Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Reliable quality-control methods for protein crystal structures

John Badger* and Jörg Hendle

Structural GenomiX, 10505 Roselle Street, San Diego, California 92121, USA

Correspondence e-mail: john_badger@stromix.com

Received 15 August 2001
Accepted 22 November 2001

The emergence of structure-determination initiatives that employ high-throughput protein crystallography emphasizes the need to establish quality-control methods for screening the resulting models prior to deposition with the public data banks. An in-house database of 26 new protein structures, associated diffraction data and high-quality experimentally determined electron-density maps have been used to develop (i) a set of minimal global quality criteria that a structure must meet before the refinement may be considered completed and (ii) a reliable set of indicators for detecting local errors in protein structures. These criteria have been applied to detecting local errors to a set of structures recently deposited in the Protein Data Bank and it is estimated that about 3% of amino acids are incorrectly modeled.

1. Introduction

Over the last few years, the number of protein structures solved and deposited in the Protein Data Bank (Bernstein *et al.*, 1977; Berman *et al.*, 2000) has risen sharply. During this period, the available software for macromolecular crystallography has evolved to a point where many structures are being solved semi-automatically and by less trained workers than in the past. In addition to the structures solved within individual laboratories, the emergence of several public and privately funded structural genomics projects is expected to greatly increase the number of solved structures deposited within the Protein Data Bank over the next few years (Norvell & Machalek, 2000). Many of these structures will be solved using highly automated methods and without the dedicated attention of a particular crystallographer. One need brought into focus by these initiatives is the definition of a set of well defined conditions under which a structure refinement may be considered completed. Despite the pressure to produce as many new structures as possible, these initiatives also offer new opportunities for systematically screening all of the resulting structures through a set of validation tests to eliminate errors prior to deposition. In short, establishing automated and effective structure-validation systems is critical for the success of high-throughput protein crystallography.

Most of the currently available software for structure validation was developed in the 1990s as a result of a need to develop standardized validation tests outside the information provided by the specific software used for refinement and model building. Impetus for the development of several validation methods (Lüthy *et al.*, 1992) was provided by the discovery of major errors in several published structures (Brändén & Jones, 1990). At the same time, crystallographic model-building methods were improved through the use of

real-space evaluations for density-map fitting (Jones *et al.*, 1991) and refinement targets became more objective through cross-validation tests (the $R_{\rm free}$ index; Brünger, 1992) and the implementation of maximum-likelihood methodology (Bricogne & Irwin, 1996; Murshudov *et al.*, 1997; Pannu & Read, 1996; Pannu *et al.*, 1998). Widely used structure-validation software developed during this period includes *PROCHECK* (Morris *et al.*, 1992) and *WHAT CHECK* (Vriend, 1990) for evaluating quantities related to protein modeling and stereochemistry, and the *SFCHECK* program (Vaguine *et al.*, 1999) for evaluating the agreement of the model with the diffraction data.

The availability of these (and other) programs solves the practical problem of computing most of the commonly used validation indices (Kleywegt, 2000), but relatively few definitive conclusions appear to have been reached regarding the usefulness of the various possible validation criteria or the specific thresholds that would indicate structural errors with a high degree of certainty. A noteworthy exception is a systematic study by Carson *et al.* (1994) using five validation criteria (temperature factor, real-space fit, geometric strain, dihedral angle value and shift from previous refinement cycle) and a data set of six structures. This work provided some useful information on the detection of gross structural errors but also revealed very different levels of usefulness in these criteria. The EU 3-D Validation Network (1998) provided case studies of eight structures refined using high-resolution diffraction data, with an emphasis on analyzing the stereochemical aspects of these structures *versus* dictionary values and refinement weights.

This paper describes an analysis of structure-quality metrics for a database containing 26 new protein structures, their associated diffraction data and high-quality electron-density maps. The results of this analysis provide a set of structure-quality indices and threshold values that we believe a fully refined protein structure should be able to satisfy. We also emphasize the importance of local structural errors in determining structure quality by applying a testing concept: the position of each amino acid in these structures is evaluated using a variety of criteria and then scored as to whether it is appears correct or not. We have applied these tests to a set of structures recently deposited in the Protein Data Bank to evaluate the numbers and types of errors that occur most frequently.

2. Materials and methods

2.1. The crystallographic database

The structure database used to define the error-validation criteria reported in this paper comprised 26 structures containing a total of 14 623 amino acids and ligands. These are the first 26 structures solved by our in-house structure determination group and therefore represent a new and independent training set of data for the development of quality-control criteria. These structures were refined at resolutions between 2.9 and 1.34 Å, with resolutions exceeding 2.3 Å in 19 cases. High-quality experimentally determined electron-density maps were available for 19 structures. These maps were obtained from SAD or MAD data from Se-Met crystals with initial phasing computations carried out by *SHARP* (de La Fortelle & Bricogne, 1997) and improved by density modification using *SOLOMON* (Abrahams & Leslie, 1996) or *DM* (Cowtan, 1994). The average phase difference between experimental phase sets and phases computed from the final models was 39° and the average correlation coefficient between the experimentally phased electron-density maps used for the initial model-building and maps computed from the final models was 0.81. The difference between experimental and model phases was greater than 50° for two structures and the correlation coefficient between the experimental and model map was less than 0.70 for one structure. Structures in this database were refined using amplitude or intensity maximum-likelihood targets with either the *CNX* program (Brünger *et al.*, 1998; Molecular Simulations Inc., 2000) or the *CCP*4/*REFMAC* program (Murshudov *et al.*, 1997). Six different crystallographers, originally trained in different laboratories with different conceptions of structure validation and refinement, produced this initial set of 26 structures.

2.2. Validation software

We have developed a single automated structure-validation script that runs *PROCHECK* (Morris *et al.*, 1992), *WHAT CHECK* (Vriend, 1990), *SFCHECK* (Vaguine *et al.*, 1999), *CCP*4/*PHISTATS* and *CCP*4/*OVERLAPMAP* (Collaborative Computational Project, Number 4, 1994) in addition to carrying out the other functions (*CCP*4 software) for writing out σ_A-weighted Fourier coefficients (Read, 1986) and experimentally phased structure factors for the display of electron-density maps with the *XtalView*/*Xfit* program (McRee, 1999). The only required inputs for this validation script are a coordinate file (PDB format) and a diffraction data file (*CCP*4/MTZ format). Additional functions of this script are to check the amino-acid sequence given in the structure file against a sequence file, output a standardized PDB coordinate file and output an mmCIF file of the diffraction data used for refinement. This validation script also runs software that parses the resulting outputs from these programs into a single 'crystallographic validation file' that we have designed by making minor local extensions to the mmCIF dictionary (Bourne *et al.*, 1997). Thus, all quantities relating to structure validation are calculated by the same method and are recorded in a consistent and easily searchable format. Furthermore, all data files and phase sets are created in a form that is convenient for inspection.

The structure-validation software was run on 26 structures prior to being uploaded into our database. For those programs that provide specific error indications (for example, amino acids in the disallowed regions of the Ramachandran plot), the structure was checked at these positions against the experimentally obtained electron-density maps and the σ_A-weighted versions of $2F_{\rm obs} - F_{\rm calc}$ and $F_{\rm obs} - F_{\rm calc}$ maps. Checks against

Table 1
Mean values for the global structure-validation criteria for the 26 structures in our in-house crystallographic database.

Where two values are given, these correspond to the sets of structures refined using data above and below 2.3 Å resolution, respectively.

R_{work}	0.203/0.223
Difference between R_{work} and R_{free}	0.043
Residues in Ramachandran core regions (%)	91.3
No. of close contacts per 100 residues	0.2/1.0
Residues with abnormal $\chi1$–$\chi2$ angles (%)	0.6/0.8

the maps calculated with experimental phases were especially valuable in providing unbiased information on the reliability and usefulness of the various validation tests.

As a result of the work described in this paper, the validation script was modified to parse information from the various output files into an additional file that lists each amino acid that appears to be in error, including the validation test that it failed.

The structure-analysis script with associated parsing programs is available from john_badger@stromix.com. This software runs under Linux operating systems and requires access to *PROCHECK* 3.5, *SFCHECK* 5.3.4 (both included in *CCP*4 4.0) and *WHAT CHECK* 4.99 installations.

3. Results and discussion

3.1. A standard for global measures of structure quality

To avoid subjectivity when deciding whether a structure is sufficiently refined for the structure determination to be considered finished, specific minimal required values for global measures of structure quality must be defined. These measures provide a set of 'gatekeeper' criteria that should not be exceeded for an acceptable structure, so they are purposely more relaxed than the typical values that are obtained for well refined structures for which there is high-resolution data (Table 1).

By examining the validation statistics for the structures in our database, we found that the following values provide a minimal level of structure quality that is both achievable (all structures in this database meet this standard) and which leads to stereochemical quality values better than earlier expectations based on structures in the PDB (Morris *et al.*, 1992).

(i) As calculated by *SFCHECK*, R_{work} must be lower than 0.225 if the resolution is higher than 2.3 Å; the *R* factor for all data must be lower than 0.250 if the resolution is lower than 2.3 Å.

(ii) As calculated by *SFCHECK*, the difference between *R* values for the working data and the test subset of data must not exceed 0.08. This threshold is set relatively high to accommodate structures solved at low resolution.

(iii) As calculated by *PROCHECK*, at least 88% of amino acids should lie in the core region (A, B, L) of the Ramachandran plot. We have focused on this core region, which covers only 11% of the Ramachandran plot (Morris *et al.*, 1992), rather than the more extended 'allowed' region in order to be consistent with more recent studies that show a tightly confined range of preferred φ–ψ angles (Kleywegt & Jones, 1996).

(iv) As calculated by *PROCHECK*, there should not be more than one abnormally close van der Waals contact per 100 residues if the resolution is greater than 2.3 Å; the number of abnormally close contacts should not exceed four contacts in 100 residues if the resolution is lower than 2.3 Å.

(v) As calculated by *PROCHECK*, there should be fewer than 2.0% of possible residues flagged with abnormal (3σ threshold) $\chi1$–$\chi2$ angles if the resolution is greater than 2.3 Å and fewer then 3% of residues flagged with abnormal $\chi1$–$\chi2$ angles if the resolution is lower than 2.3 Å.

As a further demonstration that this is a practical standard that should be achievable for fully refined structures, 34 additional structures have recently been added to our database and all structures refined at resolution better than 2.8 Å have passed these quality criteria.

The resolution dependence included in some of these criteria reflects the fact that it is usually much easier to obtain good validation statistics for structures refined with atomic resolution data. We selected 2.3 Å as the resolution threshold for some of these structure-quality metrics because this is the point at which atomic detail begins to emerge and the average numbers of close contacts show a significant increase at lower resolution. Nevertheless, where resolution-dependent criteria are necessary, it might be better to establish more smoothly varying resolution-dependent values. Work on a much larger set of structures (in progress) should be able to provide these.

3.2. Reliability of conventional measures of structure quality

Published reports of macromolecular structures generally provide a summary table of structure-determination details containing the *R* value, R_{free} and the root-mean-square deviations from ideality for covalent bond lengths or angles. Although it is obviously necessary that the values for these statistics lie within a suitable range, it may be asked how much information these numbers provide on the accuracy of the model or the number of local errors the model contains.

We can approach these questions using the 26 structures in our database at the time of the study, since the *R* value and R_{free} are calculated in identical fashion for all structures. This avoids the numerical variations arising from differing data cutoffs, scaling methods and bulk-solvent corrections that appear when trying to make detailed comparisons using the values quoted in files obtained from the Protein Data Bank (Weissig & Bourne, 1999). Similarly, all stereochemical values are calculated using the same dictionary.

We have used the diffraction precision index (DPI) (Cruickshank, 1999), which is a function of the resolution, the number of data, the number of parameters and the *R* factor, to estimate the absolute accuracy of the protein model. Perhaps unsurprisingly, both the resolution and the *R* factor correlate quite well with this index (0.85 and 0.80; Table 2). The root-mean-square deviations from ideality for bond lengths and

bond angles appear unrelated to both the estimate for the overall accuracy of a structure and the number of local errors that it contains. In fact, none of the global measures for structural quality is well correlated with the percentage of probable errors, highlighting the need for additional quality measures that are more focused on detecting local structural problems.

3.3. Detection of local structural errors

In addition to satisfying criteria relating to the overall quality of a structure, it is also necessary to ensure that all detectable local errors are eliminated from the model before the structure determination is considered completed. In selecting the best measures to use for locating errors in protein structures, we have attempted to choose criteria that list structural errors that are both significant and correctable. A 'significant error' is an error that requires a major rearrangement of atoms to correct and which might possibly result in a change of functional interpretation of the structure. For example, a rotamer error in a side-chain conformation would constitute a significant error but a slightly misplaced atom with a covalent bond length stretched by four standard deviations from its dictionary value would not.

We find that a useful set of indicators for probable local structural errors is as follows.

(i) As calculated by *SFCHECK*, an electron-density correlation coefficient for the main-chain atoms within a residue of less than 0.85 (0.80 if the resolution is worse than 2.3 Å).

(ii) As calculated by *SFCHECK*, an electron-density correlation coefficient for the side-chain atoms within a residue of less than 0.80 (0.65 if the resolution is worse than 2.3 Å).

(iii) As calculated by *PROCHECK*, a covalent bond and angle more than six standard deviations from the ideal value.

(iv) As calculated by *PROCHECK*, a residue in a disallowed region of the Ramachandran plot.

(v) As calculated by *WHAT CHECK*, a side-chain rotation for an asparagine, glutamine or histidine residue needed to optimize the hydrogen-bonding network.

(vi) As calculated by *WHAT CHECK*, a packing abnormality extending over a tripeptide.

Programs that use different calculation methods or different dictionary values might require different threshold values or definitions to these.

The extent to which these tests for probable errors leads to incorrect error indications (*i.e.* false positives, where an error is indicated but the model is correct) may be estimated by reference to Table 3. This table shows for each structure the number of probable errors indications that remained after each flagged residue had been visually checked against the available electron-density maps in the context of its crystal environment and rebuilt if necessary. Most significantly, the numbers of false error indications for the final structures are very small (79) compared with the total numbers of amino acids in these structures (14 623), leading to an average false positive error rate of 0.5%. As a practical note: at the point in time when these structure refinements were initially considered 'completed', but prior to refitting using information from this validation system, the average rate of probable errors was 2.7%. Therefore, these tests typically identified several additional amino acids per structure as being in error, and the average ratio of correctly identified errors to false error indications was slightly better than 5:1. In more recent structure determinations we have actively begun to apply these structure diagnostics during the final stages of structure

Table 2
Correlations of global structure-validation criteria (*R* factor, R_{free}, r.m.s. deviations in bond lengths and bond angles from ideality) and resolution with the structure accuracy (measured by the DPI) and the percentage number of probable local errors.

This calculation was based on the 26 structures in our in-house crystallographic database. Note that this calculation of the DPI by *SFCHECK* uses the *R* factor over all data, but the associated values for R_{free} and R_{work} exclude structure factors below 5 Å resolution and data for which the amplitudes are smaller than two standard deviations.

Type	DPI	No. of errors (%)
R factor (all data)	0.80	0.32
R_{free}	0.41	0.20
R_{work}	0.32	0.22
Resolution (Å)	0.85	0.34
R.m.s. deviation from ideality for bond lengths	0.14	−0.13
R.m.s. deviation from ideality for bond angles	0.01	0.08

Table 3
The number of local error indications remaining in the final set of refined structures after each residue flagged as a probable error had been examined and rebuilt if necessary.

1, Low main-chain density correlation; 2, low side-chain density correlation; 3, disallowed region of the Ramachandran plot; 4, strained main-chain covalent geometry; 5, Asp, Gln, His side-chain flip; 6, tripeptide packing error.

ID	No. amino acids	1	2	3	4	5	6
1	366	0	0	1	0	0	0
2	1076	0	1	0	0	0	0
3	749	0	1	0	0	0	0
4	161	2	1	1	0	0	1
5	937	2	5	0	0	0	1
6	302	0	0	0	0	0	0
7	323	0	0	0	0	0	0
8	305	0	0	0	0	0	0
9	751	1	0	1	0	0	2
10	151	0	1	0	0	0	0
11	152	0	0	0	0	0	0
12	727	4	4	0	0	0	0
13	697	2	2	0	0	0	1
14	824	1	3	3	0	0	0
15	403	0	0	0	0	0	0
16	405	1	0	1	0	0	0
17	276	0	1	0	0	0	0
18	210	0	0	2	0	0	0
19	251	0	0	2	0	0	0
20	921	2	0	0	0	0	0
21	147	0	0	0	0	0	1
22	276	1	1	0	0	0	0
23	424	1	1	0	0	0	0
24	922	3	4	0	0	0	0
25	1273	2	0	6	0	1	0
26	1234	0	3	0	0	2	3

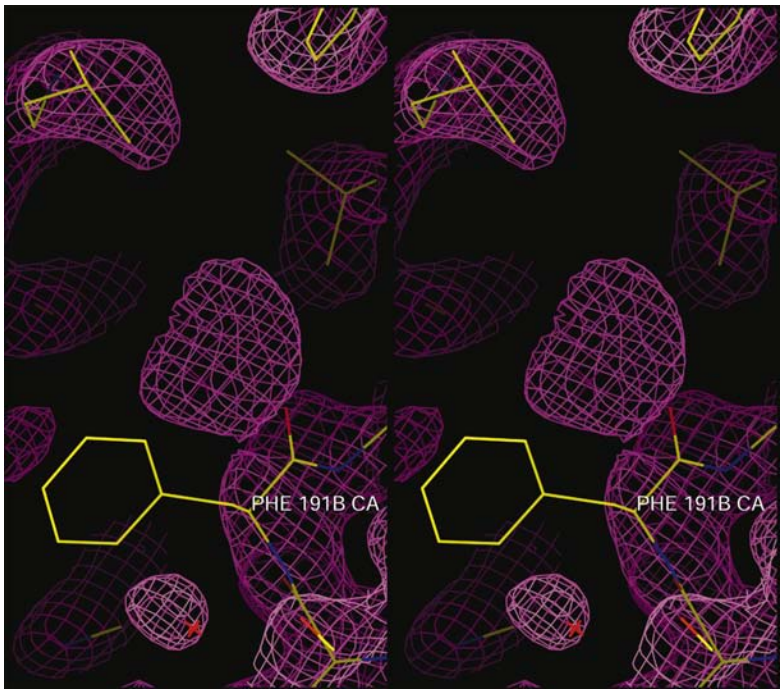

Figure 1
Portion of a protein where the main chain of residue Phe191*B* misfits the density and the side chain is incorrectly placed. These errors are identified by density correlation coefficients of 0.74 for the main-chain atoms and −0.08 for the side-chain atoms. A correct placement of the Phe aromatic ring would fit the flattened density feature in the center of the image. The experimentally determined electron-density map (purple contours) resulted from MAD phasing at 2.0 Å resolution and is contoured at the 1σ level.

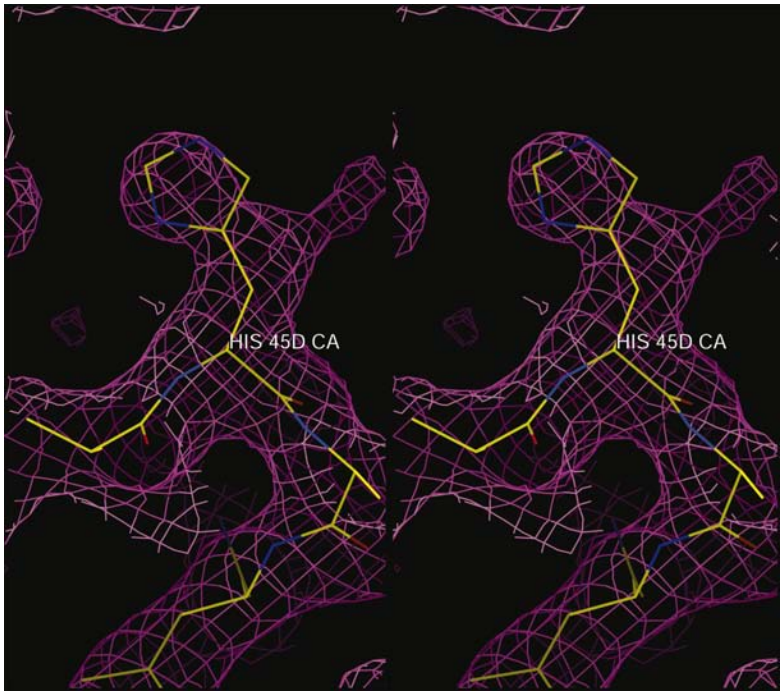

Figure 2
Portion of a protein where a relatively low value for the density correlation of the main-chain atoms (0.77) of residue His45*D* flags a probable error but visual inspection of the electron-density map indicates that the fit is probably correct. The low value for the density-correlation coefficient may arise from discrete disordering of the His side chain that is not represented by the model. The experimentally determined electron-density map (purple contours) resulted from MAD phasing at 2.35 Å resolution and is contoured at the 1σ level.

refinement, when the number of probable errors is often initially two to three times larger. In particular, we note that structures that have been automatically built by the *ARP*/*wARP* system (Perrakis *et al.*, 1999) are of generally very high quality but usually require deletion of disordered side chains and re-orientation of many asparagines, glutamine and histidine side chains to optimize hydrogen-bonding patterns.

The electron-density correlation metric identifies atomic groups that are either disordered or clearly misfit the density (Fig. 1) and the residual error indications for electron-density correlations (Table 3) were invariably minor violations of our current thresholds (Fig. 2). Amino acids in the disallowed regions of the Ramachandran plot were justified only in a few cases where the main-chain conformation was stabilized by hydrogen bonds and significant electron density supported the abnormal conformation (Fig. 3). We did not encounter any cases where severe violations of the covalent geometry could be justified and these types of error frequently exposed significant model-building errors or refinement problems (for example, inappropriate restraints for *cis* prolines). Testing the the optimal orientation of asparagine, glutamine and histidine side chains with regard to hydrogen-bonding possibilities usually gives a definite indication of the preferred orientation (Fig. 4) with just a few ambiguous cases in which the optimal orientation 'flip-flops' in the final refinement cycles. Finally, the few false positive error indications in the packing-error category occurred for surface loops and sections of protein that adjoin other molecules in the crystal.

We have examined the identification of incorrectly built side chains using deviations of the observed $\chi1$–$\chi2$ angles from the expected rotamers (Morris *et al.*, 1992) as an additional test for local structural errors. For this set of refined structures an average of 1/151 amino acids contain rotamer conformations more than three standard deviations from expected values. However, it appears that conventional model-fitting procedures (which use rotamer libraries) and refinement targets (which include torsion angles amongst the restraints) largely eliminate incorrect side-chain positions from models prior to submission to the validation process. Furthermore, the side-chain density correlation test provides a check on side chains which are inconsistent with electron density. For this reason, checks on $\chi1$–$\chi2$ angles are not part of our automated validation process, but may be used as a useful supplementary measure. We also note that the *average* number of $\chi1$–$\chi2$ violations is amongst the global criteria we use for defining an adequately refined structure (*c.f.* §3.1) and this ensures that

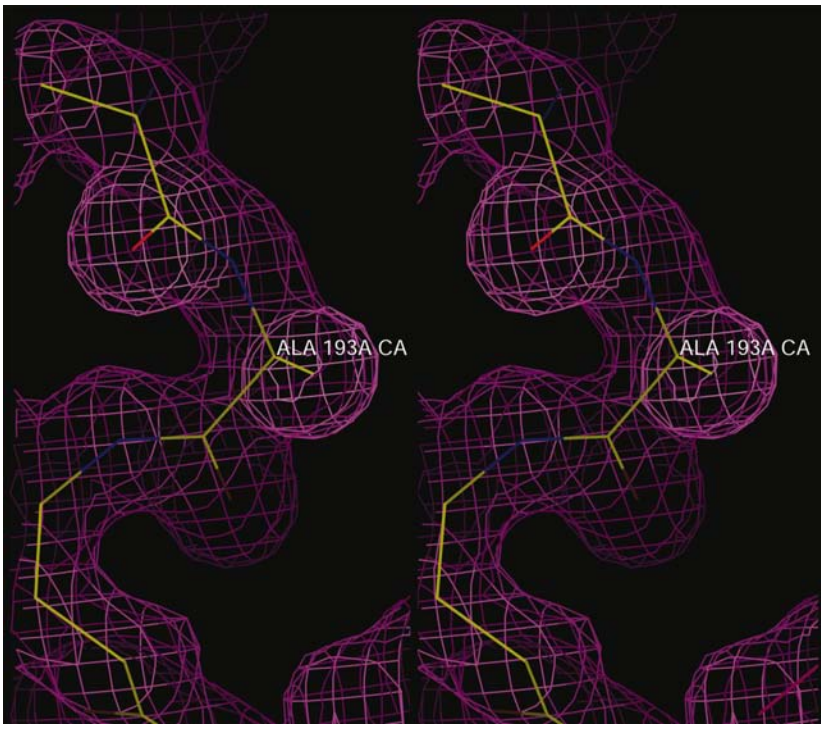

Figure 3
Example in which amino acid Ala193*A* lies in the most strongly disallowed region of the Ramachandran plot (the XX region as computed by *CCP*4/*PROCHECK*), but the unusual conformation is supported by significant electron density. The experimentally determined electron-density map (purple contours) resulted from MAD phasing at 1.7 Å resolution and is contoured at the 1σ level.

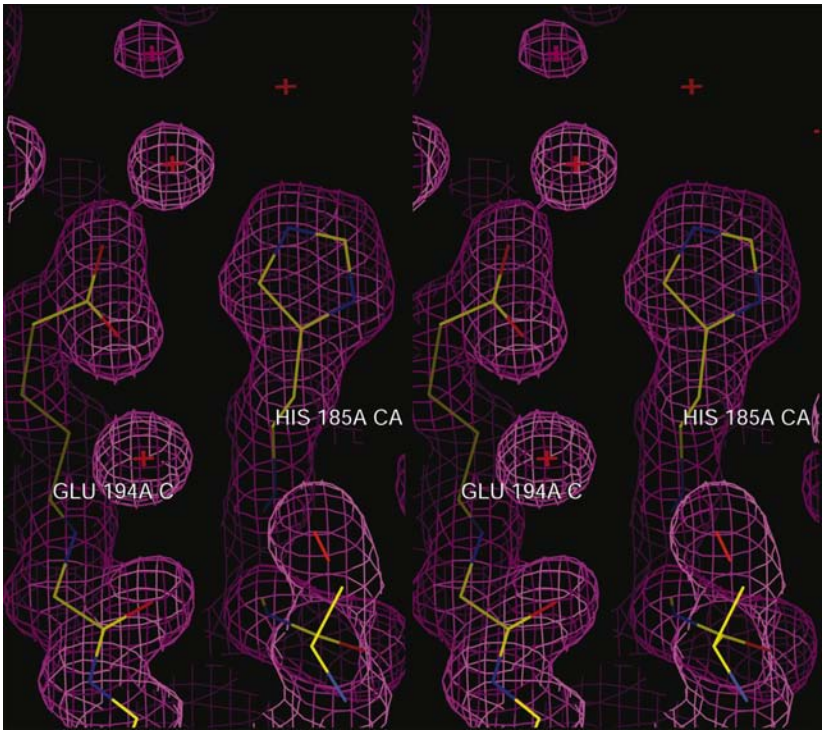

Figure 4
Example of a side-chain conformational error identified for His185*A* in which the side chain should be rotated by 180° about the χ2 angle in order to optimize hydrogen-bonding interactions with the side chain of Glu194*A*. The experimentally determined electron-density map (purple contours) was the result of MAD phasing at 1.7 Å resolution and is contoured at the 1σ level.

Table 4
Distribution of the percentage of amino acids with probable errors per structure for the 28 protein structures in the Protein Data Bank update on 6 March 2001 determined by X-ray crystallography and for which the experimental data was available.

Amino acids with probable errors (%)	No. of structures
0–1	2
1–2	4
2–3	8
3–4	2
4–5	8
5–6	3
>6	1

Table 5
No. of different types of errors found in structures selected from the Protein Data Bank based on probable errors in 374 residues over 28 structures.

The percentages total over 100% because some residues are flagged with multiple errors.

Type	No.
Main-chain density correlation	53 (14.1%)
Side-chain density correlation	113 (30.2%)
Disallowed region of Ramachandran plot	44 (11.8%)
Severe bond/angle length violations	53 (14.1%)
His/Asn/Gln, side-chain flip	142 (38.0%)
Tripeptide packing error	11 (2.9%)

there will be few side chains with abnormal torsion angles in the final structures.

An important role of this validation system is to identify structural errors that were not apparent to the crystallographer during the final stages of the structure refinement. All amino acids that obey normal stereochemical conditions and that overlap sufficiently well with the electron-density map pass the set of six tests for local errors cited above. For lower resolution structures there are ambiguities (for example, the χ1 angles of poorly ordered valines) and sub-optimal local structures (for example, large side chains with a few terminal atoms out of density) that cannot be reliably detected by the density correlation metric.

3.4. Error rates in the Protein Data Bank

To assess the quality of structures currently entering the Protein Data Bank, we retrieved all new structures released on 6 March 2001 that met the conditions that the entry (i) was a protein, (ii) was solved by X-ray diffraction and (iii) had available experimental data. This search was performed using the *SearchFields* query tool at the Protein Data Bank web site (Berman *et al.*, 2000). The resulting 28 structures

(PDB codes 1cx4, 1e3c, 1eja, 1ejj, 1ejm, 1ek3, 1ek8, 1f9c, 1fp1, 1fp2, 1fpq, 1fpx, 1fx7, 1fxm, 1g82, 1h8i, 1h96, 1ho3, 1hq8, 1hx1, 1i2e, 1i2f, 1i2g, 1i3f, 1i3i, 1i44, 1i5i and 1qhq) were then processed using our structure-validation script in order to produce validation diagnostics and lists of probable errors. For this set of structures the average error rate was found to be 3.6%. The 'best' structure contained no probable errors and the 'worst' structure had 10.6% of amino acids flagged with probable errors (Table 4).

There is considerable variation between structures as to what are the most common types of local errors, but analysis of the complete set reveals several trends (Table 5). The most common type of error is an asparagine, glutamine or histidine side chain that could be rotated by 180° to create a nearly isomorphous structure with an improved hydrogen-bond network. This type of error presumably reflects a crystallographer's blindness to local energetics once a side chain is fitted to electron density.

The second most common type of error is that some atomic groups correlate poorly with the electron-density map. In many cases this appears to be a result of including disordered side chains, for which there is no significant electron density, in the model. Poor density correlations also frequently occur in instances where a side chain has been incorrectly fitted into fragmented solvent densities and the correct side-chain density has become occupied by water molecules. In our system, the real-space density correlation is the key check between the model and the diffraction data, but some imperfections with this measure are evident and may be eventually replaced by more sophisticated approaches (Cowtan & Ten Eyck, 2000). Although it is not our own practice, we are aware that in some structures deposited with the PDB disordered side chains are included in the coordinate sets by modeling them in plausible conformations. These side chains were excluded from our analysis if the side-chain atoms had occupancy fields set to zero. Otherwise, these side chains are flagged as 'errors' if the density correlation data does not provide sufficient evidence for the modeled conformation. Our choice of the density-correlation index, rather than absolute density value, allows for acceptable placement of side chains with large temperature factors in weak density.

The other types of local error that we list (main-chain conformation in disallowed regions of the Ramachandran plot, severely strained covalent geometry and poor local packing) seem quite rare. Crystallographers may now be conscious of the usefulness of the Ramachandran plot as an error indicator, perhaps because of work by Kleywegt & Jones (1996) to popularize this index, and the availability of the *PROCHECK* program (Morris *et al.*, 1992) to calculate it. Severely strained covalent geometries are probably rare because the default weights for the refinement restraints used by the *CNS*/*CNX* program (Brünger *et al.*, 1998) for maintaining bond lengths and angles near standard values allow little deviation from ideality. However, a negative aspect of this restraint system is that overall stereochemistry for the structures refined using this program is frequently over-restrained to the target values. Severely strained geometries appear more frequently in structures refined with the *CCP*4/*REFMAC* program (Murshudov *et al.*, 1997) and these anomalies often provide a useful indication of a contradiction between the experimental data and the conformation of the model. Packing errors appear intrinsically rare and would normally indicate gross tracing errors that seldom occur for crystallographic structures.

Most of these flagged errors are easy to correct: for example, by rotating those asparagines, glutamine and histidine side chains for which the hydrogen-bonding network could be improved and deleting or refitting side chains that correlate poorly with the electron density, the number of probable errors in this set of PDB structures would be immediately reduced to ∼1.5%.

3.5. Design of automated structure-validation and deposition systems

The increasing levels of automation in the macromolecular structure-determination process (in particular, the availability of high-quality electron-density maps from anomalous scattering phasing and the development of automated model-building systems) is shifting the emphasis of the interactive aspects of structure determination to the 'finishing process', where the crystallographer completes the model by correcting minor errors, adds ligands, models discrete disorder, checks solvent atoms *etc.* Routine application of the structure-validation methods described here, together with convenient reporting of results and pre-calculation of density maps for model corrections both streamlines this process and greatly improves the quality of the final structures.

We have designed a validation system (*c.f.* §2.1) that provides a consistently calculated and uniformly represented set of validation statistics, requiring only input of coordinate and diffraction data files. We believe this system is simpler for depositors and provides much more consistent and reliable information than the system at the Protein Data Bank (Berman *et al.*, 2000), in which the depositors supply these statistics. In our own working environment we have established a quality-control staff to oversee the deposition process and to act as an independent gatekeeper to prevent incorrect structures from entering our database. We do not believe that structure depositions should be automatically triggered upon reaching a preset quality standard; even the most reliable structures require a final check by a crystallographer to ensure, for example, that electron densities corresponding to bound ligands are found and correctly modeled. It is unclear how quality control will be enforced for structures emerging from the various structural genomics initiatives, but a system of the type described here, with a well defined minimal standard and automated checking for local structural errors, is one possible model.

We echo the comments of others (Dodson *et al.*, 1996) that it is of critical importance that structure-factor data and (where available) experimentally determined phase sets are available to validate macromolecular models. At the present time (11 April 2001) only 5202 sets of diffraction data are

available in the Protein Data Bank for the 12 250 structures solved by X-ray diffraction (*i.e.* data is available for 42% of structures). Between 1 January 2001 and 11 April 2001, there were 97 diffraction data sets deposited for the 168 structures solved by X-ray diffraction (*i.e.* data is available for 58% of structures), so the current deposition rates of diffraction data fall far short of ideal. The specification and format for depositing diffraction data in the Protein Data Bank is still rudimentary but the availability of a simple reflection list in a standard layout is a useful start. One of the frequently expressed goals for the publicly funded structural genomics initiatives is that more information from the diffraction experiment will be captured and recorded in usable form. If this goal becomes reality, more sets of diffraction data will be available in the future. Automated methods for annotating diffraction data have been described elsewhere (Badger, 2001).

This work would not have been possible without the structures solved by crystallographers at Structural GenomiX (K. Gajiwala, H. Lewis, G. Louie, H.-J. Muller-Dieckmann, J. Newman, F. Park, T. Peat, J. Xu). Other members of the Structural GenomiX development team (J. Christopher, C. Kissinger) together with E. de La Fortelle and M. Milburn are acknowledged for discussions and for supporting this work. H. Hackworth and I. Miller were responsible for Oracle database administration and uploading structures into the database.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.
Badger, J. (2001). *CCP4 Newsl. Protein Crystallogr.* **39**. http://www.ccp4.ac.uk/newsletter39/05_sgx.html.
Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–242.
Bourne, P., Berman, H. M., Watenpaugh, K., Westbrook, J. & Fitzgerald, P. M. D. (1997). *Methods Enzymol.* **277**, 571–590.
Brändén, C.-I. & Jones, T. A. (1990). *Nature (London)*, **343**, 687–689.
Bricogne, G. & Irwin, J. J. (1996). *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.
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, N., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Carson, M., Bruckner, T. W., Yang, Z., Narayana, S. V. L. & Bugg, C. E. (1994). *Acta Cryst.* D**50**, 900–909.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cowtan, K. (1994). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **31**, 34–38.
Cowtan, K. & Ten Eyck, L. F. (2000). *Acta Cryst.* D**56**, 842–856.
Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583–601.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
Dodson, E., Kleywegt, G. J. & Wilson, K. (1996). *Acta Cryst.* D**52**, 228–234.
EU 3-D Validation Network (1998). *J. Mol. Biol.* **276**, 417–436.
Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Kleywegt, G. J. (2000). *Acta Cryst.* D**56**, 249–265.
Kleywegt, G. J. & Jones, T. A. (1996). *Structure*, **4**, 1395–1400.
Lüthy, R., Bowie, J. U. & Eisenberg, D. (1992). *Nature (London)*, **256**, 83–85.
McRee, D. E. (1999). *J. Struct. Biol.* **125** 156–165.
Molecular Simulations Inc. (2000). *Crystallography and NMR Explorer 2000.1.* Molecular Simulations Inc., 9685 Scranton Road, San Diego, CA 92121, USA.
Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). *Proteins*, **12**, 345–364.
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**, 931.
Pannu, N. S., Murshudov, G. N., Dodson, E. J. & Read, R. J. (1998). *Acta Cryst*. D**54**, 1285–1294.
Pannu, N. S. & Read, R. J. (1996). *Acta Cryst.* A**52**, 659–668.
Perrakis, A., Morris, R. M. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–205.
Vriend, G. (1990). *J. Mol. Graph.* **8**, 52–56.
Weissig, H. & Bourne, P. E. (1999). *Bioinformatics*, **15**, 807–831.