nature structural & molecular biology

Crystal structure of XMRV protease differs from the structures of other retropepsins

Mi Li[1,2], Frank DiMaio[3], Dongwen Zhou[1], Alla Gustchina[1], Jacek Lubkowski[4], Zbigniew Dauter[5], David Baker[3] & Alexander Wlodawer[1]

Using energy and density guided Rosetta refinement to improve molecular replacement, we determined the crystal structure of the protease encoded by xenotropic murine leukemia virus–related virus (XMRV). Despite overall similarity of XMRV protease to other retropepsins, the topology of its dimer interface more closely resembles those of the monomeric, pepsin-like enzymes. Thus, XMRV protease may represent a distinct branch of the aspartic protease family.

XMRV is a newly discovered human retrovirus and the first gammaretrovirus shown to be associated with human diseases. It has been detected in prostate cancer cells[1] as well as in individuals with chronic fatigue syndrome[2]. Although the identification of XMRV as the causal agent for these diseases is still controversial[3], it seems prudent to identify targets for drugs against this potential pathogen. Because XMRV is a retrovirus, inhibition of the three enzymes encoded in its genome (reverse transcriptase, integrase and protease) provides the most direct path to inactivation of the virus. It has already been shown that the integrase inhibitor raltegravir is a potent inhibitor of XMRV[4]. Enzyme inhibition has been a very successful route for developing therapeutic agents against human immunodeficiency virus (HIV). In particular, numerous drugs targeting HIV-1 protease have been developed in the last 20 years[5]. The success of these efforts depended very much on the availability of the structure of HIV-1 protease, both as an apoenzyme and in complexes with inhibitors[6]. Although all retroviral proteases studied to date are structurally similar[7], the fine differences in their structures allow for the development of specific inhibitors. For example, although HTLV-1 protease[8] is similar to HIV-1 protease[9], it is very poorly inhibited by most HIV-1 protease inhibitors. None of the clinical inhibitors of HIV-1 protease have EC_{50} values below 35 μM against XMRV in cell culture, which is three to four orders of magnitude higher compared to HIV-1 (ref. 4).

Although XMRV protease has not been previously isolated or expressed and characterized on a molecular level, a closely related enzyme from Moloney murine leukemia virus (MoMLV) has been isolated and its amino acid sequence determined[10]. This information served as a guide in cloning XMRV protease (**Supplementary Methods**) and particularly in deciding the location of its probable termini. The expression construct contains 125 amino acids belonging to the enzyme, as well as a N-terminal hexahistidine tag preceded by a methionine. The enzyme migrates as a dimer on a gel-filtration column (data not shown). Its activity was demonstrated by extensive autolysis (**Supplementary Fig. 1a**) and by its cleavage of maltose-binding protein (MBP) in the MBP-XMRV fusion protein (**Supplementary Fig. 1b**). This autolysis was inhibited by TL-3, a broad-specificity retropepsin inhibitor (**Supplementary Methods** and **Supplementary Fig. 1**). This construct of XMRV protease was purified and crystallized, and diffraction data were collected to 1.97-Å resolution (**Supplementary Methods**).

Because XMRV protease contains only a single methionine, near its C terminus (Met118), phasing of diffraction data by using anomalous dispersion of selenomethionine seemed unlikely without the introduction of additional methionine residues. However, the structural similarity of all known retroviral proteases[7] suggested that molecular replacement should be sufficient for solving the structure of XMRV protease. We carried out extensive trials with models built on the basis of crystal structures of several retroviral proteases but found no refinable solutions (**Supplementary Methods**). We finally solved the structure of XMRV protease through a novel application of the Rosetta refinement[11] to several highest-scoring molecular replacement models. This application of the Rosetta refinement produced sufficient improvement of these structures to enhance the molecular replacement signal and resulted in a model that could be further refined by standard means (**Supplementary Methods** and **Table 1**).

A molecule of XMRV protease is a homodimer (**Fig. 1a**), with a two-fold symmetry axis that does not coincide with the symmetry elements of the crystal. Its fold generally resembles those of other retroviral proteases (**Fig. 1b**), although with several substantial differences, especially at the two termini. Both the N and C termini are longer in XMRV protease than in most other retropepsins. The N terminus contains a helical insertion before strand β1 (**Fig. 1c**). Instead of the interdigitated N and C termini (β1 and β9 strands, **Fig. 1c**) that create the dimer interfaces in all other structurally characterized retroviral proteases, the dimer interface of XMRV protease utilizes hairpins formed by strands β10 and β11, near the C termini of both monomers (**Figs. 1c** and **2a** and **Supplementary Figs. 2a** and **3**). The flaps of each protomer (residues 48–66) are partially disordered at their tips, a situation common for the apoenzymes of retropepsins[12]. However, the ordered parts of the flaps appear to represent the open conformation

[1]Protein Structure Section, Macromolecular Crystallography Laboratory, National Cancer Institute at Frederick, Frederick, Maryland, USA. [2]Basic Research Program, SAIC-Frederick, Frederick, Maryland, USA. [3]Department of Biochemistry, University of Washington, Seattle, Washington, USA. [4]Macromolecular Assembly Structure and Cell Signaling Section, Macromolecular Crystallography Laboratory, National Cancer Institute at Frederick, Frederick, Maryland, USA. [5]Synchrotron Radiation Research Section, Macromolecular Crystallography Laboratory, National Cancer Institute, Argonne National Laboratory, Argonne, Illinois, USA. Correspondence should be addressed to A.W. (wlodawer@nih.gov).

Received 29 July; accepted 29 October 2010; published online 23 January 2011; doi:10.1038/nsmb.1964

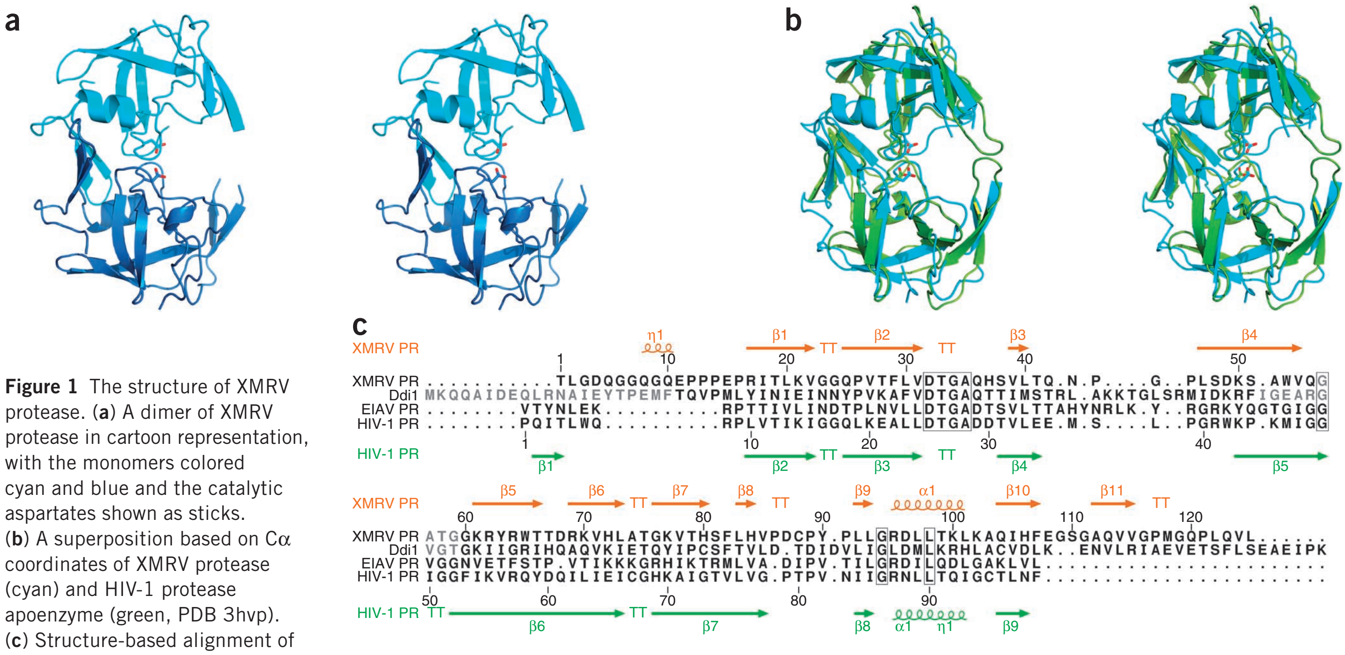
a
b
c
XMRV PR
η1
β1
TT
β2
TT
β3
β4
1 10 20 30 40 50
XMRV PR ............TLGDQGGQGQEPPPEPRITLKVGGQPVTFLVDTGAQHSVLTQ.N.P....G..PLSDKS.AWVQG
Ddi1 MKQQAIDEQLRNAIEYTPEMFTQVPMLYINIEINNYPVKAFVDTGAQTTIMSTRL.AKKTGLSRMIDKRFIGEARG
EIAV PR .........VTYNLEK..........RPTTIVLINDTPLNVLLDTGADTSVLTTAHYNRLK.Y..RGRKYQGTGIGG
HIV-1 PR .........PQITLWQ..........RPLVTIKIGGQLKEALLDTGADDTVLEE.M.S....L..PGRWKP.KMIGG
1 10 20 30 40
HIV-1 PR
β1
β2
TT
β3
TT
β4
β5
XMRV PR
β5
β6
TT
β7
β8
TT
β9
α1
β10
β11
TT
60 70 80 90 100 110 120
XMRV PR ATGGKRYRWTTDRKVHLATGKVTHSFLHVPDCPY.PLLGRDLLTKLKAQIHFEGSGAQVVGPMGQPLQVL......
Ddi1 VGTGKIIGRIHQAQVKIETQYIPCSFTVLD.TDIDVLIGLDMLKRHLACVDLK..ENVLRIAEVETSFLSEAEIPK
EIAV PR VGGNVETFSTP.VTIKKKGRHIKTRMLVA.DIPV.TILGRDILQDLGAKLVL...................
HIV-1 PR IGGFIKVRQYDQILIEICGHKAIGTVLVG.PTPV.NIIGRNLLTQIGCTLNF.......................
50 60 70 80 90
HIV-1 PR
TT
β6
TT
β7
β8
α1
η1
β9

Figure 1 The structure of XMRV protease. (**a**) A dimer of XMRV protease in cartoon representation, with the monomers colored cyan and blue and the catalytic aspartates shown as sticks. (**b**) A superposition based on Cα coordinates of XMRV protease (cyan) and HIV-1 protease apoenzyme (green, PDB 3hvp). (**c**) Structure-based alignment of the XMRV, HIV-1 and EIAV proteases, as well as Ddi1 RP. Secondary structure elements and residue numbers are marked for XMRV and HIV-1 proteases. Residues identical in all four enzymes are boxed. Panels **a** and **b** are stereoviews prepared with PyMOLI[18].

seen in the apo form of HIV-1 protease[13]. The N-terminal fragment of XMRV protease is partially helical, with residues Gly6 through Glu11 disordered in monomer B, and is quite different from its counterparts in other retroviral enzymes (**Supplementary Fig. 4**).

Although the mode of dimerization of XMRV protease shows substantial differences from those of other retropepsins (**Fig. 2b**), it is much closer to that of the putative protease (RP) domain of the eukaryotic protein Ddi1 (ref. 14). The crystal structure of the isolated RP domain of *Saccharomyces cerevisiae* Ddi1 was solved and refined at 2.3-Å resolution (PDB code 2I1A; ref. 14), revealing similarity in the overall structural fold to retropepsins. However, to our knowledge, no enzymatic activity of Ddi1 RP has been reported. The overall structural similarity (**Supplementary Fig. 5**) of XMRV protease and Ddi1 RP is reflected by the r.m.s. deviations of 1.66 Å and 1.87 Å between the equivalent 85 Cα atoms in the monomers and 174 Cα atoms in the dimers of both proteins, respectively. By comparison, an analogous alignment of XMRV protease with the apo form of HIV-1 protease (PDB code 3HVP; ref. 13) yields r.m.s. deviations of 2.18 Å for the monomers and 2.35 Å for the dimers.

Like those of XMRV protease, the N and C termini of Ddi1 RP are substantially longer than in a majority of retropepsins. The dimer interface in Ddi1 RP is formed solely by the C-terminal part of the protomer (by three consecutive β strands, β7–β9; **Fig. 1c**) and does not include the N terminus at all (**Fig. 2c**). A comparable situation is seen in XMRV protease, except that the interface uses only two β strands (β10–β11). Residues Gly119 and Gln120 make a turn after β11 and form hydrogen bonds with the O and N atoms of Gly116, thus extending the sheet, but the following segment of the C-terminal chain does not form any regular structure and points in a completely different direction (**Fig. 2a** and **Supplementary Fig. 2a**).

As noted in the description of the structure of Ddi1 RP[14], β strands that form the dimer interface in that protein are rotated by ~45° compared to their counterparts in HIV-1 protease and other retropepsins. Two of these strands in XMRV protease superimpose almost exactly on their counterparts in Ddi1 RP, retaining their angles, with only the residues at the turn between the interface strands following a slightly different path in the two proteins, despite their identical length (**Supplementary Fig. 6**). The axis of the dimer interface β-sheet in XMRV protease is aligned roughly perpendicular to the long axis of the protease dimer. The direction of the interface strands and the lack of interdigitation resembles a situation seen in pepsin-like aspartic proteases, with the caveat that the dimerization interface in the latter enzymes is six-stranded (as in Ddi1 RP), in contrast to the four-stranded interface in XMRV protease (**Fig. 2** and **Supplementary Fig. 2a**). This structure of the interface sheet results in a much smaller number of contacts with the opposite protomer in the dimer compared to other retropepsins, in which extensive intermolecular contacts are created by interdigitation of the C- and N-terminal β strands. Nonetheless, XMRV protease is dimeric in solution as well as in crystals.

The two β-strands that follow helix α1 in XMRV protease and Ddi1 RP and form the dimer interface are both topologically and structurally equivalent to the corresponding C-terminal loops of each domain of pepsin-like aspartic proteases (**Fig. 2** and **Supplementary Fig. 2a**), whereas the third strand is missing in XMRV protease. In this respect, XMRV protease seems to be closer than the other retroviral proteases to the putative common ancestor of monomeric and dimeric aspartic proteases[15], indicating divergence in their evolutionary paths.

A unique structural organization of N and C termini in XMRV protease leads to differences in the intersubunit interactions within the dimer interface compared to other retroviral enzymes. An important interaction stabilizing the dimers of retroviral proteases is created by an ion pair involving Arg8 of one protomer and Asp29′ of the other one (HIV-1 protease numbering) (**Supplementary Fig. 2b**). In contrast to all other characterized retropepsins, in XMRV protease these two residues are not conserved. A residue equivalent to Arg8 is Glu15 (**Fig. 1c**), but its side chain faces an opposite direction because the following Pro16 adopts a *cis* conformation. Although Pro16 is conserved among retroviral proteases, the *trans* conformation of this residue in most of these enzymes leads to observed differences in topologies in the N-terminal strand. Gln36 in XMRV protease is equivalent to Asp29 in HIV-1 protease, and their respective side chains, in addition to differing in their ionic state, are also oriented differently. Although simian foamy

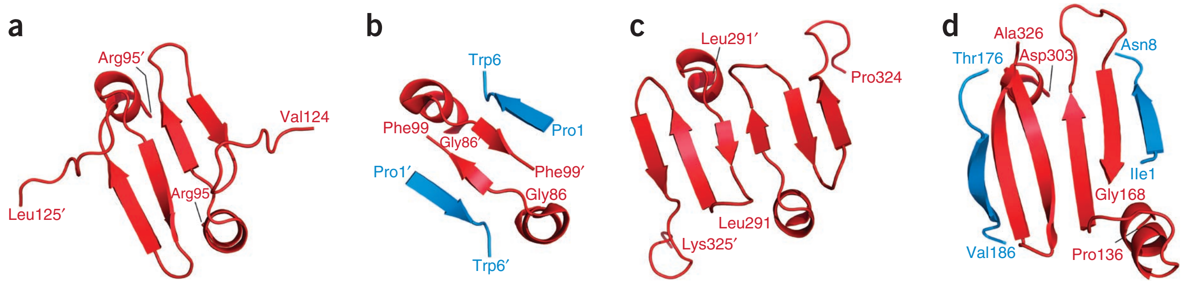

Figure 2 Dimer interface regions in aspartic proteases. Strands belonging to the N-terminal regions of the molecules (or domains in pepsin) are blue, and the C-terminal regions are red. (**a**) XMRV protease. (**b**) HIV-1 protease. (**c**) Ddi1 RP. (**d**) Pepsin.

virus protease also lacks a corresponding ion pair, its structure has been characterized by NMR only for a monomer[16], and thus its dimer interface cannot be analyzed. These intersubunit ionic interactions are substituted by hydrophobic contacts in XMRV protease (**Supplementary Fig. 2b**), thus modifying the network of interactions within the dimer interface. It must be pointed out, however, that mutation R8Q in HIV-1 protease, which replaces the ion pair with polar interactions, leads to only small differences in the activity of the enzyme[17], indicating that the presence of an ion pair may not be necessary to stabilize the dimer.

As for other retropepsins crystallized in the absence of ligands, a water molecule bridges the two catalytic aspartates. The architecture of the active site in XMRV protease, particularly the hydrophobic lining of the binding site area, also resembles those of other retropepsins, suggesting that this enzyme might have similar substrate-recognition preferences. As an example, the loop Leu83–Leu92, equivalent to the so-called polyproline loop in HIV-1 protease (residues Leu76–Ile84), adopts a conformation in XMRV protease that is very similar to that in other retroviral enzymes, but contrasts with the one found in Ddi1 RP (**Supplementary Fig. 7**). As revealed by numerous structures of inhibitor complexes of retropepsins, residues of this loop are involved in extensive interactions with the ligands. Therefore, although the only structure of XMRV protease currently available is that of the apoenzyme form, overall conservation of the structural features of retropepsins in the active site area allows prediction of the putative subsites for the residues of substrates and/or peptidic inhibitors. The residues predicted to form subsites S1–S4 in the monomer of XMRV protease are compared with their equivalents in HIV-1 and EIAV proteases in **Supplementary Table 2**. Although the predominantly hydrophobic character of the binding sites is well preserved as a result of the conservative nature of a majority of substitutions, the presence in XMRV protease of unique polar residues such as His37 in S2 and S4, Tyr90 in S1 and S3, and Gln36 and Gln55 (presumably, since the fragment of the flap with this residue is disordered) in S3 and S5 provides clues for the design of specific inhibitors against XMRV protease. The other important difference observed in pocket S3 is due to the lack of conservation in XMRV protease of the previously mentioned Arg8 and Asp29 that form part of this pocket in the other retroviral enzymes. Further studies with substrates and inhibitors of XMRV protease will be necessary to define the specificity of this enzyme and to design more effective inhibitors.

Accession codes. Protein Data Bank: Coordinates and structure factors have been deposited with the accession code 3NR6.

Note: An enhanced version of this article and supplementary information are available on the Nature Structural & Molecular Biology website.

ACKNOWLEDGMENTS
We acknowledge the use of beamline 22-ID of the Southeast Regional Collaborative Access Team (SER-CAT), located at the Advanced Photon Source (APS), Argonne National Laboratory. Use of the APS was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, under contract no. W-31-109-Eng-38. This work was supported in part by the Intramural Research Program of the US National Institutes of Health, National Cancer Institute, Center for Cancer Research and with federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does the mention of trade names, commercial products or organizations imply endorsement by the US Government.

AUTHOR CONTRIBUTIONS

M.L. produced the protein and grew crystals; M.L. and Z.D. collected and processed crystallographic data; M.L., F.D.M., D.Z., A.G., J.L., and Z.D. performed calculations, structure refinement and analysis; D.B. and A.W. supervised the project. All authors discussed the results and participated in writing the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/nsmb/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Schlaberg, R., Choe, D.J., Brown, K.R., Thaker, H.M. & Singh, I.R. *Proc. Natl. Acad. Sci. USA* **106**, 16351–16356 (2009).
2. Lombardi, V.C. *et al. Science* **326**, 585–589 (2009).
3. Groom, H.C. *et al. Retrovirology* **7**, 10 (2010).
4. Singh, I.R., Gorzynski, J.E., Drobysheva, D., Bassit, L. & Schinazi, R.F. *PLoS ONE* **5**, e9948 (2010).
5. Wlodawer, A. & Vondrasek, J. *Annu. Rev. Biophys. Biomol. Struct.* **27**, 249–284 (1998).
6. Wlodawer, A. & Erickson, J.W. *Annu. Rev. Biochem.* **62**, 543–585 (1993).
7. Dunn, B.M., Goodenow, M.M., Gustchina, A. & Wlodawer, A. *Genome Biol.* **3**, S3006.1–S3006.7 (2002).
8. Gustchina, A., Jaskolski, M. & Wlodawer, A. *Cell Cycle* **5**, 463–464 (2006).
9. Li, M. *et al. Proc. Natl. Acad. Sci. USA* **102**, 18322–18337 (2005).
10. Yoshinaka, Y., Katoh, I., Copeland, T.D. & Oroszlan, S. *Proc. Natl. Acad. Sci. USA* **82**, 1618–1622 (1985).
11. DiMaio, F., Tyka, M.D., Baker, M.L., Chiu, W. & Baker, D. *J. Mol. Biol.* **392**, 181–190 (2009).
12. Miller, M., Jaskólski, M., Rao, J.K.M., Leis, J. & Wlodawer, A. *Nature* **337**, 576–579 (1989).
13. Wlodawer, A. *et al. Science* **245**, 616–621 (1989).
14. Sirkis, R., Gerst, J.E. & Fass, D. *J. Mol. Biol.* **364**, 376–387 (2006).
15. Tang, J., James, M.N.G., Hsu, I.N., Jenkins, J.A. & Blundell, T.L. *Nature* **271**, 618–621 (1978).
16. Hartl, M.J., Wohrl, B.M., Rosch, P. & Schweimer, K. *J. Mol. Biol.* **381**, 141–149 (2008).
17. Louis, J.M., Clore, G.M. & Gronenborn, A.M. *Nat. Struct. Biol.* **6**, 868–875 (1999).
18. DeLano, W.L. *The PyMOL Molecular Graphics System* (DeLano Scientific, San Carlos, California, USA, 2002).

Supplementary Figures, Tables, and Methods for:

Crystal structure of XMRV protease differs from the structures of other retropepsins

Mi Li[1,2], Frank DiMaio[3], Dongwen Zhou[1], Alla Gustchina[1], Jacek Lubkowski[4], Zbigniew Dauter[5], David Baker[3] and Alexander Wlodawer[1]*

[1]Protein Structure Section, Macromolecular Crystallography Laboratory, National Cancer Institute at Frederick, Frederick, MD 21702, USA

[2]Basic Research Program, SAIC-Frederick, Frederick, MD 21702, USA

[3]Department of Biochemistry, University of Washington, Seattle, WA 98195, USA

[4]Macromolecular Assembly Structure and Cell Signaling Section, Macromolecular Crystallography Laboratory, National Cancer Institute at Frederick, Frederick, MD 21702, USA

[5]Synchrotron Radiation Research Section, Macromolecular Crystallography Laboratory, NCI, Argonne National Laboratory, Argonne, IL 60439, USA

*To whom correspondence should be addressed:
National Cancer Institute, MCL
Bldg. 536, Rm. 5
Frederick, Maryland 21702-1201
Phone: +1-301-846-5036
Fax: +1-301-846-6322
email: wlodawer@nih.gov

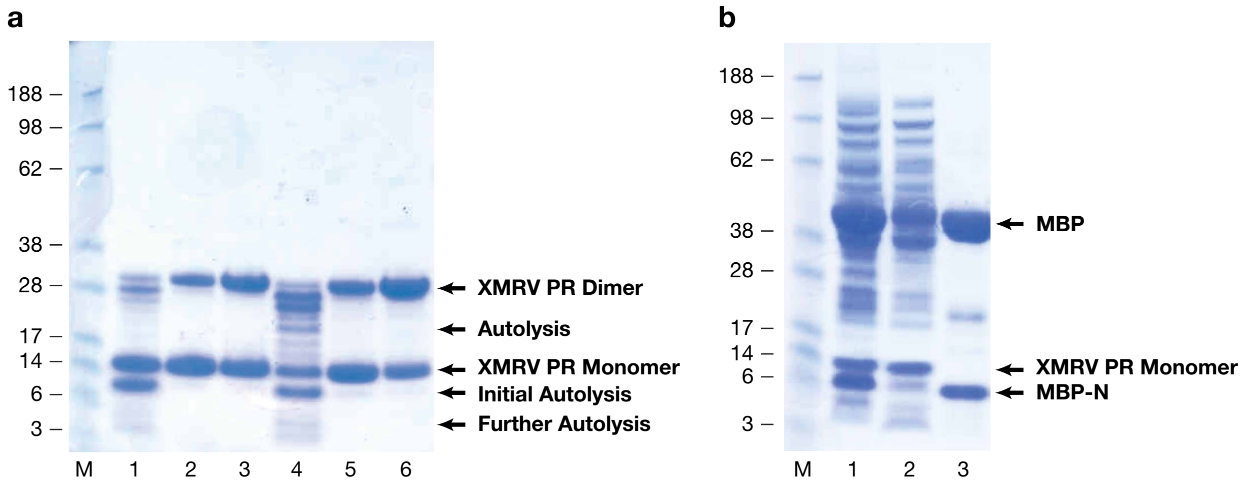

Figure S1. XMRV PR is enzymatically active. **a)** A gel showing that His6-tagged construct of XMRV PR is capable of autolysis. The enzyme was incubated for 2.5 h (lanes 1-3) or overnight (lanes 4-6) at 25°, pH 5.5, in the absence (lanes 1, 4) or presence of inhibitor TL-3 at concentration of ~1 mM (lanes 2, 5) or ~5 mM (lanes 3, 6). Protein was denatured with SDS before application to the gel. Identity of the bands was established by sequencing and mass spectrometry. Lane M – molecular mass standards. Under such circumstances the protein should be migrating as a monomer, thus the presence of the band representing a dimer is due to creation of a disulfide bond between Cys88 of two monomers during the procedure. The dimer band is absent when DTT is added (not shown). **b)** Autolytic cleavage of a construct consisting of the N-terminally His6-tagged maltose binding protein followed by XMRV PR. Lane M – molecular mass standards. Lane 1, supernatant after cells were broken. Lane 2, flow-through of a Ni column. Lane 3, main fraction eluting from Ni column. Identity of the bands was established by sequencing and mass spectrometry.

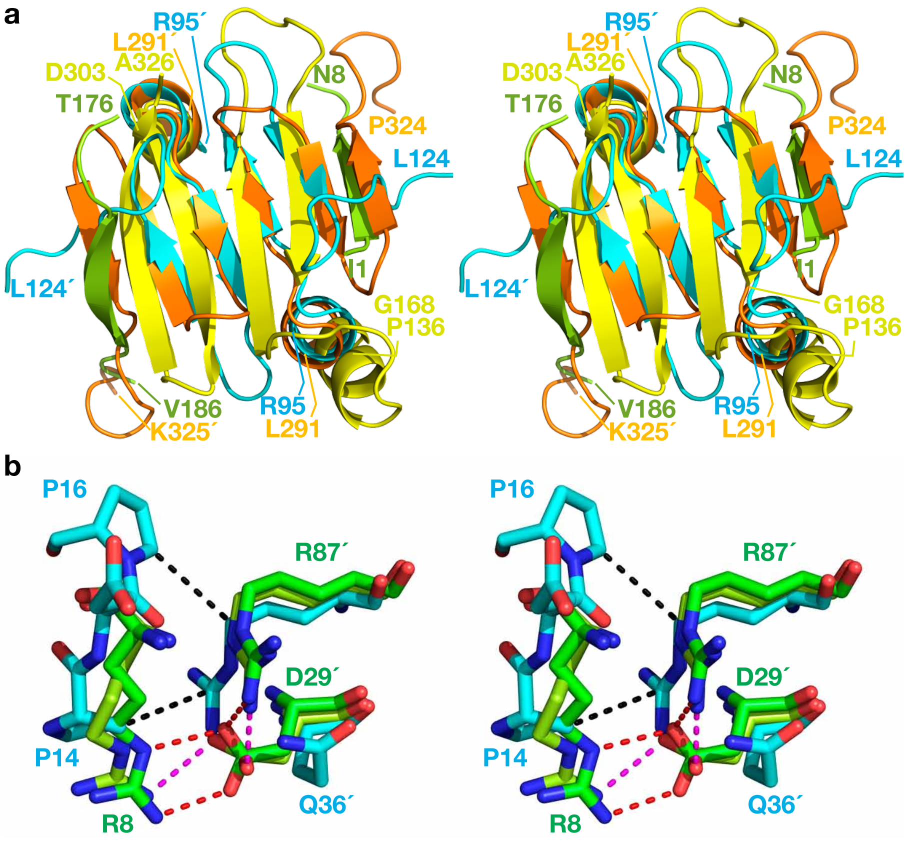

Figure S2. Unique structural features of XMRV PR. **a)** Stereoview of the dimer interface of XMRV PR (cyan), pepsin (yellow for the C-terminal loops and green for the N-terminal strands in each of the two domains, PDB code 4pep), and Ddi1 RP (orange, PDB code 2i1a). Superposition of XMRV PR and Ddi1 RP was based on their Cα coordinates, whereas the superposition of XMRV PR and pepsin was done using the structural template for the family of aspartic proteases. **b)** Stereoview of a fragment of the dimer interface in the superimposed structures of XMRV PR (cyan) and HIV-1 PR, as apoenzyme (lemon green, PDB code 3hvp) and complexed with an inhibitor (green, PDB code 8HVP). Hydrophobic interactions unique to XMRV PR are shown in black, whereas ionic interactions conserved in all retropepsins are colored magenta for the apoenzyme, and red for the inhibitor complex.

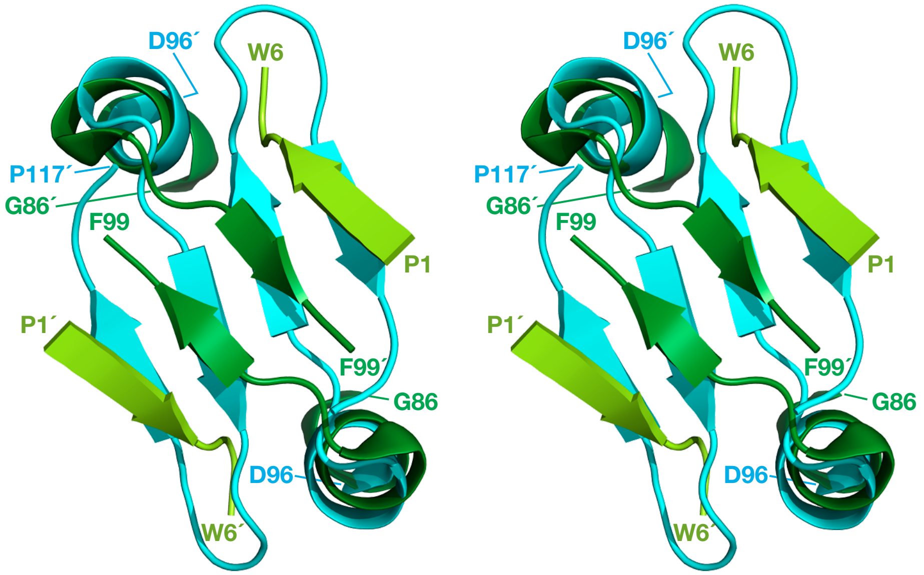

Figure S3. Stereoview of the dimer interface β sheet for the superimposed XMRV PR (cyan) and the apoenzyme form of HIV-1 PR (two shades of green, PDB code 3hvp). This superposition is based on Cα coordinates. The N-terminal and C-terminal strands of HIV-1 PR are shown in lemon and darker shade of green, respectively.

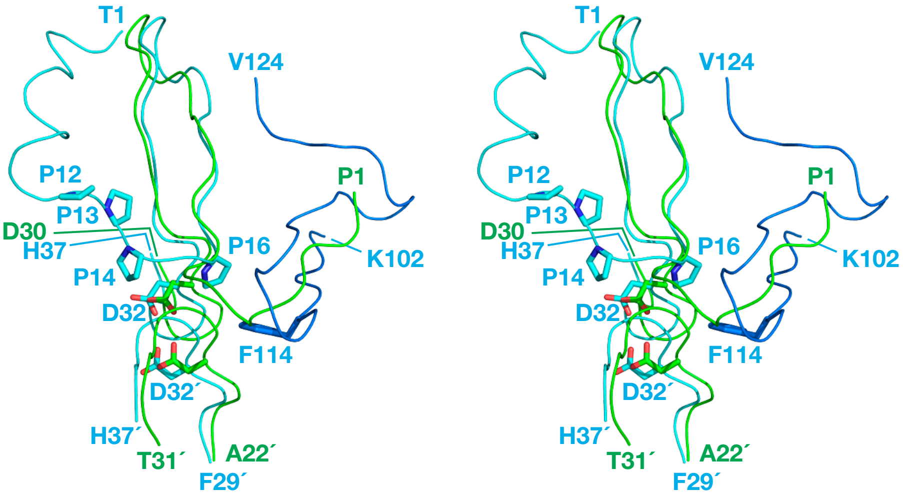

Figure S4. A superposition of the N-terminal regions of XMRV PR (cyan) and HIV-1 PR (lemon green), also including the loops carrying the catalytic aspartates (shown as sticks). The C terminus of XMRV PR is shown in dark blue. Several proline residues in the unusual turn of the N-terminal fragment of XMRV PR and Phe114 near the C terminus, responsible for the shift of the loops with the catalytic aspartates in XMRV enzyme when compared to those in HIV-1 PR, are shown in sticks.

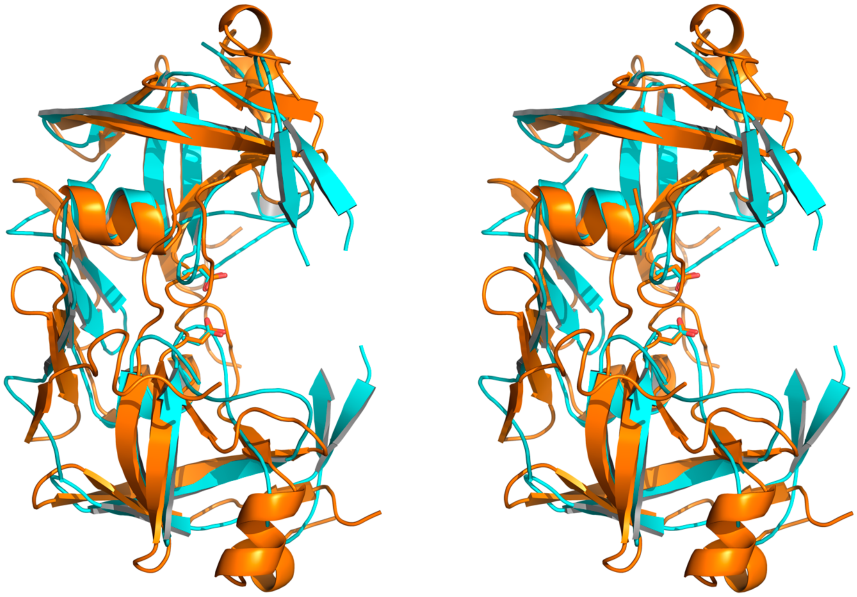

Figure S5. A superposition of the structures of XMRV PR (cyan) and Ddi1 RP (orange, PDB code 2i1a) based on the Cα coordinates and shown in a cartoon representation.

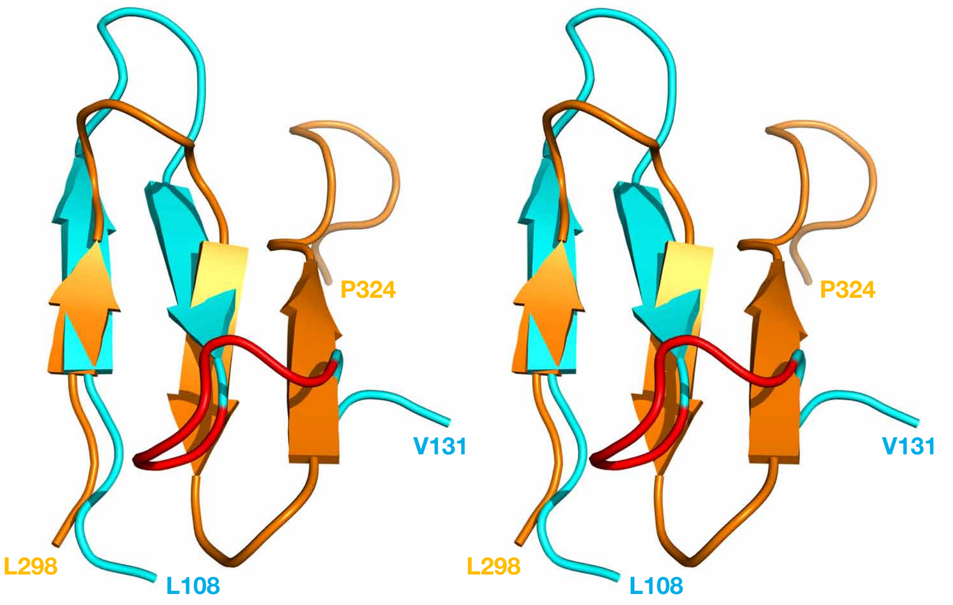

Figure S6. A stereoview of the dimer interface β sheet for the superimposed XMRV PR (cyan) and Ddi1 RP (orange). This superposition is based on the Cα coordinates. A different conformation of the fragment of XMPR PR shown in red is responsible for the absence of the third interface β strand in this enzyme.

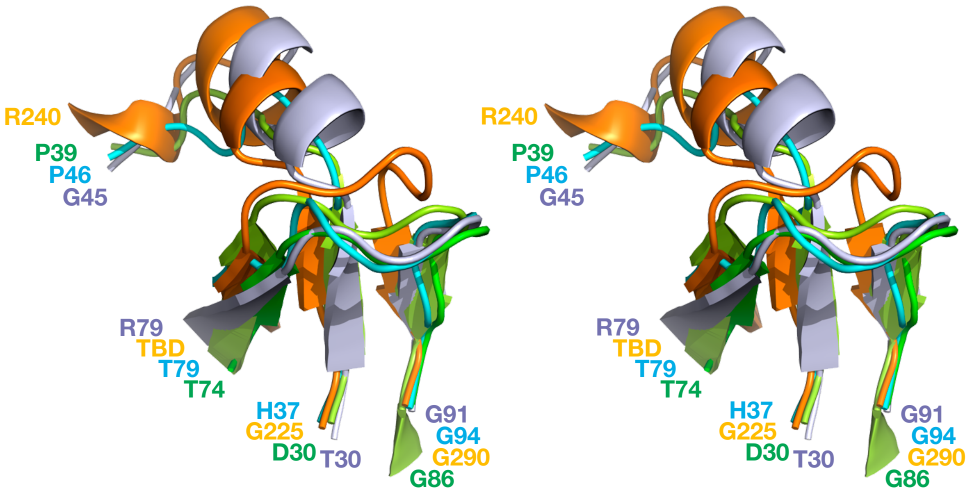

Figure S7. Superposition of two structural fragments in XMRV PR, HIV-1 PR, and Ddi1 RP. A loop that is extensively involved in the interactions with the inhibitors in all retroviral enzymes adopts a similar conformation in XMRV PR, but a different one in Ddi1 RP. The structures of XMRV PR (cyan), Ddi1 RP (orange), HIV-1 PR (apoenzyme in lemon green and the inhibited form in darker green), and EIAV PR (lilac) were superimposed based the Cα coordinates. The labels for the residues comprising the cartoon fragments are shown in matching colors.

Supplementary Table 1. Data collection and refinement statistics

	XMRV PR
Data collection	
Space group	P422
Cell dimensions	
a, b, c (Å)	63.9, 63.9, 105.5
α, β, γ (°)	90, 90, 90
Resolution (Å)	50-1.97 (2.0-1.97) *
R_{sym} or R_{merge}	0.044(0.257)
I / σI	32.1 (3.2)
Completeness (%)	98.5 (82.1)
Redundancy	4.3 (2.9)
Refinement	
Resolution (Å)	40-1.97
No. reflections	15336
R_{work} / R_{free}	0.196/0.233
No. atoms	1883
Protein	1811
Ligand/ion	6
Water	66
B-factors	30.6
Protein	30.2
Ligand/ion	60.1
Water	40.1
R.m.s. deviations	
Bond lengths (Å)	0.020
Bond angles (°)	1.8

Diffraction data were collected on one crystal.

*Values in parentheses are for highest-resolution shell.

Supplementary Table 2. Predicted substrate binding sites in XMRV PR based on a comparison with inhibitor complexes of HIV-1 and EIAV PRs. The primed residues correspond to the second protomer and subsites S1’-S4’ can be created by reversing the primed and unprimed designation of the residues. Residues 54, 55, and 57 are disordered in XMRV PR and were included only on the basis of their sequence equivalence.

Subsite	XMRV	HIV-1	EIAV
S4	Gln36’	Asp29’	Asp29’
	Val54’	Ile47’	Ile53’
	Leu83’	Leu76’	Leu81’
	His37’	Asp30’	Thr30’
	Ala52’	Lys45’	Thr51’
	Trp65’	Gln58’	Ser64’
S3	Pro14	-	-
	-	Arg8	Arg8
	Leu30	Leu23	Leu23
	Gln36’	Asp29’	Asp29’
	Gln55’	Gly48’	Gly54’
	Pro89	Pro81	Pro96
	Tyr90	Val82	Val87
S2	Ala35’	Ala28’	Ala28’
	Leu92’	Ile84’	Ile89’
	Val39’	Val32’	Val32’
	Val54’	Ile47’	Ile53’
	Ala57	Ile50	Val56
	Leu83’	Leu76’	Leu81’
	His37’	Asp30’	Thr30’
S1	Leu30	Leu23	Leu23
	Asp32	Asp25	Asp25
	Gly34’	Gly27’	Gly27’
	Leu92	Ile84	Ile89
	Ala57’	Ile50’	Val56’
	Tyr90	Val82	Val87

Supplementary Methods

Protein Expression and Purification

Recombinant XMRV protease was expressed in *E. coli*. The gene encoding XMRV PR was amplified from the clone pCDNA3.1-XMRV-VP63 (provided by the laboratory of Robert Silverman) by PCR with primers that included a 5' *E. coli* ribosome binding site and 6-His purification tag. The region of DNA corresponding to bp 3169-3543 was amplified, and a stop codon at bp 3181-3183 was converted from a TAG to a CAG within the 5' primer. Primers were flanked by Gateway™ recombination sites, and the PCR amplicon was cloned by BP recombination into pDonr253 according the manufacturer's instructions (Invitrogen). Clones were sequence verified and transferred by Gateway LR recombination into pDest-521, a T7 promoter *E. coli* expression vector based on the Novagen pET43 vector. The vector was transformed into BL21(DE3)tonA pRARE. The cell culture was grown in Superbroth media overnight at 16 °C and at 220 rpm. Cells were collected by centrifugation at 6000 rpm for 15 minutes and resuspended in a lysis buffer containing 50 mM Tris, 250 mM NaCl, and 250 mM $MgCl_2$. Lysis reaction at 4 °C was initiated by adding Bugbuster™ to the lysis buffer at the ratio of 1/50 to its stock solution. After 30 minutes the lysate was centrifuged at 16K rpm for 30 minutes. The supernatant was loaded to a nickel column and the column was washed with 30 mM Tris buffer at pH 8.0, also containing 0.3 M NaCl, and eluted with the same buffer with 0.5 M imidazole added. The eluate was concentrated and loaded onto Superdex 75 column. The column was eluted with 20 mM Tris buffer at pH 8.0, also containing 0.2 M NaCl. The eluate was concentrated to ~8 mg/ml and used for crystallization.

Enzymatic activity

The purified enzyme was enzymatically active, as shown by two separate lines of evidence. When incubated at 25° at pH 5.5 for 2.5 h, the enzyme showed extensive autolysis, with the first and most prominent cleavage between residues His106 and Phe107 (**Supplementary Fig. 1a**, lane 1), as established by sequencing of the bands on the gel and by mass spectrometry. The autolysis was more extensive after overnight incubation (Lane 4). Autolysis could be almost completely prevented by addition of TL-3, an inhibitor of retropepsins that has been shown to exhibit very broad specificity[1]. Addition of the inhibitor at a concentration ~1mM prevented autolysis (lanes 2,3 and 5,6), although the effects of residual activity could be seen after longer incubation.

In another experiment we utilized a construct consisting of N-terminally His_6-tagged maltose binding protein followed by the sequence of mature XMRV PR. When this construct was expressed, no fusion protein was observed, although a band corresponding to free XMRV PR could be identified based on its amino acid sequence (**Supplementary Fig. 1b**, lane 1). When the supernatant was passed through a nickel column and bound protein was eluted with imidazole, we could identify the intact maltose binding protein (cleaved at its C terminus after the sequence YFQG and before TLGD, the N terminus of XMRV PR), and its N-terminal fragment cleaved between the sequences VEAL and SLIY (these residues are part of an extended β strand).

Crystallization and data collection

Crystals of XMRV PR were obtained by the vapor diffusion method in hanging drops mixed from 3 μl of protein and 3 μl reservoir solution, consisting of 0.06 M KH_2PO_4 and 1.34 M Na_2HPO_4 at pH 8.0. The crystals appeared in two days and grew to the final size of 0.1x0.05x0.02 mm^3 in a week. The crystals belong to the relatively rare space group *P*422 with unit cell parameters a=b=63.9 Å, c=106.5 Å. Each asymmetric unit contains a single dimer of XMRV PR. X-ray data at 1.97 Å were collected at SER-CAT 22-ID at APS.

Structure determination

Initial efforts to solve the structure of XMRV PR utilized the molecular replacement (MR) method and search models derived from the crystal structures of such retroviral proteases as HIV, FIV, EIAV, and SIV. The models included both monomeric and dimeric forms of proteins. In selected models, fragments with the largest structural differences among various retroviral proteases were removed. Other models represented polyalanine approximations of the original proteases structures. Various models for B factors (i.e. the original B factor distributions, overall B's, etc.) were also utilized. Most MR calculations were conducted with the program Phaser[2]. During the calculations, the default parameters in the program were adjusted to include nearly all peaks from both rotation and translation functions in the refinement of resulting potential solutions. While all models were subjected to packing analysis, the collision allowance was increased to 30. For each search model, calculations were repeated using various resolution ranges. Despite their extensive scope, all initial MR searches failed to identify a correct solution. In all cases the Z-score and Log Likelihood values were very low and the resulting electron density maps were uninterpretable.

The structure was ultimately solved by application of the Rosetta comparative modeling constrained by poorly phased density data[3]. The HHPred server[4] was used to generate a set of alignments to XMRV PR. The top six templates (PDB codes: 1fmb, 2b7f, 2hah, 3ec0, 2hs1, 3fsm) corresponded to proteases from EIAV, HTLV, FIV, HIV-2, and HIV-1, respectively. The poorly aligned residues were removed from the resulting models. These models, all representing monomeric forms of the proteins, were used in MR searches using Phaser with a very permissive rotation function cutoff (1000 peaks were subjected to the translation search) and a collision allowance (of 10 short contacts between Cα atoms). Additionally, three templates, representing dimeric proteases, were also used in MR searches. For each template, up to five potential MR solutions were utilized for phase generation. For each set of phases obtained from Phaser a separate model was generated with Rosetta. In this process, gaps introduced after the initial HHPred alignments were rebuilt, followed by the torsion-space refinement according to Rosetta's *looprelax* protocol[5]. Models of dimers were constructed with symmetric modeling in Rosetta[6], in which only the symmetrical degrees of freedom are explicitly modeled; however, the density constraints were applied over the entire molecule.

For each putative MR solution (30 for monomeric and 10 for dimeric templates), 2000 models were constructed from independent Monte Carlo trajectories. The top 5% of monomer and dimer models as scored by Rosetta energy function (this score includes the density constraint score) were selected and re-scored using Phaser's *refinement-and-*

phasing rigid-body minimization. In contrast to the monomeric models, the top-scoring models of dimers were characterized by approximately the same translation and orientation, suggesting them as the correct MR solution. The best solution was characterized by the Phaser LLG value of 90 and was chosen for subsequent refinement.

This initial "Rosetta" model was subjected to ten cycles of refinement with Refmac[7], which lowered the R factor from 54.0 to 38.5% and R_{free} from 52.6 to 45.4%. The phases resulting from this refinement, together with the amino acid sequence of XMRV PR, were used during the automatic ARP/wARP procedure[8] executed in the "warpNtrace" mode. The total of 50 cycles of ARP/wARP resulted in the model consisting of 147 residues in six fragments with properly assigned sequence, one 12-residue fragment of polyGly, and 117 water molecules. The corresponding R factor was 32.1%. Large parts of the the ARP/wARP model agreed relatively well with the "Rosetta" model (with the rmsd of 0.96 Å between 133 equivalent Cα atoms), but orientations of the six C-terminal residues in the chains of the two models were considerably different.

Finally, the ARP/wARP model was refined using Refmac in conjunction with manual rebuilding aided by the program COOT[9]. This procedure led to the model characterized by R and R_{free} of 19.6 and 23.2%, respectively. The final model includes residues 1-55 and 60-124 in monomer A, 1-5, 12-54, and 60-125 in monomer B, one partially occupied K^+ cation, a phosphate ion with half occupancy, and 66 water sites. Other residues of the mature protease, as well as the initial methionine and 6-His tag appear to be disordered and were thus not modeled. The characteristics of the model and its refinement are included in **Supplementary Table 1**.

Analysis of the structure of the active site

All structures were superimposed with the program Align[10]. A recent comparison of the binding sites in retroviral proteases[11] provided the basis for predicting possible substrate/inhibitor binding sites in XMRV PR. The residues forming specific pocket(s) in each of the two proteases used as a reference and listed in **Supplementary Table 2** were identified by visual inspection of the individual structures.

Supplementary References

1. Li, M., *et al.* Structural studies of FIV and HIV-1 proteases complexed with an efficient inhibitor of FIV protease. *Proteins: Struct. Funct. Genet.* **38,** 29-40 (2000).

2. McCoy, A.J. Solving structures of protein complexes by molecular replacement with Phaser. *Acta Crystallogr.* **D63,** 32-41 (2007).

3. DiMaio, F., *et al.* Increasing the radius of convergence of molecular replacement by density and energy guided protein structure optimization. *Nature* **in press** (2011).

4. Soding, J., Biegert, A. & Lupas, A.N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33,** W244-W248 (2005).

5. Qian, B., *et al.* High-resolution structure prediction and the crystallographic phase problem. *Nature* **450,** 259-264 (2007).

6. Andre, I., Bradley, P., Wang, C. & Baker, D. Prediction of the structure of symmetrical protein assemblies. *Proc. Natl. Acad. Sci. USA* **104,** 17656-17661 (2007).

7. Murshudov, G.N., Vagin, A.A. & Dodson, E.J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr.* **D53,** 240-255 (1997).

8. Perrakis, A., Morris, R. & Lamzin, V.S. Automated protein model building combined with iterative structure refinement. *Nature Struct. Biol.* **6,** 458-463 (1999).

9. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr.* **D60,** 2126-2132 (2004).

10. Cohen, G.E. ALIGN: a program to superimpose protein coordinates, accounting for insertions and deletions. *J. Appl. Crystallogr.* **30,** 1160-1161 (1997).

11. Li, M., *et al.* Crystal structure of human T-cell leukemia virus protease, a novel target for anti-cancer drug design. *Proc. Natl. Acad. Sci. USA* **102,** 18322-18337 (2005).