Procedures for Assessing the Quality of X–Ray Structures of Macromolecules

S. J. Wodak
Unité de Conformation de Macromolécules Biologiques,
Université Libre de Bruxelles, 50 av. F.D. Roosevelt, CP160/16,
B–1050 Brussels Belgium,
e–mail: shosh@ucmb.ulb.ac.be
url: http://www.ucmb.ulb.ac.be
and
EBI–EMBL, Hinxton, Cambridge CB10 1RQ, UK
shosh@ebi.ac.uk
url: http://www.ebi.ac.uk

Joan Pontius, Alexei Vaguine, Jean Richelle
Unité de Conformation de Macromolécules Biologiques,
Université Libre de Bruxelles, 50 av. F.D. Roosevelt, CP160/16,
B–1050 Brussels Belgium,
e–mail: joan@ucmb.ulb.ac.be, alexei@ucmb.ulb.ac.be, jean@ucmb.ulb.ac.be
url: http://www.ucmb.ulb.ac.be/~joan – url: http://www.ucmb.ulb.ac.be/~jean

Abstract

Two approaches for assessing the quality of protein X–ray structures are presented. One, implemented in the software PROVE, is based on the calculation of atomic volumes and uses the deviation from standard volumes, computed from a reference set of very accurate protein structures, to assess the quality of an atomic model, as a whole, and in specific regions of the model. The other, implemented in the software SF–CHECK, uses several objective criteria for evaluating the experimental structure factor data, when the latter are deposited, and for assessing the agreement of the atomic coordinates with these data, both for the model as a whole and on a per–residue basis. A combination of such tools with existing procedures like PROCHECK may represent the backbone of routine structure validation protocols in the future.

1 Introduction

Recent years have witnessed an exponential growth of data on the 3D structures of macromolecules, and in particular proteins. Managing this information is a challenging problem. It requires efficient ways of storing, cross referencing and accessing these data and the information that can be obtained from them, commonly referred to as 'databases'[1]. Such databases can only be useful if the data they contain are consistent and as error free as possible. This applies in particular to the atomic coordinates of the macromolecules. Owing to the lack of atomic resolution in X–ray and NMR experiments, the data they provide may not be sufficient to define the model of a macromolecule accurately enough, and this model represents a compromise between the fit to the experimental data and to our knowledge of chemistry. Procedures and criteria for assessing the quality of the atomic coordinates, both overall and in specific regions of the structure, are hence of prime importance.

Procedures such as PROCHECK[2], often used in the crystallographic community, focus on the validation of geometric and stereo–chemical parameters of the molecular models. They work mostly by evaluating how these parameters deviate from their standard values, derived from a reference set of high quality protein structures or crystals of small molecules. However, X–ray refinement procedures, as well as methods used for deriving models from NMR data, often use the same parameters as constraints or restraints. For example, least squares refinement algorithms such as PROLSQ[3] or TNT[4] use restraints on covalent geometry, whereas procedures such as XPLOR[5], based on molecular dynamics methods, apply in addition, restraints on non–bonded contacts. These restraints and constraints can leave their mark on the final model[6], and measuring the quality of a structure in terms of how well certain parameters match the standard values may thus in fact evaluate how different standard values compare with one another [2] [7]. Hence, there is a need for objective methods for assessing the quality of a protein model. These methods should use quality measures based on parameters that are not directly used in generating the model, and more importantly still, they should evaluate the agreement between the model and the experimental data.

Here we present two different approaches to the quality assessment of protein crystal structures. One is based on the calculations of atomic volumes. Atomic volumes are clearly influenced by a number of parameters (bond distance, bond angles, non–bonded contacts) that are subjected to restraints in many refinement procedures. But volumes, as such, are not restrained during refinement. The software PROVE[8] uses the deviation from standard volumes, computed from a reference set of accurate protein structures, to assess the quality of an atomic model, as a whole, and in specific regions of the model. The other approach, implemented in the software SF–CHECK, proposes standard procedures for analyzing structure factor data, when the latter are deposited, and for assessing the agreement of the atomic coordinates with the electron density, both for the model as a whole and on a per residue basis. Several of the quality measures and criteria used by SF–CHECK are already computed in one form or another in existing refinement programs, but several are novel. SF_CHECK applies these different measures to a given structure completely automatically, and provides a concise pictorial output in PostScript format.

2 Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures

2.1 Computation of Atomic Volumes

Atomic volumes were computed in a reference set of 64 highly resolved (better than 2.0Å) and refined protein structures (listed in [8]) using the classical Voronoi method[9] as implemented in the program SurVol[10]. Unlike other commonly used variants of the Voronoi procedures [11] [12] [13], the classical method does not require assigning atomic radii (Figure 1). It is therefore particularly well suited for large surveys, because there is little hope for obtaining a consistent set of atomic radii for all ligands and co–factors encountered in protein crystal structures.

The computations considered only buried atoms, defined as those with zero surface area accessible to solvent. Atoms with non–zero accessible surface area could not be handled, because they are not completely surrounded by other atoms, and their Voronoi volume can therefore not be defined. Water molecules, DNA, RNA and hetero group atoms were excluded from the volume and accessible surface calculations, as were hydrogen atoms. Protein atoms lining cavities within the structure, such as those created by these excluded groups, or cavities that are empty even when these groups are included, were identified and treated as surface atoms, and excluded from the analysis.

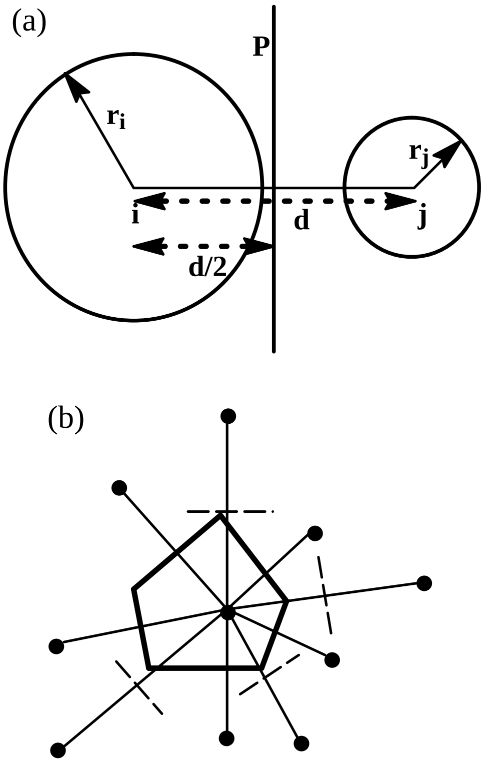


Figure 1 The Voronoi procedure

(a) The Classical Voronoi procedure positions the plane P half way (d/2) between the centers of atoms i and j, with radii r_i and r_j, whose center–to–center distance is d.
(b) A 2D illustration of the Voronoi polygon (polyhedron in 3D) defining the space (volume) occupied by the central atom when it is surrounded by neighboring atoms in a structure. Vectors are drawn from the central atom to all its neighbors within a given radius, and the planes P perpendicular to these vectors are positioned, as illustrated in (a). The smallest polygon constructed in this way is the Voronoi polygone.

2.2 The Relevant Volume Distributions

To derive standard volumes, it is first of all necessary to identify the relevant volume distributions. In this work atoms were assigned to a total of 23 atom chemical types, used in the BRUGEL package[14] (see also [8]). Since the atom chemical type reflects its bonding properties and chemical character, volume distribution according to these chemical types were computed (Figure 2). It was found that the computed volumes correlate better with their bonding properties than with their van der Waals (vdW) radii. For example the CH3 group, which is usually considered as having the same vdW radius as the CH1 and CH2 groups[11] [15] [16] [17], but is bonded to only one atom, has a larger volume than all other groups. Similarly, the backbone carbonyl oxygen, being bonded to only one atom, has a larger volume than the backbone amide (NH1), even though it is considered to have a smaller vdW radius than other backbone atoms. The influence of the number and nature of the covalently bonded neighbors on the computed atomic volumes was also analyzed and found to be important. Based on these findings, mean volumes of buried atoms and the corresponding standard deviations were computed for atoms grouped according to their atom type, defined by their residue, IUPAC code[18] and chemical type. These data can be found in Table 1 of Pontius et al.[8]

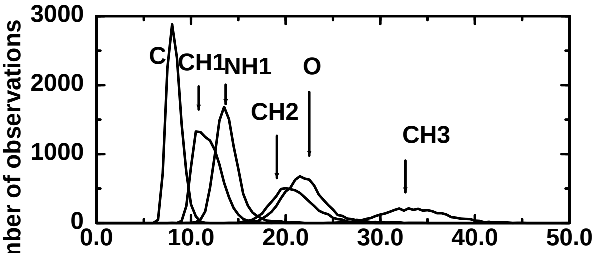


Figure 2. Volume distributions for different sub−populations of atoms

Distributions of atomic volumes computed with the Voronoi procedure for atom sub−populations segregated according to their chemical types. We use the 23 chemical types defined in the modelling package BRUGEL[14] and detailed in Pontius et al.[8].

2.3 Scoring the Deviations from the Expected Values

The deviation of the atomic volume from the standard value is evaluated by the volume Z–score:

$$Z\ score_i = \frac{[V_i^k - \overline{V^k}]}{\sigma^k}$$

V_i^k is the atomic volume of atom i, having atom type k, calculated using SurVol. $\overline{V^k}$ denotes the mean volume of buried atoms with the same atom type k, and σ^k denotes its associated standard deviation. A negative Z–score means that the atom has a smaller than average volume, whereas a positive score indicates that an atom has a larger than average volume. The expected average Z–score is zero.

The Z–score rms deviation from ideality is used as a global measure of departure from the expected behavior in a given set of N atoms, which can be all the atoms of a given protein structure, or atoms with specific attributes, such as the same B–factor range:

$$Z\ score\ rms = \sqrt{\frac{\sum_{i=1}^{N}[Z\ score_i]^2}{N}}.$$

2.4 Global Measures of Structure Quality

Since the resolution and the R factor are good guides for the overall quality of a structure determined by X–ray diffraction, the correlation between these parameters and the average volume irregularity of a protein structure was investigated in a test set of 900 proteins (see [8] for details). An analysis performed on a larger set of 3000 proteins, and a set of 8 atomic resolution structures is also briefly quoted.

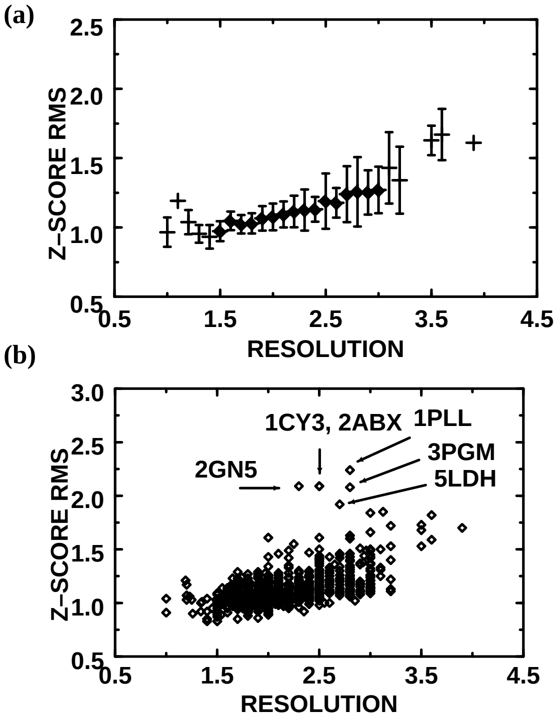


Figure 3. Z−score rms variation with resolution in a test set of 900 protein structures

(a) Z−score rms as a function of the resolution. Average Z−score rms is computed for structures having the same resolution (± 0.1Å). The vertical bars indicate the magnitude of the standard deviation of the Z−score rms in individual resolution ranges. Graph points are derived from less than 10 structures (horizontal lines) and from more than 10 structures (filled diamonds).
(b) Z−score rms as in (a), displayed for individual protein structures as a function of resolution. The furthest 6 outlier proteins are marked by the PDB codes.

Figure 3a displays the average Z–score rms computed for structures of a given resolution range as a function of resolution in our test set of 900 protein structures. The overall considered resolution range was from 1 to 3.9Å, and the averages were computed for bins of 0.1Å resolution. For resolutions of 1.6Å or better, the average Z–score rms is essentially constant (~1.0). Lower resolution structures display, on the average, a larger Z–score rms, with the average Z–score rms increasing steadily as the resolution decreases. The correlation factor between the average Z–score rms and the resolution is 0.89 over the entire range of considered resolution, and 0.98 for resolutions between 1.5–3.0Å. The standard deviations of the Z–score rms in each resolution bin is displayed as vertical bars in Figure 3a. They delimit the expected spread of the Z–score rms for a given resolution. A structure determined at a given resolution, whose Z–score rms falls outside the expected spread, is likely to exhibit problems.

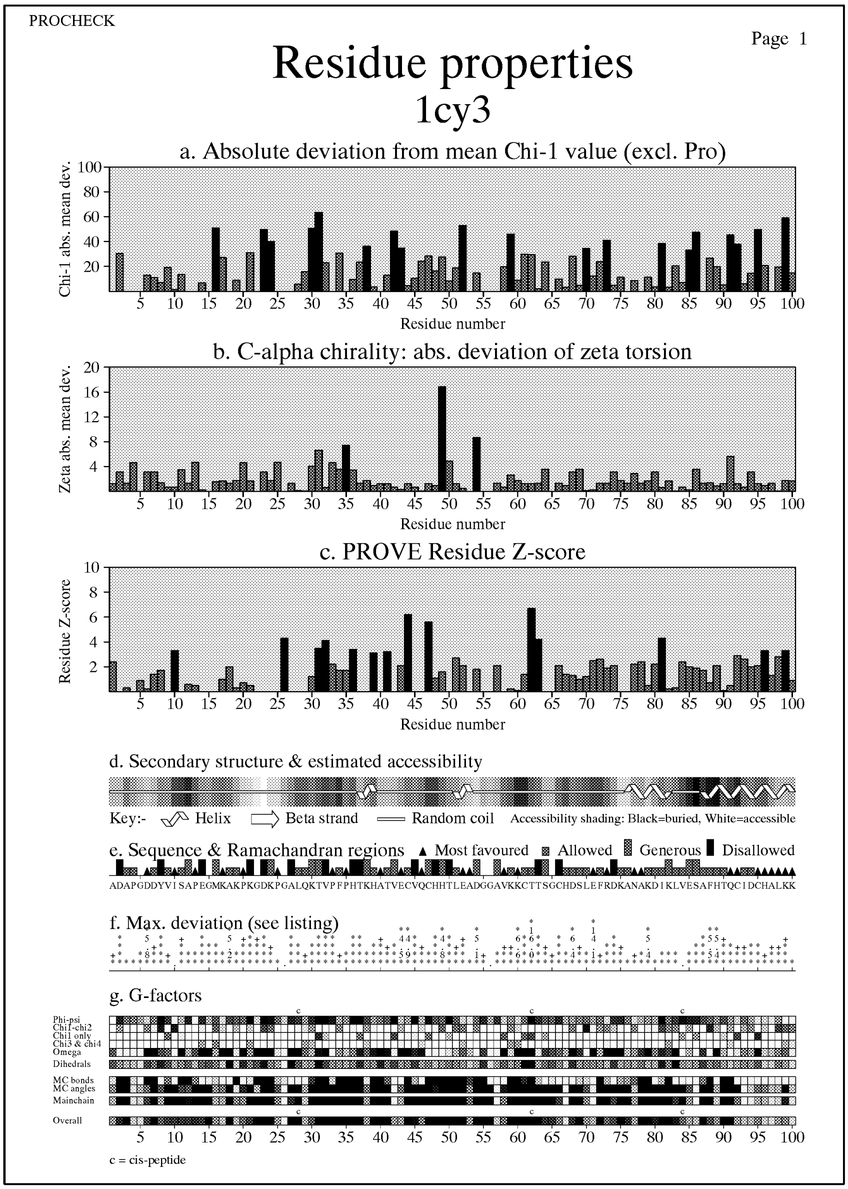
It is noteworthy that the spread in Z–score rms values for a given resolution can be as much as 0.4, indicating that the correlation of the Z–score rms of individual protein structures with resolution is poorer than that of the average Z–score rms. This spread is clearly illustrated in Figure 3b, which displays the Z–score rms of individual proteins in our test set as a function of their resolution. A number of structures have Z–score rms values well outside the expected spread. Six of the farthest outliers are marked (Figure 3b). They correspond to the entries 1PLL (oncogene protein), 2GN5 (gene 5 DNA binding protein), 3PGM (phosphoglycerate mutase), 1CY3 (cytochrome C3), 2ABX (alpha–bungarotoxin), and 5LDH (lactate dehydrogenase). All these structures were also found to be severe outliers with regard to their stereochemical parameters, as discussed in [8].

A significant correlation between the Z–score rms and the crystallographic R–factor was also established (data not shown), but the correlation coefficient between the average Z–score rms with the R–factor was poorer (0.76) than with the resolution. This is not unexpected, given that the R–factor is a versatile parameter that can be computed for various subsets of data. It hence reflects more the agreement between the model and those data subsets than the quality of the model itself.

2.5 Local Measures of Structure Quality

To analyze the deviations of the atomic volumes in specific regions of the polypeptide, the atomic volumes Z–scores in individual residues were surveyed, and compared with the behavior of structural parameters analyzed by PROCHECK. Figure 4 displays part of the standard PROCHECK output together with the volume Z–score plots for cytochrome C3 (1CY3), one of the severe outlier structures in Figure 3b. The PROVE plots show for each residue the maximum absolute volume Z–score displayed by the buried atoms of this residue. It represents the largest departure from the standard volume displayed in a single atom within a residue and is therefore not an average property of the residue. Figure 4c shows that residues 26, 32, 44, 47, 62, 63, and 81 of cytochrome C3 (1CY3) contain atoms with absolute volume Z–scores, greater than 4. These high Z–scores belong to the backbone carbonyls of Gly 26, Cys 47, and Thr 62, the Cβ of Val 32 and Ile 81, and the backbone oxygens of Cys 44, and Thr 63. We find that the same residues, or their close neighbors, also have unusual Chi–1 values , unusual omega values , or distorted Cα chirality. Residues 45, 48, 50, 63 and 66, for example, are also in disallowed regions of the Ramachandran map (Figure 4e). Similar observations were made for the other proteins of our test set.

We observed that residues found to be outliers on the basis of the volume Z–score of one of their atoms were not always outliers by the PROCHECK measures. This is not surprizing considering that the volume of a given atom can be affected by the position of its spatial neighbors, some of which may belong to residues far apart along the sequence. Unusual volumes may therefore result from errors occurring in several parts of the atomic model, and thereby have more complex origins than the deviations of geometric parameters such as the Cα chirality or the Chi 1 angle, which are due to local modelling errors. They may also result from a combined effect of small irregularities in several parameters, which taken individually go undetected by PROCHECK.



1cy3_01.ps

Figure 4. PROCHECK and PROVE outputs for the first 100 residues of cytochrome c3 (1CY3)

(a) Absolute deviation from mean Chi–1 value, computed by PROCHECK. Highlighted residues are those that deviate by more than 2 standard deviations from ideal.
(b) Cα chirality: absolute deviation of zeta torsion, computed by PROCHECK. Highlighted residues are those that deviate by more than 2 standard deviations from ideal.
(c) Maximum absolute Z–score of atomic volumes in individual residues along the sequence, computed by PROVE, in this study. Highlighted residues are those with Z–scores >3. (d) Standard PROCHECK output for Secondary Structure and estimated accessibility.
(e) Standard PROCHECK output for sequence and backbone ϕ, ψ values relative to the Ramachandran regions.

2.6 Influence of the Refinement Procedure on the Volume Deviations

Although atomic volumes are not directly restrained in refinement procedures, these procedures usually restrain bond distances and angles. Programs such as XPLOR, and others, also use energy parameters, which impose restraints on non–bonded contacts. All this may affect atomic volumes and packing in the crystal.

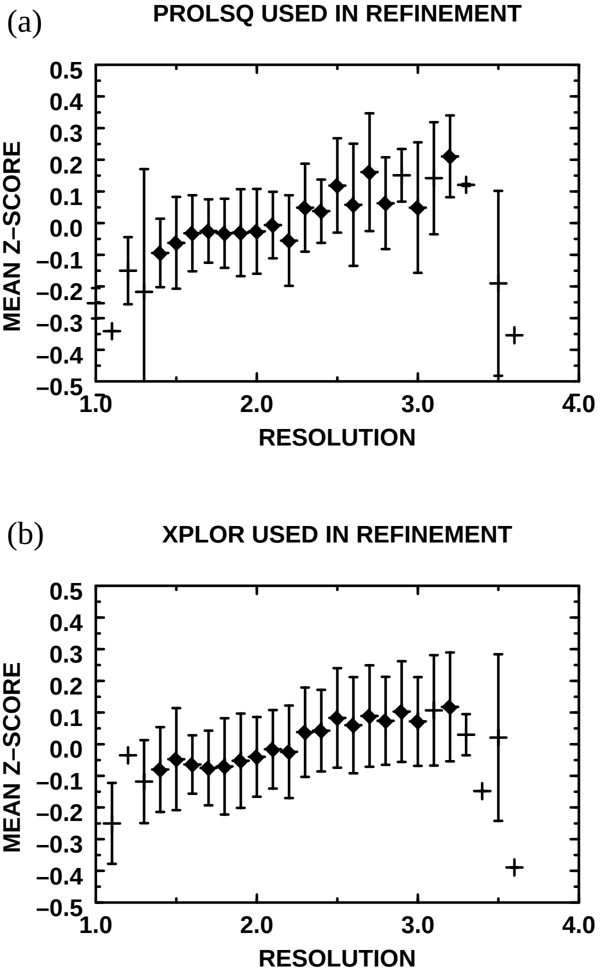


Figure 5. Variation of the average mean Z−score with resolution in structures refined with by different methods

(a) The average mean Z−score as a function of resolution in structures refined by PROLSQ[3]. The mean Z−scores of structures having the same resolution (± 0.1Å) is averaged. The vertical bars indicate the magnitude of the standard deviation of the mean Z−score in individual resolution ranges. Graph points are derived from less than 10 structures (horizontal lines) and from more than 10 structures (filled diamonds).

(b) The average mean Z−score as a function of resolution in structures refined by XPLOR[5].

To investigate such effects, the volume Z–scores were analyzed in groups of proteins structures refined with different procedures. These groups were taken from a larger test set of 3000 protein structures in a more recent release of the PDB. A first inspection of the results showed that atomic coordinates derived by some of the most commonly used refinement procedures such as PROLSQ and XPLOR display very similar volume deviations as a function of resolution (Figure 5).

Figure 5 also reveals the following intriguing trend: High resolution structures tend to have negative mean Z–scores, whereas medium and low resolution structures have positive mean Z–scores. Recalling that the mean Z–score is the algebraic mean of the atomic Z–scores in a given structure averaged over all the structures in a given resolution range, this result indicates that atoms in high resolution structures tend to occupy smaller volumes than expected , whereas the opposite is true for the medium and low resolution structures. A volume shrinkage has also been observed in an independent analysis of 8 structures solved at atomic resolution [Pontius et al., unpublished]. The origins of these observations is presently not understood. It has been suggested by Gérard Bricogne [personal communication, and public remark during the presentation] that shrinkage could result from problems in modelling rigid–body motion in current protein refinement protocols.

3 SF_CHECK: A Set of Standard Procedures for Evaluating Structure Factor Data and the Agreement between the Atomic Model and the Electron Density

The quality assessment of a deposited protein model is not complete without evaluating the quality and completeness of the experimental data, and measuring the agreement of the derived model with those data. In the case of crystal structures, the experimental data refer to the structure factor amplitudes, which are derived by processing the raw diffracted intensities. The quality and completeness of these data are usually evaluated during various stages of the structure determination process by different programs, whereas the agreement of the model with the experimental data is evaluated at the refinement stage based on the commonly used quantities such as the R–factor, or the Free R–factor[19]. Though these parameters, which qualify the model as a whole, are nearly always reported by the authors, they are not computed in the same way by everyone, and can therefore not be meaningfully compared between structures. In addition, protein structures often have regions that are less reliably modelled than others. Authors usually know very well where these regions are, but this information is only partially passed on in the deposited entries, by the occupancy and B–factor parameters, or by the author's comments in text form. Ad–hoc means are then needed to relate this information to the atomic coordinates.

Following these considerations it appeared useful to undertake the development of SF_CHECK, a stand–alone package containing a set of standard procedures, which may be applied to to the deposited atomic coordinates and structure factor data in order to (1) validate the experimentally derived structure factors data and (2) evaluate the agreement between the atomic coordinates and these data. With the mounting pressure to make the deposition of diffraction data mandatory[20], a standard tool such as SF_CHECK, which is independent from any specific refinement program, computes a range of quality checks and produces an easy to read pictorial summary, should be an extremely welcome addition to the panoply of structure validation tools.

In what follows we briefly describe the first version of SF–CHECK, and summarize some of its major features. A detailed description of the package will appear elsewhere shortly.

The flow–chart of SF_CHECK is depicted in Figure 6. SF_CHECK reads in the structure factor data written in the mmCIF format[21], or in the files currently deposited in the PDB, and reads the atomic coordinates, provided either in the PDB or the mmCIF formats. Next, it computes statistics on the structure factor data (see below), generates an electron density map from the atomic coordinates, computes F_{calc} by FFT, scales the $|F_{obs}|$ and $|F_{calc}|$. Then it uses FFT to compute two electron density maps, with calculated phases and respectively, observed and calculated amplitudes. Lastly, it calculates the gradients of the difference map with respect to the atomic coordinates, compares the observed and calculated structure factor amplitudes, and computes various quantities that are used to asses the local agreement between the observed and model electron densities. A more detailed description of selected tasks performed by SF–CHECK is given below. Additional information may be found in the legend of Figure 6.

3.1 The main tasks performed by SF_CHECK

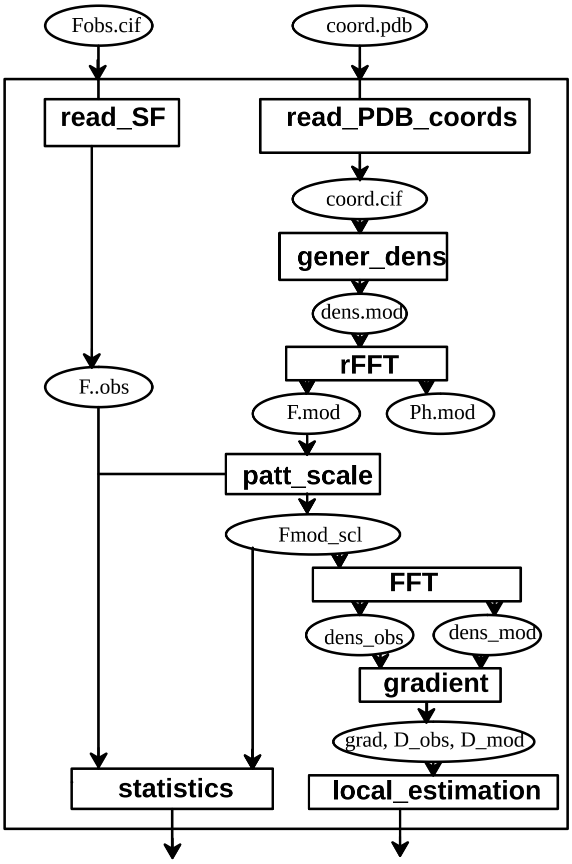


Figure 6. SF_CHECK flowchart

read_SF: reads formatted structure factor amplitude files in CIF format, or the formats of existing structure factor files in the PDB.
read_PDB_coords: reads standard PDB coordinate files. These are then translated to CIF format (coord.cif) by SF_CHECK (task not shown).
gener_dens: generates the electron density map from the atomic model.
rFFT: computes the calculated structure factor amplitudes from the model, by FFT.
patt_scale: scales the calculated structure factor amplitudes by Patterson. This involves deriving the overall B−factor from the width of the origin peak of Patterson maps computed using respectively, the calculated and observed amplitudes.
FFT: computes 2 electron density maps us-

ing phases from the model, and respectively, observed and calculated amplitudes
gradient: for each atom, computes the gradient of the difference electron density map with respect to the atomic coordinates, and the curvature of the map at the atomic center.
statistics: computes quantities concerning the model and the observed amplitudes, such as the R–factor, the R–factor as a function of resolution, observed amplitudes in resolution bins etc.
local_estimation: for each residue computes the 5 quantities plotted in Figure 7(b). These quantities are defined in the text.

3.2 Analysis the structure factor data and global agreement of the model with these data

SF_CHECK performs a detailed analysis of the deposited structure factor amplitudes, and evaluates the global agreement between the model and the structure factor data using a number of standard criteria. The SF_CHECK output, presented in Figure 7(a) summarizes the results of this analysis performed on the structure of the HIN recombinate (DNA binding domain C) (1HCR). This structure was chosen as an example, because it is declared by its authors to represent a 'preliminary coordinate set', with 'refinement still in progress'. Accordingly, it could display problems not commonly featured by well refined structures, but which SF_CHECK should readily detect.

The SF_CHECK output displays a total of 7 panels. The four panels at the top summarize numerical data. The lower three panels display (from left to right and from top to bottom): the distribution of the R–factor and the Luzatti plot; the distribution of the structure factor amplitudes, and the Wilson plot; a plot giving the completeness of the data, and the ratio <σ(F)>/<F>, as a function of the resolution. Further details about the various numerical quantities listed in the different panels is given in the legend of Figure 7.

A quick inspection of Figure 7(a) allows to identify several features which could a be source of difficulties in the structure determination. It shows, for example, that the R_factor distribution at high resolution is unusual, and that even though the reported resolution is 1.8Å, the |F|'s are very weak beyond 3.0Å resolution. The low R_stand values at resolutions higher than 3.0Å, directly confirm the poor quality of the diffraction data at those resolutions. SF_CHECK can also be helpful in validating numerical information provided by the authors, or annotations by database curators. We see for example, that there is a discrepancy between the reported R–factor (0.22) and those computed by SF–CHECK considering all acceptable reflections (0.33) or considering only a reflection subset (|F|> 1.5 σ, and resol <6Å) (0.325). The origins of this discrepancy is not clear. Our R–factor calculations performed with different subsets of the deposited structure factor data, including the subset allegedly used by the authors to compute the repoorted R–factor (this will be a default feature in future versions of SF_CHECK), suggest that there may have been a misprint in the PDB entry.

3.3 Evaluation of local agreement between the model and the electron density

Figure 7(b) illustrates the results of the analysis performed by SF_CHECK on the local agreement between the model and the crystallographic data for 1HCR. The figure displays the following five quantities for each residue along the polypeptide chain:

1) The normalized average displacement of atoms in each residue *Shift*:

$$Shift = \frac{1}{N}\sum_{i=1}^{N}\frac{\Delta i}{\sigma i}$$

with Δ_i equals:

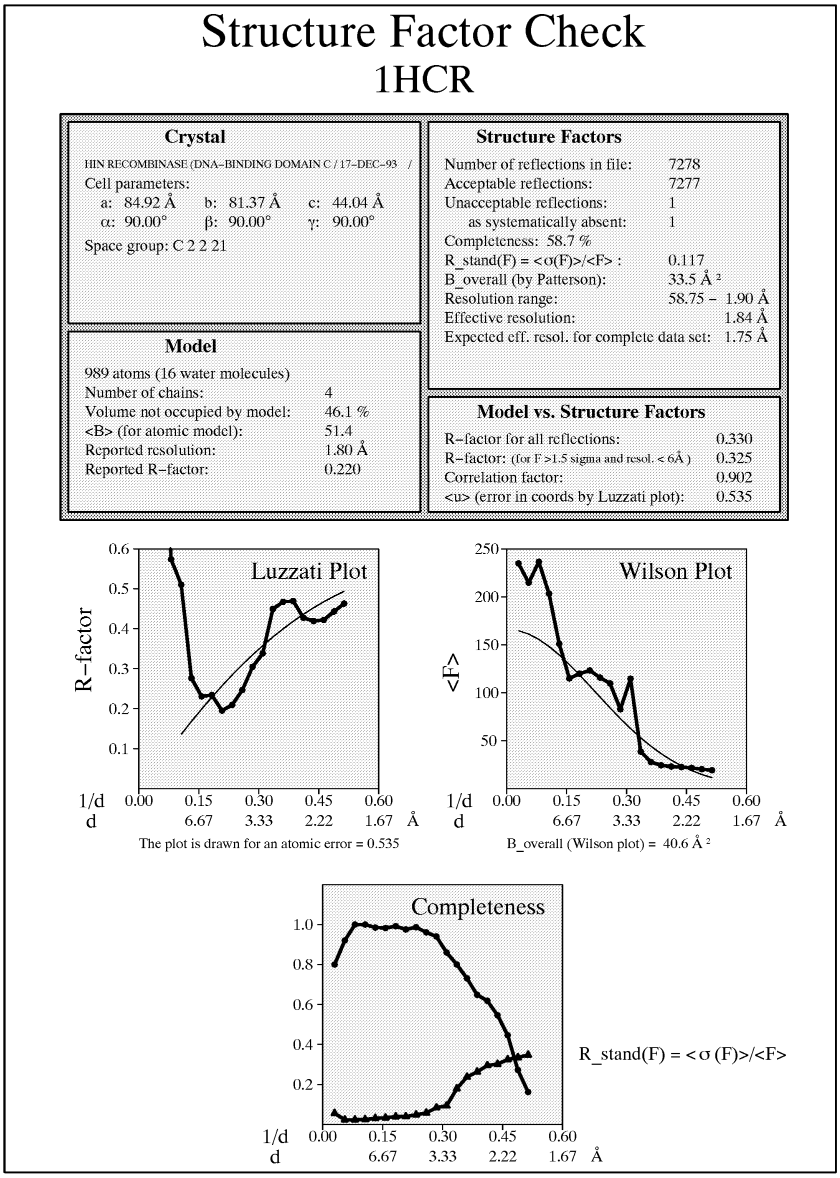
$$\Delta i = \frac{Gradient_i}{Curvature_i}$$

where *Gradient*$_i$ is the gradient of the (F_{obs}–F_{calc}) map with respect to the atomic coordinates and *Curvature*$_i$ is the curvature of the model map computed at the atomic center and N, the number of atoms. σ is the standard deviation of the Δ_i values computed in the structure. *Shift* indicates the tendency of the model to move away from its current position, with large values of *Shift*, corresponding to regions where this tendency is high.

2) *D_Corr*, the electron density correlation coefficient for the atoms in each residue:

$$D\_corr = \frac{\sum_{i=1}^{N} \varrho_{calc}(x_i)\ [2\varrho_{obs}(x_i) - \varrho_{calc}(x_i)]}{\sqrt{\left[\sum_{i=1}^{N} \varrho_{calc}(x_i)^2\right]\left[\sum_{i=1}^{N}(2\varrho_{obs}(x_i) - \varrho_{calc}(x_i))^2\right]}}$$

where $\varrho_{calc}(x_i)$ and $\varrho_{obs}(x_i)$ are respectively the electron density computed from calculated and observed structure factor amplitudes, and the summation is performed over the N–backbone atoms or sidechain atoms separately. *D–Corr* measures the agreement between the model and the electron density. Small values of *D–Corr*, depicted by large bars in Figure 7(b), indicate that the model of the corresponding backbone or sidechain agrees poorly with the electron density.



Structure Factor Check

1HCR

Crystal

HIN RECOMBINASE (DNA–BINDING DOMAIN C / 17–DEC–93 /

Cell parameters:

a: 84.92 Å	b: 81.37 Å	c: 44.04 Å
α: 90.00°	β: 90.00°	γ: 90.00°

Space group: C 2 2 21

Model

989 atoms (16 water molecules)

Number of chains:	4
Volume not occupied by model:	46.1 %
<B> (for atomic model):	51.4
Reported resolution:	1.80 Å
Reported R–factor:	0.220

Structure Factors

Number of reflections in file:	7278
Acceptable reflections:	7277
Unacceptable reflections:	1
as systematically absent:	1
Completeness: 58.7 %	
R_stand(F) = <σ(F)>/<F> :	0.117
B_overall (by Patterson):	33.5 Å ²
Resolution range:	58.75 – 1.90 Å
Effective resolution:	1.84 Å
Expected eff. resol. for complete data set:	1.75 Å

Model vs. Structure Factors

R–factor for all reflections:	0.330
R–factor: (for F >1.5 sigma and resol. < 6Å)	0.325
Correlation factor:	0.902
<u> (error in coords by Luzzati plot):	0.535

The plot is drawn for an atomic error = 0.535

B_overall (Wilson plot) = 40.6 Å ²

Figure 7a. SF_CHECK output for the HIN recombinase (1CHR)

(a) Results of the analysis of structure factor data and of the global agreement of the model with that data.

The panel **’Crystal’** summarizes the crystal information taken from the PDB records

The **’Model’** panel summarizes data about the model The total number of atoms and the unit volume not occupied by the model are quantities computed by SF_CHECK from the atomic coordinates, and the crystal data.

The **’Structure Factors’** panel summarizes data on the deposited structure factor amplitudes. All listed quantities are self explanatory, except for the following: ”B_overall (by Patterson)” is the overall B–factor computed from the width of the Patterson origin peak; The ”effective resolution” is defined as the expected minimum distance between 2 well resolved peaks in the electron density map. It is computed as $\frac{2}{\sqrt{2}} \cdot \Delta_{Patt}$, where Δ_{Patt} is the width of the Patterson origin peak.

The ’Expected eff. resol. for complete data set’ is the expected effective resolution computed as above, but taking into account all reflections, with values for missing reflections equalling the average value of the reflections in the corresponding resolution bin.

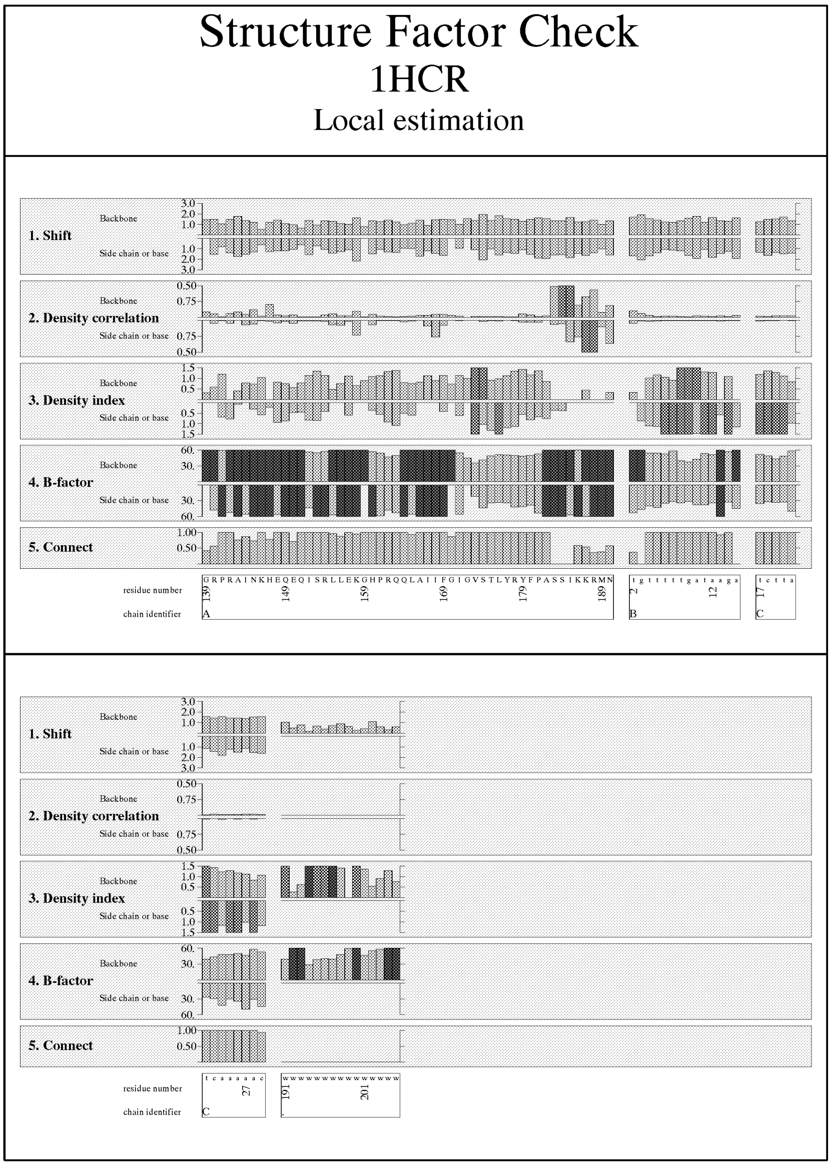


Figure 7b. SF_CHECK output for the HIN recombinase (1CHR)

(b) Results of the analysis of local properties of the model, the electron density, and the agreement between both.
The 5 plotted quantities are defined in the text. The amino–acid or nucleic acid sequence is listed below the 5th plot, using the one letter code, with w denoting water molecules. The residue numbering and chain identifiers are given below the sequence. Blackened rectangles indicate outliers.

3) The residue *Density–index* is defined as follows:

$$Density\_index \ = \frac{\sqrt[N]{\prod_{i=1}^{N} \varrho \ (x_i)}}{< \varrho >_{all\ atoms}}$$

where N is the total number of considered atoms in the sidechain or backbone groups; the numerator of the above equation is the geometric mean of the (2_{Fobs}–$\mathrm{F_{calc}}$) electron density of the considered atom subset and the denominator is the average electron density of the atoms in the structure. The *Density_index* reflects the level of the electron density at the backbone or side chain atoms of a given residue, and thereby provides a local measure of the density level. For regions with high electron density, the value of the *Density_index* nearly always exceeds 1. For regions with low electron density, this value will be < 1. Such regions may be problematic for model fitting.

4) The *B–factor* is computed as the average of the atomic B–factors of the backbone and sidechain atoms of each residue. Comparison of the *B–factor* and *Density_index* plots, can be useful for detecting regions with errors in the model. One would expect that in a well refined model, atoms with large B–factors would lie in regions with low density, characterized in our plot by a low *Density_index*. Therefore, when such atoms occur in high density regions, one may suspect problems with either the model or the refinement procedure.

5) The connectivity index, *Connect*, is the same quantity as the *Density_index*, but computed for the backbone atoms excluding the carbonyl oxygens in proteins, and considering the P, O5’,C5’,C3’,O3’ atoms in nucleic acids. *Connect* measures the level of the electron density along the macromolecular skeleton and can be used to assess the continuity of the electron density along the polymer chain. A low *Connect* value indicates locations where this continuity is broken. Such locations may occur in loops lying in regions with low electron density, or in places where errors in model tracing occurred.

Inspection of the SF_CHECK plots for 1HCR (Figure 7(b)) reveals several features which clearly suggest that the refinement of the proposed model ’is still in progress’. The *Shift* value (plot 1) is large (>1Å) for both the backbone and side chain atoms, indicating that the refinement has not converged. In addition, the B–factors (plot 4) are very high for most backbone atoms, where they generally exceed 60Å^2, as witnessed by the large number of black rectangles. Since the *Density_index* of many residues is quite low in both the main chain and side chains (plot 3), the large B–values could result from attempts by the refinement program to fit a model into low density regions. Interestingly, the *Connect* value (plot 5) is rather high throughout, except at residues 83–85, where it is zero. Since the *Connect* parameter, just like the *Density_index*, measures the density level for backbone atoms, but excluding the carbonyl oxygens, this indicates that the latter atoms, in particular, tend to lie outside the electron density.

We thus see that a quick glance of the SF–CHECK summary can identify problem regions in the model, and help formulate hypotheses on the origins of these problems. Such hypotheses must of course be investigated further by a more detailed analysis using the usual panoply of tools.

4 Concluding Remarks

This paper presented two different procedures for evaluating the quality of protein X–ray structures. One (PROVE), analyzes the departures of atomic volumes from standard values compiled from known protein structures, and therefore belongs to the category of procedures that measure how unusual a protein model is in comparison with other protein models derived previously. The other (SF_CHECK) belongs to a different category, in that it uses a number of objective criteria to measure the quality of the X–ray data, and to assess the agreement between the model and that data. The latter task, in particular, is notoriously difficult, and the criteria proposed here by SF–CHECK should be considered only as a starting point. The main bottleneck to the generalization of procedures such as SF_CHECK is that diffraction data are not available for most of the structures in the PDB. However, when, as some of us hope, the deposition of these data will become mandatory, routine structure validation protocols will most likely combine both types of procedures.

The program PROVE is accessible through the World Wide Web as part of the European Biotech validation server (http://biotech.embl–ebi.ac.uk:8400/, in Europe and http://biotech.pdb.bnl.gov:8400/, in the US.

Acknowledgements

We thank all members of the European consortium on Structure Validation for stimulating discussions, and H. Bermann and her colleagues for their critical comments on the SF_CHECK output. The work on PROVE was part of the European BIOTECHNOLOGY project BIO2–CT92–0524 on 3D Macromolecular Structure Validation. We moreover gratefully acknowledge support from the Department Of Energy, USA; the Belgian programme of Interuniversity Poles of Attraction initiated by the Belgian State, Prime Minister’s Office for Science, Technology and Culture; the Université Libre de Bruxelles fellowship fund, and the Fund for Joint Basic Research (Belgium).

References

[1] EU BRIDGE Database project consortium: P.M.D. Gray, G.J.L. Kemp, C.J. Rawlings, N.P Brown., C. Sander, J.M.Thornton, C.M. Orengo, S.J. Wodak & J. Richelle ”Molecular structure information and databases”, TIBS, Vol. 21, pp. 251–256, 1996.

[2] R.A. Laskowski, M.W. MacArthur, D.S. Moss, & J.M. Thornton, ”PROCHECK: a program to check the stereochemical quality of protein structures”, J. Appl. Cryst., Vol. 26, pp. 283–291, 1993.

[3] W.A. Hendrickson & J.H. Konnert, Computing in Crystallography (Diamond, R., Ramaseshan S. & Venkatesan K., eds.), pp. 1301, Indian Acad. Sci., Bangalore, 1980.

[4] D.E. Tronrud, L.F. Ten Eyck & B.W. Matthews, ”An Efficient General–Purpose Least–Squares Refinement Program for Macromolecular Structures”, Acta Cryst. Vol. A43, pp. 489–501, 1987.

[5] A.T. Brünger, J. Kuriyan. & M. Karplus ”Crystallographic R–factor Refinement by Molecular Dynamics” Science, Vol. 235, pp. 458–460, 1987.

[6] D.E. Stewart, A. Sarkar & J.E. Wampler, ”Occurrence and Role of Cis Peptide Bonds in Protein Structures”, J. Mol. Biol. Vol. 214, pp. 253–260, 1990.

[7] R.A. Laskowski, D.S. Moss & J. Thornton, ”Main–chain Bond Lengths and Bond Angles in Protein Structures”, J. Mol. Biol. Vol. 231, pp. 1049–1067, 1993.

[8] J. Pontius, J. Richelle & S.J. Wodak, ”Deviations from standard atomic volumes as a quality measure for protein crystal structures” J. Mol. Biol. (in press)

[9] G.F. Voronoi, ”Nouvelles applications des paramètres continus à la théorie des formes quadratiques”, J. Reine Angew. Math. Vol. 134, pp. 198–287, 1908.

[10] P. Alard, PhD Thesis Dissertation, Calculs de surface et d’énergie dans le domaine des macromolécules. Université Libre de Bruxelles, 1991.

[11] F.M. Richards, ”The Interpretation of Protein Structures: Total Volume, Group Volume Distributions and Packing Density”, J. Mol. Biol., Vol. 82, pp. 1–14, 1974.

[12] F.M. Richards, ”Calculation of Molecular Volumes and Areas for Structures of Known Geometry”, Methods Enzymology, Vol. 115, pp. 440–464, 1985.

[13] B.J. Gellatly & J.L. Finney, ”Calculation of Protein Volumes: An Alternative to the Voronoi Procedure”, J. Mol. Biol. Vol. 161, pp. 305–322, 1982.

[14] P. Delhaise, D. Van Belle, M. Bardiaux, P. Alard, P. Hamers, E. Van Cutsem & S. Wodak, ”Analysis of data from computer simulations on macromolecules using the CERAM package” J. Mol. Graphics, Vol. 3, pp. 116–119, 1985.

[15] J.L. Finney, ”Volume Occupation, Environment and Accessibility in Proteins. The Problem of the Protein Surface”, J. Mol. Biol. Vol. 96, pp. 721–732, 1975.

[16] C. Chothia, ”Structural invariants in protein folding’, Nature (London), Vol. 254, pp. 304–308, 1975.

[17] B.R. Brooks, R.E. Bruccoreri, D. Olafson, D. States, S. Swaminathan & M. Karplus, ”CHARMM: a program for Macromolecular Energy, Minimization, and Dynamics Calculation”, J. Comp. Chem. Vol. 4, pp. 187–217, 1983.

[18] IUPAC–IUB Commission on Biochemical Nomenclature, ”Abbreviations and Symbols for the Description of the Conformation of Polypeptide Chains. Tentative Rules (1969)”, Biochemistry, Vol. 9, pp. 3471–3479, 1970.

[19] A.T. Brünger, ”Free R value: a novel statistical quantity for assessing the accuracy of crystal structures”, Nature (London) Vol. 355, pp. 472–474, 1992.

[20] T.A. Jones, G.J. Kleywegt & A.T. Brünger, ”Storing diffraction data”, Nature (London) Vol. 383, pp. 18–19, 1996.

[21] P.E.Bourne, H.M.Berman, B. McMahon, K. Watenpaugh, J. Westbrook & P.M.D. Fitzgerald, ”The Macromolecular CIF Dictionary (mmCIF)”, Methods in Enzymology, accepted, 1996.