Retroviral protease

A computer-based exercise

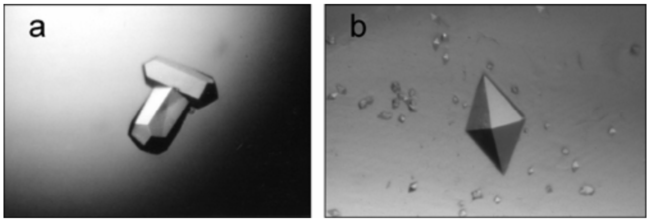
Mariusz Jaskolski

The structure of retroviral protease (PR) was discovered in the end of 1988. It is a striking fact that the 20th anniversary of retroviral protease structure almost coincides with the 50th anniversary of the first protein structure, myoglobin, announced in 1958 by Kendrew. The discovery of retroviral protease structure was very important because it has lead to structure-guided design of protease inhibitors that are now used as efficient drugs for treating HIV infection.

HIV (human immunodeficiency virus), the causative agent of AIDS (acquired immunodeficiency syndrome) is a retrovirus, meaning that during its life cycle, after invasion of a T4 cell of the immune system, a viral enzyme, called reverse transcriptase (RT) retro-transcribes the viral RNA genome into DNA. "Retro" means backward, because this is a process that violates the normal flow of genetic information: DNA→RNA→protein. In the next step, another of the viral-encoded enzymes, integrase (IN), carries the viral DNA into the cell nucleus and incorporates it into the host genome. From this moment, the infection of the T4 cell is permanent and cannot be cured, because the virus is part of the cell's genome, meaning, for example, that it will be inherited by all daughter cells (vertical spread of the virus). But the infection is also spread horizontally, from cell to cell (and form patient to patient), by nascent viral particles, which multiply within the infected cell, and then leave it in a process called budding and maturation, ready to infect new T4 cells. During the budding/maturation process, the third retroviral enzyme, protease takes action, cleaving the viral polyproteins into mature products. Without the protease, the nascent virions are not infective.

The retroviral genome contains only three genes: *gag* (coding a number of structural proteins), *pol* (coding PR, RT and IN, the three retroviral enzymes), and *env* (coding the glycoproteins of the viral envelope). The retroviral protease is necessary to cleave the fusion polyproteins into the mature products.

The first team to solve the crystal structure of a retroviral protease was formed by Alex Wlodawer at the National Cancer Institute (USA) and was concerned with the enzyme from Rous sarcoma virus (RSV), now also known as avian sarcoma virus (ASV). Proteins from HIV were very difficult to come by at that time, especially in quantities required for crystallography, thus the decision to work on the enzyme form the closely related chicken virus, RSV, seemed very reasonable. As it turned out later, parallel work was being done on the authentic HIV-1 protein, at Merck Sharp and Dohme Research Laboratories.



Single crystals of RSV protease (a) and synthetic HIV-1 protease (b).

The RSV protease was composed of 124 residues and it behaved like an aspartic protease, for instance it could be inhibited by pepstatin, a standard inhibitor of these enzymes. Pepsin (a digestive enzyme present in the gastric juice) and other cell-derived aspartic proteases have a single polypeptide chain composed of about 350 residues and folded into two similar domains. Each domain contributes a DTG sequence motive to the active site, which is located at the bottom of a deep cavity. The most important catalytic residues are the two aspartates. They are used by the enzyme to bind in a symmetric fashion a water molecule, which becomes the catalytic nucleophile that attacks the carbonyl C atom of the cleaved peptide substrate. The active site cavity is covered by a single flap arm contributed by the N-terminal domain.

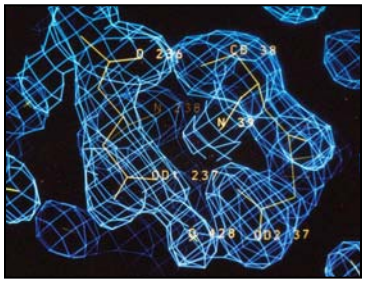
Please sketch the scheme of the protease reaction that leads to the hydrolysis of a peptide bond.

Mostly because the amino acid sequence contained the DTG signature (or its variant DSG in the case of RSV), the retroviral enzyme indeed looked like an aspartic protease, but there was only one copy of the DTG motif and the enzyme was three times smaller than a typical pepsin-like aspartic protease.

However, there were papers by Jordan Tang, by then 10 or even 13 years old, which predicted that cell-encoded two-domain pepsin-like aspartic proteases might have evolved via gene duplication from much smaller, homodimeric ancestral enzymes. When the sequences of retroviral proteases became available, a hypothetical model of HIV-1 protease was built by Laurence Pearl and William Taylor in 1987, in a bold modeling exercise performed despite almost nil sequence conservation between cell-derived and retroviral proteases.

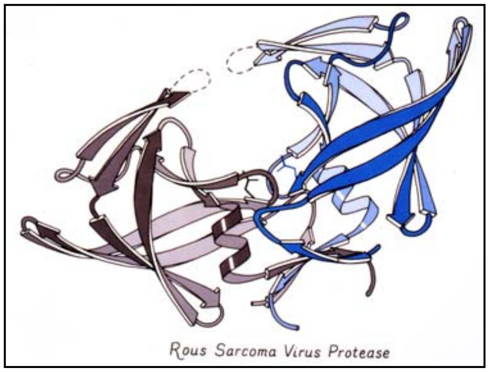
Please try to locate through the PubMed database the articles by Jordan Tang and by Laurence Pearl and William Taylor. Read the abstracts of those papers.

The experimental approach used to solve the structure of RSV PR was the MIR, or heavy-atom method. Several heavy-atom derivatives of the RSV PR crystals were produced. One of them, obtained with the application of uranyl acetate, was especially important because the uranyl cation $(UO)_2^{2+}$ has a tendency to bind to acidic groups. When the phase problem was solved, the electron density maps revealed two segments of helical density and a very clear electron density at the uranium site, symmetrically flanked by two carboxylates. It was clear that those carboxylates were the aspartates of the two-fold symmetric DSG/DSG active site.



First electron density map of RSV protease showing a water molecule (O428) hydrogen-bonded between two aspartate side chains (Asp37 and Asp237).

The structure revealed that the enzyme was a homodimeric aspartic protease, composed of two identical subunits, resembling the monomeric two-domain pepsin. The active site had the same architecture and there was a water molecule between the aspartates. But there were also very significant differences with pepsin. Because of the symmetry, the two flaps were of the same length and both prominent, although the tips in their elevated position over the empty active site were disordered. The subunit interface was formed by a tight four-stranded antiparallel β-sheet weaved from all the termini in the order: NA-CB-CA-NB. The first RSV protease model was immediately used to build a homology model of the HIV-1 enzyme. The model looked very plausible, had all the features of the template, with differences limited to loop regions.

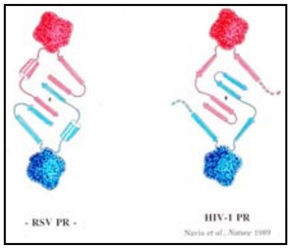


First model of RSV protease, hand-drawn by Dr. Jane Richardson (Duke Univ.).

The structure of RSV protease was announced in *Nature* early in February, 1989. A week later, in the same journal, the crystal structure of HIV-1 protease was unveiled by Manuel Navia and coworkers from Merck Sharp and Dohme. In the same week, the homology model of HIV PR was published in *Science*.

Please find the original papers mentioned above and read their abstracts.

After the first burst of joy, there was suddenly consternation because the crystal structures of the RSV and HIV-1 proteases, while similar in the basic features, also showed some perplexing differences, especially in the C-terminal part of the molecule. Where the RSV model had a clear α-helix, the HIV-1 structure had a β-strand, and the topology of the dimer interface was completely different. Instead of the interlaced termini with three inter-subunit β-sheet connections, the HIV-1 structure had a hairpin with only one area of intersubunit contact, and a disordered N-terminus. The latter difference was not trivial at all because it had profound consequences for the dimer stability and for the way the protease is able to liberate itself from the gag-pol fusion polyprotein. Moreover, it was not a purely academic question because accurate HIV-1 protease model was badly needed for structure-guided design of inhibitors that might be developed into AIDS drugs.



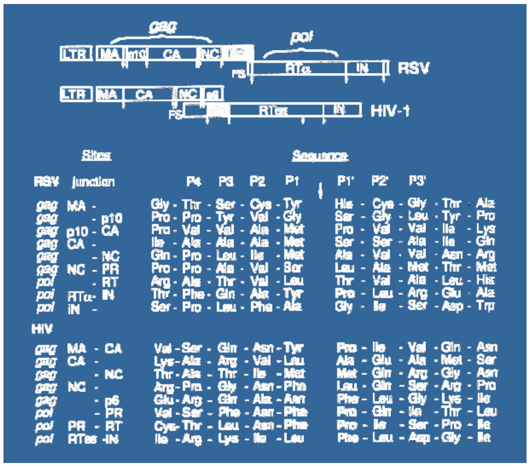
Comparison of the topology of the interdomain β-sheet formed by the N- and C-termini of the original RSV and HIV-1 protease models.

Find the first structures of protease from RSV and HIV in the Protein Data Bank (PDB) and save their codes for future structural comparisons.

It may be appropriate to mention that the HIV-1 protease is translated as part of a huge polyprotein containing the structural (gag gene products) and enzymatic (pol gene products) proteins. (In the case of RSV, the protease is part of both gag and gag-pol expression products.) For maturation of the virion particles, all the proteins, including the protease, must be liberated from the precursors. This maturation process is carried out by the viral protease itself, which poses a topological puzzle, how the protease can fold properly while still embedded in the polyprotein, form active dimer, and ultimately cut itself out, and all this in the restricted confinement of the budding viral particle. The disorder of the N-terminus suggested by the Merck model would allow protease self-excision not only in *trans* but even in *cis*!

Looking at the organization of the retroviral polyproteins, one realizes the unusual property of the protease, which must be able to cleave a number of different sequences with high fidelity. Loosing specificity for even one of those sites would render the virus crippled, unable to mature and to propagate infection. This is why efficient inhibitors of retroviral protease are excellent candidates for anti-HIV drugs. On the other hand, the virus can efficiently develop drug resistance by mutations. Those mutations are often localized in the protease sequence but can also affect the nature of the cleavage sites.

Contemplating their genetic and enzymatic features, one is struck by the extrodinary parsimony combined with high effectiveness of retroviruses. Not only is the number of genes minimized by making fusion products together with a tool to process them, but also the tool itself is cleverly miniaturized by assembly from two equal parts.



Sites | Sequence

RSV	junction		P4	P3	P2	P1	↓	P1'	P2'	P3'		
gag	MA -	Gly	Thr	Ser	Cys	Tyr		His	Cys	Gly	Thr	Ala
gag	- p10	Pro	Pro	Tyr	Val	Gly		Ser	Gly	Leu	Tyr	Pro
gag	p10 - CA	Pro	Val	Val	Ala	Met		Pro	Val	Val	Ile	Lys
gag	CA -	Ile	Ala	Ala	Ala	Met		Ser	Ser	Ala	Ile	Gln
gag	- NC	Gln	Pro	Leu	Ile	Met		Ala	Val	Val	Asn	Arg
gag	NC - PR	Pro	Pro	Ala	Val	Ser		Leu	Ala	Met	Thr	Met
pol	- RT	Arg	Ala	Thr	Val	Leu		Thr	Val	Ala	Leu	His
pol	RTα- IN	Thr	Phe	Gln	Ala	Tyr		Pro	Leu	Arg	Glu	Ala
pol	IN -	Ser	Pro	Leu	Phe	Ala		Gly	Ile	Ser	Asp	Trp
HIV												
gag	MA - CA	Val	Ser	Gln	Asn	Tyr		Pro	Ile	Val	Gln	Asn
gag	CA -	Lys	Ala	Arg	Val	Leu		Ala	Glu	Ala	Met	Ser
gag	- NC	Thr	Ala	Thr	Ile	Met		Met	Gln	Arg	Gly	Asn
gag	NC -	Arg	Pro	Gly	Asn	Phe		Leu	Gln	Ser	Arg	Pro
gag	- p6	Glu	Arg	Gln	Ala	Asn		Phe	Leu	Gly	Lys	Ile
pol	- PR	Val	Ser	Phe	Asn	Phe		Pro	Gln	Ile	Thr	Leu
pol	PR - RT	Cys	Thr	Leu	Asn	Phe		Pro	Ile	Ser	Pro	Ile
pol	RTαs -IN	Ile	Arg	Lys	Ile	Leu		Phe	Leu	Asp	Gly	Ile

Organization of retroviral genomes and the cleavage sites of the retroviral protease.

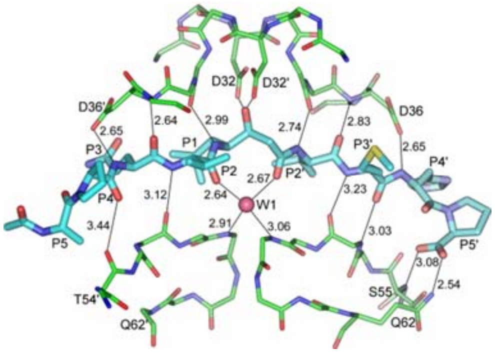
The situation in February 1989 became rather uncomfortable: which HIV-1 protease model should be used for design of AIDS drugs? Which was correct? The dilemma was resolved experimentally by the NCI team. The protein for this independent experiment was obtained by total chemical synthesis from amino acids. This was a great accomplishment in itself because it showed, as proof of principle, that synthetic proteins can be properly folded outside of any biological context, and crystallized.

The result confirmed beyond any doubt that the model of the RSV enzyme was correct. The dimer had fully visible, elevated flap arms, a tightly interlaced intersubunit β-sheet, the C-terminal α-helix, and a well defined water molecule between the catalytic aspartates. The definite structure of HIV-1 protease was published in *Science* in August, 1989.

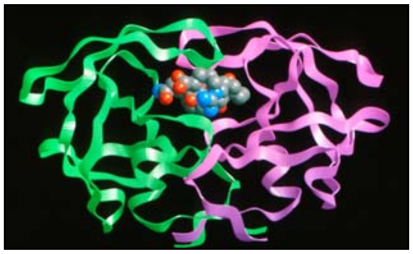
The first correct model of HIV-1 protease.

Please find this article and read the abstract.

The next goal was the structure of retroviral protease in complex with inhibitors. The first inhibitors were the obvious choice: oligopeptides with substrate sequence, but with the scissile peptide bond replaced with a non-hydrolyzable surrogate, such as reduced peptide, or various hydroxylated ethyl groups. Also, the existing inhibitors of cell-derived aspartic proteases, such as pepstatin, could be immediately tried. However, if selective inhibitors of retroviral proteases were to be found, they should not interfere with the host enzymes. Instead, they should exploit the unique features of retroviral protease, such as the perfect symmetry of the binding site, the existence of two flaps, or the presence of a structural water molecule in complexes with peptidic inhibitors. This water molecule, with perfect tetrahedral coordination at the inhibitor/flap interface, was first observed in the crystal structure of a complex of HIV-1 protease with MVT-101 inhibitor. Later, this interface water molecule was included in a novel class of inhibitors, based on cyclic urea scaffold.

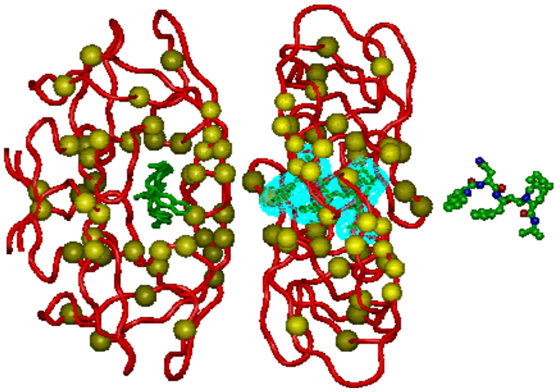


Binding of a peptidomimetic inhibitor (turquoise) in the active site of retroviral protease (green). Note the extended conformation of the inhibitor which is wedged between the active site marked by the two aspartates (D32 and D32', top) and the flap arms (bottom) that are lowered towards the active site. Note the tetrahedrally coordinated water molecule W1 at the flaps-inhibitor interface. H-bond distances in Å.



The first inhibitor (space-filling model) complex of HIV-1 PR, here shown in green and pink depicting the two subunits. Note that the flap arms are lowered and locked over the inhibitor.

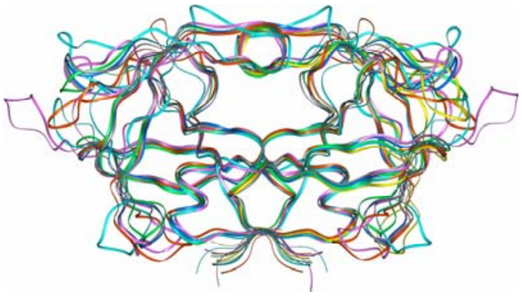
Some of the inhibitors have been developed into very potent drugs for treating HIV infection. The first protease inhibitor, Saquinavir, was approved for clinical use in December 1995. It was only six years after the structure of the first inhibitor complex had been published and less than seven years from the moment an experimental model of the protein saw the light of day. This marks a real triumph of structural biology, now having the power to quickly lead to efficient therapies against a disease which only a few years earlier was considered a global threat. So far, 11 protease inhibitors gained FDA approval for the treatment of HIV infection. All those molecules are competitive inhibitors, meaning that they compete for the active site with protease substrates. The first competitive inhibitor was characterized by a submicromolar dissociation constant. Picomolar inhibitors were subsequently developed through fine-tuning to the enzyme binding pockets.



The first HIV protease inhibitor, Saquinavir (right), approved as AIDS drug, and its complex with the protease shown in two projections. The balls indicate the protease residues whose mutations lead to drug resistance.

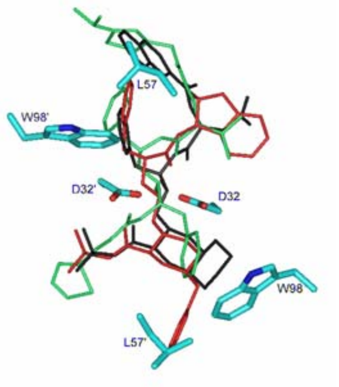
Theoretically, competitive inhibition is not the only option. One might also imagine irreversible modification of the active site, or binding of the inhibitor molecule in a place different form the active site, for instance, to hinder flap closure or to disrupt dimer formation. However, all those options, even when tried, have not resulted in useable pharmacological agents.

Cornered with a potent drug, the virus counteracts, and resistance to protease inhibitors arises either as selection of existing variants or as mutations. An important aspect of the continuing structural research is to understand the mutations of this arms race and design even more clever drugs or their combinations. Another area of activity is concerned with the structure of protease from different retroviruses. In addition to RSV and HIV-1, the enzymes corresponding to HIV-2, SIV (simian immunodeficiency virus), FIV (feline immunodeficiency virus), or EIAV (equine infectious anemia virus) have also been studied. All the proteins share the same fold and domain organization. However, knowledge of the structural details in which they differ, especially in the context of inhibitor complexes, is also contributing to our understanding of drug resistance through sequence alterations. The most recent addition to the collection of retroviral protease structures is the enzyme from the HTLV-1 (human T-cell leukemia) retrovirus. Thus, suddenly, the efforts to cure AIDS and cancer have a common structural point. When the structure of HTLV-1 protease was solved, it became obvious why the AIDS drugs tried on HTLV patients have no effect. It is interesting to note in this context that, although HTLV was discovered before HIV (and in fact was connected with a naming confusion), the protease from HTLV had resisted structural characterization for a long time, partly because of various crystallographic obstacles. For instance, the r.m.s. (root-mean-square) deviation between the Cα traces of HTLV-1 protease and the molecules from other retroviruses is as high as 1.93 Å (RSV protease) and 1.72 Å on average, showing that, on closer look, there are indeed significant variations of the canonical retroviral protease fold. Incidentally, similar r.m.s. deviations are obtained in comparisons with pepsin, albeit for a smaller number of superposed atoms. However, when only the atoms of the active site are compared, the match is nearly perfect, with an r.m.s.d. of about 0.5 Å in superposition of retroviral and cell-derived aspartic proteases.

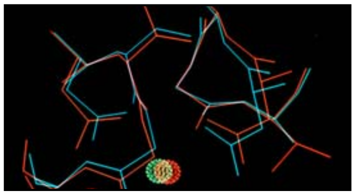


Superposition of the Cα traces of proteases from several retroviruses, HIV-1 (green), HIV-2 (blue), SIV (olive), RSV (pink), FIV (red), EIAV (orange), HTLV-1 (turquoise). Since the coordinates have been taken from complexes with inhibitors (which occupy the catalytic cavity), the flap arms are locked in a lowered position over the active site.

Can you indicate the reasons why AIDS drugs are not effective against HTLV infection?

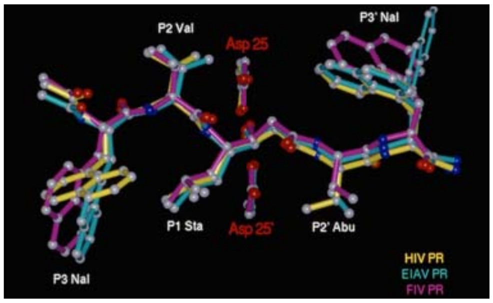


Three AIDS drugs functioning as HIV-1 protease inhibitors, saquinavir (black), indinavir (red) and ritonavir (green) docked in the active site of HTLV-1 protease (turquoise).



Superposition of the active sites of RSV protease (blue) and Rhizopuspepsin (red). The corresponding green and red spheres represent the position of the nucleophilic water molecule between the catalytic aspartates.

Structural studies of a broad range of retroviral proteases have the added advantage that they allow one to look at bottlenecks and obstacles from a different perspective. One such difficulty stems from the mixed blessing of the two-fold symmetry of retroviral proteases. With HIV-1 protease, this has led to ambiguity of space-group symmetry assignment and to two-fold disorder of the bound inhibitors. This drawback has been turned to an advantage when C_2 symmetric (or pseudosymmetric) inhibitors were synthesized. Many of the active site inhibitors, including pepstatin, are mechanism-derived, in the sense that they contain a central hydroxyl group that mimics the catalytic water molecule in a tetrahedral transition state analog. One such universal inhibitor, LP130, has been studied in complex with a variety of retroviral proteases.



Superposition of the LP130 inhibitor from its complexes with three retroviral proteases. Note that the conformation and orientation of the inhibitor in these complexes are essentially identical. The OH group in the center of the inhibitor molecule is H-bonded by the active site aspartates in a fashion imitating the nucleophilic water molecule in unliganded protease.

A serious problem, as with many other proteases, is autodigestion on prolonged incubation. This difficulty is, of course removed by the use of inhibitors, another option being a mutation, usually D→N, in the active site. With regard to mutation, in the simplest variant, this approach leads to simultaneous change of both catalytic aspartates. However, asymmetric mutations became possible through a clever engineering trick, whereby the two subunits are tethered through a CA-NB linker.

The structural studies of retroviral proteases over the last 20 years have generated a tremendous stimulus for structural biology and an enormous amount of information. HIV protease has become the most studied protein, in structural terms, in the Universe. The number of structure determinations goes into hundreds and count has been lost, despite efforts at bookkeeping in a designated database. The overwhelming majority of the structures have been determined by protein crystallography, but there are also NMR structures, including a monomeric form of the protein.

HIV protease has helped to advance the frontiers of structural biology in many different ways. We've mentioned already the demonstration that total chemical synthesis can be an option for making proteins for structural characterization. The method of chemical synthesis was also used to obtain the D-enantiomer of HIV-1 protease and to demonstrate that this mirror twin of the natural enzyme behaves identically in a looking-glass world. Recently, huge single crystals of HIV-1 protease have been grown in preparation for a neutron-diffraction experiment, which, by visualization of hydrogen atoms, will hopefully help to answer the persisting questions about the catalytic mechanism. HIV-1 protease has been already characterized at a breathtaking ultra-high resolution of 0.84 Å.

ADVANCED LEVEL Why is neutron diffraction better for visualization of H atoms than X-ray diffraction?

It is generally recognized that HIV protease has been the platform for the development of rational drug design strategies from a flimsy dream to reality. The fact that through this approach effective therapies for an incurable disease have been found, is among the major scientific achievements of the last century.