Acta Crystallographica Section D
**Biological
Crystallography**

ISSN 0907-4449

Checking nucleic acid crystal structures

Ujjwal Das,[a] Shengfeng Chen,[b] Monika Fuxreiter,[b]† Alexei A. Vaguine,[a]‡ Jean Richelle,[a] Helen M. Berman[b] and Shoshana J. Wodak[a,c]*

[a]Service de Conformation de Macromolécules Biologiques et Bioinformatique, Université Libre de Bruxelles, Avenue F. D. Roosevelt 50, CP160/16, B-1050 Bruxelles, Belgium, [b]Department of Chemistry, Rutgers University, 610 Taylor Road, Piscataway, NJ 08854-8087, USA, and [c]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England

† Present address: Department of Physiology and Biophysics, Mount Sinai School of Medicine, 1 Gustave L. Levy Place, New York, NY 10029-6574, USA.
‡ Present address: University of York, Department of Chemistry, York YO1 5DD, England.

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

Received 26 April 2000
Accepted 1 March 2001

The program *SFCHECK* [Vaguine *et al.* (1999), *Acta Cryst.* D**55**, 191–205] is used to survey the quality of the structure-factor data and the agreement of those data with the atomic coordinates in 105 nucleic acid crystal structures for which structure-factor amplitudes have been deposited in the Nucleic Acid Database [NDB; Berman *et al.* (1992), *Biophys. J.* **63**, 751–759]. Nucleic acid structures present a particular challenge for structure-quality evaluations. The majority of these structures, and DNA molecules in particular, have been solved by molecular replacement of the double-helical motif, whose high degree of symmetry can lead to problems in positioning the molecule in the unit cell. In this paper, the overall quality of each structure was evaluated using parameters such as the *R* factor, the correlation coefficient and various atomic error estimates. In addition, each structure is characterized by the average values of several local quality indicators, which include the atomic displacement, the density correlation, the *B* factor and the density index. The latter parameter measures the relative electron-density level at the atomic position. In order to assess the quality of the model in specific regions, the same local quality indicators are also surveyed for individual groups of atoms in each structure. Several of the global quality indicators are found to vary linearly with resolution and less than a dozen structures are found to exhibit values significantly different from the mean for these indicators, showing that the quality of the nucleic acid structures tends to be rather uniform. Analysis of the mutual dependence of the values of different local quality indicators, computed for individual residues and atom groups, reveals that these indicators essentially complement each other and are not redundant with the *B* factor. Using several of these indicators, it was found that the atomic coordinates of the nucleic acid bases tend to be better defined than those of the backbone. One of the local indicators, the density index, is particularly useful in spotting regions of the model that fit poorly in the electron density. Using this parameter, the quality of crystallographic water positions in the analyzed structures was surveyed and it was found that a sizable fraction of these positions have poorly defined electron density and may therefore not be reliable. The possibility that cases of poorly positioned water molecules are symptomatic of more widespread problems with the structure as a whole is also raised.

1. Introduction

The atomic coordinates of biological macromolecules obtained by X-ray diffraction and NMR experiments represent a compromise between the fit to the experimental data and our knowledge of chemistry, because the data provided by these techniques usually lack atomic resolution. This compromise is achieved with the help of refinement procedures. The quality of the model produced by these procedures is not uniform. It not only depends on the quality of the experimental data but also on the refinement protocol and the reference values used to describe the chemistry of the macromolecular components. Additionally, a model of good overall quality may still contain regions where the atomic coordinates are less accurately defined. This may become a particularly important issue if these regions are associated with biological function.

Several software packages have been developed in recent years for evaluating the quality of protein structures. The most widely used programs are *PROCHECK* (Laskowski, MacArthur *et al.*, 1993) and *WHAT-IF* (Hooft *et al.*, 1996). These programs assess the quality of the geometric and stereochemical parameters of protein molecular models (*e.g.* covalent bonds and angles, main-chain and side-chain dihedral angles, geometry of chiral centre *etc.*). This is performed by evaluating how these parameters deviate from their standard values, which are derived from crystals of small molecules (Engh & Huber, 1991) and from a reference set of high-quality protein structures (Laskowski, Moss *et al.*, 1993; Hooft *et al.*, 1996).

More recently, standard values for the valence geometry of the nucleic acid bases (Clowney *et al.*, 1996) and sugar–phosphate backbone (Gelbin *et al.*, 1996) have also been derived from small-molecule crystal structures. These were used as the basis for the parameter files in nucleic acid refinement (Parkinson *et al.*, 1996) in *X-PLOR* (Brünger, 1992*b*) and *CNS* (Brunger *et al.*, 1998). The program *NUCHECK* (Feng *et al.*, 1998) checks the deviation of the covalent geometry from these standard values, the torsion-angle ranges (Schneider *et al.*, 1997) and chirality, among other features, in all nucleic acid structures that are deposited in the NDB (http://ndbserver.rutgers.edu/; Berman *et al.*, 1992).

The model compliance with the standard geometric parameters is however inadequate for detecting errors, since these parameters are usually also used as restraints in the refinement and hence may leave their mark on the final model (Stewart *et al.*, 1990). Furthermore, compliance with the standard parameters does not reflect the agreement with the experimental data, but with the standard refinement practices (Laskowski, MacArthur *et al.*, 1993; Laskowski, Moss *et al.*, 1993).

The stereochemical quality measures must therefore be supplemented 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 derived from the X-ray diffraction of the crystal.

The quality and completeness of the experimental data are usually evaluated during various stages of the structure-determination process by different programs. However, assessing the agreement of a given model with the experimental data is still performed on a rudimentary level for macromolecules.

Global indicators of the agreement between the experimental data and the model, such as the R factor, can be misleading, especially when there is limited data. The 'free R factor' ($R_{\rm free}$; Brünger, 1992*a*), which is based on standard statistical cross-validation techniques (Brünger, 1997), is more informative. However, there are so far no clear guidelines on what an 'acceptable' $R_{\rm free}$ value should be (Kleywegt & Brünger, 1996), although several ways of estimating it have been proposed (Dodson *et al.*, 1996; Tickle *et al.*, 1998, 2000).

Macromolecular structures feature different levels of precision in different regions and there needs to be some local quality indicators to help pinpoint regions with likely errors in the structure. In small-molecule crystals this is achieved by computing the estimated standard uncertainty (e.s.u.) for the atomic coordinates and the B factors from the variance–covariance matrix obtained by inverting the full normal equation matrix (Cruickshank, 1965). This poses problems in macromolecules, mainly because the quality of the refinement of structures below atomic resolution is dependent on including restraints and hence the X-ray-based normal matrix alone does not reflect the quality of the fit between the model and the experiment.

Lately, however, an increasing number of macromolecular structures, primarily those solved at atomic resolution, have had their e.s.u.s computed (Deacon *et al.*, 1997; Harata *et al.*, 1998), often using the program *SHELXL* (Sheldrick, 1993; Sheldrick & Schneider, 1997), which was recently extended to proteins.

Other methods for determining the relative precision of atoms in macromolecular structures involve calculating the agreement between the model and the electron density in specific regions. The newer approach by Zhou *et al.* (1998) is related to the real-space R factor of Jones *et al.* (1991), but computes the electron density in a different way (Chapman & Rossmann, 1995).

We present here the application of a unified set of criteria for evaluating the experimental data and the agreement of the model with those data in biological macromolecules. These procedures are collated in the stand-alone software package *SFCHECK* (Vaguine *et al.*, 1999), which performs these evaluations on a given structure completely automatically and provides a concise pictorial output of the results in PostScript format. This offers the opportunity of surveying and comparing the results obtained for a large number of structures using the same criteria.

In this work, we use *SFCHECK* to survey the quality of deposited nucleic acid structures. These structures, particularly DNA, present a special challenge with respect to validation. The diffraction data is often limited and the frequently observed pseudo-continuous helix generated by crystallographic symmetry is one cause for making errors with respect to helix register, especially when molecular-replacement methods are used for structure determination. Such errors are usually quite difficult to detect (Joshua-Tor *et al.*, 1992; Vojtechovsky *et al.*, 1995).

A total of 145 nucleic acid entries have been processed from the NDB and the Protein Data Bank (PDB; Bernstein *et al.*, 1977; Berman *et al.*, 2000; http://www.pdb.org/) for which both structure-factor data and atomic coordinates were available. From these, a subset of 105 structures is selected for which a meaningful quality assessment can be performed. The quality of these structures is evaluated both at the global and local levels. At the global level, we use descriptors such as the R factor, correlation coefficient and various atomic coordinate error estimates. At the local level, a set of local quality measures is computed for each residue or atom group in individual structures. These measures comprise the per-residue (or per-group) atomic displacement, density correlation and B factor, as well as a measure of the local electron-density level.

Table 1
Parameters computed for the analysis of the structure-factor data.

The leftmost column lists the parameter, the next column gives the formula or definition of the parameter and the third column contains a short description of the meaning of the parameter, whenever warranted.

Parameter	Formula/definition	Meaning
Completeness (%)	Percentage of the expected number of reflections for the given crystal space group and resolution	
$B_{\rm overall}$ (Patterson)	$8\pi^2\sigma_{\rm Patt}/2^{1/2}$†	Overall B factor
$R_{\rm stand}(F)$	$\langle\sigma(F)\rangle/\langle F\rangle$‡	Uncertainty of the structure-factor amplitudes
Optical resolution	$[2(\sigma^2_{\rm Patt}+\sigma^2_{\rm sph})]^{1/2}$†§	Expected minimum distance between two resolved atomic peaks
Expected optical resolution	Optical resolution computed considering all reflections	
CC_F	$\dfrac{\langle F_{\rm obs}\cdot F_{\rm calc}\rangle-\langle F_{\rm obs}\rangle\cdot\langle F_{\rm calc}\rangle}{[(\langle F^2_{\rm obs}\rangle-\langle F_{\rm obs}\rangle^2)\cdot(\langle F^2_{\rm calc}\rangle-\langle F_{\rm calc}\rangle^2)]^{1/2}}$	Correlation coefficient between the observed and calculated structure-factor amplitudes
S	$\left\{\dfrac{\sum(F_{\rm obs}\cdot f_{\rm cutoff})^2}{\sum[F_{\rm calc}\cdot\exp(-B^{\rm overall}_{\rm diff}\cdot s^2)\cdot f_{\rm cutoff}]^2}\right\}^{1/2}$¶	Factor applied to scale $F_{\rm calc}$ to $F_{\rm obs}$
$f_{\rm cutoff}$	$1-\exp(-B_{\rm off}{\cdot}s^2)$††	Function applied to obtain a smooth cutoff for low-resolution data

† $\sigma_{\rm Patt}$ is the standard deviation of the Gaussian fitted to the Patterson origin peak. ‡ F is the structure-factor amplitude and $\sigma(F)$ is the structure-factor standard deviation. The brackets denote averages. § $\sigma_{\rm sph}$ is the standard deviation of the spherical interference function, which is the Fourier transform of a sphere of radius $1/d_{\rm min}$, with $d_{\rm min}$ being the minimum d spacing. ¶ $B^{\rm overall}_{\rm diff}=B^{\rm overall}_{\rm obs}-B^{\rm overall}_{\rm calc}$ is added to the calculated overall B factor $B_{\rm overall}$, so as to make the width of the calculated Patterson origin peak equal to the observed one; s is the magnitude of the reciprocal-lattice vector. †† $B_{\rm off}=4d^2_{\rm max}$, where s and $d_{\rm max}$ are the magnitude of the reciprocal-lattice vector and the maximum d spacing, respectively.

Our study focuses on identifying general trends across structures and typical trends across groups within individual structures chosen as examples. A detailed account of the quality of individual structures examined in this work may be obtained by consulting the *SFCHECK* outputs available on the World Wide Web (http://ndbserver.rutgers.edu/NDB/archives/).

2. Methods

2.1. Analysis performed by *SFCHECK*

In this section, we give a succinct description of the parameters computed by *SFCHECK*, as well as the conditions used by this software in performing the analysis described in this work. A detailed description of these parameters and the various features of the software can be found elsewhere (Vaguine *et al.*, 1999).

2.1.1. Treatment of structure-factor data and scaling. The following operations are performed on the structure-factor data. Reflections are excluded if they are systematically absent, negative or have flagged σ values (99.9). Equivalent reflections are merged. The amplitudes of missing reflections are approximated by taking the average value for the corresponding resolution shell.

From the model coordinates, *SFCHECK* calculates structure factors and scales them to the observed structure factors. The scaling factor S is computed using a smooth cutoff for low-resolution data (Vaguine *et al.*, 1999) as shown in Table 1. This involves the calculation of the observed and calculated overall B factors from the standard deviations of the Gaussian fitted to the Patterson origin peaks (see Table 1 and Vaguine *et al.*, 1999).

To assess the quality of the structure-factor data, the program computes four additional quantities: the completeness of the data, the uncertainty of the structure-factor amplitudes, the optical resolution and the expected optical resolution (see Table 1 for details).

2.1.2. Global agreement between the model and experimental data. To evaluate the global agreement between the atomic model and the experimental data, the program computes three commonly used quality indicators of X-ray structures of macromolecules. These are the classical R factor, $R_{\rm free}$ (Brünger, 1992*a*) and the correlation coefficient CC_F between the calculated and observed structure-factor amplitudes (see Table 1). The first two quantities are computed using (i) all the considered reflections (except those approximated by their average value in the corresponding resolution shell) and (ii) applying the same resolution and σ cutoff as those reported by the authors. The correlation coefficient is computed using all reflections from the reported high-resolution limit, applying the smooth low-resolution cutoff but no σ cutoff.

2.1.3. Estimations of errors in atomic positions. The errors associated with the atomic positions are expressed as standard deviations (σ) of these positions. *SFCHECK* computes three different error measures. 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. The first two error measures are the expected maximal and minimal errors, respectively, and the third measure is the diffraction precision indicator

Table 2
Estimations of errors in atomic coordinates.

The leftmost column lists the parameter, the next column gives the formula or definition of the parameter and the third column contains a short description of the meaning of the parameters, when warranted.

Parameter	Formula/definition	Meaning
$\sigma(x)$	σ(slope)/curvature†	Standard deviation of the atomic coordinates following Cruickshank (1949) for the minimal and maximal errors (Vaguine *et al.*, 1999)
σ(slope) for maximal error	$\frac{2\pi \cdot \{\sum[h^2 \cdot (F_{\rm obs} - F_{\rm calc})^2]\}^{1/2}}{V_{\rm unit\_cell} \cdot a}$‡	Expression for σ(slope) in the expected maximal error following Cruickshank (1949)
Curvature	$\frac{2\pi^2 \cdot \sum[h^2 \cdot F_{\rm obs}]}{V_{\rm unit\_cell} \cdot a^2}$	Expression for the curvature following Murshudov *et al.* (1998)
σ(slope) for minimal error	$\frac{2\pi \cdot \{\sum[h^2 \cdot \sigma(F_{\rm obs})^2]\}^{1/2}}{V_{\rm unit\_cell} \cdot a}$§	Expression for σ(slope) in the expected minimal error following Cruickshank (1949)
DPI	$\sigma(x) = \left(\frac{N_{\rm atoms}}{N_{obs} - 4 \cdot N_{\rm atoms}}\right)^{1/2} \cdot c^{-1/3} \cdot d_{\rm min} \cdot R_{\rm factor}$¶	Atomic coordinate error estimate following Cruickshank (1996)

† σ(slope) and curvature are the slope and curvature of the electron-density map at the atomic centre in the x direction; for spherically symmetric peaks, $\sigma(x) \simeq \sigma(y) \simeq \sigma(z)$. ‡ a is the crystal unit-cell length, h is the Miller index and $V_{\rm unit\_cell}$ is the unit-cell volume. § $\sigma(F_{\rm obs})$ is the standard deviation of the structure-factor amplitude. ¶ c is the structure-factor data completeness expressed as a fraction (0–1); $R_{\rm factor}$ is the conventional R factor, $N_{\rm atoms}$ is the total number of atoms in the unit cell, $N_{\rm obs}$ is the total number of observed reflections and $d_{\rm min}$ is the minimum d spacing.

Table 3
Parameters computed by *SFCHECK* to assess the quality of the model in specific regions.

The leftmost column lists the parameter, the next column give the formula or definition of the parameter and the third column contains a short description of the meaning of the parameters, when warranted.

Parameter	Formula/definition	Meaning
Shift	$(1/N\sigma)\sum_i^N \Delta_i$; Δ_i = gradient$_i$/curvature$_i$†	Normalized average atomic displacement computed over a group of atoms or residue; reflects the tendency of the group of atoms to move from their current position
D_corr	$\frac{\sum \rho_{\rm calc}(x_i)[2\rho_{\rm obs}(x_i) - \rho_{\rm calc}(x_i)]}{\left([\sum \rho^2_{\rm calc}(x_i)]\{\sum [2\rho_{\rm obs}(x_i) - \rho_{\rm calc}(x_i)]^2\}\right)^{1/2}}$‡	Electron-density correlation coefficient computed over a group of atoms or residue; reflects the local agreement of the model with the electron density
Density_index	$[\prod \rho(x_i)]^{1/N} / \langle\rho\rangle_{\rm all\_atoms}$ §	Reflects the level of the electron density for a group of atoms; is a local measure of the density level
Connect		Same as density_index, but considering only backbone atoms¶

† Gradient$_i$ is the gradient of the $(F_{\rm obs} - F_{\rm calc})$ map with respect to the atomic coordinates, 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 group under consideration and σ is the standard deviation of the Δ_i values computed in the structure. ‡ $\rho_{\rm calc}(x_i)$ and $\rho_{\rm obs}(x_i)$ 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. For the calculation of the electron density at the atomic centre, see Vaguine *et al.* (1999). § $[\prod \rho(x_i)]^{1/N}$ is the geometric mean of the $(2F_{\rm obs} - F_{\rm calc})$ electron density of the considered atom subset and $\langle\rho\rangle_{\rm all\_atoms}$ is the average electron density of the atoms in the structure. For water molecules or for ions which are represented by a unique atom the above expression reduces to the ratio $\rho(x_i)/\langle\rho\rangle_{\rm all\_atoms}$. ¶ Backbone atoms are N, C, C$_\alpha$ for proteins and P, O5′, C5′, C3′, O3′ for nucleic acids.

(DPI). The mathematical expressions for these error measures are given in Table 2 and further details can be found in Vaguine *et al.* (1999).

2.1.4. Local agreement between the model and the experimental data. In addition to the global structure-quality measures, *SFCHECK* also assesses the quality of the model in specific regions by computing several quality estimators for each residue in the macromolecule and, whenever appropriate, also for solvent molecules and groups of atoms in ligand molecules. These estimators are the normalized atomic displacement, shift; the electron-density correlation coefficient, density correlation; the density index; the average B factor and the connectivity index, connect. These quantities are computed for individual atoms and averaged over those composing each residue or group (see Table 3 and Vaguine *et al.*, 1999 for details).

2.1.5. Data sets. With the goal of assessing the quality of the nucleic acid structures archived in the NDB and the PDB, all entries for which both structure-factor data and atomic coordinates were available at the time of this study were processed by *SFCHECK* (information on these entries is available as supplementary material[1]). Since *SFCHECK* recomputes a number of parameters reflecting the agreement between the experimental data and the molecular model, it was necessary to exclude from the analysis a number of structures for which such comparison could not be meaningfully performed. Of the total of 145 processed structures, we excluded those for which the experimental structure-factor σ values were missing, those that were refined anisotropically (at the time of this study, *SFCHECK* was capable of handling only isotropic refinement) and structures that had major disordered portions. This reduced the number of analyzed structures to 105, which are referred to as set I throughout this study.

From data set I, six high-quality structures were selected in order to compare the quality of specific portions of the model: the bases and the sugar–phosphate backbone. These structures, referred to as set II, fulfilled the following conditions: R factor < 0.20, d spacing < 2.5 Å, completeness > 0.70,

[1] Supplementary data have been deposited in the IUCr electronic archive (Reference: li0353). Services for accessing these data are described at the back of the journal.

structure-factor σ values must be available. Set II comprised the following structures: ADH007, ADH038, BDJ017, BDJ031, ZDG054 and ZDG056. Information on all 145 structures, including citation information, is available as supplementary material.

3. Results

3.1. Validation of the *SFCHECK* scaling procedure

In evaluating the agreement of the molecular model with the experimental data, an important but difficult issues one must grapple with is the scaling of $F_{\rm calc}$ to $F_{\rm obs}$. In *SFCHECK*, this scaling is performed using a somewhat different scaling procedure than in many of the routinely used refinement programs. For both the calculated and observed amplitudes, the $B_{\rm overall}$ parameter is derived from the standard deviation of the Gaussian fitted to the Patterson origin peak (see §2 and Vaguine *et al.*, 1999) rather than from the Wilson plot (Wilson, 1949). This scaling method yields more uniform $B_{\rm overall}$ values over different d-spacing limits than the Wilson plot, as illustrated in Fig. 1. This figure compares the $B_{\rm overall}$ computed by both methods for data set I comprising the 105 nucleic acid structures examined in this study. Fig. 1(*a*) displays its variation with the d spacing. At small d spacing (high resolution) both methods yield quite similar $B_{\rm overall}$ values, differing by less than 15 Å^2, whereas at larger d spacing (>2.0 Å) the $B_{\rm overall}$ calculated by Wilson and Patterson scaling tend to deviate significantly (20–40 Å^2). Interestingly, the Patterson $B_{\rm overall}$ varies linearly across the entire range of optical resolution of the electron-density map, whereas the $B_{\rm overall}$ from the Wilson scaling shows different behaviours below and above the optical resolution of 1.6 Å (Fig. 1*b*). On the basis of these results, we conclude that Patterson scaling yields more reliable estimates of the $B_{\rm overall}$ than the Wilson scaling does, especially when only low-resolution data are available.

3.2. Quality assessment of structures in data set I

This section presents results on the analysis of the 105 nucleic acid structures from the NDB with deposited structure-factor files and coordinate files referred to as data set I. To perform the analysis, *SFCHECK* was run on each of the 105 structures individually, yielding a comprehensive *SFCHECK* output for each structure (available at http://ndbserver.rutgers.edu/NDB/archives/). A detailed description of the contents of the *SFCHECK* output can be found in Vaguine *et al.* (1999). Various global and local quality indicators output by *SFCHECK* were then used to evaluate the quality of the analyzed structures at the global and local levels.

3.2.1. Global structure quality. Fig. 2 plots the R factor reported by the authors *versus* that calculated by *SFCHECK* using the reported d-spacing limits and σ thresholds for the structure-factor data. We find that the recomputed R factor often tends to be larger than the reported one. In general, the difference between the two quantities does not exceed 5%. This can be considered as a reasonable discrepancy, given that *SFCHECK* uses a different scaling procedure, applies a smooth low-resolution cutoff and sometimes considers a slightly different number of reflections than the authors (see §2).

Larger deviations between the recomputed and reported R factors usually indicate problems with the coordinates or structure-factor files, which for the most part occurred in the course of data submission or processing. Several formatting and syntax errors or missing and incorrect data were detected and corrected based on the observation of R-factor discrepancies. In some cases authors were contacted to verify these corrections. Overall, in only six structures out of the total of 105 (5%) did the recalculated and reported R factor differ by more than 5%; in three of these (NDB IDs ADFB72, ADF073 and ADJ081) the difference exceeded 10%.

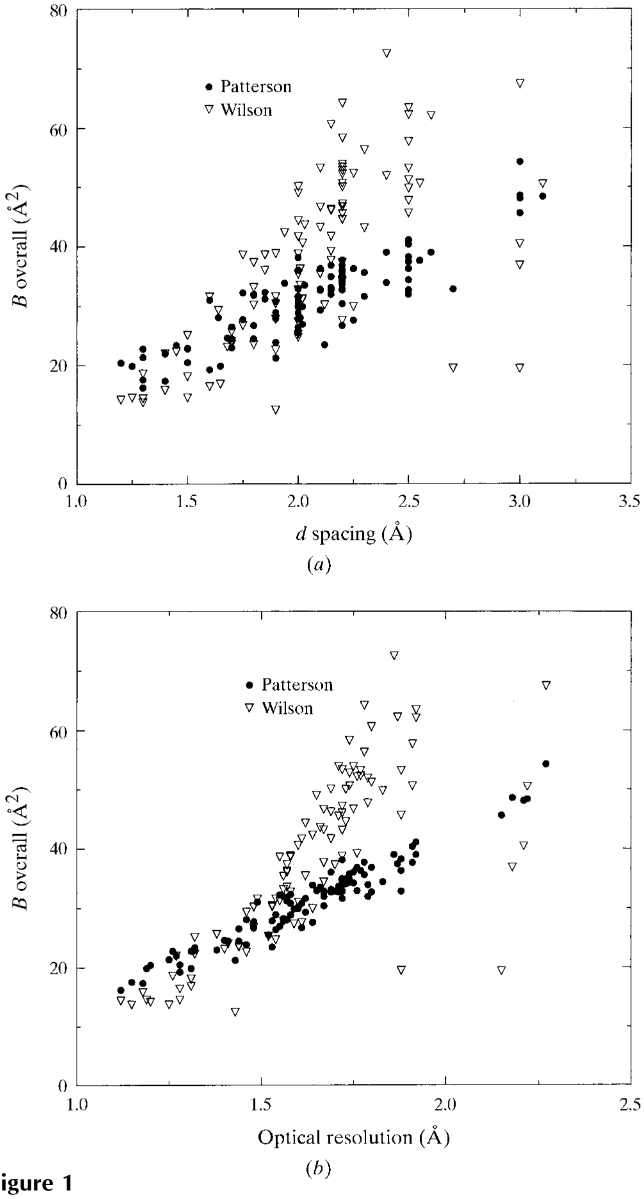


Figure 1
$B_{\rm overall}$ computed by *SFCHECK* from the Wilson plot (open triangles) and Patterson scaling (filled squares) as a function of (*a*) the d spacing and (*b*) the optical resolution . Each point represents the $B_{\rm overall}$ of one of the 105 nucleic acid structures of set I (see §2 for details).

Fig. 3 shows the distributions of the d spacing, completeness and R-factor values in the 105 analyzed structures. The d-spacing distribution is rather broad, with a mean around 2.1 Å and a standard deviation of 0.55 Å. Only a handful of structures are resolved to better than 1 Å resolution and no structures are determined at a resolution lower than 2.8 Å. Another useful indicator of the data quality is the completeness of the structure-factor data, expressed in percent. Its distribution for the 105 nucleic acid structures (Fig. 3b) shows that in one third of these the completeness is less than 80%. The R-factor values (recomputed by *SFCHECK*) are rather narrowly distributed, with only five structures (three A-DNA structures, one Z-DNA and one RNA structure) having an R factor above 0.25 (see legend of Fig. 3 for details).

Fig. 4 illustrates how the key global quality indicators, the completeness, the R factor, the correlation coefficient between $F_{\rm calc}$ and $F_{\rm obs}$ and the maximal error in atomic coordinates estimated from the difference between $F_{\rm calc}$ and $F_{\rm obs}$ (see Table 2), vary as a function of the d spacing. The R factor remains rather constant over the resolution range of the analyzed structures, but the values display a large spread (Fig. 4a). An even larger spread is displayed by the completeness values (Fig. 4b), but the majority is above 80%, especially in the 2 Å d-spacing range in which the largest number of the structures have been determined. The correlation coefficient (Fig. 4c) and maximal error (Fig. 4d) vary roughly linearly with the d spacing. The correlation coefficient decreases from 0.97 on average at 1.25 Å d spacing to about 0.87 at 3 Å d spacing, whereas the maximal error increases about tenfold with d spacing, ranging from 0.05 to 0.45 Å in the d-spacing range 1.25–3 Å.

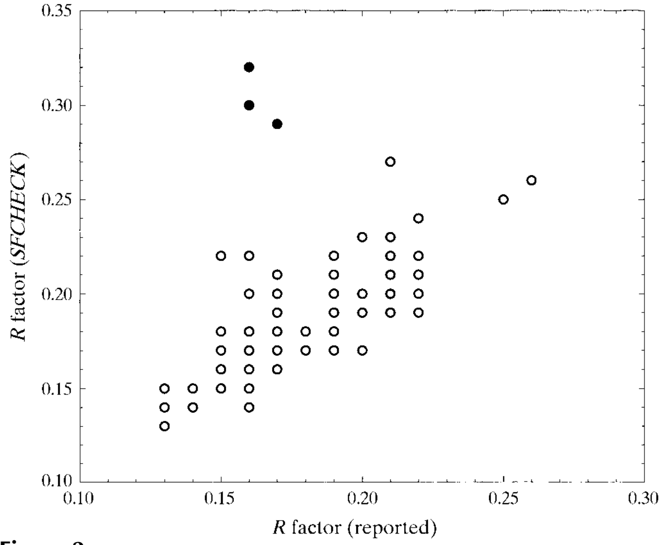
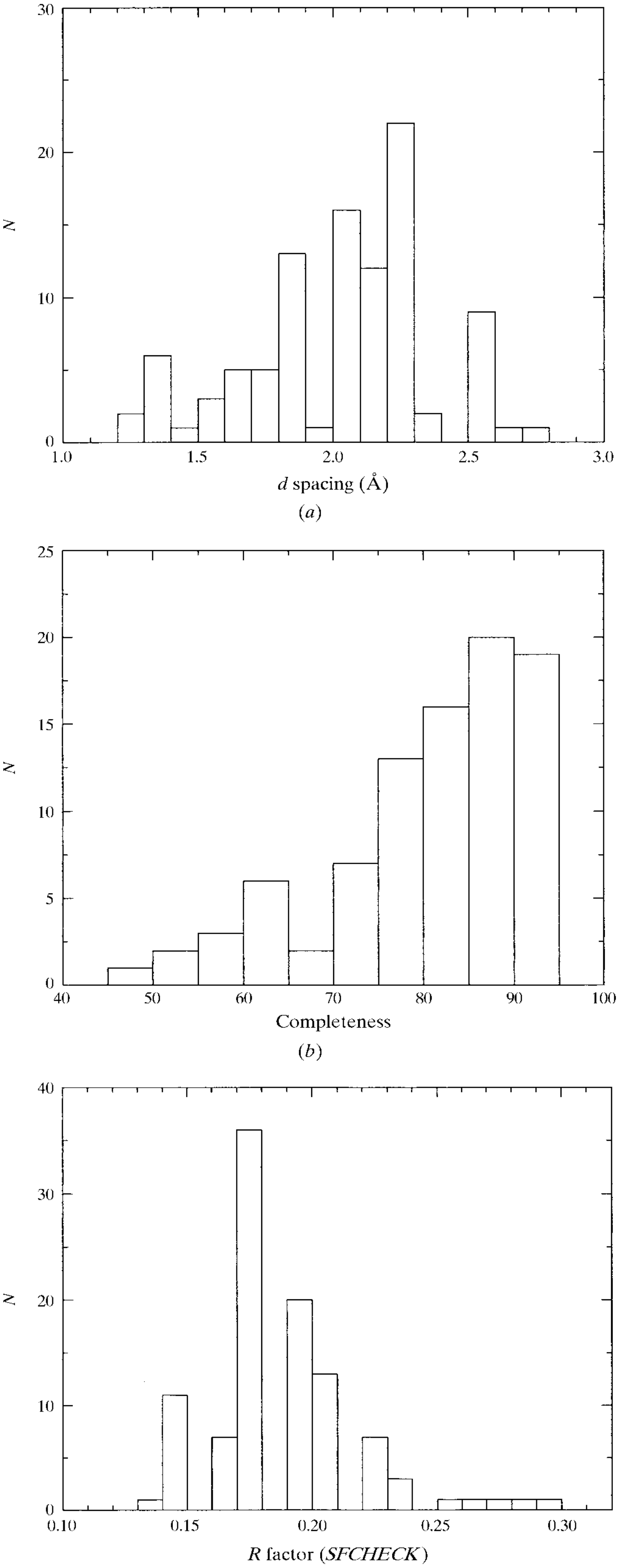


Figure 2
The reported R factor [R factor (reported)] *versus* the R factor computed by *SFCHECK* [R factor (*SFCHECK*)] for the 105 nucleic acid structures in data set I (see §2 for details). The computed R factor was obtained using structure factors within the same d-spacing range and σ values as those reported by the authors. The filled dots correspond to the three structures for which the difference between the reported and recomputed R factors exceeded 10%. The NDB (PDB) IDs for these structures are ADFB72 (256d), ADF073 (257d) and ADJ081 (320d).



(c)

Figure 3
Distributions of the values for three parameters measuring the overall quality of individual structures in data set I with 105 nucleic acid crystal structures. (a) Distribution of the d spacing (Å); (b) distribution of the data completeness; (c) distribution of the R factor computed by *SFCHECK*.

In addition to the analysis of global quality indicators, we also computed the average values of three local quality indicators computed by *SFCHECK* for each structure. These are the density correlation, the atomic displacement and the density index. The first of these quantities is the electron-density correlation coefficient computed over a group of atoms or residue (see Table 3). The second is the normalized atomic displacement which indicates the tendency of the considered atom or group of atoms to move away from their current position. The density index of an atom or group of atoms is a measure of the local level of the $2F_{\rm obs} - F_{\rm calc}$ electron density computed by *SFCHECK* at the corresponding atomic positions (Table 3).

Fig. 5 shows how these three quantities, duly averaged over all the groups or atoms in each structure, vary with the *d* spacing and *R* factor. In general, the variation with *d* spacing is roughly linear, whereas that with the *R* factor is more diffuse. The average density correlation, which is a finer measure of the agreement between the observed and calculated structure factors than the *R* factor, decreases approximately linearly with the *d* spacing (Fig. 5*a*), albeit by a narrow range of about 4%. However, the scatter is large, indicating that other factors in addition to data resolution also influence this quality indicator. A roughly linear decrease and large scatter is also apparent for the average density correlation as a function of the *R* factor (Fig. 5*b*).

The average atomic displacement increases with *d* spacing and *R* factor (Figs. 5*c* and 5*d*). The variation with *d* spacing is roughly linear, indicating that the convergence reached during refinement, reflected by this measure, is significantly influ-

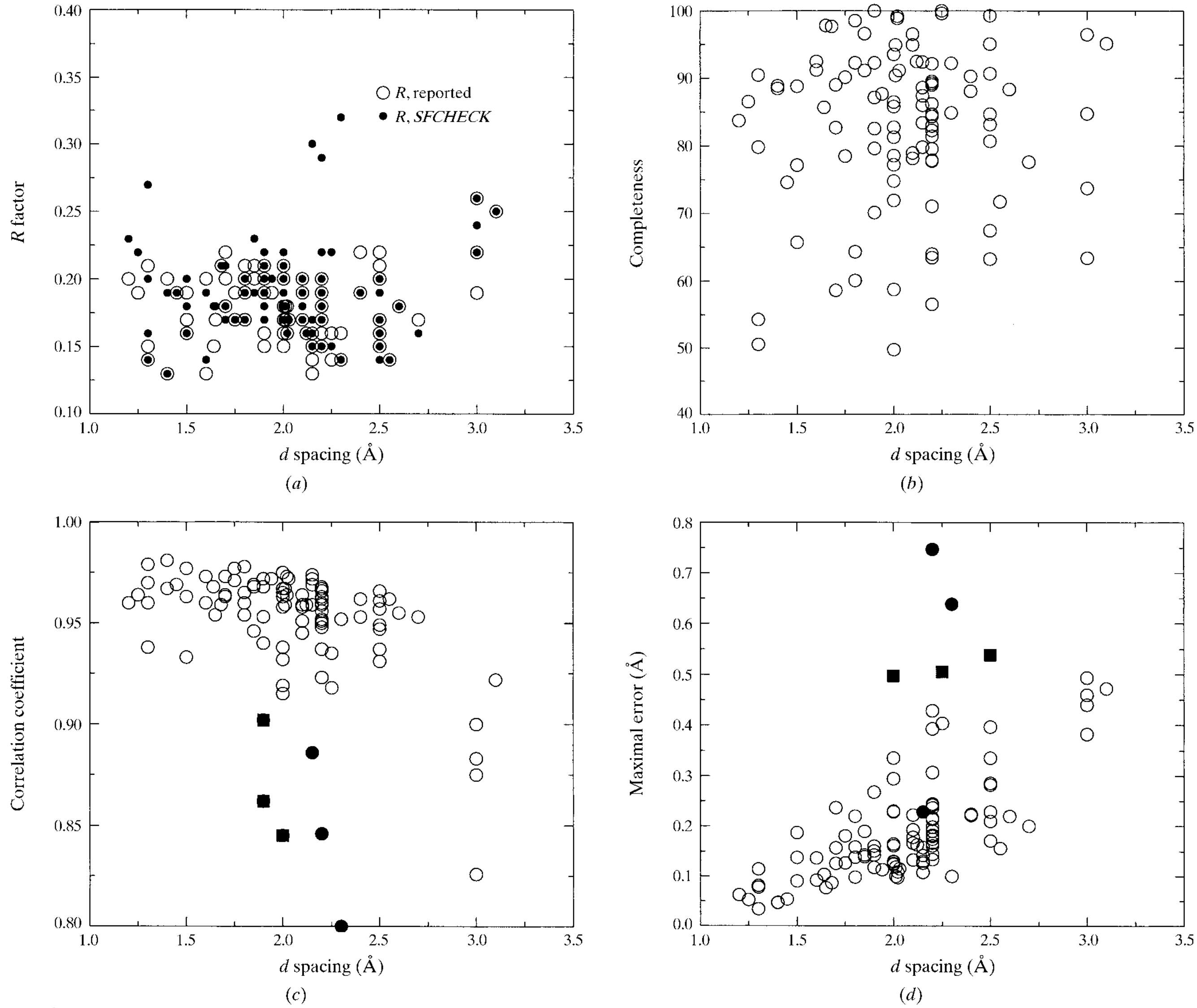


Figure 4
Variation of global quality indicators with the nominal resolution (*d* spacing) of the crystallographic data. The quality indicators were computed by *SFCHECK* for each of the 105 nucleic acid crystal structures in data set I; (*a*) *R* factor, (*b*) completeness, (*c*) correlation coefficient and (*d*) maximal error. For the meaning of the various quantities see §2. The three structures for which the reported and recomputed *R* factors differ by more than 10% are highlighted as black circles. The NDB (PDB) codes for these structures are ADFB72 (256d), ADF073 (257d) and ADJ081 (320d). Other outliers, identified visually, are marked as filled squares in plots (*c*) and (*d*). No outliers are marked on plot (*b*), owing to the large scatter in this plot.

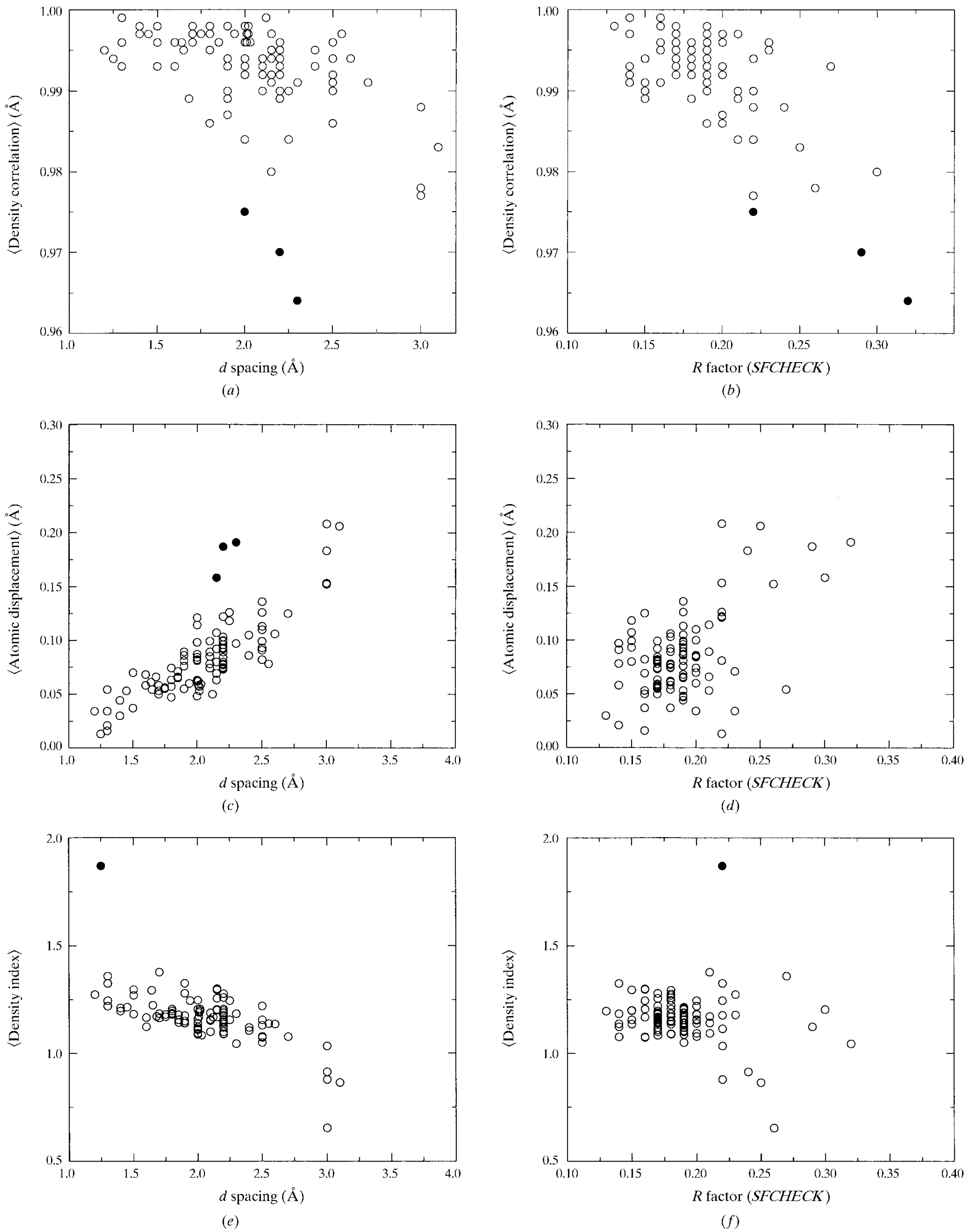


Figure 5
Variation of average local quality indicators with the crystallographic *d* spacing and *R* factor, respectively. The analyzed indicators are the density correlation (*a*, *b*), the atomic displacement (*c*, *d*) and the density index (*e*, *f*). Filled circles indicate outlier structures, which also display the largest difference between their reported and computed *R* factors.

enced by the resolution of the X-ray data. However, the corresponding *R*-factor plot displays a much larger scatter and a less clear trend.

For the average density index, a linear decrease with *d* spacing is observed (Fig. 5*e*) and the dependence on the *R* factor is also weak (Fig. 5*f*). The decrease with *d* spacing reflects the fact that lower resolution of the X-ray data yields models with poorer fit to the electron-density map. It is noteworthy that the scatter of the average density-index values in the *d*-spacing plot (Fig. 5*e*) is significantly lower than for the other average local quality indicators. This suggests that this parameter should be a particularly useful indicator for the quality of a structure as a whole.

3.2.2. Identifying structures with unusual properties. The analysis presented above could in principle be used to flag outlier structures, those for which one or more global quality indicators display unusual values. To define such values in a rigorous manner requires a sound statistical analysis. Unfortunately, the limited number of structures in data set I (105) as well as the variable deposition procedures (incomplete data sets and no cross-validation data assigned) precluded at this stage the computation of meaningful distributions as a function of resolution or other useful parameters.

Nevertheless, visual inspection of Figs. 4 and 5 allowed us to qualitatively identify the most prominent outliers. Note that in doing so the few structures determined at 3 Å resolution were ignored. We found that four A-DNA and B-DNA structures (ADF073, ADFB72, GDL013, ADJ081) are outliers in many plots. Three of these (ADF073, ADFB72 and ADJ081; marked as filled circles) also displayed the largest difference between their reported and computed *R* factors (Fig. 2). It is thus very likely that these structures or the corresponding structure-factor files have a problem.

Six additional structures stand out in some of the plots. The outliers in the correlation coefficient *versus* *d*-spacing plot (Fig. 4*c*) include two DNA structures (ZDFB21, GDL012). The maximal error *versus* *d*-spacing plot (Fig. 4*d*) features a further three outliers (BDL042, BDL029, ADFB63). The density-index plots in Figs. 5(*e*) and 5(*f*) show a Z-DNA structure (ZDFB31) as standing out owing to its very high average density index (>1.75). This structure is not an outlier in any of the other plots of Fig. 4 and 5; detailed inspection of the full *SFCHECK* output for this structure indicates that it is of quite high resolution (1.3 Å) and of unusually high quality with, among other things, a data completeness of 90.5%.

3.2.3. Structure quality at the local level. To evaluate the quality of the structures of data set I in specific regions of the model, we analyze the local quality measures, those computed for individual residues or groups of atoms in each structure. *SFCHECK* computes four different local measures: the three local measures mentioned above, namely the density correlation, the atomic displacement and density index, as well as the well known *B* factor. A legitimate question to ask, therefore, is the degree of correlation between them, since two highly correlated measures could reasonably be considered as redundant.

To investigate this question, we analyze the pairwise relationships between all four measures, first within individual structures of set I that contain a sufficiently large number of residues and then across all 105 structures in this set.

Figs. 6(*a*)–6(*f*) display the scatter plots for pairs of local quality measures in one of the largest structures of our set, that of the hammerhead ribozyme catalytic RNA loop (URX035; PDB code 1mme; Scott *et al.*, 1995), which contains 82 residues. Of particular interest are the relationships between the average residue *B* factor and the other local measures. Not unexpectedly, the average *B* factor is clearly anticorrelated to the residue density index (Fig. 6*a*), with a roughly exponential behaviour. The *B* factor rises quite steeply as the value of the density index, which measures the electron-density level, decreases. Interestingly, the points in Fig. 6(*a*) follow two independent curves. The points in the lower curve (in red) correspond to the nucleic acid bases, whereas those in the upper curve (black) represent the sugar–phosphate backbone. This indicates that the bases are in general better defined in the electron density than the backbone, most certainly owing to the constraints imposed by base-pair formation.

The correlation of the residue *B* factor with other local measures is in general poorer. It shows virtually no correlation with the residue atomic displacement (Fig. 6*b*) and hence the latter quality measure also correlates poorly with the density index (Fig. 6*c*). On the other hand, the fourth residue quality indicator, the electron-density correlation, appears to display weak linear correlation with the residue *B* factor (Fig. 6*d*), density index (Fig. 6*e*) and atomic displacement (Fig. 6*f*), respectively. This trend is clearer for the bases than for the backbone: an additional indication that the former are generally better defined than the latter.

The local quality indicators computed by *SFCHECK* thus provide information on the model which is complementary to that given by the residue *B* factor. This information is particularly helpful for finding problem regions in a structure which would not be detectable on the basis of a high *B* factor alone.

To further illustrate this point, two individual B-DNA structures (BDJ037 and BLDB76) are examined in detail. In the first structure (Fig. 7*a*) the backbone of residue 11 (Cyt in chain *B*) has only a marginally higher *B* factor than residues 8 and 16, but its density index is <0.5, is the lowest value encountered in the entire structures, an indication that at least one of its atoms has very low electron density. In the second structure, the backbone of residue 15 (Cyt in chain *B*) has a zero density index (Fig. 7*b*), which means that at least one of its atoms lies completely outside any density. The average *B* factor for the same group is again rather low (~20 Å^2) and is lower than that of the backbone of several other residues whose density indices are higher.

Inspection of the *SFCHECK* results for other structures in data set I revealed similar trends and confirmed that the value of the average density index of a given group of atoms is particularly helpful for identifying problem regions.

In order to double-check the *SFCHECK* results, $2F_o - F_c$ electron-density maps were computed for several structures featuring residues with low density-index values. This was performed using the *CCP*4 package (Collaborative Computational Project, Number 4, 1994) and the maps were displayed using the program *O* (Jones *et al.*, 1991). The typical result obtained is illustrated in Fig. 8, which displays the electron density at and around guanine residues 19 and 20 in the structure BDJB27. In this region, which corresponds to the termini of the DNA duplex, there are breaks in the electron

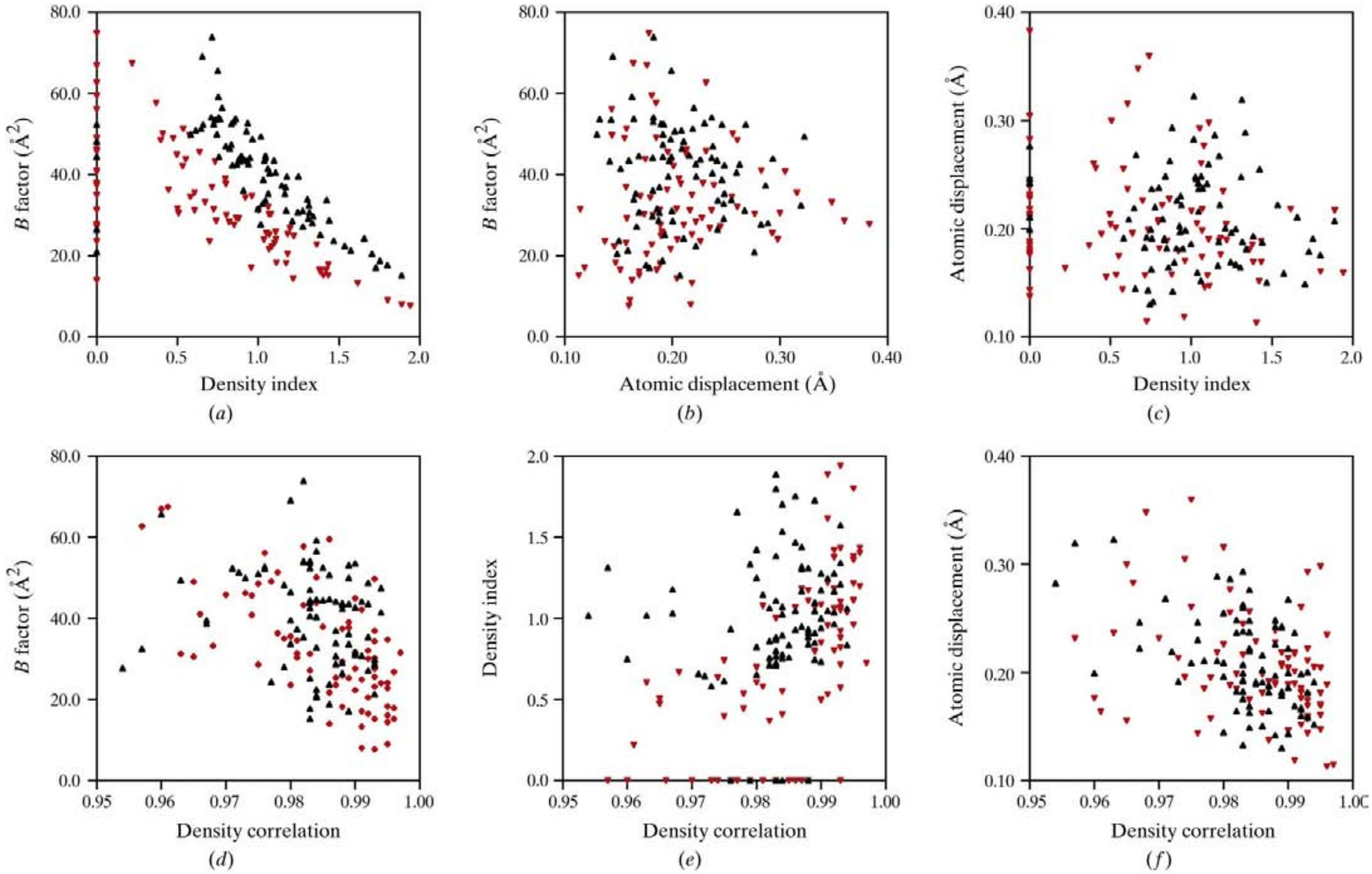


Figure 6

Pairwise relations between the various local quality indicators computed by *SFCHECK*. (*a*)–(*f*) display the plotted values for the crystal structure of the hammerhead ribozyme catalytic RNA loop, URX035 (PDB code 1mme), which contains 82 residues (Scott *et al.*, 1995). The meaning of the displayed parameters is described in §2. Two outliers have been taken out of (*a*)–(*f*) in order to display a meaningful figure. The red triangles are for the bases, whereas the black triangles are for the sugar–phosphate backbone. An appreciable number of groups, corresponding primarily to bases, have zero density index and hence appear on the abscissa or on the ordinate in some of the plots.

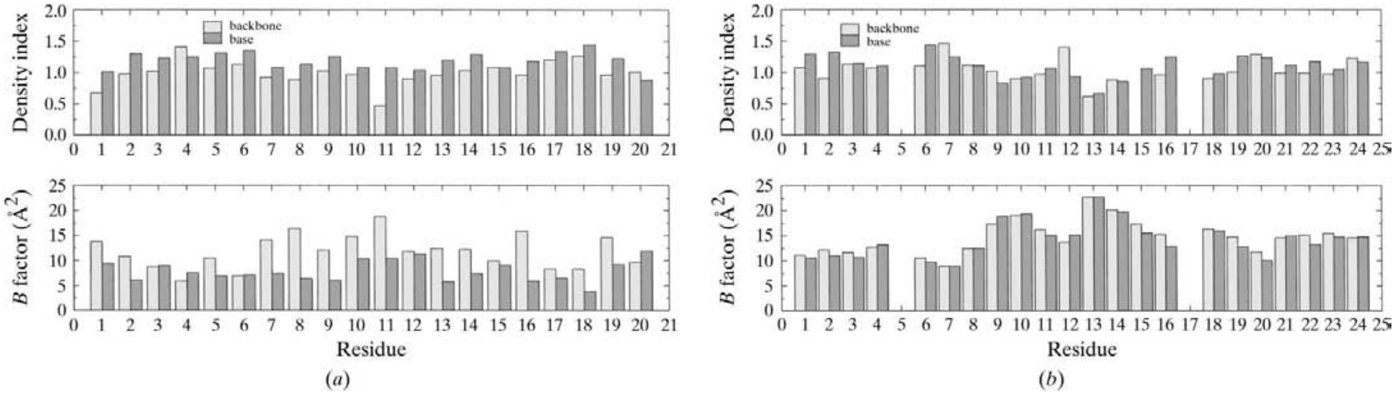


Figure 7

Local quality indicators of individual nucleic acid residues along the chain. The plotted local quality indicators are the density index (top) and the *B* factor (bottom). The residue number is indicated horizontally. (*a*) Plots for the B-DNA structure BDJ037 (1d57; Yuan *et al.*, 1992). Note that in this structure residues 5 and 17 contain modified bases which have not been analyzed. (*b*) Plots for the B-DNA structure BDLB76 (285d; Shatzky-Schwatz *et al.*, 1997).

Table 4

Summary of the quality assessment results for structures in data set II.

The leftmost column gives the NDB code of the structure. The second column lists the atom groups over which the quality indicators were averaged. The quantities in the last three columns refer to local quality indicators averaged over the atoms included in each group. The atomic displacement and *B* factor are given in Å and Å^2. The density index is a dimensionless ratio (see §2).

Entry	Atom groups	Atomic displacement	Density index	*B* factor
ADH007	All nucleic acid atoms	0.083	1.073	18.1
	Backbone atoms	0.082	1.052	19.5
	Base atoms	0.083	1.095	16.6
ADH038	All nucleic acid atoms	0.044	1.211	17.8
	Backbone atoms	0.048	0.992	20.4
	Base atoms	0.040	1.431	15.2
BDJ017	All nucleic acid atoms	0.039	1.224	16.5
	Backbone atoms	0.042	1.047	20.3
	Base atoms	0.036	1.401	12.6
BDJ031	All nucleic acid atoms	0.070	1.183	10.7
	Backbone atoms	0.078	1.046	13.3
	Base atoms	0.062	1.320	8.0
ZDG054	All nucleic acid atoms	0.074	1.181	13.6
	Backbone atoms	0.073	1.140	14.3
	Base atoms	0.075	1.222	12.9
ZDG056	All nucleic acid atoms	0.086	1.139	17.6
	Backbone atoms	0.089	1.103	19.3
	Base atoms	0.083	1.176	16.0

density at the 1σ level, in agreement with the low density index of the sugar–phosphate backbone of these residues, which is 0.4 and 0.75, respectively.

However, these are definitely not the worst errors that *SFCHECK* is able to detect. Analysis of the $2F_o - F_c$ electron-density maps of ADF073, one of the outliers in the plots of Figs. 4 and 5, shows regions with a particularly poor agreement between the model and the data. Several atoms of residue G10 in chain *B* are out of density, in complete agreement with the low density index (zero) and the density-correlation values computed by *SFCHECK*. The solvent region is also poorly modelled, with many electron-density peaks unaccounted for and some of the modelled solvent outside the electron density, even at the 1σ level. Analysis of the complete *SFCHECK* output for this structure reveals that there is a 20% discrepancy between the number of reflections reported to be used in refinement and those that *SFCHECK* retrieves from the deposited structure-factor file, using the resolution limits and σ value indicated by the authors. The observed anomalies may thus stem from a mismatch between the deposited model and structure-factor files.

3.2.4. Differences in model quality between the sugar–phosphate backbone and bases across DNA structures of data set II. The trends in three key local quality indicators, the atomic displacement, density index and *B* factor were analyzed separately for the nucleic acid backbone and bases in the six DNA structures of data set II, with the results given in Table 4.

We see that in the examined structures the atomic displacement of the sugar–phosphate backbone is on average larger than the atomic displacement of the bases. Three of the six selected DNA structures show about the same atomic displacement for the backbones and bases (ADH007, ZDG054 and ZDG056); the others have higher atomic displacement for the backbones than the bases. This is also observed in all the structures in data set I (data not shown). In addition, the average density index of the sugar–phosphate backbone is barely above 1 for most structures, even though the backbone includes the more electron-dense phosphates. In contrast, the density index of the bases is consistently higher in all the structures in data set II as already pointed out. This confirms that the bases tend to be better positioned in the electron density than the backbone atoms, a finding in turn corroborated by the *B*-factor values, which are consistently higher for the nucleic acid backbone than the bases (Table 4).

Analysis of the density index of the bases and sugar–phosphate moieties in the 2056 individual nucleic acid residues of the structures in data set I confirms this conclusion. It shows that on average the density index of the bases (1.45) is higher that of the backbone atoms (1.2), as seen from the distributions depicted in Fig. 9. A very small fraction of both moieties have a density index of zero, which means that at least one of their atoms lies outside any electron density altogether. Twice as many bases (66) as backbone groups (34) display this property. However, whereas the backbone groups with zero density index are evenly distributed amongst DNA and RNA structures, nearly all the poorly

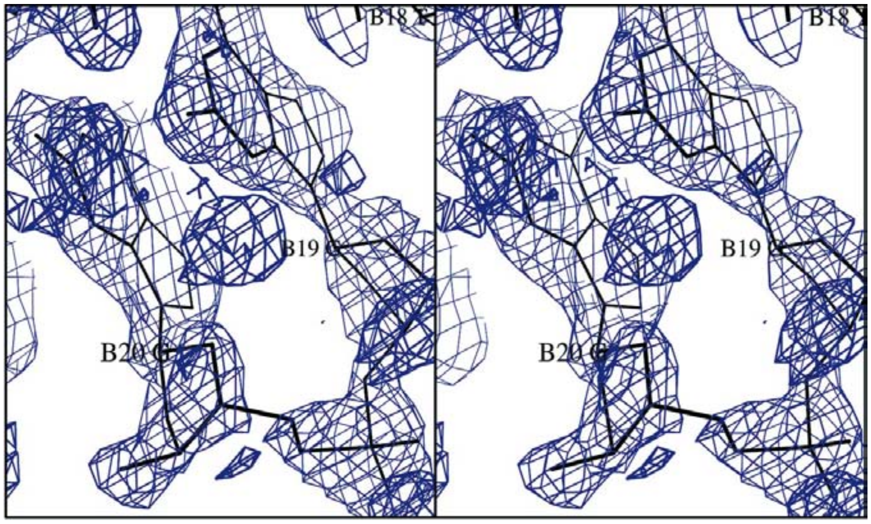


Figure 8

The $2F_o$ – F_c electron-density map (blue contours) and model (bold black lines) in the region around residues 19 and 20 in BDJB27 (2d25; Heinemann & Hahn, 1992). These residues have an average and above average density-index values for the bases, but low density indices for the backbone (0.4 and 0.75 for residues 19 and 20, respectively). Inspection of the figure reveals that there are breaks in the electron density at the 1σ contour level of the map computed using the *CCP*4 program suite and displayed using the program *O*. Figures are generated using *Molscript* (Kraulis, 1997).

determined bases (65 of 66) belong to RNA molecules; closer inspection indicates that they occur mainly in globular macromolecular RNA structures. This is most likely to be because the positions of the bases in these structures are not as restricted as in helical structures.

3.2.5. Quality of water positions. Having determined that the density index is a particularly useful indicator of how well the model fits the electron density, we use it to analyze crystallographic water positions in the structures of data set I.

Fig. 10 displays the distribution of the density index values of the 6720 water positions in data set I. This distribution is very different from those of the density index of nucleic acid atoms (Fig. 9) and displays several striking features. Its median equals 0.75, a much lower value than for the nucleic acid groups, with a sizable fraction of water molecules (29%) having a density index of 0.5 or lower. Of these, 469 molecules have a density index exactly equal to zero, indicating that they are positioned outside the electron density altogether. Moreover, the distribution displays a very long tail towards density index values as high as 6, suggesting that the corresponding water positions are located in regions with electron-density values much higher than average and conceivably higher than those expected for water molecules.

In order to gain insight into the possible causes of these unusual properties, we selected a set of 22 structures from those containing a particularly high fraction of water molecules with low density indices (Table 5). $2F_o - F_c$ electron-density maps were computed for all 22 structures using the *CCP*4 software package (Collaborative Computational Project, Number 4, 1994) and the results were examined using the program *O* (Jones *et al.*, 1991). In addition, we examined the high-quality Z-DNA structure (NDB code ZDFB31), which was taken as a reference (see above). We could verify that its water molecules all had density-index values in the same range as those of the nucleic acid groups, the atomic model showed good agreement with the electron-density map across the entire structure and no excessive density was observed outside the model. Data on the 23 structures are given in Table 5.

The examined water molecules in the 22 selected structures fell into three main categories. The first and largest category included the vast majority (98%) of the cases. They corresponded to water molecules with very low density-index values (<0.1) which were found to have no electron density in the computed map. The other two categories grouped the remaining 2% of the cases. One comprised water molecules with low density indices but medium to high electron density in the computed maps and the other grouped water molecules with a normal or high density index but weak or missing electron density in the map.

A typical example of the observations made for waters in category 1 is illustrated in Fig. 11(*a*), which displays the position of water molecule 85 in the electron density of structure ADJB83. We see that while the atomic positions of the bases fit well into the electron-density contours, the water molecule has no density at all. A possible reason that these water molecules are positioned in such a way is that different kind of maps (such as $3F_o - F_c$ map) may have been used by the authors when adding water molecules during refinement. Alternatively, some of the water molecules may have moved away from the electron-density peaks during the last stages of refinement, but have not been removed from the coordinate file.

Figs. 11(*b*) and 11(*c*) illustrate typical observations made for the small fraction of water molecules in categories 2 and 3. Waters in category 2, having a density index of zero or <0.1 but weak to medium electron density, were found in the $2F_o - F_c$ electron-density maps of a few structures. A good example of this situation is water molecule 33 in structure ADHP36 (Fig. 11*b*). It has a density index of 0.04 but medium electron density in the map. This apparent discrepancy with the *SFCHECK* results might be because of differences in the

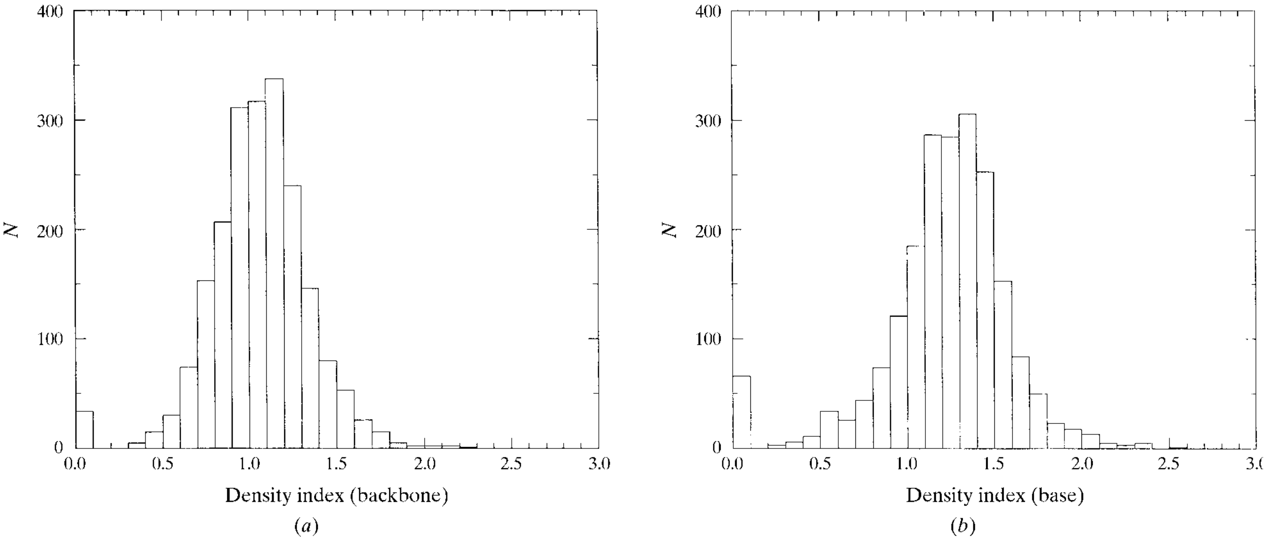


Figure 9
Distribution of the density-index values for (*a*) nucleic acid backbone and (*b*) bases in the 105 structures analyzed in this work.

Table 5
Summary of the 24 nucleic acid structures whose water positions were analyzed in detail.

Iso. Refin., isomorphous refinement; Mol. Repl., molecular replacement; MIR, multiple isomorphous replacement; ISRR, iterative single isomorphous replacement. The last structure in the list (ZDFB31) is a high-quality Z-DNA structure (see text), taken as reference.

NDB ID	PDB ID	Program	Space group	*d*-spacing range (Å)	Phasing method	Authors
ADH047	118d	*NUCLSQ*	$P4_32_12$	8.0–1.64	Iso. Refin.	C. Bingman, X. Li, G. Zon, M. Sundaralingam
ADHP36	1d26	*NUCLSQ*	$P4_32_12$	6.0–2.12	Iso. Refin.	U. Heinemann, L.-N. Rudolph, C. Alings, M. Morr, W. Heikens, R. Frank, H. Bloecker
ADJ081	320d	*X-PLOR*	$P2_12_12_1$	8.0–2.15	Mol. Repl.	D. B. Tippin, M. Sundaralingam
ADJ082	321d	*X-PLOR*	$P2_12_12_1$	8.0–2.15	Mol. Repl.	D. B. Tippin, M. Sundaralingam
ADJB79	318d	*X-PLOR*	$P6_1$	8.0–2.2	ISIR	D. B. Tippin, M. Sundaralingam
ADJB80	319d	*X-PLOR*	$P2_12_12_1$	8.0–2.5	Mol. Repl.	D. B. Tippin, M. Sundaralingam
ADJB83	322d	*X-PLOR*	$P6_1$	8.0–2.15	ISIR	D. B. Tippin, M. Sundaralingam
ADJB84	323d	*X-PLOR*	$P2_12_12_1$	8.0–2.15	Mol. Repl.	D. B. Tippin, M. Sundaralingam
ADJB86	325d	*X-PLOR*	$P6_1$	8.0–2.5	ISIR	D. B. Tippin, M. Sundaralingam
ADJB87	326d	*X-PLOR*	$P2_12_12_1$	8.0–2.5	Mol. Repl.	D. B. Tippin, M. Sundaralingam
ADJB88	327d	*X-PLOR*	$P6_122$	8.0–1.94	Mol. Repl.	D. B. Tippin, B. Ramakrishnan, M. Sundaralingam
AHH071	246d	*X-PLOR*	$R3$	10.0–2.2	Mol. Repl.	M. C. Wahl, C. Ban, C. Sekharudu, B. Ramakrishnan, M. Sundaralingam
ARL048	157d	*NUCLSQ*	$P2_1$	7.0–1.8	Mol. Repl.	G. A. Leonard, K. E. McAuley-Hecht, S. Ebel, D. M. Lough, T. Brown, W. N. Hunter
BDL038	1d65	*NUCLSQ*	$P2_12_12_1$	8.0–2.2	Mol. Repl.	K. J. Edwards, D. G. Brown, N. Spink, S. Neidle
GDL032	102d	*X-PLOR*	$P2_12_12_1$	8.0–2.2	Mol. Repl.	C. M. Nunn, S. Neidle
GDL033	109d	*X-PLOR*	$P2_12_12_1$	8.0–2	Mol. Repl.	A. A. Wood, C. M. Nunn, A. Czarny, D. W. Boykin, S. Neidle
GDL045	289d	*X-PLOR*	$P2_12_12_1$	8.0–2.2	Mol. Repl.	J. O. Trent, G. R. Clark, A. Kumar, W. D. Wilson, D. W. Boykin, J. E. Hall, R. R. Tidwell, B. L. Blagburn, S. Neidle
GDL048	303d	*X-PLOR*	$P2_12_12_1$	8.0–2.2	Iso. Refin.	G. R. Clark, C. J. Squire, E. J. Gray, W. Leupin, S. Neidle
GDL052	311d	*X-PLOR*	$P2_12_12_1$	8.0–2.2	Iso. Refin.	G. R. Clark, D. W. Boykin, A. Czarny, S. Neidle
TRNA08	3tra	*NUCLSQ*	$C222_1$	10.0–3		E. Westhof, P. Dumas, D. Moras
UHJ055	1fix	*X-PLOR*	$P4_322$	5.0–2.3	MIR	N. C. Horton, B. C. Finzel
URL050	280d	*X-PLOR*	$P1$	8.0–2.4	Mol. Repl.	S. E. Lietzke, C. L. Barnes, C. E. Kundrot
ZDFB31	1d76	*NUCLSQ*	$P2_12_12_1$	8.0–1.3	Mol. Repl.	B. Schneider, S. L. Ginell, R. Jones, B. Gaffney, H. M. Berman

protocols for electron-density calculation (*e.g.* scaling procedure) used in *SFCHECK* and *CCP*4. It may also be a consequence of the fact that water 33 is positioned in a rather narrow isolated peak, surrounded from all sides by regions with low electron density. Indeed, to compute the electron density at a given position, *SFCHECK* averages electron-density values of neighboring grid points within a 2.5 Å distance limit and applies a smooth cutoff (Vaguine *et al.*, 1999). Low or missing electron density at these points could contribute to lower significantly the computed density for the water position and hence its density index.

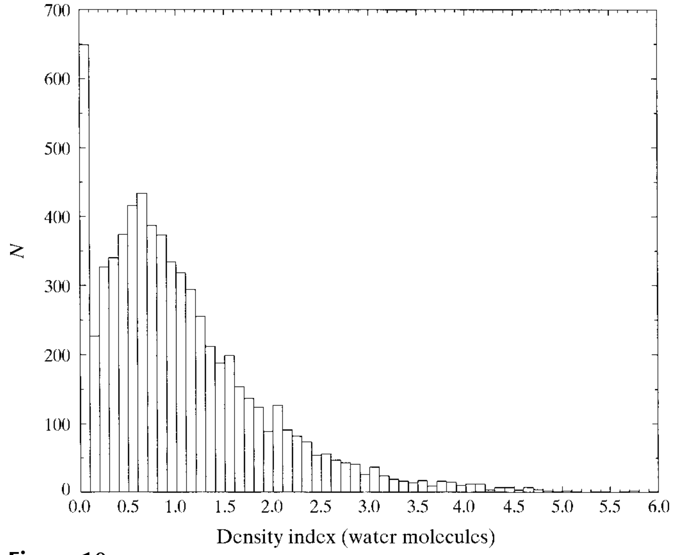


Figure 10
Distribution of the density-index values of the 6720 water molecules in the 105 nucleic acid structures analyzed in this study.

The corollary of the above behaviour is displayed by the waters in category 3. These have rather high density indices, but most surprisingly show no electron density at all or only very weak density in the $2F_o - F_c$ maps (*e.g.* ADJB83 and ADJB86). A typical example of this observation is water 34 in structure ADJB83. This water has a density index of 1.58, but the $2F_o - F_c$ electron-density map shows no electron density at the corresponding position (Fig. 11*c*). On the other hand, there is extra electron density very close by which is unaccounted for, as highlighted in Fig. 11(*c*). When *SFCHECK* computes the electron density for this water molecule, it sums over this excessive electron density, yielding a high density-index value. A similar result would be obtained if a water molecule is too close to other atoms such as a phosphate group or an O atom, or generally when it is off-centred from its electron-density peak.

Very large density-index values for water positions could also result from incorrectly assigning water molecules to positions actually corresponding to metal ions. Such cases can be identified by comparing the geometry of the contacts

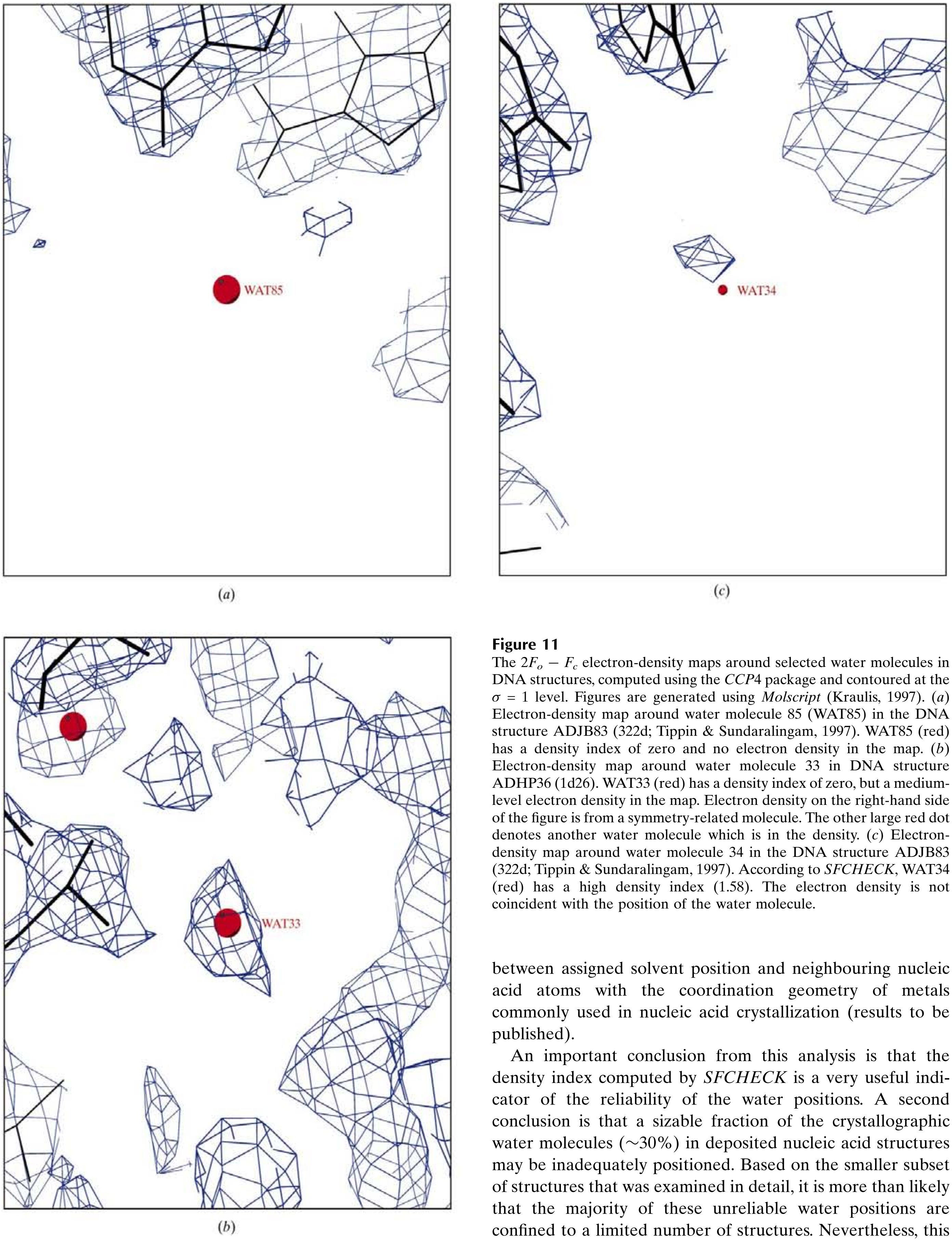


Figure 11
The $2F_o - F_c$ electron-density maps around selected water molecules in DNA structures, computed using the *CCP*4 package and contoured at the σ = 1 level. Figures are generated using *Molscript* (Kraulis, 1997). (*a*) Electron-density map around water molecule 85 (WAT85) in the DNA structure ADJB83 (322d; Tippin & Sundaralingam, 1997). WAT85 (red) has a density index of zero and no electron density in the map. (*b*) Electron-density map around water molecule 33 in DNA structure ADHP36 (1d26). WAT33 (red) has a density index of zero, but a medium-level electron density in the map. Electron density on the right-hand side of the figure is from a symmetry-related molecule. The other large red dot denotes another water molecule which is in the density. (*c*) Electron-density map around water molecule 34 in the DNA structure ADJB83 (322d; Tippin & Sundaralingam, 1997). According to *SFCHECK*, WAT34 (red) has a high density index (1.58). The electron density is not coincident with the position of the water molecule.

between assigned solvent position and neighbouring nucleic acid atoms with the coordination geometry of metals commonly used in nucleic acid crystallization (results to be published).

An important conclusion from this analysis is that the density index computed by *SFCHECK* is a very useful indicator of the reliability of the water positions. A second conclusion is that a sizable fraction of the crystallographic water molecules (~30%) in deposited nucleic acid structures may be inadequately positioned. Based on the smaller subset of structures that was examined in detail, it is more than likely that the majority of these unreliable water positions are confined to a limited number of structures. Nevertheless, this calls for great caution in studies aimed at analyzing water–nucleic acid interactions in crystal structures and for much greater attention in monitoring the addition of water molecules during the last stages of refinement.

Lastly, the possibility that the detected problems with water positions are symptomatic of more widespread problems with the model as a whole should be seriously considered. With the majority of the nucleic acid structures and particularly those in Table 5 being solved by molecular-replacement methods, incorrect positioning in the unit cell of the highly symmetrical double-helical motif could be a much more common error than suspected previously. Validation with *SFCHECK* might be a convenient means of detecting such errors. However, a systematic analysis is needed to establish the combination of symptoms detected by *SFCHECK* that reliably reflect these errors.

4. Conclusions

In this study, we have presented a survey of the quality of 105 nucleic acid crystal structures for which structure-factor data were deposited and which contained enough information to allow a meaningful comparison with the information deposited by the authors.

Unlike most previous surveys of the quality of crystal structures, which have been mainly geared to evaluating the quality of the atomic coordinates, this one analyzes the quality of the deposited structure-factor data, as well as the agreement of the atomic coordinates with the electron-density map. The latter analysis is performed (i) for each structure taken as a whole, by computing the values of a number of global quality indicators, and (ii) for specific regions of each model, by computing the values of several local quality indicators.

Based on the survey of the global quality indicators, we conclude that the analyzed nucleic acid structures are of rather uniform quality, with only very few structures exhibiting what seem to be unusual values for these indicators. However, the limited number of analyzed structures and the fact that the deposited experimental data is often incomplete and contains no information on cross-validation data, did not permit at this stage the computation of resolution-dependent ranges of expected values for these indicators.

The survey of the local quality indicators was mainly focused on investigating the relationships between the various indicators. This showed that they were generally not redundant with, but complementary to, the commonly used *B* factor.

Finally, we also showed that the density-index parameter computed by *SFCHECK* is particularly useful for examining the quality of the model and particularly of crystallographic water positions and found that quite often the quality of these positions in nucleic acid crystals was not optimal. It was also suggested that a high number of unreliable water positions could be symptomatic of problems with the model as a whole, which may have various origins, including the incorrect positioning of the model by molecular-replacement techniques, an aspect currently under investigation. The fact that many of the entries examined in this study do not include low-resolution data may also have serious deleterious effects on the quality of the corresponding maps, particularly in the solvent-boundary regions.

These results taken together show that *SFCHECK* should be a useful complement to validation procedures based on geometric and stereochemical criteria alone, such as *PROCHECK* or *WHAT-IF*, which do not take into account the X-ray data.

Presently, the main bottleneck to the improvement and generalization of procedures such as *SFCHECK* is that the diffraction data are not available for most of the publicly deposited structures or are of rather uneven quality for the structures for which they have been provided. This needs to change if we wish to see the emergence of more effective structure-validation protocols which combine geometrical and X-ray-based quality-assessment measures.

The full *SFCHECK* output (as per Fig. 1) for the 105 nucleic acid structures analyzed in this study is available at http://ndbserver.rutgers.edu/NDB/archives/.

This work was carried out with funding from the Department of Energy, USA (grant number DE-FG02096ER62166.A000) and from the European BIOTECHNOLOGY project BIO2-CT92-0524 on 3D Macromolecular Structure Validation. All members of the European Consortium on Structure Validation, especially R. Laskowski, are thanked for stimulating discussions. We are furthermore grateful to the anonymous referees for very useful suggestions and comments.

References

Agarwal, R. C. (1978). *Acta Cryst.* A**34**, 791–809.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A. R. & Schneider, B. (1992). *Biophys. J.* **63**, 751–759.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Brünger, A. T. (1992*a*). *Nature (London)*, **355**, 472–474.

Brünger, A. T. (1992*b*). *X-PLOR Version* 3.0: *A System for Crystallography and NMR.* Yale University, New Haven, Connecticut, USA.

Brünger, A. T. (1997). *Methods Enzymol.* **277**, 366–396.

Brunger, A. T., Adams, P. D., Clore, G. M., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Chapman, M. S. & Rossmann, M. G. (1995). *Structure*, **3**, 151–162.

Clowney, L., Jain, S. C., Srinivasan, A. R., Westbrook, J., Olson, W. K. & Berman, H. M. (1996). *J. Am. Chem. Soc.* **118**, 509–518.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Cruickshank, D. W. J. (1949). *Acta Cryst.* **2**, 65–82.

Cruickshank, D. W. J. (1965). *Computing Methods in Crystallography*, edited by J. S. Rollet, pp. 112–116. Oxford: Pergamon.

Cruickshank, D. W. J. (1996). *Proceedings of the CCP*4 *Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 11–22. Warrington: Daresbury Laboratory.

Deacon, A., Gleichmann, T., Kalb (Gilboa), A. J., Price, H., Raftery, J., Bradbrook, G., Yariv, J. & Helliwell, J. R. (1997). *J. Chem. Soc. Faraday Trans.* **93**, 4305–4312.

Dodson, E., Kleywegt, G. J. & Wilson, K. S. (1996). *Acta Cryst.* D**52**, 228–234.

Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Feng, Z., Westbrook, J. & Berman, H. M. (1998). *NUCHECK*. NDB Report 407. Rutgers University, Piscataway, NJ, USA.

Gelbin, A., Schneider, B., Clowney, L., Hsieh, S.-H., Olson, W. K. & Berman, H. M. (1996). *J. Am. Chem. Soc.* **118**, 519–528.

Harata, K., Abe, Y. & Muraki, M. (1998). *Proteins*, **30**, 232–243.

Heinemann, U. & Hahn, M. (1992). *J. Biol. Chem.* **267**, 7332–7341.

Hooft, R. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Joshua-Tor, L., Frolow, F., Appela, E., Hope, H., Rabinovich, D. & Sussman, J. L. (1992). *J. Mol. Biol.* **225**, 397–431.

Kraulis, P. (1997). *J. Appl. Cryst.* **24**, 946–950.

Kleywegt, G. J. & Brünger, A. T. (1996). *Structure*, **4**, 897–904.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.

Laskowski, R., Moss, D. & Thornton, J. (1993). *J. Mol. Biol.* **231**, 1049–1067.

Murshudov, G. N., Davies, G. J., Isupov, M., Krzywda, S. & Dodson, E. J. (1998). *CCP4 Newsl. Protein Crystallogr.* **35**, 37–42.

Parkinson, G., Vojtechovsky, J., Clowney, L., Brünger, A. T. & Berman, H. M. (1996). *Acta Cryst.* D**52**, 57–64.

Schneider, B., Neidle, S. & Berman, H. M. (1997). *Biopolymers*, **42**, 113–124.

Scott, W. G., Finch, J. T. & Klug, A. (1995). *Cell*, **81**, 991–1002.

Shatzky-Schwartz, M., Arbuckle, N. D., Eisenstein, M., Rabinovitch, D., Bareket-Samish, A., Haran, T. E., Luisi, B. F. & Shakked, Z. (1997). *J. Mol. Biol.* **267**, 595-623.

Sheldrick, G. M. (1993). *SHELXL*93. *J. Phys. Chem.* **98**, 9700–9711.

Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.

Stewart, D. E., Sarkar, A. & Wampler, J. E. (1990). *J. Mol. Biol.* **214**, 253–260.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (1998). *Acta Cryst.* D**54**, 243–252.

Tickle, I. J., Laskowski, R. A. & Moss, D. S. (2000). *Acta Cryst.* D**56**, 442–450.

Tippin, D. B. & Sundaralingam, M. (1997). *J. Mol. Biol.* **267**, 1171–1185.

Vaguine, A. A., Richelle, J. & Wodak, S. J. (1999). *Acta Cryst.* D**55**, 191–205.

Vojtechovsky, J., Eaton, M. D., Gaffney, B., Jones, R. & Berman, H. M. (1995). *Biochemistry*, **34**, 16632–16640.

Wilson, A. (1949). *Acta Cryst.* **2**, 318–321.

Yuan, H., Quintana, J. & Dickerson, R. E. (1992). *Biochemistry*, **31**, 8009-8021.

Zhou, G., Wang, J., Blanc, E. & Chapman, M. (1998). *Acta Cryst.* D**54**, 391–399.