Acta Crystallographica Section D

Biological
Crystallography

ISSN 0907-4449

On the fitting of model electron densities into EM reconstructions: a reciprocal-space formulation

J. Navaza, J. Lepault, F. A. Rey, C. Álvarez-Rúa and J. Borge

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On the fitting of model electron densities into EM reconstructions: a reciprocal-space formulation

J. Navaza,[a]* J. Lepault,[a] F. A. Rey,[a] C. Álvarez-Rúa[b] and J. Borge[b]

[a]LVMS, CNRS-GIF, 91198 Gif-sur-Yvette, France, and [b]Departamento de Química Física y Analítica, Universidad de Oviedo, 33006 Oviedo, Spain

Correspondence e-mail: jorge.navaza@gv.cnrs-gif.fr

Received 15 April 2002
Accepted 30 July 2002

A fast method for fitting model electron densities into EM reconstructions is presented. The methodology was inspired by the molecular-replacement technique, adapted to take into account phase information and the symmetry imposed during the EM reconstruction. Calculations are performed in reciprocal space, which enables the selection of large volumes of the EM maps, thus avoiding the bias introduced when defining the boundaries of the target density.

1. Introduction

X-ray crystallography provides atomic resolution models of subunits of large biological assemblies. However, the technique requires crystals, which are difficult or practically impossible to obtain for large molecular complexes. On the other hand, electron microscopy (EM) allows the imaging of complete biological assemblies. Unfortunately, the images only extend to a limited resolution, which does not allow the distinction of atomic details of the three-dimensional structures of the particles. The combination of the information provided by both techniques often enables the interpretation of the EM reconstruction in terms of atomic models. This relies on the possibility of docking models into EM maps.

Sometimes, the EM density presents strong features which allow a direct fitting by eye of the X-ray structures into the complete EM reconstruction (Hewat *et al.*, 1998) or into the difference EM map obtained by subtraction of already positioned components of the biological assembly (Rossmann, 2000). This manual docking only provides a qualitative idea of the correctness of the fit. Usually, the procedure is complemented with a quantitative refinement of the parameters that specify the positions of the fitted molecules.

Several quantitative methods for docking have been developed (Stewart *et al.*, 1993; Cheng *et al.*, 1995; Che *et al.*, 1998; Wriggers *et al.*, 1999; Volkmann & Hanein, 1999; Belnap *et al.*, 2000; Roseman, 2000; Rossmann, 2000; Thouvenin & Hewat, 2000). The method we present here is directly inspired by a crystal structure solution technique.

2. A molecular-replacement approach

Basically, the problem of fitting an atomic model into an EM map may be addressed using the concepts of the molecular-replacement method (MR) of X-ray crystallography (Rossmann & Arnold, 2001). However, some important differences exist between both problems which hinder the use of conventional molecular-replacement programs with EM data. The most important difference is that EM images suffer from lack of resolution (typically below 15 Å, with the exception of two-dimensional crystals) and low signal-to-noise ratio. Fortunately, phase information is available.

A consequence of the low resolution (and sometimes the low quality) of the EM maps is that very often the frontiers of individual molecules cannot be distinguished easily. In this case, the extraction of volumes containing single molecules is inevitably biased. It is then highly desirable to consider large volumes containing several copies of the independent molecules without making any assumption concerning their shapes. These copies are related by the symmetry imposed during the EM reconstruction (icosahedral, helical, point-group symmetry), so that equivalent molecules will, in general, be sampled at non-equivalent grid points. Indeed, most of the imposed symmetries are not compatible with equally spaced Cartesian grids, which is the standard way EM maps are presented. Therefore, even if the information content of a whole map is the same as that of its asymmetric part, the form in which data is available determines the procedures to use the information in an efficient way.

Most MR procedures can be performed either in direct space or in reciprocal space. When the boundaries of isolated molecules can be determined, the advantage of reciprocal-space over real-space formulations is minimal. However, reciprocal space enables the use of large volumes, if necessary, while keeping coarse grid spacing. The actual size of the selected volume results from a compromise between the amount of computation and the number of symmetry elements included.

The image of the biological assembly can be used to guess initial positions of the search models. Hence, the MR problem can be reduced to the application of a rigid-body refinement protocol starting from putative locations of the model molecules, instead of performing exhaustive six-dimensional searches or separate rotational and translational searches as in the standard MR procedure. Indeed, phase information is straightforwardly derived from the EM reconstruction. Its presence dramatically increases the radius of convergence of the refinement procedures as compared with the standard, phaseless, MR case.

We have adapted to the EM case a rigid-body refinement method used in X-ray crystallography. The algorithms and their implementations are essentially those described in the program *FITING* (Castellano *et al.*, 1992).

2.1. Formulation of the fitting problem

We want to compare the EM map, $\rho^{\rm em}(\mathbf{r})$, and a model-based electron density, $\rho^{\rm mod}(\mathbf{r})$, within a selected volume of the EM reconstruction. This region should include a reasonable number of copies of the independent components of the EM map. The selected region will be called the 'EM box'.

The goodness of fit will be measured by the normalized quadratic misfit

$$Q = \int [\rho^{\rm em}(\mathbf{r}) - \lambda\rho^{\rm mod}(\mathbf{r})]^2 {\rm d}^3r \Big/ \int \rho^{\rm em}(\mathbf{r})^2 {\rm d}^3 r, \qquad (1)$$

the integration being extended over the EM box. λ is the scale factor. By using Parseval's theorem, this expression is written in terms of reciprocal-space variables as

$$Q = \sum_{\mathbf{H}} \left|F^{\rm em}_{\mathbf{H}} - \lambda F^{\rm mod}_{\mathbf{H}}\right|^2 \Big/ \sum_{\mathbf{H}} \left|F^{\rm em}_{\mathbf{H}}\right|^2. \qquad (2)$$

$F^{\rm em}$ and $F^{\rm mod}$ are the Fourier transforms of $\rho^{\rm em}$ and $\rho^{\rm mod}$, respectively.

$F^{\rm mod}$ is efficiently calculated in terms of the individual molecular scattering factors $f(\mathbf{s})$, *i.e.* the Fourier transform of the electron density corresponding to the isolated molecule. Although all molecules within the integration volume contribute to $F^{\rm mod}$, only a few of them are independent, the others being generated by application of a subset G of the symmetry operations imposed during the EM reconstruction. G must be chosen so that the generated molecules cover the EM box as much as possible.

Each independent molecule is considered as a rigid body. Its position within the EM box is specified by a rotation matrix $\mathbf{R}$ parameterized by the Euler angles (α, β, γ) and a translation vector $\mathbf{T}$ with fractional coordinates (x, y, z) in the EM box. The zero value of the six positional variables (the reference position) corresponds to the model with its centre of mass at the origin and its principal axes of inertia parallel to the box edges.

In the case of only one independent molecule, $F^{\rm mod}$ is written as

$$F^{\rm mod}_{\mathbf{H}}(\mathbf{R},\mathbf{T}) = \sum_{g\in G} f(\mathbf{H}\mathbf{M}_g\mathbf{R}) \exp[2\pi i\mathbf{H}(\mathbf{M}_g\mathbf{T} + \mathbf{t}_g)], \qquad (3)$$

where $\mathbf{M}_g$ and $\mathbf{t}_g$ denote the transformation matrix and the translation vector corresponding to the gth symmetry operation of the EM reconstruction. If there are several independent molecules, the calculated structure factors will be given by a sum of individual contributions like (3). Note that some independent molecules may correspond to the same molecular model.

It is worth noting that the model scattering factors may well be the Fourier transform of any electron density or even an electron-microscopy reconstruction.

2.2. The optimization protocol

The quadratic misfit Q is a function of the positional variables of the independent molecules and the scale factor. Their optimal values are obtained through a minimization process starting from initial positions determined by visual inspection or with the help of a phased-translation function (see, for example, Navaza, 2001). The minimization proceeds by cycles, where the positions of all the independent molecules are sequentially refined, keeping the others fixed. The minimization stops when the change in positions observed during a whole cycle are smaller than a given threshold.

The choice of the subset G of symmetry operations must guarantee that the generated molecules lay inside the EM box. In this case, the truncation of the density at the box edges has no effect on the target function. Indeed, minimizing the quadratic misfit is strictly equivalent to maximizing the correlation coefficient

$$\mathrm{CC} = \frac{\sum_{\mathbf{H}} \overline{F^{\mathrm{em}}_{\mathbf{H}}} F^{\mathrm{mod}}_{\mathbf{H}}}{(\sum_{\mathbf{H}} |F^{\mathrm{em}}_{\mathbf{H}}|^2)^{1/2} (\sum_{\mathbf{H}} |F^{\mathrm{mod}}_{\mathbf{H}}|^2)^{1/2}}, \qquad (4)$$

where the overline means 'complex conjugate'. CC is independent of the scale factor. In the absence of overlap of the moving molecules, the denominator remains unchanged. Using Parseval's theorem, the numerator of CC can be written as

$$\int \rho^{\mathrm{em}}(\mathbf{r})\, \rho^{\mathrm{mod}}(\mathbf{r})\, \mathrm{d}^3 r. \qquad (5)$$

The integral and hence the driving force of the minimization procedure is determined by the values of ρ^{em} in the regions where the molecules exist.

3. Methodology

The whole docking process consists of a certain number of steps, including the use of graphics, map manipulation and optimization procedures. In this section, we describe the methodology while applying it to a real case that illustrates most of the difficulties encountered in practice. We will discuss the fitting of the major capsid protein of rotavirus (VP6; PDB code 1qhd; Mathieu *et al.*, 2001) into a helical EM reconstruction (Lepault *et al.*, 2001). This structural protein forms the middle layer in the triple-layered viral capsid. When isolated, the protein VP6 self-assembles into spherical or helical particles mainly depending upon pH. Two types of helical assemblies have been observed: large and small tubes with diameters of 75 and 45 nm, respectively. We will use the small-tube data.

3.1. Extraction of the EM box and computation of its Fourier coefficients

A helical assembly can be considered as an infinite periodic arrangement. The selection rule that defines the symmetry of the small tubes is $l = -9n + 103m$, with an axial repeat of 820 Å. Obviously, only a portion of the helix is selected to carry out the fitting procedure. An orthogonal box (the EM box; Fig. 1) of dimensions 252 × 332 × 300 Å was extracted from the whole reconstruction using a combination of programs widely used in X-ray crystallography (*O*, *AL_MAPMAN* and *CCP*4; Jones *et al.*, 1991; Kleywegt & Jones, 1996; Collaborative Computational Project, Number 4, 1994). The grid spacing of the map was 4 Å; the nominal resolution of the reconstruction was 15 Å.

The EM image magnification may be inaccurate by as much as 5%, so that the absolute scale of the reconstruction had to be determined. This was performed simultaneously with the refinement process, as explained later.

Calculations were performed using data to 20 Å, a value close to the true resolution of the reconstruction. A Fourier transform of the EM density yielded a set of 6800 Fourier coefficients.

3.2. Calculation of the molecular scattering factors and setting the initial molecular positions

The molecular scattering factors were calculated by fast Fourier transforming the electron density generated from the VP6 atomic coordinates (Ten Eyck, 1977). First, the model was placed at its reference position, as previously explained. The molecular scattering factors were finely sampled to allow model structure factors and gradients of the rotating model to be accurately interpolated, as required by the optimization procedure.

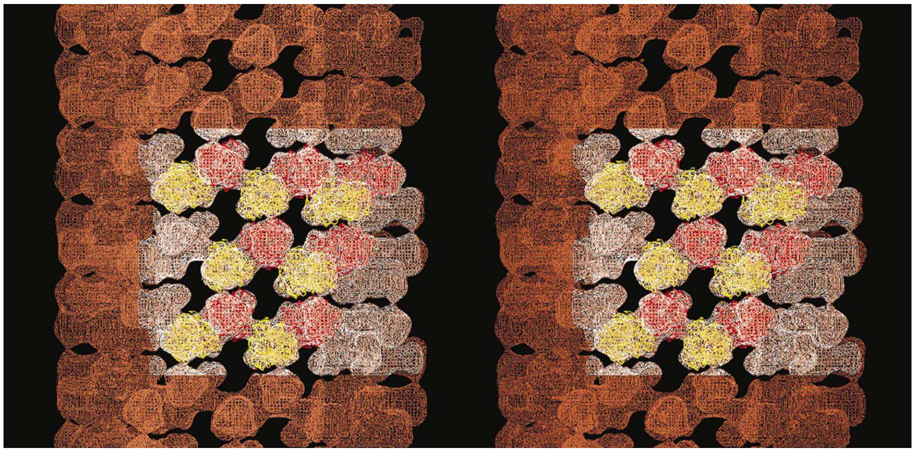

Figure 1
Original EM map showing the selected box (the EM box) used for fitting and the whole set of fitted molecules, including the independent ones. Images were generated using *PyMol* (DeLano, 2002).

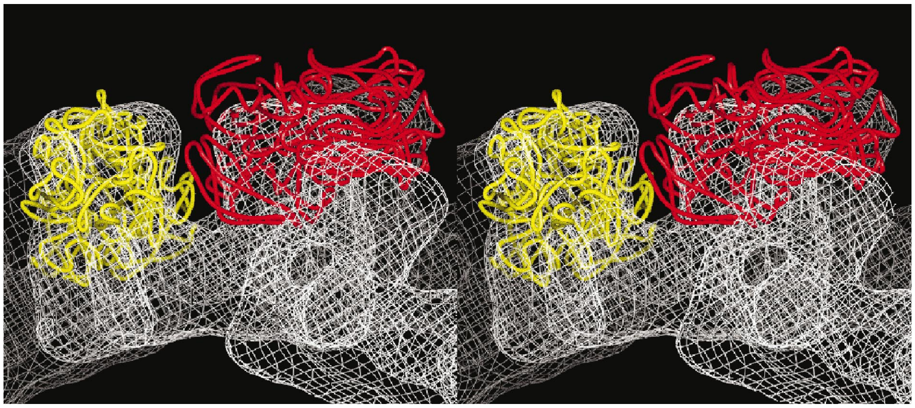

Figure 2
View of the two independent molecules in their initial positions.

Original EM map		U			V		
	Initial CC, R, Q	19.6			66.2		63.5
	Cumulated shifts (r, t, a)	9.81	16.74	19.40	28.18	18.78	33.87
Cycle 1	Final CC, R, Q	44.3			66.2		45.2
....							
	Cumulated shifts (r, t, a)	2.87	0.64	2.94	1.61	0.96	1.87
Cycle 4	Final CC, R, Q	63.5			64.5		31.0
....							
	Cumulated shifts (r, t, a)	0.02	0.01	0.02	0.00	0.01	0.01
Cycle 8	Final CC, R, Q	63.6			64.6		30.9

(*a*)

Masked map		U			V		
	Cumulated shifts (r, t, a)	0.30	0.29	0.41	0.43	0.41	0.59
Cycle 1	Final CC, R, Q	94.1			33.4		5.8
....							
	Cumulated shifts (r, t, a)	0.00	0.00	0.00	0.01	0.00	0.01
Cycle 3	Final CC, R, Q	94.1			33.4		5.8

(*b*)

Figure 3
(*a*) Results of the optimization process for the two independent molecules (U and V). Cumulated shifts are expressed in r.m.s. deviation units (Å). Types of shifts: *r*, rotational; *t*, translational; *a*, all [$a = (r^2 + t^2)^{1/2}$]. (*b*) Results of the optimization process using a mask based on the previously refined positions.

The repeating unit of the helix contains a pair of trimers. Therefore, the positions of these two molecules had to be independently treated. Initial positions were visually obtained by grossly placing the model within the EM density, as shown in Fig. 2, with the help of the program *O*.

3.3. Selection of the symmetry operators

All the protein molecules in the assembly can be generated by application of the helical symmetry operations imposed during the EM reconstruction to the repeating unit of the helix. However, the EM box is filled by applying a subset *G* of these transformations to the initial positions of the independent molecules. An ample subset was generated. As a consequence, some of the symmetry mates laid outside the box and the corresponding operators had to be pruned away from the original subset. This was accomplished with the help of the program *O*. Necessarily, some portions of the EM map were not covered by any molecule. For the selected region shown in Fig. 1, the subset *G* consisted of seven elements, including the identity operator.

It is possible that during the refinement some of the symmetry mates may move outside the selected fitting region. It is then necessary to update *G* according to the actual positions of the fitted molecules.

3.4. Refinement procedure

The positional parameters of the independent molecules were refined following the optimization procedure described above. Data from 400 to 20 Å were used in the minimization process. Several optimization cycles were performed until no shifts in the coordinates higher than a certain threshold were observed.

At this stage, the magnification of the EM image was determined by changing the values of the EM grid spacing, Fourier transforming the rescaled EM map and repeating the optimization process. The value of the correlation coefficient was used to assess the different scales. Detailed results of the refinement corresponding to the optimal magnification are shown in Fig. 3(*a*) and Figs. 1 and 4. The r.m.s. deviations between the initial and the refined positions were quite high: 53.8 and 38.3 Å.

Fig. 3(*a*) shows values of the correlation coefficient which are rather low. This is because there are regions of density in the EM box which are not covered by any molecule. After refinement, these regions were masked off in order to obtain a meaningful value of CC. The original EM map was replaced by this new masked map and the minimization process continued. The results are shown in Fig. 3(*b*). We see a substantial improvement in the magnitudes of the target function, the CC and the crystallographic *R* factor, but only minimal shifts in coordinates.

3.5. Radius of convergence

Tests were carried out to assess the radius of convergence of the minimization procedure. Random shifts were applied to the refined coordinates and the optimization protocol was restarted. The results suggest that the radius of convergence of the procedure, for data in the 400–20 Å resolution range, is slightly smaller than 30 Å r.m.s. deviation. Indeed, only 3% of trials failed when initiated with positions shifted by 30 Å, whereas all converged to the correct solution for shifts below 27.5 Å r.m.s. deviation. By using data to 25 Å, all trials succeeded. As expected, the radius of convergence increases when the high-resolution limit is lowered.

3.6. Splitting the refined models into domains

VP6 forms a tight trimer composed of two domains: a β-barrel domain, which we call the head, and an α-helical domain, which we call the base. When assembled, some domains of the atomic structure may move. The resulting

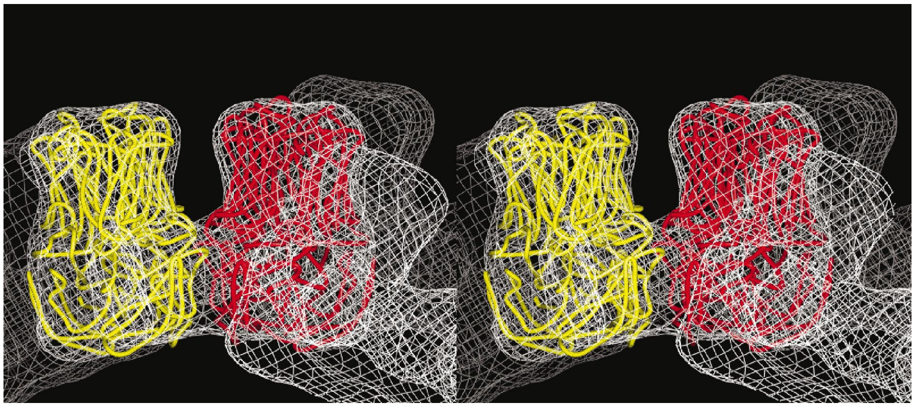

Figure 4
Refined positions of the independent molecules for the same view as in Fig. 2.

Original EM map		U_1	U_2	U_3	U_4	V_1	V_2	V_3	V_4
	Initial CC, R, Q		63.6		64.6			30.9	
Cycle 1	Cumulated shifts (r, t, a)	1.33 1.30 1.86	1.77 1.93 2.62	3.26 2.35 4.02	8.26 2.31 8.57	1.46 1.08 1.82	2.07 2.04 2.90	3.44 2.69 4.37	10.05 2.30 10.31
	Final CC, R, Q		65.2		62.3			29.9	
....									
Cycle 9	Cumulated shifts (r, t, a)	0.01 0.00 0.01	0.04 0.02 0.05	0.06 0.02 0.06	0.05 0.03 0.05	0.02 0.01 0.02	0.04 0.02 0.05	0.03 0.01 0.03	0.06 0.02 0.06
	Final CC, R, Q		65.2		62.2			29.8	

Masked map		U_1	U_2	U_3	U_4	V_1	V_2	V_3	V_4
Cycle 1	Cumulated shifts (r, t, a)	0.38 0.64 0.74	1.42 1.16 1.84	0.84 1.53 1.74	0.46 1.02 1.12	0.40 0.61 0.73	1.53 1.14 1.91	2.07 1.63 2.64	0.51 1.30 1.40
	Final CC, R, Q		95.7		29.1			4.3	
....									
Cycle 5	Cumulated shifts (r, t, a)	0.00 0.00 0.00	0.02 0.01 0.02	0.03 0.02 0.03	0.02 0.01 0.02	0.01 0.00 0.01	0.03 0.00 0.03	0.04 0.03 0.05	0.02 0.01 0.02
	Final CC, R, Q		95.7		29.1			4.2	

Figure 5
Results of the optimization process after splitting each one of the independent molecules (U and V) into four domains. The subindex 1 corresponds to the head and 2, 3 and 4 correspond to the legs of VP6. Cumulated shifts are expressed in r.m.s. deviation units (Å). Types of shifts: *r*, rotational; *t*, translational; *a*, all [$a = (r^2 + t^2)^{1/2}$].

structure may thus differ from the crystallographic one. In our case, three different types of lateral contacts have been observed. The threefold symmetry of the trimer is broken at the level of the contacts with adjacent molecules, which occur mainly between the α-helical domains. This suggests that the structure of these domains may slightly change with respect to the X-ray structure.

Therefore, each trimer was split into four different domains: a head, composed of the β-barrel domain, and three 'legs', composed of the three α-helical domains. The optimization process was performed using the original EM map in order to avoid the bias toward the refined model which may have been introduced in the calculation of the mask. The initial coordinates of the eight independent subunits were obtained from the refined positions of the two trimers.

The refinement was carried out following the steps described in the preceding section. Fig. 5 shows slight improvements of the correlation coefficient and the crystallographic *R* factor, even when using a mask. The r.m.s. deviations between the coordinates of the refined molecules before and after splitting were 4.55 and 4.57 Å, respectively. Fig. 6 shows the final positions of the split trimers. Note how arbitrary it would be to extract volumes containing single molecules, especially when one is interested in recovering geometrical information about the assembling mechanisms.

4. Conclusions

The procedure for fitting molecular models into EM reconstructions has been implemented in a package called *URO* which is available free of charge from the authors upon request. It has been used to fit atomic models into different EM reconstructions, including whole icosahedral particles. The main characteristics of the procedure are as follows.

(i) It is carried out in reciprocal space. Important volumes of the reconstruction containing several symmetry mates may be taken into account. Symmetry is incorporated in a simple way into the optimization algorithm.

(ii) Many independent bodies can be simultaneously fitted. The package provides a tool for automatically splitting the search models into subunits, which are then independently refined. Obviously, there is a limit to the number of domains into which the original model can be split. This limit is determined by the resolution of the EM map and the size of the domains.

(iii) Reciprocal space offers the possibility of changing the resolution of the data included in the calculations. This can be exploited to increase the radius of convergence of the optimization procedure.

(iv) Last but not least, the whole procedure is fast. The most time-consuming part is map and symmetry manipulations to extract the EM box and select the corresponding symmetry operations, a trivial task for an experienced *O* user. For the above example, the whole procedure takes about 1 h of user time and about 1 min of CPU on a Digital XP1000, 500 MHz.

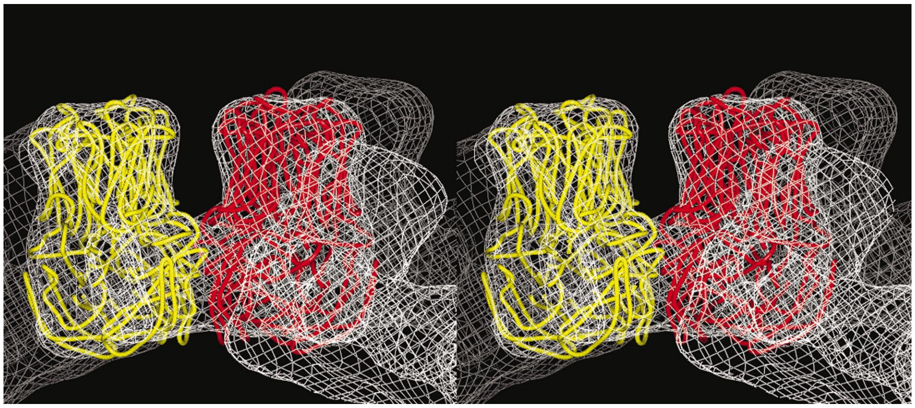

Figure 6
Refined positions of the independent domains. Each independent molecule is split into four domains. Same view as in Fig. 4.

One of the authors (JN) is grateful to York Structural Biology Laboratory, where he wrote a preliminary version of the package, for the stimulating scientific atmosphere and financial support.

References

Belnap, D. M., Filman, D. J., Trus, B. L., Cheng, N., Booy, F. P., Conway, J. F., Curry, S., Hiremath, C. N., Tsang, S. K., Steven, A. C. & Hogle, J. M. (2000). *J. Virol.* **74**, 1342–1354.

Castellano, E. E., Oliva, G. & Navaza, J. (1992). *J. Appl. Cryst.* **25**, 281–284.

Che, Z., Olson, N. H., Leippe, D., Lee, W.-M., Mosser, A. G., Rueckert, R. R., Baker, T. S. & Smith, T. J. (1998). *J. Virol.* **72**, 4610–4622.

Cheng, R. H., Kuhn, R. J., Olson, N. H., Rossmann, M. G., Choi, H.-K., Smith, T. J. & Baker, T. S. (1995). *Cell*, **80**, 621–630.

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

DeLano, W. L. (2002). *The PyMol Molecular Graphics System.* DeLano Scientific, San Carlos, CA, USA; http://www.pymol.org.

Hewat, E. A., Marlovits, T. C. & Blaas, D. (1998). *J. Virol.* **72**, 4396–4402.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

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

Lepault, J., Petitpas, I., Erk, I., Navaza, J., Bigot, D., Dona, M., Vachette, P., Cohen, J. & Rey, F. A. (2001). *EMBO J.* **20**, 1498–1507.

Mathieu, M., Petitpas, I., Navaza, J., Lepault, J., Kohli, E., Pothier, P., Prasad, B. V. V., Cohen, J. & Rey, F. A. (2001). *EMBO J.* **20**, 1485–1497.

Navaza, J. (2001). *Acta Cryst.* D**57**, 1367–1372.

Roseman, A. M. (2000). *Acta Cryst.* D**56**, 1332–1340.

Rossmann, M. G. (2000). *Acta Cryst.* D**56**, 1341–1349.

Rossmann, M. G. & Arnold, E. (2001). Editors. *International Tables for Crystallography*, Vol. F. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Stewart, P. L., Fuller, S. D. & Burnett, R. M. (1993). *EMBO J.* **12**, 2589–2599.

Ten Eyck, L. F. (1977). *Acta Cryst.* A**33**, 486–492.

Thouvenin, E. & Hewat, E. (2000). *Acta Cryst.* D**56**, 1350–1357.

Volkmann, N. & Hanein, D. (1999). *J. Struct. Biol.* **125**, 176–184.

Wriggers, W., Milligan, R. A. & McCammon, J. A. (1999). *J. Struct. Biol.* **125**, 185–195.