Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Locating proper non-crystallographic symmetry in low-resolution electron-density maps with the program *GETAX*

Clemens Vonrhein and Georg E. Schulz*

Institut für Organische Chemie und Biochemie, Albertstrasse 21, D-79104 Freiburg im Breisgau, Germany

Correspondence e-mail:
schulz@bio5.chemie.uni-freiburg.de

Received 6 April 1998
Accepted 4 June 1998

Non-crystallographic symmetry averaging for improving and extending an initial set of phases can be crucial at an early stage of a protein structure analysis. A method is described which detects the position of a proper rotation axis in a surprisingly poor electron-density map and is fast enough to run through a large number of axis orientations. It uses a simple *multimer* mask to define the searching unit, which is then shifted through the whole unit cell looking for the position with the highest correlation coefficient between the interrelated parts. Appropriate weighting and averaging enhances the signal-to-noise ratio. Examples of the application of this algorithm are given. The use of the local rotation axis for phasing is commented on. A search of the Protein Data Bank showed that 27% of the unique crystal forms contain proper local n-fold axes, which could have been located with the presented method.

1. Introduction

Non-crystallographic symmetry (NCS) averaging is a powerful tool for phase refinement and extension to full resolution of a native data set (Bricogne, 1974; Kleywegt & Read, 1997). Together with other density-modification techniques like solvent flattening (Wang, 1985) and histogram matching (Zhang & Main, 1990*a*,*b*), NCS averaging can refine and extend very poor starting phases from low to high resolution (Seemann & Schulz, 1997).

The rotational part of the NCS can be determined without any phase knowledge from a self-rotation function (Rossmann & Blow, 1962). In contrast, the translational part usually requires some phase information. Occasionally it can be deduced from the relationship between heavy-atom binding sites of a derivative. In exceptional cases, part of the translational component of an NCS operator can be determined directly through examination of Patterson vectors (Rossmann *et al.*, 1964; Eagles *et al.*, 1969; Epp *et al.*, 1971; Schirmer *et al.*, 1995; Stubbs *et al.*, 1996). Here we describe a general method to locate an n-fold axis in poor electron-density maps. The method requires a crude estimate of the multimer shape and size. It can be run in a fully automated manner and is fast enough to check a large number of axis orientations.

2. Method

2.1. NCS-related monomers

Two NCS-related monomers A and B obey the equation

$$B = \mathbf{R}A + \mathbf{T},$$

where $\mathbf{R}$ is a rotation matrix and $\mathbf{T}$ a translation vector. The rotation is defined by the orientation of its axis and the rotation angle κ. If the NCS is an n-fold rotation axis with symmetry C_n, this equation can be written as

$$B = \mathbf{R}(A - X) + X,$$

where X is a position on the rotation axis and κ is a multiple of $360°/n$. In order to locate the NCS in an electron-density map, all grid points X of such a map can be examined for the best correlation of **R**-related densities within a given mask around X.

2.2. Setting up the multimer mask

Fig. 1 gives a flow chart for a protein structure analysis using NCS averaging. The space group and the cell dimensions of the crystal, together with the monomer mass (M_r), yield an estimate for the number of monomers in the asymmetric unit (Matthews, 1968). The estimate may be improved by a crystal-density measurement. This number is taken into account when inspecting the self-rotation function for an n-fold rotation axis, which is calculated from a native data set. A set of phases is required to produce an initial electron-density map. In general, the phases are derived from heavy-atom-derivative data. For the program *GETAX*, the electron-density map is calculated at about 6 Å resolution for the whole unit cell using a grid spacing of about 2 Å.

From a rough estimate of the multimer shape and size (deduced from M_r using a protein density of 1.35 g cm^{-3}), a simple multimer mask is constructed around the z axis of an orthogonalized system and centred at the origin. The program *GETAX* allows for a spherical or a cylindrical mask. The mask is then subdivided into n sets of points (each describing one monomer) interrelated by the n-fold axis. For the first set, these points are located on grid points of the electron-density map, whereas those of the other sets are on general positions, as they are generated by rotating the first set around the z axis. The points on the z axis are exempted and labelled for later use in line averaging (see §2.3). For all space groups except $P1$ these points are on the grid, whereas for $P1$ they are on general positions.

The rotation axis is then tilted to the given orientation, moving all points to new positions which, in general, do not coincide with the grid. In order to simplify the computation, all points of each set and of the axis are subsequently shifted to the nearest grid point so that the interrelation concerns sets of grid points. As an example, the resulting sets for a cylindrical mask around a twofold axis are shown in Fig. 2(*a*).

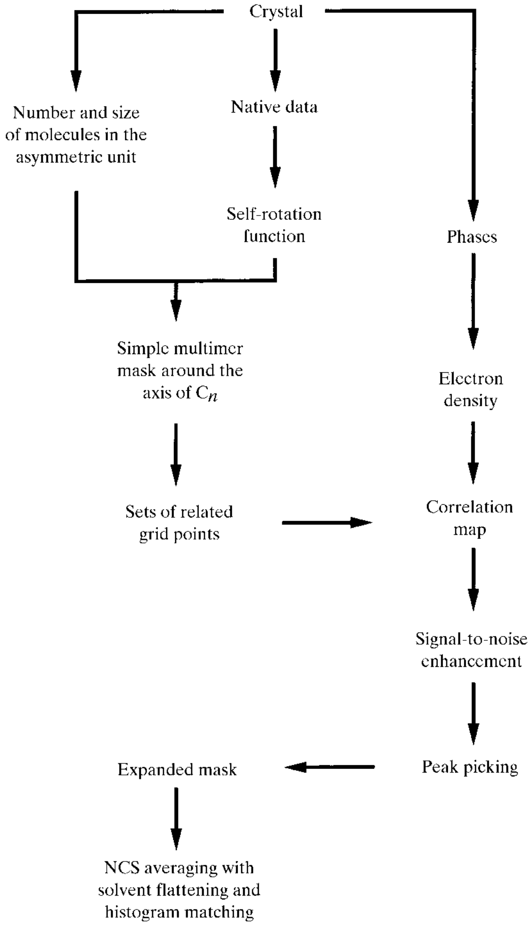

Crystal
Number and size of molecules in the asymmetric unit
Native data
Self-rotation function
Phases
Simple multimer mask around the axis of C_n
Electron density
Sets of related grid points
Correlation map
Signal-to-noise enhancement
Expanded mask
Peak picking
NCS averaging with solvent flattening and histogram matching

Figure 1
Flow chart describing the procedure for locating and applying NCS. Note that the expanded-mask approach is recommended after the NCS has been established.

2.3. Correlation map and signal-to-noise enhancement

The interrelated point sets are then shifted by a vector **X** over the grid of the whole unit cell, calculating the average correlation coefficient CC(X) at position X by a similar method to Vellieux *et al.* (1995) using

$$\mathrm{CC}(X) = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} C_{ij}(X)}{n(n-1)/2},$$

where $C_{ij}(X)$ is the pairwise density correlation coefficient of point sets i and j representing monomers A and B. CC(X) then constitutes a correlation map in the same unit cell and on the same grid as the density map. In the example of Fig. 2(*a*), $n = 2$ and there is only a single coefficient $C_{12}(X)$.

For symmetry C_n with $n > 2$ or for D_n, the signal can be enhanced by subtracting a penalty term for discrepancy, $\sigma_{\mathrm{CC}(X)}$, which is defined as the standard deviation of the $n(n-1)/2$ values $C_{ij}(X)$,

$$\mathrm{CC}_E(X) = \mathrm{CC}(X) - \sigma_{\mathrm{CC}(X)}.$$

The resulting $\mathrm{CC}_E(X)$- or CC(X)-correlation map can be examined either automatically or visually. Usually, a search for C_n symmetry results in long stretches of high correlations (see Fig. 2*a*) because the local symmetry is not very sensitive to shifting the mask along the rotation axis.

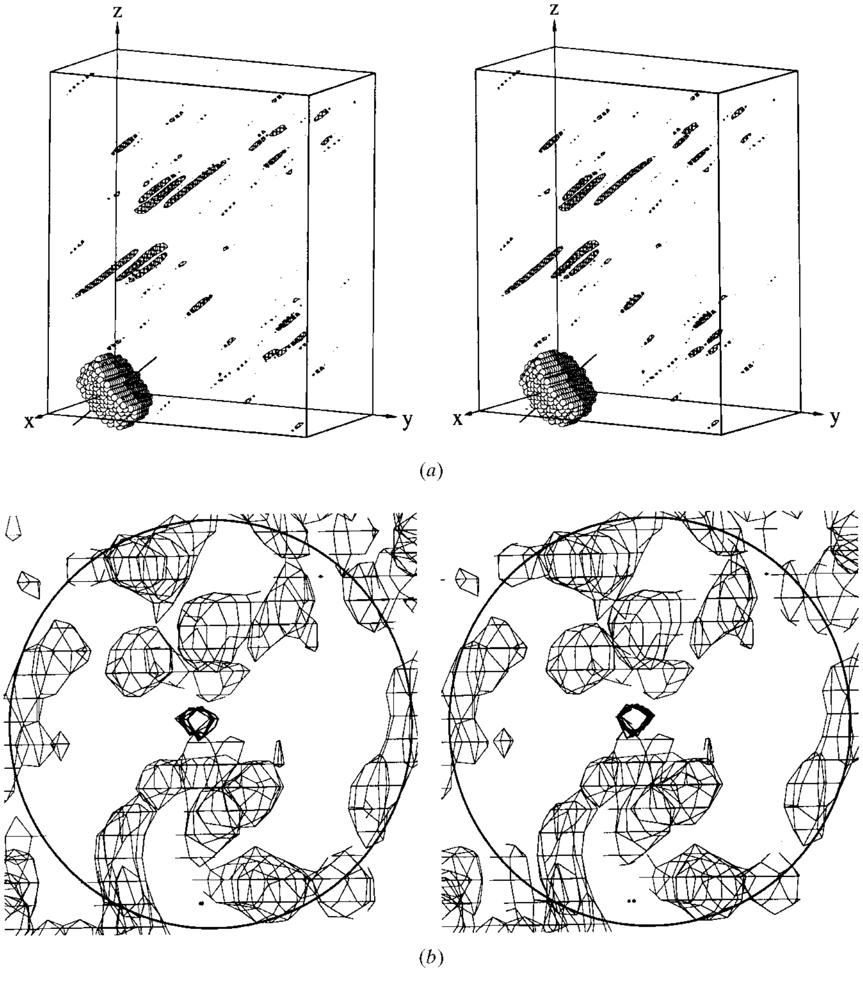
This insensitivity of C_n to such longitudinal shifts is used for a further improvement of the signal-to-noise ratio. For this purpose the m labelled positions of the axis (see §2.2) are used to calculate a line-averaged correlation $\mathrm{CC}_{\mathrm{LA}}(X)$ according to

$$\mathrm{CC}_{\mathrm{LA}}(X) = \sum_{k=1}^{m} w_k \, \mathrm{CC}_E(X_k),$$

Table 1
Performance of *GETAX* as a function of phase quality and mask.

	Phase set†			Cylindrical mask‡						Difference for phase set§		
	1S (%)	2S (%)	3M (%)	Radius (Å)	Height (Å)	$V_{\rm rel}$ (%)	$V_{\rm prot}$ (%)	$V_{\rm sol}$ (%)	$V_{\rm other}$ (%)	1S ($\Delta\sigma$)	2S ($\Delta\sigma$)	3M ($\Delta\sigma$)
Completeness	68	98	99	—	—	—	—	—	—	—	—	—
F. o. m.¶	36	42	62	—	—	—	—	—	—	—	—	—
Mask-1	—	—	—	18	15	10	52	22	26	−1.1	2.0	2.2
Mask-2	—	—	—	26	14	20	51	25	24	0.5	3.0	4.2
Mask-3	—	—	—	30	14	26	51	28	21	0.8	3.3	3.3
Mask-4	—	—	—	30	22	38	49	31	20	0.9	0.6	1.6

† Calculated for the resolution range 40–6 Å using *MLPHARE* (Collaborative Computational Project, Number 4, 1994). The multiple isomorphous-replacement phase set 3M is the result of a parameter refinement of all three derivatives. Phase sets 1S and 2S are based on two different single heavy-atom derivatives using the heavy-atom parameters from 3M. ‡ Mask-1 was used in the structure analysis. $V_{\rm rel}$ is the mask volume as related to the protein volume of the two interrelated adenylate kinase trimers in the asymmetric unit. The protein volume was calculated using *NCSMASK* (Collaborative Computational Project, Number 4, 1994). $V_{\rm prot}$, $V_{\rm sol}$ and $V_{\rm other}$ are the volume fractions of the mask that are occupied by NCS-related protein, solvent and non-related protein, respectively. § All correlation maps were line averaged. The difference between the best correct and the highest wrong solution is stated in σ units of the respective map. ¶ Average figure of merit.



(*a*)

(*b*)

Figure 2
Real-space correlation search for a local C_2 symmetry using *GETAX* with crystals of adenylate kinase from *S. acidocaldarius* in space group $P2_12_12_1$. (*a*) Line-averaged correlation map $CC_{\rm LA}(X)$ based on the multiple isomorphous-replacement density map of phase set 3M (Table 1). The unit cell is drawn out, the grid size is about 2 Å, the contour level is 3σ. The mask containing the two sets of interrelated grid points (light and dark, mask-1 in Table 1) is shown together with its axis as centred at the origin $X = 0$. Placing the mask centre on all grid points results in the correlation map. (*b*) The analyzed 6 Å electron-density map contoured at 1σ (thin lines) with mask-1 (Table 1) at its best position. The view is along the local twofold axis; the $CC_{\rm LA}(X)$ map at a contour level of 3σ is overlaid with thick lines. The rim of the cylindrical mask is drawn out. The map section is as deep as the height of the cylinder.

where X_k is a labelled position of the axis when the mask is centred at X and w_k is the reciprocal distance between X_k and X (in units of Å^{-1} with $w_k = 1$ for the centre point $X_k = X$). The maximum of this map locates the best position for the mask centre, which is the best centre of the NCS.

3. Results and discussion

3.1. Establishing the NCS for a trimeric adenylate kinase

The structure of the adenylate kinase from *Sulfolobus acidocaldarius* could not be solved by molecular replacement, and despite solvent flattening and histogram matching, the available heavy-atom derivatives did not yield an interpretable electron-density map (Vonrhein *et al.*, 1998). Solution studies had indicated a dimer (Lacher & Schäfer, 1993). Space group $P2_12_12_1$, unit-cell dimensions (73.5, 145.7, 172.3 Å) and an M_r of 21 110 per monomer pointed to four dimers per asymmetric unit. At this stage the program *GETAX* was written to establish the NCS.

For structure analysis, a 6 Å resolution map based on phase set 3M (Table 1) was calculated on a 2 Å grid. A cylinder with a radius of 18 Å and a height of 15 Å was used as the multimer mask (mask-1 in Table 1). A self-rotation function yielded the highest peak at the polar angles $\omega = 50.5$, $\varphi = 90$ and $\kappa = 180°$. The maximum of the line-averaged correlation map $CC_{\rm LA}(X)$ was 2.2σ above the highest wrong peak. The map is depicted in Fig. 2(*a*). The electron density within mask-1 at its best position is shown in Fig. 2(*b*).

Table 2
Map qualities at several stages of an analysis.

	Map correlation for resolution range† (Å)		Deviation from final NCS‡	
Phases	20–4.5	20–3.0	Rotation (°)	Translation (Å)
3M§	0.49	–	–	–
S-H-3M¶	0.56	0.38	–	–
LS-S-H-3M††	0.70	0.51	1.0 (2.7)	2.2 (5.2)
CD-S-H-3M‡‡	0.55	0.39	2.3 (2.7)	4.6 (5.2)

† Map correlation coefficients to the final model were calculated using *OVERLAPMAP* (Collaborative Computational Project, Number 4, 1994). ‡ The deviation is the difference between the refined NCS resulting from program *DM* and that of the final model. In both cases, the phase refinement/extension started from the best position of *GETAX*, the deviation of which is given in parentheses. § Phase set based on all three heavy-atom derivatives of the adenylate kinase from *S. acidocaldarius* (Table 1). ¶ After 100 cycles of phase refinement/extension using solvent flattening and histogram matching and starting from phase set 3M. †† As S-H-3M, but adding NCS-averaging to the density-modification procedure. The applied mask was large and simple. It was produced by placing the searching mask of *GETAX* at the best position and expanding it until packing overlap in the crystal. ‡‡ As LS-S-H-3M, but using the mask from the correlation-dependent automatic mask-generation option of program *DM* (Cowtan, 1994).

At its best position, mask-1 was then expanded until it overlapped with its neighbours related by crystal symmetry. Subsequently, the phases were refined and extended to 3 Å resolution by NCS averaging within this mask, in conjunction with solvent flattening and histogram matching using *DM* (Cowtan, 1994). The resulting electron density was interpretable, revealing that the C_2 symmetry interrelated two trimers (there were, in fact, no dimers). The map improvement at the initial stage is demonstrated in Fig. 3. After structure refinement at 2.57 Å resolution, it could be shown that the best *GETAX* position was 5.2 Å away from the true centre of NCS, but only 0.6 Å away from the true axis. The orientation of the true axis deviated by 3° from the result of the self-rotation function.

3.2. Other examples

The use of *GETAX* was essential for the structure analysis of L-fucose isomerase from *Escherichia coli* (Seemann & Schulz, 1997). Starting from a poor-quality single isomorphous-replacement phase set at a resolution of 7.3 Å (figure of merit = 0.31 for the phases to 6 Å resolution), the position of a local twofold axis could be determined and used for establishing an interpretable map showing a hexamer with D_3 symmetry. The phases were then extended to high resolution using NCS averaging over symmetry D_3 together with solvent flattening and histogram matching.

The structure of rhamnulose aldolase from *E. coli* started from a single-site derivative at 6 Å resolution with a figure of merit of 0.26. Using the described method, a total of 20 monomers organized as two and a half D_4 octamers could be located in the asymmetric unit and eventually refined to high resolution (Krömer, Thoma, Vonrhein & Schulz, unpublished results).

3.3. Robustness against inaccuracy of phases and mask

The available data for the trimeric adenylate kinase from *S. acidocaldarius* were used to investigate the influence of low-quality input on the resulting correlation map (Table 1). Three phase sets were generated and used to produce electron-density maps at 6 Å resolution on a 2 Å grid. The highest peak of the self-rotation function was used. The phase sets were calculated from the worst heavy-atom derivative (1S), from the best heavy-atom derivative (2S) and from all three derivatives after refinement (3M). Mask-1 was used in the initial analysis and contained only 10% of the volume of the two trimers in the asymmetric unit (Fig. 2*b*). Mask-2, mask-3 and mask-4 were larger. In all cases, the correlation map yielded the correct position with excesses of up to 4.2σ above the first wrong peak, except for the combination of worst phases (1S) with the smallest mask-1 (Table 1). This demonstrates that *GETAX* is rather robust, despite its coarse grid spacing and interpolation.

3.4. Procedure after the initial localization of the NCS

As indicated in Fig. 1, we propose to expand the *GETAX* searching mask at its best position until packing overlap in the crystal, and to use this large simple multimer mask at the early stages of phase refinement/extension by NCS averaging/solvent flattening/histogram matching. In our hands, the method of automatic correlation-dependent mask generation (Vellieux *et al.*, 1995) as implemented in program *DM* (Cowtan, 1994) resulted in less accurate phases. The inferior quality was quantified by map correlations with the final model. The data for six tests are given in Table 2. Amongst these, only the large simple-mask procedure yielded an interpretable map. The same experience was encountered by Seemann & Schulz (1997) and by Krömer, Thoma, Vonrhein & Schulz (unpublished results). In Table 2 we also give the deviation of the refined NCS (result of program *DM*) from that of the final model, showing that the large simple mask fares better than the correlation-dependent mask.

3.5. Applicability of program *GETAX*

It is known that NCS occurs frequently in protein crystals (Wang & Janin, 1993; Kleywegt, 1996). In order to quantify this knowledge we used an automated procedure to search for NCS in the Protein Data Bank (Abola *et al.*, 1987). First, protein crystal structures with a minimum of 20 residues per chain were selected from the 7097 available entries (status at March 6, 1998), discarding NMR and DNA/RNA structures as well as pure models. The resulting subset of 5770 structures was then grouped, combining those with identical space groups, Z values (number of copies of most populous chain per unit cell), unit-cell volumes (± 3%) and unit-cell parameters (± 1%). Only one entry was taken from each group, reducing the set to 2982 entries, among which 1205 or 40% had multiple copies of a chain within the asymmetric unit, *i.e.* NCS, and a subgroup of 820 or 27% had proper rotation axes (Table 3). This large subgroup could have been tackled by *GETAX* if some initial phases were available.

A classification into C_n types showed a total number of 1347 local n-fold symmetries with a clear dominance of twofold axes. There were only 90 n-fold rotation axes with translations larger than 3 Å, which are unlikely to be recognized and located by *GETAX*. Taken together, these numbers demonstrate that *GETAX* is widely applicable.

Table 3
Frequency of NCS in the Protein Data Bank.

Total number of unique entries	2982
Entries with proper rotation axes†	820
Entries with other NCS	385
Number of rotation axes‡	
with n = 2	1134 (87)
with n = 3	79 (2)
with n = 4	10 (1)
with n = 5	28 (0)
with n = 6	0 (0)
with $n \geq 7$	6 (0)

† Proper rotation axes are defined as relating n monomers by rotation angles κ within 5° of 360°/n and translations smaller than 3 Å along the rotation axis with each other; they correspond to symmetry C_n. ‡ Given is the number of proper rotation axes, whereas the number of screw rotation axes is in parentheses. The screw rotations are defined in the same way as the proper rotations, but have translations larger than 3 Å.

3.6. Conclusions

A major advantage of *GETAX* is its speed, which allows for an extensive search with various orientations deduced from self-rotation maps. A 3° spacing appears to be suitable for that purpose. A further increase in speed can be achieved by diminishing the interrelated point sets to about 500 statistically selected positions, which usually suffice for a reliable correlation map. The program is available, either from the authors or as part of the *CCP*4 suite of programs (Collaborative Computational Project, Number 4, 1994). It is written in standard Fortran77 using subroutines from the *CCP*4 library. It can be used with all relevant space groups and with different orthogonalization conventions.

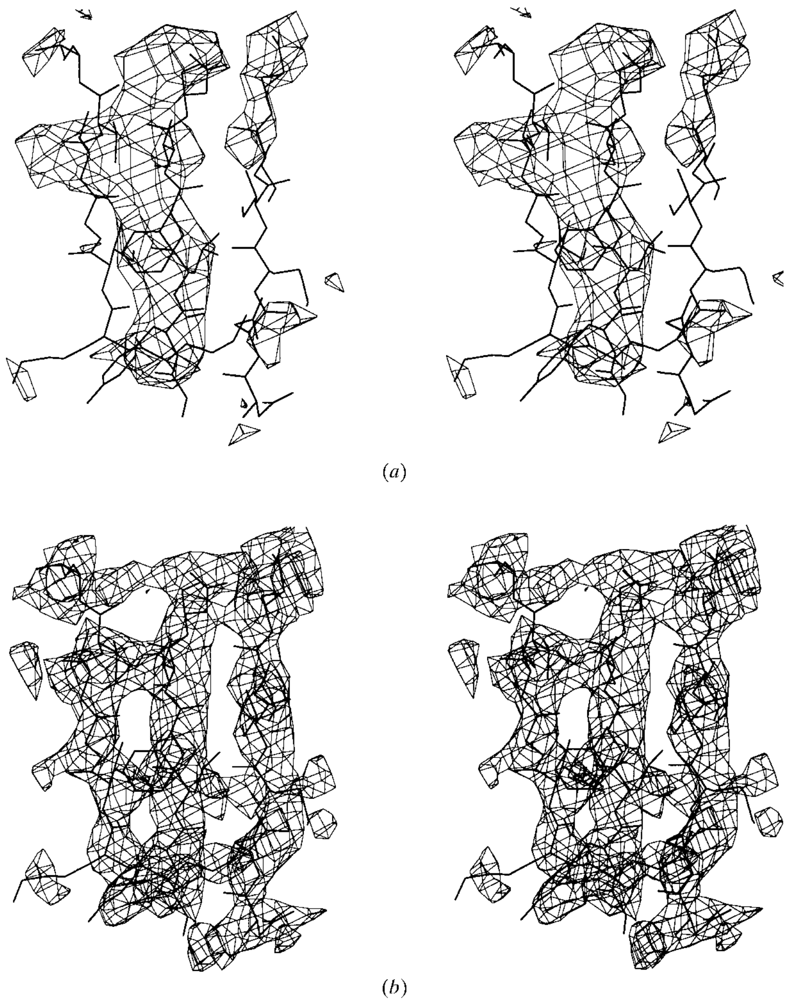


Figure 3
Electron density for strands β1, β2 and β3 of the central parallel sheet of subunit E of adenylate kinase from *S. acidocaldarius* together with the final model (Vonrhein *et al.*, 1998). (*a*) The initial 4.5 Å multiple isomorphous replacement map contoured at 1.2σ. (*b*) 3.0 Å density map contoured at 1.2σ. The phases were refined and extended by 100 cycles of solvent flattening, histogram matching and twofold averaging within the expanded simple multimer mask placed at the best position of *GETAX*.

The work was supported by the Deutsche Forschungsgemeinschaft under SFB-428.

References

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987). *Crystallographic Databases – Information Content, Software Systems, Scientific Applications*, edited by F. H. Allen, G. Bergerhoff & R. Sievers, pp. 107–132. Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography.

Bricogne, G. (1974). *Acta Cryst.* A**30**, 395–405.

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

Cowtan, K. D. (1994). *Jnt. CCP4 ESF-EACBM Newslett. Protein Crystallogr.* **31**, 34–38.

Eagles, P. A. M., Johnson, L. N., Joynson, M. A., McMurray, C. H. & Gutfreund, H. (1969). *J. Mol. Biol.* **45**, 533–544.

Epp, O., Steigemann, W., Formanek, H. & Huber, R. (1971). *Eur. J. Biochem.* **20**, 432–437.

Kleywegt, G. J. (1996). *Acta Cryst.* D**52**, 842–857.

Kleywegt, G. J. & Read, R. J. (1997). *Structure*, **5**, 1557–1569.

Lacher, K. & Schäfer, G. (1993). *Arch. Biochem. Biophys.* **302**, 391–397.

Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.

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

Rossmann, M. G., Blow, D. M., Harding, M. M. & Coller, E. (1964). *Acta Cryst.* **17**, 338–342.

Schirmer, T., Keller, T. A., Wang, Y.-F. & Rosenbusch, J. P. (1995). *Science*, **267**, 512–514.

Seemann, J. E. & Schulz, G. E. (1997). *J. Mol. Biol.* **273**, 256–268.

Stubbs, M. T., Nar, H., Löwe, J., Huber, R., Ladenstein, R., Spangfort, M. D. & Svensson, L. A. (1996). *Acta Cryst.* D**52**, 447–452.

Vellieux, F. M. D. A. P., Hunt, J. F., Roy, S. & Read, R. J. (1995). *J. Appl. Cryst.* **28**, 347–351.

Vonrhein, O., Bönisch, H., Schäfer, G. & Schulz, G. E. (1998). *J. Mol. Biol.* **282**, 167–179.

Wang, B. C. (1985). *Methods Enzymol.* **115**, 90–112.

Wang, X. & Janin, J. (1993). *Acta Cryst.* D**49**, 505–512.

Zhang, K. Y. J. & Main, P. (1990*a*). *Acta Cryst.* A**46**, 41–46.

Zhang, K. Y. J. & Main, P. (1990*b*). *Acta Cryst.* A**46**, 377–381.