Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Pattern-recognition methods to identify secondary structure within X-ray crystallographic electron-density maps

Thomas Oldfield

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

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

Received 17 September 2001
Accepted 7 January 2002

The interpretation of macromolecular crystallographic electron-density maps is a difficult and traditionally a manual step in the determination of a protein structure. The visualization of information within an electron-density map can be extremely arduous owing to the amount and complexity of information present. The ability to see the overall fold and structure of the molecule is usually lost among all the detail, particularly with larger structures. This paper describes a novel method of analysis of electron density in real space that can determine the secondary structure of a protein within minutes without any user intervention. The method is able to work with poor data as well as good data at resolutions down to 3.5 Å and is integral to the functionality of *QUANTA*. This article describes the methodology of the pattern recognition and its use with a number of sets of experimental data.

1. Introduction

Protein crystallography is evolving from a method in its own right that is hypothesis-driven towards a tool for biologists in the field of structure genomics. There is a change from the solution of a single protein structure to understand its function towards the generation of results quickly, correctly and preferably with as little user intervention as possible for many structures. All aspects of protein crystallography, from expression of genes to the analysis of the final macromolecular structure, are becoming more automated. A part of this process that is not the rate-limiting step, but requires significant expertise and time, is that of experimental electron-density map interpretation.

Interpretation of electron-density maps is necessary when phases are obtained using *de novo* phasing techniques and little knowledge of the macromolecular structure is known prior to the structure elucidation. Any method that provides reliable and rapid map interpretation from just the phased map can also provide an unbiased technique suitable for molecular-replacement (MR) map reinterpretation.

The problem of map interpretation is compounded because the electron density contains both systematic and random error. The quality of an experimental map depends on the resolution and the amount of error, which is predominantly the result of indeterminacy of phase. This error has a direct impact on the amount of effort required to interpret an electron-density map. Both proteins and nucleic acids are unbranched polymers; therefore, the interpretation of an electron-density map involves tracing a pathway.

It is clear that such a complex three-dimensional graphic is much easier to interpret if some visual clues can be offered to the crystallographer to provide a starting point of analysis. The lowest resolution information available is the solvent boundary of the macromolecule; automated determination of this is described elsewhere (Cowtan & Main, 1998; Jones, 1992). The next useful information is the secondary structure; typically, 50–75% of residues are in recognisable secondary structure. The process of structure determination can be greatly eased if secondary structure within an electron-density map can be identified automatically and displayed to the crystallographer as a graphic. This detail can subsequently be used to aid tracing and also as a method of searching the Protein Data Bank for homologous proteins before any expenditure of time by the crystallographer. Secondary-structure search methods that use X-ray crystallographic data have been described in the literature (Kleywegt & Jones, 1997; Cowtan, 1998) and these determine the correlation coefficient for the fitting of structure templates at all points and orientations within a map. Kleywegt & Jones' method carries this out in real space by convolution, while Cowtan's method uses a reciprocal-space method and is therefore more efficient. The problem with these methods is the slowness of the calculations. Additionally, there is a problem of low signal-to-noise ratio in large asymmetric units, particularly where there is significant phase error. The template used is also of fixed size, which means that secondary structure smaller or larger than the search template reduces the signal. This article presents a method that usually completes in minutes, is highly sensitive even at low resolution and with poor phases and is independent of structure-element size.

2. Methods

The secondary-structure search algorithm presented here is based on pattern recognition of the electron-density skeleton formed by data reduction of the electron density (Greer, 1974) using an algorithm that has some modifications (Oldfield, in preparation). The result of this data-reduction algorithm is a set of pseudo-atoms at map grid points that are generally known as bones. These bones are highly processed and the salient features of this processing are the identification of bones that correspond to residue side-chain atoms and main-chain atoms and a ringlet analysis that cuts up the small rings of bones that form in regions of noisy data.

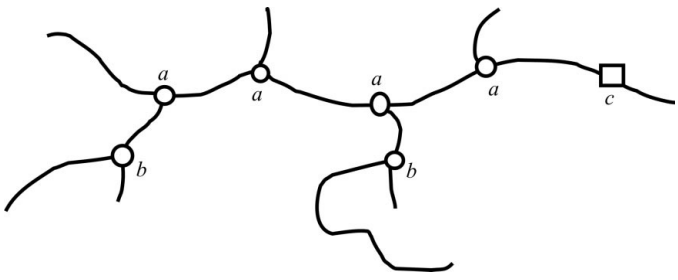

Figure 1
A stylized view of bones. Points marked *a* are branch points with two main-chain connections and one side-chain connection. Points marked *b* are branch points in side-chain bones and points *c* are false points added to sample the bones skeleton at regular intervals. Only points *a* and *c* are used in the depth analysis during pattern recognition.

2.1. The mathematical description of bones

The bones skeleton generated from the map can be considered to be a mathematical tree. The points of interest in the analysis are the branch points within this tree and termini; therefore, the algorithm only considers points in the bones that do not have two neighbours (Fig. 1). It is therefore possible to use a depth-search algorithm that, for a starting point, identifies all points that are branch points and are N depth neighbours from this starting point. This involves starting at a branch, traversing the bones net and storing a stack of points that neighbour this starting point. The first-order depth points are all those points that are closest to the starting point and are not the starting point (the case when a ring is present). The information associated with each N depth point is the bones pathway, the number of neighbours this point has – unless it is the same point as the $(N - 1)$ point, when it is flagged as such – and the bones path length from $(N - 1)$ to point N. The true path length is determined from the map grid spacing and the number of grid points in a path.

To allow the described algorithm to proceed efficiently, it is necessary to process the bones using properties of the path depth and the path length. Bones are classified as main chain or side chain based on which atomic structure they represent. All bones are initially defined as main chain, but are converted to side chain by depth analysis from termini (branch points with a single neighbour). A fragment of bones is defined as side chain for all points that are $N_{\rm side}$ depth from a terminus and where the total path length is less than $R_{\rm side}$. Deleting any fragment using the same definition but using the parameters $N_{\rm delete}$ and $R_{\rm delete}$ can clean the bones representation. The algorithm to generate bones can produce ringlets in regions of noisy electron density. A ringlet is defined as a main-chain bones path where $(1 < N < N_{\rm ring})$ depth analysis can return to the starting point within a total path length of less than $R_{\rm ring}$. The longest path for a ringlet analysis is set to side chain. The values of $N_{\rm side}$, $N_{\rm delete}$ and $N_{\rm ring}$ are user-defined and have default values of three, one and three, respectively. The values of $R_{\rm side}$, $R_{\rm delete}$ and $R_{\rm ring}$ have default values of 7.0, 2.0 and 3.0 Å, respectively.

Since the branch points within the bones skeleton are critical to the analysis, it is necessary to add artificial branch points to the bones skeleton (Fig. 1). This is to make sure the bones tree is sampled at a certain frequency within the analysis, particularly at lower resolution, where the maps can be featureless. The artificial branch points are added to main-chain bones at 3.8 Å steps in both directions from existing points in the bones skeleton. If an integer number of artificial points cannot be generated that traverse the featureless section, then the nearest integer number of branch points are added so that the number added is 80% weighted towards oversampling.

2.2. Pattern recognition

The analysis of the map is started using a pattern-recognition algorithm that is based on a loose description of what might be considered a β-strand or helix in the electron density. Starting from each branch point in the map that is not a terminus, the depth-search algorithm traverses the bones net. After adding at least six points to a current path and for every new point added, an analysis is carried out to determine the shape of the current path. This is then compared with the expected description of a helix and strand.

The first analysis determines the overall shape of the current path. This is determined by calculating the principal components (PC) using a tensor generated from the unweighted points in the current path (1).

$$I[V] = \text{Diagonalize}\begin{bmatrix} \sum x_r^2 & \sum x_r y_r & \sum x_r z_r \\ \sum y_r x_r & \sum y_r^2 & \sum y_r z_r \\ \sum z_r x & \sum z_r y_r & \sum z_r^2 \end{bmatrix}, \quad (1)$$

where $x_r = x - \langle x \rangle$, $y_r = y - \langle y \rangle$, $z_r = z - \langle z \rangle$ and $I[V] = \lambda[i]\cdot V[i]$ (for $i = 1, 3$). $I[V]$ is the diagonalized inertia tensor matrix for bones branch points in the current path. x, y and z are the real-space coordinates of bones branch points in the current path and $\langle x \rangle$, $\langle y \rangle$ and $\langle z \rangle$ are the mean values for all x, y and z coordinates in the current path, respectively.

The PCs of the bones path are given by the eigenvectors $V_{[i]}$ sorted by the eigenvalues $\lambda_{[i]}$ and must have values within the limits shown in Table 1 for the analysis to continue down the current path. If the PCs are within the range then further analysis is carried out using the branch points. The current path is considered a possible strand if the restraints shown in Fig. 2 are satisfied for consecutive points in the current path which are more than 1.5 Å apart. If satisfied, then the new

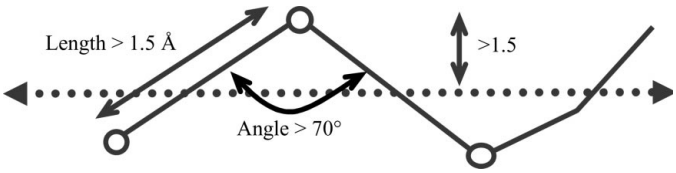

Figure 2
The figure shows the restraints used for the bones analysis to describe a β-strand. The double arrow dotted line shows the position of $V_{[1]}$ for the strand and the circles mark the branch points in the bones, with two length restraints and one angle restraint.

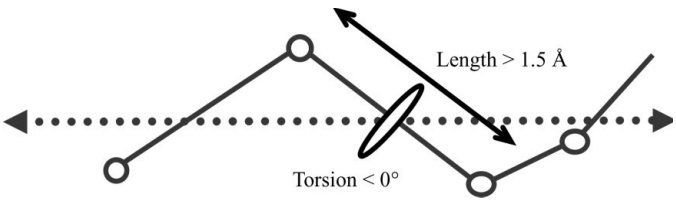

Figure 3
The figure shows the restraints applied to define a helical path in the bones. The double arrow dotted lines indicate $V_{[1]}$ for the helix. The restraints include a minimum separation for the error analysis, an ideal length and a torsion angle. The circles mark branch points in the bones path marked with solid straight lines.

Table 1
Limits of PC values that define a strand and helix.

Principal component	Strand	Helix
$i = 1$	No limit	No limit
$i = 2$	$\lambda^{1/2} < 4.0$	$1.8 < \lambda^{1/2} < 6.0$
$i = 3$	$\lambda^{1/2} < 2.5$	$1.8 < \lambda^{1/2} < 6.0$

point is added to the tensor matrix and a new PC set calculated. If the current path has more than ten points then this is considered a possible helix if the restraints (Fig. 3) are true for all points more than 1.5 Å from depth-search neighbours and all defined torsion angles less than zero. An error weighting is defined (2) for points in the path with a torsion angle > 0°; a threshold of 5 is required for the entire path. The 1.5 Å limit on point separation is used as a filter for noise. Branch points that are close together with the bones path are probably the result of noise and therefore the local path is likely to be unrepresentative of the region.

Error weights for torsion > 0° for each four point set in the bones path are

$$1.5\ \text{Å} < L < 3.0\ \text{Å} : E = E + 1, \quad (2a)$$

$$3.0\ \text{Å} < L < 4.5\ \text{Å} : E = E + 3, \quad (2b)$$

$$4.5\ \text{Å} < L : E = E + 5, \quad (2c)$$

where L is the separation between points 2 and 3 for a four-point definition of a torsion angle and E is the error sum.

The analysis of a strand determines a path that is approximately a straight line, while a helix is defined as those points that form a left-handed helical path most of the time with greater emphasis placed on points that are further apart. This initial analysis is *ad hoc* in its basis and was designed empirically to provide information from very noisy maps. The algorithm returns solutions as matrices that contain the PC of the current path; subsequent analysis is based on these matrices. There is no limit to the extent of the analysis and it can identify any length of secondary structure above minima which are two residues for a strand and five residues for a helix (in ideal bones) with two branch points per residue. It also produces a large excess of possible example helices and strands and as such requires pruning.

If a new matrix is determined it will be merged with any other existing matrix if it has the properties defined by (3)–(5) and is the same secondary-structure type. The merging supersedes the existing matrix with a new matrix as defined in (6).

$$|V(\text{new})_{[1]}| > |V(\text{existing})_{[1]}|, \quad (3)$$

$$|\langle V(\text{new})_{[1]} \rangle - \langle V(\text{existing})_{[1]} \rangle| < 2.5\ \text{Å}, \quad (4)$$

$$\{V(\text{new})_{[1]} \cdot V(\text{existing})_{[1]}\} < \cos 10^\circ, \quad (5)$$

$$V(\text{merged})_{[i]} = (2/3)V(\text{new})_{[i]} + (1/3)V(\text{existing})_{[i]}. \quad (6)$$

The third stage involves the principal component real-space rigid-body refinement of a backbone trace to the electron density defined by each vector as described in Oldfield (2001*a*). The atoms for refinement are generated as an ideal

secondary-structure trace of C^{α} atoms about $V_{[i]}$ of each inertia matrix. The number of C^{α} atoms generated is determined by the length of the $V_{[1]}$ vector and defined in (7) and (8). The length of the vector is based on the variance of the bones points along the largest PC and is calculated as $2.0\lambda_{[1]}^{1/2}$. For strand

$$N = \frac{2\lambda_{[1]}^{1/2}}{3.0} \tag{7}$$

and for helix

$$N = \frac{2\lambda_{[1]}^{1/2}}{2.1}, \tag{8}$$

where N is the number of C^{α} atoms in ideal secondary structure.

Atomic positions of the main-chain N and carbonyl C atoms are approximated by coordinates 1/3 and 2/3 along the $C^{\alpha}-C^{\alpha}$ pseudo-bond and with 0.5 occupancy. The resulting atomic model has undefined direction. In the case of a helix a screw refinement along $V_{[1]}$ is also applied. This refinement provides an initial weight for each vector that is the overlap integral between the atoms and the electron-density map.

The next analysis applies further weighting to each vector by analysis of the super-secondary structure. Multiple sets of vectors are compared with standard super-secondary-structure motifs and if any motif is observed within the vector list then the current weight defined by refinement is increased fourfold. The current available release of the algorithm only uses the β-sheet as a super-secondary-structure definition, but common feature definitions found by data mining (Oldfield, 2001*b*) are being studied as an aid to the pattern recognition. All combinations of $V_{[1]}$ vectors determined so far from the analysis are compared with the expected motif set by checking the deviations of centroids and angles using the PCs as defined in (4) and (5).

The final analysis of the $V_{[1]}$ vectors uses the principle that secondary structure does not overlap in space. The aim is to find the largest set of non-overlapping $V_{[1]}$ vectors that has the largest weight determined by refinement and super-secondary-structure analysis. Exclusion in space between a pair of vectors is defined if the closest point between the two $V_{[1]}$ vectors is less than 2.1 Å and not within the end 5% of the $V_{[1]}$ vector. The analysis of the central 90% of the vector is used to allow vectors to be placed end to end and weighted by length. The overlap analysis starts by considering the highest weighted vector and merging other consistent vectors with this. If an inconsistency is found then a new group is generated to include this new vector and the overlap analysis restarted with the new group. At the end of this process, the largest cluster with the highest overall weight is considered to be the correct solution.

The final vectors are displayed graphically within the context of the experimental data. The vectors are colour coded as a function of whether they represent β-strands or α-helices and have a line-thickness differential depending on the modified residual of the principal component refinement. The final part of using this algorithm involves the conversion of the vectors into a C^{α} trace using principle component refinement (Oldfield, 2001*a*). These ideal secondary-structure elements can then be used as a starting point for automated tracing methods.

Table 2
Results of secondary-structure search algorithm.

The proteins used are: Rnase(1), ribonuclease SA (Sevcik *et al.*, 1991) MIR phases; Rnase (2), as Rnase(1) but after density modification using *DM* (Cowtan, 1998); CysB (Tyrrell *et al.*, 1998); EMTA (Gliubich, personal communication), *endo*-specific membrane-bound *trans*-glycosylase; OMPLA, outer membrane phospholipase A (Snijder *et al.*, 1999); PA(1) and PA(2) penicillin acylase results using two different σ values (Duggleby *et al.*, 1995).

	Res (Å)†	N_{res}†	FOM‡	Helix§	Strand¶	Extra††	Best σ‡‡	Time (min:s)§§
Rnase(1)	2.5	96	0.5	1/1	3/3	1s	1.3	0:29
Rnase(2)	2.5	96	0.7	1/1	3/3	0	1.2	0:12
CysB	1.8	236	0.53	1/4	4/4	0	1.3	4:10
EMTA	2.75	184	0.57	8/8	0/0	4s	1.2	1:17
OMPLA	2.7	256	0.75	0/0	12/17	0	1.2	2:15
PA(1)	2.5	750	0.52	12/24	17/27	1s	1.4	4:12
PA(2)	2.5	750	0.52	9/24	25/27	5s	1.3	13:20

† The resolution of the phases used in the map calculation. ‡ Figure of merit. § The number of helices found and the total number of helices. ¶ The number of strands found and the total number of strands. †† The number of extra secondary-structure elements found that were not correct: 's' indicates strand and 'h' indicates helix. ‡‡ The bones start level used that gave the best results for vector search. §§ Time for the calculation running on a SGI O2 R5000 computer.

3. Results

The algorithm described was used on a number of experimental sets of data. All data [except Rnase(2)] used for the analysis are the same as those used to determine the published structures with no resolution cutoff. Data described as Rnase(2) were produced from the original data Rnase(1) by the application of density modification (Cowtan, 1998); the technique of density modification was not available for the original structure solution of Rnase. Data ranged in resolution from 1.8 to 2.75 Å and figures of merit (FOM) from 0.7 to 0.5. In each case, the experimental map was opened within the program *QUANTA*2000 (Accelrys Inc.) and displayed. *QUANTA* bones were calculated at the σ value shown in Table 2 and auto-edited using tools available within the X-AUTOFIT application of *QUANTA*2000. A map mask was generated from the bones of each protein and used for the analysis as a calculation-bounding surface. The secondary-structure search algorithm was run and the results are shown in Table 2.

Figs. 4 and 5 shows the maps of ribonuclease SA from data sets 1 and 2 with the vectors at the best σ value superposed. These figures show that density modification (DM; Cowtan & Zhang, 1999) makes an enormous difference to the quality of the map in this sheet region. The two views of ribonuclease SA are approximately the same in these figures, but there is a discrepancy in the angle of the right-hand vector of the sheet for map at FOM 0.5 owing to a shift in the electron density as a result of data error. Therefore, although the pattern-recognition algorithm can identify structural information where there is a lot of data noise, it is always sensible to apply DM tech-

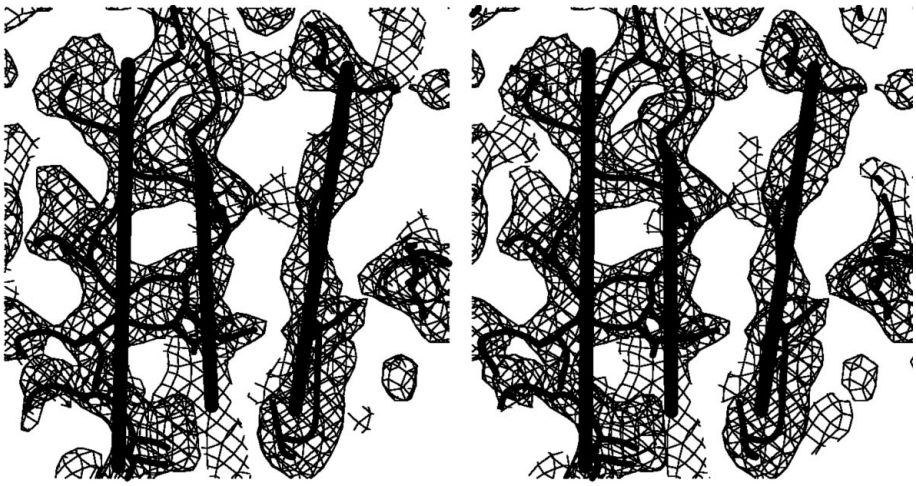

Figure 4
2.5 Å electron density for ribonuclease SA with an FOM of 0.5. The figure shows the main three-stranded β-sheet of the molecule with the bones and vectors superposed. The electron density is contoured at 1.3σ, the same value as used for the secondary-structure pattern recognition.

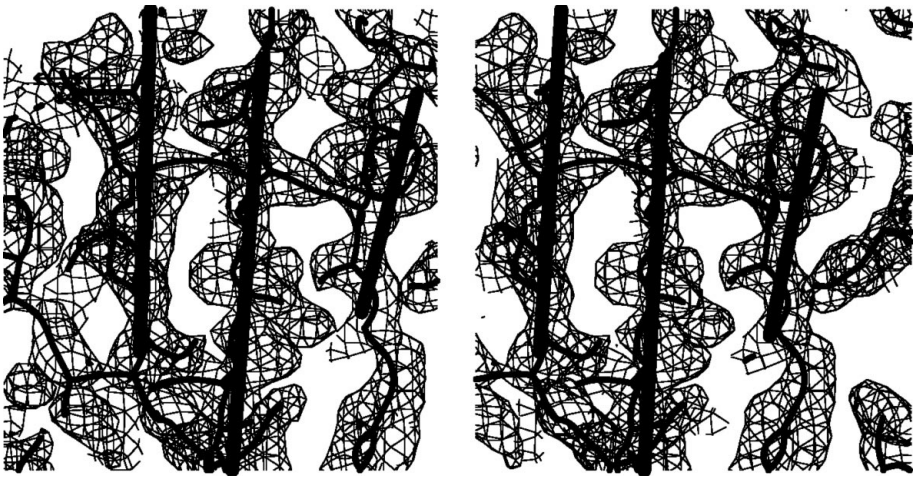

Figure 5
2.5 Å electron density for the DM ribonuclease SA with an FOM of 0.7. The figure shows the main three-stranded β-sheet of the molecule with the bones and vectors superposed. The electron density is contoured at 1.2σ, the same value as used for the secondary-structure pattern recognition.

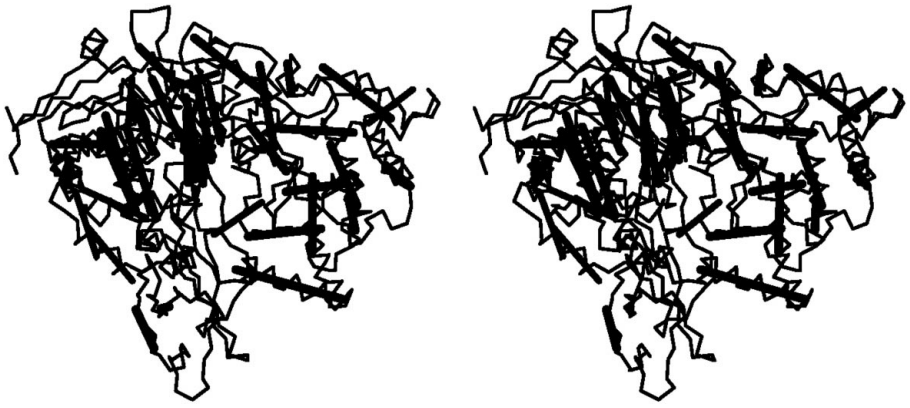

Figure 6
The C^{α} trace from the structure 1pnl (Duggleby *et al.*, 1995) with the vectors found by the pattern recognition superposed.

niques. The use of DM results in the sharpening of data, so it is necessary to use a slightly lower σ-cutoff value for the bones in order to compensate for the slight fragmentation of the bones as a consequence of this sharpening.

It can be seen from the two penicillin acylase (PA) examples that there is a significant calculation-time difference as a function of the bones start value selected; it is generally sensible to try higher σ-cutoff values for larger maps. The information found is different in the two PA examples and in part complementary and it would probably be sensible to combine these two results. Fig. 6 shows the result of the vector analysis of PA(2) and shows that a reasonable knowledge of the structure can be inferred from just this vector information. The results indicate that the number of vectors found is dictated more by data quality than resolution. Only when data has a good FOM (0.7 or better) can the algorithm identify all the secondary structure with no additional false-positive results. At lower FOM the algorithm tends to include some false-positive and false-negative results, while below an FOM of 0.5 there is a rapid fall-off in the amount of structure found as the electron density tends to become disjointed. This algorithm cannot find structure within a map containing no information.

The analysis of the secondary structure is sensitive to the bones start parameter (Greer, 1974). An analysis was carried out for the two ribonuclease examples to test the effect of changing the start value on the quality of the secondary-structure determination. Table 3 shows the results of the analysis, where the correct solution is three-stranded β-sheet, a single fourth strand and one helix. At an FOM of 0.5 the algorithm does not find the correct number of secondary-structure elements, while at an FOM of 0.7 the correct solution is found over a wide range of start values of σ for bones generation. Since the calculation is faster when using higher values of σ, it would be recommended to start with a σ of 1.3 to 1.4, although a map-quality index (Oldfield, in preparation) can aid in this selection.

Table 3
Results of interpreting the two ribonuclease maps at different bones σ start values.

Bones σ	Rnase(1)	Rnase(2)
1.0	—	—
1.1	—	Correct
1.2	+1 strand	Correct
1.3	+1 strand	Correct
1.4	−1 strand	Correct
1.5	−1 strand	−1 strand
1.6	—	−2 helix and strand
1.7	—	—

4. Discussion

The results obtained with the six different example maps shows that the algorithm can be used at a number of different resolutions and map qualities. Most of the secondary structure can be determined from a map very quickly (within minutes). The algorithm cannot find secondary-structure information within a map where none exists and only identifies information found in proteins. In the case of the PA experimental map, there is a significant fraction of secondary structure which the algorithm fails to find [PA(1) in Table 2]. This is because the experimental electron-density map was of very poor quality in some regions (P. Moody, personal communication) owing to data noise and the poor phasing power of the derivatives used. Fig. 7 shows a region of the map contoured at 1.3σ and indicates that although the density is of reasonable quality within this volume, it is broken in a number of places. The algorithm cannot determine secondary-structure information if the bones skeleton is broken as the pathway analysis and depth analysis is defined by the connectivity of the bones. The map is connected within this volume when contoured at 1.1σ, but at that level the majority of the experimental map is very noisy and not ideal for an overall molecule interpretation. The algorithm can determine secondary-structure information where bones are over-connected (Fig. 4) and where there is significant phase error. The parameterization of the bones represents a limitation of the described method. The method is not limited by the size of the protein molecule and does not have an upper bound on the length of the secondary-structure elements, as the calculation is in real space and is therefore a local property of the experimental map. The shortest strand that can found is limited to more than three residues (7) and the shortest helix is limited to more than five residues (8) to prevent propagation of noise during the calculation. Whether the end-points of the vectors represent the end-points of the secondary structure within the final coordinates is dependent on the quality of the map. Where the experimental map is a reasonable representation of the final coordinates, these end-points are well defined. It should be noted that the length of a vector is based on a statistical definition using $2\lambda_{[1]}^{1/2}$ which assumes that the bones path points are evenly distributed. Since the aim of map interpretation is not the generation of vectors but rather the placement of all atoms correctly into an experimental map, it is necessary that the method described provides a good visual clue as to the secondary-structure content to a crystallographer. It is also necessary that subsequent automated map-interpretation methods based on these vectors works correctly. The number of C^{α}-trace atoms generated using (7) and (8) therefore overestimates of the rise-per-residue values for helix and strands (Dickerson & Geis, 1969). This will underestimate the number of residues generated, so that only the core region is fitted during the principal component real-space rigid-body refinement.

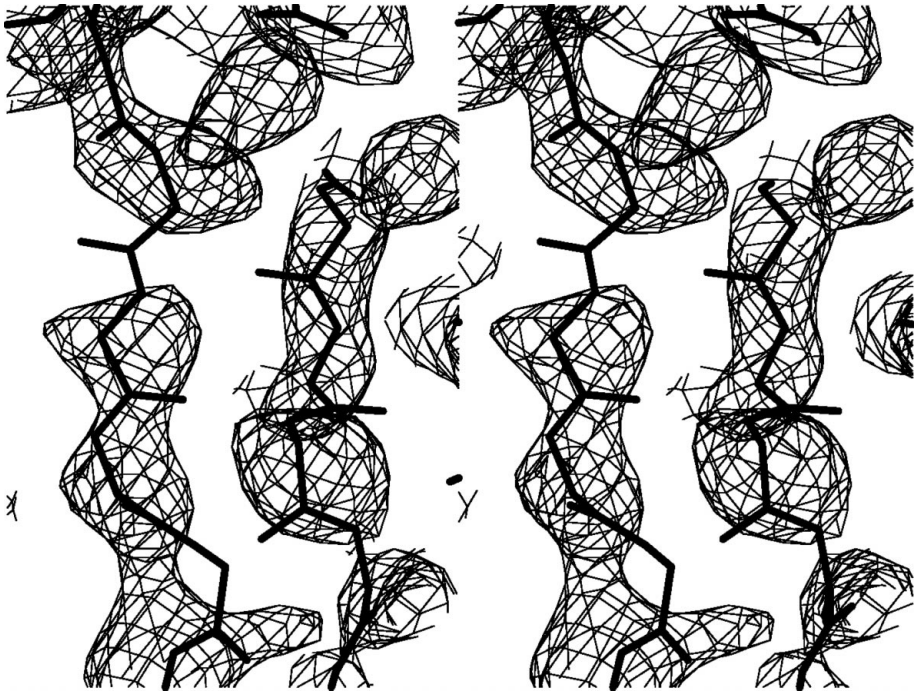

Figure 7
Section of electron density contoured at 1.3σ overlaid with the final main-chain coordinates of PA. This region of the map is part of a β-sheet and two strands are shown in the figure. No secondary structure is found for these two strands owing to gaps within the bones pathway.

The vector information is displayed graphically within the X-ray application X-POWERFIT of *QUANTA*2000 as vectors to provide a visual aid to map tracing. The results are also used as prior information for automated tracing techniques (in preparation) and in this case the vectors generated are automatically converted to a C^{α} trace (7 and 8) and the remaining trace generated. When using auto-tracing it is not imperative for all the secondary structure to be identified. The vectors generated can be used to search the PDB (or a subset) for similar structure (or subset) using secondary-structure element analysis (SSE; Mizuguchi & Go, 1995) using the algorithm derived from *SQUID* (Oldfield, 1992). An

interface runs an external alignment program with the current set of vectors from the map and allows the inclusion of a C^{α} trace from matched proteins into *QUANTA* from any search solution. Owing to its sensitivity at detecting real structure in experimental maps, the algorithm has also been used as a means to access the quality/interpretability of a map before any manual expenditure in time. The analysis is able to quantify whether it would be more successful to trace the map or determine better phases first.

The method is quick and can handle large maps or parts of maps and is useful up to a lower resolution limit of 4 Å. Beyond this resolution, the bones skeleton for helices and β-sheets is no longer correctly modelled by the analysis described. It should be noted that a helix merges to become a cylinder at lower resolutions with a single straight bones path. The algorithm has been used at 6 Å with the knowledge that strand vectors actually indicate the presence of helices (J. Wilson, personal communication). The high-resolution limit of the algorithm is approximately 1.5 Å as the method is based on pathway analysis. At higher resolution the experimental data is discrete and no pathway can be observed. It is possible to calculate maps at lower resolution than the phase information available, but these are generally sharp with breaks that prevent observation of the secondary structure. Since the algorithm is designed to identify left-handed α-helices using a torsion-angle description, it is possible to identify the correct phase solution where ambiguity exists, as right-handed helices are not marked. A map that obviously contains helical local structure where the algorithm only finds strands is a candidate for phase inversion.

The one main disadvantage of the algorithm is that it is based on the analysis of electron-density bones that requires correct parameterization for a successful analysis. It is also necessary to highly process the bones to make the algorithm efficient, as many ringlets within a bones skeleton result in a large magnification in the number of trees to search during the analysis. Table 2 shows that the analysis is reasonably robust to the σ level used to define the start value for the bones and as long as no scaling is carried out during map calculation a σ value of 1.2–1.4 is a good estimate of the starting point for analysis.

5. Availability

The algorithm is implemented within the program *QUANTA* (latest release 2000) and is activated by a single tool within the X-POWERFIT application. The application also forms the basis of an entirely automated map-interpretation method that is under development.

I am indebted to Francesca Gliubich for carrying out some of the analysis and providing statistics for this paper. I would like to thank the crystallographers Joseph Sevcik, Koen Verschueren, Arian Snijder, Francesca Gliubich and Peter Moody, who provided data to test the algorithm, and members of the YSBL for suggestions.

References

Cowtan, K. (1998). *Acta Cryst.* D**54**, 750–756.
Cowtan, K. & Main, P. (1998). *Acta Cryst.* D**54**, 487–493.
Cowtan, K. D. & Zhang, K. Y. J. (1999). *Prog. Biophys. Mol. Biol.* **72**, 245–270.
Dickerson, R. E. & Geis, I. (1969). *The Structure and Actions of Proteins*, p. 28. Menlo Park, CA, USA: Benjamin/Cummings.
Duggleby, H. J., Tolley, S. P., Hill, C. P., Dodson, E. J., Dodson, G. G. & Moody, P. C. E. (1995). *Nature (London)*, **373**, 264–268.
Greer, J. (1974) *J. Mol. Biol.* **82**, 279–284.
Jones, T. A. (1992). *Proceedings of the CCP4 Study Weekend: Molecular Replacement*, edited by E. J. Dodson, S. Gover & W. Wolf, pp. 91–105. Warrington: Daresbury Laboratory.
Kleywegt, G. J. & Jones, T. A. (1997). *Acta Cryst.* D**53**, 179–185.
Mizuguchi, K. & Go, N. (1995). *Protein Eng.* **8**, 353–362.
Oldfield, T. J. (1992). *J. Mol. Graph.* **10**, 247–252.
Oldfield, T. J. (2001*a*). *Acta Cryst.* D**57**, 82–94.
Oldfield, T. J. (2001*b*). *Acta Cryst.* D**57**, 1421–1427.
Sevcik, J., Dodson, E. J. & Dodson, G. G. (1991). *Acta Cryst.* B**47**, 240–353.
Snijder, H. J., Ubarretxena-Belandia, I., Blaauw, M., Kalk, K. H., Verheij, H. M., Egmond, M. R., Dekker, N. & Dijkstra, B. W. (1999). *Nature (London)*, **401**, 717–721.
Tyrrell, R., Verschueren, K. H. G., Dodson, E. J., Murshudov, G. N., Addy, C. & Wilkinson, A. J. (1997). *Structure*, **5**, 1017–1032.