Acta Cryst. (1996). D52, 842–857

Use of Non-crystallographic Symmetry in Protein Structure Refinement

GERARD J. KLEYWEGT

Department of Molecular Biology, Biomedical Centre, Uppsala University, Box 590, S-751 24 Uppsala, Sweden. E-mail: gerard@xray.bmc.uu.se

(Received 26 July 1995; accepted 5 December 1995)

Abstract

Several methods to assess the (dis)similarity of protein structures objectively are described, some of which, when applied to non-crystallographically related protein models, are able to discriminate between significant differences and 'random noise'. Some of these methods have been used to investigate a sample of several hundred protein structures which have been solved by means of X-ray crystallography in order to investigate the extent to which non-crystallographically related protein models differ from one another. It is shown that the extent of such differences is largely dependent on the resolution of the data used for the determination and refinement of the structure and, measured by some statistics, even varies essentially linearly with the resolution. The implications of these findings for the strategies used to refine structures with non-crystallographic symmetry, in particular at low resolution, are discussed. Finally, two examples are given of recent structure determinations from this laboratory in which the presence (and employment) of non-crystallographic symmetry was crucial to the solution and refinement of the structure.

1. Abbreviations

CBH I, cellobiohydrolase I; EG I, endoglucanase I; F_o, F_c, observed and calculated structure-factor amplitudes, respectively; LLDH, L-lactate dehydrogenase; NCS, non-crystallographic symmetry; r.m.s.(d.), root-mean-square (distance/deviation); iSOD, iron superoxide dismutase; MIR, multiple isomorphous replacement; PDB, Protein Data Bank.

2. Introduction

Chothia and Lesk demonstrated a decade ago that the r.m.s. deviation between protein structures is related to the degree of conservation of their amino-acid sequences (Chothia & Lesk, 1986). The lower the percentage of identical residues in two sequences, the more their three-dimensional structures will differ. One puzzling observation is that their curve, which plots the observed r.m.s.d. as a function of percent residue identity, passes through the point (100%, 0.5 Å). In other words, even the structures of proteins with identical sequences are apparently not identical. This observation may often be attributed to different crystallization conditions, different space groups and crystal packing, different data-collection conditions (*e.g.*, temperature) and resolution, the use of different refinement programs and different protocols for building and refining structures in general, and perhaps an inherent limitation on the accuracy with which protein structures can be determined by means of X-ray crystallography (ignoring mobility). Many case studies have appeared in the literature which describe and analyse multiple determinations of the same protein structure in different laboratories and, with the advent of protein-structure determination by means of NMR spectroscopy, structures determined with different techniques.

In this paper, a special case is considered, in which proteins with 100% sequence identity are subjected to identical crystallization, data-collection, building and refinement procedures by the same crystallographer. This is the case for protein molecules which are related by non-crystallographic symmetry. In this case, the only obvious difference between the molecules lies in their environment within the crystal, which one would expect might lead to conformational differences for side chains and loops at the surface of the protein, and sometimes to global domain movements. However, as we will show, in many cases NCS-related molecules look more like distant cousins than like identical twins. We shall also show that the extent of the random component of such differences increases almost linearly with the resolution of the data set that was used in the refinement of the structure. The only exceptions to this observation are the cases in which the structure was refined with a protocol appropriate for the resolution of the data (Kleywegt & Jones, 1995*a*).

3. Assessing structural similarity

Traditionally, the similarity of biomacromolecules is assessed through r.m.s. deviations in the coordinates of subsets of atoms (calculated as the root-mean-square distance between sets of corresponding atoms after two structures have been superimposed). There are, however, a number of drawbacks to this approach.

First, the actual number obtained depends rather critically on the set of atoms that is chosen for the calculation. In the case of NCS-related protein molecules, one has a choice of using, for example, only core Cα atoms (*e.g.*, of those residues which obey the NCS well and lie within, say, 3.5 Å from one another after superpositioning), all Cα atoms, all main-chain atoms, or all non-H atoms. Usually, the value of the r.m.s.d. will increase in the same order.

Second, there may be a problem in defining the most suitable superpositioning of two molecules, for instance for multiple-domain structures in which the relative orientation of the domains differs in the NCS-related copies.

Third, there may be isolated regions (*e.g.*, flexible loops) which display differences and cause high r.m.s.d. values, whereas, on the whole, the structures are very similar.

Fourth, the use of only Cα atoms, which is common practice, involves loss of detail with respect to the actual similarity of the geometry of the main chain.

Finally, when all non-H atoms are used in a comparison, there are trivial naming conventions to cope with which are easily overlooked. For instance, if a tyrosine side chain was 'flipped' during simulated-annealing refinement in one molecule, but not the other, the two residues may have an r.m.s.d. exceeding 1 Å, even though their conformations are chemically and structurally identical. Examples of some of these drawbacks will be discussed later.

A set of more powerful statistics becomes available if one decides to assess how well a set of NCS-related molecules adheres to the principle of conservation of secondary structure. This principle states that 100% homologous proteins normally have identical secondary structures. To our knowledge, there are very few exceptions known to this principle, and most exceptions involve copies of a protein in distinctly different biological or physical conditions. For example, an α-helix may be unwound in the unligated state, whereas it gets ordered in a complex (different biological conditions). Also, some small peptides may be forced into a predominantly α-helical or β-strand conformation by dissolving them in appropriate solvents such as trifluoroacetic acid or dimethyl sulfoxide (different physical conditions). However, in most cases, NCS-related molecules are observed in highly similar biological and physical conditions [although, sometimes, one molecule in the asymmetric unit may bind a ligand, whereas another does not (Sevcik, Dauter, Lamzin & Wilson, 1996)].

If one allows for local deviations from the principle of conservation of secondary structure because of different crystal environments, one may still expect that the following corrolaries hold for most NCS-related molecules.

(*a*) The main-chain φ and ψ angles for corresponding residues in the NCS-related molecules are very similar, and any major exceptions are expected only in loop and hinge regions.

(*b*) The packing in the core of the proteins is very similar, which means that corresponding residues in the core should have very similar side-chain torsion angles, and a similar pattern of temperature factors.

Based on these assumptions, there are many ways in which the similarity of NCS-related molecules can be assessed, and some of these provide insight as to whether or not differences between molecules are genuine or artefacts, probably due to over-fitting (Kleywegt & Jones, 1995*a*). In general, one would expect to find a certain noise level in the variations of torsion angles between different molecules, with some isolated spikes for those residues which perhaps display genuine differences. In the following discussion, three published structures will be used as examples.

CBH I (Divine *et al.*, 1994) (PDB entry 1CEL), which contains twofold NCS and was refined at 1.8 Å without imposing the NCS.

iSOD (Stoddard, Howell, Ringe & Petsko, 1990) (PDB entry 3SDP), which also contains twofold NCS, and was also refined without imposing the NCS, at 2.1 Å.

LLDH (Wigley *et al.*, 1992) (PDB entry 1LDN), which contains eightfold NCS which was initially constrained during the refinement (at 2.5 Å), but later released.

3.1. *Ramachandran plot*

The most obvious, but often overlooked, place to check if a structure adheres to the principle of conservation of secondary structure is the Ramachandran plot (Ramakrishnan & Ramachandran, 1965). Obviously, for a structure with *N*-fold NCS, one expects that most points in the graph occur as tight clusters of *N* points, at least those that lie in the areas typical of α-helices and β-strands. To emphasize this feature in the visualization, one can calculate the angle-averaged values of φ and ψ for every residue, and connect this centroid point with lines to the individual residues in each of the NCS-related molecules, to obtain a multiple-model Ramachandran plot. In calculating the average angle, one should of course take into account that the angles have a period of 360° (*e.g.*, the average of +178° and −174° is not +2°, but −178°).

In Fig. 1(*a*), a multiple-model Ramachandran plot is shown for the structure of CBH I. It is obvious that this model adheres well to the principle of conservation of secondary structure. The largest outlier is Ser99, which is situated in a loop with relatively poor density (this residue is also an outlier in the Ramachandran plot). Fig. 1(*b*), on the other hand, shows the same type of plot for iSOD. In this case, there are many residues which lie in the α- or β-area of the Ramachandran plot in one molecule, but in a forbidden area in the other molecule. Based on the principle of conservation of secondary structure, this appears to be highly unlikely to be real. Fig. 2 shows a superposition of the Cα traces of the two NCS-related iSOD molecules. Fig. 1(*c*) shows the multiple-model Ramachandran plot for LLDH. This structure represents an intermediate case, in which the majority of the NCS-related residues cluster fairly tightly.

3.2. *Linear φ and ψ plots*

Another way to assess the level of 'random scatter' and any possible genuine differences in the values of φ and ψ for NCS-related residues, is by plotting these angles in different ways (in particular when the multiple-model Ramachandran plot becomes cluttered). For the case of two NCS-related molecules, Sevcik *et al.* (1996) have plotted φ *versus* φ and ψ *versus* ψ scatter graphs. Perfect adherence to the principle of

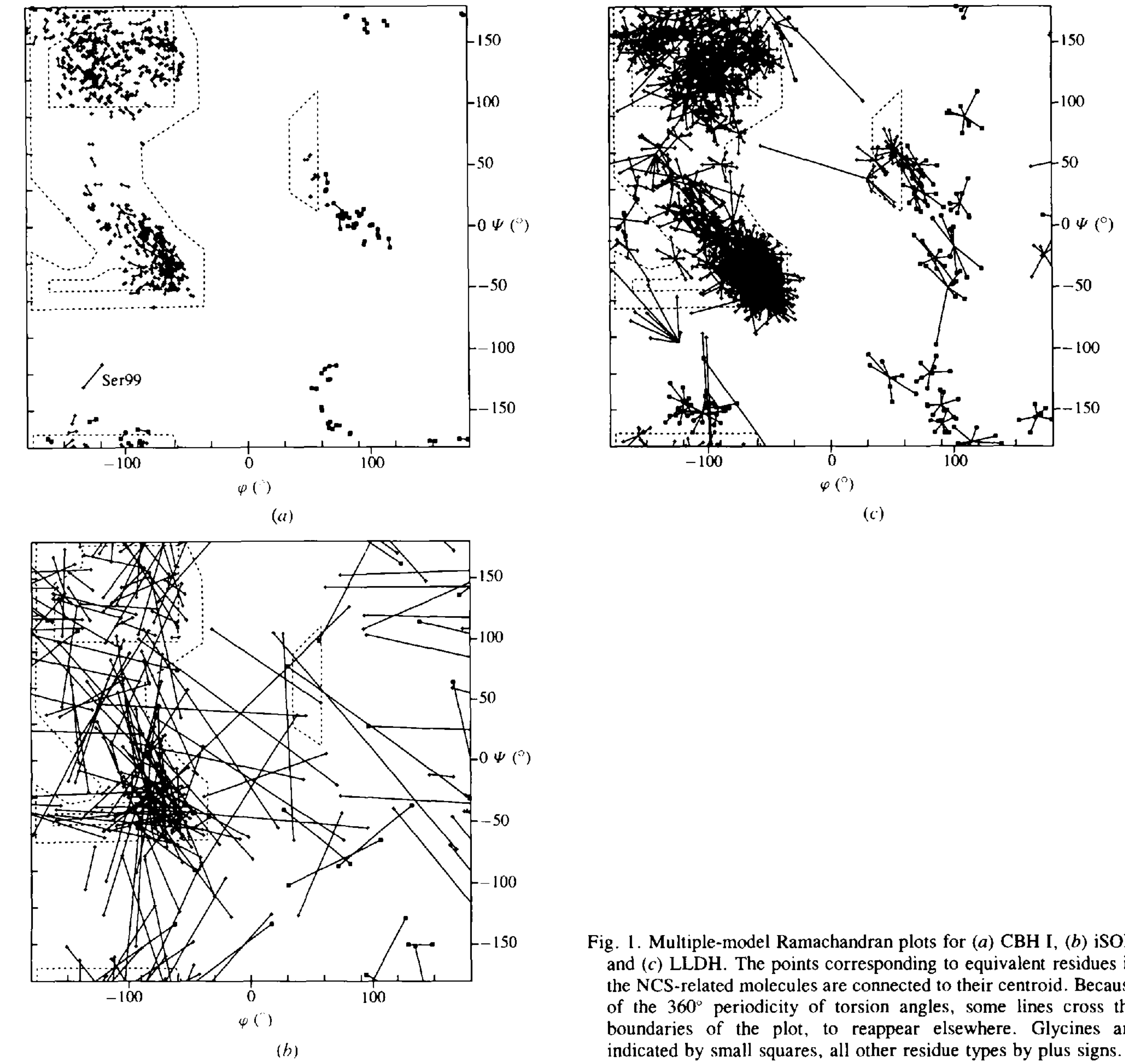

Fig. 1. Multiple-model Ramachandran plots for (*a*) CBH I, (*b*) iSOD and (*c*) LLDH. The points corresponding to equivalent residues in the NCS-related molecules are connected to their centroid. Because of the 360° periodicity of torsion angles, some lines cross the boundaries of the plot, to reappear elsewhere. Glycines are indicated by small squares, all other residue types by plus signs.

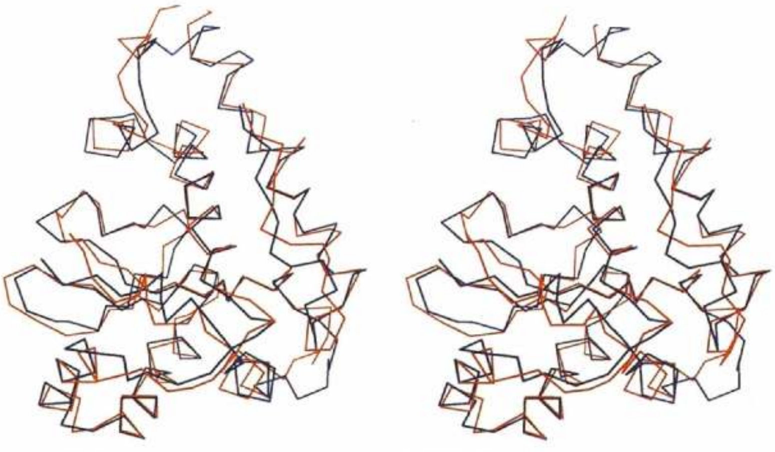

Fig. 2. Superimpositioning of the Cα-traces of the two NCS-related molecules in the structure of iSOD which shows that the two molecules are peculiarly different.

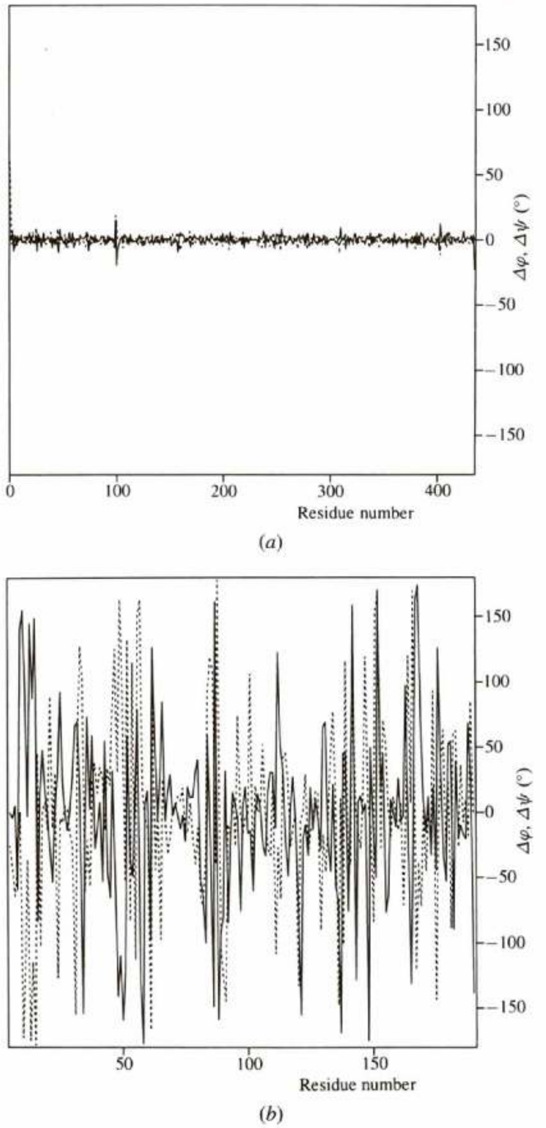

Fig. 3. $\Delta\varphi$, $\Delta\psi$ plots for (*a*) CBH I, and (*b*) iSOD. The solid curve shows the difference between the φ angles of corresponding residues in the two NCS-related molecules; the dashed curve shows the difference between their ψ angles.

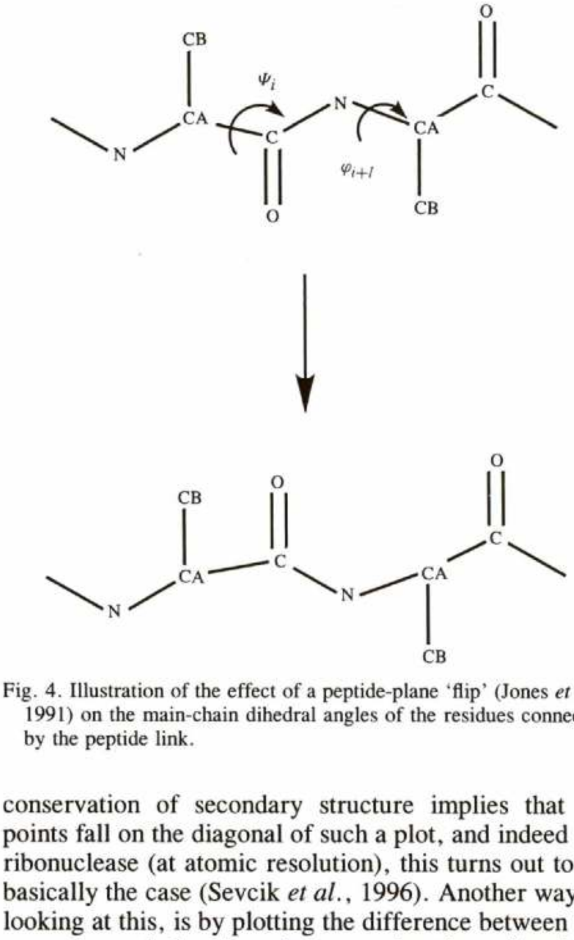

Fig. 4. Illustration of the effect of a peptide-plane 'flip' (Jones *et al.*, 1991) on the main-chain dihedral angles of the residues connected by the peptide link.

conservation of secondary structure implies that all points fall on the diagonal of such a plot, and indeed for ribonuclease (at atomic resolution), this turns out to be basically the case (Sevcik *et al.*, 1996). Another way of looking at this, is by plotting the difference between the φ angles and the ψ angles as a function of residue number [$\Delta\varphi$, $\Delta\psi$ plots (Korn & Rose, 1994; Kleywegt & Jones, 1995*a*)]. Fig. 3 shows such curves for CBH I and iSOD. The conclusion to be drawn from these plots is the same as that which was based on the multiple-model Ramachandran plot.

Note that some of the spikes in a $\Delta\varphi$, $\Delta\psi$ plot may be due to different orientations of the peptide O atoms (Kleywegt & Jones, 1995*a*). As illustrated in Fig. 4, a peptide 'flip' (Jones, Zou, Cowan & Kjeldgaard, 1991) alters the ψ angle of residue i and the φ angle of residue $i+1$ by ~ 150–180°. This also means that outliers in a Ramachandran plot can sometimes be explained (and corrected) by closer inspection of the orientation of the peptide plane.

In cases where there are more than two NCS-related molecules (or other multiple models, such as with 'families' of structures derived by NMR), a more useful way of visualizing the variation in the main-chain dihedral angles is by plotting the standard deviation of φ and ψ, again as a function of residue number. Fig. 5 shows such a plot for iSOD and LLDH. Note that the plot for LLDH reveals that there are really only two places in the sequence where the eight different molecules display significantly non-random variations, whereas the differences between the two iSOD molecules are distributed throughout the entire model. Although the overall fold of the two iSOD molecules is more or less the same, Fig. 2, large local variations in the φ and ψ angles occur, which also result in a high value for the r.m.s.d.

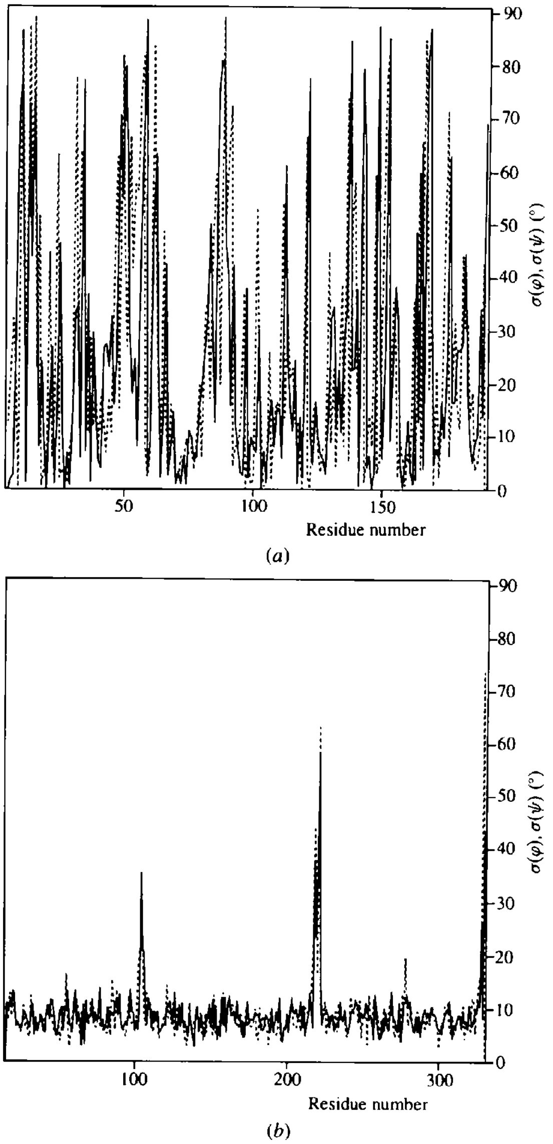

Fig. 5. $\sigma(\varphi)$, $\sigma(\psi)$ plots for (*a*) iSOD, and (*b*) LLDH. The solid curve shows the standard deviation of the φ angles of corresponding residues in all NCS-related molecules; the dashed curve shows the standard deviation of their ψ angles.

3.3. χ_1 and χ_2 plots

Any plot that can be produced for φ, ψ angles can also be made for χ_1, χ_2 side-chain torsion angles, albeit that such angles are conventionally mapped into the range [0°, 360°>, rather than [−180°, +180°>. Fig. 6 shows a multiple-model χ_1, χ_2 plot for CBH I in which the individual residues in both NCS-related molecules have been connected. This plot again emphasizes that CBH I is a well behaved and well refined protein. Fig. 7 shows the standard deviation of the side-chain torsion angles for the structures of CBH I and iSOD as a function of residue number.

3.4. Cα coordinates

In the absence of full coordinate sets, adherence to the principle of conservation of secondary structure can still be assessed by comparison of the geometry in 'Cα space' (Oldfield & Hubbard, 1994). If two structures have identical secondary structure, then the angles and dihedrals formed by subsequent Cα atoms should be the same for both structures. Fig. 8 shows plots of the

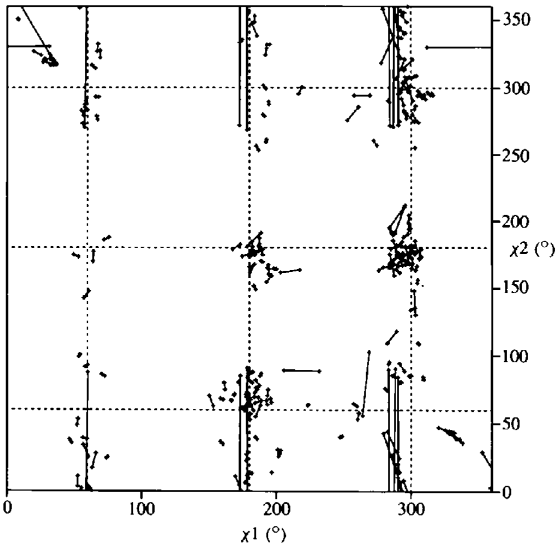

Fig. 6. Multiple-model χ_1, χ_2 plot for CBH I. The points corresponding to equivalent residues in the NCS-related molecules are connected to their centroid. Because of the 360° periodicity of torsion angles, some lines cross the boundaries of the plot, to reappear elsewhere. See also the legend of Fig. 7.

differences in the Cα angles and dihedrals for CBH I and iSOD. Clearly, such plots smooth out much of the noise associated with local differences in φ and ψ.

3.5. *Temperature factors*

A direct comparison of temperature factors between NCS-related molecules is not always valid, since the molecules sometimes display different average temperature factors. However, the correlation coefficient of the temperature factors of two sets of NCS-related core Cα atoms should be high. One may plot the standard deviation and the observed range of temperature factors as a function of residue number, calculated after the average temperature factor for each individual molecule has been subtracted. Fig. 9 shows such a plot for the eight NCS-related molecules of LLDH. Note the rather high level of noise (~ 10 Å^2) and the high average level of the spread (~ 30 Å^2). Both would appear to be

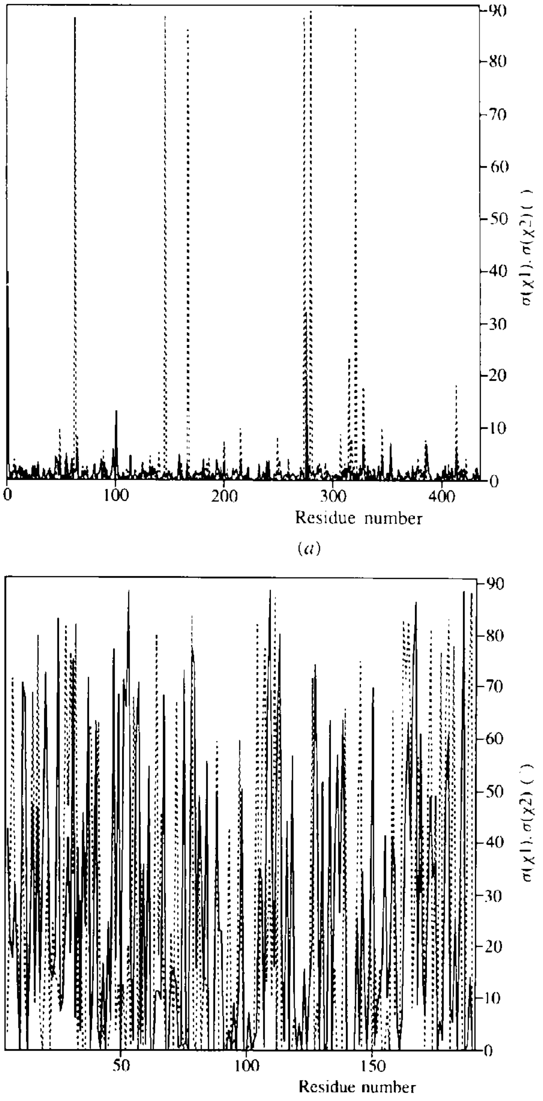

Fig. 7. $\sigma(\chi_1)$, $\sigma(\chi_2)$ plots for (*a*) CBH I, and (*b*) iSOD. The solid curve shows the standard deviation of the χ_1 angles of corresponding residues in all NCS-related molecules; the dashed curve shows the standard deviation of their χ_2 angles. In CBH I there are only two residues for which $\sigma(\chi_1)$ exceeds 10° (Gln101 and Pro276). The six residues whose χ_2 angles differ by almost 180° are all phenylalanines, tyrosines and aspartates, which means that these side chains are in chemically indistinguishable conformations. Apart from these six, there are only five residues in CBH I for which $\sigma(\chi_2)$ exceeds 10° (Pro276, Asn315, Asp328, Asp345 and Asn413).

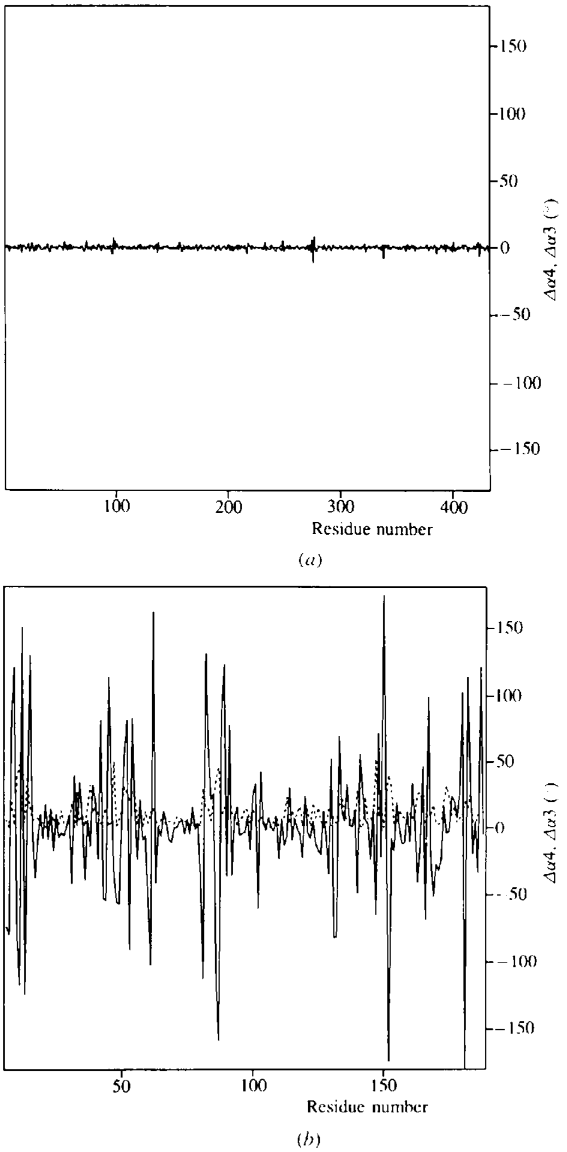

Fig. 8. Cα-geometry difference plots for (*a*) CBH I, and (*b*) iSOD. The solid curve shows the difference of the Cα—Cα*—Cα—Cα (α4) dihedrals of corresponding residues in the NCS-related molecules; the dashed curve shows the differences between the Cα—Cα*—Cα (α3) angles. Angles and dihedrals are calculated relative to the residue marked with an asterisk (*).

unrealistic and are probably a result of over-fitting of the low-resolution (2.5 Å) data. For comparison, for CBH I the average value of $\sigma(B)$ is 0.6 Å^2, and the average magnitude of the B-factor range is 1.2 Å^2.

3.6. *Validation statistics*

In addition to these plots, several overall statistics can be calculated. For example, when comparing the main-chain dihedral angles of two molecules one could quote the r.m.s. value of $\Delta\varphi$ and $\Delta\psi$, or the percentage of residues for which $|\Delta\varphi|$ or $|\Delta\psi|$ exceeds a certain threshold (for instance, 10°). Table 1 lists some of these statistics for CBH I, iSOD and LLDH. Note that iSOD has a very low value for the r.m.s. ΔB of NCS-related core Cα atoms. However, the correlation coefficient of the temperature factors of these atoms is also low, which means that, although the variations are small, they do not correlate very well in the two NCS-related molecules.

Since not many people appear to have checked these properties of their models in the past, they constitute a set of powerful structure-validation tools. However, one should realise that any model property which is monitored and restrained during rebuilding and refinement, cannot be used to validate the structure afterwards [the conventional R factor being the most notorious example of this (Kleywegt & Jones, 1995*a*)]. For example, a model which is refined with restraints on the temperature factors of bonded atoms or on the dihedral angles of NCS-related residues will obviously end up with acceptable values for any criterion which tests these properties. However, if the NCS is ignored during the refinement (either throughout, or in the final stages), the statistics and plots described here are useful in assessing whether this has lead to artefactual differences between the NCS-related molecules or not.

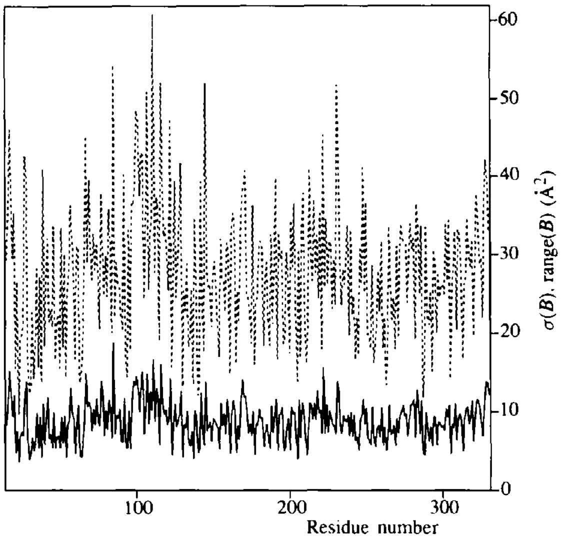

Fig. 9. Plot of the standard deviation (solid curve) and the magnitude of the observed range (dashed curve) of temperature factors for the Cα atoms of all eight NCS-related molecules in the structure of LLDH.

Table 1. *Statistics measuring the degree of similarity of NCS-related molecules (see the text for details)*

Statistic	CBH I	LLDH	iSOD
PDB code	1CEL	1LDN	3SDP
NCS units	2	8	2
Residues per unit	434	316	186
Nominal resolution (Å)	1.8	2.5	2.1
R.m.s.d., all Cα atoms (Å)	0.09	0.32–0.68	1.84
R.m.s.d., all atoms (Å)	0.26	0.55–1.06	2.73
No. of core residues*	434	313	172
R.m.s.d., core Cα atoms (Å)*	0.09	0.45	1.49
R.m.s. ΔB, core Cα atoms (Å^2)*	1.5	14.0	1.2
Correlation coefficient B's, core Cα atoms*	0.97	0.26	0.38
R.m.s. $\Delta\varphi$ (°)*	3.3	20.5	69.0
$\langle\|\Delta\varphi\|\rangle$ (°)*	2.4	12.7	49.6
Residues with $\|\Delta\varphi\| > 10°$ (%)*	0.9	45.9	80.6
R.m.s. $\Delta\psi$ (°)*	4.2	22.2	71.1
$\langle\|\Delta\psi\|\rangle$ (°)*	2.4	13.1	51.3
Residues with $\|\Delta\psi\| > 10°$ (%)*	1.2	48.4	79.6
R.m.s. Δ(Cα—Cα—Cα—Cα dihedral) (°)*	1.8	19.3	54.3
$\langle\|\Delta$(Cα—Cα—Cα—Cα dihedral)$\|\rangle$ (°)*	1.3	8.2	35.8
Residues with $\|\Delta$(Cα—Cα—Cα—Cα dihedral)$\| > 10°$ (%)*	0.2	16.3	65.6
R.m.s. Δ(Cα—Cα—Cα angle) (°)*	1.1	5.9	15.9
$\langle\|\Delta$(Cα—Cα—Cα angle)$\|\rangle$ (°)*	0.8	4.7	11.8
Residues with $\|\Delta$(Cα—Cα—Cα angle)$\| > 5°$ (%)*	0.2	39.3	67.8
$\langle\sigma(\varphi)\rangle$ (°)	1.2	9.0	24.8
$\langle\|\varphi$ range$\|\rangle$ (°)	2.4	28.4	49.6
$\langle\sigma(\psi)\rangle$ (°)	1.2	8.9	25.6
$\langle\|\psi$ range$\|\rangle$ (°)	2.4	28.0	51.3
$\langle\sigma(\chi_1)\rangle$ (°)	1.2	10.8	25.9
$\langle\|\chi_1$ range$\|\rangle$ (°)	2.4	34.6	51.8
$\langle\sigma(\chi_2)\rangle$ (°)	2.3	13.1	22.5
$\langle\|\chi_2$ range$\|\rangle$ (°)	4.6	39.9	45.0
$\langle\sigma(B)\rangle$, all Cα atoms (Å^2)	0.6	8.9	0.5
$\langle\|B$ range$\|\rangle$, all Cα atoms (Å^2)	1.2	27.7	1.0

* Statistics which compare two molecules; for these the two least similar chains in LLDH were used (chains *A* and *C*).

4. Quality of structures with NCS

In order to investigate the effect of limited amounts of data on the quality of the resulting models, we have compiled a *Quality DataBase* (QDB). This QDB contains statistics pertaining to 476 protein entries from the PDB (Bernstein *et al.*, 1977), all of which have either been solved at low resolution, or contain NCS, or both (however, no virus structures have been included). They have in common that the ratio of experimental diffraction observations to adjustable model parameters was low. The sample contains 220 structures with NCS in the resolution range 1.5–2.5 Å, and 256 structures

Table 2. *Statistics pertaining to a sample of 476 protein structures from the PDB which contain NCS, and/or were solved at low resolution (see the text for details)*

Statistic	Average	Minimum and/or maximum	Remarks
Resolution (Å)	2.4	1.5–3.5	
R factor	0.186	0.110–0.370	
Temperature factors			
$\langle B \rangle$, all atoms (Å^2)	23.5	Max. 66.5	Max. for 1CGP (3.0 Å)
$\langle B \rangle$, all waters (Å^2)	25.5	Max. 84.9	Max. for 2DRC (1.9 Å)
Maximum *B*, water (A^2)	49.5	Max. 156	Max. for 1DRA (1.9 Å)
R.m.s. ΔB, bonded atoms (Å^2)	4.1	0.09–32.2	Max. for 2HIP (2.5 Å)
Non-crystallographic symmetry*			
R.m.s.d., core Cα atoms (Å)	0.46	0.0–1.63	Max. for 8FAB (1.8 Å)
$\langle \vert\Delta\varphi\vert \rangle$ (°)	10.2	0.0–49.6	Max. for 3SDP (2.1 Å)
Residues with $\vert\Delta\varphi\vert > 10°$ (%)	29.1	0.0–80.6	Max. for 3SDP (2.1 Å)
$\langle \vert\Delta\psi\vert \rangle$ (°)	10.3	0.0–51.4	
Residues with $\vert\Delta\psi\vert > 10°$ (%)	28.4	0.0–79.6	
$\langle \vert\Delta$(Cα—Cα—Cα—Cα dihedral)$\vert \rangle$ (°)	6.2	0.0–35.8	
Residues with $\vert\Delta$(Cα—Cα—Cα—Cα dihedral)$\vert$ >10° (%)	14.7	0.0–71.6	Max. for 2TUN (3.1 Å)
$\langle \vert\Delta$(Cα—Cα—Cα angle)$\vert \rangle$ (°)	3.5	0.0–11.8	
Residues with $\vert\Delta$(Cα—Cα—Cα angle)$\vert$ > 5° (%)	20.7	0.0–70.9	
Geometry and side-chain packing			
Residues in core Ramachandran-plot regions (%)†	86.3	50.0–98.0	Min. for 4RCR (2.8 Å)
Residues in disallowed Ramachandran-plot regions (%)†	0.5	0.0–6.7	Max. for 2GLS (3.5 Å)
Residues in α-helices and β-strands (%)‡	60.7	22.9–87.8	Min. for 7WGA (2.0 Å)
Omega dihedral standard deviation (°)†	4.0	0.5–12.9	Max. for 1BAA (2.8 Å)
Zeta virtual torsion angle standard deviation (°)†	2.3	0.5–9.5	Max. for 4HHB (1.74 Å)
Overall *G* factor†	−0.45	−7.7– +0.40	Min. for 4HHB (1.74 Å)
DACA score§	−0.79	−2.8– +0.14	Min. for 1PI2 (2.5 Å)
Residues with unusual peptide-plane orientation (%)‡	2.6	0.0–13.3	Max. for 3AAT (2.8 Å)
Residues with non-rotamer side-chain conformations (%)‡	13.1	1.8–44.5	Max. for 1RFB (3.0 Å)

* Values calculated with *LSQMAN* (this work). † Values calculated with *ProCheck* (Laskowski *et al.*, 1993; Laskowski, MacArthur & Thornton, 1994). ‡ Values calculated with *O* (Jones *et al.*, 1991; Zou & Mowbray, 1994). § Values calculated with *WhatIf* (Vriend & Sander, 1993).

(with and without NCS) in the resolution range 2.5–3.5 Å. Of the 476 structures, 20 were deposited with the PDB between 1976 and 1988 (included for reasons of historical interest), 76 in 1989/90, 83 in 1991, 157 in 1992 and 140 in 1993. The structures contain between 52 and 825 residues per chain, and between 320 and 46000 atoms in total. 131 structures did not contain NCS (which means that roughly half of the low-resolution structures do contain NCS), 267 structures contain twofold NCS and 78 structures contain three or more copies per asymmetric unit, the maximum being 12. For each of these structures, statistics and information pertaining to the refinement, *etc.*, were derived from the PDB coordinate entries (human inspection), and by running *ProCheck* (Laskowski, MacArthur, Moss & Thornton, 1993), *O* (Jones *et al.*, 1991), *WhatIf* (Vriend & Sander, 1993), *LSQMAN* (this work) and several local programs. Table 2 shows some of the statistics obtained. Although some of the other results deserve closer study as well, here only the ones that pertain to the similarity of NCS-related molecules will be discussed.

Fig. 10(*a*) shows the distribution of r.m.s.d. values between core NCS-related Cα atoms, as a function of the resolution of the study. There is a certain spread which gets wider as the resolution decreases, as one might expect (due to the increasingly infavourable data-to-parameter ratio). However, some of the largest differences are observed at the high-resolution end of the scale, and some of these are discussed below. For comparison, the value for CBH I is 0.09 Å, for iSOD 1.5 Å, and for LLDH it varies between 0.31 and 0.45 Å. On the other hand, some of the smallest deviations are found for structures solved at low resolution. These are structures which have apparently been refined with NCS restraints or constraints, as we consider one should do when there are relatively few experimental observations and the danger of over-fitting is at its greatest (Kleywegt & Jones, 1995*a*).

Fig. 10(*b*) shows a similar plot of the average value of $\vert\Delta\varphi\vert$ over all residues *versus* resolution. This statistic does not suffer from ambiguities in the definition of optimal superpositioning operators (and, hence, from domain movements, except for a few residues in hinge regions). Except for a handful of outliers and the conservatively refined low-resolution models, it is clear that the value of this statistic is correlated with the resolution: crystals which happen to diffract to better than 2 Å yield models with average $\vert\Delta\varphi\vert$ values of less than 15°, whereas crystals which diffract to 3 Å result in models which are almost twice as 'different'.

An even more convincing demonstration of the relationship between the resolution of a study and the extent of the 'observed' differences between NCS-related molecules is given in Fig. 10(*c*). This plot shows the percentage of residues for which $|\Delta\varphi|$ exceeds 10° as a function of resolution. The value of 10° was chosen rather arbitrarily, but it does mean that, when exceeded, the two NCS-related residues are clearly separated in the Ramachandran plot. In addition, for many high-resolution structures in which the NCS was not restrained during refinement (such as CBH I), the values for r.m.s.$(\Delta\varphi)$ and r.m.s.$(\Delta\psi)$ are of the order of 3-4°, which means that a cut-off value of 10° is a generous estimate of 'random scatter'. Therefore, this particular statistic can be interpreted as the fraction of residues for which the differences between the φ angles are non-random (but note that 'non-random' is not necessarily the same as 'significant' or 'real'). Fig. 10(*c*) shows that the relationship is linear for all intents and purposes (the corresponding plot for the ψ angles looks the same; data not shown). The structure with the highest value for this statistic in our sample is iSOD, which has 80.6% of its residues differing by 10° or more in their φ angles. Overall, for more than half of the structures in our sample this value is greater than 25%.

The outliers at the high-resolution end of the plot of r.m.s.d. *versus* resolution (Fig. 10*a*) have been labelled with their PDB code. 3SDP is the iSOD structure used as one of the examples in this paper. For the other three

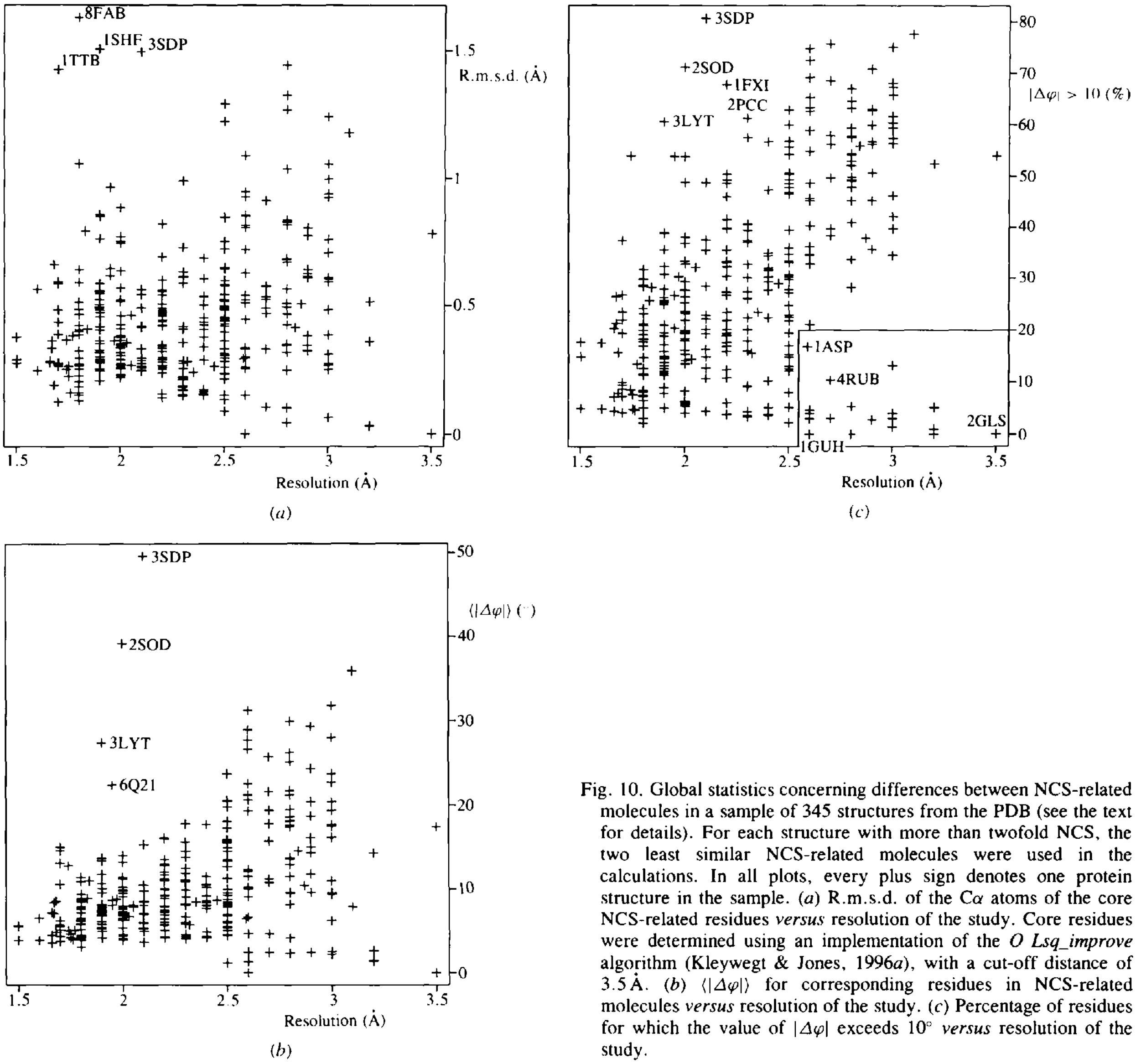

Fig. 10. Global statistics concerning differences between NCS-related molecules in a sample of 345 structures from the PDB (see the text for details). For each structure with more than twofold NCS, the two least similar NCS-related molecules were used in the calculations. In all plots, every plus sign denotes one protein structure in the sample. (*a*) R.m.s.d. of the Cα atoms of the core NCS-related residues *versus* resolution of the study. Core residues were determined using an implementation of the *O Lsq_improve* algorithm (Kleywegt & Jones, 1996*a*), with a cut-off distance of 3.5 Å. (*b*) $\langle|\Delta\varphi|\rangle$ for corresponding residues in NCS-related molecules *versus* resolution of the study. (*c*) Percentage of residues for which the value of $|\Delta\varphi|$ exceeds 10° *versus* resolution of the study.

outliers there are good explanations for the differences: in the case of 8FAB [an immunoglobulin structure (Saul & Poljak, 1992)], the high r.m.s.d. is due to a difference in the relative orientation of the two domains, Fig. 11(*a*), which shows up nicely in a $\Delta\varphi,\Delta\psi$ plot as an isolated region with several spikes, Fig. 11(*b*). In the SH_3 domain of human Fyn [1SHF (Noble, Musacchio, Saraste, Courtneidge & Wierenga, 1993)] there is one loop (residues 112–118) which is different in both molecules, Fig. 12. 1TTB is a transthyretin (prealbumin) mutant (Hamilton *et al.*, 1993) which shows three isolated areas of differences between the two NCS-related molecules (data not shown). All three structures appear to be well refined and they obey the NCS to a large extent, as evidenced by the fact that none of these structures is an outlier in any of the other plots in Fig. 10. 3SDP, on the other hand, is also an outlier in Figs. 10(*b*) and 10(*c*).

Outliers at the high-resolution end of Fig. 10(*b*) include, besides 3SDP, 2SOD [an older Cu,Zn superoxide dismutase structure (Tainer, Getzoff, Beem, Richardson & Richardson, 1982)], 3LYT [a 100 K structure of hen egg-white lysozyme (Young, Dewan, Nave & Tilton, 1993)] and 6Q21 [the catalytic domain of Ras P21 protein (Privé *et al.*, 1992)]. The differences between the NCS-related molecules of 2SOD appear to be of the same type as those of 3SDP. In the case of 3LYT, there are four regions which display main-chain torsional differences. The four NCS-related molecules of 6Q21 all contain a disordered region of approxi-

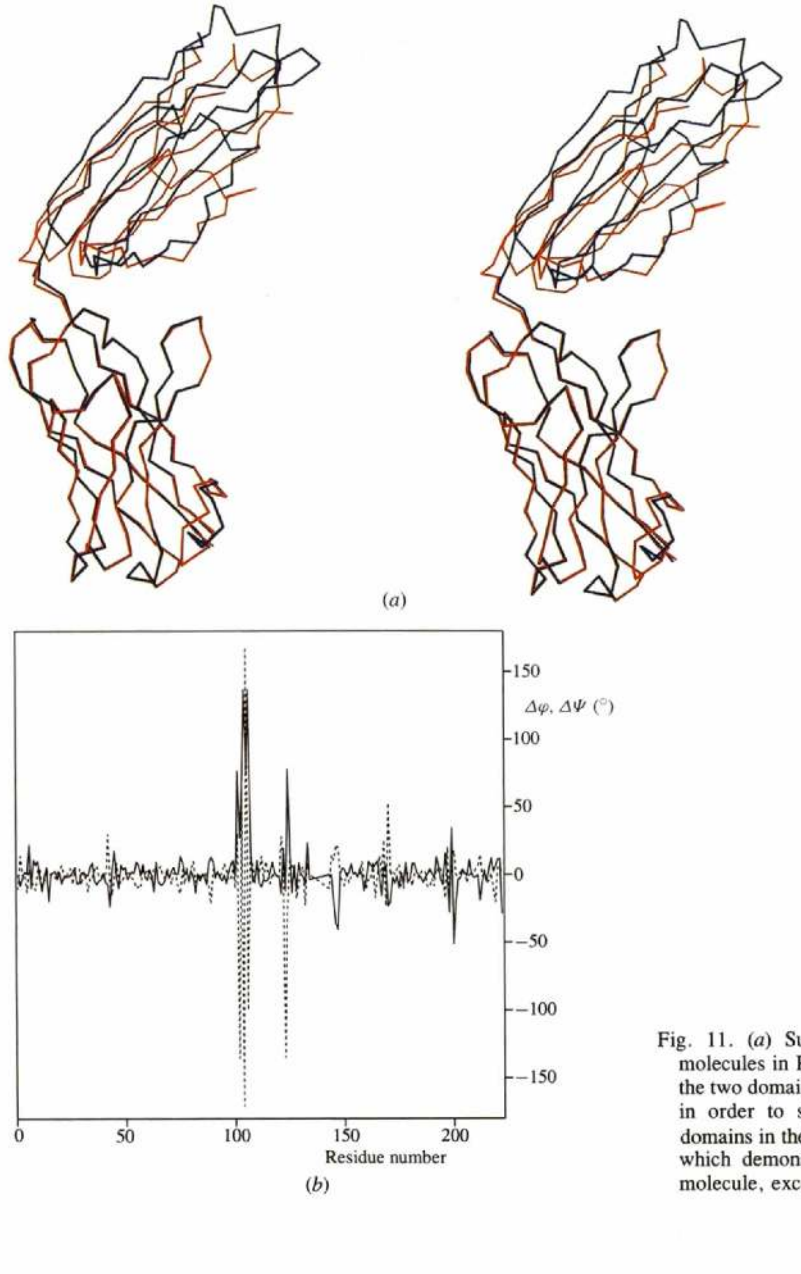

Fig. 11. (*a*) Superpositioning of the Cα traces of the two F_{AB} molecules in PDB entry 8FAB (Saul & Poljak, 1992); only one of the two domains was used to calculate the superpositioning operator in order to show the different relative orientation of the two domains in the two molecules. (*b*) $\Delta\varphi$, $\Delta\psi$ plot for the same model which demonstrates that the NCS is obeyed everywhere in the molecule, except in the interdomain loop.

mately ten residues which, lacking well defined density, have been built in different ways.

Outliers at the high-resolution end of Fig. 10(c) are 3SDP, 2SOD, 3LYT, 1FXI and 2PCC. 1FXI is the structure of ferredoxin I, with fourfold NCS (Tsukihara *et al.*, 1990). Interestingly, there are no spikes at all in the $\sigma(\varphi)$, $\sigma(\psi)$ plot (data not shown); the high average value is entirely due to a much higher noise level than usual at this resolution (the average absolute values of $\Delta\varphi$ and $\Delta\psi$ are $\sim 15°$, roughly twice the average value of $\sim 8°$ for all structures between 1.5 and 2.2 Å). Something similar is observed in the case of 2PCC, a complex of cytochrome c peroxidase and iso-1-cytochrome c (Pelletier & Kraut, 1992). Again, there is a fairly high noise level (in particular for the iso-1-cytochrome c molecules), but there are no spikes in the $\Delta\varphi$, $\Delta\psi$ plot of either the peroxidase or the cytochrome (data not shown).

On the other end of the resolution scale of Fig. 10(c), there are only 18 structures in this survey which have been refined at a resolution lower than 2.5 Å and which have fewer than 20% of their residues differing by more than 10° in their φ angle. Of these 18 structures, two were refined with NCS constraints [2GLS, glutamine synthetase (Almassy, Janson, Hamlin, Xuong & Eisenberg, 1986; Yamashita, Almassy, Janson, Cascio & Eisenberg, 1989), and 1GUH, human alpha class glutathione S-transferase (Sinning *et al.*, 1993)]. (There are more models that have been refined with constrained NCS in the PDB, but for these the coordinates of only one molecule were deposited, which meant that they were not recognised by our automated procedure to select PDB entries containing NCS.) One structure was refined with a mixture of NCS constraints and restraints [4RUB, tobacco Rubisco (Schreuder *et al.*, 1993)]. 13 structures were refined with NCS restraints during the entire or most of the refinement process, and for the entire structure or most parts of the structure. For one structure no mention is made of the NCS model (neither in the PDB file, nor in the original paper), and only one of the 18 structures (the one with the highest fraction of

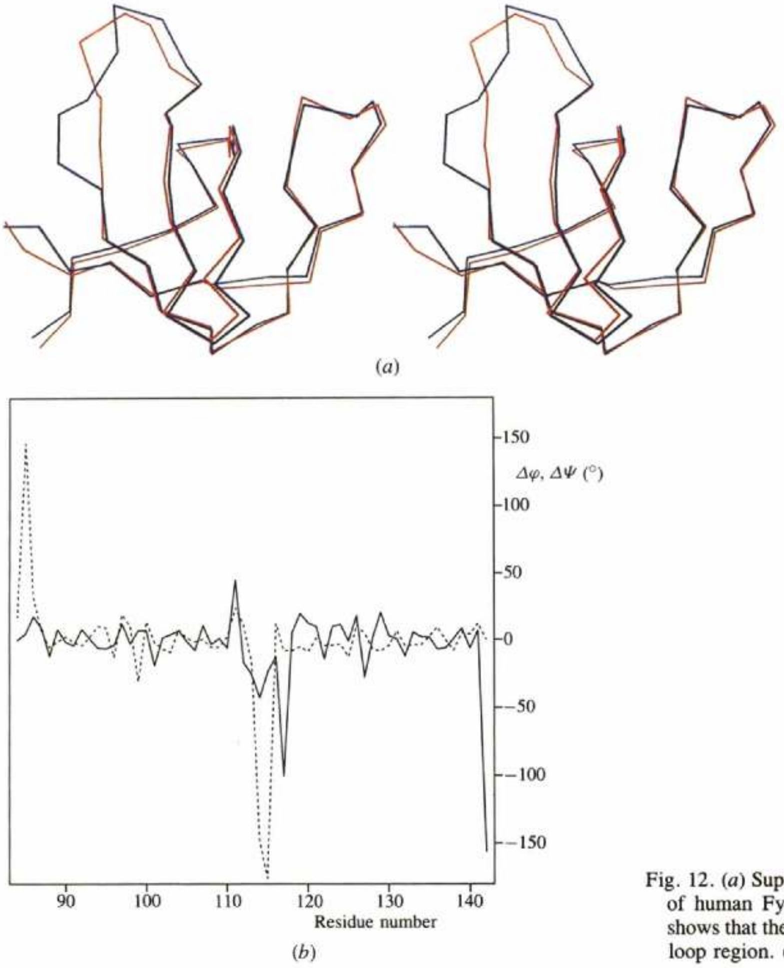

Fig. 12. (a) Superpositioning of the Cα traces of the two SH_3 domains of human Fyn in PDB entry 1SHF (Noble *et al.*, 1993), which shows that the molecules are fairly similar with the exception of one loop region. (b) $\Delta\varphi$, $\Delta\psi$ plot for the same model.

residues, namely 16.8%, differing by more than 10° of these 18 structures) has not been refined with any NCS constraints or restraints at all [entry 1ASP, the peroxide form of ascorbate oxidase, refined at 2.59 Å (Messerschmidt, Steigemann, Huber, Lang & Kroneck, 1992)].

5. Implications for refinement

The observation that the extent to which NCS-related molecules differ is (linearly) related to the resolution at which they were refined begs the question whether this is a reflection of a genuine phenomenon or not. One could, for instance, postulate that proteins which form poorly diffracting crystals are inherently more flexible, which leads to larger differences between NCS-related molecules. Alternatively, one could argue that in cases in which the ratio of the number of experimental diffraction observations to the number of refined model parameters is low, the refinement program is invited to over-fit the model, and that this over-fitting leads to artefactual differences between the NCS-related molecules (Kleywegt & Jones, 1995*a*,*b*). We believe that the second provides a considerably more plausible explanation for the majority of cases than the former, for the following reasons.

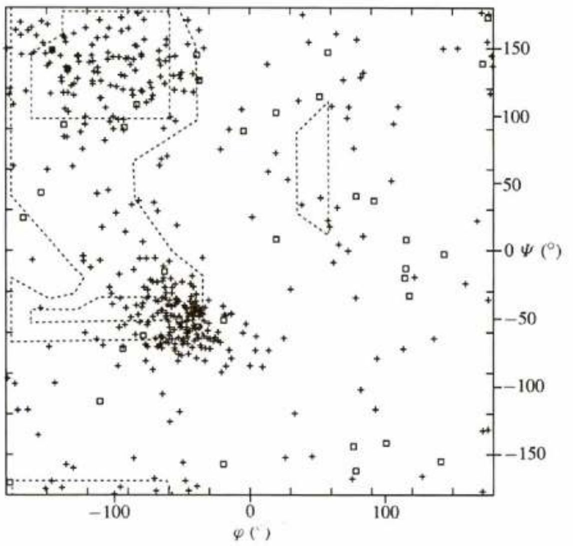

Fig. 13. Ramachandran plot for the structure of glutamine synthetase (one monomer). Even though this structure was refined (at 3.5 Å) with constrained 12-fold NCS, the Ramachandran plot seems to indicate that there are problems with this structure.

Our QDB includes a rough estimate of the experimental diffraction data-to-parameter ratio for each structure. Based on this, it appears that approximately one-fifth of the structures were refined with a diffraction-data-to-parameter ratio smaller than one, whereas for about half the structures the ratio was smaller than 1.5. This means that one may safely assume that many of these structures have been over-fitted.

We have shown previously for chloromuconate cycloisomerase that even molecules which are identical because they are related by crystallographic symmetry can be refined in a lower symmetry space group to end up looking quite different from one another (Kleywegt, Hoier & Jones, 1996). In this case, the r.m.s.d. on Cα atoms between two (crystallographically related) chloromuconate cycloisomerase molecules was 0.86 Å, the r.m.s.d. on all atoms was 1.5 Å, and the r.m.s. $\Delta\varphi$ and $\Delta\psi$ values were ~38°. This demonstrates that even an r.m.s.d. of 1.5 Å after refinement does not exclude the possibility that the molecules are actually identical (or very similar).

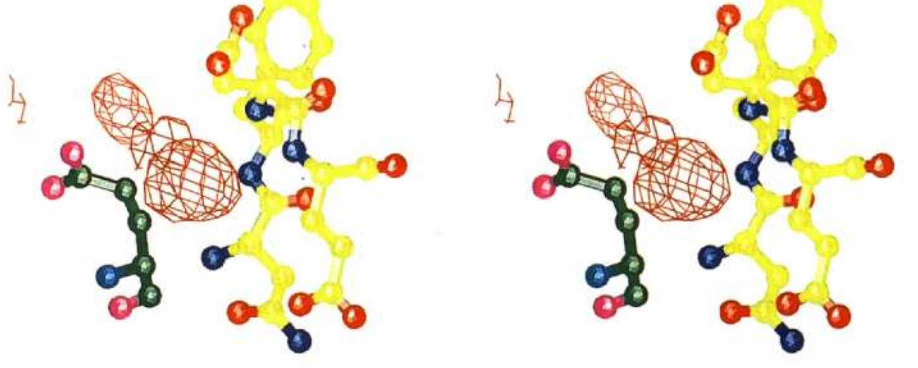

Fig. 14. Example of NCS breakdown. The structure of cellular retinoic-acid-binding protein (Kleywegt *et al.*, 1994) in complex with AM-80, a synthetic retinoid, was refined at 2.7 Å with twofold NCS constraints (unpublished results). Since the two molecules have different environments, some differences between them are to be expected. In this case, Glu13 in a '*B*' molecule (on the left, with C atoms coloured green) clearly has a wrong conformation. The difference density shows that the side chain ought to be rotated, which would bring the carboxylate O atoms within hydrogen-bonding distance from three backbone amide N atoms in an '*A*' molecule (on the right, with C atoms coloured yellow).

The best way to investigate how different NCS-related molecules really are, is to solve the same structure (or very related structures) at high resolution. We have shown previously (Kleywegt & Jones, 1995*a*) for the case of Rubisco that refinement of a 1.8 Å structure in a complex at 2.6 Å leads to fairly large differences between the NCS-related molecules (r.m.s. $\Delta\varphi \simeq 46^\circ$) that were not present in the high-resolution structure (r.m.s. $\Delta\varphi \simeq 18^\circ$). Conversely, we showed that in the case of human alpha class glutathione S-transferase the refinement with constrained NCS at 2.6 Å yielded a model that could easily withstand the comparison with a (non-isomorphous) 2.0 Å model of a mutant protein that was solved later.

Although the evidence is largely anecdotal (it would be impractical to do a large-scale systematic study of this phenomenon, even if the experimental diffraction data were available from the PDB which, in the majority of cases, they are not), it all points to the same conclusion: at low resolution, a failure to exploit the redundancies introduced by the presence of NCS leads to models which display over-estimated, artefactual differences between the NCS-related molecules. Unfortunately, for the large majority of the outliers in the plots of Fig. 10, no structure factors have been deposited with the PDB.

There are many ways to exploit the redundancies because of NCS, when tracing, refining and rebuilding a model. In the presence of NCS, at present most people use molecular averaging in order to improve the electron-density maps in which the initial model is to be traced. As for refinement, most programs allow the use of positional NCS restraints, either by restraining positional differences [*X-PLOR* (Brünger, 1992*a*), *TNT* (Tronrud, Ten Eyck & Matthews, 1987), *PROLSQ* (Hendrickson & Konnert, 1980)], or by restraining corresponding 1-4 distances [*SHELXL* (Sheldrick & Schneider, 1996)]. In addition, some refinement programs allow one to restrain the temperature factors of NCS-related atoms to be similar. When one decides to use NCS restraints, it is probably a good idea to design an appropriate weighting scheme. Typically, the atoms would be divided into at least four classes: main-chain and side-chain atoms, each divided into sets that obey the NCS almost strictly and others that deviate from it.

NCS may also be constrained, for example in *X-PLOR* (Brünger, 1992*a*), which means that in effect only one copy of the molecule is ever refined (the others are generated implicitly for structure-factor calculations only). Alternatively, one may use a mixture of constraints and restraints. For example, if one has two dimers in the asymmetric unit, one could constrain the dimers to be identical, but restrain the monomers that make up a dimer to be similar.

Finally, one may release the NCS entirely, but the dangers involved when this is done at too low a resolution (or too early a stage of the refinement) are well illustrated by many of the figures in this paper.

Clearly, if one has *N*-fold NCS, the diffraction-data-to-parameter ratio can be improved by a factor *N* (if the NCS is constrained; almost *N* if restraints are used instead), which may make the difference between successfully refining a low-resolution structure and getting stuck. If one is lucky enough to work on a well behaved protein (in the sense of its obeying the NCS virtually throughout the entire structure), NCS constraints can and should be used. We also tend to use NCS averaging of the density maps for use in rebuilding. We have noticed on several occasions that at resolutions up to ~2 Å such averaged maps tend to be superior to the unaveraged ones obtained when one replaces the NCS constraints by restraints (GJK, unpublished observations).

The final argument in favour of NCS constraints will appeal to those who are in a hurry to publish a structure or, as we are, lazy: both the rebuilding of a model and its refinement can be carried out almost *N* times faster if *N*-fold NCS constraints are used.

The fact that we advocate the use of NCS constraints at low resolution does not imply that we think that all NCS-related protein molecules are necessarily identical. However, we do feel that at low resolution there is often insufficient experimental data to model any differences that may exist. Once differences are allowed to be modelled (especially, when the NCS is ignored completely), any refinement program will take this freedom and introduce differences (Kleywegt & Jones, 1995*a*). The question one has to ask is if the differences obtained are related to any real differences. The evidence in the case of Rubisco and chloromuconate cycloisomerase is against this.

Fig. 10(*a*) shows that high-resolution structures have an average r.m.s.d. on core Cα atoms of ~0.4 Å, and the average r.m.s.d. on all atoms is roughly 1.0 Å (data not shown). This does not mean that one should aim to obtain such differences at very low resolution as well since, although any refinement program can undoubtedly be tuned to produce such differences, there is no guarantee that the observed differences are related to real differences between the molecules.

On the other hand, one may ask if an equally simplifying assumption as constrained NCS leads to better models. The only example we have so far involves glutathione S-transferase, and in that case the answer is a resounding 'yes', since the conservatively refined 2.6 Å model is very similar to the 2.0 Å structure solved later (Kleywegt & Jones, 1995*a*). However, constraining the NCS is no guarantee for a high-quality model. For example, the structure of glutamine synthetase (Almassy *et al.*, 1986; Yamashita *et al.*, 1989) (PDB code 2GLS) was refined at 3.5 Å with constrained 12-fold NCS. Nevertheless, the Ramachandran plot of this structure, Fig. 13, indicates that there may still be some local problems with the final model. This is indicative of the problems involved in correctly identifying the orientation of peptide planes at low resolution without the use of main-chain databases (Jones *et al.*, 1991).

6. Breakdown of NCS

Sometimes there are clear indications that NCS breaks down. Such indications may stem from biochemical experiments which assign different roles to multiple copies of a subunit in a complex, such as in the case of F_1 ATPase (Abrahams, Leslie, Lutter & Walker, 1994). If this is not the case, one should be very critical and try to prove that the assumed NCS breaks down. The best way of doing this, is by collecting very high resolution data, but this, of course, is not always possible. In that case, one should inspect the averaged and unaveraged maps closely and look for places where the averaged density is poor, and where the unaveraged and difference maps show indications of different conformations in different molecules (see Fig. 14 for an example). Absence of averaged density alone in our experience often indicates general (crystallographic) disorder of a loop or side chain and is no reason for relaxation of NCS constraints or restraints.

However, even in the case where constrained NCS is not a valid assumption, NCS restraints can still be used for parts of the model which do obey the NCS. For example, if the orientation of two domains is different in NCS-related molecules, each of the individual domains can still be restrained to be similar to its counterparts in the other molecules.

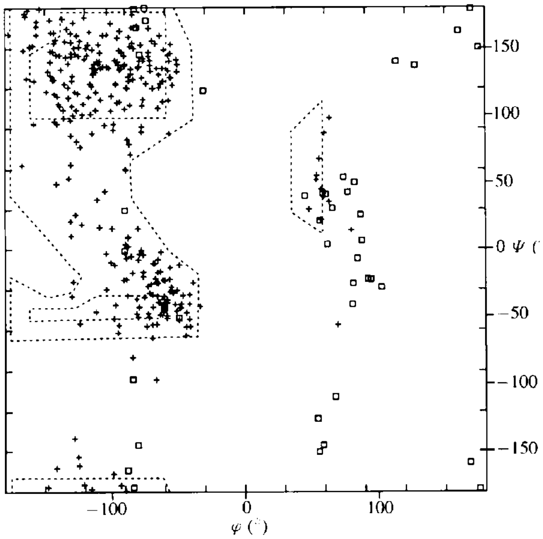

Fig. 15. Ramachandran plot for the final 3.6 Å model of *Trichoderma reesei* EG I (Kleywegt, Zou *et al.*, 1996). Despite the low resolution, the quality of the Ramachandran plot is high.

To assess if the release of constraints or the relaxation of restraints has yielded a better model, one should check if the value of the free *R* factor (Brünger, 1992*b*, 1993) decreases, and of course if the quality of the maps is higher afterwards than it was before. If neither is the case, one is better off retaining the previous constraints or restraints. If one has independent phase information (*e.g.*, derivatives or anomalous data) this should of course be used as well in the calculation of maps to assess how realistic any modelled differences are. However, in cases where such independent phase information is unavailable (*e.g.*, in molecular-replacement studies and for complexes or mutants which are isomorphous to a previously determined structure), $R_{\rm free}$ is at present the best available statistic for assessing which NCS model is the most appropriate (Kleywegt & Jones, 1995*a*, 1996*b*; Dodson, Kleywegt & Wilson, 1996). Relationships between reflections (through the *G* function) may lead to artificially low values of $R_{\rm free}$ in the case of NCS, and one could conceive that these relationships might be so strong as to lead to acceptable values for the free *R* factor even for completely incorrect models. In order to investigate this, we have carried out an experiment using the structure of an orthorhombic crystal form of the protein α2u-globulin, which has fourfold NCS (Kleywegt, Björkman *et al.*, 1996). We intentionally traced the structure of this protein backwards and refined it (at 3.0 Å) using different protocols (Kleywegt & Jones, 1996*b*). The lowest value of $R_{\rm free}$ we obtained was 0.465, which is high enough to indicate a problem with the model. Moreover, this experiment demonstrated another benefit of using constrained or restrained NCS. Whereas the conventional *R* factor was easily brought down to ~0.27 if the NCS was ignored, it was impossible to accomplish this when NCS restraints ($R \simeq 0.35$) or constraints ($R \simeq 0.36$) were used. In other words, with a conservative NCS model at low resolution, even the conventional *R* factor cannot be brought into the realm of respectability if the model is completely wrong.

One final matter concerning refinement with strict NCS concerns the modelling of temperature factors. We have noticed in several refinements that the use of grouped temperature factors (in which, for example, all main-chain and all side-chain atoms of a residue get the same temperature factor) tends to give lower $R_{\rm free}$ values than individual isotropic temperature-factor refinement, even at moderately high resolution (~2.0–2.2 Å). One possible explanation for this phenomenon lies in the fact that there are no restraints on the temperature-factor differences between neighbouring groups. This, in turn, enables a single side chain which is (NCS) disordered to obtain an extremely high temperature factor, whereas nearby atoms which are well ordered have normal temperature factors. If restraints on the temperature factors of bonded atoms are used, such high temperature factors are forced to be 'smeared out' over an entire loop, for instance, and may give a false impression of (NCS-) disorder and generally deteriorate the model.

7. Examples

Historically, NCS is employed most often to average maps at the stage where a structure is traced in an MIR map. However, also in molecular-replacement cases the use of NCS in both (re-)building and refinement can be extremely useful. Perhaps it is even more important to employ the redundancies introduced by the NCS in the refinement of molecular-replacement structures, since there is usually no independent phase information available. To illustrate this, we shall briefly discuss two recent structure determinations in our laboratory (both by means of molecular replacement) in which the availability of NCS was of great importance in enabling the tracing, rebuilding and refinement of the structures.

The structure of *Trichoderma reesei* EG I was solved by molecular replacement at 4.0 Å resolution (Kleywegt, Zou *et al.*, 1996) using the structure of EG I from another organism (Davies, 1995) as the search model. There are two molecules in the asymmetric unit, but this was sufficient to (*a*) confirm the correctness of the molecular-replacement solution, (*b*) enable the tracing of large parts of the model prior to any refinement, (*c*) enable high-temperature simulated-annealing refinement (Brünger & Rice, 1996) to proceed successfully. The molecular-replacement solutions were weak, but thanks to the twofold NCS their correctness could be verified as follows: the solutions were changed into polyalanine models and used to calculate a $2F_o - F_c$ map. This map was uninterpretable, but 15 cycles of twofold averaging produced a map of surprisingly good quality. In this map, major parts of the model (~75%) could be traced and assigned to the sequence. Subsequently, we subjected the model to several simulated-annealing protocols (with NCS constraints and grouped temperature factors), although we did not expect any of these to be successful at this resolution, based on previous experiences (Sauer-Eriksson, Kleywegt, Uhlén & Jones, 1995). However, the refinement was successful, both in terms of the free R factor, and in terms of the quality of the ensuing averaged map in which an additional 60 residues could be traced. Eventually, the synchrotron data were reprocessed to 3.6 Å, and refinement was completed using that data set. The final model is of surprisingly good quality judged by most criteria, despite the low resolution of the data (Kleywegt, Zou *et al.*, 1996). The Ramachandran plot for the final model is shown in Fig. 15.

The structure of acyl-coenzyme A binding protein (Zou, Kleywegt & Jones, 1996) was recently solved by molecular replacement, using a search model composed of the most conserved parts of the 14 most similar models from the family of 20 NMR structures of this protein (Andersen & Poulsen, 1992). We had two different crystal forms, one in space group $P2_12_12_1$ (no NCS, 2.0 Å data), and one in space group $P4_1$ (threefold NCS, 2.4 Å data). A weak solution could only be obtained for the orthorhombic crystal form, but because of the absence of NCS the correctness of the solution could not be verified in the same fashion as for EG I. Therefore, we subjected the (incomplete) model to high-temperature simulated-annealing refinement and tested if this had improved the model sufficiently to enable the solution of the other crystal form. The refinement of the orthorhombic model stalled at a free R factor of ~46% ($R \simeq 36\%$), but still it turned out to be possible to solve the tetragonal crystal form with this model. However, despite the fairly high resolution data for the orthorhombic crystal form, some parts of the map were very poor and the missing parts of the structure could not be built. The threefold averaged density for the tetragonal crystal form, on the other hand, was very clear and enabled unambiguous tracing of the missing residues to yield a complete model. In this case, 2.4 Å data with threefold NCS was clearly more powerful than 2.0 Å data without NCS. The current model has an r.m.s.d. of 2.2 Å on Cα atoms to the starting NMR model, which explains why the molecular-replacement problem was so difficult to solve. The refinement of both crystal forms is currently in progress, and the details and results will be published elsewhere (Zou *et al.*, 1996).

8. Availability

The *Quality DataBase* (including a program for querying and analyzing it and to produce lists of structures sorted by any quality criterion or statistic, as well as some plots) is available freely to anyone interested *via* anonymous ftp from rigel.bmc.uu.se, directory pub/qdb. The program *LSQMAN*, which was used to analyse differences between NCS-related molecules, is available free of charge to academic researchers from the author (e-mail: gerard@xray.bmc.uu.se).

This work was supported by Uppsala University. I would like to thank Professor Alwyn Jones (Uppsala) for encouragement, discussions, and critical comments on this manuscript, as well as Drs Rob Hooft and Gert Vriend (EMBL, Heidelberg) for providing a list of Directional Atomic Contact Analysis scores for most of the structures included in this survey.

References

Abrahams, J. P., Leslie, A. G. W., Lutter, R. & Walker, J. E. (1994). *Nature (London)*, **370**, 621-628.

Almassy, R. J., Janson, C. A., Hamlin, R., Xuong, N. H. & Eisenberg, .D. (1986). *Nature (London)*, **323**, 304-309.

Andersen, K. V. & Poulsen, F. M. (1992). *J. Mol. Biol.* **226**, 1131-1141.

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*). *X-PLOR. A system for crystallography and NMR*, Yale University, New Haven, CT, USA.

Brünger, A. T. (1992*b*). *Nature (London)*, **355**, 472-475.

Brünger, A. T. (1993). *Acta Cryst.* D**49**, 24-36.

Brünger, A. T. & Rice, L. M. (1996). *Methods Enzymol.* In the press.

Chothia, C. & Lesk, A. M. (1986). *EMBO J.* **5**, 823-826.

Davies, G. J. (1995). Personal communication.

Divne, C., Ståhlberg, J., Reinikainen, T., Ruohonen, L., Pettersson, G., Knowles, J. K. C., Teeri, T. T. & Jones, T. A. (1994). *Science*, **265**, 524-528.

Dodson, E. J., Kleywegt, G. J. & Wilson, K. S. (1996). *Acta Cryst.* **D52**, 228-234.

Hamilton, J. A., Steinrauf, L. K., Braden, B. C., Liepniks, J., Benson, M. D., Holmgren, G., Sandgren, O. & Steen, L. (1993). *J. Biol. Chem.* **268**, 2416-2424.

Hendrickson, W. A. & Konnert, J. H. (1980). In *Computing in Crystallography*, edited by R. Diamond, pp. 13.01-13.25. Bangalore: Indian Academy of Science.

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., Le Motte, P., Gsell, B., Shudo, K. & Jones, T. A. (1994). *Structure*, **2**, 1241-1258.

Kleywegt, G. J., Björkman, J., Uppenberg, J., Ogg, D., Lehman-McKeeman, L. D., Oliver, J. D. & Jones, T. A. (1996). In preparation.

Kleywegt, G. J., Hoier, H. & Jones, T. A. (1996). *Acta Cryst.* **D52**, 858-863.

Kleywegt, G. J. & Jones, T. A. (1995*a*). *Structure*, **3**, 535-540.

Kleywegt, G. J. & Jones, T. A. (1995*b*). *Making the Most of your Model*, edited by S. Bailey & W. N. Hunter, pp. 11-24. Warrington: Daresbury Laboratory.

Kleywegt, G. J. & Jones, T. A. (1996*a*). *Methods Enzymol.* In the press.

Kleywegt, G. J. & Jones, T. A. (1996*b*). *Methods Enzymol.* In the press.

Kleywegt, G. J., Zou, J. Y., Divne, C., Sinning, I., Ståhlberg, J., Davies, G. J., Teeri, T. T. & Jones, T. A. (1996). In preparation.

Korn, A. P. & Rose, D. R. (1994). *Protein Eng.* **7**, 961-967.

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

Laskowski, R. A., MacArthur, M. W. & Thornton, J. M. (1994). *From First Map to Final Model*, edited by S. Bailey, R. Hubbard & D. A. Waller, pp. 149-159. Warrington: Daresbury Laboratory.

Messerschmidt, A., Steigemann, W., Huber, R., Lang, G. & Kroneck, P. M. H. (1992). *Eur. J. Biochem.* **209**, 597-602.

Noble, M. E. M., Musacchio, A., Saraste, M., Courtneidge, S. A. & Wierenga, R. K. (1993). *EMBO J.* **12**, 2617-2624.

Oldfield, T. J. & Hubbard, R. E. (1994). *Proteins*, **18**, 324-337.

Pelletier, H. & Kraut, J. (1992). *Science*, **258**, 1748-1755.

Privé, G. G., Milburn, M. V., Tong, L., de Vos, A. M., Yamaizumi, Z., Nishimura, S. & Kim, S. H. (1992). *Proc. Natl Acad. Sci. USA*, **89**, 3649-3653.

Ramakrishnan, C. & Ramachandran, G. N. (1965). *Biophys. J.* **5**, 909-933.

Sauer-Eriksson, A. E., Kleywegt, G. J., Uhlén, M. & Jones, T. A. (1995). *Structure*, **3**, 265-278.

Saul, F. A. & Poljak, R. J. (1992). *Proteins*, **14**, 363-371.

Schreuder, H. A., Knight, S., Curmi, P. M. G., Andersson, I., Cascio, D., Sweet, R. M., Brändén, C.-I. & Eisenberg, D. (1993). *Protein Sci.* **2**, 1136-1146.

Sevcik, J., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1996). *Acta Cryst.* **D52**, 327-344.

Sheldrick, G. M. & Schneider, T. (1996). *Methods Enzymol.* In the press.

Sinning, I., Kleywegt, G. J., Cowan, S. W., Reinemer, P., Dirr, H. W., Huber, R., Gilliland, G. L., Armstrong, R. N., Ji, X., Board, P. G., Olin, B., Mannervik, B. & Jones, T. A. (1993). *J. Mol. Biol.* **232**, 192-212.

Stoddard, B. L., Howell, P. L., Ringe, D. & Petsko, G. A. (1990). *Biochemistry*, **29**, 8885-8893.

Tainer, J. A., Getzoff, E. D., Beem, K. M., Richardson, J. S. & Richardson, D. C. (1982). *J. Mol. Biol.* **160**, 181-217.

Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). *Acta Cryst.* A**43**, 489-501.

Tsukihara, T., Fukuyama, K., Mizushima, M., Harioka, T., Kusunoki, M., Katsube, Y., Hase, T. & Matsubara, H. (1990). *J. Mol. Biol.* **216**, 399-410.

Vriend, G. & Sander, C. (1993). *J. Appl. Cryst.* **26**, 47-60.

Wigley, D. B., Gamblin, S. J., Turkenburg, J. P., Dodson, E. J., Piontek, K., Muirhead, H. & Holbrook, J. J. (1992). *J. Mol. Biol.* **223**, 317-335.

Yamashita, M. M., Almassy, R. J., Janson, C. A., Cascio, D. & Eisenberg, D. (1989). *J. Biol. Chem.* **264**, 17681-17690.

Young, A. C. M., Dewan, J. C., Nave, C. & Tilton, R. F. (1993). *J. Appl. Cryst.* **26**, 309-319.

Zou, J. Y., Kleywegt, G. J. & Jones, T. A. (1996). In preparation.

Zou, J. Y. & Mowbray, S. L. (1994). *Acta Cryst.* **D50**, 237-249.