Acta Cryst. (1997). D**53**, 179–185

Template Convolution to Enhance or Detect Structural Features in Macromolecular Electron-Density Maps

GERARD J. KLEYWEGT AND T. ALWYN JONES*

Department of Molecular Biology, Biomedical Centre, Uppsala University, Box 590, S-751 24 Uppsala, Sweden. E-mail: alwyn@xray.bmc.uu.se

(Received 20 May 1996; accepted 2 October 1996)

Abstract

A conceptually simple real-space convolution method has been developed which can be used to detect or enhance structural features in experimental macromolecular electron-density maps. The method has been implemented in a computer program (*ESSENS*). One application of the method is in selectively visualizing secondary-structure elements in multiple isomorphous replacement (MIR) maps of proteins, prior to map interpretation. This application is demonstrated for MIR maps of P2 myelin protein [Jones, Bergfors, Sedzik & Unge (1988). *EMBO J.* **7**, 1597–1604; Cowan, Newcomer & Jones (1993). *J. Mol. Biol.* **230**, 1225–1246] and glyoxalase I [Cameron, Olin, Ridderström, Mannervik & Jones (1997). In preparation]. Another application is in finding the optimal orientation and position of a known structural fragment (*e.g.* a protein domain or a ligand) in any type of electron-density map (real-space or phased molecular replacement). This application is demonstrated for the complex of acetylcholinesterase and the snake toxin fasciculin II [Harel, Kleywegt, Ravelli, Silman & Sussman (1995). *Structure*, **3**, 1355–1366] where the toxin was located in a map phased using the molecular-replacement solution for the acetylcholinesterase alone.

1. Abbreviations

AChE, acetylcholinesterase; *CCP*4, Collaborative Computational Project, Number 4; GBD, glutathione-binding domain; MIR, multiple isomorphous replacement; NCS, non-crystallographic symmetry; PDB, Protein Data Bank; SSE, secondary-structure element.

2. Introduction

Electron-density map interpretation is a critical, and usually exciting, event in the crystallographic analysis of the structure of a biological macromolecule. This process requires that the crystallographer maintains both a detailed and an overview description of the density. One way of achieving this is now widely used and makes use of displayed contoured portions of density for a detailed view, and a skeletonized representation to create the overview (Greer, 1974, 1985; Jones & Thirup, 1986; Jones, Zou, Cowan & Kjeldgaard, 1991). The job of the crystallographer is then to evaluate the detailed view to remove errors in the skeleton (Jones & Kjeldgaard, 1994, 1997). Gradually, a picture emerges where the crystallographer is able to recognise macromolecule-like features. For proteins, this recognition usually takes the form of helix and strand identification. Non-protein entities, such as metal ions, heme groups *etc.*, may also be recognised at this stage. In this paper, we describe a technique to aid in this recognition process that makes use of the known structures of the basic macromolecular building blocks.

The method is based on using standard structural entities. Known molecular structures can be used to solve crystallographic problems when no phase information is available. The method is now well established and known as the molecular replacement method. It consists of a multidimensional search in Patterson or reciprocal space to correctly place the fragment. Usually, the search is carried out in two passes, first a rotational search (Rossmann & Blow, 1962), and then a translational search in which the correctly oriented molecule is moved around, relative to the crystallographic symmetry axes (Crowther & Blow, 1967; Rossmann, 1990). Many search functions, operating in real and/or reciprocal space, have been proposed and used with great success. Very early on, Buerger (1950, 1951, 1959) suggested the use of three real-space image search functions, the so-called sum, product and minimum functions. With these functions, a set of interatomic vectors that have been determined from the coordinates of a known molecular fragment, can be compared with a calculated Patterson function. Nordman further improved on the scoring qualities of the minimum function by using just a subset of the set of lowest values (Nordman & Schilling, 1969).

When phase information is available, the search for a particular molecular fragment becomes a six-dimensional affair. Similar real-space scoring functions to those proposed by Buerger can be used, namely sum, product or minimum functions.

3. Algorithm

The present algorithm is fairly simple: a set of atoms (the template) is rotated around a pivot point for each point in the map, and a score is calculated which reflects how well the atoms fit the density for each orientation and at each point. Although we have experimented with several scoring methods, the best results (as judged by inspection of the resulting maps at the graphics) were consistently obtained with a derivative of the minimum function of Nordman (Nordman & Schilling, 1969) in which for each atom the average of the density values at the eight nearest grid points is calculated. These values are sorted, and the K lowest values are summed and divided by K (if the template contains N atoms, $K < N$). Because of the sorting step, this method is slower than calculating the sum of density values for all atoms, but it is much less sensitive to noise. The rationale for using this scoring function is that if a template fits the density well, even the worst-fitting atoms will still fit reasonably well. However, if only a part of the template fits the density, the score will be low. At the end of the calculations, a new map is created in which the value at each grid point contains the best score found for all attempted rotations of the template at that point. This map, when contoured and viewed on the graphics system, will show places where some rotation of the template about its pivot point best fits the density.

The input to the program consists of an electron-density map in *CCP*4 format (Collaborative Computational Project, Number 4, 1994), an atomic coordinate file of the template or search model in PDB format (Bernstein *et al.*, 1977), and some parameter settings (see below). Optionally, a mask (molecular envelope) file can be provided in *MAMA* format (Jones, 1992; Kleywegt & Jones, 1994); only map points which are set in the mask will then be considered by the program.

If the template contains one or more Cα atoms, the pivot point is the Cα atom closest to the centre of gravity of the template; otherwise the actual centre of gravity is used as the pivot point. The rotational search is carried out in Euler or polar angle space, and the user can supply the range of angles to be tested and their step sizes. Normally, a full asymmetric unit of rotational space would be searched (Euler angle ranges: $0 \leq \alpha < 360°$, $0 \leq \beta < 180°$, $0 \leq \gamma < 360°$) unless the template contains internal symmetry (*e.g.* a heme group), or if one wants to fine tune the solution found in a previous run of the program. For protein feature-enhancement calculations we find that two templates are particularly useful: (*a*) a penta-alanine in 'ideal' α-helical conformation, and (*b*) a penta-alanine in 'ideal' β-strand conformation. Increasing the length of the peptide template may help in recognising longer structural units, but would hinder seeing smaller ones. Obviously, the choice of template is not limited to regular secondary-structure elements. By using the indole ring of a tryptophan residue, for instance, the program can be useful to locate aromatic residues. Similarly, a template consisting of a single alanine residue may be helpful for tracing the chain in loop regions, as well as for finding the positions of residues with long side chains. For structures containing RNA, DNA, carbohydrates or other regular polymers, suitable templates can be derived as well. In general, any structural fragment which is (or is expected to be) tightly restrained geometrically (*i.e.* rigid) can be used as a template.

In order to speed up the calculations further, a density cut-off can be applied (*e.g.* to only evaluate map points with positive density if the average density in the cell equals zero), although this may be hazardous for large templates and templates which do not contain any Cα atoms. All map points which (*a*) lie outside the mask, or (*b*) fail the density cut-off, or (*c*) are too close to the border of the map to allow full rotation of the template, are rejected. For each of the remaining points, the rotational search is carried out and the best score stored. If one wants to find the optimal fit of a larger template (*e.g.* a domain or whole protein), the best global solution can be stored; at the end of the calculation the optimal rotation and translation will be applied to the template coordinates and it will be saved in a new PDB file. When the rotational calculations are finished, statistics of the scores for all selected points (minimum, maximum, mean and standard deviation) are calculated and listed; these can be used to decide on an appropriate contour level at the graphics system. Finally, the output map (containing the best scores for each grid point considered) is saved in *CCP*4 format. A useful side-effect of using a Cα atom as the pivot point is that if the template is a small penta-peptide (in α-helical or β-strand conformation) with the central Cα atom as the pivot point, the resulting map, when contoured on the graphics, will reveal helix-shaped density for α-helices, and up-and-down strand-type density for β-strands.

In practice we find that the raw map produced by *ESSENS* can be made less noisy and more aesthetically pleasing by applying an extra filter step, in which each map point is replaced by the average of the five highest values of its 27 immediate neighbours (including itself). Although we have implemented many other digital image filters (Niblack, 1986) in *MAPMAN* (Kleywegt & Jones, 1996), this simple (perhaps counterintuitive) one gave the best qualitative results for our present purposes.

4. Applications

4.1. *Feature enhancement, P2 myelin*

P2 myelin is a member of a family of fatty-acid-binding proteins. Its structure was solved using MIR techniques (Jones, Bergfors, Sedzik & Unge, 1988), but without the use of molecular averaging, even though there are three molecules per asymmetric unit. Both the original unaveraged 2.7 Å MIR map, and a map calculated after ten cycles of threefold NCS averaging were used for test calculations during the development of the program (grid spacing ~0.9 Å). Table 1 shows some statistics of the helix and strand detection calculations with *ESSENS*. Figs. 1 and 2 show the helix and strand maps together with a cartoon of the structure of the protein. A contour level of ~3σ above the mean signal was found to be appropriate for the filtered maps. As Fig. 2 shows, the results obtained with the original map are almost as good as those obtained with the averaged map.

P2 myelin has a compact structure of ten up-and-down β-strands arranged to form a pair of orthogonal sheets (Jones *et al.*, 1988; Cowan, Newcomer & Jones, 1993). The first strand is kinked allowing it to be involved in both sheets. Adjacent strands are linked by short reverse turns except strands 1 and 2 which are linked by an antiparallel helix–turn–helix motif. The hydrogen-bonding pattern between strands 4 and 5 is broken. In the threefold averaged map, *ESSENS* locates all strands and helices. The shortest and somewhat irregular strands (1 and 5) show smaller features in the scoring map.

The results obtained with the unaveraged map, around the subunit used to make the original trace, are almost as good. This map has been much used for teaching purposes in crystallographic courses, which means that the regions that students find difficult to manage are known. Usually, helix 2 is not recognised as a helix by the students, but it is by *ESSENS*. Similarly, errors are frequently made in strand 2 because of strong side-chain density pointing towards the protein interior, as well as small breaks in the main-chain density. This region is much clearer in the *ESSENS* map. The features in the map give a strong signal for the position of the main chain. Confusion arising from interacting side chains and an unexpected fatty acid ligand are either eliminated or much reduced.

4.2. *Feature detection and enhancement, glyoxalase I*

The glyoxalase system plays a role in scavenging methylglyoxal and other toxic 2-oxoaldehydes. The structure of glyoxalase I was solved recently by MIR methods and using phase improvement by fourfold averaging at 2.7 Å resolution (Cameron, Olin, Ridderström, Mannervik & Jones, 1997). Various *ESSENS* calculations were carried out during the process of map interpretation. With the averaged MIR map, we initially attempted to find a fit for the canonical glutathione-binding domain (GBD) (Sinning *et al.*, 1993) using *ESSENS* to find the optimal rotation and position. However, even the best solutions were very weak, and they positioned the GBD near the solvent boundary. The experimental map was then used to find helices and strands (grid spacing ≃0.7 Å). Statistics of the calculations are shown in Table 1. Fig. 3 shows the resulting maps. *ESSENS* located four of the five α-helices, missing the C-terminal one. Of the eight β-strands in the structure, *ESSENS* located all but one (which is short and distorted). The lack of success in finding a GBD was easy to explain: glyoxalase I does not contain such a domain. For the glyoxalase I calculations, only map points with positive density were evaluated. Not only did this reduce the CPU-time requirements of the calculations by a factor of ~2, the resulting maps were also less noisy than those calculated for P2 myelin. A contour level of ~2.5σ above the mean signal was found to be appropriate for the filtered map.

Table 1. *Statistics of the helix and strand detection calculations for P2 myelin (averaged map) and for glyoxalase I*

In all calculations an appropriate mask file used. Scores were calculated using the *K*-minimum sum function (with $K = 10$). Full rotational calculations were carried out with 10° steps for all three Euler angles. All calculations were carried out on a DEC Alphastation 600 with a 275 MHz EV5 processor, operating essentially in single-user mode.

	P2 myelin		Glyoxalase I	
Density cut-off	None		0.0	
No. of mask points	88991		170877	
No. of map points evaluated	88991		75789	
	Helix	Strand	Helix	Strand
CPU time (s)	49300	49500	44600	44200
SSE's present in monomer	2	10	5	8
SSE's located by *ESSENS*	2	10	4	7

4.3. *Feature detection, acetylcholinesterase–fasciculin II complex*

The crystal structure of the complex between *Torpedo californica* acetylcholinesterase (AChE) and green mamba fasciculin II was recently reported (Harel, Kleywegt, Ravelli, Silman & Sussman, 1995). The structure was solved at 3.0 Å resolution by molecular replacement techniques, using the structure of AChE as the search model. In a map calculated with AChE phases alone, the structure of fasciculin I (Le Du, Marchot, Bougis & Fontecilla-Camps, 1992) could be positioned and rebuilt. However, the refinement of the complex stalled at a fairly high free *R* value (Brünger, 1992), and the toxin maintained several regions with poor electron density. Hence, the question arose if the toxin model had been positioned correctly in the initial map. At that stage, the crystal structure of fasciculin II was available (Le Du, Housset, Marchot, Bougis, Navaza & Fontecilla-Camps, 1996), and this was used in *ESSENS* calculations. A search model was created comprising 30 (of 61) residues with all side-chain atoms beyond Cβ removed. This model (containing 148 atoms; the Cα atom of Gly38 was the pivot point) was then used to obtain the optimal position and orientation in the 3.0 Å map calculated with AChE phases alone (grid spacing ~1.0 Å). All map points were included (since the pivot atom might accidentally have poor or no density at its correct position) and the *K*-minimum sum function was used (with $K = 100$). In the initial calculation, a 10° step size was used for all angles; the entire calculation took ~10.5 CPU h on a DEC Alphastation 600. The best solution had a signal of 9.0σ above the mean (second best solution, 4.1σ above the mean). The orientation was then fine-tuned by carrying out limited searches around the optimum orientation using smaller angular step sizes (5, 2 and 1°, respectively). These fine-tuning calculations require only of the order of 15–30 CPU min to complete. The final solution differs by only 0.5° from the orientation of the final refined model.

5. Discussion

The results obtained with the original experimental maps of both P2 myelin and glyoxalase I demonstrate the power and usefulness of the feature-enhancement technique. In both cases, the α-helices and large parts of the β-strands can be identified prior to more detailed map interpretation. Using this information obviously greatly simplifies the tracing problem and it may reduce the risk of tracing errors. Moreover, once secondary-structure elements have been identified, even if their direction and connectivity is as yet unknown, they can be used in database searches to look for protein structures which show structural similarities (Kleywegt & Jones, 1994), for instance using *DEJAVU* (Kleywegt & Jones, 1997). If such proteins can be found (or even a single domain), they can be positioned into the density as a rigid body, either manually or, in case of ambiguity, using *ESSENS* in feature-detection mode.

The use of α-helix and β-strand templates can potentially also be used to judge if an experimentally phased protein map is interpretable or not. If use of both templates fails to reveal readily interpretable secondary-structure features, it may be better to start searching for additional heavy-atom derivatives rather than to attempt to interpret the present map (unless the protein at hand does not contain any helices or strands, *e.g.* a kringle domain).

Sometimes one expects or suspects a protein to have a certain fold, or to contain a certain well defined structural domain, ligand or co-factor. In such cases, *ESSENS* can be used in feature-detection mode to try and find the best fit of such a structural entity in a map. This might possibly even work in cases where the MIR phasing is poor and it is doubtful if the map is interpretable. If a large enough partial structure can be positioned, this might then jump start the phase-improvement and extension process by allowing the use of phase combination.

Another potential application of feature detection is in the analysis of very large complexes such as ribosomes, liposomes and cellulosomes. It is to be expected that smaller components (proteins, protein fragments, tRNA *etc.*) of such systems are more amenable to structural analysis. Therefore, by the time crystallographic information is available for a large complex, the structure of several of its components may well be known, and their models can then be positioned into maps of the complex. Reciprocal-space molecular replacement techniques often fail in such cases, since the search models contain too small a fraction of the scattering matter for the signal to be distinguishable from the noise. Real-space search techniques while requiring the availability of external phase information, do not suffer this drawback because they operate locally and employ the phase information. In effect, all applications discussed here can be viewed as 'phased molecular replacement' calculations. Clearly, one would hesitate to carry out reciprocal-space molecular replacement calculations with search fragments as small as penta-alanine. Naturally, the real-space calculations consume a fair amount of computer time, but they need to be carried out only once during the structure determination process, and they have the potential to actually save the crystallographer a lot of time at the map interpretation stage.

Finally, the method can be used in any search where a set of coordinates must be matched to a three-dimensional function. For example, different structures can be compared by generating a pseudo-density for one of them and using *ESSENS* to fit the other molecule to that density. For macromolecules, we have done this by generating a mask around the Cα atoms of one molecule and then fitting the Cα trace of the other molecule into this map (data not shown). In this fashion, even molecules with very low topological similarity can be superimposed. For small molecules, we have used a different approach (results not shown) in which a map is calculated for one molecule which at every grid point contains the distance of that grid point to the nearest atom (multiplied by −1, since *ESSENS* maximizes its score). *ESSENS* is then used again to fit the other molecule into this map. Both methods optimize the fit of two molecules (implicitly using information about the shapes of the molecules), but whereas the first one (using a map which is essentially a mask) will maximize the number of matched atoms, the second one (using a distance map) will minimize the root-mean-square distance between the atoms in both molecules. In neither case does one need to specify any correspondence between certain atoms or atom types, nor is the sequentiality of protein residues used as a constraint.

6. Software

The program *ESSENS* is available [as part of the *RAVE* package (Kleywegt & Jones, 1994)], free of charge to academic users. For further information, contact GJK (e-mail: gerard@xray.bmc.uu.se).

This work was supported by the Swedish Natural Science Research Council and Uppsala University. We would like to thank Dr Alex Cameron (Uppsala) for providing us with the glyoxalase I maps and models, and Drs Michal Harel, Israel Silman and Joel Sussman

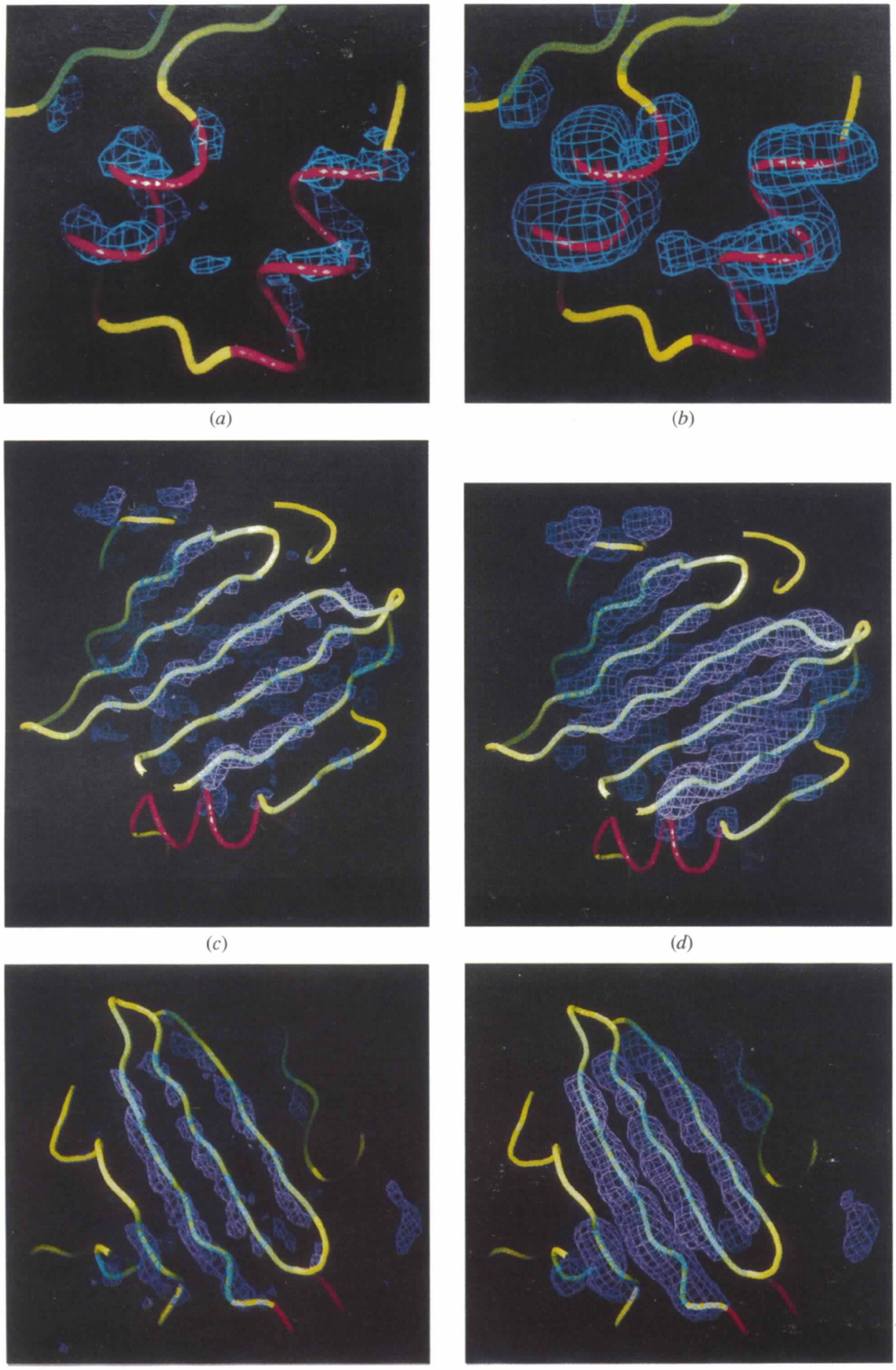
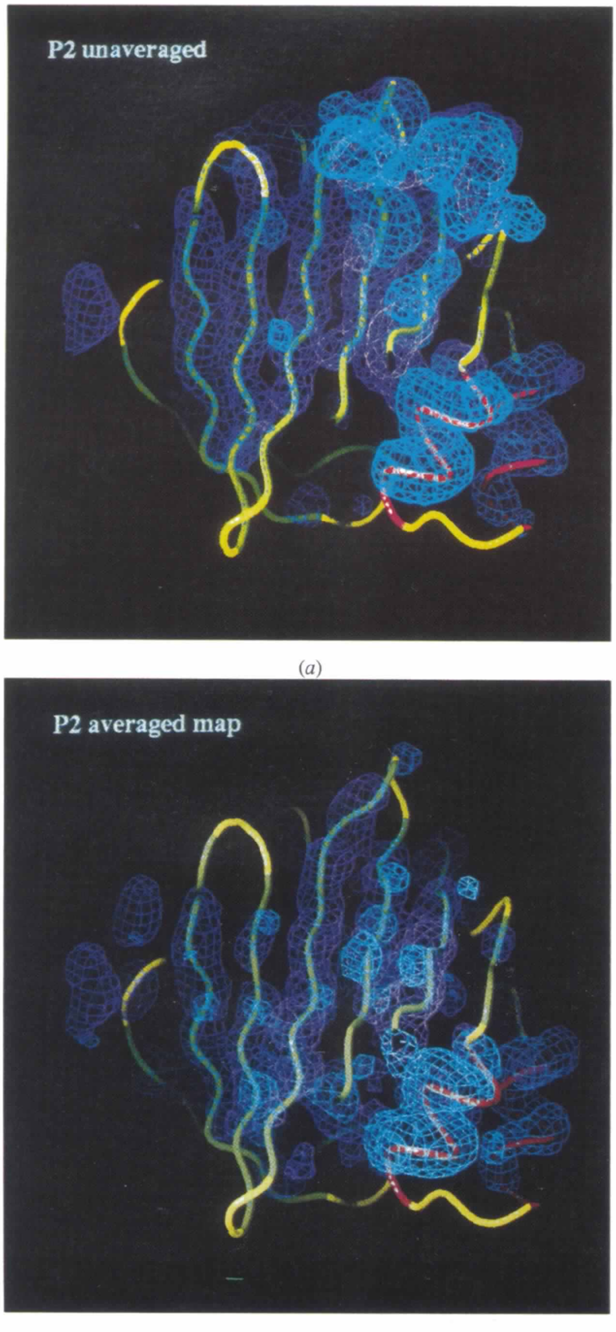


Fig. 1. Results of the template convolution calculations with *ESSENS* using the averaged MIR map of P2 myelin (Jones *et al.*, 1988; Cowan, Newcomer & Jones, 1993). The Cα trace of the final model of P2 myelin is shown as a rattler, with helices coloured red, strands green, and the remaining residues yellow. Refer to the text for more details. (*a*) The scoring map obtained with an α-helical penta-alanine fragment as the template, and (*b*) the same map after filtering. (*c*) The scoring map obtained for one of the two β-sheets with a penta-alanine fragment in a β-strand conformation as the template, and (*d*) the same map after filtering. (*e*) The β-strand scoring map for the other β-sheet, and (*f*) the same map after filtering. All figures in this paper were prepared with *O* (Jones, Zou, Cowan & Kjeldgaard, 1991).

(Weizmann Institute of Science, Rehovot, Israel) for the collaboration on the acetylcholinesterase–fasciculin complex.



(b)

Fig. 2. Results of the template convolution calculations with *ESSENS* using the unaveraged (*a*) and the averaged MIR map (*b*) of P2 myelin (Jones *et al.*, 1988; Cowan, Newcomer & Jones, 1993). The filtered α-helical scoring maps are shown in cyan, and the filtered β-strand scoring maps in magenta. The colouring of the Cα trace of P2 myelin is the same as in Fig. 1. The extra features in the unaveraged map represent signal from symmetry-related molecules.

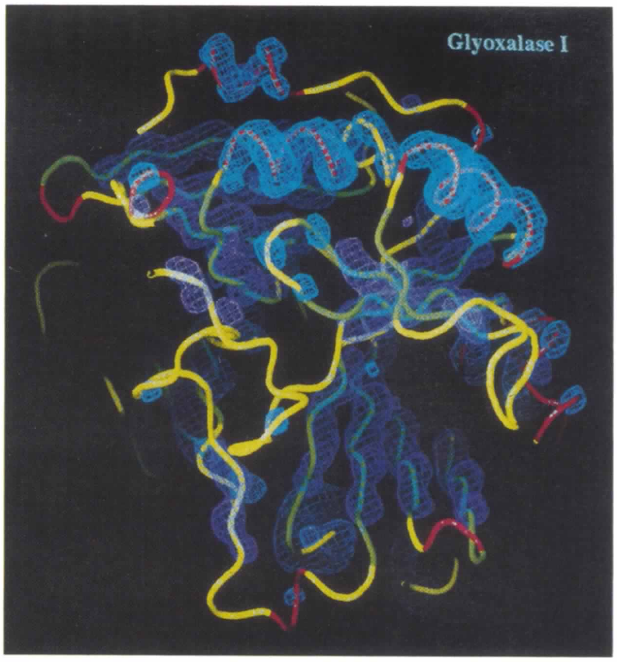


Fig. 3. Results of the template convolution calculations with *ESSENS* using the averaged MIR map of glyoxalase I (Cameron, Olin, Ridderström, Mannervik & Jones, 1997). The Cα trace of the final glyoxalase model is shown as a rattler, with helices coloured red, strands green, and the remaining residues yellow. The filtered α-helical scoring map is shown in cyan, and the filtered β-strand scoring map in magenta.

References

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.

Buerger, M. J. (1950). *Acta Cryst.* **3**, 87–97.

Buerger, M. J. (1951). *Acta Cryst.* **4**, 531–544.

Buerger, M. J. (1959). *Vector Space and Its Applications in Crystal-Structure Investigation*, ch. 10, pp. 218–251. New York: John Wiley & Sons.

Cameron, A., Olin, B., Ridderström, M., Mannervik, B. & Jones, T. A. (1997). In preparation.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Cowan, S. W., Newcomer, M. E. & Jones, T. A. (1993). *J. Mol. Biol.* **230**, 1225–1246.

Crowther, R. A. & Blow, D. M. (1967). *Acta Cryst.* **23**, 544–548.

Greer, J. (1974). *J. Mol. Biol.* **82**, 279–301.

Greer, J. (1985). *Methods Enzymol.* **115**, 206–224.

Harel, M., Kleywegt, G. J., Ravelli, R. B. G., Silman, I., & Sussman, J. L. (1995). *Structure*, **3**, 1355–1366.

Jones, T. A. (1992). In *Molecular Replacement*, edited by E. J. Dodson, S. Glover & W. Wolf, pp. 91–105. Warrington: Daresbury Laboratory.

Jones, T. A., Bergfors, T., Sedzik, J. & Unge, T. (1988). *EMBO J.* **7**, 1597–1604.

Jones, T. A. & Kjeldgaard, M. (1994). *From First Map to Final Model*, edited by S. Bailey, R. Hubbard, & D. A. Waller, pp. 1–13. Warrington: Daresbury Laboratory.

Jones, T. A. & Kjeldgaard, M. (1997). *Methods Enzymol.* In the press.

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**, 110–119.

Kleywegt, G. J. & Jones, T. A. (1994). *From First Map to Final Model*, edited by S. Bailey, R. Hubbard, & D. A. Waller, pp. 59–66. Warrington: Daresbury Laboratory.

Kleywegt, G. J. & Jones, T. A. (1996). *Acta Cryst.* D**52**, 826–828.

Kleywegt, G. J. & Jones, T. A. (1997). *Methods Enzymol.* In the press.

Le Du, M. H., Housset, D., Marchot, P., Bougis, P. E., Navaza, J., & Fontecilla-Camps, J. C. (1996). *Acta Cryst.* D**52**, 87–92.

Le Du, M. H., Marchot, P., Bougis, P. E., & Fontecilla-Camps, J. C. (1992). *J. Biol. Chem.* **267**, 22122–22130.

Niblack, W. (1986). *An Introduction to Digital Image Processing*. London: Prentice-Hall.

Nordman, C. E. & Schilling, J. W. (1969). *Crystallographic Computing*, edited by F. R. Ahmed, pp. 110–114. Copenhagen: Munksgaard.

Rossmann, M. G. (1990). *Acta Cryst.* A**46**, 73–82.

Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.

Sinning, I., Kleywegt, G. J., Cowan, S. W., Reinemer, P., Dirr, H. W., Huber, R., Gilliland, G. L., Armstrong, R. N., Ji, X., Board, P. G., Olin, B., Mannervik, B., & Jones, T. A. (1993). *J. Mol. Biol.* **232**, 192–212.