Between objectivity and subjectivity

Carl-Ivar Brändén and T. Alwyn Jones

Protein crystallography is an exacting trade, and the results may contain errors that are difficult to identify. It is the crystallographer's responsibility to make sure that incorrect protein structures do not reach the literature.

STRUCTURAL knowledge of biological macromolecules is essential for fully understanding their function as well as for changing that function by protein engineering or drug design. There are two powerful, complementary methods to obtain the structure of large molecules — X-ray crystallography and two-dimensional nuclear magnetic resonance spectroscopy. For good reasons, the biological community has regarded X-ray structures as gospels of truth because X-ray diffraction data in principle contain sufficient information to extract the correct structure. This faith has been shaken by studies[1-5] showing that there are serious errors in several recently published X-ray structures[6-10]; among the structures concerned are an electron-transport molecule, ferredoxin[1,6], the glycolytic enzyme enolase[2,7], the small subunit of the CO_2-fixing enzyme Rubisco[3,8] (ribulose-1,5-bisphosphate carboxylase), the HIV-proteinase[4,9], and the Ha-*ras* oncogene product p21 (refs 5, 10). In view of the rapid evolution of macromolecular crystallography, one may well wonder why this situation has arisen. Here we explain some of the problems that crystallographers face in interpreting their data, and argue for the need to devise better checks on their results.

Technical developments. In the past ten years there has been a minor revolution in structural molecular biology. Developments in gene technology now allow expression of the large amounts of protein (what the cloner likes to call 'gene product') usually needed for structural work. This improves the chance of obtaining crystals suitable for structure determination simply by allowing us to try a much larger variety of growth conditions. Crystallization robots are now used in a number of laboratories in the hope of speeding up this process. Other technical advances have had an even more profound influence. Collection of diffraction data has been revolutionized by new X-ray detectors and X-ray sources. Electronic area detectors of various types allow the rapid collection of diffraction data in a form immediately suited for computer processing. High-energy electron (or positron) synchrotrons and storage rings produce very intense X-ray radiation, thereby allowing the rapid collection of data and increasing crystal lifetime. Bombarding a crystal with X-rays always destroys the crystallinity of the sample. But because the process is time-dependent, short and intense exposure gives better results (and also allows the use of smaller crystals).

Many aspects of treating diffraction data before arriving at the 'final' model are computer intensive, and crystallographers have benefited enormously from the 'silicon revolution' in which the price: performance ratio of computers is halved every 18 months or so. Developments in computer graphics have been almost as impressive.

Together, these improvements have allowed us to take on and solve many incredibly complex structural problems. But at a price.

The crystallographic problem. The basic difficulty of protein crystallography remains unchanged — the problem of phases. In diffraction experiments we are able to collect only part of the data needed to reconstruct the repeating unit within the crystal. Diffraction data can be considered as complex numbers where we can measure only the amplitude but not the phase of each number. In small-molecule crystallography this problem has been solved by so-called direct methods (recognized by the award of a Nobel prize in chemistry to Karle and Hauptman). Where brains fail, protein crystallographers have stayed at the laboratory bench using techniques pioneered by Perutz, Kendrew and co-workers to circumvent the phase problem. This method, called multiple isomorphous replacement (MIR), requires the introduction of new X-ray scatterers into the crystal unit cell. These additions should be heavy atoms (so that their presence can be felt in the diffraction process); there should not be too many of them (so that we can find them); and they should not change the structure of the molecule or of the crystal cell. This final requirement is often incompatible with the others.

If a number of heavy-atom derivatives can be obtained, it is then possible to calculate an electron-density map showing the repeating unit of the crystal. The map is a three-dimensional set of numbers that has to be interpreted as a polypeptide chain of a particular amino-acid sequence. But although it is compiled from objective information, the map contains numerous errors; its interpretation remains to some extent subjective. It should be emphasized that there is a fine line between an uninterpretable map and one from which a reasonable model can be built. Errors in the model introduced by the crystallographer's interpretation depend on the map's quality, which in turn depends on several factors.

(1) Accuracy of the diffraction measurements — small differences between large numbers are used to obtain the phases.

(2) The number and usefulness of the heavy-atom derivatives — metal atoms that bind isomorphously to a few different sites are ideal.

(3) Resolution of the diffraction data, which refers to how well ordered the crystals are and directly influences the image that can be produced. Native crystals (that is, those with no heavy atoms added) often have better quality diffraction than the derivatives, limiting the quality of the initial map but providing hope for ultimately determining a high-quality model. At 4 Å resolution there is little side-chain detail. At 3 Å it should be possible to decide between the density for an alanine side chain and a leucine; at 2 Å, between a leucine and an isoleucine; and at 1 Å one sees atoms as balls of density (Fig. 1; see over). As far as we are aware, the lowest resolution at which an essentially correct model has been built is for satellite tobacco necrosis virus[11] at ~3.7Å. In this study use was made of the high symmetry of the virus capsid to produce phases with essentially no errors. This is an unusual case and normally phase errors are much larger for the higher resolution data. In normal studies we aim for an MIR map based on phases determined to ~3 Å resolution or better.

Building a model. Building an initial model is a two-stage process. First, one has to decide how the polypeptide chain weaves its way through the electron-density map. The resulting chain trace constitutes a hypothesis, where one tries to match the density to the sequence of the polypeptide. An initial model is then built to give the best fit of the atoms to electron density. Both processes use computer graphics to present the data and manipulate the model building.

Tracing the chain has yet to be automated and is often one of the most difficult steps in solving a new structure. In practice the process requires recognizing some part of the sequence in the map by the

© 1990 Nature Publishing Group

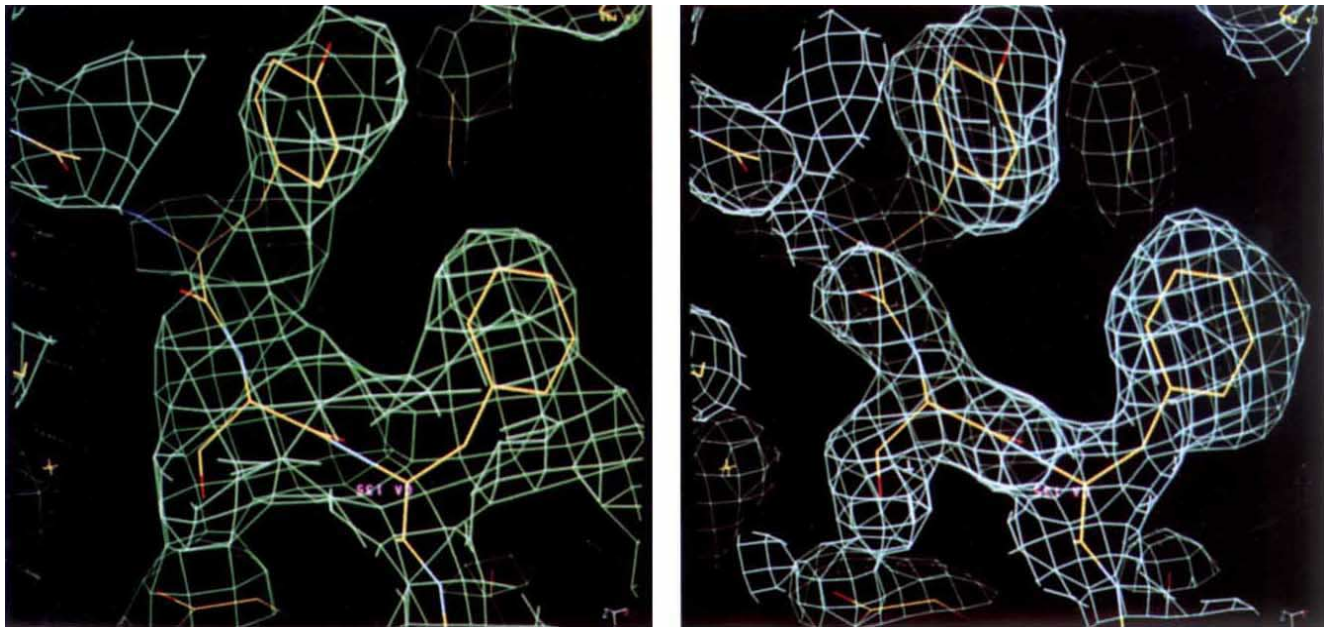

FIG. 1 Views of part of the electron density for the retinol-binding protein with superimposed final model (left) after model refinement to 2.0 Å resolution, and (right) from the MIR map at 3.1 Å resolution.

shape of the electron density. One then tries to extend this alignment in both directions in the sequence until the quality of the map or the matching to the sequence breaks down. This sounds easy but it is not; a map showing continuous density from N terminus to C terminus is rare. More usually one produces a number of matches of discontinuous regions of the sequence that may initially account for a small fraction of the molecule and may be internally inconsistent. For example, two hypothetical segments may come together representing entirely different parts of the sequence and even be in different directions. Chain direction can also be determined by recognizing bumps in the density representing carbonyl oxygens, but this requires high-resolution data.

At this stage one needs an overview of how the chain trace is developing, and also a detailed view to identify where we are in the sequence. Three-dimensional contoured nets[12] are used to portray a detailed view of the density. An overview can be produced with a skeletonized representation of the map[13] (analogous to a thread of string passing through the main features of the density), which is easy to change by removing incorrect connections or adding new ones and which can be coloured to illustrate the current state of the chain trace[14].

At any stage in this process, the chain trace has to relate to what is known about protein structure. The α-helices and β-strands are easily recognizable because of their distinct pattern of adjacent side chains. α-helices are invariably right-handed; it was the presence of left-handed α-helical features in the electron-density map of ferredoxin that provided the essential clue[1] for the cause of error[6] in that structure.

The chain trace must also relate to chemical and physical information about the protein. Knowing the positions of the heavy atoms used to determine the structure is helpful. These metal ions should have proper protein ligands[15] — mercury atoms usually bind to cysteine residues, platinum atoms to methionine residues and the uranyl group to oxygen ligands. If endogenous metal ions are present in the protein they should also have proper protein ligands which are usually conserved in homologous proteins from other species. Important active-site residues should also be conserved. Deletions and insertions within a family of related proteins should occur in loop regions and not in the middle of α-helices or β-strands.

Correct connections of the α-helices and β-strands are usually the most difficult parts of map interpretation. The loop regions that connect these secondary-structure elements are on the outside of the molecule and are frequently more flexible; in consequence their electron densities are weaker and more distorted. Errors in such connectivity occurred in the incorrect solution of the structures of both Rubisco[8] and p21 (ref. 10).

The α-helices and β-strands combine to form domain structures with specific folding patterns, some of which occur more frequently than others. So it can be helpful in map interpretation to see whether simple reconnections result in a known, common fold.

One of the most frequently observed domain structures is the α/β barrel which comprises eight parallel β-strands arranged like the staves of a barrel and surrounded by eight α-helices. In the early days of protein crystallography, Jane Richardson[16] suggested that the preliminary model of the enzyme KDPG-aldolase[17] could be reconnected to produce such a regular α/β barrel. High-resolution work confirmed this proposal and the structure was corrected[18]. On the other hand, it can be very dangerous to force a model to conform to a known domain structure. The enolase structure was first interpreted as a regular α/β barrel[7], but further refinement established that some connections were wrong[2] and that one of the β-strands was in fact antiparallel to the other seven strands.

Having obtained a full trace (or perhaps a partial trace to check how sure one really is) various methods have been developed to build a model. We believe the use of databases offers a number of advantages. Fragments from a database of highly refined structures can be extracted to best fit the skeleton[14]. This forces the use of stereochemically correct pieces of main chain[14]. Most of the amino acids adopt only a few preferred side-chain conformations, called rotamers[19]. For example, 67 per cent of all valine residues adopt a single conformation. We therefore build our initial models fully automatically from the Cα trace using the main-chain and side-chain databases.

Removal of errors in the initial model. The final model coming from the MIR map will contain errors. Provided the native crystals diffract to high enough resolution, most of them can be removed by crystallographic refinement of the model. In this process, the model is changed to improve the fit of the structure factor amplitudes calculated from the model ($|F_c|$) with the observed amplitudes ($|F_o|$). This process is computationally

© 1990 Nature Publishing Group

intensive. The goodness of fit is expressed as the R factor ($R=\Sigma||F_o|-|F_c||/\Sigma|F_o|$) where the sum is taken over the diffraction measurements. The results are then reviewed at the graphics terminal, and the model changed accordingly. The model is then subjected to continued refinement until no further improvement is obtained. The refinement techniques have been based on least squares methods[20], but recently methods based on molecular dynamics algorithms[21] have become widely used. These are even more computationally expensive.

The errors vary in severity:

(1) The model may be totally wrong (although that is unlikely if one has a sequence) because the whole chain trace is incorrect; individual secondary-structure elements may be correctly identified, but their directions may be wrong. Such a trace would inevitably contain incorrect connections between secondary-structure elements. The error may also be due to a more basic mistake such as wrong space group[6].

(2) In multi-subunit molecules, one subunit could be totally wrong[8]. Such a structure can only be corrected by returning to the initial MIR map. In this case, combining the MIR phase information with phases calculated from the correct part of the model may provide the basis for a better map.

(3) Main-chain connectivity may be partly wrong[7,9,10,22], which is the most common serious error and is usually the result of incorrect connections between secondary-structure elements. Such errors can sometimes be identified during refinement of high-resolution data and be corrected. The refinement of the incorrect enolase model stalled at an R value of 26.6 per cent for data to 2.25-Å resolution. The model was then revised and the correct structure easily refined to an R factor of 20.2 per cent[2]. In p21 and HIV protease, connectivity errors were not recognized during refinement and became apparent only when independent structure determinations were carried out. In both these structures, the incorrect models could be refined to R factors of ~25 per cent.

(4) Sequence out of register with the density over a small part of the structure. This is very common and should be correctable during refinement.

(5) Locally misbehaving model — some parts of the main chain may just be badly built and be beyond the radius of convergence of the refinement.

(6) Incorrect side-chain conformation. Once the major flaws have been corrected, these errors may be easier to spot in the new maps. If the correction is outside the radius of convergence of the refinement algorithm, the model may still be pushed into some distorted conformation mimicking the correct one. At each refitting stage we compare side-chain conformations with the rotamer database[19].

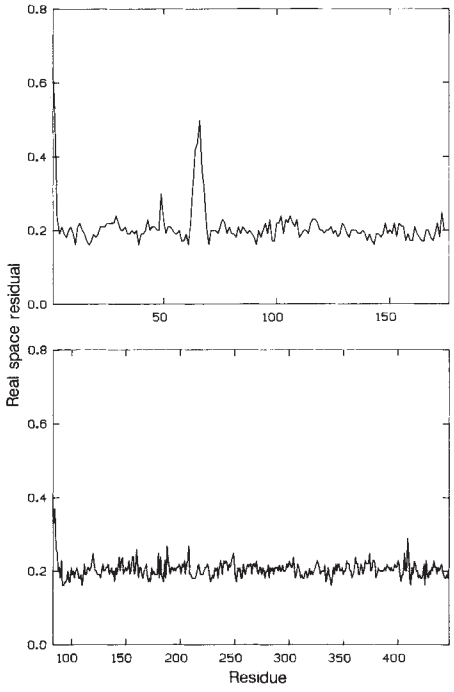

FIG. 2 The main-chain real-space R factor for (top) retinol-binding protein; and (bottom) cellobiohydrolase II. For the main-chain atoms of each residue the function calculates $\Sigma|\varrho_o-\varrho_c|/\Sigma|\varrho_o+\varrho_c|$ on the grid of the observed density (ϱ_o), for all grid points contained in the map calculated from the atoms (ϱ_c). This function can define the fit of the whole residue, the main chain or the side chain. It can unambiguously show the problem areas in a model. The top figure shows clearly a localized, poorly fitting section of the chain. The proteins have crystallographic R factors to 2.0 Å resolution of 18 and 15 per cent, respectively.

(7) Main-chain conformational errors (peptide flips). Although Cα atoms of neighbouring residues may be correctly positioned, the carbonyl oxygen of the linking peptide may be pointing in the wrong direction. This should be corrected during the refinement. We have found that using the database to build models reduces this problem. At each refitting stage we check the deviation of carbonyl oxygens from those of the best-fitting fragments in the database[23].

Refinement is a gradual process; as errors are located and corrected it becomes easier to correct remaining errors.

Assessment of the final model. There are a number of general points concerning assessment of reliability of a structure determination.

(1) The better the MIR map, the better the initial model, the easier is the refinement and the more reliable is the structure.

(2) The higher the resolution the better. Unfortunately, not all crystals diffract as well as we would like.

(3) The quoted R factor has to be viewed carefully for a number of reasons. It is affected by the choice of low-resolution cutoff and by the removal of weak reflections. It is also insensitive to errors in main-chain connectivity. Alarm bells should ring if the refinement stalls at an R value of about 25 per cent for high-resolution data. We advocate the use of a real-space R factor[23] that shows, on a residue-by-residue basis, how good the fit is between the model and the map (Fig. 2).

(4) The final model should be tightly restrained to standard bond lengths, bond angles and fixed dihedral angles (reasonable values are 0.02 Å, 2° and 2°, although each refinement program has its own fingerprint).

(5) Main-chain dihedral angles should conform to the 'allowed' regions of the Ramachandran plot.

(6) The number of water molecules added during refinement should be realistic. For high-resolution structures a reasonable value is one water molecule for each residue, and these molecules should have plausible hydrogen bonds.

Protein crystallography is a highly competitive field where the same or similar structures are being worked on in a number of different laboratories. Although this may result in an urge to publish quickly and prematurely, it is the responsibility of crystallographers to check their preliminary models carefully before rushing into print. Journals must insist on the publication of enough data for crystallographers to convince the reader that they have a correct structure, and the readership should be sophisticated enough to judge the quality of the data. We strongly object to publication of structural work where authors supply a minimum amount of detail in the form of a cartoon. □

Carl-Ivar Brändén and T. Alwyn Jones are in the Department of Molecular Biology, Uppsala Biomedical Center, PO Box 590, S-751 24 Uppsala, Sweden.

1. Stout, G.H. *et al. Proc. natn. Acad. Sci. U.S.A.* **85**, 1020–1022 (1988).
2. Lebioda, L., Stec, B. & Brewer, J.M. *J. biol. Chem.* **264**, 3685–3693 (1989).
3. Knight, S., Andersson, I. & Brändén, C.-I. *Science* **244**, 702–705 (1989).
4. Wlodawer, A. *et al. Science* **245**, 616–621 (1989).
5. Pai, E.F. *et al. Nature* **341**, 209–214 (1989).
6. Ghosh, D. *et al. J. molec. Biol.* **158**, 73–109 (1982).
7. Lebioda, L. & Stec, B. *Nature* **333**, 683–686 (1988).
8. Chapman, M.S. *et al. Science* **241**, 71–74 (1988).
9. Navia, M.A. *et al. Nature* **337**, 615–620 (1989).
10. deVos, A.M. *et al. Science* **239**, 888–893 (1988).
11. Lijas, L. *et al. J. molec. Biol.* **159**, 93 – 108 (1982).
12. Jones, T.A. *J. appl. Crystallogr.* **11**, 268–272 (1978).
13. Greer, J. *J. molec. Biol.* **82**, 279–288 (1974).
14. Jones, T.A. & Thirup, S. *EMBO J.* **5**, 819–822 (1986).
15. Blundell, T.L. & Johnsson, L.N. *Protein Crystallography* (Academic, New York, 1976).
16. Richardson, J.S. *Biochem. biophys. Res. Comm.* **90**, 285–290 (1979).
17. Mavridis, I.M. & Tulinsky, A. *Biochemistry* **15**, 4410–4417 (1976).
18. Mavridis, I.M. *et al. J. molec. Biol.* **162**, 419–444 (1982).
19. Ponder, J.W. & Richards, F.J. *J. molec. Biol.* **193**, 775–792 (1987).
20. Hendrickson, W.A. *Meth. Enzym.* **115**, 252–270 (1985).
21. Brünger, T.A. *et al. Science* **235**, 458–460 (1987).
22. Tong, L. *et al. Science* **245**, 244 (1989).
23. Jones, T.A. *et al. Acta Cryst.* (in the press).

© 1990 Nature Publishing Group