Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model

Alexei A. Vaguine,[a]† Jean Richelle[a] and S. J. Wodak[a,b]*

[a]Unité de Conformation de Macromolécules Biologiques, Université Libre de Bruxelles, Avenue F. D. Roosevelt 50, CP160/16, B-1050 Bruxelles, Belgium, and [b]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England

† Present address: Department of Chemistry, University of York, York YO1 5DD, England.

Correspondence e-mail: shosh@ucmb.ulb.ac.be

Received 8 January 1998
Accepted 7 May 1998

In this paper we present *SFCHECK*, a stand-alone software package that features a unified set of procedures for evaluating the structure-factor data obtained from X-ray diffraction experiments and for assessing the agreement of the atomic coordinates with these data. The evaluation is performed completely automatically, and produces a concise PostScript pictorial output similar to that of *PROCHECK* [Laskowski, MacArthur, Moss & Thornton (1993). *J. Appl. Cryst.* **26**, 283–291], greatly facilitating visual inspection of the results. The required inputs are the structure-factor amplitudes and the atomic coordinates. Having those, the program summarizes relevant information on the deposited structure factors and evaluates their quality using criteria such as data completeness, structure-factor uncertainty and the optical resolution computed from the Patterson origin peak. The dependence of various parameters on the nominal resolution (d spacing) is also given. To evaluate the global agreement of the atomic model with the experimental data, the program recomputes the R factor, the correlation coefficient between observed and calculated structure-factor amplitudes and $R_{\rm free}$ (when appropriate). In addition, it gives several estimates of the average error in the atomic coordinates. The local agreement between the model and the electron-density map is evaluated on a per-residue basis, considering separately the macromolecule backbone and side-chain atoms, as well as solvent atoms and heterogroups. Among the criteria are the normalized average atomic displacement, the local density correlation coefficient and the polymer chain connectivity. The possibility of computing these criteria using the omit-map procedure is also provided. The described software should be a valuable tool in monitoring the refinement procedure and in assessing structures deposited in databases.

1. Introduction

The body of data on the three-dimensional structures of macromolecules and, in particular, proteins has grown exponentially in recent years. 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' (EU BRIDGE Database Project Consortium; Gray *et al.*, 1996). Such databases can only be useful if the data they contain are consistent and as error-free as possible. This particularly applies 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 allow the definition of the atomic model of a macromolecule with sufficient accuracy. Atomic models provided by these techniques, therefore, represent a compromise between the fit to the

experimental data and to our chemical knowledge. 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.

Several such procedures have been developed in recent years. Packages such as *PROCHECK* (Laskowski, MacArthur *et al.*, 1993) and *WHAT-IF* (Hooft *et al.*, 1996), often used in the crystallographic community, focus on the validation of geometric and stereochemical parameters of the molecular models (*e.g.* covalent bonds and angles, main-chain and side-chain dihedral angles, geometry of chiral centres *etc.*). These procedures essentially evaluate how these parameters deviate from their standard values, which are derived from a reference set of high-quality protein structures or crystals of small molecules. However, X-ray refinement procedures and the derivation of models from NMR data often use the same parameters as constraints or restraints. For example, refinement algorithms such as *PROLSQ* (Hendrickson & Konnert, 1980; Konnert & Hendrickson, 1980), *TNT* (Tronrud *et al.*, 1987), *SHELX* (Sheldrick, 1995) and *REFMAC* (Murshudov *et al.*, 1997) use restraints on covalent geometry as well as steric restraints, whereas procedures such as *EREF* (Jack & Levitt, 1978) or *X-PLOR* (Brünger *et al.*, 1987) replace the steric restraints by more sophisticated energy functions. These restraints and constraints can leave their mark on the final model (Stewart *et al.*, 1990), and measuring the quality of a structure in terms of how well certain parameters match the standard values may, in fact, evaluate how different standard values compare with one another (Laskowski, MacArthur *et al.*, 1993; Laskowski, Moss *et al.*, 1993).

There is a clear need to supplement the stereochemical quality measures with procedures that evaluate the quality of the experimental data and the agreement of the derived atomic model with those data. In the case of crystal structures, the experimental data are 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 using routine measures such as the R factor or the free R factor (Brünger, 1992*a*). However, though these parameters, which apply to the model as a whole, are nearly always reported by the authors, they are not computed by a uniformly accepted algorithm and, therefore, cannot be meaningfully compared between structures. In addition, protein structures often have regions that are less reliably modelled than others. Although methods for evaluating the local agreement of the model with electron density on an atom or per-residue basis (Brändén & Jones, 1990; Jones *et al.*, 1991) are routinely used by crystallographers, the information they produce is only partially passed on to the deposited entries through the occupancy and B-factor parameters, or by the authors' comments in text form. *Ad hoc* methods are then needed to link this information to the atomic coordinates.

Here we describe *SFCHECK*, a stand-alone software package featuring a set of procedures for analysing and validating the deposited structure-factor data and for evaluating the agreement of the deposited atomic coordinates with those data, both for the model as a whole and on a per-residue basis.

Many of the quality measures and evaluation criteria used by *SFCHECK* are often computed in one form or another in existing structure-determination and refinement programs, but some have hitherto not been widely applied. *SFCHECK* applies these different measures to a given structure completely automatically and provides a concise pictorial output of the results as a PostScript file, in a manner similar to that of *PROCHECK*. This offers the opportunity of surveying and comparing the results obtained for a large number of structures using a unified set of criteria and allows the making of direct comparisons with the evaluations carried out by *PROCHECK* and other similar packages.

The present paper presents a detailed description of the tasks performed and the quality-assessment criteria computed by *SFCHECK* and illustrates its application to two protein structures and one nucleic acid structure. Results of extensive surveys of macromolecule structures performed with *SFCHECK* will be reported elsewhere.

2. The tasks performed by *SFCHECK*

The major tasks performed by *SFCHECK* are summarized in the flow-chart (Fig. 1). *SFCHECK* reads in the structure-factor data written in the macromolecular crystallographic information file (mmCIF) format (Bourne *et al.*, 1997) or in the file formats currently deposited in the PDB (Bernstein *et al.*, 1977). Given the diversity of the latter formats, human intervention is often necessary to process these files correctly. As for the atomic coordinates, these can be provided either in the PDB or mmCIF formats.

Next, *SFCHECK* analyses the structure-factor data (see below), generates an electron-density map from the atomic coordinates, computes $F_{\rm calc}$ using a fast Fourier transform (FFT) algorithm and scales the $F_{\rm calc}$ to $F_{\rm obs}$. It then uses FFT to compute two electron-density maps, with calculated phases and observed and calculated amplitudes, and calculates the gradients of the difference maps with respect to atomic coordinates. Furthermore, it compares the observed and calculated structure-factor amplitudes and computes various parameters which are used to evaluate the agreement between the observed and calculated electron densities in specific regions of the model.

In the following, we provide a detailed account of the structure-factor data analysis and the scaling procedures, and describe the different parameters and procedures used by *SFCHECK* in assessing the quality of the model as a whole and in specific regions.

2.1. Analysis of the structure-factor data

Having read the deposited structure-factor amplitudes, *SFCHECK* first of all provides relevant information concerning these data. This information is given in the

Structure Factors panel in Fig. 2(*a*), which displays the first page of the *SFCHECK* output for the cellular retinoic acid binding protein T (Kleywegt *et al.*, 1994) (PDB code 1CBS).

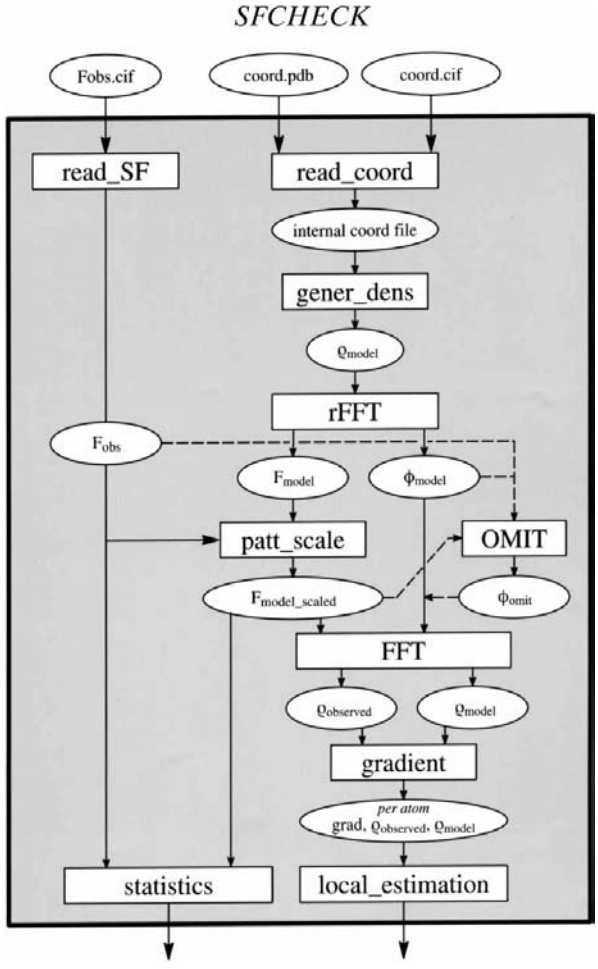


Figure 1
SFCHECK flowchart. The sequence of tasks performed by *SFCHECK* (rectangular boxes) and the corresponding input and output data (ovals) are displayed. The right-hand side of the flowchart depicts the tasks performed to evaluate the agreement between the model and the electron density. The left-hand side summarizes the processing of the observed structure-factor data. The program reads in (read_coord) the model atomic coordinates, in either the Protein Data Bank (PDB) or macromolecular Crystallographic information File (mmCIF) formats. It then generates an internal coordinates file (internal coord file). From this, it computes the model electron density (gener_dens), producing $\rho_{\rm model}$, which in turn is used to compute calculated structure factors ($F_{\rm model}$, $\varphi_{\rm model}$) using the fast Fourier transform (rFFT). Using $F_{\rm model}$ and $F_{\rm obs}$ (the deposited observed structure factors), Patterson scaling is performed (patt_scale) and scaled model structure factors ($F_{\rm model\_scaled}$) are produced. Using $\varphi_{\rm model}$, the omit procedure (OMIT) can be activated. The observed and scaled structure factors are then combined with the model phases, and Fourier transformed (FFT) to yield the observed and model electron densities, respectively. When the OMIT procedure is applied, the program uses $\varphi_{\rm omit}$ instead of $\varphi_{\rm model}$. At this stage, the program computes the gradients (gradient) of the $F_{\rm obs} - F_{\rm calc}$ map with respect to the atomic coordinates (grad), as well as the observed and model electron density at the atomic centres. The various criteria for assessing the local agreement of the model with the electron density are then computed (local_estimation). The structure-factor data (left-hand side) are read in mmCIF format. These are then processed and analysed (see text).

This PDB entry reports a high-resolution structure (1.8 Å) refined by the program *X-PLOR*. The entry lists the values of both the *R* factor and $R_{\rm free}$ parameters and is accompanied by a structure-factor file. It was therefore considered as an appropriate choice for illustrating the functionalities of *SFCHECK*.

The Structure Factors panel in Fig. 2(*a*) lists the following items: the minimum and maximum nominal resolution (*d* spacing) of the deposited structure-factor amplitudes, the total number of reflections in the file and the number with $|F_{\rm obs}| > \sigma$ and $|F_{\rm obs}| > 3\sigma$. It also gives the number of rejected reflections due to systematic absence and those with amplitudes ≤ 0. Next, it gives information on the reflections that are used in the analyses performed by *SFCHECK*. This comprises the nominal resolution range, in which the minimum *d* spacing (high-resolution) limit corresponds to that reported by the authors in the records REMARK 2 or 3 in the PDB coordinate file, and the total number of reflections used by the program that are within the minimum *d*-spacing limit.

To further assess the quality of the structure-factor data, the program computes five additional quantities: the completeness of the data, the structure-factor amplitude uncertainty, the overall *B* factor, the optical resolution and the expected optical resolution, which are detailed below. The numerical values of these quantities are given in the Structure Factors panel of Fig. 2(*a*). The corresponding graphical plots are shown on the second page of the *SFCHECK* output for the cellular retinoic acid binding protein (Fig. 2*b*). Together, the two types of information provide a comprehensive overview of the structure-factor data. *SFCHECK* can generate this information in the absence of data on atomic coordinates and could, therefore, be helpful in evaluating the structure-factor data during the very early stages of structure determination.

2.1.1. Completeness. Completeness in Fig. 2(*a*) refers to the experimental structure-factor data deposited by the authors, and is expressed as a percentage of the total number of reflections expected for the given crystal space group and minimal *d* spacing. A more detailed analysis of the data completeness is given in Fig. 2(*b*). The middle plot on the left-hand side of this figure displays completeness as a function of the *d* spacing, and the bottom plot shows a stereographic projection of completeness sampled along vectors in the reciprocal-space asymmetric unit. The latter plot is helpful in identifying regions of missing data and can be used as a guide during data collections.

2.1.2. Structure-factor amplitude uncertainty. The Structure Factors panel of Fig. 2(*a*) also lists the uncertainty of the structure amplitudes $R_{\rm stand}(F) = \langle\sigma(F)\rangle/\langle F\rangle$, where *F* is the structure-factor amplitude, $\sigma(F)$ is the structure-factor standard deviation and the brackets represent averages over the considered resolution range. The middle left-hand-side plot of Fig. 2(*b*) shows how $R_{\rm stand}(F)$ varies with *d* spacing.

2.1.3. Scaling $F_{\rm calc}$ to $F_{\rm obs}$. In *SFCHECK*, scaling of $F_{\rm calc}$ to $F_{\rm obs}$ is based on the Patterson origin peak (Rogers, 1965), which is approximated by a Gaussian. This peak is computed for both the observed and calculated amplitudes, and in each case the $B_{\rm overall}$ quantity is computed by

$$B_{\rm overall} = 8\pi^2\sigma_{\rm Patt}/2^{1/2},$$

where $\sigma_{\rm Patt}$ is the standard deviation of the Gaussian fitted to the Patterson origin peak.

The difference $B^{\rm diff}_{\rm overall} = B^{\rm obs}_{\rm overall} - B^{\rm calc}_{\rm overall}$ is then added to the calculated $B_{\rm overall}$ so as to make the width of the calculated Patterson origin peak equal to the observed peak.

The scale factor S is then calculated as

$$S = \{\sum(F_{\rm obs}\, f_{\rm cutoff})^2 / \sum[F_{\rm calc}\, f_{\rm cutoff} \exp(-B^{\rm diff}_{\rm overall} s^2)]^2\}^{1/2},$$

where $f_{\rm cutoff}$ is the function $f_{\rm cutoff} = 1 - \exp(-B_{\rm off}s^2)$, illustrated in Fig. 4, in which $B_{\rm off} = 4d^2_{\rm max}$ and s and d are the magnitudes of the reciprocal and real lattice vectors, respectively. The program always uses $B_{\rm off}$ = 256 Å^2, which corresponds to $d_{\rm max}$ = 8 Å.

The $f_{\rm cutoff}$ function is used to obtain a smooth cutoff for data with large d spacing. A large d-spacing cutoff is usually applied to remove the influence of disordered solvent molecules, which contribute to the diffraction at low nominal resolution (Tronrud, 1997). Removing these data entirely from the calculations produces series-termination effects, which introduce spurious peaks in the electron density at the surface of the macromolecule. Applying the so-called soft low-resolution cutoff to the structure factors, as performed here, significantly reduces these effects.

Finally, $F'_{\rm calc}$ is calculated as

$$F'_{\rm calc} = F_{\rm calc} S \exp(-B^{\rm diff}_{\rm overall} s^2).$$

This scaling scheme, together with the soft large d-spacing cutoff, is applied every time observed or calculated structure-factor amplitudes are computed. The Patterson origin peak-scaling method used by *SFCHECK* has advantages over conventional scaling by the Wilson plot (Wilson, 1949), particularly when only low-resolution data are available. The program computes overall B factors using both methods. The overall B factor derived from the Patterson origin peak scaling is listed in the Structure Factors panel in Fig. 2(*a*) and that obtained from the Wilson plot is listed under the Wilson plot (upper left-hand side of Fig. 2*b*).

2.1.4. Optical resolution. The optical resolution is defined as the expected minimum distance between two resolved peaks in the electron-density map. With the shape of the atomic peak being fitted by a single Gaussian, this minimum distance equals twice the standard deviation of the fitted Gaussian, or its width W.

W can be computed from the standard deviation $\sigma_{\rm Patt}$ of the Gaussian fitted to the Patterson origin peak,

$$W = (\sigma^2_{\rm Patt} + \sigma^2_{\rm sph})^{1/2},$$

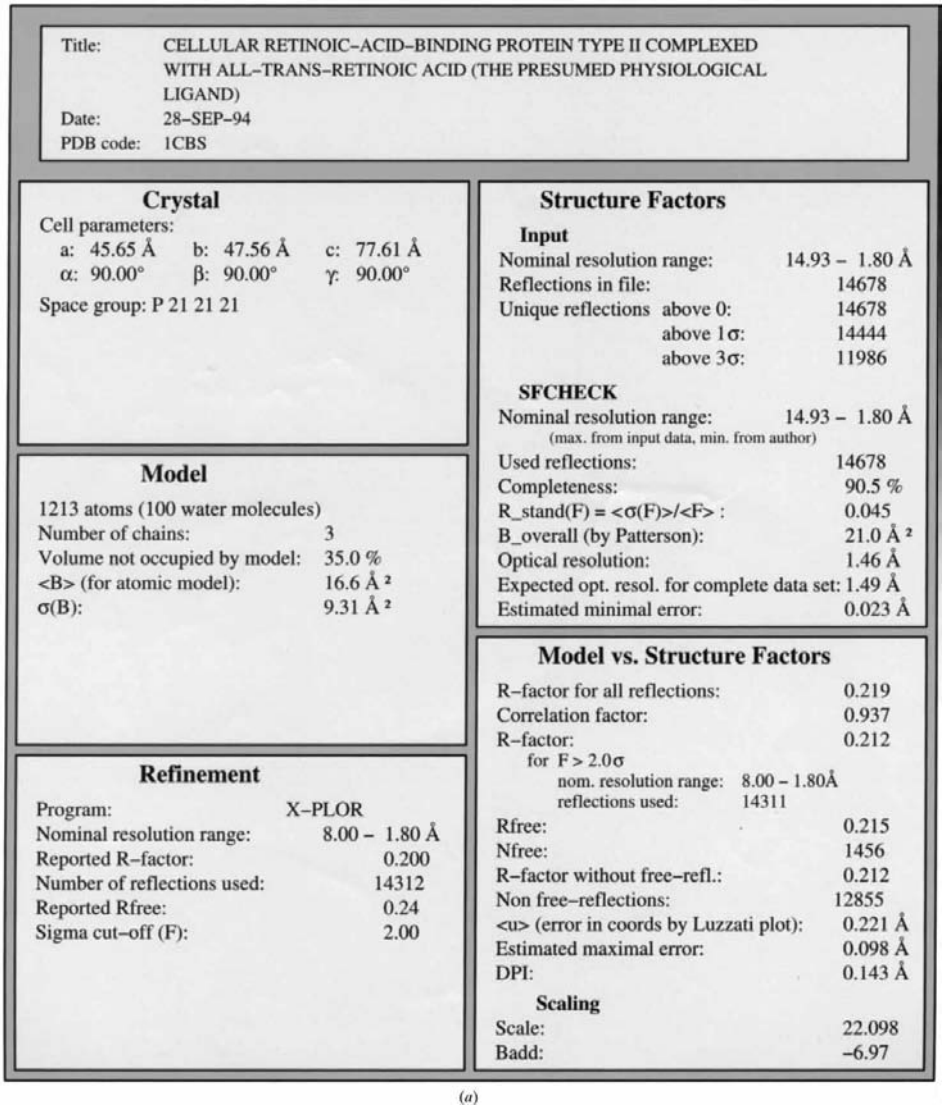
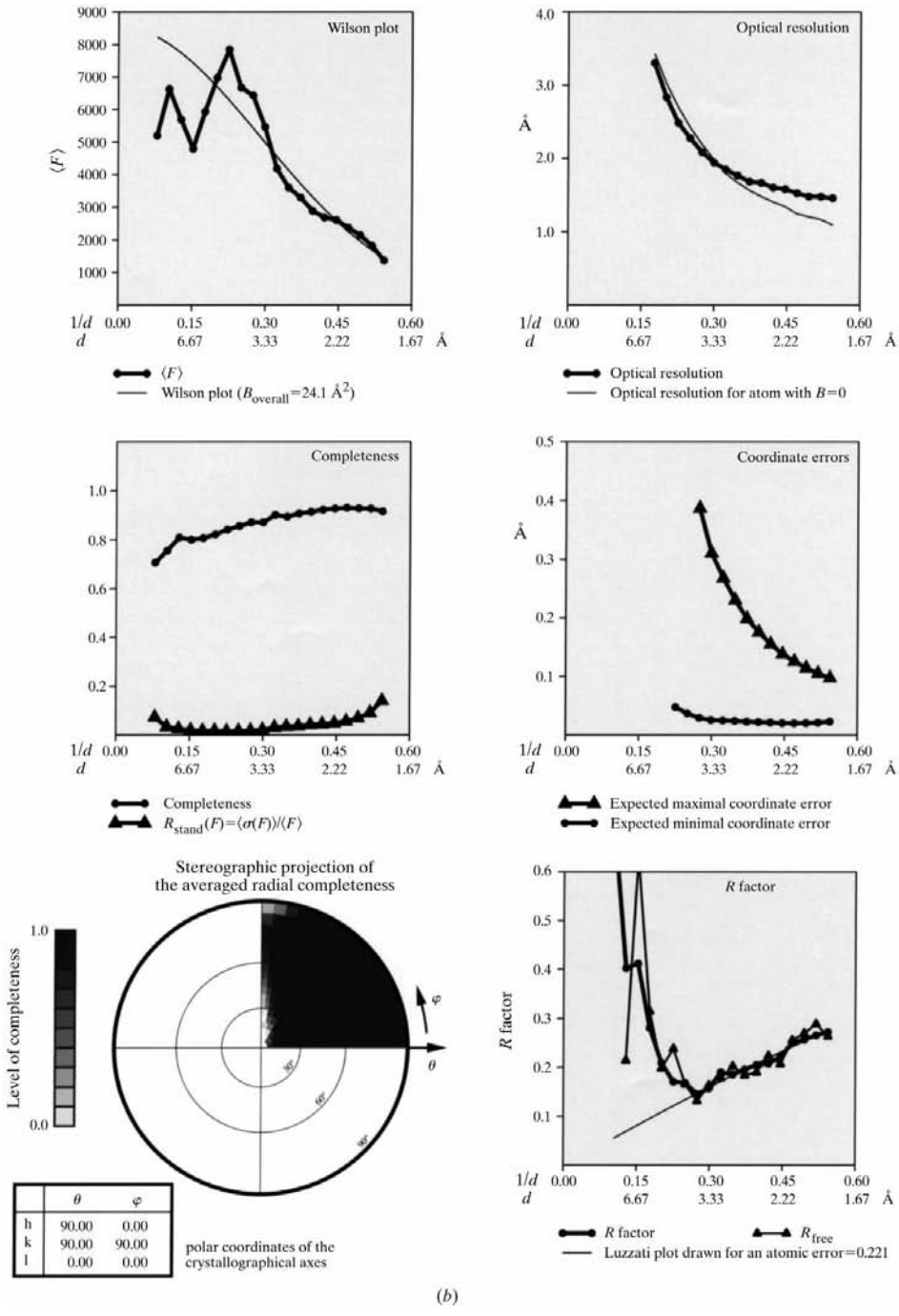


Figure 2
Typical *SFCHECK* output in PostScript is illustrated for the cellular retinoic acid binding protein (1CBS). (*a*) Summary panels displaying the numerical results from the analysis of the deposited structure-factor data and from the evaluation of the global agreement between the model and these data. The top elongated panel lists the PDB title record, deposition date and PDB code (1CBS). The Crystal panel summarizes the crystal parameters, provided by the authors, as read from the model input files. The Model and Refinement panels list the information provided by the authors on the model and the refinement procedure, respectively. This information is read from the PDB coordinates entry. The Structure Factor panel summarizes the information on the deposited structure-factor data (Input section) and on the data used and criteria computed by *SFCHECK* (SFCHECK section). The meanings of the various computed quantities in this panel are detailed in the text. The Model vs. Structure Factors panel summarizes the results of the verifications made by *SFCHECK*. The meanings of the various listed quantities are either self-explanatory or described in the text.



(*b*)

Figure 2 (continued)

(*b*) Graphical output from the *SFCHECK* analysis of global characteristics, of the structure-factor data and the model agreement with that data. From left to right and from top to bottom, it displays: the Wilson plot, the behaviour of the optical resolution as a function of the crystallographic resolution, the data completeness and structure-factor standard error as a function of the *d* spacing (nominal resolution), the maximal and minimal coordinate-error dependence on *d* spacing, the stereographic projection of the averaged radial data structure-factor data completeness and, finally, the *R*-factor dependence and Luzzati plots for a given atomic error.

This definition of optical resolution is closely related to the one used by Blundell & Johnson (1976) and James (1948) for point atoms. It takes into account factors such as errors in the data, atomic *B* factors and quality of the crystal, in addition to effects resulting from finite data. Plotting the optical resolution against the nominal resolution (upper right-hand plot in Fig. 2*b*) indicates how much the resolution of the electron-density map improves upon incorporation of reflections with smaller *d* spacing. The same plot in Fig. 2(*b*) also displays a second graph which represents the *d*-spacing-dependent behaviour of the optical resolution when the atomic *B* factors are set to zero. This latter graph depicts the contribution to the optical resolution arising solely from series termination.

Lastly, *SFCHECK* also calculates the expected optical resolution for the complete data set. This optical resolution is computed as described above, but using all the reflections. To approximate the amplitudes of missing reflections, the average value for the corresponding resolution shell is used. These reflections are not included in the *R*-factor calculations.

2.2. Global agreement between the model and the experimental data

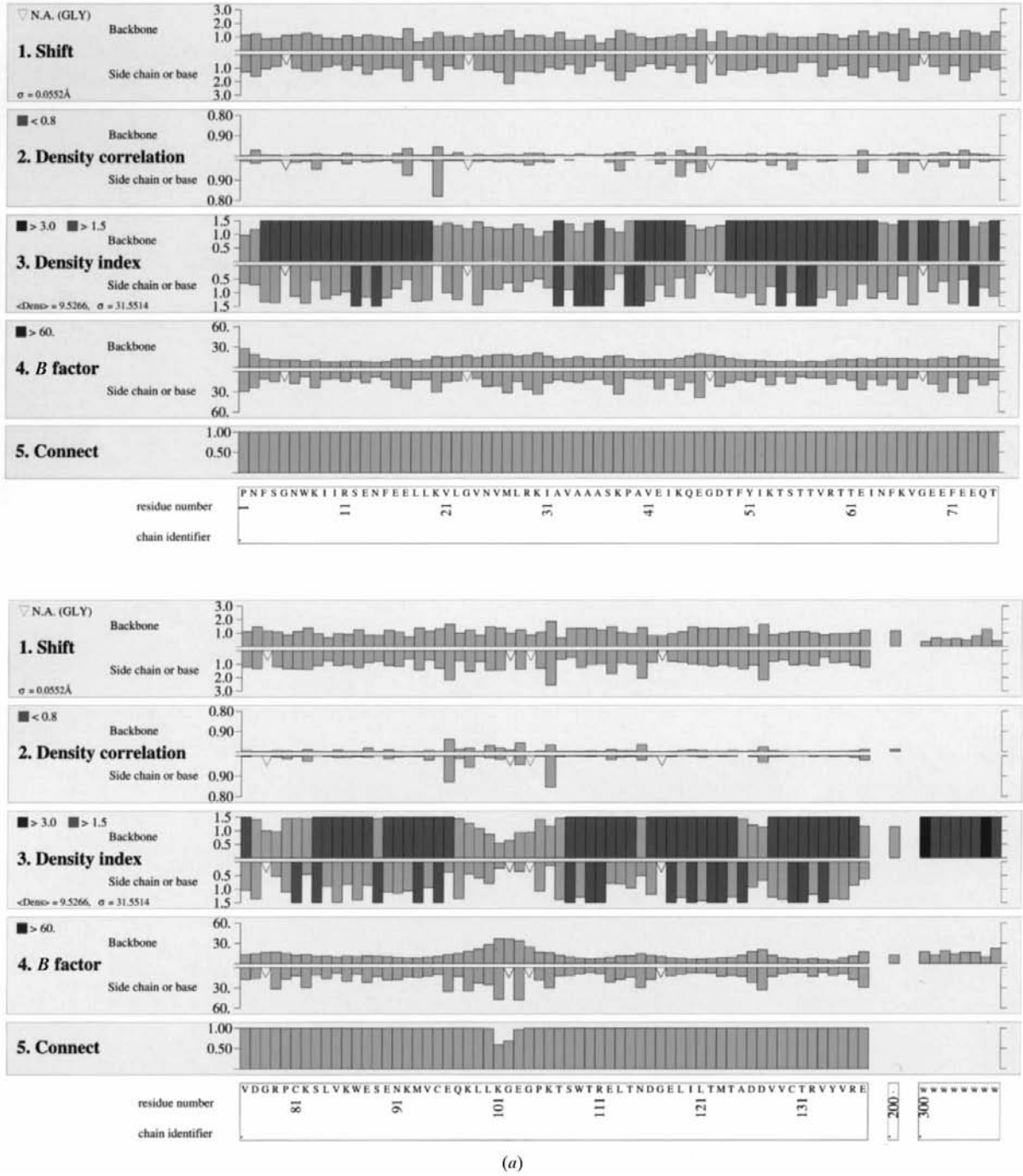
2.2.1. *R* factor, R_{free} and the correlation coefficient. To evaluate the global agreement between the atomic coordinates and the electron-density map, *SFCHECK* computes three well known parameters commonly used as indicators for the quality of X-ray structures of macromolecules. These are the classical *R* factor, the R_{free} (Brünger, 1992*a*) and the correlation coefficient CC_F between the calculated and observed structure-factor amplitudes,

$$CC_F = \frac{\langle F_{obs}F_{calc}\rangle - \langle F_{obs}\rangle\langle F_{calc}\rangle}{[(\langle F_{obs}^2\rangle - \langle F_{obs}\rangle^2)(\langle F_{calc}^2\rangle - \langle F_{calc}\rangle^2)]^{1/2}}.$$

The above quantity is computed taking into account all reflections within the reported high-resolution limit,

where σ_{sph} is the standard deviation of the Gaussian fitted to the origin peak of the spherical interference function, representing the Fourier transform of a sphere with radius $1/d_{min}$, with d_{min} being the nominal resolution (minimum *d* spacing). One can readily show that $\sigma_{sph} \simeq 0.356\ d_{min}$.

without any σ_{cutoff} and applying B_{off} = 256 Å^2. This allows comparisons across X-ray structures deposited by different authors. To permit validation of the R factor and R_{free} values deposited by the author, the program also computes these parameters considering only those reflections which the author indicated as being used in computing the reported quantity. Thus, the R factor is recomputed using the reported maximum d-spacing limit, σ_{cutoff}, and without B_{off}, whereas R_{free} is recomputed using only the specified structure-factor subset. However, some authors rightfully perform a final refinement run against all available data, precluding the validation of R_{free}. This appears to be the case for the retinoic acid binding protein 1CBS, as seen from the results in Fig. 2(*a*).



(*a*)

Figure 3
Typical *SFCHECK* output in PostScript for cellular retinoic acid binding protein. *SFCHECK* evaluation summary of the local agreement between the model and the electron density. Five criteria are plotted for each residue of the macromolecule (designated by the one-letter code), as well as for each solvent molecule (w) or heterogroup. These criteria are: (1) Shift, (2) Density correlation, (3) Density index, (4) *B* factor, (5) Connect. The definitions of these criteria are given in the text. Note that the values of the Connect parameter are truncted to a maximum of one. The *SFCHECK* outputs are generated using routines from *PROCHECK* kindly provided by R. Laskowski.

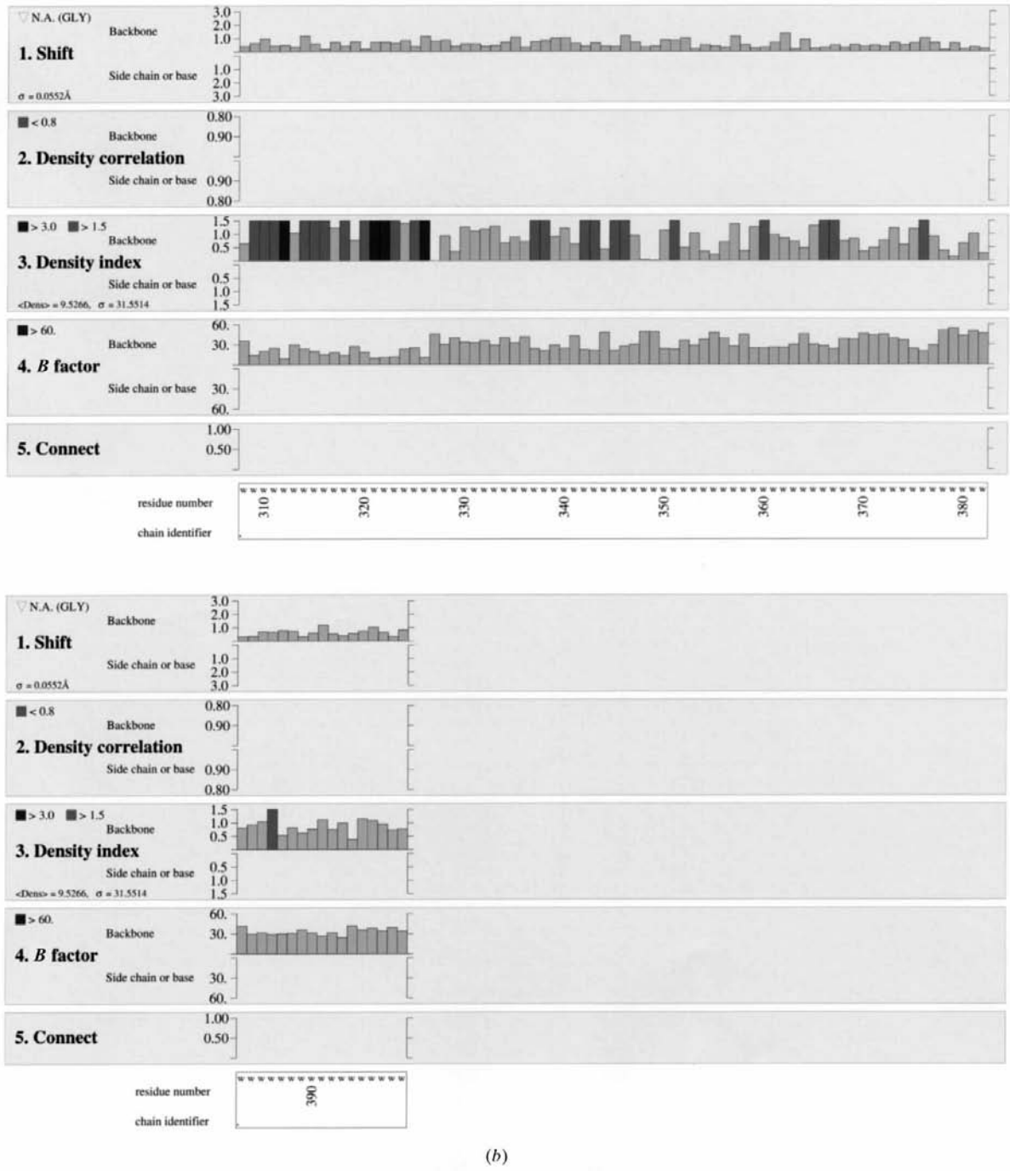


Figure 3 (continued)

Even in the ideal case, the values computed by *SFCHECK* could be expected to differ somewhat from those reported by the authors, owing to differences in the computational procedures. The most prominent of these is in the procedure used for scaling the calculated to the observed structure factors. This can lead to differences in R factor of up to 5% for low-resolution data sets or for models refined using anisotropic treatment (M. Fuxreiter *et al.*, unpublished results).

The values of the R factors and correlation factor computed by *SFCHECK* are listed in the Model *versus* Structure Factors panel of Fig. 2(*a*). The corresponding reported values are listed in the Model and Refinement panels of this figure. The behaviour of the R factor as a function of resolution is displayed in the bottom-right plot of Fig. 2(*b*), together with the Luzzati plot (Luzzati, 1952).

2.3. Estimation of errors in atomic positions

Estimating the errors associated with the atomic coordinates derived from a crystallographic experiment is an important aspect of the quality-assessment procedure. These errors, which are expressed as standard deviations (σ) of the

atomic positions are, however, not straightforward quantities to compute. This has led to a number of different error-evaluation methods.

Luzzati plots (Luzzati, 1952) have frequently been used in macromolecular refinement to estimate errors in atomic coordinates from the values of *R* factors. However, as pointed out by Cruickshank (1996), these plots in fact do not give an estimation of the error, but of how far the model is from convergence, and then only in the case of error-free data and functional forms of the model, neither of which conditions are met in macromolecular crystallography. *SFCHECK* generates Luzzati plots nevertheless, in order to allow comparisons with results obtained by other programs.

Another approach, developed by Cruickshank (1949), is based on the analysis of the accuracy of the electron-density map. Following up his original approach, Cruickshank recently derived a simpler expression for the coordinate error, termed the diffraction-data precision indicator (DPI), which depends on the *R* factor (Cruickshank, 1996). Murshudov & Dodson (1997) have extended this expression as a function of $R_{\rm free}$.

SFCHECK computes three measures of error in the atomic coordinates in addition to the Luzzati plot. One is the original error measure of Cruickshank (1949). The second is a modified version of this error measure, in which the difference between the observed and calculated structure factors is replaced by the error in the experimental structure factors (as described below). The first two error measures are termed the expected maximal and minimal errors. The third error measure is the DPI. It must be mentioned that all three error measures assume that geometric restraints are not used, a condition which is never met in macromolecular refinement.

2.3.1. Expected maximal error. Following the method of Cruickshank (1949), the standard deviation of the atomic coordinates can be derived from the properties of the electron-density map,

$$\sigma(x) = \sigma(\text{slope})/\text{curvature},$$

where σ(slope) and curvature are the standard deviation of the slope and curvature of the electron-density map at the atomic centre in the x direction. For spherically symmetric peaks, $\sigma(x) \simeq \sigma(y) \simeq \sigma(z)$.

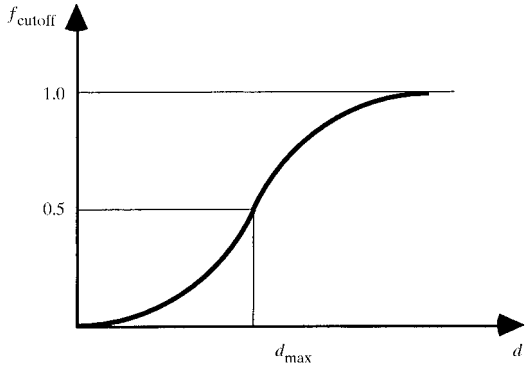


Figure 4
The function applied to obtain a smooth cutoff for high d spacing (low nominal resolution) structure-factor data. $f_{\rm cutoff}$ is applied to reduce the series-termination effects produced when a step function is used to eliminate high d-spacing data. The cutoff function varies between zero and one, in the manner shown, and equals 0.5 at $d_{\rm max}$.

σ(slope) is expressed as a function of the difference between $F_{\rm obs}$ and $F_{\rm calc}$,

$$\sigma(\text{slope}) = 2\pi\{\textstyle\sum[h^2(F_{\rm obs} - F_{\rm calc})^2]\}^{1/2}/aV_{\rm unit\_cell},$$

where a is the unit-cell length and h the Miller index.

Lipson & Cochran (1966) have shown that refinement by the method of least squares is equivalent to refinement based on difference synthesis with weights $1/f$, where f is the average atomic form factor. *SFCHECK* uses the diagonal terms of the second-derivative matrix of the least-squares equations to approximate the curvature (Agarwal, 1978; Murshudov *et al.*, 1997),

$$\text{curvature} = 2\pi^2(\textstyle\sum h^2F_{\rm obs})/a^2V_{\rm unit\_cell}.$$

For the missing reflections, the program uses the average value of $\sigma(F)$ for the corresponding resolution shell instead of $(F_{\rm obs} - F_{\rm calc})$ (see below).

2.3.2. Expected minimal error. The expected minimal coordinate error is estimated using the experimental $\sigma(F)$ instead of the difference between the observed and calculated structure factors. In this case σ(slope) is computed as

$$\sigma(\text{slope}) = 2\pi\{\textstyle\sum[h^2\sigma(F)^2]\}^{1/2}/aV_{\rm unit\_cell},$$

where a is the unit-cell length and h the Miller index.

If there are no σ values for the observed structure factor, the program uses $\sigma = 0.04F_{\rm obs}$ as the default value, which is roughly the error magnitude usually encountered.

2.3.3. DPI, diffraction-data precision indicator. DPI is the atomic coordinate error estimated by the method of Cruickshank (1996),

$$\sigma(x) = \left(\frac{N_{\rm atoms}}{N_{\rm obs} - 4N_{\rm atoms}}\right)^{1/2} c^{-1/3} d_{\rm min} R_{\rm factor},$$

where c is the structure-factor data completeness expressed as a fraction (0 to 1), $R_{\rm factor}$ is the conventional crystallographic R factor, $N_{\rm obs}$ is the number of reflections and $d_{\rm min}$ is the minimal d spacing. This expression assumes that at least four reflections per atom were included in the refinement. It thus works well for structures with resolution higher than ~2 Å, but is not valid for lower resolution structures, where the number of observables per atom is lower.

When the reflections used in $R_{\rm free}$ calculations are flagged, the program computes the $R_{\rm free}$ factor and uses the modified DPI expression (Murshudov & Dodson, 1997),

$$\sigma(x) = \left(\frac{N_{\rm atoms}}{N_{\rm obs}}\right)^{1/2} c^{-1/3} d_{\rm min} R_{\rm free}.$$

Note that the $R_{\rm free}$-based DPI works well at both high and low resolution.

2.4. Local agreement between the model and the electron density

To evaluate the local agreement between the model and the electron-density map, *SFCHECK* computes five measures for

each residue along the polymer chain (the program handles proteins and nucleic acids), as well as for groups of atoms such as the solvent molecules, whole ligands (when they are small) or portion of ligands defined as separate units in the input file. Residue-based measures of the local agreement between the model and the electron-density map have previously been shown to be very useful (Brändén & Jones, 1990; Jones *et al.*, 1991).

The following provides a detailed description of the various local quality measures computed by *SFCHECK*.

2.4.1. The normalized average displacement. The normalized average displacement of atoms, shift, is computed for each residue by

$$\text{shift} = (1/N\sigma)\sum_{i}^{N}\Delta_i$$

with

$$\Delta_i = \text{gradient}_i/\text{curvature}_i,$$

where gradient_i is the gradient of the $(F_{\text{obs}} - F_{\text{calc}})$ map with respect to the atomic coordinates and curvature_i is the curvature of the model map computed at the atomic centre (see Agarwal, 1978). N is the number of atoms in the considered group and σ is the standard deviation of the Δ_i values computed in the structure.

The quantity shift, expressed in units of σ, indicates the tendency of the considered group of atoms to move away from their current position, with large values of shift corresponding to regions where this tendency is high.

2.4.2. Density correlation. The electron-density correlation coefficient, D_{corr}, for a given group of atoms is calculated as

$$D_{\text{corr}} = \frac{\sum \rho_{\text{calc}}(x_i)[2\rho_{\text{obs}}(x_i) - \rho_{\text{calc}}(x_i)]}{\left([\sum \rho^2_{\text{calc}}(x_i)]\{\sum[2\rho_{\text{obs}}(x_i) - \rho_{\text{calc}}(x_i)]^2\}\right)^{1/2}},$$

where $\rho_{\text{calc}}(x_i)$ and $\rho_{\text{obs}}(x_i)$ are, respectively, the electron density computed from calculated and observed structure-factor amplitudes at the atomic centre. The summation is performed over all the atoms in the considered group. For polymer residues, D_{corr} is computed separately for backbone and side-chain atoms.

The value of the electron density at a given atomic position $\rho(x_a)$ is computed as

$$\rho(x_a) = \{\sum_i[\rho(x_i)\rho_{\text{atom}}(x_i - x_a)]\}/\sum_i \rho_{\text{atom}}(x_i - x_a),$$

where ρ is any electron density and ρ_{atom} is the atomic electron density, x_a is the vector of the atom centre and x_i is the vector of the ith grid point. The sum is taken over all grid points within a distance d_{lim} = 2.5 Å from the atomic centre. ρ_{atom} implies a weighting scheme which reduces the dependence of the computed density correlation on the value of d_{lim} and differentiates our approach from other published methods (Jones *et al.*, 1991).

Small values of D_{corr}, depicted by tall bars in Fig. 3(*a*), indicate that the model of the corresponding backbone or side chain agrees poorly with the electron density.

2.4.3. The residue-density index. The residue-density index is expressed as

$$\text{density index} = [\prod \rho(x_i)]^{1/N}/\langle\rho\rangle_{\text{all\_atoms}},$$

where N is the total number of considered atoms in the side chain or backbone groups, $[\prod \rho(x_i)]^{1/N}$ is the geometric mean of the $(2F_{\text{obs}} - F_{\text{calc}})$ electron density of the considered atom subset and $\langle\rho\rangle_{\text{all\_atoms}}$ is the average electron density of the atoms in the structure. For water molecules or ions which are represented by a unique atom, the above expression reduces to the ratio $\rho(x_i)/\langle\rho\rangle_{\text{all\_atoms}}$.

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 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.

2.4.4. The average *B* factor per residue. This quantity is computed as the average of the atomic B factors of the backbone and side-chain atoms of each residue. Comparison of the B-factor and density index plots can be useful for detecting regions with errors in the model. It would be expected 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, problems with either the model or the refinement procedure may be suspected.

2.4.5. The connectivity index. The connectivity index, connect, is the same quantity as the residue-density index, but computed for the backbone atoms excluding the carbonyl O atoms in proteins, and considering the P, O5′, C5′, C3′ and O3′ atoms in nucleic acids. Connect measures the level of the electron density along the macromolecule skeleton and can be used to assess the continuity of the electron density along the polymer chain. Low levels of the connect index indicate 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.

2.5. Omit procedure

An omit map is a way to reduce the model bias in the electron density calculated with model phases (Bhat, 1988). *SFCHECK* produces the so-called total omit map by an automatic procedure. First, the initial $(F_{\text{obs}}, \varphi_{\text{model}})$ map is divided into N slightly overlapping boxes. For each box, step by step, the electron density in it is set to zero and new phases are calculated from this modified map. A new map is then computed using these phases and F_{obs}, and regions of the map delimited by the current box are stored. This procedure is repeated for all boxes, yielding the total omit map. Phases calculated from this total map are combined with the initial model phases, using the Sim weighting scheme (Sim, 1960). This entire procedure may be repeated N times, but becomes rather time consuming as the value of N increases.

3. Results and discussion

In this section we illustrate a typical application of *SFCHECK* to two protein structures from the PDB and a specific application that uses the omit-map option of *SFCHECK*.

In the typical application we present the results obtained by *SFCHECK* for the cellular retinoic acid binding protein (Kleywegt *et al.*, 1994; PDB code 1CBS), representing an example of a good-quality model derived from high-quality data at high resolution (1.8 Å), and for the structure of the HIN recombinase (DNA-binding domain C)–DNA complex (Feng *et al.*, 1994; PDB code 1HCR), taken as an example of a poorer quality structure still in the process of being refined.

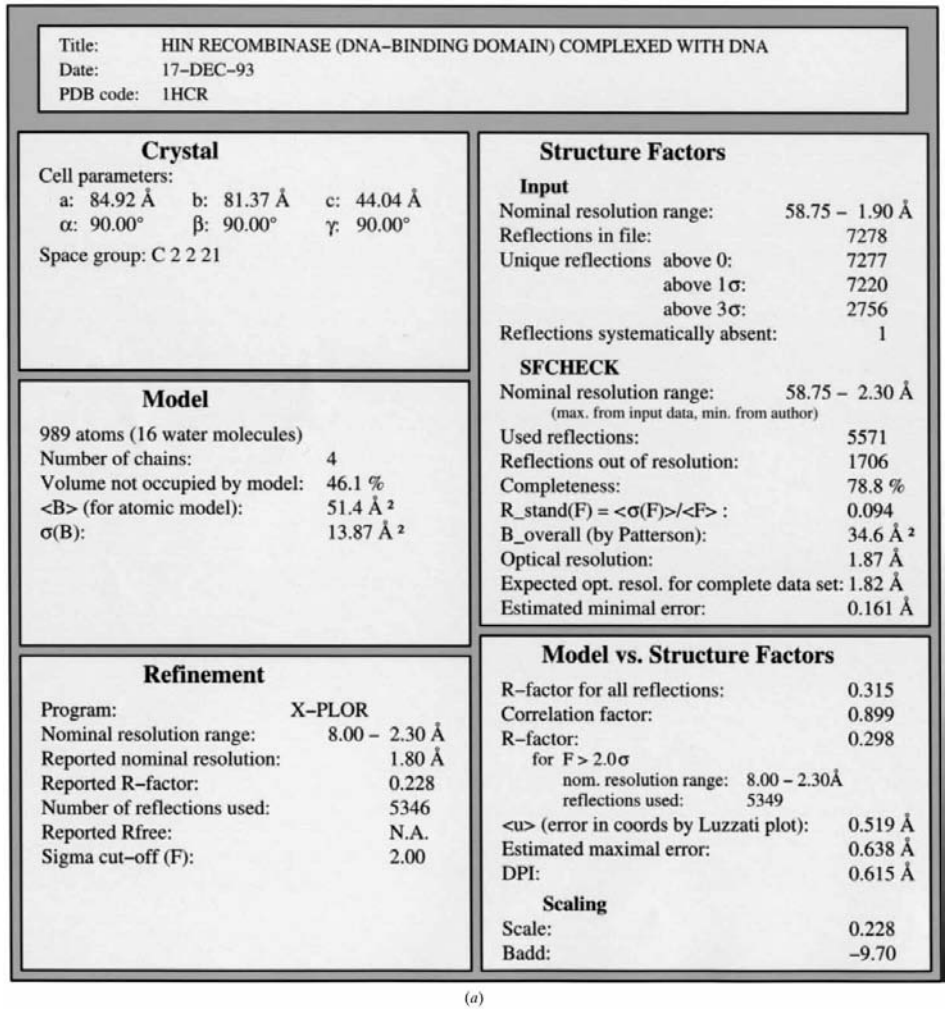
3.1. *SFCHECK* analysis of the cellular retinoic acid binding protein (1CBS)

Figs. 2 and 3 summarize the *SFCHECK* analysis of 1CBS. Fig. 2(*a*) displays the numerical results from the analysis of the structure-factor data and from the evaluation of the global agreement between the model and the data, as discussed in §2. We see that the *R*-factor and R_{free} values computed by *SFCHECK* (Model vs. Structure Factors panel) using the identical reflection subset to that reported by the authors (Refinement panel) show small differences to the reported values. The latter are 0.20 and 0.24 for the *R* factor and R_{free}, respectively, whereas the corresponding *SFCHECK* values are 0.212 and 0.215, respectively. Fig. 2(*b*) shows that the R_{free} and *R* factor display a similar resolution dependence. The fact that the R_{free} and *R*-factor values show negligible differences is most probably due to the fact that the authors performed a final run of refinement using all the reflections and, therefore, say nothing about the quality of the model. The output also validates various other numerical values reported by the authors. It shows, for example, that the resolution range of the deposited structure-factor data (14.93–1.8 Å) is consistent with the reported resolution of the model (1.8 Å).

The information in Figs. 2(*a*) and 2(*b*) also allows some judgement about the quality of the structure-factor data for this protein. We see that the relatively high resolution of this structure (1.8 Å) is accompanied by a somewhat limited data completeness (only 90.5%). This is confirmed by the completeness plot in Fig. 2(*b*). The $R_{stand}(F)$ plot on the same graph shows, furthermore, a decrease in quality of the high-resolution data (2–1.8 Å). The average radial completeness plot (bottom-right plot in Fig. 2*b*) indicates that the incomplete data primarily concerns reflections in the *KL* plane.

The effective resolution computed with the complete data set (1.49 Å) is higher than with the actual data set (1.46 Å). This is most probably due to the incomplete data at low resolution (middle left-hand-side plot in Fig. 2*b*).

Fig. 3(*a*) presents the *SFCHECK* analysis of the local agreement of the model with the electron density for 1CBS. The shift plot shows that both backbone and side-chain shifts are low (0.075 Å or less) throughout, with only a few shifts reaching 0.11 Å (residues 17, 27, 97, 105). The density correlation is excellent across the entire molecule, except for a few Lys and Glu side chains (20, 96 and 106) which display poor correlation. These side chains are thus poorly defined in the electron-density map. The density index remains high



(*a*)

Figure 5
SFCHECK output in PostScript for the HIN recombinase DNA-binding domain–DNA complex (1HCN). (*a*) Summary panels displaying the numerical results from the analysis of the deposited structure-factor data and from the evaluation of the global agreement between the model and these data. (*b*) Graphical output from the *SFCHECK* analysis of global characteristics of the structure-factor data and model agreement with that data.

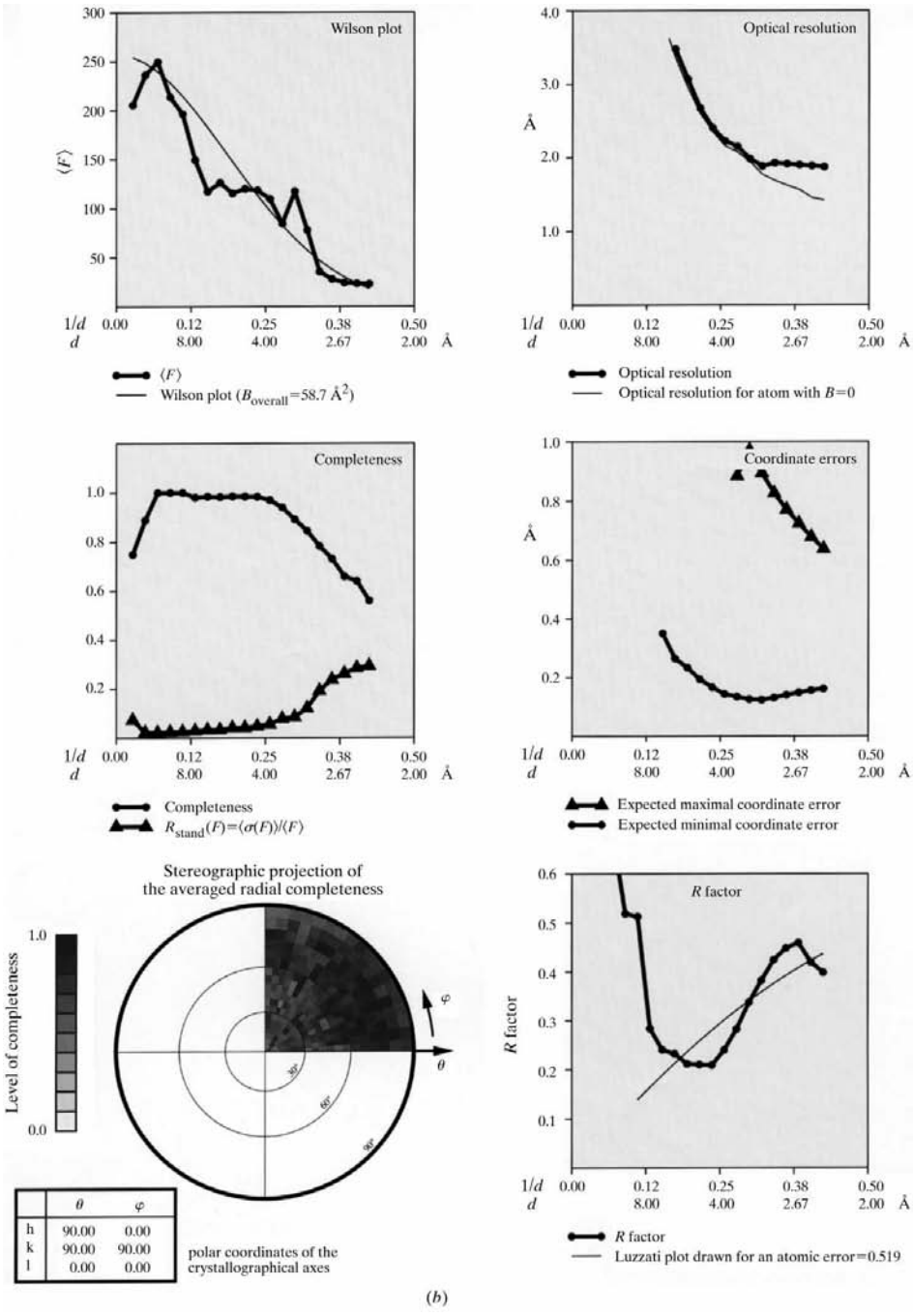


Figure 5 (continued)

throughout the backbone, except in the regions between residues 98–107. It is significantly lower for the side chains. Low values for the side-chain density index are observed mainly in loop regions delimited by residues 21–31, 41–51 and 98–107. The segment 98–107 also displays significantly higher backbone and side-chain *B* factors, as well as poorer backbone connectivity. Residues 101 and 102 are the only ones to have a connect value <1 (Fig. 3*a*).

Water molecules (labelled w in the *SFCHECK* output) are also evaluated. The relevant plots for these molecules are those of the shift, density index and *B*-factor parameters. We see that the first 25 or so water molecules in the list (appearing sequentially along the plot from left to right) display consistently the highest density index and lowest *B* factors (∼20 Å^2). They thus seem to be more reliably positioned than subsequent molecules, whose density indices sometimes drop perilously. A steady increase in *B* factor is also apparent as one goes down the water molecules list. Considering that the place of these molecules in the list probably reflects the refinement stage at which they were added, with molecules lower in the list corresponding to those added at later stages, we can conclude that the reliability of such water positions is generally poorer.

Analysis of the density index and *B* factors for individual water molecules by *SFCHECK* furthermore provides valuable information on the reliability of atomic coordinates of these molecules, which should be a very useful guide in any survey investigating the properties of crystallographic water molecules and their interactions with protein atoms.

3.2. *SFCHECK* analysis of the HIN recombinase DNA-binding-domain–DNA complex

Figs. 5 and 6 summarize the results of the *SFCHECK* analysis performed on the structure of the HIN recombinase (DNA binding domain C)–DNA complex (1HCR). This structure, declared by its authors to represent a 'preliminary coordinate set' with 'refinement still in progress', is clearly more problematic. This is confirmed by a quick inspection of Figs. 5(*a*) and 5(*b*), which summarize the global assessments. Indeed, it allows the identification of several features which could be the source of difficulties in the structure determination. We see, for example, that the behaviour of the *R*-factor distribution at high resolution is unusual (lower right-hand-side plot) and that even though the reported resolution is 1.8 Å, the *F*s are very weak beyond 3.0 Å resolution (upper left-hand-side plot). The high $R_{stand}(F)$ values at resolutions higher than 3.0 Å and the concomitant drop in completeness (middle panel of Fig. 5*b*) directly confirm the poor quality of the diffraction data at those resolutions. Inspection of the stereographic projection plot (bottom left of Fig. 5*b*) shows that the level of data completeness is rather poor and similar

throughout the reciprocal space, indicating that it is most probably limited by the quality of the crystal rather than by problems with data collection. The poorer quality of the X-ray data yields high values for the maximal and minimal coordinate errors and is the reason for the unusual resolution-dependent behaviour of the *R* factor.

We see that *SFCHECK* can also be helpful in validating numerical information provided by the authors or annotations made by database curators. For example, there is a discrepancy between the reported *R* factor (0.228) (see Model panel of Fig. 5*a*) and those computed by *SFCHECK*. The latter pertains to the *R* factor computed considering all acceptable reflections (0.315) (Model vs. Structure Factors panel in Fig. 5*a*) or considering only the reflection subset allegedly used by the authors in the refinement ($F > 2\sigma$ and resolutions between 8.0 and 2.3 Å) (0.298). The origin of this discrepancy is not clear.

The *SFCHECK* results of the analysis of the local agreement between the model and the electron density of 1HCR are displayed in Fig. 6. The plots in this figure reveal

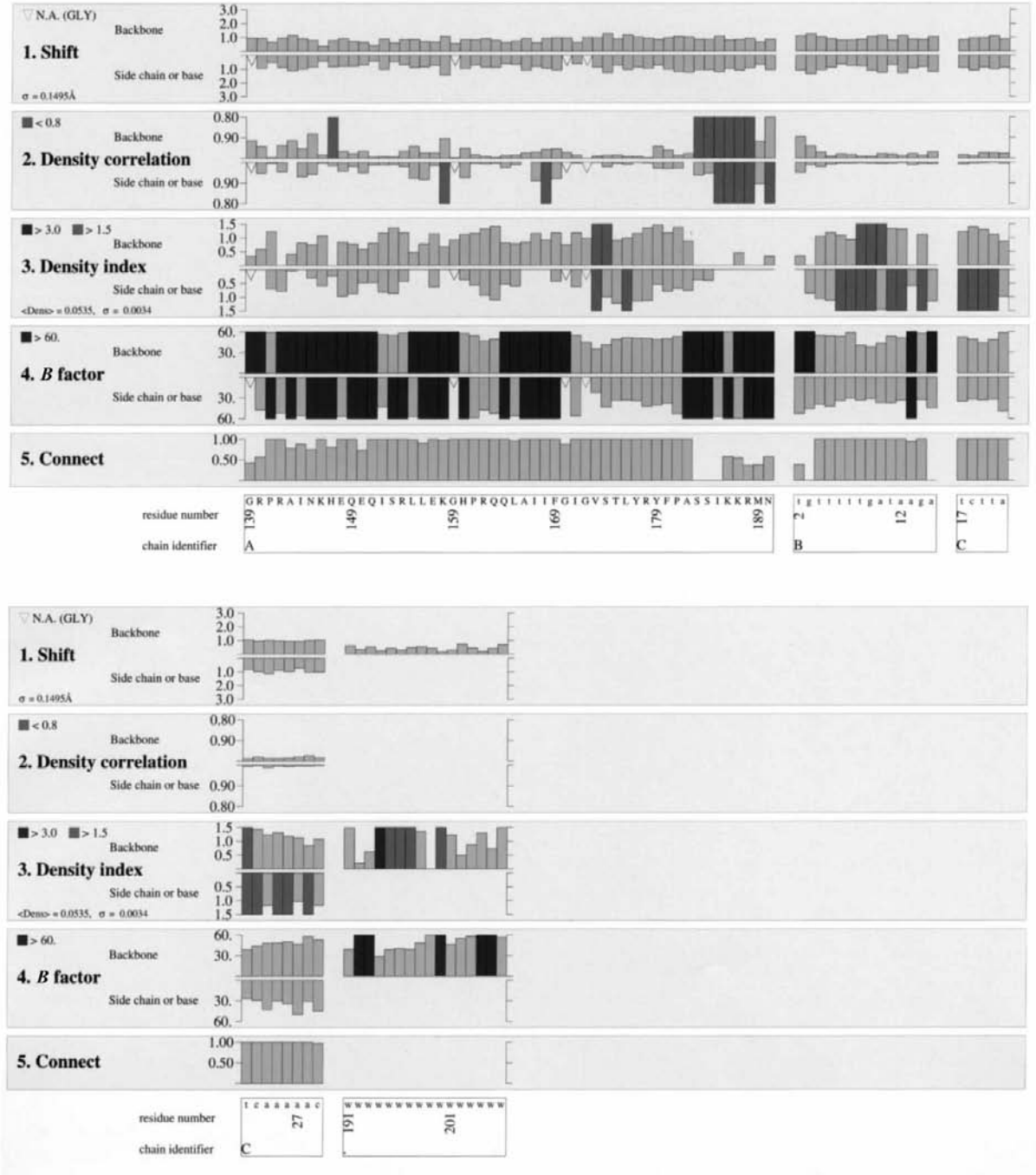


Figure 6
*SFCHECK*ouput in PostScript for HIN recombinase DNA-binding domain–DNA complex (1HCR). *SFCHECK* evaluation summary of the local agreement between the model and the electron density. See legend of Fig. 3 and the text for further details concerning the definitions of the quantities listed and plotted.

several features which clearly suggest that the refinement of the proposed model 'is still in progress'. The shift values are rather large (>0.13 Å) for both the backbone and side-chain atoms, indicating that the refinement has not converged. In addition, the *B* factors are very high for most backbone atoms, generally exceeding 60 Å^2, as witnessed by the large number of black rectangles in the plot. Since the density index of many residues is quite low in both the main chain and side chains, the large *B* values could result from attempts by the refinement program to fit a model into low-density regions.

Interestingly, the connect values are rather high throughout, except at residues 183–185 and residue 2 of the first DNA strand, which have values of zero. Since the connect parameter measures the density level for a subset of the polymer-backbone atoms, this indicates that these atoms tend to lie outside the electron density, suggesting a possible chain-tracing problem.

The density index and *B* factors of the water molecules indicate that the positions of molecules 192, 193, and 199 are probably incorrect, as they are characterized by a low density index and high *B* values (60 Å^2 or larger).

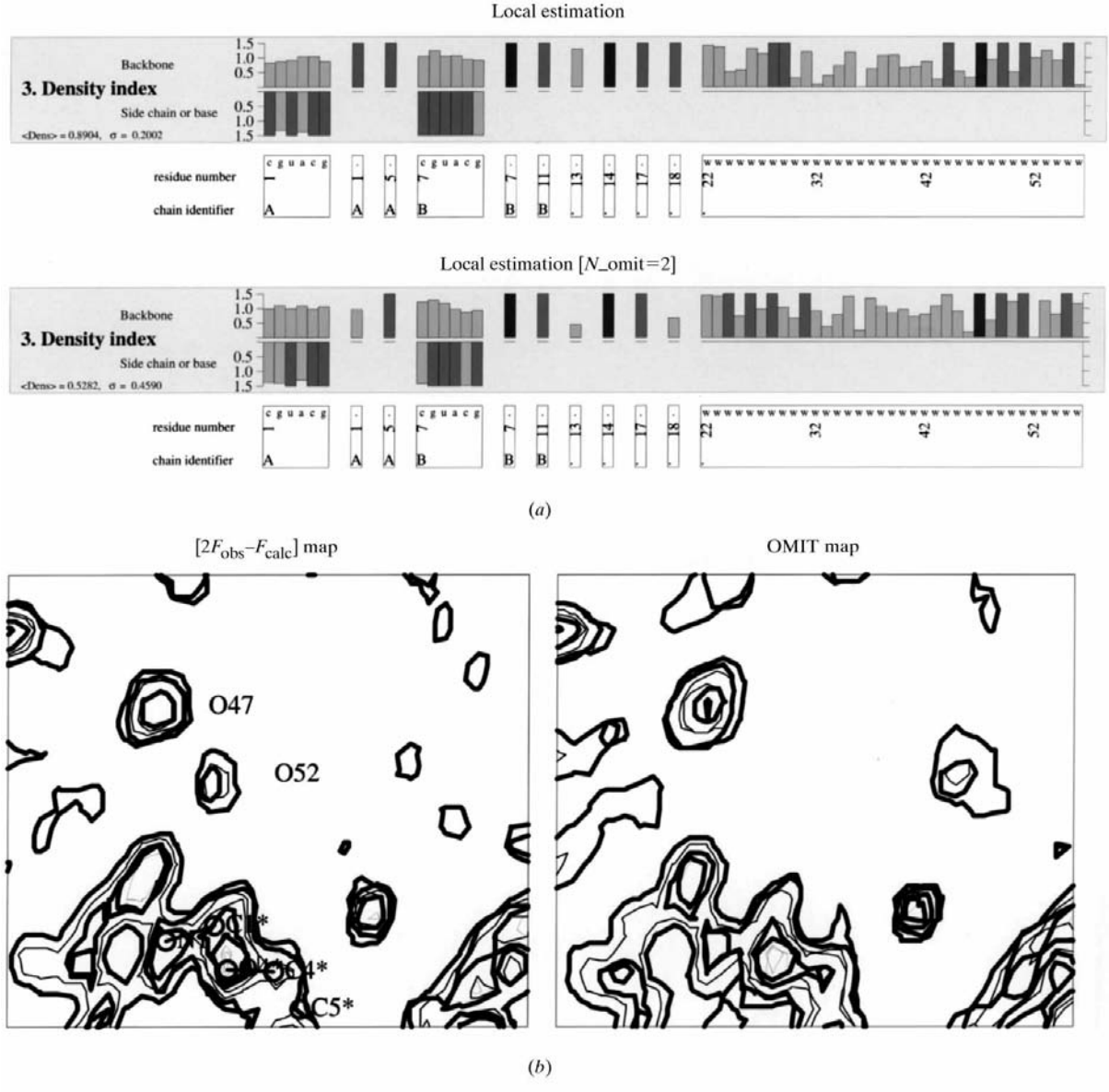


Figure 7
Sensitivity of the omit procedure in the evaluation of the local agreement between the model and the electron density, performed by *SFCHECK*, illustrated for the Z-DNA structure (PDB code 1D40; NDB code ZDFB10; Geierstanger *et al.*, 1991). (*a*) The residue or group-based density correlation, density index. *B* factor and connect (chain connectivity), computed considering the regular electron densities (see legend of Figs. 2 and 3 and text). Below these are plotted the density correlation, density index and connect parameters computed using the OMIT procedure (see text). (*b*) Sections of the electron density in the region of water molecule 52, whose density index in the OMIT map is zero. On the left-hand side are the sections of the regular electron density, and on the right-hand side those of the OMIT maps, where the density for water 52 is clearly missing.

We thus see that a quick glance at the *SFCHECK* output 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 other available tools.

3.3. Use of the OMIT-map procedure to ascertain problem regions

This procedure can, for example, be used to investigate three types of potential problems: (i) incorrect placement of water molecules, (ii) incorrect positions for atoms in part of the model and (iii) errors in the chemical structure (polymer sequence).

Fig. 7 displays the conventional *SFCHECK* density index together with that calculated from two cycles of the OMIT procedure (see §2) for the DNA-Z structure (PDB code 1D40; NDB code ZDFB10; Geierstanger *et al.*, 1991). Comparison of the two density index plots (Fig. 7*a*) shows, for example, that in the OMIT map the density index for water molecule 52 drops to zero, suggesting that this water molecule was probably positioned incorrectly. Inspection of the corresponding region of the electron-density maps, displayed in Fig. 7(*b*), confirms this suggestion. The map on the left-hand side of this figure displays the normal $2F_{obs} - F_{calc}$ density in this region, together with the corresponding atoms of the model. In this map, water molecule 52, as well as other depicted atoms, lie in the density, whereas in the OMIT map (on the right-hand side of Fig. 7*b*) no electron density appears at the corresponding water position.

The OMIT map procedure is thus much more sensitive to possible errors in the model than the conventional map. However, its computation is unfortunately still too time consuming at present to be applied systematically to a large number of structures.

4. Concluding remarks

In this paper, we have presented the program *SFCHECK*, which collates a number of objective criteria for measuring the quality of the X-ray data and assessing the agreement between the model and those data. It is geared to analyse protein and nucleic acid crystal structures which also contain water and other small molecules. We illustrate how the analysis made by *SFCHECK* can be used to readily evaluate the model as a whole, or in specific regions corresponding to residues or groups of atoms. The latter task, in particular, is notoriously difficult, and the criteria proposed here by *SFCHECK* should be considered only as a starting point for more detailed analyses. Such analyses could in some cases involve limited or extensive re-refinement of the structure, which may be required not only to correct detected problems, but also to make a correct diagnosis of what these problems may be. In this regard, *SFCHECK* analysis should be considered as giving useful hints, at best.

SFCHECK is thus a useful complement to validation procedures based on geometric and stereochemical criteria, such as *PROCHECK* or *WHAT-IF*, which do not take into account the X-ray data, but provide extremely valuable insights into how the features of a given model compare to those derived from other known structures, and may detect biases introduced in the model through the process of structure determination.

Presently, the main bottleneck to the generalization of procedures such as *SFCHECK* is that diffraction data are not available for most of the structures in the PDB. However, when, as many of us hope, the deposition of these data becomes routine, structure-validation protocols will most likely combine geometry/stereochemistry and X-ray based quality-assessment procedures.

SFCHECK is available from the authors upon request.

All members of the European Consortium on Structure Validation are thanked for stimulating discussions. Special thanks go to R. Laskowski and J. Thornton for letting us have the *PROCHECK* PostScript graphical routines. Our thanks go also to H. Berman and colleagues and to E. Dodson for critical comments on the *SFCHECK* content and output. We gratefully acknowledge support for this work from the following sources: the European BIOTECHNOLOGY project BIO2-CT92-0524 on three-dimensional macromolecular structure validation, the Department of Energy (USA), the Belgian programme of Inter-University 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

Agarwal, R. C. (1978). *Acta Cryst.* A**34**, 791–809.
Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanoushi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
Bhat, T. N. (1988). *J. Appl. Cryst.* **21**, 279–281.
Blundell, T. L. & Johnson, L. N. (1996). *Protein Crystallography.* London: Academic Press.
Bourne, P. E., Berman, H. M., McMahon, B., 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.
Brünger, A. T. (1992*a*). *Nature (London)*, **355**, 472–474.
Brünger, A. T. (1992*b*). *X-PLOR Manual, Version* 3.1. New Haven: Yale University Press.
Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.
Cruickshank, D. W. J. (1949). *Acta Cryst.* **2**, 65–82.
Cruickshank, D. W. J. (1996). *Macromolecular Refinement: Proceedings of the CCP4 Study Weekend*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 11–22. Warrington: Daresbury Laboratory.
Feng, J. A., Johnson, R. E. & Dickerson, R. E. (1994). *Science*, **263**, 348–352.
Geierstanger, B. H., Kagawa, T. F., Chen, E. L., Quigley, P. S. & Ho, P. S. (1991). *J. Biol. Chem.* **266**, 20185–20191.
Gray, P. M. D., Kemp, G. J. L., Rawlings, C. J., Brown, N. P., Sander, C., Thornton, J. M., Orengo, C. M., Wodak, S. J. & Richelle, J. (1996). *Trends Biochem. Sci.* **21**(7), 251–256.

Hendrickson, W. A. & Konnert, J. H. (1980). *Computing in Crystallography*, edited by R. Diamond, S. Ramaseshan & K. Venkatesan, pp. 1301. Bangalore: Indian Academy of Science.
Hooft, R., Sander, C. & Vriend, G. (1996). *Proteins*, **26**, 363–376.
Jack, A. & Levitt, M. (1978). *Acta Cryst.* A**34**, 931–935.
James, R. W. (1948). *The Optical Principles of the Diffraction of X-rays.* London: G. Bell & Sons.
Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Kleywegt, G. J., Bergfors, T., Senn, H., LeMotte, P., Gsell, B., Shudok, K. & Jones, T. A. (1994). *Structure*, **2**, 1241.
Konnert, J. H. & Hendrickson, W. A. (1980). *Acta Cryst.* A**36**, 344–350.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
Laskowski, R. A., Moss, D. S. & Thornton, J. (1993). *J. Mol. Biol.* **231**, 1049–1067.
Lipson, H. & Cochran, W. (1966). *The Determination of Crystal Structures.* London: G. Bell & Sons.
Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.
Murshudov, G. N. & Dodson, E. J. (1997). *Newslett. Protein Crystallogr.* **33**, 25–30.
Murshudov, G. N, Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Rogers, D. (1965). *Computing Methods in Crystallography*, edited by J. S. Rollett, pp. 133–148.
Sheldrick, G. M. (1995). *SHELXL93, a Program for the Refinement of Crystal Structures.* University of Göttingen, Germany.
Sim, G. A. (1960). *Acta Cryst.* **13**, 511–516.
Stewart, D. E., Sarkar, A. & Wampler, J. E. (1990). *J. Mol. Biol.* **214**, 253–260.
Tronrud, D. E. (1997). *Methods Enzymol.* **277**, 306–319.
Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489–501.
Wilson, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.