Acta Cryst. (1995). D**51**, 331–336

A Strategy for Rapid and Effective Refinement Applied to Black Swan Lysozyme

BY Z. RAO AND R. ESNOUF

Laboratory of Molecular Biophysics and Oxford Centre for Molecular Sciences, Oxford University, Oxford OX1 3QU, England

N. ISAACS

Department of Chemistry, Glasgow University, Glasgow G12 8QQ, Scotland

AND D. STUART

Laboratory of Molecular Biophysics and Oxford Centre for Molecular Sciences, Oxford University, Oxford OX1 3QU, England

(*Received* 26 *May* 1994; *accepted* 26 *August* 1994)

Abstract

The crystal structure of a goose-type lysozyme from the egg white of black swan has been determined at 1.9 Å resolution using a semi-automatic procedure based on the C_α coordinates of the homologous goose protein.

Introduction

The goose-type lysozyme (SEWL) from the egg white of black swan, *Cygnus atratus*, consists of a single chain of 185 amino acids with a molecular weight of 20400 (Morgan & Arnheim, 1974). SEWL closely resembles the Embden goose egg-white lysozyme (GEWL) differing in only six residues (Simpson & Morgan, 1983), and the structure of GEWL has been reported (Grütter, Weaver & Matthews, 1983).

The structure of SEWL based on a multiple isomorphous replacement (MIR) analysis at 2.8 Å resolution has been reported previously by Isaacs, Machin & Masakuni (1985). Only the peptide backbone could be traced and in some regions even this was not clearly defined. An attempt to refine a model constructed from the 2.8 Å map failed. Here we report the determination of the structure using the homologous GEWL structure. Our method was based on the assumption that the MIR model was positioned roughly correctly in the unit cell but had many substantial errors. Since only the C_α coordinates of GEWL were available to us (Weaver *et al.*, 1985) the strategy used was (i) build a main-chain and C_β model automatically using the database program *CALPHA* (Esnouf & Stuart, 1995); (ii) add side chains in preferred conformations automatically, program *SIDECHAIN* (Esnouf & Stuart, 1995); (iii) refine the structure using *X-PLOR* (Brünger, Kuriyan & Karplus, 1987) in such a way that the fine detail of the input model is preserved until the refinement is well advanced. This essentially automatic procedure yielded a structure with rather few errors.

The re-solved structure has now been further refined using data between 30 and 1.9 Å resolution to an R factor of 16.6% [$R = (\sum ||F_{obs}| - |F_{calc}||)/\sum |F_{obs}|$] on data with $I > 2\sigma(I)$. The model has good geometry, good hydrogen bonding, a sensible distribution of B factors, favourable main-chain dihedral angles and good inter- and intramolecular contacts (Rao, 1989).

Since it is likely that the situation encounted here of having limited structural information from an homologous structure will arise elsewhere, we describe the methods used, which we hope will complement the existing battery of approaches (Brünger, 1991).

Data collection and initial refinement

Crystals of SEWL were grown with orthorhombic form, space group $P2_12_12$ and $a = 91.8$, $b = 65.4$, $c = 38.8$ Å (Isaacs *et al.*, 1985). A set of data to a limit of 1.9 Å resolution was collected by the rotation method (Arndt & Wonacott, 1977). A total of 90 packs of data photographs, each covering 1.5° of rotation, were collected from these crystals. The data from these were merged to give a set of 14477 independent reflections [78% complete at the $I > 2\sigma(I)$ level] with an R_{merge} on intensity of 9.2% [$R_{merge} = (\sum |I - \langle I \rangle|)/\sum |I|$]. Unfortunately, due to the nature of the data-processing protocol all data for which $I < 2\sigma(I)$ were rejected at this stage, thus all statistics quoted below are limited to this subset of stronger data. Initial refinement of the model derived from the MIR map, using data from 6 to 2.8 Å resolution (4512 reflections) resulted (after considerable manual intervention) in a satisfactory R factor (19.1%) and reasonable stereochemistry (root-mean-square deviation in bond lengths was 0.012 Å and in bond angles was 2.6°). In spite of this, the model had serious problems.

(i) The B factor fluctuated wildly between bonded atoms (r.m.s. $\Delta B_{bond} = 19.5$ Å^2 and r.m.s. $\Delta B_{angle} = 20.0$ Å^2).

(ii) There were many missing hydrogen bonds in the regions thought to be helical.

(iii) Several of the non-glycine residues were in unfavourable conformations.

(iv) There were substantial differences between the model and that for GEWL (r.m.s. deviation of 2.49 Å for C_α atoms).

A Ramachandran plot (Ramakrishnan & Ramachandran, 1965) for this structure is shown in Fig. 1(*a*).

In an attempt to correct the model, refinement was continued using *X-PLOR*. Simulated-annealing refinement using data in the resolution range 6–2.8 Å did not improve the model (*R* factor increased to 21.3%). Clearly some radical rebuilding was needed.

Construction of a new starting model

Since SEWL and GEWL have only six sequence differences in their 185 residues, the available GEWL C_α coordinates were used in the following procedure. Firstly, the long helix (residues 109–132) was used to superimpose the GEWL C_α model onto the current model of SEWL (r.m.s. deviation of 0.42 Å for the 23 C_α atoms of the helix). A new backbone and C_β model was created automatically using the rotated and translated GEWL C_α coordinates as a guide. The program *CALPHA* enables a stereochemically satisfactory backbone structure (main-chain and C_β atoms) to be built from a set of protein segments contained in a database derived from some 80 well refined protein structures (Esnouf & Stuart, 1995). This database is essentialy a sub-set of the Protein Data Bank (Bernstein *et al.*, 1977).

The polypeptide main-chain and C_β atoms were then used as a scaffold which the program *SIDECHAIN* decorated with the correct side chains for SEWL. This program simply places each side chain in its most likely conformation (with no reference to the structure of nearby amino acids). The model produced had an r.m.s. deviation from the C_α positions of GEWL of 0.29 Å.

Database-restrained refinement

At this stage, if we assume the GEWL C_α coordinates to be reliable, we could anticipate that the model would contain errors of various sorts.

(i) The position and orientation in the cell are likely to be only approximate, being derived by fitting a single helix.

(ii) The model will contain internal inconsistencies because of the method of construction. The most serious of these are likely to be due to the very simple algorithm for placing side chains.

(iii) There may be incorrect structural prejudices brought into the model from the database.

(iv) There may be errors due to genuine differences between GEWL and SEWL.

Table 1. *Automatic refinement of the database-derived model*

Refinement stage	Final *R* factor	Notes
Initial EM	0.501	300 steps energy minimization, main-chain coordinates restrained to initial positions to relieve side-chain bad contacts (no X-ray restraints included)
Rigid body	0.418	Data 8–4 Å, 30 steps rigid-body refinement, translation 1 Å, rotation 5°
Positional	0.333	Data 6–3 Å, 60 cycles positional refinement
Positional	0.335	Data 6–2.5 Å, 60 cycles positional refinement
Preparatory EM	0.323	Data 6–2.5 Å, 500 steps energy minimization with X-ray terms
Heat simulation	0.315	Data 6–2.5 Å, 0.5 ps restrained MD, timestep 0.5 fs, temperature 1000 K
Cool simulation	0.294	Data 6–2.5 Å, 0.5 ps restrained MD, timestep 0.5 fs, temperature 300 K
Final EM	0.284	Data 6–2.5 Å, 500 steps energy minimization with X-ray terms
B refinement	0.245	Data 6–2.5 Å, 20 steps individual atom *B*-factor refinement
Heat simulation	0.314	Data 6–2.5 Å, 0.5 ps restrained MD with C_α and O atoms restrained to positions after rigid-body refinement, timestep 0.5 fs, temperature 3000 K
Heat simulation	0.275	Data 6–2.5 Å, 0.5 ps restrained MD with C_α and O atoms restrained to positions after rigid-body refinement, timestep 0.5 fs, temperature 1000 K
Cool simulation	0.255	Data 6–2.5 Å, 0.5 ps restrained MD, timestep 0.5 fs, temperature 300 K
Final EM	0.247	Data 6–2.5 Å, 500 steps energy minimization with X-ray terms
B refinement	0.234	Data 6–2.5 Å, 20 steps individual atom *B*-factor refinement

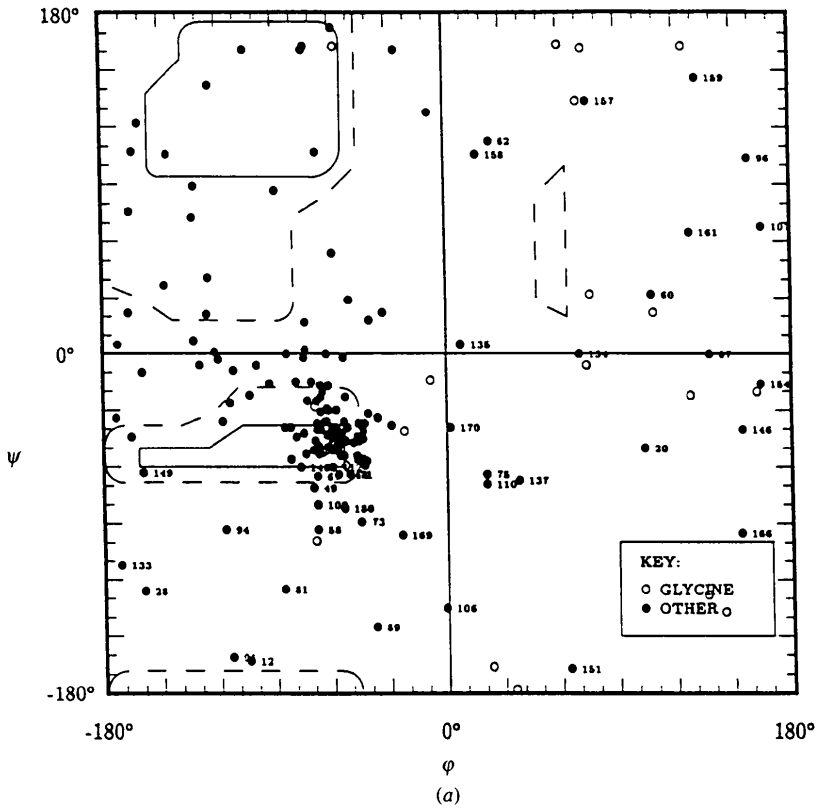
Fortunately, these four sorts of errors are, at least partially, separable and we may use a strategy that allows the mass of correct fine detail derived from the high-resolution results embodied in the database to be retained until we are well on the way to an accurate structure. The aim is to improve the phases derived from the model at these critical stages so that genuine errors in the model can be more readily detected in electron-density maps.

The protocol used various options available in the *X-PLOR* program and is described in outline in Table 1.

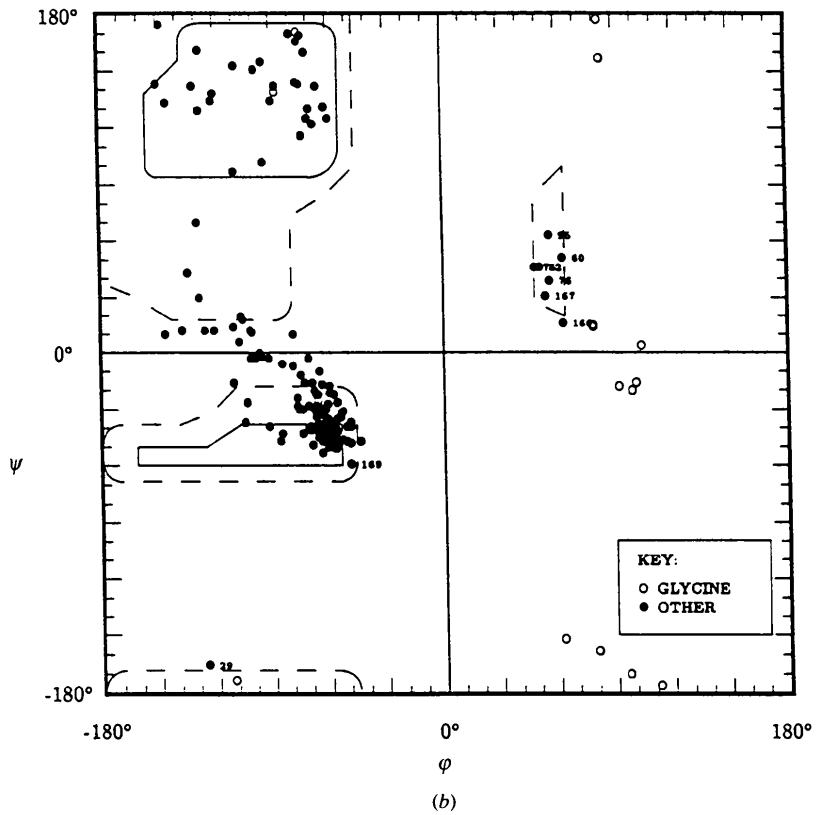
The first stage was to relax the starting model using a standard energy-minimization procedure. Since the primary objective was to relieve bad contacts caused by the incorrect positioning of side chains, the main-chain atoms were harmonically restrained to their starting positions. (The r.m.s. movements of main-chain and side-chain atoms were 0.27 and 1.52 Å, respectively, and for the relaxed model the deviations from ideal bond lengths and angles were 0.013 Å and 2.9°, respectively.)

Next, this stereochemically reasonable model was moved as a rigid body to optimize the fit to the measured data in the range 8–4 Å resolution. The model rotated by 5° and translated by 1 Å reducing the *R* factor by over 8% (to 41.8%).

Energy minimization against the X-ray data was used to correct minor infelicities in the model. Initially, data corresponding to Bragg spacings between 8 and 4 Å were included. Subsequent minimizations were per-



(a)



(b)

Fig. 1. (*a*) Ramachandran plot of the incorrect SEWL structure. (*b*) Ramachandran plot of the automatically refined database structure.

Table 2. *R factor as a function of resolution for the final model*

Resolution range (Å)	No. of reflections	*R* factor	Accum.
3.80–30.00	2199	0.123	0.123
3.02–3.80	2115	0.140	0.131
2.63–3.02	2023	0.167	0.139
2.39–2.63	1886	0.181	0.146
2.22–2.39	1769	0.187	0.151
2.09–2.22	1639	0.209	0.156
1.99–2.09	1410	0.237	0.161
1.90–1.99	961	0.269	0.166

Table 3. *Deviations from ideal stereochemistry for final model*

R.m.s. bond lengths (Å)	0.007	5 deviations ≥ 0.03
R.m.s. bond angles (°)	1.37	9 deviations ≥ 7
R.m.s. dihedral angles (°)	22.0	No deviations ≥ 65
R.m.s. improper dihedral angles (°)	1.16	No deviations ≥ 10
R.m.s. ΔB_{bond} (Å^2)	3.7	
R.m.s. ΔB_{angle} (Å^2)	5.0	

formed using data between 6 and 3 Å resolution and then using data between 6 and 2.5 Å resolution. At this stage the *R* factor stood at 32.3% and the r.m.s. deviations from ideal bond lengths and angles were 0.013 Å and 3.0°, respectively.

In brief, the remaining refinement featured a relatively high temperature (3000 K) X-ray restrained simulated-annealing stage with the C_α and backbone O atoms restrained to their positions after rigid-body refinement, allowing substantial repacking of side chains, and finished with a more 'gentle' simulation (cooling from 1000 to 300 K) followed by positional and individual atom *B*-factor refinements. The finishing model had an *R* factor of 23.4% and the r.m.s. deviations in bond lengths and angles were 0.018 Å and 2.9°, respectively. Although these values were not very different to those for the incorrect model, the evidence from a Ramachandran plot (Fig. 1*b*) and from the deviations in *B* factors between bonded atoms (r.m.s. $\Delta B_{bond} = 5.7$ Å^2, r.m.s. $\Delta B_{angle} = 6.8$ Å^2) showed that the new model was much improved. The r.m.s. deviations between the initial database-derived model and the database-restrained refined model were 0.41 Å for C_α atoms and 1.63 Å for all atoms.

Completion of the refinement

At this stage the model still contained minor errors, these were corrected by manual rebuilding, using *FRODO* (Jones, 1985). The final model had good stereochemistry and agreed well with the full set of X-ray data [*R* factor of 16.6% for data from 30 to 1.9 Å spacings with $I > 2\sigma(I)$, including 122 bound waters and a bulk solvent correction (Brünger, Kuriyan & Karplus, 1987)]. Some statistics for this model are given in Tables 2 and 3. The C_α positions for this final model are in close agreement with those for GEWL. After rigid-body superpositioning the maximum deviation is 1.11 Å at Gly184 (near the C terminus) and the r.m.s. deviation is 0.20 Å. In particular, it might have been expected that the substitution of proline for serine at position 32 would have produced a change in conformation, in fact the difference in C_α position is only 0.33 Å.

Table 4. *R.m.s. differences between four SEWL models* (Å)

Top right for C_α atoms, bottom left for all atoms.

	MIR	DB*	Refined DB†	Final
MIR	–	2.49	2.50	2.49
DB*	3.99	–	0.41	0.35
Refined DB†	3.82	1.63	–	0.30
Final	3.72	1.86	1.52	–

* Model derived from GEWL C_α's by application of *CALPHA* and *SIDECHAIN*.

† DB model subjected to automatic refinement in *X-PLOR*.

Assessment of the database-restrained refinement procedure

We now address the question, how successful was the automatic procedure? Table 4 compares some of models of SEWL. The first point to make is that the original model, derived from the MIR map, shows poor

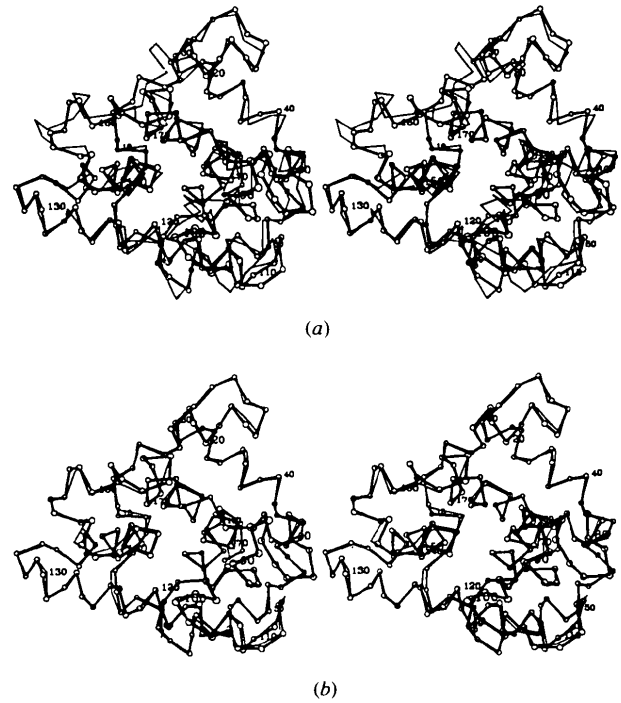


Fig. 2. (*a*) Stereoview of the superposition of C_α traces of the final refined model (thick line) and the incorrect model (thin line) using the program *MOPLOT* (Stuart, unpublished program). (*b*) Stereoview of the superposition of C_α traces of the final refined model (thick line) and the database model (thin line) using the program *MOPLOT* (Stuart, unpublished program).

agreement with the final model (Fig. 2*a*). The models are similar in overall appearance, however there are many places where the alignment of the primary sequence with the electron density is incorrect. Thus, Cys18 was placed at the position of Ser17; residues Ile67 to Ile71 were misplaced where Ala68 to Ser72 should have been; and residues Lys136 to Val154 were sited for Asp137 to Arg155. These numerous and significant errors rendered the model sufficiently incorrect to prevent normal refinement procedures leading to the correct structure.

If we now consider the unrefined database-derived model we find that the r.m.s. deviation of the C_α's from their true position was only 0.35 Å and this dropped to 0.30 Å without manual intervention. The precision of the backbone structure may be seen from Fig. 2(*b*). However, it would be wrong to conclude that the database-derived model was correct in all details. This may be seen from Figs. 3(*a*)–3(*d*). As expected the position and conformation of the side chains is imprecise in the initial database-derived model (where side chains were built in the most commonly observed rotamer). Furthermore, the automatic procedure for correcting these errors did not work completely, 11 residues having an r.m.s. error greater than 2.0 Å in side-chain atom positions.

This prompts further questions.

(i) Was it possible to detect and correct these errors in a simple and reliable way?

(ii) Where did the errors arise from, and in particular are any types of residue prone to problems of this sort?

(iii) Could any simple changes to the procedure improve matters?

We will address these sequentially.

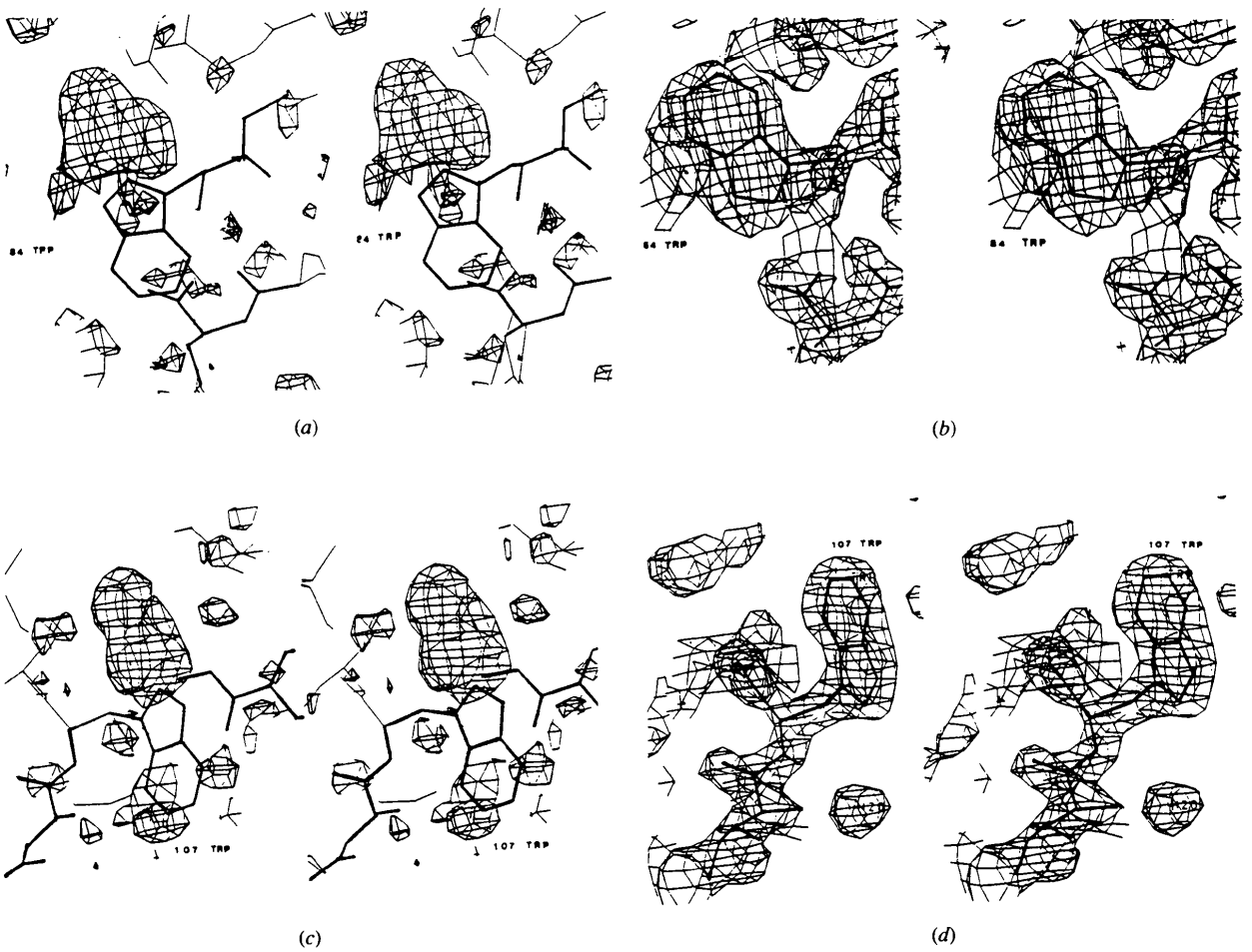


Fig. 3. (*a*) The database-refined model for the region around residue Trp84 with a *FRODO* plot of the corresponding ($F_{obs} - F_{calc}$) electron-density map contoured at 2σ and (*b*) the corrected final model for the same region with a *FRODO* plot of the corresponding ($2F_{obs} - F_{calc}$) electron-density map contoured at 1σ. (*c*) The database-refined model for the region around residue Trp107 with a *FRODO* plot of the corresponding ($F_{obs} - F_{calc}$) electron-density map contoured at 2σ and (*d*) the corrected final model for the same region with a *FRODO* plot of the corresponding ($2F_{obs} - F_{calc}$) electron-density map contoured at 1σ.

(i) *Detection of errors*

Errors in the database-restrained model were detected unambiguously in $|F_{obs}| - |F_{calc}|$ electron-density maps calculated from data between 8.0 and 2.5 Å Bragg spacings. Figs. 3(*a*) and 3(*c*) show maps and the atomic structure in the regions around residues Trp84 and Trp107. The maps are of high quality and the rebuilding was trivial.

(ii) *The nature of the errors*

Table 5 details residues requiring significant manual rebuilding. It excludes surface residues which are poorly defined (11 residues). It will be seen that the three tryptophan side chains are almost alone in having substantial errors. In each case the rings were essentially 'flipped'. It appears that the reason for this is rather subtle. Inspection of the original model (with 'freshly-grown' side chains) shows that, although all three residues showed some error in ring orientation, the side-chain torsion angles were almost correct for two of the residues. However, once the structure emerged from the energy-relaxation stage the damage was done: all three bulky side chains were locked into fundamentally incorrect conformations and even the vigorous 'heating' procedure used could not budge them! Thus, genuine information was lost at the outset in an over-hasty attempt to relieve steric clashes that, given the usual model for van der Waals interactions, were enormously unfavourable.

(iii) *Improvements to protocol*

In the light of (ii) we would eliminate the relaxation stage and use a gentler process which gives sufficient weight to the X-ray observations to 'dampen' the more extreme stereochemical gradients.

Concluding remarks

In conclusion, methods such as those presented here permit rapid and semi-automatic refinement to a reasonably advanced stage for structures where limited structural information (such as C_α coordinates) is available for an homologous molecule. As well as the ease of application of such methods, they are actually more robust than extensive rebuilding (by hand or simulated annealing) in the absence of such restraints.

Table 5. *A list of residues requiring significant manual rebuilding*

Residue	Comments
Trp84	The tryptophan rings needed to be swung 180°. C_ζ of Phe91 prevented this occurring automatically.
Arg87	The main-chain of residues 87 and 88 needed slight manual readjustment.
His101	The ring needed to be rotated by 30°
Trp107	The tryptophan rings needed to be swung 180°. O_δ of Asn108 prevented this occurring automatically.
Trp134	Trp rings were initially about 5 Å away from correct place and were refined back but with a conformation 180° different. It might have been caused by the initially incorrectly placed S_δ of Met160.
Asp168	The side chain needed to be rotated by 180°.

References

ARNDT, U. W. & WONACOTT, A. J. (1977). Editors. *The Rotation Method in Crystallography*. Amsterdam: North-Holland.

BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J. B., MEYER, E. F. JR, BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOUCHI, T. & TASUMI, M. (1977). *J. Mol. Biol.* **112**, 535–542.

BRÜNGER, A. T. (1991). *Curr. Opin. Struct. Biol.* **1**, 1016–1022.

BRÜNGER, A. T., KURIYAN, J. & KARPLUS, M. (1987). *Science*, **235**, 458–460.

ESNOUF, R. M. & STUART, D. I. (1995). In preparation.

GRÜTTER, M. G., WEAVER, L. H. & MATTHEWS, B. W. (1983). *Nature (London)*, **303**, 828–831.

ISAACS, N. W., MACHIN, K. J. & MASAKUNI, M. (1985). *Aust. J. Biol. Sci.* **38**, 13–22.

JONES, T. A. (1985). *Methods in Enzymology*, Vol. 115, edited by H. W. WYCOFF, C. H. W. HIRS & S. N. TIMASHEFF, pp. 157–171. New York: Academic Press.

MORGAN, F. J. & ARNHEIM, N. (1974). *Lysozyme*, edited by E. F. OSSERMAN, R. E. CANFIELD & S. BEYCHOK, pp. 81–87. New York: Academic Press.

RAMAKRISHNAN, C. & RAMACHANDRAN, G. N. (1965). *Biophys. J.* **5**, 909–933.

RAO, Z. (1989). PhD thesis, Melbourne Univ., Australia.

SIMPSON, R. J. & MORGAN, F. J. (1983). *Biochim. Biophys. Acta*, **744**, 349–351.

WEAVER, L. H., GRÜTTER, M. G., REMINGTON, S. J., GRAY, T. M., ISAACS, N. W. & MATTHEWS, B. W. (1985). *J. Mol. Evol.* **21**, 97–111.