STRUCTURAL AND BIOPHYSICAL CHARACTERIZATION OF RETROVIRAL POLYPROTEINS: INSIGHTS FROM PROTOTYPE FOAMY VIRUS (PFV) AND HUMAN IMMUNODEFICIENCY VIRUS TYPE 1 (HIV-1)

by

JERRY JOE EBOW KINGSLEY HARRISON

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Medicinal Chemistry

Written under the direction of

Eddy Arnold

and approved by

____

____

____

____

New Brunswick, New Jersey

October 2019

ABSTRACT OF THE DISSERTATION

Structural and Biophysical Characterization of Retroviral Polyproteins-Insights from the Prototype Foamy Virus (PFV) and Human Immunodeficiency Virus Type 1 (HIV-1)

by: JERRY JOE EBOW KINGSLEY HARRISON

Dissertation Director: Professor Eddy Arnold

Retroviruses continue to attract public health attention due to the increasing disease burden these viruses continue to have in animals. An estimated 38 million people worldwide are infected with HIV which underscores the importance of these viruses to humans. In the absence of a cure or vaccine for HIV, therapeutic interventions that have culminated in the taming of the disease, which used to be a death sentence before the discovery of these drugs, has relied on detailed understanding of key processes which underline the replication of the virus in its hosts. It is not surprising that the great majority of the currently approved drugs for treating HIV infections target the various enzymatic proteins which carry out chemical reactions that enable the virus to infect new cells.

Increasing resistance to these drugs, and problems of compliance and toxicity, have necessitated the need to search for more potent and less toxic drugs to manage the disease. It is in this regard that understanding the structural and mechanistic details of polyprotein processing is important. Synthesis of polyprotein precursors which are subsequently processed into mature enzymes is unique to pathogenic viruses such as retroviruses and RNA viruses. Developing drugs selective to these polyproteins would reduce the potential for cross-reactivity with host proteins.

The crystal structure of the PFV protease-reverse transcriptase (PR-RT), solved to 2.9 Å resolution, has revealed a monomeric protein containing individually folded subdomains. In PFV, PR-RT is the mature entity following the proteolytic release of IN from the C-terminus of the PFV Pol polyprotein. The structure of the PR resembles a monomer of the homodimeric HIV-1 PR even though the N-terminal and C-terminal residues involved in the formation of anti-parallel β-sheets which mediate the dimer interface, remain unstructured. The RT also contains the canonical subdomains: fingers, palm, thumb and connection, as well as the RNase H domain, which characterize retroviral RTs. The relative orientations of these subdomains are however different, resembling more the compact p51 subunit of HIV RT. This is an inactive RT conformation since the nucleic acid binding cleft is occluded. It however offers insights into the possible subdomain arrangements of the monomeric precursor polyproteins.

In this work also, the Pol polyproteins from the prototype foamy virus (PFV), as well as Pol and Gag-Pol polyproteins from HIV-1 have been characterized using a combination of enzymatic assays and other biophysical methods including: gel filtration, dynamic light scattering (DLS), small angle X-ray scattering (SAXS), and single particle cryo-EM. The production of these proteins, which were isolated in yields and purity suitable for biophysical studies from bacteria for the first time, relied on a new media formulation for *E. coli* growth. Fundamentally, it was discovered that at non-physiologic Mg^{2+} concentration (~50 mM) or low pH (≤ 6.0), the proteolytic degradation of these polyproteins and by extension, heterologously expressed proteins in bacteria, are highly attenuated. This decreased proteolytic processing enabled the expression and purification of these polyprotein precursors.

Single particle cryo-EM analysis of the HIV-1 Pol polyprotein revealed a dimeric RT similar to the mature heterodimer in configuration. The formation of the heterodimer at a very early stage of the maturation process offers a plausible explanation as to why only one RNase H domain of RT is cleaved in the virus during maturation, since in this configuration, the RNase H cleavage site of the p66 subunit is sequestered. The dimeric RT brings the two N-terminal PR monomer close to each other, which enables them to dimerize and cleave their substrates.

Advanced three-dimensional classification analysis of the HIV-1 Pol structures further revealed that the dimerization tendency of the PR may be reduced compared to the mature enzyme with various classes showing density for only one PR at the N-terminus of either the p51-like subunit or the p66-like subunit. This structural and configurational heterogeity offers a plausible explanation to why the immature PR is less sensitive to drugs that target the active site of the mature PR. The IN domain in the structure of the Pol is disordered and not visible. However, it binds the integrase-binding domain (IBD) of lens epithelium-derived growth factor (LEDGF/ p75) fused to the C-terminus of the maltose-binding protein (MBP), and the MBP-IBD fusion pulls down the full-length HIV-1 Pol dimer even at high salt concentrations (1.0 M NaCl). This suggests that the IN is also in a dimeric organization since IBD binds dimeric IN and not monomers. By assembling the dimeric units of these enzymes, which are the functional units or the building blocks for the functional entity, in the case of IN, very early in the maturation process, the virus embarks on an irreversible voyage committed to infect new cells by ensuring that these enzymes required later in the life cycle, are available in active forms when needed. The structures of the PFV Pol and HIV-1 Gag-Pol will provide a more comprehensive understanding of the structural underpinnings of PR activation and how maturation is orchestrated in the retroviral family.

Acknowledgments

I am grateful to God for the Grace and wisdom granted me that made the completion of this work possible. Undoubtedly, this work would not have been possible without the abundant gift of Grace from God, and for that reason, I praise his name. I would like to extend a special gratitude to my advisor, mentor and friend, Prof. Eddy Arnold. The faith you had in me against all the odds was the reason I am here today. Words will not be enough to convey my gratitude to you for your friendship, mentorship and support every step of the way. God bless you in a very unique way and may your stream of inspiration and grace be replenished forever. I would also like to thank my committee members: Prof. Longqin Hu, Prof. Edmond LaVoie, and Prof. Wilma Olson. I am particularly very grateful for your patience, generosity and kindness. Your words of encouragement resonate in my heart always and that made this journey much easier than it would have been.

I would like to also thank my numerous collaborators within and outside the HIVE Center especially; Prof. Jeffery DeStefano and Dr. Irani Ferreira of the University of Maryland, College Park, Dr. Stephen Hughes and Dr. Paul Boyer at NCI Frederick, Dr. Dmitry Lyumkis at Salk and his lab members especially Dr. Jessica Bruhn and Dr. Dario Passos, for the numerous ways each of you have contributed to the success of this work. A ship cannot set sail alone and so in that regard, I would like to also thank all the members of the Arnold lab at Rutgers, past and present; Lynda, Joe, Xavi, Anthony, Ashima, Shawn, Natalie, Chelsy, Sanjay, Matt, Kalyan, Sergio and Angela, for your friendship and warmth which made working in the lab very comfortable. Working with you is so much fun and so each one of you deserves a special praise for everything you have done for me. I would like to mention Dr. Steve Tuske, my former mentor in the Arnold lab, his wife Jenny and their family for their kind friendship to me. Indeed, without you, my Christmas would never have been the same. Thank you for all the wonderful memories. We will surely keep doing it.

I am specially indebted to my former mentors; Emeritus Prof. Addae-Mensah and Prof. Kingsford-Adaboh at the University of Ghana. You gave me so much and supported me even in the most difficult moments. Your empathy, words of wisdom and counsel, have been key to this success story. Thank you both for everything. I would also like to extend my gratitude to my colleague and friend, Dr. Enock Dankyi and his family. There is nothing more a friend can ask of a friend that you haven't already done more. I am grateful for all the errands, but above all, I am grateful for your surreal friendship. I am also thankful to the University of Ghana for the generous offer of study leave as well as all my colleagues in the Dept of Chemistry, UG, for the support and well wishes, especially, Dr. Osei-Safo.

I would also like to extend my gratitude to very special friends; Gloria, Safoa, Richard, Albert, Vitus, Rojay, Jan, Eli, Ali, Parishad, Kartheik, Tal and Adeshina, for the various ways each one of you have contributed to making this journey a lot easier than it would have been. I am thankful for all the encouragement, all the chats, games and the fun arguments. These memories would forever be with me. Gloria, thank you so much for always being there to share words of encouragement and hope. I am truly grateful.

My journey to the United States, would not have been possible without the generous financial and emotional support from the Fulbright program and its affiliated agencies. To all the workers of the public affairs directorate at the US embassy in Accra, Ghana, and all the people working for and on behalf of the Fulbright organization, I say thank you. You made this possible and you deserve special accolades. To all the families that received me and other Fulbright scholars to your homes across the United States, and treated us like kings and queens, I extend a special thank you. My experiences have been nothing short of amazing. Your hospitality is truly great, and I am eternally grateful to you all.

I would also like to thank the Rutgers University cryo-EM facility manager, Dr. Jason Kaelber and staff scientist, Dr. Emre Firlar, for all the help in optimizing these very challenging samples for data collection. I am thankful for all the knowledge you generously shared with me. CABM faculty and staff and the medicinal chemistry faculty and staff have been very receptive to me since I sojourned here. Thank you for creating an environment that fosters learning and thank you for all the help you rendered to me.

To the family to whom I belong, your love, support and prayers has been my sustenance. Without you none of this would have happened. Thank you for inspiring me every day and thank you for your genuine show of affection. You are the best gift any one can ask for, and I am grateful to God that I can call you family. May the Lord continue to bless you for all that you continue to do for me. Shalom!

Dedication

This work is dedicated to my parents,

Very Rev. S.E.K Harrison and Mrs. Hannah Harrison,

and my siblings, George, Georgina, Mercy, Samuel and Ebenezer,

for all the sacrifices you made that enabled me to learn how to read and write

Table of Contents

List of tables

List of figures

CHAPTER ONE: Transposable Elements, Retroviruses, and Disease

Synopsis

Transposable elements, which make up the vast majority of the human genome, have had a tremendous impact on the evolution of eukaryotes. While these elements continue to shape the eukaryotic genome, their evolutionary offsprings, the retroviruses such as HIV continue to wreak havoc on the human population. With about 38 million infected individuals currently living with the virus, HIV is currently the greatest public health threat to humans. Understanding of key events in the life cycle of the virus has been pivotal in the discovery of antiretroviral drugs that combat the disease and keep viral loads at manageable levels. In the absence of a cure, effective management of the disease will require further understanding of key processes currently poorly understood, such as the polyprotein processing, which will offer new avenues for designing even more potent drugs that may have higher genetic barrier to resistance.

1.1 Transposable elements, retroviruses, and the human genome

A new epoch of perhaps unprecedented reality dawned on humanity and changed its course forever upon the publication of the initial results of the human genome project (Lander et al., 2001). That 50-75% of the human genome is derived from transposable elements (mobile genes that can move from one gene loci to another) was unexpected, but that observation placed in perspective the 1940s discovery of transposable elements by Barbara McClintock and their immense contribution to the diversity and continued evolution of the eukaryotic genome (Cordaux and Batzer, 2009, Fedoroff, 2012). The estimates of the abundance of TEs in plant genomes are even higher with approximate numbers positing between 75 and 85 per cent (SanMiguel et al., 1998). These observations provide incontrovertible evidence to the ancestral heritage of retroviruses, which are ubiquitous animal pathogens responsible for numerous diseases and are hence of immense public health interest.

Classification of these mobile genetic elements is based on their replication paradigms, with two major classes recognized. The first group, DNA transposons, 'jump' by excising themselves from one genomic locus and ligating into another gene locus without any intermediate (cut and paste mechanism). While still very active in many lower organisms including bacteria and protists, they have become defunct and lost their replicative ability in humans even though they continue to shape the human genome through recombination (Cordaux and Batzer, 2009, Bradic et al., 2014). Retrotransposons/retroelements which on the other hand multiply their copy number in the genome through an RNA intermediate before ligating into other genomic loci, using a "copy and paste" mechanism, constitute the second group (Volkman and Stetson, 2014). The retrotransposon mRNAs are transcribed first by various RNA polymerases (Pol II and Pol III predominantly), translated on ribosomes, and then subsequently reverse transcribed by the transposon-encoded reverse transcriptase before being ligated into other genomic loci (Cordaux and Batzer, 2009).

Retrotransposons themselves can be further sub-classified into two groups distinguished by whether they have long-terminal repeat (LTR) sequences flanking their genome. Human LTR retrotransposons, also referred to as human endogenous retroviruses (HERVs), reminiscent of retroviruses, are indeed relics of once very active endogenous retroviruses in the human genome which seem to have lost their ability to propagate through integration of a copy of their genome into the germline, as well as to infect other cells. Hence while they remain transcriptionally active, they are not able to expand through infection of new cells; however, they continue to impact the human and other mammalian genomes through recombination and chromothripsis (Volkman and Stetson, 2014). About 8% of the human genome is constituted by these elements alone (**Fig. 1**), and are believed to have been lurking in the human genome for more than 25 million years (Cordaux and Batzer, 2009).

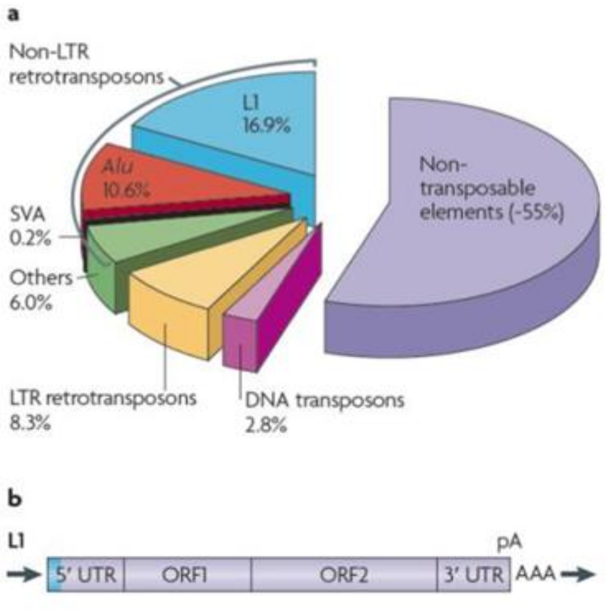


Figure 1:Distribution of TEs in the human genome.
Reproduced with permission from Cordaux and Batzer 2009.

The majority of the TEs in the human genome (at least 70%) lack long-terminal repeat sequences flanking their "proviral" genome and hence belong to the non-LTR retrotransposons family. Different non-LTR retrotransposons are further distinguished based on the length of their genome into; long interspersed nuclear elements (LINEs) or short interspersed nuclear elements (SINEs). With a genome size of ~6 kb, the long interspersed nuclear element 1 (LINE-1, or L1) remains the only retrotransposon family that is still actively replicating in the human genome, constituting ~ 17% of the human genome by mass (**Fig. 1**) (Cordaux and Batzer, 2009).

The ~6 kb genome of L1 elements contains two open reading frames ORF1 and ORF2 flanked by a 5'-untranslated region (UTR) and a 3'-UTR, harboring a polyadenylation signal (pA) and a poly-A tail. The 5'-UTR harbors an internal RNA polymerase II (Pol II) promoter site from which the genome is transcribed (Babushok and Kazazian, 2007). While ORF1 encodes an RNA-binding protein, ORF2 encodes a protein which contains both reverse transcriptase and endonuclease activities reminiscent of modern-day retroviral reverse transcriptases (RTs). The retrotransposition process known as target-primed reverse transcription (TPRT) (see **Fig. 2**), is therefore made possible only because of this molecular machinery. By encoding their own transposition machinery, the L1 elements are the only autonomous TEs/retroelements present and still active in the human genome, albeit not all L1 copies in the genome are competent for retrotransposition (Babushok and Kazazian, 2007, Cordaux and Batzer, 2009). These two proteins associate with LINE1 RNA to form a ribonucleoprotein particle which is then transported into the nucleus.

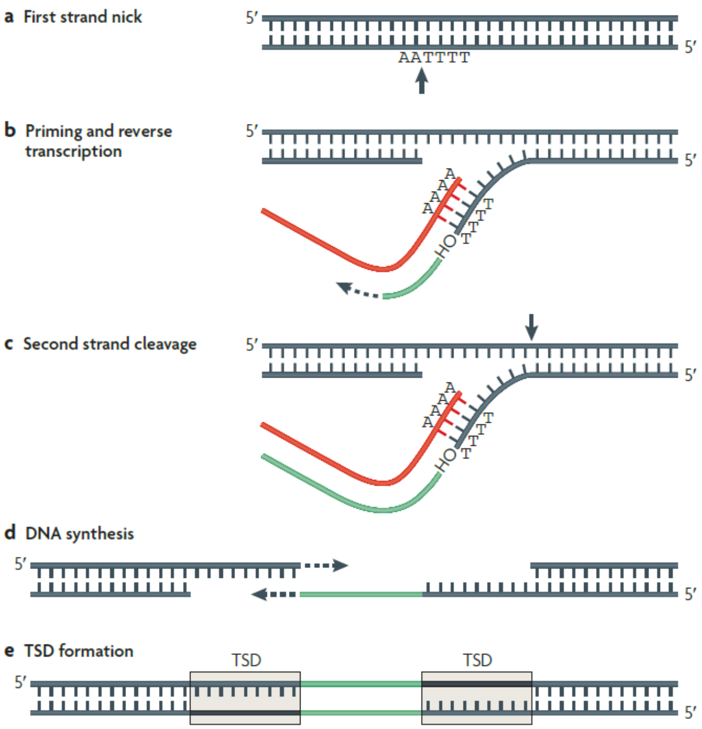


Figure 2: Schematic representation of the mechanism of transposition. Reproduced with permission from Cordaux and Batzer 2009.

1.2 Genome organization of HIV and PFV and their replication paradigms

The organization of the genomes of retroviruses closely mimics their evolutionary ancestors, the LTR retrotransposons. The ~ 10 kb + sense RNA genome of HIV for example, containing nine open reading frames that encode sixteen unique proteins (**Fig. 3**), is flanked at the 5'- and 3'-ends by long-terminal repeat (LTR) sequences which contain promoters that drive gene transcription as well as sequence elements that play regulatory roles in the life cycle of the virus (Ganser-Pornillos et al., 2008, Sundquist and Krausslich, 2012).

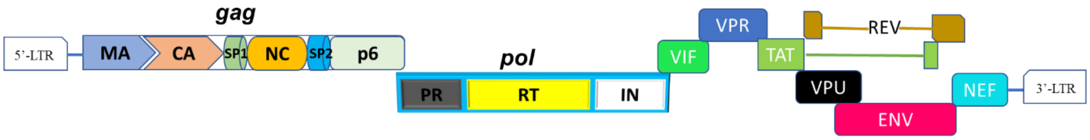


Figure 3: Schematic representation of the HIV genome.

The overlapping nature of the reading frames helps these viruses to encode the absolutely essential proteins they need for replicating in a host while economizing their genomes and, by extension their sizes, which is essential for them to infect cells. Another hallmark of these viruses, which is also common with other RNA viruses, is the synthesis of long polyprotein chains that are subsequently processed into mature enzymes (Sundquist and Krausslich, 2012, Shin et al., 2012). The Gag proteins, which constitute the structural proteins matrix (MA), capsid (CA), spacer peptide 1 (SP1), nucleocapsid (NC), spacer peptide 2 (SP2) and p6, are synthesized as a single polypeptide chain before proteolytic processing during maturation. The Pol proteins, protease (PR), reverse transcriptase (RT) and integrase (IN) which carry out all the enzymatic reactions of the virus are produced initially as part of a Gag-Pol polypeptide (Adamson and Freed, 2007, Freed, 2015).

The Gag-Pol polypeptide is produced when a -1 ribosomal frameshift at the 3'-end of the *gag* gene puts the *pol* gene in frame with *gag*. This unusual event occurs at a rate of about 5% leading to a 20:1 ratio of Gag: Gag-Pol in assembled virions of HIV (Freed, 2015). The envelope glycoproteins gp120 and gp41, which control the tropism of HIV virions and promote membrane fusion with new cells, are also initially expressed as a 160 kDa precursor before host cell endoproteases cleave it into two (Turner and Summers, 1999, Freed, 2015). Other accessory proteins which aid in the transcription of the viral genome, help evade immune response, induce cell cycle arrest, antagonize host factors that restrict HIV replication, help in the transport of spliced and unspliced RNA genome into the cytoplasm etc., are encoded by the genome as well (Nekhai and Jeang, 2006, Karn and Stoltzfus, 2012).

Even though the genome structure of PFV is similar to that of HIV, it differs significantly in how the polyprotein precursors are made and the types of accessory proteins it encodes. In PFV, separate mRNAs are used to translate the Gag and the Pol precursor polyproteins. The PFV genome encodes Bet and Tas accessory proteins, which carry out regulatory and immunomodulation functions (Bodem et al., 1997, Rethwilm and Bodem, 2013). Because only a few accessory proteins are encoded by PFV, the virus relies heavily on host machinery in carrying out transport of spliced and unspliced RNA into the cytoplasm by utilizing sequence elements encoded within the RNA to direct these processes (Berka et al., 2013, Hutter et al., 2013).

The different mechanisms of Gag and Pol synthesis by PFV compared to HIV raises questions about viral assembly mechanisms and egress. In HIV, particle assembly occurs at the plasma membrane mediated by the myristoylated Gag. Gag-Pol packaging into assembling virions is also aided by Gag molecules (Freed, 2015). Sequence elements at the C-terminus of Gag contribute to the recruitment of host factors that aid in egress of the new virions from the producer cells (Lingappa et al., 2014). On the other hand, membrane-targeting domains in HIV Gag are not present in PFV Gag. As such, assembly of virus particles occur predominantly at the centrosome or microtubule organizing center in the cytoplasm while subsequent budding of assembled virions occurs predominantly in the ER and Golgi (Mannigel et al., 2007, Hutter et al., 2013,). Sequence elements in Gag are critical for this process.

1.3 HIV life cycle

The life cycle of the HIV, an archetypical retrovirus, can be divided into early and late phases. The early phase begins with the engagement of a new virus with the surface of a host cell until a copy of the proviral genome is integrated into the host chromosome. The events that constitute the early phase can be subdivided into cell entry, reverse transcription, partial uncoating of the viral capsid, nuclear import of capsid and DNA integration (Turner and Summers, 1999, Gelinas et al., 2018).

HIV virions selectively infect cells that express CD4 glycoproteins on their cell surface such as T-helper cells, monocytes, macrophages and dendritic cells. The surface glycoprotein of HIV, gp120, recognizes and binds CD4 bearing cells, however this binding is not sufficient for fusion. The fusion of the HIV viral envelope with the host cell membrane is triggered by secondary binding to the chemokine receptors CCR5 and CXCR4, and is carried out by the HIV gp41 transmembrane glycoprotein (Turner and Summers, 1999).

Upon fusion, the capsid core containing the genetic material being reverse transcribed into DNA is docked into the cytosol of the new cell. Partial uncoating of the capsid occurs in the cytoplasm, but current knowledge suggests that the almost intact capsid containing the pre-integration complex is trafficked into the nucleus where the viral integrase bound to the genomic dsDNA catalyzes irreversible insertion of the genetic material of the virus into the host chromosome (Campbell and Hope, 2015, Marquez et al., 2018).

Transcription of the provirus, export of proviral spliced and unspliced mRNA into the cytoplasm for translation, trafficking of Gag and assembly, egress and maturation of new infectious virions constitute the late phase of the virus life cycle (Turner and Summers, 1999, Engelman and Cherepanov, 2012). Following the transcription of the HIV genome predominantly by the human RNA Pol II, unspliced mRNAs are trafficked into the cytoplasm with the help of the viral protein Rev where it is translated on ribosomes.

Post-translationally myristoylated Gag molecules are trafficked to the cell membrane where assembly of new virions occurs. The localization of Gag into the plasma membrane is also aided by the strong interaction of the matrix domain with phosphatidylinositol-(4, 5)-bisphosphate (PI(4, 5)P_2) whose concentration is very high in the inner leaflet of the plasma membrane. Gag binding to the plasma membrane has been found to induce the enrichment of cholesterol and sphingolipids in the membrane forming detergent-resistant lipid rafts which perhaps are more resistant to potential collapse as more Gag molecules are added to the assembling virion (Freed, 2015, Yandrapalli et al., 2016).

Two copies of the unspliced mRNA genome of HIV are packaged into an assembling virion in addition to several host factors necessary for replication (Nikolaitchik et al., 2013, Comas-Garcia et al., 2016). Following assembly, virions bud off the surface of the cell with the help of the host endosomal sorting complexes required for transport (ESCRT) machinery. During or immediately after budding, dimerization of the Gag-Pol takes place, which activates the embedded protease. Activation of the protease leads to the processing of the polyprotein precursors used in the viral assembly in a process called maturation, beginning with the embedded PR excising the Pol from the Gag-Pol and subsequently itself from the Pol (Pettit et al., 2005a, Pettit et al., 2005b, Engelman and Cherepanov, 2012). Maturation triggers an unprecedented structural rearrangement of the viral capsid housing the genetic elements from a spherical structure to a conical structure which is concomitant with infectivity of the progeny virus (Zhang et al., 2015, Wagner et al., 2016, Mattei et al., 2016, Dick et al., 2018,).

1.4 The ART of HIV therapeutic interventions

As of 2017, about 38 million people worldwide were estimated to be infected with HIV, the majority of whom live in sub-Saharan Africa. Out of this number, about 22 million people (59%) had access to antiretroviral therapy (ART) (hiv.gov, June 2019). The discovery of these drugs, the majority of which target and block the three main enzymatic reactions—processing, reverse transcription and integration—as well as the fusion of virions with new cells, have been key to the success story of keeping the disease at bay without the potential escalation of infections and its associated mortalities. Currently approved drugs for treating HIV and the timeline for their approval are shown in Figure **5**.

Out of the current 32 individual drugs (**Fig. 5**) approved for treating HIV infections, nine protease inhibitors (PIs) target the PR, which is responsible for the proteolytic cleavage of the polyprotein precursors into mature entities. These drugs are competitive active site inhibitors of PR which bind very tightly to the active site of the enzyme and prevent it from binding to the substrates that need to be cleaved.

Sixteen of the current drugs also target the reverse transcriptase (RT), the enzyme responsible for copying the RNA genome into a dsDNA form, underscoring its importance. There are two classes of RT drugs: nucleoside RT inhibitors (NRTIs) and non-nucleoside RT inhibitors (NNRTIs). NRTIs mimic the dNTPs used for extending nucleic acid chains and therefore bind to the dNTP pocket in the polymerase active site. However, these drugs lack a 3’-OH group that enables the incorporation of the next nucleotide, and thereby act as chain terminators. NNRTIs bind to a pocket close to the polymerase active site and cause conformational changes in both protein and nucleic acid at the polymerase active site, hence acting as allosteric inhibitors (Das et al., 2012, Das et al., 2019). Figure **4** shows the binding pockets in their respective enzymes of PIs, NRTIs and NNRTIs.

Furthermore, three drugs are currently approved that bind to the integrase-DNA complex (intasome) and prevent the concatenation reaction (strand transfer) which enables the IN to join the viral genome to the host genome (Engelman and Cherepanov, 2012, Passos et al., 2017). To increase the genetic barrier to resistance which undoubtedly arises when these drugs are used singly, various fixed-dose combination therapies of these drugs have been formulated and are currently in clinical use as well (https://www.fda.gov).

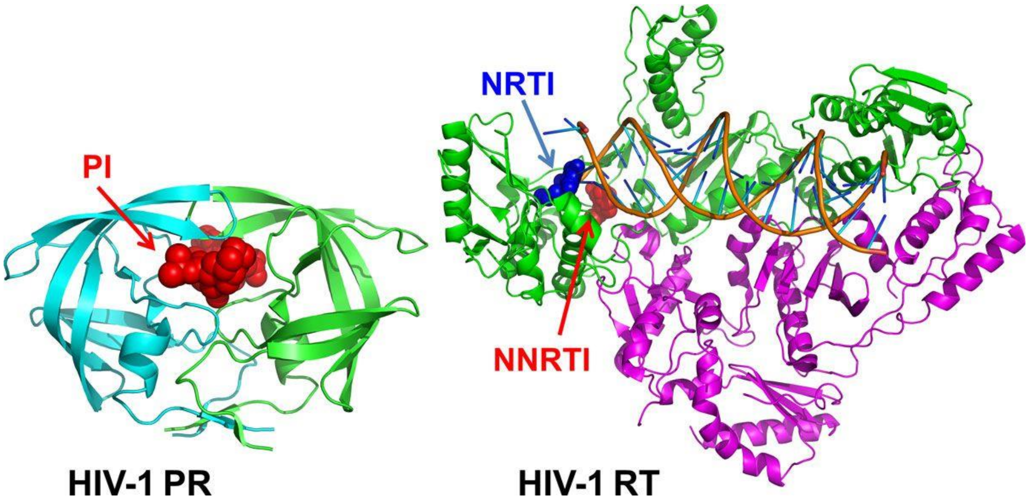
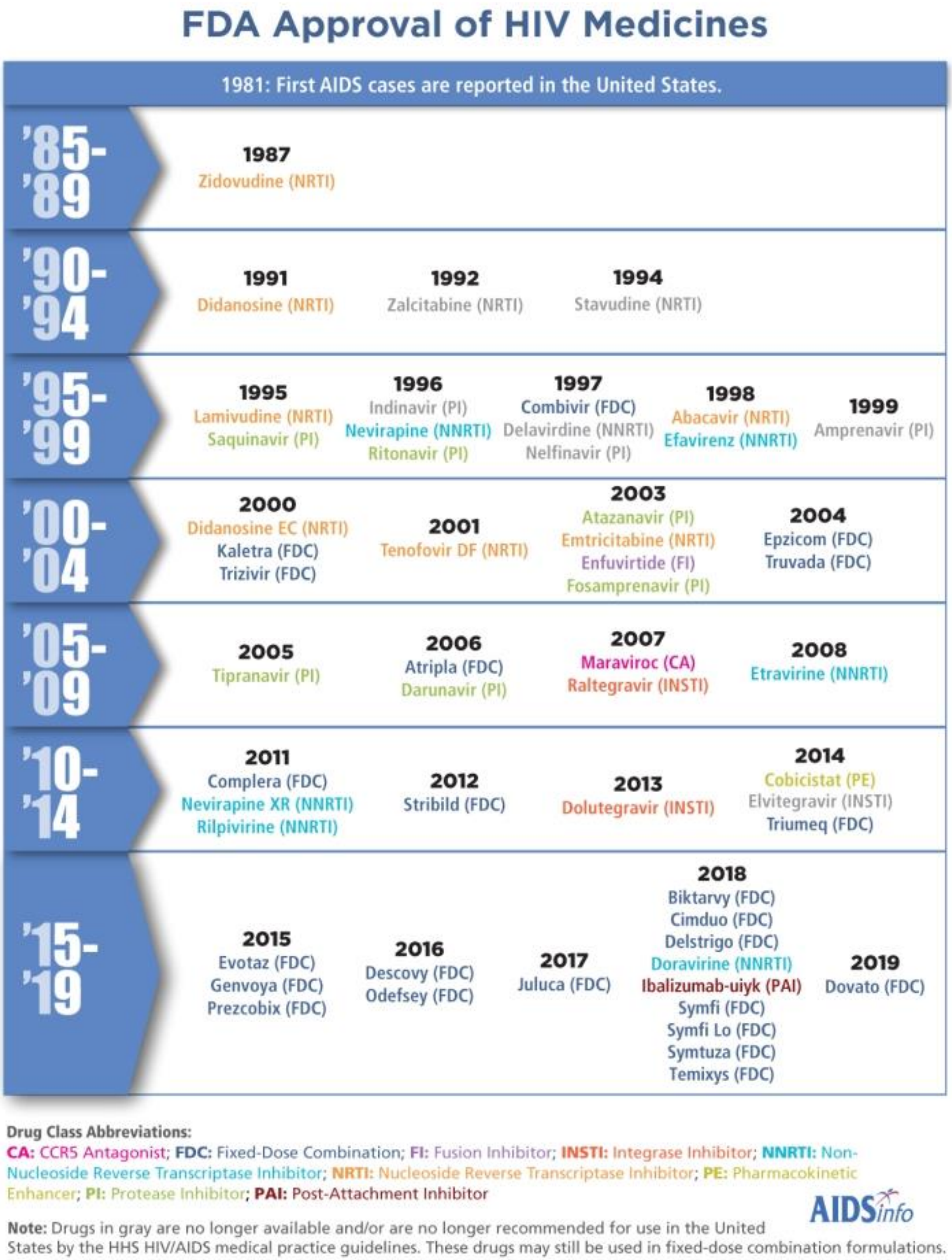


Figure 4: Reproduced from Yu et al., 2014 with the consent of BMC, the publisher.



FDA Approval of HIV Medicines

1981: First AIDS cases are reported in the United States.

’85-’89	**1987** Zidovudine (NRTI)				
’90-’94	**1991** Didanosine (NRTI)	**1992** Zalcitabine (NRTI)	**1994** Stavudine (NRTI)		
’95-’99	**1995** Lamivudine (NRTI) Saquinavir (PI)	**1996** Indinavir (PI) Nevirapine (NNRTI) Ritonavir (PI)	**1997** Combivir (FDC) Delavirdine (NNRTI) Nelfinavir (PI)	**1998** Abacavir (NRTI) Efavirenz (NNRTI)	**1999** Amprenavir (PI)
’00-’04	**2000** Didanosine EC (NRTI) Kaletra (FDC) Trizivir (FDC)	**2001** Tenofovir DF (NRTI)	**2003** Atazanavir (PI) Emtricitabine (NRTI) Enfuvirtide (FI) Fosamprenavir (PI)	**2004** Epzicom (FDC) Truvada (FDC)	
’05-’09	**2005** Tipranavir (PI)	**2006** Atripla (FDC) Darunavir (PI)	**2007** Maraviroc (CA) Raltegravir (INSTI)	**2008** Etravirine (NNRTI)	
’10-’14	**2011** Complera (FDC) Nevirapine XR (NNRTI) Rilpivirine (NNRTI)	**2012** Stribild (FDC)	**2013** Dolutegravir (INSTI)	**2014** Cobicistat (PE) Elvitegravir (INSTI) Triumeq (FDC)	
’15-’19	**2015** Evotaz (FDC) Genvoya (FDC) Prezcobix (FDC)	**2016** Descovy (FDC) Odefsey (FDC)	**2017** Juluca (FDC)	**2018** Biktarvy (FDC) Cimduo (FDC) Delstrigo (FDC) Doravirine (NNRTI) Ibalizumab-uiyk (PAI) Symfi (FDC) Symfi Lo (FDC) Symtuza (FDC) Temixys (FDC)	**2019** Dovato (FDC)

Drug Class Abbreviations:

CA: CCR5 Antagonist; **FDC:** Fixed-Dose Combination; **FI:** Fusion Inhibitor; **INSTI:** Integrase Inhibitor; **NNRTI:** Non-Nucleoside Reverse Transcriptase Inhibitor; **NRTI:** Nucleoside Reverse Transcriptase Inhibitor; **PE:** Pharmacokinetic Enhancer; **PI:** Protease Inhibitor; **PAI:** Post-Attachment Inhibitor

AIDSinfo

Note: Drugs in gray are no longer available and/or are no longer recommended for use in the United States by the HHS HIV/AIDS medical practice guidelines. These drugs may still be used in fixed-dose combination formulations.

Figure 5: https://aidsinfo.nih.gov/understanding-hiv-aids/infographics/25/fda-approval-of-hiv-medicines.

In the absence of a cure, continued efforts towards understanding in detail the various processes that enable the virus to infect cells are required to fuel the search of more potent drugs that are able to combat this disease while enhancing efforts to reach those who have no access to the drugs.

CHAPTER TWO: Structural studies of PFV PR-RT

Synopsis

Retroviruses continue to garner public health attention with an ever-increasing disease burden on the human population due to inadequate therapeutic interventions or loss of drug efficacy caused by widespread accumulation of resistance mutations in key enzymes enabling them to escape treatment. In the absence of effective vaccines or cure, therapeutic interventions for retroviral infections such as HIV, have relied on targeting key proteins in the viral life cycle. Retroviruses conventionally synthesize their proteins from polycistronic mRNAs into polyproteins, which are then proteolytically processed by the virally encoded proteases into functional entities. While the structures of some polyproteins from the structural proteins have been solved, revealing key insights into immature viral architecture and mechanisms of inhibition by maturation inhibitors, structural characterization of retroviral Pol polyproteins have generally been limited.

The crystal structure of the protease-reverse transcriptase (PR-RT) fusion, the viral reverse transcription and processing machinery of the prototype foamy virus (PFV) at 2.9 Å resolution, the first from the retroviral family is reported. Overall, this structure contains all the canonical domains observed for retroviral polymerases and proteases albeit with notable exceptions relating to the spatial positioning of the domains. While the monomeric PFV PR exhibits similar architecture as the HIV-1 PR, the N- and C-terminal residues that form a four-stranded β-sheet sandwich which stabilizes the mature dimeric HIV PR, remain unfolded in the structure. Extraneous residues between the PR and RT identified by multiple sequence alignment (MSA), a so called C-terminal extension, while unstructured at the C-terminal end of PR, folds into two helices at the base of the palm of the RT and in addition to the equivalents of helices E and F as in HIV-1, form an extended ‘floor’ of the palm. The PR is therefore anchored to the RT adjacent to the fingers and palm by these two helices via a long loop reminiscent of fishhook. This arrangement allows the PR to move unrestricted in different directions and angles to accommodate a dimerization partner without disturbing the integrity of the RT. Like HIV and other proteases, PFV PR functions as a dimer.

The structural components- (fingers, palm and thumb) are highly conserved in all retrovirus and retrotransposon RTs that carry out the addition of deoxyribonucleotides to nucleic acid strands using a common catalytic mechanism. The domain organization of the polymerase (fingers, palm, thumb, connection and RNase H) in PFV PR-RT is significantly different when compared to structures of other retroviral polymerases. Three functional entities PR-RT-RNase H are connected by flexible linkers. Notably, the C- terminus of the thumb and N-terminus of RNase H domains are linked to the connection subdomain via long flexible linkers that permit spatial rearrangement of these subdomains into a compact structure using interfaces similar to the HIV-1 RT p51. This structure therefore defines the structural organization of the “closed HIV-1 p51-like” polymerase precursor conformation of retroviral Pol prior to isomerization which has been speculated in the literature.

The novel spatial arrangements of the thumb, connection and RNase H, in the current structure compared to other retrovirus polymerase structures suggest that the PFV PR-RT would undergo significant conformational rearrangement upon nucleic acid binding (Switch and grab mechanism). This opens up the frontiers for modeling other Pol polyproteins, as well as serve as a useful model for the discovery of new anti-retroviral drugs that can stabilize this inactive RT conformation and prevent conformational maturation. This chapter describes in detail all the effort that resulted in this structure.

2.1 INTRODUCTION

Retroviruses as well as their evolutionarily related retrotransposons share conserved genome organization and architecture of structural and enzymatic proteins (Yu et al., 1996, Linial, 1999, Lee et al., 2013, Hutter et al., 2013). Synthesis of polyproteins from polycistronic mRNAs before proteolytic processing by cognate protease into functional entities, is a unique feature of many pathogens including RNA viruses, retroviruses and retrotransposons (Yu et al., 1996, Baldwin and Linial, 1998, Roy and Linial, 2007, Shin et al., 2012). The architecture of RTs and PRs remain perhaps the most conserved amongst these entities with a common mechanism of nucleic acid synthesis and proteolytic processing even in the absence of high sequence homology (Ding et al., 1998, Tozser, 2010, Nowak et al., 2014).

However, unlike lentiviruses such HIV-1 where the synthesis of the Pol polyprotein, which is subsequently proteolytically processed into mature enzymes, is synthesized as a Gag-Pol fusion from the same mRNA using translational frameshifting, spuma retroviruses such as PFV synthesize their Gag and Pol polyprotein from separate mRNAs alternatively spliced from their genomic RNA (Lochelt and Flugel, 1996, Pfrepper et al., 1998). Following their synthesis, the FV polyproteins, undergo only limited proteolysis during maturation with the integrase (IN) being the only domain processed from the Pol while a C-terminal peptide is processed from the Gag polyprotein (Yu et al., 1996). The mature polymerase therefore contains an N-terminal PR as well (PR-RT), both of which are functional.

The limited proteolysis of Gag results in infectious foamy virus (FV) particles that are morphologically indistinguishable from the immature lentiviral capsids though this mini-scale processing is absolutely required for infectivity in FVs (Yu et al., 1996, Lee et al., 2013). The RT of FVs is monomeric in solution and is responsible for copying the FV RNA 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 (Yu et al., 1999, Lee et al., 2013, Rethwilm and Bodem, 2013). These replication paradigms in addition to their apathogenicity offer unique avenues for studying processes in retroviruses such as HIV using FV as a surrogate.

The PR which is responsible for the proteolytic processing of polyproteins is functional only as a homodimer like all PRs of the *retroviridae* family (Katoh et al., 1989). Hence the PR-RT must dimerize to activate the protease. Even though the literature is replete with structures of retroviral PRs which explain the mechanism of dimerization post maturation, as well as RTs and their nucleic acid complexes, structural information on the domain organization of Pol polyproteins which contain relevant entities that offer insight into the mechanism by which retroviral proteases dimerize in the context of the Pol polyprotein which allows it to cleave other proteins as well as itself into functional entities has eluded researchers so far.

The asymmetric conformational maturation of heterodimeric HIV-1 RT 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 I report here the crystal structure of the wild-type 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 polyprotein thereby offering a glimpse into the initial events leading up to the proteolytic processing of polyproteins. The PFV RT architecture, being folded in an HIV-1 p51-like conformation, offers further insight into the structural organization of a monomeric RT precursor such as monomeric HIV-1 p66 RT prior to isomerization or maturation suggested in the literature. 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 (Trobridge, 2009, Erlwein and McClure, 2010, Lindemann and Rethwilm, 2011) 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.

2.2 Aims of the study

The precursor polyproteins remain attractive targets for the design of inhibitors of retroviral replication. This interest is due to the fact that protein synthesis in the form of polyprotein precursors are unique to these pathogenic viruses, hence potent inhibitors with little possibility of cross reactivity with host enzymes could be obtained. Such inhibitors would require low doses which will decrease the possibility of drug toxicity. While this may be a far cry, obtaining crystal structures of these proteins would pave the way for the search for the next generation of anti-retrovirals which target the nascent stages of viral assembly. This study aims therefore at structural and functional characterization the of the prototype foamy virus PR-RT as a surrogate for understanding how the mature enzymes are arranged in their precursor forms in other retroviruses such as HIV, which is of much higher public health impact.

2.3 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). This complicates purification for crystallographic studies since most of the proteolytic products have high affinity for commonly used chromatography columns. This makes purification laborious resulting in poor purity and low yield even in the presence of high doses of protease inhibitors and reducing agents (**Fig. 6**). The similarity between the full-length 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 lattice of the full-length protein. Each of these scenarios has deleterious consequences for obtaining crystals suitable for structural studies.

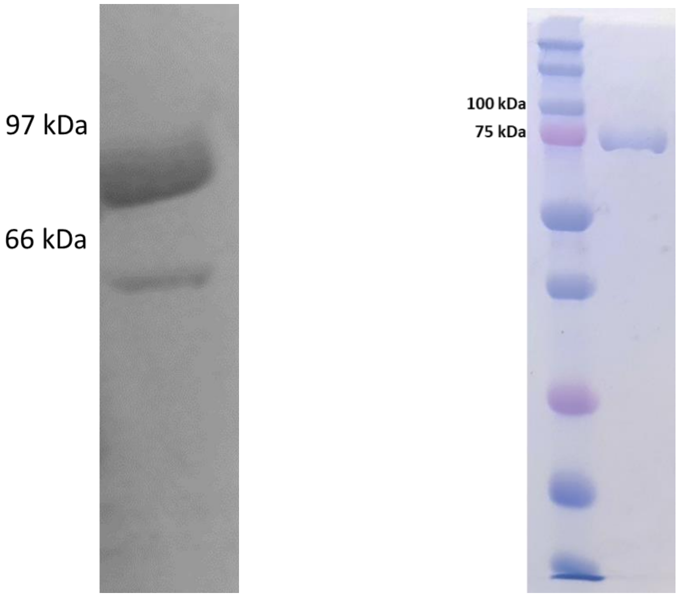


Figure 6: SDS-PAGE gel of the WT PR-RT after purification (left) and the CSH mutant (right).

In an effort to improve the proteolytic stability of WT PFV PR-RT, an *in silico* mutagenesis of the primary sequence of PFV PR-RT was carried out by replacing each amino acid at each position with the other 19 and computing an instability index on the ExPASy ProtParam server (http://web.expasy.org/protparam/) using a python script. The instability index computed is a weighted sum of the dipeptides 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 found 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 amino acids were preferred at these positions than the residues themselves. Two mutations, H507D and S584K were selected based on biophysical properties considerations in addition to C280S which is known to improve the solution behavior of HIV-1 RT. This PR-RT mutant designated CSH resulted in at least 4-fold increase in expression compared to the wild-type (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. 6**).

2.4 Enzymatic assays (In collaboration with Stephen Hughes, NCI Frederick, MD)

Changes to the primary sequence of proteins especially in the absence of structure 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 (**Fig. 7**), which was expected since all the mutations were made outside of the RNase H domain, the ability of the mutant enzyme to incorporate dNTP into a nascent chain was diminished slightly.

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 (**Fig. 7**). 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. This was not surprising since replacement of His507 with an

Asp reverses a positive charge into a negative charge which could result in nucleic acid backbone repulsion.

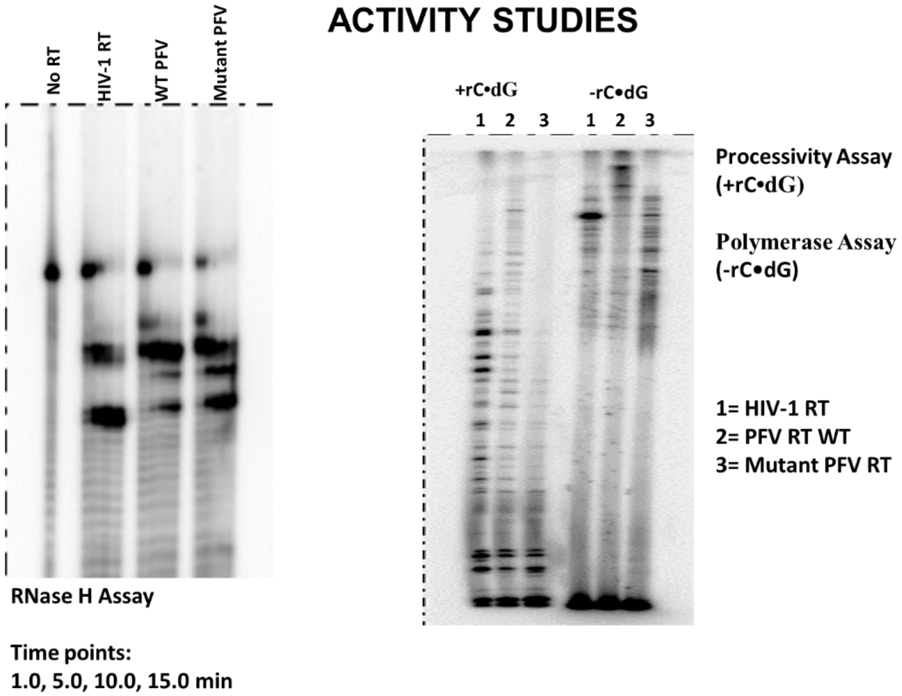


Figure 7: Ribonuclease activity of PR-RT (left) and polymerase and processivity activity using poly (rC). oligo(dG).

2.5 Crystallization trials

Crystallization screening for the CSH mutant were carried with the help of a crystallization robot (ARI Crystal Gryphon) using commercially available screening kits. Natrix, Index, PegRx and Crystal (Hampton) screen were among the kits tried. A molar ratio of 1:1 between 20 mg/ml of the CSH mutant to precipitant in an initial 96 well screening by the sitting drop vapor diffusion crystallization produced clusters of plate-like crystals which emanated from a “stalk” in Hampton’s Natrix conditions E7 and E8 as well as Index condition F5 at 4 and 20 °C within 48 hours.

Natrix E7 comprised 100 mM NaCl, 50 mM sodium cacodylate trihydrate pH 6.0, 10% PEG 4000, 0.5 mM spermine; Natrix E8; 50 mM KCl, 50 mM sodium cacodylate trihydrate pH 6.0, 10% PEG 8000, 0.5 mM spermine, 0.5 mM L-argininamide while Index screen F5 was composed of 100 mM ammonium acetate, 100 mM Bis-Tris pH 5.5, 17% PEG 10,000. Following several rounds of optimization which included changing pH, precipitant concentrations, protein concentration and precipitant ratios, micro-seeding, and crystallization under oil, the crystal form did not change compared to the initial hits.

2.6 Additive screening

In an effort to obtain crystals suitable for X-diffraction studies, an additive screening was also carried out having failed to obtain suitable crystals using conventional optimization methods. Additives are small molecule organic and inorganic compounds or ions which bind to channels, pockets, surfaces, crystal contacts etc. in ways that stabilize them and enable macromolecular crystals to grow bigger or change the morphonology or in some cases stabilize the packing to enable high resolution X-ray diffraction from an otherwise poorly diffracting crystal. The Hampton additive screen was tried with the Index F5 and the Natrix E8 conditions. Several crystals grew in many conditions with similar morphology to the initial crystals.

A careful examination of the Natrix E8 additive screening at 4 ºC let to the realization that crystals that grew in the presence of 30 mM glycylglycylglycine (GGG), and 10 mM EDTA at 20 ºC were of slightly different morphology than the initial crystal hit. Optimization was therefore carried out using these 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 the only one available in the lab at the time and so was tried. 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. The options for growing better crystals were limited at this point with no obvious path for improving these crystals forms. A 1:1 mixture of the two conditions containing GG and EDTA were mixed together in one experiment before crystallization was set up. Surprisingly, the crystals grown in these conditions were significantly thicker than those produced by their individual conditions. Thus, the GG and EDTA had a cumulatively positive effect on the crystal size even though they remained plates emanating from a single nucleation site.

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-srgininamide, 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. Larger drops produced thicker crystals. 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 enabling structural studies. Optimized crystallization conditions using this mutant enabled were subsequently used to crystallize and solve the structure of the WT as well. Some images of the crystals in the course of the optimization are shown in Figure **8a-d**.

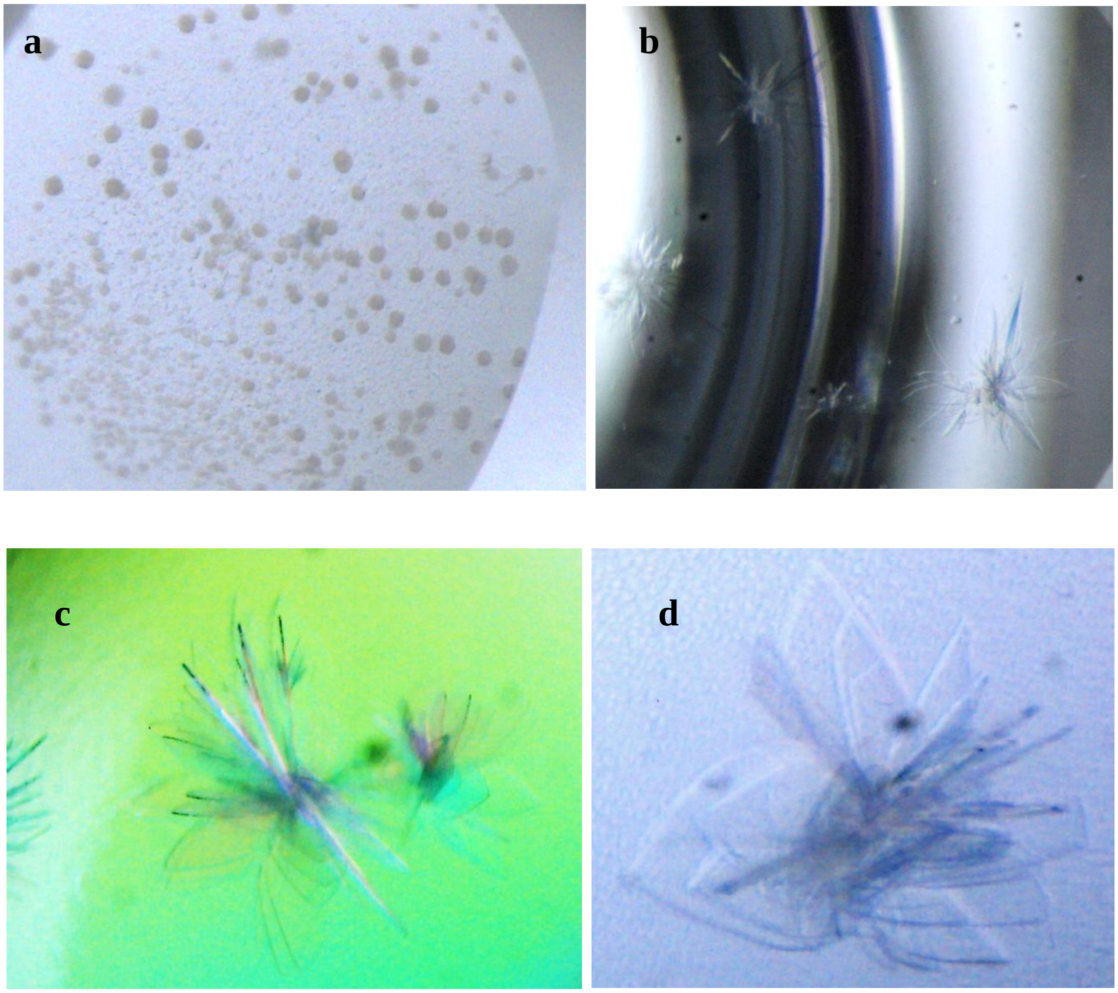


Figure 8: (A) Initial crystallization hit (B) Optimization of initial additive screening with (C) and (D) Fully optimized crystals used for structure determination.

2.7 Phasing and structure refinement

Having obtained crystals which diffracted routinely to 3.0 Å resolution, it was anticipated that phasing by molecular replacement would be simple and straight forward, but it turned out not to be the case. Models of HIV-1 RT heterodimer, p66 only, p51 only, MMLV and individual domains of these proteins were used to phase the structure without success. Soaking of heavy metals, Hg, Pt, Y, Yb, Sm, Rh, Pb, Au, W, Eu, Gd, Ir, and Ta complexes, including halogens Br and I and their pyrazole derivatives were tried without success. Co-crystallization with Br and I-pyrazole shown to useful for single anomalous diffraction (SAD) phasing of macromolecular structures (Bauman et al., 2016) as well as many of these heavy metals were also tried but none was successful at providing anomalous signal.

Selenomethionine (SeMet) labeling of the CSH mutant was carried out after initial effort at molecular replacement and heavy metal soaks failed. By growing the bacterial cell in minimal media supplemented with SeMet at the time of induction of protein expression, all the Met residues in the protein are expected to be substituted with the SeMet. Selenium has more electrons than sulfur, and scatters more strongly when incorporated into the crystal, and also has a stronger anomalous signal in the X-ray data which can be used to phase structures (Hendrickson et al., 1989, Hendrickson et al., 1990).

Mass spectroscopy analysis of the purified protein (**Fig. 9**) showed that 13 out of the 14 Met residues in the PR-RT (89%) were substituted by SeMet. After several trials, crystals of the SeMet-labeled protein were obtained which diffracted X-rays to similar resolution as the unlabeled protein. The crystals have the symmetry of space group C2. Any expectations that the structure solution would be routine with these crystals were misplaced. While individual datasets showed anomalous signal, it was too weak to phase the structure. It must be emphasized that the generally high mosaicity (>1^{o}) for these crystals further compounded the problem. Eventually, the structure was solved using single anomalous diffraction (SAD) phasing from selenomethionine labelled PFV CSH mutant to 3.0 Å resolution from data merged together from four datasets collected at two different wavelengths (0.987 Å and 0.979 Å). The merging of the datasets increased the anomalous signal, enabling structure solution. Initial phases which enabled modeling of about 40% of the residues were obtained using the Crank2 pipeline (Pannu et al., 2011) in ccp4i

(Potterton et al., 2018).The rest of the model was built and refined through several rounds of iteration and density modification using Coot (Emsley et al., 2010), PHENIX (Adams et al., 2011) and REFMAC (Murshudov et al., 1997).

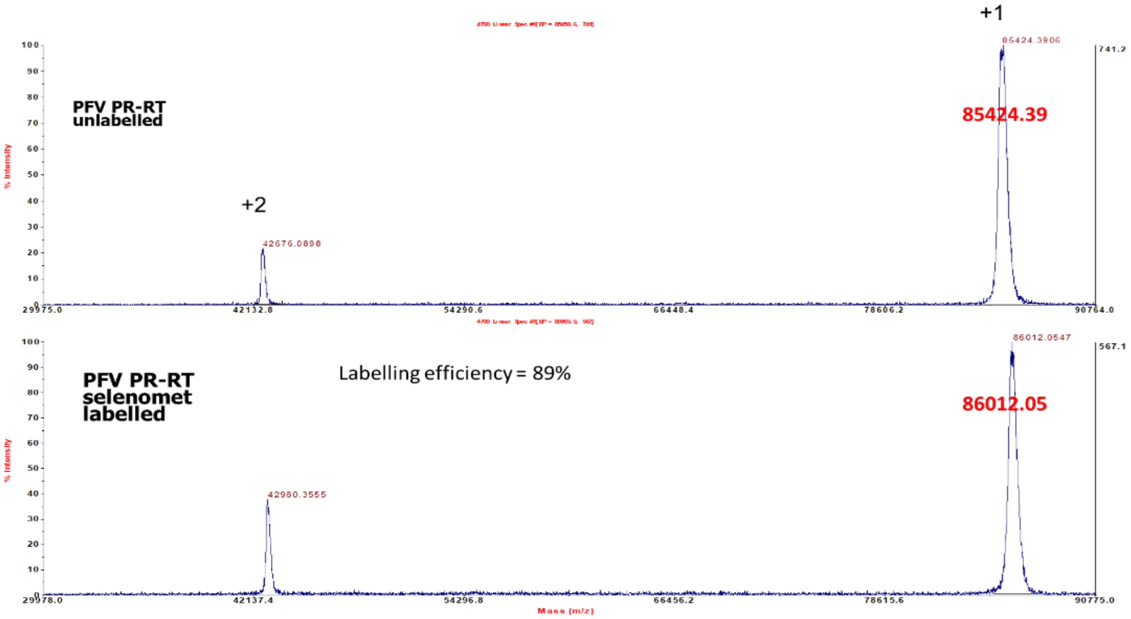


Figure 9: MALDI-TOF MS Spectra of the unlabeled PR-RT and SeMet labeled PR-RT.

2.8 Architecture of PFV PR-RT

PFV PR-RT crystallizes as a monomer with a single molecule in the asymmetric unit in a C2 space group (**Fig. 10**). The structure was solved by using the anomalous signal from the SeMet substituted CSH mutant and the model obtained used as the search model in molecular replacement with the WT data in the same space group (**Fig. 10**). Even though sequence homology of PFV with other retroviruses is generally low, typically less than 30%, individual structural elements of retroviral polymerases and proteases are highly conserved. Notable differences however exist in terms of relative domain positioning and arrangements which offer potential insights into their unique functionalities and subtle differences in substrate binding specificities, processivity and fidelity of PFV PR-RT.

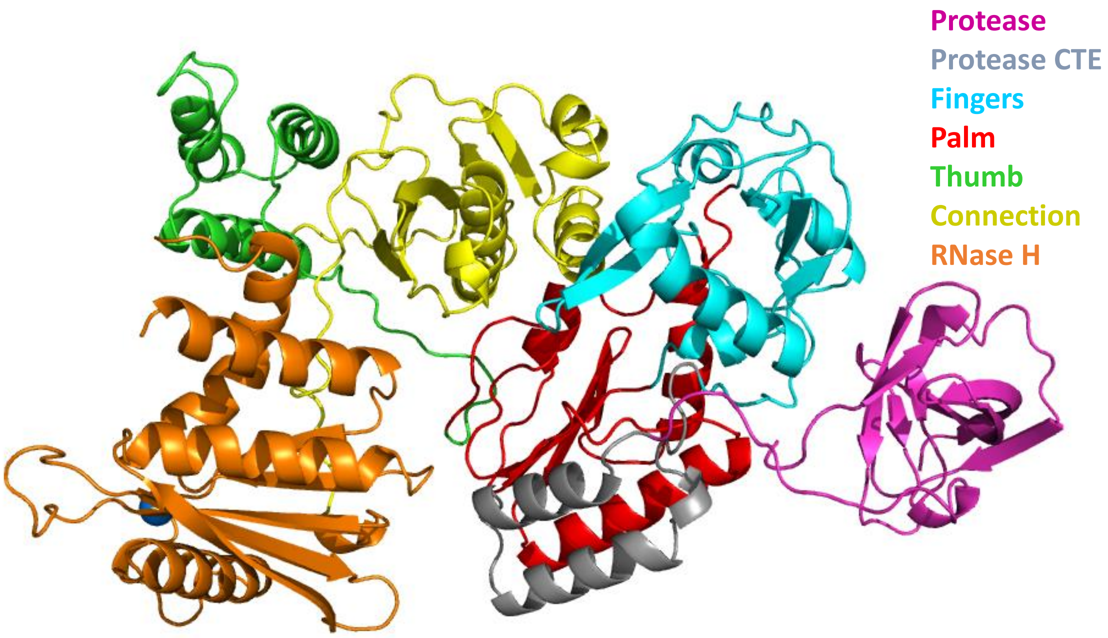


Figure 10:Architecture of the PFV PR-RT.

2.9 PR and the PR-CTE Domain

Seven beta-sheets (β1'-β7'), two short helices (αA' and αB') and random coils define the structure of the monomeric protease. The well folded single unit of the protease known to function exclusively as a homodimer exhibits the closed barrel-like core domain made up of predominantly 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 (**see Fig. 11**).

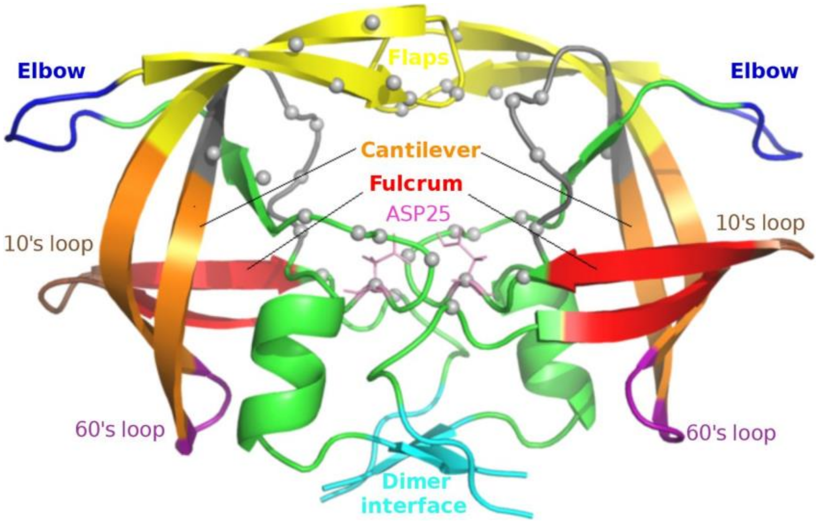


Figure 11: Architecture of the HIV-1 PR reproduced from Sheik Amamuddy et al., 2018.

Strands β1’ and β2’ which forms 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. The continuous hairpin loop bordered on each side by β2’ and β3’ define the “fireman's grip” which harbors the conserved DSG (residues 24, 25, 26) catalytic triad, at the active site. This large loop tilts slightly from orthogonality compared 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 is seen as 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 structure, 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’, that allows the tip of the flap to open and close to allow 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. 10**). The large P1-loop which lines the walls of the active site connects β7’ (residues, 85-87) and β6’ which are the only parallel sheets in the structure.

A short coil and a short helix (half-turn, αB’) further links 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 with well-defined density in this structure. Whether these would form beta sheets upon dimerization as 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.

Additional residues at the protease C-terminus have been observed through multiple sequence alignments (**Fig. 12).** The residues 101-143 which was initially unknown as either being part of 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 (**Fig. 16**). This CTE forms a long linker closer to the C-terminus of the PR. It however forms three helices (designated αC’, αD’, αE’) which pack against helices F, G and strand 4, (see **Fig. 12**) the equivalents of helices E and F and strand 6 in HIV-1 RT which forms the “floor” of the palm containing the polymerase active site. This stabilized packing with an extensive hydrophobic network forming an extended floor of the palm, also anchors the PR monomer to the RT which allow the protease to swing open or tilt to enable dimerization without any anticipated steric requirements 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 reaching dimer.

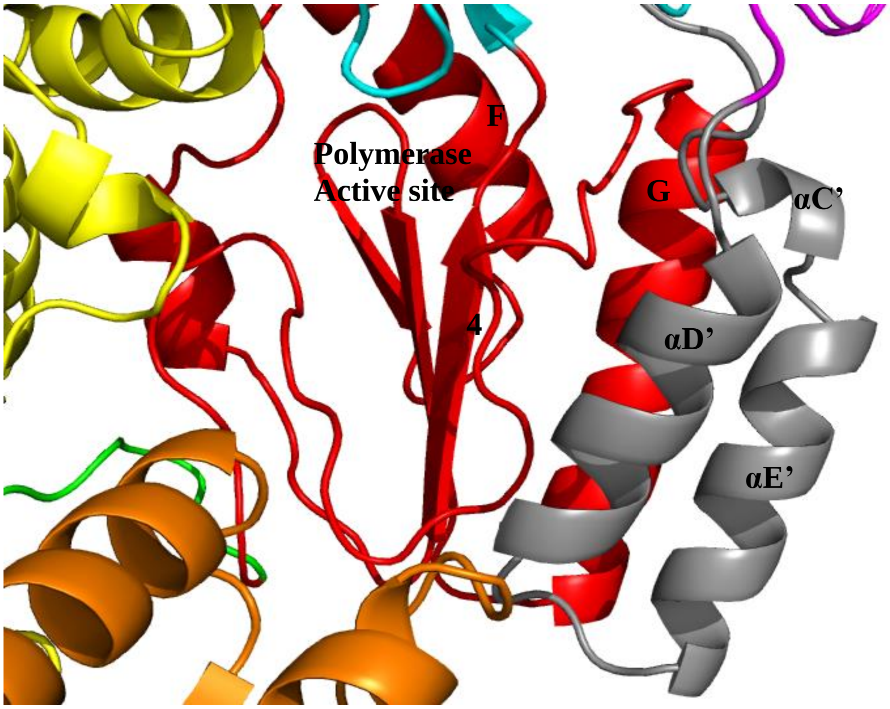


Figure 12: Interface formed by the PR-CTE and the Palm domain.

An extensive network of interactions occurs between the fingers subdomain of the RT, the unstructured N and C-terminal residues, the fulcrum, and the P1 loop of the protease which keep it closely attached to the RT even though these interactions are weak (**Fig. 13**).

Two PR-RT structures positioned *in silico* by two-fold symmetry creates dimeric PRs in a manner poised for catalysis without extensive steric clashes with the RTs (**Fig. 14, 15**) 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 protease appears much wider than that of HIV-1 which may explain why HIV-1 PR inhibitors such as tipranavir, darunavir and indinavir (Hartl et al., 2010b), do not inhibit PFV PR.

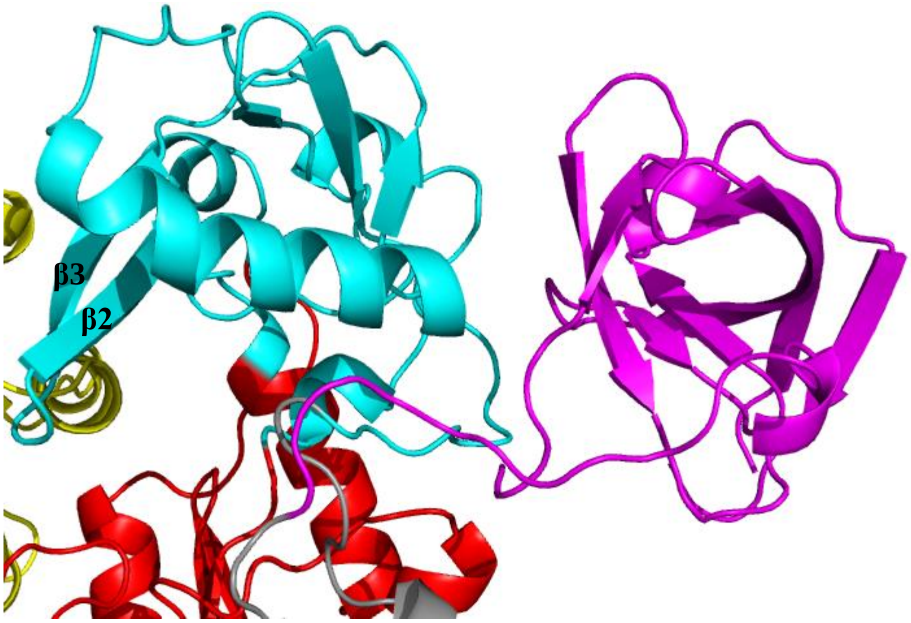


Figure 13: Interface formed by the PR monomer (pink) and the fingers subdomain (cyan).

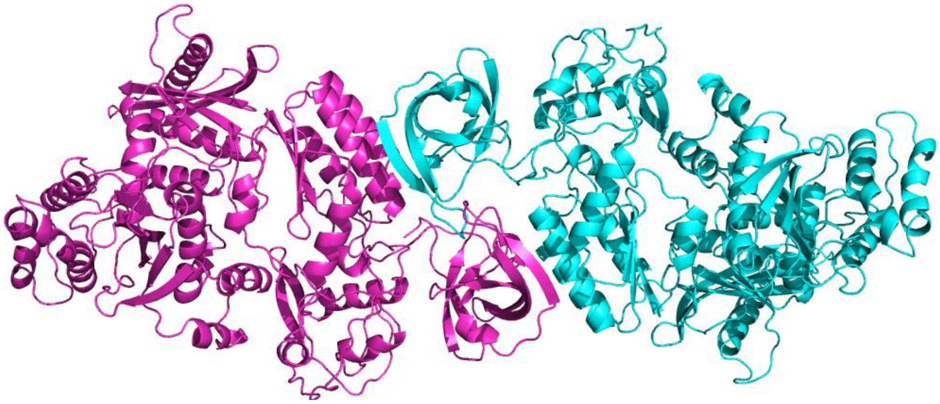


Figure 14: Dimers of PR-RT generated in silico by a 2-fold rotation

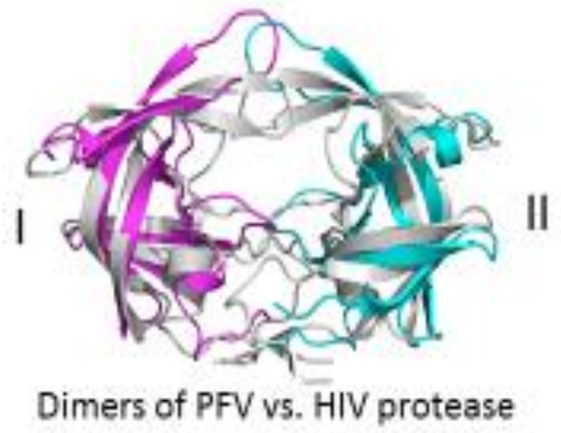
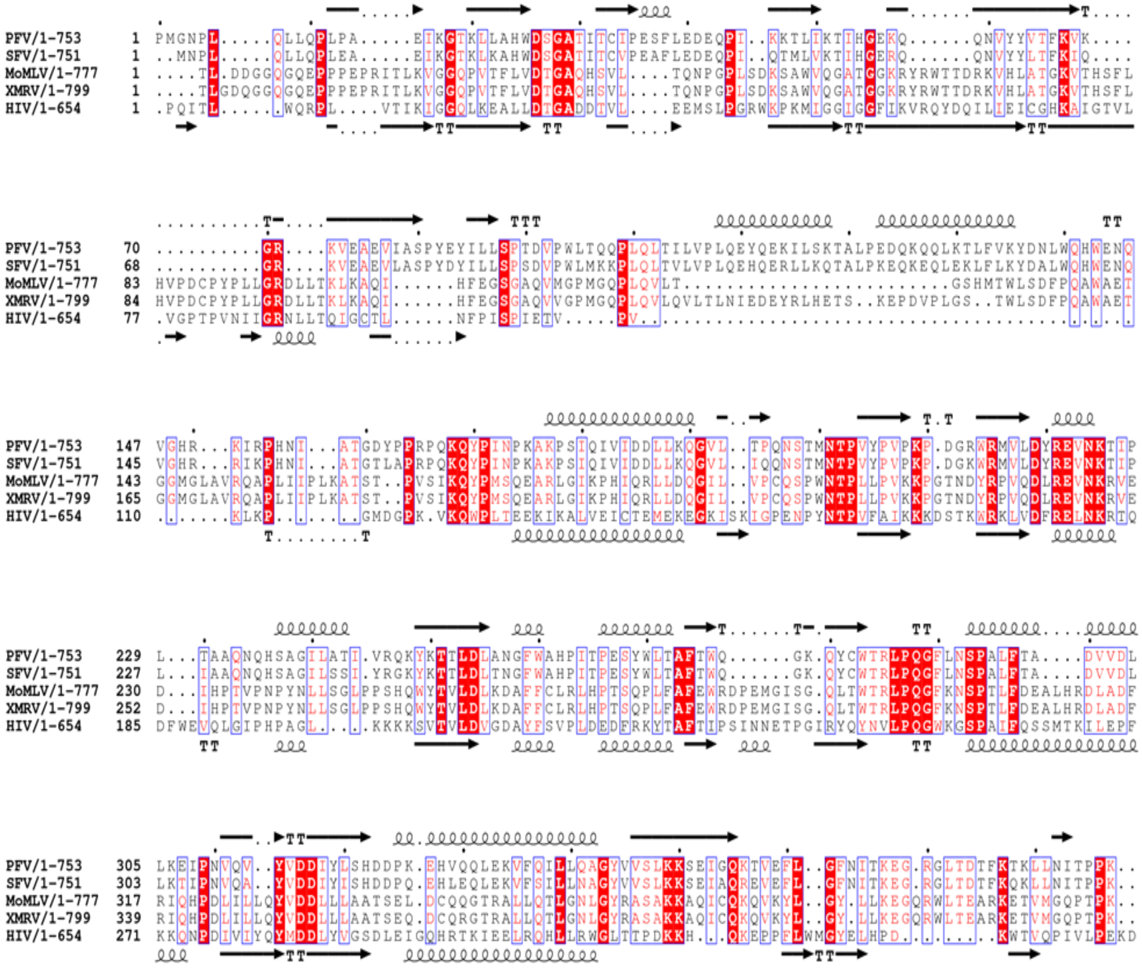


Figure 15: Superposition of dimeric HIV-1 PR (PDB ID 2HB4 on dimeric PFV PR generated in silico by 2-fold axis.

2.10 RT Domain

Retroviral RTs and other polymerase structures solved resemble the right hand and so the domains in these proteins have been historically named after the corresponding segments of the right hand. Even though the RT structure of the PFV does not resemble the right hand in its current configuration, all the canonical subdomains with their unique secondary structural features as identified in retroviral polymerase structures solved so far- fingers, palm, thumb, connection, and RNase H, domain, are present in the PFV RT (**Fig. 10**).

The relative positions of these subdomains compared to other retroviral RTs are different. The fingers subdomain of PFV 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, characteristic 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. 13**).


PFV/1-753 1 PMGNPL.....QLLQPLPA.....EIKGTKLLAHWDSGATITCIPESFLEDEQPI..KKTLIKTIHGEKQ......QNVYYVTFKVK....
SFV/1-751 1 ..MNPL.....QLLQPLEA.....EIKGTKLKAHWDSGATITCVPEAFLEDEQPI..QTMLVKTIHGERQ......QNVYYLTFKIQ....
MoMLV/1-777 1 ....TL.DDGGQGQEPPPEPRITLKVGGQPVTFLVDTGAQHSVL....TQNPGPLSDKSAWVQGATGGKRYRWTTDRKVHLATGKVTHSFL
XMRV/1-799 1 ....TLGDQGGQGQEPPPEPRITLKVGGQPVTFLVDTGAQHSVL....TQNPGPLSDKSAWVQGATGGKRYRWTTDRKVHLATGKVTHSFL
HIV/1-654 1 .PQITL......WQRPL....VTIKIGGQLKEALLDTGADDTVL....EEMSLPGRWKPKMIGGIGGFIKVRQYDQILIEICGHKAIGTVL
PFV/1-753 70 ....GR....KVEAEVIASPYEYILLSPTDVPWLTQQPLQLTILVPLQEYQEKILSKTALPEDQKQQLKTLFVKYDNLWQHWENQ
SFV/1-751 68 ....GR....KVEAEVLASPYDYILLSPSDVPWLMKKPLQLTVLVPLQEHQERLLKQTALPKEQKEQLEKLFLKYDALWQHWENQ
MoMLV/1-777 83 HVPDCPYPLLGRDLLTKLKAQI......HFEGSGAQVMGPMGQPLQVLT.........................GSHMTWLSDFPQAWAET
XMRV/1-799 84 HVPDCPYPLLGRDLLTKLKAQI......HFEGSGAQVVGPMGQPLQVLQVLTLNIEDEYRLHETS..KEPDVPLGS..TWLSDFPQAWAET
HIV/1-654 77 .VGPTPVNIIGRNLLTQIGCTL......NFPISPIETV.....PV......................................
PFV/1-753 147 VGHR...KIRPHNI...ATGDYPPRPQKQYPINPKAKPSIQIVIDDLLKQGVL..TPQNSTMNTPVYPVPKP.DGRWRMVLDYREVNKTIP
SFV/1-751 145 VGHR...RIKPHNI...ATGTLAPRPQKQYPINPKAKPSIQIVIDDLLKQGVL..IQQNSTMNTPVYPVPKP.DGKWRMVLDYREVNKTIP
MoMLV/1-777 143 GGMGLAVRQAPLIIPLKATST..PVSIKQYPMSQEARLGIKPHIQRLLDQGIL..VPCQSPWNTPLLPVKKPGTNDYRPVQDLREVNKRVE
XMRV/1-799 165 GGMGLAVRQAPLIIPLKATST..PVSIKQYPMSQEARLGIKPHIQRLLDQGIL..VPCQSPWNTPLLPVKKPGTNDYRPVQDLREVNKRVE
HIV/1-654 110 .......KLKP..........GMDGPK.VKQWPLTEEKIKALVEICTEMEKEGKISKIGPENPYNTPVFAIKKKDSTKWRKLVDFRELNKRTQ
PFV/1-753 229 L...TAAQNQHSAGILATI.VRQKYKTTLDLANGFWAHPITPESYWLTAFTWQ......GK.QYCWTRLPQGFLNSPALFTA....DVVDL
SFV/1-751 227 L...IAAQNQHSAGILSSI.YRGKYKTTLDLTNGFWAHPITPESYWLTAFTWQ......GK.QYCWTRLPQGFLNSPALFTA....DVVDL
MoMLV/1-777 230 D...IHPTVPNPYNLLSGLPPSHQWYTVLDLKDAFFCLRLHPTSQPLFAFEWRDPEMGISG.QLTWTRLPQGFKNSPTLFDEALHRDLADF
XMRV/1-799 252 D...IHPTVPNPYNLLSGLPPSHQWYTVLDLKDAFFCLRLHPTSQPLFAFEWRDPEMGISG.QLTWTRLPQGFKNSPTLFDEALHRDLADF
HIV/1-654 185 DFWEVQLGIPHPAGL.....KKKKSVTVLDVGDAYFSVPLDEDFRKYTAFTIPSINNETPGIRYQYNVLPQGWKGSPAIFQSSMTKILEPF
PFV/1-753 305 LKEIPNVQV..YVDDIYLSHDDPK.EHVQQLEKVFQILLQAGYVVSLKKSEIGQKTVEFL..GFNITKEG.RGLTDTFKTKLLNITPPK..
SFV/1-751 303 LKTIPNVQA..YVDDIYISHDDPQ.EHLEQLEKVFSILLNAGYVVSLKKSEIAQREVEFL..GFNITKEG.RGLTDTFKQKLLNITPPK..
MoMLV/1-777 317 RIQHPDLILLQYVDDLLLAATSEL.DCQQGTRALLQTLGNLGYRASAKKAQICQKQVKYL..GY.LLKEGQRWLTEARKETVMGQPTPK..
XMRV/1-799 339 RIQHPDLILLQYVDDLLLAATSEQ.DCQRGTRALLQTLGNLGYRASAKKAQICQKQVKYL..GY.LLKEGQRWLTEARKETVMGQPTPK..
HIV/1-654 271 KKQNPDIVIYQYMDDLYVGSDLEIGQHRTKIEELRQHLLRWGLTTPDKKH...QKEPPFLWMGYELHPD.........KWTVQPIVLPEKD
T
TT
TTT
T.T
T........T
T......T

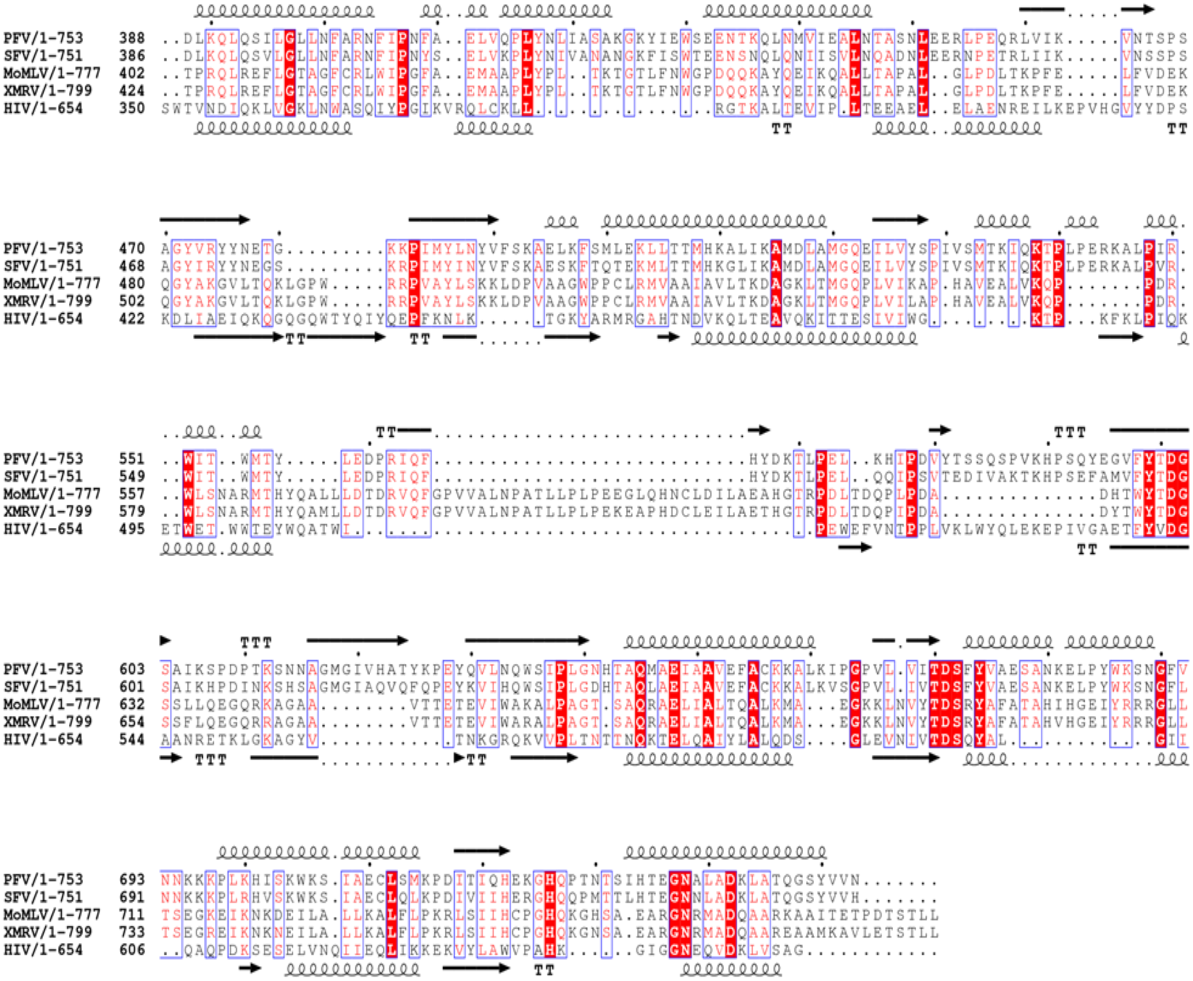


Figure 16: Secondary structural alignments of some retroviral PRs and RTs with PFV generated using Clustal Alignment.

The palm of PFV RT (229-259, 291-367) snuggled between the connection, fingers and PR-CTE subdomains comprises three helices and four anti-parallel β-sheets similar to other RTs. The sheet which contains the conserved residues YVDD (314-317) together with the primer grip (355-368) constitute the polymerase active site. The primer grip observed in other RTs as a short β-hairpin, is rather a large loop pointing towards the active site hairpin. The residues in the hairpin loop appears largely conserved with MoMLV, XMRV and HIV-1. This flexible primer grip architecture may reduce the structural constraints imposed by rigid β-sheets in other RTs which allows for easier primer orientation during nucleic acid synthesis and hence contributing to the higher processivity of PFV compared to other RTs (**Fig. 17**).

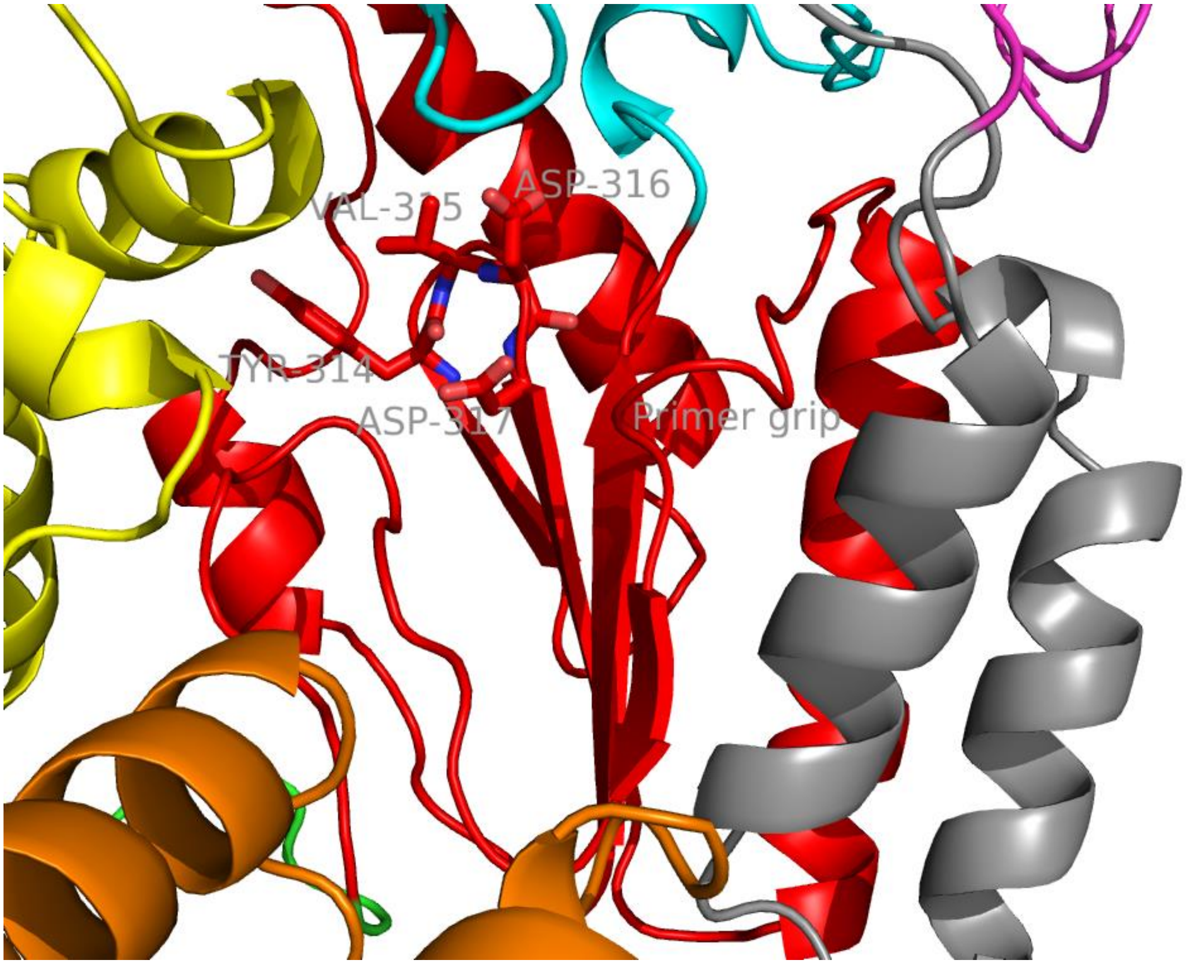


Figure 17: The PFV RT polymerase active site residues and some conserved structural elements found in retroviral RTs.

Concomitantly, mutation of V315M to mimic the active site of HIV-1 leads to 50% loss of RT activity in virions with no observable full-length cDNA detectable in transfected cells with reduced processivity (Rinke et al., 2002, Boyer et al., 2004). Apart from the possibility of steric clashes, the bulky side chain potentially reduces the flexibility of this loop hence slowing down the rate of nucleotide incorporation needed to complete cDNA synthesis. A conserved residue D254 situated at the tip of β4 together with the polymerase active site loop created by β7, β8, house D316 and 317 which together constitute the catalytic aspartate triad (β9, β10 of HIV-1) (**Fig. 17**). While helix F (E, in HIV-1 RT) is oriented about 100º relative to helix G (F in HIV-1 RT) above the active site without the kinks observed in HIV-1 p66/p51, 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 domain (450-590), making extensive contact with it (**Fig. 18**).

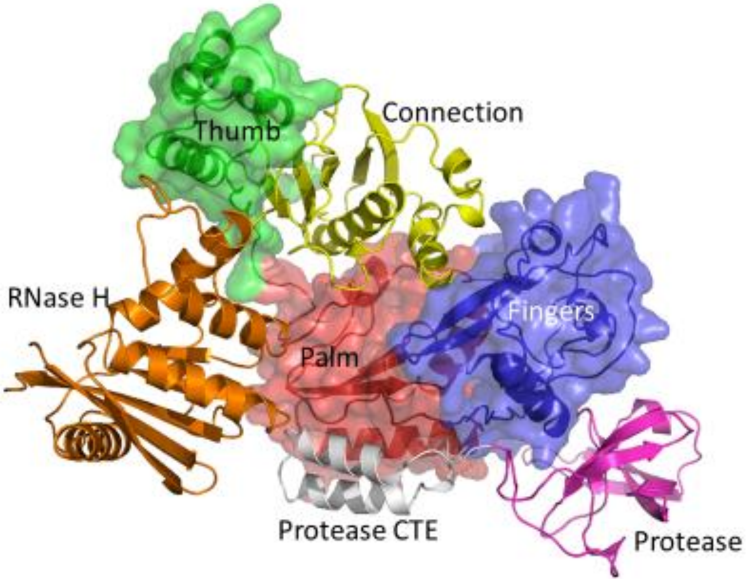
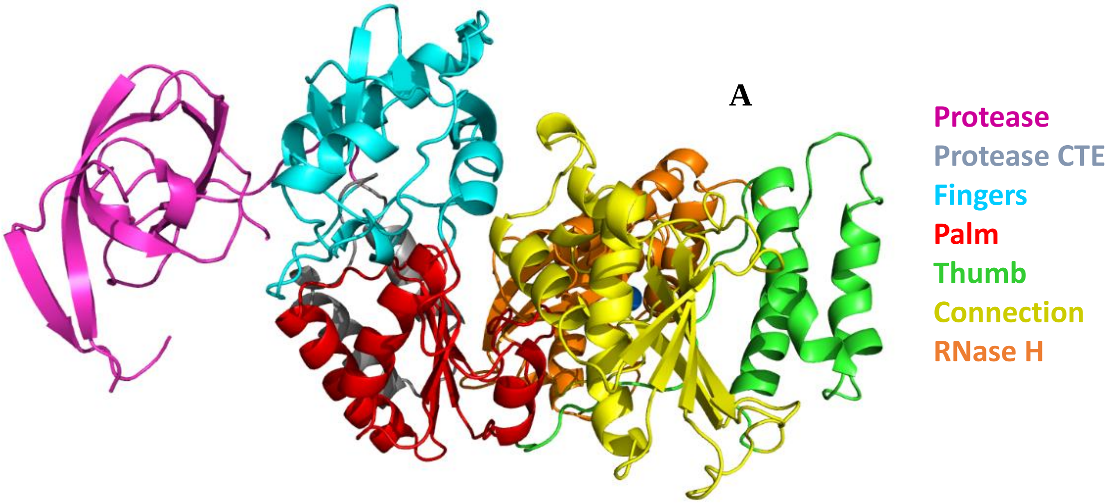


Figure 18: Surface rendering of the finger-palm-thumb subdomains in PR-RT showing the connection and RNase H backbones.

The connection domain characterized by five-stranded mixed beta-sheets stabilized by five helices sits almost parallel to β7, β8 compared to its position in HIV-1 p66 and other RTs where they are orthogonal. This compact fold allows it to snugly fit between the fingers and the RNase H subdomain on opposite sides and the thumb on the other side, tilted slightly towards the active site. Helix L (K in HIV RT) packs against H, while helix O (L in HIV RT) tilts towards the active site loop positioning T552 to form a hydrogen bond with Y314 (3.0Å) , T556 with H236 (2.3Å) while H509 forms a bifurcated hydrogen bond with S237 (2.6Å) and N364 (**Fig. 19A-C**).

The H509D mutant employed in initial studies also makes these contacts and crystallizes much faster compared to WT, while H509F/ I mutants which lose these hydrogen-bond interactions, did not crystallize after several attempts. This underscores the importance of these hydrogen bonds on stabilizing this conformation. The C-terminus of helix L also makes extensive contacts with the primer grip and helix C of the palm.



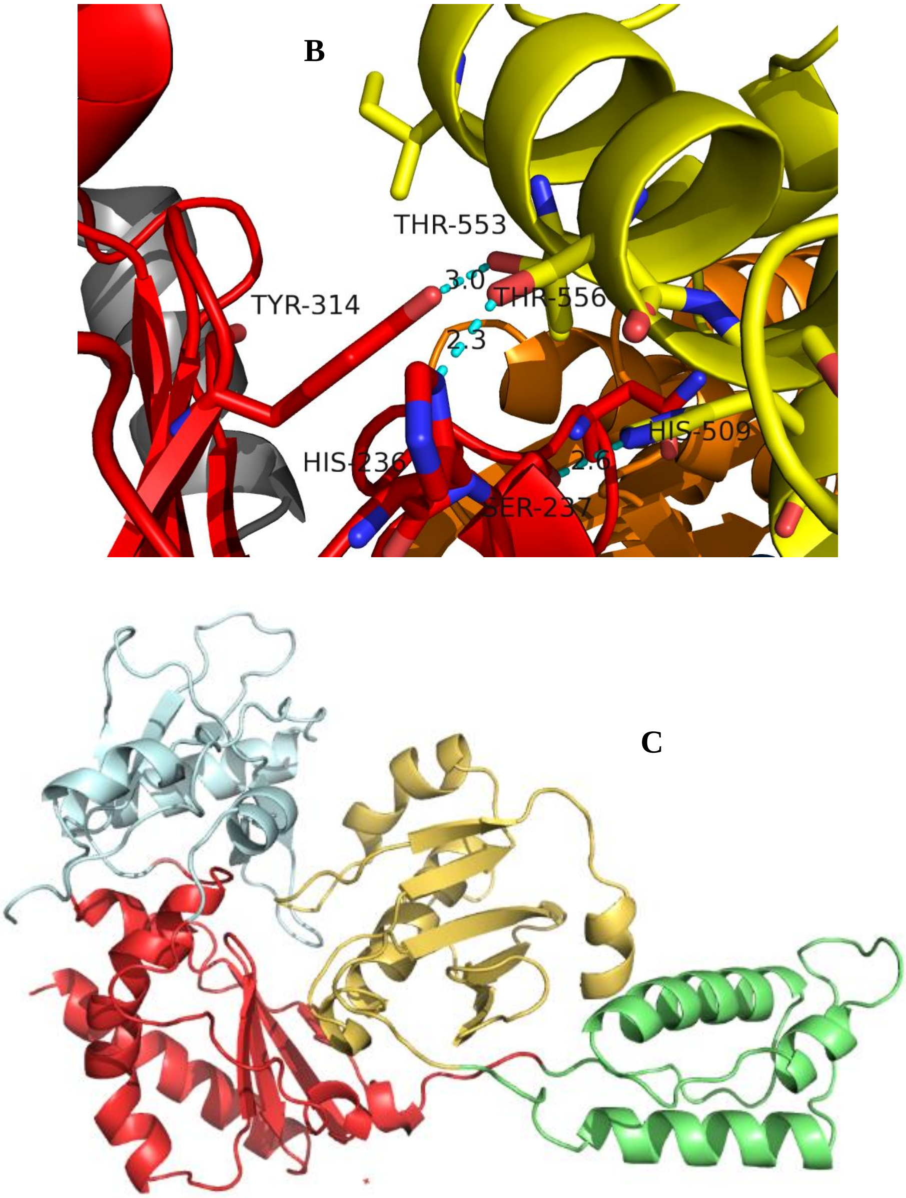


Figure 19: (A) Structure of PR-RT-orientation of the domains from N- terminus to the C-terminus (B) hydrogen-bonding interactions between connection and palm in PR-RT (C) Relative positioning of individual subdomains of HIV-1 RT p51 color coded.

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 conceivable 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. 19B**).

The three-helix bundle characteristic of the thumb (αH, αI, αJ) together with helix L, sandwich 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. 10**). Even though the antiparallel beta-sheets at the base of the thumb which help position it uprightly close to the fingers 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 is extended in a way and detached from the connection that exposes helix H to form the interface with the RNase H domain in the heterodimer. Compared to PFV thumb, this represents close to 90° rotation (**Fig. 19C** and **20**).

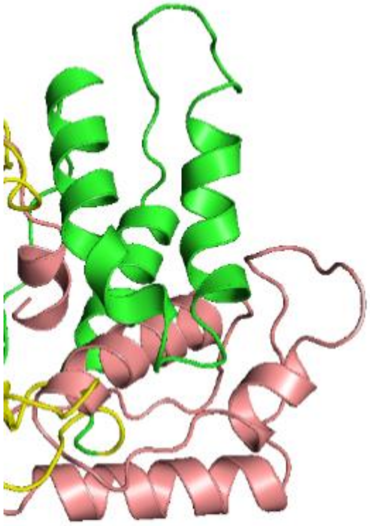


Figure 20: Relative orientation of the thumb subdomain of PFV (green) and HIV-1 p51 (salmon) based on alignment of fingers and palm subdomains.

It must be noted that while some hydrophobic residues are buried in these interfaces between the connection/palm/thumb, key anchoring interactions are actually hydrogen bonds and van der Waal's interactions. Since this machinery must undergo extensive rearrangement to function as a polymerase, these rather 'weak' interactions ensure that the energetic barrier to isomerization is not insurmountable as compared to HIV-1 p51 where these interfaces are predominantly hydrophobic.

2.11 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, 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. The RNase H domain is swiveled around the connection and the thumb, and positioned next to the fingers, and opposite the connection/thumb domain. While the general architecture is similar to other solved structures, the PFV RNase H generally has longer helices and strands compared to other RNase H domains.

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 followed by a basic protrusion/loop, is a regular feature of E.coli RNase H, human RNase H and MoMLV RNase H which are important for binding substrates and activity. 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 translocation of the thumb in that direction (**Fig. 21**).

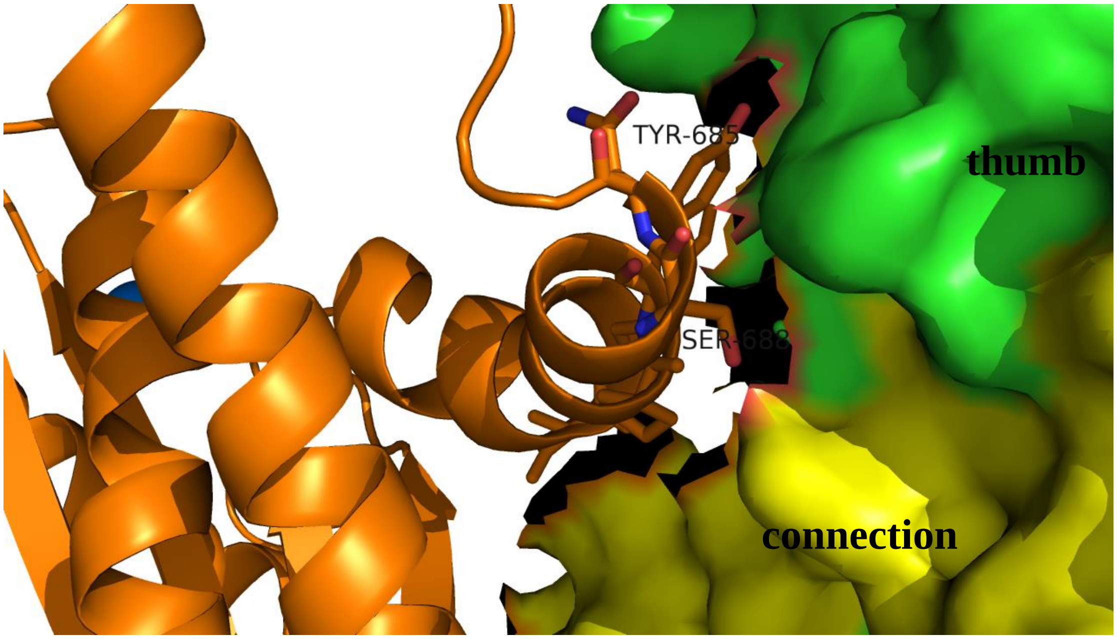


Figure 21: Surface rendering of the interface between PR-RT thumb/connection subdomains and potential interactions RNase H subdomain helix C.

A calcium ion used in the crystallization buffer of the WT PFV PR-RT is observed chelated to the side chains of residues Asp599, Glu646 and Asp669 together with carbonyl of Gly600 which define the RNase H active site (**Fig. 22**). These residues even though situated at the center of the domain, are exposed to the solvent and point away from the putative nucleic acid-binding cleft. This further suggests that this domain must undergo some rotation to bring the active site into register with the nucleic acid substrate which it cleaves when it binds it.

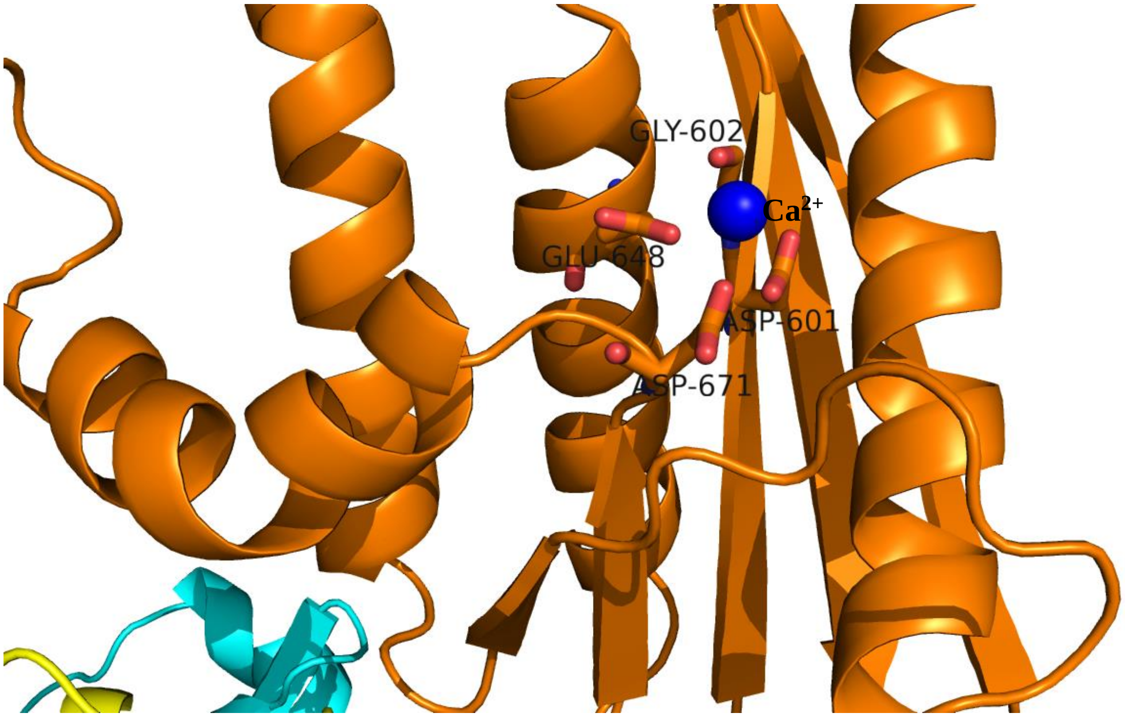


Figure 22: Active site of PFV RNase H subdomain residues chelating Ca^{2+} (Blue sphere).

2.12 Discussions

Retroviruses encode an aspartic PR and an RT in their genome. These proteins are responsible for the proteolytic processing of polyprotein precursors as well as reverse transcription of their RNA genome into a dsDNA that can subsequently be inserted into the host chromosome. In this study, the crystal structure of the PFV PR-RT fusion precursor polyprotein which is responsible for carrying out the afore mentioned processes is presented. The crystal structure of the PFV PR-RT shows a PR monomer and an RT monomer. The PR is folded similarly to a single unit of the mature dimeric enzyme.

The N- and C-terminal residues involved in the dimerization interface in the mature aspartic acid protease remain unstructured in these structures consistent with their flexibility observed in crystal structures of the mature enzymes. *In silico* placement of two PR-RT molecules related by two-fold symmetry bring two PRs 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.

On the other hand, the polymerase is in a configuration that cannot bind nucleic acid. Reshuffling of the subdomains as seen in RTs bound to nucleic acids is observed with the structure bearing a rather close resemblance to the p51 subdomain of HIV-1 RT. Hence a significant rearrangement of the subdomains is envisaged when PFV PR-RT binds nucleic acid. With the long linkers connecting the palm and the thumb as well as the connection and RNase H, such movements are easily fathomable.

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 and p51 heterodimer, Wang *et al 1994*, suggested that monomeric forms of these enzymes would adopt a more compact "p51-like" conformation to compensate for the cost of exposing hydrophobic residues. Zheng *et al 2015,* using NMR spectroscopy confirmed this initial observation by Wang *et al 1994*, 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 this structure 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 suggest that the most predominant population in solution is the "p51-like" conformation. Transient sampling of catalytically competent p66 open conformer exposes 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. This structure therefore captures that all-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 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. 23**) suggests that the connection domain (the "gate keeper") currently occupies the nucleic acid-binding cleft. For nucleic acid binding therefore, it is envisaged that the "gate keeper" must undergo close to 90º rotation followed by a translation perpendicular to helical axis of helix L of the connection (**Fig. 24 and 25**). 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. The origin of this conformational change is likely that it is sampled in solution and then the binding of nucleic acid substrates traps and stabilizes it. 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.

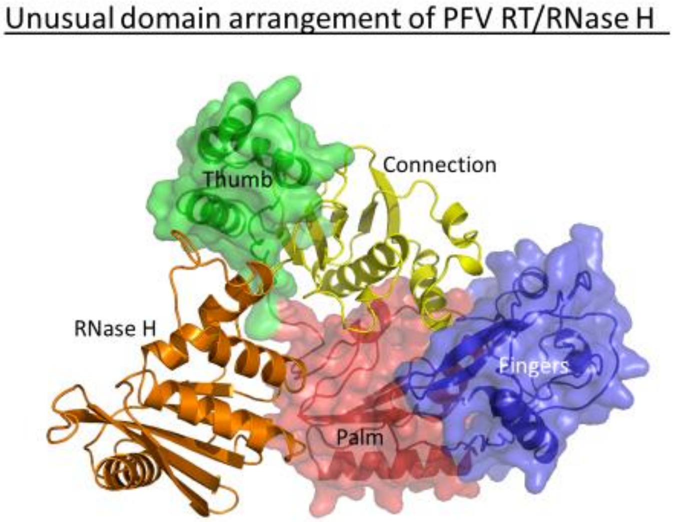

Unusual domain arrangement of PFV RT/RNase H

Figure 23: Structure of PFV RT with surface rendering of fingers-palm-thumb.

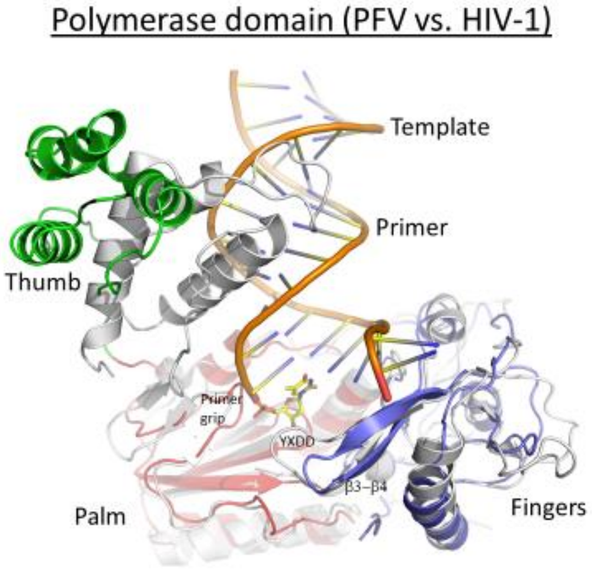

Polymerase domain (PFV vs. HIV-1)

Figure 24: Structure of PFV RT (grey) superimposed on HIV RT-DNA complex (PDB ID 1N6Q).

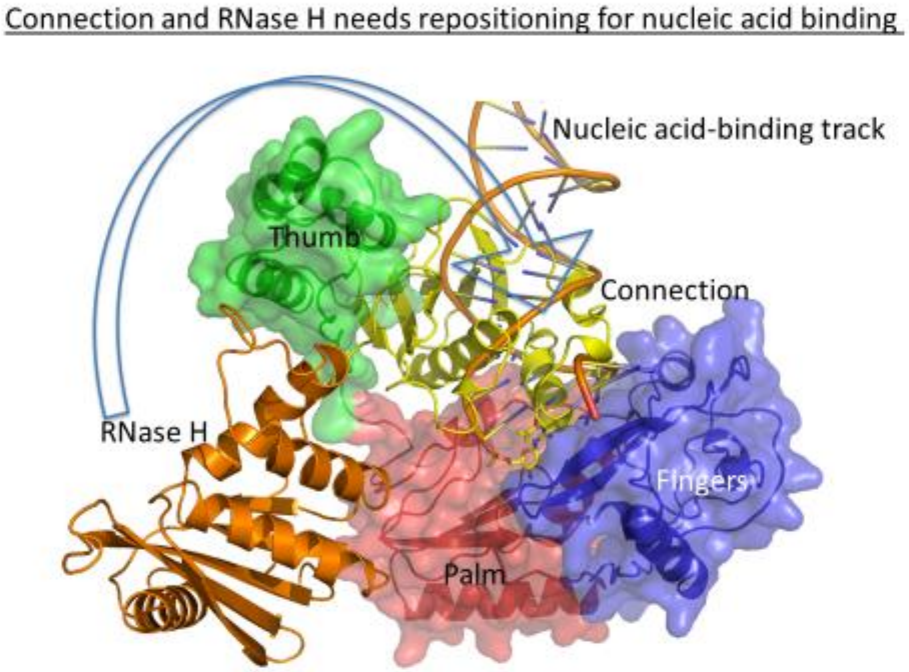

Connection and RNase H needs repositioning for nucleic acid binding
Nucleic acid-binding track
Thumb
Connection
RNase H
Fingers
Palm

Figure 25: Structure of PFV RT superimposed on HIV RT-DNA complex (PDB ID 1N6Q). Arrow shows direction of domain movement required to open the nucleic acid.

Since the 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 predominantly helps to decouple the two functions and ensures that they can be regulated from a temporal and spatial perspective. Such a high energy barrier may easily be compensated for by the extensive interactions with nucleic acids. This structure does not only capture the imaginative evolutionary genius of retroviruses beginning with its most primitive member, it offers significant insight into the evolutionary landscape of these largely pathogenic enemies.

1.13 Conclusions

A systematic study of the PR-RT precursor polyprotein of PFV has been carried out. The PR-RT is a functional reverse transcriptase and a protease. The crystal structure of this machinery was solved initially from a mutant enzyme before conditions optimized for the crystallization of the mutant enabled the structure of the WT to be solved. The mutations in the enzyme which resulted in multiple folds of purified proteins compared to the WT and crystallized much quickly were identified using the instability indices computation on the ExPASy ProtParam webserver. This demonstrates the possibility of identifying proteolytically vulnerable regions in a protein by simply querying the primary sequence using the instability index as an output for verifying how a particular amino acid is preferred in its position based on its neighbors.

Each of the domains in the protein as revealed by the crystal structure is independently folded but spatially arranged differently when compared to the mature enzymes from HIV and other retroviruses. This means that the precursor polyprotein contains independently folded domains as seen in this structure but not necessarily arranged similar to the mature enzymes in 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. Because these structures are highly conserved in retroviruses, it is anticipated that such structures from other retroviral genera would present new avenues for inhibiting the virus using inhibitors that target the precursor polyproteins rather than the mature enzymes and making the precursor polyprotein a legitimate target for drug discovery.

For the PFV 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 limited NMR spectroscopy. The compact fold 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. This structure 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.

2.14 Materials and Methods

2.14.1 Cloning, expression and purification

PFV PR-RT with a protease null mutation (D24A) WT, was cloned into a pET28a vector with an N-terminal His-tag and an HRV14 3C protease cleavage site. Quick Change site-directed mutagenesis (Liu and 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 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 5000g for 25 minutes. Cell pellet was re-suspended in 100 ml of lysis buffer (50 mM phosphate pH 7.6, 300 mM NaCl, 1 mM TCEP, 20 mM imidazole, 10% glycerol, 0.5 mM EDTA, 1 mM PMSF). Cells were sonicated for 10 minutes and spun down at 18000g 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 His6-tag cleavage with HRV14 3C 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.

2.14.2 Selenomethionine 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.

2.14.3 Crystallization

Purified protein at a concentration of at least 20 mg/mL was mixed in a 1:1 ratio with precipitant in a sitting drop vapor diffusion tray using a micro bridge in drop volumes of 1 -10 µL or 1-3 µL drop size using the hanging drop method. A reservoir volume of 500 µL -1.0 mL was used. Crystals were harvested using MiTeGen MicroLoops or MicroMesh mounts and cryo-protected in reservoir supplemented with 25% glycerol before flash freezing.

214.4 Data collection

X-ray diffraction data were collected at the Cornell High Energy Synchrotron Source (CHESS), the Advanced Photon Source (APS), and the Stanford Synchrotron Radiation Light Source (SSRL). The PFV PR-RT crystallized with C2 space group symmetry and cell parameters a, b, c: 240.2, 52.83, 74.95 Å and α, β, γ: 90, 100.54, 90°. Data processing and scaling were done using iMosFlm and Aimless, the structure was solved using Crank2, and the refinement was carried using REFMAC all in the CCP4i suite (Potterton et al., 2018). Model building was done in Coot. The model of the SeMet-labeled CSH mutant was successfully used for molecular replacement using the WT data which was found to be twinned (Xtriage) in Phaser (Phenix) and refined to 2.9 Å resolution using the twin law -h, -k, l to R-work/R-free of 22.9/24.2% in REFMAC (Murshudov et al., 1997).

2.14.5 Extension and processivity assay

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

CHAPTER THREE: Biophysical characterization of PFV PR-RT dimers and nucleic acid complexes

Synopsis

The structure of PFV PR-RT offered insight into the arrangement of the individual domains of the PR-RT but even more importantly, it shed unprecedented light on the extreme conformational flexibility of the retroviral RT. However, the structure raises key questions whose answers are not immediately apparent from the structure. The PR-RT functions as a protease as well as a reverse transcriptase and therefore it is worth asking how the PR dimerizes within the context of the PR-RT since the dimer is the functional PR entity. Also, what are the conformational changes that occur when the RT binds nucleic acid?

To answer these questions, the PR active site S25C mutation generated has enabled a disulfide crosslinked intermediate of the PFV PR-RT dimer through the PR active site for structural and biophysical studies. This cross-linking is possible only when the PR is dimeric with the active site residues in the right conformation as envisioned from numerous structures of HIV-1 PR. This is the first demonstration of the disulfide cross-linking through the active site of any aspartic protease.

Engineered site-specific cysteine at position Q391 in the thumb of the PFV RT has enabled disulfide crosslinking PFV PR-RT to a DNA template-primer using the same chemistry developed for HIV-1 RT. The cross-linking of the RT to nucleic acid traps the nucleic acid permanently in the nucleic acid-binding cleft of the RT which may increase the chances of obtaining crystals of this complex. Studies of the PR-RT binding to a DNA aptamer initially developed for HIV-1 RT with some modification has also been carried out with the hope of obtaining structures of the PR-RT dimer and the nucleic acid bound structures. Crystals of the PR-RT dimer and nucleic acid complexes have been obtained and are currently being optimized for X-ray diffraction experiments.

3.1 Introduction

The synthesis of polyprotein precursors which are subsequently proteolytically processed into mature functional enzymes and structural proteins is well conserved in all the retroviral genera. There are however differences in the method of production of these polyprotein precursors. In orthoretroviruses such as HIV, the structural proteins and the enzymatic proteins (Pol) are synthesized as one long polypeptide chain referred to as Gag-Pol through a -1 ribosomal frameshift at the 3'-end of the *gag* gene to enable a readthrough of the stop codon at the end of the gene. This frameshift is highly well regulated by sequence elements in the genome ensuring a narrow range of Gag to Gag-Pol ratios of usually between 5-10% for optimal viral infectivity (Cherry et al., 1998a, Low et al., 2014, Mathew et al., 2015).

On the other hand, the prototype foamy virus (PFV) makes separate RNA transcripts for the Gag and Pol polyprotein precursors, which are therefore translated separately (Jordan et al., 1996, Hartl et al., 2010a). This mode of protein production leads also to different mechanisms of viral assembly and budding. Hence the proteolytic processing of these precursor polyproteins which is an obligate process for infectivity of progeny virions occurs to different extents as well in these viruses using different ways of making precursor polyproteins.

The observation that structure and function of biological macromolecules have been conserved throughout the history of evolution has been very useful for structural biology. This has enabled the inference of structural and functional information for macromolecules in distantly related species and genera where due to sequence divergence, have different biophysical behavior in solution suitable for structural studies or not. This concept has been well demonstrated in the retroviral literature with the numerous structures from different subfamilies of the *Retroviridae*.

To obtain structural details of polyprotein processing, PFV was chosen as a surrogate for studying HIV and other retroviruses. This selection was in part the result of the observation that the PFV integrase was more soluble and less aggregated in solution even at a high concentration compared to other retroviral integrases which enabled valuable structural insight into retroviral intasome assembly and strand transfer reactions as well as HIV integrase inhibitor binding mechanisms (Hare et al., 2010). The similarity of the crystal structure of the PFV PR-RT to the HIV PR monomer and the domains in HIV RT further reinforced this notion of suitability of this system for these studies based on the conserved structural elements observed.

3.2 Retroviral Proteases: Structure and Function

Retroviral PRs are highly conserved aspartic acid proteinases with active the site catalytic triad of Asp-Thr-Gly (DTG) residues typified by HIV or Asp-Ser-Gly (DSG), exemplified by PFV (Wu et al., 1998, Tozser et al., 2000, Tozser, 2010). These enzymes function as homodimers with each monomer contributing a triad to the catalytic activity (**Fig. 11**). These amino acids together with others engage in a network of hydrogen bonding with highly ordered water molecules and the backbones of peptides earmarked for cleavage within the active site of the enzymes (Ozen et al., 2011, Mittal et al., 2013). The presence of serine instead of threonine has functional consequences for the PRs with the latter leading to a more compact dimeric structure and more stability than the former. The stability of these enzymes correlates with enzymatic activity as well (Tozser, 2010).

During maturation of the polyprotein precursors, retroviral proteases excise themselves from their respective precursor proteins in a temporally and spatially well-regulated fashion except in FVs where the PR remains attached to the RT as a functional unit (Roy and Linial, 2007, Wohrl, 2019). With dimerization obligately coupled to enzymatic activity, the retroviral protease embedded in the Pol polyprotein precursor either as a Gag-Pol or simply Pol must dimerize to activate the embedded protease to initiate the cascade of events that lead to the production of functionally relevant polypeptides. The critical roles played by PRs in the life cycle of retroviruses makes them highly attractive targets for drug development. The mechanism(s) underlying the protease activation mechanism has been elusive despite numerous structures of monomeric and dimeric PRs with and without substrates.

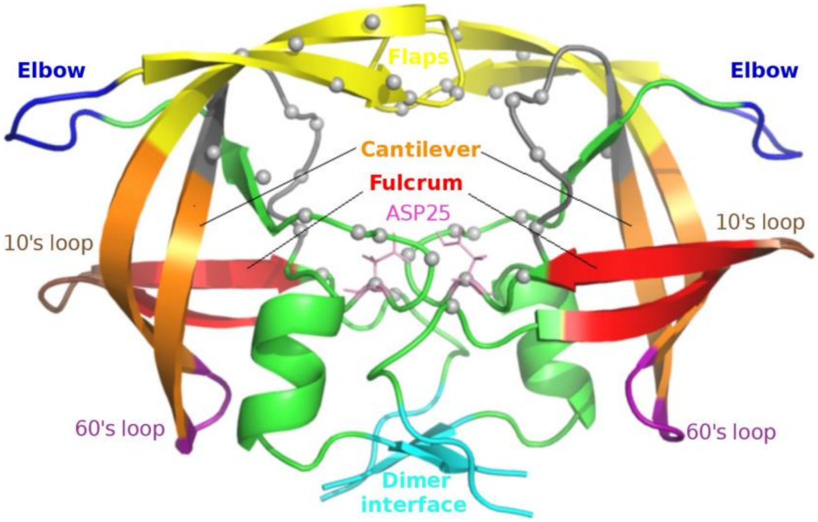


Figure 11: Architecture of the HIV-1 PR with annotations reproduced from Sheik Amamuddy *et al*., 2018.

3.3 PFV PR activation and regulation mechanism

The FV PR in the PR-RT enzyme or when expressed separately does not exhibit any enzymatic activity *in vitro* under physiologically relevant conditions. This behavior is unique to Spuma retroviruses with all the others predominantly dimeric and catalytically very active under similar conditions (Spannaus et al., 2012, Schneider et al., 2014). This observation raises two hypotheses (1) either the PR is functionally inactive in PR-RT or (2) there is an exogenous mechanism of activating the PR to ensure it does not cleave prematurely or at off-target sites. Evidence for each hypothesis has been garnered in the literature. The first school of thought suggests that the PR domain of the PR-RT is enzymatically inactive in the absence of the integrase, which has been shown to enhance the PR activity. This conclusion follows a study where the IN domain of the PFV Pol was systematically truncated while measuring PR activity in virus particles. While PR activity was completely abolished, in constructs of the PFV Pol lacking IN, replacement of the IN with a leucine zipper dimerization motif was found to restore PR activity (Lee et al., 2011, Spannaus et al., 2012).

The other school of thought suggests that the lack of proteolytic activity of the PFV PR-RT is not an intrinsic phenomenon but instead a regulatory mechanism of ensuring that activation does not lead to spurious off-target cleavage events that may harm successful replication of the virus. They posit that an exogenous agent is required to activate the PR. A protease-activating RNA motif (PARM) whose sequence overlaps with that of the open reading frame of the *pol* region as well as a cis-acting sequence II (CAS II) (**Fig. 26, 27**) of the viral RNA genome was subsequently identified as the agent responsible for stimulating PR activity under physiological conditions *in vitro* and *in vivo* (Hartl et al., 2011).

A selective 2’-hydroxyl acylation analyzed by primer extension (SHAPE) analysis of this sequence showed a well-defined secondary structure which may aid its function. It is therefore fathomable that the truncation experiments earlier carried out (Lee et al., 2011, Hartl et al., 2011) could have deleted part of this sequence thereby disrupting the secondary structure necessary for stimulating the PR activity since this sequence overlaps with that of the IN-reading frame. It was also shown that PR activity could be stimulated by a non-physiologic buffer containing at least 2.5 M NaCl (Hartl et al., 2011, Spannaus et al., 2012).

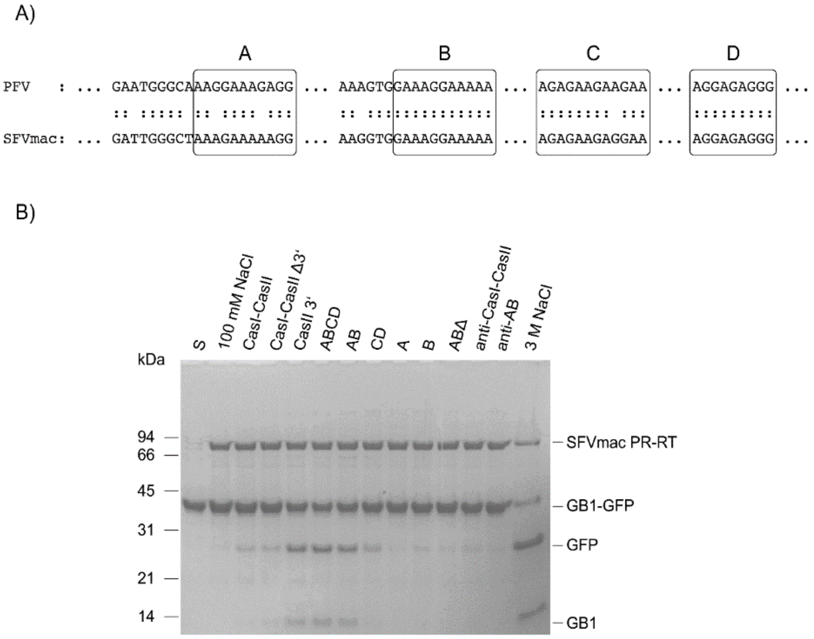


Figure 26: Sequence elements of PARM found to stimulate PFV PR activity reproduced from Hartl et al., 2011 with the consent of ASM.

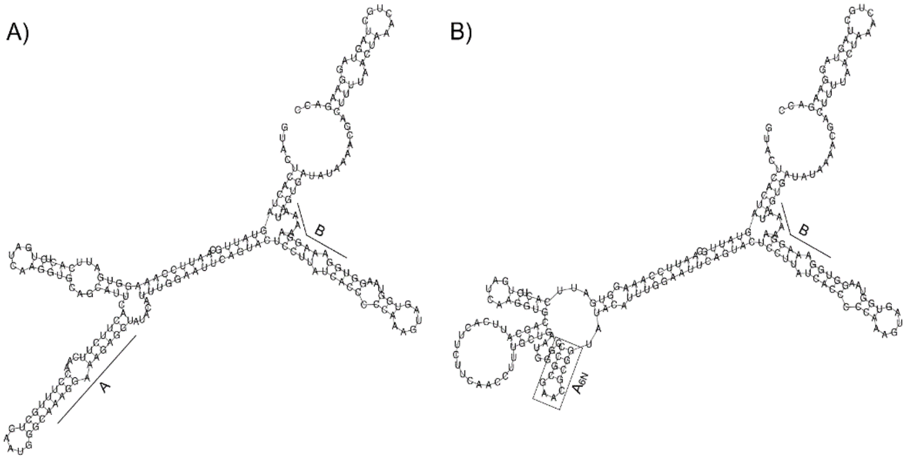


Figure 27: Predicted secondary structure of PARM found to stimulate PFV PR activity reproduced from Hartl et al., 2011 with the consent of ASM.

3.4 Conformational Maturation of Retroviral RTs

The structural asymmetry seen in heterodimeric RTs such as HIV RT and Ty3/Gypsy-like retrotransposon RT (Nowak et al., 2014) raises an important question about the mechanism(s) by which such asymmetry is achieved. Such asymmetric structures of the same primary sequence suggest extreme conformational plasticity. Since all RTs must exist as monomers at some point in their life cycle (at least a portion of immature polyprotein precursors), it is relevant to ask also the conformational preference of the monomeric state of RTs before heterodimerization or nucleic acid binding which for functionally monomeric RTs like PFV and XMRV/MoMLV triggers conformational rearrangement. Evidence in support of such a notion is gleaned from the crystal structure of PFV PR-RT (chapter two).

For the heterodimeric RTs, a single polypeptide folds into two different conformers which associate tightly to carry out enzymatic functions. It is interesting to note that the enzymatic activity resides in only one of such conformers which is usually described as having an extended or open conformation (Zheng et al., 2015) and providing a hollow cleft within which nucleic acids can bind as well as binding grooves for dNTPs to allow for subsequent polymerization to be carried out. The more compact or closed conformer appears to offer structural support to the open conformer (Kohlstaedt et al., 1992, Wang et al., 1994, Ding et al., 1994).

While in HIV a single ribonuclease domain (RNase H) domain of the compact conformer is cleaved off by the virally encoded PR, Ty3 retrotransposon RTs retain all subdomains in each of the conformers. In Ty3 RT however, the connection subdomain extant in other retroviral RTs is missing. The two asymmetric conformers in HIV-1 RT referred to as p66 and p51 reflecting their molecular weights respectively associate with an equilibrium dissociation constant, Kd of 0.4nM-4.0 µM underscoring the strong nature of their interactions mediated largely by sequestering hydrophobic residues across the interface between the two subunits (Venezia et al., 2006, Venezia et al., 2009).

Buried surface area analysis of the structure of HIV-1 RT shows that not only is the p51 subunit less solvent accessible than the p66 subunit but even more importantly, ~4500 Å^2 of surface area is buried upon heterodimerization (Wang et al., 1994). On the basis therefore of the entropic gain during the structural transition from a compact p51 conformation to a heterodimer, it has been postulated that the monomeric form of these proteins would prefer the more compact ‘p51-like’ conformation to decrease buried surface area. It is a reasonable expectation from such analysis that the monomeric RTs would also prefer such a compact conformation in the absence of nucleic acid. The existence of such compact structures in solution have been partially detected using various NMR techniques but the details of the intra-subunit interactions and secondary structural details have not been possible so far for HIV RT (Zheng et al., 2014, Zheng et al., 2015).

The crystal structure of the PFV PR-RT revealed full atomic detail of the compact structure of the monomeric RTs which is consistent with the initial hypothesis. While the subdomain orientations and interactions are not the same as seen in HIV p51, the structure is in a p51-like conformation with slightly different inter-subunit contacts (see chapter two for details). A structure of the PFV PR-RT bound to a nucleic acid substrate would therefore offer a rare insight into this conformational transition.

3.5 Aims of the study

The increasing public health burden that retroviruses continue to create in animals require new avenues for therapeutic interventions. Undoubtedly understanding key processes in retroviral biology such as polyprotein processing and conformational maturation of RTs offers new avenues for such endeavors. This has become even more important as resistance to existing drugs that treat retroviral infections continues to spread (Sarafianos et al., 1999, Ali et al., 2010, Vijayan et al., 2014).

Using a combination of biochemistry and molecular biology tools such as cloning, protein-nucleic acid cross-linking, media engineering and protein purification techniques, this study sought to generate key reagents for unraveling the processes outlined above. Through sequence and structural alignments, Q391 was identified as the equivalent of HIV-1 RT Q258. Generation of Q391C mutant of PFV PR-RT using oligonucleotide based site-directed mutagenesis followed by protein expression and purification enabled the cross-linking of a DNA T-P previously used for cross-linking studies of HIV-1 RT (Sarafianos et al., 2003) to the PFV protein for structural and biophysical studies.

Concurrently, crystallization trials of a DNA aptamer previously selected against HIV-1 RT (Miller et al., 2016) with some modification resulted in the crystal hits in commercially available crystallization screens with optimization currently being undertaken. A PR active site substitution of S25C also generated stable disulfide cross-linked dimers of the PFV PR-RT through the PR active site. Crystallization hits have been obtained for this dimer and optimization efforts are currently underway. These complexes have been characterized using biophysical methods such as, gel filtration and dynamic light scattering (DLS).

3.6 RT -Nucleic Acid complexes

RTs possess three enzymatic activities-RNA dependent DNA polymerase activity, DNA-dependent DNA polymerase activity and a ribonuclease activity which work in concert to ensure successful replication of retroviral genome (Wang et al., 1994, Tian et al., 2018). RTs cannot initiate nucleic acid synthesis *de novo* and so utilize the host tRNAs with complementary sequences to the primer binding sites (PBS) in the 5’-UTRs of their genomes to initiate reverse transcription of their positive sense (+ sense) genomic RNA into a complementary DNA (- sense) with concomitant degradation of their RNA genomes by the ribonuclease H (RNase H) domain of their respective RTs (Larsen et al., 2018, Das et al., 2019). Plus-strand DNA synthesis is initiated by a piece of RNA called polypurine tract (PPT) which is initially resistant to RNase H degradation during – sense DNA synthesis. The completion of + sense DNA synthesis is carried out bidirectionally using primers generated after the removal of the PPT sequences (Sarafianos et al., 2001)(**Fig. 28**).

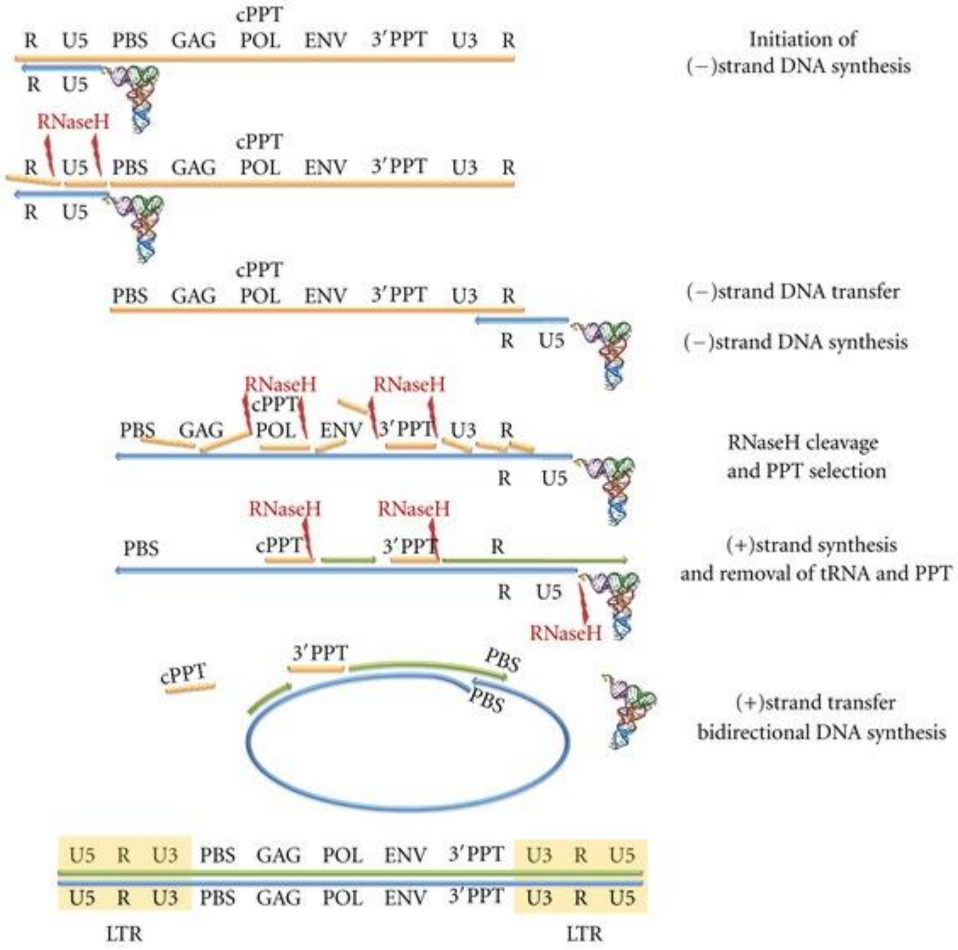


Figure 28: Schematic representation of the various steps involved in synthesis of dsDNA copy from ssRNA genome reproduced from Esposito et al., 2012.

For HIV-1 RT, structures of all of the key priming events which capture protein-nucleic acid interactions have been solved to high resolution using co-crystallization, protein-nucleic acid cross-linking chemistry as well as DNA aptamers that bind with ~15 pM affinity to RT. Because of the differences in the geometry and conformation of dsRNA and dsDNA and dsRNA/DNA, polymerases interact differently with these substrates which are reflected in the differences in their incorporation efficiencies. The initiation of reverse transcription by HIV-1 for example which involves RT binding to dsRNA is known to be very slow compared to elongation where the binding is predominantly to RNA/DNA and dsDNA (Miller et al., 2016, Larsen et al., 2018, Das et al., 2019).

Dissociation of PFV PR-RT from a dsRNA/DNA hybrid and dsDNA was measured and compared to that of HIV-1 RT to obtain insight into the binding kinetics for these substrates. While the PFV PR-RT bound dsDNA and dissociated much more slowly compared to when bound to dsRNA/DNA, HIV-RT dissociated much slower from a dsRNA/DNA than the dsDNA (**Fig. 29**). 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 or different residues may come into play depending on which substrate is bound or even both. These interactions are relevant for understanding the differences in processivity between the enzymes outlined in chapter two.

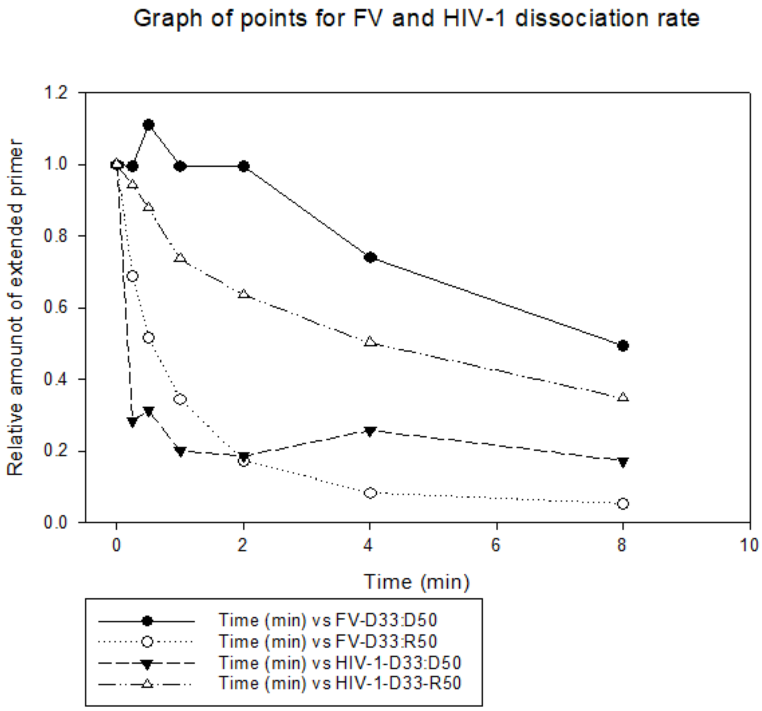


Figure 29: Dissociation rate of PFV PR-RT and HIV-1 RT from an RNA/DNA hybrid (D33-R50) and dsDNA (D33:D50)-Jeffery DeStefano, University of Maryland, MD

3.7 Crystallization Trials

In an effort to obtain structures of PFV PR-RT-nucleic acid complexes which may help explain many of the questions outlined above, co-crystallization experiments of the PFV PR-RT WT and CSH mutant with various dsDNA template-primers (T-Ps) derived from the primer binding site sequences of PFV genome of varying length, were carried out using commercially available crystallization screens such as Natrix, Index, Crystal Screen, and PegRx (Hampton) at 4 and 20 ºC. Unfortunately, no crystal hit was obtained after several trials using these nucleic acids.

3.8 Cross-linking studies

Binding Kd measurements, dissociation rates and extension assays for RTs have clearly demonstrated that the off-rates for T-P binding for RTs while variable, is generally poor. Such poor binding correlate highly with crystallizability of such complexes since high off-rates create a highly heterogenous pool of complexes which makes it difficult to organize them into a single lattice during crystallization.

In view of this observation HIV-1 RT T-P cross-linking chemistry had been developed for crystallization studies. This chemistry has relied on site-specific cysteine mutation in the protein (I63C and Q258C) and a modification at specific base(s) in the nucleic acid which enabled disulfide cross-linking of the nucleic acid and protein when mixed together under the right conditions. Based on sequence alignments (see chapter two), the equivalent mutations were carried out in PFV PR-RT to take advantage of this chemistry. Due to limited reagents for the I63C cross-linking, it was not further pursued.

Processive synthesis by RTs, like with other polymerases, involves three steps. In the presence of magnesium and deoxynucleoside triphosphate (dNTP), the RT initially binds the T-P to form a binary priming site (P-site) complex where the complex is poised for catalysis. When the dNTP and magnesium binds to the RT, conformational change in the fingers domain leads to a nucleophilic attack by the 3’-OH group of the primer on the α-phosphate of the NTP leading to phosphodiester bond formation and release of a pyrophosphate product (N-site complex). The RT subsequently translocates one nucleotide leaving the dNTP pocket open and ready for another round of incorporation.

Determination of high-resolution structures that offered insight into this cycle of events has relied on covalent trapping of the protein-nucleic acid complex at various stages of the cycle. The Q258C cross-linking chemistry was designed such that the modified base in the primer is out of register to be able to cross-link with the 258C for the binary, RT-dsDNA complex, the ternary, RT-dsDNA plus in-coming dNTP or post-incorporation but pre-translocated complex. The covalent trapping of the DNA is only possible after a single round of nucleotide incorporation followed by translocation when the modified base is positioned in register with the cysteine. The sequence of DNA used is given below where the asterisk shows the thioalkyl modified G (N^2) in the primer (Sarafianos et al., 2003, Das et al., 2012, Das et al., 2019).

```
27mer T DNA:   3’ G TGTCAGGGACAAGCCCGCGGATCG TA   5’
20mer P DNA:   5’   ACAGTCCCTGTTCGGGCGCC          3’
                                   *
```

By using a dideoxy modified dNTP, the enzyme is restricted to a single round of polymerization making it possible for obtaining both P-site and N-site complexes depending on the divalent metal used. The covalently trapped complex is further purified using tandem nickel-heparin columns or only a heparin column. The binding of the DNA to the protein reduces the number of available positive charges on the protein decreasing its affinity to heparin compared to the non-cross-linked protein. This allows separation of the cross-linked and non-cross-linked proteins using a salt gradient. A representative chromatogram from the heparin gradient and the accompanying gel (non-reducing 4-20% acrylamide gel) for a cross-linking experiment is given below (**Fig. 30**).

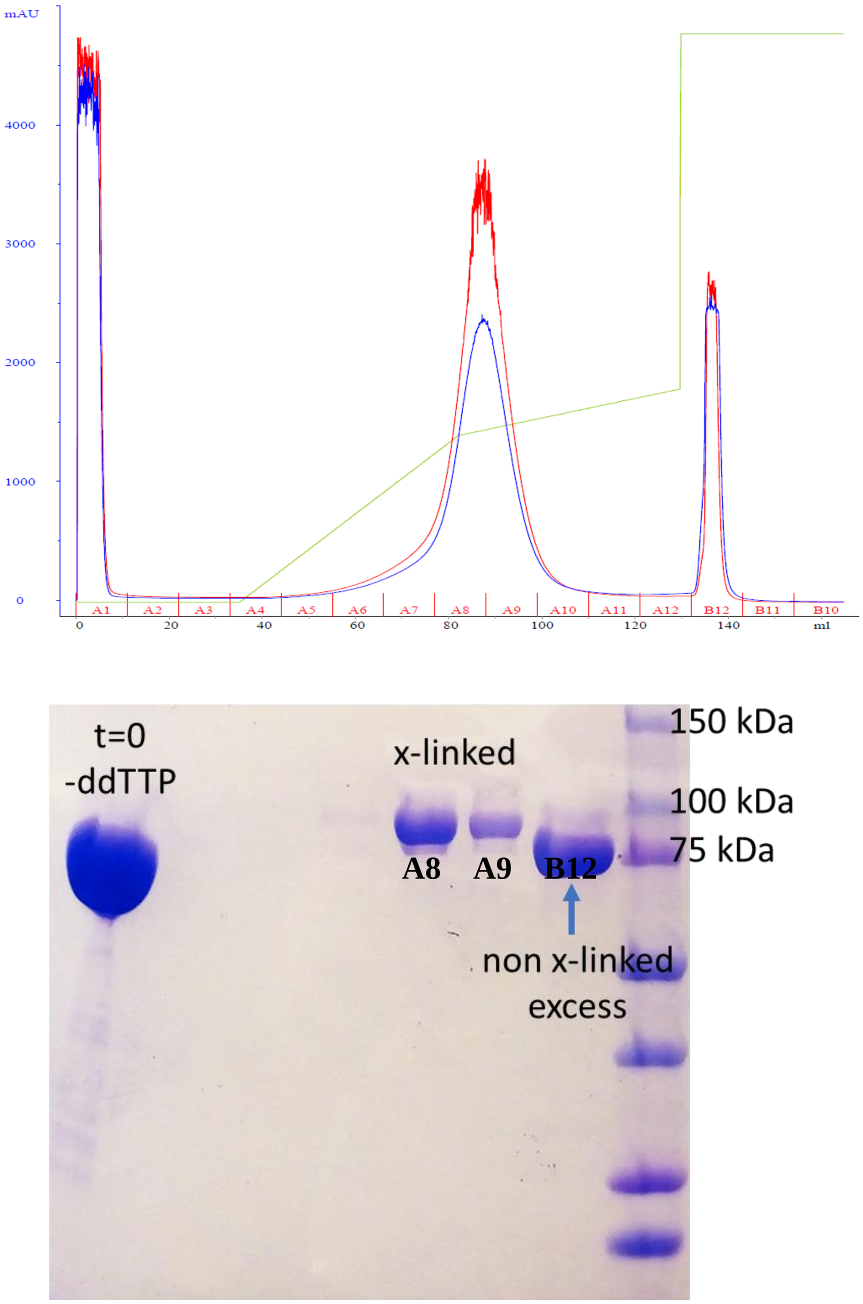


Figure 30: Heparin chromatography trace of PR-RT-dsDNA cross-linked complex with its accompanying SDS-PAGE gel

Crystallization trials of the PFV PR-RT cross-linked to T-P DNA using commercially available screens have so far been unsuccessful. The successful cross-linking of a T-P primer optimized for HIV-1 RT to PFV at an equivalent position suggests that the conformational states of nucleic acid bound molecules of the two proteins are similar. Furthermore, the distance between the polymerase active site of HIV-1 RT to the site of the cross-linking residue and that of PFV are identical as well. This is not withstanding the fact that PFV is a monomeric RT while HIV-1 is heterodimeric. This further reinforces the notion that structural details of the PR-RT nucleic acid complex will be useful for understanding the dynamics of conformational maturation in retroviral RTs. Based on this knowledge, a structural model of PFV bound to a DNA T-P was generated (**Fig. 31**). It remains to be seen if this model holds true.

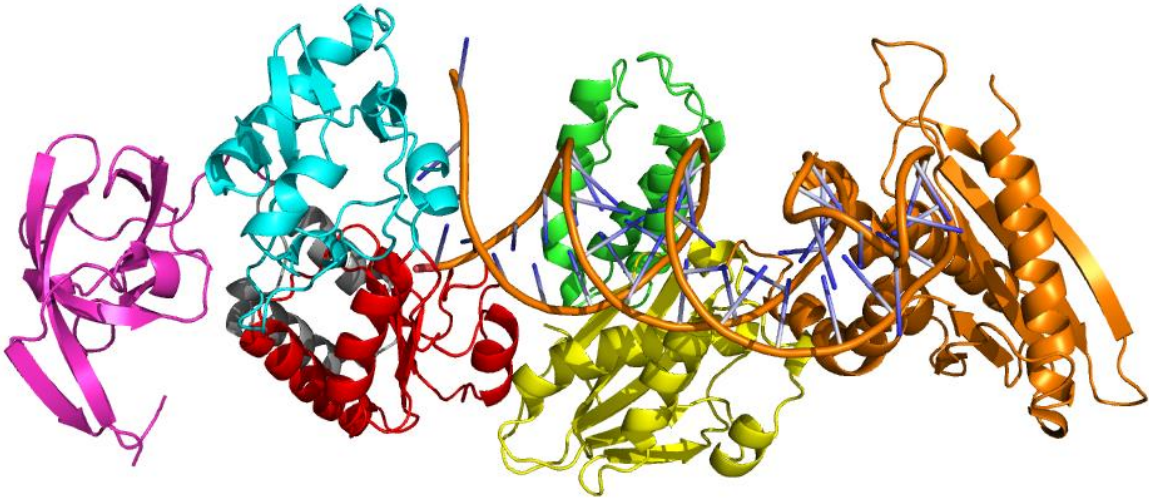


Figure 31: A model of PFV PR-RT bound to dsDNA generated by aligning individual domains in PFV PR-RT to HIV-1 RT-48mer DNA aptamer complex- Ruiz-Figueras and Arnold (unpublished).

3.9 Crystallization screening with DNA Aptamers (In collaboration with Jeffery DeStefano, University of Maryland, MD)

Oligonucleotide aptamers which are artificially engineered short stretches of nucleic acids selected through *in vitro* selection or systematic evolution of ligands by exponential enrichment (SELEX) with very high binding affinity to their targets have found widespread use in biotechnology (Stoltenburg et al., 2007, Zhuo et al., 2017). A 38NT-2’,4’-O-methyl modified DNA aptamer in a T-P configuration for HIV-1 RT obtained through SELEX with ~15 pM binding affinity enabled the crystallization of HIV-1 RT which diffracted X-rays to 2.3 Å resolution which was the highest for HIV-1 RT nucleic acid complex (Miller et al., 2016). Binding Kd measurements showed that this aptamer binds with ~300 pM affinity to PFV PR-RT which, even though ~20-fold lower than the HIV-1 RT binding, is still very tight. Crystallization screening experiments were therefore carried out with this aptamer following an unsuccessful crystallization campaign with the T-P cross-linked protein using commercially available kits. Unfortunately, no hits were obtained for this aptamer after several trials.

Other aptamers were designed based on this template in the hope of improving binding and crystallization success. The design took into consideration variability of sequences, length of duplex regions and overhang and the nature and sequence of hairpin. The core of 15 nucleotide base pairs (15 bp) present in the 38NT-2,4-aptamer was kept constant to ensure that binding affinity for the polymerase region which is covered by these sequences is maintained at the very least. Due to the long linker (30 residues) between the connection domain and the RNase H domain of PFV RT (567-596 see chapter two), it was hypothesized that the RNase H domain in PFV will be more mobile.

To reduce this mobility which could be inimical to crystallization, the length of the duplex region of the aptamer was increased to 20-23 bp. These sequences would span the RNase H region since the nucleic acid binding cleft length in PFV RT appears similar to that of HIV-1 RT based on the cross-linking studies. These longer sequences should provide a framework for the RNase H domain to latch onto thus decreasing its mobility and perhaps aiding crystallization unlike the 38NT-2,4-Apt which only spanned up to the RNase H primer grip in HIV-1 RT.

Also, beyond the first two nucleotides in the template overhang of HIV-1 RT-nucleic acid structures that have well defined electron density, all other nucleotides in the overhang do not show any electron density in the crystal structure of HIV-1 RT-38NT-2,4-aptamer structure suggesting that they likely are disordered. The length of this overhang was also varied (2-3) to further reduce any disordered residues/nucleotides that may hamper crystallization. A variation in the type and sequence of the hairpin was also tested since nucleic acid mediated crystal packing is a regular feature of protein-nucleic acid complexes. The regular three nucleotide, TTT, (3T) hairpin as well as four nucleotide hairpin (tetraloop) with TTAA (T2A2) and TTTG (T3G) configuration were made (**see Table 1**). The binding affinities of these aptamers were measured to be between 162-362 pM to the WT PR-RT.

The RNase H cleavage experiment of the PR-RT (Chapter two) showed that apart from the primary cut observed at 18-19 bp away from the polymerase active, and a secondary cut around 12-13 nucleotides away from the polymerase active site believed be directed by the 5’-end of the T-P RNA/DNA hybrid instead of the 3’-end (Gerondelis et al., 1999, Schultz et al., 2006), which is characteristic of all retrovirus ribonucleases including HIV-1, a tertiary cut around 14-15 bp away from the 3’-end is observed in PFV as well. It is not clear how this cut is produced but it is a further indication of higher mobility of the RNase H domain of this protein. Such mobility is expected to be deleterious to crystallization success.

To further anchor the RNase H domain to the nucleic acid substrate when bound, the catalytic residues D599, D669 and E646 were mutated. The aspartates were mutated to asparagine, and the glutamate to a lysine. These residues come in direct contact with the nucleic acid substrate, engages it and cleaves it using these residues. Since these are highly acidic residues, they exert repulsive forces on the nucleic acid backbone especially in the absence of divalent metal cations.

It was therefore surmised that neutralizing these charges should reduce such repulsive forces and enhance binding. The replacement of the glutamate with a lysine which is positively charged is expected to increase this binding affinity even further since the extra carbon in lysine compared to glutamate would be expected to bring this residue closer to the nucleic acid. Consistent with this observation, this RNase H mutant construct binds to these aptamers with Kd values of 50-80 pM which is 3-6-fold better than the WT enzyme. Crystal screening experiments were therefore carried out with these aptamers complexed with the WT protein and the RNase H mutant using commercially available crystallization screening solutions.

Table 1: Different aptamers designed for crystallization with PFV PR-RT

Aptamer sequence	Kd in 2 mM $MgCl_2$
38NT-2',4'-O-methyl Aptamer (38NT-2,4-Apt) **TAATA**CmCCmCCCCTTCGGTGC**TTTGCACCGAAGGGGGGG-3'**	306 pM
1. TA CmCCmCCCCTTCGGTGCGTGGG**TTTCCCACGCACCGAAGGGGGGG-3'**	362 pM
2. TA CmCCmCCCCTTCGGTGCCACCC**TTTGGGTGGCACCGAAGGGGGGG-3'**	175 pM
3. GAT CmCCmCCCCTTCGGTGCGTGGG**TTAACCCACGCACCGAAGGGGGGG-3'**	162 pM
4. TA CmCCmCCCCTTCGGTGCGTGGG**TTGTCCCACGCACCGAAGGGGGGG-3'**	287 pM
5. GAT CmCCmCCCCTTCGGTGCGGGGGGGC**TTAAGCCCCCCCGCACCGAAGGGGGGG-3'**	ND

After about 2 months, aptamer 3 produced tiny crystals in the Natrix crystallization formulations G7, G8, G9 and G10 at 20°C. These formulations have similar composition consisting of sodium cacodylate trihydrate as the buffer and (+/-)-2-methyl-2,4-pentanediol (MPD) as the precipitant (see table. 2 for details). No other hits were obtained for any of the other aptamers similar to 1 and 2 in any other screens.

Table 2: Composition of formulations in Natrix in which PR-RT-aptamer crystals were observed

Natrix crystallization formulation	Composition
G7	0.08 M strontium chloride hexahydrate, 0.02 M magnesium chloride hexahydrate, 0.04 M sodium cacodylate trihydrate pH 7.0, 20% v/v (+/-)-2-methyl-2,4-pentanediol, 0.012 M spermine tetrahydrochloride
G8	0.08 M sodium chloride, 0.04 M sodium cacodylate trihydrate pH 7.0, 30% v/v (+/-)-2-methyl-2,4-pentanediol, 0.012 M spermine tetrahydrochloride
G9	0.04 M lithium chloride, 0.08 M strontium chloride hexahydrate, 0.04 M sodium cacodylate trihydrate pH 7.0, 30% v/v (+/-)-2-methyl-2,4-pentanediol, 0.012 M spermine tetrahydrochloride

G10	0.04 M lithium chloride, 0.08 M strontium chloride hexahydrate, 0.02 M magnesium chloride hexahydrate, 0.04 M sodium cacodylate trihydrate pH 7.0, 30% v/v (+/-)-2-methyl-2,4-pentanediol, 0.012 M spermine tetrahydrochloride
Optimized condition	**0.04 M lithium chloride, 0.08 M strontium chloride hexahydrate, 0.04 M sodium cacodylate trihydrate pH 6.5, 25% v/v (+/-)-2-methyl-2,4-pentanediol, 0.012 M spermine tetrahydrochloride**

3.10 Crystallization optimization

Classic crystallization optimization techniques which explore the crystallization phase diagram to alter the crystallization kinetics by varying the components of the crystallization solution (McPherson, 2004, McPherson and Cudney, 2014) are still underway in an effort to produce crystals suitable for X-ray diffraction studies. Initial optimization efforts have included varying the protein concentration, the length of the aptamer and nature of hairpin, pH of the buffer (5.5, 6, 6.5, 7.0, 7.4, 7.8), varying the concentration of the MPD (20, 25, 30, 35 and 50) as well as the temperature (4ºC, 17ºC, 20ºC, 25ºC), changing the concentration and type of monovalent and divalent cations.

None of these parameters tested so far has had a profound impact on the crystallization kinetics and by extension the quality of the crystals obtained. A consensus condition for future crystallization trials was developed based on several observations (**see Table. 2**). Aptamer 5 with a 23 base-pair duplex region in table 2 which was designed based on aptamer 3 gave similar crystals as aptamer 3 suggesting that duplex length greater than 20 bp is not necessary for crystallization. It was however observed that high humidity at the time of setting up of the crystallization experiments which results in the reservoir becoming cloudy with condensation occurring on the lid/sealing film always led to bigger crystals and not showers as is usually the case at 20ºC.

This high humidity does not seem to alter the rate of crystallization, however. Crystallization still takes months before the impact of a parameter can be evaluated. Current optimization efforts are therefore being pursued accordingly by improvising high humidity chambers in an effort to obtain crystals suitable for X-ray diffraction studies. Initial thin crystals < 15 µM x 3 µM obtained from these optimization efforts diffracted X-rays to about 12 Å resolution for the mutant protein while that of the WT was less than 20 Å. The diffraction was however highly anisotropic and so subsequent crystallization optimization relied on the mutant and not the WT protein.

3.11 Additive screening

Small molecule additives which comprise an array of molecules ranging from inorganic molecules and metals as well as organic compounds and cofactors have been found to influence crystallization of macromolecular entities (McPherson and Cudney, 2014, McPherson, 2017). These molecules may affect the stability and solubility of biological macromolecules through direct binding, alter nucleation kinetics and space group of crystallization by perturbing crystal contacts or protein-protein interactions or in some cases provide the right turbidity, viscosity and other rheological properties in the crystallization milieu to aid crystallization or improve the diffraction limits of crystals (McPherson and Kuznetsov, 2014, McPherson, 2017). Commercially available additives have been formulated based on empirical observations and data mining from several sources including the PDB. Using the optimized crystallization condition, additive screening from Hampton was carried out. Again, none of this array of molecules had any profound impact on the nature and type of crystals obtained with more than 80% of the conditions producing crystals of similar morphology as the original hits.

In consonance with the additive screening principle, a 50:50 mixture (v/v) of the TOP96 crystallization screen from Anatrace and the optimized crystallization solution (**Table 2**) was also carried out. The TOP96 is a crystallization formulation based on the 96 most frequently reported crystallization conditions in the PDB. This was carried out largely based on the observation that the protein-nucleic acid complex precipitated even at 4 mg/mL in more than 80% of the conditions without any hits. By mixing a condition in which the complex crystallizes with these, it was hoped that some of these chemical agents may act as additives that improve crystallization kinetics as well as improve the diffraction.

Several of the drops produced crystals as was expected, but none of these additives improved the crystallization kinetics. It took over a month to for these crystals to appear. Among the several hits, it was interesting to observe that PEG 3350 was a common factor to most of them. The best 4 hits obtained were from E2, F6 H4 and H6 all of which had this PEG as the precipitant. Condition E2 contains only PEG 3350, F6 contains ammonium nitrate pH 6.3 and PEG 3350, H4 contains ammonium acetate, Bis-Tris pH 6.5 and PEG 3350 while H6 contains Tris HCl, pH 8.5 PEG 3350. The wide range of pH in which these crystals appeared is consistent with the initial optimization with pH variation, which didn’t indicate this parameter to be important for crystallization. Among these crystals, F6 crystals were the fewest and the largest measuring nearly 30 x 10 µm followed by H4 where some were about 20 x 5 µm in dimension while others were smaller. Even though H6 and E2 crystals (**Fig. 32**) were small, less than 10 microns in the longest dimension and just about 1-2 micron in the smallest dimension, they looked well separated and single and could be good candidates for micro electron diffraction.

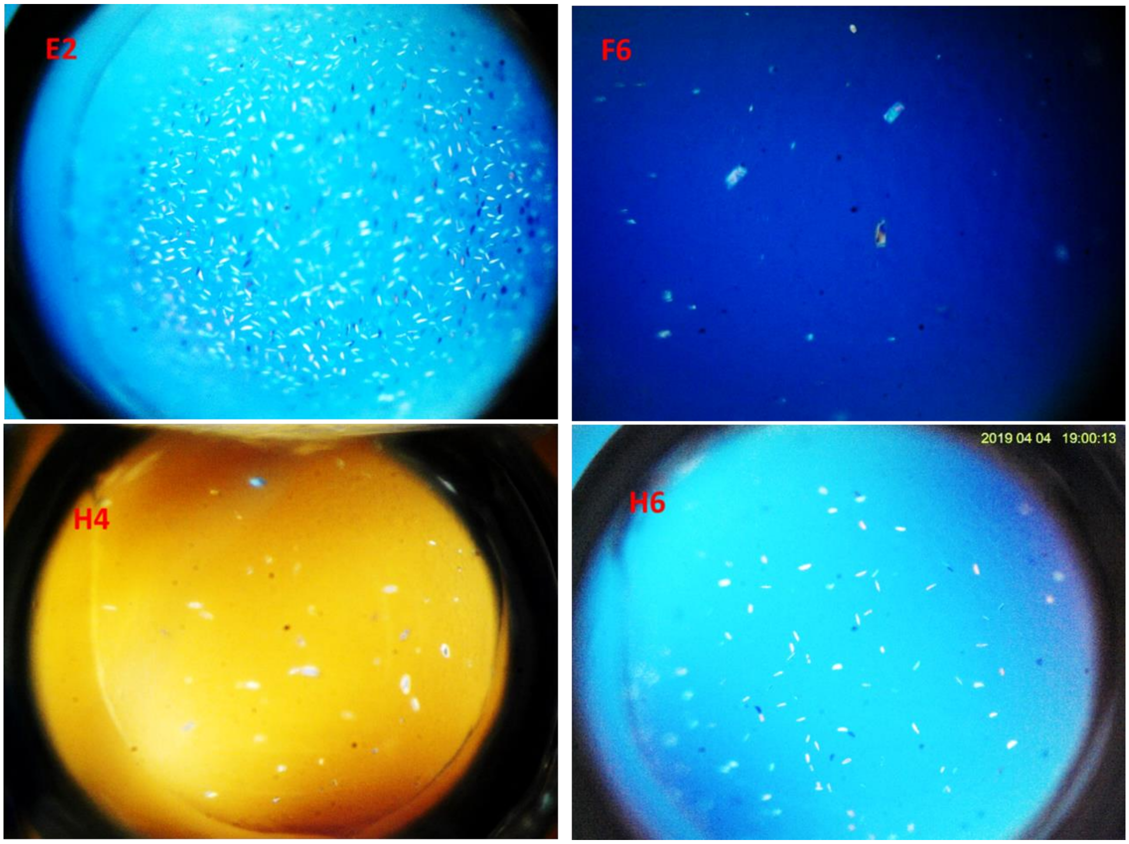


Figure 32: Images of the PR-RT-Aptamer crystals obtained in the Top96 additive screening experiment

Consistent with the crystal sizes, crystals from F6 and H4 diffract X-rays to about 3.3 Å resolution in one direction (**Fig. 33**) but the diffraction also suffers from the same anisotropy as the initial crystal hits. The unit cell in these crystals is elongated with two short axes along which X-rays are diffracted to the highest resolution limit and a very long axis along which the diffraction is poor. However, because of the very long nature of the third axis, all the diffraction spots overlap into a single streak even when the detector is pulled further back. Further optimization is currently being undertaking by varying the pH of the Tris, Bis-Tris, concentration of ammonium nitrate and ammonium acetate, as well as the PEG 3350 in the hope of obtaining bigger and better diffracting crystals.

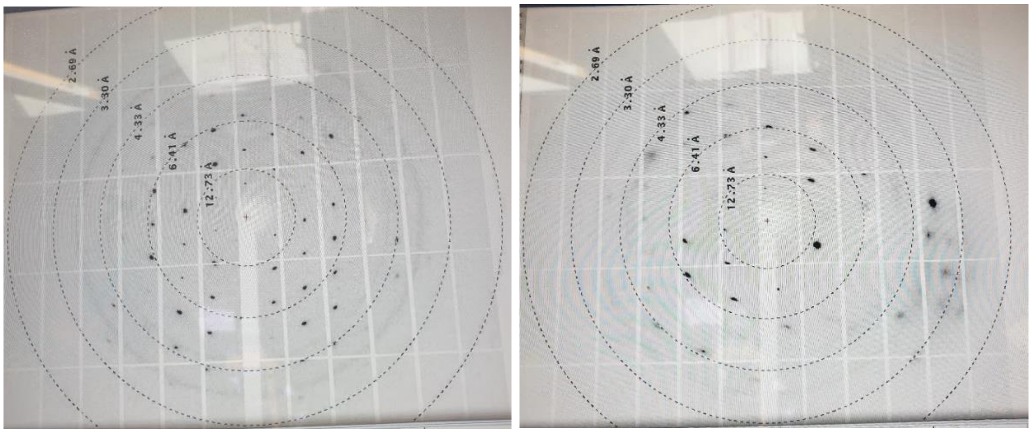


Figure 33: Diffraction spots from crystals that grew in Top96 condition H6

3.12 Pre-seeding

Nucleation precedes crystal growth during crystallization of molecules. While nucleation is generally spontaneous when supersaturation of the protein/precipitant mixture is reached, many crystallization experiments fail because this spontaneous nucleation does not occur (McPherson and Cudney, 2014, McPherson and Kuznetsov, 2014, McPherson, 2017). To overcome this situation, seeding which involves addition of preformed crystals of either the same macromolecule or a homolog or sometimes nonrelated entities may be helpful in growing crystals suitable for structure determination.

The preformed crystals which may be added to the macromolecule/precipitant mixture as large chunks or crushed into submicron sizes provide a nucleus with the right interfaces along which new macromolecular entities can associate in a lattice. The seeds must have fresh and non-poisoned surface able to accommodate new macromolecular entities for this to work. On the other hand, the crystallization kinetics may not depend on nucleation but rather the rate at which new macromolecules associate in the lattice (McPherson and Kuznetsov, 2014). Seeding is less likely to work under such circumstances. Pre-seeding of the crystallization drops of the PR-RT/aptamer complexes have also been carried out in a 96 well Natrix crystallization screen as well as the optimized condition in which the complex crystallizes without success.

This suggests that the length of time it takes for crystals to reach critical size is not necessarily a function of the nucleation but more likely due to the inability of the complexes to associate stably in the crystal growing interface. Also due to the slow kinetics of association, the crystal growth surface may become poisoned over time and unable to allow new molecules to form. Poisoning of the crystallization surface is a likely factor responsible for the failure of pre-seeding in producing better quality crystals that diffract X-rays uniformly to high resolution for structure determination.

3.13 Formation of stable PR-RT dimers

The PFV PR-RT carries out reverse transcription of the genomic RNA to dsDNA which can be inserted into the host genome. The PR is responsible for the processing of the IN from the Pol, as well as the cleavage of a short peptide from the C-terminus of the Gag which is required for infectivity (Flugel and Pfrepper, 2003, Boyer et al., 2004, Hartl et al., 2010a). The mechanism by which the PR is activated is controversial as alluded to earlier. Nonetheless, since the HIV PR-RT for which this system is a surrogate for contains an enzymatically active PR (Cherry et al., 1998a, Cherry et al., 1998b), structural studies of this PR-RT is necessary. Working with RNA in a laboratory that is not RNase free is very tricky and difficult with several avenues for RNase contamination.

While initially contemplated, the idea of producing the PARM RNA shown to activate the proteolytic activity of the PR-RT was discarded. The complexity of the RNA motif would likely not promote crystallization since there are flexible domains in it and concerns with the stability of this complex itself once it is formed made it a difficult option to take.

3.14 Disulfide cross-linking of PR-RT

Generation of a tethered dimer of PFV PR remained perhaps the only option left having abandoned the idea of using RNA. A covalently tethered dimer of the HIV-1 PR has been described (Bhat et al., 1994). This was achieved by connecting two PR monomers with a 5-residue linker which enabled self-dimerization of the PR monomers. Clearly this will not be possible in the PR-RT because of the presence of the RT at the c-terminus of PR. Disulfide cross-linked HIV-1 PR has also been described in the literature (Mittal et al., 2012). However, these cross-linking studies, which required introduction of site-specific cysteines in at least two different places in the protein, yield intramolecular disulfide cross-linked proteins. Clearly, this approach would not be useful for obtaining stable dimers of the PR-RT as well.

Retroviral PRs contain a catalytic triad of DSG as in PFV or DTG as in HIV. These residues engage in a network of hydrogen bonding which stabilizes the dimeric protein, the so-called fire man’s grip. In the crystal structures of HIV PR, the two oxygen atoms of the two-fold related T26 residues are opposite each other and separated by 3.4 Å (**Fig. 34**). This distance is close enough for disulfide bond formation if these were sulfhydryls however, the geometry can be problematic since disulfide bond formation requires close to ideal geometry of dihedral angle to enable the stereochemistry required for this covalent bond formation (Craig and Dombkowski, 2013).

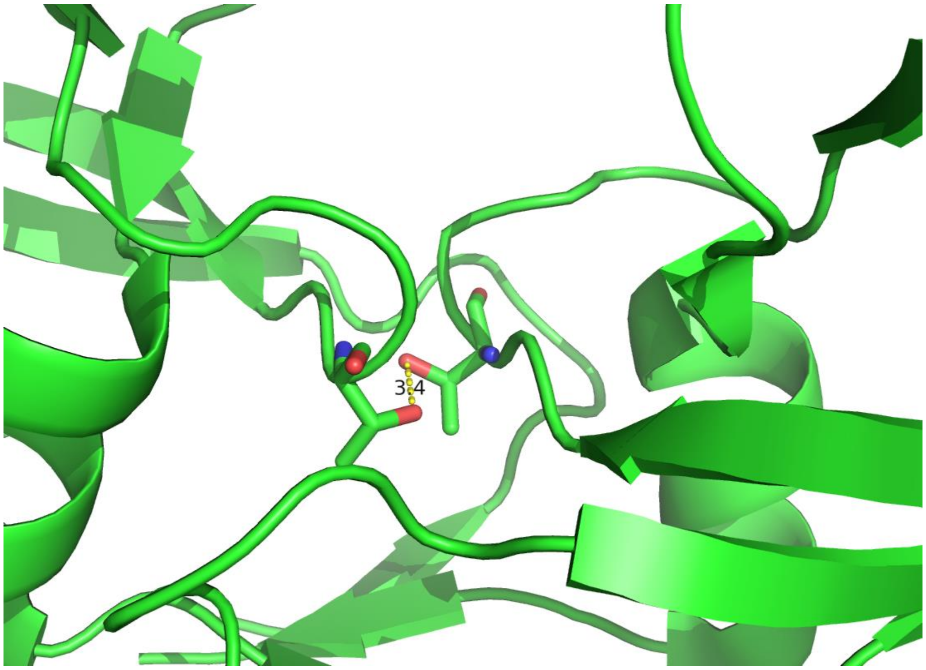


Figure 34: HIV-1 PR dimer showing the distance between T26 from the two protomers

Since in PFV this residue is a Ser, the potential impact of the loss of any stabilization interaction from the β-carbon upon substitution with Cys will not arise, however. Despite the potential geometric constraints on disulfide formation, the free rotation about the sp^3 hybridized α-carbon means the probability of the two cysteines being in the right configuration upon dimerization could not be zero. This residue was therefore very attractive for substitution since it is slightly hindered even though it is solvent accessible thus reducing the possibility of forming spurious disulfide. The S25C mutant of the PR-RT would be expected to form a disulfide bond if at all only under the right conditions. To reduce the tendency of forming any unwanted disulfide bond, mutations C31S in the PR and C280S in the RT were also carried out in the construct of the PR-RT containing the S25C.

PR-RT constructs containing the S25C mutation were expressed and purified according to the protocol discussed in chapter two of this work. To trigger the dimerization, the eluted protein from the heparin column in 50 mM Tris-Cl, pH 8.5, 1.0 M NaCl and 2 mM TCEP was concentrated to > 25 mg/mL and diluted into a buffer of 50 mM Tris-Cl, pH 8.5, 4.0 M NaCl. The volume of the solution was chosen such that the final concentration of NaCl in the solution will be at least 3.0 M. This method of inducing PR dimerization has been well characterized in the literature for SFVmac PR and PR-RT as well (Hartl et al., 2010a, Hartl et al., 2010b, Schneider et al., 2014).

Non-reducing SDS-PAGE analysis of the dimerization reaction showed that the efficiency of the disulfide bond formation is about 50% (**Fig. 35**). This result is highly reproducible if the cross-linking is induced within 24 hours of heparin purification using buffers that had been degassed for the protein purification and the high salt cross-linking buffer which reduces the oxidation of the sulfhydryl groups. It is worthy of note that the formation of the disulfide is spontaneous. The yield of the reaction therefore does not increase by incubating the protein in the 3 M salt for a longer time.

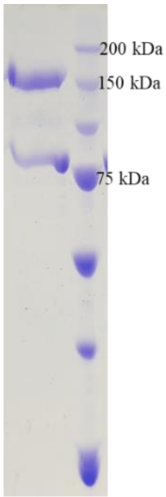


Figure 35: SDS-PAGE gel of PFV PR-RT S25C dimerization reaction

3.16 Separation of PR-RT dimers and monomers

Biophysical studies of the PR-RT dimers require removal of the monomeric proteins from the mixture. Neither hydrophobic interaction chromatography taking advantage of the protein already in a high salt buffer to separate monomers from dimers based on increased hydrophobicity or heparin purification using gradient elution was successful at separating the monomers and the dimers. Dimers and monomers of the PR-RT were therefore separated using gel filtration chromatography.

The protein in high salt buffer was diluted at least 3-fold and buffer exchanged into a 25 mM Tris-Cl pH 8.0, 150 mM NaCl buffer using an Amicon Ultra-15 centrifugal filter unit. The concentrated mixture was then loaded onto either Superose 6 Increase 10/300 or Superdex 75 preparatory grade column run using 25 mM Tris-Cl pH 8.0, 150 mM NaCl. Two overlapping peaks are obtained using either column. Non-reducing SDS-PAGE analysis shows predominantly dimeric proteins in the early peak, which gets contaminated with the monomeric protein as the elution progresses (**Fig. 36**). Successive runs of the sample on the column helps in increasing the yield of the dimers.

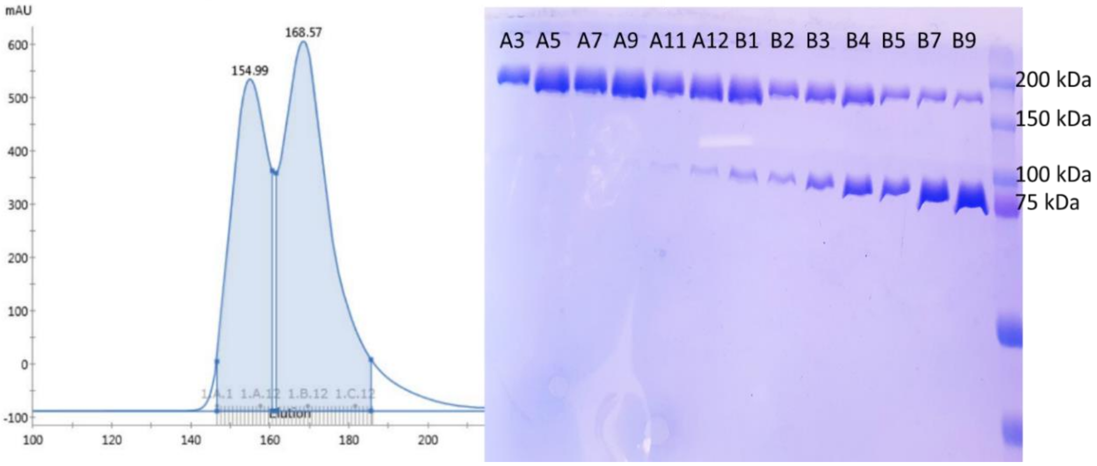


Figure 36: Superdex 75 gel filtration profile of PR-RT dimers (left) and corresponding non-reducing SDS-PAGE gel

3.17 PR-RT cross-linked dimer is partially resistant to reduction

To ascertain whether the higher molecular weight bands on the non-reducing SDS-PAGE gel are actually disulfide cross-linked proteins, 20 µg of the isolated dimer was taken and incubated with 2 mM TCEP at room temperature for 4 hours before an aliquot was taken and added to an SDS-PAGE loading dye and run on a 10% SDS-PAGE gel under non-reducing conditions. Surprisingly, only about 50% of the protein was reduced (**Fig. 31**). This is consistent with the fact that the disulfide bond is sterically inaccessible which is expected given the strategic position the disulfide bond is expected to form in these molecules.

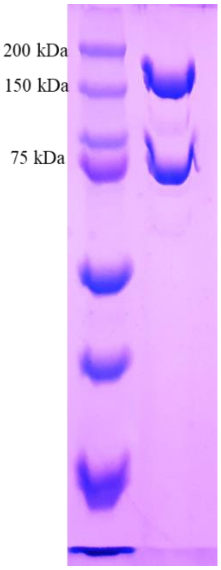
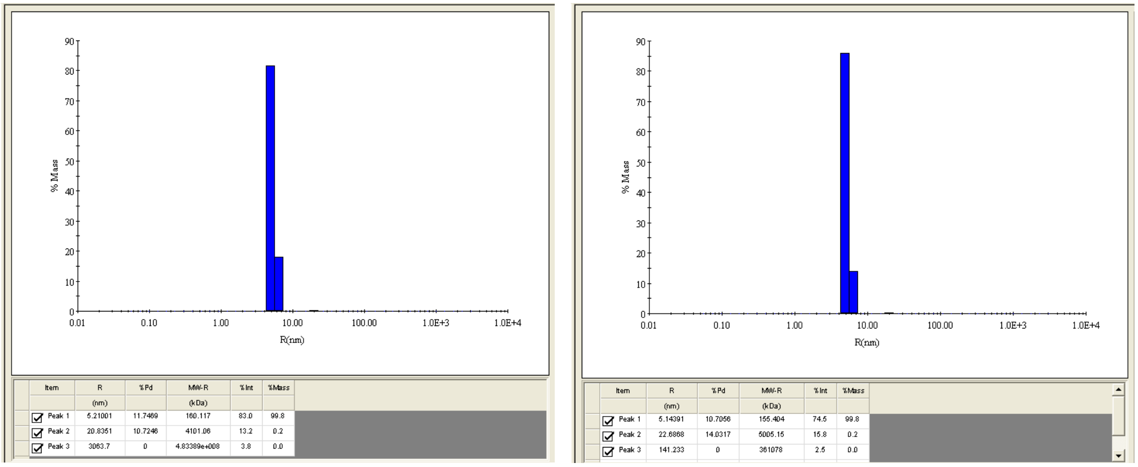


Figure 37: SDS-PAGE gel of PFV PR-RT S25C dimers + 2 mM TCEP

3.18 DLS Analysis of PR-RT dimers

The PR-RT dimers were also characterized by DLS to determine the sizes of the molecules in solution and their hydrodynamic radii. At 5.0 mg/mL of protein, the average molecular weight of the molecules in solution was found to be 155-160 kDa with a hydrodynamic radius of ~5 nm. Compared to the molecular weight of 85.5 kDa for the monomer, this is consistent with dimeric molecules of this monomer in solution (**Fig. 38**). This is a further confirmation of stable dimers in solution consistent with the cross-linking experiments. The polydispersity of the molecules was found to be < 12% suggesting that there is less conformational variability in the molecules making it a good candidate for crystallization screening.



Item	R	%Pd	MW-R	%Int	%Mass
	(nm)		(kDa)		
☑ Peak 1	5.21001	11.7469	160.117	83.0	99.8
☑ Peak 2	20.8351	10.7246	4101.06	13.2	0.2
☑ Peak 3	3063.7	0	4.83389e+008	3.8	0.0

Item	R	%Pd	MW-R	%Int	%Mass
	(nm)		(kDa)		
☑ Peak 1	5.14391	10.7056	155.404	74.5	99.8
☑ Peak 2	22.6868	14.0317	5005.15	15.8	0.2
☑ Peak 3	141.233	0	361078	2.5	0.0

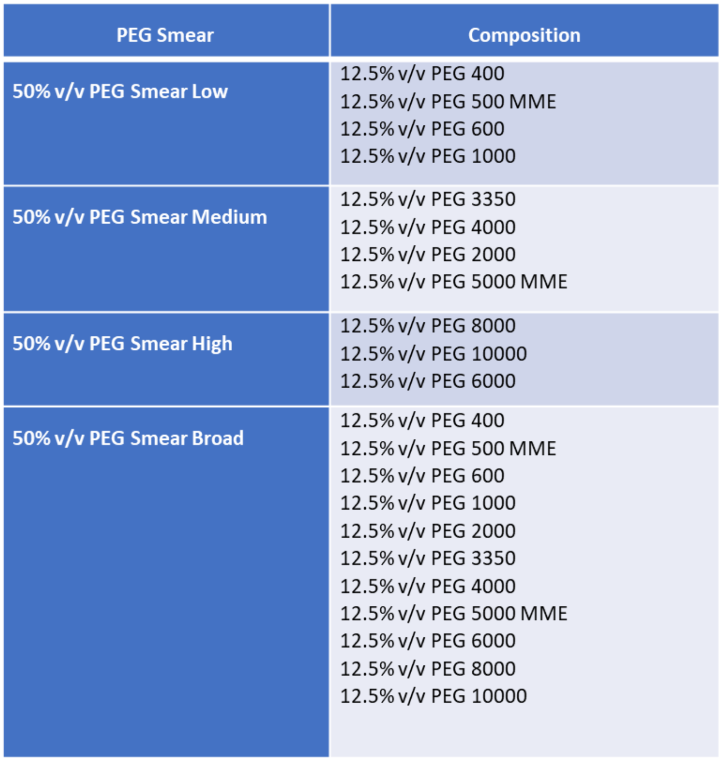
Figure 38: DLS profiles of PR-RT dimers

3.19 Crystallization screening of PR-RT dimers

PR-RT dimer at 5 mg/mL was screened for crystal hits using commercially available crystallization screening solutions and with the help of crystallization screening robot at 20 ºC. Natrix (Hampton), Morpheus I & II, BCS screen (Molecular Dimension) and Top96 (Anatrace) crystallization screens have been tried so far. One crystal hit was observed in Morpheus I condition B8 (0.09 M Halogens-NaF, NaBr, NaI, 100 mM Sodium HEPES-MOPS, pH 7.5, 37.5% MPD_PEG 1000_PEG 3350- at 12.5% each). On the other hand, several hits of similarly looking crystals were observed in the BCS screen in a variety of different conditions details of which are given in Table. 3, Fig.39 and 40. No other hit was observed in the other screens. Optimization of these hits is currently being undertaken to improve these crystals.

Table 3: BCS screening conditions in which PR-RT dimer crystal hits were observed

B2	0.1 M Tris pH 8.5, 30 % v/v PEG Smear Low
B5	0.1 M Tris pH 8.5, 25 % v/v PEG Smear Medium
B7	0.1 M HEPES pH 7.5, 20 % v/v PEG Smear High
B9	0.1 M BICINE pH 9.3, 20 % v/v PEG Smear High
B10	0.1 M HEPES pH 7.5, 22 % v/v PEG Smear Broad
B11	0.1 M Tris pH 8.5, 22 % v/v PEG Smear Broad
B12	0.1 M BICINE pH 9.3, 22 % v/v PEG Smear Broad
C12	0.1 M calcium chloride dihydrate, 0.1 M magnesium chloride hexahydrate, 0.1 M PIPES pH 7.0, 22.5 % v/v PEG Smear Medium
D10	0.2 M sodium chloride 0.1 M sodium phosphate pH 6.2, 28 % v/v PEG Smear Broad
E11	0.05 M magnesium chloride hexahydrate, 0.05 M sodium citrate tribasic dihydrate, 0.1 M Bis-Tris propane pH 7.8, 22.5 % v/v PEG Smear High
G10	0.1 M ammonium sulfate, 0.1 M sodium formate, 0.1 M HEPES pH 7.0, 25 % v/v PEG Smear Broad
H9	0.05 M ammonium sulfate, 0.05 M lithium sulfate, 0.1 M Bis-Tris propane pH 8.5, 28 % v/v PEG Smear Broad



PEG Smear	Composition
50% v/v PEG Smear Low	12.5% v/v PEG 400 12.5% v/v PEG 500 MME 12.5% v/v PEG 600 12.5% v/v PEG 1000
50% v/v PEG Smear Medium	12.5% v/v PEG 3350 12.5% v/v PEG 4000 12.5% v/v PEG 2000 12.5% v/v PEG 5000 MME
50% v/v PEG Smear High	12.5% v/v PEG 8000 12.5% v/v PEG 10000 12.5% v/v PEG 6000
50% v/v PEG Smear Broad	12.5% v/v PEG 400 12.5% v/v PEG 500 MME 12.5% v/v PEG 600 12.5% v/v PEG 1000 12.5% v/v PEG 2000 12.5% v/v PEG 3350 12.5% v/v PEG 4000 12.5% v/v PEG 5000 MME 12.5% v/v PEG 6000 12.5% v/v PEG 8000 12.5% v/v PEG 10000

Figure 39: Composition of PEG Smears

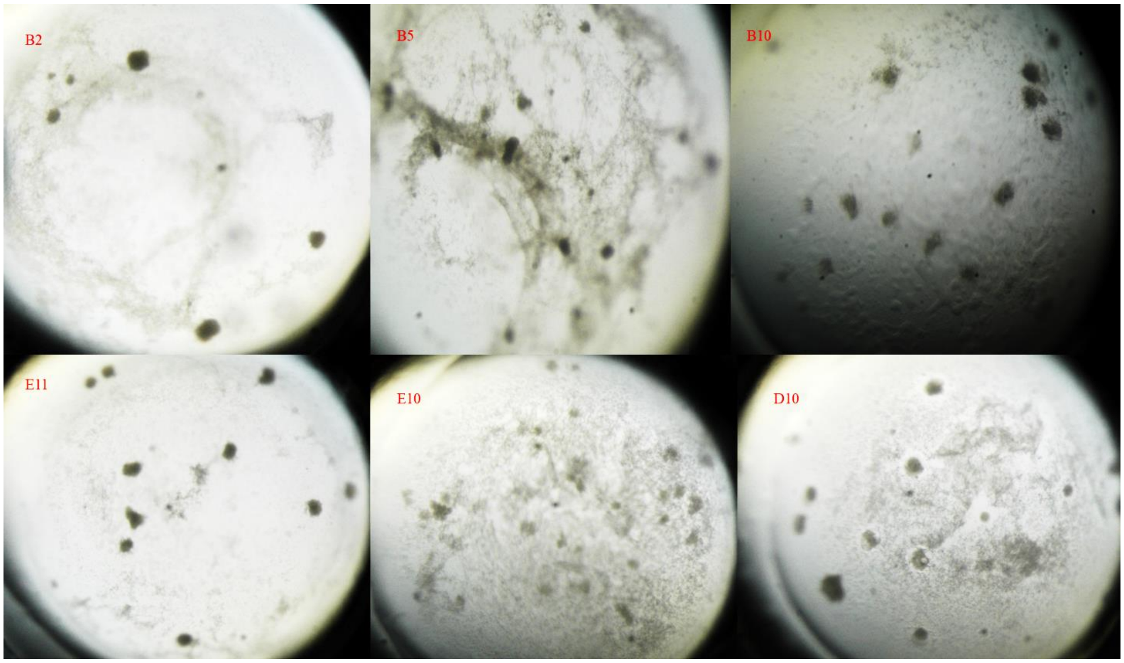


Figure 40: Nature of initial crystal hits for PR-RT dimers in BCS crystallization screen

3.20 Conclusions

In an effort to understand the mechanism by which the PFV PR-RT carries out nucleotide incorporation into a growing chain of dsDNA, structural studies of the PR-RT with T-P DNAs have been undertaken. Biochemical studies suggest that unlike HIV-1 RT, which engages dsRNA/DNA hybrids more tightly and dissociates slowly from this nucleic acid hybrid compared to dsDNA, PFV RT engages dsDNA more tightly and dissociates more slowly from the dsDNA than the dsRNA/DNA hybrid.

Crystallization experiments were hence carried out using dsDNA to identify conditions under which PFV PR-RT crystallizes with DNA. Site-specific Q391C in PFV RT enabled cross-linking of the protein-nucleic acid complex using the same T-P dsDNA previously crystallized with HIV-1 RT Q258C mutant. The success of these cross-linking experiments suggests that the distance between the polymerase active site and the thumb sub-domain residue Q391 is the same in these two proteins. This is not withstanding the fact that PFV RT is monomeric while HIV RT is heterodimeric and that the subdomain arrangements in the crystal structures of the apo proteins of these two proteins is different. It is a confirmation that the PFV PR-RT protein undergoes an extreme conformational rearrangement when it binds T-P nucleic acid which results in a configuration similar to the p66 sub-domain of HIV-1 RT. Crystallization was however unsuccessful with this complex.

Crystallization trials were also carried out using a 38NT-2,4-methyl hairpin aptamer with 13 pM affinity to HIV-1 RT initially obtained through SELEX. This aptamer binds with 300 pM affinity to PFV PR-RT. A mutant form of the PFV PR-RT carrying mutations in the RNase H domain bind with to the aptamer with about 3-6-fold increased affinity. While this 38-mer aptamer was also not successful at producing crystals of the PFV PR-RT, an extension of the 15 bp duplex region of the 38NT-aptamer to 20-23 while changing the hairpin from TTT to TTAA tetraloop resulted in crystals of the PFV PR-RT-DNA complex initially in the Natrix commercial screening conditions G7-G11.

Following series of optimizations, crystals diffracting X-rays to about 3.3 Å resolution have been obtained. This was obtained by adding ammonium nitrate or ammonium acetate to the optimized crystallization condition; 0.04 M lithium chloride, 0.08 M strontium chloride hexahydrate, 0.04 M sodium cacodylate trihydrate pH 6.5, 25% v/v (+/-)-2-methyl-2,4-pentanediol, 0.012 M spermine tetrahydrochloride. These inorganic molecules were identified through additive screening experiment carried out by mixing the optimized Natrix hit with TOP96 crystallization screen. The X-diffraction pattern is however highly anisotropic with diffraction spots along the longer cell axis forming a streak which is not resolved even when the detector is moved further away. Further optimization is currently ongoing.

To understand the mechanism of proteolytic maturation of the PFV Polyprotein precursors which has implications for HIV and other retroviruses, stable dimers of the PFV PR-RT has been generated through disulfide bond formation between two monomers of the PFV PR carrying S25C mutation in the fireman's grip. In low salt buffer, no cross-linking is observed between these protein monomers. However, in 3.0 M NaCl or higher, spontaneous disulfide bond is formed between two monomers of PR with a yield of about 50%.

This is the first demonstration of disulfide cross-linking of an aspartic protease through the fireman's grip which contributes strongly to the stability of these protease dimers in the non-cross-linked dimers. Also, despite the geometrical constraints and regiochemical requirements for disulfide bond formation, ~50% of the monomers of the PR are in a configuration suitable for disulfide bond formation. The cross-linked dimers are partially resistant to TCEP reduction consistent with the fact that this disulfide bond is at the PR dimer interface formed by the fireman's grip and therefore sterically hindered.

The monomers and dimers can be separated using gel filtration chromatography on Superose 6 Increase 10/300 GL or Superdex 75 preparatory grade gel filtration columns. Two overlapping peaks containing the dimer in the earlier peaks and mixture of dimer and monomer in the later peak were obtained. DLS analysis of the isolated dimer confirms dimeric molecules of PR-RT in solution with molecular weight ~160 kDa and hydrodynamic radius of about 5 nm. The polydispersity of these molecules is less than 12% suggesting that there is little conformational heterogeneity in these molecules. Crystallization screening has identified several hits in the BCS crystallization screening kit. These hits are currently being optimized.

3.21 Materials and Methods

Protein expression and purification was carried out as outlined in chapter one of this thesis. For protein DNA cross-linking, a slight excess (10%) of the template in this instance was used during the T-P annealing since the primer which is modified is more expensive. Annealing is done in 20 mM Tris-Cl pH 8.0 and 150 mM NaCl. The Q391C mutant protein is also buffered in 20 mM Tris-Cl pH 8.0 and 150 mM NaCl. These salt concentrations are 2-fold higher than the amount used for the cross-linking reaction with HIV-1 RT. The PFV PR-RT DNA complex precipitates when the NaCl concentration is less than 150 mM. For the cross-linking reaction, 2-fold molar excess protein was mixed with the annealed DNA in the presence of 1.0 mM BME, 5 mM $MgCl_2$, 0.2-1mM dideoxy-NTP (ddNTP) in 50 mM Tris-Cl buffer pH 8.0 and allowed to sit at room temperature for about 4 hours or 20-30 minutes at 37 ºC. This is also contrary to the cross-linking reaction with HIV-1 RT which is more efficient at 37 ºC for hours. For PFV PR-RT, it was observed that significant precipitation occurs after about 30 minutes at 37 ºC.

After cross-linking, the mixture is loaded onto a 5 mL heparin column pre-equilibrated with 25 mM Tris-Cl, pH 8.0, 150 mM NaCl. Column is washed with this buffer to remove all uncross-linked DNA and excess ddNTP. A gradient run is initiated by increasing the concentration of buffer B, which is the same as the wash buffer except it has 1.0 M NaCl. The cross-linked protein elutes around 300 mM NaCl concentration. The non cross-linked protein is eluted off the column with 1.0 M NaCl. Cross-linked protein is buffer exchanged into 25 mM Tris-Cl, pH 8.0, 150 mM NaCl after the purity has been ascertained to be satisfactory on SDS-PAG and stored at -80 ºC.

CHAPTER FOUR: Cloning, Expression and Purification of PFV Pol, and HIV-1 Pol and Gag-Pol Polyproteins for Structural and Biophysical Studies

Synopsis

Precursor polyproteins are the parent entities from which mature, and fully functional entities are begat. The structural and functional characterization of these precursor proteins ubiquitous among pathogenic organisms including retroviruses remains extremely challenging. This has largely been due to lack of simple and easy to use media for expressing and purifying these proteins. They are of high interest for therapeutic development as they represent a vulnerable stage of the life cycle of these pathogens. For HIV, all the enzymes used to catalyze various reactions that make the virus infectious reside in the Gag-Pol precursor which constitutes only 5% of the virus particle while the remaining 95% which provide the structural infrastructure which houses the genomic materials and other host macromolecules required by the virus to replicate are also made en bloc as a precursor polyprotein. The possibility of finding new interfaces and domains with rather high genetic barrier to resistance makes it even more enticing since the next generation of anti-HIV therapeutics could target these structural arrangements unique to HIV and other pathogens without any replica in eukaryotes.

Overcoming protein expression challenges in bacteria generally require a multi-faceted approach. Media engineering, which entails manipulation of the constituents of media for growing cells including bacteria, was carried out to identify conditions suitable for expression and purification of these polyprotein precursors from bacteria, which remain the most flexible and cost-effective means of protein production for biotechnological applications. The discovery of the new media conditions coupled with cloning techniques enabled the expression and purification of the full-length PFV Pol, HIV-1 Pol, and Gag-Pol polyprotein precursors in yields and purity suitable for biophysical and structural studies. The HIV-1 constructs containing IN were also successfully expressed and purified with the IN-binding domain (IBD) of lens epithelium derived growth factor (LEDGF), fused with a rigid helical linker to the C-terminus of the maltose binding protein (MBP). LEDGF is a binding partner of HIV IN that plays a key role in integration site selection. Some of these complexes have also been characterized using gel filtration, dynamic light scattering (DLS), small angle X-ray scattering (SAXS) and single particle cryo-electron microscopy (cryo-EM). The details of these experiments are described below.

4.1 Polyproteins in the retroviral life cycle

Synthesis of polyproteins and their subsequent proteolysis to produce mature proteins is highly conserved in all retroviral genera. It affords them the opportunity to control their replication in their respective hosts to a large extent. The Gag polyprotein precursor made as a precursor in all retroviruses comprises the structural proteins; the individual proteins from the N-terminus to the C-terminus are: matrix (MA), capsid (CA), spacer peptide 1 (SP1), nucleocapsid (NC), spacer peptide 2 (SP2) and p6 (Adamson and Freed, 2007, Engelman and Cherepanov, 2012, Freed, 2015). The enzymatic proteins protease (PR), reverse transcriptase (RT) and integrase (IN) are either made separately as in PFV (Enssle et al., 1996) or as part of Gag in a form of Gag-Pol (Adamson and Freed, 2007).

While the mechanisms of assembly generally differ in different retroviruses, formation of the virus particles is mediated by Gag molecules in all species (Baldwin and Linial, 1998, Freed, 2015, Hamann and Lindemann, 2016). In PFV, proteolytic maturation is limited to a single cut in Pol between RT and IN and two cuts within the last 10 kDa of the Gag C-terminus (Flugel and Pfrepper, 2003). Thus, for the entire duration of the life cycle, the PFV proteins exist like their polyprotein precursors. In HIV-1 and other retroviruses, this is not the case because each of the domains is released from their precursor albeit at a well-choreographed tempo (Lee et al., 2012). While the cleavage events are asynchronous, maturation of HIV-1 viruses begin with the dimerization of two Gag-Pol molecules followed by an intramolecular cleavage event believed to occur between SP2 and NC to liberate two precursor polyproteins. The subsequent events leading to the fully mature infectious virion can take up to 12 hours to complete with various precursor intermediates generated in the process (Pettit et al., 2004, Pettit et al., 2005b).

The HIV-1 PR recognizes eleven unique cleavage sites in Gag and Gag-Pol without any off-target cleavage events (Adamson, 2012, Konvalinka et al., 2015). A dynamic substrate envelope hypothesis has been postulated as the mechanism by which the PR recognizes these unique sites for accurate cleavage. This observation relies largely on individual peptides and PI inhibitors of the various cleavage sites bound to the mature enzyme (Ali et al., 2010, Ozen et al., 2011). For the purpose of drug design, this concept has been quite useful in overcoming drug resistance, which is a regular feature of inhibition by these very dynamic enzymes. Since most of the cleavages, especially during the initial stages, are carried out by an enzyme with extra residues at its N- and C-termini, and taking into consideration the steric effects of many of these cleavage sites, it is plausible that other means of specificity determination exist which can be deciphered by solving the structures of these precursor polyproteins. Undoubtedly, understanding this intricate processing event will be very useful for therapeutic design for inhibitors that target HIV-1 replication.

4.2 Optimization of recombinant protein expression in bacteria: *Escherichia coli* (*E. coli*)

Using bacteria strains remain the most popular means of recombinant protein production not only for academic research but also for industrial and pharmaceutical applications as well (Huang et al., 2012). This popularity is due to several reasons. First, laboratory strains of *E. coli* are generally safe, grow very rapidly and hence easy to cultivate in the lab. Having been around for a long time, it remains the most well studied system for protein production. The compatibility with different types of promoters and commercially available vectors permit tight regulation of transcription which enables induction of protein expression under optimal conditions that allows the reproducibility and scale-up of protein expression to be straight forward (del Solar et al., 1998, Francis and Page, 2010). Furthermore, the use of purification tags facilitates the extraction of expressed protein from the potpourri of bacterial proteins by using a variety of affinity purification types. However, despite all the knowledge available on this system, there are serious limitations to the expression of heterologous proteins in *E. coli*, some of which are given below (Francis and Page, 2010).

Proteolytic cleavage and breakdown of heterologous proteins in *E. coli* is a major bottleneck to a successful protein production campaign. Not surprisingly, many proteins of research interest are recognized by the bacteria host as foreign and therefore induce proteases which cleave these heterologous proteins to truncated products resulting in very low yields or sometimes complete breakdown. This is especially true for large multi-domain proteins especially with the molecular weight greater than 60 kDa (Structural Genomics et al., 2008).

This is not withstanding the knock-out of the most active proteases OmpT and Lon in many of the commercially available strains of *E. coli* earmarked for protein expression (Grodberg and Dunn, 1988). Biophysical features of proteins such as sequence complexity, hydrophobicity as well as the toxicity of proteins are also important features that may affect their expression and or successful extraction from bacteria. Hydrophobic proteins are generally prone to aggregation in the cytosol of bacteria which elicit immune-like responses (Peti and Page, 2007, Sahdev et al., 2008).

It is generally true that these proteins associate non-specifically with bacterial proteins that result in a cascade of cellular responses which generally result in the destruction of the heterologous protein or it being shunted into inclusion bodies to protect the cells against apoptotic signaling (Peti and Page, 2007, Vera et al., 2007, Huang et al., 2012). Each taxonomic group of organisms utilize sets of tRNA codons that generally vary significantly from the rest. It has been recognized that many codons very prevalent in mammalian systems and efficiently utilized are rarely utilized in bacteria. The rarity of these codons means their concentration in the pool of tRNAs in the bacterial system is low compared to the most frequently used codons. Consequently, expressing proteins of eukaryotic origin enriched in these rarely used codons in bacteria lead to very low yields if at all, because the ribosomes must wait for a very long time for these rarely used codons to arrive. These may lead to abortive translation and therefore truncated proteins (Sahdev et al., 2008). On the other hand, toxic proteins may activate/deactivate a signaling pathway that may result in apoptosis of the cells.

Several methods have been developed to overcome some of these bottlenecks of lack of expression or low yield of protein of interest in bacteria (Studier et al., 1990, Francis and Page, 2010, Ozturk et al., 2017). They include but are not limited to:

- Type of cloning vector and promoter
- Strain of *E. coli*
- Type of media
- Expression temperature
- Concentration of inducer (IPTG, Arabinose etc)
- Choice of fusion tag
- Co-expression with molecular chaperones and protein binding partners

4.3 Cloning vectors and promoters used in *E. coli*

Introduced in 1952, a plasmid, which is the most common means of expressing heterologous proteins in *E. coli*, refers to an extra-chromosomal element carrying genetic information (Lederberg, 1998). Generally present as small covalently-closed double-stranded DNA molecules, these extra-chromosomal genetic materials have characteristics that make them capable of replicating in their hosts independently using the same replication machinery as the host (del Solar et al., 1998). Plasmids are a major means of transferring genetic properties such as antibiotic resistance between organisms. They can also be introduced into cells intentionally through transformation, which means that the cells carry the self-replicating new elements. Artificially constructed circular DNA carrying all the features that make it capable of autonomous replication in a host is referred to as a vector (del Solar et al., 1998, Lederberg, 1998).

Covalently-closed circles of dsDNA or plasmid vectors are the most common vehicles through which over production of heterologous proteins is achieved both in bacteria and other cell lines where integration of a chromosomal copy of the gene of interest is not required. Plasmids generally contain a replicon (origin of replication) where the replication machinery coalesces to initiate the replication of the plasmid. The replicon also controls the copy number of the plasmid. Copy number may be low (less than 20 copies per cell), medium (20-100 copies per cell) or high (more than 100 copies per cell) (del Solar et al., 1998, Lederberg, 1998, Adamczyk and Jagura-Burdzy, 2003,). An antibiotic resistance gene which allows the selection of bacteria colonies or cells transformed with a plasmid while growing in a media supplemented with a tolerable concentration of that antibiotic is a regular feature of cloning vectors as well. Transcriptional and translational regulation of protein expression using these vectors is achieved using promoters. Several promoter types have been unraveled since the classic lactose operon of gene expression control was first described in bacteria by Francois Jacob and Jacques Monod in 1960 (Jacob et al., 2005) which opened the frontiers of gene expression and regulation.

Control of transcription is achieved by a lac repressor binding to operator sequences upstream and downstream of the promoter binding site (-10 and -35 elements) and blocking the RNA polymerase (RNAP) from gaining access to the promoter to initiate transcription of the downstream gene(s) which usually encodes the target protein(s) of interest (Studier et al., 1990). The strength of these operator sequences as well as the promoter sequences generally determines the strictness of transcription control. Derepression of transcription under the lac system is achieved using an inducer molecule such as isopropyl β-D-1-thiogalactopyranoside (IPTG) which is a non-hydrolyzable analogue of allolactose, a hydrolytic product of lactose and the inducer molecule for the lac operon. L-arabinose is the inducer for AraC regulated promoters such as *ara*BAD (Studier et al., 1990, Francis and Page, 2010, Ozturk et al., 2017).

The T7 RNA polymerase promoter coupled with the lac repressor remains the most common transcription control system in many commercially available vectors for protein expression in bacteria. Under this system, gene expression upon induction is driven solely by the T7 bacteriophage RNA polymerase either integrated into the *E. coli* genome or supplied on a separate plasmid under the lacUV5 promoter or any other promoter depending on the strain of *E. coli* used. T7 RNAP binds and transcribes downstream genes about four times faster than bacterial RNAP. Because the T7 RNAP promoter sequence is not recognized by *E. coli* RNAP, basal level expression prior to induction (leaky expression) is curtailed. This is extremely important especially for toxic proteins (Studier et al., 1990, Studier, 2005).

Hybrid promoters such as *tac* which combine elements (-10 and -35) from different operons are also available. Each of these combinations have been optimized to ameliorate leakiness of transcription and to ensure stricter regulation and scalability of protein production. Plasmids also contain a 5'-UTR harboring a Shine-Delgarno (SD) for ribosome binding immediately downstream of the promoter, a multi-cloning site which allows the insertion of the DNA of the gene(s) of interest in frame with the start ATG codon as well as terminator sequence further downstream of the cloned gene. Terminator sequences help to dissociate the ribosome from the mRNA (de Boer et al., 1983, Amann et al., 1988).

4.5 *E. coli* strains for protein production

The variety of proteins required for academic studies and industrial/pharmaceutical applications immediately suggest variety in protein behavior which makes it untenable for a single strain of *E. coli* to be useful for all purposes. *E. coli* strains K12 and B which are among the most well studied bacterial strains undoubtedly have become the most common ancestral strains from which most of the current workhorses of molecular biology and protein production have been derived (Dumon-Seignovert et al., 2004, Jeong et al., 2009, Kim et al., 2017). A B strain derivative, BL21 and its descendants have become among the most popular workhorses for protein expression in bacteria. Genetic modifications to these strains have sought to address problems with proteolytic cleavage of heterologous proteins, stricter transcription control, tRNA codon usage bias, increasing mRNA stability, post-translational disulfide bond formation, immortality to enable expression of toxic proteins, etc. These parameters have implications for protein expression and or yield, which makes it imperative for careful consideration in selection of host strains for protein production experiments (Dumon-Seignovert et al., 2004, Jeong et al., 2009, Kim et al., 2017).

4.6 Media for protein production in *E. coli*

The total amount of recombinantly expressed protein that is accumulated by an *E. coli* host depends on several factors. Under many situations, the yield may be quantified by the specific activity of the enzyme produced, which may also differ under different expression conditions. Among the many factors that impinge on the level of intracellular accumulation of a protein, the constituents of the growth media are perhaps the most important. The total amount of cell biomass obtained positively correlates with the amount of nutrients present in the growth media. Lysogeny broth, generally referred to as LB has become a standard for protein production (Bertani, 2004). The media is composed of tryptone, which supplies peptides/peptones and amino acids, yeast extract, which is a source of amino acids, soluble vitamins and trace metals, and sodium chloride, for providing osmotic balance (Bertani, 2004, Taylor et al., 2017).

The growth of the bacteria under the culture conditions in the laboratory depletes the nutrients over time and results in secretion of bacterial metabolites and by-products into the culture media which changes the amount of dissolved oxygen and pH which may have consequences for the protein being made (Peti and Page, 2007, Sahdev et al., 2008, Francis and Page, 2010, Huang et al., 2012). Various modifications have been carried out on the original recipe of LB to improve yield, solubility, stability and activity of expressed proteins. SOC Media, Terrific Broth, 2xYT and Super Broth are among the most commonly used media for protein expression in bacteria based on variations in the quantities of tryptone, yeast extract and sodium chloride, and supplemented with a combination of additives such as glycerol, divalent metals ($MgSO_4$, $MgCl_2$, $CaCl_2$, $ZnCl_2$, etc), and phosphates as a buffering agent. The starting pH for these media is generally around neutral to slightly alkaline. The impact of each media type on the expression of a protein cannot be accessed *a priori* and so must be experimentally determined (Dumon-Seignovert et al., 2004, Taylor et al., 2017).

4.7 Temperature and inducer concentration

The growth rate of *E. coli* positively correlates with temperature reaching optimal value around 37ºC with the rate of transcription and translation tightly coupled together. At its optimal growth rate, the ribosome translates ~1.0 kDa of polypeptide per second (Lorimer, 1996, Walter et al., 1996). Under induction conditions where most of the transcription and translation machinery is dedicated to the expression of heterologous protein, the cellular environment can easily be overwhelmed due to this high rate of protein synthesis. The concomitant effect is that depending on the type of protein, and its behavior, folding may be compromised leading misfolding, aggregation and off-target effects, which may result in immune-like responses. Lowering the expression temperature to reduce the rate of transcription and translation has therefore been found to be a routine means of improving the biophysical properties of proteins expressed in bacteria such as specific activity, proteolytic stability and solubility While protein expression is routinely carried out between 15ºC and 25ºC, temperatures as low as 6ºC have been used to express soluble proteins (Vasina and Baneyx, 1997, Vera et al., 2007, Song et al., 2012).

Decreasing the rate of protein synthesis, which reduces the rate of intracellular protein accumulation and thereby makes it possible for the cytosolic milieu of *E. coli* to remain sanitized, can also be achieved by modulating the inducer concentration. Low concentrations of IPTG for T7 lac based systems or L-arabinose for the *ara*BAD promoter result in lower levels of repression leading to the isolation of more active and well folded proteins albeit with a reduced yield (Vasina and Baneyx, 1997).

A derivative of the BL21 strain called Tuner (DE3) has been developed in accordance with this general observation which enables the titration of IPTG concentration and thereby enables precise control of the rate of protein expression in response to its cellular behavior (Speer et al., 2019). A variation of this induction mechanism relies an auto-induction of the heterologous protein production by modifying the LB recipe to include glucose, lactose and buffering agents which allows the allolactose produced upon depletion of the glucose to induce the gene transcription. Because the depletion of glucose and the utilization of lactose occur concurrently, the induction process appears controlled with slow rate of protein production achieving similar results to modulating inducer concentration with the added advantage of generally producing high cell biomass since the depletion of glucose is concomitant with high cell density (Studier, 2005).

4.8 Choice of fusion tag

Extraction of target proteins from the mixture of *E. coli* proteins relies commonly on some affinity column chromatography, aided by an affinity tag fused to the protein of interest (Dumon-Seignovert et al., 2004, Taylor et al., 2017). This is achieved by inserting the nucleotide sequences encoding the particular tag in frame with the gene of interest at the stage of cloning. This is particularly useful when the expression level is low even though the traditional method of ammonium sulfate precipitation is useful for highly over expressed proteins (Francis and Page, 2010).

Fusion tags can be either peptides of few amino acids such as His tag and Flag tag, or a protein of considerable size such as green fluorescent protein (GFP), maltose-binding protein (MBP), Halo, small ubiquitin-like modifier (SUMO), and glutathione-S-transferase. Many of these tags serve multifunctional purposes. MBP and GST, which rank among the commonest fusion partners for protein expression in bacteria do not only function as purification tags but also as solubility, stability and expression-enhancing partners. Because they are generally very stable molecules and more soluble than many proteins, their fusion to others can enhance their solubility and stability as well (Li, 2011, Kosobokova et al., 2016, Steinmetz and Auldridge, 2017).

The critical roles played by many of these protein tags in their natural environments makes their sequences optimized for high mRNA stability and high-level translation. Fusing these to target proteins serves as a decoy for improving the translation efficiency of the target proteins. In many instances also, these evolutionarily optimized sequences when fused to partners disrupt secondary structural elements in mRNAs of fusion partners, which hitherto makes translation less efficient.

In general, therefore, attachment of a combination of these tags especially in combination with hexa-histidine preceded by or followed by a protease cleavage sequence to enable the removal of these tags post purification offers a very useful technique in a successful protein expression and purification experiment using an *E. coli* host. The best fusion partner for a particular exercise must also be empirically determined.

4.9 Co-expression with molecular chaperones and protein binding partners

Soluble expression of many proteins in bacteria is achieved only in the presence of molecular chaperones and/or other protein binders (Francis and Page, 2010). Molecular chaperones are proteins conserved across all genera of life which assist in the proper folding of endogenously expressed proteins. Typical examples include GroEL/ES, and the DnaK-DnaJ-GrpE system in *E. coli* with homologs such as the Hsp proteins in other genera (Lorimer, 1996, Todd et al., 1996, Walter et al., 1996, Schlapschy and Skerra, 2011). Their expression is up-regulated under stress conditions which increases the propensity of endogenous protein denaturation thereby helping in the proper folding of endogenous proteins to ensure cell survival. Such stresses include high temperature, low temperature or high acetate build-up during fermentation.

These chaperones generally bind to hydrophobic regions of protein and provide a solvent- inaccessible environment within their barrel shaped cores and prevent them from mis-folding and aggregation (Todd et al., 1996, Walter et al., 1996). Co-expression of molecular chaperones or cold induction of endogenous chaperones with target proteins therefore achieves the same result of increasing the intracellular pool of these proteins, which facilitate, the folding of heterologously expressed proteins.

Multimerization of proteins can frequently bury hydrophobic surface and shield them from solvent exposure. For hetero-multimeric proteins therefore, expression of individual proteins may expose these hydrophobic surfaces to solvent as well as to other proteins in the cytosol of the bacteria, the result of which may be undesirable. Under such circumstances, co-expression of these binding partners may be a useful way of obtaining properly folded and constituted complex (Vasina and Baneyx, 1997, Schlapschy and Skerra, 2011).

Homo-multimerization and stabilization of some protein domains in multi-domain proteins may actually be mediated by another protein or for very hydrophobic proteins, patches of hydrophobic surfaces are exposed in its homo-multimers for other binding partners that can elicit the same undesirable consequences as stated above. Hence expressing these binding partners together can also ameliorate these off-target activities which may negatively impact the success of protein expression and purification in bacteria (Vasina and Baneyx, 1997, Vera et al., 2007).

Also important is toxicity associated with off target signaling activity associated with the expression of some proteins. Kinases and phosphatases, and toxin-antitoxin pairs whose targets are ubiquitously present in all genera of life are typical examples. When expressed in the absence of their authentic targets, these proteins have the tendency to act on their host macromolecular entities producing off target effects which are usually deleterious to the host. These often result in expression debacles which are often times ameliorated by coexpression of the authentic binding partners (Dumon-Seignovert et al., 2004, Peti and Page, 2007, Bird et al., 2016, Taylor et al., 2017).

4.10 Aims of the study

The role of polyproteins in the life cycle of retroviruses remains extremely critical to the fate of the virus. As the source of 'life' for the virus, it remains one of the most vulnerable stages of the viral life cycle where therapeutic intervention may potentially be less susceptible to resistance mutations. Since animals do not produce polyprotein precursors, any drug candidates specific to these polyproteins would likely be very specific requiring low doses which could reduce the possibility of toxicity. Unfortunately, there are no 3D structures of these precursors except a few structures of the Gag which contain more than one domain. The difficulty of expressing and purifying these proteins stably and in soluble form in bacteria, which is a simple expression system available in most laboratories, is largely responsible for this situation. This study therefore seeks to develop a bacterial expression system for soluble expression and purification of PFV Pol, HIV-1 Pol and HIV-1 Gag-Pol constructs for structural and biophysical studies.

4.11 Media Engineering

The type of media used in cultivating bacteria has a strong influence on whether the protein of interest can be expressed as well as the yield. Initial efforts at making polyproteins of HIV-1 origin were beset with numerous problems including proteolytic degradation and inclusion body formation. In an effort to overcome these problems, several approaches were tried including but not limited to (1) using different types of media such as LB, 2xYT, Terrific Broth (TB), Super Broth (SB), Super optimal broth with catabolite repression (SOC), and Auto-induction media (Studier, 2005); (2) screen a library of different tags: GFP, cherry, MBP, Strep, GST, and Sso7d (3); different temperatures: 17 ºC, 25 ºC, 30 ºC and 37 ºC; (4) different isolates of HIV-1, including BH10 and NL4-3 (5); on-column refolding and renaturation of inclusion bodies (6); different types of vectors and promoters: pCDF, pET28a, and pRSF in combination with T7, and TAC promoters; (6) and different strains of E. coli: BL21 DE3 RIL and RIPL (Agilent), Rosetta DE3 (EMD Millipore), BL21 star (Thermo Fisher Scientific), Origami DE3 (Novagen), OverExpress C41 and C43(DE3) (Lucigen), ArcticExpress (DE3) (Agilent).

Unfortunately, none of these showed promise in ameliorating the proteolytic degradation of HIV-1 Pol and PFV Pol polyprotein precursors which were used for these experiments. Having explored quite extensively commonly used methods of enhancing protein expression in bacteria without success, a careful examination of the contents of the media types was undertaken. Even though all these media are based on the original LB recipe (Bertani, 2004) comprising tryptone, yeast extract and NaCl, different amounts are used in different media in addition to additives such as glycerol, NZ-Amine, phosphates, Mg^{2+}, glucose, trace metals and vitamins. Based on these observations, a new media was formulated with the following recipe:

1.5 % (w/v)	Tryptone
1.5 % (w/v)	NaCl
1.5 % (w/v)	NZ-Amine
1.0 % (w/v)	Yeast extract
5 % (w/v)	Glycerol
50 mM	Mg^{2+} (Cl$_2^-$ or SO$_4^{2-}$)

pH 5.5-6.5 (optimal pH used is 6.0)

50 mM phosphate as optional ingredient

The hypotheses underlying the choice of these ingredients were to provide high amount of nutrient to support bacteria growth to high cell density. These rich nutrients were expected to lead to high level of protein expression in bacteria since starvation, which limits expression, would likely not occur under these conditions. Mg^{2+} is known to increase the growth yield of bacteria with a strong propensity to stabilize nucleic acid structure as well as being a co-factor for many enzymes. It therefore remains a popular additive to bacteria growth media (Christensen et al., 2017).

If the hypothesis that in this nutrient rich media, protein yield would be increased, then the amount of mRNA produced should be high as well to lead to high accumulation of proteins. A non-physiologic amount of 50 mM Mg^{2+} was chosen to provide enough stock of the divalent cation needed for these purposes. The impact of pH on the growth of bacteria has been well documented in the literature. By-products of metabolism, either excreted into the extracellular milieu or not, have the potential to alter the pH of the cytoplasm as well as the extracellular environment (Slonczewski et al., 2009).

Living organisms have therefore evolved elaborate mechanisms of pH homeostasis to ensure that fluctuations in the pH of the extracellular environment does not perturb the pH of the cytoplasm. In *E. coli* for example, within an extracellular pH range of 5-9, a cytoplasmic pH of 7.2-7.8 is still maintained under optimal growth conditions. The maintenance of such robust cytoplasmic pH within a very narrow range is aided by the strong buffering capacity of free amino acid pool in the cytoplasm as well as the ionizable groups on proteins and other organic and inorganic metabolites (Slonczewski et al., 1981, Pan and Macnab, 1990, Wilks and Slonczewski, 2007). As a result of this observation, media for bacteria growth and protein expression are generally kept at around neutral pH. For the same reason also, the effect of pH on the fate of heterologouly expressed proteins in bacteria remains unexplored in the literature. Nonetheless the pH of the new media was kept acidic, 5.5-6.5 in the hope that the activity of the protease(s) responsible for the proteolytic degradation of the heterologously expressed proteins that had been seen with other media types would be attenuated and therefore lead to less proteolysis.

4.12 Expression and purification of PFV Pol (PR-RT-IN)

Computation-guided mutagenesis of the PFV PR-RT had led to the identification of two key mutations that reduced the susceptibility of the protein to *in vivo* protease cleavage and enabled the crystallization and structural characterization of the PR-RT as presented in chapter two. The construct which contained an HRV14 3C cleavable hexa-histidine tag (His-tag) for affinity purification using a nickel nitrilotriacetic acid column and the H507D and S584K mutations identified through the ExPASy ProtParam webserver well as the C280S substitution known to reduce aggregation of HIV-1 RT, served as the basis for the cloning of the full-length Pol. Since the structures of the PR-RT CSH mutant and the WT were indistinguishable, it was not anticipated that the structure of the full-length Pol would be perturbed by these mutations.

The IN domain was appended therefore to the PR-RT to generate the PFV Pol CSH-IN which was the initial candidate for testing the expression of these proteins. A deletion of the last 14 residues of IN (CSH-IN delta14) which are disordered in all the structures of PFV IN available in the PDB was also generated since these floppy regions could be inhibitory to crystallization which remained the method of choice in the structural characterization of these proteins. Constructs of the PFV Pol containing N-terminal GFP and mCherry (GFP-CSH-IN, mCherry-CSH-IN) were also generated. The intrinsic fluorescence of these proteins was a visual confirmation of protein expression originating from the transcription and translation of the mRNA of the gene inserted into the vector. They were also useful in designing purification protocols for these proteins (Wilks and Slonczewski, 2007). The efficiency of elution of the proteins from various columns with different buffers containing different amounts of NaCl, glycerol, imidazole etc., was determined by the residual fluorescence on the column after elution.

To verify whether the hypothesis predicated on which the new media was formulated was valid, the CSH-IN construct was expressed using the newly constituted media which will henceforth be referred to as JJH media. Surprisingly, full-length PFV Pol was isolated intact without the concomitant proteolytic breakdown observed using other media. Various constructs of the PFV Pol were expressed and purified to >90% homogeneity as assessed by SDS-PAGE gel (see **Fig. 41**) using Ni^{2+}-affinity purification followed by heparin using the protocol as described in chapter one for the PR-RT.

4.13 Biophysical characterization of PFV Pol

4.13.1 Gel filtration chromatography

The samples of PFV Pol expressed and purified using the JJH media were analyzed by gel-filtration, dynamic light scattering, small angle X-ray scattering and single particle cryo-EM. Following elution of the protein from the heparin column using a buffer composed of 50 mM Tris-Cl pH 8.0, 1.0 M NaCl, 5% glycerol and 1 mM TCEP, the sample was concentrated using Amicon Ultra-15 centrifugal filter units with either 30 kDa or 50 kDa molecular weight cut-off. Concentrated samples were loaded onto a Superose 6 Increase 10/30 gl gel filtration column pre-equilibrated with the buffer, 25 mM Tris-Cl, pH

8.0, 250-300 mM NaCl,1.0 mM TCEP, using an Akta pure FPLC system operated at 0.4-0.5 mL/min. A representative chromatogram and gel of peak fractions are shown in Figure **41**. The retention time of the peak fractions occurred at ~16.0 mL which, based on the calibration curve of the column from the manufacturer, suggested predominant monomeric molecules in solution based on the molecular weight of the monomeric Pol of ~130 kDa. A small population of oligomers of molecular weight larger than 130 kDa can been seen as a shoulder of the main peak. Peak fractions were pooled together and concentrated to 0.5-1 mg/mL, flash frozen in LN2 and stored at -80ºC.

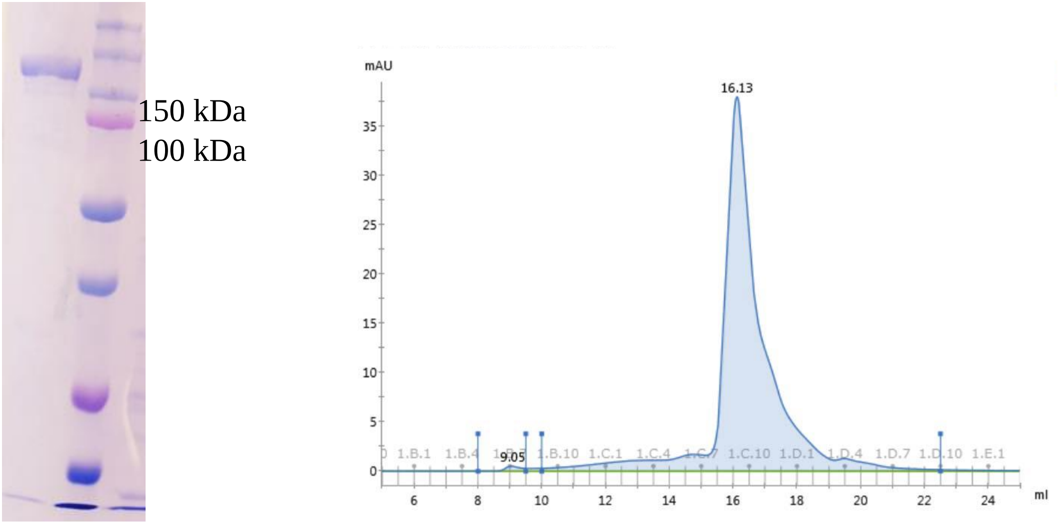
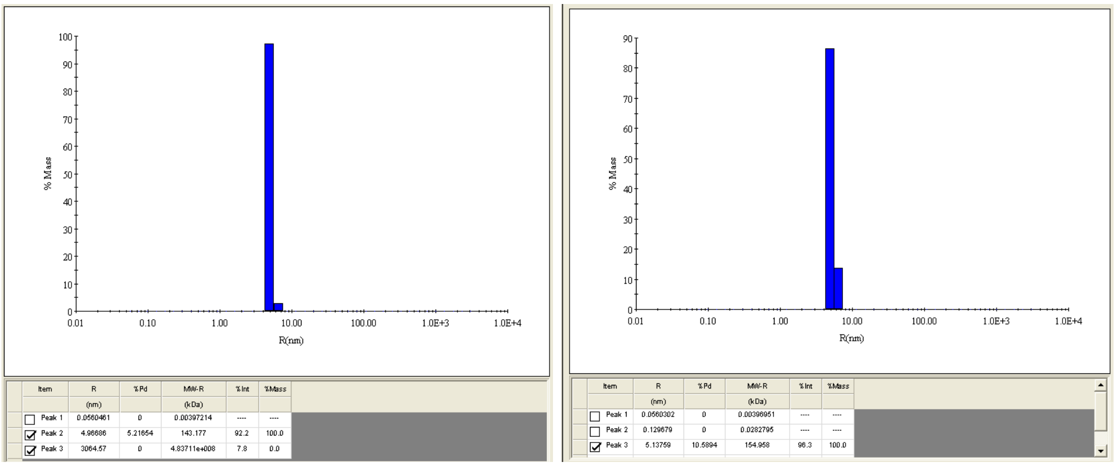


Figure 41: SDS Page gel of PFV Pol (left) before gel filtration. The gel filtration profile run on Superose 6 Increase 10/30 gl is shown on the right

4.13.2 Dynamic light scattering studies of PFV Pol

Dynamic light scattering (DLS) was used to determine the multimeric state of PFV Pol in solution as well as glean information about the hydrodynamic radii of molecules in solution and the polydispersity of the particles. The extent of structural and conformational heterogeneity which is measured by the polydispersity strongly correlates with crystallizability of macromolecular entities (Veesler et al., 1994, Laganowsky et al., 2010).

DLS experiments showed that the average molecular weight of the Pol molecules in solution at a concentration of 0.7 mg/mL was 143-155 kDa. This is consistent with monomeric Pol molecules in solution whose theoretical molecular mass is ~130 kDa as suggested earlier by the gel filtration experiments. The hydrodynamic radius (Rh) of these molecules was determined to be ~5.0 nm. Figure **42** shows representative profiles of the DLS experiments.



Item	R	%Pd	MW-R	%Int	%Mass
	(nm)		(kDa)		
□ Peak 1	0.0560461	0	0.00397214	----	----
☑ Peak 2	4.96686	5.21654	143.177	92.2	100.0
☑ Peak 3	3064.57	0	4.83711e+008	7.8	0.0

Item	R	%Pd	MW-R	%Int	%Mass
	(nm)		(kDa)		
□ Peak 1	0.0560302	0	0.00396951	----	----
□ Peak 2	0.129679	0	0.0282795	----	----
☑ Peak 3	5.13759	10.5894	154.958	96.3	100.0

Figure 42: DLS profile of PFV Pol at 0.7 mg/mL

4.13.3 Small angle X-ray scattering (SAXS)

Further characterization of the PFV Pol samples was carried out using SAXS. The size distribution of molecules in solution as well as their shapes, radius of gyration (Rg) and maximum diameter between scattering electrons in the sample (D_{max}) are some information that can be gleaned from SAXS experiments. A plot of the intensity of elastically scattered X-rays when they pass through materials including macromolecular entities as a function of their scattering angle gives a profile which contains information about the biophysical characteristics mentioned above. The linear region of the profile, the so-called Guinier region (Kikhney and Svergun, 2015, O'Brien et al., 2018, Schneidman-Duhovny and Hammel, 2018) provides information about the radius of gyration.

The determination of the maximum distance between scattering electrons in the sample (D_{max}) is not trivial for flexible proteins due to the high conformational space these electrons occupy (Bernado, 2010, Kikhney and Svergun, 2015, Schneidman-Duhovny and Hammel, 2018). A radius of gyration of ~51 Å and a D_{max} of 185 Å was determined for these samples. The pair distance distribution function (Pr) and the SAXS profile is given in Figure **43** while the envelope calculated ab initio is given in Figure **44**.

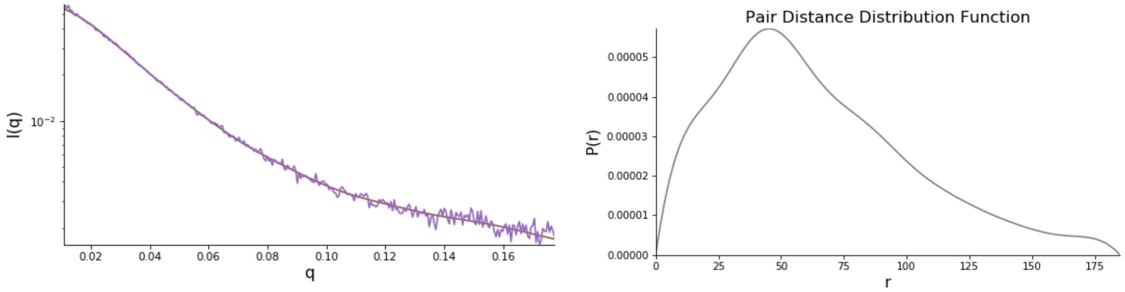


Figure 43: SAXS curve (left) and P(r) function curve of PFV Pol

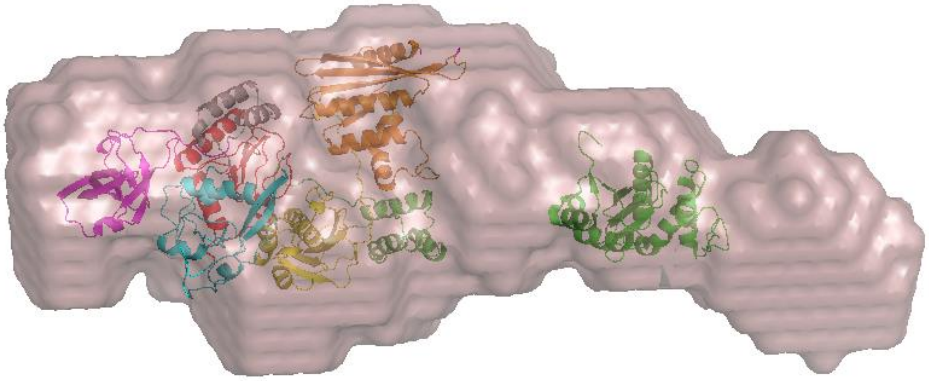


Figure 44: SAXS envelope of PFV-Pol fit with the crystal structure of PFV PR-RT and IN CCD, PDB 2X6N

The fit of the crystal structure of the PR-RT and the CCD domain of the IN domain in crystal leaves enough room for the other domains of IN in PFV (NTD, NED and CTD). The envelope and the structural fits suggest that while the RT portion may be slightly more rigid, the peripheral regions of the structure, PR and IN remain very flexible. This conformational flexibility reflects in the apparent bigger envelope than the crystal structures of the monomeric units. Furthermore, the shape of the P(r) function and how it tails off at larger r (**Fig. 43**) is characteristic of flexible and elongated proteins. The multi-domain nature of the protein is also seen in the bumps in the P(r) distribution plot at high r. This is situation arises because the scattering profile of a flexible protein as measured in a SAXS experiment is an average of all the possible conformations coexisting in solution (Bernado, 2010, Kikhney and Svergun, 2015).

4.13.4 Optimization of conditions for cryo-EM data collection for PFV Pol

Single particle cryo-EM analysis entails imaging in an electron microscope of vitrified biomolecules hydrated in a buffered solution and embedded on grids in random orientations (Czarnocki-Cieciura and Nowotny, 2016, Thompson et al., 2016, Murata and Wolf, 2018). To obtain the best datasets for structure determination, the imaging conditions must be optimized. Optimization requires manipulation of macromolecule sample and concentration, type of grid and glow discharge time, nature of buffer and additives, blotting force and time, as well as temperature. In some cases, the blotting method may also need to be optimized to ensure success. Cryo-EM structural analysis relies on the contrast between biomolecules and the background constituted by different kinds of scattering in the microscope as well as scattering from the electrons in the buffer in which the biomolecule is suspended (Czarnocki-Cieciura and Nowotny, 2016).

High concentration may lead to aggregation of the biomolecules on the grid or produce overcrowded grids where particles of the biomolecule may be overlapping, which makes it difficult to distinguish individual particles. A low concentration on the other hand, may provide very few particles per field of view requiring very long periods of data collection in the best-case scenarios. For flexible and small proteins, low concentration could lead to blurring of the signal of the protein relative to the background making it difficult to identify true particles from background artifacts. The materials used in making grids impart physical properties to them that influence how biomolecules interact with them. Strong interactions between the grid material may lead to preferred orientation of the biomolecules on the grid while strong interactions between the carbon support for the grids may lead to particles not partitioning into the holes on the grid. Common types of materials using in making grids for electron microscopy imaging include copper, gold, carbon and molybdenum. These grids come in different forms with different designs and different hole sizes (Czarnocki-Cieciura and Nowotny, 2016, Thompson et al., 2016).

Because the grids are hydrophobic in nature, they are glow discharged to make them hydrophilic before the application of biomolecules to their surfaces. Different materials require different glow discharge times to ensure uniform distribution of hydrophilicity on the surface of the grids which impacts the uniformity of particle distribution on the grids (Frank, 2016, Thompson et al., 2016). For example, gold grids require generally longer glow discharge times than copper or molybdenum. Such charge distribution is also affected by the nature and size of the holes on the grids as well. Hydrophilicity of the grids may also be modified by coating the grid with a single layer of graphene oxide (GO). Since the graphene oxide lattice underneath frozen-hydrated biomolecule is very unique, it is easily discernible and contributes less to background. The uniform nature of the charges on the GO surface also lead to a uniform spread of particles on the grids (Glaeser et al., 2016, Martin et al., 2016, Palovcak et al., 2018).

Biomolecular entities contain predominantly C, H, O, N, P and S, the same elements present in the most biological buffer systems in which these molecules are kept. While scattering from the background is inevitable, high concentrations of buffers, salts and glycerol as well as detergents and other additives, which are required by biomolecular entities to remain soluble, will also reduce the contrast between the biomolecules and the background in electron micrographs (Glaeser et al., 2016, Thompson et al., 2016). Optimal concentrations should therefore be chosen to ensure maximization of contrast in the images which will influence the final resolution of the structures of biomolecules that can be obtained while making sure the biomolecules are not negatively impacted. During freezing of grids for cryo-EM imaging, excess volume of biomolecular sample is applied to the glow-discharged or GO-coated grid. To obtain a thin layer of buffer transparent to the electrons used during imaging, excess buffer must be blotted away. The length of time and force used therefore has consequences for how thin the ice obtained after vitrification would be. Optimization of these parameters is therefore required. Manual blotting, blotting using automated cryo plungers, Vitrobot (Thermo Fisher Scientific) or EM GP (Leica) can produce different results because of the mechanics of how each of these systems encountered in the course of these studies is designed to function (Iancu et al., 2006).

After series of trials, 0.4-0.8 mg/mL of protein in 25 mM Tris-Cl, pH 8.0, 200-300 mM NaCl, GO-coated Quantifoil R 2/2 Cu or Au grids of 200-300 mesh size, blot time of 4.5-5 seconds before plunge freezing in liquid ethane on Vitrobot automated cryo-plunger, using 3.5 µL of protein solution were found to be optimal for obtaining good contrast images suitable for data collection. A representative micrograph collected on the Rutgers

Talos Arctica 200 keV equipped with a K2 direct electron detection camera is shown in Figure **45**.

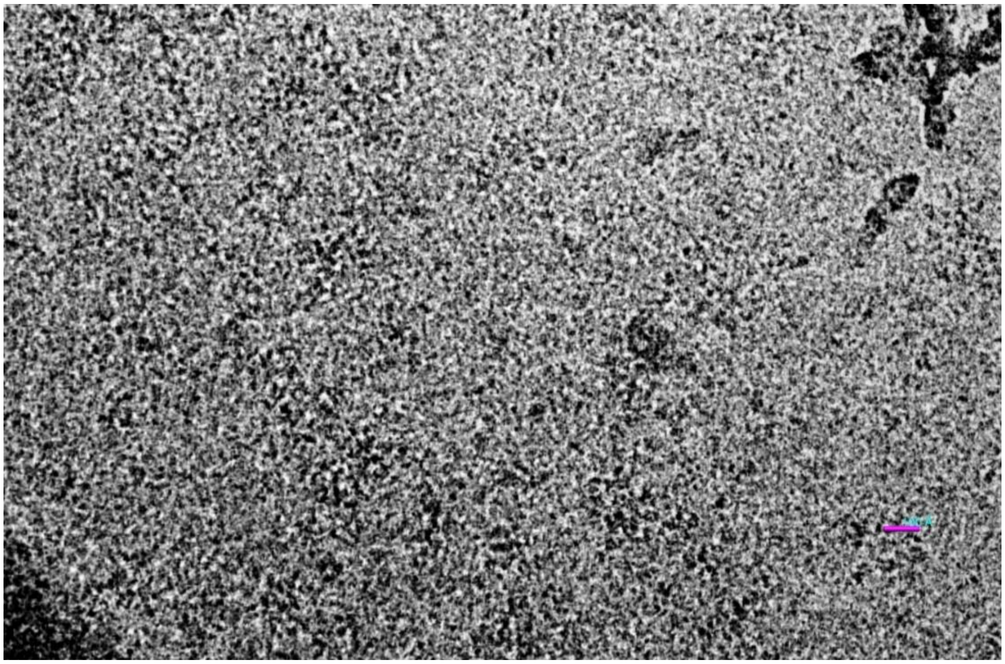


Figure 45: Example micrograph of the PFV Pol collected on Talos Arctica 200 keV electron microscope at Rutgers University

4.14 Expression and purification of HIV-1 Pol, Gag-Pol delta-IN and full-length Gag-Pol

Buoyed by the success of the JJH media in facilitating the expression and purification of PFV Pol, various constructs of HIV-1 Pol and Gag-Pol were tried to verify if the new media would have the same positive impact on these proteins that had been previously tried but failed. Constructs designed included a 3C cleavable N-terminal Sso7d with varying length of the p6* region immediately downstream of the PR in the genome of HIV. The initial construct contained 23 residues of the p6* which was the construct obtained from our collaborator Mamuka Kvaratskhelia (University of Colorado, Denver).

These constructs were designed to recapitulate early proteolytic products of HIV-1 Gag-Pol during maturation. Other constructs made included replacement of the p6* sequence of the BH10 HIV-1 isolate with a consensus Ser-Asn-Leu (SNL) as well as truncation of the CCD and CTD of the IN domain while maintaining the Sso7d. The choice of Sso7d had been based on the fact that it had been found to be important for solubilizing HIV-1 IN which enabled the structure of HIV-1 intasome to be solved (Passos et al., 2017). To further improve solubility and stability to proteolysis, the RT/IN junction was mutated from LF to DD as well. Mutations in HIV-1 IN shown previously to enhance the solubility of IN which enabled structural studies (Jenkins et al., 1996, Chen et al., 2000) and crystallization were also introduced.

Using the JJH media, expression and purification of the HIV-1 Pol for all the constructs described above were successful, enabling the structure of the HIV-1 Pol to be solved using single particle cryo-EM, the details of which are presented in chapter five. Coexpression of the Pol with the IN-binding domain of lens epithelium derived growth factor (LEDGF), fused to the C-terminus of the maltose-binding protein (MBP), MBP-IBD, enabled the pull-down of the Pol and offered a new avenue of achieving even better purity of the Pol polyprotein. LEDGF is known to bind the preintegration complex (PIC) (Singh et al., 2015, Li et al., 2015, Metifiot et al., 2016, Passos et al., 2017) and tether it to chromatin to enable the integration site to be selected (see chapter five for details).

Ultimately a more comprehensive understanding of retroviral polyprotein processing would require structures of the full-length Gag-Pol. Having succeeded therefore in expressing and purifying HIV-1 Pol that enabled its structure determination by single particle cryo-EM (see chapter five), several constructs of the HIV-1 Gag-Pol, using the

NL4-3 isolate obtained from our collaborator Alan Engelman (Dana-Farber Cancer Institute, Harvard University) were also made and tried with the new JJH media. To enable tandem affinity purification which may improve the purity of the proteins expressed, constructs of the Gag-Pol with a 3C-cleavable MBP fused to the N-terminus with a C-terminal hexa-histidine tag were constructed. It was envisioned that this would enable initial nickel affinity purification and subsequent amylose resin purification. Cleavage products lacking the C-terminal His-tag would be expected to flow through the nickel column while products lacking the MBP would also flow through the amylose column and therefore lead to purer protein preparations.

The RT/IN junction LF to DD mutation was maintained in the constructs of the full-length Gag-Pol since it had improved the solubility of the Pol samples. Constructs containing solubility mutations in IN as described for the Pol were made in the hope of increasing the solubility of the protein as well as PR active site D25A and D25N versions. The D25A was originally present in the construct obtained from our collaborator. These mutations render the PR inactive as it would undergo self-proteolysis if the protein is expressed as with an active protease in these constructs. Co-expression of the Gag-Pol with MBP-IBD was also explored since it had enhanced the expression and the purification of HIV Pol.

To further simplify the system and improve the biophysical behavior of the protein in solution, an IN-deletion mutant was also made while maintaining the tandem purification tags strategy. If the hypothesis that the IN contributes so much to the insolubility of the protein is correct, then the delta IN construct should behave much better to enable biophysical characterization. Consistent with the positive impact of the new media in enabling the expression and purification of PFV and HIV-1 Pol, these Gag-Pol constructs were also successfully expressed and purified. Representative SDS-PAGE gels are shown in Figure **46**.

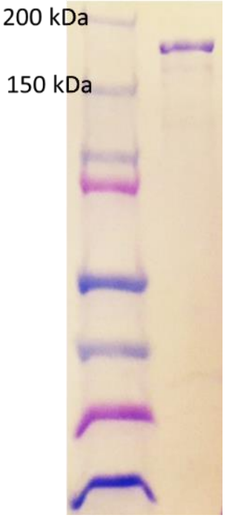


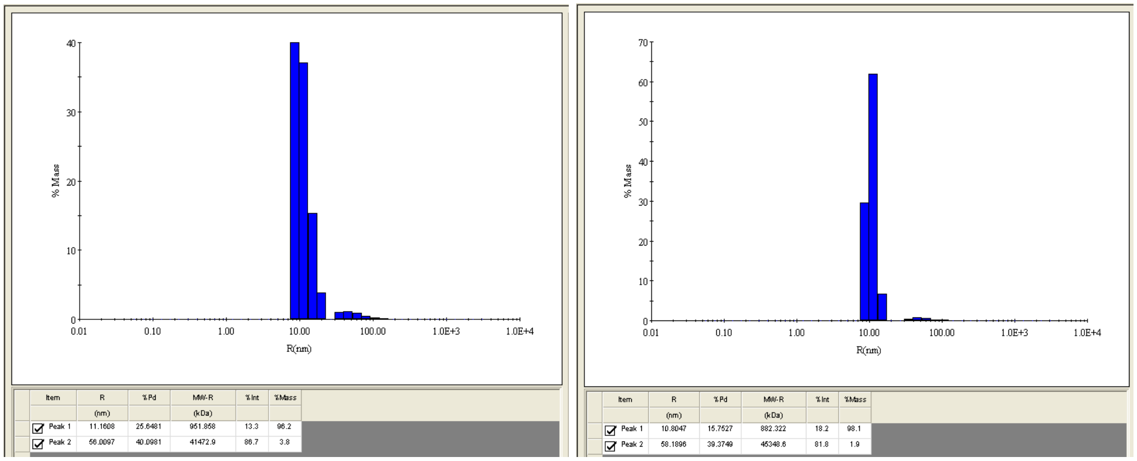
Figure 46: SDS-PAGE gel of HIV-1 Gag-Pol expressed in E. coli

Consistent with the earlier observation, the full-length Gag-Pol could not be concentrated beyond 0.1 mg/mL without aggregation in the presence of 400 mM NaCl. The poor solution behavior made biophysical characterization of the purified protein more challenging and therefore required further optimization.

4.14.1 DLS studies of HIV-1 Gag-Pol + MBP-IBD

DLS studies were carried out to assess the sizes of molecules in solution as well as the extent of polydispersity of the Gag-Pol + MBP-IBD. At a Gag-Pol concentration of about 0.1 mg/mL, the average apparent molecular weight of molecules in solution ranged from 740-950 kDa. Assuming a 1:1 ratio of Gag-Pol to MBP-IBD, this would suggest that a tetrameric arrangement predominated in solution since a Gag-Pol monomer has a molecular weight of ~160 kDa while that of MBP-IBD is ~52 kDa.

The hydrodynamic radii of these molecules were about 11 nm with polydispersity ranging from 15-26% (**Fig. 47**).



Item	R	%Pd	MW-R	%Int	%Mass
	(nm)		(kDa)		
☑ Peak 1	11.1608	25.6481	951.858	13.3	96.2
☑ Peak 2	56.0097	40.0981	41472.9	86.7	3.8

Item	R	%Pd	MW-R	%Int	%Mass
	(nm)		(kDa)		
☑ Peak 1	10.8047	15.7527	882.322	18.2	98.1
☑ Peak 2	58.1896	39.3749	45348.6	81.8	1.9

Figure 47: DLS profile of Gag-Pol-MBP-IBD complex

4.14.2 Cryo-EM Screening of HIV-1 Gag-Pol + MBP-IBD

In an effort to gain some understanding of how these proteins behaved during freezing and cryo-EM imaging, efforts were initiated to optimize the conditions for imaging these samples. The Gag-Pol + MBP-IBD construct was chosen largely in part due to the fact that it could be purified to homogeneity much easier and concentrated to about 0.2 mg/mL without the protein crashing out of solution. A representative SDS-PAGE gel of the protein preparation is given in Figure **48** together with an example micrograph (**Fig. 49**). The samples were frozen on Quantifoil R 2/2 Cu or Au grids of 200-300 mesh size and blotted for 5 seconds before being plunge frozen in liquid ethane to vitrify. Most of the protein was found to be aggregated on the grids but some individual molecules could be observed as well. Conditions of obtaining less aggregated molecules with more particles per field of view are currently being optimized for subsequent data collection. Specifically, the amount of salt and glycerol as well as detergents such as LDAO and DDM are being used to reduce the aggregation tendency of these proteins and to produce grids suitable for single particle structure determination.

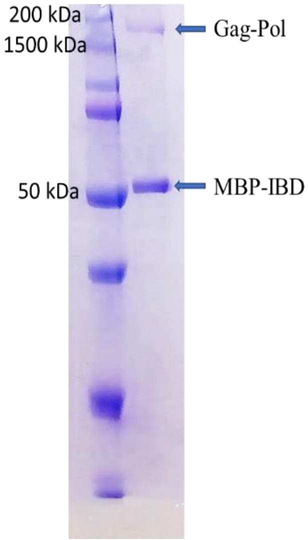


Figure 48: SDS-PAGE gel of HIV-1 Gag-Pol coexpressed with MBP-IBD in E. coli

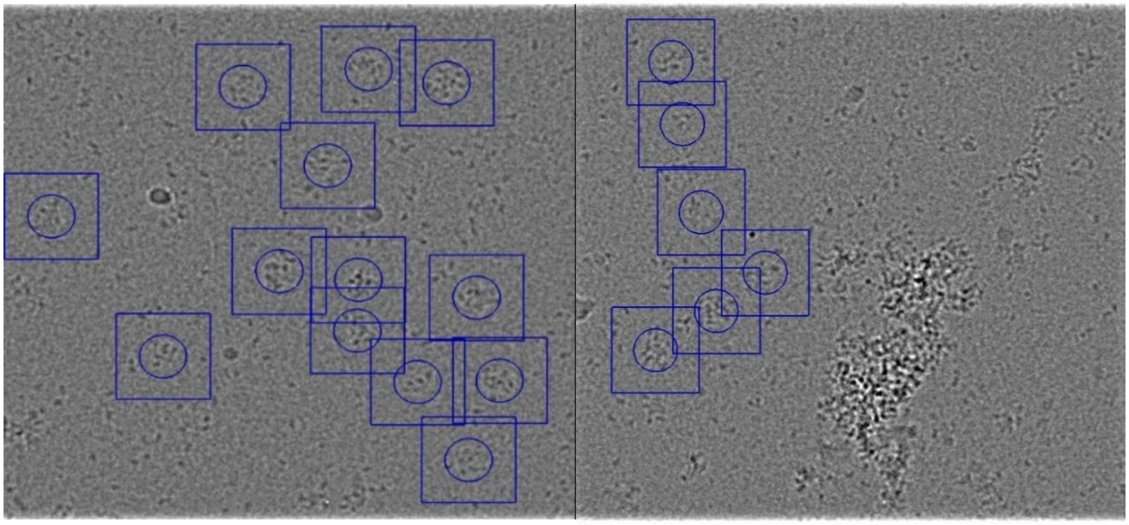


Figure 49: HIV-Gag-Pol + MBP-IBD images from screening at Rutgers collected by Jason Kaelber. Circles are 315 Å in diameter, squares have 630 Å edge length.

14.4.3 Biophysical characterization of MBP tagged Gag-Pol delta IN

The presence of large aggregates on the EM grids of the Gag-Pol + MBP-IBD complex and the generally limited solubility of the Gag-Pol prompted the search for alternative constructs that may behave better in solution and enable structure solution. IN-deleted constructs of the Gag-Pol were created while optimization of the Gag-Pol grids was being carried out. It was surmised that the aggregation is predominantly caused by the IN domain which has poor solubility behavior and had been difficult to structurally characterize compared to the other proteins of HIV.

The 3C-leavable N-terminal MBP and the C-terminal His tag (deca-histidine or His10X) allowed a C-terminal purification using the nickel column, followed by heparin and MBP trap columns respectively. The protein eluted off of the heparin column in 1.0 M NaCl was concentrated and buffer exchanged into a 25 mM Tris-Cl, pH 8.0 and 300 mM NaCl using a Superose 6 Increase 10/30 gl gel filtration column. Two main peaks eluted off the column with retention volumes of 11.61 mL and 14.42 mL. A representative profile of the gel filtration is shown in Figure **50** together with accompanying SDS-PAGE gel.

The gel showed that these two peaks contain the same molecular weight protein. This suggested that there could be different oligomeric species in solution. On the other hand, the two peaks could also correspond to different conformations of the same oligomer present in solution. A compact oligomer with a small surface area would likely elute at a longer retention time than an oligomer in an elongated state.

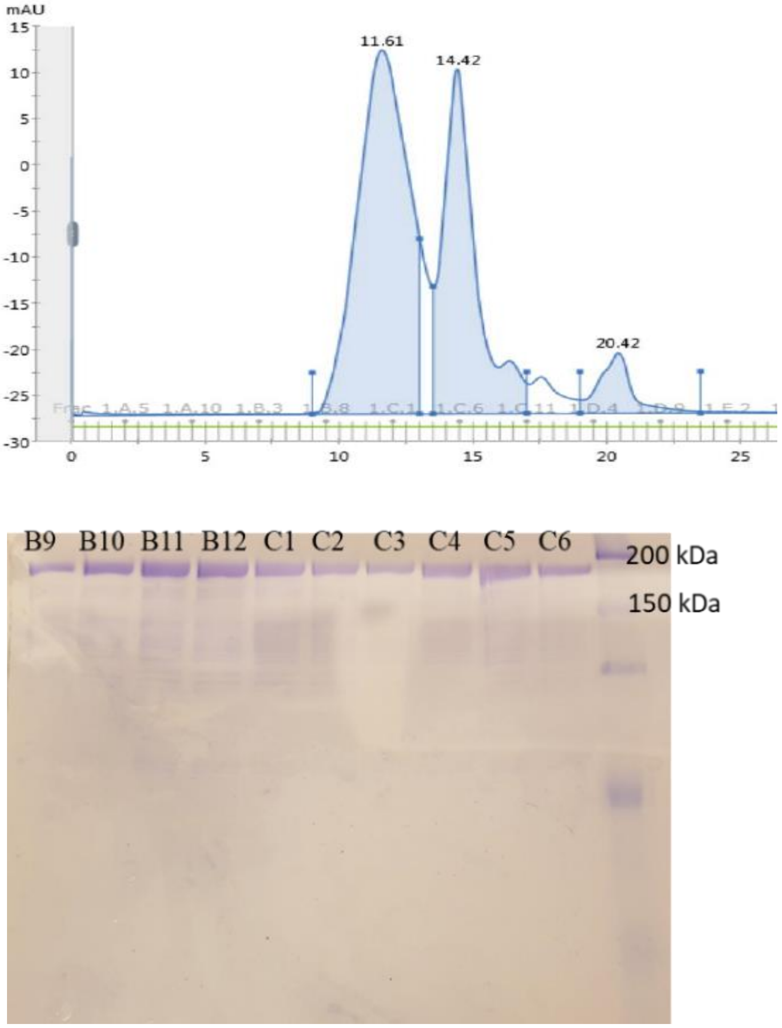
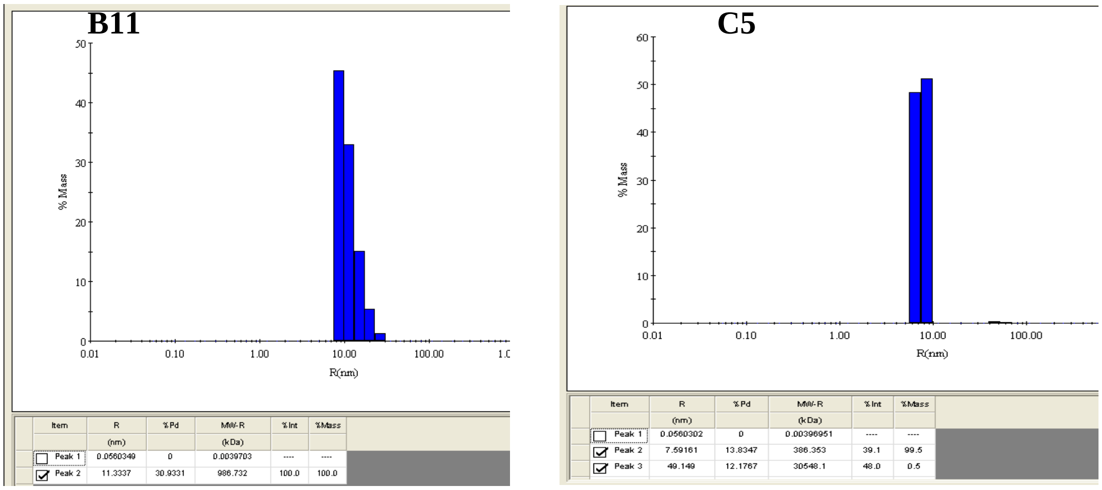


Figure 50: Gel filtration profile (above) and accompanying SDS-PAGE gel of HIV-1 Gag-Pol delta IN

14.4.4 Dynamic light scattering studies of HIV-1 MBP-Gag-Pol delta IN

To assess the size and nature of particles in the solution of MBP-Gag-Pol delta IN, a DLS analysis was carried out on peak fractions. At a concentration of 0.1 mg/mL, fraction B11 representing the first peak (11.61 mL), the DLS analysis showed that the average apparent molecular weight of particles in solution was ~1.0 MDa. Since the molecular weight of the Gag-Pol delta IN construct is ~160 kDa, this peak would correspond to a hexamer of proteins in solution. The hydrodynamic radii of these molecules as obtained from these studies was about 11 nm with polydispersity greater than 30%. Peak fraction C5 also at ~ 0.1 mg/mL, on the other hand, showed average molecular weight of particles in solution as 386 kDa with hydrodynamic radii of 7.6 nm and polydispersity of ~14%.

This molecular weight is consistent with a dimeric oligomer of the protein in solution. The low polydispersity of this peak compared to the first is consistent with the argument that peak two is more compact with less conformational variability compared to the first peak. It is not inconceivable that the oligomeric states of these proteins are the same with two very distinct conformations-compact and elongated. The DLS experimental determination of molecular weight assumes globularity of proteins in solution and therefore a high deviation from such globularity could result in significantly different estimates of the molecular weight. Figure **51** shows representative profiles of the DLS experiments.



Item	R	%Pd	MW-R	%Int	%Mass
	(nm)		(kDa)		
□ Peak 1	0.0560349	0	0.0039703	----	----
☑ Peak 2	11.3337	30.9331	986.732	100.0	100.0

Item	R	%Pd	MW-R	%Int	%Mass
	(nm)		(kDa)		
□ Peak 1	0.0560302	0	0.00396951	----	----
☑ Peak 2	7.59161	13.8347	386.353	39.1	99.5
☑ Peak 3	49.149	12.1767	30548.1	48.0	0.5

Figure 51: DLS profiles of peak fractions from the gel filtration Gag-Pol delta IN

14.4.5 Cryo-EM Screening of HIV-1 MBP-Gag-Pol delta IN

Peak fractions from the gel filtration experiments were frozen on Quantifoil R 2/2, 300 mesh glow-discharged Au grids and blotted for 5 seconds before being imaged in a

Talos Arctica microscope. Unfortunately, the grids from elution peak2 (14.41 mL) devitrified during transfer to the grid cassette for imaging. The particles on the micrographs from the elution peak1 (11.6 mL) did not show any obvious aggregation at the 0.1 mg/mL concentration but like the Gag-Pol + MBP-IBD, the concentration was too low for optimal data collection. The sizes of the particles on the grids (**Fig. 52**) did not appear to be as large as the DLS experiments had suggested. It is therefore highly likely that these particles are an elongated form of the peak2, which is likely more compact. Further optimization is currently being pursued to optimize the concentration of these samples for cryo-EM data collection.

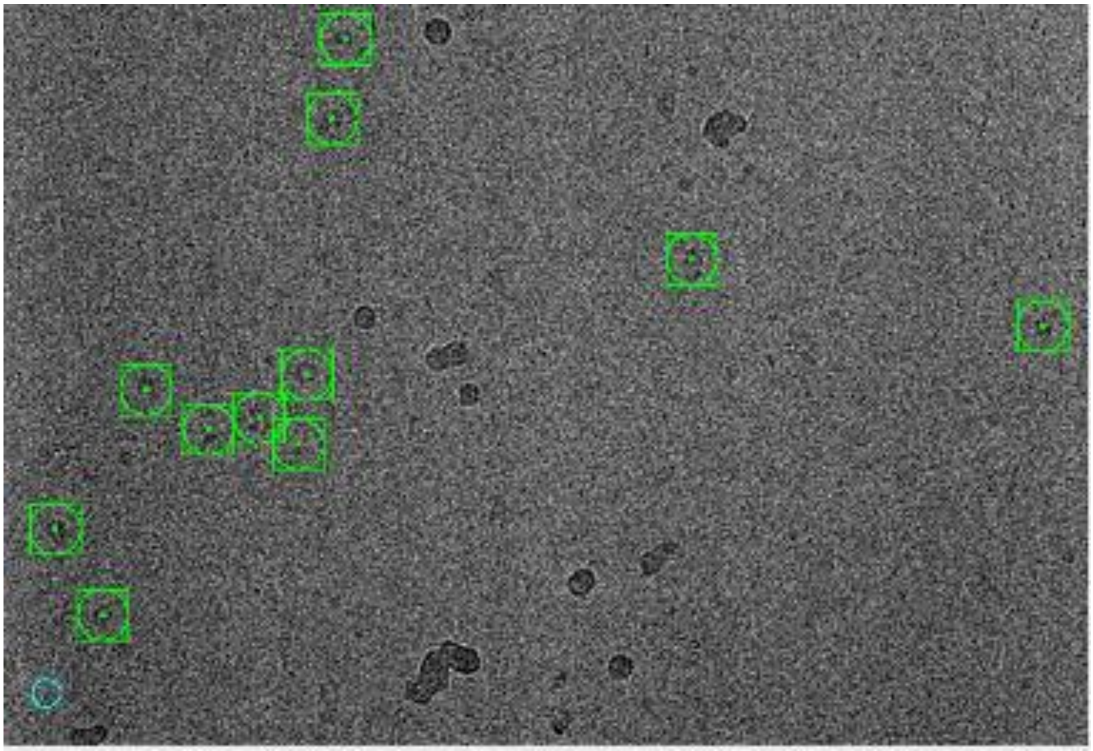


Figure 52: Example micrograph from the EM screening of Gag-Pol deltaIN

14.5 Conclusions

Many attempts of expression and purification of prototype foamy virus (PFV) and HIV-1 polyprotein precursors in bacteria were unsuccessful using conventional methods, which included variation of expression and purification tags, changes in expression temperature, inclusion body expression followed by denaturation and renaturation, as well as using different bacterial cell lines. Instead, novel media designs reconstituted from existing recipes for protein expression in bacteria proved effective. These media containing at least 50 mM Mg^{2+} and pH 6 or below, remarkably ushered in an era of stably expressing and purifying polyprotein precursors from PFV and HIV-1. This study demonstrates that various constructs of these polyproteins, none of which had been successfully expressed and purified in multi-milligram amounts from any source to the best of my knowledge and to the level of purity shown in this work was made possible by this new media. This breakthrough enabled the structure determination of HIV-1 Pol by single-particle cryo-EM, details of which are presented in chapter 5 of this thesis. It is tempting to speculate that this unusual media condition may have broad implications for ameliorating the problem of proteolytic degradation of heterologously expressed proteins in bacteria.

Biophysical analysis of the PFV Pol by gel filtration, DLS and SAXS indicate a predominantly monomeric protein of approximately 5.0 nm hydrodynamic radius in solution. The SAXS studies suggested that the radius of gyration (Rg) of the Pol molecules is about 51 Å. The Rh values from the DLS and the Rg values from the SAXS are very consistent with each other. The pair distribution function (P(r)) shows that the D_{max}, which represents the maximum particle dimension in solution, is 185 Å. The shape of the P(r) also suggests a multi-domain protein with flexible linkers in solution. The calculated envelope from the SAXS curve fits the crystal structure of the PR-RT and the CCD of IN with enough room for the rest of the domains. The apparent sizes of the PR and the IN from the SAXS envelope are bigger than the monomeric entities which is consistent with the shape of the P(r) curve.

Samples of HIV-1 Gag-Pol have also been expressed and purified either alone or co-expressed with MBP-IBD. DLS analysis of the Gag-Pol + MBP-IBD, which is the most soluble form of the protein in 400 mM NaCl, suggests that the average molecular weight of molecules in solution range from 740-950 kDa with hydrodynamic radii of ~11 nm. This is consistent with tetramers of Gag-Pol + MBP-IBD molecules in solution. Initial screening of grids on a Talos Arctica microscope shows many protein aggregates on the grids. However, isolated particles were also observed whose sizes may be consistent with these DLS data. An MBP-tagged Gag-Pol with the IN domain deleted was found to be a mixture of hexamers and dimers on a Superose 6 Increase gel filtration column as well as DLS studies. The first peak was found by DLS experiments to contain molecules of about 11 nm Rh with polydispersity greater than 30% while peak2 contained molecules of Rh = 7.6 nm and polydispersity of ~ 14%. Cryo-EM images of peak1 were however inconsistent with hexameric entities in solution suggesting that the two peaks may represent elongated and compact conformations of molecules with the similar molecular weight.

14.5 Materials and Methods

14.5.1 Protein expression and purification

PFV Pol constructs cloned into a pET28a vector or HIV-1 Gag-Pol constructs cloned into pET28a or pCDF vectors were transformed into BL21 DE3 CodonPlus RIL cells. Several colonies were selected and inoculated into 100 mL of overnight culture in a 500 mL Erlenmeyer flask containing 50 µg/mL kanamycin or streptomycin and 34 µg/mL of chloramphenicol and shaken at overnight at 37ºC. Media for overnight culture composed of 1.5% tryptone, 1.0% yeast extract, 1.5% NaCl, 1.5% NZ Amine, and 50 mM $MgSO_4$ at pH 6.5. The 100 mL culture was diluted into 1 L of media at pH 6.0, supplemented with kanamycin or streptomycin and allowed to grow to an O.D of 2-2.5 at 37 ºC before being transferred to a shaker pre-cooled to 15 ºC and allowed to grow for at least 1 hour. Induction of protein expression was carried out by the addition of 1 mM IPTG and culture allowed to grow for at least 17 hours. 50 mM phosphate buffer at pH 6.0 may be added to the media as a buffering agent from the beginning or when the culture is transferred to 15ºC prior to induction.

Cells were harvested by spinning down the culture at 4000Xg for 30 minutes, resuspended in 100 mM phosphate or Tris-Cl buffer at pH 8.0-8.5 supplemented with 600 mM NaCl, 0.5% Triton X-100 or 10 mM CHAPS or 2 mM LDAO, 10% glycerol, 30 mM imidazole, and 1 mM TCEP at a minimum ratio of 10 mL/g of cells on ice. 1 mM PMSF, and 1 µM each of pepstatin A and leupeptin were added to uniformly homogenized cells and sonicated for at least 10 minutes with 30 seconds pulse and pause cycles on ice. The cellular debris were spun down at 38,000x g for 30 minutes and the supernatant loaded onto a nickel gravity column pre-equilibrated with the resuspension buffer.

Column was subsequently washed with 10-20 column volumes (CV) of resuspension buffer followed by a 10-20 CV of high salt buffer wash containing 1.5 M NaCl in the resuspension buffer. The high salt wash is followed by chaperone wash of at least 10 CV. Chaperone wash buffer contains 5-10 mM ATP, 5 mM MgCl2, and 50 mM imidazole in the resuspension buffer. A 2 CV wash with the resuspension buffer was carried out after the chaperone wash to remove all the chaperone wash buffer before protein is eluted with at least 4 CV of 80 mM Tris pH 8.0, 600 mM NaCl, 500 mM imidazole, and 10% glycerol. Eluted protein is supplemented with 2 mM TCEP, diluted 2-fold with water and loaded onto a 5 mL HiTrap heparin column pre-equilibrated with 30 mM Tris-Cl pH 8.0, 300 mM NaCl, 5% glycerol, and 1 mM TCEP using an FPLC. The column is washed with this buffer until the background UV absorption is negligible. Elution of protein from the heparin column was carried out by washing the column with the wash buffer containing 1.0 M NaCl. Eluted proteins were concentrated and injected onto a Superose 6 Increase 10/30 gel filtration column pre-equilibrated with 20 mM Tris-Cl pH 8.0 and 250-400 mM NaCl. Fractions containing pure protein were pooled together, concentrated to 0.1-0.5 mg/L, flash frozen and stored in -80 ºC.

To cleave the tag on the protein, HRV14 3C protease prepared in-house is added to the protein at a ratio of 1:20 after the heparin step, diluted with buffer to ensure NaCl concentration is about 300 mM and kept on ice or at 4 ºC overnight. Cleaved protein is passed through a nickel column to remove His-tagged protein, and the flow-through re-purified on the heparin and gel filtration column as described above before storage. For constructs co-expressed with MBP-IBD, or tagged with MBP, glycerol was removed from the buffers during elution from the heparin column. Proteins were then eluted from the heparin using Tris buffer containing 1.0 M NaCl directly onto an MBPTrap HP column prepacked with Dextran Sepharose, washed with this buffer until the UV base line is negligible and eluted with 30 mM Tris pH 8.0, 600 mM NaCl, and 10 mM maltose before the gel filtration step.

14.5.5 SAXS data processing

SAXS data was collected at the Cornell High Energy Synchrotron Light Source. Data file averaging and buffer subtraction as well as Guinier and Gnom analysis for the determination of Rg and D_{max} were carried out using the BioXTAS RAW software package (Hopkins et al., 2017). *Ab initio* envelope reconstructions were carried out using DAMMIF on the ATSAS online web server (Petoukhov et al., 2012, Franke et al., 2017).

CHAPTER FIVE: Cryo-EM Studies of HIV-1 Pol Reveals Mature-Like Heterodimer of RT: Implications for Protease Activation

Synopsis

The production of vital structural and enzymatic proteins as long precursor polypeptides is a hallmark of many pathogens including RNA viruses and retroviruses. This ensures genetic economy in viral genomes that are often very limited in size. Subsequent cleavage of these polyproteins into mature entities are aided by cellular host factors as well as proteases encoded by the pathogens. For HIV, the structural proteins, Gag (matrix-MA, capsid-CA, nucleocapsid-NC, and space peptide 1-SP1) and the enzymatic proteins, Pol (protease-PR, reverse transcriptase-RT, and integrase-IN) are produced from the same mRNA transcript in the form of Gag and Gag-Pol polyproteins at an effective ratio of 20:1 through a programmed -1 ribosomal frameshift at the 3'-end of the *gag* gene which inserts the *pol* transcript in frame with the *gag*.

Even though some variability exists, this manner of protein production is highly conserved in all retroviruses except the prototype foamy virus (PFV), where separate mRNAs are used to make the Gag and Pol polyproteins. In retroviruses, the functionally homodimeric aspartic protease encoded by the *pol* gene is responsible for all the cleavage events that ultimately convert the polyprotein precursors into mature enzymes. This well-choreographed process is controlled temporally and spatially with different kinetics for each cleavage, which allows the polyprotein precursors to persist for several hours until maturation is completed. During activation, which occurs during or immediately after budding, two Gag-Pol polyproteins dimerize which also brings together monomeric PRs embedded in each of them. This enables each monomer to provide a catalytic aspartate for cleaving peptides. Gag-Pol dimerization is therefore concomitant with cleavage of peptides initially distal to the N-terminus of PR, effectively liberating Pol from Gag before subsequent cleavages complete the process. The architecture of the PR embedded in Gag-Pol is likely different from the mature enzyme since it is less sensitive to active site PR inhibitors compared to the mature PR. The structural underpinnings of PR activation and the architecture of the Pol polyprotein precursor as well as the puzzle of why only one RNase H domain of RT is cleaved to form the heterodimer remain poorly understood.

To understand early events of maturation of HIV-1, a bacterial expression system that yields multiple-milligram quantities of purified Pol polyproteins has been developed. There are no previous reports of efficient production of high-quality Pol in amounts suitable for biophysical studies to the best of my knowledge. Novel media were formulated with unusually high Mg concentrations and low pH which enabled various constructs of HIV-1 Pol containing inactive PR (D25A) and varying lengths of the p6* region to be expressed and purified from *E. coli*, and their structures solved using single particle cryo-EM.

The structures show that the RT portions in the Pol exhibit a dimeric organization similar to that of the mature asymmetric heterodimer. This RT dimerization provides a platform which draws the two PR monomers at the N-termini of RT into close proximity. These proximal PRs are dimerization competent with a significant class of these structures showing density consistent with this architecture. These structures are consistent with the earlier observation that the p6* region of Pol perturbs the ability of PR to form a stable tertiary structure associated with high enzymatic activity and thereby preventing premature PR activation—an observation which invokes the existence of alternative mechanism(s) of mitigating this effect since it is highly unlikely that the destabilization effect of p6* would be relieved by itself. While the IN domain remains largely disordered in our structures, it can bind the IBD domain of LEDGF/p75 when coexpressed with it and pull it down. This study therefore helps to explain the key role of RT in PR activation initially suggested through biochemical and virological studies and confirms the formation of RT p66/p66 homodimer in a p66/p51 heterodimer-like configuration as a prelude to RNase H cleavage which effectively makes only one RNase H domain available for cleavage.

*Valuable contributions from other people culminated in this work. Specifically, all the Western blots reported in this work were done by Lynda Tuberty and Joe Bauman in the Arnold laboratory at Rutgers. Lynda Tuberty also carried out the diabody binding experiment. Cryo-EM structures were solved by Jessica Bruhn and Dario Passos of the Lyumkis lab at Salk while the enzymatic assays were carried out by Jeffery DeStafano and Irani Ferreira at the University of Maryland, College Park.

5.1 Overcoming proteolytic degradation of HIV-1 polyproteins expressed in bacteria

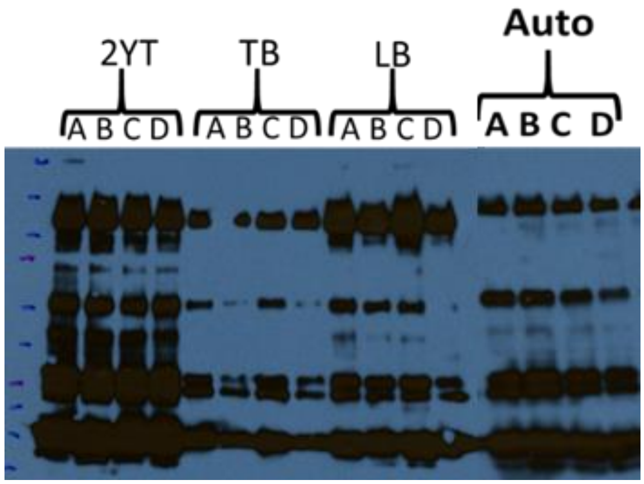
Bacteria remain the most popular expression vehicle for proteins for various biotechnological applications including therapeutics and biophysical studies (Huang et al., 2012). The quality of heterologously expressed proteins in bacteria and indeed other cell lines are often times a bottleneck to structural and biophysical studies. Degradation of proteins expressed in *E. coli* remains among the most significant challenges to its utility for protein production despite several genetic approaches developed to minimize degradation of heterologously expressed proteins in *E. coli* (Peti and Page, 2007, Sahdev et al., 2008). Several modifications of the lysogeny broth (LB) recipe originally developed by Giuseppe Bertani in the early 1950s (Bertani, 2004) for growing bacteria have also been undertaken with the view of overcoming challenges relating to yield, activity, stability, etc., of proteins expressed in bacteria. Other factors and constituents such as temperature, fusion tags, coexpression with chaperones and binding partners, have also been varied to overcome some of these problems (Vasina and Baneyx, 1997, Francis and Page, 2010, Schlapschy and Skerra, 2011, Bird et al., 2016, Christensen et al., 2017, Steinmetz and Auldridge, 2017).

While all these efforts singly or in combination have yielded success to some extent, stable and soluble expression of proteins, especially flexible and multi-domain proteins, remains a significant challenge in *E. coli* (Grodberg and Dunn, 1988). Previous attempts to express HIV-1 Pol polyproteins employing common strategies such as different fusion tags, different vectors and promoters, low temperature and different *E. coli* strains did not yield soluble expression of target proteins. Proteins were highly degraded by endogenous proteases while the rest was shuttled into inclusion bodies. It was observed however that the RT/IN junction appeared to be the most susceptible to endogenous proteases. Mutations at this cleavage junction even though reduced proteolysis at this junction, did not yield soluble protein either using available recipes.

To overcome these challenges having tried different recipes without success, modification of the original LB media recipe by varying the concentration of tryptone, yeast extract, and NaCl while adding NZ amine, glycerol, and Mg followed by adjustment of the pH of the media was undertaken. A systematic analysis of the effect of key components (glycerol, Mg, and low pH) which are not standard constituents and/or variables in the commonly used media for protein expression in *E. coli*, on proteolysis of the expressed protein was carried out using anti-integrase 8E5 antibody in Western blot analysis. For the purpose of comparison, commonly used media were constituted using their original recipes. To ensure that lack of nutrients do not become a limiting factor in these experiments, the concentrations of tryptone, NZ amine, yeast extract, and NaCl were kept quite high at 1.5%, 1.5%, 1.0% and 1.5%, respectively, to provide an unlimited supply of all nutrients required for bacterial growth. All experiments were conducted with the BL21 DE3 strain of *E. coli*.

5.2 Effect of Common Media Types (2xYT, TB, LB, and Auto-Induction Media)

The effect of media 2x YT, Terrific Broth (TB), lysogeny broth (LB), and auto-induction media (Sun et al., 2009, Lessard, 2013) on the proteolytic degradation of HIV-1 Pol construct co-expressed with MBP-IBD under different IPTG induction conditions, where necessary, was carried out to recapitulate earlier observations. Also tested was the effect of 0.5% glycerol and 2 mM $MgSO_4$ in these media as well. Mg^{2+} at these low concentrations have been reported to support bacteria growth and lead to high cell density under optimal growth conditions (Christensen et al., 2017). None of these parameters however had a significant impact on the proteolytic degradation profile of the expressed proteins (**Fig. 53**)



Induction conditions and Centrifuge Speeds

A. 0.5 mM IPTG, **4,500** x g

B. 0.5 mM IPTG **16,000** x g

C. 0.2 mM IPTG, 0.5% Glycerol 2 mM $MgSO_4$., **4,500** x g

D. 0.2 mM IPTG, 0.5% Glycerol 2 mM $MgSO_4$., **16,000** x g

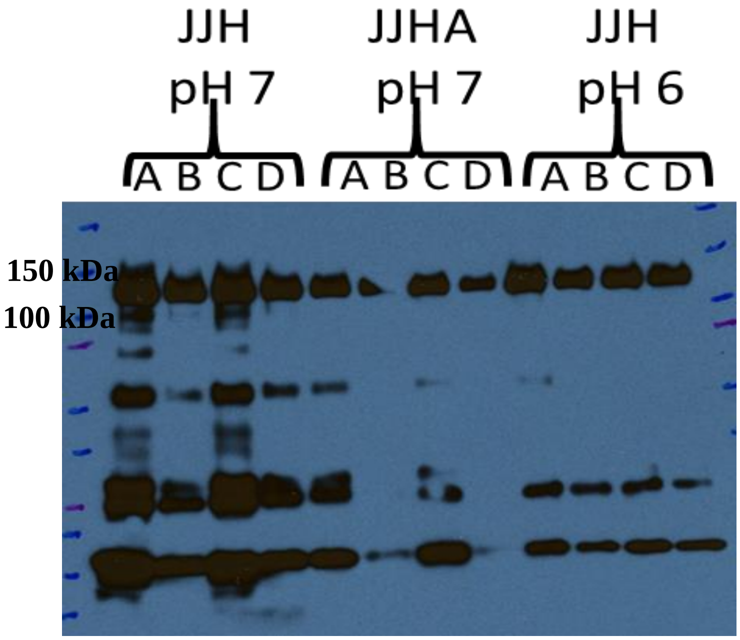
Figure 53: Anti-IN Western blot of HIV-1 Pol-MBP-IBD complex using different media commonly used for protein expression in bacteria under different conditions

5.3 Effect of pH

Living organisms including bacteria have developed very robust homeostatic mechanisms of dealing with external stress including pH fluctuation and temperature to ensure optimal growth (Slonczewski et al., 1981, Zilberstein et al., 1984, Slonczewski et al., 2009). Among the most common methods used by bacteria for dealing with acid stress include robust proton pumps for pumping protons out of the cytosol, upregulation in the synthesis of compounds that consume protons such ammonia, and increased rate of oxidative decarboxylation reactions which consume protons, as well as decreased membrane permeability through modifications of the lipid content in the membrane to reduce proton influx into the cytosol (Lund et al., 2014).

Even though proteases have characteristic pH values at which they are optimally active, the impact of pH of media on heterologous protein expression in bacteria has not been widely explored. It is understandable that such a parameter has not been explored given the robust maintenance of constant pH in bacteria (Zilberstein et al., 1984, Slonczewski et al., 2009). Having formulated a rich media with all the attributes that should aid high yield protein expression (discussed in chapter four), the impact of the pH of the media on the proteolytic profile of the HIV-1 Pol constructs was evaluated using western blot analysis. The HIV-1 Pol + MBP-IBD coexpression constructs using the JJH media reformulated media at pH values of 7 and 6 was used for these studies to decipher which ingredients in the JJH media were responsible for its remarkable ability to reduce proteolysis of the HIV and PFV polyproteins as discussed in chapter four of this thesis.

It was a surprise that the proteolysis of the expressed protein was much higher at the physiologic pH 7 than the non-physiologic pH 6.0 (**Fig. 54**). At pH 6, whether the media was buffered with 50-100 mM phosphate did not change the proteolytic profile. This suggested to that the protease(s) responsible for cleaving heterologously expressed proteins in BL21 DE3 codon plus RIL strain of *E. coli* may be less active at this non-physiologic pH. It is also possible that protein folding is much faster at this pH such that misfolding and unfolding which normally induce proteases activation is minimized.



Induction conditions and Centrifuge Speeds

A. 0.5 mM IPTG, **4,500** x g
B. 0.5 mM IPTG **16,000** x g
C. 0.2 mM IPTG, 0.5% Glycerol 2 mM $MgSO_4$., **4,500** x g
D. 0.2 mM IPTG, 0.5% Glycerol 2 mM $MgSO_4$., **16,000** x g

Figure 54: Anti-IN Western blot of HIV-1 Pol-MBP-IBD complex using JJH media at different pH

5.4 Effect of Mg^{2+}

The effect of Mg^{2+} concentrations on bacteria cell growth has long been known (Christensen et al., 2017). At low concentrations of 1-5 mM, it known to stimulate cell growth leading to high cell density compared media without it. However, the addition of Mg^{2+} to bacterial growth media at concentrations above 10 mM to the best of my knowledge has not been evaluated for its impact on protein expression in bacteria. The presence such levels of Mg in media for bacteria growth is non-physiologic. To determine whether the large amount of Mg in the media has any impact on the proteolytic profile of the proteins expressed, the HIV-1 Pol and MBP-IBD co-expression construct was expressed in the presence of 10 mM and 50 mM $MgSO_4$. The use of 2 mM $MgSO_4$ in the experiments evaluating pH did not seem to make a significant difference. At 50 mM Mg, the proteolytic degradation of the protein was also drastically reduced similar to the effect of pH 6 compared to the 10 mM concentration (**Fig. 55**).

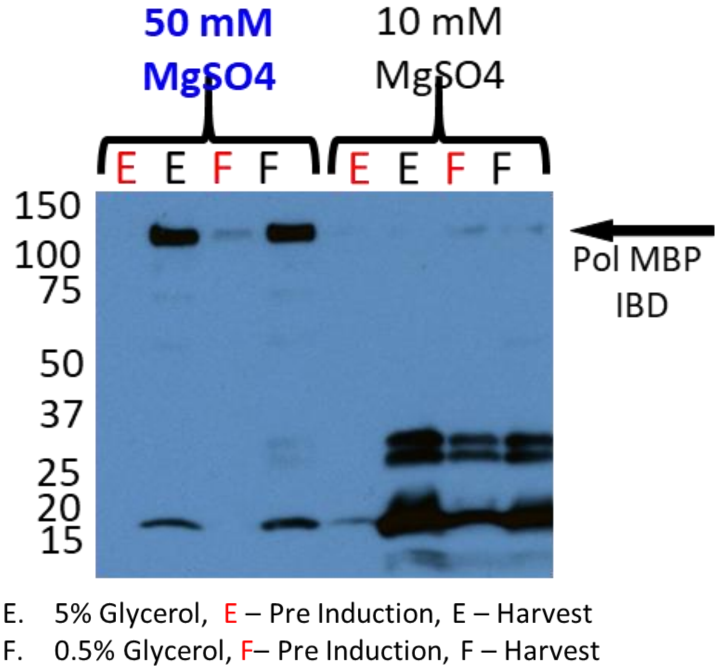


Figure 55: Anti-IN Western blot of HIV-1 Pol-MBP-IBD complex using JJH media

The combined effect of at least 50 mM Mg and pH 6.0 or less led to a significant reduction in proteolysis even at high OD leading to the accumulation of multiple milligrams of soluble HIV polyprotein products which can be extracted and purified. It is unclear how the effect of the high Mg concentration may be influencing the reduction in the proteolytic degradation. Apart from stabilizing nucleic acid structures, Mg^{2+} at these high levels may be competing with the divalent metal cofactors of metalloproteases such as Zn, Co and Mn and thereby rendering them inactive or less active. It is however not known if metalloproteases are responsible for these proteolytic cleavage events. If proteolytic degradation of these proteins is caused largely by metalloproteases, then it is possible that metal binding at low pH would be less stable than at high pH and would provide a plausible mechanism by which the combined effect of Mg and pH may be exerted. All of the HIV-1 protein expression was therefore carried out in media containing at least 50 mM Mg at pH 6 while maintaining the glycerol concentration at 5% w/v to support high OD.

5.5 Biophysical characterization of HIV-1 Pol polyprotein constructs

Biochemical data suggest that early cleavage events during maturation of HIV-1 liberates the Pol polyprotein with the TFR at its N-terminus (Pettit et al., 2003, Pettit et al., 2004, Pettit et al., 2005a). To understand the structural organization of this early cleavage product and potentially offer insight into the initial dimerization of PR in the context of Gag-Pol, we designed HIV-1 Pol constructs of the BH10 strain with different tags and a HRV14 3C-cleavage site its N-terminus, with D25A mutation in PR and 23 residues of the p6* region, designating that as WT. Initial screening of constructs which retained good expression and solubility for biophysical studies using these WT constructs with various tags did not yield a positive outcome.

While difficult to make and with very low yield, all the isolated HIV-1 Pol WT proteins aggregated even in high salt (500-700 mM) and glycerol (10%), 50 mM arginine, 50 mM glutamic acid buffer. An N-terminal Sso7d-containing construct and a non-conservative mutation of the RT/IN junction residues LF to aspartates were also introduced into the constructs and designated as HIV-1 Pol TF23. Aspartate substitution at the RT/IN junction significantly improved the solubility of the protein, enabling structural and biophysical studies. Pol TF3, which contains only 3 consensus amino acid residues, Ser-Asn-Leu (SNL) of the p6* sequence, had even better solution behavior with reduced aggregation. A version of Pol TF3 containing only the NTD of IN, Pol TF3-NTD was also made for structural and biophysical studies. Mutation of surface exposed hydrophobic residues to hydrophilic amino acids to improve biophysical behavior is a widely employed strategy in structural biology (Jenkins et al., 1996, Chen et al., 2000).

Proteins isolated after Ni affinity chromatography and HiTrap heparin purification were concentrated in the high salt elution buffer 25 mM Tris pH 8.0, 1 M NaCl, and 5% glycerol after the purity had been assessed on an SDS-PAGE gel and loaded onto a Superose 6 Increase 10/300 GL gel filtration column pre-equilibrated in 25 mM Tris pH 8.0, and 250 mM NaCl. Peak fractions were pooled and concentrated to 0.2-0.5 mg/mL, flash-frozen and stored at -80ºC.

Dynamic light scattering (DLS) was used to assess the sizes and the hydrodynamic radii of the molecules in solution prior to being frozen. The gel filtration profiles of the constructs were generally broad (**Fig. 56**), an indication of high conformational heterogeneity in the sample and perhaps high polydispersity. Consistent with this observation, DLS experiments conducted on purified samples suggested that the molecules in solution had an apparent molecular weight of ~250 kDa (**Fig. 57**), which according to the calculated molecular weight of ~119 kDa for the TF23 Pol, would suggest that dimeric protein molecules predominate in solution. The molecular weight of the TF3 Pol was found to be ~234 kDa which is again consistent with a dimer.

The average hydrodynamic radius of the molecules of the Pol TF23 was about 6.3 nm while that of Pol TF3 was approximately 6.1nm. The polydispersity indices of the two constructs were however different. While Pol TF3 showed polydispersity of about 11%, that of Pol TF23 was about 49%. This is very consistent with the fact that the longer p6* residues in TF23 introduce much higher degrees of freedom in this region and therefore greater flexibility compared to the three p6* residues in Pol TF3 (see **Fig. 58** and **59** for the gel filtration profiles and DLS data). Pol TF3-NTD made by truncation of the catalytic core domain (CCD) and the C-terminal domain (CTD) from the IN portion of the Pol TF3 produced a much better-behaved protein in solution. It could be concentrated to beyond 2 mg/mL without aggregation. At 0.5 mg/mL, it was found to be dimeric by DLS studies with hydrodynamic radius of ~5.0 nm. This significant decrease in the hydrodynamic radius is an indication that the IN CCD-CTD of the protein extend further from the N-terminal domain (NTD) without much contact between them. A MALDI-TOF mass spectroscopic analysis of the TF3 Pol-NTD (**Fig**. **60**) sample initially prompted by the apparent low molecular weight on the SDS-PAGE gel compared to the marker (see **Fig. 61**) confirmed that the molecular weight is indeed ~81 kDa consistent with the expected theoretical molecular weight calculated.

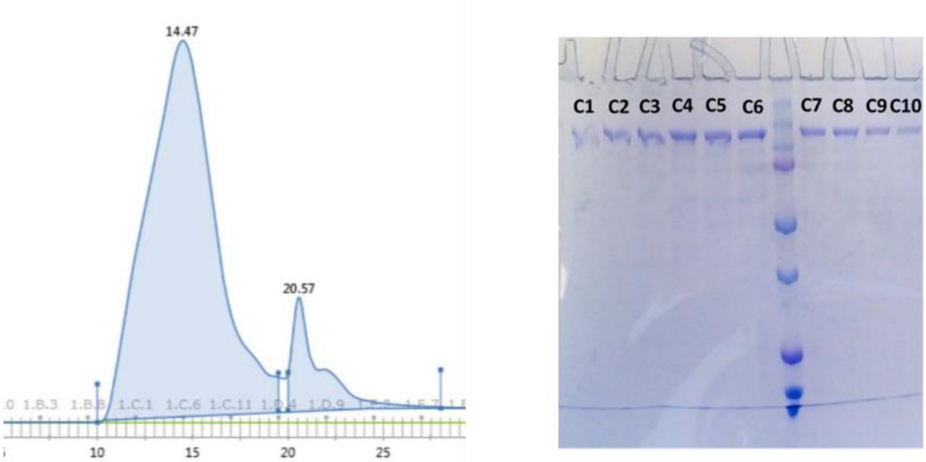
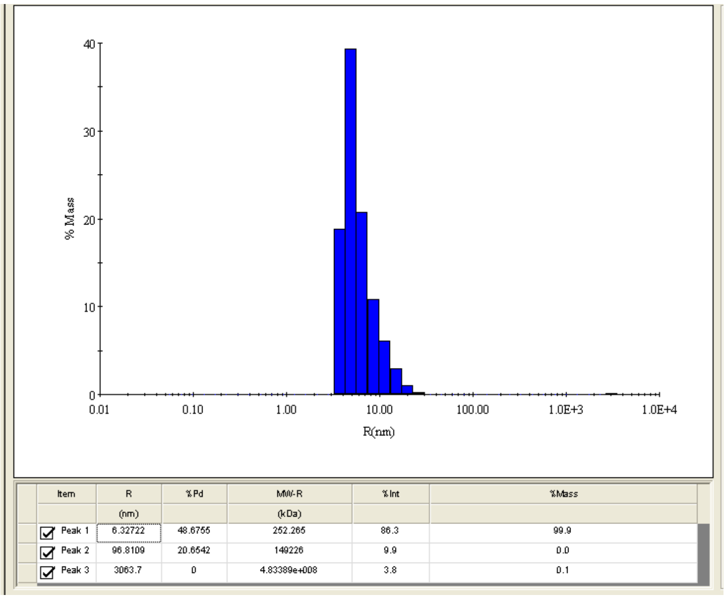


Figure 56: Superose 6 Increase 10/30 GL gel filtration profile of HIV-1 Pol TF23 run in 50 mM Tris-Cl pH 7.6, 250 mM NaCl and corresponding SDS-PAGE gel fractions



Item	R	%Pd	MW-R	%Int	%Mass
	(nm)		(kDa)		
☑ Peak 1	6.32722	48.6755	252.265	86.3	99.9
☑ Peak 2	96.8109	20.6542	149226	9.9	0.0
☑ Peak 3	3063.7	0	4.83389e+008	3.8	0.1

Figure 57: DLS profile of HIV-1 Pol TF23 at 0.25 mg/mL

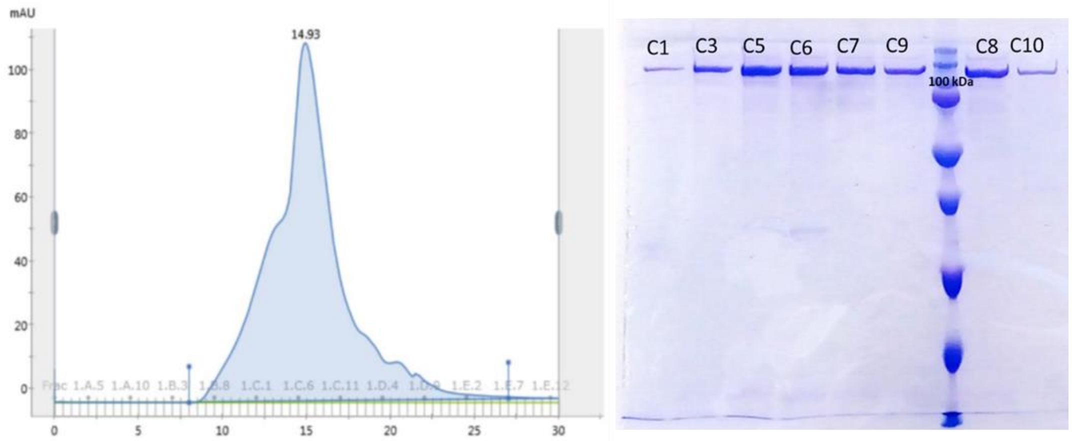
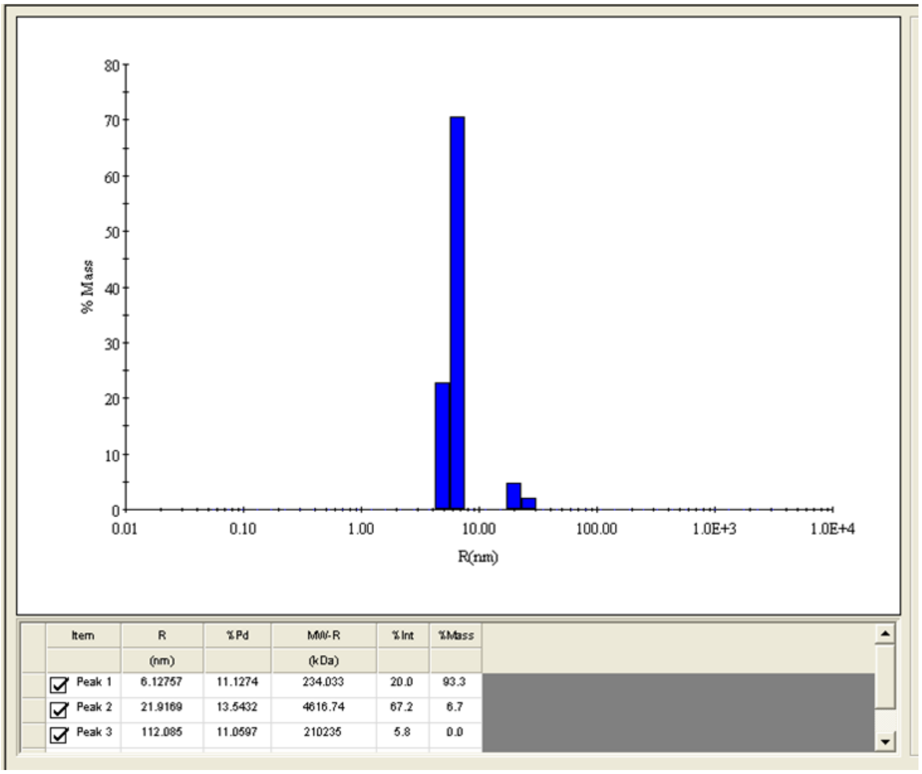


Figure 58: Superose 6 Increase 10/30 GL gel filtration profile of HIV-1 Pol TF3 run in 50 mM Tris-Cl pH 7.6, 250 mM NaCl and corresponding SDS-PAGE gel fractions



Item	R	%Pd	MW-R	%Int	%Mass
	(nm)		(kDa)		
☑ Peak 1	6.12757	11.1274	234.033	20.0	93.3
☑ Peak 2	21.9169	13.5432	4616.74	67.2	6.7
☑ Peak 3	112.085	11.0597	210235	5.8	0.0

Figure 59: DLS profile of HIV-1 Sso7d-SNL-Pol at 0.23 mg/mL

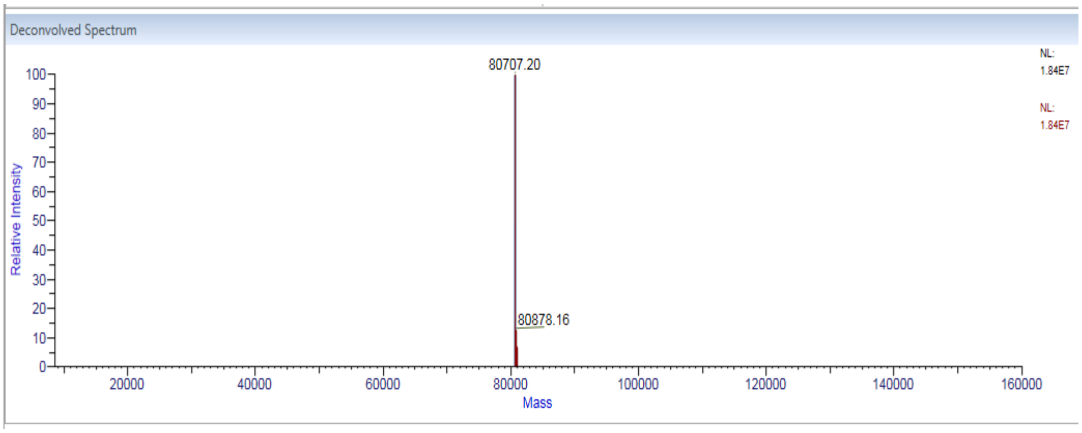
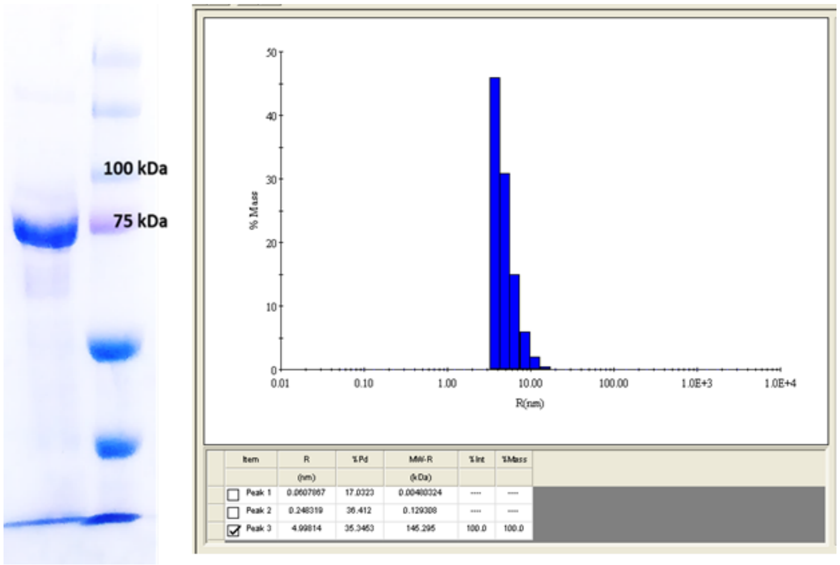


Figure 60: MALDI-TOF-MS Spectrum of HIV-1 Pol TF3-NTD



	Item	R	%Pd	MW-R	%Int	%Mass
		(nm)		(kDa)		
□	Peak 1	0.0607067	17.0323	0.00480324	----	----
□	Peak 2	0.248319	36.412	0.129308	----	----
☑	Peak 3	4.99814	35.3463	146.295	100.0	100.0

Figure 61: SDS-PAGE gel of HIV-1 Pol TF3-NTD and corresponding DLS reading at 0.5 mg/mL

5.6 Integrase domains in HIV-1 Pol polyprotein constructs bind IBD of LEDGF/p75

The existence of dimeric Pol protein molecules in solution was not too surprising since all of the HIV enzymes are functional multimers with dimers as their fundamental building blocks. PR and RT are functional dimers while IN uses multimers of dimers as a functional unit. To investigate whether the IN domain is dimeric or capable of forming dimeric entities in the constructs containing full-length Pol, a fusion of the IBD domain of LEDGF/p75 and the maltose-binding protein, MBP with a helical extension at the C-terminus of MBP and designated MBP-IBD, was coexpressed with the Pol TF23. IBD is known to bind HIV IN CCD dimers at the dimeric interface (Busschots et al., 2005, Llano et al., 2006).

When coexpressed with the Pol constructs containing the full-length IN, the MBP-IBD was retained during the heparin column purification even with a 300-400 mM NaCl wash of the column. To verify that such tight association is not an intrinsic property of the MPB-IBD itself, the heparin elution was loaded onto a 5 mL MBPTrap HP columns prepacked with Dextran Sepharose and washed with Tris buffer containing 1 M NaCl. At this stringent wash, any molecule not specifically bound to the column by itself or the MBP bound to the column would be expected to either flow through the column or be washed out since the Dextran Sepharose resin is highly specific for the MPB. Elution of the protein from the MBPTrap column was carried out using 25 mM Tris pH 8.0, 500 mM NaCl, and 10 mM maltose. The protein was subsequently concentrated and loaded onto a Superose 6 Increase 10/300 GL SEC column as before. Peak fractions showed a slight shift to a lower elution volume peak compared to the Pol only and a second peak for excess MBP-IBD. SDS-PAGE gel of the fractions showed prominent bands for the MBP-BD and the Pol in the same fractions with excess MBP-IBD peak not showing any Pol bands (see **Fig. 62**).

DLS analyses were consistent with the earlier observation of dimeric protein molecules in solution with a hydrodynamic radius of 6.5nm, which is higher the either of the Pol constructs and consistent with the increased molecular weight (see **Fig. 63**). Taken together, these suggest that the IN arrangement in the Pol constructs is either dimeric or capable of forming dimers which enable IBD binding. This dimeric organization is intramolecular since an intermolecular dimerization would result in polymers of high molecular weight stabilized by the IBD. Furthermore, intermolecular dimerization mediated by the CCD of IN would alter the molecular weight of the Pol in solution and on the SEC column.

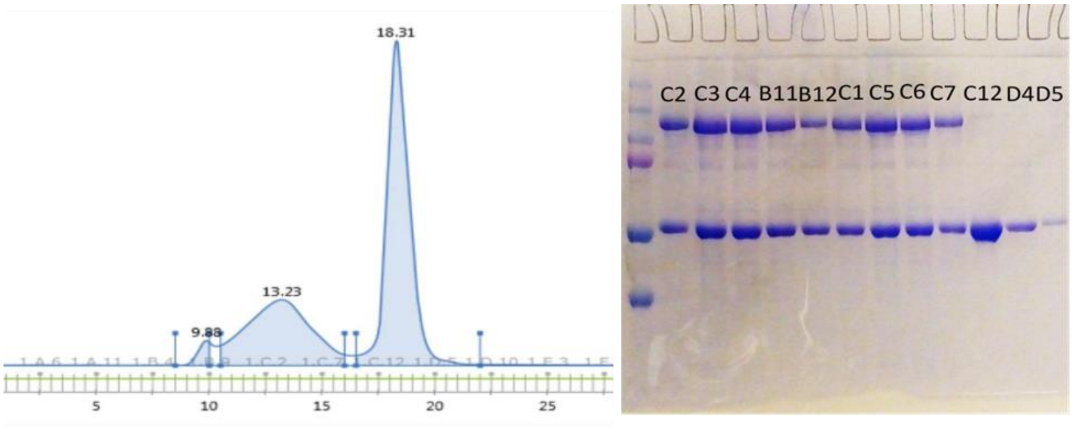
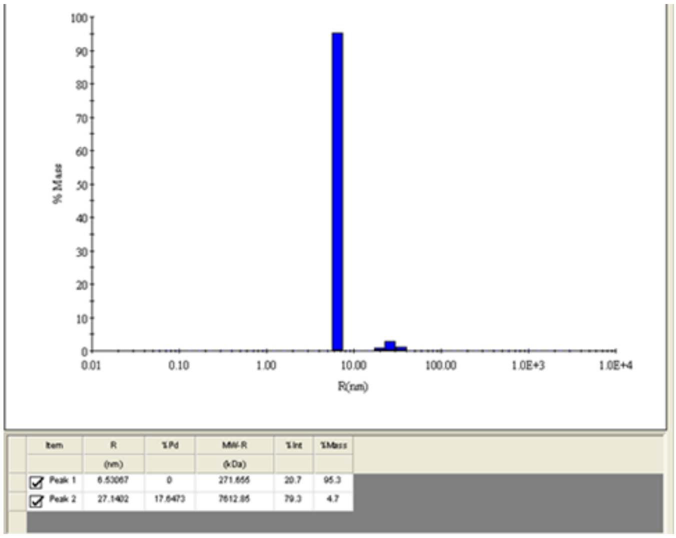


Figure 62: Superose 6 Increase 10/30 GL gel filtration profile of HIV-1 Pol TF23 coexpressed with MBP-IBD run in 50 mM Tris-Cl pH 7.6, 250 mM NaCl and corresponding SDS-PAGE gel fractions



Item	R	%Pd	MW-R	%Int	%Mass
	(nm)		(kDa)		
☑ Peak 1	6.53067	0	271.656	20.7	95.3
☑ Peak 2	27.1402	17.6473	7612.85	79.3	4.7

Figure 63: DLS reading of HIV-1 Pol TF23 -MBP-IBD complex at 0.4 mg/mL

5.7 HIV-1 Pol constructs bind diabody bearing the CDR of monoclonal antibody 28 (mAb28)

Monoclonal antibody 28 (mAb28) is an antibody that binds tightly to a flexible loop in the HIV-1 RT p51 subunit and stabilizes it (Ding et al., 1998), but mAb28 does not bind the p66 subunit since this sequence corresponds to the primer grip, which in p66 is relatively less accessible to antibody binding in the heterodimer. Chelsy Chesterman and Arnold (unpublished results) engineered a single-chain diabody, Ab62 from the CDRs of this antibody and solved crystal structures of HIV-1 RT in complex with it. The structures show the diabody binds to the same loop as mAb28 with an overall configuration very similar to crystal structures of HIV-1 RT complexes with Fab28, a Fab fragment of mAb28 (unpublished data).

To determine whether the dimeric organization of the HIV-1 Pol constructs as suggested by solution experiments involve the RT domain and if so, whether the configuration is similar to the mature HIV-1 RT heterodimer which is capable of binding Ab62, the TF3 Pol-NTD construct was incubated with this diabody and run on a Superose 6 Increase 10/30 GL gel filtration column using the same buffer as for the Pol. The peak fractions eluted at a lower retention time compared to the diabody alone or the Pol alone (see **Fig. 64**) suggesting a molecular weight larger than the individual components. SDS-PAGE analysis showed that the peak fractions contained both the Pol and the Ab62 diabody. A control experiment carried out using a diabody with specificity towards Herceptin did not show any binding through non-specific interaction the Pol construct, effectively ruling out the possibility of non-specific interaction between the Pol and Ab62. This clearly demonstrated to that not only is the IN domain in the Pol dimeric or dimerization competent, the RT domain is also dimeric, and that this dimeric organization is similar to the mature RT heterodimer capable of binding the mAb28-based diabody.

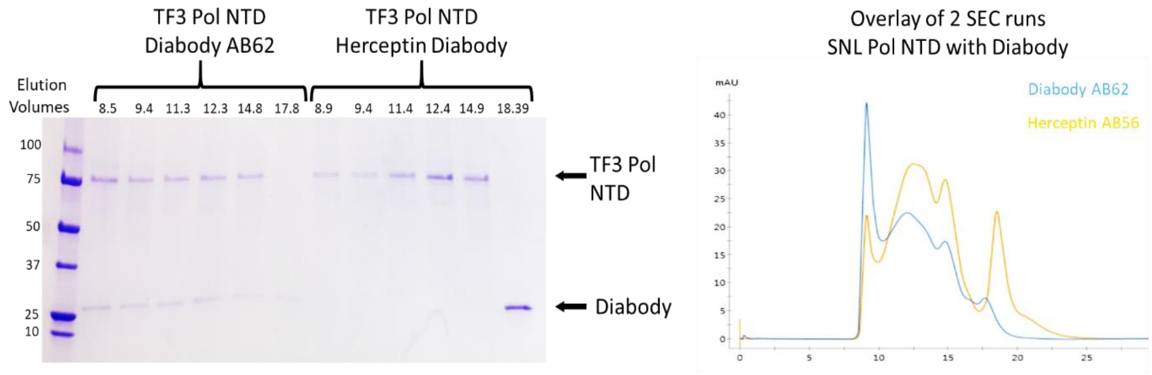


Figure 64: TF3 Pol NTD-Ab62 diabody complex run on Superose 6 Increase 10/30 GL

5.8 HIV-1 Pol polyprotein exhibits polymerase and ribonuclease H activity

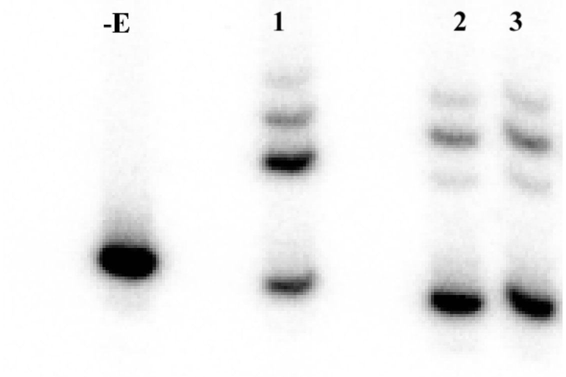
Having observed that the RT in the Pol constructs is dimeric and in a configuration similar to mature HIV-1 RT heterodimer that is able to bind Ab62, the polymerase and the ribonuclease H activities of the Pol constructs TF23 as well as the Pol coexpressed with MBP-IBD were tested to verify whether the Pol is capable of engaging nucleic acid substrates and whether it is active for polymerase and ribonuclease H compared to the WT RT. Two templates were chosen for the polymerase assay, a 38NT2,4-O-methyl DNA aptamer selected through SELEX and shown to bind WT RT with a Kd of ~15 pM (Miller et al., 2016) and a 4000 nt regular DNA template with 20 nt 5'-end labeled primer which was used not only to determine the polymerase activity but to determine the processivity of these enzymes as well.

Since the Pol constructs are prone to aggregation in low salt, 150 mM KCl was used for the extension and processivity experiments with the 4000 nt template with heparin sulfate was as a trap in the processivity assays. A 5'-end labeled 60 nt RNA with 23 nt DNA hybrid was used for assaying the ribonuclease H activity. As shown in Figure **65** (**A-C**), the RT in the Pol constructs indeed exhibits polymerase and ribonuclease H activity comparable to the WT RT while the processivity of the Pol-MBP-IBD may be slightly higher than the rest of the enzymes.

It is important to note however that, while the WT HIV-1 RT shows a significant pause after the addition of 3 nucleotides on the 38NT aptamer, the Pol constructs exhibit a significant pause only after the incorporation of 4 nucleotides on the same template. This suggests that these two enzymes interact slightly differently with the tight binding RT aptamer. It may also offer some insight into how the nucleic acid binding grooves of the two enzymes may differ. While the hairpin is perhaps able to lock the RNase H primer grip as seen the crystal structure of HIV-1 RT and the aptamer by the time the 3rd NT is added (Miller et al., 2016), significant contacts between this hairpin with the RNase H in the Pol occurs only after addition of 4th NT.

Extension of 38NT2,4-O-methyl DNA Aptamer

A



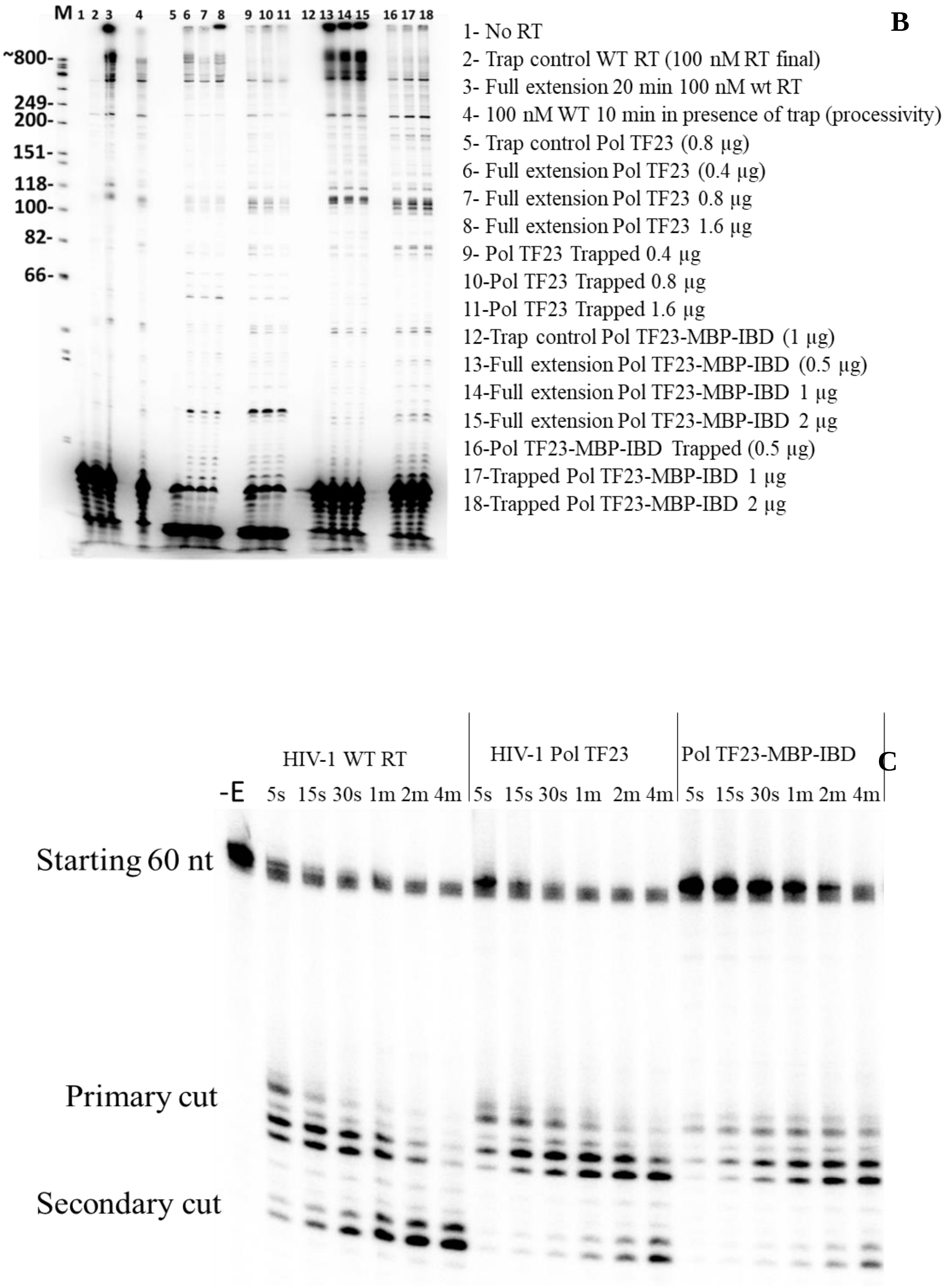
-E- No enzyme

1- HIV-1 RT WT

2- HIV-1 Pol TF23

3- HIV-1 Pol TF23-MBP-IBD complex

50 mM Tris (pH 8), 80 mM KCl, 6 mM MgCl, 1 mM DTT, 50 uM
dNTPs, 5 nM aptamer and 10 nM constructs for 10 min at 37ºC



1- No RT
2- Trap control WT RT (100 nM RT final)
3- Full extension 20 min 100 nM wt RT
4- 100 nM WT 10 min in presence of trap (processivity)
5- Trap control Pol TF23 (0.8 µg)
6- Full extension Pol TF23 (0.4 µg)
7- Full extension Pol TF23 0.8 µg
8- Full extension Pol TF23 1.6 µg
9- Pol TF23 Trapped 0.4 µg
10-Pol TF23 Trapped 0.8 µg
11-Pol TF23 Trapped 1.6 µg
12-Trap control Pol TF23-MBP-IBD (1 µg)
13-Full extension Pol TF23-MBP-IBD (0.5 µg)
14-Full extension Pol TF23-MBP-IBD 1 µg
15-Full extension Pol TF23-MBP-IBD 2 µg
16-Pol TF23-MBP-IBD Trapped (0.5 µg)
17-Trapped Pol TF23-MBP-IBD 1 µg
18-Trapped Pol TF23-MBP-IBD 2 µg

Figure 65: Comparison of the polymerase and ribonuclease activity of the HIV-1 Pol constructs compared to mature WT RT (A) extension assay on 38NT-3,4methyl Aptamer.

5.9 Architecture of HIV-1 Pol Domains

5.9.1 Reverse Transcriptase

HIV-1 RT (**Fig. 66**) is an asymmetric heterodimer with differently folded 66 kDa (p66) and 51 kDa (p51), subunits comprising the same primary sequence. However, all the enzymatic activity, both polymerase and ribonuclease H, reside in p66 while the p51 subunit, which lacks the C-terminal ribonuclease H subdomain, largely plays a structural role (Jacobo-Molina and Arnold, 1991, Wang et al., 1994). The stoichiometric ratio of p66: p51 remains equal in virions, which raises important questions regarding how this ratio is maintained. The formation of a p66/p66 homodimer as a precursor to the removal of one RNase H domain to form p51 has been proposed in the literature (Fan et al., 1995, Sluis-Cremer et al., 2004).

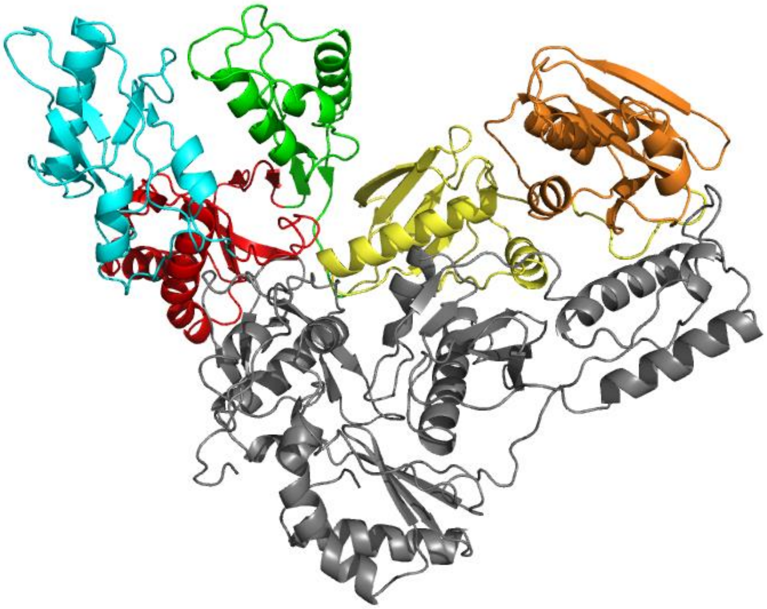


Figure 66: HIV-1 RT structure with different domains in p66 color coded and p51 in grey.

While the evidence in support of the formation of p66/p66 homodimer is incontrovertible, key questions are 1) when in the life cycle of the virus is this complex generated and 2) how is this formed? The pathway and pattern of proteolytic processing of the numerous cleavage sites in the Gag-Pol polyprotein of HIV-1 remains contentious without a clear consensus. However, the release of Pol from the Gag-Pol through an intramolecular cleavage, initially between SP1 and NC and subsequently NC/TFR junction, has garnered the most support (Pettit et al., 2004, Pettit et al., 2005a, Pettit et al., 2005b). Since Gag-Pol dimerization is a prerequisite to the PR activation which leads to these cleavage events, it is reasonable to conjecture that the large dimerization interface in Pol spanning the PR through to IN could play a role in this process.

The structures of the RT in the Pol constructs are consistent with our earlier biochemical experiments. Indeed, the structure of the RT is very similar to the mature asymmetric heterodimeric RT structure (see **Fig. 67**). The p66 thumb domain remains folded "down" and packed against the fingers domain thereby occluding the nucleic acid binding cleft. The p51 primer grip residues 219-230 remain disordered as previously seen in crystals structures without Fab28 bound suggesting an intrinsic disorder in this region of RT.

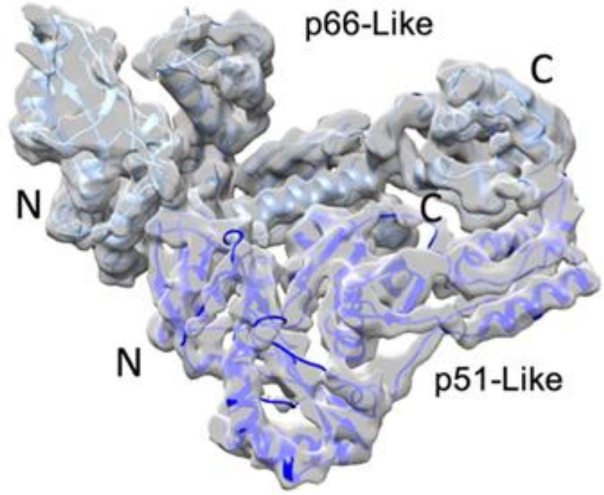


Figure 67: 3.9 Å resolution single particle cryo-EM reconstruction of the RT portion in HIV-1 Pol fit with HIV-1 RT crystal structure (PDB ID, 1DLO)

In crystal structures of the HIV RT, the N-termini of p66 and p51 remain unstructured but they are packed close to the knuckle. In the Pol structures, these N-terminal unstructured residues are “peeled away” from this usual position and stretched out in a hyper-extended configuration. The displacement of these residues can be attributed to the presence of the PR, which needs to be accommodated within this vicinity (see **Fig. 71**). The uncleaved RNase H domain of the p51 remains disordered in these Pol structures, consistent with the fact that only few of the residues intervening the connection and the RNase H subdomains in p51 are resolved in the numerous structures of RT.

The appearance of a heterodimer-like RT in these early stages of polyprotein processing suggest that RT dimerization is coupled to PR activation. It is worth noting that this coupling mechanism as suggested by these structures effectively sequesters the PR cleavage site in RT, F440/Y441 in the p66 subdomain making it inaccessible to the PR. This explains why only one RNase H domain is cleaved from the RT and offers a plausible mechanism of maintaining the stoichiometry of p66/p51. This is consistent with evidence supporting the formation of p66/p66 homodimer as a prerequisite for RNase H cleavage (Sluis-Cremer et al., 2004). Whether the cleavage of the IN domain occurs before the RNase H domain of p51 or whether removal of the IN connecting the p51 is a requirement for the removal of RNase H remains unknown. However, co-expression of the untagged p66, p51 or PR-RT fusion with the full-length Pol results in the pull down of these constructs by the Pol suggesting that the IN domain is not required for this RT dimerization (data not shown).

SAXS analysis was also carried out on the Pol samples to determine the size distribution of the molecules in solution and their radius of gyration. Unfortunately, the sticky nature of the full-length Pol and the MBP-IBD complex made the determination of the maximum particle dimensions (D_{max}) and the radius of gyration (Rg) unreliable. This is not surprising because in crystal structures of HIV-1 IN CCD-CTD (PDB ID 1EX4), such intermolecular interactions have been captured especially involving the CTD domains (**Fig. 68**). Further optimization of the sample concentration and buffer conditions is required to improve the monodispersity of the samples to make them suitable for SAXS analysis.

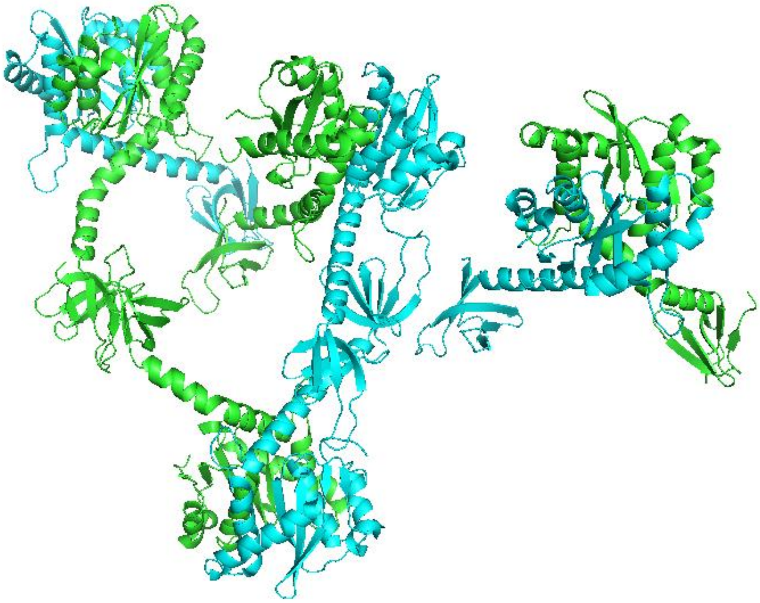


Figure 68: Crystal structure of HIV IN (PDB ID, 1EX4) showing CTD interactions

Analysis of the Pol TF3-NTD data showed that the Dmax is ~ 145 Å with a Rg value of ~ 47 Å. The P(r) function curve (**Fig. 69**) displays characteristics of multi-domain protein consistent with the construct. The envelope calculated based on these parameters (**Fig. 70**) shows a characteristic RT core which fits very well the crystal structure of HIV-1 RT (PDB 1DLO) as well as the NTD of the IN domain. Consistent with the cryo-EM structure, the RNase H domain of the p51-like subunit is not visible even in the SAXS envelope while the density of the PR is prominent with enough room to accommodate a dimeric PR. This further affirms the flexibility of the PR as well in the Pol.

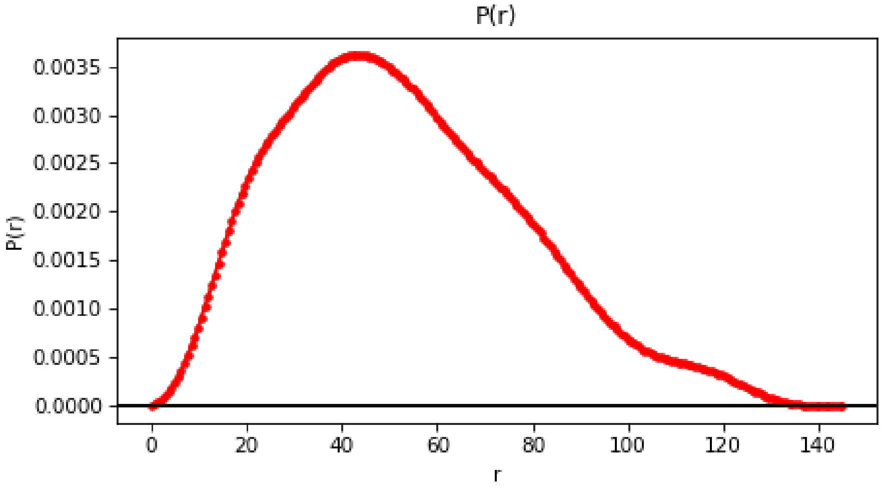


Figure 69: P(r) function profile showing Dmax

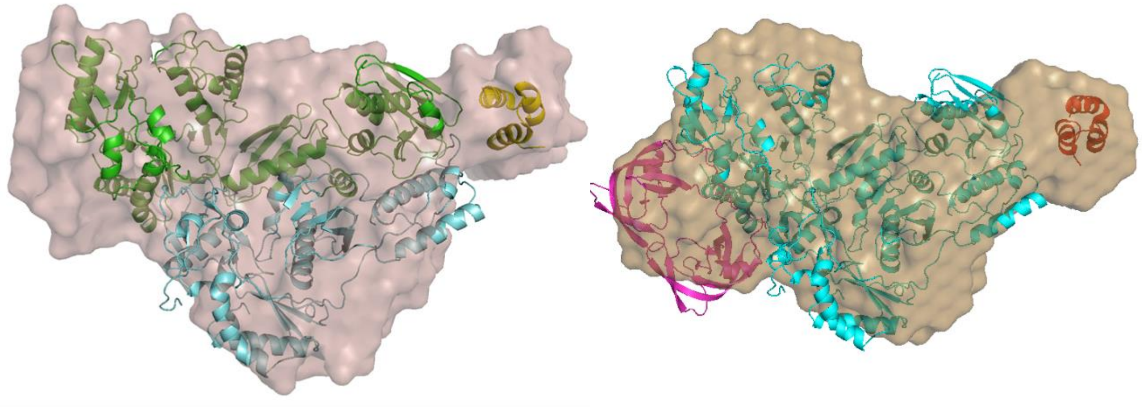


Figure 70: SAXS envelopes of HIV-1 Pol-IN-NTD fitted with HIV-1 RT (PDB ID 1DLO) IN NTD (PDB ID 1K6Y) and PR (PDB ID 2HB4)-Right.

5.9.2 Protease

The retroviral PR, functional as a homodimer, is responsible for processing all the cleavage sites in the Gag and Gag-Pol polyproteins (Tozser, 2010, Adamson, 2012).

During PR activation which precedes maturation, two Gag-Pol molecules dimerize which results in the monomeric PRs embedded in them coming together as well to generate the active site in which the various cleavage site peptides are sequestered and cleaved (Pettit et al., 2003, Pettit et al., 2004, Pettit et al., 2005a, Sadiq et al., 2012, Mattei et al., 2016). These constructs of the Pol designed therefore mimic the PR in the nascent stages of maturation. Activation of the PR which leads the production of infectious virus particles occurs only during or after budding. Delayed or premature activation are concomitantly associated with non-infectious virus particles (Tachedjian et al., 2005, Sudo et al., 2013,)

The timing of PR activation suggests inherent mechanism(s) of controlling premature PR activation since a stochastic distribution of 5% Gag-Pol in the virion during assembly may not be enough to prevent two Gag-Pol molecules from finding each other. To this end, various biochemical and structural studies have suggested that sequences immediately downstream the N-terminus of PR in the p6* region negatively modulate PR dimerization and by extension its enzymatic activity (Louis et al., 1999, Pettit et al., 2003, Tang et al., 2008). The observation that the PR embedded in the Gag-Pol also shows less sensitivity to active site PR inhibitors suggests differences in architecture of the mature PR and its immature precursor which lends further credence to the negative impact of the p6* sequences on the tertiary folding of the immature PR (Pettit et al., 2004, Davis et al., 2012).

The density of the PR in the Pol structures is consistent with a PR that is spatially and conformationally very dynamic with a solvent-accessible active site. Clearly the N- and C-terminal residues which form interweaving β-sheets that stabilize the mature enzyme, are unfolded in these immature structures. The unfolding of the C-termini β-sheets help provide a loose hinge that tethers the PR to the RT with the help of the N-terminal residues of RT which are themselves peeled out slightly from their usual position in crystal structures (see **Fig. 71**). The PR is therefore conformationally very flexible with significantly weaker density. Despite the weak density and conformational heterogeneity, it is clear that there is little interaction between the PR and the RT and therefore not much buried surface area between them. The PR is therefore held at a fixed distance from the RT, making the PR/RT cleavage junction easily accessible to PR.

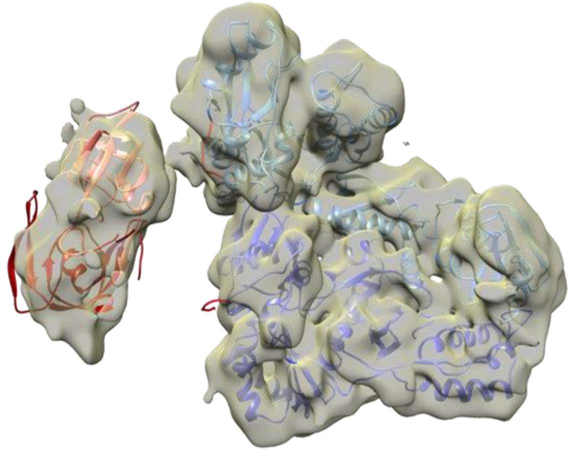


Figure 71: 8 Å resolution cryo-EM structure at of HIV-1 Pol, showing resolved density of PR-RT fitted with HIV-1 RT (PDB ID 1DLO) and HIV-1 PR (2HB4).

These structures strongly suggest an inherently diminished dimerization competency of the immature PR. The various conformational states of the PR in its efforts to overcome very strong dimer destabilization exerted by its N-terminal residues coupled with a loss of strongly stabilizing four-stranded β-sheets and weak flap-flap interactions in the absence of substrates or inhibitors are well captured in the focused classifications of the PR density (**Fig. 72**). This unstable dimerization partly explains the sluggishness of the initial cleavage events as well as the low sensitivity to active site PR inhibitors. The ability of the PR to form a stable dimeric structure capable of peptide cleavage even when tethered to the Pol in this instance or the Gag-Pol precursor is captured by these structures.

No significant differences in the PR density or the architecture of the RT was observed when the Sso7d tag was removed by HRV 14 3C protease digestion as well as the structures of the Pol TF3 and the TF3-NTD construct. This suggests that specific residues at the p6* region may not be required necessarily to perturb the PR dimerization, or the PR is simply conformationally constrained to adopt a more stable mature-like conformation when tethered to RT.

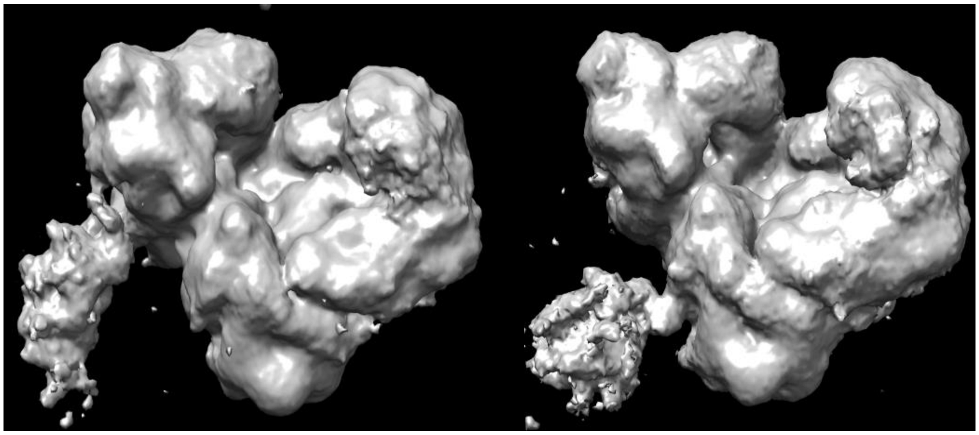


Figure 72: Low resolution focused classified cryo-EM showing density of PR on RT p66 (left) and p51

5.9.3 Integrase

IN comprises three independently folded domains NTD, CCD and CTD connected by long flexible linkers. The structure of the full-length IN without nucleic acid substrates in the context of intasomes (Passos et al., 2017) remains intractable by structural biology techniques. This is in part due to the inherent flexibility of this protein. Consistent with this observation, most of the of IN density in these structures is not visible. The C-terminus of the RT p66-like domain in the low-resolution cryo-EM maps, as well as the SAXS envelope, show extra density consistent with the NTD of IN (**Fig. 70** and **73**).

The flexibility of the IN itself is further compounded by the flexible loop at the C-terminus of RT which serves as the hinge for IN. The flexibility of the RT/IN junction suggests that this cleavage junction is accessible by the PR. Due to the low resolution of these IN-containing classes, the multimeric state of the IN cannot be confidently assigned. However, as already alluded to, the IN domain in the full-length Pol constructs binds very tightly and pulls down the IBD domain of LEDGF/p75. Since the IBD is known to bind highly conserved residues at the CCD/CCD interface, a dimeric IN is likely present in the Pol structure. It is also worthy of note that IN in solution in the absence of nucleic acid is a conglomerate of monomers, dimers and tetramers. The linker between the p51 connection domain and the RNase H domain (25 residues) is long enough to permit intramolecular CCD dimerization without potential steric clashes. This interaction is likely similar to that observed in the crystal structures of a NTD-CCD (PDB 1K6Y) which shows interaction of proximal NTDs in the same dimer as the CCD and CCD-CTD (PDB ID 1EX4), which shows the interaction of the CCDs relative to the CTDs (see **Fig. 74**). The directionality of the CCD dimerization interface is likely responsible for the lack of intermolecular dimers of IN CCDs which would result in polymeric units stabilized by the IBD in the Pol constructs.

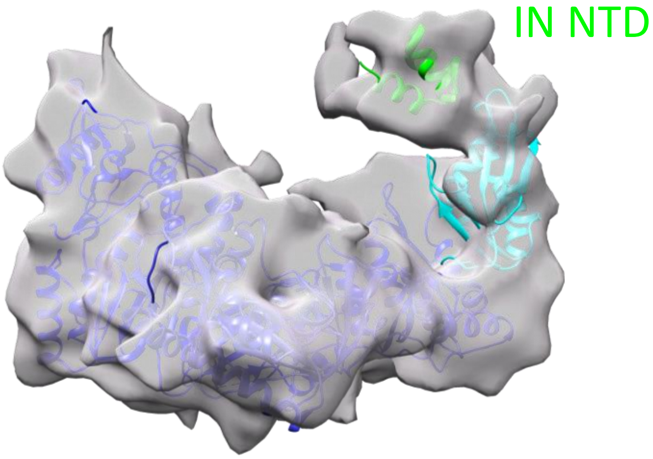


Figure 73: Low resolution EM map showing density for IN-NTD

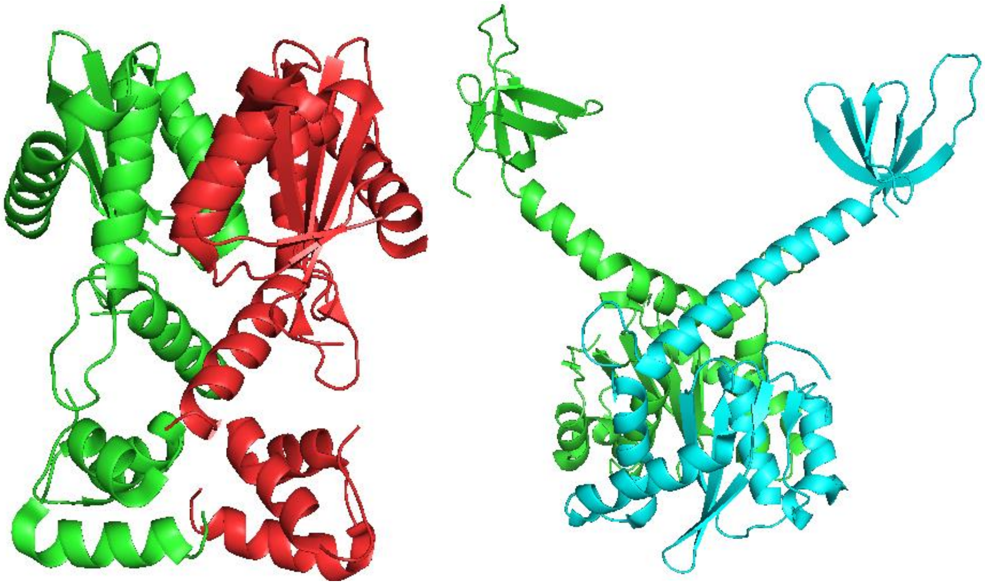


Figure 74: Structures of HIV-1 IN showing the CCD dimer in relation to NTD (left) and CTD

5.10 Conclusions

The cryo-EM structure of HIV-1 Pol constructs expressed in a soluble form on a multi-milligram scale in bacteria and other biophysical data have presented. The expression of these proteins in bacteria which hitherto had been refractory was made possible following the discovery that at low pH ≤ 6.5 and high concentrations of Mg (≥ 50 mM), proteolytic degradation of these proteins heterologously expressed in BL21 DE3 cells was greatly attenuated. A new media recipe which is a rich source of nutrients supplemented with 50 mM Mg and pH 6.0 yields multiple milligrams of soluble HIV-1 Pol polyproteins which can be purified to homogeneity using Ni affinity and heparin chromatography. Gel filtration and DLS analysis suggested that the molecules of the Pol were dimeric.

Consistent with these data, single particle cryo-EM and SAXS analysis of the Pol constructs reveals a dimeric organization of the independent domains of the HIV-1 Pol. The RT remains structurally indistinguishable from the mature enzyme with the thumb folded down into the nucleic acid binding cleft while the IN, even though largely disordered, is still visible at low resolution above the p66 subunit. Due to the disorder of the IN density, the multimerization state of IN could not be ascertained from these structures. However, since the constructs containing the full-length IN bind the IBD domain of LEDGF, it can be inferred to be a dimer since the IBD only binds dimers of IN. The PR is also conformationally dynamic but visible in these structures at moderate resolution. The HIV PR dimer is more stable at low pH (4-5.5) and less stable at physiologic pH and above (Tyagi et al., 1994, Wondrak and Louis, 1996) at which these structures were determined. Coupled with the fact that these constructs contain the D25A mutation in the PR active site which renders even the mature PR less stable than the WT or the D25N version due to loss of a hydrogen-bonding network that stabilizes the dimer at the fireman’s grip, it is remarkable that any PR density can be seen at all in these structures.

These observations suggest a positive role of RT in bringing the PR monomers into close proximity thereby enabling them to form enzymatically active dimers. Consistent with this, it has been well documented in the literature the deleterious effect of RT dimerization inhibition (Tachedjian et al., 2005, Chiang et al., 2012, Sudo et al., 2013) on the Gag and Gag-Pol processing during maturation. Conversely, stabilization of RT dimers using molecules such as efavirenz lead to premature PR activation presumably by also stabilizing enzymatically competent PR dimers (Tachedjian et al., 2005, Sudo et al., 2013). On the basis of these evidence a model of PR activation is proposed as follows: in the presence of RT or IN, Gag-Pol dimerization may be initiated and stabilized through p66/p66 homodimerization and or IN dimers. This extremely stable association anchors two PR monomers at their respective N-termini into close proximity which enables them to form enzymatically active dimers. Thus, RT and IN homodimerization partially offsets the dimerization inhibitory effect of the p6* residues at the N-terminus of PR, (Ludwig et al., 2008) which is not envisaged to be relieved on its own. By requiring RT and IN dimerization to drive PR activation, the virus commits itself to an irreversible process of maturation. It also ensures that enzymes needed during downstream processes of reverse transcription and integration are assembled and ready for the next stage of the viral life cycle. It must also be noted that the NC domains of these two Gag-Pol dimers when bound to nucleic acid simultaneously could also seed the initial dimerization and bring the RT and IN together and further stabilize the Gag-Pol/Gag-Pol dimer akin to a ‘swing’. The positive contributions of NC and RT to PR activation do not need to be mutually exclusive.

The RT embedded in the Pol exhibits both polymerase and ribonuclease H activity comparable to the mature RT. On the basis of the fact that the IN domain has very specific binding to nucleotides in the genome of the virus especially in the 5’-UTR where reverse transcription of the viral genome is initiated, it is conceivable that initiation of reverse transcription of the viral genome is carried out by remnants of the Pol polyprotein contrary to the usual notion that this process is carried out by the mature RT.

5.11 Materials and Methods

5.11.1 Protein expression and purification

HIV-1 Pol constructs of the BH10 strain cloned into a pET28a vector were transformed into BL21 DE3 CodonPlus RIL cells. Several colonies were selected and inoculated into 100 mL of overnight culture in a 500 mL Erlenmeyer flask containing 50 µg/mL kanamycin and 34 µg/mL of chloramphenicol. The culture was shaken at 230-250 rpm overnight at 37ºC. Media for overnight culture composed of 1.5% tryptone, 1.0% yeast extract, 1.5% NaCl, 1.5% NZ Amine, and 50 mM $MgSO_4$ at pH 6.5. The 100 mL culture was diluted into 1 L of media at pH 6.0, supplemented with kanamycin and allowed to grow to an O.D of 2-2.5 at 37 ºC before being transferred to a shaker pre-cooled to 15 ºC, and allowed to grow for at least 1 hour. Induction of protein expression was carried out by the addition of 1 mM IPTG and culture allowed to grow for at least 17 hours. 50 mM phosphate buffer at pH 6.0 may be added to the media as a buffering agent from the beginning or when the culture is transferred to 15ºC prior to induction.

Cells were harvested by spinning down the culture at 4 000Xg for 30 minutes, resuspended in 100 mM phosphate or Tris-Cl buffer at pH 8.0 supplemented with 600 mM NaCl, 0.5% Triton X-100, 10% glycerol, 30 mM imidazole, and 1 mM TCEP at a minimum ratio of 10 mL/g of cells on ice. 1 mM PMSF, and 1 µM each of pepstatin A and leupeptin were added to uniformly homogenized cells and sonicated for at least 10 minutes with 30 seconds pulse and pause cycles on ice. The cellular debris were spun down at 38 000Xg for 30 minutes and the supernatant loaded onto a nickel gravity column pre-equilibrated with the resuspension buffer.

The column was subsequently washed with 5 column volumes (CV) of resuspension buffer followed by a 10 CV of high salt buffer wash containing 1.5 M NaCl in the resuspension buffer. A chaperone wash of at least 10 CV is also carried out after high salt wash. Chaperone buffer contains 5 mM ATP, 5 mM MgCl2, and 50 mM imidazole in the resuspension buffer. 2 CV wash with the resuspension buffer is carried out to remove all the chaperone buffer before protein is eluted with at least 4 CV of 80 mM Tris pH 8.0, 600 mM NaCl, 500 mM imidazole, and 10% glycerol.

Eluted protein is supplemented with 2 mM TCEP, diluted 2-fold with water and loaded onto a 5 mL HiTrap heparin column connected to an FPLC and pre-equilibrated with 30 mM Tris-Cl pH 8.0, 300 mM NaCl, 5% glycerol, and 1 mM TCEP. The column was washed with this buffer until the background UV absorption is negligible. Elution of protein from the heparin column was carried out by washing the column with the wash buffer containing 1.0 M NaCl. Eluted protein was concentrated and injected onto a Superose 6 Increase 10/30 GL gel filtration column pre-equilibrated with 20 mM Tris-Cl pH 8.0 and 250-300 mM NaCl. Fractions containing pure protein were pooled together, concentrated to 0.2-0.5 mg/L, flash frozen and stored at -80 ºC.

To cleave the tag on the proteins, equimolar amounts of HRV14 3C protease prepared in-house is added to the protein after heparin step, diluted with buffer to ensure NaCl concentration is below 300 mM, passed through a nickel column, and the flow-through re-purified on the heparin and gel filtration column as described before storage. Sub-stoichiometric amounts of the 3C protease lead to partial cleavage of the expression tag that becomes resistant to cleavage. For constructs co-expressed with MBP-IBD, glycerol was removed from the buffers. Protein was eluted from the heparin column using Tris buffer containing 1.0 M NaCl directly onto an MBPTrap HP column prepacked with Dextran Sepharose, washed with this buffer until the UV base line is negligible and eluted with 30 mM Tris pH 8.0, 600 mM NaCl, and 10 mM maltose before the gel filtration.

5.6.2 SAXS data processing

SAXS data was collected at SIBYLS beamline, Advanced Light Source in Berkeley, CA using the size exclusion-in-line SAXS setup. Averaging and buffer subtraction was carried out at the beam line. Data file averaging and buffer subtraction as well as Guinier and Gnom analysis for the determination of Rg and Dmax were carried out using the BioXTAS RAW software package (Hopkins et al., 2017). *Ab initio* envelop reconstructions were carried out using DAMMIF on the Atsas online web server (Petoukhov et al., 2012, Franke et al., 2017).

References

ADAMCZYK, M. & JAGURA-BURDZY, G. 2003. Spread and survival of promiscuous IncP-1 plasmids. *Acta Biochim Pol,* 50**,** 425-53.

ADAMS, P. D., AFONINE, P. V., BUNKOCZI, G., CHEN, V. B., ECHOLS, N., HEADD, J. J., HUNG, L. W., JAIN, S., KAPRAL, G. J., GROSSE KUNSTLEVE, R. W., MCCOY, A. J., MORIARTY, N. W., OEFFNER, R. D., READ, R. J., RICHARDSON, D. C., RICHARDSON, J. S., TERWILLIGER, T. C. & ZWART, P. H. 2011. The Phenix software for automated determination of macromolecular structures. *Methods,* 55**,** 94-106.

ADAMSON, C. S. 2012. Protease-Mediated Maturation of HIV: Inhibitors of Protease and the Maturation Process. *Mol Biol Int,* 2012**,** 604261.

ADAMSON, C. S. & FREED, E. O. 2007. Human immunodeficiency virus type 1 assembly, release, and maturation. *Adv Pharmacol,* 55**,** 347-87.

ALI, A., BANDARANAYAKE, R. M., CAI, Y., KING, N. M., KOLLI, M., MITTAL, S., MURZYCKI, J. F., NALAM, M. N., NALIVAIKA, E. A., OZEN, A., PRABU-JEYABALAN, M. M., THAYER, K. & SCHIFFER, C. A. 2010. Molecular Basis for Drug Resistance in HIV-1 Protease. *Viruses,* 2**,** 2509-35.

AMANN, E., OCHS, B. & ABEL, K. J. 1988. Tightly regulated tac promoter vectors useful for the expression of unfused and fused proteins in Escherichia coli. *Gene,* 69**,** 301-15.

BABUSHOK, D. V. & KAZAZIAN, H. H., JR. 2007. Progress in understanding the biology of the human mutagen LINE-1. *Hum Mutat,* 28**,** 527-39.

BALDWIN, D. N. & LINIAL, M. L. 1998. The roles of Pol and Env in the assembly pathway of human foamy virus. *J Virol,* 72**,** 3658-65.

BAUMAN, J. D., HARRISON, J. J. & ARNOLD, E. 2016. Rapid experimental SAD phasing and hot-spot identification with halogenated fragments. *IUCrJ,* 3**,** 51-60.

BERKA, U., HAMANN, M. V. & LINDEMANN, D. 2013. Early events in foamy virus-host interaction and intracellular trafficking. *Viruses,* 5**,** 1055-74.

BERNADO, P. 2010. Effect of interdomain dynamics on the structure determination of modular proteins by small-angle scattering. *Eur Biophys J,* 39**,** 769-80.

BERTANI, G. 2004. Lysogeny at mid-twentieth century: P1, P2, and other experimental systems. *J Bacteriol,* 186**,** 595-600.

BHAT, T. N., BALDWIN, E. T., LIU, B., CHENG, Y. S. & ERICKSON, J. W. 1994. Crystal structure of a tethered dimer of HIV-1 proteinase complexed with an inhibitor. *Nat Struct Biol,* 1**,** 552-6.

BIRD, L. E., NETTLESHIP, J. E., JARVINEN, V., RADA, H., VERMA, A. & OWENS, R. J. 2016. Expression Screening of Integral Membrane Proteins by Fusion to Fluorescent Reporters. *Adv Exp Med Biol,* 922**,** 1-11.

BODEM, J., LOCHELT, M., YANG, P. & FLUGEL, R. M. 1997. Regulation of gene expression by human foamy virus and potentials of foamy viral vectors. *Stem Cells,* 15 Suppl 1**,** 141-7.

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**,** 6112-21.

BRADIC, M., WARRING, S. D., LOW, V. & CARLTON, J. M. 2014. The Tc1/mariner transposable element family shapes genetic variation and gene expression in the protist Trichomonas vaginalis. *Mob DNA,* 5**,** 12.

BUSSCHOTS, K., VERCAMMEN, J., EMILIANI, S., BENAROUS, R., ENGELBORGHS, Y., CHRIST, F. & DEBYSER, Z. 2005. The interaction of LEDGF/p75 with integrase is lentivirus-specific and promotes DNA binding. *J Biol Chem,* 280**,** 17841-7.

CAMPBELL, E. M. & HOPE, T. J. 2015. HIV-1 capsid: the multifaceted key player in HIV-1 infection. *Nat Rev Microbiol,* 13**,** 471-83.

CHEN, J. C., KRUCINSKI, J., MIERCKE, L. J., FINER-MOORE, J. S., TANG, A. H., LEAVITT, A. D. & STROUD, R. M. 2000. Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: a model for viral DNA binding. *Proc Natl Acad Sci U S A,* 97**,** 8233-8.

CHERRY, E., LIANG, C., RONG, L., QUAN, Y., INOUYE, P., LI, X., MORIN, N., KOTLER, M. & WAINBERG, M. A. 1998a. Characterization of human immunodeficiency virus type-1 (HIV-1) particles that express protease-reverse transcriptase fusion proteins. *J Mol Biol,* 284**,** 43-56.

CHERRY, E., MORIN, N. & WAINBERG, M. A. 1998b. Effect of HIV constructs containing protease-reverse transcriptase fusion proteins on viral replication. *AIDS,* 12**,** 967-75.

CHIANG, C. C., TSENG, Y. T., HUANG, K. J., PAN, Y. Y. & WANG, C. T. 2012. Mutations in the HIV-1 reverse transcriptase tryptophan repeat motif affect virion maturation and Gag-Pol packaging. *Virology,* 422**,** 278-87.

CHRISTENSEN, D. G., ORR, J. S., RAO, C. V. & WOLFE, A. J. 2017. Increasing Growth Yield and Decreasing Acetylation in Escherichia coli by Optimizing the Carbon-to-Magnesium Ratio in Peptide-Based Media. *Appl Environ Microbiol,* 83.

COMAS-GARCIA, M., DAVIS, S. R. & REIN, A. 2016. On the Selective Packaging of Genomic RNA by HIV-1. *Viruses,* 8.

CORDAUX, R. & BATZER, M. A. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet,* 10**,** 691-703.

CRAIG, D. B. & DOMBKOWSKI, A. A. 2013. Disulfide by Design 2.0: a web-based tool for disulfide engineering in proteins. *BMC Bioinformatics,* 14**,** 346.

CZARNOCKI-CIECIURA, M. & NOWOTNY, M. 2016. Introduction to high-resolution cryo-electron microscopy. *Postepy Biochem,* 62**,** 383-394.

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**,** 253-9.

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**,** 7308-7313.

DAVIS, D. A., SOULE, E. E., DAVIDOFF, K. S., DANIELS, S. I., NAIMAN, N. E. & YARCHOAN, R. 2012. Activity of human immunodeficiency virus type 1 protease inhibitors against the initial autocleavage in Gag-Pol polyprotein processing. *Antimicrob Agents Chemother,* 56**,** 3620-8.

DE BOER, H. A., COMSTOCK, L. J. & VASSER, M. 1983. The tac promoter: a functional hybrid derived from the trp and lac promoters. *Proc Natl Acad Sci U S A,* 80**,** 21-5.

DEL SOLAR, G., GIRALDO, R., RUIZ-ECHEVARRIA, M. J., ESPINOSA, M. & DIAZ-OREJAS, R. 1998. Replication and control of circular bacterial plasmids. *Microbiol Mol Biol Rev,* 62**,** 434-64.

DICK, R. A., ZADROZNY, K. K., XU, C., SCHUR, F. K. M., LYDDON, T. D., RICANA, C. L., WAGNER, J. M., PERILLA, J. R., GANSER-PORNILLOS, B. K., JOHNSON, M. C., PORNILLOS, O. & VOGT, V. M. 2018. Inositol phosphates are assembly co-factors for HIV-1. *Nature,* 560**,** 509-512.

DING, J., DAS, K., HSIOU, Y., SARAFIANOS, S. G., CLARK, A. D., JR., JACOBO-MOLINA, A., TANTILLO, C., HUGHES, S. H. & 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**,** 1095-111.

DING, J., JACOBO-MOLINA, A., TANTILLO, C., LU, X., NANNI, R. G. & ARNOLD, E. 1994. Buried surface analysis of HIV-1 reverse transcriptase p66/p51 heterodimer and its interaction with dsDNA template/primer. *J Mol Recognit,* 7**,** 157-61.

DUMON-SEIGNOVERT, L., CARIOT, G. & VUILLARD, L. 2004. The toxicity of recombinant proteins in Escherichia coli: a comparison of overexpression in BL21(DE3), C41(DE3), and C43(DE3). *Protein Expr Purif,* 37**,** 203-6.

EMSLEY, P., LOHKAMP, B., SCOTT, W. G. & COWTAN, K. 2010. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr,* 66**,** 486-501.

ENGELMAN, A. & CHEREPANOV, P. 2012. The structural biology of HIV-1: mechanistic and therapeutic insights. *Nat Rev Microbiol,* 10**,** 279-90.

ENSSLE, J., JORDAN, I., MAUER, B. & RETHWILM, A. 1996. Foamy virus reverse transcriptase is expressed independently from the Gag protein. *Proc Natl Acad Sci U S A,* 93**,** 4137-41.

ERLWEIN, O. & MCCLURE, M. O. 2010. Progress and prospects: foamy virus vectors enter a new age. *Gene Ther,* 17**,** 1423-9.

ESPOSITO, F., CORONA, A. & TRAMONTANO, E. 2012. HIV-1 Reverse Transcriptase Still Remains a New Drug Target: Structure, Function, Classical Inhibitors, and New Inhibitors with Innovative Mechanisms of Actions. *Mol Biol Int,* 2012**,** 586401.

FAN, N., RANK, K. B., LEONE, J. W., HEINRIKSON, R. L., BANNOW, C. A., SMITH, C. W., EVANS, D. B., POPPE, S. M., TARPLEY, W. G., ROTHROCK, D. J. & ET AL. 1995. The differential processing of homodimers of reverse transcriptases from human immunodeficiency viruses type 1 and 2 is a consequence of the distinct specificities of the viral proteases. *J Biol Chem,* 270**,** 13573-9.

FEDOROFF, N. V. 2012. Presidential address. Transposable elements, epigenetics, and genome evolution. *Science,* 338**,** 758-67.

FLUGEL, R. M. & PFREPPER, K. I. 2003. Proteolytic processing of foamy virus Gag and Pol proteins. *Curr Top Microbiol Immunol,* 277**,** 63-88.

FRANCIS, D. M. & PAGE, R. 2010. Strategies to optimize protein expression in E. coli. *Curr Protoc Protein Sci,* Chapter 5**,** Unit 5 24 1-29.

FRANK, J. 2016. Generalized single-particle cryo-EM--a historical perspective. *Microscopy (Oxf),* 65**,** 3-8.

FRANKE, D., PETOUKHOV, M. V., KONAREV, P. V., PANJKOVICH, A., TUUKKANEN, A., MERTENS, H. D. T., KIKHNEY, A. G., HAJIZADEH, N. R., FRANKLIN, J. M., JEFFRIES, C. M. & SVERGUN, D. I. 2017. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J Appl Crystallogr,* 50**,** 1212-1225.

FREED, E. O. 2015. HIV-1 assembly, release and maturation. *Nat Rev Microbiol,* 13**,** 484-96.

GANSER-PORNILLOS, B. K., YEAGER, M. & SUNDQUIST, W. I. 2008. The structural biology of HIV assembly. *Curr Opin Struct Biol,* 18**,** 203-17.

GELINAS, J. F., GILL, D. R. & HYDE, S. C. 2018. Multiple Inhibitory Factors Act in the Late Phase of HIV-1 Replication: a Systematic Review of the Literature. *Microbiol Mol Biol Rev,* 82.

GERONDELIS, P., ARCHER, R. H., PALANIAPPAN, C., REICHMAN, R. C., FAY, P. J., BAMBARA, R. A. & DEMETER, L. M. 1999. The P236L delavirdine-resistant human immunodeficiency virus type 1 mutant is replication defective and demonstrates alterations in both RNA 5'-end- and DNA 3'-end-directed RNase H activities. *J Virol,* 73**,** 5803-13.

GLAESER, R. M., HAN, B. G., CSENCSITS, R., KILLILEA, A., PULK, A. & CATE, J. H. 2016. Factors that Influence the Formation and Stability of Thin, Cryo-EM Specimens. *Biophys J,* 110**,** 749-55.

GRODBERG, J. & DUNN, J. J. 1988. ompT encodes the Escherichia coli outer membrane protease that cleaves T7 RNA polymerase during purification. *J Bacteriol,* 170**,** 1245-53.

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**,** 155-161.

HAMANN, M. V. & LINDEMANN, D. 2016. Foamy Virus Protein-Nucleic Acid Interactions during Particle Morphogenesis. *Viruses,* 8.

HARE, S., GUPTA, S. S., VALKOV, E., ENGELMAN, A. & CHEREPANOV, P. 2010. Retroviral intasome assembly and inhibition of DNA strand transfer. *Nature,* 464**,** 232-6.

HARTL, M. J., BODEM, J., JOCHHEIM, F., RETHWILM, A., ROSCH, P. & WOHRL, B. M. 2011. Regulation of foamy virus protease activity by viral RNA: a novel and unique mechanism among retroviruses. *J Virol,* 85**,** 4462-9.

HARTL, M. J., MAYR, F., RETHWILM, A. & WOHRL, B. M. 2010a. Biophysical and enzymatic properties of the simian and prototype foamy virus reverse transcriptases. *Retrovirology,* 7**,** 5.

HARTL, M. J., SCHWEIMER, K., REGER, M. H., SCHWARZINGER, S., BODEM, J., ROSCH, P. & WOHRL, B. M. 2010b. Formation of transient dimers by a retroviral protease. *Biochem J,* 427**,** 197-203.

HENDRICKSON, W. A., HORTON, J. R. & LEMASTER, D. M. 1990. Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *EMBO J,* 9**,** 1665-72.

HENDRICKSON, W. A., PAHLER, A., SMITH, J. L., SATOW, Y., MERRITT, E. A. & PHIZACKERLEY, R. P. 1989. Crystal structure of core streptavidin determined from multiwavelength anomalous diffraction of synchrotron radiation. *Proc Natl Acad Sci U S A,* 86**,** 2190-4.

HOPKINS, J. B., GILLILAN, R. E. & SKOU, S. 2017. BioXTAS RAW: improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis. *J Appl Crystallogr,* 50**,** 1545-1553.

HUANG, C. J., LIN, H. & YANG, X. 2012. Industrial production of recombinant therapeutics in Escherichia coli and its recent advancements. *J Ind Microbiol Biotechnol,* 39**,** 383-99.

HUTTER, S., ZURNIC, I. & LINDEMANN, D. 2013. Foamy virus budding and release. *Viruses,* 5**,** 1075-98.

IANCU, C. V., TIVOL, W. F., SCHOOLER, J. B., DIAS, D. P., HENDERSON, G. P., MURPHY, G. E., WRIGHT, E. R., LI, Z., YU, Z., BRIEGEL, A., GAN, L., HE, Y. & JENSEN, G. J. 2006. Electron cryotomography sample preparation using the Vitrobot. *Nat Protoc,* 1**,** 2813-9.

JACOB, F., PERRIN, D., SANCHEZ, C., MONOD, J. & EDELSTEIN, S. 2005. [The operon: a group of genes with expression coordinated by an operator. C.R.Acad. Sci. Paris 250 (1960) 1727-1729]. *C R Biol,* 328**,** 514-20.

JACOBO-MOLINA, A. & ARNOLD, E. 1991. HIV reverse transcriptase structure-function relationships. *Biochemistry,* 30**,** 6351-6.

JENKINS, T. M., ENGELMAN, A., GHIRLANDO, R. & CRAIGIE, R. 1996. A soluble active mutant of HIV-1 integrase: involvement of both the core and carboxyl-terminal domains in multimerization. *J Biol Chem,* 271**,** 7712-8.

JEONG, H., BARBE, V., LEE, C. H., VALLENET, D., YU, D. S., CHOI, S. H., COULOUX, A., LEE, S. W., YOON, S. H., CATTOLICO, L., HUR, C. G., PARK, H. S., SEGURENS, B., KIM, S. C., OH, T. K., LENSKI, R. E., STUDIER, F. W., DAEGELEN, P. & KIM, J. F. 2009. Genome sequences of Escherichia coli B strains REL606 and BL21(DE3). *J Mol Biol,* 394**,** 644-52.

JORDAN, I., ENSSLE, J., GUTTLER, E., MAUER, B. & RETHWILM, A. 1996. Expression of human foamy virus reverse transcriptase involves a spliced pol mRNA. *Virology,* 224**,** 314-9.

KAPLAN, A. H., KROGSTAD, P., KEMPF, D. J., NORBECK, D. W. & SWANSTROM, R. 1994. Human immunodeficiency virus type 1 virions composed of unprocessed Gag and Gag-Pol precursors are capable of reverse transcribing viral genomic RNA. *Antimicrob Agents Chemother,* 38**,** 2929-33.

KARN, J. & STOLTZFUS, C. M. 2012. Transcriptional and posttranscriptional regulation of HIV-1 gene expression. *Cold Spring Harb Perspect Med,* 2**,** a006916.

KATOH, I., IKAWA, Y. & YOSHINAKA, Y. 1989. Retrovirus protease characterized as a dimeric aspartic proteinase. *J Virol,* 63**,** 2226-32.

KIKHNEY, A. G. & SVERGUN, D. I. 2015. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett,* 589**,** 2570-7.

KIM, S., JEONG, H., KIM, E. Y., KIM, J. F., LEE, S. Y. & YOON, S. H. 2017. Genomic and transcriptomic landscape of Escherichia coli BL21(DE3). *Nucleic Acids Res,* 45**,** 5285-5293.

KOHLSTAEDT, L. A., WANG, J., FRIEDMAN, J. M., RICE, P. A. & STEITZ, T. A. 1992. Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science,* 256**,** 1783-90.

KONVALINKA, J., KRAUSSLICH, H. G. & MULLER, B. 2015. Retroviral proteases and their roles in virion maturation. *Virology,* 479-480**,** 403-17.

KOSOBOKOVA, E. N., SKRYPNIK, K. A. & KOSORUKOV, V. S. 2016. Overview of Fusion Tags for Recombinant Proteins. *Biochemistry (Mosc),* 81**,** 187-200.

KRUPOVIC, M., BLOMBERG, J., COFFIN, J. M., DASGUPTA, I., FAN, H., GEERING, A. D., GIFFORD, R., HARRACH, B., HULL, R., JOHNSON, W., KREUZE, J. F., LINDEMANN, D., LLORENS, C., LOCKHART, B., MAYER, J., MULLER, E., OLSZEWSKI, N., PAPPU, H. R., POOGGIN, M., RICHERT-POGGELER, K. R., SABANADZOVIC, S., SANFACON, H., SCHOELZ, J. E., SEAL, S., STAVOLONE, L., STOYE, J. P., TEYCHENEY, P. Y., TRISTEM, M., KOONIN,

E. V. & KUHN, J. H. 2018. Ortervirales: A new viral order unifying five families of reverse-transcribing viruses. *J Virol*.

LAGANOWSKY, A., BENESCH, J. L., LANDAU, M., DING, L., SAWAYA, M. R., CASCIO, D., HUANG, Q., ROBINSON, C. V., HORWITZ, J. & EISENBERG, D. 2010. Crystal structures of truncated alphaA and alphaB crystallins reveal structural mechanisms of polydispersity important for eye lens function. *Protein Sci,* 19**,** 1031-43.

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M., FITZHUGH, W., FUNKE, R., GAGE, D., HARRIS, K., HEAFORD, A., HOWLAND, J., KANN, L., LEHOCZKY, J., LEVINE, R., MCEWAN, P., MCKERNAN, K., MELDRIM, J., MESIROV, J. P., MIRANDA, C., MORRIS, W., NAYLOR, J., RAYMOND, C., ROSETTI, M., SANTOS, R., SHERIDAN, A., SOUGNEZ, C., STANGE-THOMANN, Y., STOJANOVIC, N., SUBRAMANIAN, A., WYMAN, D., ROGERS, J., SULSTON, J., AINSCOUGH, R., BECK, S., BENTLEY, D., BURTON, J., CLEE, C., CARTER, N., COULSON, A., DEADMAN, R., DELOUKAS, P., DUNHAM, A., DUNHAM, I., DURBIN, R., FRENCH, L., GRAFHAM, D., GREGORY, S., HUBBARD, T., HUMPHRAY, S., HUNT, A., JONES, M., LLOYD, C., MCMURRAY, A., MATTHEWS, L., MERCER, S., MILNE, S., MULLIKIN, J. C., MUNGALL, A., PLUMB, R., ROSS, M., SHOWNKEEN, R., SIMS, S., WATERSTON, R. H., WILSON, R. K., HILLIER, L. W., MCPHERSON, J. D., MARRA, M. A., MARDIS, E. R., FULTON, L. A., CHINWALLA, A. T., PEPIN, K. H., GISH, W. R., CHISSOE, S. L., WENDL, M. C., DELEHAUNTY, K. D., MINER, T. L., DELEHAUNTY, A., KRAMER, J. B., COOK, L. L., FULTON, R. S., JOHNSON, D. L., MINX, P. J., CLIFTON, S. W., HAWKINS, T., BRANSCOMB, E., PREDKI, P., RICHARDSON, P., WENNING, S., SLEZAK, T., DOGGETT, N., CHENG, J. F., OLSEN, A., LUCAS, S., ELKIN, C., UBERBACHER, E., FRAZIER, M., et al. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

LARSEN, K. P., MATHIHARAN, Y. K., KAPPEL, K., COEY, A. T., CHEN, D. H., BARRERO, D., MADIGAN, L., PUGLISI, J. D., SKINIOTIS, G. & PUGLISI, E. V. 2018. Architecture of an HIV-1 reverse transcriptase initiation complex. *Nature,* 557**,** 118-122.

LEDERBERG, J. 1998. Plasmid (1952-1997). *Plasmid,* 39**,** 1-9.

LEE, E. G., ROY, J., JACKSON, D., CLARK, P., BOYER, P. L., HUGHES, S. H. & LINIAL, M. L. 2011. Foamy retrovirus integrase contains a Pol dimerization domain required for protease activation. *J Virol,* 85**,** 1655-61.

LEE, E. G., STENBAK, C. R. & LINIAL, M. L. 2013. Foamy virus assembly with emphasis on pol encapsidation. *Viruses,* 5**,** 886-900.

LEE, S. K., POTEMPA, M. & SWANSTROM, R. 2012. The choreography of HIV-1 proteolytic processing and virion assembly. *J Biol Chem,* 287**,** 40867-74.

LESSARD, J. C. 2013. Growth media for E. coli. *Methods Enzymol,* 533**,** 181-9.

LI, Y. 2011. Self-cleaving fusion tags for recombinant protein production. *Biotechnol Lett,* 33**,** 869-81.

LI, Y., XUAN, S., FENG, Y. & YAN, A. 2015. Targeting HIV-1 integrase with strand transfer inhibitors. *Drug Discov Today,* 20**,** 435-49.

LINDEMANN, D. & RETHWILM, A. 2011. Foamy virus biology and its application for vector development. *Viruses,* 3**,** 561-85.

LINGAPPA, J. R., REED, J. C., TANAKA, M., CHUTIRAKA, K. & ROBINSON, B. A. 2014. How HIV-1 Gag assembles in cells: Putting together pieces of the puzzle. *Virus Res,* 193**,** 89-107.

LINIAL, M. L. 1999. Foamy viruses are unconventional retroviruses. *J Virol,* 73**,** 1747-55.

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.

LLANO, M., VANEGAS, M., HUTCHINS, N., THOMPSON, D., DELGADO, S. & POESCHLA, E. M. 2006. Identification and characterization of the chromatin-binding domains of the HIV-1 integrase interactor LEDGF/p75. *J Mol Biol,* 360**,** 760-73.

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**,** 1033-40.

LORIMER, G. H. 1996. A quantitative assessment of the role of the chaperonin proteins in protein folding in vivo. *FASEB J,* 10**,** 5-9.

LOUIS, J. M., CLORE, G. M. & GRONENBORN, A. M. 1999. Autoprocessing of HIV-1 protease is tightly coupled to protein folding. *Nat Struct Biol,* 6**,** 868-75.

LOW, J. T., GARCIA-MIRANDA, P., MOUZAKIS, K. D., GORELICK, R. J., BUTCHER, S. E. & WEEKS, K. M. 2014. Structure and dynamics of the HIV-1 frameshift element RNA. *Biochemistry,* 53**,** 4282-91.

LUDWIG, C., LEIHERER, A. & WAGNER, R. 2008. Importance of protease cleavage sites within and flanking human immunodeficiency virus type 1 transframe protein p6* for spatiotemporal regulation of protease activation. *J Virol,* 82**,** 4573-84.

LUND, P., TRAMONTI, A. & DE BIASE, D. 2014. Coping with low pH: molecular strategies in neutralophilic bacteria. *FEMS Microbiol Rev,* 38**,** 1091-125.

MANNIGEL, I., STANGE, A., ZENTGRAF, H. & LINDEMANN, D. 2007. Correct capsid assembly mediated by a conserved YXXLGL motif in prototype foamy virus Gag is essential for infectivity and reverse transcription of the viral genome. *J Virol,* 81**,** 3317-26.

MARQUEZ, C. L., LAU, D., WALSH, J., SHAH, V., MCGUINNESS, C., WONG, A., AGGARWAL, A., PARKER, M. W., JACQUES, D. A., TURVILLE, S. & BOCKING, T. 2018. Kinetics of HIV-1 capsid uncoating revealed by single-molecule analysis. *Elife,* 7.

MARTIN, T. G., BHARAT, T. A., JOERGER, A. C., BAI, X. C., PRAETORIUS, F., FERSHT, A. R., DIETZ, H. & SCHERES, S. H. 2016. Design of a molecular support for cryo-EM structure determination. *Proc Natl Acad Sci U S A,* 113**,** E7456-E7463.

MATHEW, S. F., CROWE-MCAULIFFE, C., GRAVES, R., CARDNO, T. S., MCKINNEY, C., POOLE, E. S. & TATE, W. P. 2015. The highly conserved codon following the slippery sequence supports -1 frameshift efficiency at the HIV-1 frameshift site. *PLoS One,* 10**,** e0122176.

MATTEI, S., SCHUR, F. K. & BRIGGS, J. A. 2016. Retrovirus maturation-an extraordinary structural transformation. *Curr Opin Virol,* 18**,** 27-35.

MCPHERSON, A. 2004. Introduction to protein crystallization. *Methods,* 34**,** 254-65.

MCPHERSON, A. 2017. Protein Crystallization. *Methods Mol Biol,* 1607**,** 17-50.

MCPHERSON, A. & CUDNEY, B. 2014. Optimization of crystallization conditions for biological macromolecules. *Acta Crystallogr F Struct Biol Commun,* 70**,** 1445-67.

MCPHERSON, A. & KUZNETSOV, Y. G. 2014. Mechanisms, kinetics, impurities and defects: consequences in macromolecular crystallization. *Acta Crystallogr F Struct Biol Commun,* 70**,** 384-403.

METIFIOT, M., JOHNSON, B. C., KISELEV, E., MARLER, L., ZHAO, X. Z., BURKE, T. R., JR., MARCHAND, C., HUGHES, S. H. & POMMIER, Y. 2016. Selectivity for strand-transfer over 3'-processing and susceptibility to clinical resistance of HIV-1 integrase inhibitors are driven by key enzyme-DNA interactions in the active site. *Nucleic Acids Res,* 44**,** 6896-906.

MILLER, M. T., TUSKE, S., DAS, K., DESTEFANO, J. J. & ARNOLD, E. 2016. Structure of HIV-1 reverse transcriptase bound to a novel 38-mer hairpin template-primer DNA aptamer. *Protein Sci,* 25**,** 46-55.

MITTAL, S., BANDARANAYAKE, R. M., KING, N. M., PRABU-JEYABALAN, M., NALAM, M. N., NALIVAIKA, E. A., YILMAZ, N. K. & SCHIFFER, C. A. 2013. Structural and thermodynamic basis of amprenavir/darunavir and atazanavir resistance in HIV-1 protease with mutations at residue 50. *J Virol,* 87**,** 4176-84.

MITTAL, S., CAI, Y., NALAM, M. N., BOLON, D. N. & SCHIFFER, C. A. 2012. Hydrophobic core flexibility modulates enzyme activity in HIV-1 protease. *J Am Chem Soc,* 134**,** 4163-8.

MURATA, K. & WOLF, M. 2018. Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochim Biophys Acta Gen Subj,* 1862**,** 324-334.

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**,** 240-55.

NEKHAI, S. & JEANG, K. T. 2006. Transcriptional and post-transcriptional regulation of HIV-1 gene expression: role of cellular factors for Tat and Rev. *Future Microbiol,* 1**,** 417-26.

NIKOLAITCHIK, O. A., DILLEY, K. A., FU, W., GORELICK, R. J., TAI, S. H., SOHEILIAN, F., PTAK, R. G., NAGASHIMA, K., PATHAK, V. K. & HU, W. S. 2013. Dimeric RNA recognition regulates HIV-1 genome packaging. *PLoS Pathog,* 9**,** e1003249.

NOWAK, E., MILLER, J. T., BONA, M. K., STUDNICKA, J., SZCZEPANOWSKI, R. H., JURKOWSKI, J., LE GRICE, S. F. & NOWOTNY, M. 2014. Ty3 reverse transcriptase complexed with an RNA-DNA hybrid shows structural and functional asymmetry. *Nat Struct Mol Biol,* 21**,** 389-96.

O'BRIEN, D. P., BRIER, S., LADANT, D., DURAND, D., CHENAL, A. & VACHETTE, P. 2018. SEC-SAXS and HDX-MS: A powerful combination. The case of the calcium-binding domain of a bacterial toxin. *Biotechnol Appl Biochem,* 65**,** 62-68.

OZEN, A., HALILOGLU, T. & SCHIFFER, C. A. 2011. Dynamics of preferential substrate recognition in HIV-1 protease: redefining the substrate envelope. *J Mol Biol,* 410**,** 726-44.

OZTURK, S., ERGUN, B. G. & CALIK, P. 2017. Double promoter expression systems for recombinant protein production by industrial microorganisms. *Appl Microbiol Biotechnol,* 101**,** 7459-7475.

PALOVCAK, E., WANG, F., ZHENG, S. Q., YU, Z., LI, S., BETEGON, M., BULKLEY, D., AGARD, D. A. & CHENG, Y. 2018. A simple and robust procedure for preparing graphene-oxide cryo-EM grids. *J Struct Biol,* 204**,** 80-84.

PAN, J. W. & MACNAB, R. M. 1990. Steady-state measurements of Escherichia coli sodium and proton potentials at alkaline pH support the hypothesis of electrogenic antiport. *J Biol Chem,* 265**,** 9247-50.

PANNU, N. S., WATERREUS, W. J., SKUBAK, P., SIKHARULIDZE, I., ABRAHAMS, J. P. & DE GRAAFF, R. A. 2011. Recent advances in the CRANK software suite for experimental phasing. *Acta Crystallogr D Biol Crystallogr,* 67**,** 331-7.

PASSOS, D. O., LI, M., YANG, R., REBENSBURG, S. V., GHIRLANDO, R., JEON, Y., SHKRIABAI, N., KVARATSKHELIA, M., CRAIGIE, R. & LYUMKIS, D. 2017. Cryo-EM structures and atomic model of the HIV-1 strand transfer complex intasome. *Science,* 355**,** 89-92.

PETI, W. & PAGE, R. 2007. Strategies to maximize heterologous protein expression in Escherichia coli with minimal cost. *Protein Expr Purif,* 51**,** 1-10.

PETOUKHOV, M. V., FRANKE, D., SHKUMATOV, A. V., TRIA, G., KIKHNEY, A. G., GAJDA, M., GORBA, C., MERTENS, H. D., KONAREV, P. V. & SVERGUN, D. I. 2012. New developments in the ATSAS program package for small-angle scattering data analysis. *J Appl Crystallogr,* 45**,** 342-350.

PETTIT, S. C., CLEMENTE, J. C., JEUNG, J. A., DUNN, B. M. & KAPLAN, A. H. 2005a. Ordered processing of the human immunodeficiency virus type 1 GagPol precursor is influenced by the context of the embedded viral protease. *J Virol,* 79**,** 10601-7.

PETTIT, S. C., EVERITT, L. E., CHOUDHURY, S., DUNN, B. M. & KAPLAN, A. H. 2004. Initial cleavage of the human immunodeficiency virus type 1 GagPol precursor by its activated protease occurs by an intramolecular mechanism. *J Virol,* 78**,** 8477-85.

PETTIT, S. C., GULNIK, S., EVERITT, L. & KAPLAN, A. H. 2003. The dimer interfaces of protease and extra-protease domains influence the activation of protease and the specificity of GagPol cleavage. *J Virol,* 77**,** 366-74.

PETTIT, S. C., LINDQUIST, J. N., KAPLAN, A. H. & SWANSTROM, R. 2005b. Processing sites in the human immunodeficiency virus type 1 (HIV-1) Gag-Pro-Pol precursor are cleaved by the viral protease at different rates. *Retrovirology,* 2**,** 66.

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**,** 7648-52.

POTTERTON, L., AGIRRE, J., BALLARD, C., COWTAN, K., DODSON, E., EVANS, P. R., JENKINS, H. T., KEEGAN, R., KRISSINEL, E., STEVENSON, K., LEBEDEV, A., MCNICHOLAS, S. J., NICHOLLS, R. A., NOBLE, M., PANNU, N. S., ROTH, C., SHELDRICK, G., SKUBAK, P., TURKENBURG, J., USKI, V., VON DELFT, F., WATERMAN, D., WILSON, K., WINN, M. & WOJDYR, M. 2018. CCP4i2: the new graphical user interface to the CCP4 program suite. *Acta Crystallogr D Struct Biol,* 74**,** 68-84.

RETHWILM, A. & BODEM, J. 2013. Evolution of foamy viruses: the most ancient of all retroviruses. *Viruses,* 5**,** 2349-74.

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**,** 7560-70.

ROY, J. & LINIAL, M. L. 2007. Role of the foamy virus Pol cleavage site in viral replication. *J Virol,* 81**,** 4956-62.

SADIQ, S. K., NOE, F. & DE FABRITIIS, G. 2012. Kinetic characterization of the critical step in HIV-1 protease maturation. *Proc Natl Acad Sci U S A,* 109**,** 20449-54.

SAHDEV, S., KHATTAR, S. K. & SAINI, K. S. 2008. Production of active eukaryotic proteins through bacterial expression systems: a review of the existing biotechnology strategies. *Mol Cell Biochem,* 307**,** 249-64.

SANMIGUEL, P., GAUT, B. S., TIKHONOV, A., NAKAJIMA, Y. & BENNETZEN, J. L. 1998. The paleontology of intergene retrotransposons of maize. *Nat Genet,* 20**,** 43-5.

SARAFIANOS, S. G., CLARK, A. D., JR., TUSKE, S., SQUIRE, C. J., DAS, K., SHENG, D., ILANKUMARAN, P., RAMESHA, A. R., KROTH, H., SAYER, J. M., JERINA, D. M., BOYER, P. L., HUGHES, S. H. & ARNOLD, E. 2003. Trapping HIV-1 reverse transcriptase before and after translocation on DNA. *J Biol Chem,* 278**,** 16280-8.

SARAFIANOS, S. G., DAS, K., DING, J., BOYER, P. L., HUGHES, S. H. & ARNOLD, E. 1999. Touching the heart of HIV-1 drug resistance: the fingers close down on the dNTP at the polymerase active site. *Chem Biol,* 6**,** R137-46.

SARAFIANOS, S. G., DAS, K., TANTILLO, C., CLARK, A. D., JR., DING, J., WHITCOMB, J. M., BOYER, P. L., HUGHES, S. H. & ARNOLD, E. 2001. Crystal structure of HIV-1 reverse transcriptase in complex with a polypurine tract RNA:DNA. *EMBO J,* 20**,** 1449-61.

SCHLAPSCHY, M. & SKERRA, A. 2011. Periplasmic chaperones used to enhance functional secretion of proteins in E. coli. *Methods Mol Biol,* 705**,** 211-24.

SCHNEIDER, A., PETER, D., SCHMITT, J., LEO, B., RICHTER, F., ROSCH, P., WOHRL, B. M. & HARTL, M. J. 2014. Structural requirements for enzymatic activities of foamy virus protease-reverse transcriptase. *Proteins,* 82**,** 375-85.

SCHNEIDMAN-DUHOVNY, D. & HAMMEL, M. 2018. Modeling Structure and Dynamics of Protein Complexes with SAXS Profiles. *Methods Mol Biol,* 1764**,** 449-473.

SCHULTZ, S. J., ZHANG, M. & CHAMPOUX, J. J. 2006. Sequence, distance, and accessibility are determinants of 5'-end-directed cleavages by retroviral RNases H. *J Biol Chem,* 281**,** 1943-55.

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**,** 17938.

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**,** 16534-9.

SINGH, P. K., PLUMB, M. R., FERRIS, A. L., IBEN, J. R., WU, X., FADEL, H. J., LUKE, B. T., ESNAULT, C., POESCHLA, E. M., HUGHES, S. H., KVARATSKHELIA, M. & LEVIN, H. L. 2015. LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. *Genes Dev,* 29**,** 2287-97.

SLONCZEWSKI, J. L., FUJISAWA, M., DOPSON, M. & KRULWICH, T. A. 2009. Cytoplasmic pH measurement and homeostasis in bacteria and archaea. *Adv Microb Physiol,* 55**,** 1-79, 317.

SLONCZEWSKI, J. L., ROSEN, B. P., ALGER, J. R. & MACNAB, R. M. 1981. pH homeostasis in Escherichia coli: measurement by 31P nuclear magnetic resonance of methylphosphonate and phosphate. *Proc Natl Acad Sci U S A,* 78**,** 6271-5.

SLUIS-CREMER, N., ARION, D., ABRAM, M. E. & PARNIAK, M. A. 2004. Proteolytic processing of an HIV-1 pol polyprotein precursor: insights into the mechanism of reverse transcriptase p66/p51 heterodimer formation. *Int J Biochem Cell Biol,* 36**,** 1836-47.

SONG, J. M., AN, Y. J., KANG, M. H., LEE, Y. H. & CHA, S. S. 2012. Cultivation at 6-10 degrees C is an effective strategy to overcome the insolubility of recombinant proteins in Escherichia coli. *Protein Expr Purif,* 82**,** 297-301.

SPANNAUS, R., HARTL, M. J., WOHRL, B. M., RETHWILM, A. & BODEM, J. 2012. The prototype foamy virus protease is active independently of the integrase domain. *Retrovirology,* 9**,** 41.

SPEER, S. L., GUSEMAN, A. J., PATTESON, J. B., EHRMANN, B. M. & PIELAK, G. J. 2019. Controlling and quantifying protein concentration in Escherichia coli. *Protein Sci,* 28**,** 1307-1311.

STEINMETZ, E. J. & AULDRIDGE, M. E. 2017. Screening Fusion Tags for Improved Recombinant Protein Expression in E. coli with the Expresso(R) Solubility and Expression Screening System. *Curr Protoc Protein Sci,* 90**,** 5 27 1-5 27 20.

STOLTENBURG, R., REINEMANN, C. & STREHLITZ, B. 2007. SELEX--a (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomol Eng,* 24**,** 381-403.

STRUCTURAL GENOMICS, C., CHINA STRUCTURAL GENOMICS, C., NORTHEAST STRUCTURAL GENOMICS, C., GRASLUND, S., NORDLUND, P., WEIGELT, J., HALLBERG, B. M., BRAY, J., GILEADI, O., KNAPP, S., OPPERMANN, U., ARROWSMITH, C., HUI, R., MING, J., DHE-PAGANON, S., PARK, H. W., SAVCHENKO, A., YEE, A., EDWARDS, A., VINCENTELLI, R., CAMBILLAU, C., KIM, R., KIM, S. H., RAO, Z., SHI, Y., TERWILLIGER, T. C., KIM, C. Y., HUNG, L. W., WALDO, G. S., PELEG, Y., ALBECK, S., UNGER, T., DYM, O., PRILUSKY, J., SUSSMAN, J. L., STEVENS, R. C., LESLEY, S. A., WILSON, I. A., JOACHIMIAK, A., COLLART, F., DEMENTIEVA, I., DONNELLY, M. I., ESCHENFELDT, W. H., KIM, Y., STOLS, L., WU, R., ZHOU, M., BURLEY, S. K., EMTAGE, J. S., SAUDER, J. M., THOMPSON, D., BAIN, K., LUZ, J., GHEYI, T., ZHANG, F., ATWELL, S., ALMO, S. C., BONANNO, J. B., FISER, A., SWAMINATHAN, S., STUDIER, F. W., CHANCE, M. R., SALI, A., ACTON, T. B., XIAO, R., ZHAO, L., MA, L. C., HUNT, J. F., TONG, L., CUNNINGHAM, K., INOUYE, M., ANDERSON, S., JANJUA, H., SHASTRY, R., HO, C. K., WANG, D., WANG, H., JIANG, M., MONTELIONE, G. T., STUART, D. I., OWENS, R. J., DAENKE, S., SCHUTZ, A., HEINEMANN, U., YOKOYAMA, S., BUSSOW, K. & GUNSALUS, K. C. 2008. Protein production and purification. *Nat Methods,* 5**,** 135-46.

STUDIER, F. W. 2005. Protein production by auto-induction in high density shaking cultures. *Protein Expr Purif,* 41**,** 207-34.

STUDIER, F. W., ROSENBERG, A. H., DUNN, J. J. & DUBENDORFF, J. W. 1990. Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol,* 185**,** 60-89.

SUDO, S., HARAGUCHI, H., HIRAI, Y., GATANAGA, H., SAKURAGI, J., MOMOSE, F. & MORIKAWA, Y. 2013. Efavirenz enhances HIV-1 gag processing at the plasma membrane through Gag-Pol dimerization. *J Virol,* 87**,** 3348-60.

SUN, Q. Y., DING, L. W., HE, L. L., SUN, Y. B., SHAO, J. L., LUO, M. & XU, Z. F. 2009. Culture of Escherichia coli in SOC medium improves the cloning efficiency of toxic protein genes. *Anal Biochem,* 394**,** 144-6.

SUNDQUIST, W. I. & KRAUSSLICH, H. G. 2012. HIV-1 assembly, budding, and maturation. *Cold Spring Harb Perspect Med,* 2**,** a006924.

TACHEDJIAN, G., MOORE, K. L., GOFF, S. P. & SLUIS-CREMER, N. 2005. Efavirenz enhances the proteolytic processing of an HIV-1 pol polyprotein precursor and reverse transcriptase homodimer formation. *FEBS Lett,* 579**,** 379-84.

TANG, C., LOUIS, J. M., ANIANA, A., SUH, J. Y. & CLORE, G. M. 2008. Visualizing transient events in amino-terminal autoprocessing of HIV-1 protease. *Nature,* 455**,** 693-6.

TAYLOR, T., DENSON, J. P. & ESPOSITO, D. 2017. Optimizing Expression and Solubility of Proteins in E. coli Using Modified Media and Induction Parameters. *Methods Mol Biol,* 1586**,** 65-82.

THOMPSON, R. F., WALKER, M., SIEBERT, C. A., MUENCH, S. P. & RANSON, N. A. 2016. An introduction to sample preparation and imaging by cryo-electron microscopy for structural biology. *Methods,* 100**,** 3-15.

TIAN, L., KIM, M. S., LI, H., WANG, J. & YANG, W. 2018. Structure of HIV-1 reverse transcriptase cleaving RNA in an RNA/DNA hybrid. *Proc Natl Acad Sci U S A,* 115**,** 507-512.

TODD, M. J., LORIMER, G. H. & THIRUMALAI, D. 1996. Chaperonin-facilitated protein folding: optimization of rate and yield by an iterative annealing mechanism. *Proc Natl Acad Sci U S A,* 93**,** 4030-5.

TOZSER, J. 2010. Comparative studies on retroviral proteases: substrate specificity. *Viruses,* 2**,** 147-65.

TOZSER, J., ZAHUCZKY, G., BAGOSSI, P., LOUIS, J. M., COPELAND, T. D., OROSZLAN, S., HARRISON, R. W. & WEBER, I. T. 2000. Comparison of the substrate specificity of the human T-cell leukemia virus and human immunodeficiency virus proteinases. *Eur J Biochem,* 267**,** 6287-95.

TROBRIDGE, G. D. 2009. Foamy virus vectors for gene transfer. *Expert Opin Biol Ther,* 9**,** 1427-36.

TURNER, B. G. & SUMMERS, M. F. 1999. Structural biology of HIV. *J Mol Biol,* 285**,** 1-32.

TYAGI, S. C., SIMON, S. R. & CARTER, C. A. 1994. Effect of pH and nonphysiological salt concentrations on human immunodeficiency virus-1 protease dimerization. *Biochem Cell Biol,* 72**,** 175-81.

VASINA, J. A. & BANEYX, F. 1997. Expression of aggregation-prone recombinant proteins at low temperatures: a comparative study of the Escherichia coli cspA and tac promoter systems. *Protein Expr Purif,* 9**,** 211-8.

VEESLER, S., MARCQ, S., LAFONT, S., ASTIER, J. P. & BOISTELLE, R. 1994. Influence of polydispersity on protein crystallization: a quasi-elastic light-scattering study applied to alpha-amylase. *Acta Crystallogr D Biol Crystallogr,* 50**,** 355-60.

VENEZIA, C. F., HOWARD, K. J., IGNATOV, M. E., HOLLADAY, L. A. & BARKLEY, M. D. 2006. Effects of efavirenz binding on the subunit equilibria of HIV-1 reverse transcriptase. *Biochemistry,* 45**,** 2779-89.

VENEZIA, C. F., MEANY, B. J., BRAZ, V. A. & BARKLEY, M. D. 2009. Kinetics of association and dissociation of HIV-1 reverse transcriptase subunits. *Biochemistry,* 48**,** 9084-93.

VERA, A., GONZALEZ-MONTALBAN, N., ARIS, A. & VILLAVERDE, A. 2007. The conformational quality of insoluble recombinant proteins is enhanced at low growth temperatures. *Biotechnol Bioeng,* 96**,** 1101-6.

VIJAYAN, R. S., ARNOLD, E. & DAS, K. 2014. Molecular dynamics study of HIV-1 RT-DNA-nevirapine complexes explains NNRTI inhibition and resistance by connection mutations. *Proteins,* 82**,** 815-29.

VOLKMAN, H. E. & STETSON, D. B. 2014. The enemy within: endogenous retroelements and autoimmune disease. *Nat Immunol,* 15**,** 415-22.

WAGNER, J. M., ZADROZNY, K. K., CHRUSTOWICZ, J., PURDY, M. D., YEAGER, M., GANSER-PORNILLOS, B. K. & PORNILLOS, O. 2016. Crystal structure of an HIV assembly and maturation switch. *Elife,* 5.

WALTER, S., LORIMER, G. H. & SCHMID, F. X. 1996. A thermodynamic coupling mechanism for GroEL-mediated unfolding. *Proc Natl Acad Sci U S A,* 93**,** 9425-30.

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**,** 7242-6.

WILKS, J. C. & SLONCZEWSKI, J. L. 2007. pH of the cytoplasm and periplasm of Escherichia coli: rapid measurement by green fluorescent protein fluorimetry. *J Bacteriol,* 189**,** 5601-7.

WOHRL, B. M. 2019. Structural and Functional Aspects of Foamy Virus Protease-Reverse Transcriptase. *Viruses,* 11.

WONDRAK, E. M. & LOUIS, J. M. 1996. Influence of flanking sequences on the dimer stability of human immunodeficiency virus type 1 protease. *Biochemistry,* 35**,** 12957-62.

WU, J., ADOMAT, J. M., RIDKY, T. W., LOUIS, J. M., LEIS, J., HARRISON, R. W. & WEBER, I. T. 1998. Structural basis for specificity of retroviral proteases. *Biochemistry,* 37**,** 4518-26.

YANDRAPALLI, N., LUBART, Q., TANWAR, H. S., PICART, C., MAK, J., MURIAUX, D. & FAVARD, C. 2016. Self assembly of HIV-1 Gag protein on lipid membranes generates PI(4,5)P2/Cholesterol nanoclusters. *Sci Rep,* 6**,** 39332.

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**,** 1579-82.

YU, S. F., SULLIVAN, M. D. & LINIAL, M. L. 1999. Evidence that the human foamy virus genome is DNA. *J Virol,* 73**,** 1565-72.

YU, X., WEBER, I. T. & HARRISON, R. W. 2014. Prediction of HIV drug resistance from genotype with encoded three-dimensional protein structure. *BMC Genomics,* 15 Suppl 5**,** S1.

ZHANG, W., CAO, S., MARTIN, J. L., MUELLER, J. D. & MANSKY, L. M. 2015. Morphology and ultrastructure of retrovirus particles. *AIMS Biophys,* 2**,** 343-369.

ZHENG, X., PEDERSEN, L. C., GABEL, S. A., MUELLER, G. A., CUNEO, M. J., DEROSE, E. F., KRAHN, J. M. & LONDON, R. E. 2014. Selective unfolding of one Ribonuclease H domain of HIV reverse transcriptase is linked to homodimer formation. *Nucleic Acids Res,* 42**,** 5361-77.

ZHENG, X., PERERA, L., MUELLER, G. A., DEROSE, E. F. & LONDON, R. E. 2015. Asymmetric conformational maturation of HIV-1 reverse transcriptase. *Elife,* 4.

ZHUO, Z., YU, Y., WANG, M., LI, J., ZHANG, Z., LIU, J., WU, X., LU, A., ZHANG, G. & ZHANG, B. 2017. Recent Advances in SELEX Technology and Aptamer Applications in Biomedicine. *Int J Mol Sci*, 18.

ZILBERSTEIN, D., AGMON, V., SCHULDINER, S. & PADAN, E. 1984. Escherichia coli intracellular pH, membrane potential, and cell growth. *J Bacteriol,* 158**,** 246-52.