J. Mol. Biol. (1974) **82**, 279–301

Three-dimensional Pattern Recognition : An Approach to Automated Interpretation of Electron Density Maps of Proteins

JONATHAN GREER†

Department of Molecular Biophysics and Biochemistry
Yale University, New Haven, Conn. 06520, U.S.A.

(*Received 14 May 1973, and in revised form 10 September 1973*)

A procedure is outlined for reducing the high resolution electron density map of a protein to a set of connected thin lines which follow the density. The side chain representations are removed from this skeleton leaving primarily main chain, disulfide bridges and very strong hydrogen bonds. Crystallographic and local operators are used to separate one protein molecule from the neighboring chains in the crystal. Provisional α-carbon positions along the skeletal main chain are derived by application of the "4 Å rule".

The application of these methods to the 2·0 Å electron density map of ribonuclease S (Wyckoff *et al.*, 1970) is described. The skeleton of the isolated molecule that is produced in this fashion provides a good over-all view of the three-dimensional folding of the protein. The results suggest that the skeleton representation can be a valuable supplement to the present methods of map interpretation and a significant step towards complete automation of the interpretation process.

The three-dimensional pattern recognition procedures described may have much broader applications than the protein structure problem for which they have been developed.

1. Introduction

The crystal structures of protein molecules are being determined at an ever increasing rate. As methods of data collection and analysis improve, the protein crystallographer is spending an increasingly large share of his time interpreting the high resolution electron density map.

Constructing a model of the protein to fit the electron density map was greatly simplified by the introduction of the optical comparator by Richards (1968). A half-silvered mirror allows the model and the map to be superimposed optically so that the model can be constructed "into" the electron density map. However, this procedure still requires months of labor by one or more competent crystallographers, who must then measure the provisional atomic co-ordinates from the model and refine them further with computer programs now available (Diamond, 1966,1971; Levitt & Lifson, 1969). As the size of the protein increases, the Richards box becomes more unwieldy, more inaccurate and more space consuming.

Currently, computer graphics systems are being developed to interpret electron

† Present address: Department of Biological Sciences, Columbia University, New York, N.Y. 10027, U.S.A.

density maps (Katz & Levinthal, 1972). Typically, a graphics simulation of a Richards box is used to fit the model to the map. The area of the molecule that can be examined at one time is limited by the size and resolution of the computer cathode ray tube and the maximum number of lines that the hardware will allow one to display. Thus a detailed examination of larger regions of the electron density map becomes impossible. The cost of a sufficiently versatile system is considerable and often beyond the means of the crystallographer.

As an alternative, we decided to try to automate the process of map interpretation. One can conceive of a program or a series of programs which, fed the electron density map and perhaps the amino acid sequence, will produce provisional atomic co-ordinates. These co-ordinates would then be refined using a real space refinement program such as Diamond's (1971). Clearly this is a long range goal, since the interpretation of a 2·0 to 3·0 Å electron density map of a protein requires considerable skill and expertise. We have set as an intermediate objective, then, the processing of the electron density map to a form where it will give the investigator, faced with an unknown structure, an over-all picture of the structure of his protein, and thereby make map interpretation by model or computer graphics easier.

Many possible strategies can be used to process an electron density map. Because crystallographers are often faced with the problem of interpreting a map without the benefit of the amino acid sequence of the protein, this information will not be used in the analysis described in this paper. The strategy, instead, is to try to obtain information about the main chain atoms and some very rudimentary data about the side chains. At a later stage, the sequence can be added and the side chains identified and positioned.

The procedures that are used to analyze the electron density maps are described in detail in the next section. There are six major programs; each is discussed chronologically in a separate subsection. All the programs are written in FORTRAN IV and have been run on the IBM 370/155 at the Yale Computer Center. For demonstrative purposes, a simulated two-dimensional electron density map is used to illustrate the operations that are normally done on a three-dimensional map. The section on Application to Ribonuclease S will describe the application of these methods to the 2·0 Å map (Wyckoff *et al.*, 1970).

While this paper deals with the interpretation of electron density maps, the three-dimensional pattern recognition aspects of skeletonization have a much broader versatility. This procedure is being used at the present time for the analysis of the dendritic network of stained nerve cells from cockroach (Pitman *et al.*, 1972). Other applications may benefit from this type of processing.

2. Method of Attack

(a) *Preparation of the electron density map*

An electron density map with resolution between 2·0 Å and 3·0 Å is the starting material. The map is transformed into a Cartesian co-ordinate system with a grid definition of around 1 Å or slightly less. The Cartesian space chosen should conform to the following requirements. The approximate center of gravity of the protein molecule (or subunit to be examined) should lie close to the center of the Cartesian space. The reason for this stipulation will become clear in section (d) below. This is

not a very stringent requirement, since the approximate center of the molecule is usually known either from packing considerations or from a low resolution map. In addition, the Cartesian space should be large enough to include the whole of one molecule. The size of the Cartesian space must be chosen so that no part of the molecule to be analyzed should touch the edge of the space, i.e. the space chosen should provide at least two extra grid units on all sides of the main molecule. Pieces of neighboring molecules will, perforce, be included in this space. They will be dealt with later in section (d) below.

(b) *Skeletonizing the density map*

The first problem encountered is the very large number of points and the high degree of redundancy in the average electron density map. The map must be transformed to an appropriate formalism that allows for efficient manipulation, yet preserves the structure sufficiently so that interpretation is still possible.

Protein molecules consist of a small number of polypeptide chains which are reasonably thin in diameter, yet very long, often up to hundreds of Ångstroms. Branches extrude from this chain at close to regular intervals (see section (f) below). Occasionally the chains are cross-linked with disulfide bridges or very electron-dense hydrogen bonds. One possible formalism is to reduce the electron density to idealized thin lines which follow the long polypeptide chains preserving at all times the connectivity of the structure. Side chains are represented by thin lines that come off the main chain and cross-bridges become thin lines connecting the main chain lines.

As the first stage in processing we decided to reduce the electron density map to a skeleton of thin lines by generalizing the method developed by Hilditch (1969) for processing micrographs of chromosome smears. I will enumerate in some detail below the skeletonizing algorithm as generalized to three dimensions for electron density maps. It is beyond the scope of this paper to explain the principles and rationale behind the various criteria used to remove points from the map in order to create the skeleton. Readers are referred to the competent introduction by Hilditch to the pattern recognition aspects of this analysis. As far as possible, the nomenclature in this section will correspond to that of Hilditch.

The points in the map are divided into three operational subsets I, N and R. Points which are above a specified minimum density level are placed in I; points with density values below this level in N; and the last subset of points which are being removed in the current pass through the map in R. All members of R are in the same density interval; the density level being removed in the current pass through the data. When the pass is completed, all points in R are transferred to N.

Any point in the map (with the exception of the outer edge of the map) has 26 immediate neighbors. A neighbor of a point is defined as any other point whose co-ordinates differ by no more than one grid unit in all major directions. The removal of a point in the map depends upon which of its 26 nearest neighbors are in I, N or R. In the equations that follow, the neighbors are numbered as shown in Figure 1.

The grid point, at position 14, will be removed from the map if all the following tests are passed in sequence:

1. Its density $= d$, where d is the density interval being examined in this pass through the map.

2. No "hole" is created by removing this point. To remove the grid point each of these three inequalities must be satisfied:

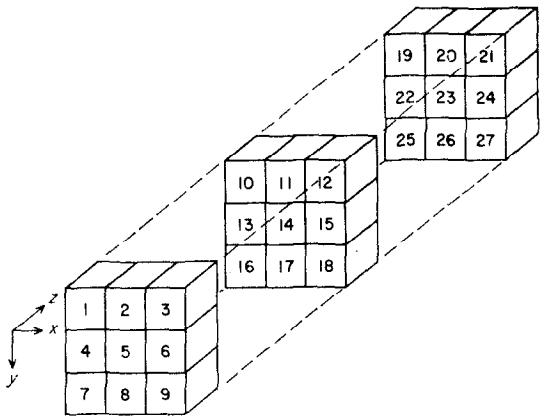


FIG. 1. A point in the map has 26 neighbors, forming a $3 \times 3 \times 3$ cube of grid points. For the equations that appear in the text, the points are numbered as shown in this Figure.

$$a(11) + a(13) + a(15) + a(17) > 0$$
$$a(5) + a(11) + a(17) + a(23) > 0$$
$$a(5) + a(13) + a(15) + a(23) > 0$$
$$\text{where } a(i) = 0 \text{ if point } i \text{ is in I or R,}$$
$$= 1 \text{ if point } i \text{ is in N.}$$

3. The point is not a "tip" or single point:

$$\sum_{i=1}^{27} c(i) > 2$$
$$\text{where } c(i) = 1 \text{ if point } i \text{ is in I,}$$
$$= 0 \text{ if point } i \text{ is in R or N.}$$

FIG. 2. Simulated 2-dimensional electron density map of a tridecapeptide. (a) This map, presented in its contour form, has been selected to highlight several common features of a protein map. The sequence, starting from the lower right-hand side of the Figure, is *N*-fMet-Lys-Ala-Gly-Val-Ser-Arg-Gly-Thr-Glu-Ile-Cys-Leu. Notice that very little of the side chains of Lys-2 and Arg-7 appears. Ser-6 hydroxyl forms a hydrogen bond with the main chain carbonyl oxygen of Ala-3, creating a loop of density. The side chains of Thr-9 and Ile-11 are disconnected from the main chain at this contour level. Negative contours are suppressed in this Figure. The scale bar is 5 Å. (b) A digital representation of the contour map. The first contour in (a) is at 0 on the scale used in this Figure. The contour interval of (a) is 2 on this same scale. The minimum density that is being considered is 1, so all points that are 0 or below are replaced by blank. The grid interval is 1 Å. (c) Skeleton map of the simulated 2-dimensional map. The connectivity of the structure has been preserved and the side chain appendages can be seen at intervals along the main chain. (d) Skeleton map with the side chains removed. The side chains of residues Lys-2, Ala-3, Val-5, Arg-7, Glu-10 and Cys-12 have been deleted and the branch points on the main chain flagged, as represented by the circles. The side chain branches at the ends of the chain have been retained since no *a priori* information is assumed as to which branch is the end of the main chain and which is the terminal side chain (see text). (e) Line drawing representation of (d). A straight line has been drawn connecting neighboring points on the skeleton. This Figure highlights the connectivity of the structure. The asterisks mark the points along the main chain where side chain branches were removed. The scale bar is 5 Å. (f) Composite drawing of (e) and (a). The skeleton line drawing has been superimposed on the original contour map to demonstrate how closely the skeleton fits the contour map.

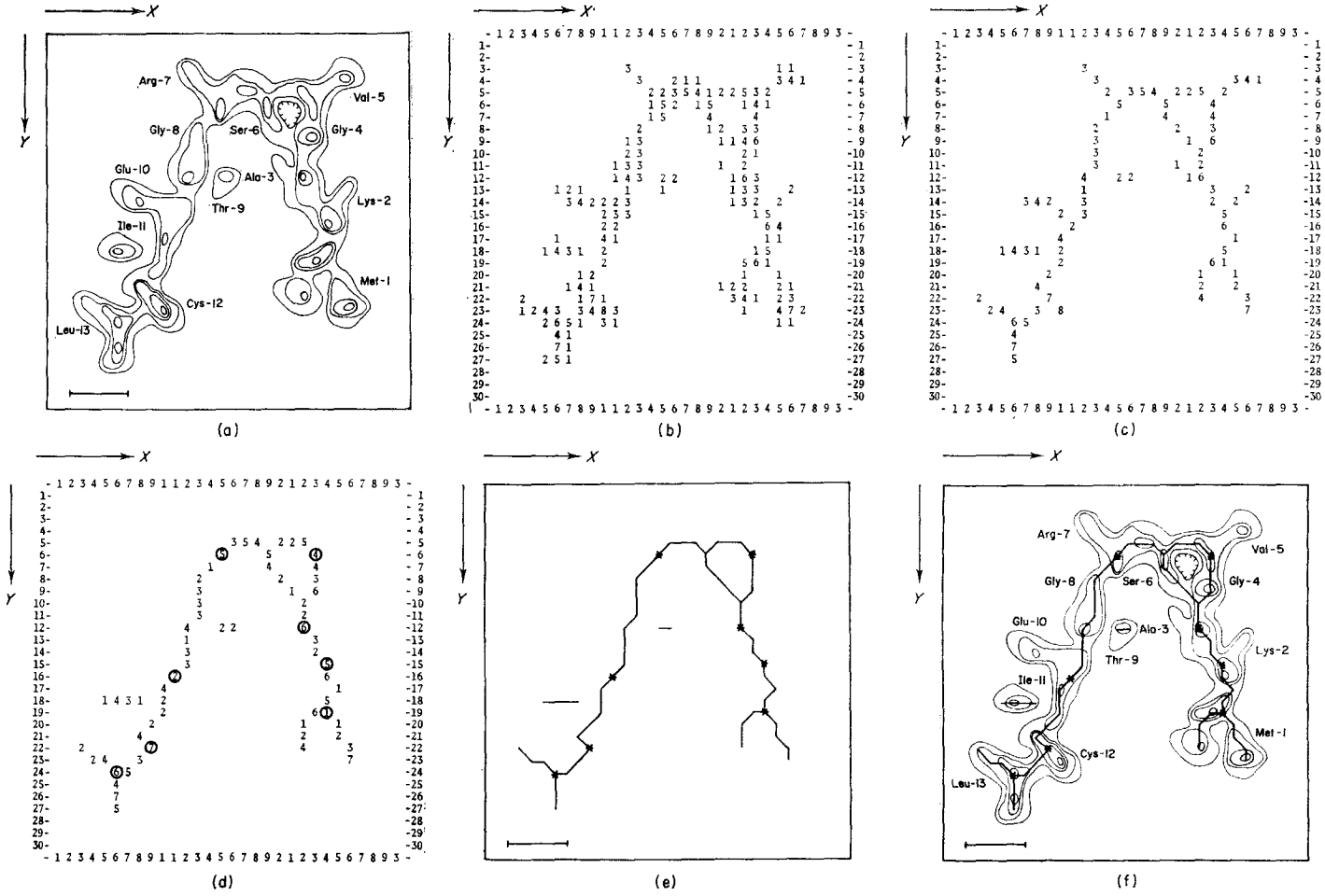


FIG. 2. *See facing page for legend.*

4. The removal of this point does not destroy the continuity of the skeleton. Operationally, this test is done as follows: point 14 is removed from the array of neighbors. The 27 members are scanned and the first point, p, which is a member of I, is stored in a subset called Q. Every other point of the 27 which is a neighbor to p and is also in I is stored in Q. This cycle is repeated for all the neighbors of the neighbors as many times as required until all of the 27 points that are connected directly or indirectly to p by the neighbor relationship are collected into subset Q. The remaining points in the 27 are then checked to see if there are any points in I which are not in Q. If there are, then the central point at position 14 is essential to preserve connectivity of the structure. If no points are found that are not in Q, then all the points remain connected in the absence of point 14 and it may be removed.

If a point passes all the above criteria, it is removed from the map. In a multi-density map, the map is divided into contour intervals and is scanned sequentially starting from the lowest density and proceeding to the highest until no more points can be removed.

In order to prepare an electron density map for skeletonization, a minimum density level or threshold must be selected. Below this value, the grid point is automatically removed. Clearly, the choice of this level is a critical decision. A contour interval must also be selected which allows the skeleton to favor higher density areas by leaving them until last for consideration, but is not so grainy that false curves are created in the skeleton. Such a choice for the simulated two-dimensional map (Fig. 2(a)) is shown in Figure 2(b). The corresponding skeleton for this map is shown in Figure 2(c). The reader is invited to apply the rules given in this section to this map in order to see how the skeletonization process works in detail.

It is worthwhile at this stage to examine some features of the skeleton that result from the application of the above rules. The exact position of the skeleton point depends to some extent upon the sequence in which the points are considered for removal. This dependence is minimized, but not eliminated entirely, by defining $a(i) = 0$ in test 2 if point i is in R as well as in I (refer to Hilditch for details). Nevertheless, the final position of the skeleton line may be one grid point away from the position one might set by visual examination of the density contours. (For example, when Fig. 2(c) is compared to Fig. 2(a), one might have chosen point (24,17) rather than (25,17) as part of the skeleton.) It remains to be seen whether this sacrifice in spatial positioning of the skeleton is too high a price to pay for the simplification of dealing with the skeleton in the next steps. The severity of this deviation of one grid point can be reduced by choosing a finer grid definition, of course, at the price of dealing with a larger number of points.

Where there are discontinuities in the chain at the minimum density level chosen, the skeletonizing procedure tends to draw the ends apart as it reduces the density to a thin line. In addition, some ellipsoidal density contours may be reduced to a single point in the skeleton if the axial ratios of the ellipsoid are not great.

(c) *Tracing the chain*

Having produced a skeletal representation of the map, the next objective is to trace the main chain of the molecule. It is useful to try to eliminate side chain branches at this stage and leave only the main chain. For this purpose, side chains

are defined operationally as the skeletal piece between a tip and branch point†. The points of the skeleton are systematically removed starting at the tip until a branch point is encountered. The branch point is then flagged so that the position of this side chain along the main chain can be used later. At the ends of the main chain, defined by the coincidence of two side chain-like branches to a single branch point, both branches are retained since we do not know at this stage which is the main chain and which the terminal side chain. This procedure is illustrated in Figure 2(d). The program then proceeds to describe the remaining skeleton as illustrated for the two-dimensional case in Table 1. Main chain cross-bridges, due either to disulfide bridges or to hydrogen bonds, are also recorded in this list. The distances between side chain and main chain branch points along the chain are listed as well.

A schematic representation of the skeleton is shown in Figure 2(e). Neighboring points on the skeleton are connected by a straight line constructing a completely connected model of the molecule. Side chain branch points that have been flagged are labeled with an asterisk. Single density points are indicated by a plus. This schematic Figure allows a visual comparison of the skeleton with the original contour map as shown in Figure 2(f). While there are some deviations from the line that one might draw by visual inspection of the contour map, by and large the fit of the skeleton to the electron density map is quite good.

(d) *Isolating the molecule, the redundancy problem*

The next major problem that must be resolved before the skeleton can be interpreted is the isolation of the single molecule from pieces of neighboring molecules. A protein molecule can have a rather odd shape and a quite irregular boundary. Often, neighboring chains touch and the density bridge connecting two molecules can be as high as main chain density. Separating neighboring molecules, therefore, can be a very difficult chore.

There are several possible ways of isolating the molecule. One obvious plan is to describe an envelope that includes one molecule only, and use this envelope to exclude pieces of other chains. Unfortunately, this envelope is often not available, since the boundary of the molecule can be ambiguous in a low resolution map. In addition, the measuring of such an envelope is exceedingly tedious and also inaccurate when used at higher resolution.

I have chosen an alternate method to eliminate neighboring molecules. It is clear that any piece of a neighboring chain must be related to the intact molecule under examination by either a crystallographic symmetry operator or by a local co-ordinate transformation. Consequently, each grid point on the skeleton, with co-ordinates, $\mathbf{X}$, in Cartesian space, can be transformed as follows:

$$\mathbf{X}' = \mathbf{MX} + \mathbf{t}$$

$$\text{where } \mathbf{M} = \mathbf{CSC}^{-1}$$

$$\text{and } \mathbf{t} = \mathbf{CSo} + \mathbf{Mv} + \mathbf{Cs} - \mathbf{Co} - \mathbf{v},$$

where $\mathbf{X}'$ are the symmetry generated transformed Cartesian co-ordinates of $\mathbf{X}$; C is the conversion matrix from crystallographic to Cartesian space; C^{-1}, the inverse of C; S is the rotation matrix of the symmetry operation; $\mathbf{o}$ and $\mathbf{v}$ are origin shifts for

† A tip is defined as a point that has only 1 neighbor. A branch point is defined as a point having more than 2 neighbors.

TABLE 1

Analysis of the main chain of the simulated two-dimensional map

Point X	Point Y	Density	Between branch points Links	Between branch points Length (Å)†	Main chain section Links	Main chain section Length (Å)†	Comments
26	23	7					
26	22	3					
25	21	2					
25	20	1					
24	19	1	4	4·8	4	4·8	Main chain branch point
24	18	5					
25	17	1					
24	16	6					
24	15	5	4	4·8			Side chain
23	14	2					
23	13	3					
22	12	6	3	3·8			Side chain
22	11	2					
22	10	2	2	2·0	9	10·7	Main chain branch point
21	9	1					
20	8	2					
19	7	4					
19	6	5	4	5·2	4	5·2	Main chain branch point
18	5	4					
17	5	5					
16	5	3					
15	6	5	4	4·8			Side chain
14	7	1					
13	8	2					
13	9	3					
13	10	3					
13	11	3					
12	12	4					
12	13	1					
12	14	3					
12	15	3					
11	16	2	10	11·7			Side chain
10	17	4					
10	18	2					
10	19	2					
9	20	2					
8	21	4					
9	22	7	6	7·7			Side chain
8	23	3					
7	24	5					
6	24	6	3	3·8	23	28·0	Main chain branch point
5	23	4					
4	23	2					
3	22	2	3	3·8	3	3·8	Tip reached
24	19	1					Start new branch
23	19	6					
22	20	1					
22	21	2					
22	22	4	4	4·4	4	4·4	Tip reached
22	10	2					Start new branch
23	9	6					
23	8	3					
23	7	4					
23	6	4	4	4·4			Side chain

TABLE 1—*continued*

Point X	Point Y	Density	Between branch points Links	Between branch points Length (Å)†	Main chain section Links	Main chain section Length (Å)†	Comments
22	5	5					
21	5	2					
20	5	2					
19	6	5	4	4·8	8	9·2	Loop completed
6	24	6					Start new branch
6	25	4					
6	26	7					
6	27	5	3	3·0	3	3·0	Tip reached

† Length is measured from grid point to grid point along the skeleton.

the Cartesian conversion; and **s** is the translation for the symmetry operation, including unit cell translations.

Using the above relations, the symmetry operations are applied to each point on the skeleton, with unit cell shifts included for true crystallographic operators where appropriate. Wherever the program finds a symmetry-related point that lies within the Cartesian space, it checks to see if that point appears on the skeleton. In this way, pairs of points on the skeleton that are symmetry-related are discovered and stored for further use.

Criteria must now be selected for determining which of the redundant points should be retained and which deleted. The simplest criterion that can be chosen is to discard the point of the pair that is furthest from the center of gravity of the molecule. (The center of gravity of the molecule is a number that is usually known, as discussed in section (a) above. Since we have chosen the Cartesian space in such a way that the molecular center is at the approximate center of the Cartesian space, the program calculates the center of the whole skeleton, including neighboring chains at this stage. This turns out to be a reasonably good approximation. This number could also be read into the program if desired.) Unfortunately, this criterion causes significant parts of the molecule under study to be eliminated, as will be seen in section 3. This is caused by the elliptical shape of many protein molecules and especially because of crevices in the molecules where neighboring chains may come closer to the center of the molecule than the symmetrically related part of the same chain.

To overcome this handicap, a modification of the above criterion was used. Rather than relate pairs of points to the center of gravity of the molecule, larger features of the skeleton are related to the center. A feature is defined as a group of connected points for each of which equivalents exist elsewhere within the skeleton. While single points on a neighboring chain can be closer to the center of gravity of the molecule under study than the same point on its own chain, this is rarely true for larger features of the molecule. The following criteria were therefore settled on. Features containing one or two points are treated on a point by point basis; for each related point pair, the point furthest from the center of gravity is removed. For the larger features, a square, symmetric correlation matrix, n_{ij}, is constructed which consists of the number of points in feature i which are equivalent to points in feature j. The mean distance of the n_{ij} points in feature i from the center of gravity and of the symmetry-related

n_{ij} points in feature j from the center of gravity is computed. That set of n_{ij} points in feature i or j which is furthest is removed. The efficacy of this criterion will be demonstrated in section 3 with ribonuclease S, a molecule which is clearly ellipsoidal and has a large crevice.

A practical difficulty inherent in this procedure is that symmetry-related points do not always appear identical. Finite grid sizes in both the original electron density map and the skeleton grid used here combined with interpolation and round-off errors produce features that are generally similar but not identical on a point by point comparison. As a consequence of this problem, discontinuities can occur between features that would otherwise be connected. An option in the program attempts to deal with this problem at the expense of a considerable increase in computing costs, but the solutions used at present will not be discussed further at this time.

An extension of the problem mentioned above relates to small pieces of the neighboring chains that are not removed during the redundancy testing program. A second criterion was introduced, therefore, to minimize this effect. After the redundancy deletions, all single points and short chains which have no branch points are re-examined. If the single point, or if both tips of a short chain are further than a specified distance from any other point in the skeleton, then that point or chain is deleted from the skeleton. A reasonable distance for this test is around 4 Å. The rationale behind this test is that any isolated point or small chain that is so far from any other part of the skeleton is unlikely to form part of the molecule under study and will just confuse interpretation if not removed.

With the criteria described above, the molecular shape occasionally causes small bits of the central molecule to be deleted. The next section will be devoted to the description of a procedure designed to reinstate these small pieces that are removed by the redundancy test. Clearly, there will be cases where the molecular shape is so complex that other criteria will have to be applied. There may also be situations when parts of the molecule loop over neighboring subunits, as in lactic dehydrogenase (Rossmann *et al.*, 1971), where even greater difficulty will be encountered in choosing the correct molecular boundary. Situations such as the latter, which often confuse the model-building crystallographer for a long time, will have to be tackled at a later and more subtle stage in the analysis of the protein structure.

(e) *Connectivity of the structure, a ligase program*

The redundancy testing program, while removing neighboring molecules, occasionally deletes small pieces of the central molecule as mentioned above. To compensate for this effect, which will occur even if more sophisticated criteria are used to differentiate between neighboring chains, a program has been created to reinstate deletions resulting from symmetry and distance criteria where these are at variance with the requirements of chain continuity. This program will be called a ligase program.

The minimum skeleton that is produced by the redundancy program is compared to the original total skeleton (from the stage just before the redundancy test). The total skeleton is examined for chains that were removed by the redundancy program and that connect two points in the minimum skeleton. In order to prevent the re-introduction of long and winding pieces of chain, which are probably part of neighboring molecules, the added chains must conform to two criteria: the ratio of the length along the new chain to the actual distance between the two end points

on the minimum skeleton must be less than d and the total number of points in the new chain must be less than N. Values for d that have been used with reasonable success vary from 1·5 to 2·5 and N around 10 to 20 (see section 3).

This program can also be used for another purpose. Protein electron density maps are notorious for having short regions where the main chain density is lower than the average main chain density in the rest of the map. Such regions may often have lower density than hydrogen bonds or other such non-covalent bridges. The resulting skeletal representation of the molecule will have a gap at this point in the main chain. To illustrate, if the minimum density level chosen for skeletonization of the map shown in Figure 2(b) had been 2 rather than 1, several discontinuities of the main chain would have resulted, at (12,13) and (24,17) for example. Rather than running all the programs again at lower density levels, with greater difficulty because of the larger number of points, the minimum skeleton that was calculated at the higher density level can be compared with the total skeleton computed at the lower density threshold. The ligase program will fill gaps in the higher density skeleton with chains that are formed in the lower density one. Discontinuities in the skeleton, such as those described above, would be corrected by this procedure. The values that should be chosen for the parameters in this step will depend on the quality of the electron density map and the degree of discontinuity in the skeleton.

(f) *Preliminary analysis of the skeleton*

The skeleton that has been derived at this stage is almost completely main chain (Fig. 2(f)). Punctuating the chain are actual main chain branch points and side chain flags. Since the intersection of the side chain with the main chain is at the α-carbon, the side chain branch points mark the α-carbon positions.

It is the property of all polypeptide chains that the distance between adjacent α-carbons along the main chain is just under 4 Å, about 3·8 Å, independent of the ψ and ϕ angles of the residues involved†. The "4 Å rule" can be used to count residues along the main chain in parts of the skeleton where branch points do not appear. Thus, the chain between (15,6) and (11,16) in Figure 2(a) and (d) and Table 1 is 11·7 Å long and represents three residues. Similarly, the 7·7 Å chain between (11,16) and (9,22) represents two residues. Table 2 shows the modified description of part of the main chain after the 4 Å rule has been applied.

The 4 Å rule will permit the extraction of a provisional set of α-carbon positions from large parts of the skeleton. In some cases, a residue may be missed or added. Nevertheless, these co-ordinates can be used as a starting set for model-building studies in a standard Richards box or graphics system. Procedures are being developed now to automate this next stage in the processing of map interpretation.

3. Application to Ribonuclease S

The 2·0 Å electron density map of ribonuclease S (Wyckoff *et al.*, 1970) was converted into a Cartesian co-ordinate system with a 1 Å grid using a skew planes program written by J. M. Baldwin at the Medical Research Council Molecular Biology Laboratory, Cambridge, England. The transformation matrix that was applied conforms to that of Richards & Wyckoff (1973) so that the skeleton that is

† This is not true for *cis* peptides. However, *cis* peptides are so rare that they can be neglected at this stage and reintroduced later in the analysis if necessary.

TABLE 2

Analysis of part of the main chain of the simulated two-dimensional map: application of the 4 Å rule

Point X	Point Y	Density	Between branch points Links	Between branch points Length (Å)	Main chain section Links	Main chain section Length (Å)	Comments
18	5	4					
17	5	5					
16	5	3					
15	6	5	4	4·8			Side chain
14	7	1					
13	8	2					
13	9	3	3	3·8			Pseudo side chain
13	10	3					
13	11	3					
12	12	4					
12	13	1	4	4·4			Pseudo side chain
12	14	3					
12	15	3					
11	16	2	3	3·4			Side chain
10	17	4					
10	18	2					
10	19	2	3	3·4			Pseudo side chain
9	20	2					
8	21	4					
9	22	7	3	4·2			Branch point
8	23	3					
7	24	5					
6	24	6	3	3·8	23	28·0	Main branch chain point
5	23	4					
4	23	2					
3	22	2	3	3·8	3	3·8	Tip reached

generated can be compared directly to their measured atomic co-ordinates with, at most, an origin shift. The matrix used is:

$$\begin{aligned} X &= xa - ya\cdot\sin 30^\circ + 36{\cdot}0 \\ Y &= \quad ya\cdot\cos 30^\circ + 26{\cdot}0 \\ Z &= \quad\quad\quad zc + 16{\cdot}0 \end{aligned}$$

where $a = 44{\cdot}65$ Å and $c = 97{\cdot}15$ Å. X, Y, Z are the Cartesian co-ordinates and x, y, z the crystallographic co-ordinates of the original electron density map. The 1 Å grid spacing, somewhat coarser than the ribonuclease S map itself, was used for two reasons. At this map definition, the main chain carbonyl oxygens are usually "absorbed" into the main chain skeleton and do not produce side chain-like branches which would complicate the application of the 4 Å rule. This grid definition also reduced the number of points that had to be processed.

For skeletonizing the map, a minimum density level of 200 was selected and a contour interval of 50. The highest density point on this scale was 841, giving 13 contours; the lowest point was −449. All grid points with density less than 200 were removed. For each contour interval, the program removed all points that passed the

TABLE 3

Number of points removed from electron density map to form skeleton

Code	Density less than	Initial	Removed	Left	Left (%)
	200	62,160	62,160	0	0·0
1	250	2039	1658	381	18·7
2	300	1409	1070	339	24·1
3	350	921	602	319	34·6
4	400	596	339	257	43·1
5	450	328	147	181	55·2
6	500	175	81	94	53·7
7	550	71	26	45	63·3
8	600	30	13	17	56·7
9	650	20	7	13	65·0
A	700	6	2	4	66·7
B	750	2	1	1	50·0
C	800	2	0	2	100·0
D	850	1	0	1	100·0
	Total	67,760	66,106	1654	
	(%)	100·0	97·6	2·4	

tests described in section 2(b). Table 3 shows the number of points removed and retained at each density level. The number of points that must be considered has been reduced from 67,760 to 1654.

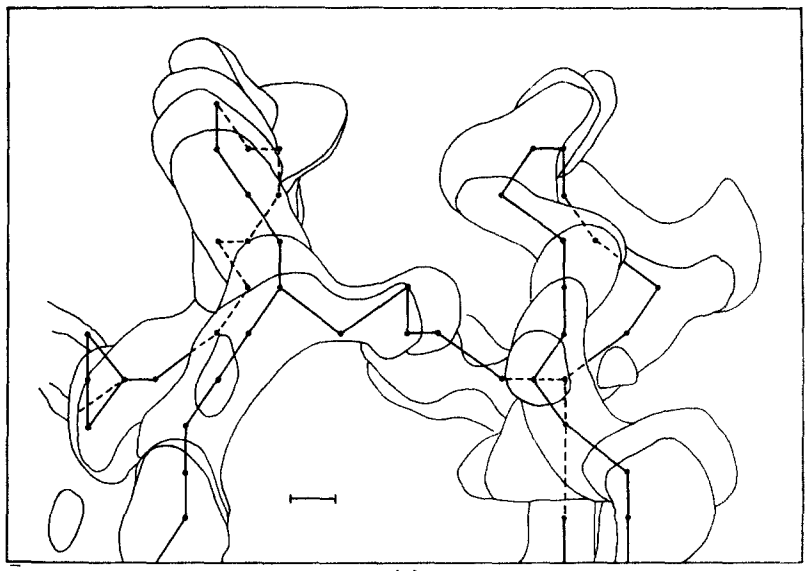
Presentation of the results of these programs on the electron density map of ribonuclease S requires the description of complex three-dimensional structures. It is not the purpose of this paper to report the structure of ribonuclease S in detail. The discussion and illustrations will therefore be limited to specific highlights of the structure as they reflect the strengths or weaknesses of the skeleton procedure. Figures of the skeleton of the whole molecule will be shown at later stages in the

TABLE 4

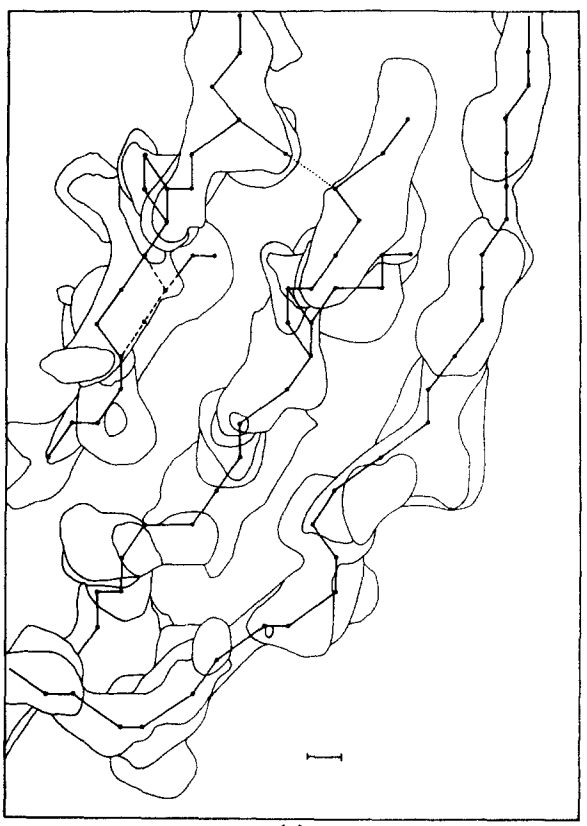
Types of small chains and branches encountered in the skeleton

Type	Number	Disposition
1. Small chains with no branches	88	Retained
2. Chains with no branches that terminate at the edge of the map, "false chains"	24	Deleted
3. Side chain branches	179	Deleted†
4. Two side chains to the same branch point	17	Retained†
5. More than two side chain branches to the same branch point (also counted in 4)	2	Retained†
6. Single density points	91	Retained

† The branch point on the main chain has been flagged.



(a)



(b)

FIG. 3. Composite drawings of the skeleton together with the electron density map of ribonuclease S. Several sections of the map contoured at the minimum density of 200 have been superimposed. The contours that are obscured by sections above them have been suppressed. The dashed lines represent the parts of the skeleton that lie just under the contours of the top sections

[*continued on facing page*

processing when they are more meaningful. Here, I will show only two areas of the map contoured at the minimum density level of 200 with the corresponding skeleton superimposed (Fig. 3(a) and (b)). The fit of the skeleton to the map is quite reasonable. The program occasionally produces three, four or five-membered rings where the observer might have placed a straight line. These rings result from the program finding a lower density point in the center of higher density. As discussed above in section 2(b), there are places where the observer might have drawn the skeletal line in a more central position on the contours. None of these deviations from the center of the density is greater than one grid unit.

The skeletal representations of the "side chains", as defined in section 2(c), were removed next. Table 4 lists the types of chains encountered by the program and their disposition. Each of the branch points on the remaining skeleton was flagged to

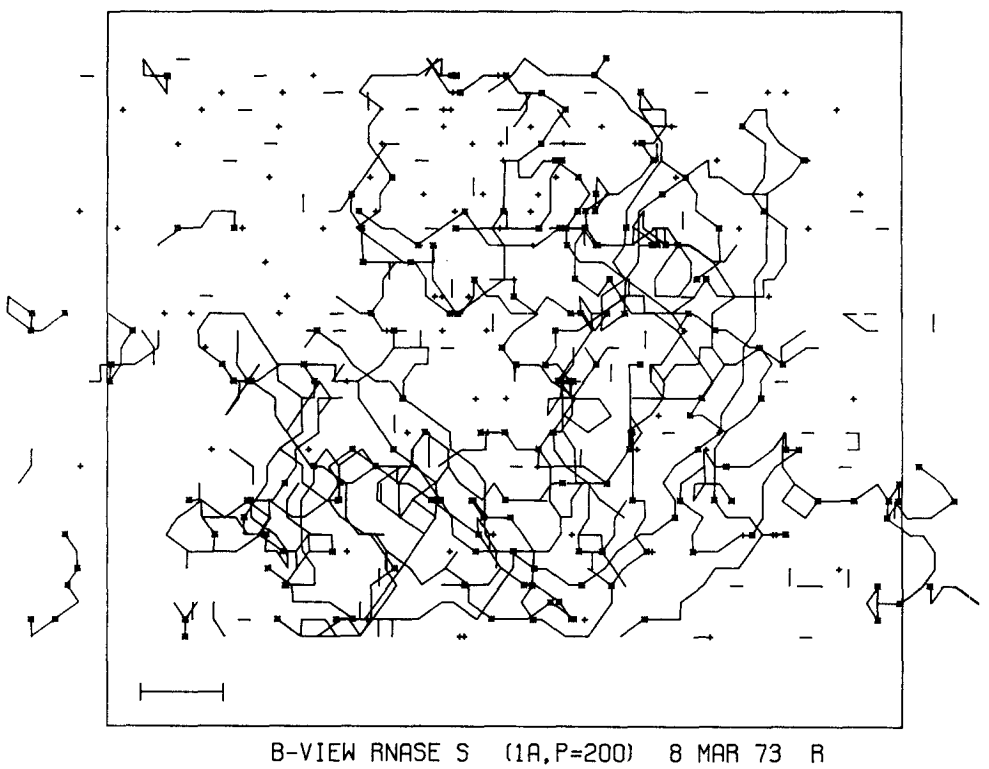


FIG. 4. Line drawing of the complete skeleton of ribonuclease S for the Cartesian space under examination. The "side chains" have been removed, as well as "false chains", i.e. short chains that have no branch points and which terminate at the edge of the Cartesian map. This plot is too complex to present in stereo form. Side chain branch points are flagged with an asterisk. Single density points are labeled with a plus. The scale bar is 5 Å. This drawing corresponds to the line drawing of Fig. 2(e).

in the drawing. The molecule is drawn in the standard B-view of Richards & Wyckoff (1973). (a) Section of the ribonuclease S map showing the disulfide bridge between residues 40 and 95. The chains from 34 to 41 and 88 to 96 are included. (b) Section of the ribonuclease S map showing one of the β-structure regions. The 3 chains are composed of residues 42–47, 80–86 and 97–104, from left to right in the Figure. The dotted line is included for continuity, even though the density contours that correspond to it are below the sections in this drawing. The scale bar is 1 Å

indicate that a side chain was removed at this point. Figure 4 shows the complete skeleton after the "side chains" and false chains (Table 4) have been removed. This plot corresponds to the line drawing described in section 2(c) and Figure 2(e). The skeleton is horrendously complicated because of the presence of neighboring chains. Consequently, the redundancy problem was tackled next.

Ribonuclease S crystallizes in space group $P3_121$ with one molecule in the asymmetric unit. There are six molecules in the unit cell and six symmetry operators† that must be applied to each point on the skeleton to search for redundancies. In addition, each of these operators must be applied with all possible combinations of the unit cell translations in order to be sure that all related points in the Cartesian space are generated, even if they are in adjacent unit cells in the crystal. Thus, 27 translation vectors (26 for the identity matrix) are applied for each symmetry operator. 1304 redundant point pairs were discovered in this way.

A simple point by point comparison of the redundant pairs of points to the center of gravity was tried first. The resulting skeleton can be seen in Figure 5(a). Comparison with a drawing of the ribonuclease S molecule as derived from the measured coordinates (see Fig. 6) demonstrates that important pieces of the molecule have been deleted and that bits of neighboring chains have been left in the skeleton. The ligase program could be used to reconnect virtually all the discontinuities in the molecule if a sufficiently large $d = 2{\cdot}5$ were stipulated. However, additional pieces of neighboring chains were also added by the ligase program under these conditions.

The more complicated criterion of first organizing the 2608 points into features was tried next. After dealing with one and two-point features on a point by point basis, 36 larger features containing 827 points remained. The correlation matrix of redundant points was compiled and the relative distances from the center of gravity were calculated for the n_{ij} points in each feature. The subfeature furthest from the center of gravity was deleted. The resulting skeleton is shown in Figure 5(b).

The ligase program was run allowing a chain length to end point distance ratio of 1·5 and a maximum length of 20 points for any new chain. The program, in fact, added only one chain of three linkages near the center of the molecule connecting two chain tips. Reference to the known structure of ribonuclease S confirms that this connection is valid. The final skeleton at this density level is shown in Figure 5(c). The improvement of this Figure over Figure 4 is enormous. The skeleton can now begin to provide useful structural information.

† This includes the identity matrix so that unit cell translations will be generated.

FIGS 5 and 6. Stereo presentations of the molecule. A stereoviewer is necessary in order to see these Figures in 3 dimensions. All of these drawings show the complete ribonuclease S molecule in the B-view of Richards & Wyckoff (1973). (Figs 3 and 4 show this same view and are, therefore, directly comparable.) The scale bar in all of these drawings corresponds to 5 Å.

FIG. 5. (a) Line drawing of ribonuclease S using a simple point by point comparison of redundant points to the center of gravity (see text). Important sections of the molecule are missing, such as: 90–95, 58–65, and 74–78. Pieces of neighboring chains have been left in the skeleton. Compare with Fig. 6. (b) Line drawing of a single ribonuclease S molecule derived using a feature by feature comparison to the center of gravity (see text). Only small and disconnected pieces of neighboring chains have not been removed. A short length of chain around residues 10 to 11 has been deleted erroneously. (c) The results of the ligase program on the skeleton of Fig. 5(b) using $d = 1{\cdot}5$ and $N = 20$ (see text, section 2(d)). The missing piece at residues 10–11 has been reinstated. No extraneous chains have been added to the skeleton.

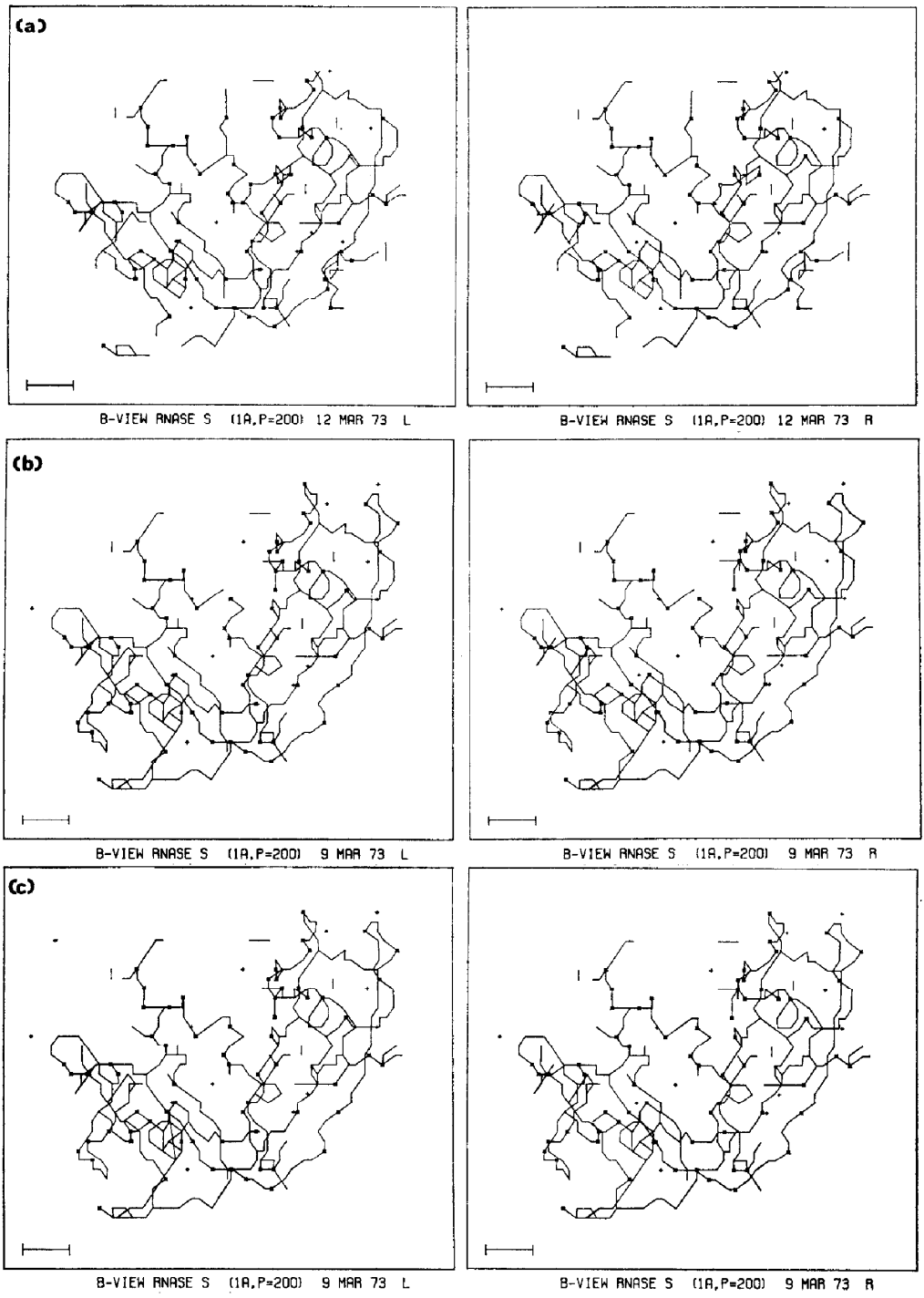
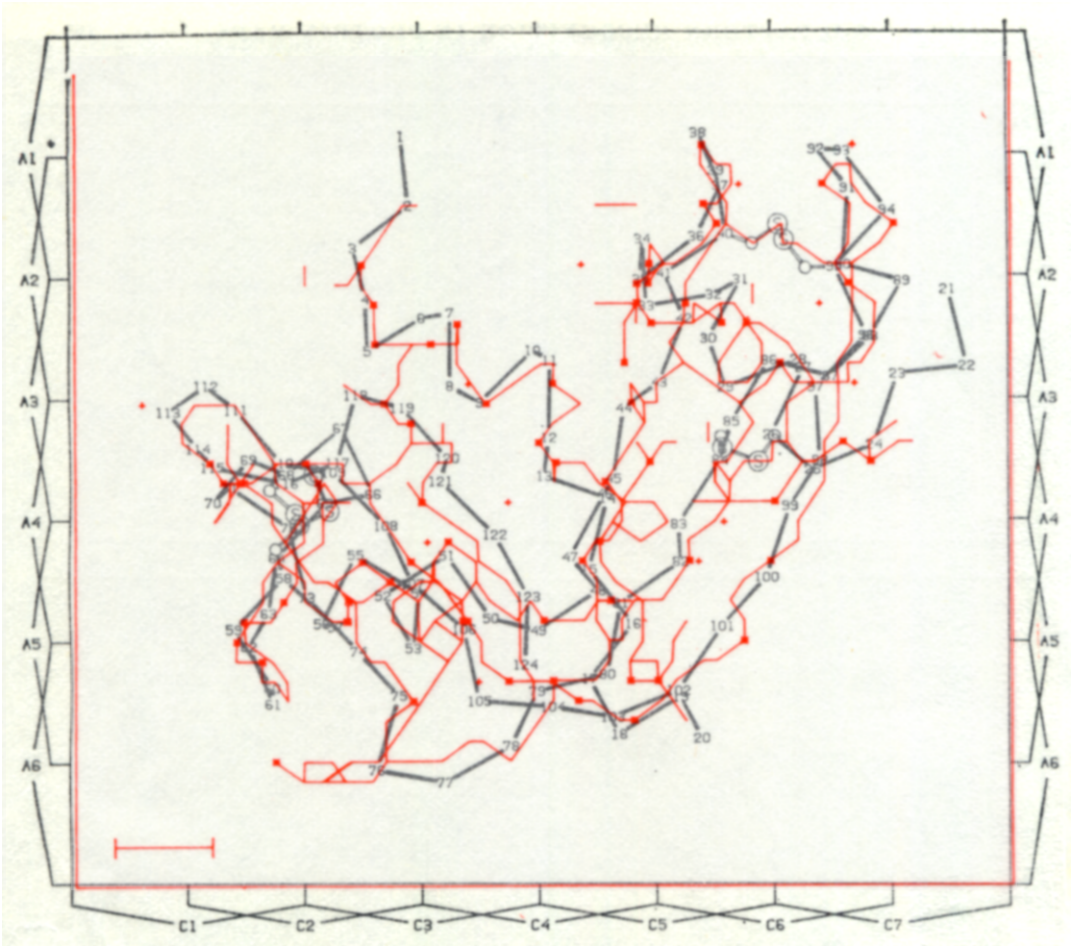
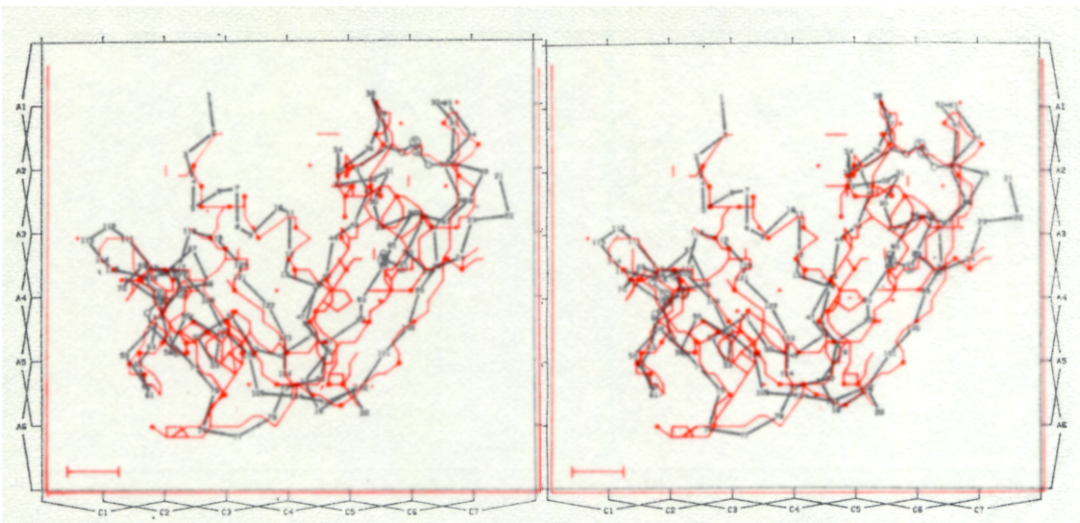


FIG. 5. *See facing page for legend*



RNASE-S.6D,VIEW B ALPHA CARBON CHAIN



RNASE-S.6D,VIEW B ALPHA CARBON CHAIN

RNASE-S.6D,VIEW B ALPHA CARBON CHAIN

FIG. 6. *See facing page for legend,*

The best demonstration of the usefulness of the skeleton at this stage is to show the skeleton superimposed on a drawing of the main chain of ribonuclease S derived from the measured co-ordinates (Richards & Wyckoff, 1973). In this drawing, Figure 6, the molecule is represented by straight lines connecting the measured α-carbon positions along the main chain. The residue numbers indicate the exact positions of the α-carbon atoms. The over-all fit of the skeleton to the molecule is quite good. Most of the skeleton follows the main chain reasonably well. The two β-structure regions in the molecule (residues 42–47, 80–86, 98–103 and 71–76, 105–110, 118–124) are clearly discernible in three dimensions. The four disulfide bridges are present: 40–95, 26–84, 58–110 and 65–72. The line connecting residues 42 to 85 is a strong hydrogen bond. There are several trouble areas that will require more subtle analysis at a later stage in the interpretation. The skeleton program occasionally leaves four or five-membered rings (see around residues 84 and 14 in Fig. 6 for examples) instead of a straight chain. These can be detected and eliminated without too much trouble. Bits of chain are missing because the main chain density is very weak in the map (residues 65–72). These situations may be resolved by using the ligase program to add the connection from lower density skeletons. α-Helical areas typically show complex cage patterns and small rings (around residues 51–53). This is due to the strong hydrogen bonds in an α-helix. This property can be used to detect α-helical regions later in the analysis of the skeleton.

Figure 6 demonstrates the over-all qualitative fit of the skeleton to the measured co-ordinates of ribonuclease S. Table 5 shows a piece of the skeleton between residues 95 and 107. The corresponding measured α-carbon co-ordinates, rounded off to the nearest grid unit, are placed next to the predicted position of the residue in the skeleton. However, it is difficult at this stage in the interpretation of the electron density map to design a quantitative comparison. In order to provide some indication of a quantitative nature, provisional α-carbon co-ordinates were deduced from the skeleton by application of the 4 Å rule. Each tip, side chain branch point and main chain branch point was listed. If the distance along the skeleton between any two of the above points exceeded 6 Å, the 4 Å rule was applied. In this way 140 points that are potential α-carbon positions were deduced for the ribonuclease S protein and 24 more for the S-peptide. (The small chains were ignored for this calculation. In fact, none forms any part of the main chain of the protein.) Of the 140 points, 95 α-carbons of the 104 residues in the S-protein were recognized, leaving nine residues in this chain unaccounted for. Of these, three are at the N-terminus and are very vague in the map and three more are in the area of residues 66–71, already noted above as having very weak density in the map. The remaining 45 points that the skeleton predicts arise mostly from the extra loops and cage structures left by the skeletonizing procedure as well as the occasional side chain (see Tyr-76 in Fig. 6). Many could be easily eliminated in a slightly more sophisticated analysis. In the S-peptide, 12 of the 20 α-carbons are predicted in a straightforward fashion. The C-terminus of the S-peptide is weak in the map and the skeleton appears to suggest

FIG. 6. The final skeleton of an isolated molecule of ribonuclease S (red lines) is superimposed upon a drawing of the molecule derived by connecting the measured α-carbon positions with straight lines (black lines). The close fit of the skeleton to the previously solved structure is demonstrated in this Figure. (See text for a more detailed comparison.)

TABLE 5

Comparison of ribonuclease S co-ordinates with skeleton map

Residue	Measured† α-Carbon positions X	Y	Z	Predicted Main chain positions X	Y	Z	Density	Length (Å)	Branch
Cys-95	13	27	11	13	28	11	4		
				13	27	11	2		
Ala-96	12	24	9	13	26	11	5		
				14	25	10	5		
				14	24	10	3		
				15	23	10	5		
				14	22	11	6		
Tyr-97	14	21	11	14	21	11	2		
				14	20	11	2		
				15	19	11	6		
Lys-98	15	18	10	15	18	11	4		
				15	17	11	4		
				15	16	12	7		
Thr-99	14	15	13	15	15	13	6		
				15	14	13	7		
Thr-100	16	12	13	16	13	14	6	18·9	Side chain
				17	12	15	3		
				17	11	16	5		
				16	10	16	2		
Gln-101	16	9	16	16	9	16	4	5·6	Side chain
				17	8	17	3		
Ala-102	18	6	17	17	8	18	2		
				18	7	19	2		
				19	6	19	3		
Asn-103	20	5	20	20	5	20	3	7·6	Side chain
				21	5	20	3		
				22	6	21	2		
Lys-104	24	5	21	23	6	21	2	3·7	Side chain
				24	7	22	2		
				25	7	22	5		
His-105	26	6	23	26	7	23	3	4·1	Side chain
				27	8	24	5		
				27	9	24	6		
Ile-106	27	9	24	27	10	25	5	4·1	Side chain
				28	11	25	6		
				29	12	25	5		
Ile-107	30	12	25	30	13	26	4	4·6	Side chain

† α-Carbon positions from co-ordinate list (Richards & Wyckoff, 1973) expressed in skeleton lattice co-ordinates rounded off to the nearest grid unit.

a somewhat different conformation from the original interpretation in this region. This region was, therefore, left out of the comparison of co-ordinates.

In total then, 107 out of 124 α-carbon positions occur in strong enough density to be predicted by the skeleton. The measured and predicted co-ordinates were compared by a least-squares program written by R. J. Fletterick. The transformation matrix and vector are:

0·9836	0·0143	0·0150	0·2396
−0·0266	0·9656	0·0021	−0·4743
0·0149	−0·0016	0·9658	0·8726

The matrix suggests an over-all volume decrease of about 8% in going from the measured to predicted co-ordinates. The translation vector is close to 1 Å. After the transformation, the mean absolute deviation of measured from predicted α-carbon co-ordinates is 1·37 Å. This value is surprisingly small considering that the skeleton

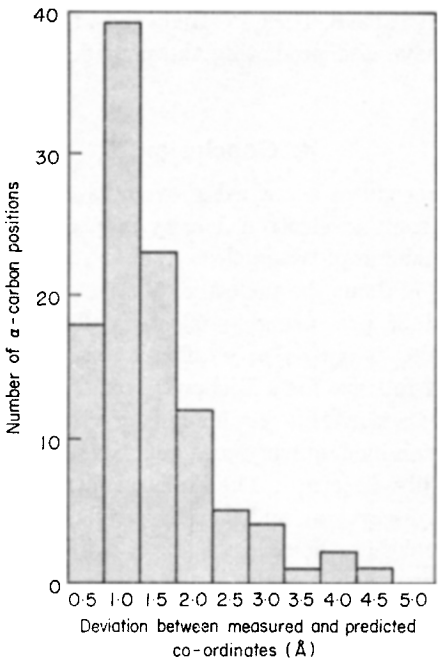


FIG. 7. A histogram of the number of α-carbon positions as a function of the size of the deviation of the measured positions from the predicted positions.

TABLE 6

Summary of program steps in the analysis of ribonuclease S and central processing unit (CPU) execution times

Step	CPU time (s)	Core (kbytes)	Section (description)
1. Conversion to Cartesian space†	211	330‡	2(a)
2. Skeletonization	102	140	2(b)
3. Removal of side chains	52	170	2(c)
4. Redundancy tests			2(d)
(a) Point-to-point	72	175	
(b) Feature-to-feature	493	180	
5. Ligase program (on 4(b))	15	162	2(e)
6. Prediction of α-carbon positions	10	130	2(f)

† Program written by J. M. Baldwin.

‡ Cartesian map was calculated in 2 parts and then merged. This step could be avoided if the map were calculated directly on a 1 Å grid with a fast Fourier program.

was calculated on a 1 Å grid and that the ribonuclease S co-ordinates have been refined only by treating the molecule as a rigid body (Fletterick *et al.*, unpublished results). A histogram of the number of α-carbon positions as a function of the size of the deviation between the measured and predicted co-ordinates is presented in Figure 7.

The feasibility of this system of programs depends, in part, upon the cost of running them. With the complicated computer systems in use today, there is no simple parameter that will summarize the total cost. As an over-all indicator, the central processing unit (CPU) time for each of the programs, as run on the IBM 370/155 for ribonuclease S, is listed in Table 6. The CPU times show that the cost of producing the skeleton of a single molecule and predicting the provisional α-carbon co-ordinates is quite reasonable.

4. Conclusion

The combination of procedures described above falls short of the goal of determining atomic co-ordinates from an electron density map without human intervention. Nevertheless, the final skeleton of ribonuclease S shows that the method can already provide useful information about the molecule.

The standard method of presenting an electron density map is on contoured sections. This format makes it particularly difficult to see larger regions of the molecule, especially if plotted full size for a Richards box. The skeletal representation of the density map provides a powerful graphic aid for visualizing the whole molecule. Large regions can be examined quickly and regular secondary structure features, such as β-structure, readily discerned. The skeleton formalism can ably supplement the Richards box or graphics system, and thereby speed the interpretation of the map.

Choosing the correct molecular boundary is often a difficult and tedious procedure. This is especially true when the molecules touch closely and the symmetry operators must be applied in order to distinguish the chains. The programs perform a valuable service by separating the central molecule from adjacent chains in the map. The list of redundant points allows one to identify quickly which part of the molecule corresponds to a piece of neighboring chain. Such identification is quite difficult to visualize in a Richards box if the plotted density map has been rotated out of the crystallographic co-ordinate system.

Most exciting of all is the ability of the skeleton to predict α-carbon positions with reasonable accuracy at this relatively crude stage of the interpretation of the map. The results generate both encouragement and incentive to continue the pursuit of automated interpretation of protein electron density maps.

I am indebted to Professor Frederic M. Richards, under whose tutelage this work was carried out, for advice and encouragement, and to Drs H. W. Wyckoff and R. J. Fletterick for many valuable discussions. I thank the Helen Hay Whitney Foundation for a postdoctoral fellowship. This research was supported in part by grants from the National Institutes of Health, GM12006 (to F. M. Richards) and GM10025 (to H. W. Wyckoff), and from Yale University.

REFERENCES

Diamond, R. (1966). *Acta Crystallogr.* **21**, 253–266.
Diamond, R. (1971). *Acta Crystallogr. ser. A*, **27**, 436–452.
Hilditch, C. J. (1969). *Machine Intelligence*, **4**, 403–420.

Katz, L. & Levinthal, C. (1972). *Annu. Rev. Biophysics Bioengineering*, **1**, 465–504.
Levitt, M. & Lifson, S. (1969). *J. Mol. Biol.* **46**, 269–279.
Pitman, R. M., Tweedle, C. D. & Cohen, M. J. (1972). *Science*, **176**, 412–414.
Richards, F. M. (1968). *J. Mol. Biol.* **37**, 225–230.
Richards, F. M. & Wyckoff, H. W. (1973). In *Atlas of Protein Structures* (Phillips, D. C. & Richards, F. M., eds), vol. 1, Oxford University Press, in the press.
Rossmann, M. G., Adams, M. J., Buehner, M., Ford, G. C., Hackert, M. L., Lentz, P. J., Jr, McPherson, A., Jr, Schevitz, R. W. & Smiley, I. E. (1971). *Cold Spring Harbor Symp. Quant. Biol.* **36**, 179–191.
Wyckoff, H. W., Tsernoglou, D., Hanson, A. W., Knox, J. R., Lee, B. & Richards, F. M. (1970). *J. Biol. Chem.* **245**, 305–328.