Crystal structure of prototype foamy virus (PFV) protease-reverse transcriptase fusion (PR-RT) reveals conformational plasticity: implications for function

Jerry Joe E. K. Harrison[1,3], Steve Tuske[1,5], Kalyan Das[1,6], Francesc X. Ruiz [1], Joseph D. Bauman[1,5], Paul Boyer[4], Jeffrey J. DeStefano[7], Stephen H. Hughes[4], Eddy Arnold[1,2,*]

[1]Center for Advanced Biotechnology and Medicine (CABM), NJ, USA; [2]Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, NJ, USA; [3]Department of Chemistry, University of Ghana, Legon, Ghana; [4]National Cancer Institute, Drug Resistance Program, Frederick, MD, USA; [5]Present address: Amicus Therapeutics, Gene Therapy Center of Excellence, 3675 Market Street, Philadelphia, PA, USA; [6]Rega Institute and Department of Microbiology and Immunology, KU Leuven, Belgium; [7]Department of Cell Biology and Molecular Genetics, University of Maryland College Park, College Park, MD, USA

*Corresponding Author

E-mail: arnold@cabm.rutgers.edu

Abstract

Proteolytic processing of the retroviral Pol polyprotein precursor produces protease (PR), reverse transcriptase (RT), and integrase (IN), except in foamy viruses (FVs) where only the IN domain is released. Here, we report the 2.9 Å resolution crystal structure of the mature PR-RT from prototype FV (PFV) needed for processing and reverse transcription. The monomeric PFV PR exhibits similar architecture as the HIV-1 PR but the N- and C-terminal residues are unstructured. A C-terminal extension of the PR folds into two helices that supports the RT palm subdomain and anchors the PR next to the RT. The subdomains of RT: fingers, palm, thumb, and connection, and the RNase H domain, are connected by flexible linkers and spatially arranged similarly to those in the HIV-1 RT p51 subunit. Significant spatial and conformational domain rearrangements are required for nucleic acid binding. This offers structural insight into retroviral RT conformational maturation and architecture of immature enzymes.

Introduction

Foamy viruses (FV), which remain the most ancient retroviruses, share conserved genome organization and architecture of structural and enzymatic proteins with their evolutionarily related retrotransposons (Hutter et al., 2013; E. G. Lee et al., 2013; Linial, 1999; Yu et al., 1999). Synthesis of polyproteins from polycistronic mRNAs before proteolytic processing by cognate protease(s) into functional entities, is a unique feature of many pathogens, including RNA viruses, retroviruses, and retrotransposons (Baldwin & Linial, 1998; Roy & Linial, 2007; Shin et al., 2012; Yu et al., 1996). Among the numerous proteins encoded in the retroviral genome, the architecture of the enzymes reverse transcriptase (RT) and protease (PR) remain perhaps the most conserved (Ding et al., 1998; Nowak et al., 2014; Tozser, 2010) with a common mechanism of nucleic acid synthesis and proteolytic processing even when their sequence identity is low. Lentiviruses such as HIV-1 synthesize their enzymatic proteins as a Pol polyprotein precursor which is subsequently proteolytically processed into mature enzymes, as part of a Gag-Pol fusion from the same mRNA using a -1 translational frameshifting. Meanwhile, Spuma retroviruses such as the prototype foamy virus (PFV) synthesize their Gag and Pol polyproteins from separate mRNAs alternatively spliced from their genomic RNA (Lochelt & Flugel, 1996; Pfrepper et al., 1998). FV polyproteins undergo only minimal proteolysis during maturation, where the integrase (IN) is the only domain processed from the Pol, by cleavage between the RNase H domain and IN, while a 3-kDa C-terminal peptide is processed from the Gag polyprotein (Yu et al., 1996). The mature PR-RT therefore contains an N-terminal PR and C-terminal RT, both of which are functional [reviewed in (Wohrl, 2019)].

The limited proteolysis of Gag results in infectious foamy virus (FV) particles that are morphologically indistinguishable from the immature lentiviral capsids, though this minimal processing is absolutely required for infectivity in FVs (E. G. Lee et al., 2013; Yu et al., 1996).

The RT of FVs is monomeric in solution and is responsible for copying the (+) FV ssRNA genome into dsDNA for integration into the host chromosome. This process occurs largely during assembly and budding in producer cells concomitantly resulting in infectious viral particles containing dsDNA (E. G. Lee et al., 2013; Rethwilm & Bodem, 2013; Yu et al., 1999).

The aspartyl PR, which is responsible for the proteolytic processing of polyproteins from retroviruses, is functional as a homodimer (Katoh et al., 1989). Hence the PR embedded in the PR-RT must dimerize to activate the protease. The mechanism(s) by which this happens remains under investigation, and potentially involves viral RNA (Hartl et al., 2010; Wohrl, 2019). Many structures of retroviral PRs both alone and in complex with relevant substrates and inhibitors are available in the literature, which help to explain mechanisms of catalysis and inhibition. Similarly, several retroviral RT structures and their complexes with nucleic acids and inhibitors have also provided important structural blueprints that elucidate the mechanism of catalysis, inhibition, and drug resistance. Structural information on the domain organization of Pol polyproteins has eluded researchers so far, resulting in a lack of clear understanding of the mechanism by which retroviral PRs dimerize (in the context of the Pol polyprotein) which allows it to cleave other proteins including itself into functional entities.

The asymmetric conformational maturation of heterodimeric HIV-1 RT, with the p66 (catalytic) and p51 (scaffolding) subunits, remains largely an unsolved puzzle in retrovirology. However, there are strong suggestions (albeit with limited structural evidence) that monomeric HIV-1 p66 would adopt a more thermodynamically stable p51-like fold in solution, occasionally sampling the open conformation seen in HIV-1 p66 in the heterodimer, prior to homodimerization and subsequent maturation (Wang et al., 1994; Zheng et al., 2014; Zheng et al., 2015). In an effort to shed light on many of the missing pieces in the retroviral structural biology puzzle, a systematic

investigation of the Pol polyprotein of PFV has been carried out and we report here the crystal structure of the full-length PFV PR-RT fusion polyprotein at 2.9 Å resolution.

This structure defines the structural organization of FV PR-RT and offers insight into the potential mechanism of dimerization of PR in the context of a polyprotein, thereby offering a glimpse into the initial events leading up to the proteolytic processing of polyproteins. The PFV RT domains are arranged relative to each other similarly to the structure of HIV-1 RT p51. This compact p51-like conformation is consistent with predictions in the literature and defines the structural organization of a monomeric RT precursor such as monomeric HIV-1 RT p66 subunit prior to isomerization to a p51-like conformation. This structure to the best of our knowledge is the first functionally relevant retroviral Pol polyprotein providing a framework for investigating other retroviral polyproteins. This structure will serve as a useful model for investigating the sources of high fidelity and processivity of FV RT especially as FV based vectors for gene therapy (Erlwein & McClure, 2010; Lindemann & Rethwilm, 2011; Trobridge, 2009) are becoming increasingly very popular. It is also anticipated that it will be a key platform for the design of new therapeutics against HIV and other retroviruses.

Results

Engineering a proteolytically resistant mutant of PFV PR-RT for bacterial expression

Wild-type PFV PR-RT is highly susceptible to proteolysis when expressed in bacterial strains (Boyer et al., 2004). The breakdown products complicate the development of purification schemes that result in good quality pure protein needed for crystallographic studies. The extra steps required to obtain useful materials make purification laborious and terrible yield of final product notwithstanding the need for large doses of protease inhibitors and reducing agents. Shared

composition between the full-length protein and the proteolytic products further complicate structural and biophysical studies. For crystallization, these impurities could selectively crystallize, crystallize alongside the full-length protein or get incorporated into the crystal lattice of the full-length protein and create lattice defects and/or structural heterogeneity that could reduce the resolution significantly. Each of these scenarios has deleterious consequences for obtaining crystals suitable for X-ray crystallographic studies.

Instability of proteins due to the composition of their primary sequence has been known for a long time (Guruprasad et al., 1990). In an effort to improve the proteolytic stability of PFV PR-RT [with a protease null mutation (D24A), designated wild-type (WT)], *in silico* mutagenesis of the primary sequence of WT PFV PR-RT was carried out by replacing each amino acid at each position with the other 19 and computing an instability index resulting from such substitution on the ExPASy ProtParam server (http://web.expasy.org/protparam/) using a Python script; see details in Materials and Methods. The instability index computed is a weighted sum of the dipeptide composition of proteins and is based on a strong correlation found to exist between stability of proteins and their dipeptide composition. It is therefore used as a qualitative measure of the *in vivo* stability of proteins. Proteins with instability estimates below 40 are considered stable while those above that threshold are considered unstable (Guruprasad et al., 1990). The instability index for the WT protein was computed to be 39.75 which is very close to the upper limit of the stability index. H507 and S584 were identified as the instability hotspots in the primary sequence where several other amino acids were preferred at these positions than the residues themselves. Two mutations, H507D and S584K, which computationally produced the most stable protein when substituted at these positions, were selected based on protein stability and biophysical considerations, in addition to C280S which is known to improve the solution behavior of HIV-1

RT. This PR-RT mutant C280S/H507D/S584K, designated CSH, resulted in at least a 4-fold increase in expression compared to the WT in bacterial strains and could be purified in two steps (nickel affinity and heparin) to at least 95% purity as judged from SDS-PAGE analysis without any protease inhibitors or reducing agents (Fig. 1).

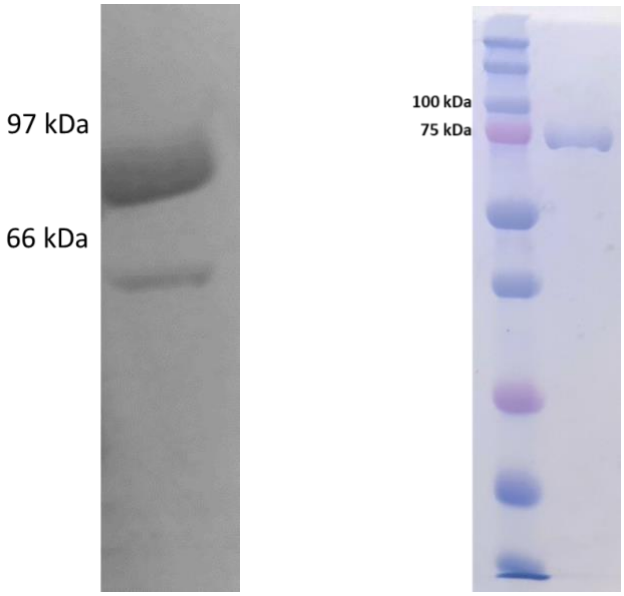
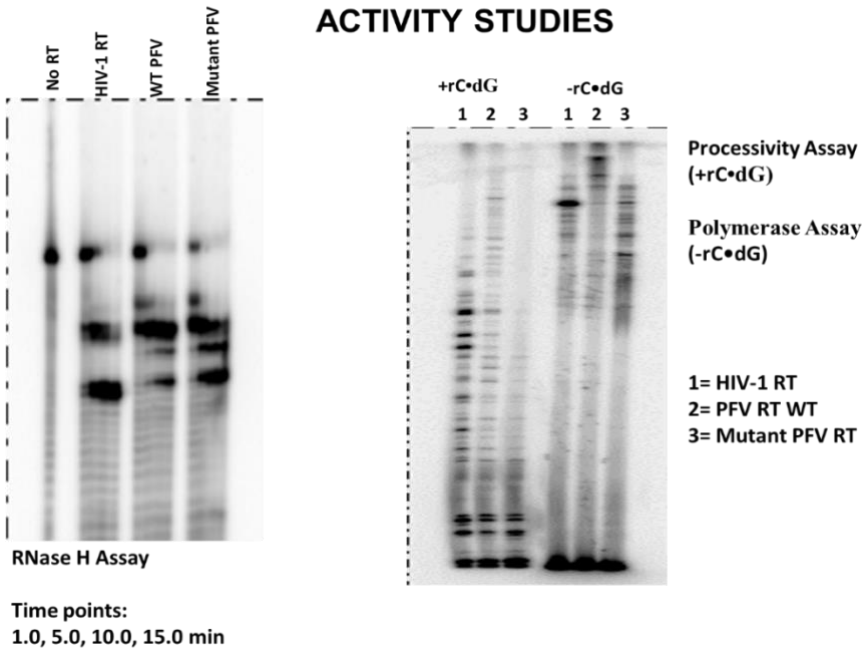


Figure 1: SDS-PAGE gel of the WT PR-RT after purification (left) and the CSH mutant (right).

Enzymatic assays

Changes to the primary sequence of proteins can have unintended consequences on the protein structure and/or the activity. To verify whether the mutations have had any impact on the enzymatic activity of the protein, the polymerase as well as the ribonuclease activities were assayed for the WT protein and the mutant using nucleic acid substrates. Compared to the WT, the CSH mutant showed only a modest reduction in polymerase activity while the ribonuclease activities were very similar. Hence while the ribonuclease activity remained similar to the WT (Supplemental Fig. 1). The processivity which measures how long polymerases engage their nucleic acid substrates enabling them to carry out consecutive dNTP incorporation into a growing

chain was also measured. It was observed that compared to the WT, the processivity of the mutant had decreased by about 5-fold (Supplemental Fig. 1). This suggests that the mutations decrease the ability of the enzyme to engage the nucleic acid for long periods without significantly affecting its catalytic ability. The rationale could be that replacement of His507 with an Asp, a reversal of positive charge to a negative charge could result in nucleic acid backbone repulsion.



Supplemental Figure 1. Ribonuclease activity of PR-RT (left) and polymerase and processivity activity (right), using poly (rC). oligo(dG) substrate.

Table 1. Data collection and refinement statistics.

	Selenomethionine PFV PR-RT CSH mutant	**Native PFV PR-RT WT D26N+CSH mutant**
PDB ID	7KSE	7KSF
Wavelength (Å)	0.98	0.98
Resolution range (Å)	50.0–3.0 (3.18–3.0)	38.03 - 2.9 (3.004 - 2.9)
Space group	C2	C2
Unit cell (Å / deg)	240.38 53.35 74.92 / 90 100 90	228.21 52.54 75.43 90 90.09 90
Unique reflections	19100 (2990)	19443 (1752)
Multiplicity	30.4 (29.7)	5.1 (5.1)
Completeness (%)	99.7 (98.0)	96.1 (88.4)
Mean I/sigma(I)	10.6 (2.7)	7.1 (1.0)
Wilson B-factor ($Å^2$)	70.312	92.105
R_{merge}	0.331 (1.973)	0.128 (1.532)
R_{meas}	0.337 (2.007)	0.157 (1.888)
R_{pim}	0.061 (0.367)	0.064 (0.812)
$CC_{1/2}$	0.996 (0.818)	0.993 (0.334)
R_{work}	0.277 (0.382)	0.224 (0.370)
R_{free}	0.312 (0.376)	0.258 (0.381)
Number of non-hydrogen atoms	5911	5915
macromolecules	5898	5905
ligands	1	1
solvent	12	9
Protein residues	743	743
RMS bonds (Å)	0.007	0.004
RMS angles (deg)	1.40	0.71
Ramachandran favored (%)	89.40	94.43
Ramachandran allowed (%)	10.05	5.30
Ramachandran outliers (%)	0.54	0.27
Rotamer outliers (%)	0.46	0.77
Clashscore	23.68	11.65
Average B-factor ($Å^2$)	85.66	103.78
macromolecules	85.71	103.78
ligands	110.53	123.65
solvent	62.54	103.14

Statistics for the highest-resolution shell are shown in parentheses

Architecture of PFV PR-RT

The aforementioned higher purity and yield of the PFV PR-RT CSH mutant *versus* the WT enabled broad robotic crystallization screening. Optimization of a crystallization condition eventually yielded suitable crystals that diffracted X-rays to around ~3 Å resolution. The condition found for the CSH mutant was successfully applied to crystallize WT PFV PR-RT. Both the PFV PR-RT CSH mutant and WT crystallize as a monomer with a single molecule per asymmetric unit in space group C2; the structures are very similar (RMSD =0.785 Å), and unless indicated the results described pertain to the WT structure. The structure was solved by using the anomalous signal from the selenomethionine (SeMet)-substituted CSH mutant, and the model obtained from the SeMet data was used as the search model in molecular replacement with the WT data (Table 1 and Fig. 2A-B).

Even though the sequence homology of PFV with other retroviruses is generally low, typically less than 30%, the individual structural elements of retroviral polymerases and proteases are highly conserved (Fig. 2). Notable differences however exist in terms of relative domain positioning and arrangements, detailed in the subsequent sections, which offer potential insights into their unique functionalities and subtle differences in substrate binding specificities, processivity and fidelity.

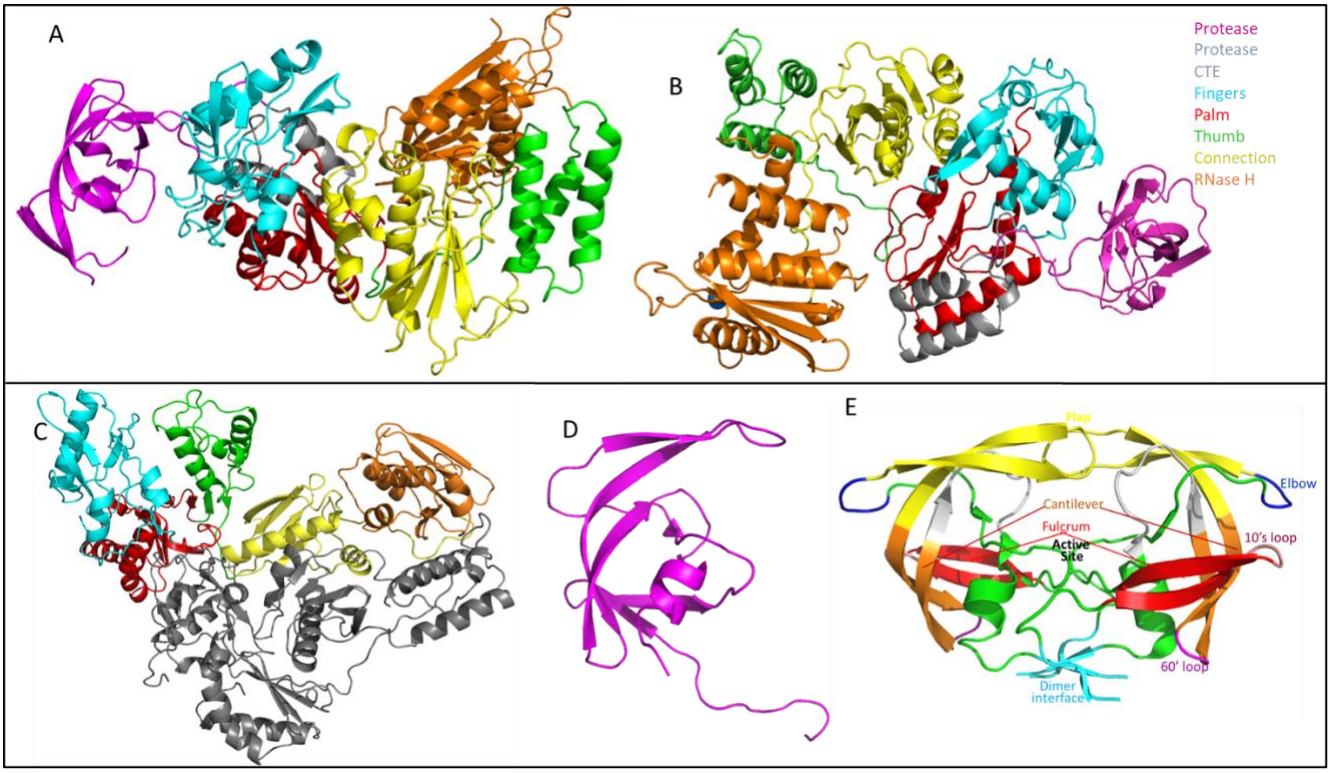


Figure 2. Architecture of the PFV PR-RT colored according to subdomains, with (B) oriented with a 2-fold rotation about the y-axis with respect to (A). C: HIV-1 RT (p66-color coded by subdomain, p51-grey). D: Structure of monomeric HIV-1 PR. E: Annotated diagram of HIV-1 PR dimer.

Seven beta-sheets (β1'-β7'), two short helices (αA' and αB'), and random coils define the structure of the monomeric protease (Figs. 2D-E). The well folded single unit of the protease which functions as a homodimer exhibits the closed barrel-like core domain constituted predominantly by β-sheets similar to a monomeric protease of HIV-1. Residues 1-4 which form part of the dimerization interface in the mature PR inferred from mature HIV-1 PR structures and residues 5-9 remain unstructured with no defined electron density for the first six residues in these structures (Fig. 2). Strands β1' and β2', which form the fulcrum of the PR (residues, 12-15, 20-25), are connected by a hairpin loop A1, (residues, 16-19) the so-called 10's loop (Sheik Amamuddy et al., 2018). The continuous hairpin loop bordered on each side by β2' and β3' define the "fireman's grip" which harbors the conserved DSG catalytic triad (residues 24-26), at the active site. This large loop tilts slightly from orthogonality relative to the core and points away from the RT,

exposing the active site to solvent. An alpha helix (αA', residues 36-38) in a typical pepsin-like protease fold extended by a loop forms the "flap elbow" of the protease following β3'. The positioning of this helix is one of the ways PFV PR differs from HIV-1 where this helix instead is a large loop. This further extends to β4' (residues 45-49) and β5' (residues 58-68) which constitute the flap domain.

The hairpin formed by these strands forms the tip of flap (residues 50-57) which opens and closes to enable substrate binding and release, respectively. Consistent with its function, this region is disordered with less well-defined electron density in this structure. β5' (residues 58-68), which traverses almost the entire length of the PR core, is connected by a short loop (residues 69-70) at its C-terminus to β6' (residues 70-80), which defines the exosite that forms the hinge, together with the mid-section of β5', allowing the tip of the flap to open and close to permit substrate binding and release. β6' (residues 70-80) is however twisted away from β5' in a way that creates a gap between the flap region and the active site loop which forms the substrate binding cavity upon dimerization (Fig. 2). The large P1-loop which lines the walls of the active site connects β7' (residues 85-87) and β6', which form the only parallel sheet in the structure.

A short coil and a short helix (half-turn, αB') further link the C-terminus of β7' to a long random coil that span the C-terminus of the protease domain. The only helix in HIV-1 PR is at the C-terminus. The C-terminus of the PR, which together with the N-terminus forms a major portion of the dimerization interface, is also unstructured without well-defined density in our structure. Whether these would form beta sheets upon dimerization as observed in HIV-1 is unknown. The first 100 residues of PFV PR-RT are sufficient to define the entire protease as observed in HIV-1 (Ishima et al., 2003).

Additional residues at the protease C-terminus have been observed through multiple sequence alignments (Supplemental Fig. 2). The residues 101-143, which were initially not known to be part of either PR or RT, are hereby designated as the protease C-terminal extension (PR-CTE). While HIV-1 does not harbor a C-terminal extension, other retroviruses like Moloney murine leukemia virus (MoMLV), mouse mammary tumor virus (MMTV) and Mason-Pfizer monkey virus (MPMV) contain this additional domain of varying lengths (Supplemental Fig. 2). This CTE forms a long linker stretching from the C-terminus of the PR to the N-terminus of RT, by forming three helices (designated αC', αD', αE') that pack against helices F, G, and strand 4 of the RT (Fig. 3), the equivalents of helices E and F and strand 6 in HIV-1 RT, which form the "floor" of the palm containing the polymerase active site. This stabilized packing with an extensive hydrophobic network forms an extended floor of the palm and also anchors the PR monomer to the RT close to the fingers without any buried surface area between them. This arrangement would allow PR to swing open or tilt to enable dimerization without any anticipated steric interference since the linker is long enough to allow the protease to detach from the RT but anchored in place to form some kind of a dimer.

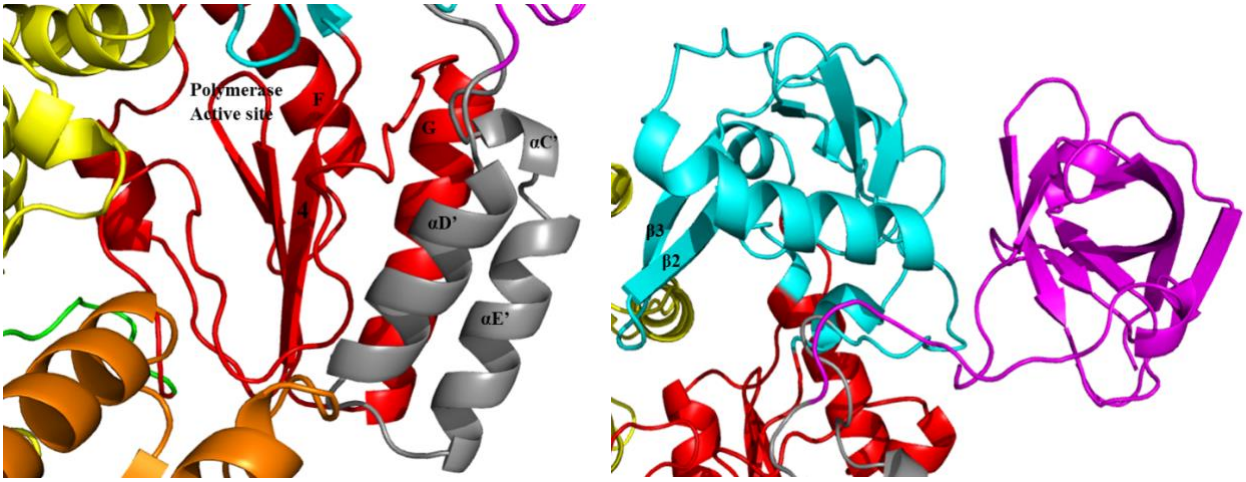


Figure 3. Left: Interface formed by the PR-CTE (in grey) and the palm subdomain (in red). Right: Interface formed by the PR monomer (pink) and the fingers subdomain of RT (cyan).

Two PR-RT structures positioned *in silico* by two-fold symmetry generates a dimeric PRs in a configuration poised for catalysis without extensive steric clashes with the RTs (Fig. 4), suggesting that this structure could be proteolytically active. While this does not offer insights into distal interactions within the RTs, it offers a glimpse of a plausible mechanism of PR dimerization and how the dimeric PR can be accommodated between the two polyproteins. Extensive interaction between helices F and G of the RT are expected to further stabilize the dimeric protease for catalysis. The substrate-binding groove of the PR generated *in silico* appears much wider than that of HIV-1 (Fig. 4), which offers a plausible reason why HIV-1 PR inhibitors such as tipranavir, darunavir, and indinavir (Hartl et al., 2010) do not inhibit PFV PR. The possibility of steric clashes between the inhibitor and side chains of residues close to the active site cannot be ruled out.

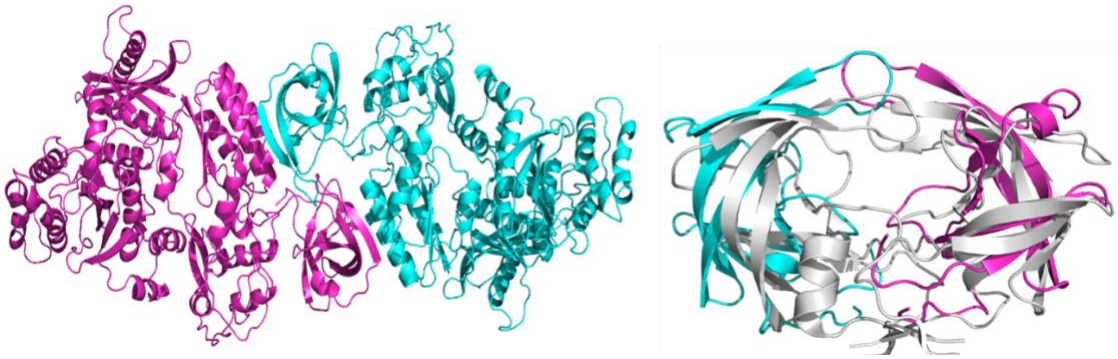


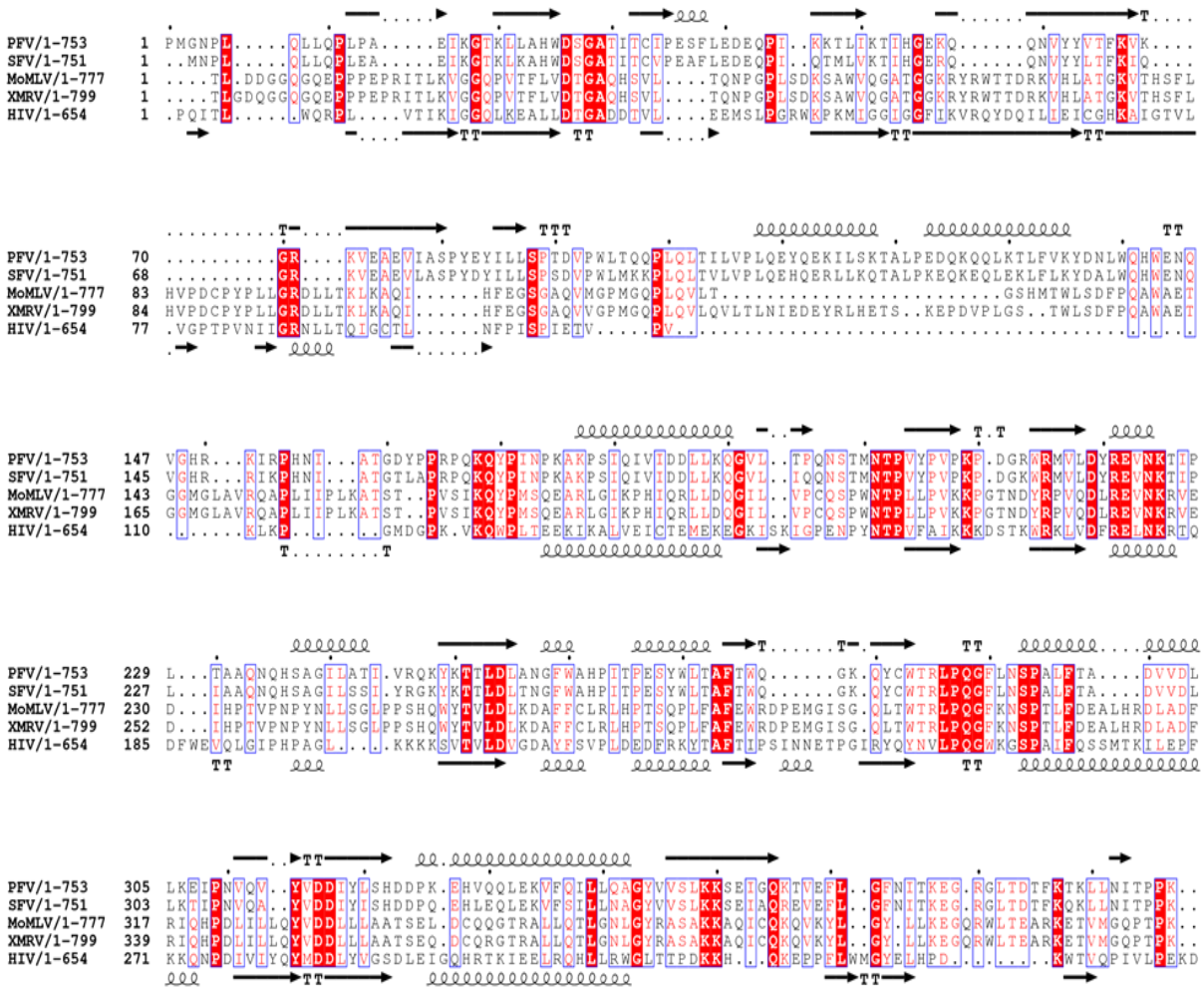
Figure 4: Dimers of PR-RT generated *in silico* by a 2-fold rotation. (left) Superposition of dimeric HIV-1 PR (PDB ID 2HB4 (grey) on dimeric PFV PR generated *in silico* by 2-fold rotation (right).

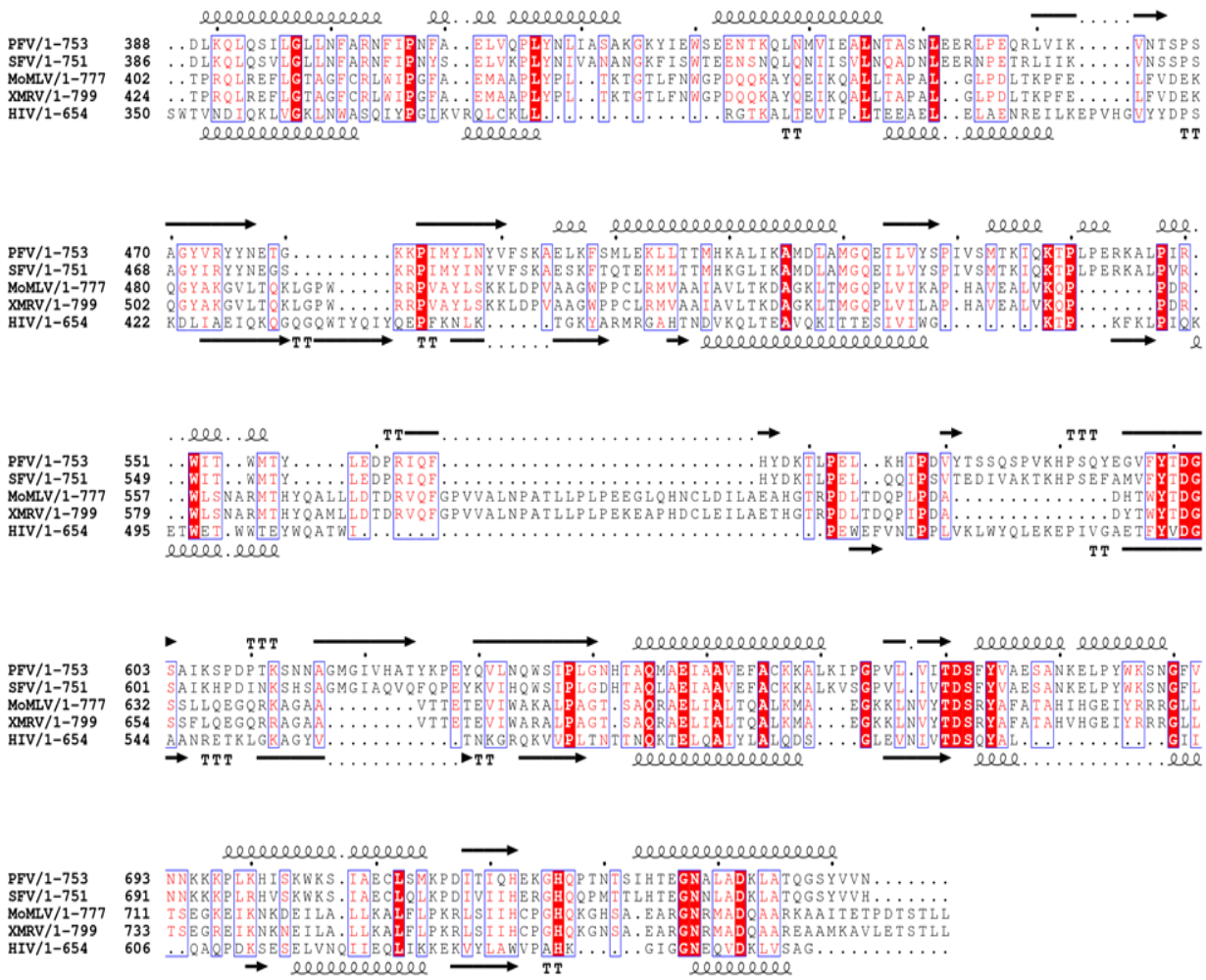
RT Domain

Retroviral RTs and other polymerase structures solved resemble a right hand and so the domains in these proteins have been historically named after the corresponding segments of a hand. Even though the RT domain of the structure of the PFV PR-RT does not resemble the right hand in its

current configuration, all of the canonical subdomains with their unique secondary structural features as identified in retroviral polymerase structures solved (fingers, palm, thumb, and connection, and RNase H domain), are present (Figs. 2A-C). The positioning of these subdomains relative to each other compared to other retroviral RTs is however different. The fingers subdomain of PFV PR-RT (144-228, 260-290) is characterized by four helices (αA, αB, αD, and αE), five strands, and random coils beginning with a long N-terminal unstructured region that terminates at helix A, a characteristic shared with other RTs such as HIV-1 and MoMLV. Three anti-parallel β-sheets are sandwiched by three helices in the hydrophobic core of the fingers similar to other RTs. β2 and β3 define a hairpin that points towards the putative nucleic acid binding cleft reminiscent of the β3-β4 hairpin in HIV-1 p66 implicated in dNTP binding (Fig. 3).

The palm of PFV PR-RT (229-259, 291-367), surrounded by the connection, fingers and PR-CTE subdomains, comprises three helices and four anti-parallel β-sheets similar to other RTs (Fig. 2). The sheets which harbor the conserved catalytic motif YVDD (314-317), together with the primer grip (355-368), constitute the polymerase active site (Fig. 5). The primer grip, observed in other RTs as a short β-hairpin, is a somewhat large loop pointing towards the active site hairpin. It must be pointed out that amino acid residues that form the primer grip hairpin loop are conserved in PFV, MoMLV, XMRV, and HIV-1. We surmise that this flexible primer grip architecture may reduce the constraints imposed by rigid β-sheets in other RTs, which would enable easier primer orientation during nucleic acid synthesis and thus contribute to the higher processivity of PFV compared to other RTs.


PFV/1-753
SFV/1-751
MoMLV/1-777
XMRV/1-799
HIV/1-654



Supplemental Figure 2. Secondary structural alignments of some retroviral PRs and RTs with PFV generated using the program Clustal.

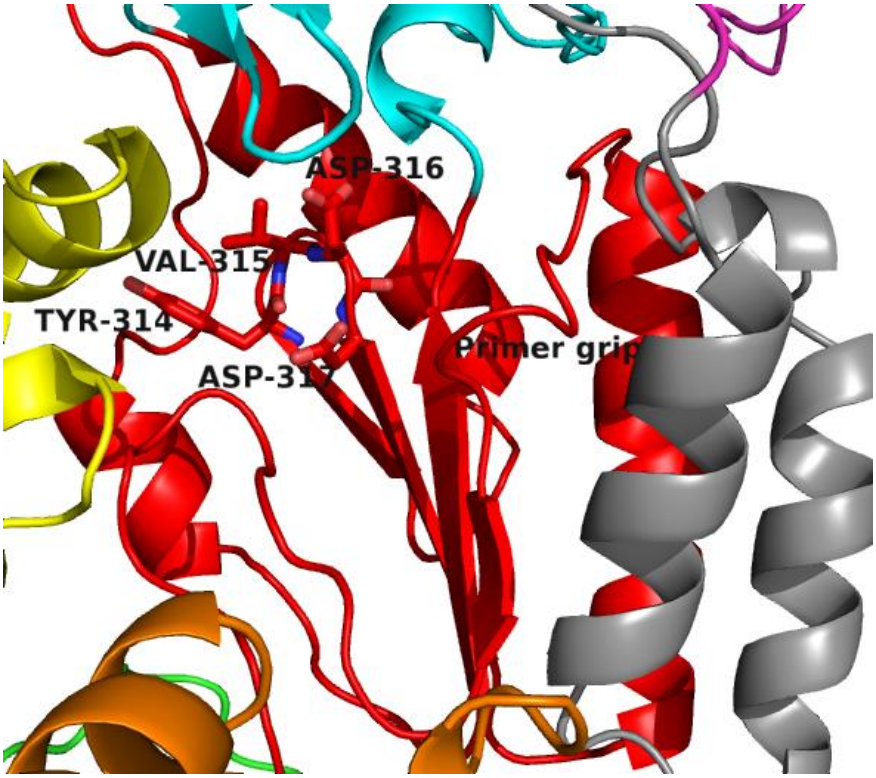


Figure 5. The PFV RT polymerase active site residues and some conserved structural elements found in retroviral RTs

It has been observed that the PFV RT mutant V315M, which mimics the active site of HIV-1 RT, leads to 50% loss of polymerase activity in virions, with no observable full-length cDNA detectable in transfected cells and the processivity greatly reduced (Boyer et al., 2004; Rinke et al., 2002). From a structural standpoint, replacement of valine with the bulky side chain of methionine could potentially distort the positioning of the primer grip, hence slowing down the rate of nucleotide incorporation needed to complete cDNA synthesis. The conserved residue D254, situated at the tip of β4 near the polymerase active site loop formed by β7, β8, house D316 and 317 which collectively constitute the catalytic aspartate triad (β9, β10 of HIV-1) (Fig. 5). While helix F (E in HIV-1 RT) is oriented ~100º relative to helix G (F in HIV-1 RT) above the active site, without the kinks observed in HIV-1 RT, helix G packs against strands 4 and 8 similar to

HIV-1. The fingers and palm subdomains remain in the same relative position compared to other RTs. However, the thumb (378-449), characterized by a three-helix bundle, is moved away from the palm and positioned next to the connection subdomain (450-590), making extensive contacts with it (Fig. 6).

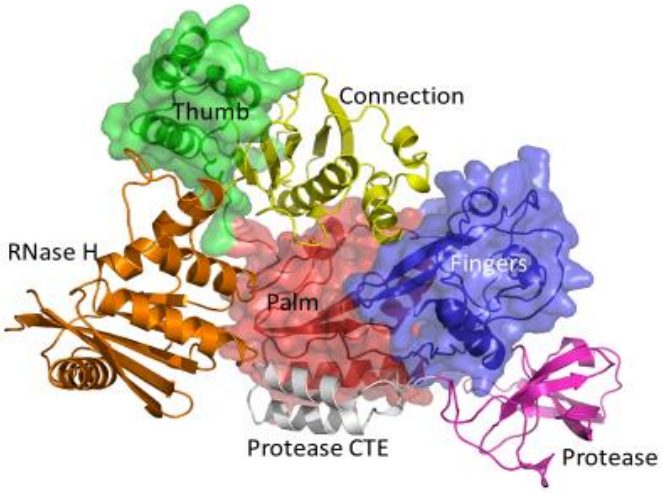


Figure 6. Surface rendering of the fingers-palm-thumb subdomains in PFV PR-RT with the connection and RNase H backbones in ribbons.

The connection domain, characterized by five-stranded mixed beta-sheets stabilized by five helices, is snuggled between the fingers/palm and the RNase H subdomain/thumb on opposite sides and tilted slightly towards the polymerase active site, sitting almost parallel to the palm of the RT compared to its position in HIV-1 p66 and other RTs where they are orthogonal (Fig. 2). This differs from its position in HIV-1 p66 and other RTs, where it lays orthogonal. This compact fold allows helix L (K in HIV-1 RT) to pack against H, while helix O (L in HIV-1 RT) tilts towards the active site loop positioning T552 to form a hydrogen bond with Y314, T556 with H236, and H507 forms a bifurcated hydrogen bond with S237 and N364 (Fig. 7). The H507D mutant employed in initial studies also makes these contacts and crystallizes much faster compared to WT, while H507F/ I mutants which lose these hydrogen-bonding interactions, did not crystallize

after several attempts. This underscores the importance of these hydrogen bonds in stabilizing this conformation. The C-terminus of helix L also makes extensive contacts with the primer grip and helix C of the palm.

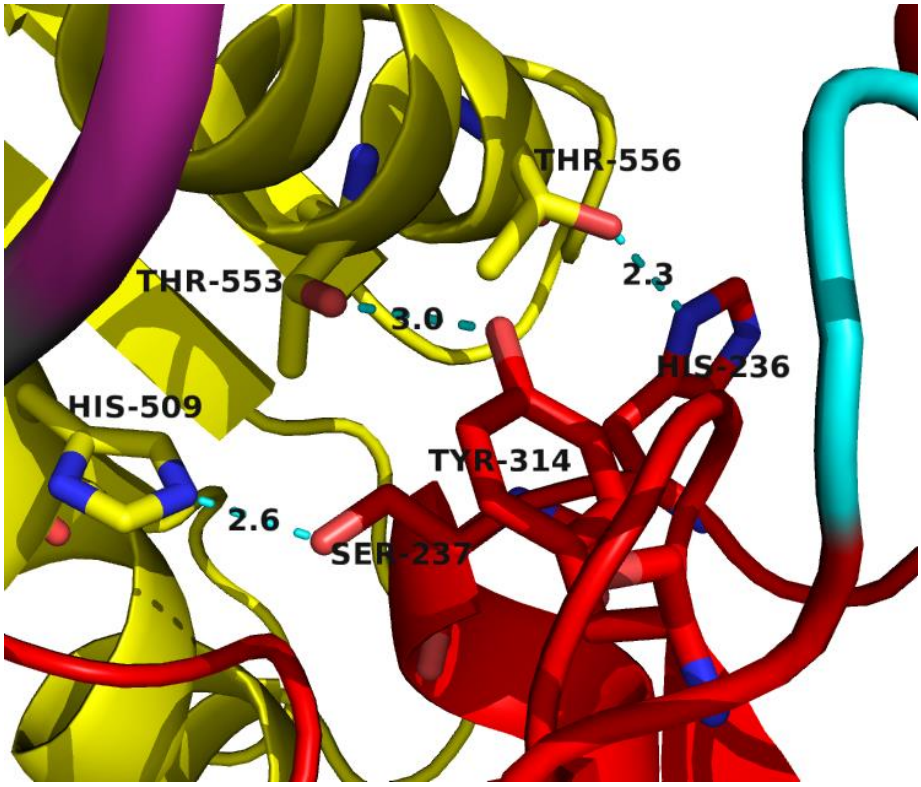


Figure 7. Hydrogen-bonding interactions between connection and palm in PR-RT. Distances shown are in Å.

The folding of the connection onto the palm in this way results not only in significant interactions with the palm and thumb but also partially blocks access to the polymerase active site. Interactions between the connection and the fingers appear very remote with the only relatively close contact occurring between R221 and L541. This is in stark contrast to HIV-1 p51 where the positioning of helix L and β-strand 20 positions the loop between them into close contact with the fingers (Fig. 8).

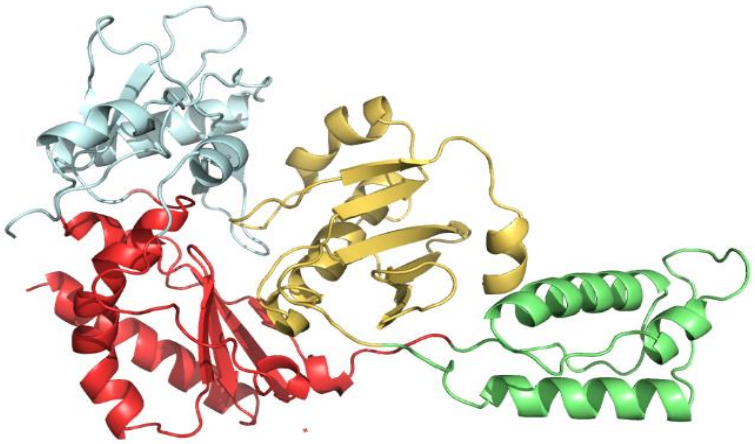


Figure 8. Relative positioning of individual subdomains of HIV-1 RT p51 (color coded as in Fig. 2).

The three-helix bundle characteristic of the thumb (αH, αI, αJ), together with helix L, sandwiches three beta-sheets (9,12,13) of the connection domain, concomitantly forming an extensive interface through a mixture of hydrophobic and van der Waals interactions and hydrogen bonds (Fig. 2). Even though the antiparallel beta-sheets at the base of the thumb which help position it uprightly close to the fingers in HIV-1 RT p66 are seen as random coils in this structure, the thumb is still in an upright position parallel to the connection domain aided by its interactions with the connection and RNase H. The thumb in HIV-1 p51 (Fig. 8) is extended and detached from the connection in a way that exposes helix H to form the interface with the RNase H domain in the heterodimer. Compared to PFV PR-RT thumb, this represents close to 90° rotation (compare Figs. 2 and 8). It must be noted that while some hydrophobic residues are buried in these interfaces between the connection/palm/thumb, the key anchoring interactions are actually hydrogen bonds and van der Waals interactions. Since this machinery must undergo extensive subdomain rearrangement to function as a polymerase, these relatively 'weak' interactions ensure that the energetic barrier to isomerization from a p51-like to a p66-like conformation is lower as compared to HIV-1 p51 where these interfaces are predominantly hydrophobic.

RNase H

The RNase H domain (593-751) also comprises an asymmetric arrangement of five-stranded mixed beta-sheets and four α-helices with the beta-sheets sandwiched between a three helix bundle formed by helices αA_R, αB_R, αC_R and αD_R, on the side facing the other domains in the RT, and a long C-terminal helix αE_R traversing the entire length of the RNase H domain on another side facing the solvent. The RNase H domain is swiveled around the connection and the thumb, and is positioned next to the fingers, and opposite to the connection/thumb domain (Figs. 2 and 6). While the general architecture is similar to other solved structures, the PFV PR-RT RNase H domain generally has longer helices and strands compared to RNase H domains of other RTs. N687 serves a hinge between helices αB_R and αC_R forming a continuous kinked helix, which is oriented across helices αA_R and αD_R. Absent in HIV-1, αC_R (also referred to as C-helix and followed by a basic protrusion) is a regular feature of *E. coli* RNase H, human RNase H, and MoMLV RNase H, critical for binding substrates and activity (Leo et al). The positioning of αB_R and αC_R is such that, while αB_R packs across helices αA_R and αD_R to stabilize them further, αC_R is positioned across the thumb and the connection, largely forming van der Waals contacts and thereby restricting the motion of the thumb in that direction (Fig. 9). A calcium ion from the crystallization buffer of the WT PFV PR-RT is observed chelated to the side chains of residues D599, E646 and D669, together with carbonyl of G600 which define the RNase H active site (Fig. 9). These residues, even though situated at the center of the domain, are solvent exposed and point away from the putative nucleic acid-binding cleft. This further suggests that this domain must undergo rotation to bring the active site into register with the nucleic acid substrate when it binds it.

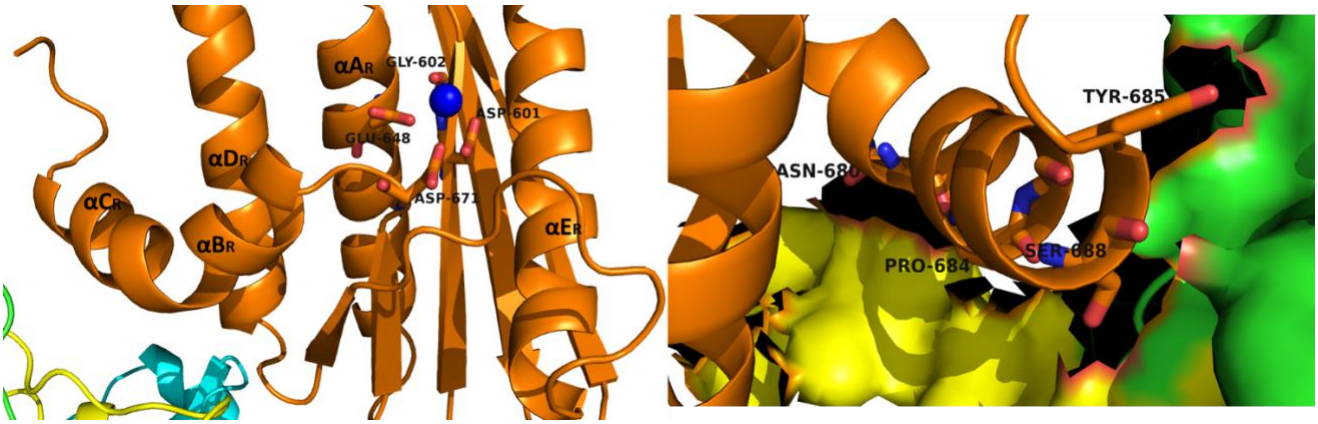
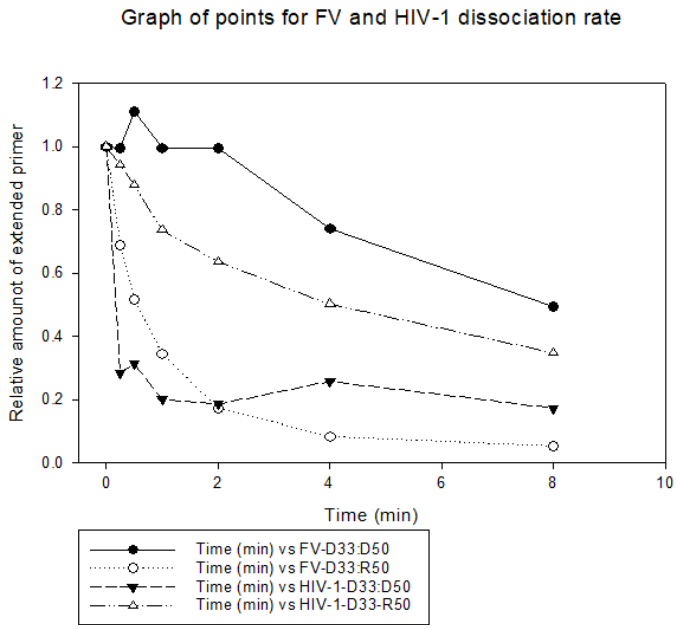


Figure 9. Left: Active site of PFV PR-RT RNase H subdomain residues chelating Ca^{2+} (blue sphere). Right: Surface rendering of the interface between PR-RT thumb/connection subdomains and potential interactions RNase H subdomain helix C.

PR-RT/DNA complexes

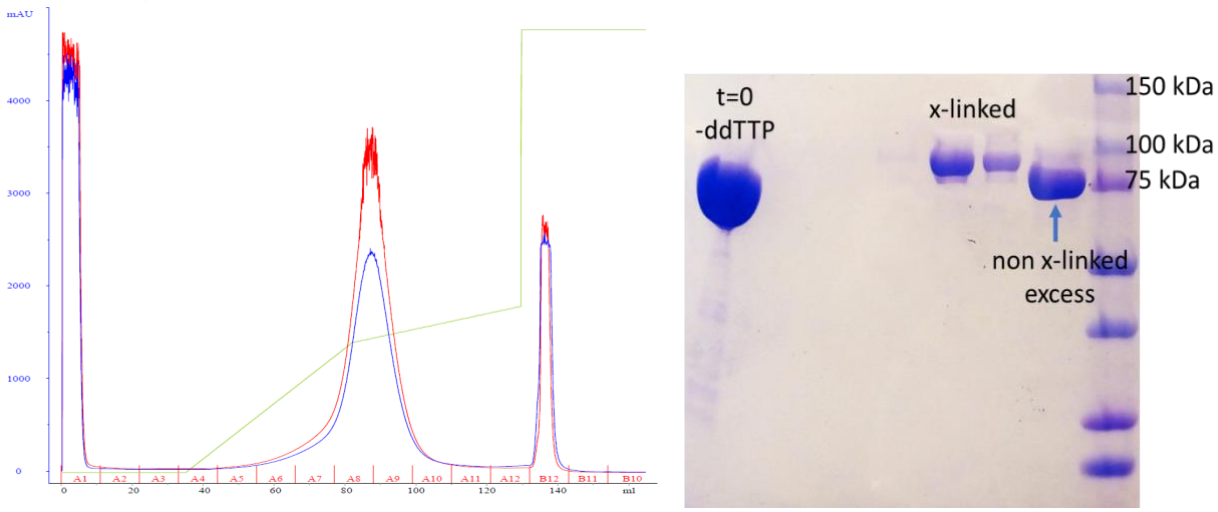
To gain insight into the nucleic acid binding kinetics and substrate specificities, the dissociation of PFV PR-RT from a template-primer RNA/DNA hybrid and a double-stranded (ds) DNA was measured and compared to that of HIV-1 RT. PFV PR-RT bound dsDNA and dissociated much more slowly compared to when bound to RNA/DNA. On the contrary, HIV-1 RT dissociated much slower from an RNA/DNA than from dsDNA (Supplemental Figure 3). This suggests that these two enzymes interact differently with these substrates. These differences may relate to the general topological differences between the RNA/DNA hybrid and the dsDNA and how these proteins engage them, or whether additional contacts come into play depending on which substrate is bound. These interactions may be relevant for understanding the differences in processivity between the enzymes.



Supplemental Figure 3. Dissociation rate of PFV PR-RT and HIV-1 RT from an RNA/DNA hybrid (D33-R50) and dsDNA (D33:D50).

To determine the similarities between the nucleic acid binding clefts of PFV PR-RT and that of HIV-1 RT, a site-specific DNA cross-linking experiment was carried out. Q258C substitution in HIV-1 RT enables the covalent trapping of nucleic acid substrates in the binding groove of the RT. The cross-linking chemistry designed for HIV-1 RT is such that covalent trapping of the dsDNA or RNA/DNA is only possible after a single round of nucleotide incorporation followed by translocation when a thioalkyl modified base G (N2) in the primer is positioned in register with the cysteine to enable disulfide bond formation (Boyer et al., 2004; Das et al., 2012). The covalently trapped complex is further purified using tandem nickel-heparin columns. PFV PR-RT Q391C, which is the equivalent of HIV-1 RT Q258C covalently traps this dsDNA designed for HIV-1 RT (Supplemental Fig. 4). Thus, even though PFV RT Q391 in the thumb is pointing away from the structure, upon rearrangement, the thumb is positioned in a configuration similar to the HIV-1 RT p66, and at a distance from the polymerase active site that is equivalent to that of p66.

This also suggests that the overall configuration of the catalytically active PFV RT is similar to that of HIV-1 RT p66.



Supplemental Figure 4. Heparin chromatography trace of PR-RT-dsDNA cross-linked complex with its accompanying SDS-PAGE gel.

Discussion

Retroviral genomes encode aspartyl PRs and RTs that are respectively responsible for the proteolytic processing of polyprotein precursors as well as reverse transcription of their RNA genome into a dsDNA, which can subsequently be inserted into the host chromosome by IN. In this work, we report the crystal structure of the PFV PR-RT fusion polyprotein is presented. The crystal structure of the PFV PR-RT shows monomeric PR and an RT. The PFV PR is folded similarly to a single subunit of the mature HIV-1 dimeric PR however, it exists in a monomeric state. The N- and C-terminal residues involved in the dimerization interface in the mature PR remain unstructured in this structure. *In silico* placement of two PFV PR-RT molecules related by two-fold symmetry bring two PR monomers together, similar to mature dimers, without significant

steric clashes between the RT domains. This offers a glimpse of how this entity might carry out proteolytic processing of peptide substrates.

The polymerase domain in the structure appears to be in an inactive configuration that would not bind a double-stranded nucleic acid substrate. The subdomains are arranged in a configuration that closely resemblances to the p51 subunit of HIV-1 RT. Hence a significant rearrangement of the subdomains is envisaged for PFV PR-RT to facilitate the binding of nucleic acid substrates. With the long linkers connecting the palm and the thumb as well as the connection and RNase H, such movements are fathomable (see movie). It must be emphasized that while for monomeric RTs the arrangement of these domains is unusual, it may not be unique to PFV. Based on the analysis of buried surface areas in the HIV-1 p66/p51 heterodimer, it was suggested that monomeric forms of these enzymes would adopt a more compact "p51-like" conformation to compensate for the cost of exposing hydrophobic residues (Wang et al., 1994). (Zheng et al., 2015), using NMR spectroscopy confirmed this initial observation, further asserting that the RNase H domain of monomeric HIV-1 p66 has a loose structure with flexible linkers connecting the thumb and RNase H domains to the rest of the structure, without any significant interaction between the thumb and the RNase H.

These observations from limited structural data are very consistent with the structure presented here. The interactions observed between the thumb and RNase H of PFV RT is through the C-helix which is lacking in HIV-1. Furthermore, monomeric HIV-1 p66 has been shown to sample predominantly two conformational states prior to homodimerization and subsequent cleavage of the RNase H domain of the p51 precursor. NMR studies (Zheng et al., 2015) suggest that the most predominant population in solution is the "p51-like" conformation. Transient sampling of the catalytically competent p66 open conformer might expose the dimerization interface leading to

homodimerization with a “p51-like” conformer prior to maturation. The asymmetric homodimer subsequently matures to the asymmetric heterodimer in mature HIV-1 RT. Structural data defining the position of the RNase H domain in the asymmetric homodimer prior to maturation has been elusive, since the structure of monomeric HIV-1 p66 is not available. Our PFV PR-RT structure reveals what could be an important isomerization intermediate of RTs that has eluded researchers for decades.

It is worth nothing that the energetic penalty of isomerization from the closed p51-like conformation with hydrophobic residues involved in extensive inter-subdomain interactions in this structure must be compensated for in the open p66-like conformation. In the absence of nucleic acid, this energy barrier is likely insurmountable, since most of these interactions are lost upon isomerization into the open catalytically competent state. It is envisaged therefore that this isomerization may only be largely triggered by nucleic acid substrates, where an extensive interface between the connection subdomain and the RNase H is expected to form, in addition to protein-nucleic acid interactions between the fingers, thumb, connection and RNase H.

Superposition of the current structure with HIV-1 RT/DNA complex (Fig. 10) suggests that the connection domain acts as the “gatekeeper”, by occupying the nucleic acid-binding cleft. For nucleic acid binding therefore, it is envisaged that this gatekeeper must undergo close to 90º rotation followed by a translation perpendicular to the helical axis of helix L of the connection. This screw movement would disrupt the interaction with the thumb, allowing it to move closer to the fingers/palm, while pushing out the RNase H domain to swivel to a position similar that observed in other RTs to form a conformation relevant for nucleic acid binding (see movie in SI). It is possible that the catalytically competent conformation is minimally sampled in solution which is trapped and stabilized by nucleic acid substrates. The sandwiching of the thumb by the RNase

H and the connection suggest that any kind of movement has to be well choreographed and that random rearrangement of the domains is unlikely.

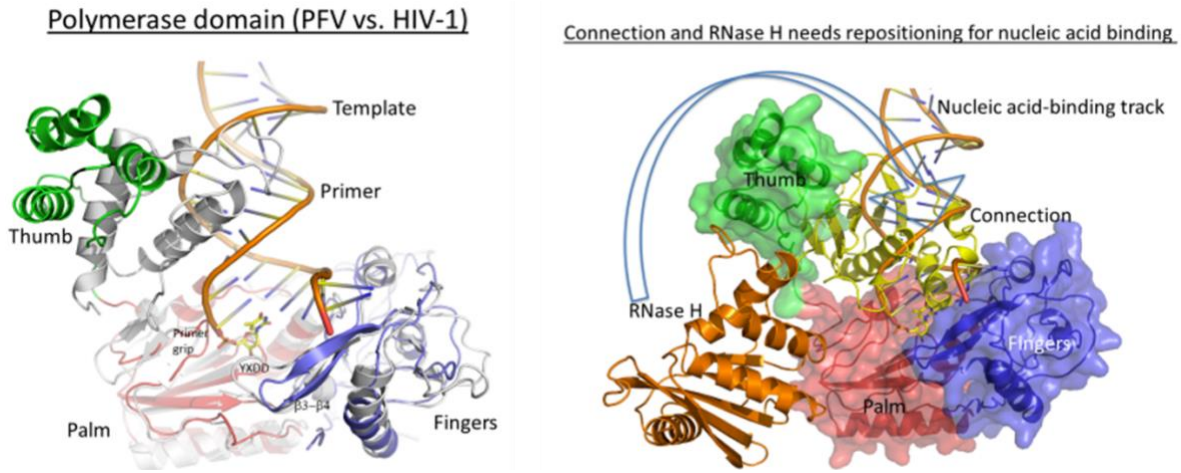


Figure 10. Left: Structure of PFV PR-RT (grey) superimposed onto HIV RT/DNA complex (PDB ID 1N6Q, subdomain coloring as in Fig. 2). Right: Surface rendering and ribbon model of the structure of PFV PR-RT (as in Fig. 6) superimposed on HIV RT/DNA complex (PDB ID 1N6Q). Arrow shows the direction of the RNase H domain movement required to engage the nucleic acid.

Since the PFV PR-RT is the machinery that carries out proteolytic processing as well as reverse transcription, a tighter control of each process is envisaged. Thus, the adoption of an inactive RT conformation with considerable energetic penalty upon rearrangement will predominantly help to decouple the two functions and might ensure that they can be regulated from a temporal and spatial perspective. Such a high energy barrier may be compensated for by the extensive interactions with nucleic acids. This structure does not only capture the imaginative evolutionary genius of the most primitive member of retroviruses but offers significant insight into the evolutionary landscape of these largely non-pathogenic viruses.

Conclusions

A systematic study of the PR-RT precursor polyprotein of PFV has been carried out. The PFV PR-RT is a functional PR and RT. The mutations in the enzyme which improved the yield of purified protein compared to the WT were identified using a Python script based on the instability indices

computation on the ExPASy ProtParam webserver. This demonstrates the feasibility of identifying unstable regions (and proteolytically vulnerable in the current case) in a protein by simply querying the primary sequence using the instability index as an output for evaluating how a particular amino acid is preferred in its position based on its neighbors. Each of the subdomains in the protein, as revealed by the crystal structure, is analogously folded but spatially arranged differently when compared to the mature enzymes from HIV and other retroviruses. The differences in the geometrical arrangement of the domains in the precursor polyproteins results in the generation of new interfaces that are not present in the mature enzymes. As previously evidenced by the surrogate use of PFV IN for deciphering the molecular basis of retroviral integration and strand transfer inhibitors resistance (G. E. Lee et al., 2019), it is anticipated that such structures from other retroviral genera, as in the case of PFV PR-RT, could present new avenues for targeting HIV using inhibitors that target its precursor polyproteins rather than the mature enzymes, and making the precursor polyprotein a legitimate target for drug discovery.

For the PFV PR-RT, the subdomains are arranged similarly to the p51 subdomain of HIV-1 RT. Such an inactive conformation of the monomeric RT had been predicted based on buried surface analysis and NMR spectroscopy. The compact fold likely ensures that the entropic cost associated with having an open nucleic acid-binding cleft in a monomeric RT is offset. It is expected that an extensive rearrangement of the domains would be required to enable nucleic acid binding. Cross-linking studies suggest that the PFV RT binds nucleic acid substrates in a mode similar to that of HIV-1 RT. This structure additionally offers insight into the pathway of conformational maturation of retroviral RTs. It is also a clever way of decoupling PR activity, which is needed early in the life cycle of the virus during maturation, from that of the polymerase to ensure they are well regulated.

Materials and Methods

Cloning, expression and purification

PFV PR-RT, with a protease null mutation (D24A), WT, was cloned into a pET28a vector with an N-terminal hexa-His-tag and an HRV14 3C protease cleavage site. Quick Change site-directed mutagenesis (Liu & Naismith, 2008) was used to introduce three other mutations: C280S, H507D and S584K to obtain the CSH mutant. These plasmids were transformed into BL21 DE3 RIL *E. coli* strain (Agilent). Expression of the proteins was carried out by inoculating 1 L of LB media containing 0.5% glycerol, 50 µg/mL of kanamycin, and 34 µg/mL of chloramphenicol with 50 mL of overnight LB cultures. Cells were allowed to grow at 37 ˚C until an OD of ~0.7 and transferred to 15 ˚C and grown until an OD of about 1. Media was supplemented with 20 mM $MgSO_4$ or $MgCl_2$ and expression induced with 1-2 mM IPTG and allowed to grow overnight at 15 ˚C overnight.

Cells were harvested and spun down at 5,000 x g for 25 minutes. Cell pellet was resuspended in 100 mL of lysis buffer [50 mM phosphate pH 7.6, 300 mM NaCl, 1 mM tris(2-carboxyethyl)phosphine (TCEP), 20 mM imidazole, 10% glycerol, 0.5 mM ethylenediaminetetraacetic acid (EDTA), 1 mM phenylmethylsulfonyl fluoride (PMSF)]. Cells were sonicated for 10 minutes and spun down at 18,000 x g for 30 minutes after which it was loaded onto a nickel gravity column pre-equilibrated with the lysis buffer. After loading, the column was washed with at least 10 column volumes of the lysis buffer followed by another 15 column volumes of a wash buffer (50 mM phosphate pH 7.6, 500 mM NaCl, 1 mM TCEP, 40 mM imidazole, 5% glycerol, 0.5 mM EDTA, 1 mM PMSF). The protein was eluted from the column using a phosphate buffer (50 mM phosphate pH 7.6, 300 mM NaCl, 300 mM imidazole, 5% glycerol, 0.5 mM EDTA). Eluted protein was either dialyzed overnight concurrently with the hexa-

His-tag cleavage with HRV14 3C protease at a ratio of 1:20 cleavage reaction in excess of Tris-Cl loading buffer (20 mM Tris, 100 mM NaCl, 0.25 mM EDTA, pH 8.0) or diluted two-fold and incubated with HRV14 3C protease overnight at 4 ºC.

The dialyzed protein was spun down and supplemented with 1 mM TCEP, loaded onto a heparin column pre-equilibrated with the Tris-Cl buffer as used for the dialysis but containing 1 mM TCEP. The column was subsequently washed with at least 15 column volumes of loading buffer followed by another 15 column volumes wash with the same buffer containing 300-500 mM NaCl. Elution of the protein from the heparin column was carried out in a single step using the Tris-Cl loading buffer containing 1.0 M NaCl and subsequently dialyzed against Tris-Cl storage buffer (20 mM Tris-Cl pH 7.4, 75 mM NaCl), concentrated to 25 mg/ml, snap frozen and stored at -80 ˚C. Purity of protein at each step was accessed by reducing SDS-PAGE gel.

Selenomethionine (SeMet) labeling

500 mL LB media was inoculated with a glycerol stock of CSH (BL21 DE3 RIL) and grown overnight in the presence of 100 mg of methionine, 50 µg/mL of kanamycin and 34 µg/mL of chloramphenicol. The cells were then pelleted at 4000g for 10 mins, washed with 20 mL minimal media and re-suspended in 1 L of minimal media supplemented with 1 mM $MgCl_2$. The re-suspended cells were grown at 15 ˚C for 30-45 mins in the presence of 50 mg each of isoleucine and leucine and 1000 mg each of selenomethionine, phenylalanine, lysine, and threonine before induction with 1 mM IPTG overnight. The cells were harvested and purified and stored the same way as unlabeled protein.

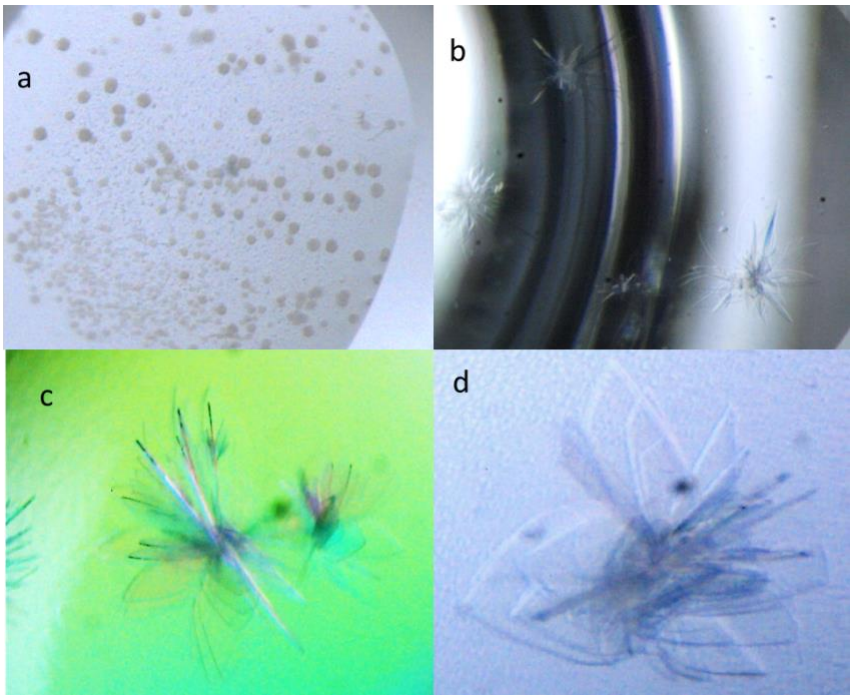
Crystallization

Initial crystallization screening was carried out using commercially purchased precipitants in a sitting-drop vapor diffusion and hanging-drop vapor diffusion experiments at a CSH protein construct at a concentration of 15-20 mg/mL at 4 ºC and 20 ºC with the help of a crystallization robot. Microcrystal hits were obtained within 48 hours at both temperatures in Hampton's Natrix HT E8 (0.05 M potassium chloride, 0.05 M sodium cacodylate trihydrate pH 6.0, 10% w/v polyethylene glycol 8,000, 0.0005 M spermine, 0.0005 M L-argininamide dihydrochloride) and F9 (0.012 M sodium chloride, 0.08 M potassium chloride, 0.04 M sodium cacodylate trihydrate pH 6.0, 50% v/v (+/-)-2-methyl-2,4-pentanediol, 0.012 M spermine tetrahydrochloride). Optimization was carried out by changing pH of precipitant, protein concentration, and variation of the concentrations of the individual components in a precipitant as well as the protein: precipitant ratio. None of these improved the crystal form significantly for structure determination (Supplemental Figs. 5A-B). The Hampton additive screen was carried out using the Natrix conditions E8 and F8. Crystals grown in 30 mM glycylglycylglycine (GGG)-C11, and 10 mM EDTA-E1 at 20 ºC were of slightly different morphology than the initial crystal hit. Further optimization was therefore carried out using glycylglycine and EDTA as additives by varying their concentrations. These conditions produced thin plates of less than 5 µm thickness in each of the individual cases in the presence of 200 mM glycylglycine or 100 mM EDTA which diffracted X-rays to about 4.5-5 Å resolution. The choice of glycylglycine (GG) was serendipitous as it was available in the lab at the time, whereas GGG was not. Crystals grown with GGG when it was finally obtained diffracted more poorly.

Even though the resolution of these datasets was sufficient in theory to solve the structure, the diffraction was very anisotropic with very high mosaicity, making it practically impossible to solve the structure using these datasets. A 1:1 mixture of the two conditions containing GG and EDTA

was carried out since the path to further optimization had been narrowed much further. Surprisingly, the crystals grown in these conditions were significantly thicker than those produced by their individual conditions. Thus, GG and EDTA together had a cumulatively positive effect on the crystal size even though they remained plates emanating from a single nucleation site (Supplemental Fig. 5).

The following conditions produced the best crystals suitable for diffraction studies: 50 mM KCl, 50 mM sodium cacodylate trihydrate pH 6.0, 12% PEG 8000, 1.0 mM spermine, 1.0 mM L-argininamide, 200 mM glycyglycine or glycylglycylglycine, and 50 mM EDTA in a sitting-drop vapor diffusion crystallization format between 10-25 °C with drop sizes ranging between 1-10 µL in micro bridges (Hampton). The EDTA could be substituted with 10 mM $MgCl_2$, 10 mM $MnCl_2$, or 100 mM $CaCl_2$ with similar results. The wild-type crystals as well as selenomethionine CSH crystals were grown with 100 mM $CaCl_2$ instead of EDTA and 2-5 mM TCEP supplemented with the protein. The CSH mutant of PFV routinely formed plate-like crystals which diffracted X-rays to at least 3.0 Å resolution enabling structural solution. Optimized crystallization conditions using this mutant were subsequently used to crystallize and solve the structure of the WT as well.



Supplemental Figure 5. (A) Initial crystallization hit in Natrix HT E8. (B) Optimization of initial additive screening with (C), and (D) fully optimized crystals used for structure determination

Crystals were harvested using MiTeGen MicroLoops or MicroMesh mounts and cryo-protected in reservoir supplemented with 25% glycerol before flash freezing.

Data collection

X-ray diffraction data were collected at the Cornell High Energy Synchrotron Source (CHESS) F1 beamline, the Advanced Photon Source (APS) 23-ID-D beamline, and the Stanford Synchrotron Radiation Light Source (SSRL) 9-2 beamline. Both PFV PR-RT WT and CSH crystallized with C2 space group symmetry. Data processing and scaling were done using iMosFlm and Aimless. The structure was solved by merging five SeMet-labeled data sets of the CSH mutant using Crank2, and the refinement was carried using REFMAC, all in the CCP4i suite (Potterton et al., 2018). Model building was done in Coot (Emsley & Cowtan, 2004). The model of the SeMet-labelled CSH mutant was successfully used for molecular replacement using the WT data which

was found to be twinned (Xtriage) in Phaser in PHENIX (Adams et al., 2011) and refined to 2.9 Å resolution using the twin law -h, -k, l to R-work/R-free of 22.4/25.8% in REFMAC (Murshudov et al., 1997) and phenix.refine (Adams et al., 2011; Murshudov et al., 1997). Publication figures were generated using PyMOL Molecular Graphics System, v. 1.7 (Schrödinger, LLC).

Extension and processivity assay

The polymerase and RNase H activity assays were carried out according to the protocol in Boyer et al., 2004 (Boyer et al., 2004; Rinke et al., 2002).

Protein-DNA cross-linking

The cross-linking experiments were conducted as reported in (Das et al., 2012; Das et al., 2019; Sarafianos et al., 2003). NaCl concentrations in the final reactions were however changed to 150-200 mM.

Acknowledgements

We are grateful to staff members for support and access to beamlines CHESS F1, APS 23-ID-D and SSRL 9-2 for data collection. Crystal structure atomic coordinates have been deposited in the Protein Data Bank under IDs 7KSE and 7KSF and will be released upon article publication. This work was supported by NIH MERIT Award R37 AI027690 and the HIVE Center Grant U54 GM103368 (both to E.A.). J.J.E.K Harrison received financial support through the United States Government Fulbright Scholarship program and the Institute of International Education for which he is grateful.

Competing interests

The authors declare that they have no conflicts of interest with the contents of this article.

References

Adams, P. D., Afonine, P. V., Bunkoczi, G., Chen, V. B., Echols, N., Headd, J. J., . . . Zwart, P. H. (2011). The Phenix software for automated determination of macromolecular structures. *Methods, 55*(1), 94-106. doi:10.1016/j.ymeth.2011.07.005

Baldwin, D. N., & Linial, M. L. (1998). The roles of Pol and Env in the assembly pathway of human foamy virus. *J Virol, 72*(5), 3658-3665. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9557646

Boyer, P. L., Stenbak, C. R., Clark, P. K., Linial, M. L., & Hughes, S. H. (2004). Characterization of the polymerase and RNase H activities of human foamy virus reverse transcriptase. *J Virol, 78*(12), 6112-6121. doi:10.1128/JVI.78.12.6112-6121.2004

Das, K., Martinez, S. E., Bauman, J. D., & Arnold, E. (2012). HIV-1 reverse transcriptase complex with DNA and nevirapine reveals non-nucleoside inhibition mechanism. *Nat Struct Mol Biol, 19*(2), 253-259. doi:10.1038/nsmb.2223

Das, K., Martinez, S. E., DeStefano, J. J., & Arnold, E. (2019). Structure of HIV-1 RT/dsRNA initiation complex prior to nucleotide incorporation. *Proc Natl Acad Sci U S A, 116*(15), 7308-7313. doi:10.1073/pnas.1814170116

Ding, J., Das, K., Hsiou, Y., Sarafianos, S. G., Clark, A. D., Jr., Jacobo-Molina, A., . . . Arnold, E. (1998). Structure and functional implications of the polymerase active site region in a complex of HIV-1 RT with a double-stranded DNA template-primer and an antibody Fab fragment at 2.8 A resolution. *J Mol Biol, 284*(4), 1095-1111. doi:10.1006/jmbi.1998.2208

Emsley, P., & Cowtan, K. (2004). Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr, 60*(Pt 12 Pt 1), 2126-2132. doi:10.1107/S0907444904019158

Erlwein, O., & McClure, M. O. (2010). Progress and prospects: foamy virus vectors enter a new age. *Gene Ther, 17*(12), 1423-1429. doi:10.1038/gt.2010.95

Guruprasad, K., Reddy, B. V. B., & Pandit, M. W. (1990). Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Engineering, Design and Selection, 4*(2), 155-161. doi:10.1093/protein/4.2.155

Hartl, M. J., Schweimer, K., Reger, M. H., Schwarzinger, S., Bodem, J., Rosch, P., & Wohrl, B. M. (2010). Formation of transient dimers by a retroviral protease. *Biochem J, 427*(2), 197-203. doi:10.1042/BJ20091451

Hutter, S., Mullers, E., Stanke, N., Reh, J., & Lindemann, D. (2013). Prototype foamy virus protease activity is essential for intraparticle reverse transcription initiation but not absolutely required for uncoating upon host cell entry. *J Virol, 87*(6), 3163-3176. doi:10.1128/JVI.02323-12

Ishima, R., Torchia, D. A., Lynch, S. M., Gronenborn, A. M., & Louis, J. M. (2003). Solution structure of the mature HIV-1 protease monomer: insight into the tertiary fold and stability of a precursor. *J Biol Chem, 278*(44), 43311-43319. doi:10.1074/jbc.M307549200

Katoh, I., Ikawa, Y., & Yoshinaka, Y. (1989). Retrovirus protease characterized as a dimeric aspartic proteinase. *J Virol, 63*(5), 2226-2232. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/2539514

Lee, E. G., Stenbak, C. R., & Linial, M. L. (2013). Foamy virus assembly with emphasis on Pol encapsidation. *Viruses, 5*(3), 886-900. doi:10.3390/v5030886

Lee, G. E., Mauro, E., Parissi, V., Shin, C. G., & Lesbats, P. (2019). Structural insights on retroviral DNA integration: learning from foamy viruses. *Viruses, 11*(9). doi:10.3390/v11090770

Lindemann, D., & Rethwilm, A. (2011). Foamy virus biology and its application for vector development. *Viruses, 3*(5), 561-585. doi:10.3390/v3050561

Linial, M. L. (1999). Foamy viruses are unconventional retroviruses. *J Virol, 73*(3), 1747-1755. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9971751

Liu, H., & Naismith, J. H. (2008). An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnol, 8*, 91. doi:10.1186/1472-6750-8-91

Lochelt, M., & Flugel, R. M. (1996). The human foamy virus *pol* gene is expressed as a Pro-Pol polyprotein and not as a Gag-Pol fusion protein. *J Virol, 70*(2), 1033-1040. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8551561

Murshudov, G. N., Vagin, A. A., & Dodson, E. J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr D Biol Crystallogr, 53*(Pt 3), 240-255. doi:10.1107/S0907444996012255

Nowak, E., Miller, J. T., Bona, M. K., Studnicka, J., Szczepanowski, R. H., Jurkowski, J., . . . Nowotny, M. (2014). Ty3 reverse transcriptase complexed with an RNA-DNA hybrid shows structural and functional asymmetry. *Nat Struct Mol Biol, 21*(4), 389-396. doi:10.1038/nsmb.2785

Pfrepper, K. I., Rackwitz, H. R., Schnolzer, M., Heid, H., Lochelt, M., & Flugel, R. M. (1998). Molecular characterization of proteolytic processing of the Pol proteins of human foamy virus reveals novel features of the viral protease. *J Virol, 72*(9), 7648-7652. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9696869

Potterton, L., Agirre, J., Ballard, C., Cowtan, K., Dodson, E., Evans, P. R., . . . Wojdyr, M. (2018). CCP4i2: the new graphical user interface to the CCP4 program suite. *Acta Crystallogr D Struct Biol, 74*(Pt 2), 68-84. doi:10.1107/S2059798317016035

Rethwilm, A., & Bodem, J. (2013). Evolution of foamy viruses: the most ancient of all retroviruses. *Viruses, 5*(10), 2349-2374. doi:10.3390/v5102349

Rinke, C. S., Boyer, P. L., Sullivan, M. D., Hughes, S. H., & Linial, M. L. (2002). Mutation of the catalytic domain of the foamy virus reverse transcriptase leads to loss of processivity and infectivity. *J Virol, 76*(15), 7560-7570. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12097569

Roy, J., & Linial, M. L. (2007). Role of the foamy virus Pol cleavage site in viral replication. *J Virol, 81*(10), 4956-4962. doi:10.1128/JVI.00104-07

Sarafianos, S. G., Clark, A. D., Jr., Tuske, S., Squire, C. J., Das, K., Sheng, D., . . . Arnold, E. (2003). Trapping HIV-1 reverse transcriptase before and after translocation on DNA. *J Biol Chem, 278*(18), 16280-16288. doi:10.1074/jbc.M212911200

Sheik Amamuddy, O., Bishop, N. T., & Tastan Bishop, O. (2018). Characterizing early drug resistance-related events using geometric ensembles from HIV protease dynamics. *Sci Rep, 8*(1), 17938. doi:10.1038/s41598-018-36041-8

Shin, G., Yost, S. A., Miller, M. T., Elrod, E. J., Grakoui, A., & Marcotrigiano, J. (2012). Structural and functional insights into alphavirus polyprotein processing and pathogenesis. *Proc Natl Acad Sci U S A, 109*(41), 16534-16539. doi:10.1073/pnas.1210418109

Tozser, J. (2010). Comparative studies on retroviral proteases: substrate specificity. *Viruses, 2*(1), 147-165. doi:10.3390/v2010147

Trobridge, G. D. (2009). Foamy virus vectors for gene transfer. *Expert Opin Biol Ther, 9*(11), 1427-1436. doi:10.1517/14712590903246388

Wang, J., Smerdon, S. J., Jager, J., Kohlstaedt, L. A., Rice, P. A., Friedman, J. M., & Steitz, T. A. (1994). Structural basis of asymmetry in the human immunodeficiency virus type 1 reverse transcriptase heterodimer. *Proc Natl Acad Sci U S A, 91*(15), 7242-7246. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7518928

Wohrl, B. M. (2019). Structural and functional aspects of foamy virus protease-reverse transcriptase. *Viruses, 11*(7). doi:10.3390/v11070598

Yu, S. F., Baldwin, D. N., Gwynn, S. R., Yendapalli, S., & Linial, M. L. (1996). Human foamy virus replication: a pathway distinct from that of retroviruses and hepadnaviruses. *Science, 271*(5255), 1579-1582. doi:10.1126/science.271.5255.1579

Yu, S. F., Sullivan, M. D., & Linial, M. L. (1999). Evidence that the human foamy virus genome is DNA. *J Virol, 73*(2), 1565-1572. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9882362

Zheng, X., Pedersen, L. C., Gabel, S. A., Mueller, G. A., Cuneo, M. J., DeRose, E. F., . . . London, R. E. (2014). Selective unfolding of one ribonuclease H domain of HIV reverse transcriptase is linked to homodimer formation. *Nucleic Acids Res, 42*(8), 5361-5377. doi:10.1093/nar/gku143

Zheng, X., Perera, L., Mueller, G. A., DeRose, E. F., & London, R. E. (2015). Asymmetric conformational maturation of HIV-1 reverse transcriptase. *eLife, 4*. doi:10.7554/eLife.06359