Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Structures of the HIV-1 capsid protein dimerization domain at 2.6 Å resolution

David K. Worthylake, Hui Wang, Sanghee Yoo, Wesley I. Sundquist and Christopher P. Hill*

Biochemistry Department, University of Utah Medical School, 50 North Medical Drive, Salt Lake City, Utah 84132, USA

Correspondence e-mail: chris@snowbird.med.utah.edu

Received 23 February 1998
Accepted 29 May 1998

PDB Reference: HIV-1 capsid protein dimerization domain, 1a43 and 1baj.

The human immunodeficiency virus type I (HIV-1) capsid protein is initially synthesized as the central domain of the Gag polyprotein, and is subsequently proteolytically processed into a discrete 231-amino-acid protein that forms the distinctive conical core of the mature virus. The crystal structures of two proteins that span the C-terminal domain of the capsid are reported here: one encompassing residues 146–231 ($CA_{146–231}$) and the other extending to include the 14-residue p2 domain of Gag ($CA_{146–p2}$). The isomorphous $CA_{146–231}$ and $CA_{146–p2}$ structures were determined by molecular replacement and have been refined at 2.6 Å resolution to R factors of 22.3 and 20.7% (R_{free} = 28.1 and 27.5%), respectively. The ordered domains comprise residues 148–219 for $CA_{146–231}$ and 148–218 for $CA_{146–p2}$, and their refined structures are essentially identical. The proteins are composed of a 3_{10} helix followed by an extended strand and four α-helices. A crystallographic twofold generates a dimer that is stabilized by parallel packing of an α-helix 2 across the dimer interface and by packing of the 3_{10} helix into a groove created by α-helices 2 and 3 of the partner molecule. $CA_{146–231}$ and $CA_{146–p2}$ dimerize with the full affinity of the intact capsid protein, and their structures therefore reveal the essential dimer interface of the HIV-1 capsid.

1. Introduction

The major structural proteins of the human immunodeficiency virus (HIV) are encoded by the viral *Gag* gene (reviewed by Kräusslich, 1996). *Gag* is initially translated as a 55 kDa polyprotein, using the unspliced viral RNA genome as a template. As Gag molecules accumulate, they capture the viral RNA, migrate to the inner cell membrane and assemble into immature viral particles that bud from the cell. Concomitant with budding, Gag is cleaved by the viral protease at five discrete sites, creating six new polypeptides: matrix (MA, residues 1–132), capsid (CA, 133–363), p2 (364–376), nucleocapsid (NC, 377–432), p1 (433–448) and p6 (449–500). Cleavage at the MA–CA, CA–p2 and p2–NC sites is essential for viral replication and is temporally controlled. The p2–NC junction is cleaved most rapidly, followed by the MA–CA junction and finally by the CA–p2 junction (Erickson-Viitanen *et al.*, 1989; Gowda *et al.*, 1989; Kräusslich *et al.*, 1989; Mervis *et al.*, 1988; Pettit *et al.*, 1994). These processing events trigger a major morphological transformation, termed viral maturation, that is essential for infectivity. During maturation, the genomic RNA–NC complex condenses into the center of the virion and the capsid protein assembles into a conical core structure that surrounds the genome.

Mutational analyses have revealed that the HIV-1 capsid polypeptide performs essential functions in a number of viral processes including virion assembly, cyclophilin A packaging, core morphogenesis and early events following infection. The 231-amino-acid capsid protein is composed of two domains: an N-terminal domain (residues 1–146) that binds cyclophilin A (Franke, Yuan & Lubun, 1994; Gamble *et al.*, 1996; Thali *et al.*, 1994) and participates in core assembly (Dorfman *et al.*, 1994; Reicin *et al.*, 1995; von Schwedler *et al.*, 1998; Wang & Barklis, 1993), and a C-terminal domain (residues 148–231) that participates in both assembly of the immature virion and the mature viral core (Carrière *et al.*, 1995; Chazal *et al.*, 1994; Dorfman *et al.*, 1994; Franke, Yuan, Bossolt *et al.*, 1994; Hong & Boulanger, 1993; Jowett *et al.*, 1992; Mammano *et al.*, 1994; Reicin *et al.*, 1995, 1996; Srinivasakumar *et al.*, 1995; von Poblotzki *et al.*, 1993; Zhang *et al.*, 1996; Zhao *et al.*, 1994). Under physiological conditions, the recombinant HIV-1 capsid protein dimerizes in solution (Brooks *et al.*, 1994; Ehrlich *et al.*, 1992, 1994; Rosé *et al.*, 1992; Yoo *et al.*, 1997) and dimerization is mediated by the C-terminal domain (Gamble *et al.*, 1997; Yoo *et al.*, 1997). The C-terminal domain also contains a highly conserved segment of 20 amino acids (residues 153–172), known as the major homology region (MHR), which is conserved across the lenti- and oncoviruses (Patarca & Haseltine, 1985; Wills & Craven, 1991), as well as in the yeast retrotransposon Ty3 (Orlinsky *et al.*, 1996). Mutations within the MHR generally block viral replication, but can have a variety of different phenotypes, including blocking viral assembly (Craven *et al.*, 1995; Mammano *et al.*, 1994; Orlinsky *et al.*, 1996; Strambio-de-Castilla & Hunter, 1992), maturation (Craven *et al.*, 1995; Mammano *et al.*, 1994) or replication (Craven *et al.*, 1995; Orlinsky *et al.*, 1996; Strambio-de-Castilla & Hunter, 1992). Thus, the precise function of the MHR is currently unknown.

The first high-resolution structural information for the capsid C-terminal domain was provided by the recent crystal structure of a protein spanning capsid residues 151–231 ($CA_{151–231}$) (Gamble *et al.*, 1997). This structure was determined at 1.7 Å resolution and revealed the helical fold of the monomeric capsid C-terminal domain in detail, including the intricate hydrogen-bonding network formed by the MHR sequence. $CA_{151–231}$ is a poor model for understanding capsid dimerization, however, since $CA_{151–231}$ is monomeric in solution, even at high micromolar concentrations (Gamble *et al.*, 1997), whereas the intact capsid protein dimerizes with a low micromolar dissociation constant (K_d = 18 μ*M*; Brooks *et al.*, 1994; Rosé *et al.*, 1992; Yoo *et al.*, 1997).

Although $CA_{151–231}$ dimerizes poorly, a slightly longer capsid protein fragment spanning capsid residues 146–231 ($CA_{146–231}$) dimerizes with an affinity that is similar to that of the intact protein (K_d = 10 μ*M*; Gamble *et al.*, 1997). $CA_{146–231}$ is therefore likely to be a good model for the dimer formed by the intact capsid protein. Consistent with this expectation, crystals of $CA_{146–231}$ exhibit a different set of protein–protein interactions to those of the $CA_{151–231}$ crystal. The initial $CA_{146–231}$ crystals diffracted poorly, however, and only allowed determination of an unrefined $CA_{146–231}$ structure at 3.0 Å resolution (Gamble *et al.*, 1997). We have now improved the crystal quality and refined the structure at 2.6 Å resolution. In an effort to determine the generality of the $CA_{146–231}$ structure and to understand the structure and function of the Gag p2 peptide, we have also crystallized and determined the structure of a longer construct that includes the p2 sequence ($CA_{146–p2}$), although, unfortunately, the p2 peptide is disordered in this crystal form. Together, these structures provide a detailed picture of the principal dimer interface of the HIV-1 capsid protein.

2. Material and methods

2.1. Protein expression and purification

The proviral pNL4-3 plasmid, containing the HIV-1_{NL4-3} genome, was obtained from M. Martin through the AIDS Research and Reference Reagent Program (Adachi *et al.*, 1986; Myers *et al.*, 1995). DNA encoding an initiator methionine followed by capsid residues 146–231 ($CA_{146–231}$) or capsid residues 146–231 followed by the 14-amino-acid p2 domain of Gag ($CA_{146–p2}$) was amplified from the pNL4-3 template using the polymerase chain reaction and subcloned into the *Nde*1/*Bam*H1 site of the expression vector pET11a (Novagen). The resulting plasmids WISP97-07 ($CA_{146-231}$) and WISP97-35 ($CA_{146–p2}$) were confirmed by DNA sequencing.

The $CA_{146–231}$ and $CA_{146–p2}$ proteins were expressed in *E. coli* strain BL21(DE3) to produce unlabelled protein or in the methionine auxotroph B834(DE3) to produce selenomethionine-substituted $CA_{146–231}$ ($SeCA_{146–231}$) (Studier *et al.*, 1990). Protein expression was induced in late log phase ($A_{600} \simeq 0.8$) by the addition of IPTG to 1 m*M*. After 4 h, the cells were lysed in a French press and the soluble supernatant sonicated to reduce viscosity. All protein purification steps were performed at 277 K in 1 m*M* PMSF to minimize proteolytic degradation. The insoluble material was removed by centrifugation for 1 h at 40000*g*. The clarified supernatant was adjusted to 40% saturation in ammonium sulfate and insoluble material was pelleted by centrifugation at 6800*g* for 10 min. At this point the purification procedures for $CA_{146–231}$ and $CA_{146–p2}$ differed.

2.2. Purification of $CA_{146–231}$ ($SeCA_{146–231}$)

The supernatant containing $CA_{146–231}$ was adjusted to 70% saturation in ammonium sulfate and subjected to a 6800*g* centrifugation for 10 min. The pellet was dissolved in buffer *A* (25 m*M* potassium phosphate pH 6.5, 5 m*M* 2-mercaptoethanol) and loaded onto an S-Sepharose column (Pharmacia), which was developed with a 250 ml linear gradient from 0 to 1 *M* NaCl in buffer *A*. Fractions containing $CA_{146–231}$ were identified by SDS–PAGE, loaded onto a preparation grade Phenyl Sepharose column (Pharmacia) and purified to homogeneity using a 300 ml linear gradient from 1 to 0 *M* ammonium sulfate in buffer *A*. The purified protein was concentrated as necessary by ultrafiltration (Amicon, YM3 membranes) and stored at 277 K.

The pure protein was analyzed by electrospray mass spectrometry, which demonstrated that the initiator methionine was retained during expression to produce a protein of molecular weight 9649 ± 1 Da (calculated M_W = 9650 Da). We have previously used analytical ultracentrifugation to demonstrate that the dimerization constant for this protein is 10 ± 3 μ*M* (Gamble *et al.*, 1997). Free thiol groups were quantified according to the method of Ellman (1959). Briefly, 2-mercaptoethanol in the storage buffer was removed by chromatography over a Superdex 75 gel-filtration column (Pharmacia) and the protein eluted in 100 m*M* NaCl, 10 m*M* potassium phosphate pH 7.0. The protein concentration was quantitated by absorption spectroscopy (ε_{280} = 8490 M^{-1} cm^{-1}) (Gill & von Hippel, 1989) and the free thiol concentration measured by absorbance at 410 nm in a buffer containing 1 m*M* EDTA and 0.4 mg ml^{-1} 5,5′-dithiobis(2-nitrobenzoic acid) (DTNB). In three repetitions of the experiment, 53 ± 5% of the Cys residues were found to be present as free thiols in freshly purified $CA_{146–231}$ protein (*i.e.*, just under half of the molecules had a disulfide bond). It is likely that this value represents a lower estimate of the actual free thiol concentration in freshly purified protein, since the protein sat at 277 K for ∼3 h in buffer that lacked 2-mercaptoethanol as the assay was being performed. After a 4 d incubation at 277 K in this buffer, the protein's free thiol content had decreased to 6 ± 1%.

2.3. Purification of $CA_{146–p2}$

The pellet containing $CA_{146–p2}$ was dissolved in buffer *B* (25 m*M* potassium morpholinopropanesulfonic acid pH 6.8, 5 m*M* 2-mercaptoethanol) and applied to a preparation-grade S-Sepharose column. The column was developed using a 300 ml 0 to 1 *M* linear gradient of NaCl in buffer *B*. Fractions containing $CA_{146–p2}$ were identified using SDS–PAGE and applied to a preparation-grade Phenyl Sepharose column. Purified $CA_{146–p2}$ was eluted using a 300 ml 1 to 0 *M* linear ammonium sulfate gradient in buffer *B*. The purified protein was analyzed by mass spectroscopy, which demonstrated that the initiator methionine was retained during expression to produce a protein of molecular weight 11092 ± 1 Da (calculated M_W = 11096 Da). $CA_{146–p2}$ dimerization was quantitated by analytical ultracentrifugation (Yoo *et al.*, 1997) in a buffer of 120 m*M* KCl, 25 m*M* potassium phosphate pH 7.2 and 2 m*M* 2-mercaptoethanol. These studies were performed at 277 K, which minimized the proteolytic degradation of $CA_{146–p2}$ but lowered the apparent K_d values of $CA_{146–231}$ and $CA_{146–p2}$. Under these conditions, the measured dimerization constants for $CA_{146–231}$ (K_d = 5 ± 2 μ*M*) and $CA_{146–p2}$ (K_d = 7 ± 1 μ*M*) were indistinguishable (data not shown).

2.4. Crystallization

Crystals of both native and selenomethionine-substituted $CA_{146–231}$ were grown by vapor diffusion at room temperature in sitting drops containing a 1:1 mixture of protein solution (2.1 m*M* $CA_{146–231}$, 10 m*M* Tris–HCl pH 8.0 and 2 m*M* 2-mercaptoethanol) and reservoir solution (0.7 *M* ammonium sulfate, 1.0 *M* lithium sulfate and 0.1 *M* HEPES pH 7.0). Crystals of $CA_{146–p2}$ were grown at room temperature by vapor diffusion in sitting drops consisting of a 1:1 mixture of protein solution (1.35 m*M* $CA_{146–p2}$, 10 m*M* Tris–HCl pH 8.0 and 2 m*M* 2-mercaptoethanol) and reservoir solution (0.1 *M* HEPES pH 7.5, 1.6 *M* KH_2PO_4/NaH_2PO_4). The measured pH of the well solution was actually 5.0 because monobasic forms of sodium and potassium phosphate were used. Crystals of each protein appeared within 2 d, and within two weeks grew to maximum dimensions of 0.6 × 0.5 × 0.5 mm for $CA_{146–p2}$ and native $CA_{146–231}$, and 0.2 × 0.2 × 0.2 mm for $SeCA_{146–231}$.

As the p2 polypeptide was not visible in electron-density maps of $CA_{146–p2}$ (see below), we used SDS–PAGE analysis to test whether this peptide had been removed by proteolysis. Ten crystals were disolved in 10 μl H_2O. 2 μl of 5X gel-loading buffer (250 m*M* Tris–HCl pH 6.8, 10% SDS, 0.5% bromophenol blue, 50% glycerol, 25% 2-mercaptoethanol) was added to the protein solution. SDS–PAGE was performed in a 20% acrylamide Tricine gel (Schägger & von Jagow, 1987). Coomassie staining of the gel revealed that the protein in the crystal was $CA_{146–p2}$, with no contaiminating $CA_{146–231}$ as judged by comparison with authentic standards of both proteins.

Table 1
Native data statistics.

Quantities in parentheses are statistics for the highest resolution shell (2.69–2.60 Å).

	$CA_{146–231}$	$CA_{146–p2}$
d_{min} (Å)	2.6	2.6
Number of observed reflections	21325	19752
Number of unique reflections	3166	3227
Completeness (%)	91.3 (92.5)	92.9 (100)
R_{sym}† (%)	7.2 (52.9)	5.6 (33.6)
$\langle I/\sigma(I)\rangle$	13.5 (5.9)	16.2 (7.9)

† $R_{sym} = 100\sum_h \sum_i |I_{hi} - \langle I_h\rangle| / \sum_h \sum_i I_{hi}$, where I_i is the intensity of an individual reflection from a symmetry-related group of reflections and $\langle I\rangle$ is the mean intensity of that group of reflections.

Table 2
Refinement statistics.

	$CA_{146–231}$	$CA_{146–p2}$
Resolution† (Å)	20.0–2.60	20–2.60
R_{free}‡ (%)	28.1	27.5
R factor§ (%)	22.3	20.7
Number of residues/number of H_2O	72/28	71/27
R.m.s. deviation bond lengths (Å)	0.009	0.008
R.m.s. deviation bond angles (°)	1.24	1.26
Ramachandran plot¶		
Most favored (%)	90.5	91.9
Allowed (%)	9.5	8.1
$\langle B\rangle$ (solvent) (Å^2)	70.3 (85.1)	66.1 (84.7)
$\langle B\rangle$ Wilson†† (Å^2)	77.1	70.6

† All observed reflections [$|F(\text{obs})| > 0$] were used for refinement. ‡ R_{free} = *R* factor for a randomly selected set of reflections that have not been used in refinement. § *R* factor = $100\sum[||F_{p(\text{obs})}| - |F_{p(\text{calc})}||] / \sum |F_{p(\text{obs})}|]$. ¶ Geometry was analyzed with *PROCHECK* (Laskowski *et al.*, 1993). †† Mean temperature factor of data set as determined from Wilson statistics using 3.5–2.6 Å data.

Table 3
$SeCA_{146–231}$ multiwavelength data statistics.

Quantities in parentheses are statistics for the highest resolution shell (2.69–2.60 Å). Selenomethionine-substituted $CA_{146–231}$ data were collected at NSLS beamline X8C.

Data set	Wavelength (Å)	Resolution (Å)	Unique reflections (N)	Redundancy†	Completeness (%)	R_{sym}‡ (%)
λ_1	0.9766	20.0–2.6	5873	2.5	94.0 (91.0)	4.8 (30.8)
λ_2	0.9763	20.0–2.6	5895	2.4	93.6 (85.1)	4.2 (32.2)
λ_3	0.9300	20.0–2.6	5798	2.5	93.2 (93.3)	4.2 (24.3)

† Ratio of the total number of measurements to the number of unique reflections. ‡ $R_{sym} = 100\sum_{hkl}\sum_i |I_i\langle I\rangle|/\sum\langle I\rangle$, where I_i is the intensity of an individual reflection from a symmetry-related group of reflections and $\langle I\rangle$ is the mean intensity of that group of reflections.

2.5. Data collection and processing

Both $CA_{146–231}$ and $CA_{146–p2}$ crystallized in space group $I4_1$ with one molecule in the asymmetric unit. $CA_{146–231}$ crystals have cell parameters a = 60.41, c = 60.43 Å, and $CA_{146–p2}$ crystals have cell parameters a = 61.33, c = 59.57 Å. Data were collected on a MAR image-plate detector at beamline X8C at the National Synchrotron Light Source, Brookhaven National Laboratory (Table 1). $CA_{146–p2}$ and native $CA_{146–231}$ data were collected at room temperature, while data from $SeCA_{146–231}$ crystals were collected at 100 K. $SeCA_{146–231}$ crystals were transferred to a cryoprotectant [0.7 *M* ammonium sulfate, 1.0 *M* lithium sulfate, 0.1 *M* HEPES pH 7.0 and 15%(*v*/*v*) glycerol], suspended in a small rayon loop and cooled by plunging into liquid nitrogen. All data were processed using *DENZO* and *SCALEPACK* (Otwinowski & Minor, 1997). The $CA_{146–p2}$ data from three separate crystals were merged, while the native $CA_{146–231}$ data were collected from a single crystal. The low-temperature multiwavelength $SeCA_{146-231}$ data were collected from a single crystal at three wavelengths: λ_1 (0.9766 Å) corresponding to the minimum in f', λ_2 (0.9763 Å) corresponding to the maximum in f'', and λ_3 (0.9300 Å) corresponding to the maximum in f' (Table 2). All data sets extend to 2.6 Å resolution.

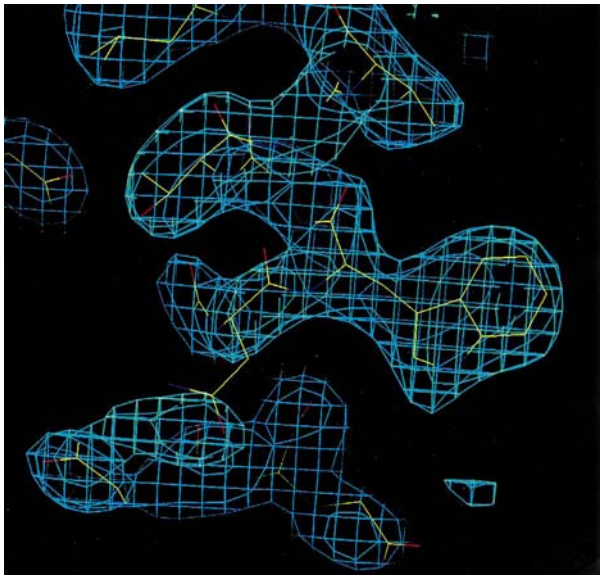

Figure 1
Experimental electron density and refined coordinates. The MAD/*DM* map of $SeCA_{146–231}$, contoured in cyan at 1.2 times the root-mean-square deviation, is calculated at 15–2.6 Å resolution.

2.6. Structure determination and refinement of $CA_{146–p2}$

The $CA_{146–p2}$ structure was determined with the molecular-replacement program *AMoRe* (Navaza, 1994) using the refined structure of $CA_{151–231}$ as a search model (Gamble *et al.*, 1997). Using data from 7–3 Å and an integration radius of 3–16 Å, the top rotation-function solution had a correlation coefficient of 15.0% (next was 14.1%) and this solution was the top translation-function solution with a correlation coefficient of 43.8% (next was 33.7%). Rigid-body refinement gave an R factor of 43.8% against all 7.0–3.0 Å data.

Rounds of positional and atomic temperature-factor refinement using the program *X-PLOR* (Brünger, 1992*b*) were interspersed with manual model building with the graphics program *O* (Jones *et al.*, 1991). 10% of the data (349 reflections) were withheld from refinement in order to monitor progress by cross-validation (Brünger, 1992*a*) and to optimize weights. The refined structure, including bulk-solvent correction, has an R factor of 20.7% (R_{free} = 27.5%) using all 20–2.6 Å data with $|F(\text{obs})| > 0$. The final model, which has good geometry (Table 3), is composed of residues 148–218 and 27 water molecules. The first two and last 27 residues are not apparent in electron-density maps and have been omitted from the model.

2.7. Structure determination and refinement of $CA_{146–231}$

The refined $CA_{146–p2}$ structure was used as a search model to determine the $CA_{146–231}$ structure by molecular replacement. Using data from 8–3.25 Å and an integration volume of radius 5–14 Å, the third highest solution from the rotation function proved to be the top translation-function solution, with a correlation coefficient of 71.2% (next was 56.8%), and this solution was rigid-body refined to a correlation of 82.4% and an R factor of 29.0% using all 8.0–3.0 Å data.

The structure was refined by the same method as described above for $CA_{146–p2}$. 10% of the data (303 reflections) were withheld from refinement in order to monitor progress by cross-validation and to optimize weights. The refined structure is composed of residues 148–219 and 28 water molecules, with an R factor of 22.3% (R_{free} = 28.1%) using 20–2.6 Å data. $CA_{146–231}$ residues 146–147 and 220–231 are not apparent in electron-density maps and have been omitted from the model. The final coordinates have been deposited in the Protein Data Bank.

2.8. $CA_{146–231}$ structure determination by multiwavelength anomalous diffraction

The $CA_{146–231}$ structure has also been determined by the MAD method (Hendrickson, 1991; Ramakrishnan & Biou, 1997) using data collected from crystals of the selenomethionine-substituted protein. Positions of the three Se atoms were determined from inspection of difference Patterson and difference Fourier maps. The selenium parameters were refined using *MLPHARE* (Otwinowski, 1991), resulting in a mean figure of merit of 0.556. The phases were then refined using solvent flattening and histogram shifting with the program *DM* (Cowtan, 1994). The resulting electron-density map clearly defines the molecular structure from Thr148 to Gln219 (Fig. 1). Although traditional measures of data quality indicate that the MAD data are superior to those collected from the native crystal, refinement of the $CA_{146–231}$ model against the individual MAD data sets was not satisfactory (R factor = 27.2% and R_{free} = 37.1% for λ_1 data; see Table 2). This was surprising given the quality of the MAD map and the success of refinement against native data. A likely explanation is that the MAD data contain errors that impair refinement, but cancel when estimating phases. Nevertheless, the MAD map is clearly interpretable and lacks density for residues 146–147 and 220–231, as do electron-density maps calculated using the native data and model phases.

3. Results and discussion

3.1. Structure determination

Recombinant $CA_{146–231}$ and $CA_{146–p2}$ proteins were expressed in *E. coli* and purified as described in §2. The proteins crystallized isomorphously under similar conditions in space group $I4_1$. There is one molecule in the asymmetric unit, with a crystallographic twofold axis generating the protein dimer. The Matthews' coefficient, V_m, is 2.9 and 2.5 $Å^3\ Da^{-1}$ for $CA_{146–231}$ and $CA_{146–p2}$, respectively, corresponding to estimated solvent contents of 57 and 51%, respectively (Matthews, 1968). Both structures were solved by molecular replacement using the previously determined structure of $CA_{151–231}$ as a search model (Gamble *et al.*, 1997). The structure of $SeCA_{146–231}$ was also determined by multiwavelength anomalous diffraction. The $CA_{146–231}$ and $CA_{146–p2}$ structures have been refined to crystallographic R factors of 22.3 and 20.7% (free R factors of 28.1 and 27.5%), respectively (Table 3).

In both structures, the first two residues (Ser146 and Pro147) as well as C-terminal residues beyond Cys218 ($CA_{146–p2}$) or Gln219 ($CA_{146–231}$) were disordered and were not included in the models. Although both crystal structures had relatively large disordered segments at their C-termini, SDS–PAGE analyses of washed crystals confirmed that these residues were still present in the crystallized proteins. We speculate that the p2 peptide may be disordered in the $CA_{146–p2}$ crystal owing to the low pH of crystallization (5.0), as the p2 peptide may exhibit pH-dependent conformational changes (Pettit *et al.*, 1994). The structures of the ordered segments of $CA_{146–p2}$ and $CA_{146–231}$ are essentially identical, with least-squares overlap giving a root-mean-square deviation of only 0.32 Å for all common atoms. Structural details are therefore only reported for the $CA_{146–p2}$ structure, except where stated.

3.2. Monomer structure

The C-terminal domain of the HIV-1 capsid is composed of a 3_{10} helix, an extended strand and four α-helices that pack into a globular structure of overall dimensions 29 × 36 × 28 Å (Figs. 2 and 3). The monomer structures of $CA_{146–231}$ and $CA_{146–p2}$ closely resemble that of the $CA_{151–231}$ search model (Gamble *et al.*, 1997), with all of the secondary structure elements conserved except for the N-terminal 3_{10} helix (which is formed by residues not present in $CA_{151–231}$). A least-squares overlap on all 68 residues common to $CA_{146–p2}$ and $CA_{151–231}$ gave a root-mean-square deviation of 0.74 Å for Cα atoms and 1.39 Å for all atoms. The structure of the $CA_{151–231}$ monomer, including the extensive hydrogen-bonding network of the MHR sequence, has previously been described in detail (Gamble *et al.*, 1997) and is therefore not discussed further here.

In all three crystal structures of the capsid C-terminal domain, the N-terminus of helix 3 is linked to the C-terminus of helix 4 by a disulfide bond between Cys198 and Cys218. This disulfide is clearly defined in simulated-annealing omit maps of $CA_{146–231}$, $CA_{146–p2}$ and $CA_{151–231}$, as well as in the experimental MAD maps of $CA_{146–231}$ and $CA_{151–231}$. It is possible that disulfide-bond formation is a crystallization artifact, as the recombinant proteins are initially isolated primarily in their reduced forms (see §2). Moreover, the majority of capsid Cys218 (but not Cys198) residues from freshly lysed viral particles can react with biotin maleimide, implying that Cys218 is present as a free thiol in freshly cultured virus (McDermott *et al.*, 1996). Nevertheless, formation of the Cys198–Cys218 disulfide bond must be a favorable reaction because the crystals are grown in the presence of 2 mM 2-mercaptoethanol and begin to appear after just 1–2 d. We therefore speculate that this disulfide bond may also

Figure 2
HIV-1 $CA_{146–231}$/$CA_{146–p2}$ primary sequence and secondary structure. Secondary structures are color-coded as follows: blue, 3_{10} helix (residues 149–152); red, α-helix 1 (residues 161–174); magenta, α-helix 2 (residues 179–192); cyan, α-helix 3 (residues 196–205); green, α-helix 4 (residues 211–217). The N-terminal 3_{10} helix is followed by an inverse γ-turn and six residues in an extended conformation. Secondary structure assignments were taken as defined by *PROMOTIF* (Hutchinson & Thornton, 1996), except that whereas the program defines residues 179–187 and 189–192 as distinct helices, we combine these as one single kinked α-helix because the two helical segments are approximately aligned with each other and the conformation of intervening residues is close to that of a standard helix. The MHR sequence is indicated with a bar and disordered residues are shown with dashed lines.

Table 4
Dimer contacts.

Interface residue†	Hydrophobic contacts (<4.0 Å)	Salt bridge/hydrogen bond‡ (<3.2 Å)	Δ accessibility§ (Å^2)
T148	W184, E187, T188		92.3
I150	T188		5.6
L151	T188, V191, Q192		105.2
D152		K203	19.6
R154			26.6
L172	W184		5.3
E175	W184		17.2
A177	W184		9.2
S178		E180	19.4
E180		S178	48.9
V181	E180, W184		63.8
W184	T148, L172, E175, A177, V181		142.4
M185			32.6
E187	T187		33.0
T188	T148, I150, L151		63.4
L189	L189		26.3
V191	L151		10.4
Q192		N193	93.4
N193		Q192	5.4
K199			9.1
K203		D152	46.1
P207			42.3

† Interface residues are defined as those that apparently bury at least 5 Å^2 of solvent-accessible surface area upon dimer formation. ‡ In all cases the intermolecular hydrogen-bonding interactions are through side-chain atoms. § Estimated change in solvent-accessible surface area upon dimer formation, calculated using *NACCESS* (Hubbard & Thornton, 1993).

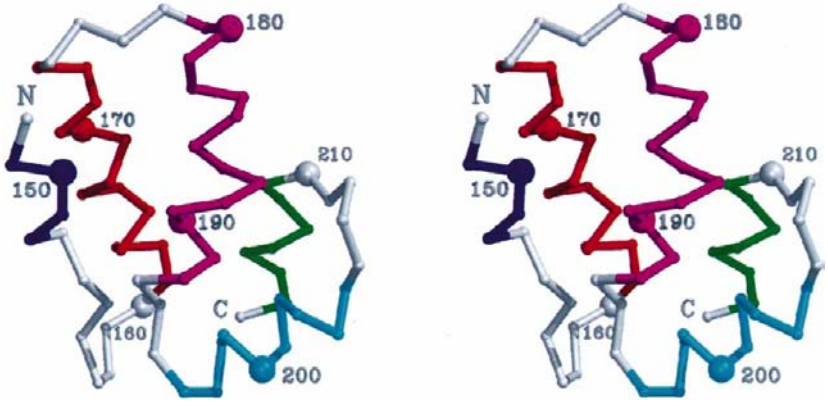

Figure 3
Stereoview of the Cα trace for $CA_{146–231}$/$CA_{146–p2}$. Every tenth residue is labeled and the secondary structure color code is the same as in Fig. 2.

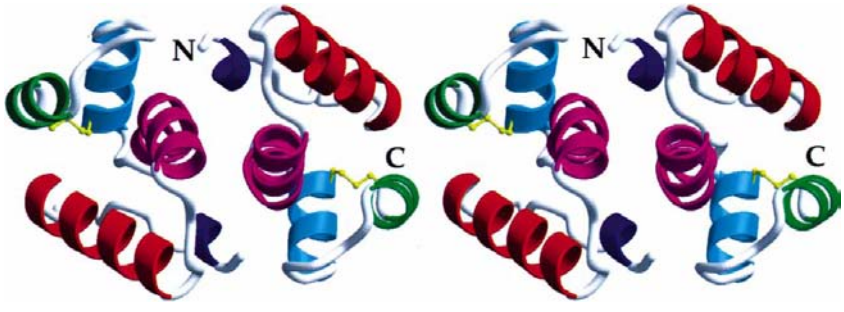

Figure 4
Stereoview ribbon diagram of the $CA_{146–231}$/$CA_{146–p2}$ dimer. The view direction is along the dimer twofold axis, which is from the top of Fig. 3. The color code is the same as for Figs. 2 and 3.

form as the virus leaves the reducing environment of the cell and enters the oxidizing environment of the bloodstream. Conceivably, the disulfide bond could modulate CA–CA interactions and thereby facilitate uncoating of the core structure upon infection. Consistent with this hypothesis, Cys198 and Cys218 are both highly conserved across various strains of HIV-1, HIV-2 and SIV (Myers *et al.*, 1995), and mutation of either residue to serine blocks HIV-1 replication in culture (McDermott *et al.*, 1996).

3.3. Dimer structure

As described above, both $CA_{146–231}$ and $CA_{146–p2}$ dimerize with affinities that are similar to that of the full-length protein (Gamble *et al.*, 1997; Yoo *et al.*, 1997). $CA_{146–231}$ and $CA_{146–p2}$ are therefore expected to exhibit all of the energetically significant dimer interactions made by the intact protein. The isomorphous $CA_{146–231}$ and $CA_{146–p2}$ crystals have one molecule in the asymmetric unit, and both contain the same dimer generated by the action of a crystallographic twofold axis. As shown in Fig. 4, $CA_{146–231}$ and $CA_{146–p2}$ dimerize *via* parallel packing of α-helix 2 (magenta) across the dimer interface. This arrangement also aligns the N-terminal 3_{10} helix (blue) of one monomer to pack into a groove between helices 2 and 3 in the partner molecule. The dimer interface buries a total of 1846 Å^2 of solvent-accessible surface area, with 1222 Å^2 contributed from non-polar side-chain atoms, 490 Å^2 from polar side-chain atoms and the remaining 134 Å^2 from main-chain atoms (Hubbard & Thornton, 1993). Hydrophilic interactions across the dimer interface include two salt bridges (Asp152–Lys203 and their symmetry equivalents) and four intermolecular hydrogen bonds (Gln192 OE1···Asn193 ND2 and Ser178 OG···Glu180 OE1 and their symmetry equivalents). Hydrophobic interactions across the interface include the complete burial of Met185, the packing of Trp184 against Thr148, Leu172, Glu175, Ala177 and Val181, and the packing of Leu151 against Thr188, Val191 and Gln192 (see Table 4 and Fig. 5 for details of dimer contacts). Consistent with the relevance of this interface for capsid dimerization, mutation of either Trp184 or Met185 to Ala abolishes detectable dimerization of the intact capsid protein as analyzed by equilibrium sedimentation (Gamble *et al.*, 1997). Moreover, the Met185

to Ala mutation also blocks capsid assembly *in vitro* and HIV-1 replication in culture (von Schwedler, 1998; von Schwedler *et al.*, 1998).

The presence of the N-terminal 3_{10} helix in the $CA_{146\text{-}231}$/$CA_{146\text{–}p2}$ proteins explains why these proteins dimerize more efficiently than the shorter $CA_{151\text{–}231}$ construct. As shown in Figs. 4 and 5, the 3_{10} helix packs into the dimer interface and orients Leu151 to make a series of hydrophobic intermolecular contacts. In contrast, although $CA_{151\text{–}231}$ molecules also close-packed in the crystal using the hydrophobic face of α-helix 2, their crystal-packing interactions are distinctly different, with a $\sim$60° change in the relative orientation of individual monomers. As expected, the $CA_{146\text{–}231}$ interface is much more extensive, exhibiting two more salt bridges, two more hydrogen bonds and burying 70% more solvent-accessible surface area than the $CA_{151\text{–}231}$ interface. Details of the side-chain packing interactions also necessarily differ between the two interfaces. For example, the χ_2 angle of Trp184 differs by approximately 110° between the two structures.

The three-dimensional structure of the N-terminal domain of the HIV-1 capsid has also been determined in solution (Gitti *et al.*, 1996) and in crystals as complexes with an Fab fragment (Momany *et al.*, 1996) and cyclophilin A (Gamble *et al.*, 1996). In each case, the last ordered residue of the N-terminal domain was Tyr145. As described above, the first ordered residue of the C-terminal domain in the structures of $CA_{146\text{–}231}$ and $CA_{146\text{–}p2}$ is Thr148. We have therefore modeled the intact capsid protein by assuming that the N- and C-terminal domains of the capsid are accurately represented by the isolated constructs and that the range of possible orientations between these two domains is constrained by covalent linkage through residues Ser146 and Pro147. Such a model for the intact capsid protein dimer is shown in Fig. 6.

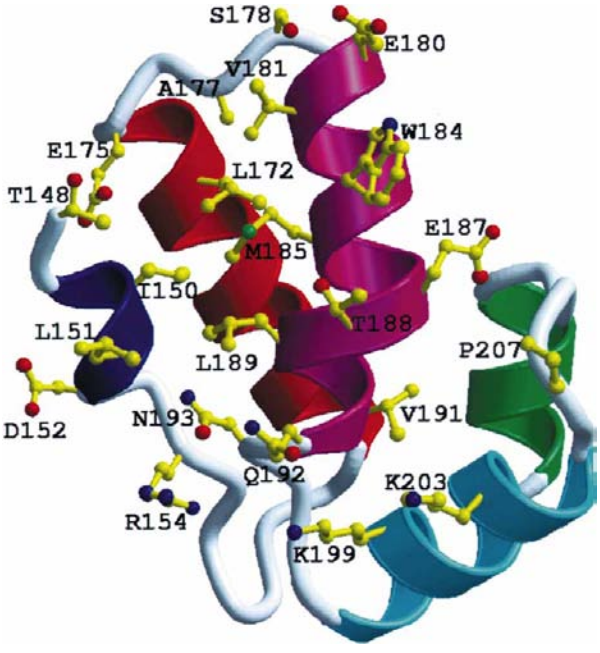

Figure 5
View of the $CA_{146\text{–}231}$/$CA_{146\text{–}p2}$ dimer-interface residues. Residues that appear to lose more than 5 Å^2 of solvent-accessible surface area upon dimer formation are shown explicitly. Orientation and color code are the same as for Fig. 3.

Figure 6
Model of the full-length capsid protein. N-terminal domains (residues 1–145) are colored magenta and C-terminal domain dimer (residues 148–219) are colored blue. The disordered residues Ser146 and Pro147 which link the N- and C-terminal domains are colored gray. The C-terminal dimer is oriented with the twofold axis vertical. Atomic models of the full-length capsid protein were built using the crystal structures of the N- and C-terminal domains of capsid (Gamble *et al.*, 1996). The last ordered residue of the N-terminal domain (Tyr145) was covalently connected to the first ordered residue of the C-terminal domain (Thr148) with a dipeptide linker (Ser146–Pro147). Plausible conformations for the Ser146–Pro147 linker allowed models to be built that exhibited a wide ($\sim$90°) range of relative rotations between the two domains.

We speculate that the Ser146–Pro147 dipeptide may function as a semi-flexible linker that allows the N- and C-terminal domains to adopt a range of relative orientations in the intact protein, as may be required if capsid molecules reside in different local environments at the narrow and wide ends of the conical viral core. The use of a flexible interdomain linker to accommodate local packing variations seems more probable than supposing that intermolecular CA–CA contacts, particularly the tight-packed C-terminal dimer interface, can adopt a range of different orientations.

In summary, the genetic, biochemical and structural data indicate that the $CA_{146\text{–}231}$/$CA_{146\text{–}p2}$ dimers represent the authentic dimer interaction formed by the C-terminal domain of the intact HIV-1 capsid protein. The detailed structure of this essential interface therefore provides a new target for the structure-based design of therapeutic agents for the treatment of AIDS.

We wish to thank Robert M. Sweet, Joel Berendzen, Leon Flaks, Michael Mathews and Su Li for assistance with data collection. This work was supported by grants to WIS and CPH from the National Institutes of Health (R01 AI40333 and R01 AI43036).

References

Adachi, A., Gendelman, H. E., Koenig, S., Folks, T., Willey, R., Rabson, A. & Martin, M. A. (1986). *J. Virol.* **59**, 284–291.

Brooks, I., Watts, D. G., Soneson, K. K. & Hensley, P. (1994). *Methods Enzymol.* **240**, 459–478.

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

Brünger, A. T. (1992*b*). *X-PLOR.* Version 3.1. *A System for X-ray Crystallography and NMR*. Yale University, Connecticut, USA.

Carrière, C., Gay, B., Chazal, N., Morin, N. & Boulanger, P. (1995). *J. Virol.* **69**, 2366–2377.

Chazal, N., Carrière, C., Gay, B. & Boulanger, P. (1994). *J. Virol.* **68**, 111–122.

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

Craven, R. C., Leure-duPree, A. E., Weldon, R. A. Jr & Wills, J. W. (1995). *J. Virol.* **69**, 4213–4227.

Dorfman, T., Bukovsky, A., Öhagen, Å., Höglund, S. & Göttlinger, H. G. (1994). *J. Virol.* **68**, 8180–8187.

Ehrlich, L. S., Agresta, B. E. & Carter, C. A. (1992). *J. Virol.* **66**, 4874–4883.

Ehrlich, L. S., Agresta, B. E., Gelfand, C. A., Jentoft, J. & Carter, C. A. (1994). *Virology*, **204**, 515–525.

Ellman, G. L. (1959). *Arch. Biochem. Biophys.* **82**, 70–77.

Erickson-Viitanen, S., Manfredi, J., Viitanen, P., Tribe, D. E., Tritch, R., Huttchinson, C. A. I., Loeb, D. D. & Swanstrom, R. (1989). *AIDS Res. Hum. Retrovir.* **5**, 577–591.

Franke, E. K., Yuan, H. E., Bossolt, K. L., Goff, S. P. & Luban, J. (1994). *J. Virol.* **68**, 5300–5305.

Franke, E. K., Yuan, H. E. & Luban, J. (1994). *Nature (London)*, **372**, 359–362.

Gamble, T. R., Vajdos, F. F., Yoo, S., Worthylake, D. K., Houseweart, M., Sundquist, W. I. & Hill, C. P. (1996). *Cell*, **87**, 1285–1294.

Gamble, T. R., Yoo, S., Vajdos, F. F., von Schwedler, U. K., Worthylake, D. K., Wang, H., McCutcheon, J. P., Sundquist, W. I. & Hill, C. P. (1997). *Science*, **278**, 849–853.

Gill, S. C. & von Hippel, P. H. (1989). *Anal. Biochem.* **182**, 319–326.

Gitti, R. K., Lee, B. M., Walker, J., Summers, M. F., Yoo, S. & Sundquist, W. I. (1996). *Science*, **273**, 231–235.

Gowda, S. D., Stein, B. S. & Engleman, E. G. (1989). *J. Biol. Chem.* **264**, 8459–8462.

Hendrickson, W. A. (1991). *Science*, **254**, 51–58.

Hong, S. S. & Boulanger, P. (1993). *J. Virol.* **67**, 2787–2798.

Hubbard, S. J. & Thornton, J. M. (1993). *NACCESS*. Department of Biochemistry and Molecular Biology, University College, London.

Hutchinson, E. G. & Thornton, J. M. (1996). *Protein Sci.* **5**, 212–220.

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

Jowett, J. B., Hockley, D. J., Nermut, M. V. & Jones, I. M. (1992). *J. Gen. Virol.* **73**, 3079–3086.

Kräusslich, H.-G. (1996) Editor. *Current Topics in Immunology*, Vol. 214, *Morphogenesis and Maturation of Retroviruses.* Berlin: Springer–Verlag.

Kräusslich, H.-G., Ingraham, R. H., Skoog, M. T., Wimmer, E., Pallai, P. V. & Carter, C. A. (1989). *Proc. Natl Acad. Sci. USA*, **86**, 807–811.

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

McDermott, J., Farrell, L., Ross, R. & Barklis, E. (1996). *J. Virol.* **70**, 5106–5114.

Mammano, F., Öhagen, Å., Höglund, S. & Göttlinger, H. G. (1994). *J. Virol.* **68**, 4927–4936.

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

Mervis, R. J., Ahmad, N., Lillehoj, E. P., Raum, M. G., Salazar, F. H., Chan, H. W. & Venkatesan, S. (1988). *J. Virol.* **62**, 3993–4002.

Momany, C., Kovari, L. C., Prongay, A. J., Keller, W., Gitti, R. K., Lee, B. M., Gorbalenya, A. E., Tong, L., McClure, J., Ehrlich, L. S., Summers, M. F., Carter, C. & Rossmann, M. G. (1996). *Nature Struct. Biol.* **3**, 763–770.

Myers, G., Korber, B., Hahn, B. H., Jeang, K.-T., Mellors, J. W., McCutchan, F. E., Henderson, L. E. & Pavlakis, G. N. (1995). Editors. *Human Retroviruses and AIDS*. Los Alamos, NM: Los Alamos National Laboratory.

Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.

Orlinsky, K. J., Gu, J., Hoyt, M., Sandmeyer, S. & Menees, T. M. (1996). *J. Virol.* **70**, 3440–3448.

Otwinowski, Z. (1991). *Isomorphous Replacement and Anomalous Scattering*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80–86. Warrington: Daresbury Laboratory.

Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.

Patarca, R. & Haseltine, W. A. (1985). *Nature (London)*, **318**, 390.

Pettit, S. C., Moody, M. D., Wehbie, R. S., Kaplan, A. H., Nantermet, P. V., Klein, C. A. & Swanstrom, R. (1994). *J. Virol.* **68**, 8017–8027.

Poblotzki, A. von, Wagner, R., Niedrig, M., Wanner, G., Wolf, H. & Modrow, S. (1993). *Virology*, **193**, 981–985.

Ramakrishnan, V. & Biou, V. (1997). *Methods Enzymol.* **276**, 538–557.

Reicin, A. S., Ohagen, A., Yin, L., Hoglund, S. & Goff, S. P. (1996). *J. Virol.* **70**, 8645–8652.

Reicin, A. S., Paik, S., Berkowitz, R. D., Luban, J., Lowy, I. & Goff, S. P. (1995). *J. Virol.* **69**, 642–650.

Rosé, S., Hensley, P., O'Shannessy, D. J., Culp, J., Debouck, C. & Chaiken, I. (1992). *Proteins*, **13**, 112–119.

Schägger, H. & von Jagow, G. (1987). *Anal. Biochem.* **166**, 368–379.

Schwedler, U. von (1998). Personal communication.

Schwedler, U. K. von, Stemmler, T. L., Klishko, V. Y., Li, S., Albertine, K. H., Davis, D. R. & Sundquist, W. I. (1998). *EMBO J.* **17**, 1555–1568.

Srinivasakumar, N., Hammarskjöld, M.-L. & Rekosh, D. (1995). *J. Virol.* **69**, 6106–6114.

Strambio-de-Castilla, C. & Hunter, E. (1992). *J. Virol.* **66**, 3232–3242.

Studier, F. W., Rosenberg, A. H., Dunn, J. J. & Dubendorff, J. W. (1990). *Methods Enzymol.* **185**, 60–89.

Thali, M., Bukovsky, A., Kondo, E., Rosenwirth, B., Walsh, C. T., Sodroski, J. & Göttlinger, H. G. (1994). *Nature (London)*, **372**, 363–365.

Wang, C.-T. & Barklis, E. (1993). *J. Virol.* **67**, 4264–4273.

Wills, J. W. & Craven, R. C. (1991). *AIDS*, **5**, 639–654.

Yoo, S., Myszka, D. G., Yeh, C.-Y., McMurray, M., Hill, C. P. & Sundquist, W. I. (1997). *J. Mol. Biol.* **269**, 780–795.

Zhang, W.-H., Hockley, D. J., Nermut, M. V., Morikawa, Y. & Jones, I. M. (1996). *J. Gen. Virol.* **77**, 743–751.

Zhao, Y., Jones, I. M., Hockley, D. J., Nermut, M. V. & Roy, P. (1994). *Virology*, **199**, 403–408.