Protein family review

Retroviral proteases

Ben M Dunn*, Maureen M Goodenow†, Alla Gustchina‡ and Alexander Wlodawer‡

Addresses: Departments of *Biochemistry and Molecular Biology and †Pathology and Experimental Medicine, University of Florida, Gainesville, FL 32610, USA. ‡Macromolecular Crystallography Laboratory, National Cancer Institute, Frederick, MD 21702, USA.

Correspondence: Ben M Dunn. E-mail: bdunn@college.med.ufl.edu

Published: 26 March 2002

Genome ***Biology*** 2002, **3(4)**:reviews3006.1–3006.7

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/4/reviews/3006

Summary

The proteases of retroviruses, such as leukemia viruses, immunodeficiency viruses (including the human immunodeficiency virus, HIV), infectious anemia viruses, and mammary tumor viruses, form a family with the proteases encoded by several retrotransposons in *Drosophila* and yeast and endogenous viral sequences in primates. Retroviral proteases are key enzymes in viral propagation and are initially synthesized with other viral proteins as polyprotein precursors that are subsequently cleaved by the viral protease activity at specific sites to produce mature, functional units. Active retroviral proteases are homodimers, with each dimer structurally related to the larger class of single-chain aspartic peptidases. Each monomer has four structural elements: two distinct hairpin loops, a wide loop containing the catalytic aspartic acid and an α helix. Retroviral gene sequences can vary between infected individuals, and mutations affecting the binding cleft of the protease or the substrate cleavage sites can alter the response of the virus to therapeutic drugs. The need to develop new drugs against HIV will continue to be, to a large extent, the driving force behind further characterization of retroviral proteases.

Gene organization and evolutionary history

Retroviral proteases are encoded by a part of the *pol* gene, for example in that of the human immunodeficiency virus (HIV). The protease gene is located between the *gag* gene (encoding structural proteins) and other enzymatic genes, such as reverse transcriptase and integrase. There are 93 sequences belonging to the retroviral protease family A2 of the aspartic peptidase clan AA at present, according to the Merops database, which provides information on viral as well as other proteases [1]. The A2 family includes the proteases of leukemia viruses, immunodeficiency viruses, infectious anemia viruses, and mammary tumor viruses, as well as those encoded by several retrotransposons from fruit flies and yeast, and endogenous viral sequences in humans and other primates. Figure 1 presents a phylogenetic tree that shows the evolutionary history of, and relationships between, selected members of the family of retroviral proteases.

The RNA of retroviruses is replicated through a DNA intermediate, the product of the virus-encoded reverse transcriptase, which is an error-prone enzyme that lacks a proofreading function. In HIV-1 (the HIV type responsible for most cases of the acquired immune deficiency syndrome, AIDS), at least one nucleotide substitution occurs on average during every round of replication. Selective pressures affect replication, cell tropism (the ability of a virus to enter particular cell types), and escape from host immunity, and contribute to genetic differences between HIV-1 isolates within an individual and between individuals [2]. Thus, there is no 'wild-type' HIV-1 protease, but rather a complex mixture of related sequences [3]. Variability is most pronounced in the HIV-1 envelope (*env*) gene, but is found in virtually all regions of the viral genome, including the protease gene. Similar variability is expected in other retroviral sequences, but much less information is available

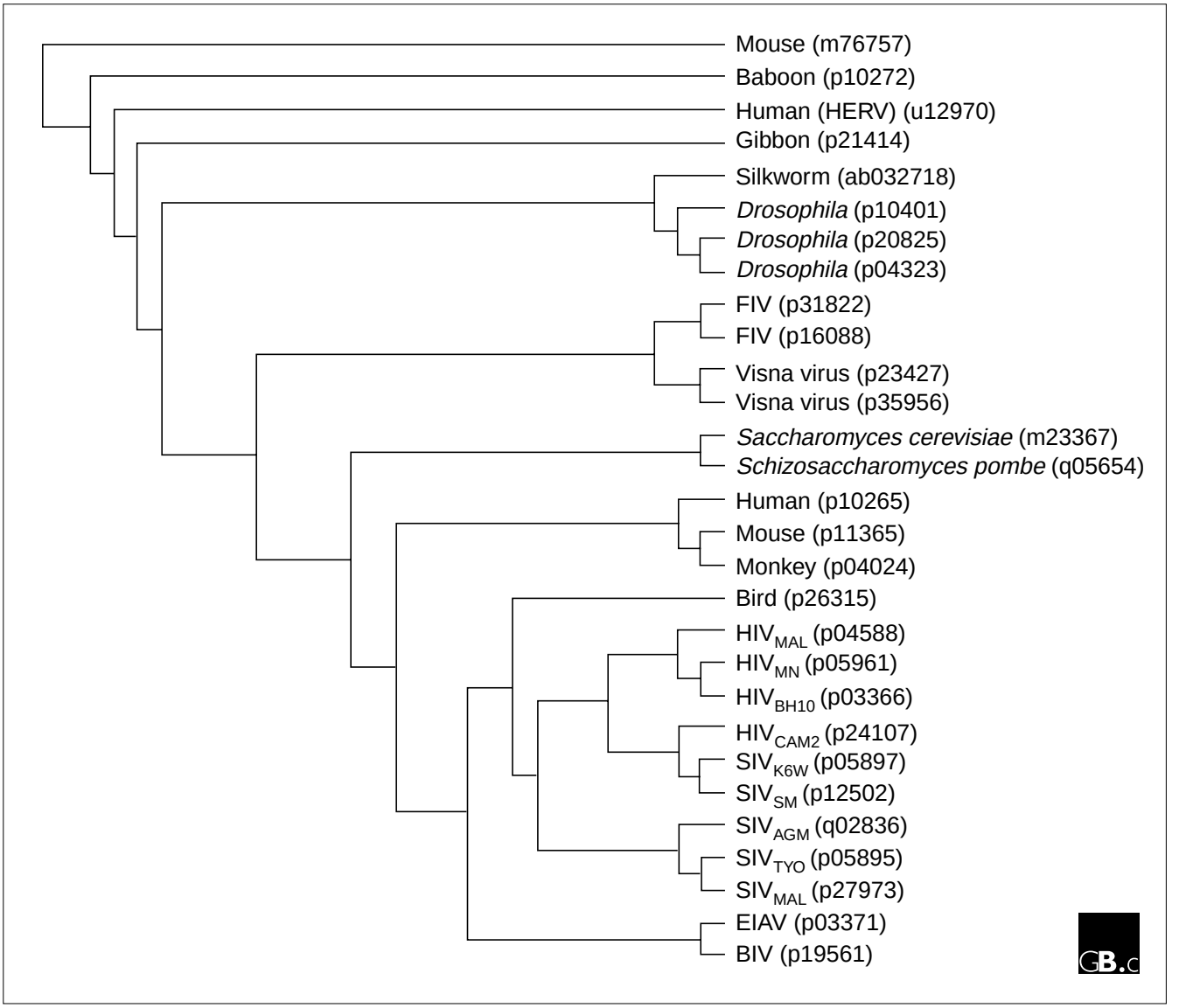

Figure 1
The relationships between retroviral proteases. Protease sequences from several different immunodeficiency viruses are compared with endogenous retroviral sequences found in various eukaryotic genomes. The numbers in brackets indicate GenBank accession numbers [18]; viral strains are indicated by subscript letters. Nucleic-acid sequences were aligned using ClustalW [19] and a Jukes-Cantor phylogenetic tree file was generated using the PHYLIP package and the programs DNADIST and FITCH. The tree was produced using the cladogram option and the programs TreeView. Abbreviations: BIV, bovine immunodeficiency virus; EIAV, equine infectious anemia virus; FIV, feline immunodeficiency virus; HERV, human endogenous retrovirus; SIV, simian immunodeficiency virus.

compared to the wealth of data that has been gathered for the HIV system. Genetic analysis of proteases from different individuals [4] is illustrated in Figure 2. Viruses from different individuals form separate branches in a phylogenetic tree of protease sequences. Each major branch develops into multiple small branches that represent the swarm, or quasispecies, of viruses within an individual. Protease sequences in viruses from children who were infected perinatally by maternal transmission differ from one another, but are closely related to sequences in viral quasispecies found in their mother or siblings. Even when individuals are unrelated, the relationship between their HIV-1 isolates and the history of infections can be detected; for example, in Figure 2, children 6 and 7 were not related but were infected by the same blood product. Individual 2 was infected by sexual transmission of HIV-1 from individual 1. The protease from the laboratory strain HIV_{LAI} is located on a separate branch in the tree, indicating that no HIV-1 protease from patient viruses is identical to this prototype protease sequence.

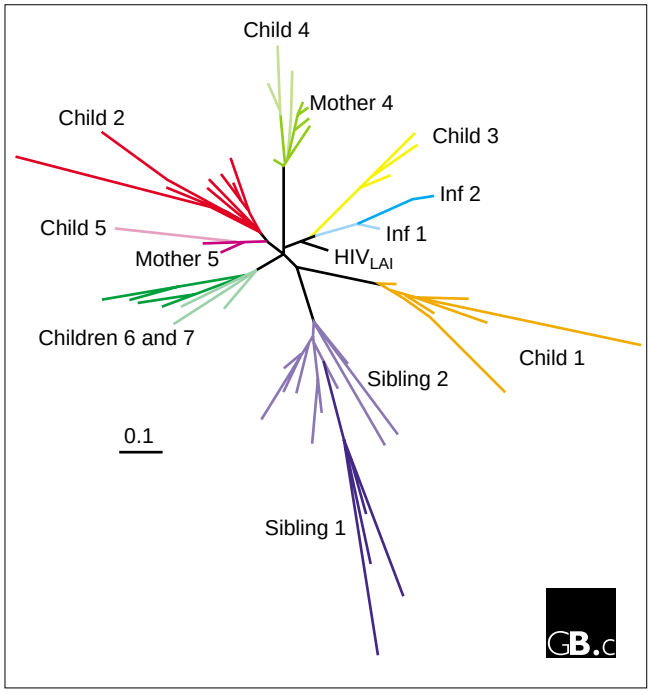

Figure 2
HIV-1 protease sequences are heterogeneous. Nucleotide sequences from protease quasispecies in peripheral blood mononuclear cells from infected children and adults were analyzed by constructing a phylogenetic tree. Major branches are black; colored branches represent sequences within individuals; individuals infected by related viruses are represented by shades of similar colors on a major branch. Inf1 and Inf2 are two unrelated individuals whose closely related virus sequences suggest that individual 2 was infected by individual 1.

Characteristic structural features

Crystal and nuclear magnetic resonance (NMR) structures are available for retroviral proteases from HIV-1 [5], HIV-2 [6], simian [7] and feline [8] immunodeficiency viruses (SIV and FIV), rous sarcoma virus (RSV) [9] and equine infectious anemia virus (EIAV) [10]; reviewed in [11]. The secondary structures of all retroviral proteases share a structural template (Figure 3) that was previously used to describe non-viral aspartic proteases [12]. Retroviral proteases form homodimers and the template structure shows that a monomer is formed by the duplication of four structural elements: a hairpin (containing loop A1), a wide loop (B1, containing the catalytic aspartic acid), an α helix (C1), and a second hairpin (D1). The second monomer contains the identical elements, named A2, B2, C2, and D2 in Figure 3. The length of loops A1 and A2 is different in various retroviral proteases, as are the length and conformation of the connecting segments between these structural elements. The α helix C1 is prominent only in EIAV protease, whereas it consists of a single helical turn in RSV and FIV proteases and is replaced by a loop in the proteases of HIV-1, HIV-2, and SIV. The flexible β loop D1, known as a 'flap' in non-viral proteases, is functionally very important, because it changes orientation during binding of the ligand (substrate or inhibitor) and forms numerous interactions with it. Two such flaps are present in the symmetric dimers of retroviral proteases. The hairpin D2 is substituted by a β strand in all retroviral proteases for which structural information is available. In addition to the four core structural elements, the amino and carboxyl termini in a dimer form a four-stranded β-sheet interface. The amino-acid sequences of retroviral proteases are significantly similar, particularly in the locations of residues that are important in preserving both structure and function.

The active site of each retroviral protease contains a pair of aspartic acid residues (Asp25 and Asp25′; amino acids are numbered according to their positions in HIV-1 protease). The conserved active-site residues - Asp25, Thr26 (replaced by Ser38 in RSV protease), and Gly27 - are located in a loop, the structure of which is stabilized by a network of hydrogen bonds similar to that found in the eukaryotic proteases (Figure 4; for a review, see [13]). The carboxylate groups of the Asp25 residues from both chains are nearly co-planar and make close contacts via their O1 atoms. The network is quite rigid as the result of a set of interactions called the 'fireman's grip', in which the Oγ atom of each Thr26 accepts a hydrogen bond from the main-chain NH group of the Thr26 in the opposing loop; Thr26 also donates a hydrogen bond to the oxygen atom of the carbonyl group of residue 24 on the opposite loop. Identical interactions have been observed in all retroviral proteases thus far examined by crystallographic methods. The carboxylate residues are bridged by a water molecule, located within hydrogen-bonding distance of the oxygen atoms of the Asp25 carboxylates. Water molecules forming similar bridges have also been reported in non-viral proteases [13]; they might correspond to the catalytic water molecule required for hydrolysis of the peptide bond in the substrate. The distances between the inner oxygen atoms of the co-planar carboxylates are 2.8 to 3 Å, indicating the presence of an acidic proton in the bridge.

Binding of inhibitors is accompanied by a large shift in the flaps of both subunits (Figure 3c). In some enzymes (for example, RSV protease), the flaps are disordered and therefore are not seen in the X-ray structure [9]. In other enzymes, the flaps are seen in an 'open' conformation when no ligands (substrates and/or inhibitors) are present. Binding to the active site induces a downward movement of the flap residues; this allows additional interactions with the ligand and strengthens the binding of both substrates (by inference) and inhibitors.

Localization and function

Translation of the retroviral gag-pol mRNA produces in most cases a Gag protein of 55 kDa, ending before the protease gene. In about 5% of the gag-pol transcripts, a translational frameshift occurs slightly upstream of the protease

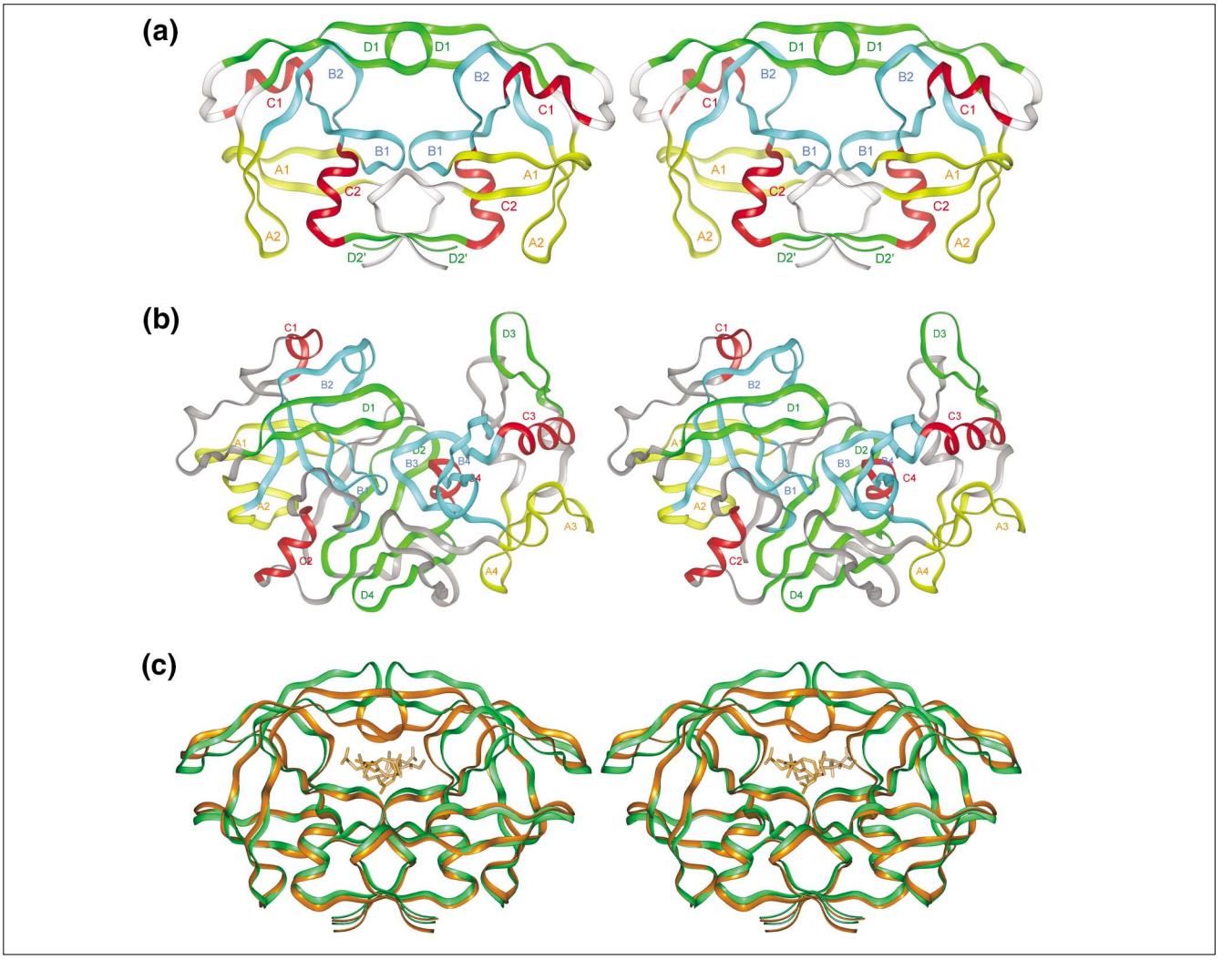

Figure 3
Structural template for **(a)** retroviral proteases compared to that for **(b)** the aspartic protease family. In the symmetrical retroviral dimer (a), loops A1 and A2 are shown in yellow in each monomer, shown as stereo pairs. In the single-chain aspartic protease (b), the corresponding loops are labeled A1 and A2 in the left-side domain and A3 and A4 in the right-side domain. Likewise, loops B1 and B2 in (a) are shown in blue in each monomer of the retroviral dimer, and the analogous loops in blue are labeled B1, B2, B3, and B4 in the single-chain enzymes. Loops B1 in the retroviral enzymes and B1 and B3 in the single-chain enzymes contain the catalytic residues. Helical segments C1 and C2 (red) in (a) are mirrored by segments C1-C4 in (b). Finally, loop D1 in the retroviral monomers provides a double flap structure in (a), whereas the 'half loops' D2' provide the four strands that form a β sheet at the bottom of the dimer. In (b), loop D1 provides the flap on one side only, whereas D3 on the other side is pointing outward. Loops D2 and D4 provide the center of the β sheet at the bottom of these enzymes. **(c)** Movement of flaps in the retroviral protease during ligand binding. This stereo pair shows the movement that occurs upon binding of a ligand to the active site of HIV-1 protease. The 'empty' enzyme structure is shown as a green ribbon and the enzyme following binding is shown as an orange ribbon. The flap residues move downward by approximately 8 Å.

gene and the stop codon after the gag locus is no longer in frame, producing a Gag-Pol fusion polyprotein (Figure 5). The protease embedded within the Gag-Pol polyprotein cleaves itself out by specifically cutting peptide bonds at either end of its sequence. The protease then cleaves additional bonds within the remaining fragment of the Gag-Pol polyprotein to yield reverse transcriptase and integrase, two other important enzymes of the virus [14]. Cleavage of Gag-Pol occurs sequentially and with high fidelity at nine separate, unrelated cleavage sites. The rates of cleavage can differ by up to 400-fold between sites [15]. These differences may be related to different steps in assembly of virions.

Important mutants

Viral species with altered protease sequences arise as a result of the high nucleotide-substitution rate during viral replication. The functional properties of these variant proteases have been the subject of intense study. Some changes occur in regions exposed at the enzyme's surface without significant alteration of the enzymatic properties of the protease;

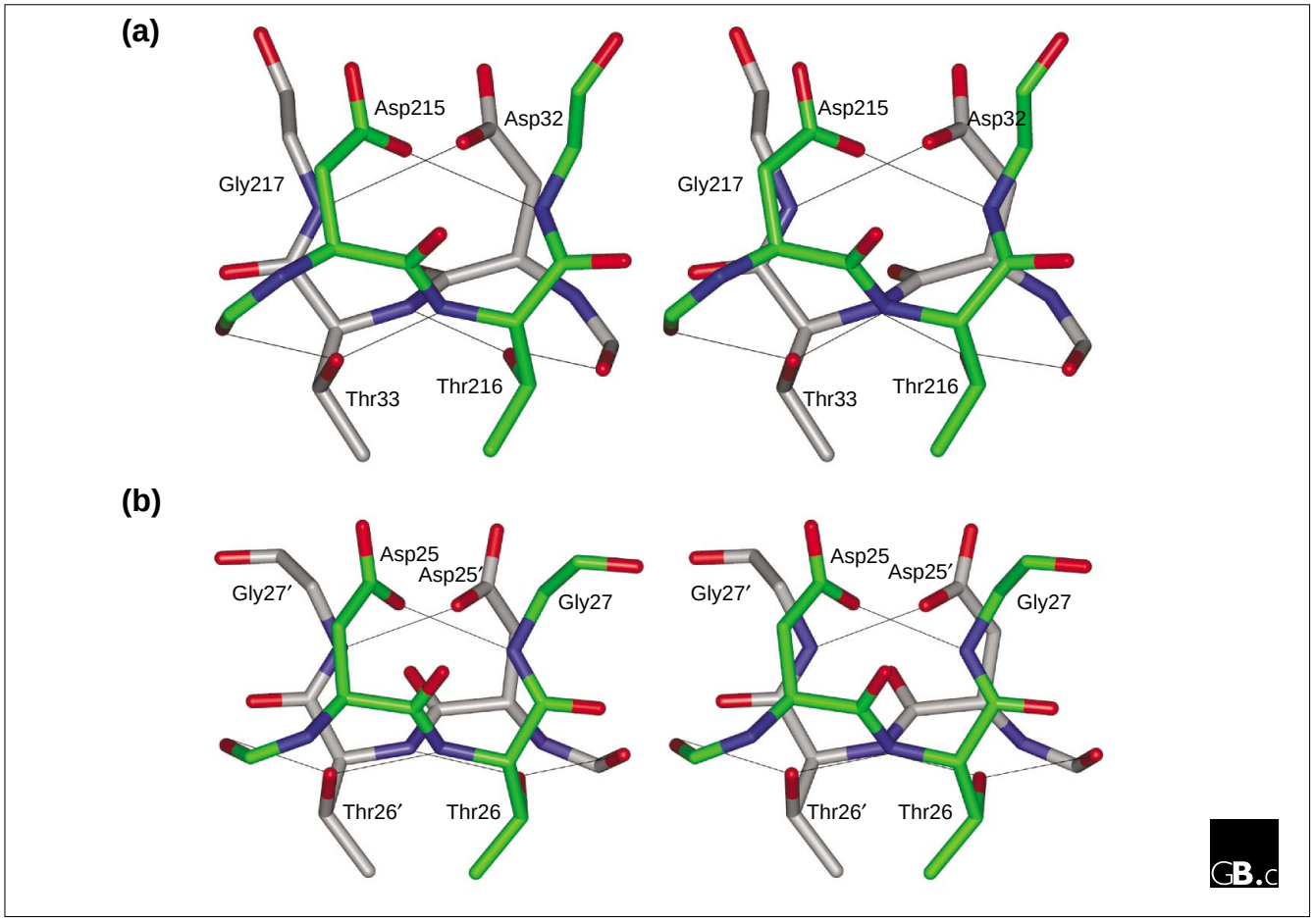

Figure 4
The 'fireman's grip', a stereotypical rigid network structure involving the Asp-Thr-Gly signature sequence in **(a)** the classical aspartic peptidases and **(b)** the retroviral proteases. Amino acids are identified by three-letter codes. In each case, the catalytic aspartic acid residue (Asp) is hydrogen-bonded to the backbone NH group of the glycine (Gly) two amino acids further along in the sequence. In addition, the OH groups of threonine (Thr) are hydrogen-bonded to two points on the opposite domain or monomer, the backbone NH group (blue) of the threonine and to the carbonyl oxygen (red) of the residue before the catalytic aspartic acid.

other changes occur within the binding cavity, leading to changes in the binding of both substrates and inhibitors. The balance between the ability to bind substrates and the interactions with inhibitors will determine the success or failure of the variant protease and hence of the variant virus. If the viral protease has lost the ability to bind an inhibitor tightly, the virus might be able to survive drug therapy with that compound; if, on the other hand, the viral protease has also lost the ability to bind to and cleave the polyprotein, the virus will be unable to replicate successfully. (Figure 6 shows those mutations that have well-defined consequences for function, leading to reduced susceptibility to protease inhibitors.)

In addition to direct effects on the binding of inhibitors to HIV protease, mutations in other positions along the polyprotein sequence can have consequences for polyprotein processing (Figure 7). These events can impact the viability of the virus in both positive and negative ways [16]. For example, it is becoming apparent that mutations in cleavage sites can compensate for changes within the binding cleft of HIV protease. Alterations in the active site will alter the cleavage specificity; alterations in the cleavage site to better match the variant protease could allow the virus to escape inhibition by antiviral compounds, while also maintaining the necessary points of cleavage to produce structural proteins.

Frontiers

Understanding protease function in polyprotein processing and viral replication remains important. Despite the early successes with the development of drugs that control HIV infection by blocking proteolytic processing, the poor bioavailability of inhibitors *in vivo* leads to suboptimal drug levels. The high turnover of the virus (two or three cycles of replication per day) coupled with the high viral load in infected individuals, and the mutation rate has led to the

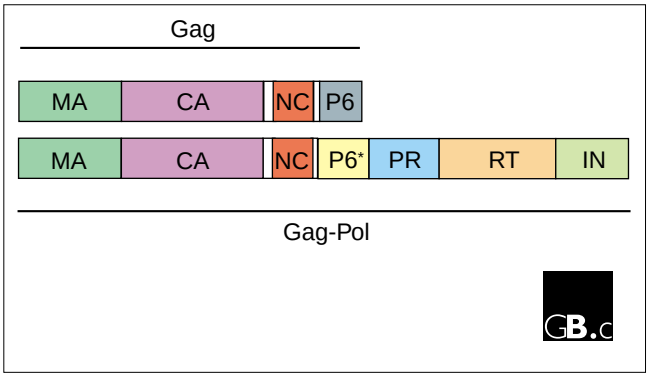

Figure 5
Translational products from the retroviral gag-pol mRNA. In most cases, translation of the gag-pol transcript results in a Gag polyprotein including structural proteins. A translational frameshift within the p6 region allows translation beyond the p6 *gag* gene, resulting in a Gag-Pol fusion protein. The Gag-Pol fusion protein contains a p6* protein, the sequence of which differs from the p6 protein as a result of the frameshift. Abbreviations: MA, p17 matrix protein; CA, p24 capsid protein; NC, p7 nucleocapsid protein; PR, protease; RT, reverse transcriptase; IN, integrase.

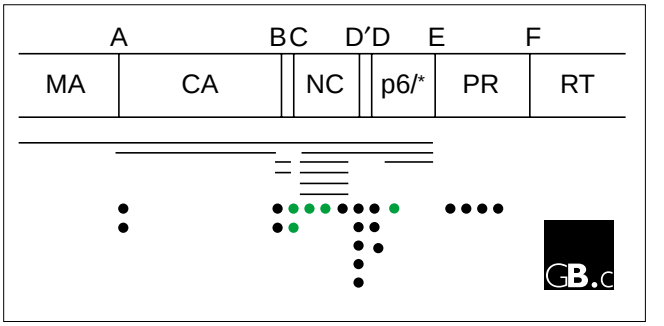

Figure 7
The impact of mutations in the Gag-Pol polyprotein on protease activity. The letters above the diagram indicate cleavage sites; cleavage at sites E and F release the protease; subsequent cleavage at sites C, A, and B produce mature structural proteins. Regions in Gag that impact protease processing have been defined by deletion analysis (underlined). Specific mutations (indicated by dots), particularly at the sites between nucleocapsid ($p7^{NC}$) and p6 or p6* (depending on the reading frame), can alter the rates of protease processing at different cleavage sites (green circles: our unpublished data; black circles: summary of published data. For abbreviations see Figure 5.

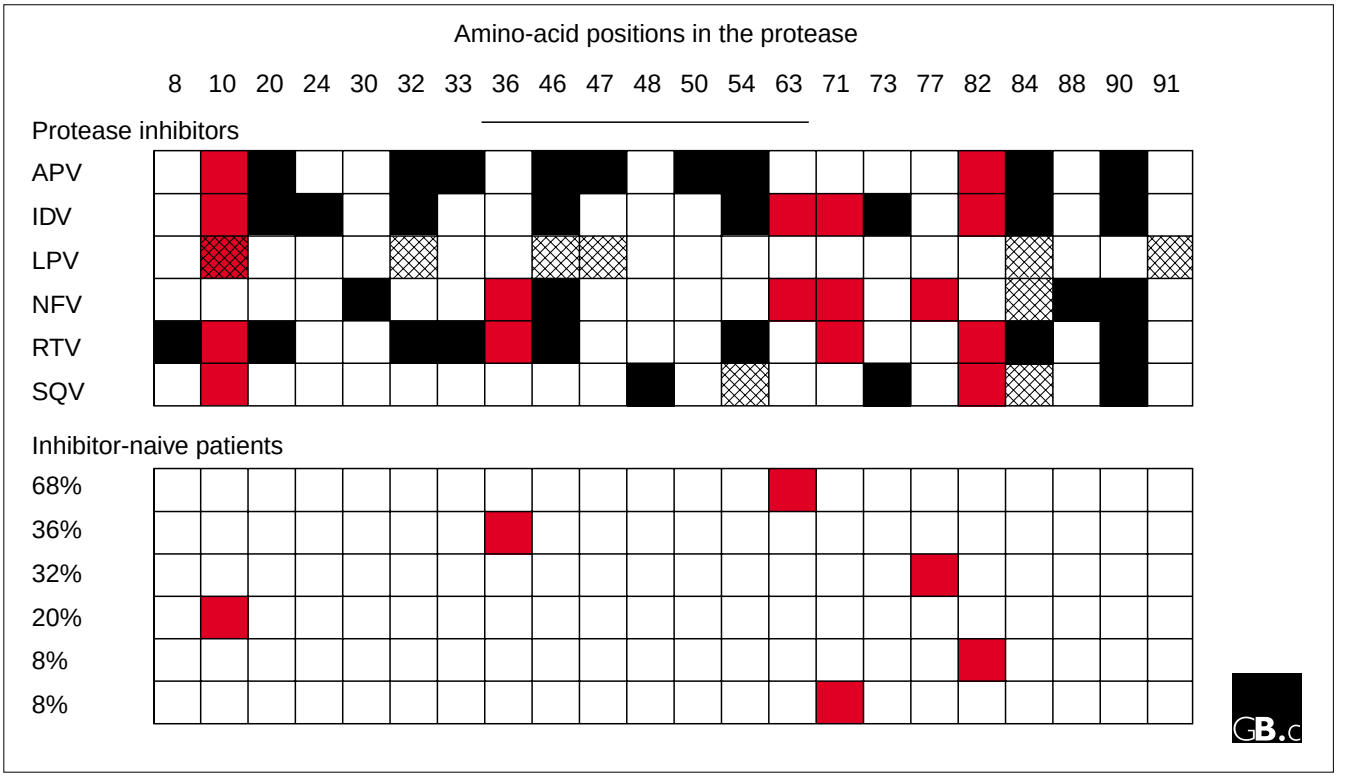

Figure 6
Drug-resistance amino-acid profiles of HIV-1 protease. Protease-inhibitor treatment leads to growth of viruses with changes in specific amino-acid positions. The numbers across the top designate amino-acid positions in HIV-1 protease; the solid line indicates the flap region. Filled boxes indicate mutations that occur in treated patients; red boxes indicate that mutation is seen in patients both before therapy starts (bottom panel) and in patients after therapy has begun (top panel); hatched boxes indicate mutations that occur during passage of virus cultured in the presence of a drug. The percentages in the bottom panel refer to the percentage of clones that contain the mutation indicated in the corresponding row. In the top panel, each row displays the profile of amino-acid changes related to high-level resistance to protease inhibitors that are approved by the US Food and Drugs Administration (FDA) for treatment of HIV-1-infected adults. Abbreviations: APV, amprenavir; IDV, indinavir; LPV, loprinavir; NFV, nelfinavir; RTV, ritonavir; SQV, saquinavir. Only RTV, NFV, and IDV have FDA approval for treatment of children and adolescents.

emergence of viruses resistant to all approved drugs [17]. The variant forms of drug-resistant protease have been expressed and studied biochemically and structurally, and a new round of drug design is underway to target variant forms. One can imagine that this cycle will continue until a universal inhibitor is found that binds tightly to all forms of the viral enzyme. Other approaches, such as the development of peptides that bind to the dimerization interface and block assembly of functional proteases, are also under extensive investigation.

References

1. **The Merops database** [http://www.merops.co.uk]
   A compilation of protease sequence information organized into mechanistic classes. The database provides information on literature, alignments, and links to other databases.
2. Barrie KA, Perez E, Lamers SL, Sleasman JW, Dunn BM, Goodenow MM: **Natural variation in HIV-1 protease, Gag p7 and p6, and protease cleavage sites within Gag/Pol polyproteins: amino acid substitutions in the absence of protease inhibitors in mothers and children infected by human immunodeficiency virus type 1.** *Virology* 1996, **219:**407-416.
   A description of the nucleotide sequence variation present in clones derived from HIV-1-infected patients.
3. **Stanford HIV RT and protease sequence database** [http://hivdb.stanford.edu]
   A compilation of RNA and protein sequences derived from patients in clinical trials undergoing anti-retroviral therapy. These data show the development of resistant protease sequences in response to treatment with specific drugs.
4. Goodenow MM, Perez EE, Sleasman JW: **Genetic variability in HIV-1 in children treated by protease inhibitors.** In *Human Retroviral Infection: Immunological and Molecular Theories*. Edited by Friedman H, Ugen K, Bendinelli M. New York: Plenum Press; 2000, 287-305.
   An analysis of variation in protease sequence in patients infected with HIV-1 in relation to their clinical and immunological status.
5. Wlodawer A, Miller M, Jaskólski M, Sathyanarayana B K, Baldwin E, Weber IT, Selk L M, Clawson L, Schneider J, Kent SBH: **Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease.** *Science* 1989, **245:**616-621.
   This paper describes the first three-dimensional structure of HIV-1 protease, or any protein, produced from chemically synthesized protein. It established the correct protein fold and gave the first glimpse of the active-site pocket.
6. Mulichak AM, Hui JO, Tomasselli AG, Heinrikson RL, Curry KA, Tomich CS, Thaisrivongs S, Sawyer TK, Watenpaugh KD: **The crystallographic structure of the protease from human immunodeficiency virus type 2 with two synthetic peptidic transition state analog inhibitors.** *J Biol Chem* 1993, **268:**13103-13109.
   A comparison of the free protease structure with the inhibitor-bound enzyme structure to illustrate changes in enzyme conformation upon ligand binding.
7. Rose RB, Rose J R, Salto R, Craik C S, Stroud RM: **Structure of the protease from simian immunodeficiency virus: complex with an irreversible nonpeptide inhibitor.** *Biochemistry* 1993, **32:**12498-12507.
   This paper presents the crystallographic structure of the retroviral enzyme that infects non-human primates. This structure is notable as it presents a complex with a large inhibitor.
8. Wlodawer A, Gustchina A, Reshetnikova L, Lubkowski J, Zdanov A, Hui KY, Angleton EL, Farmerie WG, Goodenow MM, Bhatt D, *et al.*: **Structure of an inhibitor complex of the proteinase from feline immunodeficiency virus.** *Nat Struct Biol* 1995, **2:**480-488.
   The three-dimensional structure of the enzyme from the virus that infects cats. The structure is compared with that of HIV-1 and RSV proteases.
9. Miller M, Jaskólski M, Rao JKM, Leis J, Wlodawer A: **Crystal structure of a retroviral protease proves relationship to aspartic protease family.** *Nature* 1989, **337:**576-579.
   The first retroviral protease structure to be determined. This structure established the similarity of the retroviral enzymes to the larger single-chain proteases from other species such as fungi and animals.
10. Gustchina A, Kervinen J, Powell DJ, Zdanov A, Kay J, Wlodawer A: **Structure of equine infectious anemia virus proteinase complexed with an inhibitor.** *Protein Sci* 1996, **5:**1453-1465.
    The crystal structure of the enzyme from the virus that infects horses. This enzyme has interesting variations when compared to HIV-1 protease, such as the size of surface loops and the presence of a helix at a certain position in the structure.
11. Wlodawer A, Gustchina A: **Structural and biochemical studies of retroviral proteases.** *Biochim Biophys Acta* 2000, **1477:**16-34.
    A review of the structural organization of retroviral protease and discussion of drug-resistant forms.
12. Andreeva N: **A consensus template of the aspartic proteinase fold.** In *Structure and Function of the Aspartic Proteinases*. Edited by Dunn BM. New York: Plenum Press; 1991, 559-572.
    The first description of the common structural organization of the aspartic proteinase class of enzymes. This work pre-dated the discovery of HIV-1 protease and other retroviral enzymes.
13. Davies DR: **The structure and function of the aspartic proteinases.** *Annu Rev Biophys Biophys Chem* 1990, **19:**189-215.
    A general review of all aspects of the aspartic protease family, including structure, catalytic mechanism, and inhibition.
14. Goodenow MM, Bloom G, Rose SL, Pomeroy SM, O'Brien PO, Perez EE, Sleasman JW, Dunn BM: **Naturally occurring amino acid polymorphisms in human immunodeficiency virus type 1 (HIV-1) Gag NC p7 and the C-cleavage site impact GagPol processing by HIV-1 protease.** *Virology* 2002, **292:**137-149.
    A description of alterations in the pathway by which HIV-1 protease cuts itself out of a Gag-Pol polyprotein and sequentially cleaves the other sites within the polyprotein.
15. Erickson-Viitanen S, Manfredi J, Viitanen P, Tribe DE, Tritch R, Hutchison CA, Loeb DD, Swanstrom R: **Cleavage of HIV-1 gag polyprotein synthesized *in vitro* - sequential cleavage by the viral protease.** *Aids Res Hum Retroviruses* 1989, **5:**577-591.
    This paper presents analysis of the cleavage of multiple sites within the Gag-Pol polyprotein. This work established the principle of sequential cleavage of the different cleavage junctions.
16. Perez EE, Rose SL, Peyser B, Lamers SL, Burkhardt B, Dunn BM, Hutson AD, Sleasman JW, Goodenow: **HIV-1 protease genotype predicts immune and viral response to combination therapy with protease inhibitors [PI] in PI-naive patients.** *J Inf Dis* 2001, **183:**579-588.
    An analysis of the relationship between the HIV-1 sequence in infected patients and the success or failure of combination therapy. Sequences within the gag-pol region were analyzed.
17. *HIV Resistance and Implications for Therapy*. Edited by Larder B, Richman D, Vella S. Second Edition, Atlanta: Medicom Inc; 2001.
    A study of all sequence variation found in patients undergoing therapy with analysis of the differences that occur depending on the drugs utilized.
18. **GenBank** [http://www.ncbi.nlm.nih.gov/Genbank]
    Database of protein and DNA sequences.
19. **ClustalW** [http://www2.ebi.ac.uk/clustalw]
    A computer program used to align sequences and calculate relatedness.