

Software for handling macromolecular envelopes

G. Jacob Kleywegt* and
Thomas A. JonesDepartment of Cell and Molecular Biology,
Uppsala University, Biomedical Centre, Box
590, SE-751 24 Uppsala, Sweden

Correspondence e-mail: gerard@xray.bmc.uu.se

Macromolecular phase-refinement and phase-extension calculations using real-space electron-density averaging techniques require accurate envelopes (or masks) to define the boundaries of each domain or molecule whose density is to be averaged. An extensive set of tools, implemented in four computer programs (*O*, *MAMA*, *COMA* and *MASKIT*) are described which can be used to generate such envelopes (either from an atomic model or based on local density-correlation maps), to improve them, to remove overlap owing to crystallographic or non-crystallographic symmetry, to display them and to manipulate them in a variety of manners.

Received 11 October 1998

Accepted 19 January 1999

1. Introduction

Real-space electron-density averaging is a conceptually simple yet very powerful technique for macromolecular phase refinement and extension (Rossmann & Blow, 1963; Bricogne, 1974, 1976; Jones, 1992; Vellieux & Read, 1997; Kleywegt & Read, 1997). It relies on the heuristic that multiple copies of a molecule [either related through non-crystallographic symmetry (NCS) or found in different crystal forms] will assume similar conformations and therefore engender similar electron density. In a typical cyclic averaging process, the density for each copy of the molecule is replaced by the averaged electron density, structure-factor amplitudes and phases are calculated from the averaged map and a new map is calculated. This map, in turn, can be averaged, and this cyclic process is usually carried out a number of times until no further improvement is achieved. Besides initial phase information and a set of Cartesian operators which relate the various copies of the molecule, another important prerequisite for electron-density averaging is the availability of a molecular envelope or mask. Contrary to envelopes for solvent-flattening purposes (Bhat & Blow, 1983; Wang, 1985; Leslie, 1987), envelopes used in averaging must generally be constructed such that copies of the envelope related by NCS and/or space-group symmetry do not overlap. In addition, such envelopes should not contain holes (cavities) and should cover as much as possible of the scattering matter of the NCS unit.

Here, we describe a set of software tools which were designed to facilitate the genera-

tion, optimization, handling, storage and display of macromolecular envelopes. Some of these tools have been implemented in the crystallographic modelling program *O* (Jones *et al.*, 1991), others in a general mask-manipulation program called *MAMA* (Kleywegt & Jones, 1994) and others in standalone utility programs. The tools in both *O* and *MAMA* have been available to the community for more than five years (Jones, 1992; Kleywegt & Jones, 1994) and are used in conjunction with many different software packages. The highly compressed envelope format used by both programs has also been adopted by other programs. In our approach (Jones, 1992), a mask is represented as a three-dimensional array of grid points which can only assume the values zero (not masked) and one (part of the mask). When storing a mask in a file, the three-dimensional array is considered as a one-dimensional array instead, and only the indices of the points where consecutive stretches of ones begin and end are written out. Except in particularly pathetic cases (*e.g.* unphysical masks in which zeros and ones alternate), this tends to achieve a ten to 50-fold compression of the output file compared with a file which contains the entire array.

2. Mask generation

In general, masks can be generated either from a set of (pseudo) atom coordinates (*e.g.* an all-atom model, a C^α -only model or skeletonized electron density) or directly from the electron density itself. In the former case, mask generation is a trivial operation, easily carried

out with *MAMA*. An empty grid is constructed which encompasses the entire model, and subsequently all grid points that lie within a user-specified radius of any atom

or pseudo-atom are added to the mask. For mask generation directly from an electron-density map, we have implemented Read's local correlation method (as described in

Vellieux *et al.*, 1995) in two standalone programs called *COMA* and *MASKIT*. *COMA* calculates correlation maps using non-crystallographic symmetry (Fig. 1), whereas *MASKIT* calculates correlation maps between two different crystals or crystal forms (Fig. 2), making it a useful tool in multiple crystal form electron-density averaging. The density-correlation map method tends to yield excellent masks, comparable with those generated from atomic models (Figs. 1 and 2). In addition to the two methods discussed above, crude masks can also be generated as spheres or parallelepipeds and by combining these in arbitrary fashion. Finally, existing masks (e.g. derived during work on a mutant, a complex or a different crystal form of a molecule) can be 'recycled' easily, since *MAMA* contains options to apply a Cartesian rotation/translation operator to a mask, and to copy a mask onto a new grid (e.g. a finer spaced grid if higher resolution data has become available).

3. Mask improvement

A mask which is used in electron-density averaging procedures should not normally contain any internal cavities (unless these are used specifically to mask out heavy-atom sites, for instance), should not contain isolated 'blobs' which are not connected to the main bulk of the mask, should cover all of the appropriate scattering matter (e.g. an entire molecule or domain), should be smooth and devoid of sharp extrusions or invaginations (*i.e.* have a similar 'resolution' as the data that will be used in the averaging procedure) and should have as little overlap as possible with other masks (either belonging to different domains or molecules, or to copies of itself related by crystallographic and/or non-crystallographic symmetry). *MAMA* contains options to improve masks with respect to all these criteria: it can fill cavities which may exist inside a mask, it can remove parts of the mask which are not connected to the bulk of the mask, it can list atoms which are not covered by the current mask, it can smooth masks (either by removing mask grid points which have many non-mask neighbours or by adding non-mask grid points which have many neighbours which are already part of the mask) and it can cull parts of the mask which overlap with (crystallographic and/or non-crystallographic) symmetry-related copies of itself. Overlap removal entails expansion of the mask onto a grid which covers an entire asymmetric unit or unit cell and keeping track of either *which* or *how*

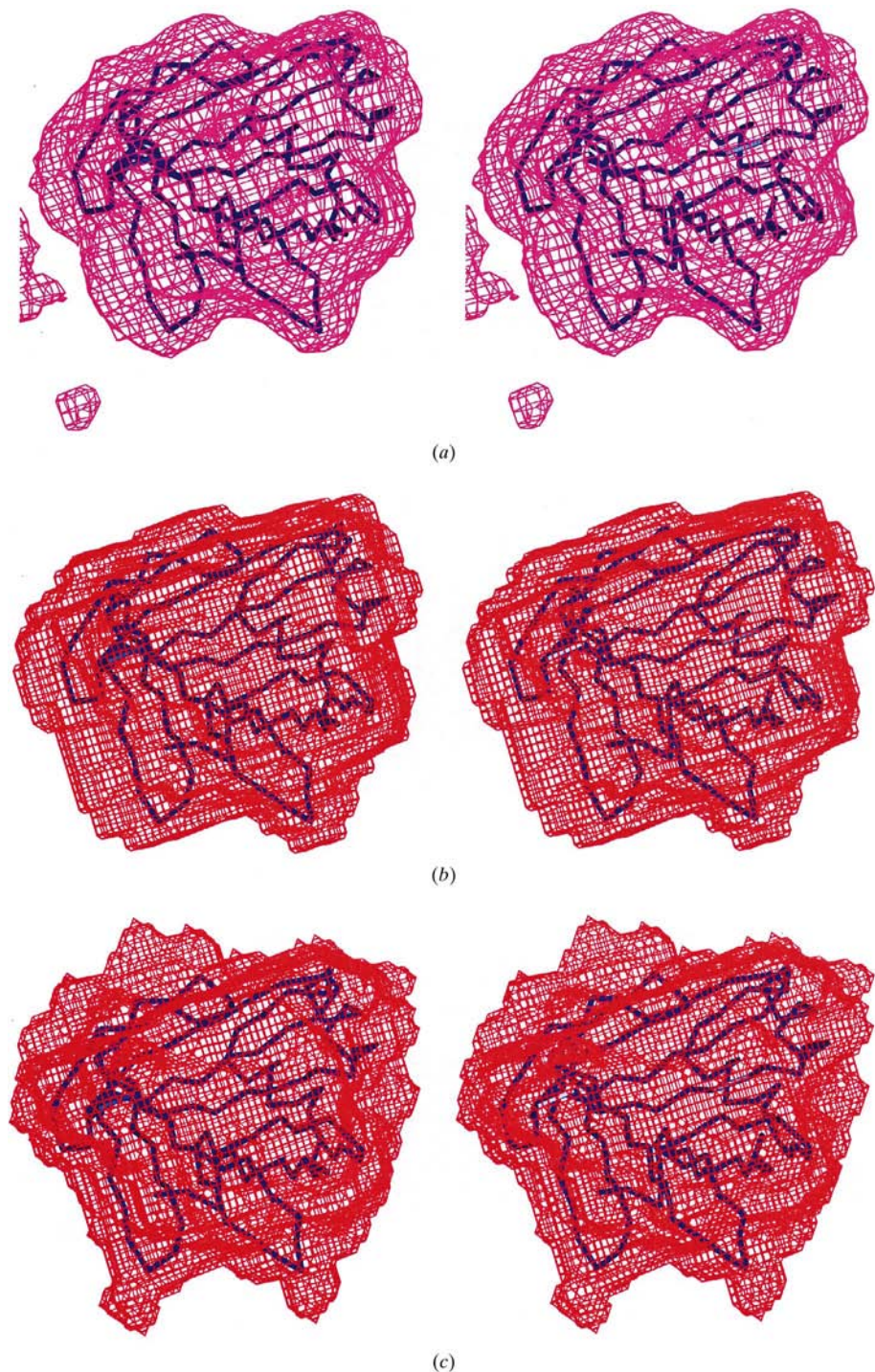


Figure 1

Use of the local density correlation method, as implemented in *COMA*, to generate a mask suitable for NCS-averaging. (a) MIR density for P2 myelin protein (Jones *et al.*, 1988) was used to calculate an R1-correlation map (Vellieux *et al.*, 1995) for the three NCS-related molecules. This map is shown here, contoured at a level of +0.50. The C α -trace of the final refined model is also shown (PDB code 1pmp). (b) A mask derived from the correlation map using *MAPMAN* (Kleywegt & Jones, 1996) and *MAMA*. (c) For comparison, a mask derived from the refined atomic model of P2 myelin protein is shown. The similarity index of the two masks (see text) is 0.89, indicating that the density-correlation method yields accurate masks. The major differences between the two masks lie in the degree of their covering some of the surface side chains and in their different resolution (level of fine detail).

many NCS operators project a mask point onto a given grid point in the asymmetric unit or unit cell. By 'averaging' back onto

the original mask, one obtains an overlap map with information regarding the degree of overlap for every mask point. Subse-

quently, all mask points which introduce overlap can be removed, but this usually removes too much of the mask. Instead, only overlapping points on the surface of the mask are removed, and a few iterations of this process usually suffice to remove most of the overlap.

4. Interactive mask editing

The crystallographic modelling program *O* (Jones *et al.*, 1991) contains facilities for interactive mask editing and display (Jones, 1992). Displaying masks on a graphics workstation makes it easy to inspect the overall quality of a mask, *e.g.* does it contain cavities? Are there isolated pieces of mask not connected to the main bulk? Are there parts of the molecule which stick out of the mask? If any such problems are found, they can be fixed rapidly with *MAMA*. In addition, a mask can be fine-tuned interactively either by adding points to it (*e.g.* to cover a loop which has newly been built or which has moved substantially during previous refinement) or by removing points (*e.g.* those which belong to a neighbouring molecule). This interactive editing can be carried out by the user, who then moves a point around in space, while *O* adds to or removes from the mask all grid points within a certain distance from this point. Once some sort of model is available (be it a skeleton or an atomic model), two simple *O* macros can be used instead which will add or remove all grid points within a certain radius of any atom that the user identifies with the mouse (Kleywegt & Jones, 1994; Fig. 3).

5. Additional tools and applications

The mask-manipulation program *MAMA* contains a large number of tools for dealing with molecular envelopes, and some of these can also be applied for other purposes. One group of commands enable output of masks (and C^α -traces of molecules) in VRML format, which is useful if one wants to publish masks on the WWW or if one does not have a copy of *O* available for visual inspection of masks. Other commands enable logical operations on (combinations of) masks, namely logical NOT, AND ($A \cap B$), OR ($A \cup B$), XOR [$A \cup B - (A \cap B)$] and 'BUTNOT' [$A - (A \cap B)$]. These, together with options to expand and contract masks, have been used successfully to implement Delaney's algorithm (Delaney, 1992) to visualize cavities and other features of protein structures. For instance, the cellulose-binding tunnel in the structure of the catalytic domain of cellobiohydrolase I

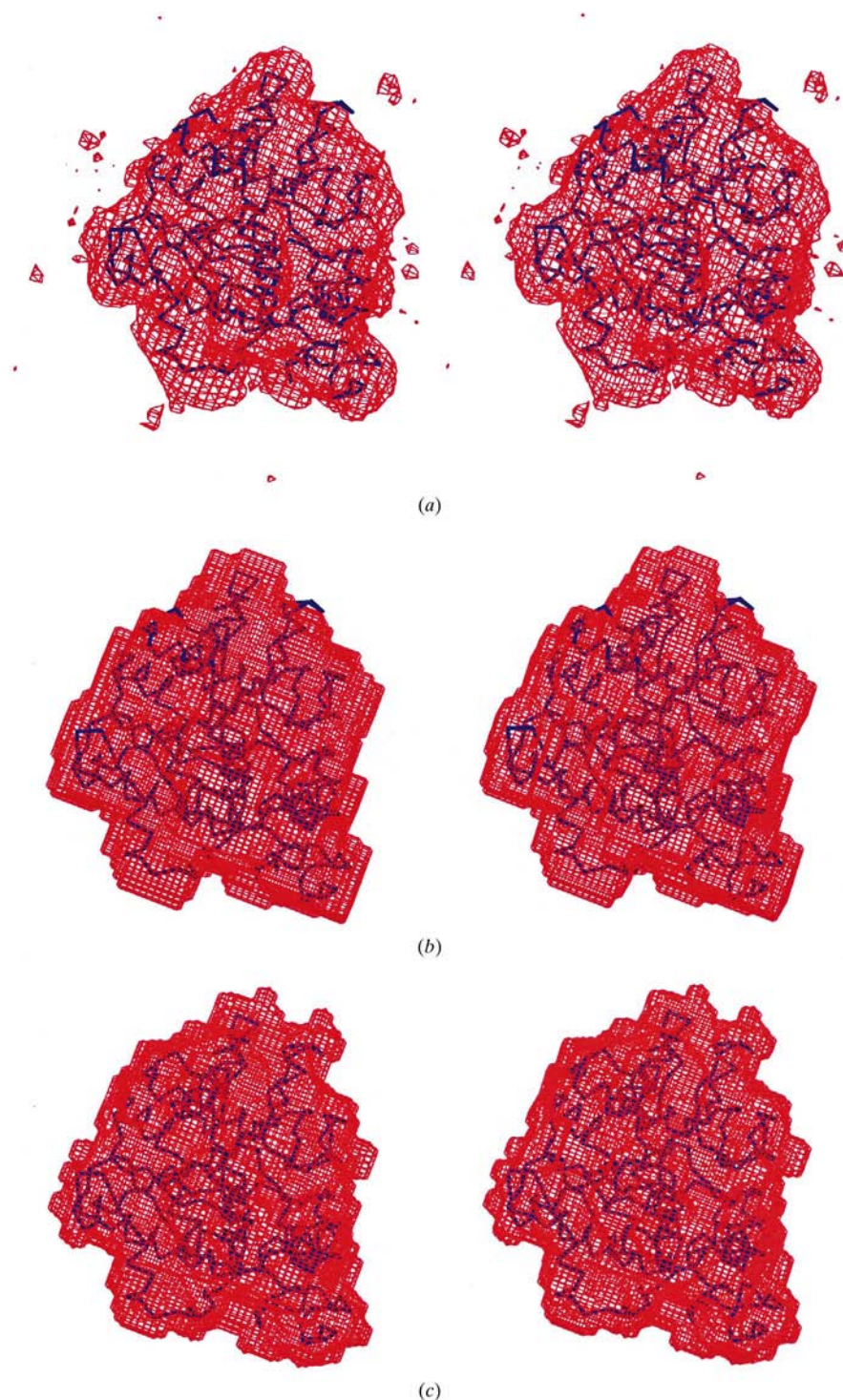
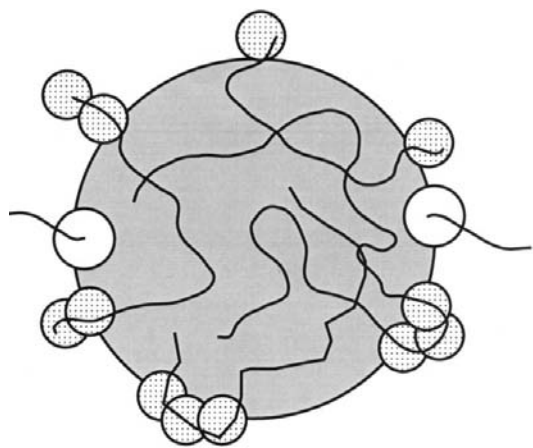


Figure 2

Use of the local density-correlation method, as implemented in *MASKIT*, to define a mask suitable for multiple-crystal form averaging. (a) MIR density of two crystal forms of *C. antarctica* lipase B (Uppenberg *et al.*, 1994, 1995) was used to calculate a local correlation map (between the orthorhombic crystal form, PDB code 1tca, and the *A* molecule of one of the monoclinic crystal forms, PDB code 1tcc). The map is contoured at a level of +0.40, and the C^α -trace of molecule *A* of the final refined model of the monoclinic crystal form is shown as well. (b) A mask derived from the local correlation map using *MAPMAN* (Kleywegt & Jones, 1996) and *MAMA*. (c) For comparison, a mask derived from the refined atomic model is shown. The similarity index of the two masks (see text) is 0.89, indicating that the density-correlation method also yields accurate masks when multiple crystal forms are involved.



```
! mask_atom.omac -> mask a small area around an atom
message 'Click on atom to MASK'
! add all points within 3.0 Å to the mask
mask_set 1 1 3.0 ; ; ;
mask_on wait_id no
mask_contour
message Done
```

```
! unmask_atom.omac -> unmask a small area around an atom
message 'Click on atom to UNMASK ...'
! remove all points within 1.0 Å from the mask
mask_set 0 1 1.0 ; ; ;
mask_on wait_id no
mask_contour
message Done
```

Figure 3

The use of mask-editing facilities in *O*. Top: while editing a rough initial mask (large circle), one may encounter parts of the molecule which are not yet covered by the mask (lightly dotted circles) or parts of the mask which cover pieces of neighbouring molecules (white circles). Bottom: two simple *O* macros which can be used to remedy such problems. The first one ('mask_atom.omac') prompts the user to identify an atom with the mouse; subsequently, all grid points within a certain radius from this atom will be added to the mask. The second macro ('unmask_atom.omac') does the opposite: it will remove all grid points within a certain radius from the mask.

was visualized in this fashion (see Fig. 3 in Divine *et al.*, 1994). *MAMA* can also be used to investigate the similarity of two molecular shapes *A* and *B*, for example by calculating their shape-similarity index (SI), which is defined as $SI = N_{A \cap B} / (N_A N_B)^{1/2}$, where N_A is the number of grid points in mask *A*, N_B is the number of points in mask *B* and $N_{A \cap B}$ is the number of points the two masks have in common. The similarity index assumes values between zero and one, where a value of zero indicates that the two masks do not intersect at all and a value of one indicates that the two masks are identical.

COMA and *MASKIT* calculate local correlation maps between NCS-related molecules (*COMA*) or between copies of a molecule in different crystal forms (*MASKIT*). Besides for the purpose of defining molecular envelopes, these correlation maps can also be used to assess how similar, for instance, NCS-related molecules

are and in which parts they may be genuinely different (G. J. Kleywegt, submitted). Such information can be useful when designing an appropriate NCS-restraint scheme for use in crystallographic refinement (Kleywegt, 1996), or when assessing whether large differences observed between two copies of a molecule related by NCS or present in two different crystal forms are supported by the experimental data. Alternatively, local density correlation maps can be used to assess whether an NCS operator is proper or improper. If an improper NCS operator is treated as though it were proper, major difficulties may be encountered during subsequent averaging procedures (Tête-Favier *et al.*, 1993). If the NCS is truly proper, then the Cartesian NCS operators form a closed group. This means that a local density correlation map in such a case will encompass the entire *N*-mer, rather than just a monomer. Finally, the local density correlation method can be used to assess if the assumption that a single set of NCS operators applies to an entire molecule is valid or not. For instance, if two domains of a protein have slightly different

relative orientations in NCS-related copies, a single set of NCS operators may yield high local density correlation values for only one of these domains. This implies that the two domains should be treated separately during averaging (*i.e.* each with their own mask and set of NCS operators).

6. Availability

Fortran-77 routines to read and write masks in *O/MAMA* format are available for use in other programs (ftp://alpha2.bmc.uu.se/pub/gerard/extras/fortran/latest_mask_io.f). *O* is a crystallographic modelling program which runs on a variety of workstations (see: <http://alpha2.bmc.uu.se/~alwyn/> and <http://imsb.au.dk/~mok/o/>). For more information about *O*, contact TAJ (<mailto:alwyn@xray.bmc.uu.se>). *MAMA*, *COMA*, *MASKIT* and *MAPMAN* are part of the *RAVE* package of real-space averaging and density-

manipulation programs. Manuals for these programs are available on the WWW (<http://alpha2.bmc.uu.se/usf/>); they describe the operation and use of the programs in detail, and provide examples and 'recipes' for inexperienced users. *RAVE* is available free of charge to the academic and not-for-profit community (<ftp://alpha2.bmc.uu.se/pub/gerard/rave/>); commercial users may contact GJK (<mailto:gerard@xray.bmc.uu.se>) for licensing details.

The authors wish to thank Dr Jonas Uppenberg for providing the MIR maps of the two crystal forms of *C. antarctica* lipase B. This work was supported by Uppsala University, the Swedish Natural Sciences Research Council, the Swedish Foundation for Strategic Research and its Structural Biology Network (SBNet).

References

- Bhat, T. N. & Blow, D. M. (1983). *Acta Cryst.* **A38**, 21–29.
- Bricogne, G. (1974). *Acta Cryst.* **A30**, 395–405.
- Bricogne, G. (1976). *Acta Cryst.* **A32**, 832–847.
- Delaney, J. S. (1992). *J. Mol. Graph.* **10**, 174–177.
- Divne, C., Ståhlberg, J., Reinikainen, T., Ruohonen, L., Pettersson, G., Knowles, J. K. C., Teeri, T. T. & Jones, T. A. (1994). *Science*, **265**, 524–528.
- Jones, T. A. (1992). *Proceedings of the CCP4 Study Weekend. 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., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* **A47**, 110–119.
- Kleywegt, G. J. (1996). *Acta Cryst.* **D52**, 842–857.
- Kleywegt, G. J. & Jones, T. A. (1994). *Proceedings of the CCP4 Study Weekend. 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.* **D52**, 826–828.
- Kleywegt, G. J. & Read, R. J. (1997). *Structure*, **5**, 1557–1569.
- Leslie, A. G. W. (1987). *Acta Cryst.* **A43**, 134–136.
- Rossmann, M. G. & Blow, D. M. (1963). *Acta Cryst.* **16**, 39–45.
- Tête-Favier, F., Rondeau, J. M., Podjarny, A. & Moras, D. (1993). *Acta Cryst.* **D49**, 246–256.
- Uppenberg, J., Öhrner, N., Norin, M., Hult, K., Kleywegt, G. J., Patkar, S., Waagen, V., Anthonsen, T. & Jones, T. A. (1995). *Biochemistry*, **34**, 16838–16851.
- Uppenberg, J., Trier Hansen, M., Patkar, S. & Jones, T. A. (1994). *Structure*, **2**, 293–308.
- Vellieux, F. M. D. A. P., Hunt, J. F., Roy, S. & Read, R. J. (1995). *J. Appl. Cryst.* **28**, 347–351.
- Vellieux, F. M. D. & Read, R. J. (1997). *Methods Enzymol.* **277**, 18–53.
- Wang, B. C. (1985). *Methods Enzymol.* **115**, 90–112.