Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Automated tracing of electron-density maps of proteins

Thomas J. Oldfield†

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

† Present address: EMBL-EBI Hinxton, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, England.

Correspondence e-mail: oldfield@ebi.ac.uk

Received 7 May 2002
Accepted 23 December 2002

The tracing of experimental electron maps in the field of protein crystallography is not a rate-limiting step for structure elucidation, but does represent the process that requires the most expertise and user time. This paper presents a method for automatically tracing the electron-density maps of proteins which can reliably generate a C^{α} trace for protein maps with data in the resolution range 1.5–4 Å. The number of C^{α} atoms placed and the precision of atom placement depends on the quality of the map, but even with poor maps (FOM $\simeq$ 0.5) the algorithm can provide a significant saving in time over conventional methods of interpretation. The interpretation of six experimental maps is presented at different resolutions and levels of phase error; these show that data with an FOM of 0.7 or better can be entirely traced with no user intervention.

1. Introduction

Protein crystallography is currently the principal technique used to elucidate the atomic detail of proteins deposited in the Protein Data Bank (PDB; Bernstein *et al.*, 1977; Berman *et al.*, 2000). There are a number of steps within the macromolecular structure-determination process using X-ray crystallography, of which the interpretation of experimental maps is only one. However, it does represent a significant expenditure of time and expertise. There are experimental limitations that mean that most data can only be collected to moderate resolutions or with significant phase error. When this is the case, the interpretation of experimental maps represents a significant bottleneck within the structure-determination process.

Experimental X-ray electron-density maps generated as part of the structure solution of protein molecules are large three-dimensional objects that contain a great deal of information. They also contain errors that result from the limitations of data collection. Two major aspects that affect the structure-determination process are resolution and phase error. Where phased data are truncated to a resolution of 2.0 Å or worse, the electron density becomes smeared so that the atomic structure is not defined by peaks within the map. At lower resolutions (3 Å), the experimental detail only gives information on the chemical grouping within the structure; therefore, the positions of atoms are defined by the internal geometry of these chemical groups or amino acids. Below 4 Å the data only show local fold structure, such as helices and β-sheets, and therefore the precise positioning of atoms becomes very difficult. The quality of the available phases affects the overall clarity of this detail within the electron density. It affects the amount of data smearing and introduces a number of erroneous details that can prevent correct interpretation. At worst, poor phase information produces

experimental maps that contain no useful information and cannot be interpreted.

A significant advance in protein-structure elucidation was provided by the molecular-graphics programs that became available when computer hardware could support the huge computational demands associated with protein crystallography. The first of these programs was *FRODO* (Jones, 1978, 1985) and this has been superseded by *MAIN* (Turk, 1992), *CHAIN* (Sack & Quiocho, 1997), *QUANTA* (Accelrys Inc.), *XTALVIEW* (McRee, 1999) and the most well known program *O* (Jones *et al.*, 1991; Jones & Kjeldgaard, 1997). Another major advance for protein-structure interpretation was the use of the technique of skeletonization (Greer, 1974) that provides data reduction of the three-dimensional map into a series of ridgelines or bones. Map tracing involves the conversion of the bones ridgelines into atoms (Jones & Thirup, 1986; Jones & Kjeldgaard, 1997). More advanced techniques defined a 3.8 Å length 'baton' (Oldfield, 1994; Jones & Kjeldgaard, 1997; McRee, 1999) that can be used by the crystallographer to map out a C^{α} trace using the bones as an information backdrop. Levitt's *MAID* program (Levitt, 2001) uses a bones analysis to identify secondary structure as three-residue constructs from the bones points and extends the polypeptide chain from these regions. Recently, methods that carry out a pattern analysis of electron-density peaks and water positions (Perrakis *et al.*, 1999; Oldfield, 2002*b*) have been shown to be extremely effective. Unfortunately, these methods are limited to higher resolution data, as the experimental detail must be sufficient to have peaks close to the correct atomic sites for convergence. Owing to diffraction limits imposed by crystal quality, most experimental data cannot be collected at resolutions suitable for this. Interpretation methods that work with lower resolution data that identify secondary structure both in real space and reciprocal space are available (Kleywegt & Jones, 1997; Cowtan, 1998; Oldfield, 2002*a*) and provide a starting point over a broad range of resolutions. In contrast to these techniques, this paper describes a method of protein electron-density map tracing using a number of rules to identify the correct position of the C^{α} atoms. It is based on a pathway analysis using the bones ridgelines and does not require the bones points to define atomic positions; it is therefore suitable for use with data below 1.5 Å.

2. Methods

The algorithm described uses a sequential trace method; that is, it places one C^{α} atom at a time based on the position of previously placed C^{α} atoms. Hence, the method of C^{α} placement must be error-tolerant even where there are deviations in previously placed atoms so as to prevent error accumulation. A tracing method that works for resolutions below 1.5 Å must be able to work where the map peaks do not necessarily define atomic positions and must be able to handle problems associated with phase error such as false connectivity. 16 different rules (Table 1) are described to place the next C^{α}-trace atom within a growing segment of trace atoms. The rules are based on the analysis of bones ridgelines, real-space fit to the experimental data and a number of short-range and long-range geometrical terms. Each rule adjusts a weight for each of a number of possible trial positions for the next atom in the trace; the best-weighted coordinate is taken as the next atom in the trace. They are certainly not an exhaustive selection of scale factors that could be used for tracing electron density, but they do cover the essential aspects of map interpretation. Exit conditions for the tracing process are also described, as it is necessary to know when to stop. The last issue that is addressed is that of defining the program parameters required for optimal tracing. The method is based on the presence or absence of bones ridgelines and therefore the correct parameterization of the calculation of these is critical to the quality of the trace results. A consensus tracing method is therefore presented that optimizes the bones start parameter on the basis of both the quality of the trace results obtained at different values and whether parts of the map are consistently traced at different parameter values.

Table 1
Decision rules used for tracing.

Rule	Property	Weight scale
1	Main-chain bones with respect to side chain	1.6
2	Branch point at nearest bones point	1.5
3	Branch point close to bones point	1.3
4	Opening angle $\leq 50^{\circ}$	1
5	$50 < \text{angle} < 70^{\circ}$ and $160 < \text{angle} < 180^{\circ}$	3
6	$70 < \text{angle} < 160^{\circ}$	9
7	Density statistics of bones path	$\rho(\text{min}) + \langle\rho\rangle$
8	Four-parameter RSR	ρ fit
9	Secondary-structure equivalence to $C^{\alpha}(i)$	2
10	Secondary-structure equivalence to $C^{\alpha}(i-1)$	If (condition 9), 2
11	Good vector entry/continue angle	3/3
12	OK vector entry/continue angle	2/2
13	Bad vector entry/continue angle	0.5/0.33
14	Non-bond contact < 2.5–2.8 Å	0.01–1.0
15	Log_{10} (occurrence) in Ramachandran space	0.1–0.01
16	Free-G fit zero	Abort
17	No path: interaction with terminal C^{α}	Successful abort
18	No path: interaction with non-terminal C^{α}	Abort: path error

2.1. The start point for tracing

The description here defines the rules necessary to place a single atom into the electron density relative to a fitted fragment of C^{α} trace and therefore a starting point of analysis must already exist. This starting point can be defined by the placement of an initial C^{α} position manually by a crystallographer or by conversion of a vector defined as the principal component of a secondary structure generated from the electron-density map (Oldfield, 2002*a*). The last C^{α} atom already placed is used to define a bones ridgeline seed point (Fig. 1, blue circle), which is the closest bones ridgeline point to the C^{α} atomic coordinate. Each analysis cycle updates this bones seed point using the last placed C^{α} atom. The addition of new atoms assumes that there are at least three existing C^{α} positions. If this is not the case, then some scaling factors described are undefined and therefore cannot contribute to the overall atom-placement probability. New C^{α} positions can

be added to either end of this starting set of C^{α} atoms, as the directionality of the map is generally unknown during much of the tracing process. All the rules described (Table 1) are therefore independent of the trace direction.

2.2. Pathway analysis: initial trial sets

Pathways within the electron-density map are defined by analysis of bones ridgelines starting from a bones seed point (Fig. 1, blue circle). The bones are analysed to determine all points that lie 3.8 Å from this start point and are joined to this seed point by the bones ridgelines (Fig. 1, red circles). Each trial point generated by this analysis has an associated bones pathway, which is the list of bones points for the shortest route to it. Initial weighting for the trial point list is based on the bones classification for the route to the trial points (Oldfield, 2002*b*); that is, whether the path is mostly main chain or side chain as defined by the bones processing algorithm (rule 1 in Table 1). Weights are adjusted by the presence of branch points in the bones that occur close to each trial point (rules 2 and 3 in Table 1) and by the opening angle defined by the three coordinates '$C^{\alpha}(I-1)$—$C^{\alpha}(I)$—trial point' (rules 4, 5 and 6 in Table 1). A trial position is therefore preferred if it is close to a bones branch point in the bones, if it runs along main-chain bones and where it contributes to a C^{α} trace with a particular curvature. A further scale factor is applied to each of the bones trial points based on the density quality along the bones pathway (rule 7 in Table 1). This scale factor is the normalized sum of the average and minimum density in the bones path, resulting in a greater weight for a good density route that has no significant minimum. It is found that these initial trial points are useful where the C^{α} trace forms an extended conformation, but where the trace has high curvature they often represent a short cut. A second set of mass-weighted trial points is therefore generated from the 3.8 Å bones points, where the weights for each point are inherited from the initial set (Fig. 1, green circles). The coordinate of each new point is defined by a 3.8 Å vector that joins the start bones point (blue circle) and passes through the unweighted centre of mass of the bones pathway to the parent trial point.

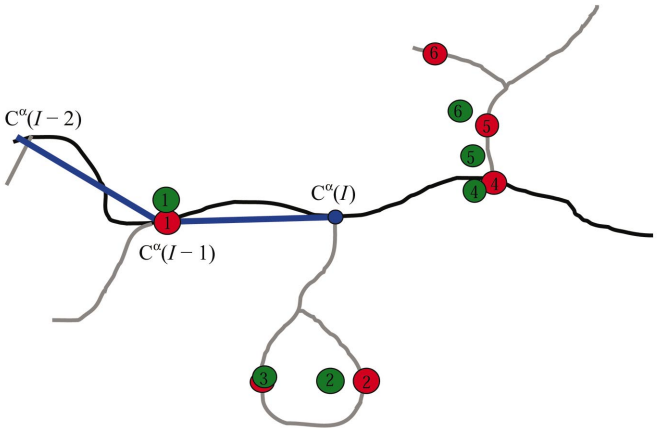

Figure 1
The black trace shows a main-chain trace of bones ridgelines and the grey trace indicates a side-chain bones trace. The blue lines indicate the current fitted C^{α}-trace atoms joined by pseudo-bonds. The blue circle indicates the nearest point on the bones trace to the current C^{α}-trace atom and the red circles are points on the bones that are 3.8 Å from this blue point. Green circles are 3.8 Å vector points determined as the unweighted average point of the bones pathway to respective red circle trial points.

2.3. Density fitting: trial point refinement

The next analysis uses the electron-density map as a refinement target to both scale each possible trial point and to improve its position. The map fit is carried out using a bounded-gradient real-space torsion-angle refinement as a function of the four parameters shown in Fig. 2. Refinement details are discussed in Oldfield (2001). The four parameters define the length of a C^{α}—C^{α} pseudo-bond, the opening angle with respect to the previous two C^{α} atoms, a torsion angle with respect to the previous three C^{α} atoms and a torsion parameter that defines the orientation of the C^{α} atom and its virtual covalent neighbours: atoms N, C and C^{β}. The length of the C^{α}—C^{α} pseudo-bond is bounded by values of 2.8 and 4.0 Å, representing a significant bias to shorter lengths than the expected 3.8 Å. This is necessary because the electron-density map volume (and therefore the bones) tends to be closer to the centre of path curvature than the real coordinate pathway that it describes. This is particularly the case at resolutions lower than 4 Å, where an α-helix ridgeline pathway collapses to a straight line. The skew of this parameter to a value less than the optimal C^{α}—C^{α} distance is important to maintain fit register while tracing bones that traverse a shorter path than the model coordinates could take. The incorrect pseudo-bond length can be corrected by C^{α}-trace real-space torsion-angle/opening-angle refinement with pseudo C^{α}—C^{α} bond parameters set to 3.8 Å (Oldfield, 2001). The fit to the map of the tetrahedral C^{α} and its covalent atom neighbours at each trial point provides a new weight which is the product of the refinement minimum value and the old weight (rule 8 in Table 1). The coordinate of each trial position is updated to the minimum found by the refinement.

2.4. Geometry screening

The next weights are defined using geometric terms calculated from the last fitted section of trace. Scaling is first applied

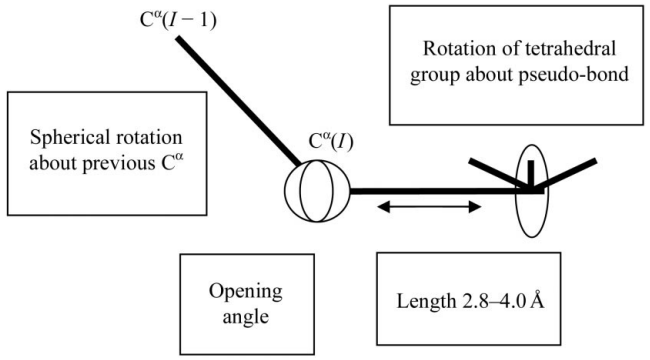

Figure 2
Refinement parameters use for the placement of a C^{α}-trace atom into the electron density.

to the trial point weights if there is continuity of either β-strand or α-helical structure within the growing C^{α} trace. The local secondary structure is determined using a C^{α} geometry map (Oldfield & Hubbard, 1994) from four C^{α}-trace atoms. If the secondary-structure type for a trial point matches the previously placed C^{α}-trace atom, then the trial point weight is doubled (rule 9 in Table 1). If it also matches the secondary structure of the previous-but-one placed C^{α}-trace atom, then the trial weight is doubled again (rule 10 in Table 1). A further secondary-structure weight is applied using information determined from a pattern-recognition algorithm that finds helices and strands within experimental maps (Oldfield, 2002*a*). Since this pattern recognition is designed to determine the complete secondary structure within an experimental map, the target values provided by this information represent a long-range target for C^{α} placement. The aim is to increase weights for trial points that are consistent with this overall packing of secondary structure. A scale factor is applied to each trial point based on entry and continuation conditions with respect to the secondary structure (rules 11, 12 and 13 in Table 1) with the parameters defined in Fig. 3. The conditions are based on the vector angle and distance between a baton, defined as the vector between the atom $C^{\alpha}(I)$ and current trial point and each secondary-structure vector. Entry to a secondary-structure element is defined when the centre of the baton is less than 4 Å from it and where the last fitted trace baton was more than 4 Å from the secondary-structure element. The scale is defined on the basis of the angle made between the secondary-structure vector and the baton (Fig. 3) and a high scale value occurs where the two vectors are nearly collinear. A secondary-structure continuation scale is generated when both the current and last fitted baton are within 4 Å of the same secondary-structure vector; scaling is defined on the basis of the same opening angle. In particular, this scaling prevents tracing perpendicular to the strands of a β-sheet even where the map quality indicates this is an ideal pathway. For a crossing condition, the entry angle to the new secondary-structure element will be relatively small (~90°), which results in a scale value less than unity and also results in a bad continuation metric (less than unity) because the vector number has changed by this trace route.

The last geometric weight is based on non-bond interactions to previously placed C^{α}-trace atoms and their symmetry. This is checked by determining distances less than 2.8 Å (rule 14 in Table 1) using a lattice non-bond search. The trial point weights are reduced 100-fold when the separation is below 2.5 Å and linear interpolation is used between 2.5 and 2.8 Å where the weight scale factor is 1 for distances greater than 2.8 Å.

2.5. Probabilistic weighting in Ramachandran space

The last rule that defines a tracing weight is based on a probability of occurrence of a path generated by addition of each trial point within protein conformational space. This is based on statistical information empirically derived from the PDB. The geometry of a polypeptide conformation can be defined by four C^{α} atoms using two opening angles and one torsion angle (Fig. 4). This information has been mapped statistically (Oldfield, 1994; Oldfield, in preparation) to the occurrence of geometry with protein backbone atoms defined by Ramachandran angles (Ramachandran & Saisiskharan, 1969). The direct mapping between C^{α} conformation space and Ramachandran conformational space is approximately 95% complete and provides information on the rate of occurrence as well as a statistical error. The $\log_{10}$ value of the normalized numerical occurrence within Ramachandran space is used to scale the current weights on each trial point. The scale value (rule 15 in Table 1) is bounded to prevent domination of the helical conformation in proteins and to prevent the return of zero weights for conformations not observed within protein structures.

2.6. The free-geometry fit

The trial point with the greatest weight represents the best new C^{α} atom and this is added to the current segment of trace. Once this atom has been added, further analysis is performed to determine the probability that the trace is a sensible fit to the electron density using a free-geometry fit (free-G). The change of this value for the last five atom placements is also determined. The free-G fit for a segment of C^{α} trace is calculated by real-space density-fitting peptide planes between each consecutive pair of C^{α} atoms (Oldfield, 2001) and then

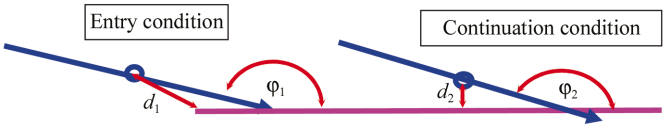

Figure 3
Parameters used to define the placement of a C^{α}-trace atom with respect to the secondary-structure element vectors. The vector is shown in purple, the current C^{α}-trace pseudo-bond is shown in blue with an arrow to show the trace direction to the current trial point and distance/angle parameters are shown in red. φ_1 is the entry angle and is defined if $d_1 <$ 4 Å, where the previous placed trace atom mean point had $d_1 > 4$ Å. The weighting is defined by the angle value, with $\varphi_1 > 140°$ being good, $140 > \varphi_1 > 110°$ being OK, $110 > \varphi_1 > 80°$ being poor and $80° > \varphi_1$ being bad. φ_2 is the continuation angle where the last placed C^{α} -trace atom mean point had $d_1 < 4$ Å with the current value $d_1 < 4$ Å. The weighting is defined by the angle value, with $\varphi_1 > 160°$ being good, $160 > \varphi_1 > 140°$ unweighted and $140° > \varphi_1$ bad. If the current value $d_1 > 4$ Å when the previous trace atom had $d_1 < 4$ Å then the trace just exited a secondary-structure element and the weight is unity. Weights are shown in Table 1.

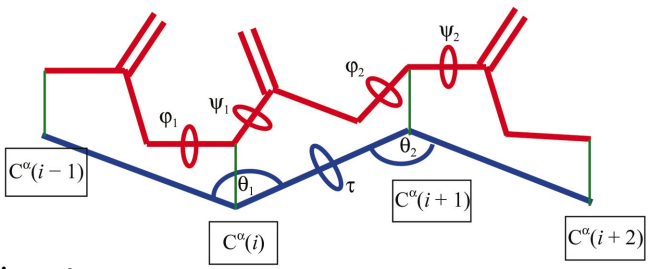

Figure 4
Figure to show the mapping of three C^{α}-trace parameters (two angle and one torsion) to two pairs of Ramachandran angles. For each four C^{α}-trace atoms, the occurrence of the mapping into Ramachandran space is used to define a probability weight that the trace conformation is credible.

calculating the sum of the deviations of the N—C^{α}—C bond angles from target values. The residual is described as a free-geometry fit because no restraint information for this opening angle is used during the calculation of the optimal fit for the peptide plane to the electron density. Since the direction of the C^{α} trace is indeterminate, the peptide-plane fitting is carried out both in a forward and a backward direction and the best value is taken as the metric. The measured value is redefined in the range 0% for geometry incorrect at each C^{α} atom to 100% for a perfect fit at each C^{α} atom. Errors in the placement of C^{α} atoms with respect to the experimental map result in peptide-plane fitting errors and a subsequent deviation of the measured opening angle from the target. Since there is correlation between consecutive geometry values along the trace, this measure is extremely sensitive to trace errors and easily falls to 0% where tracing has gone wrong. The function can be defined over different regions of a fitted trace, so it is possible to define a gradient that represents the change in the free-G fit for overlapping five-residue fragments. This information is used to abort a trace if the fit falls to zero locally and does not rise for five placed atoms. This metric is also displayed to the user as a pie chart during the tracing process to provide feedback during tracing.

2.7. Exit conditions

This completes the analysis necessary to place a single C^{α}-trace atom at the trial point with the highest weight. Tracing consists of repeated addition of C^{α} atoms until an exit condition occurs. A trace failure can occur when no path can be found in the map owing to the absence of a bones ridgeline pathway or because all possible pathways result in an atom clash or bad free-G. If an atom clash occurs with another terminal C^{α}-trace atom, then this is assumed to be a sensible join point to another trace as long as it does not result in a loop; this is the only successful trace exit condition. The following procedure is performed to handle unsuccessful exit conditions. The algorithm keeps a record of all the trial positions and weights for the two previously placed C^{α}-trace atoms in a stack of depth two. If an unsuccessful exit condition occurs, then this atom is deleted and the algorithm works back though the remaining sorted list of the possible paths and tries to extend each of these in order of weight magnitude. If no path can be found from any of these, then a further C^{α} atom is deleted and the exercise repeated for the previous path weights. Therefore, this procedure can backtrack two C^{α}-atom positions, for example where the rules have resulted in tracing along an amino-acid side chain. Where there really is no possible path available, the algorithm will stop tracing after returning one of the C^{α}-trace atoms from the depth search.

The tracing algorithm described is designed to be used after generating vectors of likely secondary structure from the pattern-recognition search of a map. The protocol of automated tracing tries extending from each end of a secondary-structure vector in turn, starting with the helix with the best fit to the density. A good map with no breaks can therefore be traced starting from a single helix, but where there are breaks in the density each secondary-structure element found from the vector-pattern search is built from both ends to give significant map coverage.

2.8. Parameter optimization: consensus tracing

A critical aspect that affects the quality of results from the tracing is the initial selection of pathway points using the bones ridgelines. Therefore, it can be expected that the parameter for the calculation of bones ridgelines will be critical to the quality of the trace obtained, which is therefore down to the experience of a user in setting this correctly. When a low value for the bones parameter is used, the bones ridgelines will have too many inter-connections, which can overwhelm the algorithm and result in a wrong trace. A high value for the bones parameter results in breaks within the bones, which can prevent tracing as no pathway will exist. Consensus tracing is used to solve this non-deterministic problem similar in method, but not in implementation, to that used within *trace* and *wARP* (Perrakis *et al.*, 1999). The method of consensus tracing involves interpreting the electron-density map nine times with a different bones start parameter, then attempting to automatically identify the ideal trace. Each of the nine traces is analysed in turn and the quality index (TQ_i) is defined for each trace as the variance along the trace atoms compared with all other traces. The equivalent atoms used within the comparison between traces is the closest atom found in each of the other traces. The best trace is defined as that with the lowest quality index and represents the trace with the lowest variance from all other traces,

$$TQ_i = \frac{1}{N_i^2}\left[\sum_{j=1,9(i\neq j)}\sum_{n=1,N_i} y_{n,i,j}^2\right],$$

where TQ_i is the trace quality for trace 'i' and $y_{n,i,j}$ is the separation between atom 'n' in the trace 'i' and the nearest atom in trace 'j'. The index is normalized by $1/N^2$ to decrease the value of TQ_i for long trace lengths, as these are more desirable. The variance of each atom C_n^{α} is defined as the sum of the distances to the nearest atoms within the eight other traces.

3. Results

Six experimental σ_A-weighted maps were generated using all the available data with no resolution cutoff. The maps were read by *QUANTA* 2000.2, bones calculated and edited using the tools to generate a single molecular object with no symmetry overlaps. A map mask was generated from the bones and used as a bounding surface for the tracing process. Secondary-structure vectors were calculated by pattern recognition at the best bones parameter (Oldfield, 2002*a*) and these were used as seed points for the tracing process. A single GUI tool was used to start tracing. Tracing involved conversion of all helix vectors into ideal C^{α} traces where the principal component refinement residual (PCRR) is better than 0.7 error weight (Oldfield, 2001) and then continues by adding

further C^{α} atoms to both ends of each helix in turn. The order of the build was defined by sorting the PCRR for each secondary-structure vector and starting with the best-fitted helix element. On completion of tracing starting from the helix vectors, any untraced volume of the map was interpreted starting from strand vectors with PCRR better than 0.7. A repeat cycle was carried out for those strand vectors where the PCRR was better than 1.4 error weight for the map volume still untraced. Tracing was complete when the algorithm added no more C^{α}-trace atoms after trying all vectors as starting points. This entire procedure was repeated nine times with different bones ridgeline start parameters. A central bones start value for the consensus tracing can be defined using a bones tree over/under-connectivity quality index (Oldfield, in preparation), although a value of 1.1σ was used for all the maps in this study. The multiple-trace analysis was performed using nine 0.04σ steps bracketed about this value of 1.1σ. The multiple-tracing procedure is implemented as a single tool. Exactly the same procedure was performed using a non-graphical server ported to a PC running Microsoft Windows and the timings for these calculations are additionally shown in Table 2.

Table 2
Basic tracing results.

The tracing time represents the time for a single interpretation of the electron density; consensus tracing takes approximately nine times the quoted value. Times are quoted for an R5000 SGI (180 MHz) (SGI) and a PC running Windows 98 (1.7 GHz Athlon). Res is the resolution of the map, N_{res} is the number of residues within the published structure and FOM is the figure of merit. N_{res} consensus defines the number of residues traced using the described algorithm within the times shown.

	Res (Å)	N_{res}	FOM	N_{res} consensus	Time SGI (min:s)	Time PC (min:s)
RNase(1)	2.5	96	0.7	89	0:55	0:04
RNase(2)	2.5	96	0.5	78	0:12	0:01
CysB	1.8	236	0.53	82	0:23	0:05
EMTA	2.75	184	0.57	174	1:17	0:05
OMPLA	2.7	256	0.75	255	2:15	0:09
PA(1)	2.5	750	0.52	527	9:12	0:52

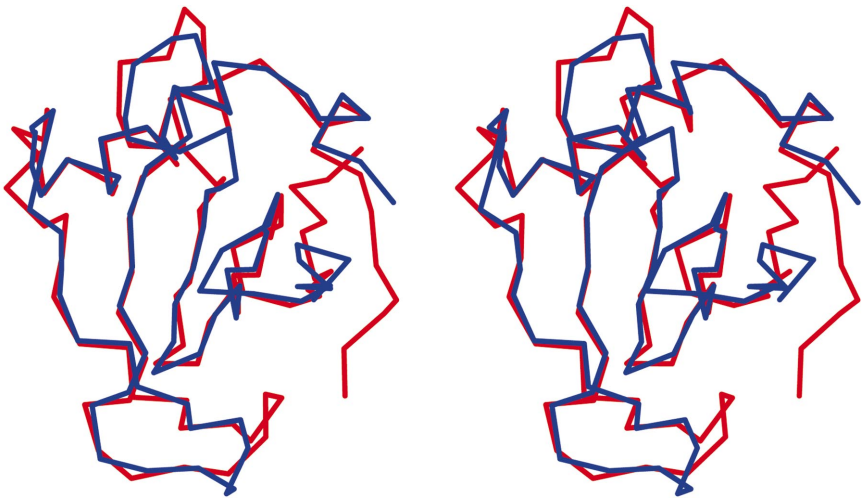

Figure 5
C^{α} trace of RNase(1) generated for the electron-density map superposed on the final coordinates. The published coordinate trace is shown in red and the C^{α}-traced atoms generated by the described algorithm are shown in blue.

The ribonuclease SA (RNase; Sevcik *et al.*, 1991) example map was traced with experimental data before (data set 2 in Tables 2 and 3) and after (data set 1 in Tables 2 and 3) the use of density modification (Cowtan, 1998) to provide maps at different extremes of phase quality. The estimated phase error is 74° before density modification and 58° afterwards, with an associated improvement in the electron density. This is a small protein with one helix and a three-stranded β-sheet. The quality of the density-modified map for data 1 is good and the trace was almost entirely generated with no incorrect trace connectivity (Fig. 5). This trace was correctly generated despite a number of false connections between the bones ridgelines within the β-sheet owing to the hydrogen bonding. The scale-factor weights associated with the presence of secondary-structure vectors prevents mistracing between the strands of the β-sheet. There was some deviation in the terminal atoms, where the trace method aborts owing to a disulfide bridge that connects residue 7 to the C-terminus. The tracing algorithm places just one C^{α}-trace atom along the disulfide bridge at the C-terminal cystine residue and then aborts because a disulfide bridge breaks a number of the geometrical rules and generates a large drop in the free-G fit. Since there was no main-chain path to continue along at the C-terminus, the single-atom error in the trace was not deleted because it represented the exit condition.

The map of RNase before density modification has a number of features that make it difficult to interpret within the β-sheet (Fig. 6), indicating the importance of the secondary-structure information and the use of this information to aid in map tracing. Only 78 of the 96 residues were traced without user intervention, although a single manual edit and subsequent continued auto-trace results fits a further 12 C^{α} atoms. The figure shows the β-sheet of the protein with the traced coordinates superposed on the electron-density map and final coordinates of the protein. As with the density-modified data, the algorithm does not trace through the disulfide bridge even though density was good because of the geometric rules.

The outer membrane phospholipase A (OMPLA) structure (Fig. 7) is composed entirely of β-sheet, while the structure of endo-specific membrane-bound transglycosylase (EMTA) (Fig. 8) is entirely α-helical. The maps for both these structures were almost completely traced by the consensus trace, indicating the ability of the algorithm to handle different structure types without problem. The atoms not traced are on the surface of the protein and were not placed because of the lack of connectivity within the electron density in these regions owing to data error.

The map for penicillin acylase (PA) has a number of errors and has a figure of merit

(FOM) of 0.52 and phase error of 78°, with the result that much of the map information was missing within the loop structure. The problem of manual interpretation of this molecule with space group *P*1 is exacerbated because of its size. This map is close to the limit of interpretability of electron density and represents a difficult test for the tracing algorithm; even so, 60% of the map was interpreted in a little over 9 min per trace cycle, although with some errors. The CYSB map was also poor quality with many breaks and so the tracing algorithm did not perform well because of the breaks in the bones pathway.

Table 3 shows an assessment of the errors that result from the maps traced without user intervention. The table shows the number of trace atoms that deviate from the coordinates published within the PDB at three different deviation limits. Table 4 shows the results of tracing the density-modified map of ribonuclease SA with σ start levels used for the bones skeletonization at different bones start levels. The amount of trace generated is rather sensitive to the parameterization of the bones, with values generally equivalent over a range of 0.5σ. This demonstrates the importance of consensus tracing, which determines the best trace over a number of different bones start parameters to optimize the results.

Table 5 provides an assessment of tracing errors based on the presence of insertions and deletions with respect to the final coordinates; these are commonly known as register errors. A register error is noted where the residue index skips within the model-built trace. Insertion and deletion events did not occur within the secondary structure, as these regions are usually well defined within the electron density and the trace was fitted as rigid-body sections. There are a number of trace errors within the loop regions. There appears to be fewer trace errors for CYSB and PA data, but tracing of these structures was not continuous owing to missing electron density in loops, so errors occur as non-tracing of the electron density rather than register errors.

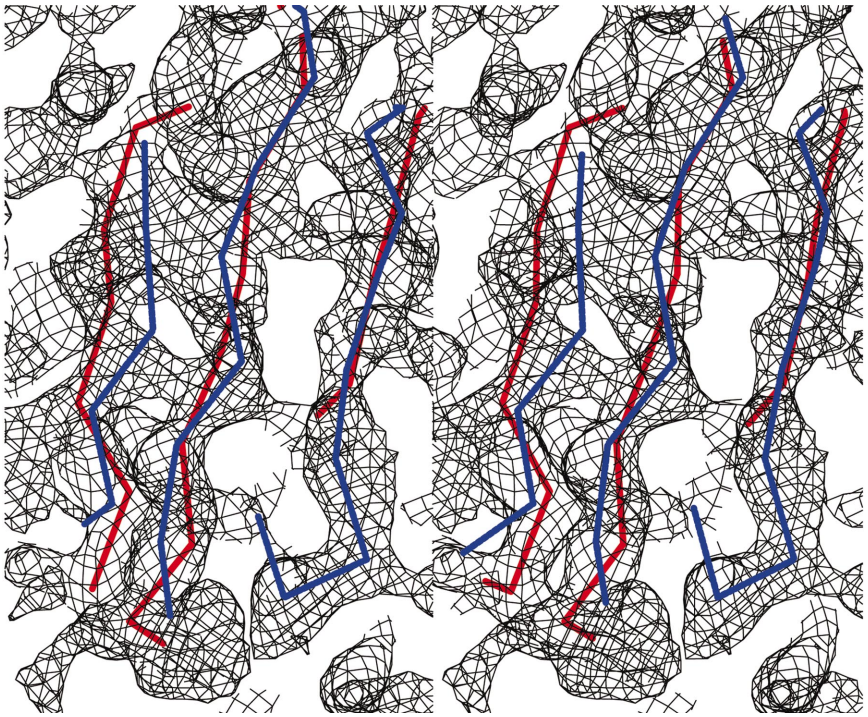

Figure 6
C^{α}-trace atoms within the β-sheet of RNase(2) superposed on the final coordinates of the structure and the map added at 1.2σ. The published coordinate trace is shown in red and the C^{α}-traced atoms generated by the described algorithm are shown in blue.

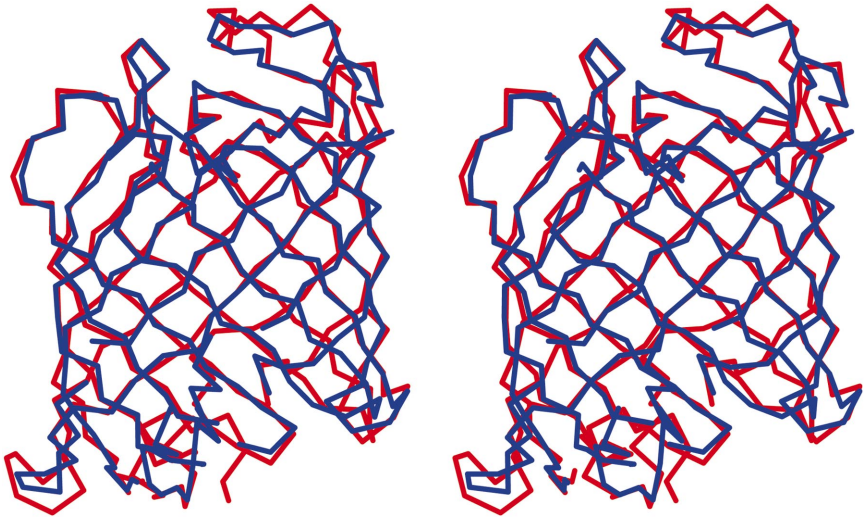

Figure 7
The C^{α}-trace of OMPLA generated for the electron-density map superposed on the final coordinates. The published coordinate trace is shown in red and the C^{α}-traced atoms generated by the described algorithm are shown in blue.

4. Discussion

The results show that the tracing method is able to work at a number of resolutions between 1.8 and 2.75 Å and with a range of different phase errors as estimated by the FOM. In each case it was not possible to use the entirely automated method of *trace* and *wARP* (Perrakis *et al.*, 1999) owing to resolution and data-quality criteria. The density-modified maps of RNase, OMPLA and EMTA were essentially completely traced starting from the secondary-structure vectors. Where the FOM is lower, it is possible to trace a significant proportion of a map without user intervention, but it also necessary for a competent crystallographer to adjust errors and restart the automated tracing. It should be noted that phase-combination methods allow the use of the

Table 3
Assessment of errors.

N_{res} is the number of residues within the protein and N_{trace} is the number traced, while the next three columns (1.0, 1.5 and 2.0 Å) define the number of traced atoms within the defined distance of the final coordinates.

	N_{res}	N_{trace}	1.0 Å	1.5 Å	2.0 Å
RNase(1)	96	89	51	69	83
RNase(2)	96	78	32	48	62
CysB	236	82	23	42	59
EMTA	184	174	70	126	159
OMPLA	256	255	122	190	230
PA(1)	750	527	159	308	407

Table 4
Tracing results for the RNase(1) map using a different bones start parameter defined at the map σ value.

Start/σ	N_{res}	1.0 Å	1.5 Å	2.0 Å
1.12	30	13	15	26
1.10	80	37	56	65
1.08	80	37	56	65
1.06	88	51	69	83
1.04	88	51	69	83
1.02	88	51	69	83
1.00	77	43	57	67
0.98	77	43	57	67
0.96	80	43	58	69
0.94	80	43	58	69
0.92	73	26	44	62

Table 5
Chain insertion and deletion errors from the traced structures determined by comparison with the final coordinate structures.

	N_{res}	N_{trace}	Insertions	Deletions
RNase1	96			31, 44–45, 61
RNase2	96			25, 30, 34, 40, 43, 51
CYSB			139	
EMTA	184		37, 90–91, 154	30
OMPLA	256		21, 43–44, 71–72, 119, 201, 237, 240	
PA			250, 346, 545, 639	

partially built structure to improve map quality, which can then be subsequently retraced without user intervention. Trace-connectivity errors must be avoided and the method described here is very resistant to producing these, particularly within secondary structure, and none are produced in RNase, EMTA and OMPLA. Insertion and deletions errors within a trace would be considered serious if any subsequent map analysis assumed that the trace generated were complete and continuous. The outcome would be register errors when the sequence information is applied. These out-of-register errors can be difficult to correct at a later stage in the structure-determination process because this error propagates as a result of the phase combination of the incorrect model and experimental phase information. Since the automated methods used to assign sequence (Oldfield, in preparation) are designed to signal the presence of deletions/insertions, chain-direction error and connectivity error this problem is not critical, but it does represent a limitation of the algorithm described. It does allow the possibility of the generation of error during structure determination if an inexperienced crystallographer overrides error analysis or where error analysis is not robust. Since unambiguous sequence alignment is only possible when tracing is perfect, these trace errors represent a constraint to complete automation of protein-structure determination at moderate data resolution and phase quality. An additional problem not handled by the tracing algorithm is the detection of *cis*-peptide bonds, as it would be unreliable to define these from maps of moderate resolution with significant phase error.

The results do indicate that the rule-based method of tracing described represents a very rapid way of tracing protein experimental data even where there is significant phase error or where the resolution is sub-optimal. It is possible to trace experimental data at lower resolutions: some success has been obtained with 6 Å data phase-extended from 10 Å electron-diffraction data using wavelet-analysis phase extension (J. Wilson, personal communication). In general, the resolution limit can be predicted to be 4 Å owing to the loss of map detail at this resolution and results here show success at 2.75 Å. Initial testing has been undertaken to determine whether the trace results are suitable as a starting point for automated crystallography. Test calculations with the EMTA, OMPLA and RNase maps show that it was possible to trace, sequence assign and model build (*QUANTA* 2000.2) without user intervention. Two subsequent cycles of automated model rebuilding (*QUANTA* 2000.2) interspersed with reciprocal-space refinement (*CNX*) can reduce the free *R* factor by 10% (data not shown). These results suggest that general automated structure solution from phased data may be possible when validation methods are powerful enough to always correctly identify wrong local structure and particularly register errors. These results

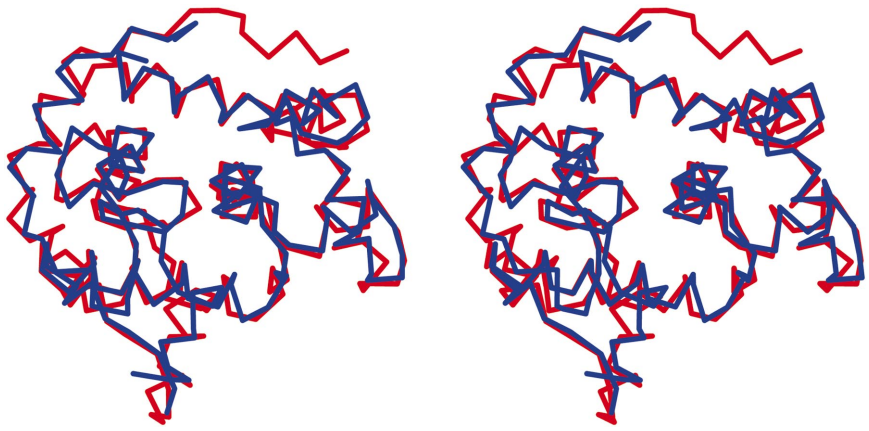

Figure 8
The C^{α} trace of EMTA generated for the electron-density map superposed on the final coordinates. The published coordinate trace is shown in red and the C^{α}-traced atoms generated by the described algorithm are shown in blue.

suggest that these trace results provide a good starting point for a subsequent user-free structure-determination process, although considerable effort will be required to provide a useful program. With this in mind, the method described here, along with all the other X-ray features within *QUANTA*, have been converted to provide a non-graphical server. It can be seen that the use of a server-based program running on a PC gives a 10–15-fold increase in speed over *QUANTA* running on an SGI R5000 (Table 2).

The method described has a number of analyses that contribute to the weighting scheme to decide on the position of a C^{α} atom. These weights are based on the initial generation of multiple pathways within the map and then a decision analysis to choose the most likely candidate pathway. The method is based on electron-density quality as well as a number of short-range and long-range geometrical restraints. The free-G fit has only been used to a limited extent to date and further analysis of the gradient would ideally provide better qualitative analysis of the tracing process. No rule appears to offer significant advantage over the others; each seems to provide a small advantage for different map conditions. The magnitude of the weights has been empirically and systematically adjusted to improve the tracing of a number of different maps, but since no particular rule dominates, optimization is a difficult balance. Many of the scale factors used are discrete values rather than continuous functions, although the implementation of the latter may improve tracing if good functional forms can be determined.

The tracing is entirely dependent on the parameterization of the electron-density bones and therefore the ability to determine the correct start values for the bones calculation is critical for successful use of these methods. This has in part been solved by the use of consensus tracing, where a number of different traces are generated for different bones skeleton start values. The best trace is taken as the one that produces atoms that are most reproducible and weighted by the length of the trace generated. Since the initial set of trial points is dependent on the presence of the bones pathway, the tracing algorithm cannot traverse a break in the map, while over-connection of the bones can result in a completely wrong trace. The consensus tracing therefore provides a means to optimize the results as a function of map connectivity. Improvements are possible and these are under investigation. The consensus-trace method could be improved, as it is necessary for all regions of the map to be identified by a single trace using the current method. Combination of the multiple traces into a single result is under investigation using the variance along each trace to mix and match the trace results. A method of predictive C^{α}-atom placement is also under investigation where the cumulative weights for one or more C^{α} atoms ahead affect the path weights of the current atom placement. Initial analysis indicates that the complex weighting scheme required is likely to be problematic for predictive tracing. The user interface of *QUANTA* 2000 includes the ability to mark the sequence table with a predicted secondary structure and it may be useful to include this prediction within the tracing algorithm to improve the analysis.

5. Availability

The methods described are implemented in the program *QUANTA* 2000.2, which is available from Accelrys Inc.

I am indebted to Francesca Gliubich for carrying out some of the initial analysis for this paper and also Francesca and Andrew English on feedback from user-free *de novo* structure solutions using the tracing algorithm described and other methods. I would like to thank the crystallographers, Joseph Sevik, Koen Verschueren, Arian Snijder, Francesca Gliubich and Peter Moody, who provided data to test the algorithm and members of the YSBL for suggestions.

References

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.
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.
Cowtan, K. (1998). *Acta Cryst.* D**54**, 750–756.
Greer, J. (1974). *J. Mol. Biol.* **82**, 279–284.
Jones, T. A. (1978). *J. Appl. Cryst.* **11**, 268–272.
Jones, T. A. (1985). *Methods Enzymol.* **115**, 115–157.
Jones, T. A. & Kjeldgaard, M. (1997). *Methods Enzymol.* **227**, 173–230.
Jones, T. A. & Thirup, S. (1986). *EMBO J.* **5**, 819–822.
Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110.
Kleywegt, G. J. & Jones, T. A. (1997). *Acta Cryst.* D**53**, 179–185.
Levitt, D. G. (2001). *Acta Cryst.* D**57**, 1013–1019.
McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.
Oldfield, T. J. (1994). *Proceedings of the CCP4 Study Weekend. From First Map to Final Model*, edited by S. Bailey, R. Hubbard & D. A. Waller, pp. 15–16. Warrington: Daresbury Laboratory.
Oldfield, T. J. (2001). *Acta Cryst.* D**57**, 82–94.
Oldfield, T. J. (2002*a*). *Acta Cryst.* D**58**, 487–493.
Oldfield, T. J. (2002*b*). *Acta Cryst.* D**58**, 963–967.
Oldfield, T. J. & Hubbard, R. E. (1994). *Proteins Struct. Funct. Genet.* **18**, 324–337.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 459–463.
Ramachandran, G. N. & Saisiskharan, V. (1969). *Adv. Protein Chem.* **23**, 283–437.
Sack, S. & Quiocho, F. A. (1997). *Methods Enzymol.* **227**, 158–173.
Sevcik, J., Dodson, E. J., & Dodson, G. G. (1991). *Acta Cryst.* B**47**, 240–353.
Turk, D. (1992). PhD thesis. Technische Universität München, Germany.