Acta Cryst. (1994). D**50**, 695–708

Core Tracing: Depicting Connections Between Features in Electron Density

BY STANLEY M. SWANSON

Biographics Laboratory, Department of Biochemistry and Biophysics, Texas A & M University, College Station, Texas 77843-2128, *USA*

(*Received* 23 *November* 1992; *accepted* 1 *March* 1994)

Abstract

Core tracing is a threshold-independent method of determining connectivity (long chains of high-density values) in electron-density maps. It gives visually sparse pictures of large volumes which are useful for initial fitting and for molecular-boundary determination. New methods for visual presentation of the traces are suggested by the way that the connectivity is parameterized in terms of local connections between maxima and the saddle (lowest) points along the connecting paths. The algorithm also partitions the density into small compact volumes containing the maxima. These volumes are useful for localization and statistical analysis.

Introduction

The interpretation of electron-density maps is one of the most labor-intensive and demanding aspects of macromolecular crystallography, especially when the map is noisy or poorly phased. Core tracing addresses this problem by highlighting dominant features in the density.

The graphical presentation of electron density has traditionally been in terms of contours: by two-dimensional projections, by layers of two dimensional sections on transparent sheets, or, with the advent of computer graphics, by a wire-frame representation of iso-density surfaces formed from superimposed two-dimensional contoured layers in three directions. In the initial stages of macromolecular analysis, when molecular packing is determined, or when the initial chain tracing of a structure is performed, contours on a typical computer-graphics screen provide a too local and too cluttered view of the density.

Various skeletonization techniques [Greer (1974); Hilditch (1969); Johnson (1977, 1978); *GRINCH* (Williams, 1982; Swanson, 1979); *BONES* (Jones & Thirup, 1986)] have been proposed to give a visually economical depiction of density connectivity in large volumes. The complexity of skeletons (the number of vectors in the picture) is approximately that of the final structural model, and at least an order of magnitude less than that of contours. Skeletons are complementary to contours: they provide an overview of large volumes and, when used with contours in small volumes, an indication of the most probable connection path. Skeletons fail to depict aspects of density such as shape and bulk which contours suggest, but it is possible to encode some of these in the rendering or presentation phase.

Implicit in the idea of skeletonization is that connectivity in the density corresponds to bond chains in the structure. In contoured density of macromolecular structures at intermediate resolution, it does. However, at either very low or very high resolution, or in the presence of noise, exceptions occur. For very low resolution, one expects only to see molecular outlines, while in the case of very high resolution (attained in small-molecule studies) one resolves individual atoms and infers bonding from distance calculations. Noise, by altering density values relative to a chosen threshold, can break or add connections. For example, one may have to interpret a string of islands (a sequence of nearby but disconnected lumps) as a continuous chain.

Core tracing is a new method of density skeletonization that has been developed to address some perceived inadequacies in previous methods. Greer's method and its descendents (*BONES*) force a pre-processing decision on the lowest connection level; it is not easy to ask what other possibilitieslie just below that threshold. In contrast, core tracing uses a threshold-independent top-down scan of the density map so that all local connections can be found. The most prominent features are noted first, and the decision about viewing threshold level can be an interactive one at the display.

Greer's method forms skeletons by connecting adjacent grid points; the paths tend to have many short lines of about atomic bond length which appear only in limited orientations (along lattice coordinates or diagonals). *GRINCH* does interpolate locally (which removes the limitations on line orientation), but still uses many short lines. Moreover, the interpolation is biased toward the grid positions. In contrast, while developing core tracing, it was found that drawing paths which connect maxima to saddles gives an adequate and less busy picture with longer line segments. Interpolation to provide diversity of orientation becomes less necessary because adjacent grid points are rarely connected. (The subject

ISSN 0907-4449

of interpolation itself deserves more careful study to find optimum methods and to determine accuracy in the face of experimental error. See the *Appendix* for more comments.)

Descriptive geometry

Mathematically, an electron density is a scalar field: a real function defined on a three-dimensional domain. Experience suggests that we will find one-dimensional paths passing through high density which correspond to bonded chains of atoms in the macromolecular structure being studied. Although an experimental density is sampled on a grid and algorithms must deal with that partial information, we now consider the ideal continuous case. The gradient operator measures spatial changes. Most points will have a non-zero slope, but a few will be critical points where the gradient is zero. There are critical points in addition to maxima and minima: the saddle points with mixed partial curvatures. The number of kinds of saddle points depends on the dimension of the space; there are two different kinds in three dimensions, but only one in two dimensions. The language used previously in crystallography has a two-dimensional bias, being based on terminology borrowed from topography [peaks, ridges, passes, pales, pits (Johnson, 1978)], and really describes only planar projections of density.

To sharpen our description, consider three dimensions specifically. Density maxima are local *concentrations* of high density, or *nodules*. One-dimensional *connections* follow the path of highest density between the nodules. Such a path travels through two-dimensional maxima in planes perpendicular to its direction, but will encounter a minimum density value somewhere between two nodules. The minimum on a path would be seen as a *constriction* or *neck* in a contour representation of the density, and corresponds to one of the intermediate non-extremal critical points. Together, the nodules, the paths connecting them and the constrictions will form the focus of this paper and be called the *core* of the density. Note that our viewpoint has been from outside the density: looking at a hard object or structure in the density. The other critical points are best thought about from the inside, as though we travel through a system of caves. Minima correspond to *voids*, connections of minimum density between voids to *passages* and narrow places in passages to *portals* (the other non-extremal critical point). In Johnson's terminology, the core of the density consists of peaks, ridges, and passes; the caves are described by pales and pits.

After having discussed three dimensions in fairly picturesque language, I will also use a dimensionally more neutral terminology: *maxima* (for the nodule centers) and *joins* (for the constrictions), and *path* or *core* or *trace* for the connections. Together, maxima and joins will be called *features* of the density. For macromolecular crystallography, we can ignore the caves and focus on the core of the density.

The algorithm

Now consider the algorithm, first in qualitative terms and then in more detail. Some more technical details which define and facilitate the implementation are discussed in the *Appendix*.

Core tracing proceeds by associating successively lower *nearby* points to maxima in the density, forming distinct, local, growing nodules. Eventually these nodules will merge; the highest point at which two (or more) nodules *touch* is the *join* between them. Line segments from a join to the connected maxima represent the core of the density. The ideas of *nearby* and *touch* are defined in terms of a neighborhood of a point (other points within a specified distance). The result is a list of connected features and a partition of the density into many small, compact volumes, each identified with a feature contained within it. A two-dimensional example is given in Fig. 1.

A neighborhood is defined by a list of nearby lattice points, sorted so that the nearest ones come first. A single loop then drives the investigation of the neighborhood of a point; a re-analysis of a map with a different neighborhood is handled by a different list of neighbors.

As concrete examples of neighborhoods, consider cubic and hexagonal lattices with equal grid intervals in all directions. The 27 points which form a cube with a maximum coordinate offset of 1 grid unit from the central point separate into four distance classes: the single point at the center, the six points along the coordinate directions (distance squared = 1 grid unit), the 12 points on edges of the cube (distance squared = 2 units), and the eight points at the cube vertices (distance squared = 3 units). Although the conventional neighborhood (or shell) is all 26 surface points, one can choose fewer (18 or 6) or more merely by specifying a new defining distance limit. For a hexagonal lattice, the corresponding set contains 21 points arranged as a stack of three hexagons, each with six peripheral points and a center. There are three classes: the single central point, the eight points at unit distance (six on the medial hexagon and two axial points) and 12 points on the periphery of the top and bottom hexagons (squared distance = 2 units). Again, the search loops need not be rewritten, one simply gets a different list of position offsets. Non-standard density sampling schemes can be accommodated (*e.g.* body-centered cubic sampling in which alternate layers are shifted by half a grid unit).

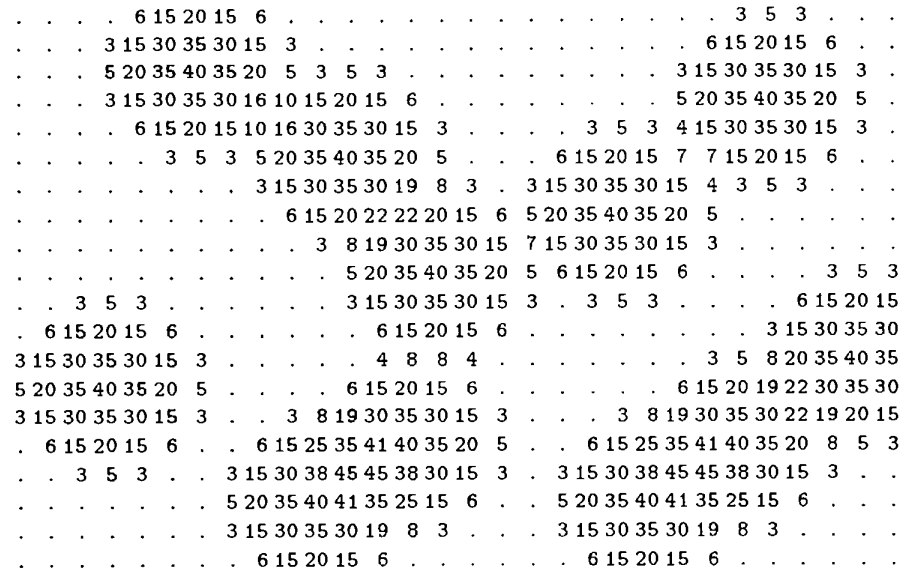
Features (maxima and joins) are found and tabulated as the map is examined, and are assigned an identifying number, with smaller numbers corresponding to higher density. We construct a list of features, their positions

and density, and a connectivity table. In the connectivity table, each join has the set of feature numbers (usually maxima) found together in its neighborhood, and each maximum has a list of all the joins which include it.

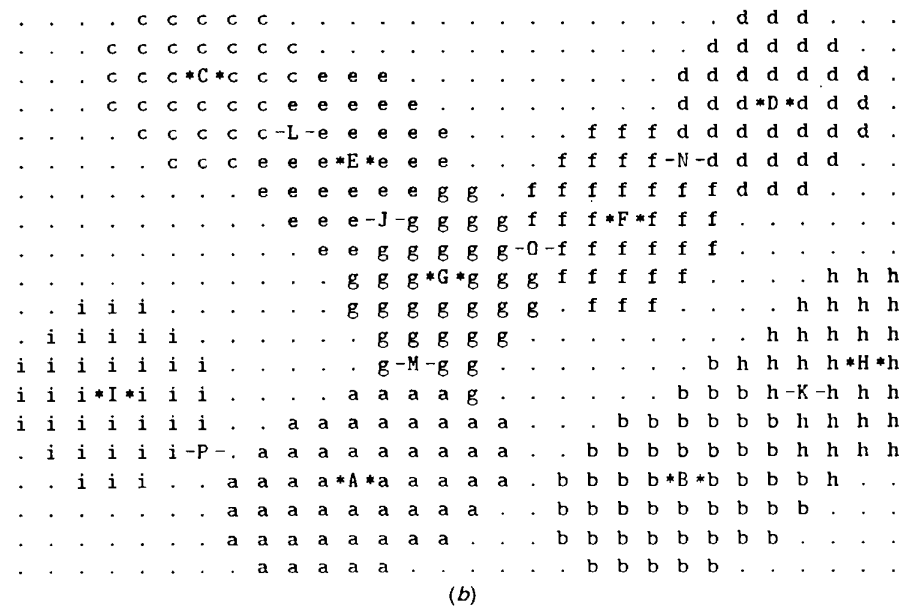
More specifically, we break the algorithm into its data structures and their initialization (0) and three steps: (1) an initial sort that determines the sequence in which the grid points of the map are considered; (2) the examination of a shell of points about each grid point for connectivity; and (3) an assessment of whether there is any new connectivity information.

(0) Start with a three-dimensional array of density values, $d(j,k,l)$ (the map) and a corresponding array of *marks*, $q(j,k,l)$. A mark is initially zero, but ultimately is the identifying number of the associated feature, which is normally a nearby maximum. This feature will not necessarily be the geometrically closest one, but the closest one among a set of features reachable by paths through equal or higher density.

(1) Sort the grid positions by their density values, to give a list of grid points ordered by density



(*a*)



(*b*)

Fig. 1. A model two-dimensional density (*a*) and its partition (*b*) with features indicated. This example is generated by and analysed by the computer code furnished as supplemental material to this paper. (*a*) 11 peaks have been summed in a plane. Resolution is four grid points; pairs of peaks have merged in features *A* and *B*. Values range from 0 to 45, with the range below 3 replaced by dots to separate the molecule from the solvent. (*b*) The partition of the density is indicated by blocks of lower case letters, except in the regions of very low density masked by dots. Features are indicated by capital letters: maxima flanked by asterisks (*A* through *I*) and joins flanked by hyphens (-J- through -P-). Thus, all points which are closer to *A* than to any other maximum are identified by 'a'. The point -M- is the join between *A* and *G*.

value, from highest to lowest. (The sort can be performed efficiently: see *Appendix*.)

(2) At each point, in the order determined in (1), examine a shell of nearby points. Define the set of unique feature marks found in the neighborhood of the central point to be the feature set of the point. The marks serve as proxies for the features, indicating that a path from the mark to the feature exists through higher density. To the central point, assign a mark which depends on the surrounding density values and previously assigned marks. If the feature set is not empty, the mark is normally the closest feature whose mark is in the feature set (2*b*, 2*c*, 2*d*).

There are several alternatives:

(2*a*) No previous marks are seen – a new local maximum. Mark the point with a new feature number, and add that position to the list of features as a maximum.

(2*b*) Only one kind of mark (one mark value) is seen – the point is part of a growing nodule which is isolated in this direction. Mark the point as associated with the feature whose mark was found.

(2*c*) A set of at least two different previous marks are seen – two or more nodules are merging, possibly a new join [see (3) below].

(2*d*) All points in the neighborhood are marked – a new local minimum.

Alternative (2*c*) requires more analysis and it is discussed as a separate step: checking for new connections.

(3) To determine which features have not previously been connected to others in the feature set, try to construct paths from each feature to the others, using the existing connectivity table. If all features are found to be interconnected, there is no new information. If one or more subsets are unconnected, the central point is tabulated as a new join, and the feature set is entered into the connectivity table. Search paths are limited in length: small rings can be excluded from the connectivity table, but larger ones will not be. Increasing the length of the search paths decreases the number of joins accepted, but may ignore direct connections in favor of alternate, more circuitous routes.

One way to limit searching is to use the minimum search depth (3) and attempt to eliminate short loops later during the rendering phase. This has emerged as a reasonable strategy since there are not many loops at thresholds of interest (1.3σ) and one does not irretrievably lose connectivity information.

Typically, most of the lower points near a join will involve the same feature set found originally for the join. As another way to eliminate excessive searching, a scheme of growing 'pancakes' (flat separating disks) out from joins is available. In this case, grid positions are also marked with join numbers, whereas previously only maxima were used as feature marks. Whenever a single join mark is found in a neighborhood (even mixed with maxima), one assumes that one is in the vicinity of that join, and marks the central point with the join mark without a connectivity search. If one sees multiple join

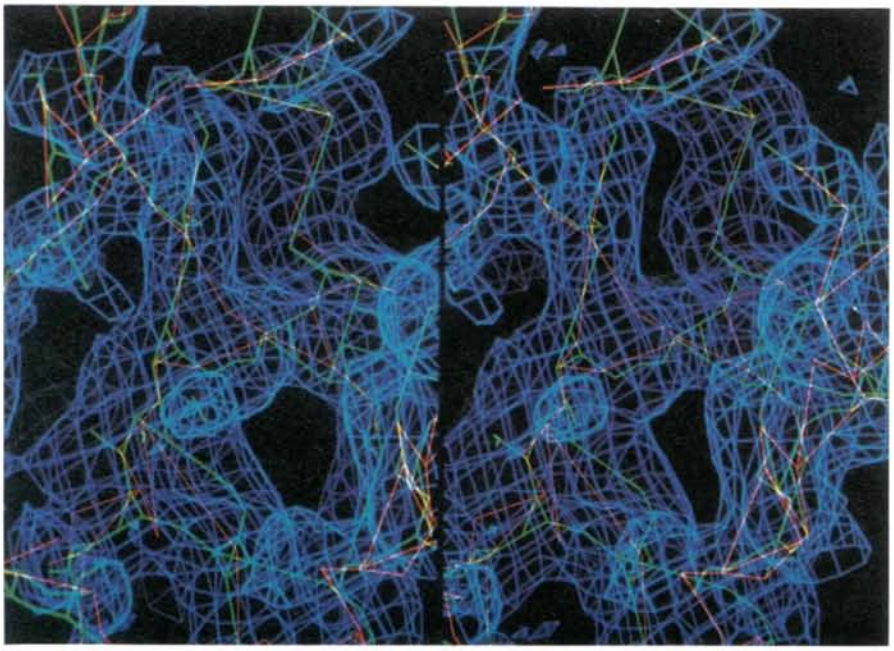


Fig. 2. Comparison of contouring (blue), core tracing (red), and Greer–Hilditch skeletonization (green). Threshold level is 1.3σ in a four-derivative MIR-phased Ht-d map with resolution of about 3 Å. Note the longer vectors in the core tracing and the jaggedness of the Greer skeleton. The Greer rendering is unsophisticated: no attempt is made to eliminate extra lines at some branch points, nor is a choice made among members of clusters of points of equal value.

marks, alone or mixed with maximum marks the point becomes a candidate for a join. Pancakes complicate the logic but are effective; they tend to separate the original nodules associated with maxima and to reduce the number and complexity of possible joins at lower density values.

Core tracing is actually a family of techniques. It is controlled by choice of neighborhood, depth of connectivity search, and use of pancakes. We need more experience to make definitive recommendations for parameters. There will also be some effect from the grid-size choice (as compared to resolution) and the amount of smoothing used to combat truncation ripple in the map. Current working defaults are as follows.

Neighbors: 26 for 'rectangular' lattices (or possibly 18), 20 for hexagonal.
Search depth: 3 (*m–j–m* paths) or possibly 5 (*m–j–m–j–m* paths).
Pancakes: use a search depth of 3.

Core tracing finds the 'total connectivity' of the density, but the nature of the joins found at lower density levels is influenced by the parameter choices just discussed. Although core tracing can find the connections at all density levels, there may not be much point in tabulating this information below the map average (well into the noise for initial maps). Such a restriction cuts the amount of computation at least in half.

Presentation and analysis

There are choices to be made concerning the graphical presentation of the core tracing. These choices are as important as knowing the connectivity because they influence one's perceptions of the apparent connectivity. Some of them have been or could be applied to connectivity determined by other skeletonization techniques. Fig. 2 compares core tracing, contouring and Greer skeletonization in a small volume of density.

Even with the economy of vectors inherent in the method, an entire asymmetric unit or unit cell can be too busy for visual analysis. More limited volumes may also need pruning. Several techniques are available to select a subset of the joins:

thresholds,
limiting the number of connections drawn to each maximum,
lower limits on the length of continuous tracing above threshold, and
volume restrictions based on closeness to a position or to a partial model.

Usually a threshold is combined with one or more of the other restrictions. Since the set of maxima and joins is threshold independent, the choice of level can be made or changed when a display is generated. Most selection criteria could be made interactive with the proper integration into a display program – the number of vectors involved is in the low thousands. To simulate dynamic change of level, core tracings have been rendered at several different thresholds and displayed sequentially or superimposed.

The traditional method of restricting the complexity of a map display is a threshold. Often the choice of threshold has been made on a visual assessment of contours: low enough to give connectivity but not too low so as to result in confusion. Typical levels are of the order of 1σ. (σ is calculated by considering the map as a statistical distribution: calculate the average density and the square root of the variance about that average.) The number of joins increases rapidly below about 1σ in the map. Since a count of the number of maxima and the number of joins is kept as a core tracing is generated, one heuristic could be the density level at which the number of maxima equals the number of joins. This gives enough joins to form a single trace with each join between two maxima and each maximum between two joins ($\cdots m–j–m–j–m\cdots$). The level at which this occurs is somewhat below 1σ (*ca* 0.9σ) and gives excessive connectivity.

However, some of the maxima are in solvent volume so that connections to and between them are irrelevant to the protein structure. Using only the number of joins equal to the number of maxima in the fraction of volume occupied by protein results in a density threshold comparable with the heuristic rule of 1.3σ found by Jones & Thirup (1986). This value is comfortably larger than one estimate of statistical noise of about 0.5σ (Swanson, 1993). In two test cases (see examples below), most of the helix backbone was evident already at 1.5σ, while β-sheet structure was seen only at a lower level (1.3–1.0σ). These observations emphasize that the optimum threshold may not be the same in all regions of the structure.

Core tracing may find six or more joins for prominent maxima; in addition to finding those connections likely to correspond to chemical structure, it connects maxima in neighboring chains, albeit at a lower level. (Indeed, the C, N and O atoms in proteins form not more than three bonds to each other, so one can expect only two or three 'real' connections to a maximum.) For a given maximum, the connection table entries order the joins by their density value, with the first entry having the largest density. The 'join order' of a join is the highest order in any of the several maxima lists in which the join appears. Thus, a join between two maxima which is first in one list and second in the other would have join order 2. Connections of join order 1 and 2 usually correspond to main chain, and those of order 3 to branches, although sometimes the 'chain' visits a strong side residue or crosses a disulfide bridge. A simpler way to include approximately the same set of connections is to draw all joins above threshold which are either first or second entries for any maximum.

Displays of only joins of order 1 and 2 in an initial MIR-phased map (solvent flattened) at 3 Å resolution showed suggestions of helices and sheets and aided in the location of molecular boundaries. With no information other than density level, one should argue that the highest joins give the most likely connections. However, one may still examine some of the lower connections for more acceptable structural shape or topology. In regions of disorder where the connections are at a lower level, the first and second joins to a maximum may give clues to the correct path, even if they are nominally below the global threshold.

By following the connectivity in the maximum-join tables and restricting paths to be above a threshold, the length of chains can be determined. Typically there are many short paths, but only a few very long ones. This provides a powerful selection method to emphasize prominent aspects of a map by focusing only on the longer paths, and is useful for finding molecular boundaries (*cf.* Fig. 5) or for studying a map on the scale of domains or whole molecules.

Two conventional means of restricting display volume are the use of hardware clipping planes (z axis) and density contouring in a small box. A density partition provides a more flexible way to define a volume of interest by combining the grid points associated with some selected subset of features. To restrict the display to the neighborhood of a partial model, determine the feature volumes which contain atoms of the model, and then optionally extend this volume outward in successive layers by adding features (and their volumes) connected to the previous subset. One need not have even a partial model: the original subset could be a few very prominent features or those features within some radius of a point. Core tracing can be restricted to such a subset of features, or can include the entirety of any path which contains a feature in the subset, thus providing clues for extending the model.

A density partition provides a framework in which to ask statistical questions. Does the density connectivity correspond to the model connectivity? Are there statistical differences between features in solvent and protein volumes? How do feature locations and connectivity change as a function of phasing or noise? Can one correlate the shape or bulk of feature volumes with certain side chains? What is the variability of density maxima near specific atom types (C_α, carbonyl O, main or side chain)?

To adequately explore these questions will require comparison of a number of structures, but Table 1 gives some preliminary results. Refined atomic coordinates were compared to MIR maps for two structures (see next section for references and more details). To get main-chain statistics a sequence of atoms ($\cdots$N—C_α—C—N$\cdots$) is transformed into a sequence of features. Pairs of maxima are determined from the density partition corresponding to atom pairs along the chain. Two atoms in the same feature volume are presumed connected and ignored. When adjacent atoms correspond to different maxima, the density level of the join between these maxima is found. The '% main-chain' statistic compares the percentage of joins above a specified level to the total number of joins required to form a complete chain. Counting terminations of sequences of joined features at some threshold gives the number of 'breaks'. The astacin and Ht-d four-derivative maps are clearly better than the three-derivative Ht-d map, but even the good maps show only 80% main-chain connectivity at 1.0σ.

Table 1. *Connection statistics*

	Breaks	% Main chain		% Outside	
Map	1.3σ	1.0σ	1.3σ	1.3σ	1.5σ
Astacin	40	81	62	18	12
Ht-d 4a	28	78	62	26	16
Ht-d 4b	25	82	69	21	14
Ht-d 3a	49	60	43	26	20
Ht-d 3b	51	56	38	31	27

Notes: Astacin has 200 residues, each molecule of Ht-d has 202 residues. 'Ht-d 4a' and 'Ht-d 4b' designate two independent molecules in a four-derivative map. 'Ht-d 3a' and 'Ht-d 3b' designate corresponding molecules in an earlier three-derivative map. 'Breaks' gives the number of discontinuities in the main-chain trace at 1.3σ. '% Main chain' gives the percentage of joins present along the main-chain path at two levels when compared with the total number of joins required for a connected main chain. '% Outside' gives the percentage of joins in the molecular volume which connect to features outside that volume.

Molecules in a crystal are not isolated. To estimate intermolecular connectivity, the maxima whose feature volumes included atoms from a single molecule were determined. Then all joins above a threshold which referenced any maximum in this set were found. The percentage of these joins which also connect to a maximum not in the set is an estimate of the connectivity between molecules ('% outside' in Table 1). No effort was made to determine whether such outside paths extend for short or long distances or actually made contact with other molecules. All of the examples show more than 10% outside connections at a relatively high threshold (1.5σ). The astacin map may have slightly lower values because it was solvent flattened.

Examples

Core tracing can be instructively compared with the original density or with the final model at several different scales: close up (Fig. 2, a few residues), at the scale of elements of secondary structure (several helices or a β-sheet, Fig. 3), or large volumes encompassing domains, whole molecules or even multiple asymmetric units when delineating molecular boundaries (Figs. 4 and 5).

Our examples are taken from two zinc metalloproteases, astacin and a snake venom type IV collagenase

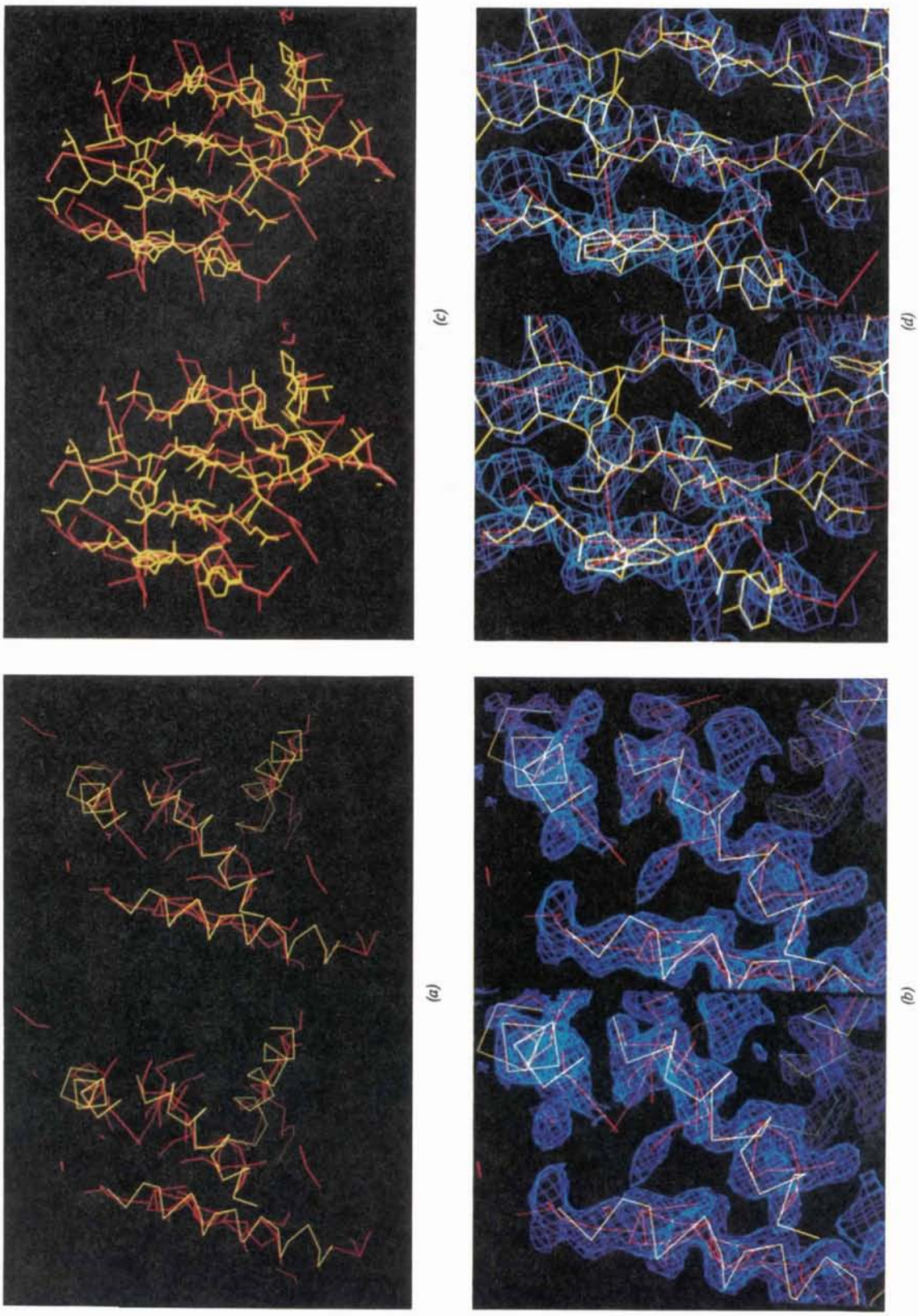


Fig. 3. (*a*) and (*b*) The four major helices in Ht-d. Threshold level is 1.5σ in a four-derivative map. Both the core tracing and the contouring have been restricted to feature volumes either containing atoms of the refined model helices or to the first guard layer of volumes surrounding these (see text). (*a*) C_α trace (yellow) and core tracing (red). (*b*) A closer view of a part of (*a*) with contours (blue). (*c*) and (*d*) β-sheet strands from astacin. Volume is restricted to model atoms as in (*a*) and (*b*). (*c*) Core tracing (red) at 1.0σ and refined model (yellow). (*d*) Detail of (*c*) with contours and core tracing at 1.3σ.

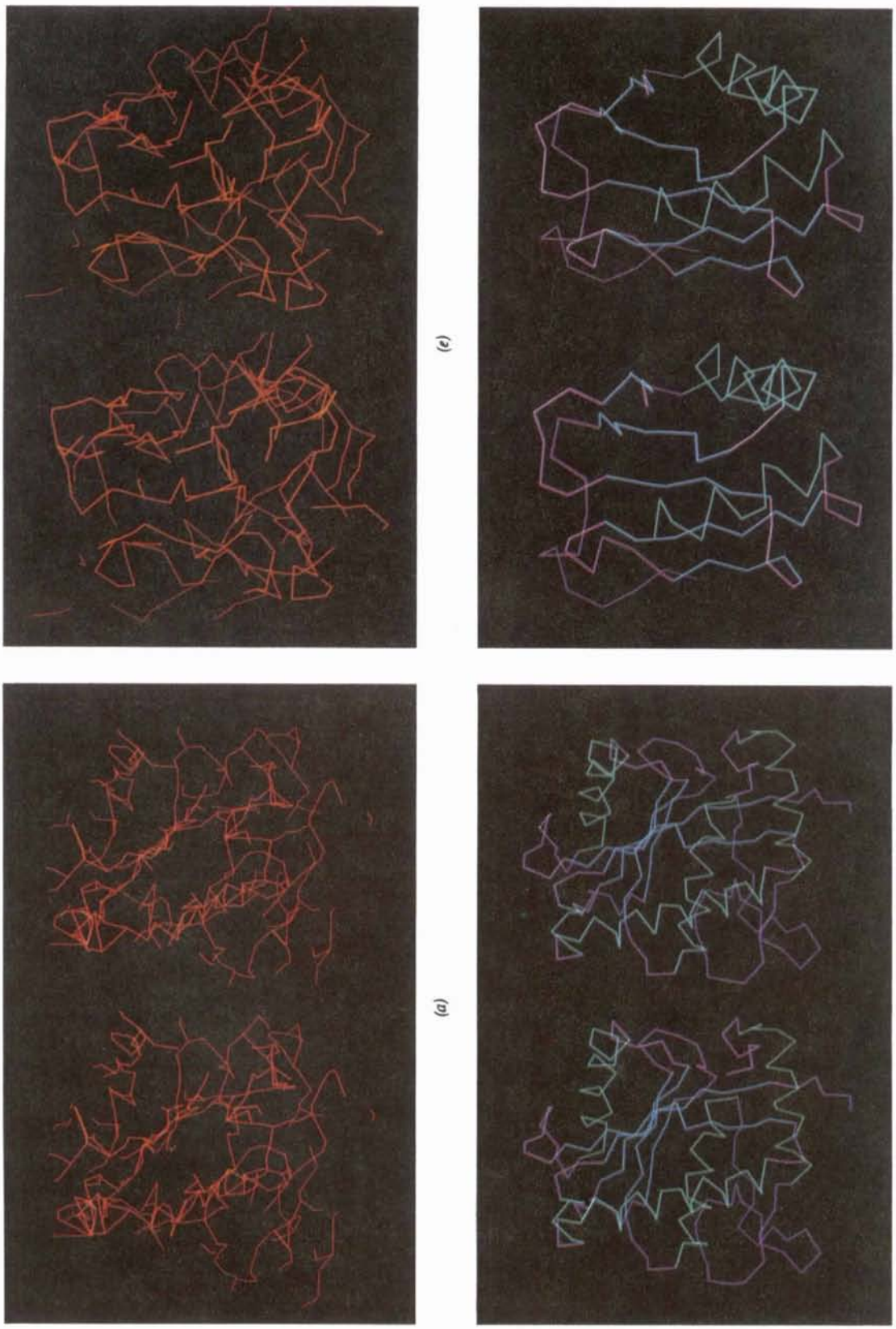


Fig. 4. (*a*)–(*d*) and (*h*) Core tracings of the two independent molecules of Hr-d compared to a C_α trace. Density volume is restricted to C_α atoms plus two guard layers. Molecule *a* has been transformed by the non-crystallographic symmetry. Core tracing includes all joins above 1.3σ. (*a*) Core tracing of molecule *b* (red). (*b*) Color-coded C_α trace with helices (cyan), β-strands (blue), and random coil (violet). (*c*) Core tracing of molecule *a* (violet).

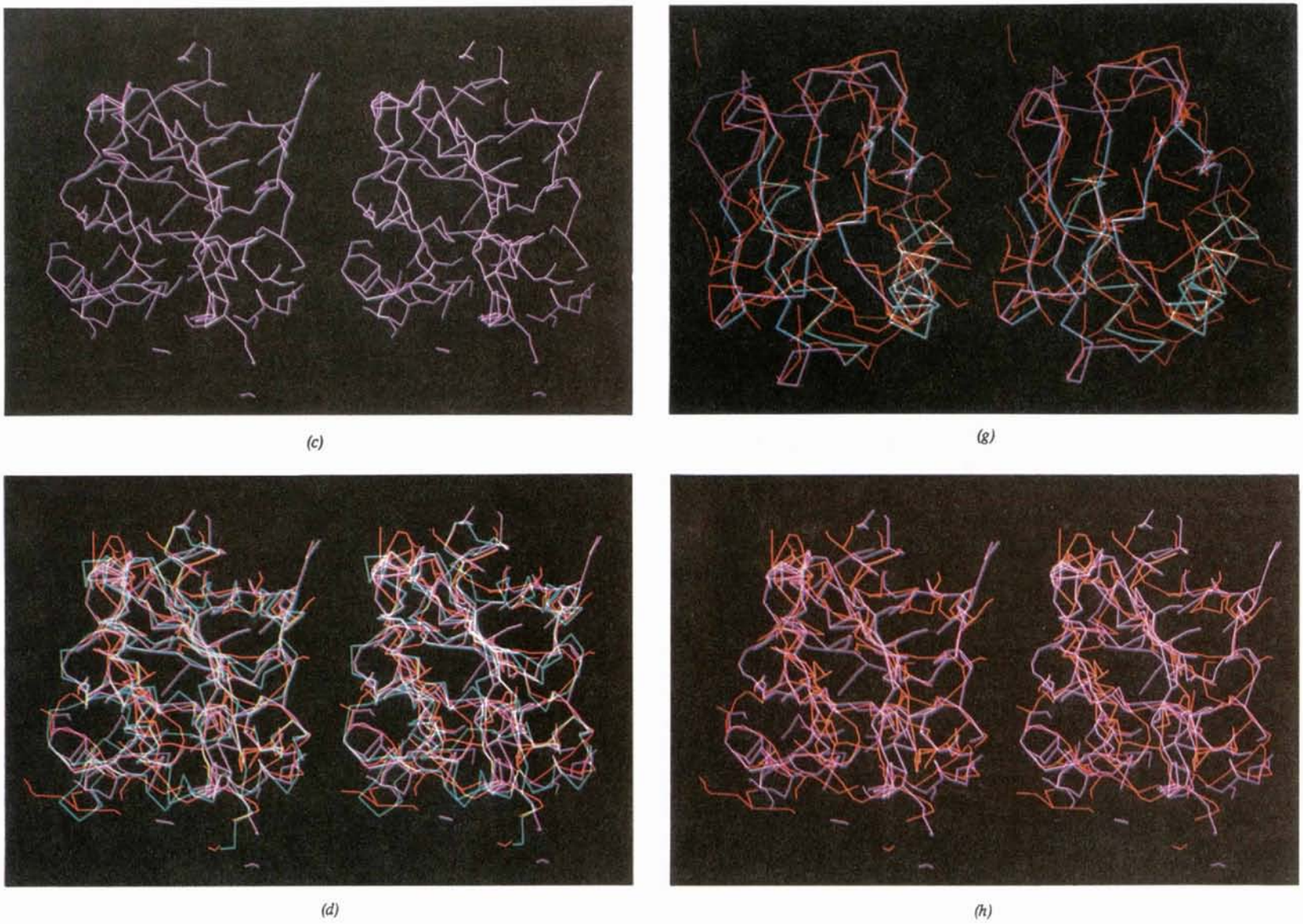


Fig. 4 (*cont.*) (*d*) Superimposed core tracing of molecules *a* and *b*. (*h*) Superposition of C_α trace (cyan), molecule *a* (violet), and molecule *b* (red). Other colors (*e.g.* white) appear due to additive mixing of overlapping lines. (*e*)–(*g*) Core tracing of the amino domain (residues 1–99) of astacin compared to a C_α trace. Density volume is restricted as in (*a*). (*e*) Core tracing (red) including all joins above 1.3σ. (*f*) Color-coded C_α trace with helices (cyan), β-strands (blue), and random coil (violet). (*g*) Superposition of (*e*) and (*f*), with some additive mixing.

(Ht-d). Although neither structure was solved using core tracing, the solution of Ht-d employed core tracing for the determination of molecular boundaries and in the initial fitting.

Astacin has a well formed 3 Å solvent-flattened MIR-density map (Gomis-Rüth, Stoĉker, Huber, Zwilling & Bode, 1993). The map (their Model 1) is based on six heavy-atom derivatives. Helices, some β-strands and a number of bulky side chains are clearly visible when the map is compared with the final model. Several figures have been made from this structure in order to have a wider representation in our examples.

The majority of our examples are taken from Ht-d, a structure under investigation in our laboratory (Zhang *et al.*, 1994). Despite extensive effort, we did not solve this structure from our MIR maps. From the beginning, there were tantalizing views of helices with enough detail to verify the handedness and fix the space group ($P6_5$). None of the partial models would successfully refine. In retrospect, the problem was a poor density, phased on only three heavy-atom derivatives. The presence of two molecules in the asymmetric unit related by non-crystallographic symmetry also caused difficulties in the phasing. The structure was solved by a graphical replacement procedure using a closely related model (Gomis-Rüth, Cress & Bode, 1993) positioned to coincide with identifiable parts of the three-derivative maps (Zn, several helicies). At about the same time, a reassessment of our data sets found a usable fourth derivative. This produced a much better MIR map which fits the refined model but is independent of it. A four-derivative map was used in most of the illustrations, resulting in clearer views than those obtainable with a three-derivative map, but being less faithful to the structure-solution process.

Helices were discernible at higher thresholds than β-sheets in both structures, although there were always some gaps in the main chain, even at 1.0σ. Figs. 3(*a*) and 3(*b*) show the four major helices of Ht-d compared, at 1.5σ, to core tracing and contours. For the illustration, the core tracing and contours have been restricted to feature volumes containing the refined atoms and a single guard layer. This is a four-derivative map; the three-derivative maps showed only one or two helices consistently.

β-strands are not clearly resolved in Ht-d, even with a four-derivative map. There tends to be more cross connectivity between strands than connectivity along the main chain. However, when viewed edge on, the sheet is separable from the rest of the molecule. Figs. 3(*c*) and 3(*d*) show the β-sheet region of astacin which is better resolved into strands, and which shows some of the bulkier side chains. Even this sheet has cross connectivity at a level low enough (1.0σ) to capture most of the main-chain connections.

Sometimes a core trace or Greer skeleton is recognizable as helical, but often there are additional connections along the axis of the helix creating a rod-like bundle of lines. The sheets with cross connections may look more like a net than parallel strands. Only some pieces of random coil consistently appear chain-like. Nonetheless there are probably characteristic signatures for helices and sheets which we can learn to recognize.

Volumes containing a few hundred residues are not too cluttered for viewing. Thus, it is possible to compare a core tracing to a C_α trace of an entire molecule. The two independent molecules of Ht-d are shown in Figs. 4(*a*)–4(*d*) and 4(*h*) (202 residues each). There are some differences between the two molecules but, by and large, the core tracings are similar. This is more a test of the quality of the maps and phasing than of the ability of core tracing to find connectivity. As another example, the amino domain (residues 1–99) of astacin is shown in Figs. 4(*e*)–4(*g*). For clarity in these printed figures, the rendering volume has been restricted to the neighborhood of a single molecule or domain. Although such a restriction is not possible before a structure is solved, the dynamic rotation, scaling and clipping on an interactive display compensate in part.

Finally, we explore the delineation of molecular boundaries in Fig. 5. A single static view can only suggest what can be seen interactively. We present projections perpendicular to the z axis, since the asymmetric unit in $P6_5$ is relatively thin in that direction (15 Å in Ht-d). The densities used are the two MIR maps for Ht-d analysed in Table 1. The relative quality of the maps is apparent in the differentiation between molecule and solvent regions. The selection of long paths for the core tracing reduces the clutter without degrading the signal (Figs. 5*a* and 5*b*). Some paths may appear short in the figures because they have been clipped at a boundary and continued elsewhere. Our structural references are a dot plot of C_α positions (Fig. 5*f*) and an augmented dot plot also containing C_δ positions (Fig. 5*e*) which fleshes out the molecular volume but does not obscure the solvent volume. Dot plots of joins and maxima were also examined as alternatives to core tracing. Figs. 5(*c*) and 5(*d*) show the highest 30% of the maxima for Ht-d (threshold about 2.3σ). The maxima clump in the molecular volume, especially for the four-derivative map where only 9% of the dots lie in solvent volume (Fig. 5*d*). The difference between solvent and protein is still visible for the three-derivative map although 25% of the dots lie in solvent.

Spin-offs: implications for Greer's algorithm

Some of the techniques developed for core tracing, especially neighborhoods and the sort procedure, are useful when applied to other algorithms. A sort on a 13-bit density provides several thousand bins for values, many more than the usual dozen or so value ranges used in Greer implementations. Thus, the list of points

to be removed at each step (set R) is smaller and much less likely to contain neighbors. Lists of neighbors connected to each (nearby) neighbor of the central point can be computed once and used to determine whether deletion of a specific point will disconnect a point set. The computation depends only on distance, not on the presumed shape of a neighborhood (conventionally a 27-point, $3 \times 3 \times 3$ box), and thus is adaptable to non-

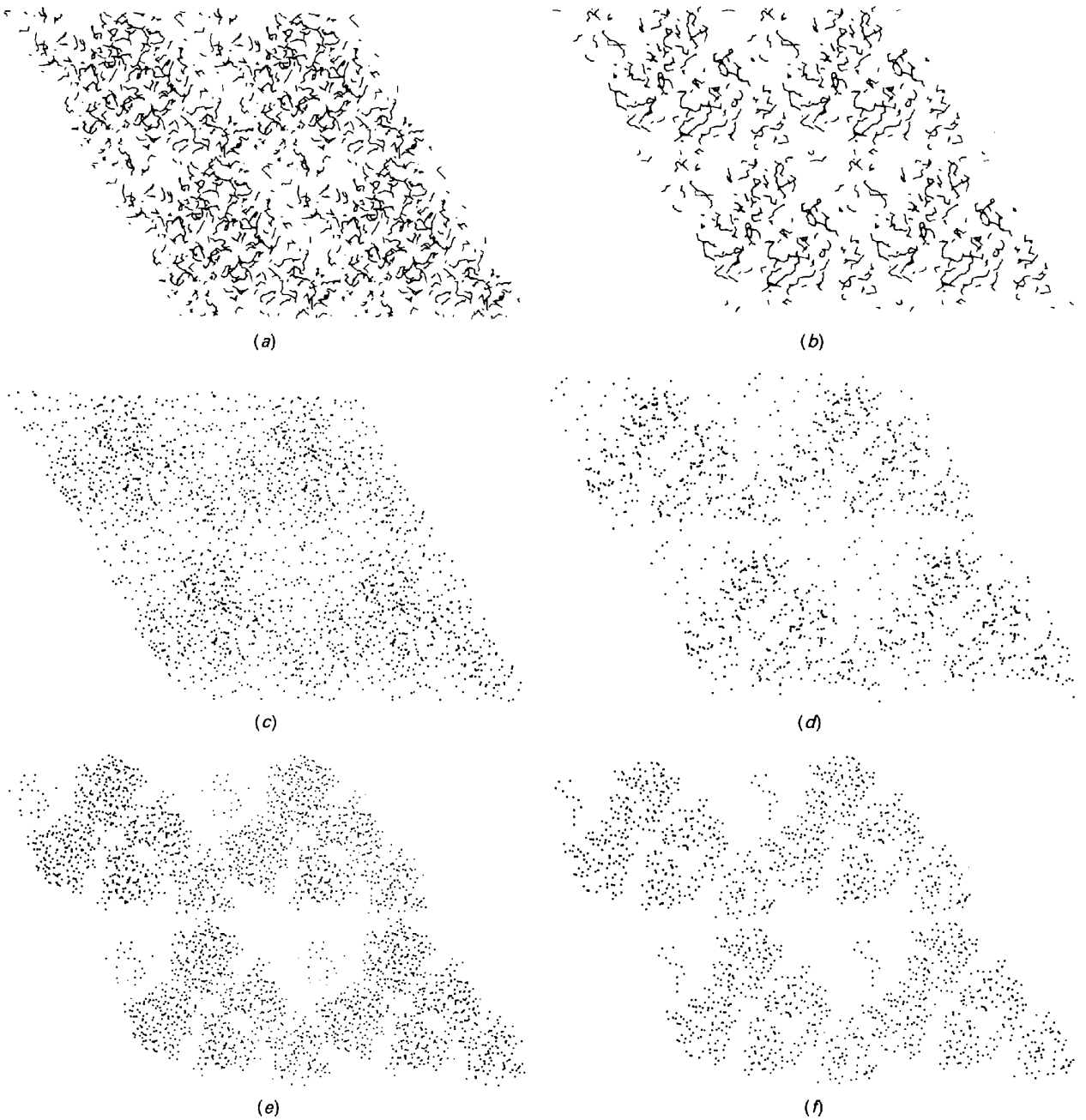


Fig. 5. Molecular boundaries and packing. Very large scale views of two MIR maps of Ht-d comparing long core-tracing segments above 1.3σ containing at least four maxima [(*a*), (*b*)] to high maxima [(*c*), (*d*)] and to selected atom positions [(*e*), (*f*)]. Views on the left [(*a*), (*c*)] are from an early three-derivative map. Views on the right [(*b*), (*d*)] are from a much better four-derivative map. Resolution is about 3 Å and neither map is solvent flattened. The space group is $P6_5$. The views extend two unit cells in x and y and are perpendicular to the z axis (1/6 of a unit cell thick), for a total of four asymmetric units containing eight protein molecules. (*a*) Core tracing of a three-derivative map (a higher threshold gives a clearer view of the solvent volume). (*b*) Core tracing of a four-derivative map. (*c*) The highest 30% of the maxima in a three-derivative map. (*d*) The highest 30% of the maxima in a four-derivative map. (*e*) C_α and C_b positions. (*f*) C_α positions alone.

orthorhombic lattices. Use of the Greer 'cube' on a hexagonal grid results in a 27-point rhomboid which is geometrically biased. Joins and maxima can be located (above the initial threshold) and rendered by core-tracing techniques, or the Greer–Hilditch trace can be constructed by a steepest upward gradient search from joins to maxima. I have not found that the test for hole creation is very useful in three dimensions and believe that it can be eliminated with no important consequences.

Implementation

The algorithm has been implemented on a VAX linked to an E & S PS330 display (program *FRODO*: Jones, 1978; Pflugrath, Saper & Quiocho, 1984) and on the E & S workstation (program *PRONTO*, a variant of *FRODO*), but not yet fully integrated with either system. Currently, a typical calculation (277 200 map points) requires 4.5 CPU min on a VAX station 3100 (five VPU's) for the identification of features. Another much quicker step produces MOL files for display with *FRODO*.

The user may then choose to display combinations of MOL files (chains, branches, at selected levels) to examine local or global volumes. Of course, the usual electron-density map contours can be displayed at any time. Because of the visually overwhelming complexity of the contour map, it is usually viewed only intermittently. The (real-time) interactive application of core tracing in the program *PRONTO* is a project for the immediate future.

This has been a long-term slowly maturing project. I wish to thank R. Swanson for numerous discussions and just for listening as I tried to explain the ideas. E. F. Meyer has provided encouragement, enthusiasm and financial support. D. Zhang and E. F. Meyer have provided initial user feedback in the application to Ht-d. F. X. Gomis-Rüth and W. Bode have kindly provided the map and model for astacin used in several of the figures. Funds have come (indirectly, as salary and laboratory support) from the National Science Foundation, the Office of Naval Research, the Robert A. Welch Foundation, the Texas Agricultural Experiment Station, ICI Americas and Schering-Plough.

APPENDIX

A simplified, two-dimensional version of the algorithm in both Fortran and C has been submitted as supplemental material.* This *Appendix* is intended to address some general implementation issues and to indicate where the algorithm must be extended beyond what was sketched in the body of the paper. Representation of density by integers permits the use of an efficient sorting algorithm but exacerbates the problem of equal values in neighborhoods. There can also be problems near the boundary surfaces of a density volume. Lastly, interpolation of feature positions may not gain much accuracy.

In order that points with the same density value are added at the 'same' time to the appropriate growing nodules all over the map, the indices of the points are sorted by the density value of the point. The points are visited in top-down density order, from highest to lowest. Since the density takes on a medium-sized (8000) range of integer values, a modified radix sort (Knuth, 1973) has been developed which takes only two passes through the density (speed of order N). A conventional radix sort is multi-digit whereas we use a single 13-bit 'digit'. The first pass counts the number of points with each value, and allocates variably sized bins in a table of indices to the density points; each bin will contain the indices of all density points of the same value. The second pass puts the indices into the corresponding variably sized bins by using and incrementing a pointer into the bin for the value. In practice, the table of indices takes twice as much space (32-bit values) as the density so that the random insertion of indices into such a large table produces excessive memory paging. Instead a multipass scheme has been adopted: by ignoring values outside of a subrange, only part of the sorted table of indices is made for a pass through the density and all of the neighborhoods for that density subrange are analysed before sorting a lower density subrange. This is effectively a variable radix two-digit sort.

Since the density values are restricted to integers, occasionally neighbors will have equal values. A local search of equal values is used to determine whether a constant region is an extended maximum, or just a flat stretch on a path. The frequency of occurrence (0.4%) and volume (two to three points) of constant regions are small with a density range of 8000, but the frequency increases to 8% of points above 1σ with a range of 250. This is one of the motivations for using a 16-bit density (together with allowing feature marks as large as 32 000) instead of a more economical eight-bit density representation. With clusters of equal density values, the radix sort may contribute a positional bias since the density indices are entered into the sorted table by a scan of the entire density in a particular sequence.

Boundaries on the density map give rise to truncated neighborhoods, and to incomplete lists of associated features (you cannot see easily beyond the boundary, although a unit-cell/symmetry continuation could be devised at the expense of much more complicated distance and neighbor calculations). I have not found a satisfactory way to find features in small volumes and

* A simplified version of the algorithm has been deposited with the IUCr (Reference: GR232). Copies may be obtained through the Managing Editor, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England.

combine only the feature lists without worrying about the completeness of the search at boundaries. Perhaps sufficiently thick guard layers would work (about eight grid points!).

Connectivity is not a purely local relation (local in the sense of depending only on fixed neighborhood of a point). Connections can be long and twisting or bent, and cannot be determined until a considerable volume of density is analysed. The model of a single line segment from maximum to join may miss the core of the density in some cases. This seems to be infrequent, since the spacing of features is usually comparable to the resolution. It can be checked during a post-processing phase, or when the display is generated, and taken care of by drawing a more complicated sequence of lines.

Density is normally sampled on a grid. One is tempted to 'gain accuracy' by interpolating positions and values from the neighboring grid points to the 'true' off-grid feature. Is this reasonable or justified when one is already sampling at one half or one third the resolution? Complicated interpolation schemes (cubic or higher order, some least-squares techniques) have seemed to violate locality in my limited testing; they have needed excessive grid span or sometimes have placed the 'interpolated' position outside of the neighborhood. I have concluded that a simple three-point quadratic scheme along axis directions is all that is justified and even that may be misleading since the density shape is not always a power law. Gradient searches of density calculated at arbitrary points (with a slow Fourier transform) suggest that at most an extra bit or two may be gained in positional accuracy. Also consider what has been implicitly used historically to position the model: with a single low contour level one fits to the midpoint of the contour cage, not to the maximum of the contained density (unless the density profile is symmetrical). Is an interpolated maximum more or less stable to noise than a midpoint of containing contours?

Finally, some remarks on coding details are given below.

Density is stored in an array of 16-bit integers. Conceptually this is a three-dimensional array, but the dimensioning is dynamic, depending on the layout of the map for a particular problem, and actual reference is done with a single index into a linear array. Density values are restricted to the range −8100 to −100 by scaling and shifting so that the same array can be used for feature marks (positive values) which replace density (negative values) as the classification of grid points proceeds.

A neighborhood is defined by a list of nearby lattice points, sorted so that the nearest ones come first. The definition is a template for all the neighborhoods in a given density map; it is given in terms of offsets from the central point, and is computed only once. For each neighbor we retain three offsets along grid axes (used to check whether the neighbor lies within the map volume), a linear offset for indexing into the density array relative to the index of the central point and information about the distance from the center. For each density point, a call to a check routine returns a list of legal neighbors within a specified distance which do not fall outside the spatial bounds of the map. A single loop then drives the investigation of the neighborhood of the point; a re-analysis of a map with a different size neighborhood is handled by a different list of neighbors, not by changing limits on three nested loops along with different boundary tests. The technique is dimension independent: we have used it in film spot analysis (two dimensions) as well.

The examples of cubic and hexagonal neighborhoods in the main text are in terms of equal real-space grid increments; actual structures often have different grid distances along each axis. An initial sorting of distances takes care of small discrepancies and warns of wildly unequal axis divisions.

A merged list of features is kept, with joins intermingled with maxima. For each maximum, a list of joins referencing it is noted; for each join, the set of features which were seen in its neighborhood. Also kept are the array index (translatable to grid indices) and the density value for each feature. Since the marks are assigned sequentially, a feature with a smaller mark has a higher (or equal) density value than one with a larger mark.

To determine whether a candidate for a join involves new information, a search of pre-existing connectivity information is made, by constructing 'fans' of connections from one of the marks seen in the neighborhood. Starting with a maximum, all its joins are added to the list, then all new maxima contained in those joins, and so on, until either all of the features in the original neighborhood are found, or the specified search depth ('remoteness') is exceeded. If the fan of connected features does not extend to all the original neighboring features, the candidate becomes a new join (and has its own feature number). Maxima connected by a common join have a remoteness of 3, those with two intervening joins and an intervening maximum have a remoteness of 5, and so on.

References

GOMIS-RÜTH, F. X., CRESS, L. F. & BODE, W. (1993). *EMBO J.* **12**, 4151–4157.

GOMIS-RÜTH, F. X., STÖCKER, W., HUBER, R., ZWILLING, R. & BODE, W. (1993). *J. Mol. Biol.* **229**, 945–968.

GREER, J. (1974). *J. Mol. Biol.* **82**, 279–301.

HILDITCH, C. J. (1969). *Mach. Intell.* **4**, 403–420.

JOHNSON, C. K. (1977). *Report of Workshop on Computer Graphics in Biology*, 22–24 June 1976, Columbia Univ., pp. 39–46. NIH, Biotechnology Resources Programs, USA.

JOHNSON, C. K. (1978). *Acta Cryst.* **A34**, S-353.

JONES, T. A. (1978). *J. Appl. Cryst.* **11**, 268–272.

JONES, T. A. & THIRUP, S. (1986). *EMBO J.* **5**, 819–822.

KNUTH, D. E. (1973). *The Art of Computer Programming*, Vol. 3, *Sorting and Searching*, pp. 170–178. Reading, Massachusetts: Addison-Wesley.

PFLUGRATH, J. W., SAPER, M. A. & QUIOCHO, F. A. (1984). *Methods and Applications in Crystallographic Computing*, edited by S. HALL & T. ASHIAKA, pp. 404–407. Oxford: Clarendon Press.

SWANSON, S. M. (1979). *J. Mol. Biol.* **129**, 637–642

SWANSON, S. M. (1993). Am. Crystallogr. Assoc. Annu. Meet., May 1993, Poster PB19.

WILLIAMS, T. V. (1982). PhD thesis, Univ. of North Carolina, Chapel Hill, USA.

ZHANG, D., BOTOS, I., GOMIS-RÜTH, F. X., DOLL, R., BLOOD, C., NJOROGE, F. G., FOX, J. W., BODE, W. & MEYER, E. (1994). *Proc. Natl Acad. Sci. USA*. Submitted.