Acta Cryst. (1991). A47, 110-119

Improved Methods for Building Protein Models in Electron Density Maps and the Location of Errors in these Models

BY T. A. JONES,* J.-Y. ZOU AND S. W. COWAN

Department of Molecular Biology, BMC, Box 590, S-75124 Uppsala, Sweden

AND M. KJELDGAARD

Department of Chemistry, University of Aarhus, DK-8000, Aarhus, Denmark

(*Received* 1 *June* 1990; *accepted* 7 *September* 1990)

Abstract

Map interpretation remains a critical step in solving the structure of a macromolecule. Errors introduced at this early stage may persist throughout crystallographic refinement and result in an incorrect structure. The normally quoted crystallographic residual is often a poor description for the quality of the model. Strategies and tools are described that help to alleviate this problem. These simplify the model-building process, quantify the goodness of fit of the model on a per-residue basis and locate possible errors in peptide and side-chain conformations.

Introduction

X-ray crystallography is the most powerful tool available to provide detailed three-dimensional structural information of macromolecules, and has led to new insights into how structure determines protein function. The last ten years have seen a technical revolution in a number of vital stages in solving a new protein structure. These include, for example, the use of modern molecular-biology techniques to overexpress proteins that are normally present in minute amounts. Our ability to collect more accurate diffraction data has improved by the development of electronic two-dimensional area detectors and powerful synchrotron-based X-ray sources. Even with rotating-anode generators, area detectors frequently allow the collection of a complete high-resolution data set from a single crystal. These factors directly affect the quality of electron density maps phased by the method of multiple isomorphous replacement (MIR).

Initial models are now routinely improved by crystallographic refinement. A number of least-squares algorithms have been described for this purpose. These include the use in the refinement of model restraints (Hendrickson & Konnert, 1985), constraints and restraints (Sussman, Holbrook, Church & Kim, 1977), explicit molecular-mechanics force fields (Jack & Levitt, 1978), fast Fourier transform methods to speed up the calculations (Agarwal, 1978) and, more recently, force-field-based molecular-dynamics algorithms (Brünger, Kuriyan & Karplus, 1987; Fujinaga, Gros & Van Gunsteren, 1989). Fortunately, the decrease in the price/performance of computers has allowed us more or less to keep up with the increased computational demands of some of these algorithms.

The interpretation of MIR maps to produce an initial molecular model is a critical step that remains problematic. At this stage in the process, errors can be introduced that either cannot be removed by refinement or require many alternating cycles of refinement and manual refitting. Incorrect models can be refined to crystallographic R factors that up to a few years ago would have been considered eminently respectable, especially for large multi-subunit structures.

Three-dimensional computer graphics workstations are now widely used for constructing models in MIR maps. One computer program in particular, *FRODO*, has been widely used (Jones, 1978) and is available on a range of workstations. In an attempt to improve the ability to interpret maps and then to construct more accurate models, Jones & Thirup (1986) introduced the use of skeletons coupled with a protein database of the best refined protein structures to build the initial model. This work suggested that all protein models could be built from fragments of existing structures. In this paper we describe our extensions to these ideas, and our initial attempts at reducing the subjectivity involved in building models. With the overall procedure shown in Fig. 1, we are able to go from an initial Cα trace to a well refined model without manual intervention. This should allow the crystallographer to spend more time considering alternative hypotheses, without worrying about most of the detailed fitting of the model to the electron density map. We are aware that, unfortunately, such a procedure can also be misused, acting as a 'black box' to produce a totally incorrect structure.

* Author to whom correspondence should be addressed.

0108-7673/91/020110-10$03.00

An introduction to *O*

Our ideas are implemented in a new computer graphics program *O*. Information is maintained in a database that can be updated by the user, by the program itself or by other utility programs. Each molecule in the database has a user-defineable name. The usual structural data associated with a molecule are converted into nine vectors of information (Table 1). We refer to some of these vectors as properties that can be associated with residues (*e.g.* the amino acid sequence) or atoms (*e.g.* the temperature factors). Properties can be used for colouring purposes or to select the atoms that are to be displayed. Any number of molecules can be stored in the database, and any number of graphical objects can be created from a molecule.

Map interpretation

Before a complete model can be built it is necessary to understand how the protein main chain folds through the experimentally determined three-dimensional matrix of the electron density map. In particular, it is necessary to decide on the correspondence between the protein sequence and the map. In parallel, one attempts to develop an idea for the overall or a significant part of the fold. During this initial stage, one frequently recognizes features in the density that may correspond to a part of the sequence. One then tries to extend the sequence assignment in either direction until the alignment breaks down. This produces a series of hypotheses that may be contradictory, and requires both an overview and a detailed description of the map. The overview can be produced with a skeleton representation of the density (Greer, 1974) that has been implemented in a computer graphics program (Williams, 1982). The detailed description can be obtained by the usual contoured net representation. These representations have been combined (Jones & Thirup, 1986) and used in our laboratory to solve a number of new protein structures: P2 myelin (Jones, Bergfors, Unge & Sedzik, 1988), two types of rubisco molecules (Schneider, Lindqvist, Brändén & Lorimer, 1986; Andersson *et al.*, 1989), PapD (Holmgren & Brändén, 1989), cellobiohydrolase II (CBHII) (Rouvinen, Bergfors, Teeri, Knowles & Jones, 1990) and ribonucleotide reductase B2 (Nordlund, Sjöberg & Eklund, 1990).

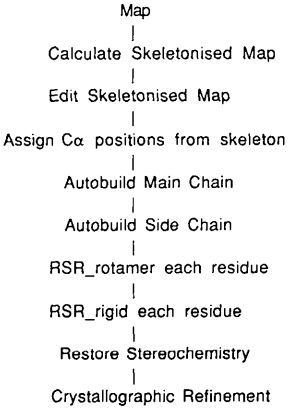


Fig. 1. Strategy overview for model building in an electron density map.

Table 1. *O datablocks used to represent protein models and skeletons*

Protein name A1	
A_ATOM_xyz	Orthogonal atomic coordinates
A1_ATOM_NAMES	Names of atoms, *e.g.* CA
A1_ATOM_B	Temperature factors of atoms
A1_ATOM_WT	Occupancies of atoms
A1_ATOM_Z	Atomic numbers
A1_RESIDUE_NAME	Name of residue, *e.g.* 75
A1_RESIDUE_TYPE	Amino acid name, *e.g.* ALA
A1_RESIDUE_POINTERS	Pointers for the residue to atom data
A1_RESIDUE_CG	Residue centre of gravity and radius
Skeleton name ANO1	
ANO1_ATOM_XYZ	Orthogonal atomic coordinates
ANO1_RESIDUE_NAME	Name of residue, just one
ANO1_RESIDUE_TYPE	Amino acid name, just one
ANO1_RESIDUE_POINTERS	Pointers for the residue to atom data
ANO1_ATOM_BONE	Skeleton status codes
ANO1_CONNECTIVITY	Skeleton connectivity codes

Our new way of working with skeletons differs mainly due to the advantages of using *O*. Each skeleton is treated as a molecule with a number of extra database vectors (Table 1). One is an atomic property ('*_atom_bone*') used to specify the status of each skeleton atom. These codes are mapped to user-defined colours when the skeleton atoms are displayed. We recommend the use of a simple classification: probable main chain, possible main chain and side chain. However, situations may arise where a more complex set of assignments may be needed. For example, when a team of people are trying to interpret a map, it may be useful to highlight changes made by different members of the team.

A second vector ('*_connectivity*') contains a description of how the skeleton atoms are connected. Because of errors in the phases used in the map calculation, the density is rarely continuous from the amino to the carboxy terminus. Editing of the skeleton connectivity is therefore required to produce a continuous trace. This is accomplished by commands that break or make connections between skeleton atoms.

One usually works with at least two objects made from a single skeleton, one showing the proposed main-chain trace within a large volume (*e.g.* a sphere of 30–50 Å), and the other showing all skeleton atoms within a smaller radius (*e.g.* 15–20 Å).

Building the model

Jones & Thirup (1986) demonstrated that, given the Cα coordinates of a protein, one could reconstruct the main chain by linking together peptide fragments of different lengths. These fragments were taken from a library of refined protein structures. They further demonstrated that an initial model could be built by using the electron density skeleton as a framework to locate peptide fragments. This was done interactively by specifying that certain skeleton atoms be used as Cα guide positions, or by allowing the program to place them along the skeleton at suitable distances. The latter method occasionally results in a residue being skipped in a turn.

Since our ultimate aim is to automate model building fully, we now introduce a stage where the position of each Cα atom in the molecule is explicitly defined. At present this is achieved interactively by the user placing a particular Cα at the position of a skeleton atom. Building a model from such a set of guide coordinates is a well known problem to protein crystallographers (Diamond, 1966, 1982; Jones, 1982) and others (Purisma & Scheraga, 1984). Likewise, numerous algorithms could be developed to use a protein structure database to reconstruct the whole protein from the Cα trace. One could, for example, use fixed length fragments, variable length fragments satisfying some cutoff criterion (Jones & Thirup, 1986; Claessens, Van Cutsem, Lasters & Wodak, 1989) or dynamic programming algorithms to locate the minimum number of fragments where each fragment satisfies some r.m.s. cutoff. The resulting model should have, as a minimum requirement, the side chains pointing roughly in the correct direction. To reduce the amount of manual intervention later in the refinement, the main-chain carbonyl O atoms should also be correctly oriented. Our experience suggests that the first qualitative requirement is met when the r.m.s. deviation between Cα guide points and the database fragment is ~1 Å. Correct carbonyl O atoms in general require a r.m.s. fit of ~0·5 Å. In Fig. 2, we show the longest fragments that can be located in our normal database of 32 structures that fit residues in retinol binding protein (RBP) with 0·5 and 1·0 Å cutoffs. This gives average fragment lengths of 7 and 10 residues for the two cutoffs. We could, therefore, build any part of RBP with a fragment of ~10 residues and be sure of getting the side-chain directions roughly correct, but we would require fragments of ~7 residues to get the correct peptide orientations. Since our aim is not to use the minimum number of fragments but to build an accurate structure, fragments of 5 residues are used. This allows a better chance of recognizing low-frequency conformations.

The complete backbone structure is built with a simple extension of our original scheme, outlined in Fig. 3, that we refer to as autobuilding. At residue i in the structure, the best fitting fragment is found that matches the Cα's of $i-2$ to $i+2$. However, only residues $i-1$ to $i+1$ have their coordinates updated from the fragment because the other main-chain atoms of residues $i-2$ and $i+2$ are not fixed by the superposition. The next fragment is chosen by stepping forward 3 residues, comparing $i+1$ and $i+5$, and repeating the process. This algorithm does not build either amino- or carboxy-terminal residues and, therefore, requires an extra residue at each end of the chain. Some deviation from standard bond lengths and angles will occur at the linkage between tripeptides and because of deviations in the structures making up the databank. These deviations are, however, small and can be initially ignored. The fitting algorithms make use of pre-calculated distance matrices to speed up the comparisons (Jones & Thirup, 1986).

To test the quality of models produced by this procedure we have taken the Cα coordinates of CBHII and rebuilt the main chain under various conditions. This structure is a suitable example because it is very well refined (14% R factor to 2 Å resolution), it is a large α/β protein (367 residues) with loops containing extensive non-regular secondary structure. The r.m.s. deviation of the rebuilt model is 0·21 Å for Cα atoms and 0·56 Å for all main-chain

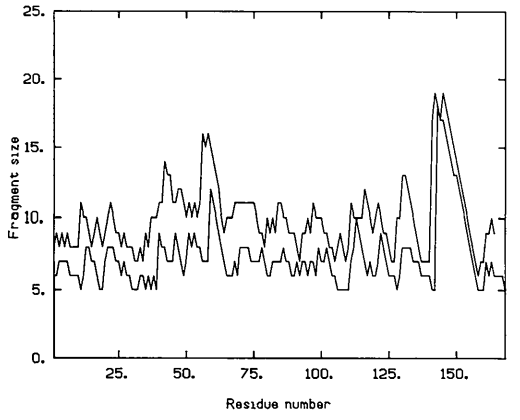


Fig. 2. The longest fragments found in a database of well refined structures that fit retinol binding protein with r.m.s. cutoffs of 0·5 Å (lower curve) and 1·0 Å (upper curve). The long fragments found around residue 150 in the sequence correspond to the start of a four-turn α-helix.

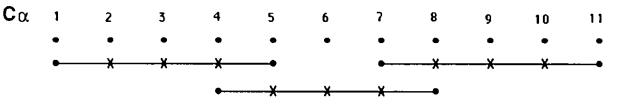


Fig. 3. The main chain of autobuilt structures is produced by identifying the best fitting pentapeptides in the database. These fragments are combined, overlapping by two residues. Only those residues marked by crosses have their coordinates updated by the transformed fragment coordinates.

Table 2. *Autobuilding of CBH* II *from Cα coordinates containing random errors*

The r.m.s. deviations refer to the fit of the autobuilt structure to the correct structure.

Cα error	R.m.s. Cα	R.m.s. main chain	R.m.s. carbonyl O atom
0·0	0·21	0·56	1·04
0·3	0·32	0·62	1·08
1·0	0·80	1·08	1·64
1·5	1·21	1·50	2·06

atoms (N, Cα, Cβ, C, O). Not unexpectedly, the carbonyl O atoms show the largest deviation, 1·04 Å. A similar value for main-chain atoms (0·51 Å) has been obtained by Reid & Thornton (1989) in reconstructing *Clostridium* flavodoxin (Smith, Burnett, Darling & Ludwig, 1977) from Cα coordinates (using the *FRODO* database).

In reality, the guide Cα atoms taken from an experimental map will contain severe errors, partly because of the skeletonizing algorithm and partly because of phase errors in the map. This has been simulated by introducing random errors into the refined coordinates of CBHII and autobuilding from the randomized Cα coordinates. The deviations of the rebuilt Cα model are compared with the correct structure in Fig. 4. Below an introduced r.m.s. error of ~0·35 Å, the autobuilt model shows a deviation from the correct structure that is slightly worse than the introduced error. Above this value, the autobuilt model Cα is a better fit to the correct Cα coordinates. Table 2 shows the deviation of the main-chain atoms and carbonyl O atoms for errors approaching values to be expected in map interpretation.

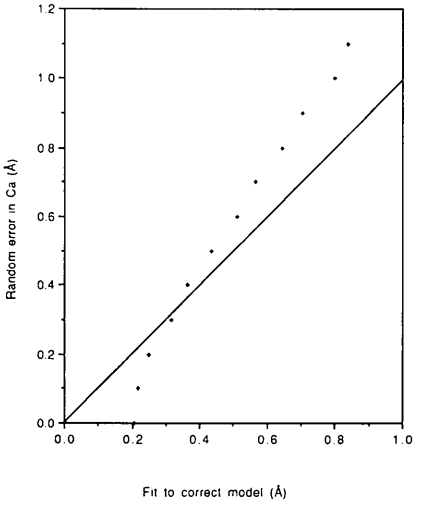


Fig. 4. The effect of random errors in guide coordinates on the autobuilt model. Below the solid line, the autobuilt struture is worse than the guide coordinates, above the line the autobuilt structure is better. The model is cellobiohydrolase II, model *A*27.

The side chains of the best refined structures show a high preference for discrete conformations (James & Sielecki, 1983). These conformations (termed rotamers) have been recently tabulated (Ponder & Richards, 1987; McGregor, Islam & Sternberg, 1987). Not all amino acids have well defined rotamers; in particular, lysine and arginine side chains are poorly modelled. However, for the remaining residues, we consider building anything but rotamers into the initial model to be a mistake. If we choose the rotamer closest to the conformation observed in our refined CBHII model, the r.m.s. fit after autobuilding from the Cα guide coordinates is 1·1 Å for all atoms. Taking the most common rotamer gives an overall fit of 2·5 Å.

A residue *R* factor

When describing the goodness of fit of an initial model, we are often forced to use vague qualitative expressions. This is directly related to the subjective nature of map interpretation. While the crystallographic *R* factor unambiguously shows how well a particular model matches the observed structure factors, it still shows relatively poor discrimination of major errors in a model (Brändén & Jones, 1990). A more useful measure would indicate the position of possible errors in a structure. Wierenga, Kalk & Hol (1987) have published how well their structure fits their map on a per-residue basis. Unaware of their efforts, we have independently developed a more quantitative function that can be used to remove much of the subjectivity of map interpretation. We believe this function will also be useful in localizing serious errors in map interpretation.

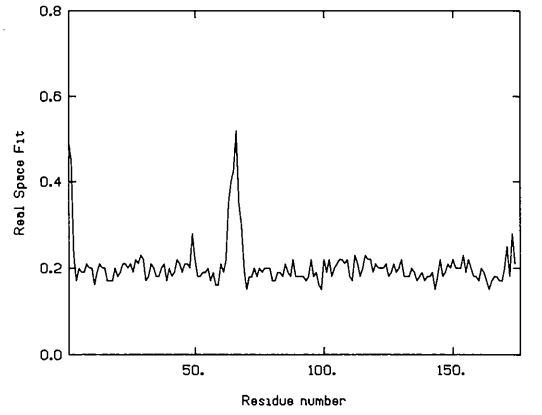
Consider an electron density map on a grid G_1. Given a set of coordinates, a calculated electron density can be built up on an identical grid G_2, by assuming a Gaussian distribution function for each atom (Diamond, 1971; Jones & Liljas, 1984). The atoms are forced to have an overall temperature factor, and the grid densities are scaled together with a single scale factor. For residue *i*, the electron density of a selected group of atoms within this residue is built on a third identical grid G_3. For every non-zero element in G_3 we then calculate the real-space fit for that residue as

$$\sum |\rho_{obs} - \rho_{calc}| / \sum |\rho_{obs} + \rho_{calc}|,$$

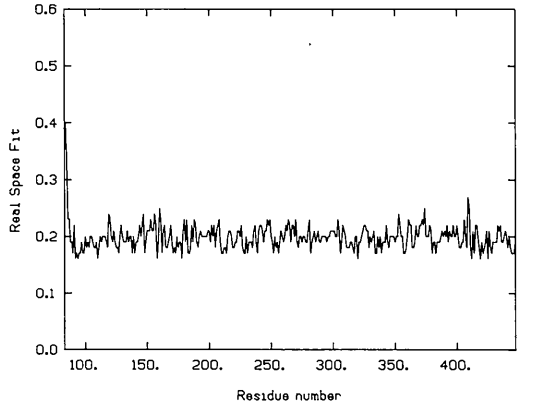
where ρ_{obs} is taken from the equivalent element in G_1 and ρ_{calc} is the equivalent element in G_2. The function may be used to demonstrate the continuity of the main chain by using just the N, Cα, Cβ, C and O atoms in the calculation. Alternatively, by defining side-chain atoms, the function can identify where the protein sequence is out of register with the density. The result of the calculation can be added to the *O* database as a residue property that in turn can be

used for colouring and/or atom selection purposes when displaying the model.

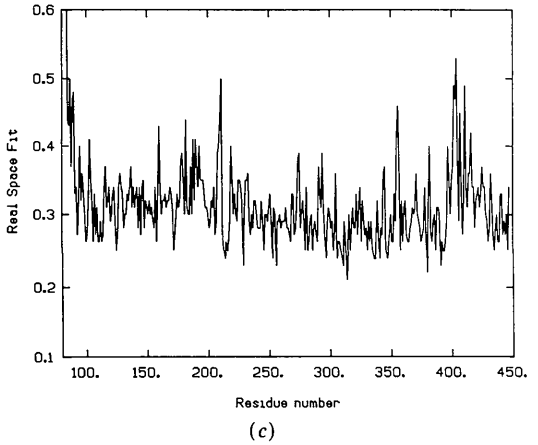
Figs. 5(*a*) and (*b*) show the main-chain fit for two well refined proteins, RBP and CBHII (*R* factors of 18 and 14% at 2 Å resolution, respectively. CBHII has two molecules in the asymmetric unit that we refer to as *A* and *B*). Clearly, all of the CBHII main chain has continuous density but in RBP there is one region with poor main-chain density. Fig. 5(*c*) shows the main-chain fit of a CBHII model obtained during its crystallographic refinement. This model, *M*7, with an *R* factor of 25·8% for all data in the resolution range 8·0–2 Å, was extensively rebuilt. Two regions were identified where the sequence was out of register with the density, and a number of localized poorly fitting areas were found in either the *A* or *B* molecule. Both of the out-of-register regions (the first 20 N-terminal residues and 402–415) can be recognized by the poor real-space residue fit. The remaining spikes mostly correspond to 2–3 residues that could be rebuilt by copying the equivalent atoms from the other chain.



(*a*)



(*b*)



(*c*)

Fig. 5. Plots of main-chain real-space-fit residuals for (*a*) refined retinol binding protein, model *M*112; (*b*) refined cellobiohydrolase II core, model *A*27; the core protein begins at residue 83 in the sequence; (*c*) partly refined cellobiohydrolase II core, model *A*7.

Table 3. *A simulation to demonstrate the use of residue real-space fits to determine out-of-register errors*

The side chains of CBHII model *A*27 were mutated, optimally fitted to the density by hand, and then further refined with *X-PLOR*. The real-space fits (RSF) are made for all atoms in each residue. The two regions were chosen at random.

Region 1							
Residue	229	230	231	232	233	234	235
Correct sequence	Asn	Leu	Gly	Thr	Pro	Lys	Cys
RSF	0·20	0·19	0·19	0·23	0·21	0·22	0·19
Shifted sequence	Asn	Gly	Thr	Pro	Lys	Cys	Cys
RSF	0·22	0·26	0·38	0·29	0·39	0·26	0·20
Region 2							
Residue	140	141	142	143	144	145	
Correct sequence	Asp	Lys	Thr	Pro	Leu	Met	
RSF	0·19	0·23	0·20	0·22	0·28	0·22	
Shifted sequence	Lys	Thr	Pro	Leu	Met	Met	
Fit	0·32	0·30	0·31	0·33	0·32	0·24	

The residue residual could also be used to search directly for out-of-register errors, We have simulated this by deliberately introducing such errors in the *A* chain of our best refined CBHII model, optimizing the fit of the side chain to the density by hand, and then crystallographically refining the structure [50 steps of Powell minimization with *X-PLOR* (Brünger *et al.*, 1987)]. The residue real-space fits for two experiments are shown in Table 3 where the correct alignments are clearly identified. An automatic procedure to search for 1, 2 and 3 residue mismatches would require procedures to optimize the mutated side chain to the density. These tools have been developed and will be described in a separate publication.

In the above formulation, the G_3 grid acts as an envelope within which to make the grid sum calculations. Other ways of forming the envelope may be better suited for certain applications. For example, when searching for out-of-register errors, the current method of defining the envelope gives poor discrimi-

Table 4. *Autobuild model statistics*

RSF_{mc} and RSF_{all} are the average residue real-space-fit factors calculated for main chain and all of the residue, respectively. R refers to the normal crystallographic R factor and is calculated for all measurements in the resolution range 7·5–2·7 Å. The model numbering is explained in Fig. 6. The r.m.s. deviations are with respect to *M*9, our best refined model. The r.m.s. deviations in brackets are calculated where arginine and lysine residues are treated as alanines.

Model	RSF_{mc}	RSF_{all}	R	R.m.s. Cα	R.m.s. O	R.m.s. Main	R.m.s. Side	R.m.s. All
*M*2	0·34	0·37	0·420	0·77	1·31	0·95	1·90	1·50
*M*3	0·34	0·37	0·225	0·51	0·95	0·67	1·69	1·28
*M*9	0·34	0·38	0·157	—	—	—	—	—
*M*100	0·42	0·47	0·516	1·05	1·70	1·32	3·32	2·49
*M*101	0·37	0·42	0·259	0·87	1·41	1·08	2·95	2·19
							(2·28)	(1·68)
*M*110	0·42	0·46	0·506	1·05	1·70	1·32	2·77	2·13
*M*111	0·36	0·40	0·241	0·73	1·24	0·92	2·38	1·78
*M*120	0·35	0·37						
*M*122	0·35	0·37	0·465	1·00	1·83	1·32	3·02	2·30
*M*124	0·36	0·39	0·246	0·96	1·38	1·13	2·53	1·93
							(1·99)	(1·53)
*M*130	0·42	0·46	0·519					
*M*131	0·44	0·49	0·275					

nation when a small side chain fits in the density meant for a longer side chain. Other envelopes can be made to improve this particular situation. Also, other grid sum calculations could be employed (such as correlation coefficients) as alternative indices of fit.

The careful monitoring of temperature factors, especially for refinements carried out at high resolution, has been widely used to monitor coordinate errors. Indeed, the large peak in the real-space fit function of RBP in Fig. 5(*a*) corresponds to a region of high temperature factors. The main advantage of using the real-space approach, therefore, is its applicability at any stage of the modelling procedure, including the initial construction of a model. It can also be used for studies at lower resolution where the temperature factors may be poorly defined. We are not aware of any disadvantage of using the real-space approach instead of studying temperature factors.

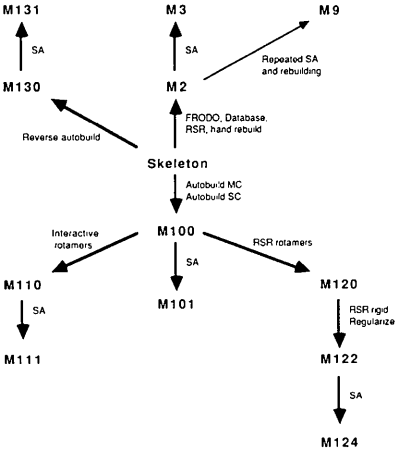


Fig. 6. Experiments on building models in the MIR map of P2 myelin. SA refers to crystallographic refinement using simulated annealing with the program *X-PLOR* and the protocol shown in Table 5.

Table 5. X-PLOR *protocol used for all model refinements*

Values as recommended in the X-PLOR manual used unless stated otherwise below.

Resolution: 7·5–2·7 Å
  Overall temperature factor: 20·0 $Å^2$
  Nbonds, dielectric constant, $\varepsilon = 4{\cdot}0$
  No charges on Lys, Glu, Asp and Arg side chains
Check stage: (to obtain the weighting terms W_A and W_P)
  40 steps Powell minimization, harmonic repulsive term
  40 steps Powell minimization, normal Van der Waals terms
  100 cycles molecular dynamics at 300 K
Preparation stage: (conjugate gradient minimization)
  Harmonic restraints on the Cα's of 83·7 kJ $mole^{-1}$ $Å^{-2}$
  40 steps Powell minimization, harmonic repulsive term
  160 steps Powell minimization, normal Van der Waals terms
Slow-cool stage:
  50 steps of molecular dynamics at each temperature starting from 5000 K and dropping 25 K until reaching 300 K
  Time step = 0·0005 ps
Final stage: (conjugate gradient minimization)
  100 steps of Powell minimization without phase restraints
Time scale:
  On a Stellar GS1000, on average 60 h of CPU.

Automatic refinement

We can now ask ourselves if it is possible to produce a completely refined structure from just a Cα trace. For a test case we have chosen a smaller crystallographic problem, *P*2 myelin protein (Jones *et al.*, 1988) and have carried out a number of experiments that have been summarized in Fig. 6 and Table 4. Models have been crystallographically refined using the same simulated annealing protocol, Table 5, and the program *X-PLOR*.

In the original study, the structure was solved from a map phased with two derivatives, using anomalous dispersion: the derivatives have identical sites. The crystal contains three molecules in the asymmetric unit, each molecule consisting of 131 amino acids. Native diffraction data have been collected to 2·7 Å resolution. Originally, a model of one chain was built

with *FRODO* from a skeleton using fragments from the database. This model was then used to build the other two chains and the three molecules refined to the density using the method described by Jones & Liljas (1984). The results were checked at the display and manually refitted where necessary. This model, referred to as *M*2, has an *R* factor of 42% (without fatty acid ligand) and has no out-of-register errors. Our current best model, *M*9, was obtained after numerous alternating cycles of crystallographic refinement by simulated annealing and manual rebuilding. Model *M*9 has an *R* factor of 15·7% for all data in the resolution range 7·5–2·7 Å. The average residue fits (both main chain and all atom) to the MIR map are approximately the same for both models. For comparison purposes, we have repeated the refinement of *M*2 using the protocol of Table 5, to give *M*3.

In our first experiment, the same edited skeleton originally used to build *M*2 was used to make an initial Cα trace. Atoms were placed where we had left a side-chain branch point in the skeleton. A complete molecule was autobuilt, using the most frequent rotamer for each side chain. The other two chains were built from this model by applying the known non-crystallographic operators to give model *M*100. This model was crystallographically refined to give *M*101. As judged by the *R* factor, this refined model is good and the *R* factor shows a significant drop from 52 to 26%. The r.m.s. fit of the models to *M*9, however, is not so impressive. The main-chain atoms show an improvement from 1·32 to 1·09 Å but the side-chain atoms still have a high r.m.s. deviation, 2·95 Å. Most of these side-chain errors are localized in *M*101 to a few residues, in particular the longest amino acids. This is serious for P2 myelin because this protein is particularly rich in arginine and lysine residues; 25 out of 131 residues. *M*101 shows a worse average fit to the MIR map than does model *M*2, despite a lower *R* factor.

Real-space refinement of a model to a map should, in theory, improve the goodness of fit (Diamond, 1971). However, with a rough initial model, great care must be taken since volume fitting algorithms have a large radius of convergence. This can easily result in large side chains (such as phenylalanine rings) moving into main-chain density. For a moderately well fitting model, the situation can be alleviated by refining into a residual map (Jones & Liljas, 1984). This is calculated by subtracting the scaled density built up using the current model from the experimental map. When an atom or a group of atoms is to be refined, its model density is first added back to the residual map.

In the strategy outlined in Fig. 1, after the main-chain autobuild, the main-chain atoms should have a reasonable fit to the density. Therefore, only these atoms should be subtracted from the experimental map. For each residue, keeping the Cα fixed and allowing a rotational search of the whole residue, we can then find the rotamer that best fits the density. All residues in this model (*M*120 in our experiment) should now approximately fit the electron density. In the next step, therefore, all atoms can be subtracted from the density to form the residual map. For each residue, we then carry out a rigid-body rotation and translation search to find the best fit to the density. This model will have relatively poor stereochemistry at the peptide linkage and should be regularized (models *M*121 and *M*122).

As judged by the residue fits, *M*122 agrees with the MIR map as well as *M*2 and *M*9. The model after refinement (*M*124) is better than *M*101 judged by both *R* factor and r.m.s. fit to *M*9. However, there are a number of obvious errors that can be recognized by inspecting those residues having the poorest fits (Fig. 7). The two worst-fitting residues in the *A* chain, Leu 86 and Met 119, are clearly wrong when viewed at the display. A model with a final lower *R* factor can be produced by interactively deciding on the choice of rotamer in the initial model. In this model, *M*110, there is a poor initial fit to the MIR map since no real-space refinement (or manual fitting) was carried out. The refined model, *M*111, shows improved fit and lower *R* factor.

Fig. 8 shows a histogram of how equivalent atoms in *M*9 and *M*124 are spatially separated. 85% of the atoms are within 1·5 Å of one another. The average residue r.m.s. deviations are plotted in Fig. 9 according to amino acid type. Not surprisingly, the worst errors occur for arginine and lysine residues, *i.e.* those residues having the worst rotamers. The behaviour of serine residues appears surprising but this is due to one residue, Ser 1, at the amino terminus (deviations of 6·6, 3·7 and 6·9 Å for each chain). Likewise, the asparagine value is influenced by errors in Asn 2 and Asn 77. These errors are also likely to be overestimates

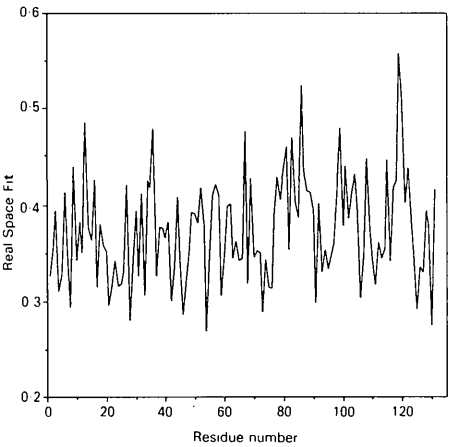


Fig. 7. The all-atom residue real-space fit of the *A* chain of P2 myelin model *M*122 to the MIR map.

because only the coordinates of the *A* chain were fitted to the density. When comparing *M*124 to model *M*9, there are significantly more large deviations observed for atoms in the *B* and *C* chains. The r.m.s. differences between the three chains in the final refined model is ~1·2 Å for all atoms and ~0·6 Å for Cα atoms.

To illustrate the inadequacy of the normal crystallographic *R* factor, we have built a completely backwards structure of P2 myelin. In this model residue *i* is built at residue 132 − *i*. The autobuilt structure, *M*130, refines to an *R* factor of 27·5%, model *M*131. The real-space-fit values start badly and do not improve upon refinement. A plot of the main-chain torsion angles does not clearly distinguish between models *M*124 and *M*131. Similarly, the r.m.s. deviations of bond lengths, angles and fixed dihedral angles have normal values.

Realistically, we cannot expect to autobuild a model better than *M*2, which was the result of many hours careful modelling. However, we had hoped that after simulated annealing the best autobuilt models would be as good as those obtained starting from *M*2. As a control, therefore, *M*2 has been refined with the same protocol. This model, *M*3, is the best model we have produced, as judged by *R* factor, r.m.s. fit and density fit. We are aware of and believe we can overcome the problems associated with modelling Lys/Arg residues. If we omit them from the comparison, the r.m.s. differences of *M*124 and *M*3 to *M*9 then differ by only 0·25 Å.

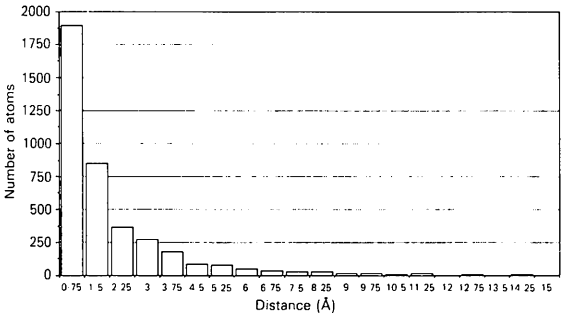


Fig. 8. Histogram of separations of equivalent atoms in models *M*124 and *M*9 of P2 myelin. The atoms in the three chains are included.

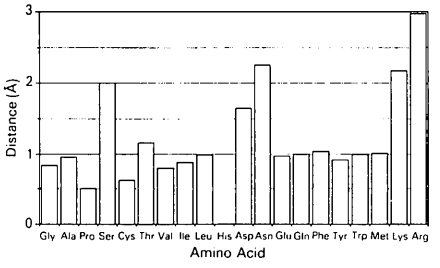
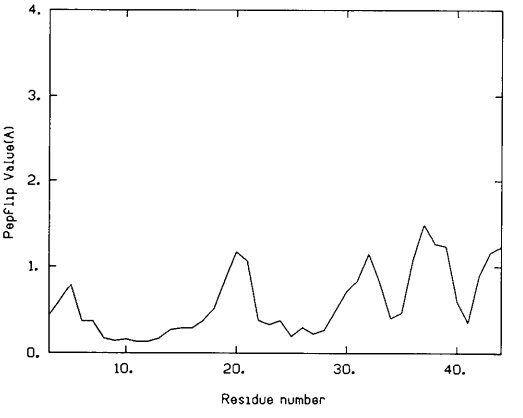


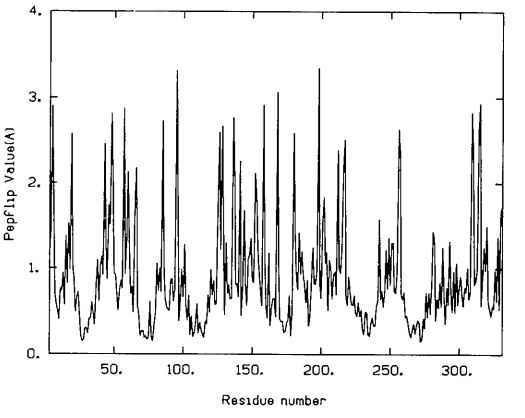
Fig. 9. R.m.s. deviations of *M*124 and *M*9 of P2 myelin according to amino acid type. The protein has no histidine residues.

Locating errors in the model during refinement

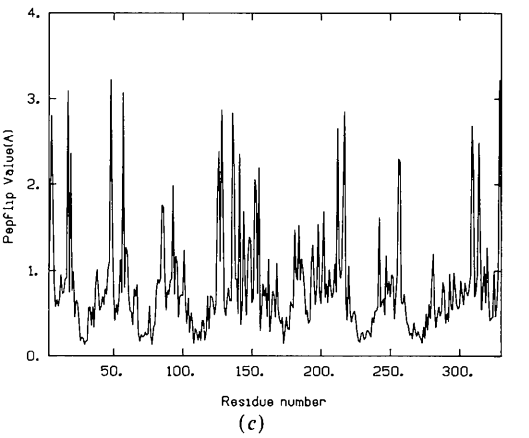
Databases can also be used to monitor the quality of structures undergoing crystallographic refinement. In



(*a*)



(*b*)



(*c*)

Fig. 10. Carbonyl O-atom error indications (Pepflip) for (*a*) a well refined high-resolution structure, crambin; (*b*) and (*c*) models 39 and 56 of the cytochrome of the reaction centre.

this publication we are concerned with errors in peptide orientation and side-chain conformation.

To monitor peptide errors, each pentapeptide in the structure is compared with the database. At position i in the sequence, therefore, we locate the best-fitting fragments to the zone $i-2:i+2$. The r.m.s. deviation of the carbonyl O atom of residue i to the equivalent O atoms of these database fragments is then used as an index of fit. The distribution for a well refined structure crambin (Hendrickson & Teeter, 1981) is shown in Fig. 10(a). The small peaks in this function occur at loops connecting secondary structure elements and are the result of some fanning in the orientation of the peptide planes.

Fig. 10(b) shows the result obtained from a partly refined structure, model 39, of the cytochrome of the reaction centre from *Rhodopseudomonas viridis* (Deisenhofer & Michel, 1989). This plot is more representative of the results obtained while monitoring a refinement. The very sharp peaks having r.m.s. deviations >3 Å correspond to peptides where the O atom points in the opposite direction from the database structures, Fig. 11. We have compared the suggestions made from this calculation with the independent actions taken by Deisenhofer & Michel during their refinement. Table 6 shows a good correlation between the suggested errors and the actions taken in producing model 56 in their refinement.

Our experience with this method suggests that every residue showing a deviation >2·5 Å is worth investigating. The method gives false peaks for structures with *cis*-peptides because there are too few such structures in the database. It also highlights conformations that have a moderate but not absolute requirement for a glycine at the next residue in the sequence. Such conformations frequently have carbonyl O atoms orientated in one direction for the glycine and in the opposite direction for non-glycine residues. Thus, model 56 of the cytochrome also shows spikes that persist to their final model.

The side-chain conformations can be monitored to find the r.m.s. deviation to each possible rotamer for the residue. The lowest value is taken as the index of fit. High values may correspond to errors in the structure.

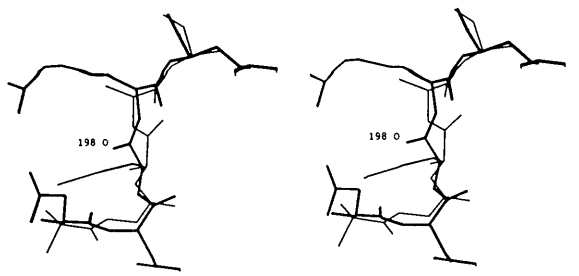


Fig. 11. The worst carbonyl of cytochrome model 39 (thick lines, including side-chain atoms) with the database structure overlaid (thin lines, main-chain atoms). The carbonyl O atom of residue 198 points in the opposite direction compared to the database structure.

Table 6. *Sorted list of proposed peptide flips for reaction centre cytochrome model* 39

The column D_{39-56} gives the distance separating the carbonyl O atoms in models 39 and 56. The action column states the action carried out by Deisenhofer during his refinement.

Residue	R.m.s. O	D_{39-56}	Action
198	3·35	3·64	Peptide flip
95	3·32	3·28	Peptide flip
168	3·07	2·74	Peptide flip
315	2·94	2·79	Peptide flip
158	2·92	3·23	Peptide flip
5	2·91	0·12	No action but polyproline region
57	2·88	0·28	No action
309	2·83	0·26	No action
48	2·82	0·39	No action but 43–47 was a region of many errors
136	2·77	0·20	No action but 137 peptide flip
85	2·73	1·88	Half peptide flip (90°)
128	2·67	0·18	No action
314	2·66	0·36	No action but 315 peptide flip
256	2·64	0·40	No action
126	2·60	0·39	No action
180	2·59	2·57	Peptide flip
19	2·58	0·41	No action but 17 peptide flip
217	2·51	1·25	Half peptide flip

As implemented, both algorithms calculate residue properties that can be used with *O* for colouring and selection purposes. The implementation is noteworthy in another respect. The programs generate files of *O* commands that can be activated, one by one, to place the user at the trouble spot, activating the necessary commands to illustrate the problem. The user then simply agrees that there is an error and corrects it, or moves on to the next problem. We intend to develop this idea further to other problems so that the user is presented with a suggestion, the reasons and then has to decide.

This work has been supported by Uppsala University and the Swedish Natural Science Research Council. MK was supported by the Bioregulation Centre at Aarhus University. We thank Dr Johan Deisenhofer for making his coordinates available to us and Dr Rik Weiranga for pointing out to us his independent use of real-space fitting.

References

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

ANDERSSON, I., KNIGHT, S., SCHNEIDER, G., LINDQVIST, Y., LINDQVIST, T., BRÄNDÉN, C.-I. & LORIMER, G. H. (1989). *Nature* (*London*), **337**, 229–234.

BRÄNDÉN, C.-I. & JONES, T. A. (1990). *Nature* (*London*), **343**, 687–689.

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

CLAESSENS, M., VAN CUTSEM, E., LASTERS, I. & WODAK, S. (1989). *Protein Eng.* **5**, 335–345.

DEISENHOFER, J. D. & MICHEL, H. (1989). *EMBO J.* **8**, 2149-2169.
DIAMOND, R. (1966). *Acta Cryst.* **21**, 253-266.
DIAMOND, R. (1971). *Acta Cryst.* A**27**, 436-452.
DIAMOND, R. (1982). In *Computational Crystallography*, edited by D. SAYRE, pp. 318-325. Oxford: Clarendon Press.
FUJINAGA, M., GROS, P. & VAN GUNSTEREN, W. F. (1989). *J. Appl. Cryst.* **22**, 1-8.
GREER, J. (1974). *J. Mol. Biol.* **82**, 279-302.
HENDRICKSON, W. A. & KONNERT, J. (1985). *Methods Enzymol.* **115**, 252-270.
HENDRICKSON, W. A. & TEETER, M. (1981). *Nature (London)*, **290**, 107-113.
HOLMGREN, A. & BRÄNDÉN, C.-I. (1989). *Nature (London)*, **342**, 248-251.
JACK, A. & LEVITT, M. (1978). *Acta Cryst.* A**34**, 931-935.
JAMES, M. N. G. & SIELECKI, A. R. (1983). *J. Mol. Biol.* **163**, 299-361.
JONES, T. A. (1978). *J. Appl. Cryst.* **11**, 268-272.
JONES, T. A. (1982). In *Computational Crystallography*, edited by D. SAYRE, pp. 303-317. Oxford: Clarendon Press.
JONES, T. A., BERGFORS, T., UNGE, T. & SEDZIK, J. (1988). *EMBO J.* **7**, 1597-1604.
JONES, T. A. & LILJAS, L. (1984). *Acta Cryst.* A**40**, 50-57.
JONES, T. A. & THIRUP, S. (1986). *EMBO J.* **5**, 819-822.
MCGREGOR, M. J., ISLAM, S. A. & STERNBERG, M. J. E. (1987). *J. Mol. Biol.* **198**, 295-310.
NORDLUND, P., SJÖBERG, B. M. & EKLUND, H. (1990). *Nature (London)*, **345**, 593-598.
PONDER, J. W. & RICHARDS, F. M. (1987). *J. Mol. Biol.* **193**, 775-791.
PURISIMA, E. O. & SCHERAGA, H. A. (1984). *Biopolymers*, **23**, 1207-1224.
REID, L. S. & THORNTON, J. M. (1989). *Proteins: Struct. Function Genet.* **5**, 170-182.
ROUVINEN, J., BERGFORS, T., TEERI, T., KNOWLES, J. & JONES, T. A. (1990). *Science*, **249**, 380-386.
SCHNEIDER, G., LINDQVIST, Y., BRÄNDÉN, C.-I. & LORIMER, G. H. (1986). *EMBO J.* **5**, 3409-3415.
SMITH, W. W., BURNETT, R. M., DARLING, G. D. & LUDWIG, M. L. (1977). *J. Mol. Biol.* **117**, 195-226.
SUSSMAN, J. L., HOLBROOK, S. R., CHURCH, G. M. & KIM, S. H. (1977). *Acta Cryst.* A**23**, 800-804.
WIERENGA, R. K., KALK, K. H. & HOL, W. G. J. (1987). *J. Mol. Biol.* **198**, 109-121.
WILLIAMS, T. V. (1982). Thesis, Univ. of North Carolina at Chapel Hill, USA.

Acta Cryst. (1991). A**47**, 119-127

Near-Coincidence Orientations in Hexagonal Materials: from a Unified Twin Approach to a Quasiperiodic Description

BY SERGE HAGÈGE*

Centre National de la Recherche Scientifique, Centre d'Etudes de Chimie Métallurgique, 15, Rue Georges Urbain, 94407 Vitry-sur-Seine CEDEX, France

(*Received* 3 *August* 1990; *accepted* 24 *September* 1990)

Abstract

In materials belonging to the hexagonal crystal family (hexagonal or trigonal crystal systems), for which the irrationality arises primarily from the lattice parameters, the concept of near-coincidence orientation has to be introduced in order to characterize experimental grain boundaries. The practical use of this concept can be simplified if a twin approach is introduced: high-Σ specific coincidence orientations are described as a deviation from very low-Σ twin orientations defined among a unique set of limiting Σ. Consequently, for real hexagonal or trigonal materials, each orientation relationship defined by a quaternion (m, u, v, w), all relatively prime integers, can be described, for any c/a, uniquely by a quasiperiodic arrangement of elementary 'twin' coincidences. Experimental cases of interfaces in hexagonal and rhombohedral crystals (h.c.p. metals, tungsten carbide, alumina) are analysed.

Introduction

In the past few years great interest has been dedicated to the study of grain boundaries in materials described in the hexagonal crystal family. Both theoretical and experimental results presented have outlined an emerging field of research where, for instance, mathematical calculation of coincidence orientations [Bleris, Nouet, Hagège & Delavignette (1982), Grimmer & Warrington (1987), Hagège & Nouet (1989) for hexagonal; Doni, Fanides & Bleris (1986), Grimmer (1989*a*) for rhombohedral], relaxation of the structure at the interface (Serra, Bacon & Pond, 1988; Hagège, Mori & Ishida, 1990), grain-boundary dislocation analysis (Antonopoulos, Karakostas, Komninou & Delavignette, 1988; Chen & King, 1988;

* Also at Ecole Nationale Supérieure de Chimie de Paris, Laboratoire de Métallurgie Structurale, 11 Rue P. et M. Curie, 75231 Paris, France.

0108-7673/91/020119-09$03.00