[10] Electron-Density Map Interpretation

By T. A. JONES and M. KJELDGAARD

Introduction

The life of a macromolecular crystallographer is sometimes exciting. We would list the high points of a crystallographic project as follows:

Material → crystals → phases → model → publication

Each step is crucial to the successful completion of the project and most certainly worth celebrating in some way, as it is likely to be followed by a period in which little progress is made. Each step can be rate determining. Fortunately, significant progress has been made in arriving at each step, except perhaps for the last. Interpreting a new electron-density map as a detailed molecular model is perhaps the most exciting step in a crystallographic project. Seeing the culmination of the efforts of many people appearing before one's eyes is a moment that is rarely forgotten. For a short time, you may be the only person in the world who knows what the structure actually looks like, or at least you think that you know.

Any errors that occur in a crystallographic project usually will be found before publication. Often, if an error has been made, the project will stall and there will be no publication. Introducing a serious error in a model can be different. In this chapter we discuss the kinds of error that might be made (with examples taken from our own work) and why these errors are made. We discuss some of the features of the crystallographic model-

Copyright © 1997 by Academic Press
All rights of reproduction in any form reserved.
METHODS IN ENZYMOLOGY, VOL. 277 0076-6879/97 $25.00

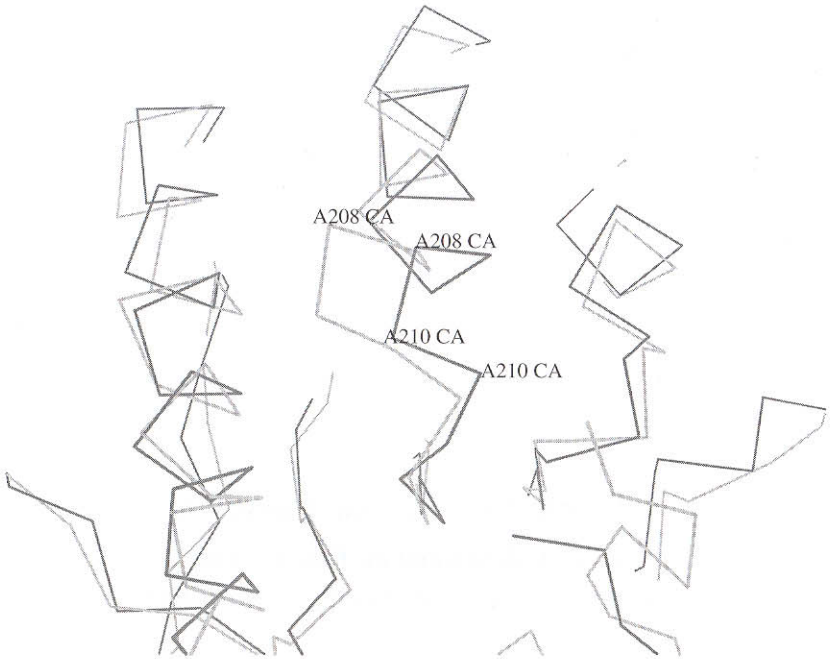

FIG. 1. Locally wrong structure in CBHII.

building program O[1] and then present an outline of how we believe maps should be interpreted.

Different Kinds of Model Error

Our primary concern is not with simple deviations from expected stereochemistry, and inconsistencies with database definitions. Real errors in models occur with frequencies that are, fortunately, inversely proportional to the seriousness of the error. They can include the following.

Totally wrong fold: If proper steps are taken, it is difficult to do and requires some determination. As yet, we do not have our own example of this kind of error. In an experiment in which we deliberately traced the chain backward through the map (i.e., we placed the correct N-terminal residue of the sequence at the C terminus of the structure and then followed the fold to the bitter end, placing the correct C-terminal residue at the N terminus), we could phrase a description of the usual quality criteria of the

[1] T. A. Jones, J. Y. Zou, S. W. Cowan, and M. Kjeldgaard, *Acta Crystallogr.* **A47,** 110 (1991).

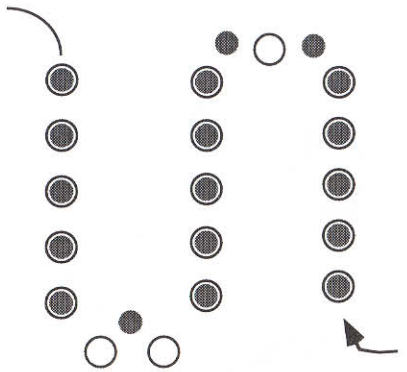

FIG. 2. Schematic out-of-register error.

structure so as not to cause alarm to a potential referee or reader.[2] This model had an R factor of 21.4% after refinement at 3-Å resolution. The free R factor[3] could not, however, be fooled, and had a value of 61.7%.

Locally wrong fold: In a multisubunit structure, for example, the folding of one domain could be totally incorrect. As another example, the connectivity between a subset of secondary structure elements may be wrong. If a large part of the structure is essentially correct, the situation may be hard to correct if the experimental map is ignored after the first model has been built. Provided that a sensible refinement strategy is used, it should be straightforward to recognize that a problem exists. Solving the problem may not be so easy.

Locally wrong structure: The main chain could, for example, be built through strong side-chain density or an unexpected metal ligand, resulting in a small region of structure that is totally incorrect. An example of this kind of error is shown in Fig. 1 and is taken from our work on cellobiohydrolase II (CBHII).[4] The lightly drawn C_α trace is the very first model built into the multiple isomorphous replacement (MIR) map, phased to 2.7 Å. The darker line is a trace of the structure after refinement to 1.8-Å resolution. There was a main-chain density break at the C terminus of this helix that prompted us to make a detour through strong density for the local side chains. This error was picked up and corrected in the third macrocycle of refinement and rebuilding.

Out-of-register errors: Although the local fold is correct, the placement of the sequence into the density may be out of register by a number of

[2] G. J. Kleywegt and T. A. Jones, *Structure* **3,** 535 (1995).
[3] A. T. Brünger, *Nature* (*London*) **355,** 472 (1992).
[4] J. Rouvinen, T. Bergfors, T. Teeri, J. K. C. Knowles, and T. A. Jones, *Science* **249,** 380 (1990).

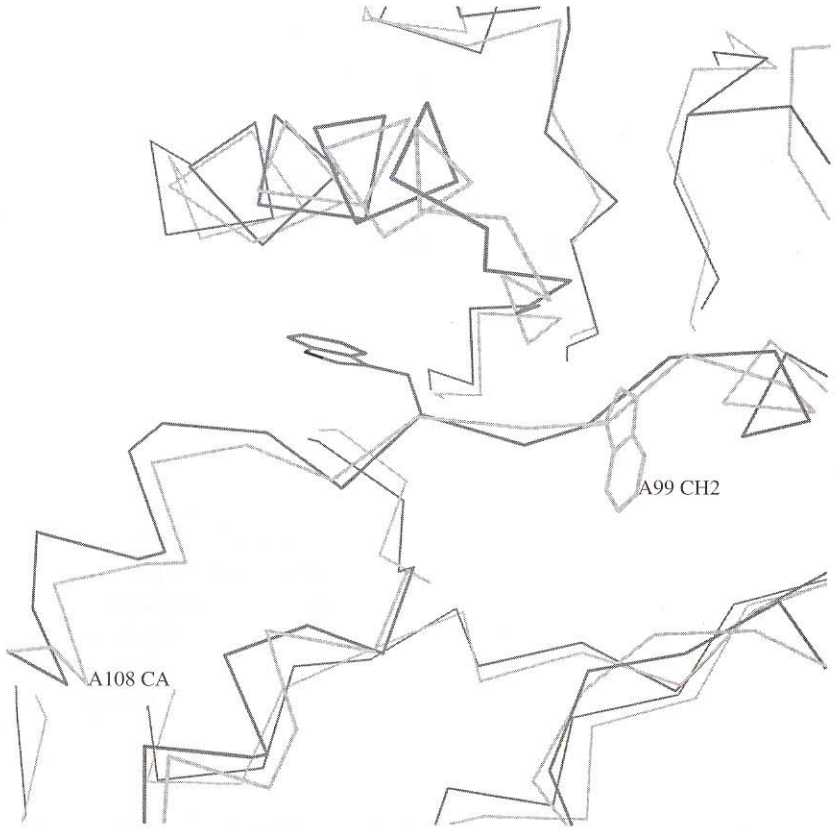

FIG. 3. Out-of-register error in CBHII.

residues. If two placement errors are made, the problem can become self-correcting so that only a zone of connected residues is wrong. Placement errors often occur in loop regions connecting secondary structural elements, as schematically illustrated in Fig. 2. Figure 3 shows a two-residue out-of-register error that occurred in the CBHII study. The first model was massaged to place the indole ring of residue 99 into very strong density. In the correct structure (darker line, Fig. 3), this density is occupied by an unexpected glycosylation of Thr-97. Forcing the model into the wrong density led to a distortion at the N terminus of the next helix, but the sequences came back into register in the middle of this helix at residue 108. Again, the error was corrected in the third macrocycle of refinement.

Wrong side-chain conformation: Wrong side-chain conformation errors are very frequent in the first model. The situation is exacerbated by changing

FIG. 4. Wrong side-chain conformation CBHII.

side-chain torsion angles in the model-building program instead of using rotamers. Figure 4 comes from the CBHII study and shows two valine residues on neighboring strands. The first model was traced with O, but because the program did not have full functionality at the time, the detailed side-chain placements were made with Frodo.[5] Both residues were built in energetically unfavorable conformations whereas in the final refined structure they both adopt the most frequent rotamer conformation. Many incorrect side-chain conformations will be corrected automatically by the refinement program, and those that are not usually can be identified during careful rebuilding.

[5] T. A. Jones, *J. Appl. Crystallogr.* **11,** 268 (1978).

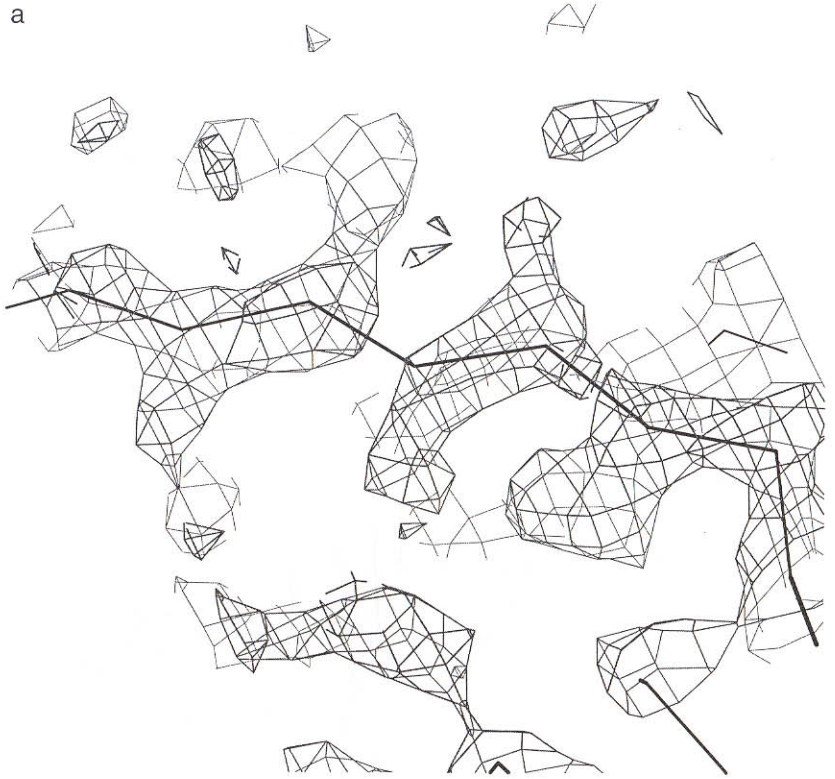

FIG. 5. (a) Experiment MIRAS map of P2 myelin protein. (b) Same region after cyclic threefold averaging.

Wrong main-chain conformation: The placement of the peptide linkage is a critical step in correctly defining the main-chain conformation. If the plane of the linkage is correctly oriented, the coordinate error of the carbonyl oxygen atom can still be more than 3 Å.

Most of the errors described above can be located and corrected, provided a careful refinement and rebuilding protocol is used (see [11] in this volume[6]).

Why There Are Errors in Models

The main causes and reasons why errors are made during map interpretation are noted in this section. Often a combination of events may lead to the error.

[6] G. J. Kleywegt and T. A. Jones, *Methods Enzymol.* **277,** [11], 1997 (this volume).

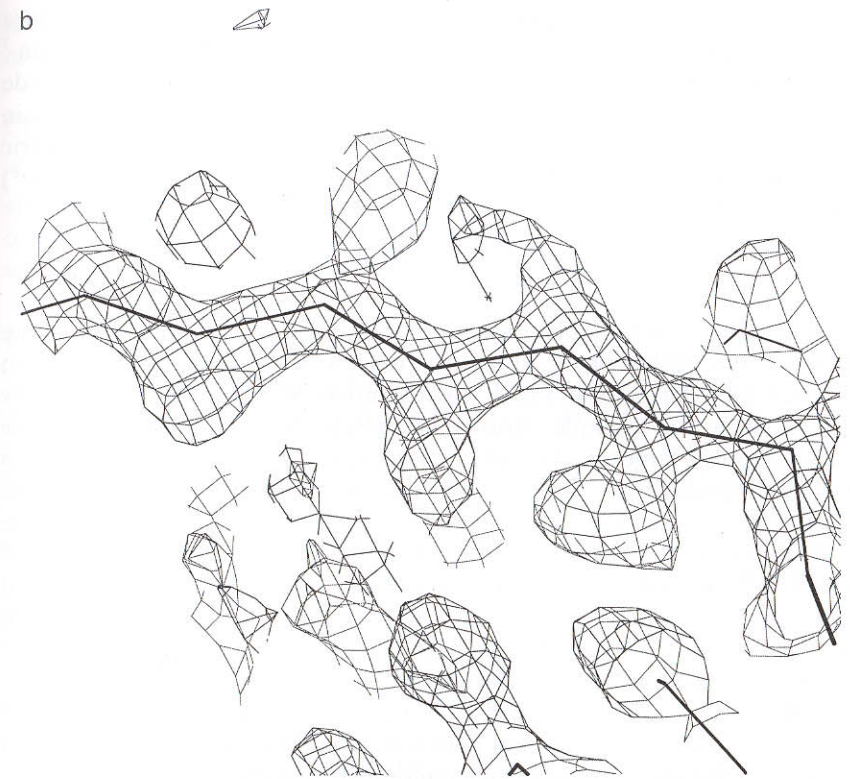

FIG. 5. (*continued*)

Phase Errors

The experimental techniques that are presently employed, be they MIR/MIRAS, multiple anomalous diffraction (MAD), or molecular replacement, cannot prevent phase errors in the diffraction data. Depending on the severity of these errors, the maps that are generated may be easy, difficult, or quite impossible to interpret. In our terminology, we call such maps "good, bad, and ugly" maps. In the last few years, we have seen substantial technical improvements in data collection (e.g., use of area detectors, synchrotron radiation, crystal cryocooling) and in phasing methods (e.g., MAD, site-directed heavy-atom incorporation, improved heavy-atom refinement algorithms, density modification techniques, and easier-to-use averaging programs). Together, this has meant that structures are now being solved where previously success would have been much more difficult.

Even good maps will have bad or ugly bits of density. This could be

functional, a result of mobility or disorder, or due to phase error. Figure 5a shows a part of the map used to solve the structure of P2 myelin protein.[7] It corresponds to a β strand, and the up-and-down direction of the side chains can perhaps be recognized. Breaks in the density for the main chain are due to phase errors. This structure has three molecules in the asymmetric unit, and after cyclic averaging with A (a forerunner of the RAVE package[8]) the breaks disappear (Fig. 5b).

Resolution

The level of detail in an electron-density map will depend also on the quantity and quality of the diffraction data. With no phase error, a chain can be traced correctly at 4 to 4.5-Å resolution. Such examples are usually special cases, for example, viruses, for which phase refinement by cyclic averaging of a large number of noncrystallographically related units can give good-quality maps. In the work on satellite tobacco necrosis virus, the chain was traced at a nominal resolution of 3 Å,[9] but it turned out that the phasing was essentially perfect to around 3.8 Å and random between 3.8 and 3 Å.[10] The tracing of the first model was correct but with major local errors caused, in particular, by unexpected ion-binding sites. In the work on bacteriophage MS2, the first model was constructed at 4.2-Å resolution.[11]

After the first model has been built, the resolution of the diffraction data should be the single factor most likely to determine the accuracy of the final model. The diffraction data sets should be complete, with a high signal-to-noise ratio, and have high multiplicity of individual measurements.

Lack of Experience

The most experienced people in a research group usually do not carry out the model building. It is also true that students must learn the trade. In our opinion, the research leader/advisor must allow the student to make the chain trace and build the structure, but the work must be monitored carefully. We have often seen people start working on their own maps before learning the pitfalls by studying examples. Unless a determined effort is made, this problem will get worse as more molecular biologists

[7] T. A. Jones, T. Bergfors, J. Sedzik, and T. Unge, *EMBO J.* **7,** 1597 (1988).
[8] G. J. Kleywegt and T. A. Jones, *in* "From First Map to Final Model" (S. Bailey, R. Hubbard, and D. Waller, eds.), p. 59. SERC Daresbury Laboratory, Warrington, U.K. 1994.
[9] L. Liljas, T. Unge, T. A. Jones, K. Fridborg, S. Lövgren, U. Skoglund, and B. Strandberg, *J. Mol. Biol.* **159,** 93 (1982).
[10] T. A. Jones and L. Liljas, *J. Mol. Biol.* **177,** 735 (1984).
[11] K. Valegård, L. Liljas, K. Fridborg, and T. Unge, *Nature* (*London*) **344,** 36 (1990).

use crystallography, and as crystallographic software takes on the properties of a "black box."

Competition

In the rush to be first with a new exciting structure, the published structure may not have been refined as carefully as it should have been, or the study may be incomplete. For example, the crystals in question may actually diffract to better resolution than was collected.

Belief in a Number

Except at the highest resolution, the crystallographic R factor is not a sufficiently good indicator by itself of the correctness of a model.[12] With the efficient minimization algorithms available in refinement programs, it is not difficult to reduce the R factor, without actually improving the model. The reason for this is the large number of parameters in a refinement of a macromolecular structure, compared to the relatively small number of diffraction measurements. Therefore, after a certain stage, the refinement algorithm will begin to fine-tune the parameters to fit the noise in the data. Common bad practices that might be used in an attempt to reduce the crystallographic R factor include removal of diffraction data, manipulation of the resolution range, removal or reduction in the weight of stereochemical restraints, increasing the number of parameters being refined [for example, by removing noncrystallographic symmetry (NCS) constraints and/or restraints, or by the use of an inappropriate temperature factor model], and uncritical addition of water molecules. These types of mistakes can be reduced if the free R factor[3] is used to monitor the progress of the refinement, instead of the "classic" crystallographic R factor. For a further discussion of refinement practice, see Ref. 2 and [11] in this volume.[6]

Lack of Equipment

Computers and graphics workstations are now less expensive, but the number of projects has also increased. This means that getting access to the equipment remains a problem in many laboratories.

Bad Refereeing

Referees take some responsibility for the scientific content of what gets published. Unfortunately, the amount of time spent on each paper may

[12] C.-I. Brändén and T. A. Jones, *Nature* (*London*) **343,** 687 (1990).

vary enormously. The ultimate responsibility for the structure must remain with the authors.

Bad Journals

The most prestigious journals aim to sell to a general audience and, therefore, do not want papers to be swamped with crystallographic detail. If a journal refuses to publish sufficient crystallographic detail, the referee still should insist on seeing it.

Data Deposition

Although most journals now insist on the deposition of coordinate data in the Protein Data Bank (see [29] in this volume[12a]), few journals insist on the deposition of diffraction data. Insisting on the immediate deposition of all experimental data would be the single most useful way to improve the quality of crystallographic models.

What Crystallographers Need to Do

Crystallographers should try to collect good-quality diffraction data. There can be no substitute for accurate, complete, multiply recorded, high-resolution diffraction data. The following should also be remembered:

1. Learn how to use the tools. Whatever model-building program is to be used, try to solve the example structure distributed with O. This includes the experimental MIRAS map and skeleton used to solve P2 myelin protein at 2.7-Å resolution.[7] This example has many of the errors that can be expected in an experimental map, but the protein is small enough to be traced in a few days, even by a beginner. This structure has three molecules in the asymmetric unit, so the map can be averaged. The resulting map is easy to interpret.
2. Every model should be treated as a hypothesis. Be aware of the tendency to believe too early in a particular trace.
3. The model should make chemical sense and satisfy all that is known about the macromolecule. Remember, however, that what has been published previously about the molecule could be wrong.
4. Keep the experimental map. It is not tainted by your model, nor beliefs. During the refinement process, this map can be used to check any major changes in the model.

[12a] E. E. Abola, J. L. Sassman, J. Prilusky, and N. O. Manning, *Methods Enzymol.* **277,** [29], 1977 (this volume).

5. Expect the unexpected.
6. Adopt good refinement practices.

What O Offers Crystallographers

It is outside the scope of this chapter to describe too many details about O, but the main functions of the program are shown in Table I.

TABLE I
MAIN FUNCTIONS AND KEYWORDS OF O, GROUPED ACCORDING TO CATEGORY

Function and keyword	Comment
Display functions	
Draw	Create and display molecular objects
Paint	Color atoms, residues according to properties. Color objects
Select	Select/display atoms/residues according to properties
Crystallographic tools	
Manip	Move atoms or groups of atoms interactively
Lego	Database modeling
Bones/Trace	Manipulate and modify skeletonized electron density
Baton	Build protein main chain using a dipeptide
Symmetry	Display symmetry and packing
Map/Qmap	Display electron-density maps
Slider	Determine where sequence matches density.
Mask	Display and manipulation of molecular envelopes
Patterson	Solve Patterson functions
RSR	Real-space refinement into electron density
Refi	Hermans and McQueens regularization
Sam	Coordinate input/output in common formats. Water adder
MolRep	Evaluate and modify MR solutions interactively
Merge	Merge coordinate data from different molecules
Structure analysis	
LSQ	Least-squares alignment of related molecules. Object transformations
Trig	Analyze distances, angles. Display H-bond pattern
Yasspa	Make assignments of secondary structure
RS_fit	Analyze real-space electron-density fit
Pep_flip	Analyze peptide plane orientation
RSC_fit	Analyze side-chain rotamers
Presentation of structures	
Sketch	Make stylized pictures of proteins/nucleic acids
Plot	Generate hardcopy output
Graph	Interactive graphing of database entries (including Ramachandran plots)
Miscellaneous	
OHeap	Read, write, copy, add, delete datablocks

One design philosophy of O is to address one of the problems of building a correct structure by enforcing a systematic mode of model construction, one that makes use of bits and pieces of existing, well-refined structures. Users are encouraged to use side-chain rotamers instead of changing dihedral angles. Other facilities are available for spotting and analyzing potential trouble spots in the structure. This, in an effective yet unobtrusive way, improves on the crystallographer's knowledge of protein structure. It also reduces the number of degrees of freedom in the model. Access to a fold library is also useful during the initial interpretation stage.

O incorporates the use of residue-based checking criteria. Many useful verification criteria are available and some are described in more detail in [11] in this volume.[6] They can easily be added to a user's database, and used for atomic selection or coloring. The fit of the structure to the electron density is particularly interesting during the map interpretation and building stage. The residue-based residuals of Jones *et al.*[1] are explained in Fig. 7. The matrix ρ_{obs} is the experimental map with a particular set of grid spacings. The existing model is used to generate ρ_{calc} on an identical grid. In O this is done by assuming a Gaussian electron-density distribution for each atom

G1	G2	G3
ρ_{obs}	ρ_{calc}	ρ_{env}
experimental	from model	envelope from model, containing fragment of interest

For all nonzero ρ_{env}, calculate:

$$\text{RSRF} = \frac{\Sigma|\rho_{obs}-\rho_{calc}|}{\Sigma|\rho_{obs}+\rho_{calc}|}$$

or

RSCC = Correlation coeff. (ρ_{obs}, ρ_{calc})

FIG. 7. Quantitative fit of a model to a map (real-space *R* factor and real-space correlation coefficient). The electron-density functions G1, G2, G3 are calculated on identical grids.

that makes use of resolution-dependent parameters. Each atom is assumed to have the same temperature factor. The third density, ρ_{env}, is also a calculated electron density, but built from only a subset of the current model. For every nonzero value in ρ_{env}, it then becomes possible to evaluate how well the observed density fits the calculated density. Either an *R* factor-like formulation can be used or a correlation coefficient. The subset of atoms used to create the envelope ρ_{env} can be changed according to what the crystallographer wants to do.

O incorporates 3-D notes. One of the few advantages of a wire model over a computer graphics model is the ability to stick notes on the wireframe! This can be useful during map interpretation, when it may be desirable to jot down such things as chain directionality, or where one is in the sequence. In O, the 3-D notes are associated with a point in space, so that it is possible to display all notes within 10 Å of the current center, for example.

The creation of "any number" of graphic objects is allowed. An object consists of a collection of graphics items such as lines, text, or polygons. Two general object types are available: pickable and nonpickable. Although it is normally desirable to be able to identify atoms by clicking on them with the mouse, it is usually not so interesting to do the same with the vertices of the electron-density wireframe representation. For each object created, its name appears on the screen, and the visibility of the object can be toggled by clicking. Most objects are created explicitly by the user, whereas other objects are created by the program. Any number of objects can be made from a molecule stored in the O database. A molecular object, however, is built from only one molecule. An object description language allows the user to create objects by using a text editor or with a suitable standalone program (e.g., MAMA[8]).

Steps in Building a Model

We identify five key steps in building a model: (1) generating a main-chain trace, (2) determining where the sequence matches the density, (3) building a rough model, (4) optimizing the fit of the model to the density, and (5) evaluating the model. These steps may not lead to a model that completely matches the full sequence of the molecule being studied. Indeed, there are many cases in the literature in which the first model to be built was just a partial polyalanine chain. Frequently, part of the structure can be built that matches the sequence while the rest is polyalanine. Such structures will usually become more complete after another round of calculations that could include phase combination, for example.

Using Skeletons

Map interpretation requires both an overview and a detailed representation of the electron density. This can be achieved by simultaneous viewing of skeletonized and contoured density, respectively. Skeletonized density can be thought of as a piece of string passing through the map. This representation was introduced by Greer[13] in early attempts to automate map interpretation. We use this representation as a means of indicating where we have been in the map and to show our interpretation of the density in terms of a main-chain trace. When a skeleton is used instead of a contoured map, one experiences a great loss of detail. The benefit of this representation, however, is that one is able to view a much larger volume of space without clutter. The skeleton data structure also allows us to associate extra information with this representation, such as our current hypothesis for the fold. In O, the skeletonized density data structure is similar to that used to describe molecular structures.

A number of different skeletonization algorithms have been described.[14,15] The final O data structures are straightforward and easy for someone with another algorithm to adopt [K. Cowtan (University of York, U.K.), for example, can generate O-style skeletons in the density modification program *dm*]. The Greer algorithm requires the definition of both a base level and a density step value. In this algorithm, all points below the base level are removed. The density is then searched for points with values of base+step. They will be removed unless they are needed to preserve continuity or are end points. This is repeated at a level of base+2*step, and so on, until just a skeleton of connected points remains. In the program *bones,* we first calculate a skeleton with base and step values typically 1.25 and 1.0 times the root-mean-square (r.m.s.) deviation level of the map in the whole asymmetric unit. If there are too many connections when viewed in O, it is necessary to increase the base level and recalculate the skeleton. If we see that too few skeleton atoms are connected, the base level needs to be decreased and the skeleton recalculated. The step value is not a sensitive parameter. It should be noted that the value of the density at each skeleton point is not part of the data structure and therefore is not passed on to O. This was an early design decision that 10 years of use seems to have justified. Instead, the crystallographer worries about skeleton connectivity and status. The *bones* program calculates initial status codes that are assigned to skeleton atoms based on the length of the connected fragment of which the atom is a member. Within O, these codes can be used to view and paint the trace according to any scheme that may be

[13] J. Greer, *J. Mol. Biol.* **82,** 279 (1974).
[14] C. K. Johnson, *Acta Crystallogr.* **A34,** S-353 (1977).
[15] S. M. Swanson, *Acta Crystallogr.* **D50,** 695 (1994).

helpful. In the simplest case, one may merely change the codes to mark clearly where one has passed through the map, deciding what is main chain and what is not. In a more complicated scheme, one could use one color for part of the trace where one has placed the sequence in the density, another color where one is sure of the main-chain trace, a third color where a main-chain branch could be wrong, etc.

When the skeleton is being modified, it is vital that the crystallographer view the 3-D contour representation of the density to provide extra detail such as size and shape. This representation is sensitive to the density level; if necessary, therefore, the contoured map can be viewed at higher or lower values. Two sets of commands exist within O to view contoured maps. One works with density brick files (*map* commands), the other reads in the whole map file (*qmap* commands; various map formats are supported). The latter commands maintain a database of contoured brick objects so that if the level is left unchanged, the recontouring time is much reduced.

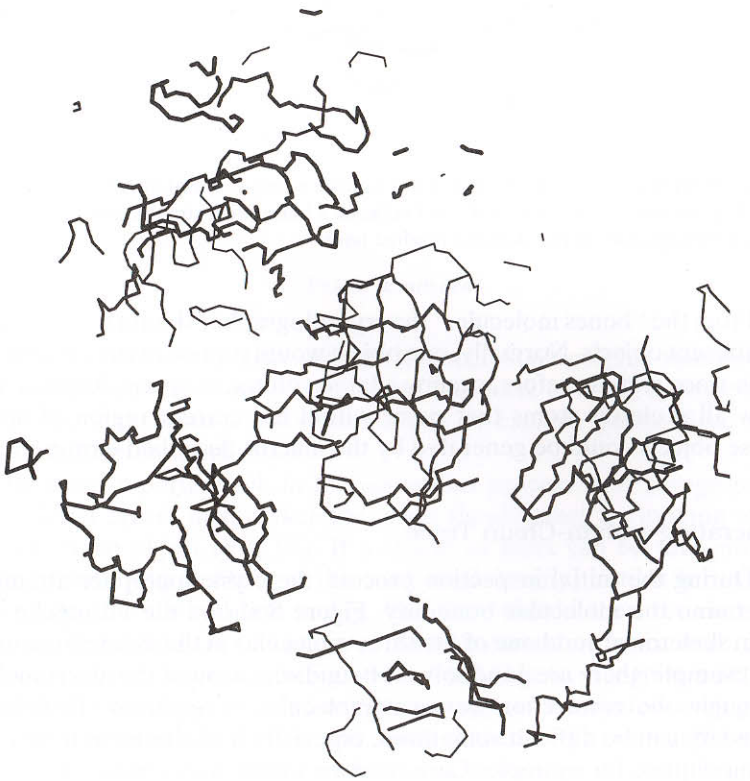

FIG. 8. Overview skeleton from P2 myelin MIRAS map. All "main-chain" bone atoms are drawn that are within 50 Å of the screen center.

FIG. 9. (a) Starting a main-chain skeleton with the wrong connections. (b) Two connections have been broken (between strands, and between a helix and a strand), and a path has been defined through part of the skeleton (dashed line).

From the "bones molecule," the crystallographer should make a number of different objects. Normally, one object would represent the current main-chain trace and, therefore, occupy a large volume of space. Another would show all skeleton atoms that occur within the current region of interest. These objects could be generated by the macros described earlier in Fig. 6.

Generating a Main-Chain Trace

During the initial inspection process, the crystallographer attempts to determine the molecular boundary. Figure 8 shows the P2 myelin main-chain skeleton around one of the three molecules in the asymmetric unit. In this example, there are good solvent boundaries around the macromolecule although one connection between molecules is apparent. Defining the boundary can be difficult sometimes, especially if chains form tightly interacting dimers, for example. Care must be taken not to move into another molecule. It may, therefore, be useful to generate a symmetry object of

FIG. 9. (*continued*)

the current main-chain trace object. If noncrystallographic symmetry is present, one can also work at improving the mask defining a single copy. This is best done with a macro to include or exclude points close to an identified skeleton atom.

The most important job in the inspection process is to recognize and correct local errors in the skeleton. One should start by looking at the overview trace object (Fig. 9a). If a strand or helix can be recognized it should be followed (looking at the contoured density) until it is not clear what to do (the dashed line in Fig. 9b). One should then return to the overview object and try to recognize another strand or helix, and repeat the process. From the trace in Fig. 9a, it is possible to recognize a set of β strands immediately. Two of these strands appear connected, owing to a main-chain hydrogen bond (Fig. 10). Such connectivity errors in the skeleton should be corrected immediately, either by making or, in this case, by breaking bonds between the bone atoms (Fig. 9b). At the same time,

FIG. 10. Close-up view of the connection error between strands, caused by a main-chain hydrogen bond connection.

classification errors should also be corrected, because in many places the main chain may be incorrectly marked as a side chain and vice versa. This is illustrated in Fig. 11 where a side chain from a neighboring α helix interacts with a β-strand side chain. The path taken along the trace can be recognized in Fig. 11 because some of the skeleton atoms have been reclassified already as part of the likely main chain. Methods have been developed that assist in recognizing structural entities such as helices, strands, and other atomic groupings.[16] These are based on the real-space convolution of the experimental map with a structural template. The resulting map has density where the correctly oriented template fits the experimental density.

[16] G. J. Kleywegt and T. A. Jones, *Acta Crystallogr.* **D53,** 179–185 (1997).

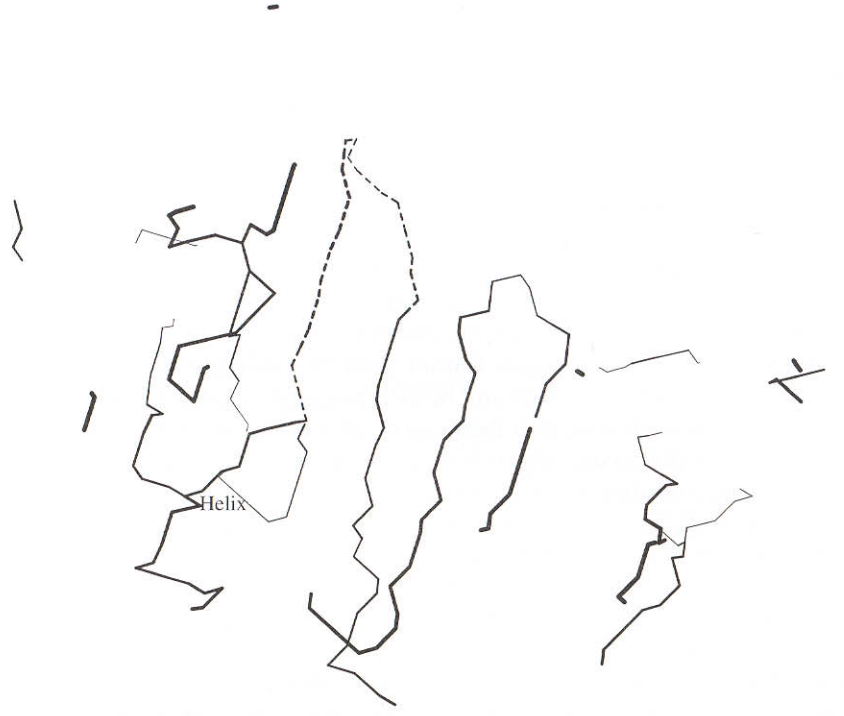

FIG. 11. Close-up of the connection error between a helix and a strand. The helix is to the left and the error is a series of links to the end of the dashed trace.

The level of care taken in repositioning the bone atoms depends on how one plans to build the structure. If the main chain is to be constructed by the placement of C_α guide points in the skeleton (see below), the skeleton atom at the intended position must be placed as accurately as possible. The r.m.s. error in placement is approximately the same as the error in the coordinates of the main-chain atoms that are built using databases,[1] and so it pays to be careful. If the main chain is to be built with the *baton* commands (see below), it is not necessary to position the guide atoms very accurately. With both methods it is a good idea to make the trace continuous and to have a branch point at the (possibly rough) C_α guide point. If the density is smooth at this position, there is a special command in O to add a skeleton branch point. In this case one should not be concerned with the direction of the branch. It is not necessary that the side-chain skeleton atoms accurately mimic the real side chain. We place a branch point even for glycine residues, because this helps later when constructing the main-chain trace.

In an earlier program, Frodo,[5] a number of different ways were introduced to build a structure in an electron density that did not depend on the use of skeletons.[17] In O, the skeleton plays an important role in the construction process. In one method, this can be a direct role wherein the crystallographer creates a so-called bones C_α trace that accurately mimicks a C_α trace in terms of the connectivity and the number of atoms. In this trace, each skeleton atom in a connected segment will be associated with the C_α coordinates of a particular residue in the structure under construction. By specifying a start and end point in a portion of skeleton, one can merge the coordinates of each bone atom into consecutive C_α coordinates in a portion of the protein. It is not necessary to edit the skeleton extensively to produce this bones C_α trace. Rather, once the skeleton has been edited to remove connectivity errors and branch points are added (or removed if in error), a connected section then can be processed to produce the bones C_α trace. The processing algorithm keeps all skeleton atoms that have a branch point and then investigates the separation between these atoms. If this is too long, an atom is added so that neighboring skeleton atoms are approximately separated by the interresidue C_α–C_α distance of 3.8 Å. The connection between the two identified atoms in Fig. 12a is made up of many short bonds between skeleton atoms positioned at grid points. After processing (Fig. 12b), only atoms that are roughly 3.8 Å apart are left. Note that when this trace is merged into the molecule of interest, a decisive step has been taken, namely a decision has been made as to where the electron density and sequence are to be matched. Tools to assist in this process are described below.

When editing the skeleton, one must pay attention to possible false connections such as hydrogen bonds between strands, interacting side chains that can make the density continuous, or disulfide bridges. The latter usually have rather strong density and can be mistaken for the main chain. One will frequently see breaks in the density. In some circumstances, one can make use of the local secondary structure to decide where the side chains are or should be pointing (Fig. 5a). When in doubt, one should proceed in the direction of the secondary structure element.

When moving along the skeleton, one should be attempting to decide on chain directionality and where the sequence of the molecule can be recognized in the density. It then becomes useful to jot down ideas on 3-D notes. In O, these are related to positions so that later one can display all notes within 10 Å of the current screen center, as in Fig. 11, for example, where we were able to recognize a helix. The local chain direction can be

[17] T. A. Jones, *in* "Computational Crystallography" (D. Sayre, ed.), p. 303. Clarendon Press, Oxford, 1982.

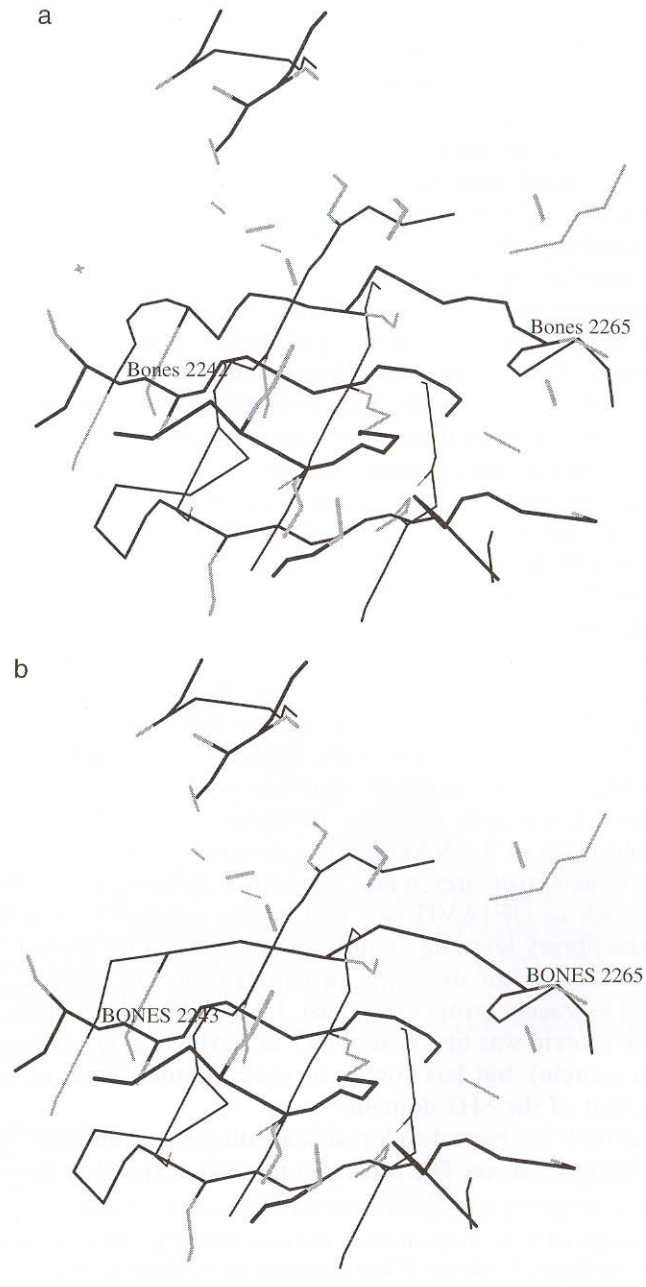

FIG. 12. Creation of a bones C_α trace. In (a) the skeleton is drawn showing the assignments that have been made for main and side chains. In (b) the connection between the identified atoms has been contracted to just those atoms at branch points or separated by ~3.8 Å.

determined in a number of ways. One way is based on the α-helix Christmas tree effect. This is a result of the direction of the C_α–C_β bond relative to the helix axis, so that side chains tend to point toward the N terminus of the helix. This can also be done quantitatively, by measuring the goodness of fit of a polyalanine segment to the electron density when built in either direction. A second, more general, approach requires good phasing and high resolution, ~2.5 Å or better. It then becomes possible to recognize peptide branching. The separation of branch-point pairs corresponding to atoms C_{α_i} and C_i is different from that of pairs C_{α_i} and C_{i-1}, and this allows one to determine the chain directionality.

When thinking about sequence placement, as well as when studying electron density (see below), one can sometimes make use of other information. For example, it may be easy to determine the active site by the location of a heavy atom if one is present, or by binding studies where the experimental phases may be good enough to locate ligands before solving the structure. In many cases, the identity of residues in the active site may be known already. In an MIR map, the known preference of some heavy-atom compounds for particular amino acid side chains may provide useful information. When using the MAD phasing method or when heavy-atom binding sites have been engineered by introducing cysteine residues, there is even more useful sequence information available.

As one moves through the map, one should try to recognize local protein-like features such as α helices, β strands, special side chains (for example, large aromatics), cofactors, ligands, or any metal-binding sites. After a while, one may recognize supersecondary structure motifs such as β–α–β units. Eventually one may recognize one of the more common domain folds such as the NAD-binding domain or a TIM barrel. With the avalanche of new structures, it may be useful to interrogate databases using programs such as DEJAVU (see [27] in this volume[18]). In general, time spent in the library learning about proteins will not be wasted. However, care must be taken for structural variations from the "classic" fold since this has led to tracing errors in the past. In one case, for example, photoactive yellow protein was first described[19] as having a β-clam structure (like P2 myelin protein), but has now been resolved and shown to have a fold more like that of the SH2 domain.[20]

A set of tools has been developed in an attempt to automate the production of a bones C_α trace. The tools aim to make a skeleton more like a C_α trace by applying a set of filters to an existing skeleton to produce a new, "better" skeleton. These tools need either a good map, or require that the skeleton be edited to remove most of the serious connectivity errors. The pruning filter keeps the connectivity only between atoms with branch points. The fill filter places C_α atoms along the trace at suitable spacings (~3.8 Å). Applying just these commands to the skeleton obtained from the averaged P2 myelin map in Fig. 5b allowed us to create automatically a polyalanine model that matched 115 of 131 residues of P2 with an r.m.s. deviation on C_α atoms of 1.2 Å. The errors in this model result from defects in the skeletonization algorithm and because some side chains interact to produce continuous density. We have stopped the development of this approach for the moment, because the interactive method described above gives more control to the user and the *baton* method is even more popular with users.

[18] G. J. Kleywegt and T. A. Jones, *Methods Enzymol.* **277,** [27], 1997 (this volume).
[19] D. McRee, J. Tainer, T. Meyer, J. Van Beeuman, M. Cusanovich, and E. Getzoff, *Proc. Natl. Acad. Sci. U.S.A.* **86,** 6533 (1989).
[20] G. E. O. Borgstahl, D. R. Williams, and E. Getzoff, *Biochemistry* **34,** 6278 (1995).

Placing the Sequence in the Density

Placing the sequence in the density is the crucial step in building a model, and this is where the qualitative aspects of model building are most apparent. The main reason for this is that the quality of the density for an amino acid side chain depends on where it is in the structure. This means that an external, floppy, tryptophan residue could have just smooth main-chain density, thus looking very much like a glycine (note that a glycine residue cannot look like a tryptophan). Glycines can be useful markers during this important step. Unfortunately, glycines are often found in loops that are usually external and frequently have bad density. Glycine residues in the middle of some solid density are more useful.

Large aromatic residues and disulphide bridges are usually the most useful markers for locating the sequence in the electron density. One must, however, be on the lookout for the unexpected, including unexpected glycosylation sites such as the one that caused the error in Fig. 3, and unexpected ligands. The latter may often have strong density, such as the endogenous fatty acid in the barrel of P2 myelin,[7] and the mixed peptide population in the first HLA structure.[21]

Slider Commands

To help in deciding where the sequence fits the density, O can assist the crystallographer with the *slider* commands.[22] One group of commands is

[21] P. J. Bjorkman, M. A. Saper, B. Samraoui, W. S. Bennett, J. L. Strominger, and D. C. Wiley, *Nature (London)* **329,** 506 (1987).
[22] J. Y. Zou and T. A. Jones, *Acta Crystallogr.* **D52,** 833 (1996).

qualitative in nature, whereas another group is quantitative. The qualitative *slider* commands allow the user to enter a guess of the sequence on the basis of the shape of the density. This guess is compared with the real sequence, using a scoring matrix to decide where we are in the sequence. The matrix is a look-up table that is needed to evaluate how well the guess scores for each of the 20 amino acids. In a simple system we could use the letters *b*, *m*, and *s* to indicate big, medium, and small residues, for example. With such a system, a tryptophan in the sequence would have a high score if it were guessed as *b*, lower if *m*, and even lower if *s*. Similarly, a glycine in this system would score low, higher, and even higher for guesses *b*, *m*, and *s*, respectively. The standard matrix distributed with O is more complicated and the guesses appear at first sight like the one-letter amino acid code (the matrix can, of course, be modified by the user). If the user

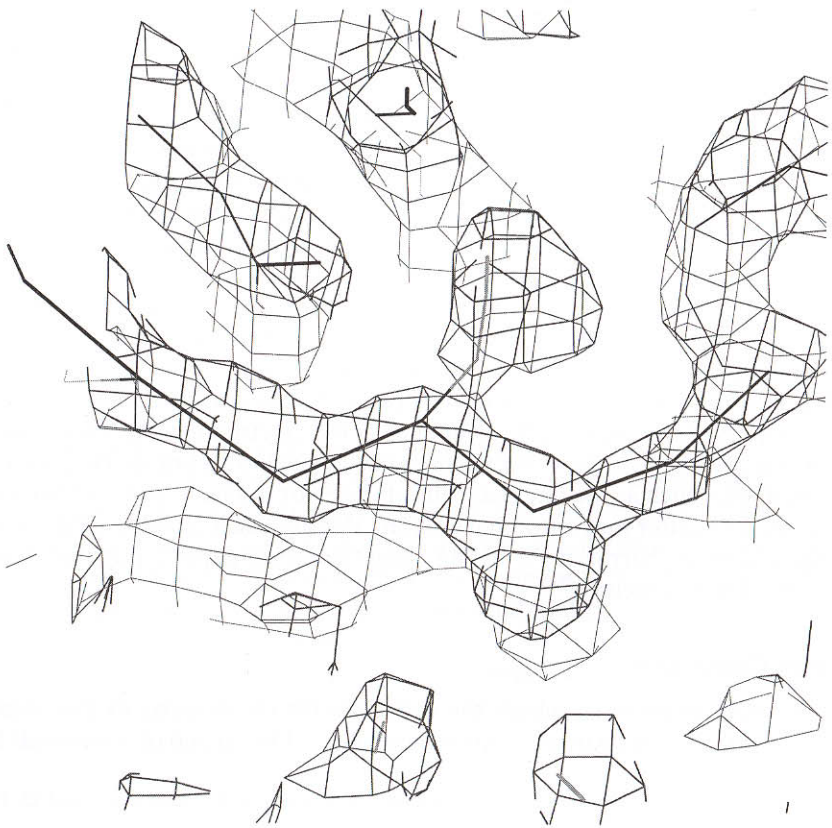

FIG. 13. Electron density and skeleton around a well-defined aromatic ring.

wishes, the result can be associated with a portion of a molecule and stored in the database. It may often be worthwhile to build a portion of a polyalanine chain in the density of interest to see how well the different amino acids fit the density. The *slider_rotamer* command then allows one to display each kind of amino acid side chain and its equivalent set of rotamers. Especially for beginners, this helps in judging the size of different side chains relative to the density of interest.

One of the problems with *slider* is that a long stretch of residues is needed to be sure that the correct result appears at the top of the scoring list. Unfortunately, a long stretch may have an insertion or deletion in it that has already been missed. Therefore, once a number of guesses have been made, they can be combined with the introduction of a variable-

```
Slid> Estimated sequence  : gvwa      Slid> Estimated sequence  : awvg
Slid> Average=0.44,rms=0.12           Slid> Average=0.44,rms=0.12
Slid>                        GVWA     Slid>                        AWVG
Slid>  Fit   1 0.875 A6      GIWK     Slid>  Fit   1 0.800 A96     KWNG
Slid>  Fit   2 0.725 A33     GNLA     Slid>  Fit   2 0.750 A3      KFLG
Slid>  Fit   3 0.700 A84     VILA     Slid>  Fit   3 0.725 A33     GNLA
Slid>  Fit   4 0.675 A26     GLAT     Slid>  Fit   4 0.675 A112    KMVV
Slid>  Fit   5 0.675 A55     SPFK     Slid>  Fit   5 0.650 A82     STVT
Slid>  Fit   6 0.675 A122    VVCT     Slid>  Fit   6 0.625 A84     VILA
Slid>  Fit   7 0.650 A99     GNET     Slid>  Fit   7 0.600 A23     LGVG
Slid>  Fit   8 0.650 A82     STVT     Slid>  Fit   8 0.600 A53     TESP
Slid>  Fit   9 0.650 A111    GKMV     Slid>  Fit   9 0.600 A40     VIIS
Slid>  Fit  10 0.625 A28     ATRK     Slid>  Fit  10 0.600 A25     VGLA
Slid>  Fit  11 0.625 A89     GSLN     Slid>  Fit  11 0.600 A80     TKST
Slid>  Fit  12 0.600 A62     ISFK     Slid>  Fit  12 0.600 A26     GLAT
Slid>  Fit  13 0.600 A40     VIIS     Slid>  Fit  13 0.600 A21     KALG
Slid>  Fit  14 0.600 A72     ETTA     Slid>  Fit  14 0.575 A64     FKLG
Slid>  Fit  15 0.600 A24     GVGL     Slid>  Fit  15 0.575 A111    GKMV
Slid>  Fit  16 0.600 A50     TIRT     Slid>  Fit  16 0.575 A81     KSTV
Slid>  Fit  17 0.600 A22     ALGV     Slid>  Fit  17 0.575 A60     TEIS
Slid>  Fit  18 0.575 A102    TTIK     Slid>  Fit  18 0.575 A123    VCIR
Slid>  Fit  19 0.550 A79     KTKS     Slid>  Fit  19 0.575 A9      KLVS
Slid>  Fit  20 0.550 A73     TTAD     Slid>  Fit  20 0.575 A72     ETTA
```

FIG. 14. *Slider* example from the tetrapeptide density shown in Fig. 13. The two sets of output search for the guess in both directions in the sequence. The output above is a pasting together of two runs with the *slider_guess* command.

length gap between them. One other drawback remains, and that concerns the chain directionality. If this is unknown, the reverse sequence must also be tried. For example, in the portion of density from P2 myelin shown in Fig. 13, one may guess that the sequence is GVWA after careful inspection and use of the *slider_rotamer* command. The comparison of both directions with the sequence gives the result shown in Fig. 14. In this case, the correct direction gives the best score.

A quantitative approach has been developed[22] that looks promising. Given a well-fitting polyalanine model, in this method we mutate each residue to each of the 20 different amino acids. For each amino acid, the fit of each rotamer is optimized to the density. Only a rotational search is carried out, pivoting the whole residue about its C_α atom. The real-space fit of the best-fitting rotamer is then used as an index of fit to determine how well that particular amino acid type fits the density. The resulting scoring matrix can then be used in the *slider* system. For the averaged P2 myelin map, this method works rather well. If one searches for a fragment of five residues, the correct answer is the top score 35% of the time. Increasing the length to 8 and 15 residues gives 67 and 86% correct scores, respectively. The main problem is caused by interacting side chains, often resulting in better fits for incorrect but longer side chains. As one develops the model, this problem should become less serious, because more and more density is occupied by fitting atoms. It should also be noted that an amino acid such as alanine will fit snugly into the density of a phenylalanine. This problem can be overcome by using envelopes for evaluating the goodness of fit that are larger than normal for the small side chains.

Generating the First Rough Model

The first rough model is usually built from a C_α trace, which is then "fleshed out" using databases. The C_α trace can be made either from a skeleton (or from multiple skeletons) as described above or by use of the *baton* commands. In its simplest mode, this command bears some resemblance to the methods used in Frodo[5] and to a prototype program developed by J. Pflugrath. In Frodo, one could "pop" along the sequence, adding the coordinates for the next residue in a standard conformation, and connected to the previous residue with standard bond lengths and angles. One could then rotate and translate the residue so that the main chain fitted the density and the side chain pointed off toward appropriate density. The *baton* method is similar but only a dipeptide is drawn and normally only the C_α atoms of the dipeptide are viewed. The rotation operation is now a pivot around the first C_α of the dipeptide and the position of the second residue is where the next residue to be built will be placed. As with Frodo,

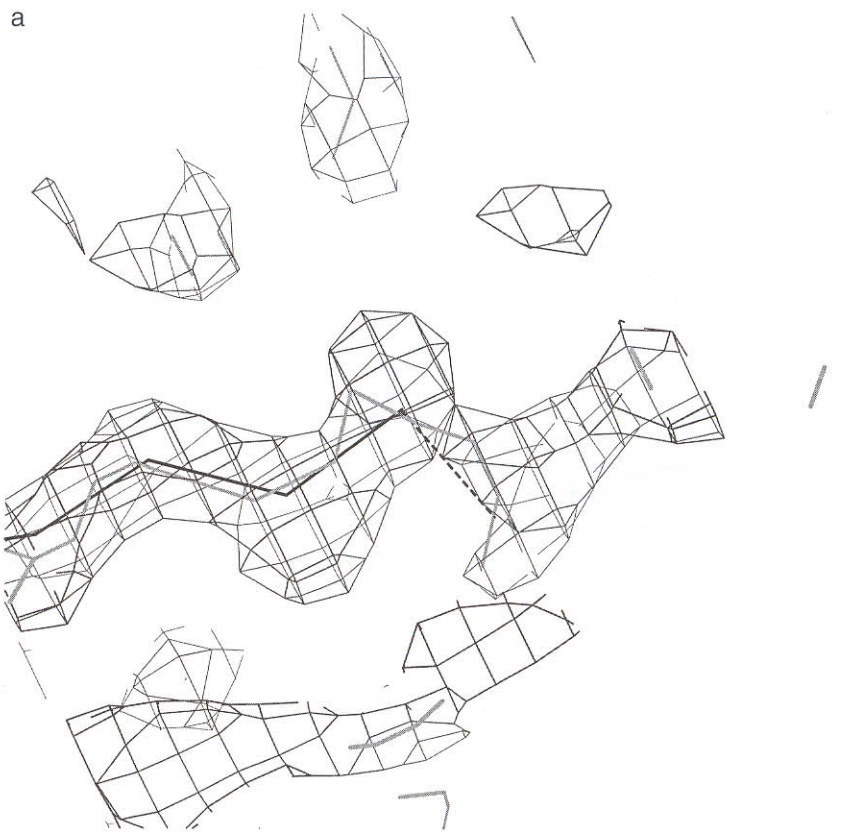

FIG. 15. Use of the baton to build the C_α trace. In (a) the dipeptide baton follows the skeleton, while in (b) it is positioned to continue building a β strand.

a buildup of errors occurs that requires an occasional translation of the dipeptide to maintain the fit to the density. In the present implementation, the user is not concerned with the placement of the side-chain atoms because this will be approximately correct after later database building. When accepted, the coordinates of the second residue of the dipeptide are written into the main chain of the structure being built. The dipeptide is repositioned so that it now pivots about the C_α position of this residue. The advantage of this method over the skeleton method described above is that one has better control in placing adjacent residues at ~3.8-Å intervals. As one gains experience with this method, the dipeptide indeed spins through the density much like a baton cast into the air.

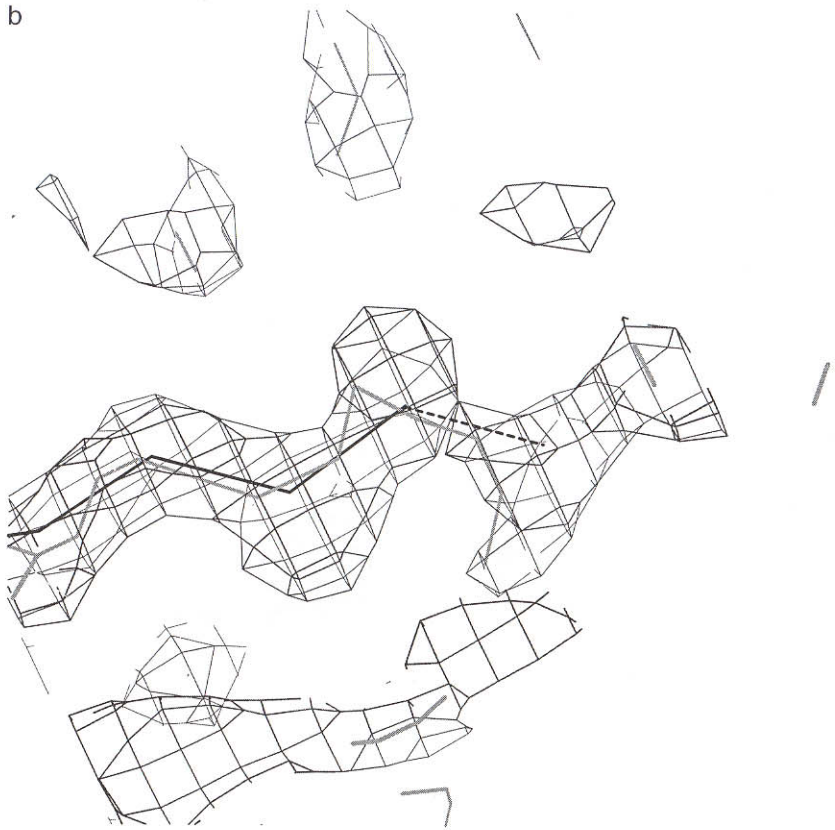

FIG. 15. (*continued*)

The next placement of the baton after the user has accepted the current position actually depends on a mode switch. In the simplest mode, it is always placed in a fixed orientation and the user then needs to change some dials to point it into the density. Alternatively, depending on the mode switch, it can be positioned to "follow" a particular skeleton or to build a standard piece of secondary structure (Fig. 15). When "skeleton sniffing," the baton may have a number of alternative possibilities due, for example, to branching. These can be cycled one at time via an algorithm that shows what are considered to be the most likely alternatives. In every alternative, the second C_α of the baton will be placed on a line joining connected skeleton atoms (Fig. 15a). The order of preference is based on the codes of these linked skeleton atoms, so that a placement between a

pair of main-chain skeletons is taken as the most likely possibility, and a placement between linked side-chain atoms as the most unlikely. The baton tip will not be placed between skeleton atoms in the direction from which the structure has just been built. When building in secondary structure mode, the last four residues that have been built are used to decide where to place the dipeptide after making a least-squares comparison with short α-helix and β-strand units. Whichever mode is in use, the crystallographer is free to decide where the dipeptide is ultimately to go. In Fig. 15a, a break in the main-chain density results in an incorrect initial placement of the baton. On changing to secondary structure mode, the baton appears at roughly the correct place (Fig. 15b).

Whether the C_α trace has been made from skeletons or using the *baton* commands, the main chain is then (re-)made from a database of well-refined structures. Jones and Thirup[23] showed that the protein main chain could be generated from short fragments taken from a library of well-refined structures. For retinol-binding protein (RBP), on average, fragments of ~7 residues could be found that locally matched with r.m.s. deviations of <0.5 Å on C_α atoms. With such small deviations, the peptide planes of these fragments were well aligned with those in the RBP structure. Therefore, provided the correct fragment is located, one can be fairly confident that the carbonyl oxygens will be correctly oriented. In a refinement of the basic algorithm for use in O, Jones *et al.*[1] have discussed how well the method works when errors are introduced into the coordinates of the C_α guide atoms. Not surprisingly, the largest deviations were found for the carbonyl oxygen atoms. The reconstructed main chain had an r.m.s. error approximately equal to the r.m.s. error introduced into the guide coordinates. The benefits of using the database approach in map interpretation have been described by Zou and Mowbray,[24] who compared three structures of the periplasmic glucose/galactose-binding protein (GBP). One was built and refined with databases at 2.4 Å, and then further refined to 1.7 Å. The third structure was from a related GBP refined by more traditional methods. They demonstrated that the use of databases both speeded up the modeling process and improved the quality of the models.

The algorithm in O (referred to as autobuilding) searches for the best-fitting pentapeptide in the current window being built, but keeps only the main-chain coordinates of the central three residues. The window is then moved forward three residues (i.e., ensuring a two-residue overlap with the previous window) and the next pentapeptide is determined. Eventually, only the first and last residues in the region of interest remain to be built.

[23] T. A. Jones and S. Thirup, *EMBO J.* **5,** 819 (1986).
[24] J. Y. Zou and S. L. Mowbray, *Acta Crystallogr.* **D50,** 237 (1994).

This can be done with *baton* or by extending the region to be built artificially by one residue at both ends, prior to autobuilding. By choosing a five-residue comparison window, we ensure that a similar conformation will be found in the database. It is likely that the peptide orientations would be "correct" even if a somewhat bigger zone is used, but there is no need to take that risk. Nothing is gained with a longer window except some minor speed improvements.

At first sight, it might appear surprising that the fragment approach works so well. However, it has become apparent as more and more high-resolution structures are determined that protein molecules adopt energetically preferred conformations. This results in tight clustering in their Ramachandran plots and is a good indicator of the accuracy of a protein model.[25] If we restrict ourselves to just the center of the regions defining the α helix, β strand, and left-handed α helix, the number of possible fragments needed to define a pentapeptide from C_α atoms alone becomes $2 \times 3 \times 3 \times 3 = 54$ pentapeptides (note that the first peptide contributes just a ψ variation). This is a somewhat smaller set than that defined by Sussman and co-workers.[26]

The side-chain atoms are generated from a rotamer database. Initially they are added in the most likely rotamer conformation. The database used in O is based on an unpublished analysis made in 1996 by G. Kleywegt of all high-resolution structures, and includes all conformations that have a frequency greater than 5%. In some of the longer side chains, the torsional angles do not form strong rotamer groupings. Rotamers for arginine, for example, have only χ_1 and χ_2 restrictions.

Optimizing the Fit of the Model to the Density

Provided care has been taken in the placement of the C_α atoms making up the initial trace, the main chain of the rough model should have a good fit to the density. The peptide planes should be "correct" provided the C_α atoms have been placed with an accuracy of 0.5–1 Å. The side chain should be pointing in the right direction, but for many side chains it will be necessary to optimize their fit to the density.

The great success and widespread adoption of Frodo for the construction of a model was due in part to the strategy of allowing the crystallographer to rip apart the model by a combination of group or single atom movements, and/or localized or extended dihedral rotations; in essence an electronic

[25] G. J. Kleywegt and T. A. Jones, *Structure* **4,** 1395–1400 (1996).

[26] R. Unger, D. Harel, S. Wherland, and J. L. Sussman, *Proteins Struct. Funct. Genet.* **5,** 355 (1989).

"Richards box." After placement in the density, the model could then be regularized to adopt standard geometry. While giving great freedom to the crystallographer, this approach could also lead to models that had bad stereochemistry. O has a large set of tools for changing a model (Table II) but in the first rounds of building and refinement, we try to persuade users to restrict themselves to the stereochemically most reasonable conformations of main and side chains. This is accomplished for side chains by using the rotamer library described above. An improved fit to the density can often be made without ripping the model apart. This can be accomplished by moving the whole residue as a rigid body and then selecting the best-fitting rotamer. If one uses the *move_zone* command and double clicks on an atom, that atom becomes the pivot point. Frequently, one would pivot about the C_α atom to point the C_α–C_β bond correctly and then the correct side-chain conformation could be chosen from the rotamer library. This can also be done automatically with a real-space density-optimizing command (*RSR_rotamer*) that pivots around the C_α atom, trying each rotamer in turn. The usefulness of automated methods depends a great deal on the quality of the map.

Whether fitted manually or automatically, the resulting chain still becomes distorted and requires regularization. The deviation of the peptide units from "standard" values can be monitored at any time by determining the "pepflip value," an indication of how much each carbonyl oxygen

TABLE II
MODEL MANIPULATION COMMANDS IN O

Interactive	Real-space refinement
move_atom	*RSR_zone*
move_zone	*RSR_rigid*
move_fragment	*RSR_rotamer*
flip_peptide	*RSR_dgnl*
tor_residue	
tor_general	
grab_atom	
grab_group	
lego_side_chain	
lego_ca	
lego_auto_mc	
lego_auto_sc	
baton_build	
refi_zone	
bond_make	
bond_break	

deviates from similar conformations in the database.[1] Some of the longer side chains will still need to be fitted because they do not have preferred rotamer conformations. This is accomplished with a single click on an atom to define torsional adjustments (*tor_residue* command) and/or fragment moves (the *move_fragment* command). Groups of connected atoms can also be generated by breaking bonds; these then are either moved as a rigid unit (the *group_grab* command) or changed by dihedral rotations around some of the remaining bonds (the *tor_general* command). Once regularized, the *rsc_fit* command can be used to monitor how much each side-chain conformation deviates from one of the preferred rotamers.[1] By using the graphing features of O, the user can also plot one residue error indicator against another as suggested by Zou and Mowbray.[24]

Eventually a quantitative measure of how well each residue fits the density can be evaluated[1] (Fig. 7). In the original description, we used an R factor-like grid summation, but one can also calculate a correlation coefficient. The latter approach has the disadvantage that if a group of atoms fits snugly in weak density, it will still score well. Once calculated, the residue-based properties can, of course, be used to color atomic objects, making it easy to recognize the good or badly fitting portions of the molecule. The property can be evaluated for any set of atoms within each kind of residue. By choosing to evaluate the goodness of fit of just the main-chain atoms, one can see how well the main-chain trace follows through the density. By evaluating just the side-chain atoms, it may be possible to detect out-of-register errors by identifying clusters of poorly fitting adjacent residues.

Evaluate the Model Continuously

As more of the sequence is fitted to the density, the process should become easier if the interpretation is correct. More frequently, the crystallographer should see aspects of what is known about the molecule become clearly understood from the model under construction. Interacting side chains from residues that are far apart in the sequence should make chemical sense, e.g., an arginine residue should form an ion pair with a glutamic acid residue, instead of being buried in a patch of hydrophobic side chains. Interactions between noncrystallographically and/or crystallographically related molecules should also make chemical sense. It may be useful to evaluate the overall distribution of some residues, looking for buried charges, patches of exposed hydrophobic residues (they might be functional and not errors), glycine, and proline distributions. Any program that evaluates residue- or atom-based environments or properties could be modified easily to produce O data structures or macros.

At this stage, the crystallographer must begin checking if the model satisfies what is known about the molecule from biochemical data. If certain residues have been identified as being in the active site, are they close together in the model? If the molecule has disulfide bridges, are they formed in the model? However, the crystallographer must be aware that this kind of published data can also contain errors. In our work on CBHI,[27] for example, an aspartic acid that had been identified as being involved in catalysis, using epoxide labeling techniques, turned out to be ~25 Å away from the active site. Crystallographers must also be careful with their notes. In our attempts to fit the sequence of RBP[28] to the electron density, for example, it was impossible to fit the expected disulfide linkages until it became apparent that they had been written down incorrectly in the notebook (the structure was then retraced in a couple of hours). Similarly, in the structure determination of α-glutathione transferase (GST),[29] the wrong sequence was entered for one residue. The first refinement macrocycles were made with an aspartic acid side chain where there should have been a glycine. This could actually be a good way of evaluating a refinement because in the GST refinement, the side chain had a group temperature factor of more than 60 Å^2 when we detected the error, while the temperature factor for the main chain was 2 Å^2.

If the structure is related at the primary sequence level to a family of proteins, it will be necessary to evaluate the sequence similarities in light of the structure. A conserved hydrophobic core would be good to see; large deletions mapping to central β strands could indicate problems.

Learning to Use the Tools

The skeletons and maps described in this chapter are available from the authors. These were used to solve the structure of P2 myelin protein and are a good start for learning about map interpretation. It should be possible to make a successful interpretation after about 2 days work. A series of macros are also available that take one through the map, pointing out different kinds of errors in various regions of interest. Other introductions to the program (*O for Morons* and *O for the Structurally Challenged*)

[27] C. Divne, J. Ståhlberg, T. Reinikainen, L. Ruohonen, G. Pettersson, J. K. C. Knowles, T. T. Teeri, and T. A. Jones, *Science* **265,** 524 (1994).

[28] M. E. Newcomer, T. A. Jones, J. Åqvist, J. Sundelin, U. Eriksson, and P. A. Peterson, *EMBO J.* **7,** 1451 (1984).

[29] I. Sinning, G. J. Kleywegt, S. W. Cowan, P. Reinemer, H. W. Dirr, R. Huber, G. L. Gilliland, R. AN. Armstrong, X. Ji, P. G. Board, B. Oln, B. Mannervik, and T. A. Jones, *J. Mol. Biol.* **232,** 192 (1993).