Journal of Structural and Functional Genomics **6:** 1–11, 2005.
DOI 10.1007/s10969-005-3138-4

A test of enhancing model accuracy in high-throughput crystallography

W. Bryan Arendall III[1], Wolfram Tempel[2], Jane S. Richardson[1], Weihong Zhou[2], Shuren Wang[1,4], Ian W. Davis[1], Zhi-Jie Liu[2], John P. Rose[2], W. Michael Carson[3], Ming Luo[3], David C. Richardson[1,]* & Bi-Cheng Wang[2,]*

*[1]Department of Biochemistry, Duke University Medical Center, Durham, NC, 27710-3711, USA; [2]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, 30602-1747, USA; [3]Center for Biophysical Sciences and Engineering, University of Alabama at Birmingham, Birmingham, AL, 35294-4400, USA; [4]Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, VA, 22908-0736, USA; *Authors for correspondence (e-mails: dcr@kinemage.biochem.duke.edu; wang@BCL1.bmb.uga.edu; fax +1-919-684-8885; +1-706-542-3077)*

Received 21 January 2005; accepted in revised form 10 February 2005

Key words: all-atom contacts, MolProbity, Ramachandran plot, sidechain rotamers, structural genomics, structure validation

Abstract

The high throughput of structure determination pipelines relies on increased automation and, consequently, a reduction of time spent on interactive quality control. In order to meet and exceed current standards in model accuracy, new approaches are needed for the facile identification and correction of model errors during refinement. One such approach is provided by the validation and structure-improvement tools of the MolProbity web service. To test their effectiveness in high-throughput mode, a large subset of the crystal structures from the SouthEast Collaboratory for Structural Genomics (SECSG) has used protocols based on the MolProbity tools. Comparison of 29 working-set and 19 control-set SECSG structures shows that working-set outlier scores for updated Ramachandran-plot, sidechain rotamer, and all-atom steric criteria have been improved by factors of 5- to 10-fold (relative to the control set or to a Protein Data Bank sample), while quality of covalent geometry, R_{work}, R_{free}, electron density and difference density are maintained or improved. Some parts of this correction process are already fully automated; other parts involve manual rebuilding of conformations flagged by the tests as trapped in the wrong local minimum, often altering features of functional significance. The ease and effectiveness of this technique shows that macromolecular crystal structures from either traditional or high-throughput determinations can feasibly reach a new level of excellence in conformational accuracy and reliability.

Introduction

Protein crystallography is a mature method that greatly enhances biological understanding with remarkably complete, detailed, and objective structural information. The well-ordered parts of high-resolution structures are among the most solid evidence obtainable in science. But every crystallographer knows that some regions are ambiguous and difficult, while comparisons of independent determinations of the same structure show coordinate errors much larger than theoretical estimates [1, 2].

The discipline of structure validation has developed as a cross-check on the reliability of macromolecular structures. Programs such as ProCheck [3], WhatIf [4], Oops [5], and MolProbity [6, 7] provide both overall statistical evaluations and flags of local problem areas, concentrating primarily on geometrical measures that can be analyzed from the model. Other validation utilities analyze aspects of the model-to-data agreement, such as SfCheck [8], real-space residuals [9], water-peak analysis in DDQ [10], and the now almost universally utilized R_{free} value [11]. Global validation measures serve the function of judging whether a structure meets accepted current practice, while local measures are especially important to users of structures, since no level of global quality can guarantee protection against a large local error in the region of your specific interest.

One shortcoming of current validation practice is that the evaluations are largely relegated to the final stages of refinement or even to the process of database deposition, as in the ADIT tool of the Protein Data Bank (PDB) deposition process [12]. That timing means that although globally poor structures are sometimes not made public, only rarely are identified local problems actually corrected.

A second issue is the challenge created by the high-throughput mandate of the structural genomics effort [13] to provide rapid, efficient, and thorough coverage of protein structure space. While developing and deploying new technology that reduces the interactive attention needed on each structure target, the quality of the models should nevertheless reflect their 'reference' status as lone representatives of their respective protein families. Important high-throughput technology developments relevant to structural accuracy have taken place in real-time validation of synchrotron data collection [14, 15] and in the process of automated refinement [16]. Model validation is a required component for all structural genomics centers, and high-throughput 'gatekeeper' criteria have been proposed for acceptability [17]; however, model validation for structural genomics has in general remained quite traditional and relatively minimal. The most notable exceptions are the Joint Center for Structural Genomics [18], which uses an unusually broad array of validation tools (http://www.jscg.org) and a 'second opinion' system; the NorthEast Structural Genomics center [19], which is working on NMR validation [20]; and the SouthEast Collaboratory for Structural Genomics or SECSG [21], as described here.

As part of technology development at the SECSG, we are investigating the benefits of a more tightly integrated approach to structure validation, which both utilizes new geometric and steric criteria and also applies them as an early and integral part of the pipeline's refinement-refitting cycle so that identified problems can be corrected to improve the accuracy of the structures. As a controlled test of the benefits of this protocol, one subset of the SECSG crystal structure output applied the validation and structure-improvement tools of the MolProbity web service (at http://kinemage.biochem.duke.edu), while another similar set used standard procedures. This study reports the comparison of those working vs. control sets, demonstrating that a dramatic improvement was obtained in structure quality.

Methods

This study involved 48 of the SECSG target structures solved by October, 2004. Of those, 29 underwent validation and structure improvement using MolProbity and related tools (the 'working set') and 19 did not use MolProbity (the 'control set').

In the working set, 17 targets were from the *Pyrococcus furiosus* genome, along with human, clostridial, nematode, and other sources. Their Protein Data Bank [12] codes, target ID's, and resolutions are listed in the first section of Table 1. All working-set data were collected at cryogenic temperatures, principally at the SER-CAT 22-ID beamline (South East Regional Collaborative Access Team) of the Advanced Photon Source at Argonne National Laboratories. Resolutions range from 1.2 to 2.7 Å, with an average of 1.91 Å and a median of 1.90 Å. Three structures used molecular replacement, while the other 26 were experimentally phased by SAS, MIRAS, or SIRAS methods, with heavy-atom substructures solved using either Solve [22] or Shelxd [23] and starting phases calculated either with Solve, ISAS [24] or MlPhare [25]. Automated tracing of an initial model was performed with Resolve [26] or ARP/wARP [27]. Refinement was done either in Refmac5 [28], using the CCP4i interface [29] of CCP4 [30], or in CNS [31]. Model rebuilding was primarily done in XFit [32].

Table 1. SECSG structures used in this study.

Working dataset			Control dataset		
PDB id	Target ID	Resolution	PDB id	Target ID	Resolution
1mjf	Pfu-132382	1.80	1l7l	Cth-FAEZ	1.50
1nnh	Pfu-1801964	1.65	1jjf	Pae-lectin1	1.75
1nnq	Pfu-1210814	2.35	1lpl	F53F4.3	1.77
1nnw	Pfu-1218608	1.90	1mo0	AAA79846	1.70
1pry	Pfu-65527	1.97	1nxc	Mum-ManA1	1.51
1ryq	Pfu-263306	1.38	1ooe	T03F6.1	1.65
1s36	Oob_CaW92F	1.96	1ooj	T21H3.3	2.11
1sen	095881	1.20	1pgv	C06A5.7	1.80
1sgw	Pfu-867808	1.70	1pzv	F58A4.10	2.52
1she	Pfu-871755	1.85	1q34	C35B1.1	2.90
1sl7	Oob_CaH_obelin	2.20	1qwk	C07D8.6	1.60
1sl8	Aae-Aeq	1.70	1r7j	Ss0-10a	1.47
1tov	F53F4.3	1.77	1r9h	F31D4.3	1.80
1twl	Pfu-264096	2.20	1row	C55C2.2	2.00
1ups	CpeEBGal	1.82	1spx	D1054.8	2.10
1vjk	Pfu-562899	1.51	1t7s	AAD16125	2.80
1vk1	Pfu-392566	1.20	1t9f	R12E2.13	2.00
1vka	q15691	1.60	1xhl	F25D1.5	2.40
1vkc	Pfu-35386	1.99	1xkq	R05D8.7	2.10
1xe1	Pfu-880080	2.00			
1xg7	Pfu-877259	1.88			
1xg9	Cth-2336	2.05			
1xhc	Pfu-1140779	2.35			
1xho	Cth-682	2.20			
1xi3	Pfu-1255191	1.70			
1xk8	060888	2.70			
1xma	Cth-833	2.30			
1xrg	Cth-2968	2.20			

For the working-set structures, the validation and structure-improvement tools of the MolProbity web service and related utilities [6, 7, 33] were used throughout the refinement process, typically starting immediately after initial chain tracing. MolProbity evaluates each side-chain conformation against an updated rotamer library and smoothed multidimensional χ-angle database distributions [34], each ϕ,ψ main-chain torsion pair against updated Ramachandran-plot distributions [6], and the deviation of each Cβ atom vs. the ideal position calculated from local backbone [6]. All-atom contact criteria are then addressed in MolProbity by first using Reduce [35] to add and optimize all hydrogen atoms in their local H-bond networks, with automatic resolution of ambiguous Asn/Gln/His orientations, producing a modified and commented PDB-format coordinate file. It then uses Probe [36] to calculate all-atom steric clashes, van der Waals contacts and H-bonds, producing residue-clustered lists of clashes ≥0.4 Å and graphic displays on the 3D structure. Subsequent model rebuilding focused on sections of the model where severe deficiencies were found in these tests: <1% probability for side-chain rotamers, <0.2% probability for ϕ,ψ values, or all-atom van der Waals overlaps > 0.5 Å. The residue-by-residue real-space fit to the experimental data, as shown by SfCheck [8], also guided the refitting process. Cycles of validation, refitting, and refinement were repeated until further changes failed to improve the steric parameters or the fit to electron density.

After convergence, the coordinates and structure factors were sent to the validation group at

Duke University. They examined the remaining problem areas on a MolProbity 'multi-criterion' kinemage in the KiNG [7] or Mage [37] display programs, along with $2F_o$–F_c and F_o –F_c electron density, and did further rebuilding where feasible. Changes included correction of side-chains trapped in the wrong local minimum conformation (much less common toward the end of this study, when they were mostly being corrected earlier), redefining solvent or ligands, and adding further parts of the structure that could be built reliably with the tools in KiNG and Mage for small backbone movements and for interactive rotamer and contact evaluation while remodeling [38]. If further changes were made, then one more brief round of refinement was done using the same options and parameters as before.

The control set consists of 19 SECSG target structures that did not make any use of the MolProbity protocols. Their PDB codes, target IDs, and resolutions are listed in the second section of Table 1. They are mostly from *Caenorhabditis elegans*, at resolutions ranging from 1.47 to 2.9 Å, with an average of 1.97 Å and a median of 1.80 Å. Data collection, phasing, chain tracing, and refinement were very similar to the working set, with the exception that they mostly used CNS rather than mostly using Refmac. Rebuilding of the control-set structures was done primarily with Quanta (from Accelrys) and Chain [39], however, and validation was done with ProCheck [3] not MolProbity. It should be noted that four SECSG crystal structures during this same time period were omitted from the study because they made partial use of MolProbity; their validation scores were indeed intermediate, but they could not fairly be assigned either to the working group or to the control group.

As a second comparison set, a representative sample of 1784 structures was chosen from fairly recent PDB depositions (their median date is approximately 01/01/00). To avoid redundancy or influence from molecular-replacement starting models at higher resolution, that sample consisted of the highest-resolution example in each family from release 1.65 of the SCOP fold classification [40], omitting any at resolution poorer than 3.5 Å. Only 1616 of the sample structures had R_{free} values, but all 1784 could be evaluated for the other criteria. This PDB-sample set is meant to approximate the recent state of the art in standard crystallographic practice.

The final overall quality evaluations from the working, control, and PDB-sample structures are reported as 2D plots of each validation criterion vs. resolution, since resolution is by far the most influential independent variable determining quality at the level of entire structures. The quantitative quality criteria used are the R_{work}, R_{free}, and bond-angle ideality as reported in the PDB file header, the overall real-space residual as reported by the Electron Density Server (http://fsrv1.bmb.uu.se/eds), and the overall MolProbity scores reporting outliers of all-atom steric clashes, side-chain rotamers, Ramachandran values, and Cβ deviations. Clash and rotamer scores were calculated with Asn/Gln/His orientations as in the deposited coordinates.

Data were tabulated, functions fitted, and plots produced in proFit 5.6 from QuantumSoft. Linear fits were not satisfactory for any of the validation parameters. R_{free} for the PDB sample was fit with a general quadratic polynomial, improved slightly by smoothing the data with a moving window of 5 points. The three new criteria are known to asymptote or plateau to very low values as resolution approaches zero [6, 41]; therefore the constant term in their fit is fixed at zero for clashscore and Ramachandran outliers, and at 0.5% for sidechain non-rotamericity (see Results section). The first-order term is omitted, giving a single-parameter fit on the quadratic term, which is quite satisfactory for the working and PDB datasets in Figure 1a–c (giving scattered but essentially flat residual plots, not shown) and acceptable for the small control set.

Results

These new geometrical and all-atom evaluation tools led to proposed changes in most of the working-set structures which not only improved the global quality scores but which can be seen individually, once accomplished, to provide a clearly better local match to the crystallographic data. Most of these refittings are major ones at the local scale, which add or delete groups of atoms or change their positions by several Å, as in traditional model rebuilding. The difference is

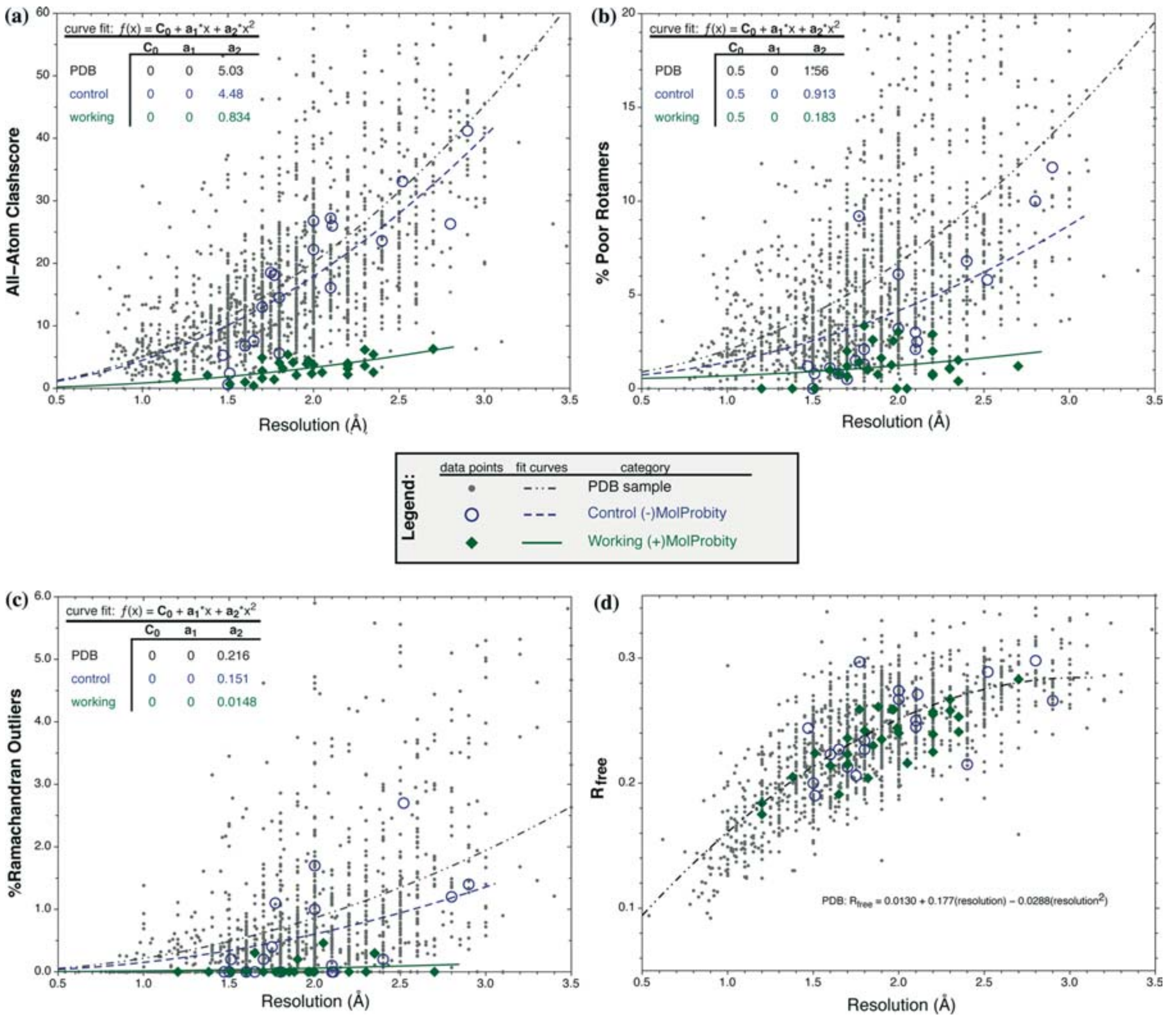

Figure 1. Plots of quality criteria vs. resolution, with quadratic fit lines, for the three datasets of the PDB sample (gray dots), the SECSG control set (open circles), and the SECSG MolProbity working set (solid diamonds). (a) All-atom clashscore (number of steric overlaps ≥0.4 Å per 1000 atoms) [36]; (b) percent poor sidechain rotamers (those with rotamer quality outside the 99th percentile contour defined by low *B*-factor, high-resolution data) [34]; (c) percent Ramachandran outliers (residues outside the 99.95% contour for general-case amino acids or the 99.8% contour for Gly, Pro, or pre-Pro) [6]; (d) R_{free} value [11].

that there is now additional information both for locating the problem and for evaluating which potential solution is the best to try.

The first and simplest corrections made in the working set are to the orientations of side-chain amides or imidazoles ambiguous by 180° (Asn/Gln/His 'flips'), which are assigned automatically in Reduce by optimizing both the H-bond networks of the polar atoms and the all-atom steric contacts of the amide NH_2 and the imidazole CH groups [35]. These flips have essentially no effect on R or R_{free}, but are well worth making since they have major consequences for electrostatics and H-bonding. In the control-set structures that did not use Reduce, 18.9% of Asn/Gln/His side chains would benefit (by more than the default threshold) from being flipped by 180°; this is very close to the PDB-sample average of 19.3%, as given in Table 2. The MolProbity working-set structures all have 0% Asn/Gln/His flips, since they used the same procedure as for the scoring, and since in this case it is always possible to make the suggested changes without hurting any other geometrical, real-space, or reciprocal-space measure. Therefore, although the Asn/Gln/His flip score is indeed a measure of structure quality, we do not cite it as an independently verified one. Recent confirming evidence, however, has come from nuclear magnetic resonance study of hen egg-white lysozyme, where all of Reduce's amide flip assignments not complicated by crystal contacts or space-group differences were confirmed by residual dipolar coupling measurements and other NMR data [42].

Tetrahedrally branched side chains can also quite often be fit backwards into electron density that appears too straight-across, resulting in eclipsed χ angles and distorted bond-angle geometry, such as for the clashing valine refit between Figure 2a and b. The case of leucine has been studied in detail [6], showing that fully staggered, non-clashing rotamers are strongly preferable by all criteria to the backward-fit, eclipsed conformations. The case of threonine has been illustrated by three refit and re-refined examples at differing resolutions [33]. Early structures in the working set often showed such backwards side chains, but in the later working protocols bad rotamers were not fit in the first place unless the data thoroughly constrained them. Such constraints do indeed occur for about 0.5–1% of high-resolution, low *B*-factor residues, but only where physical forces such as H-bonds can hold them out of the low-energy conformation. The ideal value for % non-rotameric side chains is therefore not zero but a bit less than 1%. The SECSG working set has an average non-rotamericity of 1.2%, compared with 3.7% for the control set (see Table 2).

Side-chain rotamericity, like all the criteria used here, is a strong function of crystallographic resolution; such a relationship is a necessary property of any meaningful measure of overall structure accuracy. Figure 1b, therefore, plots rotamericity as a function of resolution for the working set (filled diamonds), the control set (open circles), and the PDB sample (gray dots). A quadratic function is fit to each dataset for purposes of comparison (see Methods section). It is clear that both working and control sets score better than the PDB sample, and that the working-set structures refit by these procedures have been greatly improved. Working-set average rotamericity score is 3-fold better than average control-set score (with comparable resolutions for the two datasets). As measured by the quadratic fit parameter, working-set rotamericity is 5-fold better than for the control set and 9-fold better than for the PDB sample.

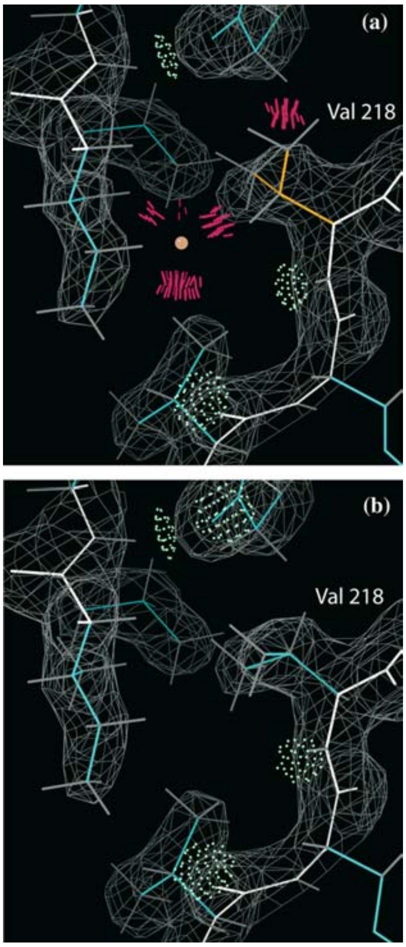

Figure 2. An example of using these methods to refit a backwards valine sidechain and an incorrect 'water' (central ball). In part (a) Val 218 is flagged in gold as a bad rotamer (at the 0.1% level) with an almost perfectly eclipsed χ_1, and has serious clashes (hot pink spikes) with both the water and the Cβ methylene of Asn 150. The water has no density above zero in the $2F_o$–F_c map (shown contoured at 1.2σ); it has serious clashes with all four of its surrounding nonpolar neighbors but makes no H-bonds (pale green dots). In part (b) the water was removed and the Val sidechain idealized and fit slightly better into the density, rotated 180° into a good rotamer. The Cβ moves 0.5 Å, the Cγ atoms each move 2.6 Å, and the backbone shifts slightly. After this and other changes, some much larger, refinement produced the map shown in part (b) and the coordinates deposited in the PDB as 1TOV.

Ramachandran-plot criteria can be used in ways similar to side-chain rotamericity, although making the corrections is somewhat harder. In order to distinguish more cleanly between unusual but possible ϕ,ψ conformations and really worrisome local outliers, the MolProbity Ramachandran criteria have been updated, filtered by resolution and backbone B-factor, smoothed with a density-dependent algorithm, and separated into Gly, Pro, pre-Pro, and general cases [6]. The 'favored' region encompasses 98% of the high-quality data, for all four cases. Forty percent of the general-case plot area now encompasses 99.95% of the high-quality data (the dividing line between allowed and outlier), so that only 1 residue in 2000 should be a valid outlier. The SECSG working set attains the ideal average of 0.05% ϕ,ψ outliers, while the control set is 10-fold higher at 0.54% outliers. The working set has 98.3% favored residues, while the control set has 96.9% favored and the PDB sample 95.9% (Table 2). Figure 1c plots the Ramachandran outlier percentage (for all residue types) vs. resolution, for all three datasets. For the quadratic fit parameters of Ramachandran outlier percentage, the working set is 10-fold better than the control set and 15-fold better than the PDB sample. The control and PDB structures only had the benefit of traditional Ramachandran criteria that are either less accurate because unfiltered and very old [3], or are accurate but less discriminating [43].

Cβ deviations from ideal can be very useful indicators of local fitting problems, and any >0.25 Å were checked in this work. However, like bond length and bond angle deviations, they are not a good measure of overall structural accuracy. At low resolution they can be constrained ideal if desired and at very high resolution they are sometimes poor because geometrical terms are downweighted, while at intermediate resolutions the overall values are more sensitive to methodological choices than to conformational accuracy. Large Cβ deviations in the working set of structures were therefore examined for potential sidechain misfittings, but their statistics are not tabulated here.

Table 2. Summary statistics and quality scores for the three datasets of crystal structures.

Parameter	Working set	Control set	PDB sample
No. of structures	29	19	1784
Resolution (avg.)	1.91 Å	1.97 Å	1.86 Å
Resolution (median)	1.90 Å	1.80 Å	1.80 Å
R_{work}	20.1%	20.9%	
R_{free}	23.4%	24.4%	
RMS bond angle	1.26°	1.34°	
Real-space R	12.8%	13.4%	
Asn/Gln/His flips	0.0%	18.9%	19.3%
Non-rotamericity	1.2%	3.7%	6.04%
Rama outliers	0.05%	0.54%	0.68%
Rama favored	98.3%	96.9%	95.9%
Clashscore	3.16	17.6	18.5

The most powerful new criterion for identifying and repairing local misfittings is all-atom contact analysis [36], which relies upon adding and optimizing all hydrogen atoms and then analyzing the steric and H-bond interactions of the complete set of atoms. H-bonds, or their absence, are very useful flags, but the primary problem indicators are serious steric clashes (that is, overlap >0.4 Å of non-H-bonding atoms). The clashscore (number of serious clashes per 1000 atoms) gives an overall evaluation, while lists or 3D displays of the individual clashes direct attention and guide rebuilding in specific directions to eliminate them.

The average clashscore for the working-set structures is 3.16 while for the control set it is 17.63 (see Table 2), a 5.6-fold improvement. Figure 1a plots clashscore vs. resolution for the three datasets. The quadratic fit parameter for the working set is 5.4-fold better than for the control set and 6-fold better than for the PDB sample. Rotamer and Ramachandran criteria only flag the specific problems for which they were designed, but all-atom clashes are very general indicators. As well as usually being present at side-chain (Figure 2a) or backbone misfittings [6,33], in this study they also were found to flag ligand misorientations, *cis* vs. *trans* peptides [44], incorrect switches between backbone and side-chain near chain termini, and places where a 'water' had been built into density that was really an ion, an alternate conformation [33], or a noise peak. The 'water' in Figure 2a clashes with non-polar methyl groups on all sides; it was presumably placed in a difference peak caused by the backwards valine, and there is no remaining density or difference density there at all in the final maps. All-atom contacts were also used in the positive direction, either to validate correct but unusual conformations [6] or to fill in missing parts of the model: when density was weak or ambiguous but excellent contacts could be achieved in just one good rotamer, then refinement nearly always confirmed that new position with improved density and lower R and R_{free}. As with any successful model rebuilding, the interpretability of electron density often improved in distant regions as well, after these procedures.

A requirement for accepting the validity of these structure corrections is that they should maintain or improve R_{work} and especially R_{free}, which are respectively the refined-set and the reserved-set agreement between model-calculated and experimental data amplitudes [11]. Figure 1d plots R_{free} vs. resolution for the three datasets, with a polynomial fit to the PDB-sample data. Both working and control sets are slightly but not significantly better in general than the PDB sample, thus evaluated overall as comparable. Average R_{free} for the working set is 23.4%, which is 1% below the control-set average of 24.4%. Average R_{work} is 20.1% for the working set and 20.9% for the control set (Table 2), down by 0.8%. Those differences are consistent with individual changes that occurred during refitting in the 6 early working-set examples for which a direct comparison can be made, since in those cases the MolProbity procedures were applied to pre-existing structures that had already been refined to convergence. Those R_{free} values improved typically about 1%, with a range of 0–4%, and R_{work} improved slightly less. Another measure of model-to-data correlation is the real-space residual, available from the Electron Density Server; that value averages 12.8% for the working set and 13.4% for the control set. It is thus clear that the order-of-magnitude improvements achieved in this study for clash, rotamer, and Ramachandran scores have not been at the expense of R_{free}, R_{work}, or real-space R. In fact all measures have improved somewhat (Table 2), so that these protocols win by all criteria, both new and traditional.

Discussion

We conclude that both the intervention to identify and correct problems in the working set of SECSG crystal structures and also the controlled comparison to evaluate that intervention were successful. The comparison shows that the structures using MolProbity-based protocols reduced their outliers for Ramachandran, sidechain rotamer, and all-atom clashes by dramatic factors of 5- to 10-fold while maintaining or improving all traditional crystallographic criteria. The value of making these changes should be judged by improvement in the local fit to the electron density, the quality of that density and difference density, and the improvement in structural and biological plausibility and significance of the new local conformations, which in most cases has changed atom positions and/or dihedral angles by large amounts (e.g., for the Val in Figure 2). The variety, categorization, and refinement history of other specific examples from this study will be described elsewhere, but many such similar improvements from other structures have been described in previous publications [6, 7, 33, 37, 44, 45]. When such changes are in functionally important regions their significance is unassailable. However, their importance can be very strongly argued anywhere in a structure for two reasons: first, the Fourier transform relationship makes increased accuracy anywhere improve interpretability everywhere else, and second, when a structure is initially solved one cannot predict what regions may turn out to be important to later-appreciated functions. Endless agonizing over the details cannot be justified in general, but large improvements from modest investments, such as the present techniques, are very clearly worthwhile.

The most independent and most sensitive part of this new method for finding and fixing local problems is all-atom contact analysis [36], which relies upon adding and optimizing hydrogen atoms and then analyzing the steric and H-bond interactions of the complete set of atoms. The analysis is geometrical rather than energy-based, which for this application has the two advantages of not over-reacting to model errors that produce severe overlaps, and of directly evaluating the screened contacts between atomic surfaces rather than using pairwise interactions defined between atomic centers. The filtered, high-quality empirical data on which our all-atom contacts and other scoring functions are based appear to be a close match to high-level quantum calculations where those are feasible, although less close to the results from molecular-mechanics force fields [6, 41, 46]. These methods should be reliable as a guide to structure improvement, and that has proven to be true for this study.

Serious clashes are very rare in the well-ordered parts of high-resolution structures, but are more prevalent than one would like elsewhere. The plot of Figure 1a shows that at a respectable resolution of 2 Å the average clashscore is about 20. Some bad clashes cluster and include common atoms, but each clash involves two atoms; at 2 Å resolution this means that on average nearly one atom in 25 has a physically unreasonable steric overlap indicating a local error in the model. The present study shows that it is feasible to correct 80–85% of those model errors in the 1.5–2.5 Å resolution range, at the same time producing a modest improvement in crystallographic parameters. At lower resolutions, it is extremely helpful to restrict model fitting to good sidechain rotamers and allowed Ramachandran angles, in the same spirit as giving increased weight to covalent geometry. Steric clashes can be used to distinguish between alternative conformations (for instance, for a ligand). It may eventually prove possible to achieve accurate models at 3–3.5 Å resolution, but that will probably require both explicit H atom contacts in the refinement calculation and a combinatorial treatment of interacting residues.

With the accurate, filtered empirical distributions now available, it is clear that most outlier residues by the Ramachandran or sidechain rotamer criteria represent fitting errors. Extrapolation of dihedral-angle distributions to low crystallographic *B*-factors or to high resolution show that the intrinsic variability of most sidechain χ angles is quite small, with an RMSD of about 9° [41]. Real rotamer outliers do definitely occur, but when they are unambiguously required by clear electron density there are always multiple H-bonds or tight packing interactions that hold the conformation far away from the rotameric local energy minimum. Partially disordered surface sidechains, in particular, cannot justifiably be fit as non-rotameric. Corrections that only change torsion angles slightly would seldom matter very much and would usually not survive later refinement, but the frequent outliers caused by some sort of backwards or switched fitting of the model are always worth correcting.

An important question is the desirable strategy and timing for making best use of this new information, since there is an inherent conflict between on one hand keeping these criteria out of the refinement process so as to maximize their power as independent validation criteria, and on the other hand maximizing their utility for actual improvement of the structures. After the extensive experience of this test study, we can say that early use and tight integration into the refinement/rebuilding process (so that the validation criteria effectively become additional restraints) promotes smoother convergence and more accurate final results than waiting until the end, presumably by avoiding incorrect local minima and their resultant model bias effects. Although criteria are weakened as validation measures if less independent, they are still very useful because it is not in fact a trivial process to satisfy them, and in particular it is difficult to satisfy them without actually making the model better. The best answer to this conflict is presumably some degree of compromise. However, as users of these structures, we decidedly value improved accuracy over improved knowledge about the degree of accuracy.

In a comparison between datasets, there are issues of statistical sample size and there are possibilities of bias due to systematic factors. The latter are typically the more worrisome issue in a complex system, and in this case would include factors such as the date, the size, the non-crystallographic symmetry, the refinement targets, the number, effort, and experience of the crystallographers, and the extent to which weakly diffracting regions were omitted from the coordinates. The systematic factor of deposition date turns out to produce no effect whatsoever on clashscore or Asn/Gln/His flips and only a small change in ϕ,ψ criteria, but it does help produce the larger difference of control from PDB structures in rotamericity. We used two comparison sets because the PDB sample is large enough for good statistics, while the SECSG control set has relatively few systematic differences from the working set. In interpreting the results of this study, the direction but certainly not the magnitude is reliable for the modest differences between PDB sample and control set in the plots of Figure 1a–c. However, the primary contrast tested here, between the MolProbity(+)

working set and the others, is a consistent, order-of-magnitude change. In clashscore, only one of the control structures and less than 1% of the PDB sample reach the midline fit of the working set, so that the distributions are almost non-overlapping. The improved accuracy achieved here is dramatic, and it would be highly desirable to work toward making such procedures a standard part of crystallographic practice.

Acknowledgements

This work was supported by National Institutes of Health PSI-1 Grant P50-62407 to the SECSG. Synchrotron use was supported by the U.S. Department of Energy, Office of Basic Energy Sciences, contract W-31-109-Eng-38. Institutions supporting the Southeast Collaborative Access Team (SER-CAT, Advanced Photon Source sector 22) are listed at www.ser-cat.org/members.html. Earlier development of the MolProbity tools was supported by National Institutes of Health Grants GM-61302 and GM-15000, and by a Howard Hughes Medical Institute graduate fellowship to IWD.

References

1. Mowbray, S.L., Helgstrand, C., Sigrell, J.A., Cameron, A.D. and Jones, T.A. (1999) *Acta Cryst. D* **55**, 1309–1319.
2. Kleywegt, G.J. (1999) *Acta Cryst. D* **55**, 1878–1884.
3. Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) *J. Appl. Crystallogr.* **26**, 283–291.
4. Vriend, G. (1990) *J. Mol. Graph.* **8**, 52–56.
5. Jones, T.A., Zou, J.-Y., Cowan, S.W. and Kjeldgaard, M. (1991) *Acta Cryst. A* **47**, 110–119.
6. Lovell, S.C., Davis, I.W., Arendall, W.B. III, de Bakker, P.I.W., Word, J.M., Prisant, M.G., Richardson, J.S. and Richardson, D.C. (2003) *Proteins* **50**, 437–450.
7. Davis, I.W., Murray, L.W., Richardson, J.S. and Richardson, D.C. (2004) *Nucleic Acids Res.* **32**(Web Server Issue), W615–W619.
8. Vaguine, A.A., Richelle, J. and Wodak, S.J. (1999) *Acta Cryst. D* **55**, 191–205.
9. Kleywegt, G.J., Harris, M.R., Zou, J.Y., Taylor, T.C., Wahlby, A. and Jones, T.A. (2004) *Acta Cryst. D* **60**, 2240–2249.
10. van den Akker, F. and Hol, W.G.J. (1999) *Acta Cryst. D* **55**, 206–218.
11. Brunger, A.T. (1992) *Nature* **355**, 472–475.
12. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) *Nucleic Acids Res.* **28**, 235–242.
13. Norvell, J.C. and Machalek, A.Z. (2000) *Nat. Struct. Biol.* **7**, 931.
14. Dauter, Z. (2002) *Acta Cryst D* **58**, 1958–1967.
15. Fu, Z.-Q., Rose, J.P. and Wang, B.-C. (2004) *Acta Cryst. D* **60**, 499–506.
16. Adams, P.D., Grosse-Kunstleve, R.W., Hung, L.-W., Ioerger, T.R., McCoy, A.J., Moriarty, N.W., Read, R.J., Sacchettini, J.C., Sauter, N.K. and Terwilliger, T.C. (2002) *Acta Cryst. D* **58**, 1948–1954.
17. Badger, J. and Hendle, J. (2002) *Acta Cryst. D* **58**, 284–291.
18. Lesley, S.A., Kuhn, P., Godzik, A., Deacon, A.M., Mathews, I., Kreusch, A., Spraggon, G., Klock, H.E., McMullan, D., Shin, T., Vincent, J., Robb, A., Brinen, L.S., Miller, M.D., McPhillips, T.M., Miller, M.A., Scheibe, D., Canaves, J.M., Guda, C., Jaroszewski, L., Selby, T.L., Elsliger, M.-A., Wooley, J., Taylor, S.S., Hodgson, K.O., Wilson, I.A., Schultz, P.G. and Stevens, R.C. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 11664–11669.
19. Montelione, G.T., Zheng, D., Huang, Y.J., Gunsalus, K.C. and Szyperski, T. (2000) *Nat. Struct. Biol.* **7**, 982–985.
20. Huang, Y.J., Powers, R. and Montelione, G.T. (2005) *J. Am. Chem. Soc.*, **127**, 1665–1674.
21. Adams, M.W.W., Dailey, H.A., Delucas, L.J., Luo, M., Prestegard, J.H., Rose, J.P. and Wang, B.-C. (2003) *Acc. Chem. Res.* **36**, 191–198.
22. Terwilliger, T.C. and Berendzen, J. (1999) *Acta Cryst. D* **55**, 849–861.
23. Schneider, T.R. and Sheldrick, G.M. (2002) *Acta Cryst. D* **58**, 1772–1779.
24. Wang, B.-C. (1985) *Methods Enzymol.* **115**, 90–112.
25. Collaborative Computational Project, No. 4 (1994) *Acta Cryst. D* **50**, 760–763.
26. Terwilliger, T.C. (2002) *Acta Cryst. D* **58**, 1937–1940.
27. Perrakis, A., Morris, R. and Lamzin, V.S. (1999) *Nat. Struct. Biol.* **6**, 458–463.
28. Murshudov, G.N., Vagin, A.A. and Dodson, E.J. (1997) *Acta Cryst. D* **53**, 240–255.
29. Potterton, E., Briggs, P., Turkenburg, M. and Dodson, E.J. (2003) *Acta Cryst. D* **59**, 1131–1137.
30. Winn, M.D. (2003) *J. Synchron. Rad.* **10**, 23–25.
31. Brunger, 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. and Warren, G.L. (1998) *Acta Cryst. D* **54**, 905–921.
32. McRee, D.E. (1999) *J. Struct. Biol.* **125**, 156–165.
33. Richardson, J.S., Arendall, W.B. III and Richardson, D.C. (2003) In *Methods in Enzymology: Macromolecular Crystallography, Pt. D* (Eds., Carter, C.W. Jr., Sweet, R.M.), Vol. 374 Academic Press, New York, pp. 385–412.
34. Lovell, S.C., Word, J.M., Richardson, J.S. and Richardson, D.C. (2000) *Proteins* **40**, 389–408.
35. Word, J.M., Lovell, S.C., Richardson, J.S. and Richardson, D.C. (1999) *J. Mol. Biol.* **285**, 1735–1747.
36. Word, J.M., Lovell, S.C., LaBean, T.H., Taylor, H.C., Zalis, M.E., Presley, B.K., Richardson, J.S. and Richardson, D.C. (1999) *J. Mol. Biol.* **285**, 1711–1733.
37. Richardson, J.S. (2003) In *Structural Bioinformatics* (Eds., Bourne, P.E., Weissig, H.), John Wiley & Sons, Inc., New York, pp. 305–320.

38. Word, J.M., Bateman, R.C. Jr., Presley, B.K., Lovell, S.C. and Richardson, D.C. (2000) *Protein Sci.* **9**, 2251–2259.
39. Sack, J.S. (1988) *J. Mol. Graph.* **6**, 224–225.
40. Murzin, A., Brenner, S.E., Hubband, T. and Chothia, C. (1995) *J. Mol. Biol.* **247**, 536–540.
41. Butterfoss, G., Richardson, J.S. and Hermans, J. (2005) *Acta Cryst. D* **61**, 88–98.
42. Higman, V.A., Boyd, J., Smith, L.J. and Redfield, C. (2004) *J. Biomol. NMR* **30**, 327–346.
43. Kleywegt, G.J. and Jones, T.A. (1996) *Structure* **4**, 1395–1400.
44. Videau, L.L., Arendall, W.B. III and Richardson, J.S. (2004) *Proteins* **56**, 298–309.
45. Richardson, J.S. and Richardson, D.C. (2003) In *Computational Biology and Genome Informatics* (Eds., Wang, J.T.L., Wu, C.H., Wang, P.P.), World Scientific Publishing Company, London, pp. 139–161.
46. Hu, H., Elstner, M. and Hermans, J. (2003) *Proteins* **50**, 451–463.