Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

High-resolution crystallographic map interpretation

Thomas J. Oldfield

Accelrys Inc., Department of Chemistry, University of York, Heslington, York YO10 5DD, England

Correspondence e-mail: tom@ysbl.york.ac.uk

Received 19 November 2001
Accepted 25 March 2002

This article describes a method for rapid interpretation of high-resolution crystallographic electron-density maps. The implemented algorithm searches for fragments of structure found in proteins and links these together to identify uniquely the atomic structure and sequence order of the atoms within a protein. The algorithm uses a two-dimensional sub-graph isomorphism method with a subsequent post-processing step to screen the results to find correct three-dimensional solutions to each search fragment. The final screening of ambiguous solutions found is performed using incomplete difference distance matrices. The algorithm has an intrinsic error-correction technique that is necessary for analysis of experimental data and should be applicable to a number of fields of bioinfomatics.

1. Introduction

The process of high-resolution macromolecular crystallography requires that we interpret an electron-density map to uniquely identify each peak in the map in terms of its atom name, which residue it belongs to and its sequential order within the protein. This is a pattern-recognition problem and requires that we identify known entities within the experimental data and thus interpret the protein structure. The interpretation of electron density has been approached using a number of methods. The most used and successful map-interpretation program for high-resolution data is *trace* and *wARP* (Perrakis *et al.*, 1999), although it is rather slow as it takes many hours and requires cyclic progressive interpretation. Other methods of automated map interpretation include coordinate conformation matching (Wang, 2000) and pathway analysis (Jones & Kjeldgaard, 1997; McRee, 1999; Oldfield, 1996). The method presented is based on two-dimensional sub-graph isomorphism with subsequent screening using incomplete difference distance matrices. Its main advantage over other density-interpretation methods is its ability to interpret data at approximately 3000 atoms per second. The method is approximately linearly scalable with respect to the number of residues within the protein. This is because the number of start points for each search is dependent on the number of branch points within the map, although errors will have some effect as they can increase the graph complexity. The method can also be adapted to search for nucleic acids or ligand structures within experimental data.

The aim of the method described is to identify fragments of chemical structure as rapidly as possible without restriction arising from internal conformational flexibility of the fragments. The two-dimensional (2D) sub-graph isomorphism method employed is based on 2D connectivity matrices analysed using a cross-peak search method [similar to that used to interpret nuclear magnetic resonance (NMR) data]. The identification of the correct 2D graph of a molecular structure does not uniquely identify three-dimensional (3D) structure. Therefore, the recognition of 3D molecular coordinates requires that further analytical techniques be implemented to remove ambiguity from the 2D solutions. Incomplete difference distance matrices (Wenger & Smith, 1982; Crippen & Havel, 1988) are used for this purpose, where terms for 1–2, 1–3 and restrained 1–4 bonding interactions are used to screen the results. This does not represent a rigorous mathematical solution to structure identification, but provides a rapid way of solving the problem of high-resolution map interpretation. Since the application was designed for the analysis of experimental data, it has a number of ways of handling missing and extra points within this data. These incomplete solutions are given lower weight than a complete solution, but enable maps with error to be correctly interpreted. The method described here has been designed for interpretation of electron-density maps, but is likely to be suitable for a number of problems where the edge density (Babel, 1991) is low. It is therefore likely to have an application to database searching of small molecules and other areas of bioinformatics.

High-resolution electron-density maps of proteins are large, generally containing thousands of peaks. If the peaks in the map are joined up using a bonding algorithm, then a mathematical graph is obtained containing nodes and edges. The graph also contains a number of errors. The main error arises from data measurement and manifests itself as missing and additional peaks within the graph; these must be handled to complete the analysis of electron density. There are additionally errors in the position of the peaks that represent atoms, but this does not pose a problem for the analysis. The sub-graphs that need to be found within the protein crystallographic data are groups of atoms that form the basic building blocks of a protein. The example sub-graphs implemented are for a partial tripeptide (Fig. 1) and 18 amino-acid side-chain atom groups (*R* groups) found in proteins. In fact, any small group of atoms can make up a sub-graph, for example nucleic acid residues or ligands, as the analysis is general for a small number of connected atoms. Fig. 1 shows the covalent bond structure of a partial tripeptide chemical group, where the atoms (nodes) are joined by covalent bonds (edges). The partial tripeptide represents an extremely important sub-graph construct of a protein, because in addition to finding the position of the basic polymer units, amino acids, it provides the information necessary to create the polymer pathway. This is because solutions of a search with this sub-graph must overlap, allowing the stitching together of the polymeric structure of a protein.

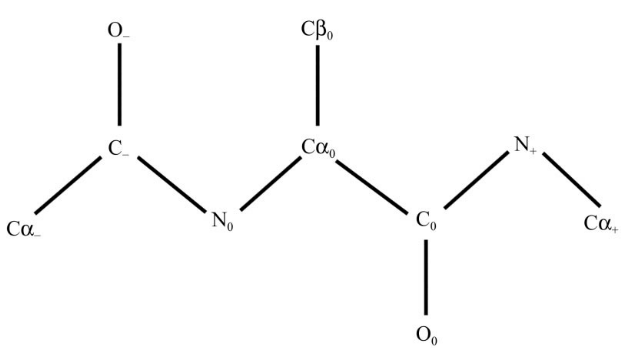

Figure 1
Atom names and bonding for the partial tripeptide search sub-graph.

2. Methods

The interpretation of high-resolution crystallographic data assumes that we have significant phase information with resolution better than about 2.2 Å (Perrakis *et al.*, 1999) to be able to carry out interpretation by peak searching. If the experimental information is satisfactory, then we need only carry out a peak search of the electron density to determine as many atom sites as there are to be expected from the known amino-acid content of the protein. It is invariably sensible to overestimate the number of atom sites to take into account the presence of localized solvent positions within the data and also because it is easier to reject nodes in a graph than identify missing nodes. These possible atom sites or nodes are then connected up by bonds, or edges, using lattice non-bond searching in the range 0.8–1.8 Å. This range of distances will correctly define the chemical structure of most molecules without joining atoms that are only in van der Waals contact. The exception to this rule is when H atoms are involved, but since X-ray protein crystallography does not provide significant detail on the position of H atoms they can be neglected.

Depth			1			2			3		
Atom		$C\alpha_0$	N_0	C_0	$C\beta_0$	C_-	N_+	O_0	O_-	$C\alpha_-$	$C\alpha_+$
	$C\alpha_0$		●	●	●						
1	N_0	●				●					
	C_0	●					●	●			
	$C\beta_0$	●									
2	C_-		●						●	●	
	N_+			●							●
	O_0			●							
3	O_-					●					
	$C\alpha_-$					●					
	$C\alpha_+$						●				
Column Sum		3	2	3	1	3	2	1	1	1	1

Figure 2
Connectivity matrix for the partial tripeptide search sub-graph in Fig. 1. The start point is $C\alpha_0$ and all points directly bonded are considered as depth = 1 nodes.

2.1. Preparation steps

An initial preparation step is required for each search sub-graph. There are three properties of the search sub-graph that need to be determined: the first is a starting point, the second is the connectivity matrix and the third is the incomplete distance matrix. The starting point can be any point in the sub-graph, but it is useful to define one that has significantly unusual connectivity and which minimizes the overall depth of subsequent analysis. If the partial tripeptide (Fig. 1) chemical group is considered, then the central atom $C\alpha_0$ is defined as the starting point and this node generally has three edges attached and is the mid-point within the sub-graph. The second stage is to define the connectivity matrix for this sub-graph from the starting point or $C\alpha_0$ atom. The node-connectivity matrix for the partial tripeptide is shown in Fig. 2 and is simply calculated using a depth-search algorithm using the connectivity list from the starting point selected. This figure shows a grid with a black circle where there is an edge between two nodes determined from the row and column index. The connectivity matrix for the partial tripeptide atomic structure has a third-order depth; that is, the furthest node from the starting point is three edges away. The 'column sum' is defined as the number of black circles in a column and defines the number of edges to each node. The preparation steps are complete with the determination of the incomplete distance matrix from the 3D coordinates of the search sub-graph. The tripeptide is a 3D chemical structure with certain fixed distances between some atoms. All 1–2 distances (bonds) are restrained and so are all 1–3 distances (angles) owing to the covalent nature of the constituent atoms. In many groups of atoms, we can also define a number of 1–4 (torsional) interactions that are highly restrained. A 1–4 interaction is restrained if the chemical bond between atoms 2 and 3 has bond order greater than 1. For the partial tripeptide example in Fig. 1, this includes the delocalized peptide bonds $C_{-}{-}N_0$ and $C_0{-}N_{+}$, resulting in 1–4 target values for the interactions $C\alpha_0{-}C\alpha_{-}$, $C\alpha_0{-}C_{+}$ and $C\alpha_0{-}O_{-}$. From the bonding information, we can therefore construct an incomplete distance matrix; that is, a matrix of distances between all pairs of atoms defined by the 1–2, 1–3 and 1–4 restraints.

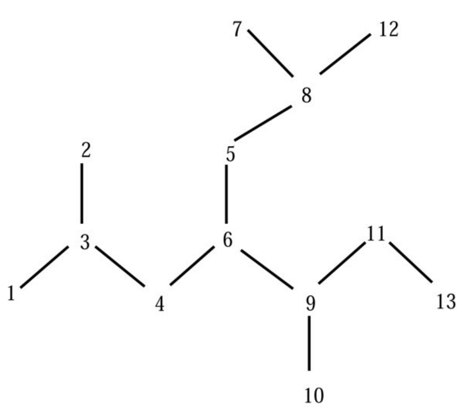

Figure 3
A tripeptide graph with a leucine residue side chain attached to the $C\beta$ position with nodes numbered from the top left of the figure. Node 6 is a hypothetical start point and all points have been collected up to a depth order of 3 from this point.

2.2. 2D isomorphous graph analysis

The first part of the analysis of the graph (protein electron-density peaks) requires a search for all nodes that have the same number or more edges attached to the sub-graph starting point. For a search with the partial tripeptide, these points have three or more edges attached. For each start node that satisfies this criterion it is necessary to collect a list of nodes attached to this with a depth less than and equal to the maximum depth of the sub-graph. The graph node list determined from the experimental data is usually longer than ten because there are 18 different R groups that can be attached to the $C\beta_0$ atom and these atoms will be included within the graph to be analysed. It is also possible to have less than ten points owing to experimental data errors that result in missing peaks and also when the R group is glycine or alanine. Fig. 3 shows a 13-node graph that is analogous to the partial tripeptide with three further nodes (7, 8 and 12). If node 6 is a $C\alpha_0$ atom then the sub-graph (5, 7, 8 and 12) has the same connectivity as that found within the amino-acid side chains of leucine, asparagine, and aspartic acid. The aim is to uniquely identify a partial tripeptide within this, and reject the points 7, 8 and 12 from the analysis. The list of nodes from the graph is collected by depth order, so that the first node in the list is the hypothetical start point (node 6 = $C\alpha_0$), and the next points in the list are those connected directly to this start point. The depth ordering of the list is naturally obtained from a depth search of the electron-density map graph and therefore does not represent a computational overhead. Fig. 3 shows the node ordering for the partial tripeptide with leucine at the central

Depth			1			2				3				
	Atom	6	4	5	9	3	10	11	8	1	2	13	7	12
	6		●	●	●									
1	4	●				●								
	5	●							●					
	9	●					●	●						
2	3		●							●	●			
	10				●									
	11				●							●		
	8			●									●	●
3	1					●								
	2					●								
	13							●						
	7								●					
	12								●					

Figure 4
The connectivity matrix for the graph in Fig. 3. Five pathways are shown that traverse the matrix from the start point (node 6) to a point within depth level 3. This represents a total of four ambiguous solutions in 2D.

Table 1
The four possible solutions for the sub-graph (Fig. 1) interpretations of the graph in Fig. 3.

Solution 1 is the correct solution; solutions 3 and 4 are rejected using the 1–4 distance restraint during analysis of the longest route. Solution 2 exceeds the final rejection threshold for the different distance matrix analysis.

	1			2			3			
$C\alpha_0$	N_0	C_0	$C\beta_0$	C_-	N_+	O_0	O_-	$C\alpha_-$	$C\alpha_+$	Residual
6	4	9	5	3	11	10	2	1	13	0.007
6	4	9	5	3	11	10	1	2	13	0.138
6	5	9	4	8	11	10	12	7	13	0.041
6	5	9	4	8	11	10	7	12	13	0.093

Table 2
Times to interpret different graphs using search sub-graphs.

Timing values were obtained using 1 000 000 repeated calls to the same algorithm for the non-protein analyses. Calculations were performed on a Compaq alpha DS10; calculations on a SGI R10000 are similar. PTP, partial tripeptide graph (see Fig. 1).

Search sub-graph	Graph	Nodes	Time (s)
PTP	PRP + alanine	10	1.5×10^{-5}
PTP	PTP + serine	11	2.6×10^{-5}
PTP	PTP + leucine	13	1.6×10^{-4}
PTP	Protein	2943	2.5×10^{-1}
PTP + amino acids	Protein	2943	9.1×10^{-1}

position starting from the top left of the figure. Fig. 4 shows a possible connectivity matrix for Fig. 3. Other possible node lists can result with a different order only within each depth. For example, the first depth node points can have order (4, 5, 9), (4, 9, 5), (5, 4, 9), (5, 9, 4), (9, 4, 5) or (9, 5, 4) depending on the order they were collected from the connectivity list of the experimental data peaks. Node 6 is the starting point and interpretation requires that nodes are identified and map to Fig. 1, where nodes 7, 12 and 8 are rejected points even though they are ambiguous in 2D with points 1, 2 and 3. Table 1 shows the four possible solutions for the graph in Fig. 3 using the sub-graph in Fig. 1.

The first optimization step is to recast the analysis so that instead of collecting results by depth order, the 2D solution results are determined directly by traversing the connectivity matrix. Crossing the leading diagonal at each step with increasing depth order provides the solution and this is shown in Fig. 4. This figure shows that there are a total of five routes from the start point 6 to points of depth three. The two blue pathways ambiguously identify the atoms ($C\alpha_0$, N_0, C_-, $C\alpha^-$) because the number of edges to node 4 is two and the number of edges at node 3 is three, which correctly identifies the atoms N_0 and C_-, respectively. The $C\alpha_-$ atom can either be node 1 or 2. The red pathways are incorrect solutions for the atoms ($C\alpha_0$, N_0, C_-, $C\alpha_-$), but the path has the correct connectivity within 2D. The green route is a unique solution for the atoms ($C\alpha_0$, C_0, N_+, $C\alpha_+$) because the number of edges to node 9 is three and the number of edges to node 11 is two, correctly identifying atoms C_0 and N_+. The remaining atoms not yet assigned by the pathway analysis are mapped using the connectivity matrix for each of the possible four solutions; these are shown in Table 1.

2.3. 3D screening

The final stage is to remove ambiguity within the 2D analysis using the incomplete difference distance matrix. That is, from all known 1–2, 1–3 and 1–4 distances within the search sub-graph (Fig. 1) the residual is determined by comparison with each incomplete distance matrix generated by 3D positions of the nodes in Fig. 3. The four different node lists in Table 1 are used to generate these incomplete distance matrices by mapping graph node coordinates to the search sub-graph. The correct solution, row 1 in Table 1, is selected as it results in the lowest residual from this analysis. To reduce the number of difference distance matrix calculations it is possible to reject one solution (row 2 in Table 1; blue pathway) because the 1–4 distance from node 6 to node 2 is not 3.8 ± 0.21 Å. In general, the ambiguity between the assignment of the O_- and $C\alpha_-$ in a protein can resolved using a single calculation of the 1–4 distance. This restraint requires a separate analysis for *cis*- and *trans*-peptide conformations to determine both structure types. It should be noted that a number of amino acids are branched at the γ atom, increasing the ambiguity with the N-terminal direction backbone atoms within the 2D search.

Since all information necessary to carry out the interpretation of a graph is determined from the search sub-graphs (start point, depth, connectivity, longest routes, incomplete distance matrix), the algorithm can take any search subgroup of atoms and carry out interpretation using this. When interpreting an electron-density map of a protein the algorithm requires the partial tripeptide sub-graph and sub-graphs for each *R* group in turn to complete the analysis. The partial tripeptide is searched first and overlap of these determines the connectivity of backbone trace of the protein. Searching for side-chain sub-graphs does not require a full search, because the position of the $C\beta_0$ atom is known from the partial tripeptide analysis. Therefore, these define the starting points for the side-chain analysis.

2.4. Error analysis

It was noted that electron-density maps contain error and this manifests itself as extra and missing nodes as well as node-position error. This affects the implementation of the described method in a number of ways. The first is that the number of edges at a node in the graph cannot be assumed to be identical to that of the sub-graph. Therefore, if we cannot find a solution, it is necessary to carry out the analysis without using the connectivity restraint of each node. Missing nodes not on a pathway to the highest depth of analysis (the atoms O_-, $C\beta_0$ and O_0 in Fig. 1) are also handled by ignoring the connectivity restraint and flagging missing nodes with a large residual within the difference distance matrix analysis. Missing nodes on a pathway (all other atoms in the partial tripeptide) prevent the finding of solutions. Extra nodes in the graph can also change the depth level of the points. That is, a third-order depth point can be elevated to a first-order or second-order point. For example, an extra node that has edges directly to $C\alpha_0$ and $C\alpha_+$ will elevate the atom $C\alpha_+$ to a second-level node; this completely defeats the analysis. This problem can be handled using a knockout procedure on remaining areas of the graph on completion of the tripeptide search. Each point in the search-node list generated from a start point is removed in turn and the calculation is reperformed on the remaining points. Since the volume of electron density that remains after a standard search is small, this knockout procedure does not represent a significant overhead.

Error in the position of the experimental map peaks will effect the residual of the incomplete difference distance matrix. It is therefore necessary to define a residual depending on the expected error of the data, noting that low tolerance will result in over-interpretation. It should be noted that expected error is greater in a 1–N interaction than with a 1–(N − 1) interaction and therefore the difference distance matrix should be weighted accordingly. There are a number of specific macromolecular properties that can help where over-interpretation occurs and these are described elsewhere (in preparation).

3. Results and discussion

Table 2 shows the time taken to solve three different small graphs using the partial tripeptide sub-graph (Fig. 1). The times taken to interpret the data include the 2D sub-graph search followed by the 3D matrix screening. To demonstrate the analysis of a high-resolution electron-density map, a 1.0 Å set of data of the protein 1box (Herbert *et al.*, in preparation) was interpreted using phases determined with *ACORN* (Foadi *et al.*, 2000). This contained 2924 density peaks or nodes. The partial tripeptide sub-graph analysis was complete in 0.24 s, while full analysis of the protein was complete in less than 1 s. This compares well with a time of about 8 h for a similar calculation using *trace* and *wARP*. Approximately 85% of the atomic structure was interpreted without the use of protein-specific analysis techniques (Oldfield, in preparation), although without these the algorithm is not complete. It should be noted that the rate-limiting step is the calculation of the difference distance matrix. Approximately 10% of the quoted times are for the generation of ambiguous 2D graph solutions and 90% of the time is used for screening these 2D solutions using the difference distance matrices. The algorithm is not a rigorous 3D graph method but does always determine the correct solution without false positives where the data are ideal. It represents a rapid method for the interpretation of high-resolution electron-density maps. None of the individual aspects of the methods described are new within the context of general data analysis. The novel feature of the algorithm is using the individual methods in such a way as to interpret atomic structure from high-resolution electron-density maps. The method is fast since the number of computational steps to identify the 2D sub-graph is small when traversing the connectivity matrix. The additional screening using restrained 1–4 distances within this matrix as well as the number of neighbours analysis also screens out most of the results that are not correct solutions. This leaves a small number of 2D results to be rejected within the 3D difference distance matrix. The algorithm should be adaptable to a number of problems that require the interpretation of molecular data independent of the conformation of the search molecule. It would be simple to adapt this method to search a graph defined by a series of likely atom positions within an active site of a target protein. A list of ligands could then be screened against this active site and ligands that match, regardless of their conformation, would be found as those with a low residual for the incomplete difference distance matrix.

3.1. Availability

The algorithm described has been implemented within *QUANTA* (Accelrys) and also as a stand-alone program. Validation and analysis techniques necessary to provide a useful map-interpretation program have been developed for low-resolution map interpretation (in preparation) and need to be adapted for use with this new interpretation method to provide a useful tool.

References

Babel, L. (1991). *Computing*, **46**, 321–341.

Crippen, G. M. & Havel, T. F. (1988). *Distance Geometry and Molecular Conformation.* Taunton: Research Studies Press.

Foadi, J., Woolfson, M. M., Dodson, E. J., Wilson, K. S. & Jia-xing, Y. (2000). *Acta Cryst.* D**56**, 1137–1147.

Jones, T. A. & Kjeldgaard, M. (1997). *Methods Enzymol.* **227**, 173–230.

McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.

Oldfield, T. J. (1996). *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 67–74. Warrington: Daresbury Laboratory.

Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 459–463.

Wang, C. E. (2000). *Acta Cryst.* D**56**, 1591–1611.

Wenger, J. C. & Smith, D. H. (1982). *J. Chem. Inf. Comput. Sci.* **22**, 29–34.