Acta Cryst. (1996). D**52**, 345–355

How to Escape from Model Bias with a High-Resolution Native Data Set – Structure Determination of the PcpA-S6 Subunit III

D. Pignol,[a] C. Gaboriaud,[a] J. C. Fontecilla-Camps,[a]* V. S. Lamzin[b] and K. S. Wilson[b]

[a]*Laboratoire de Cristallographie et de Cristallogénèse des Protéines, Institut de Biologie Structurale Jean-Pierre Ebel (CEA-CNRS), 41 Avenue des Martyrs, 38027 Grenoble CEDEX, France, and* [b]*European Molecular Biology Laboratory (EMBL), c/o DESY, Notkestrasse 85, D 22603 Hamburg 52, Germany*

(Received 5 May 1995; accepted 12 July 1995)

Abstract

The structure of procarboxypeptidase A-S6 subunit III, a truncated zymogen E, has been determined by molecular replacement using as search model porcine elastase 1 which, as revealed by crystallographic analysis, contained about 20% of the amino acids in a radically different orientation. Two monoclinic crystal forms were used: the first one diffracts to 2.3 Å resolution and contains one molecule per asymmetric unit; the second diffracts to 1.7 Å resolution and contains two molecules per asymmetric unit. Molecular replacement and conventional *X-PLOR* refinement led to a model for which 20% of the chain was ill defined in both crystal forms. To remove the bias introduced by the initial model, an automated refinement procedure [Lamzin & Wilson (1993). *Acta Cryst.* D49, 129–147] was applied successfully to the second crystal form, which diffracts to high resolution. The resulting dramatic improvement of the electron-density map led to extensive rebuilding of some surface loops. The reliability of the modified model was confirmed by refinement of the first crystal form. For the two forms, the final *R* factor is 18.8% for data between 8.0 and 2.0 Å resolution, and 18.4% for data between 8.0 and 1.7 Å, respectively.

1. Introduction

Subunit III is a defective serine endopeptidase-like pancreatic protein secreted as a complex with two zymogens, procarboxypeptidase A and a C-type chymotrypsinogen called proCPA-S6 (Keil-Dlouha, Puigserver, Marie & Keil, 1972). The presence of the ternary complex in the pancreatic juice has been tentatively associated with the protection, at least partially, of procarboxypeptidase A against the acidic conditions in the duodenum of ruminants (Kerfelec, Cambillau, Puigserver & Chapus, 1986). Subunit III is reversibly dissociated from the ternary complex under mild conditions (Puigserver & Desnuelle, 1975; Kerfelec, Chapus & Puigserver, 1984).

Subunit III is made up of a single polypeptide chain of 240 amino acids (Venot, Sciaky, Puigserver, Desnuelle & Laurent, 1986). The protein contains five disulfide bridges, four of which are strictly conserved among pancreatic serine endopeptidases. The remaining disulfide bridge (Cys98-Cys99b) is also present in protease E, a separate member of the elastase family (Cambillau, Kerfelec, Sciaky & Chapus, 1988). Indeed, analysis of structural data including amino-acid sequence comparison and disulfide bridge patterns, clearly shows that subunit III is closely related to porcine protease E (86% sequence identity). However, subunit III differs from protease E and from the other pancreatic serine endopeptidases in that its N-terminus is two residues shorter than that of the active enzymes. The absence of N-terminal residues Ile16/Val16-Val17, known to stabilize the active conformation of these enzymes by forming an ion pair between the N-terminal α-amino group and Asp194, makes subunit III a truncated protease E (Cambillau *et al.*, 1988). Thus, it is quite likely that the lack of specific activity of subunit III results mainly from a defective substrate-binding site.

We have recently reported the X-ray structure of subunit III refined at 1.7 Å (Pignol *et al.*, 1994). Here we focus on the strategy used to solve this structure by molecular replacement with a search model in which about 25% of the amino acids were in a radically different orientation.

2. Crystallization and data collection

Subunit III was purified after mild dissociation of the PcpA-S6 complex as previously reported (Kerfelec *et al.*, 1984). Two different crystal forms, here called forms 1 and 2, have been reported (Cambillau, Kerfelec, Foglizzo & Chapus, 1986; Abergel *et al.*, 1991). Both forms belong to the monoclinic space group $P2_1$ and were obtained using hanging-drop vapor diffusion (Wlodawer & Hodgson, 1975) at 293 K. Thin plate-like crystals of 'form 1' can be obtained at pH 4.2, with ammonium sulfate as a precipitant. They contain one molecule per asymmetric unit and diffract to a limit of 2.2 Å resolution at the synchrotron source at LURE. Large prismatic crystals of 'form 2' grow at pH 4.5 when using PEG 6000/NaCl as a precipitant. They contain two molecules per asymmetric unit and diffract to 1.7 Å resolution at LURE. Statistics characteristic of the two crystal forms are summarized in Table 1.

2.1. *Data collection for the first crystal form*

X-ray intensity data were collected on a MAR Research image plate at 275 K at LURE (station W32). Two crystals were used to collect two sets of 50 images each with a wavelength of 0.98 Å, a crystal-to-film distance of 200 mm, and an oscillation range of 1.5° per image. Data were reduced with the *MOSFLM* film-processing software (Leslie, Brick & Wonacott, 1986) using profile fitting and background averaging. The internal scaling was carried out after post-refinement using the programs *ROTAVATA* and *AGROVATA* from the *CCP*4 package (Collaborative Computional Project, Number 4, 1994).

A total of 12 890 observations of 5465 unique reflections were extracted from the first data set. The resulting $R_{\rm sym}$ [calculated as $\sum I(i) - \langle I \rangle / \sum I(i)$] was 0.058. For the second data set, 5582 reflections were obtained from 14 890 observations collected and the resulting $R_{\rm sym}$ was 0.098. The two data sets were subsequently merged together to give an overall data set of 9937 reflections extending to 2.2 Å resolution, with a $R_{\rm merge}$ of 0.109. The completeness was 92.3% for the data between infinity and 2.4 Å and 48.6% in the 2.4–2.2 Å resolution range.

2.2. *Data collection for the second crystal form*

A total of 70 images was collected at LURE (station W32) from a single crystal on a MAR Research image plate at 275 K to a maximum resolution of 1.7 Å. The wavelength was 0.98 Å, the crystal-to-film distance was 150 mm, and the oscillation range was 1.5° per image. Processing of 124 545 observations with profile fitting and background averaging resulted in a data set containing 40 074 unique reflections. Subsequently, this data set was scaled with *ROTAVATA* and *AGROVATA* to an $R_{\rm sym}$ of 0.048. About 89% of the reflections had intensities $I > 2\sigma(I)$, while in the outermost shell (resolution 1.73–1.69 Å) the proportion with $I > 2\sigma(I)$ was still 71.2%.

A second, lower resolution, data set was collected on a Xentronics/Siemens area detector coupled to a Rigaku RU 200 X-ray generator equipped with a graphite monochromator and a 3 × 0.3 mm focal spot size. Data were processed with the *XENGEN* set of programs (Howard *et al.*, 1987), and scaled using *ROTAVATA* and *AGROVATA*. From a total of 28 030 reflections measured to 2.0 Å resolution, 18 264 unique reflections were obtained with a $R_{\rm sym}$ value of 0.049. Finally, the rotating anode and synchrotron data sets were scaled together with *PROTEIN* (Steigemann, 1974), with an $R_{\rm merge}$ of 0.090. A final data set was obtained by completing the low-resolution data missing from the data set collected at LURE, due to count saturation, with those of the scaled rotating anode set using an in-house program (D. Housset, personal communication). The final data set consisted of 45 121 unique reflections, representing 92% of the theoretical data to 1.7 Å resolution.

Table 1. *Crystallization and crystal data of the two crystal forms of subunit III*

	Form 1	Form 2
Crystallization conditions		
Buffer	20 m*M* ammonium acetate pH = 4.2	100 m*M* ammonium acetate pH = 4.5
Precipitant	Saturated ammonium sulfate 20%	PEG 6000 25%(*w*/*v*) NaCl 8%(*w*/*v*)
Protein concentration (mg ml^{-1})	5	10
Temperature (K)	293	293
Crystal data		
Space group	$P2_1$	$P2_1$
Cell parameters (Å, °)	$a = 47.9$ $b = 61.1$ $c = 38.8$ $\beta = 95.0$	$a = 49.3$ $b = 82.7$ $c = 58.7$ $\beta = 97.9$
Number of molecules per asymmetric unit	1	2
Crystal size (mm)	0.4 × 0.2 × 0.05	0.7 × 0.7 × 0.5
Resolution limit (Å)	2.2	1.7

3. Structure solution and conventional refinement of the first crystal form

Comparisons of serine proteinase amino-acid sequences indicated that the three-dimensional structure of porcine elastase 1 was likely to be very close to that of bovine subunit III (56% identity). The only possible exception was the additional disulfide bridge between Cys98 and Cys99b in the latter. Porcine elastase 1 was, therefore, considered to be the most appropriate search model for molecular replacement.

3.1. *An unambiguous molecular-replacement solution*

AMoRe (Navaza, 1987, 1990), an integrated computer-program package of molecular-replacement programs, was used to solve the structure of subunit III. The model of porcine elastase 1 (Meyer, Cole & Radhakrishnan, 1988), refined to an *R* value ($\sum||F_o| - |F_c||/\sum|F_o|$) of 0.169 at 1.65 Å resolution, PDB (Bernstein *et al.*, 1977) code 3EST was used as a search model. The calcium ion and all water molecules were removed. The rotation search was carried out from 15 to 3 Å resolution. Table 2(*a*) shows the correlation values for different orientations of the model. The first rotation solution appears to be the correct one: its correlation value was 0.165, as compared with a value of 0.065 for the second solution.

Translation-function searches were carried out at the same resolution range, for the first rotation solution. Table 2(*b*) shows the correlation and *R*-factor values obtained for the different translation vectors with the model in this orientation. This procedure gave an unambiguous solution that was subsequently confirmed by using the rigid-body refinement procedure of *AMoRe* (Castellano,

Table 2. *Details on the rotation and translation functions for the first crystal form*

The three angles (α, β, γ) and the three translations (dx, dy and dz) are given for the best four solutions, selected according to the R factor (R) and the correlation value (C) (Navaza, 1994).

(*a*) Ro'ation functions

	α	β	γ	C
1	223.66	56.9	143.5	0.165
2	214.2	54.97	198.48	0.065
3	108.6	45.00	48.74	0.064
4	272.5	36.46	355.6	0.062

(*b*) Translation functions

	α	β	γ	dx	dy	dz	C	R
1	223.66	56.9	143.5	0.3	0.0	0.24	0.409	0.469
2	—	—	—	0.46	0.0	0.44	0.328	0.499
3	—	—	—	0.12	0.0	0.42	0.323	0.497
4	—	—	—	0.19	0.0	0.04	0.323	0.498

Olivia & Navaza, 1992). The initial model, oriented and positioned according to the molecular-replacement solution, was displayed on a graphics system with the program *O* (Jones, Zou, Cowan & Kjeldgaard, 1991) in order to check the packing contacts in the crystal lattice.

3.2. *A conventional refinement procedure*

The initial crystallographic R factor of the model, oriented and positioned according to the molecular-replacement solution, was 0.46 for data between 15.0 and 3.0 Å. Refinement was performed using the simulated-annealing and energy-minimization protocols of *X-PLOR* (Brünger, Kuriyan & Karplus, 1987).

A standard slow-cool procedure going from 3000 to 400 K with data between 8.0 to 2.2 Å lowered the R factor to a value of 0.32 The structure was subjected to several rounds of restrained refinement and individual B-factor refinement, followed by several cycles of rebuilding and inspection of electron-density maps using computer graphics. At the end of this stage, the R factor was 0.26, for a model without water molecules and similar to the starting elastase 1 structure. This model will be called 'form 1' in the following discussion.

3.3. *A disordered molecule?*

After the last refinement cycle, more than 75% of the model had well matching density in the 2.2 Å resolution electron-density map. The electron-density map was poor for the first 12 N-terminal residues, residues 130–145 (called the autolysis loop in the following), and the activation domain (residues 180–200) (Fig. 1*a*).

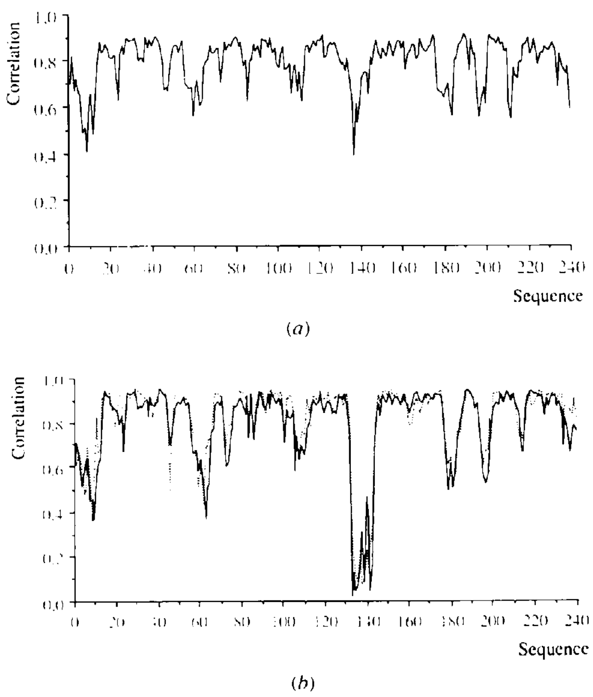
At this stage, the refinement procedure appeared to have converged. As we were quite confident that the molecular-replacement solution was correct, we attributed the poor match to the electron density of several parts of the molecule to disorder. These regions are known to be partially or completely disordered in trypsinogen (Fehlammer, Bode & Huber, 1977) and chymotrypsinogen A (Wang, Bode & Huber, 1985). It was clear, however, that the apparent disorder might also indicate misinterpretation of the electron density. In order to clarify this point we turned to the second crystal form.

4. Structure solution and conventional refinement of the second crystal form

4.1. *True or false molecular-replacement solution?*

The 'form 1' structure was used as a search model in the molecular-replacement procedure using *AMoRe*. The rotation search was carried out with data between 15 and 3 Å resolution. Two major solutions, corresponding to the two molecules in the asymmetric unit, appeared with correlation values of 0.206 and 0.122, respectively (Table 3*a*). These two orientations were consistent with the solution of the self-Patterson function.

One-body translation searches were carried out in the same resolution range. For each of the two orientations depicted in Table 3(*a*) one major translation solution was found, with a better contrast for the first orientation. The correlation values for the two positions were 0.308 and 0.221, respectively (Table 3*b*). In order to find the relative position of these two solutions, a two-body translation search was carried out. The positions of the two molecules of the asymmetric unit were then refined with the rigid-body option of *AMoRe* (Castellano *et al.*,



(*a*)

(*b*)

Fig. 1. Correlation between the ($2F_o - F_c$) electron-density map and (*a*) the form 1 model (*b*) molecule *A* and *B* (dashed line) of the initial form 2 model. The correlation is calculated by the program *O*, including main-chain and side-chain atoms.

Table 3. *Details of (a) the rotation and (b) translation functions for the second crystal form.*

(*a*) Rotation functions

	α	β	γ	C
1	316.15	78.64	92.28	0.206
2	89.79	48.61	161.27	0.122
3	228.14	15.52	261.33	0.062
4	224.22	15.32	268.30	0.062

(*b*) One body translation functions

	α	β	γ	dx	dy	dz	C	R
1	316.15	78.64	92.28	0.036	0.0	0.078	0.308	0.515
2	—	—	—	0.190	0.0	0.259	0.244	0.536
3	89.79	48.61	161.27	0.101	0.0	0.019	0.221	0.548
4	—	—	—	0.009	0.0	0.281	0.198	0.557

(*c*) Two body translation function

	α	β	γ	dx	dy	dz	C	R
	316.15	78.64	92.28	0.036	0.0	0.078		
1 + 2	89.79	48.61	161.27	0.597	0.449	0.524	0.442	0.471

1992). As a result, the R factor and the correlation values were 0.471 and 0.442, respectively.

The initial model, oriented and positioned according to the molecular-replacement solution, was displayed on a graphics system with the program *O*. Although the rotation and translation functions led to one unambiguous solution, the two molecules of the asymmetric unit showed extensive bad contacts resulting from the interpenetration of their respective autolysis loops (residues 130–145). Fig. 2 displays the two molecules in the asymmetric unit according to the molecular-replacement solution.

As an additional check, the molecular-replacement procedure was carried out, in the same way, using the structure of porcine elastase 1 as the starting model. This procedure led to the same solutions as when the 'form 1' structure was used.

4.2. *A useful heavy-atom derivative*

In order to obtain an unbiased source of phases, a systematic search for heavy-atom derivatives was begun. A problem of lack of isomorphism between native data sets (unpublished observations) made the search especially laborious. The only useful heavy-atom derivative was obtained by soaking a crystal for 24 h in a solution containing 0.5 m*M* K_2PtCl_4, 28%(*w*/*v*) PEG 6000 and 0.1 *M* ammonium acetate, pH 4.5. Data for this crystal were collected with a Xentronics/Siemens area detector at 275 K and processed with the *XENGEN* package. 35 158 observations were collected for a final data set of 12 493 reflections ($R_{\rm sym} = 0.080$), representing 78% of the possible reflections to 2.4 Å resolution. Scale and temperature factors were applied to the derivative data resulting in a fractional change $[\sum(F_{\rm PH} - F_P)/\sum F_P]$ of 0.29 between 20 and 3.5 Å resolution. Lack-of-isomorphism appears at 3.8 Å resolution as indicated by the increase of the fractional change, but it was considered acceptable up to 3.5 Å. A difference Patterson map using $(F_{\rm PH} - F_P)^2$ coefficients was calculated using different resolution ranges. A large peak corresponding to one heavy-atom site was identified in the Harker section ($v = 1/2$). The heavy-atom site was then refined with centric reflections, using the program *REFINE*, as implemented in the *CCP*4 package (Table 4*a*). The y coordinate was arbitrarily fixed to a value of 0.25.

Table 4. *(a) Statistics on the refinement of the platinium coordinates (x, y, z) and (b) the coordinates of the highest peak in the cross Fourier corresponding to the origin-translated position of the heavy-atom refined coordinates*

	x	y	z	Figure-of-merit	Rc*	Phasing† power
(*a*) Peak of the Patterson difference $(F_{\rm PH} - F_{\rm P})^2$	0.89	0.25	0.90	0.39	0.55	1.4
(*b*) Peak of the Fourier $(F_{\rm PH} - F_{\rm P})\exp(i\varphi{\rm calc})$	0.9	0.175	0.916			

* $Rc = \sum|F_{{\rm PH}_{\rm calc}} - F_{{\rm PH}_{\rm calc}}|/\sum|F_{{\rm PH}_{\rm obs}}|$, the Cullis R factor for centric reflections. † The phasing power is defined as F_H/E were F_H is the heavy-atom structure-factor amplitude and E the r.m.s. residual lack of closure.

In order to confirm the molecular-replacement solution, structure factors and phases of the model ($F_{\rm calc}$ and $\varphi_{\rm calc}$) were calculated with *X-PLOR*. A difference Fourier map using $(F_{\rm PH} - F_P)$ coefficients and $\varphi_{\rm calc}$ phases lead to a peak corresponding to an origin-translated position of the heavy-atom refined coordinates (Table 4*b*). This result confirmed the correctness of the molecular-replacement solution.

4.3. *Conventional refinement procedure*

The starting model, oriented and positioned according to the molecular-replacement solution, gave an R-factor value of 0.48 for data between 15.0 and 2.0 Å. A standard 'slow-cooling' procedure going from 4000 to 400 K' for data between 8.0 and 2.0 Å resulted in a model with an R value of 0.32. An identical refinement procedure was carried out with a model in which the autolysis loops of the two molecules of the asymmetric unit were deleted. In this case the resulting R factor was 0.31.

Inspections of the $(2F_o - F_c)$ and $(F_o - F_c)$ electron-density maps did not furnish any clear information for rebuilding the problematic loops. Subsequently, the model was subjected to several rounds of restrained refinement completed with individual B-factor refinement and several parts of the molecule were manually corrected. In addition, *SIGMAA* maps (Read, 1986) and simulated-annealing omit maps were calculated in order to decrease the model bias. All these approaches were without success. Possibly because of the lack-of-isomorphism, combination of the model phases with those of the platinum derivative calculated to 3.5 Å resolution did not improve the electron density corresponding to the unmodelled loops.

At the end of this procedure, the R factor was 0.34 using data to 1.7 Å resolution. Although about 80% of the model was well defined in the electron density, there remained four ill defined regions (Fig. 1*b*): (1) the first ten N-terminal residues; (2) residues 58–70, called the calcium binding-site loop in the following; (3) the autolysis loop (from 130 to 145); and (4) the activation domain (from 189 to 195). This model will be called here 'form 2'; it is similar to 'form 1' with the ill defined regions being equivalent in both forms. Moreover, the Ramachandran plots reveal that 30% of dihedral angles correspond to energetically unfavorable regions in both models (Ramakrishnan & Ramachandran, 1965).

5. The automated refinement procedure

Since a classical approach appeared to be unable to solve the structure of subunit III we turned to the automated refinement procedure (ARP) of Lamzin & Wilson (1993). This procedure is comparable to the iterative least-squares/Fourier synthesis approach used in small-molecule crystallography. It requires high-resolution data of good quality (better than 2.0 Å), and a starting model with at least 75% of the atoms in the correct position. Initially the atomic model is subjected to unrestrained least-squares minimization against the X-ray data in the whole range of resolution. In a second step, the ARP model is updated. $(3F_o-2F_c)$ and $(2F_o-2F_c)$ electron-density maps are calculated using model phases. A small percentage of atoms are rejected if the interpolated value of the $(3F_o-2F_c)$ electron density at the atomic center is less than 1σ above the mean electron-density value. Subsequently, new atoms found in the positive difference electron density are progressively added. If the resolution is high enough, the ARP atoms are approximately located at the true protein atomic positions. The final ARP map, calculated in the last refinement round with the ARP model, can be used to rebuild the initial model.

5.1. *The ARP protocol used on crystal 'form 2'*

'Form 2' was used as the initial model in the ARP procedure. All data in the 20–1.7 Å resolution range were included from the start. Out of the 4310 atoms of the model, around 1000 were considered to be misplaced in the electron density of $(2F_o-F_c)$ maps calculated after crystallographic refinement with *X-PLOR* (Fig. 1*b*). The number of atoms expected in the final model can be estimated to be about 20% higher than the total number of non-H protein atoms (ARP manual). This includes both ordered and disordered solvent molecules. In our case, the final ARP model was set to contain a maximum of 5000 atoms.

Each cycle consisted of one step of unrestrained refinement followed by removal of the worst 15 atoms in the $(3F_o-2F_c)$ electron density (corresponding to 0.35% of the whole protein) and addition of O atoms to the 30 highest $(2F_o-2F_c)$ electron-density peaks (corresponding to 0.7% of the model). Fig. 3 depicts the evolution of the R factor and the free R factor (calculated on 5% of the diffraction data randomly omitted) at each cycle of the automated refinement procedure. This parameter decreased very rapidly during the first five cycles. After 40 such cycles its value was 0.162.

5.2. *Concerted rebuilding of the ill defined parts of the model*

The initial 'form 2' model was inspected in the $(3F_o-2F_c)$ electron-density map calculated after ARP using computer graphics. The 80% of the model which was in good matching density in the $(2F_o-F_c)$ maps after refinement with *X-PLOR* was also in good density in the ARP $(3F_o-2F_c)$ map. The remaining 20% of the 'form 2' model was clearly incorrect, and was extensively rebuilt: the two problematic loops were simultaneously refitted (Fig. 4). In consequence, the new coordinate set presented large differences from the 'form 2' model: about 16 Å for the tip of the autolysis loop, and 18 Å for the tip of the 58–70 residue loop. The N-terminal extremity was clearly disordered and protruded into the solvent medium.

The initial 'form 2' model was successfully rebuilt without ambiguity thanks to the dramatic improvement of the initial electron-density map in the course of the automated refinement procedure. Fig. 5 depicts the differences between the original *X-PLOR* maps [$(2F_o-F_c)$ electron-density map and simulated-annealing omit map] and the ARP map for the region around Tyr58. As expected, in the *X-PLOR* maps the refinement procedure has located a maximum of atoms in the electron density. However, it is clear that the region has been misplaced (Figs. 5*a* and 5*b*). Fig. 5(*c*) shows the final 'form 2' model in this map. Although most of the density can be explained by the correct model, the incorrect atomic coordinates have introduced a great deal of bias. It would have been difficult to trace correctly the polypeptide chain, even by using simulated-annealing omit maps which are supposed to get rid of model bias. Fig. 5(*d*) shows the final 'form 2' model in the ARP map before refinement. The great improvement in the quality of the electron density and the excellent fit to the model are obvious.

It is also of interest to examine some of the intermediate steps during the ARP procedure. Figs. 6(*a*), 6(*b*) and 6(*c*) depict, in more detail, the electron density of Tyr58 in the various stages. After five cycles, the initial model had been significantly distorted and several 'solvent' atoms have been added to what will become the side chain of the tyrosine residue. After 40 cycles of ARP, the electron density is essentially correct and the original model has been completely modified and makes no stereochemical sense. Nevertheless, the positions of the added atoms are extremely accurate and close to the side-chain atom positions of the refined final model (Fig. 6*d*).

5.3. *The final 'form 2' model*

A total of 300 water molecules were automatically assigned to the model by the automatic procedure. The criteria used for selecting water molecules were that the electron density should be well defined, and that the molecules should participate in stereochemically feasible hydrogen bonding. The model was then refined using *X-PLOR*. The *R* factor for data between 8.0 and 1.7 Å dropped to 0.201. Alternate conformations were built for 11 residues (Val39, Val53, Val88, Val152, Val168, Val202, Leu114, Ile41, Ile166, Thr47 and Arg134) and a total of 465 water molecules were added. At this point, the *R* factor was 0.184 for all the data between 8.0 and 1.7 Å resolution (Table 5).

The final model comprises two molecules of 232 residues each and 470 water molecules. The first eight N-terminal residues are disordered and have not been included. The overall folding of the bovine PcpA-S6 subunit III, shown as a ribbon diagram in Fig. 7, is characteristic of the trypsin-like proteins. This model gives excellent agreement with the X-ray data, as shown by the real-space correlation plot calculated with *O* (Fig. 8*a*). Only two residues (106 and 60), located at the end of flexible loops, are poorly defined. The r.m.s. deviations from standard bond lengths and angles are presented in Table 5. The dihedral angles (φ, ψ) are clustered on the allowed region of conformational space (Fig. 9*a*). The r.m.s. coordinate error, estimated from a Luzzati plot (Luzzati, 1952) is approximately 0.2 Å.

Superimposition of the two molecules of the asymmetric unit with the program *ALIGN* (Cohen in, Satow, Cohen, Padlan & Davies, 1986) resulted in r.m.s. deviation of 0.4 Å for 230 Cα positions. For the polypeptide chain the agreement is poorest at the N-terminus, where the density is weak and the local conformation appears to differ between the two molecules, and at some of the loops connecting secondary-structure elements (Fig. 8*a*). The averaged temperature factors for main-chain atoms are 12.2 and 17.9 Å^2 for the two molecules of the asymmetric unit. The temperature-factor differences are evenly distributed throughout the two structures.

5.4. *The reliability of the model: refinement of crystal 'form 1'*

In order to refine the first crystal form, the final 'form 2' model was superimposed on the 'form 1' model with the program *O*. The refinement was carried out for data between 8.0 and 2.2 Å resolution using the conventional restrained refinement protocol of *X-PLOR*. During this procedure the *R* factor decreased from 0.3 to 0.23. Inspection of the ($2F_o - F_c$) map revealed that the regions manually rebuilt into the electron-density map generated by ARP for 'form 2' were equivalent, and as well defined in the three dimensional model of 'form 1'. Water molecules were included gradually applying the same criteria used in the case of 'form 2'. One extra residue relative to 'form 2' was built at the N-terminus.

Table 5. *Statistics on the refined models for the two crystal forms*

	Form 1	Form 2
No. of residues	233	464
No. of water molecules	120	470
B factor (Å^2)	29.5	19.2
Resolution (Å)	8–2.2	8.0–1.7
R factor (%)	18.8	18.4
Free *R* factor (%)	23.0	22.7
Standard deviation from bond lengths (Å)	0.014	0.012
Standard deviation from bond angles (°)	1.61	1.37

The final 'form 1' model consists of 233 residues and 120 water molecules. The *R* factor is 0.188 for data between 8.0 and 2.2 Å resolution. Fig. 8(*b*) depicts the real-space correlation plot as calculated with *O* (Jones *et al.*, 1991). The agreement of the model is reasonable, but the correlation values are slightly lower than those calculated for the 'form 2' crystal. This may be due to the intrinsically poorer quality of the crystals and/or to the lower resolution of the analysis. The r.m.s. deviations from standard bond lengths and angles are presented in Table 5. The dihedral angles φ and ψ are clustered in the allowed region of the conformational space, except for four residues (D178, W163, F212 and S130) (Fig. 9*b*).

Superimposition of the final 'form 1' model with molecule A of the final model of the 'form 2' crystal results in an r.m.s. value of 0.54 Å for 221 residues. This value increases to 0.78 when all the Cα positions are considered. The main differences are located at the end of flexible loops, stabilized by different packing contacts in the two crystal forms.

6. Concluding remarks

In this paper, we describe in detail the strategy used to solve the bovine subunit III structure, using two crystal forms. About 20% of the 'form 1' model obtained by molecular replacement and conventional refinement presented ill defined density. The problematic regions could be disordered (as already observed in numerous zymogens) or misinterpreted. A molecular-replacement solution, calculated for the second crystal form and confirmed by heavy-atom derivative phases proved the necessity for rebuilding the model. The automated refinement procedure applied to the high-resolution data of the second crystal form improved the electron density. The 'form 2' initial model was rebuilt without ambiguity in the resulting map, by concerted movements of the problematic loops.

Superimposition of the initial and final 'form 2' models leads to an r.m.s. deviation of 4.9 Å in the Cα positions for the 464 residues of the two molecules in the asymmetric unit. This value mainly reflects the different positions of the four regions that were manually rebuilt at the end of the ARP procedure. The major differences are located in the autolysis loop (residues 140–155) and in the calcium-binding loop (residues 58–70), with distances of 14.1 and 18.7 Å , respectively (Figs. 4 and 10). These rebuilt regions have well defined, unique conformations, and suggest that the bovine subunit III corresponds to a truncated version of a new class of highly structured elastase-like zymogen molecules (Pignol *et al.*, 1994).

The contribution of ARP to the definition of the electron density corresponding to the misplaced loops of subunit III is dramatic. In our case it is clear that the information contained in the phases obtained from the partially correct molecular-replacement model was not sufficient for unambiguous interpretation of the electron density in the ill defined regions either before or after refinement with *X-PLOR*. The important model bias is similar to the one observed for human factor D, a trypsin-like protein presenting a unique disposition of the catalytic site (Narayana *et al.*, 1994; Carson, Bugg, DeLucas & Narayana, 1994). Because of the non-isomorphism observed between native data sets, it was not possible to solve the subunit III structure by multiple isomorphous replacement. The use of a very

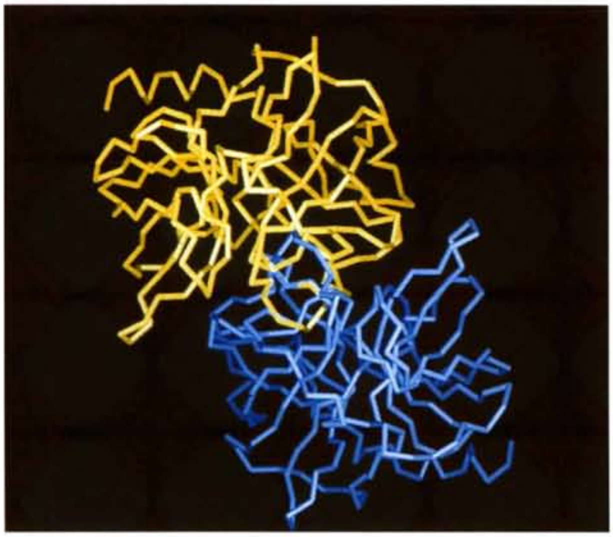


Fig. 2. Second crystal form: Cα backbone representation of the molecule *A* (in blue) and *B* (in yellow), oriented and positioned according to the molecular-replacement solution. The autolysis loops of each molecule clearly interpenetrate.

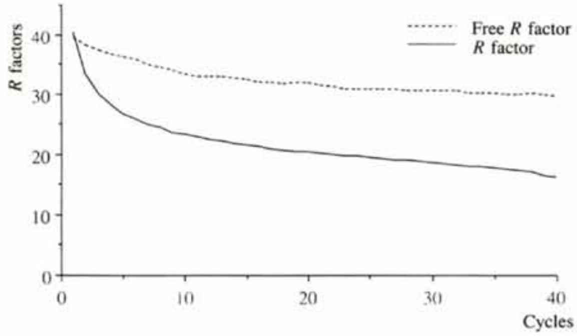
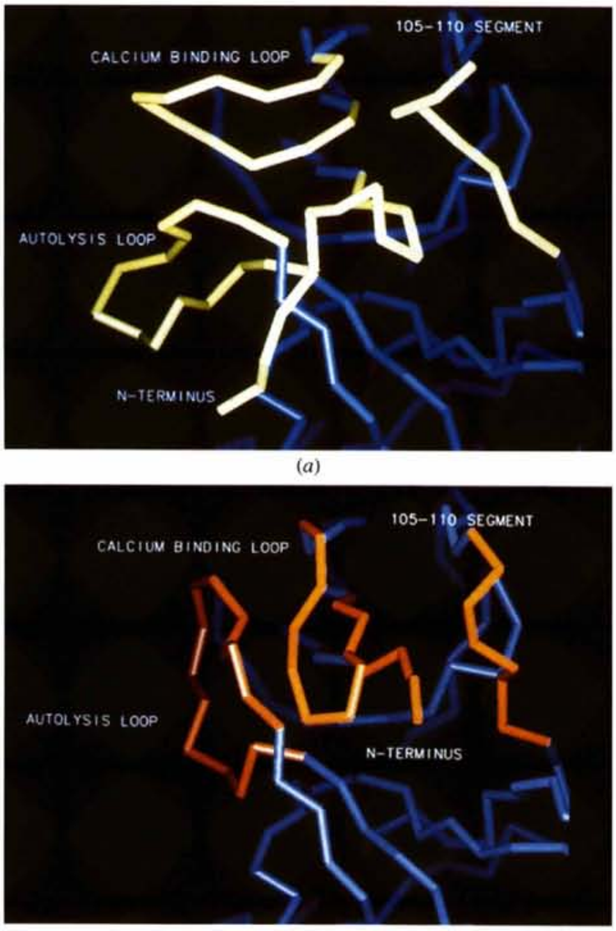


Fig. 3. Second crystal form: evolution of the *R* factor and the free *R* factor (dashed line) (Brünger, 1992) at each cycle of the automated refinement procedure.



(*a*)

(*b*)

Fig. 4. Cα backbone of (*a*) the form 2 initial model (*b*) the final form 2 model. The conserved parts of the two models are painted in blue, the main differences are drawn in (*a*) yellow or (*b*) red. The autolysis and the calcium-binding loops (60–70) are in a completely different orientation.

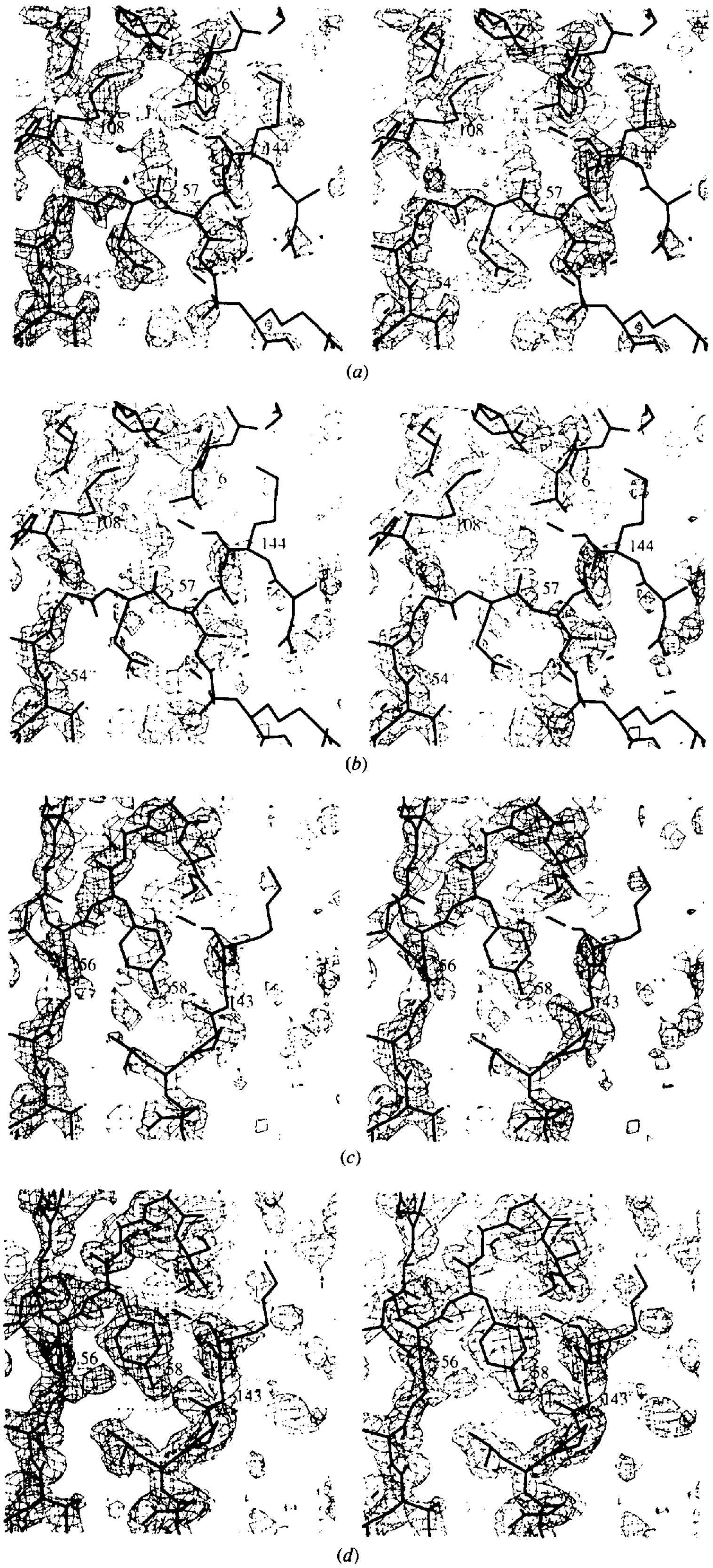


Fig. 5. Evolution of the electron-density maps around Tyr58 in the course of the automated refinement procedure: (*a*) The $(2F_o - F_c)$ *X-PLOR* electron-density map is superimposed on the initial form 2 model. The simulated-annealed omit map (*X-PLOR*), calculated with an omit region of 8 Å around Tyr58 is superimposed on (*b*) the initial form 2 model (*c*) the final form 2 model. (*d*) The $(3F_o - F_c)$ ARP electron-density map is superimposed on the final form 2 model. The electron-density maps were contoured at 1σ level.

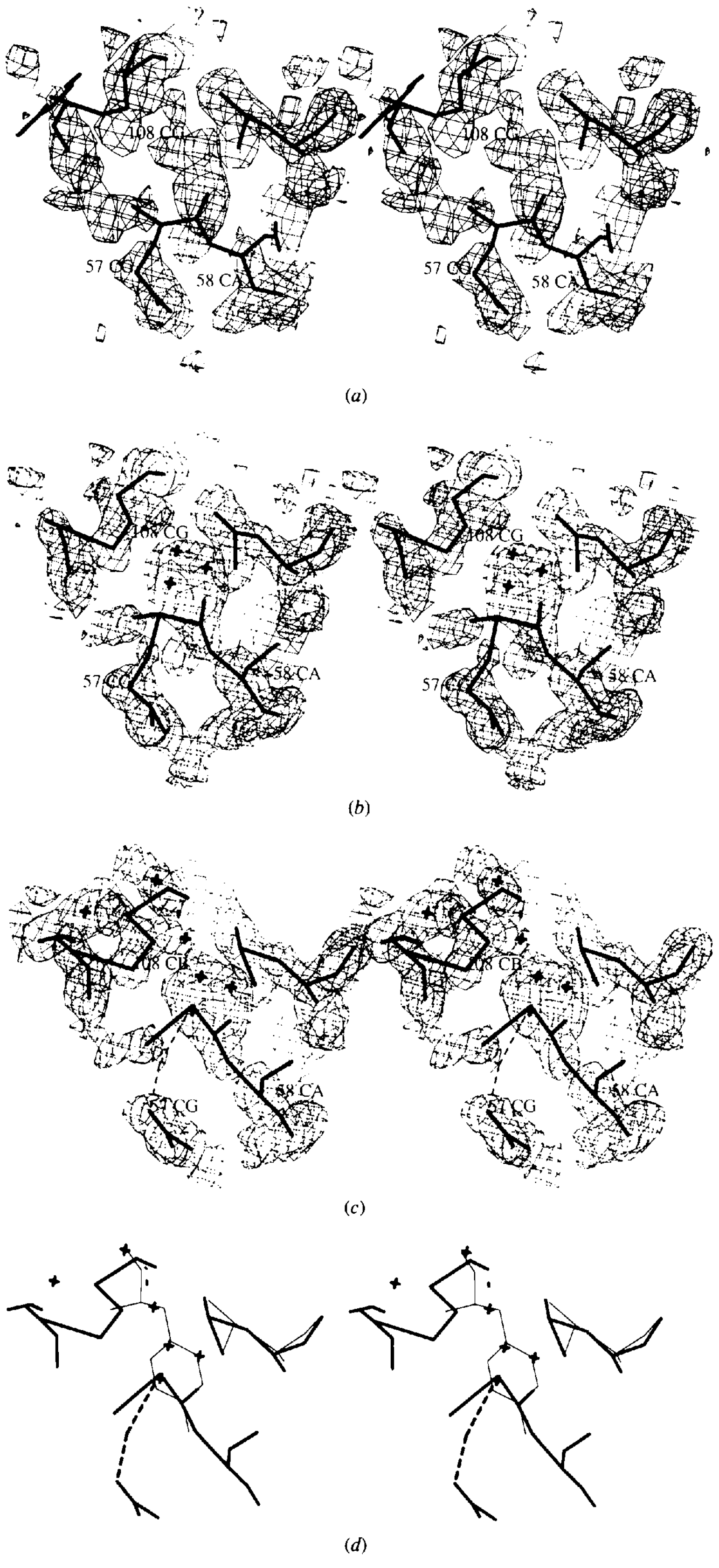


Fig. 6. Intermediate steps during the ARP procedure: (*a*) superimposition of the form 2 model on the *X-PLOR* ($2F_o - F_c$) electron-density map for the Tyr58. The intermediate models are superimposed on the corresponding ($3F_o - F_c$) ARP electron-density maps (*b*) after five cycles and (*c*) 40 cycles of the ARP procedure. The models are distorted, but the electron-density map has been dramatically improved. (*d*) The 40 cycles ARP model (thick lines) superimposed on the final refined model (thin lines) shows that the added water molecules are close to the real atom positions of the final Tyr58 side chain.

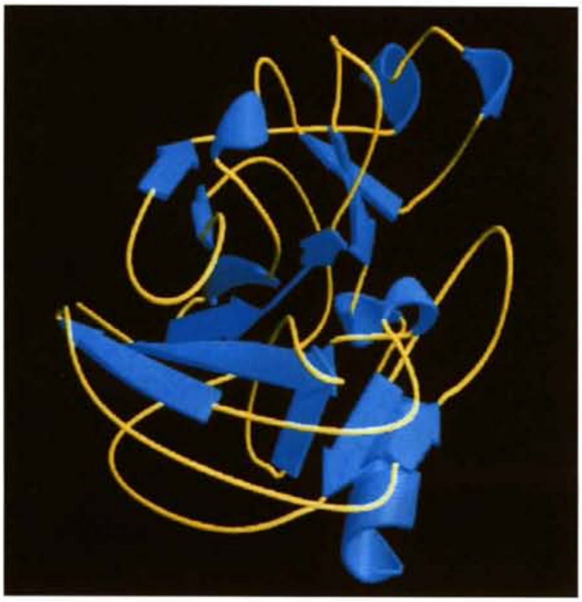
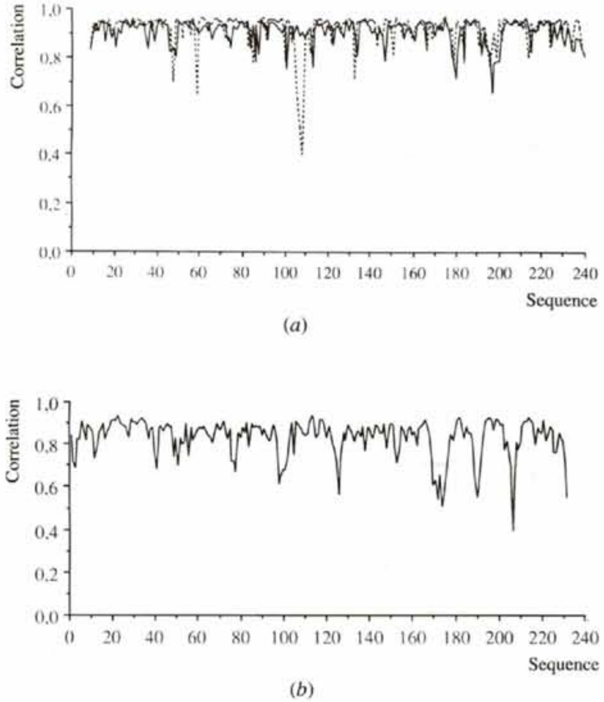


Fig. 7. The overall folding of bovine subunit III has an architecture characteristic of the trypsin-like proteins.



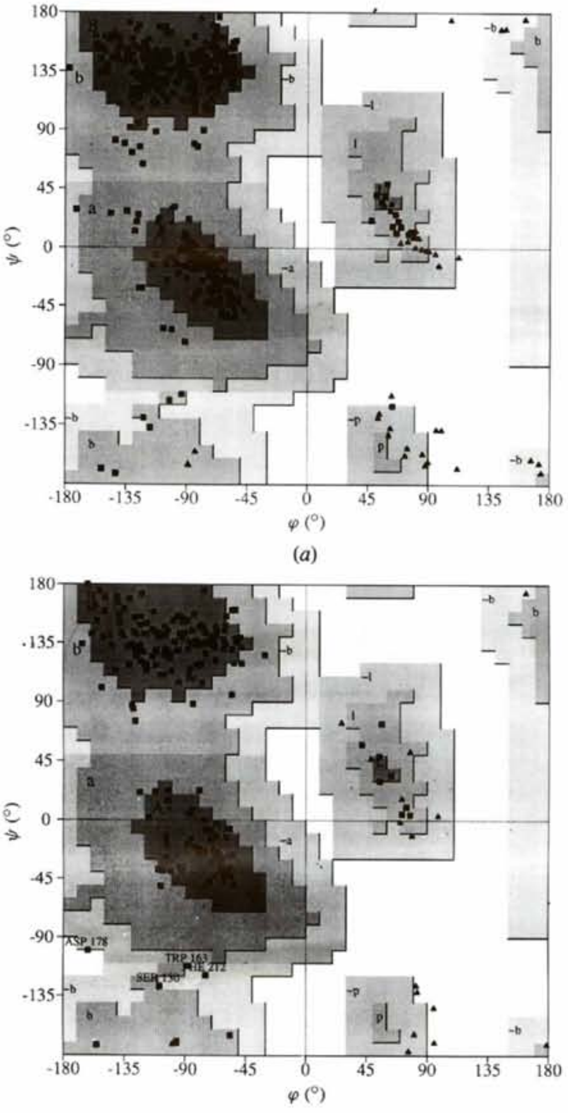
(*a*)

(*b*)

Fig. 8. Correlation plot between the $(2F_o - F_c)$ electron-density maps and the final models for (*a*) molecule *A* and *B* (dashed line) of the second crystal form and (*b*) for the first crystal form.

effective molecular-replacement package such as *AMoRe* in conjunction with the ARP procedure appears as a very powerful tool for solving structures with a search model that may have only about 75% structural similarity.*

* Atomic coordinates and structure factors have been deposited with the Protein Data Bank, Brookhaven National Laboratory. Free copies may be obtained through The Managing Editor, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England (Reference: GR0449).



(*a*)

(*b*)

Fig. 9. Ramachandran plot of the (φ, ψ) angles, obtained with the program *PROCHECK* (Laskowski, MacArthur, Moss & Thornton, 1993) for (*a*) the refined model of the second crystal form and (*b*) for the refined model of the first crystal form. Over 90% of the non-glycine residues (shown as squares) lie in the most favoured regions (A, B, L) defined by Morris, MacArthur, Hutchinson & Thornton (1992). Glycines are shown by triangles.

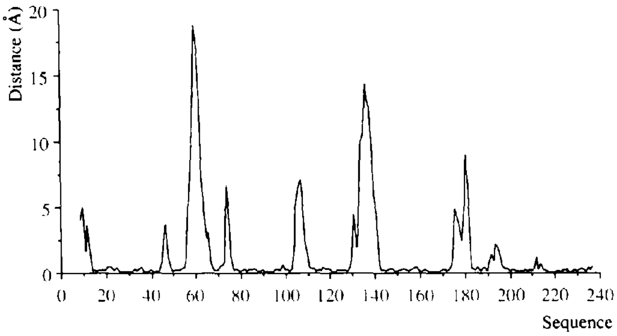


Fig. 10. Distances between Cα positions after superimposition of the initial and final form 2 models. The main differences are located in the calcium-binding loop (60–70), the autolysis loop (130–145) and the activation domain (180–195).

We thank T. Michon, B. Kerfelec and C. Chapus for provide us with pure subunit III samples. We wish to thank R. Fourme and J. P. Benoit for help with data collection at LURE. This work was supported by the Commissariat l'Energie Atomique and the Centre National de la Recherche Scientifique.

References

Abergel, C., Moulard, M., Moreau, H., Loret, E., Cambillau, C. & Fontecilla-Camps, J. C. (1991). *J. Biol. Chem.* **266**, 20131-20138.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rogers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 534–542.

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

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.

Cambillau, C., Kerfelec, B., Foglizzo, E. & Chapus, C. (1986). *J. Mol. Biol.* **189**, 709–710.

Cambillau, C., Kerfelec, B., Sciaky, M. & Chapus, C. (1988). *FEBS Lett.* **232**, 91–95.

Carson, M., Bugg, C. E., DeLucas L. & Narayana, S. V. L. (1994). *Acta Cryst.* **D50**, 889–899.

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

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

Fehlammer, H., Bode, W. & Huber, R. (1977). *J. Mol. Biol.* **111**, 415–438.

Howard, A. J., Gilliland, G. L., Finzel, B. C., Poulos, T. L., Ohlendorf, D. H. & Salemne, F. R. (1987). *J. Appl. Cryst.* **20**, 383-387.

Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* **A47**, 110–119.

Keil-Dlouha, V., Puigserver, A., Marie, A. & Keil, B. (1972). *Biochem. Biophys. Acta*, **276**, 531–535.

Kerfelec, B., Cambillau, C., Puigserver, A. & Chapus, C. (1986). *Eur. J. Biochem.* **157**, 531–538.

Kerfelec, B., Chapus, C. & Puigserver, A. (1984). *Biochim. Biophys. Res. Commun.* **121**, 162–167.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283-291.

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

Leslie, A. G. W., Brick, P. & Wonacott, A. J. (1986). *CCP4 Newslett.* **18**, 33–39.

Luzzati, V. (1952). *Acta Cryst.* **5**, 802–819.

Morris, A. L., MacArthur, M. W., Hutchinson, E. G. & Thornton, J. M. (1992). *Proteins*, **12**, 345-364.

Meyer, E., Cole, G. & Radhakrishnan, R. (1988). *Acta Cryst.* **B44**, 26–38.

Narayana, S. V. L., Carson, M., El-Kabbani, O., Kilpatrick, J. M., Moore, D., Chen, X., Bugg, C. E., Volanakis, J. E. K. & De Lucas, L. J. (1994). *J. Mol. Biol.* **235**, 695–708.

Navaza, J. (1987). *Acta Cryst.* **A43**, 645–653.

Navaza, J. (1990). *Acta Cryst.* **A46**, 619–620.

Navaza, J. (1994). *Acta Cryst.* **A50**, 157–163.

Pignol, D., Gaboriaud, C., Michon, T., Kerfelec, B., Chapus, C. & Fontecilla-Camps, J. C. (1994). *EMBO J.* **13**(8), 1763–1771.

Puigserver, A. & Desnuelle, P. (1975). *Proc. Natl Acad. Sci. USA*, **72**, 2242–2445.

Ramakrishnan, C. & Ramachandran, G. N. (1965). *Biophys. J.* **5**, 909–933.

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

Satow, Y., Cohen, G. H., Padlan, E. A. & Davies D. R. (1986). *J. Mol. Biol.* **190**, 593–604.

Steigemann, W. (1974). PhD thesis, Technische Universität, München, Germany.

Venot, N., Sciaky, M., Puigserver, A., Desnuelle, P. & Laurent, G. (1986). *Eur. J. Biochem.* **157**, 91–99.

Wang, D., Bode, W. & Huber, R. (1985). *J. Mol. Biol.* **185**, 595-596.

Wlodawer, A. & Hodgson, K. O. (1975). *Proc. Natl Acad. Sci. USA*, **72**, 398–399.