BUTTERWORTH HEINEMANN

EXTRACT: A program to extract three-dimensional coordinates from stereo diagrams of proteins

T.J. Oldfield and R.E. Hubbard

Department of Chemistry, University of York, Heslington, York, U.K.

The program EXTRACT has been developed to extract accurate three-dimensional coordinates from published stereo α-carbon diagrams of protein structures. The approach is based on the display of scanned images of the left and right eye views of the diagram on a stereo-equipped workstation, allowing construction of a molecular model using the diagram as a guide. A number of structural checks assess the building, including probability maps derived for α-carbon geometry in protein structures. The procedure has also been extended to produce less accurate models from mono images.

INTRODUCTION

With the rapid development of modern methods of macromolecular structure determination, there is an increasing number of research articles describing the three-dimensional structure of proteins. One of the frustrations in reading these articles is that often the coordinates on which diagrams are based are unavailable for local study, comparison, and modeling. The Brookhaven Data Bank[1] acts as a central repository for protein structure data, and coordinates for most protein structures eventually appear in the database. However, this can take some time. In most cases, the coordinates are caught up in the process of being deposited, validated, and published in the database. Occasionally, the coordinates are unavailable for commercial reasons or because the research group wishes to exploit the structure further. Whatever the reason, the result is that structural information reported in a research paper is not available immediately for further analysis.

These problems were recognized many years ago by Michael Rossman, who produced a computer program for generating a three-dimensional structure from the information contained in C_α stereo diagrams (the program STEREO is distributed as an unsupported program alongside the Protein Data Bank from Brookhaven[1]). This involved measuring or digitizing the two-dimensional coordinates of a left and right eye view image (where the stereo angle is refined), and from these coordinates refining a consistent three-dimensional model of the protein molecule.

In this article we present a novel technique for generating three-dimensional coordinates from both stereo and mono diagrams. This technique centers around the viewing of a digitized image of the diagram on a computer graphics workstation. Underneath this image, a computer-generated molecular model of the structure can be built and manipulated. The model can be assessed against a variety of structural principles encoded within the program. The main advantage of the procedure is that it is reliable, robust, and rapid and gives an accurate representation of the coordinates from the stereo diagram. This is demonstrated by successfully using the coordinates as a model in X-ray crystallographic molecular replacement calculations.

METHODS

Outline of the EXTRACT procedure for stereo images

Figure 1 is a flow chart showing the basic steps of the procedure. The left and right eye view parts of a published stereo diagram are digitized and transferred to a workstation. The images are scaled and displayed on the appropriate parts of the screen. When the workstation is switched to stereo mode, the published diagram then appears on the screen in stereo. A three-dimensional model is then constructed to coincide with the stereo image, using a C_α backbone. After initial scaling and positioning of one end of the C_α backbone, additional C_αs are added, assuming that the distance between successive C_αs in a sequence is 3.8 Å. Once the complete chain is built, a manual refinement phase is entered, in which each C_α position and the appropriate pieces of the diagram are shown on a larger scale. The user

Color plates for this article are on p. 52.

Address reprint requests to Dr. Hubbard, Department of Chemistry, University of York, Heslington, York Y01 5DD U.K.
Received 10 May 1994; revised 21 June 1994; accepted 28 June 1994.

655 Avenue of the Americas, New York, NY 10010

0263-7855/95/$10.00
SSDI 0263-7855(94)00006-E

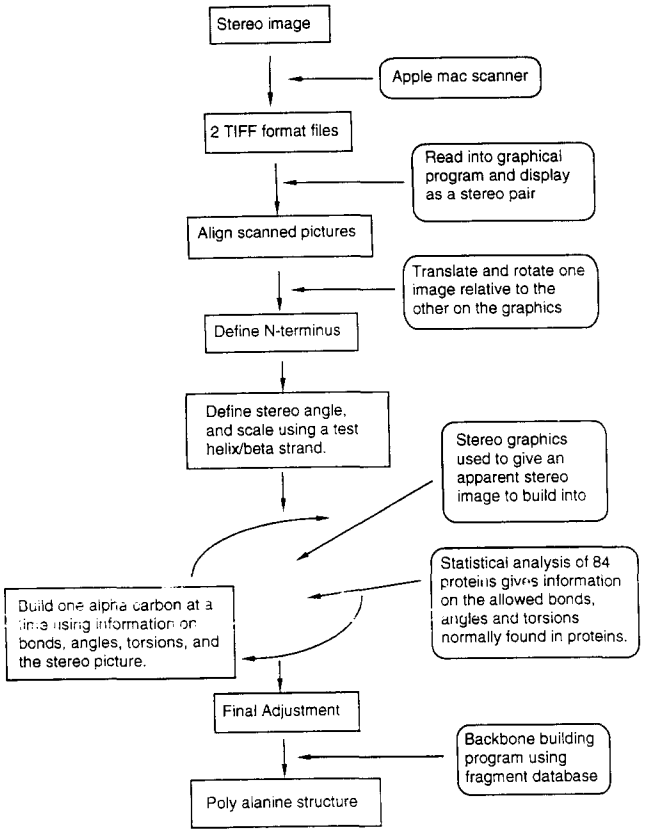

Figure 1. Schematic overview of the procedure for extracting three-dimensional coordinates from stereo images.

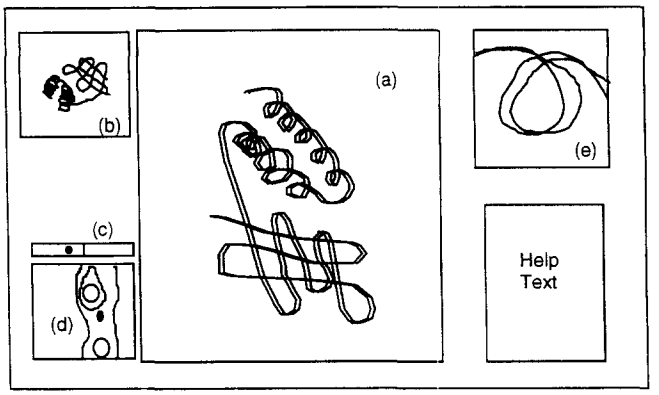

Figure 2. Schematic of the representative screen of the EXTRACT program while fitting a stereo diagram. The various parts of the screen are labeled (a)–(e). (a) The scanned image is drawn in the overlay planes in red, and the growing C_α trace built in green. (b) A transformable view of the currently built 3D coordinates, allowing a visible assessment of the model. (c) The bar and pointer marks the relative z position of the current C_α atom relative to the last C_α atom. The center line indicates the same z coordinate, and the extreme values indicate 3.8 Å. (d) A contour plot of the distribution of C_α geometry (see Figure 4 for more details). The pointer on the plot indicates the current conformation of the C_α being placed. (e) An enlarged view of the last C_α atom and a 5-Å radius.

can then adjust the position of the C_α atom, this time with the length of the C_α–C_α bond being allowed to vary as well. A polyalanine chain is then generated from the C_α backbone from a database of protein fragments, and side chains (if required) added using the program QUANTA (Molecular Simulations, Inc, Burlington, MA.).

The details of the EXTRACT procedure are described below, with reference to a number of figures. Color Plate 1 shows the screen while the program is operating. Figure 2 is a schematic representation of this screen with the various components labeled.

1. Display of digitized images. The first stage in the procedure is to digitize the parts of the figure corresponding to the left and right eye views of the molecular system with an image scanner. The EXTRACT program produces the best results from images that are more than 10 cm in length and width, but any size of figure can be used as the images can be scaled to be displayed at any size. The images are displayed on the workstation screen as illustrated in Figure 2. The left eye image is at the top of the screen and the right eye image at the bottom. The images are of single-bit resolution and are placed in write protected display memory (or overlay planes) as they are continuously displayed.

The user aligns and scales the images to each other by identifying two equivalent positions in the left and right eye images as shown in Figure 3. Typically, the N terminus and an atom as far as possible from the N terminus in the horizontal screen axis are used. The first position identified is aligned exactly; the second is aligned in the vertical screen axis only. This alignment is important as the quality of the apparent stereo image is degraded if there is an alignment error.

Once the left and right eye images are aligned, they can be viewed in stereo. Unless explicitly stated, the remainder of the procedures are performed while viewing the screen in stereo.

2. Viewing a stereo image on a workstation. The most novel aspect of this technique for extracting coordinates exploits a particular feature of the hardware used to produce

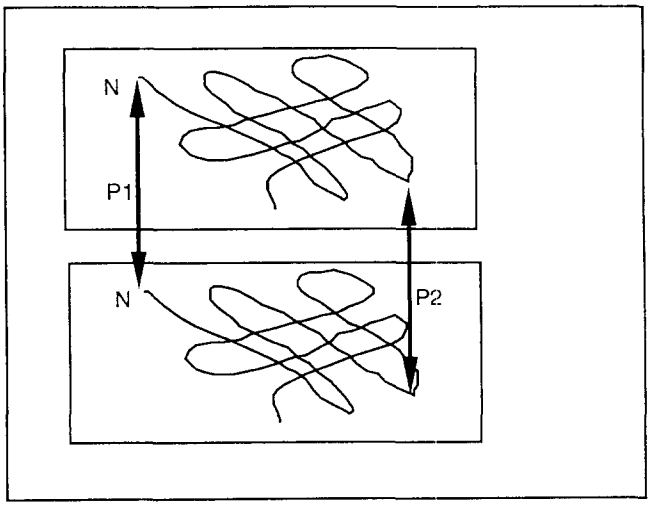

Figure 3. Schematic diagram to show the alignment of the left and right eye images. The user picks equivalent positions in both images. Typically this is the N terminus and an atom as far away as possible horizontally.

dynamic stereo in modern graphics workstations. If the left eye view of a molecule is drawn in the top half of the screen and the right hand view in the bottom half, and the monitor is then run at double standard frequency, then the individual images are lengthened in the y direction and the left and right eye images each appear at the normal frequency. The appearance of the two eye images can be synchronized with a pair of liquid crystal glasses so that when the left eye view is on the screen, the user's left eye can see the screen, and when the right eye view is on the screen, the right eye can. This produces very effective dynamic stereo perception, with the only disadvantage being that the resolution of the image is reduced in the y direction.

3. Refining the alignment and scaling. Before building of the C_α backbone can begin, it is necessary to define the scale of the built coordinates against the displayed images, and to provide a reasonable approximation to the stereo angle. Although the two parameters are in theory independent of each other, in practise they are related, particularly when they are close to the correct values.

Defining the correct scale and stereo angle is one of the most arduous aspects of the procedure. Tools are available to change the scale and stereo angle manually, using a template piece of β sheet or α helix with idealized geometry. These can be moved and rotated on the screen and the scale and stereo angle adjusted so that the piece of secondary structure appears to fit well with an appropriate section of the scanned image.

If an appropriate scale and stereo angle are not established at this stage, then subsequent building of an C_α trace will be difficult. This is usually apparent after about five or six C_α positions have been built and is seen in the inability to build the next C_α position satisfactorily within the geometric restraints imposed by the program. If this occurs, then it is necessary to return to this step to modify the scale and stereo angle.

4. Building a C_α backbone. The first stage is to define the N terminus for the protein in the stereo image from which building starts. The user then goes on to add consecutive C_α atoms, one at a time. As each atom is manipulated into the correct position and accepted, the program introduces a new C_α atom. Each C_α atom is represented by a cross, connected to the previous C_α by an C_α–C_α bond of length 3.8 Å. The position of the C_α relative to the growing backbone chain is altered to coincide with the stereo image by manipulation of two variables, the C_α–C_α–C_α angle and C_α–C_α–C_α–C_α pseudotorsion angle (C_α geometry is discussed in Section 8 below). Color plate 1 shows a typical session after fitting the first 99 C_α positions of a stereo figure of myoglobin; the various components of the system are summarized in the schematic of Figure 2.

5. Graphical aids to building. Four additional visual cues are presented to help resolve ambiguities and guide the model building. These can be seen in Figure 2. First, the piece of the diagram currently being fitted is shown much enlarged in a corner of the screen so that small adjustments in the position of the new C_α atom can be made. Second, the elevation of the new C_α (above or below the z position of the previous C_α) is represented by the color of the atom (yellow for above, blue for below). Third, a rotatable view of the growing peptide backbone is displayed to allow the user to assess how the z coordinate is being interpreted. Finally, a contour plot of the allowed regions for the C_α geometry (see Section 8) is shown together with the currently assigned values. This prompts the user when an unusual conformation is being generated.

6. Final adjustments of C_α backbone. Once all the backbone has been built, it is possible to make small adjustments to the position of the C_α atoms with no constraints on angles or bond distances. As atoms are moved, changes in geometry and atom–atom close contacts are displayed as a guide.

7. Building a polyalanine chain. The result of the building process is the generation of an C_α main chain trace for the extracted protein. This is converted into a polyalanine backbone chain, using the program BACKBONE (written by T.J. Oldfield). The program reads in the C_α atom chain trace generated by EXTRACT and, for each segment of five C_α atoms, searches a database for a five-residue fragment of structure that best fits the segment. The database contains inter-C_α distances for the protein backbone of 74 high-resolution (2.0 Å or better) structures solved by restrained least-squares refinement. The program finds the best match between the C_α distances for five atoms from the C_α trace and the C_α atoms from a five-residue segment in the database. This is done for all possible five-residue segments in the C_α chain trace, that is, for residues 1 to 5, 2 to 6, etc. The result is a list of overlapping five-residue segments for each C_α position, which gives five possible positions for each C_α atom, and the corresponding coordinates for a set of five backbone atoms. In this way, five possible alanine positions and conformations are generated for each C_α atom in the C_α trace (with the exception of the first and last four C_α atoms in the trace). The program then selects the best alanine residue for each position on the basis of C_α root-mean-square (RMS) from the set of overlapping five-residue segments and generates a complete polyalanine chain. This building procedure was checked by constructing a polyalanine model from the C_α coordinates of a known protein structure (mucolipase, containing 256 residues[2]) that is not one of the proteins in the fragment database. The resulting built polyalanine structure had an RMS with the original structure of between 0.5 to 1.0 for each residue for the main chain and C_β atoms.

The error in the fit of the alanine residue to the original C_α trace is recorded for each position, indicating the sections of the protein that are likely to have larger errors. It is common at this stage to find, particularly in the loop regions, that the positions of the alanine residues do not make a continuous polypeptide chain. The structure was minimized with the program CHARMM[3] with the position of the C_α atom constrained and constraints set to ensure a *trans* conformation for peptide bonds. The result of minimization forms the final set of polyalanine coordinates.

8. Analyses of C_α protein geometry. The EXTRACT program uses three pieces of geometric information to aid in the building process. These are the pseudobond length, bond angle, and torsion angle between consecutive C_α atoms. The geometric data are based on empirical values determined from a database of 84 protein structures (see legend to Figure 4) using a statistical database program

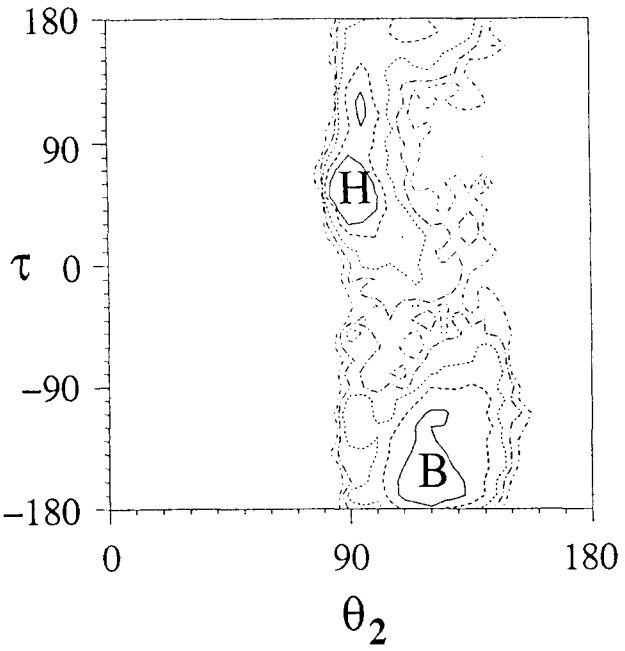

Figure 4. Probability density plot of the distribution of C_α geometry in a database of 83 high-resolution protein structures. For each successive C_α atom in a protein chain as C_α (i − 3), C_α(i − 2), C_α(i − 1), C_α (i) it is possible to compute a torsion angle τ from the positions of atoms 1–2–3–4, and an angle θ_2 from the positions of atoms 2–3–4. For 18 503 separate values in the database, the plot is contoured at 100, 50, 25, 12, and 6 values for a distribution binned in 10° increments of τ and 5° increments of θ_2.

(written by T.J. Oldfield). These pseudogeometric parameters are well defined within limits and enable the user to check that the chain of residues follows a pathway normally found in proteins.

An analysis of the C_α pseudobond length in the database of proteins shows that this is remarkably constrained to 3.81 Å with a standard deviation of 0.046. The empirically derived probability map of bond angle and torsion angle between consecutive C_α atoms was generated from the database of 84 proteins. This plot can be seen in the bottom left corner of the screen shown in Figure 2, and is shown in more detail in Figure 4. This representation of C_α geometry contains a wealth of information on the conformational preferences in proteins, which is discussed in detail elsewhere.[7] For the purpose of this description of the EXTRACT building process, it is only necessary to recognize that this plot gives a good indication of the likely and unlikely conformations available to a C_α trace. As can be seen in Figure 2, the user can follow the change of angle and torsion as a pointer moves within the map as the current atom is moved relative to the rest of the built chain. It is generally found that when building an atom into the stereo image, the atom being built can have only two possible positions. The positions have the same *x* and *y* coordinates, but differ in the *z* coordinate. The two possible *z* coordinates are given by the *z* coordinate of the previously built atom, plus or minus a distance that is less than or equal to an C_α–C_α distance. It can usually be seen that one of these positions produces an angle and torsion angle not normally found in proteins, whereas the alternative *z* coordinate produces an acceptable conformation. Hence most atoms can be build into the stereo image with no ambiguity.

Table 1. Results of the RMSD between coordinates generated using the EXTRACT program, and the original coordinates for the myoglobin or interleukin 1 β molecule used to generate the stereo figure[a]

Protein	Match atoms	Number of atoms	RMSD (Å)
Myoglobin	C_α	153	0.72
Myoglobin	Ala	750	1.01
Inerleukin	C_α	151	0.64
Interleukin	Ala	474	0.848

[a]Comparisons were made using either the C_α atoms as output from the program, or all the coordinates for the alanine polypeptide generated by the backbone-building program. Note that the number of atoms for the alanine polypeptide is not necessarily five times the number of C_α atoms because glycine residues have only four atoms.

The procedure for mono images

Often, suitable C_α stereo diagrams are not available, and a mono figure is the only information available. We have developed procedures that allow coordinates to be extracted from such figures, using pieces of structure with ideal secondary structure geometry.

The diagram is digitized and transferred to the workstation in TIFF format as described above for stereo diagrams. In this case, the image is scaled to fit the whole screen. After establishing a suitable scale, the identifiable pieces of secondary structure can be fitted using appropriate templates, and these pieces of structure manipulated against each other with bump distance monitors. These monitors, together with the transformable graphics representation of the growing model, can be used in combination to assess how well the *z* coordinate is being interpreted. Further refinement is possible by comparison of the proposed model against other views or representations of the structure that may be available.

Color Plate 2 shows the screen during the use of the EXTRACT program to fit a cartoon representation of the myoglobin molecule. The scale of the mono picture is established using the same method as for the stereo diagrams. A template is used (either α helix or an extended strand) that can be rotated, translated, and scaled so as to give an apparent good fit to the figure. As more strand and helix templates (of adjustable length) are added to the image, it is possible to refine the scale. The bump monitor checks all the distances between the structural element being fitted and all other fitted elements. If any component of the fitted secondary structural element is closer than 6 Å to a previously fitted element then a marker is drawn between the elements. In practise, it is necessary to fit the *x* and *y* position of all the elements possible and then allow the bump monitors to guide the positioning of the elements in *z*. As with the stereo program the user has a rotatable view of the protein being built, which can be zoomed to fill the full screen if necessary. It is important that the user of the mono

Table 2. Results for the molecular replacement calculation using the program AMORE on interleukin 1β and myoglobin molecules[a]

Molecule and model	Resolution shells for which successful solution obtained				*r* factor	Corr.
Interleukin 1β,	7-3	7-4	6-3	5-2	39	60
real structure	6-4	5-3	4-2			
Interleukin 1β,	7-3	—	6-3	5-2	43	48
model structure	6-4	5-3	4-2			
Myoglobin,	6-4	5-3	4-2		50	30
real structure	7-4	6-3				
Myoglobin,	6-4	5-3	—		52	20
model structure	7-4	6-3				

[a]Shown here for both proteins (original and built coordinates) are the resolution shells for which a successful solution was obtained for the rotation function, translation function, and rigid body refinement. The calculations were carried out using C^{α}, C, N, O, and C^{β} (except for glycine, where the C^{β} was absent). The values of the *R* factor and correlation coefficient (Corr.) indicate the quality of the results obtained for all the calculations.

version of the program has a good understanding of protein structures, as the intuitive packing of secondary structure elements and bumps are the only information the user has to define the *z* coordinate. This is an iterative procedure, often requiring refinement of the position of a number of elements to produce a consistent fit.

RESULTS AND DISCUSSION

The program EXTRACT was written because of the frustration of finding articles on proteins of interest where the coordinates were not available. Attempts to digitize coordinate positions from stereo images, and then refining these stereo atom positions, produced unusable results. The ability of a researcher to view a stereo pair of images with complete clarity, even though the *z* coordinate information is too small to manually digitize accurately, led to the method employed by the program EXTRACT. The main advantage of this procedure is the precision with which coordinates can be reproduced. The use of stereochemical information, and the natural ability of the brain to perceive stereo, makes this program an extremely powerful tool. Once the scale and stereo angle have been determined the building progress is actually very simple and rapid. Several structures in excess of 500 residues have been determined by this method, indicating its ability to decode large proteins.

The effectiveness of the program to produce a reasonable set of coordinates from published diagrams was assessed with three examples. These were stereo diagrams of myoglobin[4] and interleukin 1β[5] and a mono diagram of a cartoon representation of myoglobin. These figures were produced from the known coordinates for these proteins. The diagrams were scanned and transferred to a workstation and coordinates built using the EXTRACT program. A comparison was then made with the original coordinates used to produce the diagrams.

The results for the stereo diagrams are presented in Table 1 as root-mean-square distance (RMSD) values calculated for the difference between the original and extracted C_{α} positions. It can be seen that the program is equally able to produce a reasonable set of coordinates from stereo diagrams of both predominantly helical and sheet proteins. For the mono diagram, an RMSD of 4.3 Å was obtained. This large reduction in quality of the model is not surprising as much more of the *z* coordinate component of the image had to be inferred.

As a further test of the effectiveness of the procedure for stereo diagrams, we attempted to use the coordinates to determine the structure of myoglobin and interleukin 1β in molecular replacement calculations against structure factors available for these proteins in our laboratory. The search for the solution to the rotation function was also carried out using the refined coordinates taken from the protein databank files, but with side chains truncated at the β-carbon. These control "correct" structures were used to see if a "correct" polyalanine structure of a protein contains enough information to solve the rotation function. In each case, different resolution shells of data were used to determine the best data subset for the solution.

The program AMORE[6] was used to carry out the molecular replacement calculations. For both proteins, the control calculations (using the truncated real coordinates) and the polyalanine built models, a rotation function, a translation function, and a rigid body refinement were carried out. Table 2 shows the results for the four test cases. Note that for the myoglobin example there were two molecules in the asymmetric unit, and hence two solutions.

The results indicate that the molecular replacement calculations produced a satisfactory result with either the truncated polyalanine structures, or using modeled polyalanine coordinates generated from the stereo diagrams. As expected there is some deterioration in the quality of the results owing to the inherent error in the extracted coordinates, but the quality of the models was still sufficient for the molecular replacement calculations to find the correct solution.

CONCLUSIONS

EXTRACT has been found to be successful in determining three-dimensional (3D) coordinates from published stereo diagrams of protein C_{α} traces, even for proteins over 500 residues. For smaller proteins it has provided sufficient information for the determination of a solution to a rotation function, allowing a route to crystal structure determination. Most of the interpretation of depth is carried out by the viewer, as the perception of depth by a human user appears

much more accurate than that achievable using measurement of small differences of distances between left and right eye images. The program provides several visual clues as to the correct conformation based on expected geometric parameters for a protein, and these are normally enough to overcome any ambiguities in the building process.

The program is available from the authors.

ACKNOWLEDGMENTS

This work was supported by Glaxo Research and Development. We thank in particular Dr. Peter Murray-Rust for stimulating discussions during the development of the approach.

REFERENCES

1 Bernstein, F.C., Koetzel, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, K., and Tasumi, M.J. A computer based archival file for macromolecular structures. *J. Mol. Biol.* 1977, **112,** 535–542

2 Brady, L., Brzozowski, A.M., Derewenda, Z., Dodson, E., Dodson, G.G., Tolley, S., Turkenburg, J.P., Christiansen, L., Huge-Jensen, B., Norskov, L., Thim, L., and Menge, U. A serine protease triad forms the catalytic centre of triacylglycerol lipase. *Nature (London)* 1990, **343,** 767–770

3 Brooks, R.B., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, K. CHARMM: A program for macromolecular energy minimization and dynamics calculations. *J. Comp. Chem.* 1982, **4,** 187–217

4 Smerdon, S., Oldfield, T.J., Dodson, E.J., Dodson, G.G., Hubbard, R.E., and Wilkinson, A.J. The determination of the crystal structure of recombinant pig myoglobin by molecular replacement and its refinement. *Acta Cryst.* 1990, **B45,** 327–332

5 Finzel, B.C., Clancy, L.L., Holland, D.R., Muchmore, S.W., Watenpaugh, K.D., and Einspahr H.M. Crystal structure of recombinant human interleukin 1β at 2.0 Å resolution. *J. Mol. Biol.* 1989, **209,** 779–791

6 Castellano, E.E., Oliva, G., and Navaza, J. Fast rigid body refinement. *J. Appl. Cryst.* 1992, **25,** 281–284

7 Oldfield, T.J. and Hubbard, R.E. Analysis of Ca geometry in proteins. *Proteins Struct. Funct. Gen.* 1994, **18,** 324–327