added to the set of independent variables, the atomic coordinates. To be precise, one may minimize

\[ F = F_0 + q_i^2 WF_i \]  

(36)

with \( W \) a constant, and \( q_i \), a variable; selection of a small value of \( W \) serves to retard the decrease of \( q_i \) during minimization.

Evaluation

In this chapter two methods have been described by which one can obtain a model of perfect geometry that is an optimal fit to a set of experimental or target coordinates. The first method uses a model of perfect geometry that is adjusted by internal rotation about single valence bonds until the fit is optimal. In the second method, the local geometry of the experimental model is made to gradually approach perfection, as a result of which the model coordinates move away minimally from the initial target coordinates. Use of the second method is preferable for several important reasons.

Speed. For a large molecule, minimization of geometric error by the second method proceeds much more rapidly than does optimization of the fit of two sets of coordinates by the first method. The principal reason for this is that the length of the calculation of the derivatives of the geometric error with respect to the atomic coordinates is linear in the number of atoms, \( N \), while evaluation of the derivatives of the fit with respect to the variable dihedral angles requires a number of steps that varies as the square of \( N \). A second reason is an apparent difference of the ease with which at least one minimization routine, the conjugate gradient program, performs the two minimizations.

Only under special circumstances is it preferable to optimize the fit by internal rotation about single bonds. This is the case if the bonds about which rotation is possible are few in number, and the atoms are many, or if particularly large atomic displacements are required in order to achieve the fit.

Accuracy of ring closure. We have discussed the difficulties of maintaining proper ring geometry that arise when a model of rigid geometry is used. In contrast, when geometry is improved by local adjustment of a flexible model, both inherent geometric strain of small rings and geometry of large rings are appropriately treated, without need of any special measures.

Size. Because routines, such as the conjugate gradient minimizer, require storage for derivatives and auxiliary arrays, minimal program size is obtained with use of the method of Hermans and McQueen,\(^3\) in which only one atom is moved in each step.

Convenience. When program size is not a consideration, model rationalization is most conveniently performed by simultaneous adjustment of the coordinates of all atoms in each step of a minimization of geometric strain with use of the conjugate gradient minimizer.\(^{11,14,15}\) The constraints may be defined either in terms of bond lengths, bond angles, and dihedral angles,\(^2\) or in terms of interatomic distances.\(^{14,15}\) In the latter case, special constraints must be provided that will maintain planarity and chirality.

Acknowledgments

This work was supported by the National Science Foundation, Grant PCM81-12234, Computations done with support by the National Institutes of Health, Division of Research Resources, Grant RR-00898.

[14] Interpretation of Electron Density Maps

By Jane S. Richardson and David C. Richardson

In the best of cases no real "interpretation" of an electron density map is necessary—a well-phased 2 Å protein map is a joy to work with and will show what the appearance of various side chains and conformations should be. However, maps are unfortunately the exception; therefore the following section offers indications of what to expect and strategies for interpretation of protein maps at lower resolution or with less accurate phasing.

Low Resolution (5–6 Å)

The resolution level around 5–6 Å is one of the traditional milestones in the solution of a protein structure: it is a low point in most radial distributions of diffraction intensity (see Fig. 1), it is sufficient for location of heavy-atom positions, and it provides information about some overall features of the structure. The simplest information commonly obtained at low resolution is the distinction between protein and solvent density, which shows the molecular shape and packing. Subunit arrangement and symmetry can be seen, as can domains within a subunit if they are well separated. The ease and certainty with which molecular boundaries can be drawn at low resolution vary with crystal packing density (tight packing produces ambiguous contacts), crystallization medium (contrast is
Fig. 1. Radial distribution (as a function of resolution) of diffracted intensity ($F$) for a representative crystal from each of the four major classes of protein tertiary structure. The $\alpha$ and $\alpha/\beta$ proteins have the highest peaks at 10 Å resolution (mainly from helix–helix contacts) and the antiparallel $\beta$ protein the lowest, while the small protein has the most relative intensity at high resolution. All types have a trough at about 6 Å and a peak at about 4.5 Å.

better in alcohols or PEG than with concentrated salts), and accuracy of phasing (which can be improved, for instance, by averaging over noncrystallographic symmetry). Figure 2 is a sample slice through a low-resolution map showing the boundary between protein and solvent.

Within the molecule, at low resolution the density is being averaged over a volume greater than that of an individual side chain or backbone. The result is that most side chains disappear altogether (below average solvent density), especially the hydrophobic ones which contain little N or O and a rather high proportion of low-density hydrogens. Large, apparently empty, regions inside the protein as much as 10–15 Å across (as in Fig. 4) usually represent well-packed hydrophobic side chains rather than solvent-accessible holes. Sometimes small peaks appear at the positions of denser side chains such as Cys or Trp, but they are not usually enough

Fig. 2. A slice through the low-resolution (5.5 Å) map of Cu,Zn superoxide dismutase, with dense regions for a pair of subunits in the upper left and lower right of the outlined asymmetric unit, and solvent regions at low density with small, isolated peaks.
above the random noise level to allow interpretation. For the backbone, at
this resolution there is a critical distinction between the main chains within
H-bonding distance of one another (such as within helix or between β-
strands), which show up as strong positive features, and places where the
backbone is separated by side-chain packing or even by van der Waals
contact of backbone atoms, in which case the average density is low
enough that such “isolated” main chains often disappear.

The classic low-resolution feature is the α-helix, which appears as a
high-density cylindrical rod 4–6 Å in diameter (depending, of course, on
count level). Typical spacing between helices is 10 or 12 Å. Figure 31
shows a hemoglobin subunit at 6 Å resolution. The globin fold can easily
be followed and the heme located, but in the absence of such outside
information there are almost always ambiguities that prevent chain trac-
ing at low resolution even in all-helical structures. Connections from one
helix to the next are often weak or missing because of series termination
effects (which produce a negative ripple in the map around the strong
positive rod of helix density); chain termini are very difficult to identify;
and there are often extra bridges of density through such features as
hemes or metals, groupings of heavy side chains, or even unusually close
helix contacts. Myohemerythrin1b is one simple case in which an unam-
biguous and correct chain tracing was possible at low resolution for a

---

1 E. D. Getzoff, J. A. Tainer, J. S. Sack, D. Bickar, J. S. Richardson, and D. C. Richard-
72, 2160 (1975).

---

novel protein structure (the first four-helix cluster). The number and
placement of helices seen at low resolution are sometimes diagnostic of
the general category of tertiary structure (see the “mini-atlas” of illus-
trations in [23], this volume), as is the case for the domain of glyco-
late oxidase that has been identified as having a triose-P-isomerase-type barrel
mainly on the evidence of the outer ring of eight helices.2

A region of β-sheet can often be recognized at low resolution, as a
smooth isotropic sheet of somewhat lower density than in a helical rod
(see Figs. 4 and 13). The sheet looks essentially identical along the strands
and across them, so that there must be additional information in order to
tell which way the strands run. This can be done in a barrel, or in a
nucleotide-binding domain, by matching the overall twist and placement
of features with those in known structures. If the sheet density has a clear
twist and definite corners, then strand direction can be determined: if
upper left and lower right corners twist upward toward you (as in Fig. 5),
then the strands run vertically. Also, if one opposite pair of sheet edges
looks smooth and parallel while the other pair is jagged or far from parallel,
then the smooth pair probably follow the edge strands of the sheet. If
such a pair can be identified, then the number of β-strands (n) in the sheet
can be estimated from the distance d along a curving line (within the sheet
density) perpendicular to those two edges: $d = 4.8(n - 1) + 3$ Å
(approximate) (see Fig. 5). Spacing can vary slightly, but probably not

enough to give the wrong integer for \( n \). If a \( \beta \)-sheet has helical or irregular density next to both its faces, it is almost certainly predominantly parallel, while if it is exposed to solvent on one face, then it is almost surely predominantly antiparallel.

At 5–6 Å resolution, an isotropic sheet of density represents that portion of the \( \beta \)-sheet which has regular hydrogen bonding. An irregularity in the middle of the \( \beta \)-sheet produces a hole in the sheet of density, and where the \( H \) bonding stops between neighboring strands at the end of the \( \beta \)-sheet, the continuity of electron density is broken along the strand direction as well as between strands.

Medium Resolution (2.5–3.5 Å)

The region around 4–4.5 Å resolution is not discussed specifically here because it should generally be avoided for protein maps (although not necessarily for nucleic acids). This region is near a maximum in the radial intensity distribution (see Fig. 1) so that it produces particularly severe series termination errors, and the resulting maps are exceptionally difficult to interpret correctly. Helices are rather undistinguished wiggles rather than spirals (see Fig. 11c), and historically have quite frequently been overlooked in this resolution range. If structure factors are for some reason available to 4 Å resolution, it is probably worth calculating and examining an electron density map at 5 or 6 Å as well, as a check on the location of secondary structure features.

Somewhere in the 2.5–3.5 Å resolution range there is usually a point at which a correct overall chain tracing is first possible. That point depends strongly on whether or not the amino acid sequence is known, because the sequence allows one to resolve a limited number of ambiguities. At better than 2.5 Å resolution the sequence is not necessary for correct chain tracing (although of course it greatly enhances the usefulness of the result), while at something like 3.5 Å the sequence information ceases to help because so few of the side-chain shapes can be reliably recognized in the map. A plausible chain tracing must of course use all parts of the molecular density once and only once, and not make closed loops or tie knots. In order to keep track of such criteria and to see continuity of secondary structures, it is necessary to work with a fairly large three-dimensional portion of the map at once. Initial chain tracings are traditionally done on "minimaps" (stacked plastic sheets at small scale) rather than on computer graphics. However, an interactive graphics system based on ridgeline representation of density rather than contours promises to allow large-scale pattern recognition, with real-time change of scale and contour level (see Fig. 6). The risk of making an incorrect choice somewhere is of course less for small structures with fewer choice points, and is usually less for a structure with clear analogy to one of the

---

recognized common folding patterns (see the "mini-atlas," this volume [23]).

In the intermediate resolution range it is crucial that the sequence be actively used if it is available. One simple way to do so is to position independently within the sequence a number of the clearest stretches in the map. For a run of at least six or seven residues, preferably in either a helix or an extended strand (where one can be sure of not inserting or omitting a residue), the shape and location of each side chain in the map are evaluated and a probability assigned (say, on a scale of 1 to 5) for that density representing each of the 20 side-chain types. A well-formed tryptophan shape, for instance, might rate 5 for Trp, 2 or 3 for Tyr or Phe, and only 1 for anything else. An apparent alanine on the surface might rate 5 for Ala, 4 for Ser, 3 for Lys (since it could be disordered, as explained below), etc., but on the inside of the protein an alanine shape would rate only 1 for Lys. For the branched side chains one needs to evaluate whether the branch looks flat or tetrahedral, how far out it is from the Cα, and the hydrophobicity of the environment. If carbonyl oxygens are visible a probability can be estimated for which direction the chain is going, perhaps nearing certainty for a clear α-helix. Then a simple computer program can evaluate the overall probability (by multiplying all relevant numbers) that this stretch in the map represents any given stretch in the sequence. A run of six or seven residues will normally fit at only one place in the sequence, and a possible chain tracing which is tied down to the sequence on each side of each ambiguity is quite sure to be correct.

The possible sources of ambiguity at medium resolution include those mentioned below for high-resolution maps, plus several more. A disulfide (Fig. 7) or a salt link (especially an Arg-Asp or Arg-Glu with a double H bond) can look just as dense and continuous as the polypeptide backbone, and connectivity may also be confusing around metal sites. It is not always easy to distinguish the peaks or side branches in the density that are caused by the side chains from those caused either by carbonyl oxygens or by close solvent molecules. Even α-helical regions can be confusing because the density at this resolution is sometimes continuous along the hydrogen bonds as well as, or instead of, along the main chain (see Fig. 12).

At medium resolution, perhaps even if the phases are very good, there are almost always a few deep breaks in the density along the backbone. These occur most often at either the ends of loops or turns at the surface, or in β-sheets at points of local irregularities in the H bonding or strand spacing (Fig. 13). Unfortunately, for such regions in a β-sheet it is apparently common for the density to be continuous perpendicular to the chains so that the overall connectivity of the density is incorrect. This has

FIG. 7. A slice through the 3 Å map of Cu,Zn superoxide dismutase, contoured on glass sheets with some labels and triangular beads marking the α-carbons. Density for the disulfide, which bridges between a loop at the top and the central β-barrel, is approximately as heavy as the main-chain density.
led to many of the revisions that have been made in chain tracings.\textsuperscript{4-11} The chances of being correct can be improved by avoiding features such as shown on the left in Fig. 8 and also by avoiding left-handed crossover connections\textsuperscript{12} and topologies that skip back and forth frequently.\textsuperscript{12} However, even the unusual forms occur sometimes in well-established protein structures, and some map ambiguities will give plausible arrangements for either choice. Therefore, it seems wise to treat any chain tracing as tentative if it was made at much less than 3 Å resolution with a sequence or at less than about 2.5 Å without a sequence. The overall shape and organization of the structure will still be correct, as well as most of the secondary structure assignments.

In helices and β-strands, the distortions common at medium resolution almost invariably lead to underestimation of secondary structures. Therefore it is appropriate in interpreting such initial maps to be prejudiced in the direction of fitting repetitive structures wherever possible, and always trying as one alternative tracing to continue a regular structure through apparent local breaks.

Medium resolution maps contain many potential pitfalls, but they are also of great importance because (especially if the sequence is known) they can provide chain tracings and initial model fittings for new protein structures.

**High Resolution** (2–2.3 Å, or better)

At high resolution the chain tracing is normally unambiguous even if the amino acid sequence is unknown. Phase errors produce some breaks in the continuity of the main chain, but such breaks are usually quite local and the correct backbone positioning can be deduced from the shape of the density on either side. Many side chains can be identified from their characteristic shapes, of which a few examples are shown in Fig. 9. Six-membered rings can be oriented quite accurately at high resolution, and even the rounder, fatter histidine rings usually show some flattening.

A resolution of 2 Å is traditionally the highest resolution to which data is collected for heavy-atom derivatives, so that it is normally the highest resolution at which initial interpretation of electron density maps is done.

---


\textsuperscript{10} I. M. Mavridis, and A. Tulinsky, *Biochemistry* 15, 4410 (1976).


---

Fig. 9. Appearance of side chains for an Arg and a Phe in the original MIR map of staphylococcal nuclease at 2 Å resolution, contoured on stacked glass sheets and with brass Kendrew models lying on top of the stack to show the interpretation. The Phe ring has a slight dimple, and the hammerhead shape of the guanidinium can be seen to hydrogen bond both to a main-chain carbonyl and to a dense, triangular shape that is a phosphate of the bound inhibitor.\textsuperscript{13}
One exception is when resolved anomalous phasing is used, which may become commoner for small proteins and perhaps with synchrotron data, both of which encourage high-resolution data collection. Initial interpretation at, for example, 1.5 Å resolution has real advantages, since in good parts of the map it allows recognition of resolved atoms in such features as aromatic side chains or proline rings.

At 2–2.5 Å resolution individual atoms do not show separate peaks, but they produce clear effects on the shape of the density contours. Probably the most important atom type for accurate fitting is the carbonyl oxygen of the main chain: its orientation directly determines the $\phi$ and $\psi$ angles on either side, and in a well-phased map at high resolution most of the carbonyl oxygens should be visible as distinct bumps. In an $\alpha$-helix, their orientation immediately determines the direction of the polypeptide chain, since all the carbonyl oxygens point toward the COOH-terminal end of the helix. They can sometimes be used to determine direction in other parts of the chain by measuring which $\alpha$-carbon they are closest to. Their position is the major feature which allows assignment of hydrogen bonding, either directly for the O end of a bond or indirectly by determining the orientation of the peptide plane and therefore the NH position.

A specific example of the usefulness of carbonyl oxygens in map interpretation is the assignment of types of tight turns. The types are generally defined in terms of $\phi$ and $\psi$ angles for the two central residues, exact values for which cannot usually be obtained until later. However, turn types I and II and their mirror images I' and II' can be assigned from carbonyl oxygen orientations CO(1) (in the H bond) and CO(2) (in the central peptide) (see Fig. 31 of Ref. 14). Those characteristic orientations are summarized in the table, along with the dihedral angle formed by the four $\alpha$-carbons, which provides an additional check that can be seen in the map. Although there is considerable variation in the exact CO position within each turn type, genuinely intermediate conformations are quite strongly disfavored: for example, it is almost impossible to form a type I turn with a flat dihedral angle and a straight hydrogen bond. Type III turns differ from type I by only $30^\circ$ in one conformational angle, which cannot be distinguished at this initial level unless there is a second overlapping turn making it into a short piece of $3_{10}$ helix. If the carbonyl oxygens are not visible in the map, then it is possible to identify the region as a turn but not to assign the type.

Although, in general, fitting can proceed quite easily at high resolution, there are still some problems, mostly caused by either inaccurate phasing or actual disorder in the structure. Sometimes a stretch of several residues, including the backbone, can be partially or completely disordered if there are no hydrophobic side chains near a chain terminus or in a loop at the surface (e.g., Ref. 15). These regions either disappear altogether or appear at a low density in the map, and can only be fit approximately if at all. They will usually show up at least a bit more clearly after refinement of the rest of the structure.

Individual side chains at the molecular surface frequently show disorder. About three-fourths of the lysines and one-fourth to one-third of the

---

other charged side chains are disordered,\textsuperscript{14} most often indicated by density visible only out to $C_β$ or $C_γ$, but sometimes by having an unconnected peak for the end of the side chain. This disorder may often represent the superposition of two discrete, definite conformations,\textsuperscript{13} but that fact can only occasionally be seen in the initial map.

Either phase errors or series termination effects can distort the local appearance of the map. At high resolution the most susceptible place seems to be the $β$-carbon. Instead of showing the "elbow" that should result from the tetrahedral angle at $C_β$, it is not uncommon to have the density there rather weak and extending in a straight line from $C_α$ to $C_γ$, which usually makes the $χ_1$ and $χ_2$ angles ambiguous. With a branched $C_β$, especially for valine, the map may show a straight bar of density joining the $C_γ$ atoms, extended more or less parallel to the backbone but with no indication to which side the $C_β$ should protrude. Such situations illustrate a basic conflict in map interpretation: for instance, valine has a very strong preference for the staggered conformation of $χ_1$,\textsuperscript{16} and choosing that alternative in ambiguous cases greatly improves one's chances of being correct; however, that undermines later attempts to get unbiased statistics for $χ_1$ frequencies. Perhaps the best strategy for such problems is to use the expected value in initial fittings but make note of the ambiguity as one that should be explicitly checked later during refinement.

Some particular residues are especially prone to fitting difficulties. It is hard to determine the correct $ϕ$ and $ψ$ angles for glycine, since there is no $C_β$ to give a "lever arm" on what are otherwise quite subtle differences in shape along the backbone. For instance, it is quite possible for a glycine in the center of a $β$-sheet to be in left-handed (i.e., with $ϕ$ positive and $ψ$ less than $−40°$) rather than in normal $β$ conformation, although with good hydrogen bonds on both sides and only a change in the "pleat" of the $C_α$ position\textsuperscript{17} (Fig. 10). Disulfides are difficult to fit because they must simultaneously satisfy so many nonorthogonal constraints and also because they are especially likely to show indecisive directionality at $C_β$; again,


Fig. 11. Stereo views of the electron density for an $α$-helix in staphylococcal nuclease\textsuperscript{15} at (from top to bottom) 2, 3, 4, 5, and 6 Å resolution. All maps were made using $F_{obs}$, the original MIR phases, and the same grid spacing. Viewpoint is the same, and contour levels were adjusted to be approximately equivalent. All carbonyl oxygens are clear at 2 Å, but almost all of them are absent at 3 Å, although side-chain shapes can still be judged. At 4 Å, density has begun to coalesce along the helix axis, and there is a false connection between side chains at the lower left. Figures 3, 11, 12, and 13 were made using the GRIP molecular graphics system developed by Wright, Pique, Britton, Lipscomb, Brooks, and others at the University of North Carolina, Chapel Hill.
Fig. 12. Side views of the same helix as in Fig. 11, at 2, 3, and 3.5 Å resolution. At intermediate resolution the density connects through a hydrogen bond (lower right) more strongly than through the nearby helical main chain, although the connectivity is correct at both higher and lower resolutions.

Fig. 13. Stereo views of the electron density for two strands of antiparallel β-sheet in staphylococcal nuclease at 2, 3, 4, 5, and 6 Å resolution. In this case the strands separate correctly at 4 Å, but that would not always be true. At 5 and 6 Å the density is sheetlike, but with holes in variable locations. At 6 Å the right-hand side extends further out because it is no longer separated from a third strand.
expected conformations can be chosen in initial fitting and reexamined later. For prolines, one should keep in mind that about 15% of them have cis rather than trans peptides (examples in Refs. 14 and 19).

However, all of the above problems are local and only sometimes present. In general, at 2–2.5 Å resolution a quite accurate fitting of the model to the electron density map can be made, even when the amino acid sequence is not known. When it is possible to collect data out to 2 Å, that resolution can provide both a quite respectable structure from initial fitting and also the most suitable starting point for refinement.

The final set of figures (11, 12, and 13) summarize this discussion by displaying the appearance of the electron density of an α-helix and of a piece of β-sheet as a function of resolution, starting from a multiple isomorphic replacement initial map of a real protein structure at 2 Å resolution. It can be seen that there are some particular stages along the way at which the density coalesces in ways that do not quite match the covalent continuity of the chain. Although one should not necessarily assume that incorrect connectivities will always occur in the same ranges of resolution that they do here, these illustrations give some concrete examples of what level of clarity to expect and what features to look for at a given resolution.

18 There are 11 cis prolines and 63 trans prolines in a sample of 12 highly refined protein structures.


By Jonathan Greer

Introduction

The past decade has seen the three-dimensional structures of proteins solved by X-ray diffraction methods at an ever increasing pace. In addition, the sizes of the molecules studied have also increased dramatically with larger proteins being analyzed all the time. This situation has placed greater pressure on the ability to interpret the electron density maps of these proteins with an atomic model that can then be used for further refinement. The introduction by Richards of the optical comparator1,2

2 F. M. Richards, this volume [10].

moved electron density map interpretation out of the dark ages into the modern era. Yet, as the molecules studied became larger and new structures more frequent, sheer size and space limitations, as well as length of time required and accuracy of model building, have increased the need for further advances in map interpretation techniques.

Two approaches were possible to achieve improvement in map interpretation methods: computer graphics-based Richards box systems and an automated approach to map interpretation. Several groups have now developed useful computer graphics map interpreting programs. These allow protein crystallographers to interpret their map with the considerable aid and accuracy of computer graphics which can be directly coupled to a larger computer for immediate coordinate refinement. Such programs are user friendly and crystallographers have rapidly adapted to their use. This chapter deals with the alternative—automated electron density map interpretation3,4 utilizing basic pattern recognition and other advances in computer science and artificial intelligence. Because this approach is less familiar to the protein crystallographer, this chapter is intended to show the crystallographer how such a system works, give examples of the results it has achieved, and suggest future directions in which it should be expanded. Thereby, the investigator community may become acquainted with the automated system and allow its use to expand.

Method of Analysis

Preparation of the Electron Density Map

The electron density map should be between 2.0 and 3.5 Å in resolution. At this resolution, continuity of the main chain should be optimal for analysis. The map is interpolated, if necessary, into a Cartesian coordinate system of definition 1 Å. The portion of the map to be examined should contain the whole of one molecule with a border of at least 2 Å around the outer edge of the molecule. Ideally, the center of gravity of the molecule should be near the center of the map space being analyzed. In order to illustrate the various steps performed on an electron density map, a simulated two-dimensional electron density map is presented and operated upon by each successive step in the process. This map is shown in Fig. 1a in a typical contour form.