces overly large phase errors at higher resolution. A simple random perturbation (range 0–0.5 Å) combined with 10% random atom deletion (D) yields a smoothly increasing phase perturbation over the whole resolution range. A second set of graphs in Fig. 1 shows the phase error between the (correctly placed) initial model 1cpq and the final structure (1gqa). After the first round of rebuilding, the phase error between the first rebuild and the final structure is comparable to the phase error introduced by perturbation, indicating to what significant degree any starting model becomes perturbed. As would be expected in the case of bias removal, the correlation between the final model 1gqa and the *S&W* map from the first rebuild (0.89) is much better than the correlation between the *S&W* map from the first rebuild and the first rebuild model itself (0.76).

For poor MR solutions (models with an initial correlation coefficient in the range 0.3–0.35) automated sequence correction and an initial step of *CNS* slow-cool simulated annealing and torsion refinement (Adams *et al.*, 1999) can be used to assure convergence. The weakest MR model we have been able to rebuild using *S&W* and manual rebuilding had a correlation coefficient (CC) of 0.32 (Rv3465, 1.6 Å data; Fig. 2). The absolute value of the CC, however, depends critically on the quality of the low-resolution data (Dauter & Wilson, 2001) and does not necessarily provide a good predictor for the convergence of the *S&W* procedure against high-resolution data. For the purpose of bias removal and/or structure validation, a single run of *S&W* is sufficient to provide reliable local analysis by real-space correlation plots as described in the corresponding section.

3. Implementation of the *S&W* bias-removal service

Although the interactive interface of *CCP*4 (Potterton *et al.*, 2003) allows even novice users to navigate through the *CCP*4 program suite with relative ease, the scripting of a complex distributed routine such as *Shake&wARP* would be a daunting task. Therefore, we have implemented *S&W* as an easy-to-use web service incorporating several utility routines which clean up and convert the input coordinate files, standardize the PDB model file and select proper parameters for the ~20 different *CCP*4 programs called. We also incorporated run-time routines to produce GIF-format plots of real-space correlation and *B*-factor plots, *B*-factor histograms and, provided intensities were supplied, Wilson plots. The web routine provides the most common options, but does not allow extensive experimentation with all parameters. The set of parameters we

have selected as defaults are based on significant experience with the program on several platforms and work for the vast majority of cases. Stability and load limitations of the web service impose certain limitations on parameter choice.

The web service (http://tuna.tamu.edu) was implemented on a Linux cluster (RedHat Linux 7.3) controlled by one four-CPU main web and Condor job-queuing server; six dual CPU nodes execute the jobs distributed *via* Condor version 6.2.2.

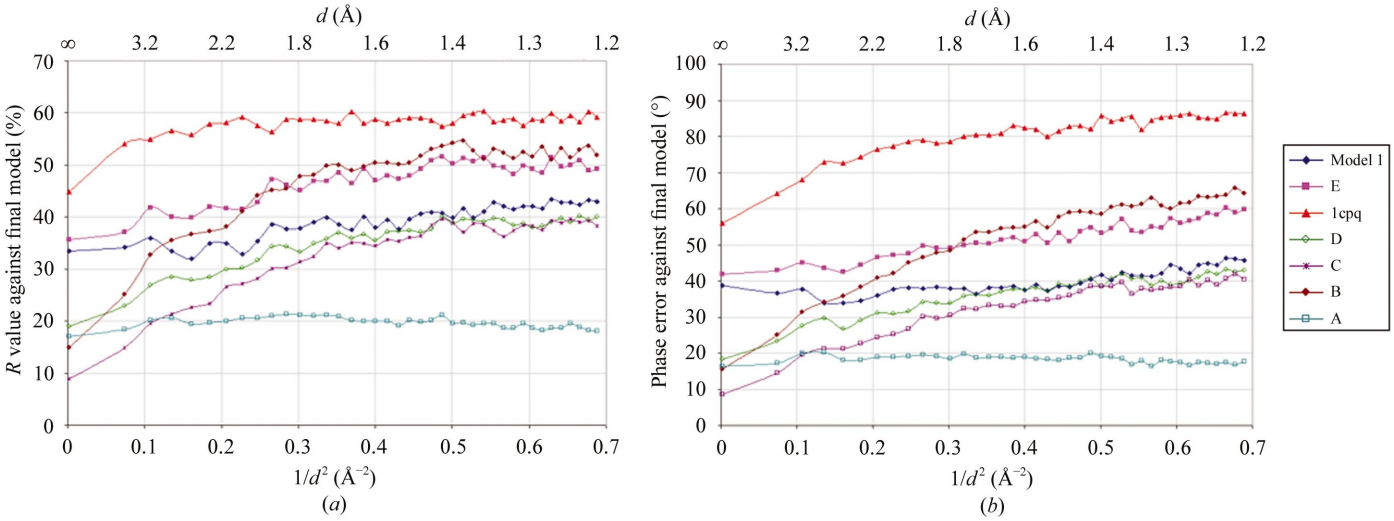


Figure 1
R value (*a*) or phase error (*b*) between the final RSCP structure model (1gqa) and various models used in *S&W*. Graph legends are as follows. 1cpq, the replaced initial MR model; model 1, sequence-corrected and CNS simulated-annealing torsion-angle refinement of the 1cpq model; A, 10% of the atoms of the final model deleted at random; B, error-function perturbation of the final model with coordinate deviation σ of 0.25 Å; C, linear random perturbation of the final model between 0 and 0.5 Å (mean = 0.25 Å); D,10% random atom deletion and linear random perturbation between 0 and 0.5 Å; E, model 1 perturbed as in D. While error-function perturbation mode B alone yields very high perturbation at higher resolution and has little effect at low resolution, combination D appears to be an optimal compromise and yields smoothly increasing model perturbation with increasing resolution. Phase error ($\Delta\varphi$) in (*b*) is given in degrees, corresponding figure of merit (FOM) equals $\cos(\Delta\varphi)$. While at 12–3 Å 1cpq is accurate enough as a model to yield a weak but clear MR solution, phases for the 1cpq model are practically random beyond 3–2.5 Å, despite the C^{α} r.m.s.d. of only 1.44 Å for alignment of 1cpq with the final model. Note that good experimental MAD/SAD phases have FOMs approaching 0.8–0.9 even at high resolution (better than 2.0 Å), which emphasizes how weak and biased MR phases are in comparison.

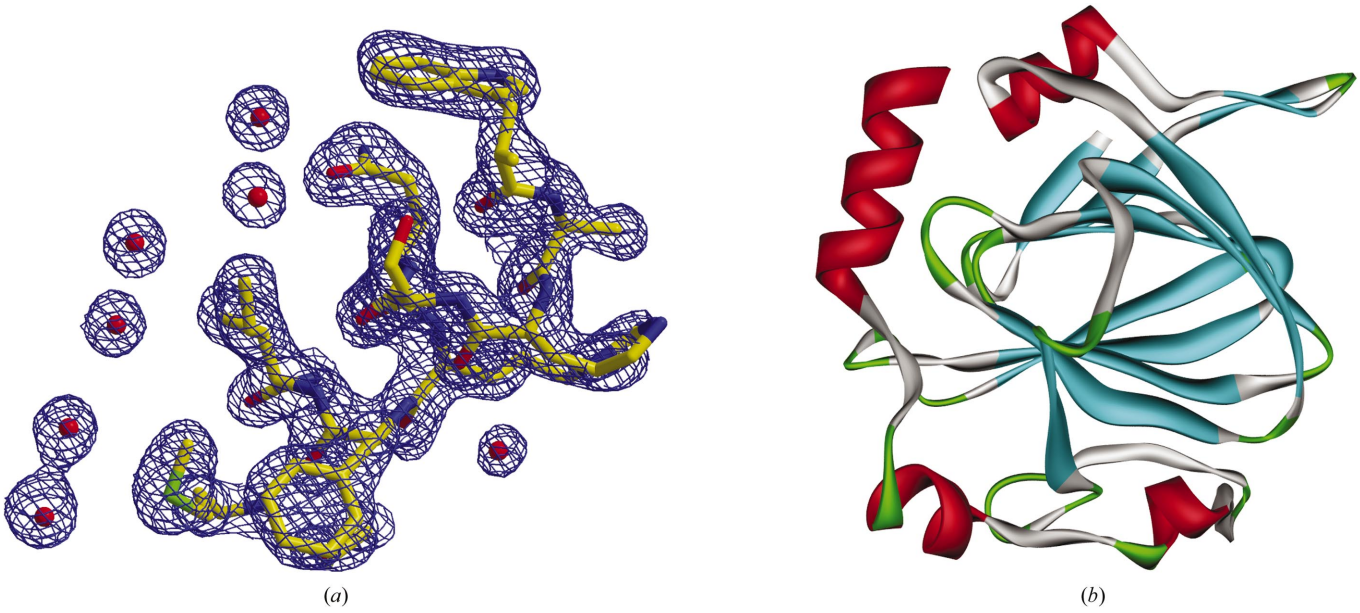


Figure 2
Monomer of *M. tuberculosis* d-TDP-4-dehydrorhamnose 3,5-epimerase (Rv3465 rmlC, PDB entry 1upi). RmlC was the first structure entirely processed by the facilities of the TB Structural Genomics Consortium using the *Shake&wARP* protocol from a poor starting model with an *EPMR* solution correlation coefficient of 0.32. (*a*) Bias-minimized *Shake&wARP* electron-density map contoured at the 1σ level demonstrating the clarity of the map and solvent definition. The final model of the missing regions (not used in map calculations) is superimposed on the map. (*b*) Ribbon diagram of the final molecular structure. Intermediate steps of map improvement and comparison with *REFMAC* $2mF_o - DF_c$ maximum-likelihood maps are shown in Rupp (2003). All figures showing electron density have been created with *XtalView*/*Xfit* (McRee, 1999) and *Raster3D* (Merritt & Bacon, 1997).

The web-server scripts were written using Perl-Cgi and Perl v.5.6.1 as the scripting language. When a job is submitted through the web server, a validation program (F90, described below) is executed and a shell script splits the submission into six parallel sub-jobs in the Condor queue, sending the jobs to free CPU nodes. The queue control then waits for all six jobs to complete and the main server continues post-processing and finalizing the output. The general program flow is depicted in Fig. 3.

3.1. Input preparation

Only two data files and a few control parameters need to be provided to start *S&W*: the model in PDB format and reflection data in *SCALEPACK* (Otwinowski & Minor, 1997), *X-PLOR*/*CNS* (Brünger *et al.*, 1998) or ASCII text format. Other required input includes unit-cell parameters (if the data are not in *SCALEPACK* format), the number of molecules per asymmetric unit and the number of residues per asymmetric unit (used to determine the optimum number of atoms to be placed/removed in each *ARP_WATERS* cycle and for F_{000} estimates in *TRUNCATE*). The following control parameters are selectable: optional molecular replacement and multi-segment rigid-body refinement with *EPMR* (up to three molecules per asymmetric unit are allowed), use of a poly-alanine model which can be advantageous for sharpness of *EPMR* solutions (Kissinger *et al.*, 2001), removal of water atoms (automatically if MR was selected) and the standard option of executing bias removal and creating plots. Fig. 4 shows the simple input panel of the server front page.

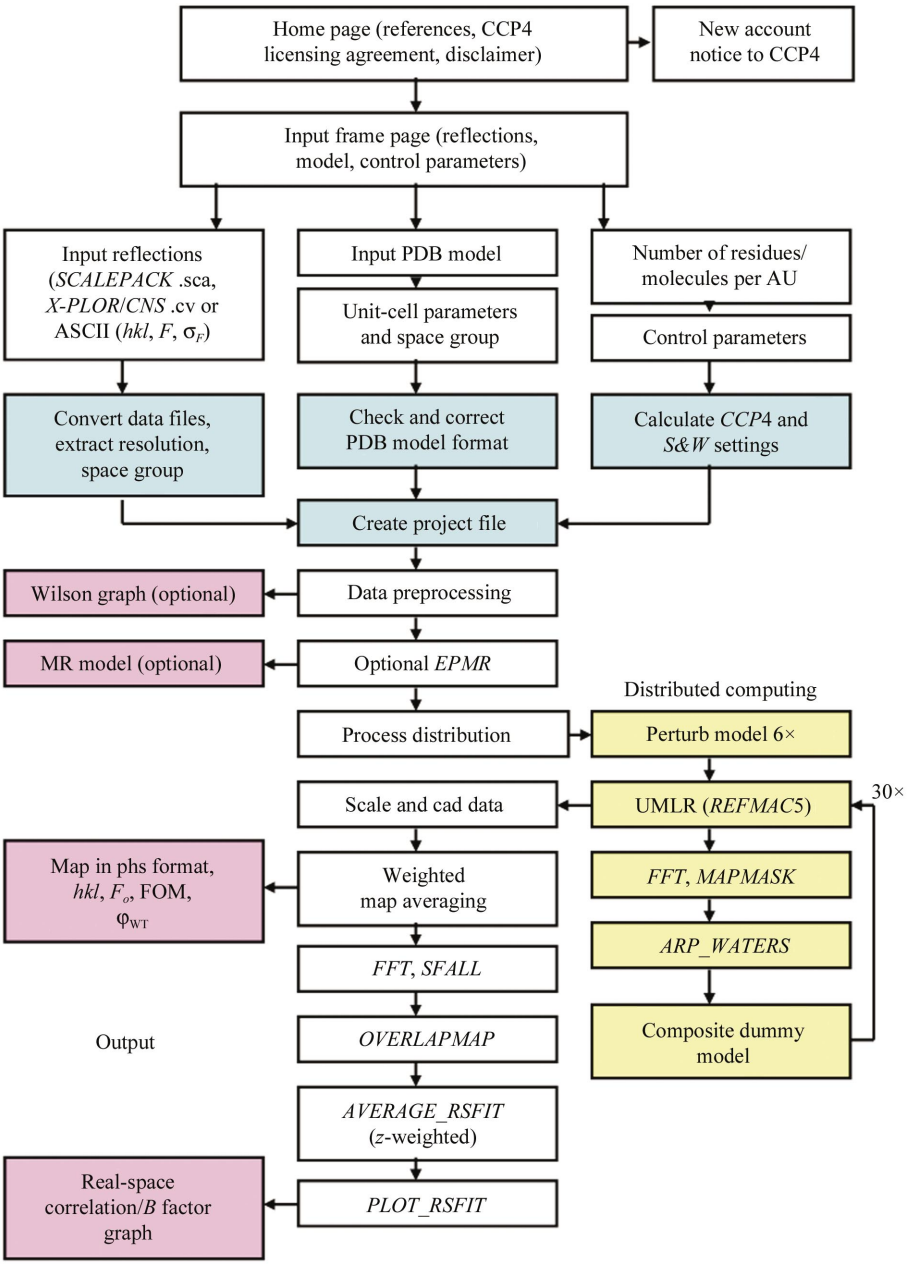


Figure 3
Schematic program flow of a *S&W* web submission. Blue, initial input preparation and validation; yellow, iterative steps conducted on cluster server members; magenta, output files.

User data and control selections are initially checked for consistency at the client level using Java scripts and after the initial input validation a preparation routine checks data for consistency, prepares and converts data files into *CCP*4 format and checks and standardizes the PDB file. A number of additional control parameters are derived from the user input by the setup routine and written to a project file. In the stand-alone version, this project file can be edited in order to perform special tasks with non-standard parameter combinations. The input-validation program derives the following *CCP*4 settings: resolution limits, FFT grid spacings according to resolution and space group, limits for *FFT*, *ARP_WATERS* and *SFALL*, and the number of atoms to remove/rebuild according to the asymmetric unit-cell contents. In the web implementation, the number of *S&W* cycles is fixed at 30, although the slope of the *R*-value convergence is reported and allows automated termination.

A report of the check and the parameter settings is returned to the web client and the input can be corrected or execution of the initial script started. After further data preparation and standardization of the files, optional MR is executed on the main server (20 cycles of *EPMR* if convergence is not reached, 12–4 Å data), followed by multi-segment rigid-body refinement against data to 2.8 Å. Subsequently, six scripts are generated and submitted to the Condor queuing system.

3.2. Output of results

After initial validation, the user is prompted whether to continue executing the job. Once the user confirms, an e-mail notification is sent and it takes 1–20 h depending on the complexity of the problem and the server workload for the results to become available. In the meantime, the user is able to see (or download, if curious) the temporary files and logs generated while the job is progressing. If the job succeeds, an e-mail notice is sent and the following can be retrieved from the server *via* the results web page.

(i) A *.phs file (*h*, *k*, *l*, *F*, FOM, PHWT) to create a bias minimized map with f*fom and phwt as Fourier coefficients [best viewed in *XtalView*/*Xfit* (McRee, 1999) by selecting f*fom as map type].

(ii) The replaced input model, if MR was selected.

(iii) For each chain in the model, publication-quality GIF images of real-space correlation plots between the fit of the model to the electron density combined with per-residue *B*-factor plots as well as an accumulative *B*-factor histogram (examples are given in the subsequent sections).

(iv) The data file converted into *.sca, *.fin and MTZ format (5% free flags set).

(v) The phased data in MTZ format (HKL, FP, SIGFP, FOM, PHWT, FreeR_Flag). The free set is used internally only to estimate α_A in the *REFMAC* ML dummy-atom refinement and except in the case of a new MR solution one should continue using the original free set.

The website introduction page includes further detailed description regarding usage, file formats, control parameters, job control, interpretation of results and licensing (only a *CCP*4 license is required to use the service; no components of the separate *ARP*/*wARP* package are used. The program *ARP_WATERS* is part of the standard *CCP*4 distribution. *EPMR* does not require licensing).

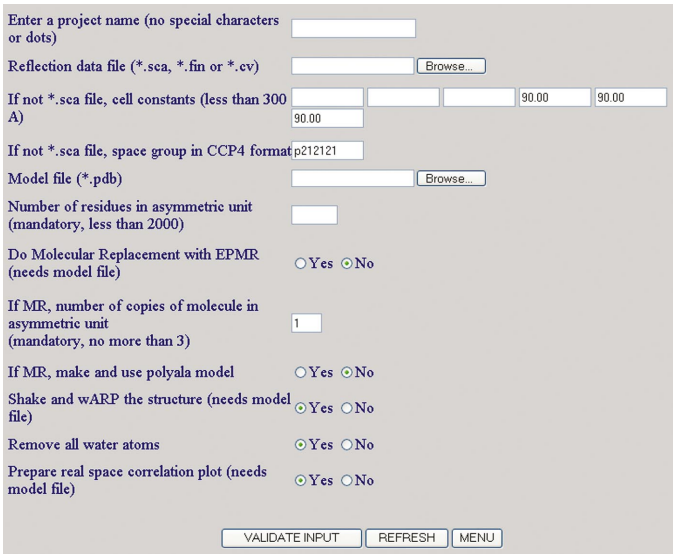


Figure 4
The simple user interface of the TB Consortium Bias Removal Service. All other program parameters are calculated from the input data and if the validation results are consistent then the user can submit the job.

3.3. Real-space correlation plots

Global indicators of structural quality such as the *R* value and $R_{\rm free}$ (Brünger, 1992) convey very little about the actual correctness of the structure and numerous examples exist of partially (or purposefully for demonstration) incorrectly traced structures with unsuspicious statistical descriptors (Dodson *et al.*, 1996; Kleywegt & Jones, 1995). In the case of molecular-replacement structures, even checks based on the plausibility of the local geometry such as those implemented in *WHAT-IF* or *PROCHECK* may not immediately trigger strong warning signs, particularly at low resolution and when refined with molecular-dynamics protocols, where geometric restraints dominate the refinement (Dodson *et al.*, 1996). In general, careful inspection of regions flagged in geometry checks, particularly Ramachandran plots (Sasisekharan, 1962; Ramachandran *et al.*, 1963), nearly always reveals problems with a structure (Kleywegt & Brünger, 1996; Rupp & Segelke, 2001). However, the most comprehensive and fastest assessment of local quality, provided structure-factor amplitudes are available, is the real-space correlation coefficient (RSCC) between the calculated model map and the 'experimental' map calculated from observed intensities (Branden & Jones, 1990), particularly when the map contains a minimum of model bias. The RSCC has the benefit of being scale-independent compared with real-space *R* values and atoms placed correctly in weak density still correlate highly. Areas with low real-space correlation coinciding with areas of high *B* factors indicate that model tracing in these areas is in all likelihood genuinely ambiguous owing to lack of electron density. Deviations from the anti-correlation of *B* and RSCC nearly always indicate problem areas worth investigating (examples are provided in the next section). *SFCHECK* (Vaguine *et al.*, 1999) and *OVERLAPMAP* from the *CCP*4 suite provide real-space correlation analysis. From a survey of the literature, however, it appears that RSCC plots are not as frequently used as they probably should be. The electron-density server (EDS) at the University of Uppsala (http://portray.bmc.uu.se/eds/) is a very useful web tool to locate potential problem areas in deposited structures. Such analytical web tools can be further enhanced to allow users to submit their coordinates and structure-factor files. Application of map improvement and phase-bias reduction routines such as the *Shake&wARP* service with return of corresponding RSCC plots and weighted Fourier map coefficients to the submitter for further refinement and rebuilding would probably promote the use of RSCC plots and contribute to increasing the quality of deposited structures.

4. Map improvement and bias reduction at work

The following section provides examples of *Shake&wARP* maps as produced by the web service. Examples include map improvement at various resolutions and states of completeness and reconstruction of absent parts or removal of questionable model parts or ligands. Even less spectacular improvements in map quality can make the difference

between a clearly traceable map and a frustrating refinement stalled at high R values, in particular for less experienced model builders, who are then more likely to succeed and to avoid some of the mishaps we show in the examples. The clarity of the averaged maps obtained from nearly finished models allows unambiguous identification of ligands and detailed fine-tuning of structural models.

4.1. Model correction and improvement

4.1.1. Sequence correction at 1.8 Å in cytochrome c′ from *R. sphaeroides*. In the crystal structure solution of a cytochrome c' dimer (PDB code 1gqa) from *R. sphaeroides* (Ramirez *et al.*, 2003), initial phases were obtained from a modest MR solution (CC = 0.44 after rigid-body refinement) using *EPMR* (Kissinger *et al.*, 1999) with the coordinates of the *R. capsulatus* cytochrome c' (1cpq) as a search model (Tahirov *et al.*, 1996). After the first round of sequence adjustment during model building into maps generated by *Shake&wARP*, a mismatch of the sequence became evident (Fig. 5). Note that the most significant improvement occurs after averaging of the six *Shake&wARP* runs, attesting to the power of map averaging for density improvement (Kleywegt & Read, 1997). It must be also noted that improvements over the *REFMAC* ML coefficient map come at a substantial price in computational effort: *Shake&wARP* spawned a total of 150 runs of unrestrained *REFMAC* ML refinements.

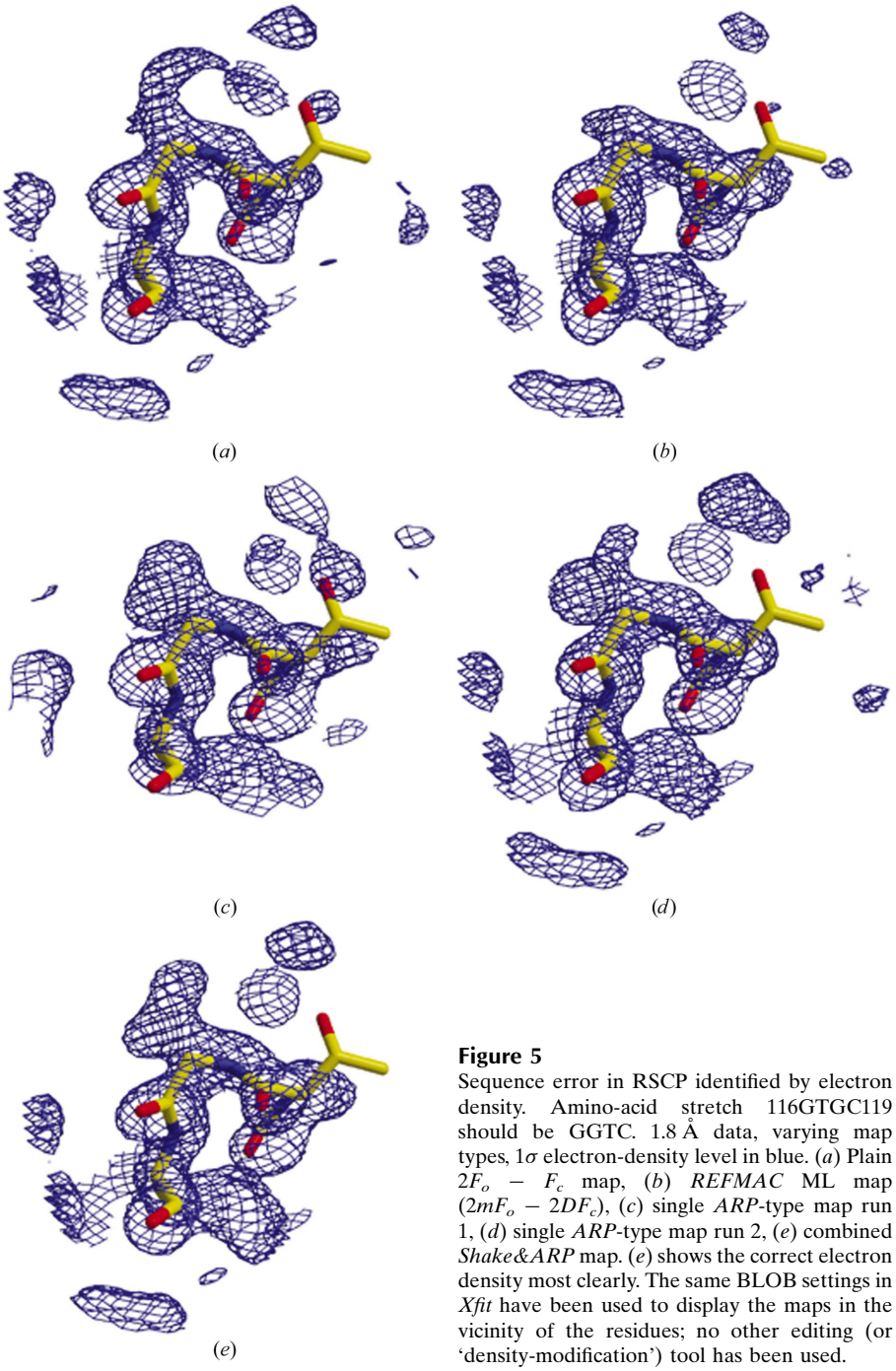


Figure 5
Sequence error in RSCP identified by electron density. Amino-acid stretch 116GTGC119 should be GGTC. 1.8 Å data, varying map types, 1σ electron-density level in blue. (*a*) Plain $2F_o - F_c$ map, (*b*) *REFMAC* ML map ($2mF_o - 2DF_c$), (*c*) single *ARP*-type map run 1, (*d*) single *ARP*-type map run 2, (*e*) combined *Shake&ARP* map. (*e*) shows the correct electron density most clearly. The same BLOB settings in *Xfit* have been used to display the maps in the vicinity of the residues; no other editing (or 'density-modification') tool has been used.

4.1.2. The elusive N-terminus of calmodulin at 1.8 Å. In a near-final model of calmodulin, the N-terminal three residues could not be unambiguously built into *CNS* simulated-annealing omit maps (provided by R. Skeene and B. Phipps, unpublished). When this model was subjected to a full bias-removal run (automated MR using *EMPR* followed by *Shake&wARP*), the correct connecting electron density became clearly visible (Fig. 6*b*) and the previously unmodelled initial three residues could be unambiguously built backwards from the fourth residue (Fig. 6*d*). While an experienced model builder might recognized the missing residues, the incorrect connectivity apparent in the SA omit map (Fig. 6*a*) is likely to complicate model building in the N-terminal region.

4.2. Low-resolution data

4.2.1. Apolipoprotein E4. The applicability of the *ARP* procedure, which in turn determines the resolution limit beyond which *Shake&wARP* can be used, has been discussed in detail (Lamzin & Wilson, 1993). Subject to the effects of map noise, data completeness and other effects that impair map quality, even a reasonable 2.5–2.8 Å MR model should allow the application of *Shake&wARP*. Apolipoprotein E4 (ApoE4) was solved in a fully automated manner from an ApoE3 search model (PDB code 1bz4; Segelke *et al.*, 2000) using *EPMR*, followed by rigid-body refinement against the 2.5 Å data

set and *Shake&wARP*. Unambiguous visibility of the ApoE3/ApoE4 isoform difference (Cys112Arg) between ApoE3 (model) and ApoE4 (electron density) is evident even at a resolution approaching the limit of applicability of the underlying *ARP* program (Fig. 7).

4.2.2. LysA from *Mycobacterium tuberculosis*. LysA is an essential gene of *M. tuberculosis* involved in the last step of lysine biosynthesis through stereospecific decarboxylation of *meso*-diaminopimelic acid (Gokulan *et al.*, 2003). Strong data to 2.8 Å were available and the initial protein-only structure model was submitted to the web service. The resulting averaged electron density (Fig. 8*a*) clearly showed soaked PLP (vitamin B_6) covalently bound as the cofactor and the product lysine (added in excess to the crystallization cocktail). A further low-resolution example is provided below (BABIM complex).

4.3. (Un)ambiguous ligands

4.3.1. PcaA. In a 2.2 Å structure of PcaA, an *S*-adenosyl-L-methionine-dependent methyltransferase from *M. tubercu-*

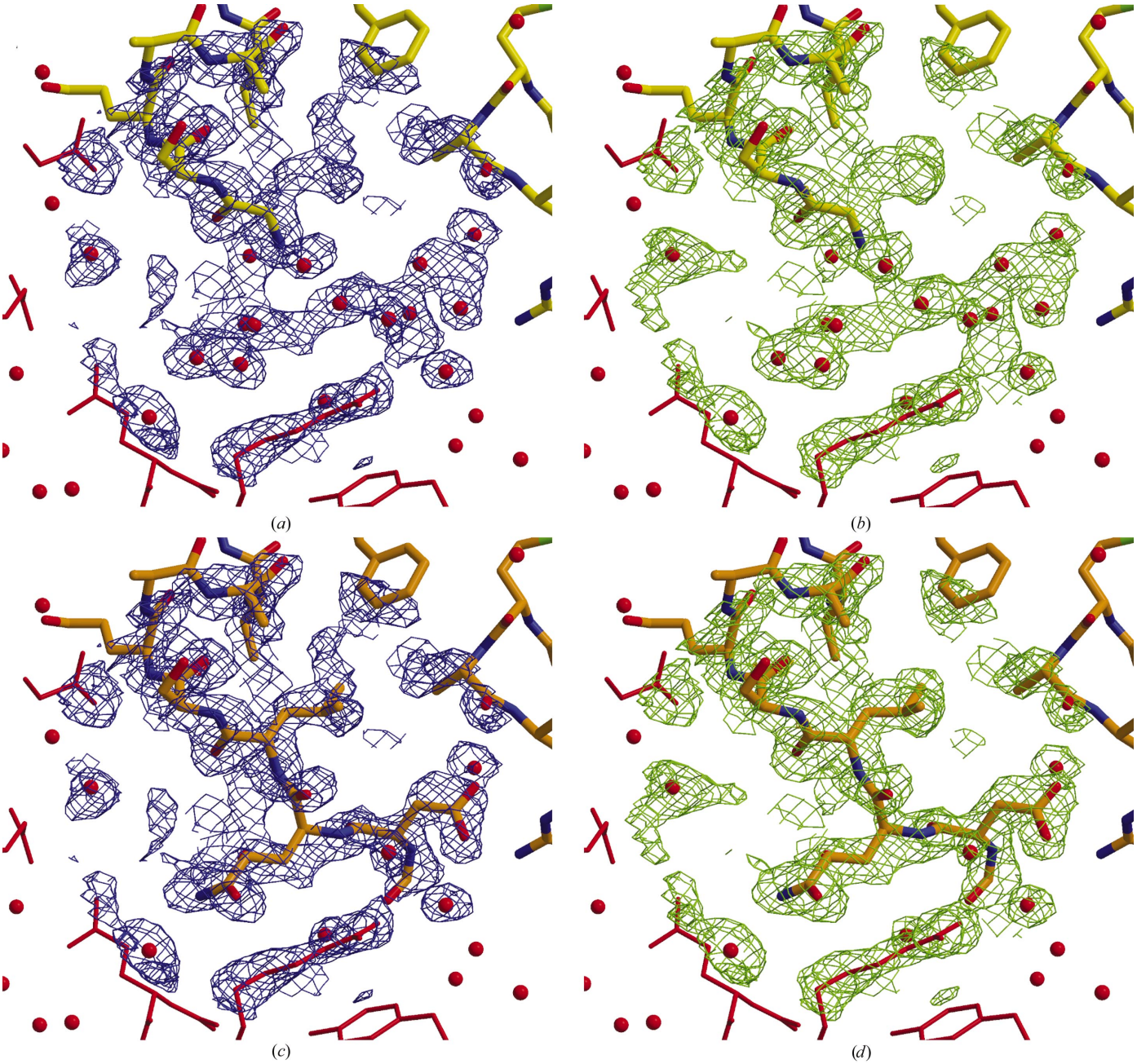


Figure 6
Corrected connectivity for the three N-terminal residues in calmodulin. (*a*) Incomplete model placed in the *CNS* $2F_o - F_c$ map (blue contours, 1σ). (*b*) Incomplete model placed in the *Shake&wARP* map (green contours, 1σ). (*c*) Complete model fitted to the *CNS* $2F_o - F_c$ electron density. (*d*) Completed model fit to the *Shake&wARP* electron density. Although the three N-terminal residues could have been placed correctly into the *CNS* $2F_o - F_c$ electron density by an experienced crystallographer, the incorrect map connectivity before residue Leu4 would likely stall automated chain tracing and model building (or a less experienced model builder). The correct connectivity of the electron density is quite clear in the *Shake&wARP* map.

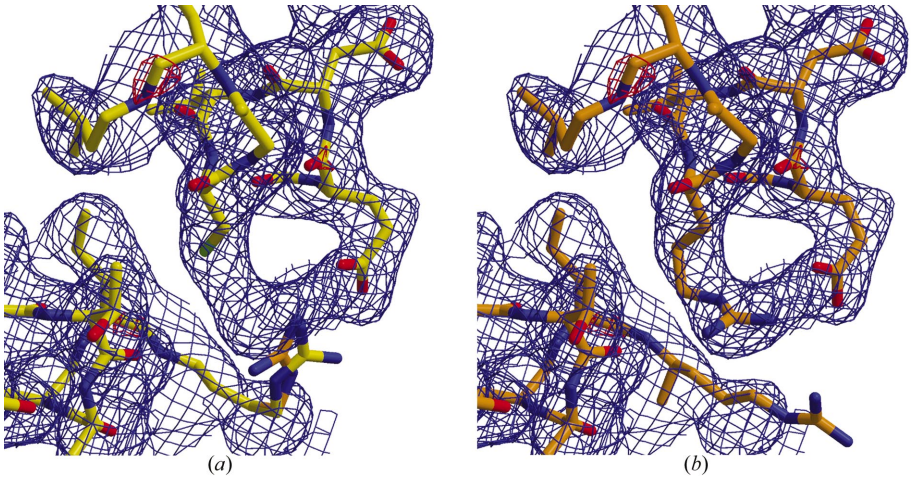


Figure 7
ApoE3/4 isoform difference Cys112Arg. The density shown is that for the ApoE4 isoform, which was solved by MR in a new crystal form (PDB entry 1gs9). The ApoE3 model (Cys112) is shown in (*a*). In the E3 isoform, Glu109 is hydrogen bonded to Arg61. The ApoE4 model is shown in (*b*). Arg112 clearly fits into electron density, making a new hydrogen bond to Glu109 and disrupting the hydrogen bond to Arg61. Arg61 adopts an extended conformation, changing the charge disposition of the helix 2–helix 3 surface and affecting very-low-density lipoprotein binding (Dong *et al.*, 1994).

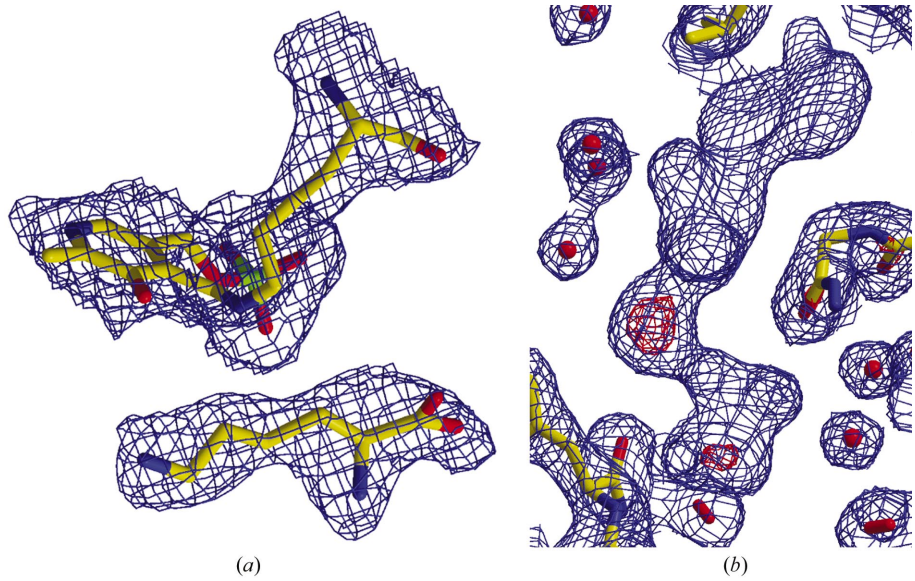


Figure 8
Shake&wARP map of *M. tuberculosis* drug-target structures. (*a*) Covalently bound cofactor PLP and product lysine were clearly recovered in the active site of LysA, a diaminopimelate decarboxylase, at 2.8 Å (Gokulan *et al.*, 2003). (*b*) PcaA, an *S*-adenosyl-L-methionine-dependent methyltransferase from *M. tuberculosis*. The standard *Shake&wARP* protocol at 2.2 Å clearly recovers the electron density of the *S*-adenosyl-L-homocysteine (Huang *et al.*, 2002).

losis (Huang *et al.*, 2002), we demonstrated the capability of *Shake&wARP* to recover ligands in complex structures. The presumed ligand *S*-adenosyl-L-homocysteine (SAH) was excluded from the model and the remainder of the model was submitted to the TB Bias Removal Server. The *Shake&wARP* map in Fig. 8(*b*) clearly recovers the SAH ligand.

4.3.2. *Clostridium botulinum* serotype B neurotoxin light-chain protease–BABIM complex. The BABIM complex of *C. botulinum* neurotoxin serotype B light-chain (BotLCB) protease (Hanson *et al.*, 2000) was submitted to the web service. The data reportedly extend to only 2.7 Å, but 2.5 Å data were deposited in the PDB (PDB code 1fqh) and these were used without any σ cutoffs for *Shake&wARP*. As an additional control for the recovery of electron density, a residue close to the BABIM inhibitor (Glu170, $B = 38$ Å^2; the average B of the protease is also 38 Å^2) and the catalytic Zn atom were also removed. The *Shake&wARP* map in Fig. 9 clearly recovers the omitted residue (perhaps not quite unambiguously built), as well as the Zn atom and the O atom of the catalytic water. However, despite 0.5σ map contouring, there is no indication of the BABIM ligand. Given its reported excessive average B factors of 130 Å^2, the inhibitor BABIM, which exhibits very few contacts to the protease, unfavorable geometry and little if any electron density, is not likely to be present in any substantial amount in this structure. Based on these findings, a correction has been published (Hanson *et al.*, 2002).

4.3.3. *C. botulinum* serotype B neurotoxin light-chain protease–synaptobrevin-II complex. A dramatic example of where the use of a real-space correlation plot would have provided early warning signs of an incorrect model is the complex of BotLCB with synaptobrevin (1f83). The plot created by the web service (Fig. 10) reveals extremely poor real-space correlation and excessive B factors for the ligand. Severe problems with the ligand refinement, including the absence of the ligand, must be expected. It is worthwhile mentioning that the deposition of structure factors for both BotLCB complexes indicates that an honest mistake was made. Suppression of structure factors when obvious warning signs are present may shed serious doubt on the validity of a structure, as recently discussed by Kleywegt & Jones (2002).

4.3.4. Buffers make excellent ligands. A taste receptor binds a number of molecules, some of which taste up to 20 000 times as sweet as sugar (K. Gokulan, unpublished work). A refined and completed model of the structure with omitted ligand was submitted to *Shake&wARP* and the resulting map shed serious suspicion about the presence of the ligand. While the *CNS* SA omit ML map may have sufficed to convince the

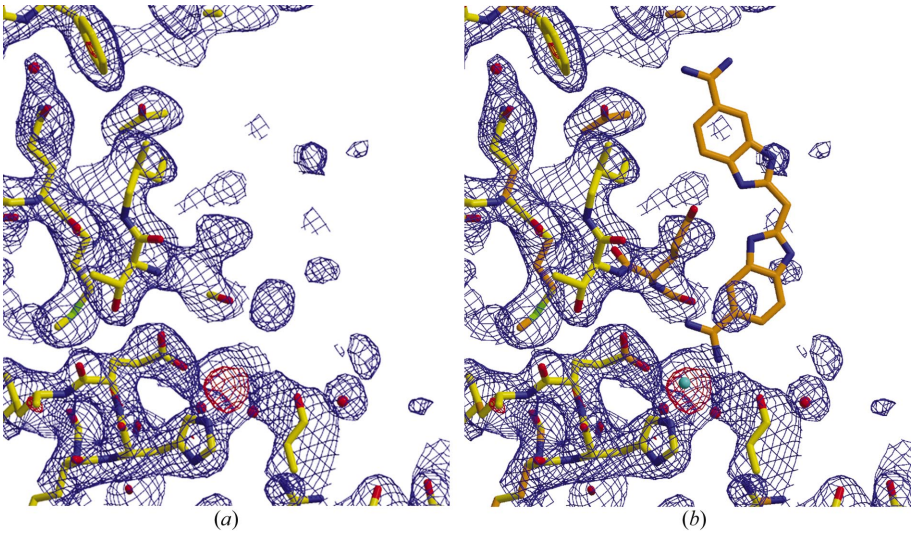


Figure 9
C. botulinum serotype B neurotoxin light-chain protease–BABIM complex. This structure (PDB code 1fqh; Hanson *et al.*, 2000) was subjected to the standard *Shake&wARP* procedure. Data were used as deposited (2.5 Å), without any σ cutoffs. BABIM, Glu170 and the catalytic Zn ion were omitted. Glu170 and the Zn ion are clearly recovered by the *Shake&wARP* procedure. However, little if any electron density is evident for the planar BABIM ligand, despite 0.5σ map contouring.

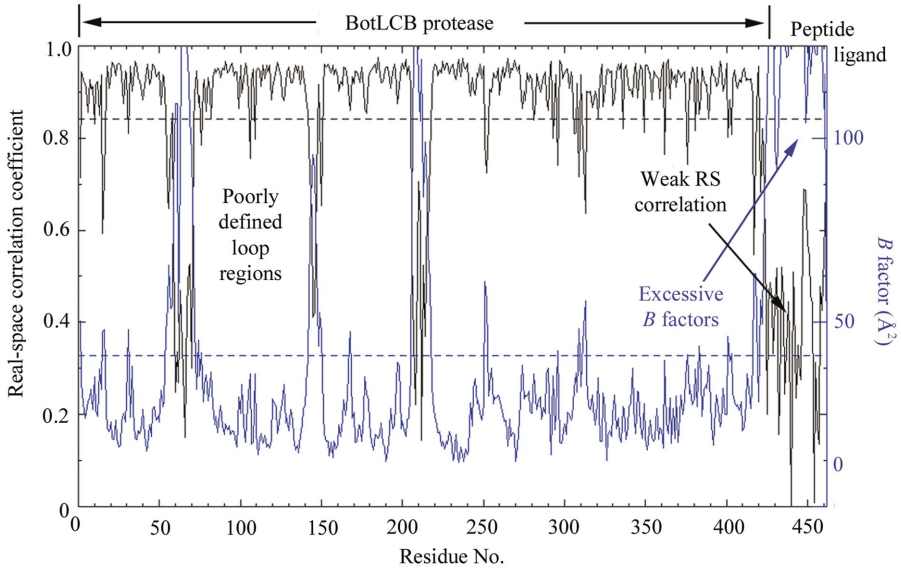


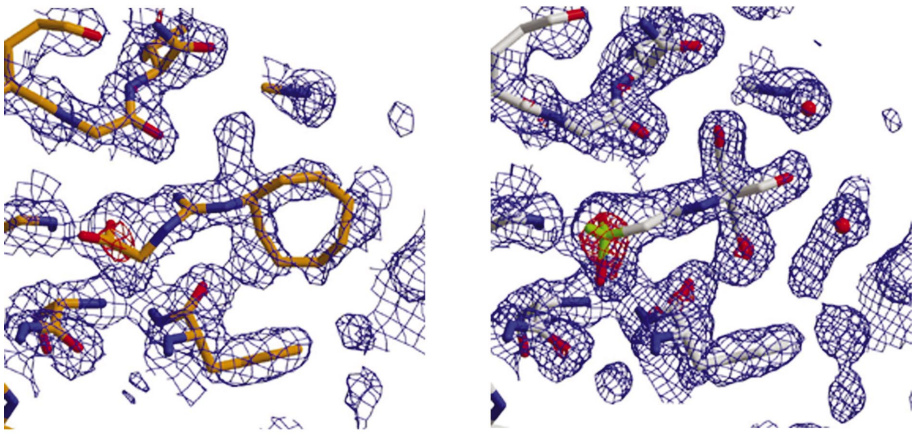
Figure 10
Real-space correlation coefficient and *B*-factor plot. PDB entry 1f83 (2.0 Å) contains the model coordinates for the BotLCB protease–synaptobrevin-II complex (Hanson & Stevens, 2000). Shown in black (upper curve) is the residue-by-residue real-space correlation coefficient; the *B* factors are plotted in blue for each residue. The left part of the figure corresponds to the protease, which, with the exception of three loop regions, shows normal behavior. The synaptobrevin-II ligand peptide at the right figure edge, however, shows a very worrying crossover between abysmal real-space correlation and excessive *B* factors. A simple plot of this nature, inspected beforehand or submitted with the manuscript, would have raised sufficient flags to prevent the public discourse regarding the validity of the results (Rupp & Segelke, 2001). The plot (less descriptive labeling) was created by the *S&W* service.

proficient crystallographer that the ligand modeling was dubious (Fig. 11*a*), the enhanced clarity of the *Shake&wARP* map proves it beyond doubt (Fig. 11*b*), thus overcoming even wishful mental bias. Based on electron density, the ligand was identified as sulfonate buffer. Subsequent consultation of the crystallization protocol confirmed the presence of the zwitterionic TES buffer {2-[2-hydroxy-1,1-bis(hydroxymethyl)ethylamino]-ethanesulfonic acid} and the corresponding structure has been modeled in the density.

5. Conclusions

Consistent use of map-validation tools, including real-space correlation plots, can prevent the great majority of bias-caused errors commonly found in crystallographic models. Although these validation methods exist and some were introduced more than a decade ago, they are not as widely used as one would expect. We hope that public availability of our web service will make it convenient to use structure-factor-based validation techniques and thus contribute to an increased quality of protein structures. The concerning trend for structure factors to be absent when global quality indicators are poor has been pointed out recently (Kleywegt & Jones, 2002) and we also hope that deposition of structure factors and their use for structure validation become prevailing practice, as has been the case in small-molecule crystallography for many years. (At the time of writing, less than 50% of deposited coordinates in the Protein Data Bank were accompanied by corresponding structure-factor entries.)

We thank the laboratory members of Jim Sacchettini and Joel Sussmann for kindly supplying many coordinate and data sets for blind test cases and for evaluating the *Shake&wARP* results. BR wishes to thank LLNL and James Sacchettini for supporting his sabbatical leave at Texas A&M University. Lawrence Livermore National Laboratory is operated by the University of California under contract W-7405-ENG-48 from the US Department of Energy. This work was sponsored by NIH-NIGMS Grant No P50 GM62410 (TB Structural Genomics Consortium) and the Robert Welch Foundation at Texas A&M University.



(*a*) (*b*)

Figure 11
TES buffer in the ligand-binding site. 2.1 Å maps contoured at 1σ (blue) and 5σ (red). (*a*) Presumed ligand built into *CNS* ML $2F_o - F_c$ map; (*b*) *Shake&wARP* map with TES buffer built into density. This map has less noise and cleaner connectivity and reveals the true nature of the ligand. A questionable van der Waals contact is also obvious between 'ligand' and protein in (*a*).

References

Adams, P. D., Pannu, N. S., Read, R. J. & Brünger, A. T. (1999). *Acta Cryst.* D**55**, 181–190.
Bhat, T. N. (1988). *J. Appl. Cryst.* **21**, 279–281.
Bhat, T. N. & Cohen, G. H. (1994). *J. Appl. Cryst.* **17**, 244–248.
Blundell, T. L., Jhoti, H. & Abell, C. (2001). *Nature Rev. Drug Discov.* **1**, 45–54.
Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography.* London: Academic Press.
Branden, C. I. & Jones, T. A. (1990). *Nature (London)*, **343**, 687–689.
Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.
Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54,** 905–921.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Dauter, Z. & Wilson, K. S. (2001). *International Tables For Crystallography*, Vol. *F*, edited by M. Rossmann & E. Arnold, pp. 177–195. Dordrecht: Kluwer Academic Publishers.
Dodson, E. J., Kleywegt, G. J. & Wilson, K. S. (1996). *Acta Cryst.* D**52**, 228–234.
Dong, L.-M., Wilson, C., Wardell, M. R., Simmons, T., Mahley, R. W., Weisgraber, K. H. & Agard, D. A. (1994). *J. Biol. Chem.* **269**, 22358–22365.
Gokulan, K., Rupp, B., Pavelka, M. S. Jr, Jacobs, W. R. Jr & Sacchettini, J. C. (2003). *J. Biol. Chem.* **278**, 18588–18596.
Hanson, M. A., Oost, T. K., Rich, D. H., Stevens, R. C. & Sukonpan, C. (2002). *J. Am. Chem. Soc.* **124**, 10248–10248.
Hanson, M. A., Oost, T. K., Sukonpan, C., Rich, D. H. & Stevens, R. C. (2000). *J. Am. Chem. Soc.* **122**, 11268–11269.
Hanson, M. A. & Stevens, R. C. (2000). *Nature Struct. Biol.* **7**, 687–692.
Hendrickson, W. A. & Ogata, C. M. (1997). *Methods Enzymol.* **276**, 494–516.
Hodel, A., Kim, S.-H. & Brünger, A. T. (1992). *Acta Cryst.* A**48**, 851–858.
Huang, C. C., Smith, C. V., Glickman, M. S., Jacobs, W. R. Jr & Sacchettini, J. C. (2002). *J. Biol. Chem.* **277**, 11559–11569.
Islam, S. A., Carvin, D., Sternberg, M. J. E. & Blundell, T. L. (1998). *Acta Cryst.* D**54**, 1199–1206.
Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.
Kissinger, C. R., Gehlhaar, D. K., Smith, B. A. & Bouzida, D. (2001). *Acta Cryst.* D**57**, 1474–1479.
Kleywegt, G. J. & Brünger, A. T. (1996). *Structure*, **4**, 897–904.
Kleywegt, G. J. & Jones, T. A. (1995). *Structure*, **3**, 535–540.
Kleywegt, G. J. & Jones, T. A. (1997). *Methods Enzymol.* **277**, 208–230.
Kleywegt, G. J. & Jones, T. A. (2002). *Structure*, **10**, 129–147.
Kleywegt, G. J. & Read, R. J. (1997). *Structure*, **5**, 1557–1569.
Lamzin, V. S. & Wilson, K. S. (1993). *Acta Cryst.* D**49**, 129–147.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.
Matthews, B. W. (2001). *International Tables For Crystallography*, Vol. *F*, edited by M. Rossmann & E. Arnold, pp. 293–298. Dordrecht: Kluwer Academic Publishers.
Merritt, E. A. & Bacon, D. J. (1997). *Methods Enzymol.* **277**, 505–524.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Norvell, J. C. & Zapp-Machalek, A. (2000). *Nature Struct. Biol.* **7**, *Suppl.*, 931.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **267**, 307–326.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Perrakis, A., Sixma, T. K., Wilson, K. S. & Lamzin, V. S. (1997). *Acta Cryst.* D**53**, 448–455.
Potterton, E., Briggs, P. J., Turkenberg, M. & Dodson, E. J. (2003). *Acta Cryst.* D**59**, 1131–1137.
Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963). *J. Mol. Biol.* **7**, 95–99.
Ramirez, L., Axelrod, H., Herron, S., Rupp, B., Allen, J. & Kantardjieff, K. A. (2003). *J. Chem. Crystallogr.* **33**, 413–424.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Read, R. J. (1990). *Acta Cryst.* A**46**, 900–912.
Read, R. J. (1997). *Methods Enzymol.* **278**, 110–128.
Rupp, B. (2003). *Acc. Chem. Res.* **36**, 173–181.
Rupp, B. & Segelke, B. W. (2001). *Nature Struct. Biol.* **8**, 643–664.
Sasisekharan, V. (1962) *Stereochemical Criteria for Polypeptide and Protein Structures.* Madras, India: Wiley & Sons.
Segelke, B. W., Forstner, M., Knapp, M., Trakhanov, S. D., Parkin, S., Newhouse, Y. M., Bellamy, H. D., Weisgraber, K. H. & Rupp, B. (2000). *Protein Sci.* **9**, 886–897.
Tahirov, T. H., Misaki, S., Meyer, T. E., Cusanovich, M. A., Higuchi, Y. & Yasuoka, N. (1996). *J. Mol. Biol.* **259**, 467–479.
Terwilliger, T. C. (1999). *Acta Cryst.* D**55**, 1863–1871.
Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.
Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–120.
Vriend, G. (1990). *J. Mol. Graph.* **8**, 52–56.
Zeng, Z.-H., Castaño, A. R., Segelke, B. W., Stura, E. A., Peterson, P. A. & Wilson, I. A. (1997). *Science*, **277**, 339–345.