Acta Cryst. (1994). D**50**, 155-166

Molecular Scene Analysis: Application of a Topological Approach to the Automated Interpretation of Protein Electron-Density Maps

BY LAURENCE LEHERTE*

Department of Computing and Information Science, Queen's University, Kingston, Canada K7L 3N6

SUZANNE FORTIER

Department of Chemistry, Queen's University, Kingston, Canada K7L 3N6

JANICE GLASGOW

Department of Computing and Information Science, Queen's University, Kingston, Canada K7L 3N6

AND FRANK H. ALLEN

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England

(*Received* 1 *June* 1993; *accepted* 14 *October* 1993)

Abstract

Methods to assist in the spatial and visual analysis of electron-density maps have been investigated as part of a project in molecular scene analysis [Fortier, Castleden, Glasgow, Conklin, Walmsley, Leherte & Allen (1993). *Acta Cryst.* D**49**, 168-178]. In particular, the usefulness of the topological approach for the segmentation of medium-resolution (3 Å) maps of proteins and their interpretation in terms of structural motifs has been assessed. The approach followed is that proposed by Johnson [Johnson (1977). *ORCRIT. The Oak Ridge Critical Point Network Program*. Chemistry Division, Oak Ridge National Laboratory, USA] which provides a global representation of the electron-density distribution through the location, identification and linkage of its critical points. In the first part of the study, the topological approach was applied to calculated maps of three proteins of small to medium size so as to develop a methodology that could then be used for analyzing maps of medium resolution. The methodology was then applied to both calculated and experimental maps of penicillopepsin at 3 Å resolution. The study shows that the networks of critical points can provide a useful segmentation of the maps, tracing the protein main chains and capturing their conformation. In addition, these networks can be parsed in terms of secondary-structure motifs, through a geometrical analysis of the critical points. The procedure adopted for secondary-structure recognition, which was phrased in terms of geometry-based rules, provides a basis for a further automated implementation of a more complete set of recognition operations through the use of artificial-intelligence techniques.

1. Introduction

With the growing reservoir of information on crystal and molecular structures, the crystallographic image reconstruction exercise can now be reformulated as a scene-analysis exercise (Duda & Hart, 1973), as earlier envisaged by Feigenbaum, Engelmore & Johnson (1977). The concept of scene analysis, which has emerged from the area of machine vision, refers to the processes associated with the reconstruction, classification and understanding of complex images. Such analyses rely on the availability of *a priori* information, both in the form of templates and in the form of rules and heuristics, to locate and identify features in a scene and to provide a full interpretation of the scene. By analogy, we use the phrase molecular scene analysis (Fortier *et al.*, 1993) to refer to the processes associated with the reconstruction and interpretation of crystal and molecular structures. For example, crystal structure determination, molecular structure classification and prediction, and studies of recognition processes, can all be viewed as exercises in molecular scene analysis.

Over the last few years, we have been involved with the design and implementation of a knowledge-based system for molecular scene analysis. One of the main goals in building this system is to implement an approach to protein crystal structure determination that incorporates direct-methods and artificial-intelligence strategies and models protein crystal structure determination as an iterative and hierarchical scene-analysis task. Such an approach, however, requires the inspection and interpretation of a possibly large number of electron-density maps which, to be realizable, calls for substantial computer assistance. In particular, what are needed are techniques to assist in the spatial and visual interpretation of electron-density maps at all levels of resolution.

* Permanent address: Laboratoire de Physico-Chimie Informatique, Facultés Universitaires Notre-Dame de la Paix, Rue de Bruxelles 61, B-5000 Namur, Belgium.

Crucial to the interpretation of these maps are the location, identification and comparison of molecular features, particularly molecular shapes. Thus, we have embarked on a study of models to be used for the representation and segmentation of molecules, or molecular fragments, into primitive shapes.

The problem of finding accurate and efficient models for describing three-dimensional objects, for the purpose of pattern recognition, is one that has received considerable attention in machine vision as well as in chemistry. In their classical paper entitled 'Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes', Marr & Nishihara (1978) have defined three criteria useful for judging the effectiveness of a shape representation: (1) *Accessibility*. The representation should be derivable from the initial image at reasonable computing costs. (2) *Scope and uniqueness*. The representation should provide a unique description of all possible shapes in the domain of interest. (3) *Stability and sensitivity*. The representation should capture both the more general (less variant) and the subtle (of finer distinction) properties of shape.

Several models for capturing three-dimensional shape information have been proposed. We note, in particular, the general cylinder model introduced by Binford (1971) and expanded by Marr & Nishihara (1978) and Biederman (1985), and the skeletonization approach (Hilditch, 1969) which has been used extensively in protein crystallography following the work of Greer (1974). In this paper, we report on the use of a topological approach that has been applied both in chemistry and machine vision. In chemistry, the approach is used to characterize the shape properties of the electron-density distribution through the location and attributes of its critical points. Until now, it has been used principally for the detailed depiction of very high resolution theoretical and experimental electron-density maps (*e.g.* Bader, 1990; Howard, Hursthouse, Lehman, Mallinson & Frampton, 1992) although Johnson (1976, 1977*a*) had already recognized its potential usefulness for lower resolution maps of proteins.

The work described in this paper follows the approach proposed by Johnson (1976, 1977*a*) and indeed rests on the computer algorithm, *ORCRIT*, written by Johnson (1977*b*). The main goal of our study was to assess the usefulness of the topological approach for the interpretation of medium (~3 Å) resolution maps of proteins. Specifically, we wished to assess whether it could be used for both the segmentation of electron-density maps and the recognition of structural motifs in the segmented units. The study was divided into two parts. In the first part, the topological approach was applied to ideal electron-density maps of three proteins so as to develop a methodology that could then be used for the analysis of maps at a comparable resolution. This part of the study is presented in §2 of the paper. In the second part of the study, the developed methodology was applied to both theoretical and experimental maps of penicillopepsin. These results are presented in §3. The paper concludes with a discussion of the results, an assessment of the topological model against the criteria set forward by Marr & Nishihara (1978) and a description of future work.

2. Methodology for the topological analysis of protein electron-density maps

The interpretation of a protein electron-density map can be divided into two stages. The first stage consists of segmenting the map into distinct regions or 'blobs'. For example, depending on the resolution, the map may be segmented into protein and solvent regions and, within the protein region, into main and secondary chains. The second stage focuses on identifying the distinct regions/blobs through geometrical analysis and/or pattern matching with structural templates. Thus, the interpretation of a protein electron-density map can be viewed as a scene analysis whose goal is to find both the location and the identification of features of interest in the scene. In this section, we describe a suitable topological methodology and, in particular, attempt to identify rules, heuristics and/or templates that can then be used for the topological analysis of protein electron-density maps.

2.1. *Background theory and terminology*

The analysis of a mathematical function is often carried out by determining its first and second derivatives in order to describe its variations. Indeed, when equal to zero, first derivatives (or gradients) help to locate the extrema of a function, while second derivatives differentiate between minima and maxima which respectively lead to positive and negative values. The shape properties of an electron-density distribution can thus be represented through the location and identification of its so-called critical points [points $\mathbf{r}$ where the gradients $\nabla\rho(\mathbf{r})$ vanish]. Such an analysis requires the computation of the gradients, Hessians ($\mathbf{H}$), and Laplacians $\nabla^2\rho(\mathbf{r})$ of the electron density (Bader, 1990; Smith, Price & Absar, 1977). The eigenvalues of $\mathbf{H}$ provide information on the local curvature while the Laplacians give details about the local concentration (sign < 0) or depletion (sign > 0) of electron density.

The Hessian matrix of a continuous three-dimensional function such as the electron density, $\rho(\mathbf{r})$, is built from its second derivatives:

$$\mathbf{H}(\mathbf{r}) = \begin{vmatrix} \partial^2\rho/\partial x^2 & \partial^2\rho/\partial x\partial y & \partial^2\rho/\partial x\partial z \\ \partial^2\rho/\partial y\partial x & \partial^2\rho/\partial y^2 & \partial^2\rho/\partial y\partial z \\ \partial^2\rho/\partial z\partial x & \partial^2\rho/\partial z\partial y & \partial^2\rho/\partial z^2 \end{vmatrix}. \quad (1)$$

This real and symmetric matrix can be put in a diagonal form which corresponds to finding a rotation of the original coordinate system that aligns the new coordinate axes with the principal axes of the critical points

$$\mathbf{H}'(\mathbf{r}) = \begin{vmatrix} \partial^2\rho/\partial x'^2 & 0 & 0 \\ 0 & \partial^2\rho/\partial y'^2 & 0 \\ 0 & 0 & \partial^2\rho/\partial z'^2 \end{vmatrix}. \quad (2)$$

The three non-zero diagonal elements of $\mathbf{H}'$ are the eigenvalues of $\mathbf{H}$ and the Laplacian is the trace of $\mathbf{H}'$. The rank of $\mathbf{H}'$ is the number of non-zero eigenvalues, and the signature, s, is the algebraic sum of their sign. When the rank of $\mathbf{H}$ is 3, then four possible cases arise: $s = -3$ corresponds to a local maximum or *peak*, *i.e.* the electron-density function adopts maximum values along each of the three principal directions x', y' and z'; $s = -1$ corresponds to a saddle point or *pass* and there are only two negative eigenvalues; $s = +1$ corresponds to a saddle point or *pale* characterized by only one negative eigenvalue; and $s = +3$ corresponds to a local minimum or *pit*, *i.e.* the electron-density function adopts minimum values along each of the three principal directions.

The critical-point approach is thus a natural way to 'catch' the fluctuations of a three-dimensional electron-density function. This approach has been implemented in the computer program *ORCRIT* (Johnson, 1977*b*), which locates the critical points and their connectivity, thus providing a global representation of electron-density maps. It was developed as part of the Crysalis project (Terry, 1983), the goal of which was to build an expert system for the automated interpretation of protein electron-density maps. For the present study we have used a slightly modified version of *ORCRIT*, which has been linked to the *Xtal* program package (Hall & Stewart, 1990) and to the graphics program *KEMIT* (Vanderveken & Vercauteren, 1989). We have also applied the approach to maps of significantly lower resolution than were used in the Crysalis project.

In graph theory (see *e.g.* Wilson, 1990), a graph is a representation of a set of points and their connectivity. In particular, a connected graph consists of a set of points (vertices) in which any pair may be connected by a path formed by one or several successive edges in the graph. Many connected graphs contain rings of points and edges (circuits). In this case, it is possible to remove one edge from each circuit to leave a graph in which all points still remain connected. This graph, from which ring-closure connections are removed, is termed a minimal spanning tree. A set of spanning trees is described as a forest - a term which will be used later in this paper.

2.2. *Calculation of electron-density maps and critical-point networks*

Three protein structures retrieved from the Brookhaven Protein Data Bank (PDB; Bernstein *et al.*, 1977) were used in this analysis. Their chemical and structural parameters are reported in Table 1. From the available protein and solvent atomic coordinates and thermal parameters, structure factors were calculated and used to generate electron-density maps at a resolution of 3 Å. Since the structure-factor data were generated without any random errors, the only input parameters at the origin of the noise in the electron-density maps were those controlling the resolution and the grid size. All these calculations were performed using programs from the *Xtal* package.

Table 1. *Chemical and structural parameters of the proteins under study*

Name	Phospholipase A2 (bovine pancreas)[a]
Code	1BP2
Resolution (Å)	1.7
R factor	0.170
Cell parameters (Å,)	47.070 64.450 38.158 90 90 90
Space group	$P2_12_12_1$
Solvent/heteroatoms	109 H_2O, 2 $C_6H_{14}O_2$, Ca^{2+}
No. of amino acids	123
% α structure	54
% β structure	8
No. of turns	10
No. of disulfide bridges	7
Name	Ribonuclease T1 complex (*Aspergillus oryzae*)[b]
Code	1RNT
Resolution (Å)	1.9
R factor	0.180
Cell parameters (Å,)	46.810 50.110 40.440 90 90 90
Space group	$P2_12_12_1$
Solvent/heteroatoms	91 H_2O, $C_{10}H_{14}N_5O_8P$
No. of amino acids	104
% α structure	16
% β structure	35
No. of turns	10
No. of disulfide bridges	2
Name	Trypsin inhibitor (bovine pancreas)[c]
Code	4PTI
Resolution (Å)	1.5
R factor	0.162
Cell parameters (Å,)	43.100 22.900 48.600 90 90 90
Space group	$P2_12_12_1$
Solvent/heteroatoms	60 H_2O
No. of amino acids	58
% α structure	17
% β structure	33
No. of turns	3 as reported by Kabsch & Sander (1983)
No. of disulfide bridges	3

References: (*a*) Dijkstra *et al.* (1981); (*b*) Arni *et al.* (1987); (*c*) Marquart *et al.* (1983).

Using *ORCRIT*, a search for critical points was performed using density values greater than a cut-off value, ρ_{min}, which was set to approximately one third of the maximum density value, ρ_{max}. In the three cases studied, ρ_{max}, ranged from 2.4 to 2.7 e Å^{-3}. Johnson (1977*b*) observed that adjacent critical-point ...peak-pass-peak-pass-peak... sequences trace out ridge lines which tend to follow the polymeric protein molecule. Similarly, ...pit-pale-pit-pale-pit... sequences trace out valley lines which tend to separate adjacent molecules. The cut-off value was selected to allow the recognition of repeating ...peak-pass-peak-pass-peak... motifs and associate them with the polypeptide chain. The parameter ρ_{min} must be sufficiently high to limit the number of critical points and facilitate their interpretation, but it must also be sufficiently low to make the passes in between the peaks detectable. The appearance of two successive peaks without the presence of a pass may be symptomatic of chain segments that are close in space but not chemically linked. Thus the

detection of passes, together with peaks, facilitates the spanning tree interpretation. With the selected cut-off value, the search focused on peaks and passes since the density values of pits and pales were generally lower than ρ_{min}. Residue side chains are also detectable using this cut-off value and appear as short ...peak-pass-peak motifs jutting out from the ridge line (Fig. 1).

Again using *ORCRIT*, the collection of critical points was analyzed so as to generate a forest of minimal spanning trees. In building the trees, two parameters are used: Δr_{max} is the maximum distance between critical points below which they are considered connected; W is the weight of a connection between two critical points, calculated using:

$$W = 1/(\Delta r \Delta \rho F' \Delta s'), \quad (3)$$

where Δr is the distance between the two critical points; $\Delta \rho$ is the difference between the density values of the two critical points; $\Delta s'$ is equal to 1.1, 1.0, or 5.0 depending on whether the difference in the signature of the two critical points is equal to 0, 1, or greater than 1, respectively; F' is equal to $(3 - F_1 - F_2)$ where F_i denotes the normalized projection of the eigenvector of point i along the vector connecting the two critical points. In cases where two critical points are related by symmetry, the weight was calculated using

$$W = 1/\Delta r, \quad (4)$$

in order to ensure their connectedness.

2.3. *Map segmentation*

The first goal in the analysis of the collection of critical points is to segment the electron-density map. Specifically, at 3 Å resolution, we want to locate and identify the main chain of the protein, any disulfide bridges present in the structure and, if possible, side chains associated with the larger residues. As discussed previously, the main chain is characterized by a regular ...peak-pass-peak-pass... pattern. Thus, it is important to find optimal cut-off values for the density (ρ_{min}), the distance (ρ_{max}), and for W (W_{min}), so as to obtain a main spanning tree which is as long as possible and has only short branches jutting out of it (Fig. 1). Several sets of values for ρ_{min}, W_{min} and Δr_{max} were tested for the three proteins under study. For ρ_{min} it was found that a value corresponding to one third of ρ_{max}, *i.e.* 1.0 e Å^{-3}, was optimal. Values of 0.25 for W_{min} and 4 Å for Δr_{max} were selected. A description of the critical-point networks obtained using *ORCRIT* is presented in Table 2 for all three proteins. The table shows that the entire amino-acid sequence can be captured by the critical points in the main branch of a single principal spanning tree. Moreover, as shown in Fig. 2, the protein backbone as it appears in the crystal structure is well traced by the main branch of the principal spanning tree. Successive peaks in the main branch of the principal spanning tree correspond to successive residues with, in general,

Table 2. *Description of the critical-point (peaks and passes) networks for the calculated electron-density maps of 1BP2, 1RNT and 4PTI, at 3 Å resolution, as obtained using ORCRIT*

	1BP2	1RTNT	4PTI
Total No. of critical points (without elimination of symmetry-related points)	752	421	234
No. of critical points after elimination of symmetry-related points	532	365	196
Total No. of spanning trees	25	19	10
Total No. of critical points in the principal spanning tree	718	385	219
Total No. of peaks in the main branch of the principal spanning tree	134	109	60
No. of residues represented in the principal spanning tree	123	104	58
No. of residues in the main branch of the principal spanning tree represented by one and only one peak*	114	99	56

* Computed after elimination of symmetry-related points and merging of peaks separated by a distance shorter than 2 Å.

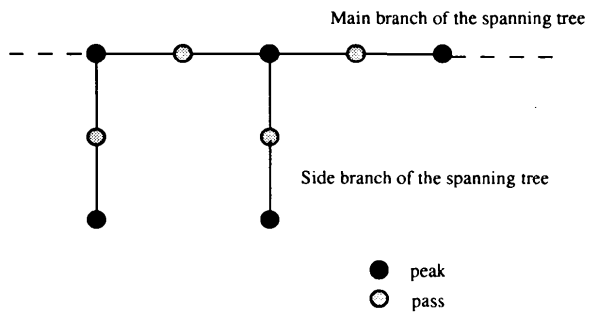

Fig. 1. Planar schematization of a typical critical-point motif as obtained using *ORCRIT* for the topological analysis of a calculated protein electron-density map at 3 Å resolution.

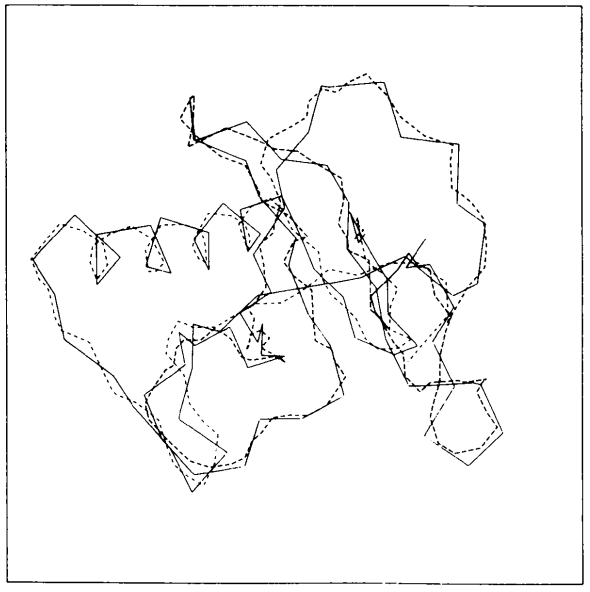

Fig. 2. Perspective view depicting the superimposition of the Cα chain for the protein structure 1RNT (solid line) to the corresponding main spanning tree (dotted line) as obtained using *ORCRIT*.

each residue being associated with only one peak. The superimposition of the Cα chain of the protein 1RNT to the corresponding main branch of the principal spanning tree shows that the topological representation of a 3 Å electron-density map effectively 'captures' the protein backbone conformation (Fig. 2). The presence of short branches along the principal spanning tree, associated with the residue side chains, is illustrated in Fig. 3, which was obtained by focusing on a portion of Fig. 2. Smaller spanning trees, composed of 1-11 critical points, were also detected. They reflect the presence of solvent molecules in the crystal structures.

The sequence of residues represented by the critical-point network is sometimes cross-linked by jumps from one residue to another, where the second residue is not directly connected to the first one. The jump frequencies for the three proteins under study are given in Table 3. Jumps tend to appear at the following three locations. (1) At the disulfide bridges. It is quite easy, however, to

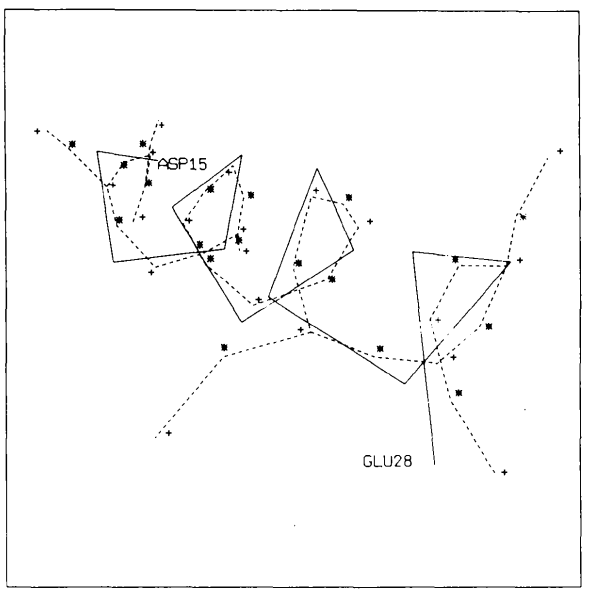

Fig. 3. Perspective view depicting a Cα helix of the protein structure 1RNT (solid line) and the corresponding critical-point network (dotted line) as obtained using *ORCRIT*. + and * denote peaks and passes, respectively.

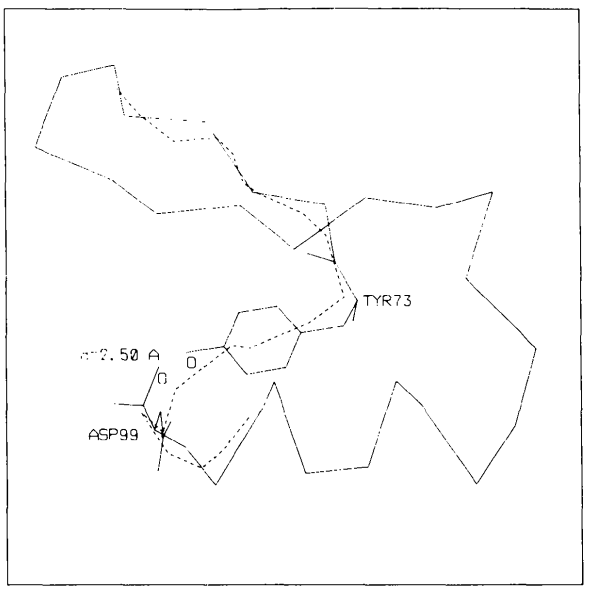

Fig. 5. Perspective view detailing a jump between hydrogen-bonded residues for the protein structure 1BP2 (solid line) and the corresponding critical-point network (dotted line) as obtained using *ORCRIT*.

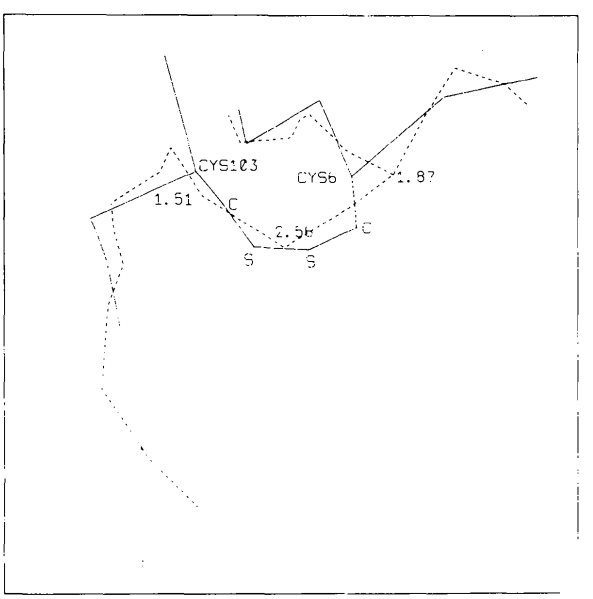

Fig. 4. Perspective view detailing a disulfide bridge for the protein structure 1RNT (solid line) and the corresponding critical-point network (dotted line) as obtained using *ORCRIT*.

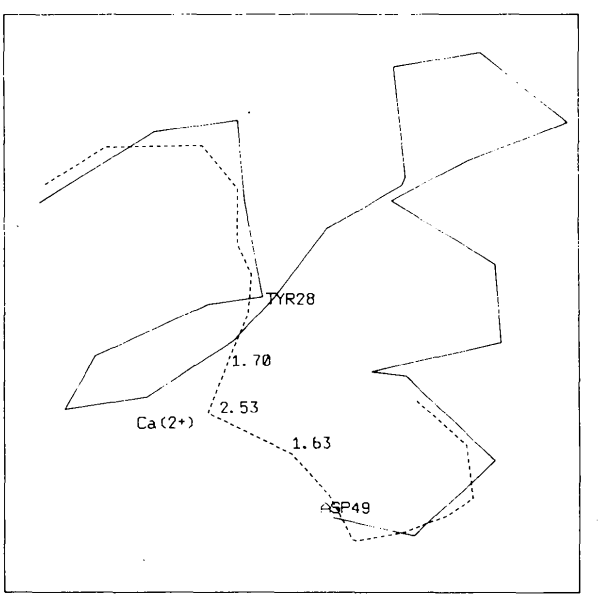

Fig. 6. Perspective view detailing a jump through a heteroatom for the protein structure 1BP2 (solid line) and the corresponding critical-point network (dotted line) as obtained using *ORCRIT*.

Table 3. *Jump frequencies in the main spanning trees derived from the calculated maps of the proteins 1BP2, 1RNT and 4PTI, at 3 Å resolution, as obtained using ORCRIT*

	1BP2	1RNT	4PTI
Total No. of jumps	4	1	1
Jumps occurring at disulfide bonds	2	1 (Fig. 4)	1
Jumps between non-adjacent residues	2 (Figs. 5 and 6)	0	0

detect such bridges as they are characterized by higher density values. Fig. 4 provides an example derived from the protein structure 1RNT. (2) At the 'contact zones' between a residue whose side chain is big and an unconnected but spatially close part of the protein structure. Fig. 5 displays such a case for the protein structure 1BP2. This particular jump occurs between two residues (Tyr73 and Asp99) interacting *via* a hydrogen bond with a short (2.5 Å) O—O distance. (3) Through a third partner, such as a heteroatom. This occurs, for example, between Asp49 and Tyr28 in the protein 1BP2 as represented in Fig. 6. Again, this type of jump can be detected by the higher density value of the peak associated with the heteroatom.

2.4. *Identification of structural motifs in electron-density maps*

2.4.1. *Secondary-structure differentiation.* The geometry of the critical-point patterns was analyzed in order to determine which parameters could help differentiate and recognize secondary structures. Specifically the aim is to be able to parse the main branch of the spanning trees in terms of helical and sheet motifs. The peaks of the spanning trees were assigned to secondary-structure motifs (helix, sheet or turn) according to the definition provided in the PDB files. Taking four, three or two adjacent peaks at a time in the main branch, torsion angles, intervector (bond) angles and distances were computed and statistically analyzed (Figs. 7-10). It is found that, taken alone, the distances (Fig. 7) and angles (Fig. 8) involving adjacent peaks are too similar to provide a clear differentiation between α-helices and β-sheets. However, distances between non-adjacent peaks, p_i and p_{i+3}, are shown to be useful (Fig. 9). Differentiation between secondary structures is, however, most easily made *via* the torsion angles (Fig. 10). Torsion angles for helices are almost all centered around 50-60°, while for β-sheets they are characterized by values that are either larger or smaller than +90 or -90° respectively (Fig. 10). However, turns have similar values to helices and the inspection of a larger neighbourhood of critical points is required in order to differentiate between these two types of secondary structure. To verify our geometrical results, the following simple calculations were carried out. Assuming that each peak corresponds to one and only one residue, the geometry of the peaks can be compared to that of the Cα atoms. Thus, idealized α-helix and β-sheet structures were built using the parameters reported in Table 4 (Schulz & Schirmer, 1988). Distances, bond angles and torsion angles for these idealized conformations are reported for compar-

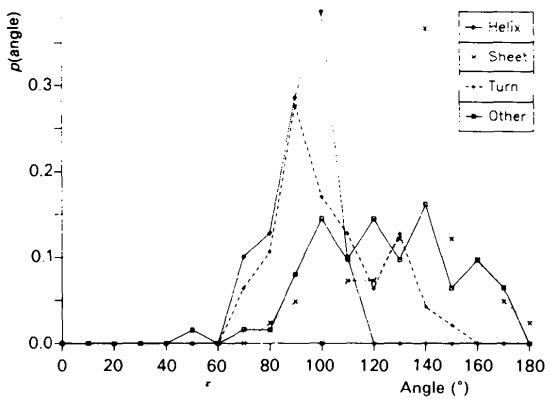

Fig. 8. Occurrence probability functions p(angle) of the bond angles (°) computed from the critical-point networks for the three proteins studied as obtained using *ORCRIT*. Functions are represented for helix, sheet and turn structures, and for all other motifs.

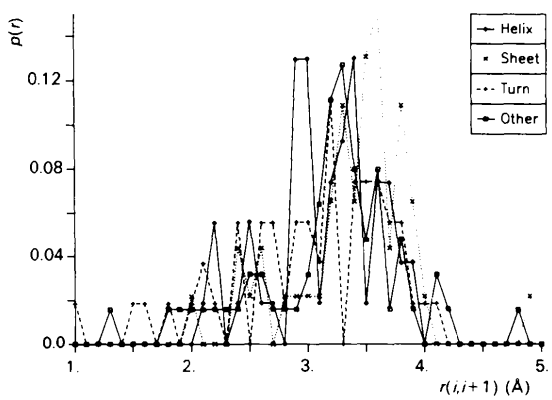

Fig. 7. Occurrence probability functions $p(r)$ of the peak-peak distances (Å) computed from the critical-point networks for the three proteins studied as obtained using *ORCRIT*. Functions are represented for helix, sheet and turn structures, and for all other motifs.

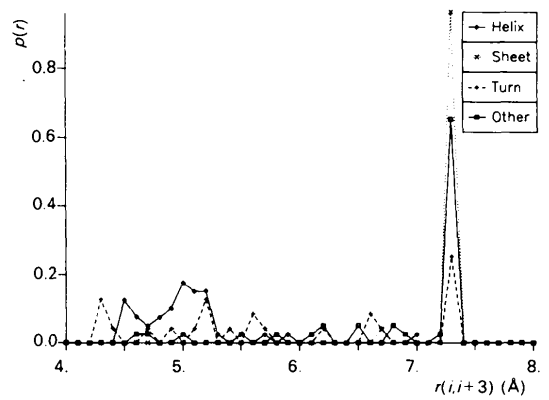

Fig. 9. Occurrence probability functions $p(r)$ of the $p_i - p_{i+3}$ distances (Å) computed from the critical-point networks for the three proteins studied as obtained using *ORCRIT*. Functions are represented for helix, sheet and turn structures, and for all other motifs.

Table 4. *Ideal α-helix and β-sheet structures as reported by Schulz & Schirmer* (1988)

	α-helix	β-sheet
Parameters		
No. of residues per turn	3.6	2.0
Rise between two residues along the helix axis (Å)	1.5	3.3
Helix radius (Å)	2.3	1.0
Results		
Distance between residues i and $i+1$ (Å)	3.8	3.9
Distance between residues i and $i+3$ (Å)	5.1	10.1
'Bond' angle ( )	90.2	117.6
Torsion angle ( )	50.1	180.0

ison (Table 4). The close correspondence between the idealized values (Table 4) and those resulting from our analysis (Figs. 7-10) confirms that indeed, in general, each peak in the main branch of the spanning trees corresponds to one and only one residue. However, since the peaks are not necessarily located precisely on the Cα positions, the values calculated from the peak positions fluctuate around the values calculated with the Cα positions.

2.4.2. *Identification of residues.* As mentioned earlier, disulfide bridges and heteroatoms can be located and identified easily through the higher density values of their corresponding peaks. In the present analysis, the Ca^{2+} ion in the 1BP2 structure and the disulfide bridges in all three proteins were located and identified readily within the networks of critical points (see Figs. 4 and 6). It is also easy to locate residues since, in general, each residue is associated with only one peak in the main branches of the spanning trees derived from 3 Å resolution maps. Further, many of the residue side chains appear as short branches in the trees. A preliminary analysis of these short branches was carried out to see if it was possible to identify specific residues from the sets of critical points. For each residue the occurrence frequency of short branches jutting out of the main backbones is listed in Table 5. Only a very broad distinction between

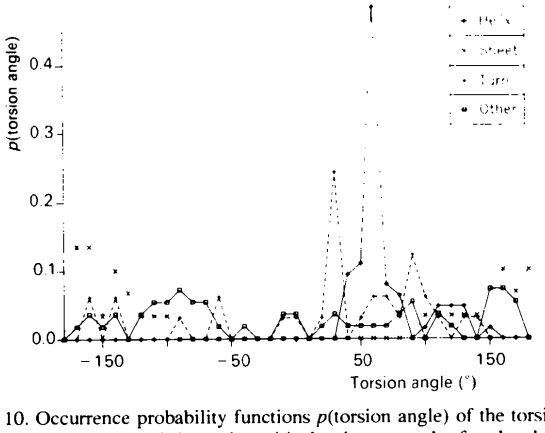

Fig. 10. Occurrence probability functions p(torsion angle) of the torsion angles (°) computed from the critical-point networks for the three proteins studied as obtained using *ORCRIT*. Functions are represented for helix, sheet and turn structures, and for all other motifs.

Table 5. *Occurrence frequency of short side branches jutting out of the main branch of the principal spanning trees for the protein structures 1BP2, 1RNT and 4PTI*

Three- and one-letter amino-acid codes	Ratio of branched residues over the total No. of residues	Percentage (%)
Ala (A)	0/19	0
Gly (G)*	1/24	4
Gln (Q)	1/6	17
Pro (P)	3/13	23
Ser (S)	7/26	27
Cys (C)	8/24	33
Glu (E)	6/13	38
His (H)	2/5	40
Lys (K)	7/17	41
Trp (W)	1/2	50
Arg (R)	5/9	56
Thr (T)	8/13	62
Ile (I)	7/9	78
Asn (N)	22/28	79
Tyr (Y)	16/20	80
Asp (D)	14/17	82
Phe (F)	10/12	83
Leu (L)	11/13	85
Val (V)	12/13	92
Met (M)	2/2	100

* This instance of a branched residue is due to the erroneous linkage of the peak associated with Gly making it appear as a short branch jutting out of the main branch of the spanning tree.

'small' and 'large' residues is possible from these data. A more detailed analysis on a larger sample of proteins is needed to provide definite and conclusive results. For many of the residues, the number of instances in the three proteins used here is too small to yield significant results. Also, thermal vibrations of the side-chain atoms must be taken into consideration.

3. Application of the topological methodology to the theoretical and experimental electron-density maps of penicillopepsin at 3 Å resolution

Theoretical and experimental maps of penicillopepsin, a 323-residue protein, were generated and analyzed in order to check on the applicability of the topological methodology outlined in §2. The resolution of the maps was selected at 3 Å, since we were primarily interested in assessing the usefulness of the topological approach for the interpretation of medium-resolution electron-density maps. The PDB parameters describing penicillopepsin are listed in Table 6.

3.1. *Map segmentation*

The theoretical electron-density map of penicillopepsin was generated using programs from the *Xtal* package, following the procedure outlined in §2.2. The experimental map was calculated using the Groningen *BIOMOL* crystallographic program package. The calculation included all reflections in the resolution range 10-3 Å. Observed structure-factor amplitudes and multiple isomorphous replacement phases with their

Table 6. *Chemical and structural parameters of penicillopepsin as reported in the Brookhaven Protein Data Bank (Bernstein et al., 1977)*

Name	Penicillopepsin (fungus) (James & Sielecki, 1983)
Code	3APP
Resolution (Å)	1.8
R factor	0.136
Cell parameters (Å, °)	97.370 46.640 65.470 90 115.4 90
Space group	*C*2
Solvent/heteroatoms	320 H_2O
No. of amino acids	323
% α structure	15
% β structure	52
No. of turns	21 as reported by Kabsch & Sander (1983)
No. of disulfide bridges	1

Table 7. *Description of the critical-point (peaks and passes) networks for the theoretical and experimental electron-density maps of penicillopepsin, at 3 Å resolution, as obtained using ORCRIT*

	Theoretical	Experimental
Total No. of critical points (without elimination of symmetry-related points)	2574	2141
No. of critical points after elimination of symmetry-related points	1576	1544
Total No. of spanning trees	37	223
No. of spanning trees considered as principal*	1	7
Total No. of critical points in the principal spanning trees	2528	1724
Total No. of peaks in the main branch of the principal spanning trees†	353	325
No. of residues represented in the principal spanning trees‡	321 (322)	278 (316)
No. of residues in the main branch of the principal spanning trees represented by one and only one peak†	267	176

* To be considered as a principal spanning tree, a sequence of critical points must represent at least five residues.

† Computed after elimination of symmetry-related points and merging of peaks separated by a distance shorter than 2 Å.

‡ Values in parentheses are the total number of residues represented in the complete critical-point networks.

figures of merit were used. The phases were derived from eight heavy-atom derivatives and had an overall figure of merit of 0.9 for all observed data to 2.8 Å resolution (Hsu, Delbare, James & Hofmann, 1977; James & Sielecki, 1983). *ORCRIT* was used to locate critical points in the electron-density maps and to generate forests of minimal spanning trees. Following the approach described in §2.2, cut-off parameters of W_{min} = 0.25, Δr_{max} = 4 Å and ρ_{min} = one third of ρ_{max} (ρ_{max} values ranged between 2.1 and 2.3 e Å^{-3}) were used for the topological analysis of both the experimental and theoretical maps. As was expected, the two maps displayed slightly different critical-point networks since the theoretical map is highly idealized while the experimental one is noisy. In particular, the noise present in the experimental map leads to additional clusters of critical points making the analysis of the minimal spanning trees more difficult. The critical-point networks obtained using *ORCRIT* are described in Table 7 and Fig. 11 for both the theoretical and experimental maps. In Fig. 11, residues represented within very small spanning trees (four peaks or less) are designated by

```
Seq. 1-50     A-A-S-G-V-A-T-N-T-P-  T-A-N-D-E-E-Y-I-T-P-  V-T-I-G-G-T-T-L-N-L-  N-F-D-T-G-S-A-D-L-W-  V-F-S-T-E-L-P-A-S-Q-
Theor.        Tree01.....01  * Tree01.....01 Tree01.....
Exper.        ***** Tree01.....01 * ***** * * *** *  Tree02.....02 *** Tree02.....02 Tree02....

Seq. 51-100   Q-S-G-H-S-V-Y-N-P-S-  A-T-G-K-E-L-S-G-Y-T-  W-S-I-S-Y-G-D-G-S-S-  A-S-G-N-V-F-T-D-S-V-  T-V-G-G-V-T-A-H-G-Q-
Theor.        .....tree01 Tree01.....
Exper.        ...02 Tree02.....02 * ******* Tree02..... .....*.....

Seq. 101-150  A-V-Q-A-A-Q-Q-I-S-A-  Q-F-Q-Q-D-T-N-N-D-G-  L-L-G-L-A-F-S-S-I-N-  T-V-Q-P-Q-S-Q-T-T-F-  F-D-T-V-K-S-S-L-A-Q-
Theor.        .....01 Tre01.....01 Tree01.....01 Tree01.....
Exper.        .02 Tree.02 Tree02***** * * *** Tree02.....02  Tree02.....02 Tree03

Seq. 151-200  P-L-F-A-V-A-L-K-H-Q-  Q-P-G-V-Y-D-F-G-F-I-  D-S-S-K-Y-T-G-S-L-T-  Y-T-G-V-D-N-S-Q-G-F-  W-S-F-N-V-D-S-Y-T-A-
Theor.        .....01 Tre01 Tree01..
Exper.        .....03 Tree02.....

Seq. 201-250  G-S-Q-S-G-D-G-F-S-G-  I-A-D-T-G-T-T-L-L-L-  L-D-D-S-V-S-Q-Y-Y-  S-Q-V-S-G-A-Q-Q-D-S-  N-A-G-G-Y-V-F-D-C-S-
Theor.        .....01 Tree01.....01 Tree01.....
Exper.        .....02 Tree04.....04 T04  Tree05..... .05  Tree...06 Tree.07

Seq. 251-300  T-N-L-P-D-F-S-V-S-I-  S-G-Y-T-A-T-V-P-G-S-  L-I-N-Y-G-P-S-G-D-G-  S-T-C-L-G-G-I-Q-S-N-  S-G-I-G-F-S-I-F-G-D-
Theor.        .....01 Tree01.....01 Tree01.....
Exper.        Tree04...04  Tree02.....02  ******* *** * * * Tre07 Tree04.....

Seq. 301-323  I-F-L-K-S-Q-Y-V-V-F-  D-S-D-G-P-Q-L-G-F-A-  P-Q-A
Theor.        .....01
Exper.        04 Tree02.....02
```

Fig. 11. Description of the critical-point networks of the theoretical and experimental electron-density maps of penicillopepsin, at 3 Å resolution, as obtained using *ORCRIT*. Principal spanning trees and smaller spanning trees are represented by dotted lines and asterisks, respectively. Jumps between non-connected residues are represented by spaces.

asterisks. It is seen that there is only one principal tree in the theoretical map, Tree01, which captures almost all the protein backbone. However, this tree is fragmented into pieces due to the fact that *ORCRIT* does not allow closed paths. Thus, if a jump occurs somewhere along the protein backbone, a break must appear in compensation. As expected, the results deteriorate when the approach is applied to the experimental map of penicillopepsin. Nevertheless, as shown in Table 7, 316 or 98% of the residues are captured by the critical points. Their linkage in a spanning tree representation results in several large trees which trace long continuous sections of the polypeptide backbone. The procedure yields, however, a large number of additional smaller trees and several jumps and breaks are detected. The breaks occur again because closed paths are not allowed in the current version of *ORCRIT*. The jumps, on the other hand, result from the presence of noise in the experimental map which causes residues no longer to be strictly associated with one peak only (as they were in the calculated maps), but rather with a small cluster of close peaks which are therefore spread over a larger area. This favors linkage between close but non-adjacent segments of the protein sequence (see *e.g.* Tree04 and Tree07). A three-dimensional representation of the critical-point networks for the experimental map of penicillopepsin is given in Fig. 12 where the Cα chain tracing is also depicted. It can be seen that the critical-point networks capture well the overall shape of the protein and trace accurately large sections of the main chain. Some of the larger discrepancies between the two representations arise from the presence of breaks or jumps as explained earlier. It should be noted, though, that small differences between the two representations are to be expected, even in ideal cases, because the peaks detected by *ORCRIT* are closer to the center of mass of the —Cα—CO—N— group of atoms than to the Cα atoms. A detailed analysis of the fit between the critical-point peaks and the polypeptide chain, including r.m.s.d.'s, is in progress and will be reported shortly.

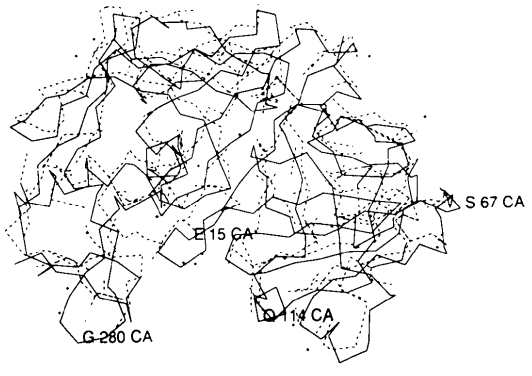

Fig. 12. Visual representation of the critical-point networks (dotted line) of the experimental electron-density map of penicillopepsin, at 3 Å resolution, as obtained using *ORCRIT*. The protein Cα backbone is shown for comparison.

3.2. *Identification of structural motifs: secondary-structure differentiation*

The next step in the study was to analyze the minimal spanning trees so as to parse them into secondary-structure motifs. For secondary-structure differentiation, it is the principal branches of the minimal spanning trees that must be analyzed. Before attempting to parse the trees into secondary-structure motifs, the following preprocessing was performed. Firstly, symmetry-coincident critical points were removed. Distances were then computed for sets of adjacent peaks, and peaks separated by a distance smaller than 2 Å were merged. The location of the resulting critical point, $\mathbf{r}_m$, was calculated using

$$\mathbf{r}_m = \sum_i^N \rho_i \mathbf{r}_i / \sum_i^N \rho_i \qquad (5)$$

where $\mathbf{r}_i$ and ρ_i are the location vector and the density value of the critical point i, respectively. N is the number of merged critical points.

The procedure adopted for the recognition of secondary-structure motifs was based on the sequential application of the following rules: (1) if the main branch of a principal spanning tree contains more than four peaks, then the peaks will be considered for secondary-structure recognition; (2) if two adjacent peaks are separated by a distance ≤ 5 Å, then the peaks will be considered to be connected (the value of 5 Å should not be confused with Δr_{max}, which applies to any pair of critical points - peaks and passes - during the construction of the spanning trees); (3) if four connected peaks have a torsion angle in the range 40-90° and a $p_i - p_{i+3}$ distance in the range 4.5-6.0 Å, then the internal peaks of the sequence will be classified as belonging to class *A* motifs (helix or turn); (4) if four connected peaks have a torsion angle in the ranges 100 to 180° or -100 to -180° and a $p_i - p_{i+3}$ distance in the range 7-8 Å, then the internal peaks of the sequence will be classified as belonging to class *B* motifs (sheet or other); (5) if three or more adjacent peaks are classified as belonging to class *A*, then these peaks will be subclassified as belonging to the helical motif subclass; (6) if three or more adjacent peaks are classified as belonging to class *B* motifs, then these peaks will be subclassified as belonging to the sheet motif subclass. Further, several peaks may be left unidentified after application of the recognition procedure if they fail to satisfy a proper subset of the rules leading to identification.

As described previously, the parameters that best differentiate between the secondary structures are the torsion angles and the $p_i - p_{i+3}$ distances. Before applying the recognition procedure, these parameters were calculated, statistically analyzed and compared to those of the three proteins reported in §2.4.1. As shown in Figs. 13-16, the results obtained were consistent with the earlier observations (Figs. 7-10), with larger

deviations being observed for the experimental map data. For example, in Fig. 13, the torsion angle values calculated for sets of four connected critical points belonging to an α-helix are not as tightly clustered around 60°, as previously observed for the calculated maps.

The recognition procedure was applied to both the theoretical and experimental maps of penicillopepsin and the results are listed in Tables 8 and 9. Table 8 shows that only a percentage of the peaks, between 55 and 58%, satisfies an appropriate subset of rules leading to identification. Of the unidentified peaks, about two thirds are associated with residues for which no secondary-structure labels were provided in the PDB. The majority of the peaks for which an identification was possible were, in fact, correctly identified. For the theoretical map, 28 peaks (14%) were wrongly identified. Of those, 24 were located at the extremities of an identified segment of peaks while the remaining four were associated with short jumps between non-connected residues. For the experimental map, 33 peaks (18%) were wrongly identified with 20 of those peaks being located at the extremities of identified segments and the remaining 13 being associated with jumps between non-

Table 8. *Statistics resulting from the application of the recognition procedure to the theoretical and experimental electron-density maps of penicillopepsin at* 3 Å *resolution*

	Theoretical	Experimental
Total No. of peaks in the principal spanning trees	1395	965
No. of peaks used in the identification procedure	714	477
No. of peaks after merging	353	325
No. of identified peaks	206	180
No. of correctly identified peaks	178	147
No. of incorrectly identified peaks	28	33
No. of unidentified peaks	147	145
No. of unidentifiable peaks	101	90
No. of peaks for which an identification exists	46	55

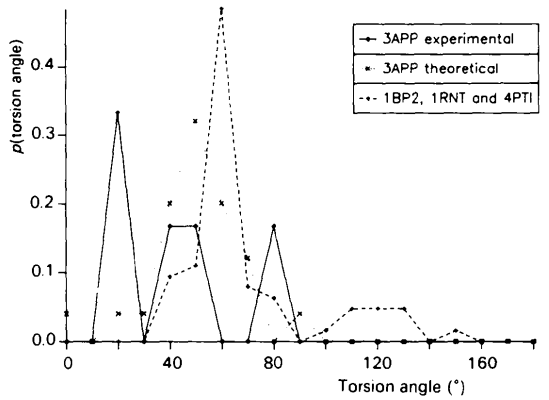

Fig. 13. Occurrence probability functions p(torsion angle) of torsion angles (°) for the 3APP helix structure (both experimental and theoretical) compared with the theoretical data obtained from the three previous protein structures.

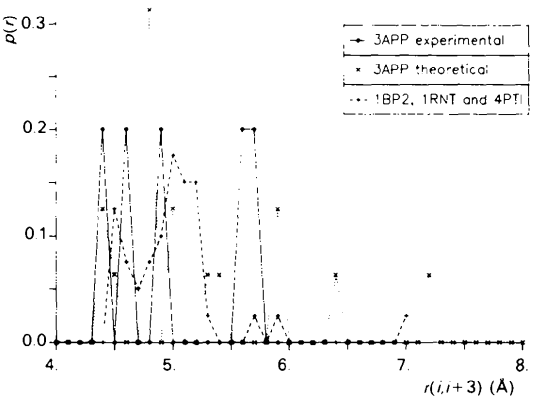

Fig. 15. Occurrence probability functions $p(r)$ of p_i - p_{i+3} distances (Å) for the 3APP helix structure (both experimental and theoretical) compared with the theoretical data obtained from the three previous protein structures.

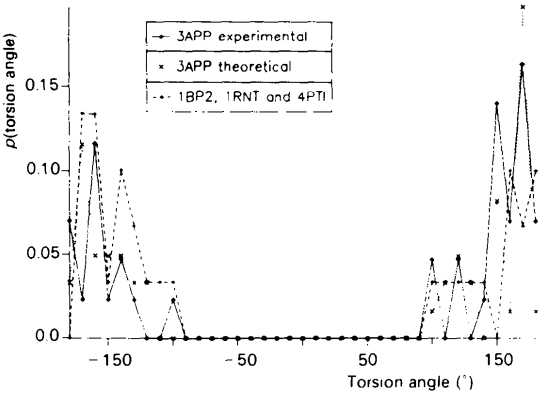

Fig. 14. Occurrence probability functions p(torsion angle) of torsion angles (°) for the 3APP sheet structure (both experimental and theoretical) compared with the theoretical data obtained from the three previous protein structures.

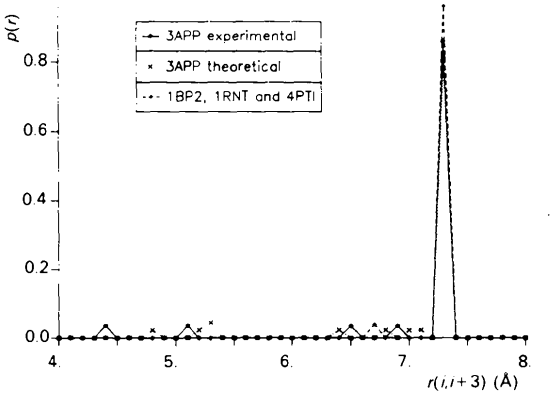

Fig. 16. Occurrence probability functions $p(r)$ of p_i - p_{i+3} distances (Å) for the 3APP sheet structure (both experimental and theoretical) compared with the theoretical data obtained from the three previous protein structures.

Table 9. *Analysis of the critical-point networks of the theoretical and the experimental electron-density maps of the penicillopepsin structure, at 3 Å resolution, as a function of the secondary structure*

	Theoretical		Experimental	
	α-helix	β-sheet	α-helix	β-sheet
No. of peaks belonging to a secondary structure and to a sequence of more than four peaks	46	178	38	159
No. of correctly identified peaks belonging to a secondary structure and to a sequence of more than four peaks	43	135	18	123

connected residues. For the experimental map, all of the incorrectly identified peaks were wrongly classified as sheet motifs, while for the theoretical map, the corresponding figure was 75%. The remaining incorrectly identified peaks were classified as helix motifs and were all located at the extremities of critical-point sequences. Table 9 summarizes the results of the recognition procedure for helix and sheet motifs comprising more than four residues. Again, it can be seen that several peaks remain unidentified. For the helices, the 3 and 20 unidentified peaks in the theoretical and experimental maps, respectively, were all located at the extremities of a critical-point segment. For the sheet motifs, the 43 and 36 unidentified peaks were either located at the extremities of a critical-point segment or else did not meet the geometrical requirements of sheet motifs.

4. Discussion and concluding remarks

Our results on the application of critical-point mapping to the calculated maps of four proteins at 3 Å resolution show that this approach has the potential to be a useful tool for the segmentation and interpretation of medium-resolution electron-density maps of proteins. The spanning trees derived from the sets of critical points accurately trace the main chains of the proteins and capture their conformation. Furthermore it is observed that at 3 Å resolution, each residue in the polypeptide chain is, in general, associated with only one peak in the main branch of spanning tree. The topological approach also permits the location and identification of secondary-structure motifs through the analysis of the geometrical parameters - distances, bond angles and torsion angles - of the peaks in the main branches of the spanning trees. Not all of the secondary-structure motifs are recognized. It should be noted, though, that the procedure does not attempt to recognize all the motifs but rather a portion for which a reasonably accurate identification might be possible.

With respect to the criteria defined by Marr & Nishihara (1978), the topological approach can be assessed as follows.

Accessibility. The representation is readily derivable from electron-density maps through the algorithm coded in *ORCRIT*. A table of computation times is given in Table 10 for a test case with a unit cell comprising 108 × 54 × 80 grid points. Although computer intensive, the calculations are certainly feasible.

Table 10. *Computation time for a test case using ORCRIT*

The calculations were carried out on the theoretical electron-density map of an asymmetric unit of the penicillopepsin unit cell, at 3 Å resolution. The parameter ρ_{min} was set to 2 e Å $^{-3}$ in the test.

	SPARCstation		IBM-RISC6000	
	Sun 3/50	SLC	Model 730	Cray S-MP/11
Time (s)	6090	238	104	35

Scope and uniqueness. The depiction of electron-density maps in terms of critical points is able to capture any physically significant arrangements of electron-density values. The location and attributes of the critical points are uniquely defined for each map. However, the presence of noise introduces additional critical points making the comparison between maps difficult. The connectivity between the critical points and the construction of spanning trees depend on the input parameters. Thus, the spanning tree representation is a unique representation, provided that these input parameters remain constant.

Stability and sensitivity. The critical-point representation is able to capture coarse and fine characteristics of molecular shapes. In this study at 3 Å, it was shown that the peaks in the main branches of the trees capture shape information at the residue level. At higher resolution, peaks become associated with individual atoms and the critical-point networks capture shape information at the atomic level.

The initial test on the application of the topological approach to a 3 Å resolution experimental map of penicillopepsin has identified some problems arising from the presence of noise and errors in the map. Several jumps between critical points associated with non-connected residues were detected resulting in breaks and shorter spanning trees. Nevertheless, the approach was shown to capture well the global shape features of the protein and large sections of its polypeptide chain. In applying the secondary-structure recognition procedure, it was seen that a smaller fraction of secondary-structure motifs could be located and identified. However, a high percentage of the peaks that were classified were classified correctly. Thus, the procedure can provide useful partial structure information which can then serve to facilitate the iterative image reconstruction exercise.

Our work is now concentrating on further improving the performance of critical-point mapping on experimental maps. Thus, a more in-depth analysis of the parameters affecting the detection and linkage of critical points in these maps is being carried out. We are reviewing the methodology used for the construction of the spanning trees with the aim of identifying rules

or heuristics that could improve the performance of the linkage procedure in maps where noise may be present. Also, modifications are being introduced in *ORCRIT* so that the program can automatically focus in regions where a more detailed (lower ρ_{min}/ρ_{max} ratio) analysis is needed. In addition, we will investigate the possibility of combining the critical-point approach with the skeletonization one, using the latter as a guide for the selection of regions for which a more detailed analysis would be warranted. We also intend to assess the performance of the critical-point mapping approach on maps of lower and higher resolutions. An initial calculation has already been carried out on a hexadecapeptide whose secondary structure consists of an α-helix followed by a 3_{10}-helix followed by three β-bends (Karle, Flippen-Anderson, Sukumar & Balaram, 1987). As expected, at high resolution atomic level information is well captured by the networks of critical points. At 5 Å resolution, in agreement with the observations reported by Lesk (1991), it is not possible to identify unambiguously the sheet motifs. However, helices can be located and identified in the spanning trees.

A simple procedure for the recognition of secondary-structure motifs was formulated in this study. The refinement and improvement of this procedure will be the subject of future studies. In particular, the analysis leading to the formulation of rules will be carried out on a larger sample of proteins. The procedure itself will also be applied to a larger sample of cases so that it can be refined, for example by attaching weights or 'degrees of confidence' to the rules. Finally the secondary-structure recognition procedure will be implemented using artificial-intelligence techniques.

The aim of the work presented in this paper is to develop an approach for the automated interpretation of protein electron-density maps. Critical-point mapping provides a representation through which automated segmentation and motif-recognition routines can be implemented. Additional work is presently being carried out in order to refine and assess these routines so that they may be applied successfully to experimental maps.

We wish to thank Carroll K. Johnson for providing the *ORCRIT* program and for many helpful discussions and Marie Fraser for making available to us the penicillopepsin experimental map data. We are also grateful to the Belgian National Foundation for Scientific Research (FNRS) and IBM-Belgium, and the Facultés Universitaires Notre-Dame de la Paix for the use of the Namur Scientific Computing Facility. Financial assistance from the Natural Sciences and Engineering Research Council of Canada in the form of operating grants to JG and SF, and from NATO in the form of a collaborative research grant to JG, SF and FHA is gratefully acknowledged. LL thanks the FNRS and the NATO Scientific Division for financial support.

References

ARNI, R., HEINEMANN, U., MASLOWSKA, M., TOKUOKA, R. & SAENGER, W. (1987). *Acta Cryst.* B**43**, 548-554.

BADER, R. W. (1990). *Atoms in Molecules - A Quantum Theory*. Oxford: Clarendon Press.

BERNSTEIN, F. C., KOETZLE, T. F., WILLIAMS, G. J. B., MEYER, E. F., BRICE, M. D., RODGERS, J. R., KENNARD, O., SHIMANOUCHI, T. & TASUMI, M. (1977). *J. Mol. Biol.* **112**, 535-542.

BIEDERMAN, I. (1985). *Comput. Vis. Graph. Image Process.* **32**, 29-73.

BINFORD, T. O. (1971). *IEEE Systems Science and Cybernetics Conference*, Miami, Florida, USA.

DIJKSTRA, B. W., KALK, K. H., HOL, W. G. J. & DRENTH, J. (1981). *J. Mol. Biol.* **147**, 97-123.

DUDA, R. O. & HART, P. E. (1973). *Pattern Recognition and Scene Analysis*. New York: John Wiley.

FEIGENBAUM, E. A., ENGELMORE, R. S. & JOHNSON, C. K. (1977). *Acta Cryst.* A**33**, 13-18.

FORTIER, S., CASTLEDEN, I., GLASGOW, J. I., CONKLIN, D., WALMSLEY, C., LEHERTE, L. & ALLEN, F. H. (1993). *Acta Cryst.* D**49**, 168-178.

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

HALL, S. R. & STEWART, J. M. (1990). Editors. *Xtal 3.0 Users Manual*. Univs. of Western Australia, Australia, and Maryland, USA.

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

HOWARD, S. T., HURSTHOUSE, M. B., LEHMAN, C. W., MALLINSON, P. R. & FRAMPTON, C. S. (1992). *J. Chem. Phys.* **97**, 5616-5630.

HSU, I.-N., DELBARE, L. T. J., JAMES, M. N. G. & HOFMANN, T. (1977). *Nature (London)*, **266**, 140-145.

JAMES, M. N. G. & SIELECKI, A. R. (1983). *J. Mol. Biol.* **163**, 299-361.

JOHNSON, C. K. (1976). Proc. Am. Crystallogr. Assoc. Meet., Evanston IL, USA. Abstract B1.

JOHNSON, C. K. (1977*a*). Proc. Am. Crystallogr. Assoc. Meet., Asilomar CA, USA. Abstract JQ6.

JOHNSON, C. K. (1977*b*). *ORCRIT. The Oak Ridge Critical Point Network Program*. Chemistry Division, Oak Ridge National Laboratory, USA.

KABSCH, W. & SANDER, C. (1983). *Biopolymers*, **22**, 2577-2637.

KARLE, I. J., FLIPPEN-ANDERSON, J., SUKUMAR, M. & BALARAM, P. (1987). *Proc. Natl Acad. Sci. USA*, **84**, 5087-5091.

LESK, A.M. (1991). *Protein Architecture*, edited by D. RICKWOOD & B. D. HAMES. Oxford: IRL Press.

MARQUART, M., WALTER, J., DEISENHOFER, J., BODE, W. & HUBER, R. (1983). *Acta Cryst.* B**39**, 480-490.

MARR, D. & NISHIHARA, H. K. (1978). *Proc. R. Soc. London Ser. B*, **200**, 269-294.

SCHULZ, G. E. & SCHIRMER, R. H. (1988). *Principles of Protein Structure*. New York: Springer-Verlag.

SMITH, V. H. JR, PRICE, P. F. & ABSAR, I. (1977). *Isr. J. Chem.* **16**, 187-197.

TERRY, A. (1983). Thesis, Stanford Heuristic Programming Project, Stanford Univ., USA.

VANDERVEKEN, D. J. & VERCAUTEREN, D. P. (1989). *Kemit - a Molecular Graphics System*. Release 1.2. Facultés Univs. Notre-Dame de la Paix, Namur, Belgium.

WILSON, R. J. (1990). *Introduction to Graph Theory*. New York: Longman.