Acta Cryst. (1996). D52, 923–936

Structure Determination of the Human Protective Protein: Twofold Averaging Reveals the Three-Dimensional Structure of a Domain Which was Entirely Absent in the Initial Model

GABBY RUDENKO,[a] ERIK BONTEN,[b] ALESSANDRA D'AZZO[b] AND WIM G. J. HOL[a,c]

[a]Department of Biological Structure, Biomolecular Structure Center, Box 357742, School of Medicine, University of Washington, Seattle, WA 98195, USA, [b]Department of Genetics, St Jude Children's Research Hospital, 332 North Lauderdale, Memphis, Tennessee 38105, USA, and [c]Howard Hughes Medical Institute, Box 357742, School of Medicine, University of Washington, Seattle, WA 98195, USA. E-mail: hol@gouda.bmsc.washington.edu

(Received 17 October 1995; accepted 9 April 1996)

Abstract

Mutations in the human 'protective protein' result in the human lysosomal storage disease galactosialidosis. The structure of the human 'protective protein' has been determined using X-ray crystallography to a resolution of 2.2 Å. Initial phases were obtained from molecular replacement calculations. A very partial search model comprising 30% of the scattering mass, was constructed from the atomic model of the wheat serine carboxypeptidase. This truncated probe was used to find the position of two monomers in the asymmetric unit. Subsequently, 'bootstrapping' cycles, consisting of twofold averaging and model expansion, retrieved the electron density for residues initially missing. In particular, it proved possible to add a domain (more than 110 residues) to the initial partial search model. In total, 314 residues per asymmetric unit were added to the 588 residues of the initial model. Factors contributing to our success are discussed.

1. Abbreviations

$|F_{\rm obs}|$, observed amplitude; $|F_{\rm calc}|$, amplitudes calculated from an atomic model; $|F_{\rm inv}|$, amplitudes calculated by map inversion; $\alpha_{\rm calc}$, phases calculated from an atomic model; $\alpha_{\rm inv}$, phases calculated by map inversion; R factor $= \sum \big| |F_{\rm obs}(h)| - |F_{\rm calc}(h)| \big| / \sum |F_{\rm obs}(h)|$; $\mathrm{CC} = \langle |F_{\rm obs}| \cdot |F_{\rm calc}| - \langle F_{\rm obs}\rangle\langle F_{\rm calc}\rangle\rangle \ / \ (\langle F^2_{\rm obs} - \langle F_{\rm obs}\rangle^2\rangle \times \langle F^2_{\rm calc} - \langle F_{\rm calc}\rangle^2\rangle)^{1/2}$ where the angular brackets denote an average over all selected Miller indices; NCS, non-crystallographic symmetry; r.m.s.d., root mean square deviation.

2. Introduction

The human 'protective protein', (HPP), also known as protective protein/cathepsin A (PPCA), forms a multienzyme complex with β-galactosidase and neuraminidase in the lysosomes protecting these two glycosidases from premature degradation (reviewed by d'Azzo, Andria, Strisciuglio & Galjaard, 1995; Okamura-Oho, Zhang & Callahan, 1994). Deficiencies of HPP in humans result in loss of lysosomal β-galactosidase and neuraminidase activity, leading to the lysosomal storage disease galactosialidosis (d'Azzo, Hoogeveen, Reuser, Robinson & Galjaard, 1982). The clinical symptoms of this hereditary disease are severe and include skeletal abnormalities, central nervous system involvement, short life-expectancy and in some cases mental retardation (see above mentioned reviews).

HPP is a multifunctional enzyme having in addition to the protective function, enzymatic activity as well. The precursor monomer consists of 452 residues with a molecular weight of 54 kDa (Galjart *et al.*, 1988). Structure determination of the dimeric form of HPP has revealed the presence of two glycosylation sites and four disulfide bridges (Rudenko, Bonten, d'Azzo & Hol, 1995). Conversion of the precursor to the mature form upon excision of a 2 kDa peptide from within the polypeptide releases serine carboxypeptidase activity identical to that of cathepsin A (Galjart *et al.*, 1988; Bonten *et al.*, 1995), hence the name protective protein/cathepsin A.

HPP has 30% sequence identity with the wheat and yeast serine carboxypeptidases (CPW and CPY) (Galjart *et al.*, 1988). The structures of the dimeric CPW and monomeric CPY have been solved to high resolution (Liao, Breddam, Sweet, Bullock & Remington, 1992; Endrizzi, Breddam & Remington, 1994). Sequence comparisons between HPP, CPW and CPY, as well as structural information from CPW and CPY, suggested that the central part of HPP comprising about 330 residues (1–180 and 301–452), would be structurally similar (Fig. 1). The HPP core is approximately 36% identical to CPW and 34% identical to CPY and belongs to the hydrolase fold family (Ollis *et al.*, 1992). CPW and CPY have, in addition to the central core domain, a second domain containing a three helical bundle. It was apparent that the HPP amino-acid sequence would contain a second (but different) domain as well, formed by a large insertion of about 120 residues (roughly residues 181–300). The sequence identity of this domain with other serine carboxypeptidases was estimated prior to the structure determination to be 10% or less, rendering prediction of the fold uncertain. In general, the

Acta Crystallographica Section D
ISSN 0907-4449 © 1996

relatively low sequence homology between HPP, CPY and CPW as well as the large insertion might normally have ruled out molecular replacement for the structure determination, making multi-isomorphous replacement the method of choice. However, severe difficulties were encountered in obtaining a sufficient number of crystals for a heavy-atom derivative search. For this reason we decided to attempt to solve the HPP precursor structure with the molecular-replacement method. Our search model, derived from the atomic coordinates of the wheat serine carboxypeptidase (Liao *et al.*, 1992), consisted of only 30% of the scattering mass in the asymmetric unit. This paper describes the successful application of the molecular-replacement technique using this model to find the positions of the two monomers in the asymmetric unit, thus roughly placing 60% of the scattering mass. We further demonstrate that starting from these very poor molecular-replacement phases, iterative cycles of twofold density averaging and model expansion, could retrieve the electron density for the remaining 40% of the scattering mass.

3. Experimental methods and results

3.1. *Native diffraction data*

The precursor form of HPP was produced in the baculovirus overexpression system (Bonten *et al.*, 1995; Rudenko, Bonten, Hol & d'Azzo, 1996). Crystals belong to the space group $P2_12_12$ with cell dimensions $a = 115.04$, $b = 148.11$ and $c = 80.97$ Å (Rudenko *et al.*, 1996). Native diffraction data extending to 2.0 Å resolution, were collected at the SSRL synchrotron (beamline 7-1) on a MAR image plate using cryocooling techniques (see Table 1). The V_m was calculated to be 3.2 Å^3 Da^{-1} for two monomers per asymmetric unit, giving an estimated solvent content of 62%.

3.2. *Molecular replacement*

3.2.1. *Search model.* Four search models were tried in the molecular-replacement calculations. Model *A*, a monomer comprising 375 out of a total of 412 residues for the CPW monomer and all CPW side chains; model *B*, the dimer of model *A*; model *C*, the complete CPY

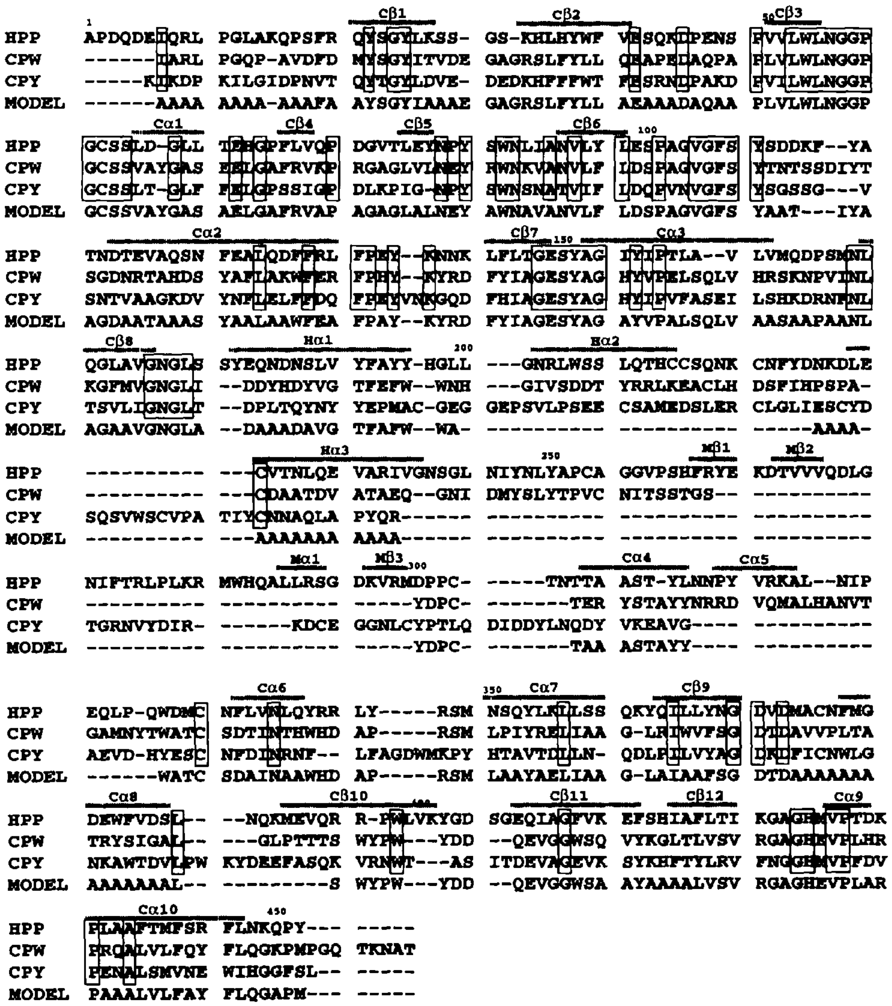

Fig. 1. Sequence alignment between HPP, CPW and CPY (top three sequences shown). Identical residues in all three sequences are boxed. Residue numbering is included for the HPP amino-acid sequence. The alignment was made using the *GCG* program *PILEUP* (*GCG* version 8), then manually adjusted using three-dimensional structural knowledge from the superposition of the CPW (Liao *et al.*, 1992) and CPY (Endrizzi *et al.*, 1994) structures. The multi-Ala search probe (model *D*) is shown in the fourth sequence indicated as 'MODEL'. The structure determination of HPP showed two domains: a 'core' domain (residues 1–182 and 303–452) and a 'cap' domain (residues 183–302). The secondary-structure elements for the HPP precursor are depicted with shaded bars (for details on the assignment and nomenclature, see Rudenko *et al.*, 1995).

Table 1. *X-ray data-collection statistics*

Resolution (Å)	32.27–2.2
Wavelength (Å)	1.08
Space group	$P2_12_12$
Unit cell (Å)	$a = 115.04$ $b = 148.11$ $c = 80.97$
Temperature of data collection (K)	95
No. of observed reflections	436709
No. of unique reflections	67740
Completeness of all data (%)	95.7
$R_{\rm sym}$* for all data (%)	5.1
Completeness of outer shell 2.26–2.20 Å (%)	87.0
$R_{\rm sym}$ in outer shell 2.26–2.20 Å (%)	13.0

* $R_{\rm sym} = \sum\sum |I_i(h) - \langle I(h)\rangle / \sum\sum I_i(h)$ where $I_i(h)$ is the ith observation for reflection h and $\langle I(h)\rangle$ is the weighted mean of all the observations.

monomer; and model *D*, a truncated monomer comprising a 'multi-Ala core'. The atomic *B* factors were set to the values obtained for the X-ray structure the search model was derived from.

The best molecular-replacement results were obtained using model *D* (the 'multi-Ala core') as a search probe and are described in this section. An evaluation of the degree of success using the other three search models is given later in the *Discussion* section.

The 'multi-Ala core' search model was constructed from the atomic coordinates of the CPW monomer (Liao *et al.*, 1992), based on the sequence alignment in Fig. 1. In designing the core model, regions expected to deviate in structure between HPP and CPW were deleted from the model (*i.e.* polypeptide stretches with low sequence identity or located in loops). The 125 residues identical in HPP and CPW were left in the model; 112 residues were truncated to alanine. The remaining 94 residues though differing between CPW and HPP, were considered sufficiently similar in size and the CPW was residue left as such in the model. The resulting 'multi-Ala core' monomer consisted of 331 residues, constituting a large portion of the core domain and little atomic information for the second domain (see Fig. 1). The model contained 30% of the expected protein scattering mass given the fact that there are two monomers in the asymmetric unit. The sequence identity between this search model and the true HPP structure is 37.7%.

3.2.2. *Rotation function, PC refinement and translation function.* Native data between 8 and 4 Å resolution was used in the molecular-replacement calculations. The rotational searches utilized a real-space Patterson search method, as implemented in *X-PLOR* (Steigeman, 1974; Huber, 1985; Brünger 1992*a*) with a Patterson vector cutoff of 21 Å. Unexpectedly, the self-rotation function failed to reveal any non-crystallographic twofold symmetry relating two monomers in the asymmetric unit. The native self Patterson maps showed no significant non-origin peaks and hence did not reveal the presence of a non-crystallographic twofold axis parallel to any crystallographic axis either. However, the crystal volume is such that one monomer per asymmetric unit would lead to 6.4 Å^3 Da^{-1}, which is most unlikely. Therefore, we proceeded to search for two monomers in the asymmetric unit.

The orientation of the two monomers in the asymmetric unit was determined by the cross-rotation function in the following way. Patterson vector sets were calculated for the search model and the native data, and the 8000 strongest Patterson vectors were used in the rotation function. The rotational space to be searched was restricted to the asymmetric unit of the rotation function according to Rao, Jih & Hartsuck (1980). This angular space was finely sampled by rotating the Patterson vectors from the search model around Eulerian angles $\theta1$, $\theta2$ and $\theta3$, using a constant angular grid interval of 2.5°. The 5000 highest rotation-function grid points, resulting from the product function of the two Patterson vector sets, were selected. Those grid points differing less than 8° around any given axis were clustered. The result was a list of 169 possible solutions for the rotation function. The two top solutions were 3.9 and 3.8σ above the mean, but hardly significant above the first error

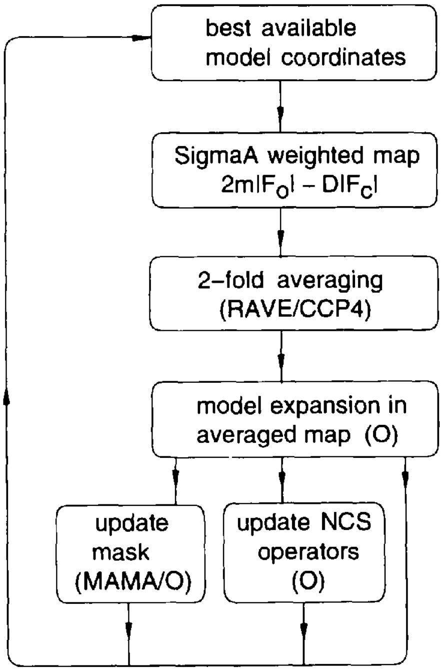
best available model coordinates
SigmaA weighted map 2m|F$_O$| - D|F$_C$|
2-fold averaging (RAVE/CCP4)
model expansion in averaged map (O)
update mask (MAMA/O)
update NCS operators (O)

Fig. 2. Steps incorporated in a 'bootstrapping' cycle. A *SigmaA*-weighted $2m|F_{\rm obs}| - D|F_{\rm calc}|$ map (Read, 1986) was calculated from the latest model. Twofold averaging with the program *RAVE* (Kleywegt & Jones, 1994) was subsequently used to reduce the phase error and to produce a better map allowing for manual model expansion. The twofold averaging was carried out along the guidelines provided by the *RAVE* documentation. In our case, a bootstrapping cycle generally entailed 13 steps of iterative averaging. In between steps, the averaged map was inverted to obtain $|F_{\rm inv}|$ and $\alpha_{\rm inv}$ and subsequently combined with $|F_{\rm obs}|$ to generate the $(2|F_{\rm obs}| - |F_{\rm inv}|){\rm exp}\alpha_{\rm inv}$ map for the next *RAVE* averaging step. The density outside the mask was set to an average value and no positivity constraints were applied inside the mask.

Table 2. *Course of model building*

Summary of the bootstrapping procedure. The resulting models have been listed chronologically starting with the molecular-replacement solution; *i.e.* mr (molecular replacement), bm (best monomer core), and the bootstrapping cycles bmc1 to bmc6. The following statistics are given for the various models: the number of C^{α} atoms built per monomer; the number of correct side chains incorporated per monomer and the volume of the molecular mask used during the averaging if applicable. The quality of each model is assessed using the *R* factor, $R_{\rm free}$, CC and $CC_{\rm free}$ calculated by *X-PLOR* for data between 8.0 and 3.0 Å. After positional refinement of model bmc3, both monomers were made equivalent by taking one monomer and generating the non-crystallographically related one.

Model	No. of C^{α}s	No. of side chains	$V_{\rm mask}$ (10^4 Å^3)	*R* factor	$R_{\rm free}$	CC	$CC_{\rm free}$
				(statistics using data between 8.0 and 3.0 Å)			
Molecular replacement (mr)	331	125	—	54.2	55.3	0.243	0.244
Rigid-body refinement (rmr)				52.6	52.9	0.287	0.318
Calculate NCS matrix							
Best monomer (bm)	294	228	—	55.9	57.4	0.228	0.216
Rigid body ref.				53.5	55.0	0.320	0.328
Update NCS matrix							
bmc1 (mask 1)	373	258	10.8	49.9	51.3	0.403	0.424
bmc2 (mask 1)	405	277	10.8	48.6	48.4	0.443	0.478
bmc3 (mask 2)	411	307	10.0	47.1	48.6	0.471	0.491
Rigid body refinement				46.9	48.4	0.476	0.492
Positional refinement (pbmc3)				39.4	44.7	0.622	0.562
Update NCS matrix							
bmc4 (mask 1)	412	327	10.8	41.7	43.1	0.584	0.585
bmc5 (mask 3)	435	387	8.9	39.8	40.6	0.621	0.623
bmc6 (mask 4)	442	413	9.1	38.4	40.2	0.647	0.637

peak of 3.7σ. PC refinement (Brünger, 1990) was carried out to optimize each of the 169 possible solutions using the complete search model as a single rigid body. This yielded two orientations with a PC index of 0.043 and 0.051, respectively (discussed later in Fig. 8, Tables 4 and 5). In contrast, the rest of the possible solutions yielded an average PC index of 0.022.

Individual translation-function calculations were performed on a 1 Å grid. A translational solution was found for each orientation, 7.7 and 8.8σ above the mean, respectively. The *R* factor for the individual solutions was 55.6 and 54.8% in the resolution range 8.0–4.0 Å, with correlation coefficients (CC) of 0.095 and 0.114. A combined translation function was calculated to place each solution relative to the same crystallographic origin, resulting in an *R* factor of 52.8% for data between 8.0 and 4.0 Å and a CC of 0.191. Rigid-body refinement using data between 8.0 and 4.0 Å, brought the *R* factor down to 51.3% and increased the CC to 0.22. The molecular packing was assessed on a graphics workstation and revealed no clashes between the placed search probes. However, a very large amount of empty space was present, where the search model was expected to be incomplete. The packing showed that the asymmetric unit contained two monomers and judging from their very extensive contacts, were associated to form a dimer related by a twofold with spherical polar angles $\varphi = 285$ and $\psi = 35°$. It is unclear why the self-rotation function was unsuccessful.

3.3. *Iterative model building and twofold averaging*

3.3.1. *Initial map.* A $2m|F_{\rm obs}| - D|F_{\rm calc}|$ *SigmaA*-weighted map (Read, 1986) was calculated using $|F_{\rm calc}|$'s and phases from the molecular-replacement solution. The map was contoured at 1σ and showed good density

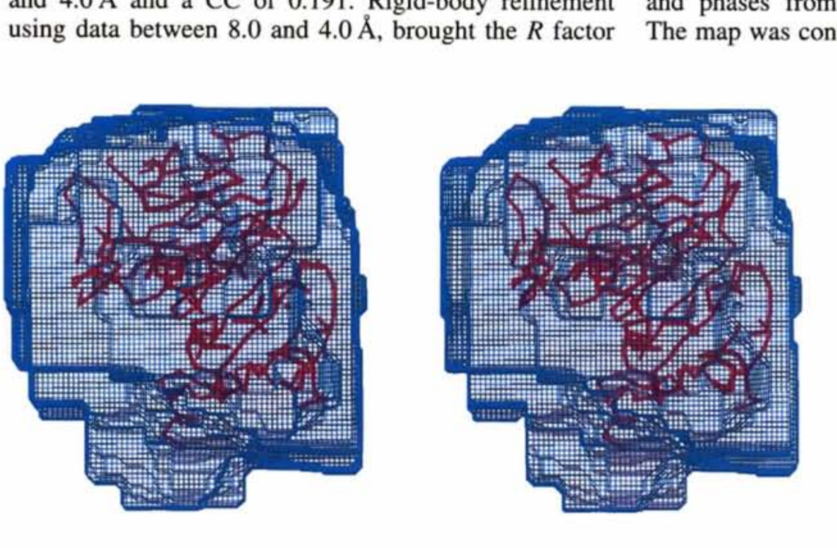

Fig. 3. Stereoview of mask 1, enlarged in areas where the bm model was thought to be incomplete. The program *MAMA* (Kleywegt & Jones, 1996) was used to calculate the mask and mask editing options in *O* (Jones, Zou *et al.*, 1991) were used to extend the mask. The Cα trace of model bm is shown in magenta.

for most of the core. Density emerged for many side chains where the input model residue had been an Ala, indicating that the molecular-replacement solution was correct.

3.3.2. *Initial model building.* The two rotated and translated search probes formed the starting point for model building of the HPP precursor. The non-crystallographic symmetry (NCS) matrix was determined between the two cores using the *lsq_explicit* option in *O* (Jones, Zou, Cowan & Kjeldgaard, 1991). Subsequently a 'best monomer' was built by superimposing the electron densities from each monomer core, and adjusting the model accordingly. Residues were only incorporated in the model where the electron density was visible for the complete side chain. Residues from the search model for which no density was visible were removed. An alanine was built in the model at places where electron density for a side chain was partial. In this manner 294 residues, *i.e.* 65% of the Cα atoms incorporated in the 'best monomer' core. The second monomer was generated from the 'best monomer' model using the NCS operator relating the two monomers in the asymmetric unit. At this point the data set was partitioned in a working set and a test set consisting of 5% of the reflections between 8 and 2.2 Å resolution, to monitor the $R_{\rm free}$ (Brünger, 1992*b*). The working data set was used for rigid-body and positional refinement. For averaging and map calculations the unpartitioned data set was used. 25 cycles of refinement using the two 'best monomer cores' positioned in the asymmetric unit as rigid bodies and data between 8.0 and 3.0 Å, resulted in an *R* factor of 53.5% for this resolution range. The atomic coordinates of this partial model were used to calculate a new $2m|F_{\rm obs}| - D|F_{\rm calc}|$ *SigmaA*-weighted map which was called the 'best monomer map'.

3.3.3. *Averaging: search for missing density.* The phasing power from the rigid-body refined 'best monomer cores', consisting of 294 residues per core was insufficient to bring back interpretable electron density for the missing part of the model, 158 residues per monomer. To overcome this, a 'bootstrapping' procedure was applied, entailing density averaging using *RAVE* (Kleywegt & Jones, 1994) and model expansion. The 'best monomer map' and the rigid-body refined 'best monomer cores' served as the starting point for this procedure.

Six bootstrapping cycles were carried out, called bmc1 to bmc6, permitting the model to be extended in stepwise increments. Fig. 2 shows a scheme of the steps incorporated in one bootstrapping cycle. After one cycle in which the model had undergone major expansion, a new molecular mask was calculated with *MAMA* (Kleywegt & Jones, 1996) for use in the subsequent bootstrapping cycle. No phase recombination was applied between bootstrapping cycles. At the end of each cycle the inverted phases $\alpha_{\rm inv}$ and inverted amplitudes $|F_{\rm inv}|$ were discarded. The NCS operator was re-optimized after cycle bmc3. The resolution range of the data included in the bootstrapping cycle started with 15–3.0 Å for bmc1 and was gradually extended to 15–2.7 Å in bmc6. To optimize the bootstrapping procedure (summarized in Table 2) consideration was given to the molecular mask, the model-building strategy and the refinement procedure.

3.3.4. *Molecular masks.* Four different masks were constructed in total. As expansion of the model progressed we were able to use increasingly more complete models for the mask calculations. Erring on the conservative side, the atomic radius of all atoms was set to 4 Å during the mask calculations by *MAMA*. The volume of each mask is given in Table 2. The masks were then manually modified using mask-editing options in *O* (Jones, Zou *et al.*, 1991). Mask 1 was constructed around the 'best monomer core'. The missing regions of the model (called 'insertions' here) were located by finding the termini of the loose polypeptide segments in the model. The amount of protein mass comprising each

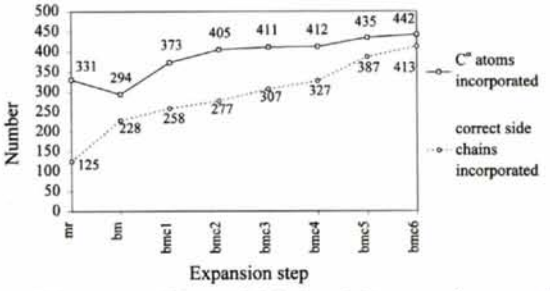

Fig. 4. Enlargement of the model during the bootstrapping procedure plotted as a function of the expansion step. The number of Cα atoms incorporated in the model per monomer is given (□) as well as the number of correct side chains (○). Note that after the first round of building in the molecular-replacement map (expansion step 'mr'), 37 residues had to be deleted from the model reducing the number of Cα atoms to 294 per subunit. Subsequent cycles permitted the model to be expanded by small increments.

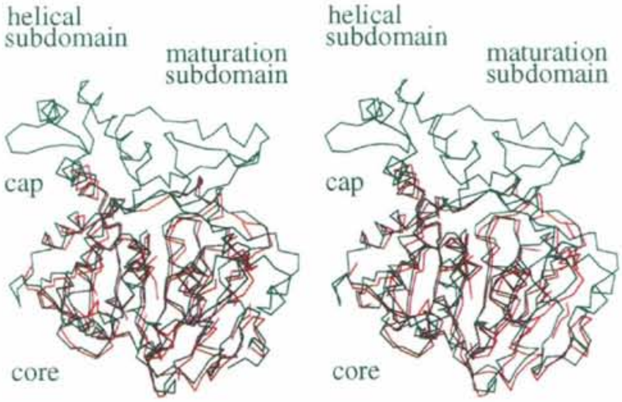

Fig. 5. Stereoview showing the comparison of the Cα trace from the best monomer core model bm (shown in magenta) and the complete HPP monomer (shown in green). The model bm contained only 294 Cα atoms compared with 452 residues for HPP. Domain and subdomain structures are labelled.

insertion was estimated from the amino-acid sequence. Subsequently the mask was greatly enlarged in these insertion regions by multiple blocks of 10–15 $Å^3$ (Fig. 3). This was crucial to prevent the density in these areas from being flattened during the averaging step. Approximately one half of the dimer interface was estimated to be formed by regions from the missing cap domain. Major expansions of the mask in this area were made to accommodate for this. This resulted in a serious overlap problem when the mask was duplicated to cover a complete dimer. The mask was reduced where overlap occurred with the *overlap_trim* option of *MAMA*. After several bootstrapping cycles, new incorporated polypeptide fragments were carefully assigned to one of the two

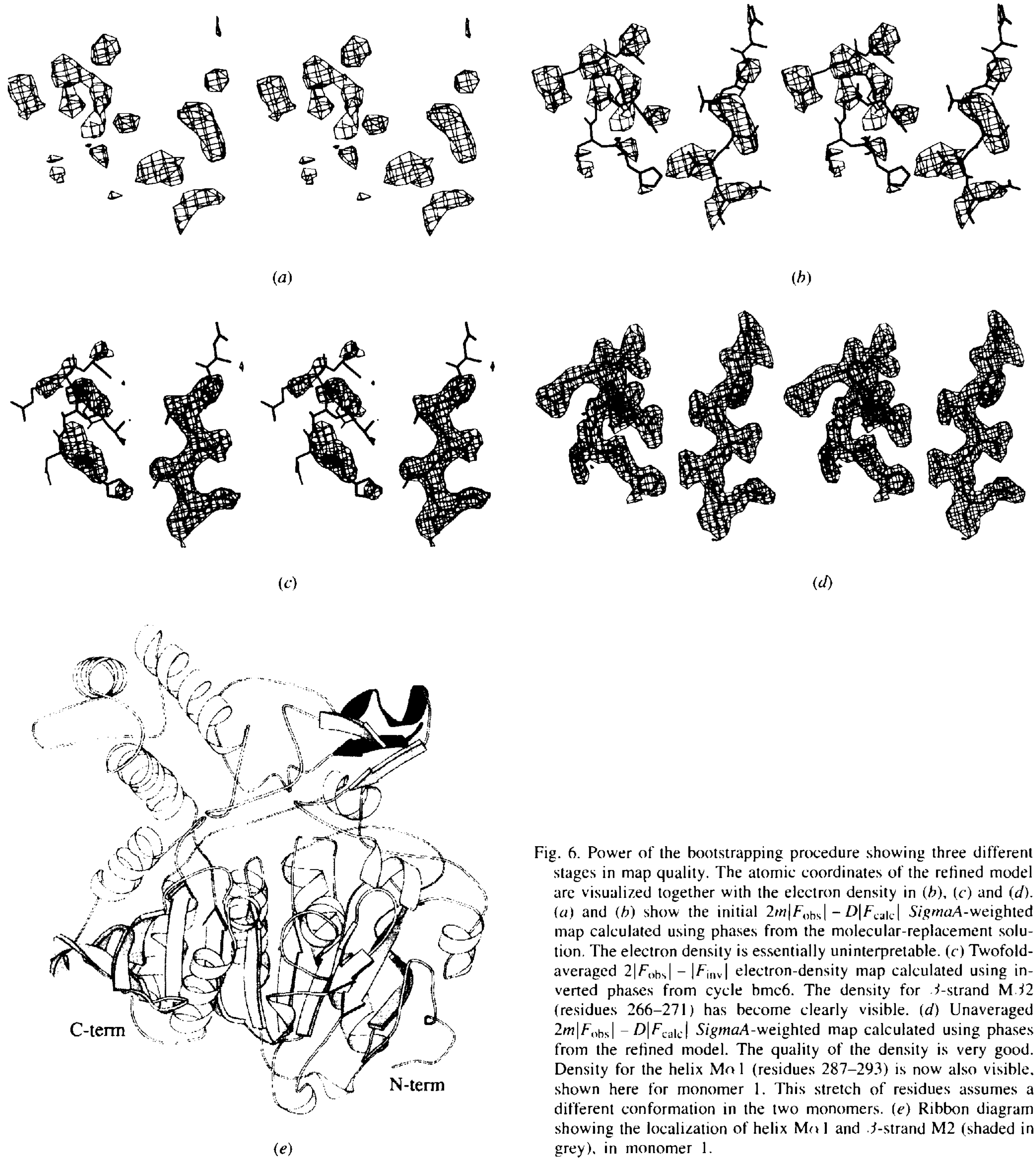

Fig. 6. Power of the bootstrapping procedure showing three different stages in map quality. The atomic coordinates of the refined model are visualized together with the electron density in (*b*), (*c*) and (*d*). (*a*) and (*b*) show the initial $2m|F_{\rm obs}| - D|F_{\rm calc}|$ *SigmaA*-weighted map calculated using phases from the molecular-replacement solution. The electron density is essentially uninterpretable. (*c*) Twofold-averaged $2|F_{\rm obs}| - |F_{\rm inv}|$ electron-density map calculated using inverted phases from cycle bmc6. The density for β-strand Mβ2 (residues 266–271) has become clearly visible. (*d*) Unaveraged $2m|F_{\rm obs}| - D|F_{\rm calc}|$ *SigmaA*-weighted map calculated using phases from the refined model. The quality of the density is very good. Density for the helix Mα1 (residues 287–293) is now also visible, shown here for monomer 1. This stretch of residues assumes a different conformation in the two monomers. (*e*) Ribbon diagram showing the localization of helix Mα1 and β-strand M2 (shaded in grey), in monomer 1.

monomers forming the dimer and the mask at the dimer interface was manually adjusted accordingly. Essentially the masks were kept far too large in regions where the model was missing in order to avoid erroneous flattening of electron density (see Fig. 3 and *Discussion*). In contrast the masks were tightened around the areas of the molecule where the model was complete.

3.3.5. *Model building.* A conservative model-building strategy was adopted. During the first six manual rebuilding cycles, side chains were mutated in the core region to fit the HPP amino-acid sequence. Where the density was clear for three or more consecutive residues in the insertion areas (loops and the cap domain), polyalanine fragments were built. Newly included atoms were given a B factor of 20 Å^2. Only once models bmc5 and bmc6 were obtained, was the electron density of sufficient quality to permit side chains for residues 190–303 to be incorporated confidently in the cap domain. At this stage the Cα trace was virtually complete for the whole dimer and the sequence could be fit unambiguously.

3.3.6. *Refinement.* Positional refinement of individual atoms was carried out only after three cycles of bootstrapping to a model containing 91% of the Cα atoms. 40 steps of positional refinement greatly improving the geometry of the model as well as the statistics between $|F_{\rm obs}|$ and $|F_{\rm calc}|$. Subsequently, only one of the refined monomers was taken and the other generated using updated NCS operators. The rationale for delaying the positional refinement is addressed in the *Discussion*.

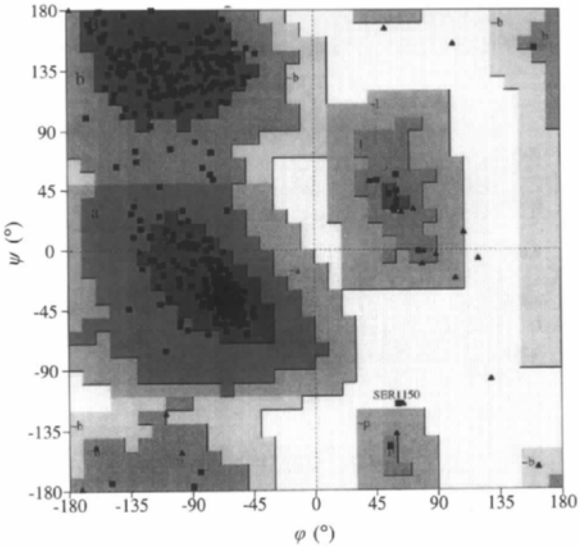

Fig. 7. Ramachandran plot calculated by *PROCHECK* (Laskowski, MacArthur, Moss & Thornton, 1993) for monomer 1 from the refined model of the HPP precursor. Both monomers in the asymmetric unit give essentially equivalent plots. Ser150 forms part of the catalytic triad. Glycine residues are shown as triangles; most favored regions [A, B, L], additional allowed regions [a, b, l, p], generously allowed regions [~a, ~b, ~l, ~p].

Table 3. *Current status of the model*

Statistics for the refinement (using 57 704 out of 66 191 reflections).

Resolution (Å)	R factor (%)	Cumulative completeness (%)
8.0–4.2	22.0	93.5
4.2–3.4	19.3	93.1
3.4–3.0	20.6	91.6
3.0–2.8	21.3	89.8
2.8–2.6	22.0	87.6
2.6–2.4	22.2	85.2
2.4–2.3	22.7	82.3
2.3–2.2	24.0	79.2
8.0–2.2	21.3	

Model	
Molecules in the asymmetric unit	2
Residues (out of 904 possible)	902
Sugars	6
Waters	296
R.m.s.d bond lengths (Å)	0.012
R.m.s.d. bond angles (°)	1.72
Average B values	
Main-chain atoms (Å^2)	16.6
Side-chain atoms (Å^2)	18.3

3.3.7. *Completing the model: deviations from twofold symmetry.* It was possible to add 148 residues and 185 side chains per monomer after a total of six bootstrapping cycles. At this stage, each subunit contained 442 residues and 413 side chains, *i.e.* 98% of the Cα atoms and 91% of the side-chain atoms. The gradual model expansion as a function of the bootstrapping cycle is shown in Fig. 4.

Ten residues were still missing per monomer at this stage. These were localized in two stretches per monomer (260–263 and 286–291). With most of the scattering mass incorporated, each monomer from model bmc6 was individually refined with *X-PLOR* (Brünger, 1992*a*) in an attempt to retrieve electron density for the still missing residues. After 40 steps of positional refinement using data from 8.0–2.6 Å and a full weight W_A on the crystallographic term, the R factor dropped significantly from 40.2 to 33.2%. The model was further positionally refined and data included in the refinement gradually extended to 2.2 Å resolution. At 2.4 Å resolution individual B factors were refined and the distribution checked as a function of atom location (*i.e.* low B factors in the core and high B factors on the surface). Cycles of refinement and refitting permitted 18 missing residues to be added. Essentially almost the complete cap domain was retrieved using the bootstrapping procedure, as shown in Fig. 5. It became apparent from the refined maps that the region around the two stretches of missing amino acids adopted very different conformations in the two monomers (with as much as an average r.m.s.d. of 5.6 Å for residues 281 to 295, and not less than 7.8 Å for the six Cα atoms of residues 287–292). For this reason electron density for these regions had not been retrieved in the twofold averaging process. The stepwise improvement of the electron-density maps as a result of averaging, model expansion and refinement is shown in Fig. 6.

The program *ARP* was used to check our model, in particular the area at the dimer interface (Lamzin & Wilson, 1993). Before the final round of positional refinement, an $|F_{\rm obs}|/\sigma$ cutoff was applied to reject about 10% of the weakest data and an anisotropic scale factor (with diagonal elements $B_{11} = -6.0$, $B_{22} = 2.0$, $B_{33} = 4.0$ calculated by *X-PLOR*) was applied to offset the decreased resolution along the crystallographic *a* axis. The final model is of good geometry with a final *R* factor of 21.3% ($R_{\rm free}$ of 26.8%) for data between 8.0 and 2.2 Å resolution (see Table 3). A Ramachandran plot is given in Fig. 7. The estimated r.m.s. coordinate error is 0.28 as calculated by *SigmaA* (Read 1986). The average phase difference between the initial molecular-replacement model and the currently refined model is calculated to be 71° for data between 10 and 2.2 Å resolution. The differences between the CPW and HPP structures are very large as described in Rudenko *et al.* (1995).

4. Discussion

The structure determination of HPP is special in that only twofold averaging could be applied to improve very poor molecular-replacement phases. Even though we had no experimental phases, we were able to retrieve electron density for 148 residues and 185 side chains per monomer missing from the initial model, by cycles of iterative averaging and model expansion. In total 314 complete residues were added per asymmetric unit, equivalent to about 35 kDa of protein. In retrospect we feel that a number of factors contributed to a successful structure determination.

4.1. *Molecular replacement and model choice*

The outcome of the molecular-replacement calculations depended heavily on the model used as a search probe. Four models were tried; the results are shown in Fig. 8, Tables 4 and 5. The clearest solutions were obtained using the multi-Ala core monomer. The CPW monomer (model *A*) and the multi-Ala core monomer

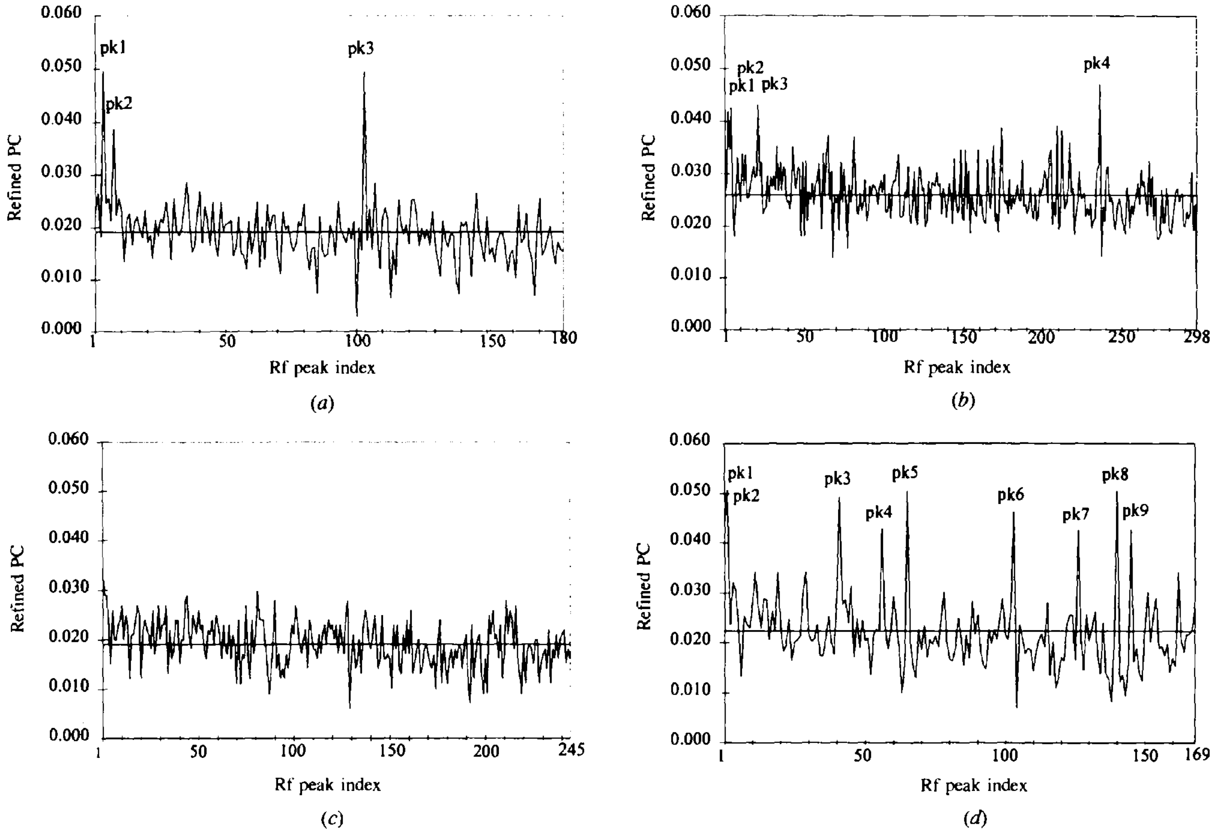

Fig. 8. Results from the rotation-function calculations and subsequent PC refinement obtained using the different models. The numerical values are given in Table 4. (*a*) Solutions obtained using the CPW monomer (model *A*). Three peaks were obtained, of which pk1 and pk2 are equivalent being related by a rotation of of 180°. The mean peak height for 180 solutions is 0.0192. (*b*) solutions obtained using the CPW dimer (model *B*). Peaks 1, 3 and 4 represent essentially one orientation. The mean peak height for 300 solutions is 0.0258. (*c*) No solutions were found using the CPY monomer as a search model (model *C*). (*d*) Multiple solutions were obtained using the multi-Ala search probe (model *D*). Peaks 2, 3, 5, 6 and 8 are equivalent, differing either less than 1° or are related by roughly 180°. Peaks 1, 4, 7 and 9 are related to each other as well. The mean peak height for 169 solutions is 0.0224.

Table 4. *Summary of the PC-refinement results*

Solutions for the rotation function after PC refinement as shown in Fig. 8. The set of orientations found for model *A* (CPW-mono) and model *D* (multi-Ala) differ by a κ value of 0.5 and 1°, respectively. The set of orientations found for model *B* (CPW dimer) which best resembled the set found for the multi-Ala model differed by a κ of 2.5 and 12°.

Rotational solutions

Model	Peak	Rotation function	After PC refinement	PC index ($\times 10^{-2}$)	σ/mean
A (CPW-mono)	1	(16.3 45.0 203.0)	(17.9 46.5 203.8)	4.96	2.6
	2	(31.0 45.0 17.7)	(17.9 46.5 23.8)	4.96	2.6
	3	(260.7 35.0 149.3)	(261.3 36.4 146.7)	3.88	2.6
B (CPW-dimer)	1	(251.0 45.0 157.7)	(250.9 45.7 157.8)	4.19	1.6
	2	(17.0 47.5 203.4)	(18.9 45.2 202.4)	4.27	1.7
	3	(260.1 42.5 156.3)	(249.0 43.3 160.1)	4.32	1.7
	4	(255.9 52.5 157.7)	(252.5 48.7 155.7)	4.70	1.8
C (CPY)	No solutions found				
D (multi-Ala)	1	(260.7 35.0 149.3)	(261.2 36.2 147.3)	4.29	1.9
	2	(20.2 47.5 20.2)	(18.5 47.4 23.2)	5.07	2.3
	3	(17.8 50.0 209.8)	(19.1 48.0 202.3)	4.94	2.2
	4	(251.5 27.5 156.2)	(261.3 36.2 147.2)	4.29	1.9
	5	(10.9 50.0 202.9)	(18.6 47.3 203.2)	5.06	2.3
	6	(30.3 45.0 10.3)	(20.2 47.7 23.7)	4.64	2.1
	7	(246.9 37.5 160.8)	(261.6 36.8 146.8)	4.25	1.9
	8	(27.9 47.5 21.4)	(18.5 47.4 23.2)	5.06	2.3
	9	(263.5 30.0 153.5)	(261.3 36.3 147.2)	4.29	1.9

Table 5. *Translation-function results using different models*

The peak number corresponds to the orientational solution as given in Fig. 8 and Table 4. For models *A* and *D*, the *R* factor and correlation coefficient (CC) is given for both the individual solutions containing one monomer per asymmetric unit, as well as for the combined solutions containing two monomers per asymmetric unit placed with respect to the same crystallographic origin.

Translation solutions

	Peak	Translation function*	Peak height	*R* factor	CC
Model	(No.)	(Å)	(σ/mean)	(8–4 Å)	(8–4 Å)
A (CPW-mono)	2	(33.3 52.0 12.8)	8.0	55.0	0.107
	3	(25.2 28.6 22.0)	7.3	55.4	0.097
	Combined translation function pk2 and pk3			53.6	0.175
B (CPW-dimer)	1	(19.2 44.2 11.4)	4.3	56.2	0.065
	2	(53.5 48.1 36.9)	4.4	55.7	0.078
	3	(12.1 57.2 14.2)	4.3	56.0	0.071
	4	(24.2 37.7 2.1)	4.7	56.0	0.078
C (CPY)	No solutions found				
D (multi-Ala)	1	(24.2 28.6 24.9)	7.7	55.6	0.095
	2	(32.5 54.6 13.5)	8.8	54.8	0.114
	Combined translation function pk1 and pk2			52.8	0.191

* The translational vectors for the different orientations are not necessarily with respect to the same crystallographic origin.

(model *D*) search probes yielded a very similar packing, placing 5954 and 4606 non-H atoms, respectively. But the initial *R* factor for the multi-Ala core packing was slightly lower, 52.8% for data between 8 and 4 Å resolution, compared to the *R* factor of 53.6% obtained for the CPW monomer packing; even though the latter contained roughly 30% more scattering mass.

The CPW dimer (model *B*) performed poorly compared with the multi-Ala core (model *D*). PC refinement attempted to optimize the model by altering the configuration of the CPW dimer, rotating one monomer with respect to the other by up to 3.5°. Shifts as great as 14 Å were introduced, pulling the dimer apart. The translation function using these PC-refined models did not give twice the signal obtained for the monomer search probe, as would be expected (see Table 5). Dimer solution pk1 generated a packing reminiscent of the multi-Ala/CPW monomer packing. Dimer solution pk4 placed one of the monomers in the dimer correctly, simultaneously mispositioning the other monomer. The dimer solutions pk2 and pk3 were grossly incorrect.

Quite surprisingly the CPY monomer (model *C*) failed completely to give any significant solutions for the rotation function even after PC refinement.

In retrospect, the three-dimensional structures of dimeric HPP, dimeric CPW and monomeric CPY can explain the differing degree of success of the search models while the core domains of CPW and CPY are essentially the same as their HPP counterpart sharing an r.m.s.d. of 1.2 Å for 293 and 271 Cα atoms, respectively; the cap domains between the three enzymes are very different (Rudenko *et al.*, 1995). HPP has an extra 49-residue 'maturation' subdomain and the configuration of the helices in the helical subdomain deviates significantly from the arrangements found in CPW and CPY. These helices superimpose on the CPW counterpart, for example, with an average r.m.s.d. of 3.7 Å for 40 Cα atoms (Rudenko *et al.*, 1995). Therefore, the multi-Ala model containing less atomic information for the cap domain and missing several deviating loops could be expected to be slightly more successful than the CPW monomer (model *A*) or the CPY monomer (model *C*), with such large differences in the cap domain apparent that it failed altogether as a search model, despite the very similar core domain.

The reason for the failure of the dimer *versus* the monomer search model is also clear. The configuration of the monomers forming the respective dimers is very different between HPP and CPW. When the core domains from one monomer in an HPP and CPW dimer are superimposed, the two other monomers making up the dimers are rotated by 15° with respect to each other (Rudenko *et al.*, 1995). PC refinement attempted to improve the dimer search model by twisting and shifting the monomers, but the differences were too great to find the correct solution.

Thus, a truncated search probe comprising less scattering mass but atoms positioned correctly gave a better signal-to-noise ratio for solutions, as compared to a larger model including incorrect structural elements and side chains (although the difference in *R* factor between the multi-Ala model *D* and the CPW model *A* is admittedly small). Care was taken to select the best possible search model to calculate molecular positions, as the quality if the initial phase set and NCS operators derived would determine if averaging could be used successfully to improve the phases or not.

4.2. *Bootstrapping procedure*

During bootstrapping various molecular masks were constructed with different goals in mind. Mask 1 encompassed a volume of 1.08×10^5 Å^3, and was designed to be as large as possible. The theoretical protein volume of the HPP monomer was calculated to be 6.17×10^4 Å^3 (Harpaz, Gerstein & Chothia, 1994). Calculating a mask for the final model using an atomic radius of 2 Å for all atoms gave a similar value of 6.82×10^4 Å^3. Using the theoretical protein volume, mask 1 was estimated to envelope 1.8 times the volume of the protein and reduced the solvent content from 63 to 37%. Mask 2 was 7.5% smaller than mask 1 (with a reduction of 8.1×10^3 Å^3 which is equivalent to 13% of the monomer volume). Mask 3 was tightened even more, being 17.6% smaller than mask 1 (with a reduction of 1.9×10^4 Å^3 or 31% of the monomer volume). The solvent content increased from 37 to 42% going from mask 1 to mask 2 and subsequently increased to 49% in mask 3. As bootstrapping progressed, the masks were tightened around the completed regions of the model, and enlarged in the areas where residues were still missing. The final mask, mask 4 (1.5 times the protein volume) was enlarged in an attempt to reveal density residues still obstinately missing, and was very carefully manually edited to correctly construct the dimer interface. Use of four different masks at the various stages of model building was most likely to be crucial to the success of the bootstrapping procedure.

The bootstrapping procedure resulted in a very gradual improvement of the model. On average 25 Cα atoms and 30 side chains were added per cycle (refer to Fig. 4). The most difficult task was building the polyalanine trace for the missing cap domain. The very large initial mask enabled us to localize the backbone of the missing cap domain, with little risk of flattening protein density. The maps from bmc4 permitted the Cα tracing to be virtually completed, and the directionality (N-term *versus* C-term) unambiguously determined by virtue of connections to helices. Almost miraculously the averaged map from bmc5 (using the tightest mask, mask 3), permitted a large stretch of 45 side chains to be incorporated in the cap domain, whereas the map from bmc4 (though revealing clues to which residues might have a small or a large side chain) definitely did not permit the amino-acid sequence to be incorporated. Apparently tightening the mask by 18% of the protein volume drastically improved the quality of the averaged map and permitted for most of the sequence of the cap domain to be unambiguously built in one sitting.

In lieu of experimental phases, the bootstrapping procedure proved particularly powerful in reducing model bias. It became clear during the first bootstrapping round that when incorrect atoms were incorporated in the model on purpose, the unaveraged electron density (either $2|F_{\rm obs} - |F_{\rm calc}|$ or $2m|F_{\rm obs}| - D|F_{\rm calc}|$) was heavily biased. However, the density around the incorrectly incorporated atoms was completely flattened by the iterative steps averaging and phase improvement that the electron-density maps were subjected to during a bootstrapping cycle. Two examples are given in Fig. 9. Therefore, to reduce the risk of model bias averaged maps were used as long as they enabled model expansion.

4.3. *Refinement*

It was a point of concern whether or not to incorporate positional refinement into the bootstrapping cycles, minimizing the difference between $|F_{\rm obs}|$ and $|F_{\rm calc}|$. At 3 Å resolution and higher, geometrical constraints would allow for sensible refinement. At low resolution (> 4 Å) however, the partiality of the model might introduce model bias as atoms might move to compensate for missing scattering mass. Positional refinement was, therefore, postponed until after cycle bmc3. The model was then 91% complete with respect to Cα atoms. Refinement of this model resulted in a tremendous improvement in statistics, greater than any step where the model had merely been expanded (see Table 2). The *R* factor dropped by 7.5% and the CC rose by 0.146. Curiously this had little effect on the electron-density maps calculated subsequently. The map contours were smoother, yet no extra density was retrieved for missing residues. The averaged maps improved greatly, independent of whether the initial map had been phased from the positionally refined, simulated annealed or unrefined bmc3 model. The number of missing residues which could be incorporated in the three different maps was virtually identical. Unaveraged *SigmaA*-weighted maps (contoured at 1σ) started to yield significant information on how to improve the model only after cycle bmc6. The model contained at this stage 98% of the Cα atoms and 91% of the side chains. Positional refinement of model bmc6 lowered the *R* factor to 31.6% (for data between 8 and 3 Å resolution). From this point onwards unaveraged maps were used and the two monomers were positionally refined and refitted independently.

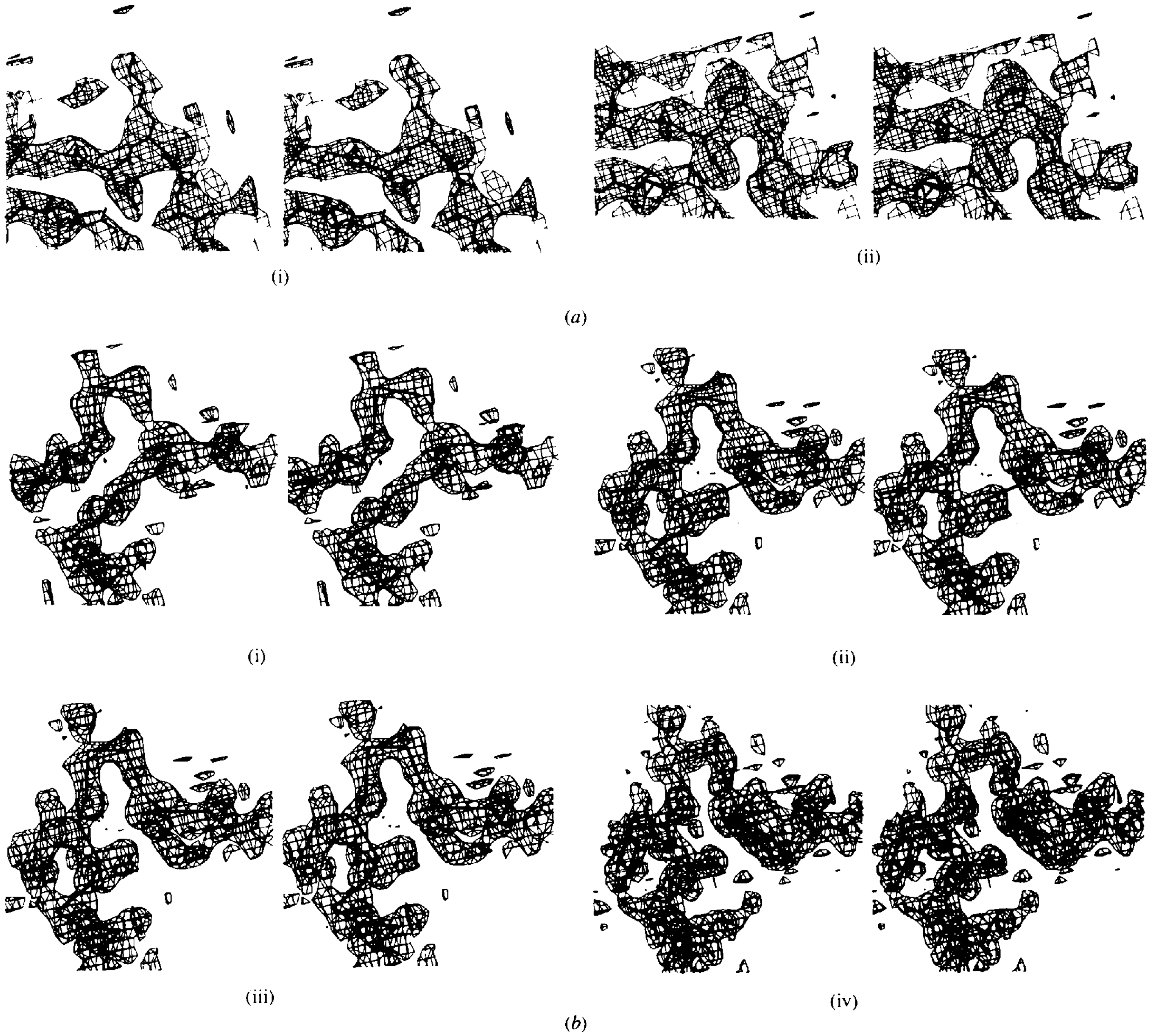

Fig. 9. Use of an averaging and phase-improvement cycle to remove model bias from electron-density maps. For the definition of the models depicted, see Table 2. (*a*) A lysine residue was incorrectly incorporated in model rbm. The subsequently calculated *SigmaA*-weighted map, contoured at 1.25σ, showed reasonable density for a lysine (i). Averaging this map showed unambiguous density for the correct residue, a proline (ii). (*b*) Model pbmc3 contained several polyalanine fragments. The *SigmaA*-weighted map, contoured at 1.25σ, followed the density reasonably well (i). Averaging this map flattened density for incorrect atoms, and revealed a serious error in the connectivities (ii). The trace could be corrected in the averaged map (iii), and later when the sequence was assigned it became clear from the *SigmaA*-weighted map calculated from the final model, contoured at 1.1σ, that we had been fooled into building a polyalanine fragment into electron density for an *N*-linked *N*-acetyl glucosamine moiety (iv).

4.4. *Assessment of model improvement*

Model improvement was followed using the R factor, R_{free}, CC and phase shift (Fig. 10). During the bootstrapping cycles, the R factor and the R_{free} gave very similar values; the same applied for the CC and the CC_{free} (Table 2). The most useful tool to monitor our progress during model expansion was the CC as a function of resolution. Being independent of the scale factor between $|F_{obs}|$ and $|F_{calc}|$ (unlike the R factor), the CC is particularly useful at low resolution where the scale factor is usually inaccurate, or in the case of a very partial or poor model with very inaccurate $|F_{calc}|$'s (Fujinaga & Read, 1987; Rossmann, 1988). Looking at Fig. 10(*b*), it appears that the initial molecular packing yielded useful information predominantly between 6 and 4 Å resolution. Manual rebuilding into model bm increased the R factor to 55.9%, while the CC decreased to 0.228 (see Table 2). By looking at the CC as a function of resolution however, it was clear that the correlation between our model and the diffraction data rose for data between 7.7 and 3.3 Å, decreasing only at higher resolution (< 3.3 Å) as the geometry of the model deteriorated somewhat upon manual rebuilding. It is also apparent from Figs. 10(*a*) and 10(*b*), that the bootstrapping procedure was a four-stage process consisting of the stages mr/rmr/bm, bmc1/bmc2/bmc3, pbmc3/bmc4/bmc5 and finally bmc6.

Beautiful examples are available in the literature where the limits have been stretched by crystallographers, eager to reveal the answer to biochemical questions hidden in the protein structure. Very partial mod-

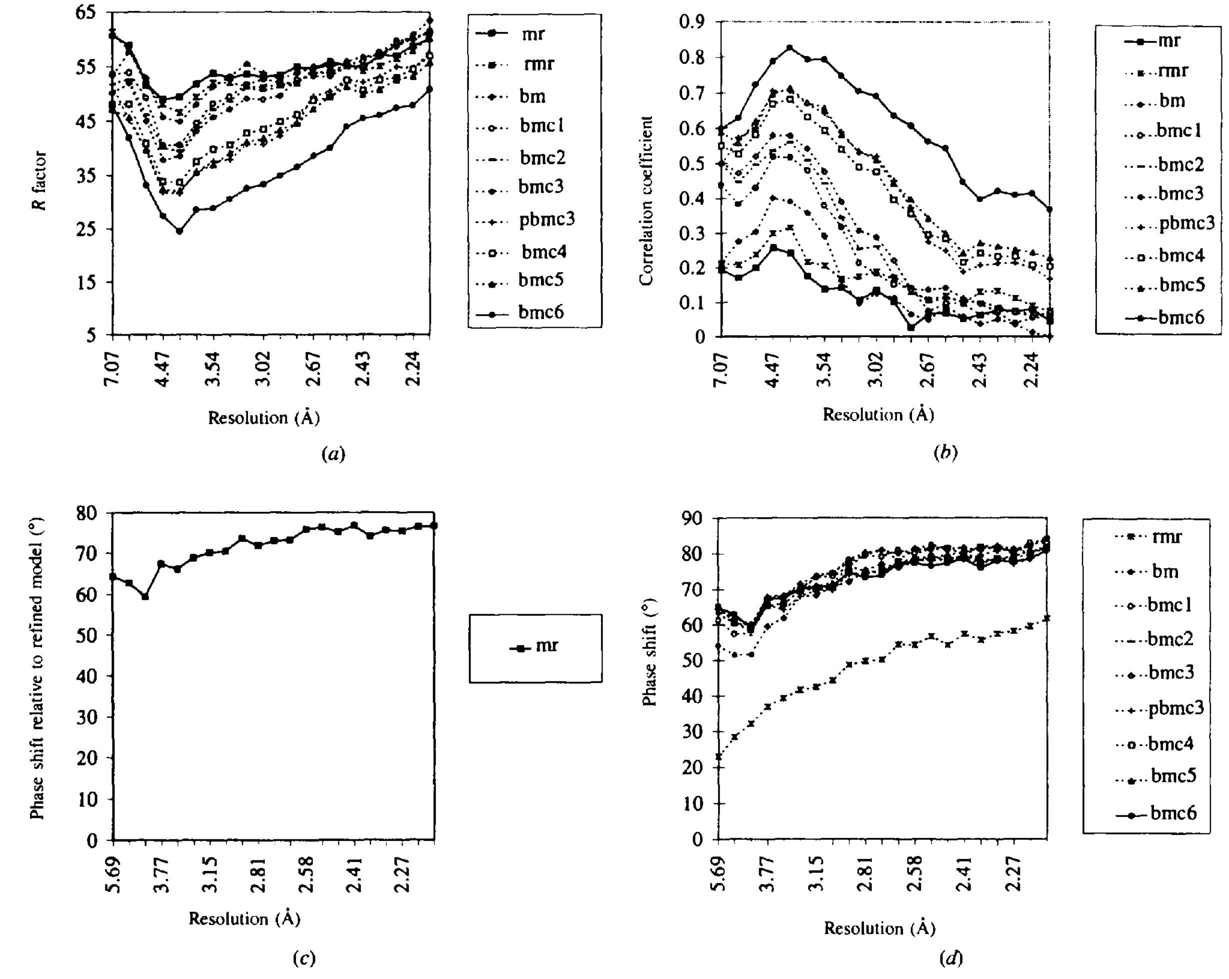

Fig. 10. Statistics monitored during the bootstrapping procedure. (*a*) R factor as a function of resolution, (*b*) CC as a function of resolution, (*c*) phase shift between the initial molecular-replacement phases and the refined model as a function of resolution, (*d*) phase shift between phases from the various models during the bootstrapping procedure and the initial set of molecular-replacement phases. Scaling of F_{calc} with respect to F_{obs} as well as the calculation of the R factor and the CC was carried out with *RSTATS* (Collaborative Computational Project, Number 4, 1994). Data was used between 10 and 2.2 Å resolution, binned in widths of $0.01(4 \sin^2\theta/\lambda^2)$. The average phase shifts were calculated using *SFTOOLS* (written by B. Hazes, *BIOMOL* suite, unpublished programs). Equalizing the two monomers after pbmc3 by using one monomer to generate the other one in the asymmetric unit caused the R factor to increase significantly again.

els have been used to solve the molecular-replacement problem (*e.g.* Oh, 1995; Pflugl *et al.*, 1993). Comprehensive studies have been carried out investigating the effect of various parameters on the success of the molecular-replacement calculation (Daresbury Proceedings on Molecular Replacement, Huber, 1985). Factors influencing the outcome of the calculations include: (*a*) the quality of the diffraction data in terms of accuracy, completeness, resolution; (*b*) the resolution range used; (*c*) the maximum and minimum length of the Patterson vectors included in the vector sets; (*d*) the quality of the search model (Rayment, 1983, as well as many excellent articles compiled in the Daresbury Proceedings on Molecular Replacement, Huber, 1985). There are also many cases where local symmetry has proven invaluable for the improvement of very poor initial phases. This is especially clear in the determination of virus structures possessing a high degree of local symmetry (Bloomer, Champness, Bricogne, Staden & Klug, 1978; Harrison, Olson, Schutt, Winkler & Bricogne, 1978; Unge *et al.*, 1980; Abad-Zapatero *et al.*, 1980; Valegård, Liljas, Fridborg & Unge, 1990). However, poor phase sets (usually from heavy-atom derivatives) have also been successfully improved and in some cases even extended, by exploiting density averaging with fewer local symmetry operators (Gaykema *et al.*, 1984; Gaykema, Volbeda & Hol, 1986; Jones, Walker & Stuart, 1991, sixfold; Tête-Favier, Rondeau, Podjarny & Moras, 1993, fourfold, and other articles compiled in the Daresbury Proceedings on Improving Protein Phases, Rossmann, 1988). It has been shown that the accuracy of the NCS operators, the molecular envelope, the quality of the data, the quality of the initial phase set as well as the redundancy of the local symmetry all determine the success to which molecular averaging can be applied (Jones, Walker *et al.*, 1991; Tête-Favier *et al.*, 1993, and articles compiled in the Daresbury Proceedings on Improving Protein Phases, Rossmann, 1988). In our case we were able to improve poor molecular-replacement phases derived from a partial model, by twofold molecular averaging. By paying careful attention to the molecular mask and a conservative model building strategy, missing electron density for 40% of the scattering mass could be retrieved.

We are deeply indebted to Drs G. Kleywegt and T. A. Jones for making the *RAVE* averaging software and documentation available prior to publication. Discussions with Dr K. Zhang on the weighted phase error are gratefully acknowledged. These studies were supported in part by the NIH Cancer Center Support CORE Grant P30-CA21765 and the American Lebanese Syrian Associated Charities (ALSAC). Support from Professor H. Galjaard and the Clinical Genetics Foundation Rotterdam in the initial stages of the structure determination has been invaluable. WH acknowledges receipt of a major equipment grant from the Murdock Charitable Trust. We are most grateful to SSRL for access to beamline 7-1 and for assistance given.*

* Atomic coordinates and structure factors have been deposited with the Protein Data Bank, Brookhaven National Laboratory (Reference: 1IVY, R1IVYSF). Free copies may be obtained through The Managing Editor, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England (Reference: AM0035). At the request of the authors, the atomic coordinates will remain privileged until 1 December 1996 and the structure factors will remain privileged until 1 December 1999.

References

Abad-Zapatero, C., Abdel-Meguid, S. S., Johnson, J. E., Leslie, A. G. W., Rayment, I., Rossmann, M. G., Suck, D. & Tsukihara, T. (1980). *Nature (London)*, **286**, 33–39.

d'Azzo, A., Andria, G., Strisciuglio, P. & Galjaard, H. (1995). *The Metabolic and Molecular Bases of Inherited Disease.* edited by C. R. Scriver, A. L. Beaudet, W. S. Sly & D. Valle, 7th. ed., ch. 91, pp. 2825–2837. New York: McGraw Hill Inc.

d'Azzo, A., Hoogeveen, A., Reuser, A. J. J., Robinson, D. & Galjaard, H. (1982). *Proc. Natl Acad. Sci. USA*, **79**, 4535–4539.

Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R. & Klug, A. (1978). *Nature (London)*, **276**, 362–368.

Bonten, E. J., Galjart, N. J., Willemsen, R., Usmany, M., Vlak, J. M. & d'Azzo, A. (1995). *J. Biol. Chem.* **270**, 26441–26445.

Brünger, A. T. (1990). *Acta Cryst.* **A46**, 46–57.

Brünger, A. T. (1992*a*). *X-PLOR version* 3.1. *A System for X-ray Crystallography and NMR*, Yale University Press, New Haven, CT, USA.

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

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* **D50**, 760–763.

Endrizzi, J. A., Breddam, K. & Remington, S. J. (1994). *Biochemistry*, **33**, 11106–11120.

Fujinaga, M. & Read, R. J. (1987). *J. Appl. Cryst.* **20**, 517–512.

Galjart, N. J., Gillemans, N., Harris, A., van der Horst, G. T. J., Verheijen, F. W., Galjaard, H. & d'Azzo, A. (1988). *Cell*, **54**, 755–764.

Gaykema, W. P. J., Hol, W. G. J., Vereijken, J. M., Soeter, N. M., Bak, H. J. & Beintema, J. J. (1984). *Nature (London)*, **309**, 23–29.

Gaykema, W. P. J., Volbeda, A. & Hol, W. G. J. (1986). *J. Mol. Biol.* **187**, 255–275.

Harpaz, Y., Gerstein, M. & Chothia, C. (1994). *Structure*, **2**, 641–649.

Harrison, S. C., Olson, A. J., Schutt, C. E., Winkler, F. K. & Bricogne, G. (1978). *Nature (London)*, **276**, 368-373.

Huber, R. (1985). *Molecular Replacement. Proceedings of the CCP4 Study Weekend, 15–16 February, 1985*, edited by P. A. Machin, pp. 58–61. Warrington: Daresbury Laboratory.

Jones, E. Y., Walker, N. P. C. & Stuart, D. I. (1991). *Acta Cryst.* A**47**, 753–770.

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

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

Kleywegt, G. & Jones, T. A. (1996). *Acta Cryst.* **D52**, 826–828.

Lamzin. V. S. & Wilson, K. S. (1993). *Acta Cryst.* **D49**, 129–147.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 282–291.

Liao. D. I., Breddam, K., Sweet, R. M., Bullock, T. & Remington, S. J. (1992). *Biochemistry*, **31**, 9796–9812.

Oh, B.-H. (1995). *Acta Cryst.* **D51**, 140–144.

Okamura-Oho, Y., Zhang, S. & Callahan, J. W. (1994). *Biochim. Biophys. Acta*, **1225**, 244–254.

Ollis, D. L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S. M., Harel, M., Remington, S. J., Silman, I., Schrag, J., Sussman, J. L., Verschueren, K. H. G. & Goldman, A. (1992). *Protein Eng.* **5**, 197–211.

Pflugl, G., Kallen, J., Schirmer, T., Jansonius, J. N., Zurini, M. G. & Walkinshaw, M. D. (1993). *Nature (London)*, **361**, 91–94

Rao, S. N., Jih, J.-H. & Hartsuck, J. A. (1980). *Acta Cryst.* **A36**, 878–884.

Rayment, I. (1983). *Acta Cryst.* **A39**, 102–116.

Read, R. J. (1986). *Acta. Cryst.* **A42**, 140–149.

Rossmann, M. G. (1988). *Improving Protein Phases. Proceedings of the CCP4 Study Weekend. 5–6 February 1988*, edited by S. Bailey, E. Dodson & S. Philips. Warrington: Daresbury Laboratory.

Rudenko, G., Bonten, E., d'Azzo, A. & Hol, W. G. J. (1995). *Structure*, **3**, 1249–1259.

Rudenko, G., Bonten, E., Hol, W. G. J. & d'Azzo, A. (1996). Submitted.

Steigeman, W. (1974). PhD thesis, Technical University Munich, Germany.

Tête-Favier, F., Rondeau, J.-M, Podjarny, A. & Moras, D. (1993). *Acta Cryst.* **D49**, 246–256.

Unge, T., Liljas, L., Strandberg, B., Vaara, I., Kannan, K. K., Fridborg, K., Nordman, C. E. & Lentz, P. J. Jr (1980). *Nature (London)*, **285**, 373–377.

Valegård, K., Liljas, L., Fridborg, K. & Unge, T (1990). *Nature (London)*, **345**, 36–41.