Université de Montréal

Linker-scanning analysis of the HIV-1 Integrase protein

par

Tan Wang

Département de chimie

Faculté des arts et des sciences

Mémoire présenté à la Faculté des Études Supérieures
en vue de l'obtention du grade de Maître ès science (M.Sc.)
en Bio-informatique

Avril, 2006

Université de Montréal
Faculté des Études Supérieures

Ce mémoire intitulé:

Linker-scanning analysis of the HIV-1 Integrase protein

présenté par:
Tan Wang

a été évalué par un jury composé des personnes suivantes :

Président du jury : Prof. Andreea R. Schmitzer
Directeur de recherche : Prof. Joelle Pelletier
Membre du jury : Prof. Luc DesGroseillers

Mémoire accepté le : ……………….

Résumé

L'intégrase du VIH-1 (IN) est l'enzyme rétrovirale responsable de l'intégration d'une copie d'ADN viral à double brin dans le chromosome hôte. Ceci constitue une étape essentielle au cycle de vie viral. Il n'existe actuellement aucune structure de l'IN du VIH-1 de pleine longueur ni aucune structure d'une IN en complexe avec un ADN substrat. En l'absence de ces informations, des études de modélisation moléculaire et des études de mutagenèse, telles celles présentées ici, pourraient constituer une stratégie propice à l'obtention de nouvelles informations structure-fonction sur cette importante cible pharmacologique. Une banque de balayage par insertion avait été préalablement produite dans le gène de l'intégrase du VIH-1. La banque a été produite en utilisant un système de transposase Tn5. Nous avons obtenu une série de vecteurs contenant chacun une insertion de 57 paires de bases disposées aléatoirement dans le gène et donnant lieu à une insertion de 19 acides aminés au cours de l'expression protéique, peu importe le cadre de lecture. Au total, 55 mutants d'insertion uniques ont été analysés: 2 insertions dans le domaine *N*-terminal, 29 insertions dans le noyau catalytique et 24 insertions dans le domaine *C*-terminal. Les effets de l'insertion sur l'activité enzymatique ont été déterminés *in vitro*. Nous avons identifié trois régions qui ont fonctionnellement toléré diverses insertions. Celles-ci correspondent à la jonction entre le domaine *N*-terminal et le noyau catalytique, à la jonction entre le noyau catalytique et le domaine *C*-terminal et au domaine *C*-terminal de l'IN. Ces résultats corrèlent avec des études de délétion délimitant les limites des domaines et des sous-domaines de diverses INs.

Mots-clés : VIH-1, intégrase, transposon Tn5

Summary

HIV-1 integrase (IN) is the retroviral protein responsible for the insertion of a double-stranded DNA copy into host chromosome, which is an essential step during the viral life cycle. As of today, there is presently no structure available for any IN complexed with a DNA substrate, and no full-length experimental structure of HIV-1 IN is available. In the absence of this information, molecular modeling studies and mutational studies, such as the ones presented here, might be a good strategy to explore a possible solution for obtaining structure-function information on this pharmacologically important enzyme. A linker-scanning library was previously generated within the HIV-1 integrase gene. The library was generated using a Tn5 transposition system and resulted in a series of vectors each containing a single 57 base pair insertion at random locations. Insertions resulted in 19-amino acid in-frame insertions. A total of 55 unique insertion mutants were analyzed: 2 insertions within the *N*-terminal domain, 29 insertions within the catalytic core and 24 insertions within the *C*-terminal domain. The effects of the insertions on enzymatic activity were determined *in vitro*. Three regions were identified that functionally tolerated various linker-insertions. These correspond to the *N*-terminal domain/catalytic core junction, the junction between the IN catalytic core and the *C*-terminal domain and the *C*-terminal domain of IN. These results correlate with deletional studies mapping the domain and sub-domain boundaries of various IN.

Keywords : HIV-1, integrase, Tn5 transposon

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABREVIATIONS

Å	Angstrom
AIDS	Acquired Immunodeficiency Syndrome
ASV	avian sarcoma virus
DTT	dithiothreitol
EDTA	Ethylene-Diamine-Tetra-acetic-Acid
Env	env gene (coding env precursor: gp160 protein)
FIV	Feline Immunodeficiency Virus
HTLV	human T-cell leukemia virus
HIV-1	human immunodeficiency virus type 1
Gag	gag gene (coding for HIV p55)
IN	integrase
IPTG	isopropyl-beta-D-thiogalactopyranoside
kDa	kilo Dalton
LTR	long terminal repeat
NMR	Nuclear magnetic resonance
μl	microliter
μM	micromolar
M-MuLV	Moloney murine leukemia virus
MOPS	morpholinepropanesulfonic acid
NaCl	sodium chloride
PAGE	PolyAcrylamide Gel Electrophoresis

pmole	picomole
pol	pol gene (that codes for the protease, reverse transcriptase and integrase)
RSV	Rous sarcoma virus
RT	reverse transcriptase
SH3	Src-homology 3
SIV	Simian immunodeficiency virus

Chapter I

Introduction

Acquired immunodeficiency syndrome (AIDS) is caused by the human immunodeficiency virus (HIV). HIV is a retrovirus belonging to the lentivirus family that was unknown until the early 1980's, but since that time has been spread around the world and has infected millions of persons[1]. The mature HIV retrovirus core has two short strands of RNA 9200 nucleotide bases long--along with the enzymes reverse transcriptase, protease, ribonuclease, and integrase (IN). The core of the retrovirus is encased in an outer lipid envelope containing an antigen, gp120, which plays a role in the binding of the virus to the target cells with CD4 receptors[2]. The gene content of the HIV genome, similar to other retroviruses, has three major genes--*gag*, *pol*, and *env*[3]. After entering the body, HIV retrovirus enters the target cell via the CD4 receptor. The HIV particle uncoats from the envelope to release its RNA into the cellular cytoplasm. Its RNA genome is transcribed into proviral DNA by reverse transcriptase bound to the HIV RNA, which is one of the enzyme products of the pol gene. HIV proviral DNA is then inserted into host cell genomic DNA by the IN enzyme. Once the HIV proviral DNA is within the infected cell's genome, it becomes part of the host genome. The host cell can replicate the HIV provirus. The infected cell can undergo lysis with release of new HIV virions, which in turn can infect additional cells. Current antiviral drugs are either inhibitors of HIV reverse transcriptase (RT) or protease (PR) but no drugs against IN are available yet. HIV IN is a good target for drug discovery, since IN is essential for retroviral replication and production of new virus; moreover, it has no obvious functional analogue in the host (Fig.1.1).

In this thesis, the HIV-1 IN gene was subjected to random insertion mutagenesis. Individual constructs, selected from the library, were assayed for the effects on IN functions *in vitro*. The effect of individual insertion on enzymatic activity were analyzed in the context of HIV-1 monomer, dimer and tetramer model with/without DNA to gain insight into the organization of the HIV-1 integrase complex with DNA. One aim of this mutational analysis was to identify sites within the IN protein that may tolerate small insertional tags whose function may alter the target-site selection of the viral integrases. Another aim is to compare available tetramer models with our experimental data to see how they agree with ours and existing biochemical data. Moreover, the domain boundaries defined in this study might be useful in expressing minimized IN constructs for crystallization studies.

1.1 HIV-1 Integrase

Integrase catalyzes the integration of a double-stranded DNA copy of the retroviral RNA genome into the genome of a host[4]. This DNA integration reaction requires specific sequences at the 3' and 5' termini of the viral DNA (referred to as U3 and U5, respectively, in reference to their unique character) and can be done using purified IN alone *in vitro* [5]. The integration reaction is carried out in two steps: 3'-end processing, which is a hydrolytic cleavage, occurs two bases from the 3' end of the U3 and U5 termini, just 3' of a conserved CA dinucleotide, and strand transfer: the 3'-end processing reaction exposes the free 3'-hydroxyl, which is then used to perform a nucleophilic attack on the target DNA. The sites of nucleophilic attack on the two strands of the target DNA are separated by 5 bp. The 5' ends of the viral DNA are left unjoined in the resulting integration intermediate. The removal of the two unpaired nucleotides at each 5' end of the viral DNA, gap filling, and ligation are likely to be carried out by host enzymes[6, 7]. An *in vitro* reconstructed system using purified IN enzyme and model DNA substrates that consist of short oligodeoxynucleotide duplexes has been employed to explore the details of the biochemistry of IN and its catalytic mechanism (Fig. 1.2). This short oligodeoxynucleotide duplex substrate mimics the U5 or U3 ends of the viral DNA [6, 8]. Blunt-ended viral substrates can be properly processed at their 3' ends *in vitro*, and the processed substrates can subsequently be used as a strand transfer substrate and inserted into a target DNA. In addition, HIV IN also can carry out the disintegration reaction, in which a substrate that mimics one end of the viral DNA joined to the target DNA is cleaved into its viral and target DNA parts [5, 6]. The disintegration reaction is the reverse of the strand transfer reaction. There is no evidence that disintegration is biologically relevant *in vivo* but it is a useful assay to test the ability of IN to catalyze polynucleotidyl transfer reactions (Fig. 1.2).

The homooligomeric HIV-1 IN protein is 288 amino acids in length. Structural studies, amino acid sequence alignments, limited proteolysis, site-directed mutagenesis studies and complementation experiments [9-13] have revealed the presence of three distinct domains per monomer: the *N*-terminal domain, the catalytic core domain, and the *C*-terminal domain. Each domain can independently fold within each monomer. The first two domains are structurally highly conserved among retroviral and retrotransposon INs. The *N*-terminal domain (1-50) binds zinc, the core domain (50-212) contains the catalytic triad motif (D, D, 35E), and the *C*-terminal domain (213-288), binds DNA non-specifically (Fig.3). Insolubility of full-length HIV-1 IN has limited structural analyses to individual domains or two contiguous domains.

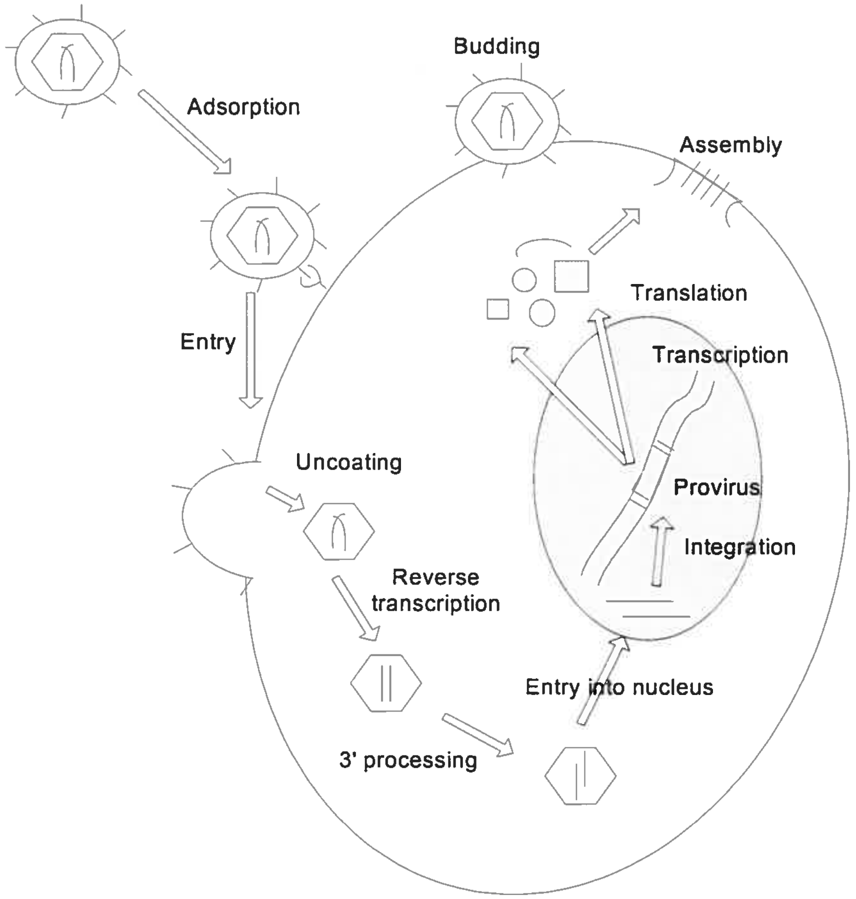

FIGURE 1.1. Overview of the retroviral life cycle. Attachment of the viral envelope surface protein to specific receptors on the surface of a host cell results in fusion and release of the viral core into the host cell cytoplasm. Reverse transcription then generates a double stranded DNA copy of the RNA genome. The viral DNA undergoes 3'-processing in the cytoplasm and subsequently travels into the nucleus, where strand transfer results in the integration of the viral DNA into the host genome to form the provirus. Transcription generates messenger RNAs as well as viron RNAs. mRNAs are translated in the cytoplasm. Virus proteins and progeny RNA assemble and bud off at the plasma membrane and subsequently mature into infectious particles.

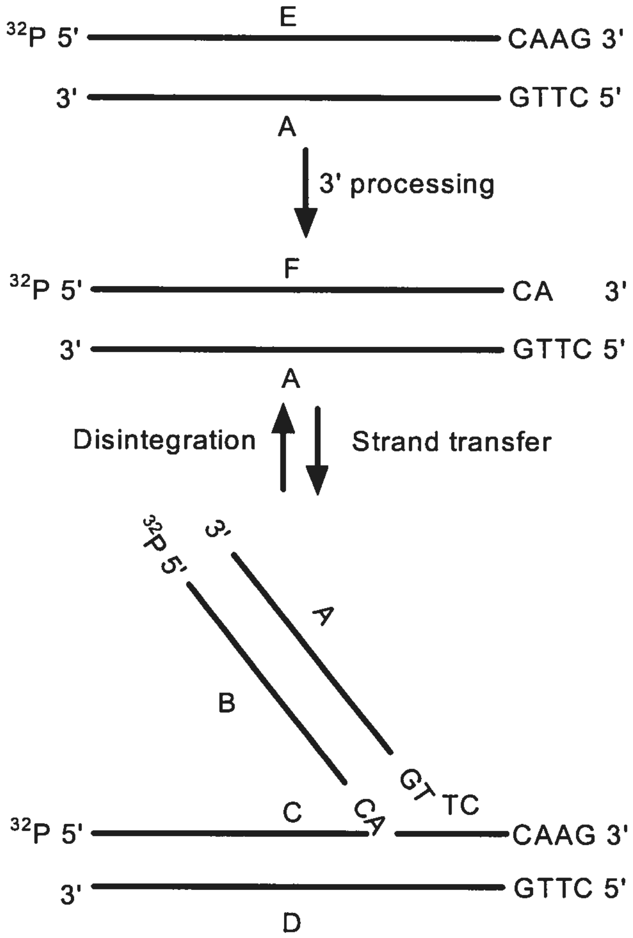

FIGURE 1.2. Schematic representation of the ^{32}P-labelled substrates and products of *in vitro* analysis of the enzymatic activities catalyzed by the retroviral IN for a single U5 end. Step 1: 3'-processing; step 2: strand transfer; and step 3: disintegration, which is the reverse of the strand-transfer reaction.

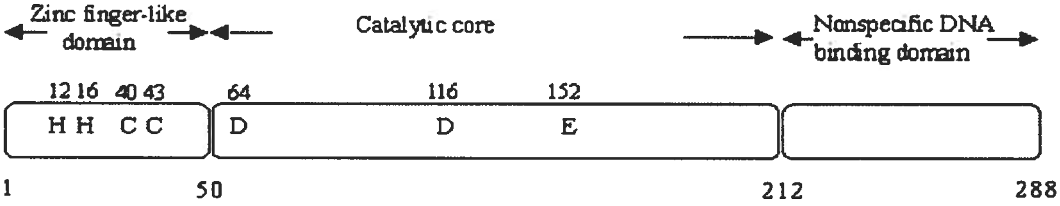

FIGURE 1.3. A schematic model of HIV-1 IN. It shows the three independently folding domains: the *N*-terminal domain, the catalytic core and the *C*-terminal domain. The conserved and catalytically active residues are indicated and the corresponding residue numbers are shown above each residue. Adapted from FIGURE 10.1 of Asante-Appiah and Skalka, 1999 [14].

The structures of individual domain structures have been determined by X-ray diffraction [15-23] or by solution NMR [24-28]. Structures of two contiguous domains, both the *N*-terminal domain plus the core domain [13] and the core domain plus the *C*-terminal domain [12, 29, 30] have also been resolved. Both individual domain and two-domain structures are also available for several other retroviral INs, which are structurally very similar to HIV-1 INs. However, there is presently no structure available for any IN complexed with DNA substrate. In the absence of this information, molecular modeling studies and mutational studies, such as the one presented here, constitute an alternative method to obtain structure-function information on this pharmacologically important enzyme.

1.1.1 Structure of the catalytic core domain

The catalytic core domain of HIV IN is well conserved among retroviral INs,shares significant structural similarity with the transposase proteins and contains the characteristic D, D (35) E motif found in polynucleotidyl transferases [26, 28, 31]. These invariant residues, Asp-64, Asp-116, and Glu-152, are key residues of the active site. Mutagenesis studies show that substitution of any of the three catalytic residues abolishes all three reactions [10, 11, 32, 33]. The core domain alone can carry out the disintegration reaction [10, 11, 34-36]. However, truncated IN proteins lacking either the *N*-terminal domain or the *C*-terminal domain cannot catalyze 3' processing and strand transfer [33, 34, 36-39]. A number of structures of the catalytic core domain of HIV-1 IN exist [12, 13, 15-18, 21, 40]. The overall topology of all these structures is similar to the structure in (Fig. 1.4). It is roughly spherical in shape and two core domains associate in the crystal to form a two-fold axis-related dimer. The dimer interface is quite large with approximately 1300 Å^2 per monomer excluded from solvent. The interface is quite hydrophobic and the primary contacts between subunits in both structures involve α-helices 1 and 5. As the corresponding helices are not involved in protein–protein contacts in the Tn5 transposase-DNA complex, which has been recently resolved [41], and as the DNA plays an important role in Tn5 transposase dimerization, one might argue against the biological relevance of this interface. However, functional IN has been suggested to be active as a tetramer [14, 42] or even an octamer [43]. This crystallized IN dimer and its interface may act as a whole to bind the same viral end and function as the single Tn5 transposase monomer does.

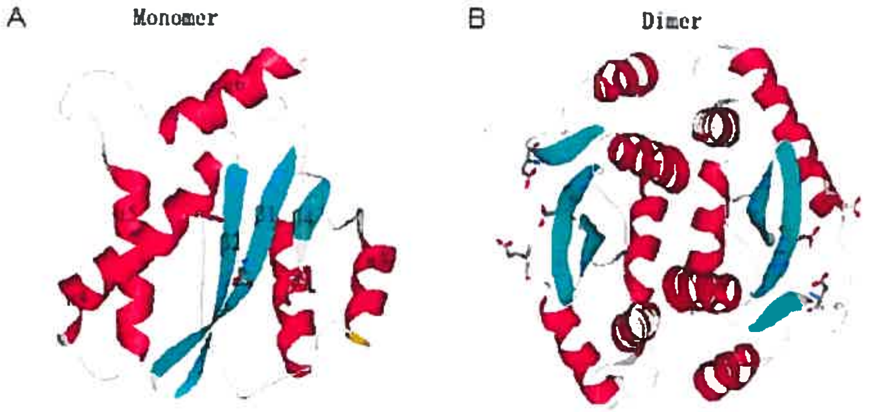

FIGURE 1.4. Structure of the catalytic core domain of HIV-1 integrase. Corresponding β-strands and α-helices are labeled in panel A, Side chains of conserved acidic residues of the active site are shown in a ball and stick representation in panel B. Coordinates from a HIV-1 IN core domain (accession codes 1ITG in the Brookhaven Protein Data Bank) was used to generate this Figure; middle panels were created using published coordinates.

Each monomer consists of a five-stranded β-sheet together with six α-helices. The three catalytic residues of the core domain are located in close proximity in the structure and define the position of the active site (Fig. 1.4). However, the two catalytic sites are on opposite sides of the spherical crystal dimer and are separated by approximately 30Å. On the other hand, the integration sites on target DNA are generally separated by 5 bp, which is equivalent to roughly 15 Å in B-form DNA. Thus, the distance observed in the dimer structure is not compatible with catalysis of the integration event. Either this crystal dimer structure is not biologically relevant or else higher order multimers are formed *in vivo* based on the dimer structure, such that two active sites could be positioned closer in a higher order multimer [15].

1.1.2 Structure of the *N*-Terminal Domain

The crystal structure of the *N*-terminal domain reveals a dimeric structure. Conserved residues His 12, His 16, Cys 40 and Cys 43, located in a three-helix bundle, ensure tetrahedral coordination of Zn^{2+}. At the monomer level, the structure of the *N*-terminal domain determined by X-ray diffraction as part of the *N*-terminal domain plus core structure is very similar to the previously determined NMR structure of the isolated *N*-terminal domain. However, the dimer interface is entirely different (Fig. 1.5) [13]. Whereas it is dominated by interactions between the third helix in the NMR structure, the dimeric interface in the IN^{1-212} crystal structure (*N*-terminal domain plus core) comprises the *N*-termini of the first and third α-helix. The surface of the dimer interface in the two-domain structure is smaller and more hydrophobic than in the dimer of the isolated domain in solution.

The biological relevance of Zn^{2+} and the *N*-terminal HHCC domain in HIV IN has been well documented. Indeed, utilizing a combination of techniques including UV visible absorption, circular dichroism, and fluorescence spectroscopies, it has been demonstrated that metal ions (Zn^{2+}, Co^{2+}, or Cd^{2+}) are bound with equimolar stoichiometry by the isolated *N*-terminal domain [44]. A mutation in the HHCC motif abolishes zinc-binding capacity of HIV-1 IN. The isolated *N*-terminal domain is disordered in the absence of zinc [45]. The *N*-terminal domain is necessary for 3'-processing and strand transfer activity in HIV-1: deletion of this domain or mutation of any of the four conserved HHCC residues abolishes 3'-processing and strand transfer activity of HIV-1 IN [10].

However, in a study of RSV IN, 3'-processing and strand transfer products are detectable in reactions after deletion of the HHCC domain, but is much less efficient than with the wild type in the presence of Mn^{2+} but not Mg^{2+}. Most strikingly, when a mutant of RSV IN lacking the *N*-terminal HHCC domain is fused to various short peptides, efficient strand transfer activity can be restored to the level of wild type RSV IN [46]. Similarly, deletion of this domain in Visna virus [47] IN has no effect on 3'-processing. Substitution of His12 and His16 in RSV IN does not significantly impair 3'-processing or strand transfer [48]. When the *N*-terminal domain of IN expressed independently, it does not bind DNA [9, 48], but it has been suggested that it interacts with DNA in the context of the intact protein [48-51]. Furthermore, several studies showed that Zn^{2+} promotes multimerization, which should thus occur through the HHCC motif and enhances the catalytic activity of HIV-1 IN [45, 52], however, mutants lacking this domain can still form tetramers [53, 54]. Based on the above observations; the exact role of the *N*-terminal domain in these reactions is not clear. However, the high conservation of this motif and results from genetic experiments suggest that this domain is functionally important. Further biochemical studies are needed to conclude on its *in vivo* importance.

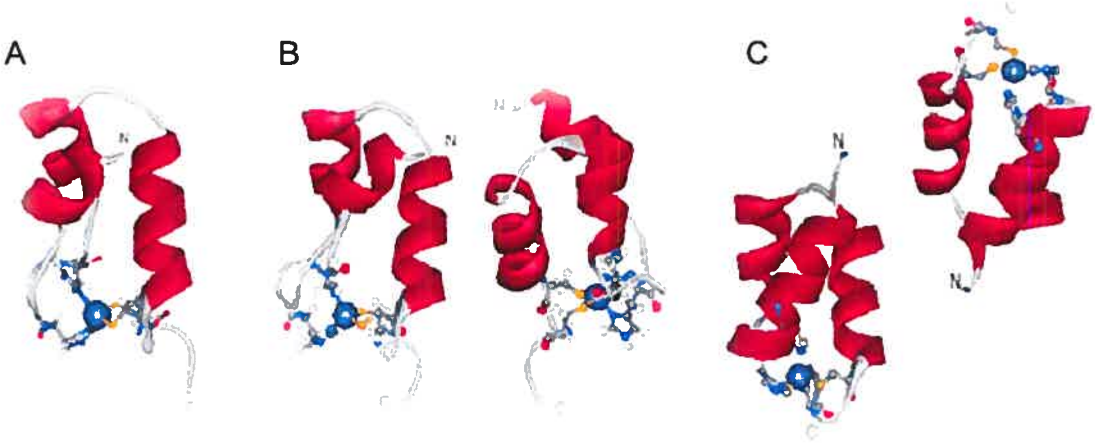

FIGURE 1.5. Ribbon drawings of *N*-terminal domain structures. Side chains of conserved HHCC residues of the Zn^{2+}-chelating HHCC motif are shown in a ball and stick representation. The Zn^{2+} cation is shown as CKP. The *N*-terminus and *C*-terminus are indicated. (A) and (B) Coordinates from PDB file 1WJA determined by NMR; (A) shows the monomer and (B) the dimer. (C) The dimer interface is shown for the *N*-terminal domain from PBD file 1K6Y determined by X-ray diffraction. Note that the dimerization interface in this crystal structure is completely different from the solution structure in (B).

1.1.3 Structure of the *C*-Terminal Domain

The amino acid sequences of the *C*-terminal domain, which are approximately 80–100 amino acids long, are not conserved among INs from different retroviruses. Mutation and deletion analyses with both avian sarcoma virus (ASV) and HIV-1 INs indicate that the *C*-terminal domain contains nonspecific DNA binding activity [9, 48, 55-57]. With the exception of Feline Immunodeficiency Virus (FIV) IN [23] in which deletion of the *C*-terminus (residues 236-281) resulted in a mutant that retained efficient 3'-end processing and disintegration activities but weak 3'-end joining activity, all INs with a deleted *C*-terminal domain lose 3' processing and strand transfer activities [11, 34, 36, 39, 46, 58-60]. This inability presumably results from loss of capacity to correctly position and orient the viral LTR ends at the active site. However, when the viral DNA ends have been correctly prepositioned, such as with a synthetic disintegration substrate during *in vitro* assays, catalysis can occur with mutant INs that have been truncated at the *N*-terminus or the *C*-terminus or both [36, 54]

Two studies using NMR [24, 61] on solutions of HIV-1 IN *C*-terminal domain showed homo-dimeric structures with each subunit composed of a five-stranded β -barrel that is topologically very similar to structures of SH3 domains, which occur in many signal transduction proteins [11, 46, 60, 62, 63]. These two solution structures agree well with each other (Fig. 1.6). The structure of the *C*-terminal domain monomer resolved by X-ray diffraction is also similar to the structures resolved by NMR. The two *C*-terminal subunits in the dimeric structure are related to each other by a 90° rotation relative to their two-fold axis.

Structures also exist for the *N*-terminal domain plus the core domain [13] and for the core domain plus the *C*-terminal domain [12, 29, 30]. The core domains in those two domain structures are almost identical to the structure of the isolated core domain and the two-fold symmetrical dimer interface is also similar. Superimposition of the catalytic core of these two-domain structures results in a proposed structure of the full-length IN dimer, which might be biologically relevant [13] and will be further discussed in this thesis. Residues 271–288 in the *C*-terminal domain are either not clearly resolved or deleted. The hinge region connecting the core and the *C*-terminal domain is different among HIV-1, Simian immunodeficiency virus (SIV) and Rous sarcoma virus (RSV) IN, which probably reflects flexibility in this region. Moloney murine leukemia virus (M-MULV) IN contains a sequence insertion of unknown function in this region [64].

The interactions at the dimer interface are predominantly hydrophobic and localize to *β*-strands 2, 3, and 4 , with the two triple-stranded antiparallel *β*-sheets, one from each subunit, oriented antiparallel to each other. One surface of the dimer is a saddle-shaped groove with dimensions of approximately 24 x 23 x 12 Å in cross section, which could accommodate a duplex DNA molecule [24]. This groove contains basic residues favorably positioned to contact DNA. Lys264, which has been shown from mutational data to be involved in DNA binding, protrudes from this surface. The diversity and hydrophobic character of the protein–protein interactions forming dimer of the *C*-terminal domains from HIV-1, RSV [30], and SIV [29] suggest that they are weak and nonspecific [12].

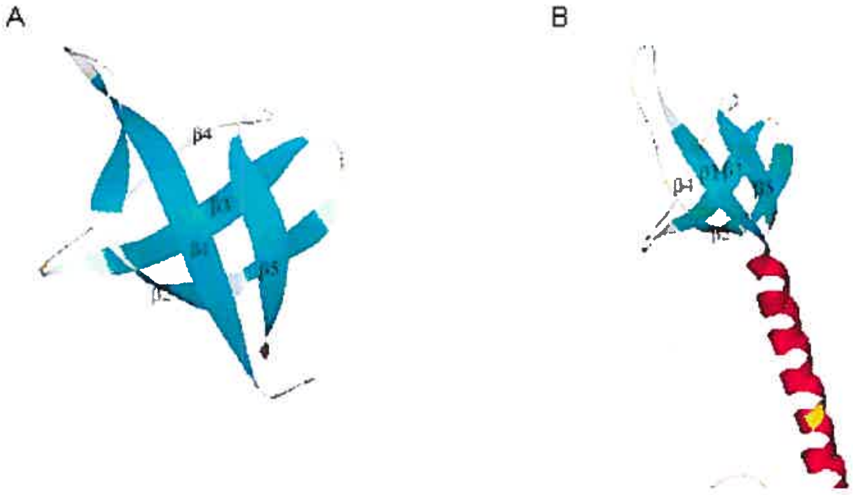

FIGURE 1.6. Ribbon drawings of *C*-terminal domain structures shown in monomeric form. Coordinates from PDB file 1IHV determined by NMR was used to generate (A). Coordinates from PDB file 1EX4 determined by X-ray diffraction was used to generate (B).

1.1.4 Integrase-DNA interaction

There are currently no IN structures available with substrate DNA docked in the active site. However, available IN structures and biochemical data provide considerable information about IN-DNA interaction. This section provides a review of the information available to date, which will be used as a basis for interpretation of the results presented in this thesis.

In the preintegration complex, IN must make specific interactions with viral DNA sequences, and after transporting the preintegration complex into the nucleus, IN must also interact with the target DNA for the integration reaction to occur. It is clear that IN can distinguish between viral DNA ends and other oligonucleotides [65-68] since IN protein requires both the subterminal and the distal position of its viral DNA recognition sequence for efficient cleavage to occur; however ultraviolet (UV) cross-linking studies [55, 69], filter binding assays [37, 70], Southwestern blots [56, 59], and electrophoretic mobility shift assays [51] show that IN binds to substrate DNAs with affinities similar to those of nonsubstrate DNAs *in vitro*. In other words, although the 3' processing and strand transfer activities of retroviral IN are sequence-dependent on both the distal and proximal sequences [67, 71-75], the binding of IN to DNA seems nonspecific. [55, 76, 77]. Experiments suggest that specificity in catalysis is achieved by nucleotides both distal and proximal to the conserved CA [74, 75] and metal cofactor [77-80]. For example, experiments showed that a stable complex of IN and viral DNA is formed in the presence of Mn^{2+} and the IN-viral DNA complex is resistant to challenge by an excess of competitor DNA [78]. A comparative study shows that each IN from M-MULV, human T-cell leukemia virus (HTLV)-1, HTLV-2 and HIV-1 required specific terminal LTR sequences for optimal catalysis of 3'-processing reactions, while strand transfer and disintegration reactions do not. Furthermore, in the 3'-processing reaction, sequence specificity for each IN was traced to the three nucleotides proximal to the conserved CA [81] in the presence of metal Mn^{2+}. Another study, by *in vitro* selection and specific photocrosslinking in the presence of Mg^{2+}, identified that distal positions in the LTR termini interact with the *C*-terminal domain of IN, providing evidence for the role of that domain in stabilization of viral DNA binding, while the terminal LTR interaction is mapped to the disordered loop of the IN core domain, specifically residues Q148 and Y143 [82]. Integrases need to fray viral LTR double-stranded DNA ends for 3'-processing to proceed since adding nucleotides to the 3'-end of the LTR sequence severely reduces 3' processing while increased cleavage by IN was detected when the nucleotides 3' to the CA-3' dinucleotide were present as single-stranded DNA [66]. Further evidence comes from a nucleotide analog substitution study, in which substitutions increasing the hydrogen bonding between the 'plus' and the 'minus' strand decreases 3' processing activity, while those which reduce or disrupt base pairing in the conserved CA dinucleotide increase activity [76, 79, 83-87]. This requirement for base-pair disruption may account for the inability of IN to use internal sites on DNA molecules as viral att (attachment) sites. Binding of IN to U5 LTR DNA is tighter, exhibiting a prolonged half-life in the presence of Mn^{2+} cations compared to Mg^{2+}. The preference observed for Mn^{2+} in standard *in vitro* integration assays can be attributed entirely to the augmentation in the DNA binding affinity of the IN [77].

The core domain contains the active site and is the only part of the IN protein capable of independent nucleotidyl transfer. This domain interacts with viral LTR ends [43, 67, 71, 72, 76, 79, 82, 88-90]. For example, cross-linking data have demonstrated that conserved Lys156 and Lys159 residues are involved in binding of the adenosine of the conserved CA [71]. Furthermore, three active site residues of the DD35E motif in the core must contact both the viral LTR end and the target DNA for the integration reaction to occur. Cross-linking experiments also suggest that the adjacent conserved cytosine and the 5' dinucleotide on the noncleaved strand also make contact with regions of the core domain, in and around the flexible loop [82, 88]. These data support a clear role in viral DNA end binding by the core domain. However, IN is known to function as a multimer (see section 1.1.5), and it remains to be determined which specific DNA contacts are in cis or trans with respect to the active site.

The *N*-terminal domain is in close proximity to target DNA 5' to the site of integration as shown by crosslinking data [88]. By constructing chimeras between HIV-1 and Visna virus INs, it is suggested that the *N*-terminus of IN does not contribute to viral DNA specificity and is not involved in determining substrate specificity for 3'-processing and strand transfer activities. Thus, this function must reside in the central region or *C*-terminus of IN [47]. Furthermore, the first 26 residues of RSV and HIV-1 IN, which include the first two histidines in the HHCC motif, are not required for DNA binding [9, 76].

Experiments showed that the *C*-terminal domain of IN interacts with bases distal to the terminal bases of the LTRs [71, 82, 88]; this interaction may play an important role in stabilization of viral DNA binding [82]. This may help to explain several results in which mutations in the *C*-terminal domain affect the 3'-processing activity [57, 91], which does not require the binding of target DNA. Structurally, it is suggested that a strip of positively charged amino acids from both monomers extending from each active site of the dimer to the *C*-terminal domain of the other monomer may act as dimeric platform for binding each viral DNA end. This strip potentially may stabilize the viral att site DNA for 3'-processing and strand transfer. This putative DNA binding site involves residues from both monomers: the core from monomer A with the *C*-terminal domain from monomer B in the dimer, implying that a viral end cleaved in the active site of one monomer is stabilized by residues from the *C*-terminus of the other monomer. This is consistent with *in vitro* complementation experiments [11, 39]. Previously, the *C*-terminal domain has been presumed to be involved in target DNA binding and this IN-target DNA interaction has been presumed to be nonspecific [57, 91], which is suggested by target sites of integration known to exhibit very little sequence specificity [92, 93]. Mutation of conserved lysine-264 has a dramatic effect on the nonspecific DNA binding activity of the isolated *C*-terminal fragment, as judged by ultraviolet (LTV) cross-linking, and by 3'-processing by full-length IN containing a K264E mutation [24, 57]. Structural modeling based on NMR *C*-terminal structure, discussed in section 1.1.3, illustrates that the dimensions of the saddle-shaped groove consisting of amino acids 220-270 in the dimer are appropriate for DNA binding. In this model, amino acids 220-270 in the dimer can readily bind to the major groove of DNA, where the side-chains of Ser230, Pro261, Lys258, and Lys264 interact with the sugar-phosphate backbone, and the side chain of Arg231 interacts with the bases [91]. Some work has suggested a role for Arg-262 and Leu-234 in DNA binding [91]. Mutagenic analysis shows that Ser-230 and Arg-263 are involved in enzymatic activity and DNA binding [91]. Although the *C*-terminal domain has been implicated in binding of target DNA, certain work from chimeric IN proteins shows that the core domain plays a more important role in target binding than the *N*- or *C*-terminal domains [47, 58]. Activity assays show that these chimeras exhibit the target site preference of the core domain, but not flanking domains. Some cross-linking experiments have crosslinked target DNA to portions of both the core and *C*-terminal domains, as well as a region of the *N*-terminal domain [94]. The exact function of the *C*-terminal domain is still not quite clear, but available experimental data suggest the idea of a complex network of DNA binding rather than a model in which individual domains are unilaterally responsible for binding to viral or target DNA [42]. Transposases and retroviral INs share a structurally related catalytic domain. They are members of the larger superfamily of polynucleotidyl transferases. Homology modeling based Tn5-DNA complex structure is likely to give us insights about IN-DNA interactions.

1.1.5 Multimeric organization of HIV-1 IN

Many evidences show that IN functions as a multimer. However, there is conflicting evidence with respect to the number of units composing the active multimer and with respect to the actual dimerization interfaces. Clearly, identification of the multimerization interfaces must take into account the surfaces involved in DNA binding, and vice-versa. In the absence of structural data of multimers with bound DNA, our study set out to explore these potential binding surfaces using biochemical methods. The following section describes the information available to date, which will be used as a complement for interpretation of the results presented in this thesis.

Biochemical studies have revealed that multimerization determinants reside in the core domain, as well as in the *N*-terminal and *C*-terminal regions of HIV-1 IN [95]. Deletion mutants of HIV-1 IN that lack either the *N*-terminal or *C*-terminal domain have no 3'-processing or strand transfer activity [34]. However, if an *N*-terminally truncated IN is mixed with a *C*-terminally truncated IN, 3'-processing and strand transfer activities can be restored [34] [11]. Further evidences for multimerization come from mutagenesis and deletion studies which show that full-length IN can multimerize to form both dimers and tetramers in solution [45, 52, 96, 97]. Furthermore, the *N*-terminal domain of IN can function in trans but not cis to the core domain [32, 98], while the *C*-terminal domain can function in cis [34] or trans [11, 34, 43, 89].The above results also suggest that the core domain contributes the active site enzymatic activity in partnership with an *N*-terminal domain from a different monomer of IN.

The difficulty of determining the multimeric organization of HIV-1 IN comes from the fact that purified recombinant INs can exist in a dynamic equilibrium including monomers, dimers, tetramers, and even higher order oligomers [50, 96, 99-101], and any species can be an active form of IN except the monomer [96]. The stoichiometry of retroviral IN complex is not known *in vitro*. Several structural studies show that a tetramer (dimer-of-dimers) or an octamer of IN would be necessary to carry out concerted integration of both LTRs on target DNA [13, 30, 43]. For example, the overall structure of an IN tetramer formed by crystal lattice contacts from the *N*-terminal and core two-domain structure [13]is structurally similar to a related bacterial transposase Tn5 dimer complex with its DNA substrate, which can be considered as evidence supporting the SIV tetrameric model. Furthermore, this tetramer model exhibits positively charged channels suitable for DNA binding [13]. A recent study that used IN complexes present in nuclear extracts from human cells suggested that the minimal cellular IN complex is a homotetramer, implying that at least an octamer of IN is required to carry out concerted integration of both LTR ends into target DNA [102]. Within this tetramer, it will be only one of the two active sites in each dimer that would be actually involved in the chemical reactions.

1.2 Purpose and Scope of the Study

1.2.1 Research objectives

The integration process is an obligate part of the retroviral life cycle. Retroviral IN is both essential and sufficient to catalyze this integration reaction. The overall aim of this study is to gain a finer understanding of the biochemistry of the integration reaction. The specific research objectives of this dissertation include:

- The expression, purification and characterization of the previously obtained insertional mutants of HIV-1 IN.

- To gain insights into the role of previously predicted unstructured loops.
- To map functionally tolerant region of insertions of HIV-1 IN.
- To gain insights into IN-DNA interaction.
- To compare available tetramer models with our experimental data to see how they agree with our new data and with existing biochemical data.
- To define domain boundaries, which might be useful to express minimized IN constructs for crystallization studies.

1.2.2 **Significance of the study**

The development of effective inhibitors of HIV replication targeted to HIV reverse transcriptase and HIV protease has demonstrated the potential effectiveness of antiviral therapy for the treatment of AIDS, which benefits from the foundation of basic knowledge in understanding the mechanism of retroviral reverse transcription and the structure of the protease. Drugs targeted to HIV IN would be a valuable complement to reverse transcriptase and protease inhibitors. However, the lack of detailed structural information about IN/substrate interactions has so far hindered the search for strong and selective IN inhibitors. Although the structures of all three domains of IN have been individually determined, their spatial arrangement in the active complex with DNA substrate is unknown. The studies I present herein using a linker-scanning approach will provide a better insight into the functioning of this enzyme. Results from this study will therefore provide valuable information for those concerned with the design of effective inhibitors of the retroviral IN.

Chapter II

Comprehensive Linker-scanning Analysis of the HIV-1 Integrase Protein

2.1 Context of the work relative to results obtained previously

The work described in Chapter 2 was begun prior to registering as a M.Sc. student in the Bio-informatics program at Université de Montréal. Specifically, Sections 2.2.2: Mutagenesis: *In vitro* transposon-based linker insertion into pINSD·His; 2.2.3: Selection of clones with insertions within the HIV-1IN coding region; and 2.2.4: Generating the IN frame 19-codon insertion, were performed by myself while I was a research associate in the laboratory of Prof. C. Jonsson, currently at the Southern Research Institute, Department of Biochemistry and Molecular Biology, Birmingham AL. As this work is not yet published, it is described herein in full, with the approval of Prof. Jonsson. Sections 2.2.5: Expression and purification of mutant and wild-type IN; 2.2.6: Substrate preparation; and 2.2.7: *In vitro* integration and disintegration assays, were initiated prior to registering as a M.Sc. student and were completed during the course of this degree. Section 2.2.8: Structural model of HIV-1 IN monomer were initiated and completed during the course of this degree.

The work presented in Chapter 2 has been submitted as part of a collaborative manuscript entitled: Comprehensive Linker-Scanning Analysis of the MuLV RNase H, MuLV and HIV-1 Integrase Proteins, Author : Jennifer Puglia[1], Tan Wang[2], Christine

Smith-Snyder[1], Marie Cote[1], Michael Scher[1], Joelle Pelletier[4], Sinu John[3], Colleen B. Jonsson[2] and Monica J. Roth[1]

[1]Department of Biochemistry, Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey 675 Hoes Lane Piscataway, NJ 08854.

[2] Department of Biochemistry and Molecular Biology, Southern Research Institute, 2000 9th Ave S, Birmingham, AL 35205.

[3]Graduate Program in Biochemistry and Molecular Genetics, University of Alabama at Birmingham, Birmingham, AL 35294.

[4]Département de Chimie, Faculté des Arts et Sciences, et Département de Biochimie, Faculté de Médecine, Université de Montréal, C.P. 6128, Succursale Centre-Ville, Montréal, Québec H3C 3J7, Canada

2.2 Introduction

Various methods have been developed for the comprehensive analysis of a gene by construction of a saturating or near-saturating library of mutants [103-105]. These studies have defined domain boundaries, provided functional maps, and insights into previously predicted unstructured loops [106] [103-105, 107-109]. In this study, the method of insertional functional mapping is applied to the HIV-1 IN (IN) protein.

The integration of retroviral particles is a complex process. Preintegrative complexes (PICs) has been purified and characterized from infected cells [110-122]. Despite extensive study, the assembly of this complex is not well understood. These efforts have been assisted by structural studies of related retroviral IN subdomains [12, 13, 16, 19, 26, 29, 30, 123, 124]. However, to date, neither a structure of a complete retroviral IN protein nor one of a subdomain in complex with DNA has been obtained.

The linker-insertion genetic footprint provides a means to identify non-essential regions within proteins capable of withstanding insertions. This method relies on retroviral-mediated insertion of a specific DNA sequence at various, random locations within target DNA. The insertion of a large DNA fragment is followed by excision of part of - but not all – the inserted sequence, leaving a shorter (57 base pairs), specific DNA sequence within the target gene. Upon expression, the inserted fragment is translated within the target protein, resulting in disruption of the native sequence. Disruption of function indicates that the area of insertion does not tolerate structural disruption, for any of a variety of reasons: the disrupted area may be directly required for function, may be required for oligomerization or may be essential to the folding of the target protein.

In this study, the HIV-1 IN gene was subjected to Tn5-based random insertion mutagenesis. Individual constructs, selected from the library, were assayed for the effects on *in vitro* IN enzymatic activity assays. The observed activities of the resulting IN mutants provide insights into the possible roles of the various parts of the HIV-1 IN protein. Using this approach, three regions that are functionally tolerant of insertions were identified within IN. These regions correlate with domain and protein junctions.

2.3 Experimental Procedures

2.3.1 Materials

Oligonuleotide PCR primers and oligonuleotide substrates were synthesized by Integrated DNA Technologies (Coralville, Iowa). Restriction enzymes, Taq polymerase, T4 DNA ligase and kinase were purchased from New England Biolabs (Beverly, Mass.). The EZ: TN In Frame Linker Insertion Kit was purchased from EPICENTRE (Madison, WI). All chemicals were purchased from Sigma.

Bacterial Strain and Plasmid: Plasmid pINSD·His containing HIV-1 IN was obtained from the AIDS Research and Reference Reagent Program. It was used to construct HIV-1IN mutants and was propagated in *Escherichia coli* strain DH5α. *E coli* strain BL21 (DE3) was used to express wild type HIV-1 IN and its mutants.

2.3.2 Mutagenesis: In vitro transposon-based linker insertion into pINSD·His

The EZ: TN In Frame Linker Insertion Kit was used to insert a nucleotide linker at random into the target plasmid pINSD·His, as described in FIGURE 2.1. The reaction mixture contained 1 μl EZ::TN 10x reaction buffer, 0.4 μg pINSD·His plasmid DNA, 1μl (0.1pmol/μl) EZ::TN<Not/KAN-3>Transposon, 1μl EZ::TN Transposase in a final volume of 10μl. The reaction mixture was incubated for 2 hours at 37 °C and the reaction was stopped by adding 1 μl EZ::TN 10x Stop Solution. 1 μl – 10 μl of the insertion reaction mixture was used to transform MAX Efficiency *E. coli* DH5α Competent Cells from Invitrogen (Carlsbad, CA) and plated and selected on kanamycin-containing plates by overnight growth at 37°C. Each individual colony was picked and transferred into 96 well culture blocks from Qiagen (Valencia, CA) with 1ml LB containing 50μg/ml kanamycin in each well. After overnight growth at 37 °C with shaking, 50% glycerol was added to each well to bring the final concentration of glycerol to 10% and the blocks were stored at –80 °C.

2.3.3 Selection of clones with insertions within the HIV-1IN coding region

The <Not I/KAN-3>Transposon should be randomly inserted into the target DNA plasmid pINSD·His. A PCR strategy was employed to determine whether the transposon insertion sites were within the HIV-1 IN coding region and their distribution.

The pinsdBsreen primer (5'-CGG GCT TTG TTA GCA GCC GG -3') and pinsdFscreen primer (5'-GGT GCC GCG CGG CAG CC –3') were used to amplify the IN and encoding sequence from 301 to 320 and 335 to 351 of the pET15b plasmid, which is parent plasmid of pINSD. The PCR reactions contain 1xPCR buffer of Expand^{TM} long Template PCR system from Boehringer Mannheim (Gmbh, Germany), 2.25mM $MgSO_4$, 0.2mM dNTP, 2 μM primers, 2 units Taq polymerase and traces of the glycerol stocks stored at-80 °C .The reactions were carried out in 96 well PCR plates from Robbins Scientific Corporation (Suuyval, CA). PCR conditions were as follow: 94 °C for 4 minutes, followed by 35 cycles of: denaturing at 94 °C for 30 seconds, annealing at 69 °C for 30 seconds followed by an extension step at 72 °C for 2minutes and 15 seconds. This cycle was followed by a final extension period at 72 °C for 4 minutes, which was followed by a hold at 4 °C. After PCR, the samples were loaded onto a 1.5 % agarose gel and examined to determine which of the clones are positive for linker insertion within the IN gene.

2.3.4 Generating the IN frame 19-codon insertion

The clones with a transposon insertion site within the IN coding region were individually digested with NotI according to manufacturer's recommendations. Purification of the linearized DNA from the 1100 bp kanamycin–containing fragment was carried out by agarose gel electrophoresis. Religation of the linearized DNA by T4 DNA Ligase regenerated a single Not I restriction site and created the 57 nucleotide (19 codon) insertion into any of the three reading frames (Fig. 2.1). The religated DNA was transformed into *E. coli* DH5α cells and selection was done on Kanamycin-containing plates. DNA sequencing of transposon insertion clones was performed with the ABI

PRISM® BigDye™ Primer v3.0 Cycle Sequencing Ready Reaction Kit with AmpliTaq® DNA Polymerase, FS. (Foster City, CA) to determine the site of the 19-codon insertion. The integrated sequence analysis software VectorNTI from InforMax Inc. (Frederick, MD) was used to analyze the sequence data.

2.3.5 Expression and purification of mutant and wild-type IN

Wild-type HIV-1 IN and insertion mutants were expressed in *E. coli* BL21 (DE3) cells in 50 ml of medium and purified as hexahistidine-tagged fusion proteins under denaturing conditions as described previously [83]. 50 ml cultures purification yielded approximately 2 mg of 90–95% homogenous protein (Fig. 2.3). The protein fraction refolded at a concentration of 5 mg/ml exhibited greatest enzymatic activity. HIV-1 IN precipitated upon addition of Buffer C (0.2 M NaCl). The precipitated protein was re-suspended in Buffer D (0.5 M NaCl) to a final concentration of 1 mg/ml.

2.3.6 Substrate preparation for in vitro activity assays

Oligonucleotides were purified on 20% denaturing polyacrylamide gels, ^{32}P-labeled at the 5' end with T4 polynucleotide kinase, and hybridized to complementary strands as previously described (Jonsson et al., 1993). Unincorporated radioactivity was removed from labeled integration and disintegration substrates with G-25 or G-50 Quick Spin columns (Boehringer, Mannheim, IN).

2.3.7 *In vitro* integration and disintegration assays

Strand transfer and disintegration reactions were performed as described previously [83]. Reaction products were separated on 20% polyacrylamide denaturing gel and subjected to autoradiography or PhosphorImager screens (Molecular Dynamics). Products were quantified with ImageQuant software (Molecular Dynamics).

2.3.8 Structural model of HIV-1 IN monomer. The structural model of the HIV-1 IN monomer was constructed from a combination of two X-ray crystal structures, represented by PDB files 1k6y and 1ex4. The "A" molecule of 1k6y was superimposed onto the "A" molecule of 1e4 using the program O. The 1ky structure is comprised of residues 1-46, 56-139, and 149-210; and 1ex4 is comprised of residues 56-141 and 145-270. Thus, the superpositioning consisted of overlaying the Cα atoms of all common core residues (RMSD-0.83 angstroms). Where the model contained disordered regions (residues 47-55 and 142-144, inclusively), polyalanine segments containing the correct number of amino acids were created and moved into the appropriate linking positions in the model. The Ala residues were then changed to the proper residues, and the regions were subjected to least squares minimization.

2.4 RESULTS

Fig. 2.1 outlines the series of steps undertaken to generate the linker-insertion library within the HIV-1 IN construct pINSD·His. Briefly, the Tn5 mutagenesis system results in the random insertion of the transposon encoding the kanamycin resistance gene throughout the plasmid. To make the screening process high throughput, each individual colony was picked and transferred into 96 wells culture blocks. In total, 1056 colonies were picked into 96 well culture blocks to screen mutants with a transposon inserted into the IN coding region by PCR. Any insertion within the IN will be amplified to a 2.1 kb product while an insertion within the vector sequence will amplified to a 0.9kb product. Each clone with a transposon inserted into the IN coding region was digested with NotI and religated to create the 57 nucleotide (19 codon) insertion. The amino acid sequence of the HIV-1 IN encoded by the target DNA pINSD is conserved on both sides of the 19-codon insertion. These constructs with 19 amino acid insertions were transformed into *E. coli* and selected using the ampicillin-resistance gene (*β*-lactamase) that is present on the original cloning vector. Selected insertion mutant proteins were expressed and purified. Strands transfer activity and disintegration activity were performed on these purified insertion mutants to evaluate the effect of insertions in different sites. The results obtained in each step of the procedure are described below.

2.4.1 Selection of clones with insertions within the HIV-1IN coding region by PCR

To determine the transposon insertion sites, PCR was used to screen the clones that have a transposon inserted into the HIV-1 IN coding region. As shown in Fig. 2.2, insertion within IN coding region resulted in a 2100 bp PCR product; otherwise a 910 bp PCR product was obtained. Of 1056 colonies (11 x 96 well plates) screened, 111 of them were positive for having one insertion within the IN gene. The 57-base pair insertion generated by the Tn5 In-Frame Linker Insertion scanning system resulted in insertions into all three reading frames of the original clone. Nine of the 57 nucleotides are the result of a 9 bp sequence duplication immediately flanking the transposon insertion site. The amino acid sequence of the protein encoded by the target DNA is conserved on both sides of the 19 codon insertion. Sequencing from the library (Table 1) showed that insertions were distributed throughout the whole IN encoding region. The insertion sites, however, were not randomly distributed, with clustering of insertions within the *C*-terminal domain. This could indicate a preference for specific structure within the plasmid DNA by the transposase enzyme since a large number of duplicate isolates were identified within the population examined. The composition of 19-amino acids inserted are determined by the target site selected and two 9 bp target site sequence duplication during the transposition as well as the sequences encoding the NotI restriction site and two 19-bp inverted repeat mosaic end sequences, which are recognized by Tn5 transposase. Depending on the reading-frame, the insertions will encode XCLLYTSCGRKMCTRD(S/R)XX, LSLVHILRPQDVYKRQXXX or XVSCTHLAAARCVQETXXX.

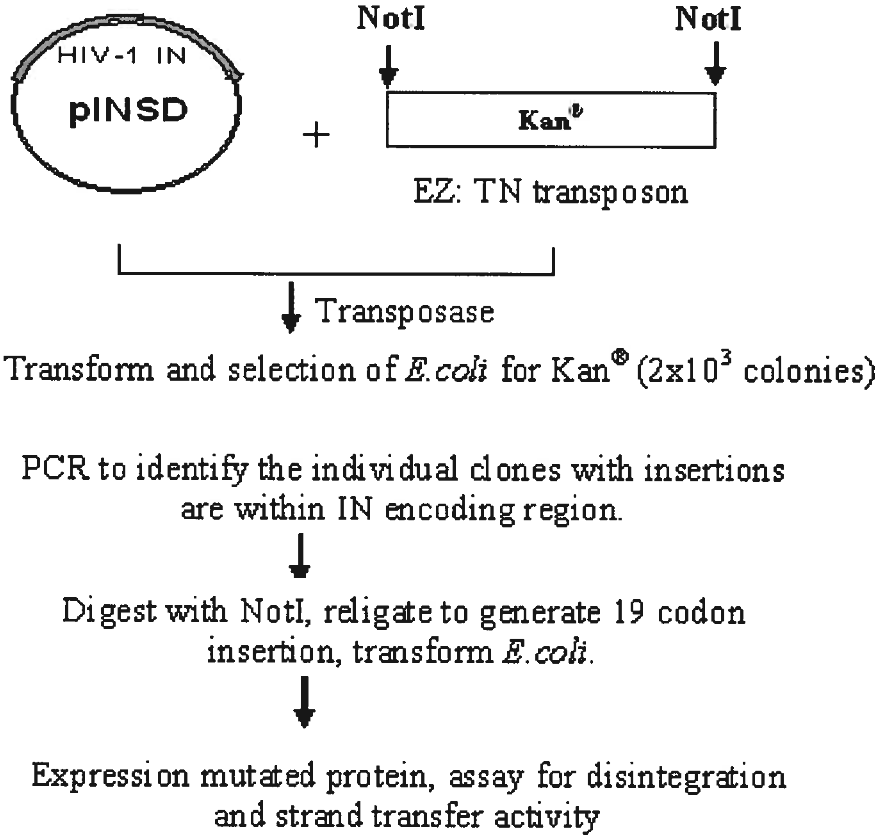
HIV-1 IN
pINSD
+
NotI
NotI
Kan^r
EZ: TN transposon
Transposase
Transform and selection of E.coli for Kan® (2x10³ colonies)
PCR to identify the individual clones with insertions are within IN encoding region.
Digest with NotI, religate to generate 19 codon insertion, transform E.coli.
Expression mutated protein, assay for disintegration and strand transfer activity

FIGURE 2.1. Generation of the HIV-1 insertional library. The three steps required to generate the insertional library within the construct are outlined. The region of IN gene encoding the Integrase (IN) protein subjected to TN5 insertional mutagenesis and transposon Tn5 DNA, which contained the kanamycin resistance gene between its short 19 basepair Mosaic End (ME) Tn5 transposase recognition sequences, are shown. Restriction sites Not I flanked by the ME also are shown.

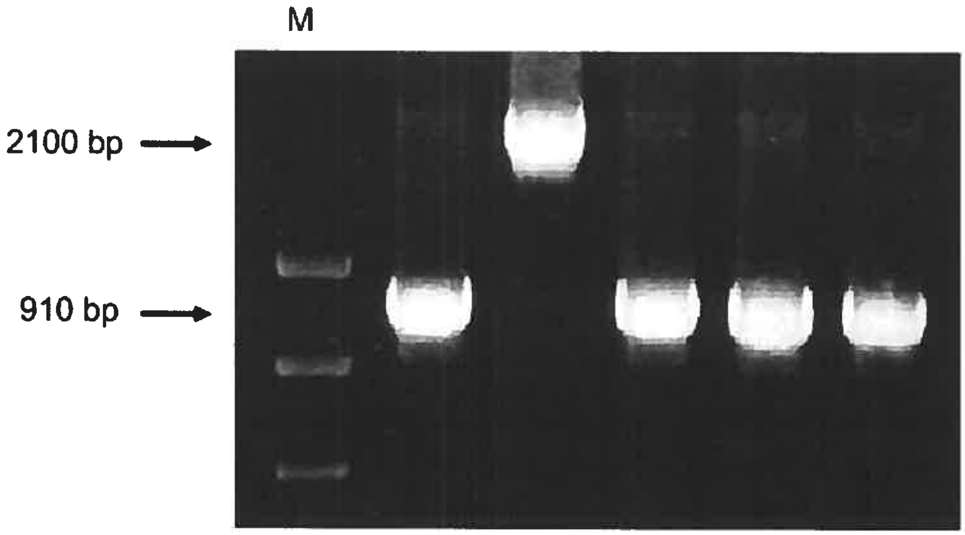

FIGURE 2.2. PCR products in the clones with insertion and without insertion within the IN coding region. A 2100-bp band indicates the presence of one single insertion within the IN coding region. A 910-bp product indicates the presence of an insertion within the vector pINSD but outside of the IN coding region.

2.4.2 Insertion Sites of HIV-1 Individual mutants

The final insertion library for HIV-1 IN was characterized by analyzing individual isolates. Isolates of the final library were subjected to sequencing analysis. Of the 111 insertions, 2 were within the *N*-terminal domain, 35 were within the catalytic core and 74 within the *C*-terminal domain. After eliminating the duplicate clones, where insertions are in the same position with same insertion sequence, 55 clones have unique insertion positions and correct sequence (summarized in Fig. 2.6. and Table 1). Each mutant IN isolate was transformed into *E. coli* BL21 (DE3) and each mutant integrase was purified and further characterized by *in vitro* enzymatic analysis of individual clones.

2.4.3 Expression and purification of the wild type and mutant IN

Fifty-five insertional mutant proteins were expressed, purified and assessed for strand transfer and disintegration activity (Table 1 and Fig. 2.6). Mutants with a variety of activity levels were identified in each domain, the *N*-terminal, core and *C*-terminal.

E. coli BL21 (DE3) cells were used to express wild type IN or the insertion mutant INs by inducing with IPTG. Following the induction period, total protein was extracted under denaturing conditions. Solubilized protein can be isolated by nickel affinity chromatography. This one-step affinity purification yielded approximately 2-3 mg of 90–95% homogenous protein from 50 ml of cells, and is well suited for manipulating multiple purifications in parallel. As can be observed following migration of the samples on 4-12% SDS polyacrylamide gel electrophoresis (PAGE) (Fig. 2.3), HIV-1 IN insertion mutants with the extra 19 amino acids migrated a little slower than wild type IN. Column fractions containing the most enriched INs were pooled and diluted to ascertain the optimal concentration for refolding as measured by the activity of the HIV-1 IN enzyme. The protein fraction refolded at a concentration of 5 mg/ml exhibited greatest enzymatic activity. HIV-1 IN precipitated upon addition of Buffer C (0.2 M NaCl). The precipitated protein was resuspended in Buffer D (0.5 M NaCl) to a final concentration of 1 mg/ml.

Table 1: Summary of HIV-1 IN insertions

Insertion position[1]	Inserted amino acid sequence[2]	DS[3,5]	ST[4,5]	Insertion position[1]	Inserted amino acid sequence[2]	DS[3,5]	ST[4,5]
N27_L	LSLVHILRPQDVYKRQDFN	††	±	A205_T	VSCTHLAAARCVQETDIIA	†	–
N27_L	PVSCTHLAAARCVQETDFN	†††	±	D207_I	LSLVHILRPQDVYKRQATD	†	–
D55_C	LSLVHILRPQDVYKRQQVD	†††	††	E212_L	SVSCTHLAAARCVQETAKE	††	††
D55_C	CLLYTSCGRKMCTRDRQVD	†††	††	R228_D	AVSCTHLAAARCVQETDYR	†	–
P58_G	VSCTHLAAARCVQETDCSP	†††	†††	S230_R	SCLLYTSCGRKMCTRDRDS	†	–
D64_C	SVSCTHLAAARCVQETELD	±	–	A239_K	TVSCTHLAAARCVQETGPA	†††	–
I73_L	LSLVHILRPQDVYKRQKVI	–	–	K240_L	PVSCTHLAAARCVQETAAK	††	–
Y83_I	TVSCTHLAAARCVQETGGY	–	–	W243_K	NCLLYTSCGRKMCTRDSLW	††	–
E96_T	LSLVHILRPQDVYKRQGQE	–	–	G245_E	CLLYTSCGRKMCTRDSWG	††	–
T115_D	CLLYTSCGRKMCTRDRVHT	–	–	G247_A	LSLVHILRPQDVYKRQGEG	†††	†††
D116_N	TVSCTHLAAARCVQETDTD	–	–	A248_V	CLLYTSCGRKMCTRDSEGA	††	–
T125_V	AVSCTHLAAARCVQETGTT	–	–	V250_I	CLLYTSCGRKMCTRDRAVV	††	–
A128_A	CLLYTSCGRKMCTRDRVKA	–	–	V250_I	LSLVHILRPQDVYKRQAVV	††	–
W131_W	LSLVHILRPQDVYKRQACW	–	–	I251_Q	PVSCTHLAAARCVQETVVI	†††	–
I135_K	CLLYTSCGRKMCTRDRAGI	††	–	D253_N	LSLVHILRPQDVYKRQIQD	†††	–
N144_P	LSLVHILRPQDVYKRQPYN	†	–	S255_D	CLLYTSCGRKMCTRDRDNS	†††	–
S147_Q	CLLYTSCGRKMCTRDSPQS	±	–	V259_V	CLLYTSCGRKMCTRDSIKV	†††	–
G149_V	AVSCTHLAAARCVQETGQG	–	–	I268_R	SCLLYTSCGRKMCTRDRII	†††	††
V165_R	LSLVHILRPQDVYKRQGQV	–	–	R269_D	AVSCTHLAAARCVQETVIR	-††	††
D167_Q	CLLYTSCGRKMCTRDRVRD	†	–	G272_K	CLLYTSCGRKMCTRDRDYG	†††	††
A169_E	CLLYTSCGRKMCTRDRDQA	–	–	K273_Q	PVSCTHLAAARCVQETDGK	†††	††
E170_H	PVSCTHLAAARCVQETEAE	†	–	D279_C	LSLVHILRPQDVYKRQGDD	†††	††
K173_T	LSLVHILRPQDVYKRQHLK	±	–	V281_A	CLLYTSCGRKMCTRDSDCV	†††	††
T174_A	CLLYTSCGRKMCTRDSLKT	†	–	R284_D	PVSCTHLAAARCVQETASR	†††	†††
M178_A	LSLVHILRPQDVYKRQVQM	±	–	Q285_D	AVSCTHLAAARCVQETGRQ	†††	†††
A196_G	CLLYTSCGRKMCTRDRYSA	±	–	E287_D	AVSCTHLAAARCVQETEDE	†††	†††
I200_V	CLLYTSCGRKMCTRDRERI	†	–	D288_	CLLYTSCGRKMCTRDRDED	†††	†††
V201_D	AVSCTHLAAARCVQETGIV	†	–				

[1] Positions are based on the protein sequence of wild-type HIV-1 IN. An arrow marks the insertion between the two amino acids indicated. All isolates are independent insertions. [2]Sequence of the 19 amino acid insertion. [3]Disintegration assay [4]Strand-transfer assay [5]Activity is based on WT IN activity. Symbols: ¯ 0%; ± 0 to 5%; †† 6 to 35%; ††† 36 to 75%, ††† 76 to 100%

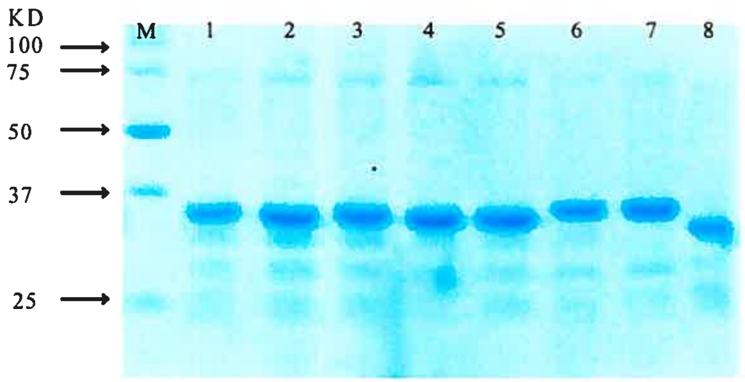

FIGURE 2.3. Purification of wild type HIV-1 IN and insertion mutant INs as observed by 4–12% SDS-PAGE. M indicates the molecular weight markers. Lanes 1-7, various insertion mutant INs. Lane 8, wild type IN.

2.4.4 *In vitro* Analysis of Individual HIV-1 IN Mutants

HIV-1 wild type IN and its mutants were assayed for strand transfer and disintegration activities with LTR-specific substrates. Strand transfer activities were detected by the use of a precleaved duplex DNA substrate. Disintegration activities were tested by a Y-shaped substrate that resembles an integration intermediate. Oligonucleotides used in the synthesis of the substrates in these experiments are listed in Fig.2.4B.

***N*-terminal domain mutants.** The HIV-1 *N*-terminal domain is made of a three-helix bundle structure. Two of the insertions at N27_L retained full disintegration activity however their integration activity was barely detectable. Notably, these two mutants were inserted into same position, but with different amino acids sequences. Insertions D55_C and P58_G fall into the hinge region between HHCC domain and core domain. In the two-domain crystal structure, which was crystallized in a tetrameric form, this connecting region (residues 47–55) is disordered in all four monomers [125]. These two insertions have retained high disintegration activity and have moderate to full integration activity. The D55/C56/S57 sequence is proposed to be involved in close proximity with the HIV LTR positions 1-4, based on a structural tetramer model [126].

Core Mutants. In the core domain, insertions D64_C, I73_L, Y83_L, E96_T, T125_V, A128_A and W131_W had completely diminished or very low disintegration and integration activity. These insertions are located within elements of secondary structure of the highly compact core consisting of a five-strand sheet together with six helices. These insertions most likely will disrupt the secondary structure. Insertions T115_D and D116_N also completely diminished both disintegration and integration activity. Two of the insertions are located in the loop between $\beta 4$ and $\alpha 2$, and hence, would not be expected to disrupt the packing of the core domain. Because D116 is part of the catalytic triad, these two insertions most likely disrupt the conformation of the catalytic triad. A group of insertions - I135_K, N144_P, S147_Q and G149_V – had different levels of disintegration activity, from trace or weak to moderate disintegration activity, and are all located near the loop between $\beta 5$ and $\alpha 4$. This agrees with a mutagenesis study on the Gly residues at 140 and 149 impaired catalysis of HIV-1 IN. Another group of insertions (V165_R, D167_Q, A169_E and E170_H, K173_T and T174A), which retained weak or trace or no disintegration activity, are located in the loop between α 4 and α 5. These two loop regions correspond to an extended loop (residues 137-156) and a flanking region (residues 161-173), which are protected from proteolysis upon metal binding [127]. Insertions A196_G, I200_V, V201_D, A205_T, D207_I and E212_ are distributed within $\alpha 6$. The activities of these insertion mutants range from low disintegration activity to low to moderate integration activity as one moves toward the *C*-terminal end of the helix. Of considerable interest, insertion E212_L, maintained nearly full disintegration and integration activity.

***C*-Terminal domain**. The *C*-terminal domain has been suggested to be involved in target DNA binding as mutations in this domain can abolish nonspecific DNA binding [128, 129]. Insertion R228_D is located at the end of $\beta 1$ strand in the *C*-terminal domain, and S230_R is located in the hairpin connecting $\beta 1$ and $\beta 2$. Neither of these two mutants had strand transfer activity and both had greatly decreased disintegration activity. Since the *C*-terminal IN can tolerate large deletions and can still have considerable level disintegration activity [33, 130], this implies that the effect of these two mutants may not simply be in disruption of folding of the *C*-terminal domain.

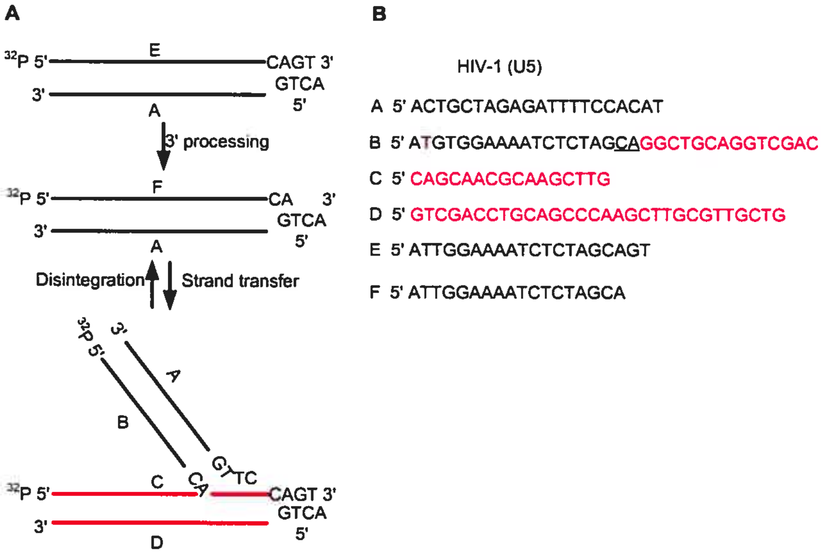

FIGURE 2.4.(A) Schematic representation of the enzymatic activities catalyzed by the retroviral IN *in vitro*: 3'-processing; strand transfer; disintegration. Strand transfer substrates are prepared by hybridizing the ^{32}P-labeled F strand with the A strand. The substrate is identical to the 3'-processing substrate, except for the absence of two terminal nucleotides proximal to the CA dinucleotide. Strand transfer activity generates products both larger and smaller than the substrate, since integration occurs at random sites along the phosphate backbone of the target substrate DNA. The disintegration substrate represents a hypothetical strand transfer intermediate. The substrate is prepared by hybridizing the ^{32}P-labeled C strand with the A, B, and strands. Disintegration reaction results in joining of the 3'-OH end of the C strand to the B strand, resulting in formation of the 30-nucleotide product. (B) Nucleotide sequences of HIV-1 LTR substrates used in the assays.

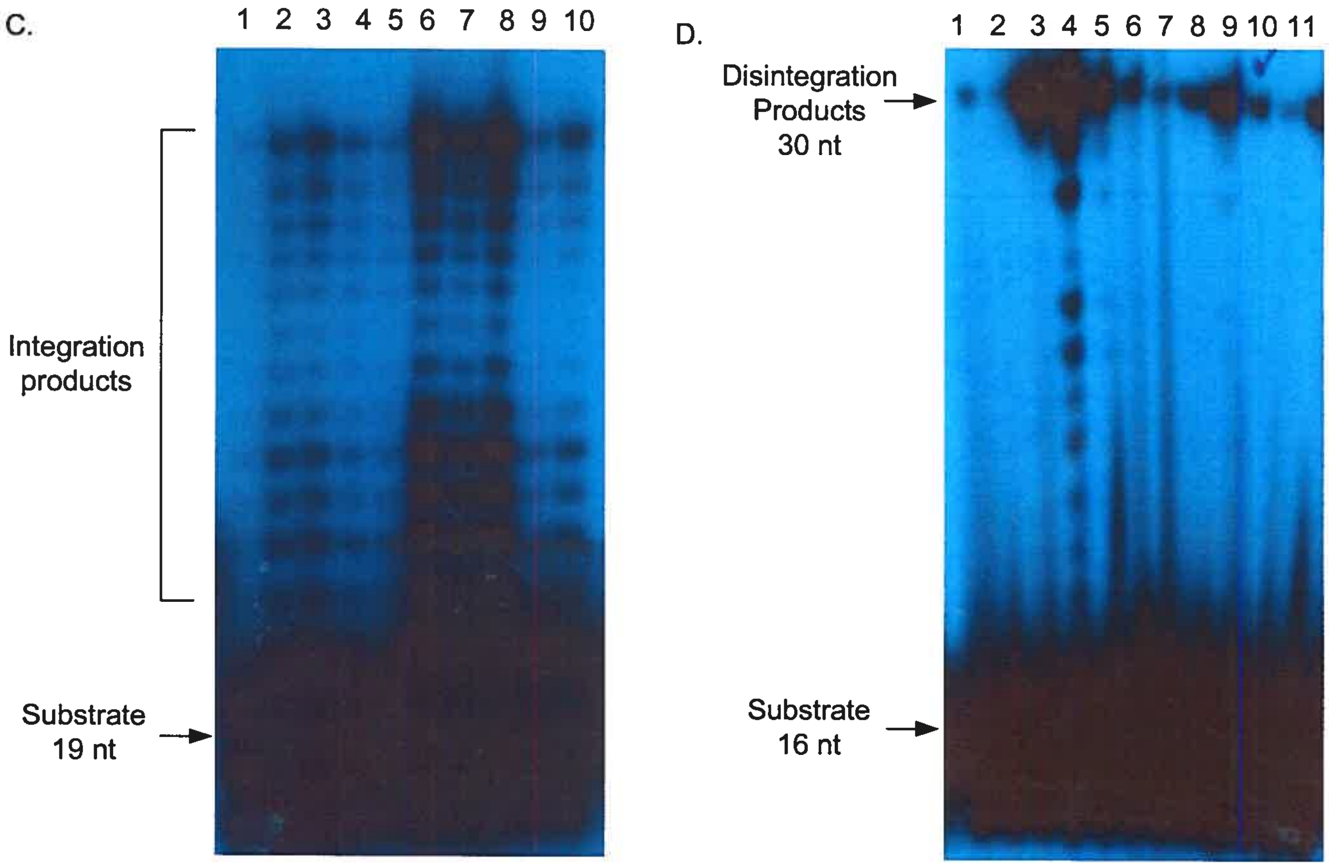

FIGURE 2.5 (A) Integration activities of the mutant INs. (B) Disintegration activities of the mutant INs. Enzymatic activities of the mutant INs were assayed as described under Experimental Procedures using HIV-1 U5 LTR-derived substrates (Fig. 2.4 B). Enzymatic activity is shown for integration (Panel A, lanes 1-10); and disintegration (Panel B, lanes 1–11) Assays were done in 15 μl reaction volumes with a final substrate concentration of 0.07 pmol/ml. Reaction with substrate in the absence of protein is represented in lane 1. In A and B, reactions were incubated at 37 °C for 60 min and terminated with proteinase K treatment for 30 min at 37°C.

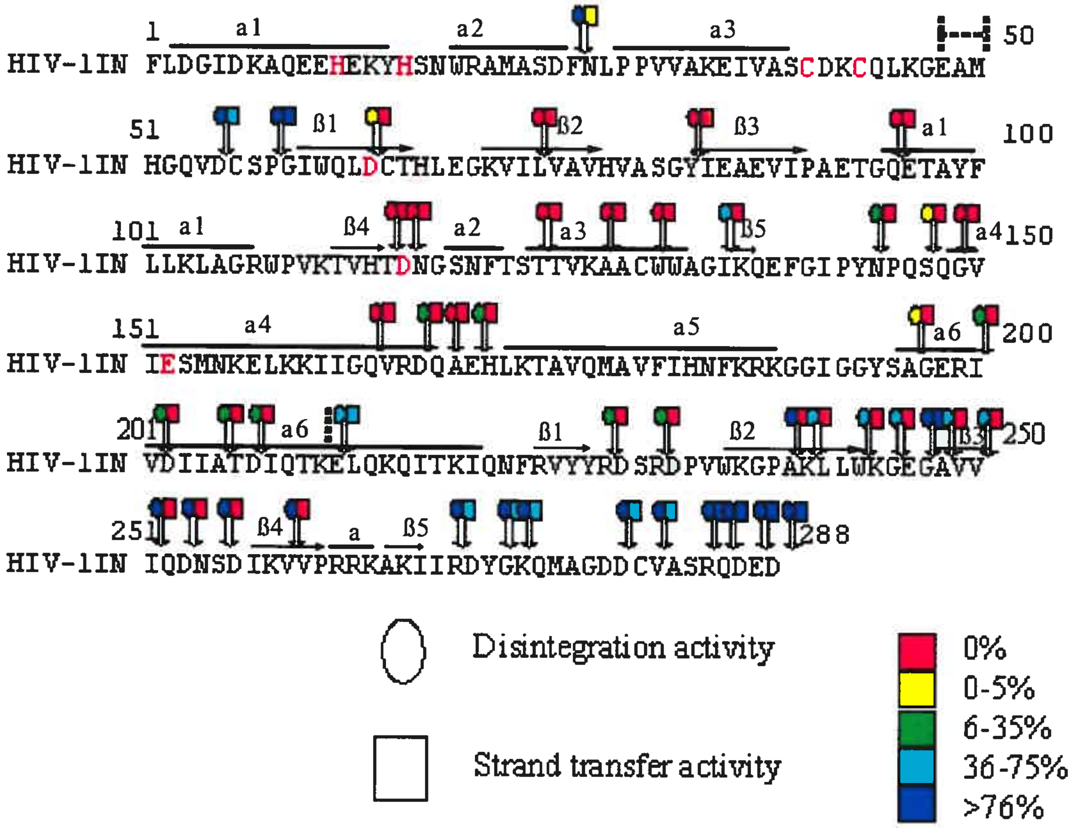
1 a1 a2 a3 50
HIV-1IN FLDGIDKAQEEHEKYHSNWRAMASDFNLPPVVAKEIVASCDKCQLKGEAM
51 ß1 ß2 ß3 a1 100
HIV-1IN HGQVDCSPGIWQLDCTHLEGKVILVAVHVASGYIEAEVIPAETGQETAYF
101 a1 ß4 a2 a3 ß5 a4 150
HIV-1IN LLKLAGRWPVKTVHTDNGSNFTSTTVKAACWWAGIKQEFGIPYNPQSQGV
151 a4 a5 a6 200
HIV-1IN IESMNKELKKIIGQVRDQAEHLKTAVQMAVFIHNFKRKGGIGGYSAGERI
201 a6 ß1 ß2 ß3 250
HIV-1IN VDIIATDIQTKELQKQITKIQNFRVYYRDSRDPVWKGPAKLLWKGEGAVV
251 ß4 a ß5 288
HIV-1IN IQDNSDIKVVPRRKAKIIRDYGKQMAGDDCVASRQDED
Disintegration activity
Strand transfer activity
0%
0-5%
6-35%
36-75%
>76%

FIGURE 2.6. Positions of each insertion (indicated by arrow) and their activity (using different color scheme) relative to disintegration (circle) and strand transfer activity (square) are shown for the amino acid sequence of HIV-1 IN protein. Numbering from the *N*-terminus of HIV-1 IN. Known structural elements of HIV-1 IN, determined by crystallography of recombinant HIV-1 IN [125, 131], are also shown (bold horizontal lines) above the respective homologous segments. Their PDB accession numbers are 1K6Y and 1ex4, respectively. `HHCC' and `DDE' motifs are highlighted in red.

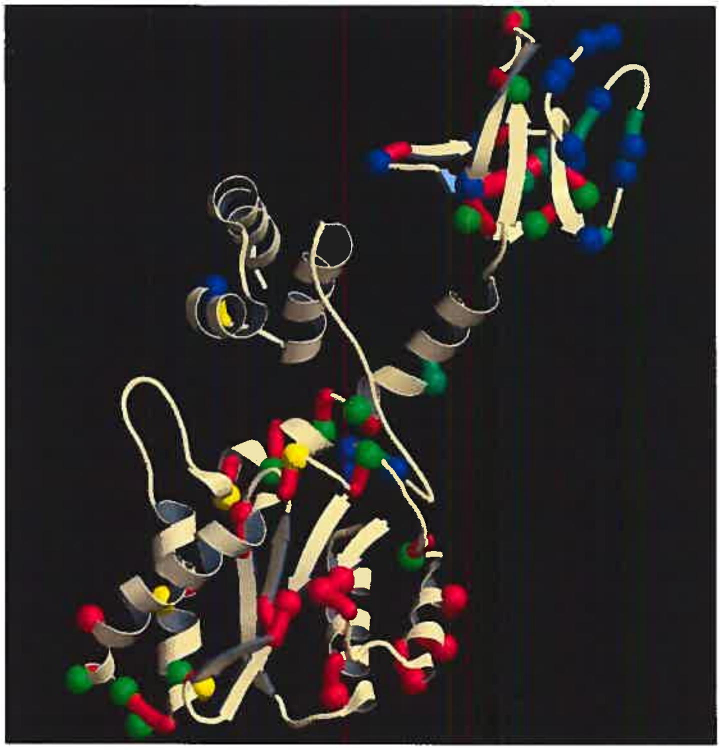

FIGURE 2.7. A three-dimensional structural model of the HIV-1 monomer. The location of the insertional mutations and their subsequent effects on disintegration and strand transfer activity are shown using the color scheme corresponding to Fig.2.6. The large spheres denote disintegration activity and the widened colored linear portions denote strand transfer activity.

Insertions after amino acids 239 (mutants A239_K, K240_L, W243_K, and G245_E in β2; V250_I, I251_Q in β3, D253_N and S255_D in loop between β3 and β4, V259_V in β5) all lost strand transfer activity while exhibiting full or slightly decreased disintegration activity. This agrees with the observation that the *C*-terminal deletion of 25 or 45 amino acids results in complete loss of integration and 3'-processing activities [33] and also agrees with the observation that the *C*-terminal deletion mutants (1-248, 55-248, 1-206) exhibit higher or same level activities in disintegration [130]. Mutant G247_A was an exception as it retained full integration activity and disintegration activity. Interestingly, the insertions in β2 and β3, which are right before and after 247 had no integration activity and were decreased in disintegration activity. Insertions after I268 and before Q284 had similar levels of activity in disintegration and retain moderate activity compared to wild type IN. This agrees with the observation that deletion of 15 amino acids from end of the *C*-terminal domain retain similar or slightly reduced 3'-processing activities and decreased integration activities in the presence of Mn^{2+} (but not Mg^{2+})[33] . Insertions after R284 retain full activity both in the disintegration and integration reactions. This again agrees with the observation that *C*-terminal deletion of 5 amino acids exhibited wild type levels of 3'-processing activity and partial activity of integration in the presence of Mn^{2+} (but not $Mg^{2+)}$[33].

In summary, three regions retained full integration activity in this *in vitro* study of HIV-1 IN. These correspond to the hinge region connecting the *N*-terminal and the core domains, within the α6 helix connecting the core and *C*-terminal domains, and the extreme *C*-terminus of the IN.

2.5. DISCUSSION

The retroviral genome has evolved to encode multifunctional proteins expressed within polyproteins. These compact viral particles must assemble, infect, replicate and integrate the viral genome using limited enzymatic functions. In this study, we have used a transposon based mutational system (Tn5) to create a functional map of the HIV-1 IN protein. Analysis of 57 insertional mutations in HIV-1 IN indicates the presence of limited, non-essential regions tolerant of amino-acid insertions. These localize to protein and domain boundaries between the *N*-terminus and the core of IN, at the *C*-terminus of IN and between the core and *C*-terminus of IN. Although these results are non-saturating, the data indicate functional conservation even within disordered regions within crystallographic structures.

The effects of insertion mutants in these studies are in general agreement with previous biochemical studies. For example, the *N*-terminal 39 amino acid deletion mutant completely abolished 3'-processing and integration activity [33]. We observed that insertions at N27_L almost abolished integration activity. The loss of integration activity agrees with the previous observations that showed that a monoclonal antibody which bound to amino acids 27-29, destabilizes the helix bundle and decreases both 3'-processing and transfer activities of HIV-1 IN *in vitro* [132]. N27 is located at one end of helix bundle in the loop connecting the second and third helices of the HIV-1 IN *N*-terminus. The third α-helix contains amino acids C (40), C (43) which interact through Zn with H (12), H (16) in the loop connecting first α-helix. The substitution of Asn for His-12 rendered HIV-1 defective at the replication step *in vivo*. In contrast,zinc binding, 3'-processing and strand transfer activities were reduced only a few fold from those of wild-type IN *in vitro* [133]. The mechanism by which the insertions, antibody and amino acids substitutions may all act is through destabilization of the helix bundle.

The effects of insertions (I135_K, N144_P, S147_Q and G149_V) within the surface loop (residues 141-148) agrees with a previous observation that substitution of Gly residues at 140 and 149 with more constrained Ala residues impaired catalysis of HIV-1 IN, suggesting that the degree of conformational flexibility at these positions is correlated with catalytic activity [17]. These two loops are believed to undergo significant movement in order to aid in the coordination of a metal ion by the catalytic triad [134, 135]. It is likely that the insertions in these two loops prevent the conformation change of these loops upon the metal binding. Interestingly, residues 168-171 are also reported to interact with the host factor LEDGF [136].

E212_L, which retains both activities at a good level, is within the region connecting the *C*-terminus and core, and consists of an extended alpha-helix with a bend, or kink, at the center. This result suggests that considerable flexibility in the linkage between the catalytic core and *C*-terminal domain can be functionally tolerated since the net result of the 19 amino acid insertion would be to lengthen the distance between the two domains and/or increase the discontinuity of the extended alpha-helix. This agrees with the observation in the solved two domain structure of the catalytic core and *C*-terminal of HIV-1, SIV-1 and RSV IN, in which the arrangement of the *C*-terminal domain relative to catalytic core differs among these structures [30, 131, 137]. Similarly, disintegration requires only the core domain (residues 50-186 [39]). Results from these studies are in agreement, with our results that showed disintegration was diminished with insertions between D64_C and S230_R.

Insertions R228_D and S230R, within the *C*-terminal domain, are interesting in that their losses of the disintegration activity are more than result from the entire deletion of *C*-terminal domain of IN. In a tetramer model of IN, these two sites flank the target DNA and are also very close to both viral DNA ends [138]. This suggests that these two insertions may prevent or distort target DNA and/or viral DNA access. Insertions after amino acid 239 generally agree with the previous *C*-terminus deletion studies in both disintegration and integration activities. Interestingly, the insertion G247_A retains both full disintegration and integration activities, whose placement seemingly does not present any structural interference on a tetramer model of HIV-1 IN [126], but may interfere with some amino acids in LEDGF when bound to HIV-1 and placed in a molecular model of the HIV-1 tetramer plus LEDGF (Monica Roth, personal communication). It would be interesting to test the effect of this insertion mutant *in vivo*. Full activities in both disintegration and integration reaction of insertions after amino acid 284 suggests that the extreme *C*-terminus of the IN protein is non-essential, which has been reported previously [128] [33]. Although it is possible that the linkers are substituting for natural amino acids at that position, we did not observe instances where two in-frame insertions at the same position resulted in differential effects toward HIV-1 IN function. This might have been predicted, as the insertions frequently encode Cys, which could alter the protein folding. In this study, insertions at the same coding sequence were identified that behaved identically, indicating there was not a positive selection for a Cys residue to, for example, stabilize the region. The two insertions at D55_C encode LSLVHILRPQDVYKRQQVD and CLLYTSCGRKMCTRDRQVD and at V250_I CLLYTSCGRKMCTRDRAVV and LSLVHILRPQDVYKRQAVV.

One aim of this mutational analysis was to identify sites within the IN protein that may tolerate small insertional tags whose function may alter the target-site selection of the viral INs. Protein domains and tags have been inserted both into the *N*-terminus [139-141] and *C*-terminus [140-145] of retroviral IN constructs. The identification of the regions between the *N*-terminus, the core and *C*-terminus of IN as functional in the presence of a variety of linker-insertions strongly suggests that this region could serve as a third potential insertion site for short tags within the IN protein. The ability of this site to function in alternative protein-protein or protein-DNA interactions depends on its accessibility within the synaptic complex. Further biochemical and structural studies are required to address this question.

Chapter III

Comparison of Two Available Tetramer Models

3.1 Introduction

Although no full-length experimental structure of HIV-1 IN is available, structures of each individual domain and two consecutive domains of HIV-1 integrase are available, as discussed in Chapter 2. However, these results by themselves are fragmentary and much less direct experimental evidence exists for overall quaternary structure of IN *in vivo*. Clearly, knowledge of the quaternary structure is required in order to better understand how integrase recognizes and binds it cognate DNA targets and also to target drug design to the appropriate solvent-exposed surfaces.

To date, there exists sufficient evidence indicating that functional IN acts as a multimer, most likely as a tetramer. Previously, two tetramer models were proposed based on biochemical experiments and detailed structural information [13, 138]. The question we address in this chapter is: how do these models agree with, or differ from each other? How well do these models comply with the available experimental data? To this end, the coordinates of the models, henceforth referred to as Model A [13] and Model B [138] were obtained from the originating authors or downloading from the Protein Data Bank. They were compared with each other as well as with the reported biochemical results.

3.2 General description of Models A and B

The active form of HIV-1 IN is likely to be a tetramer, as suggested by previous biochemical and structural studies (described in Section 1.1.5). The final form of both Model A and Model B is tetrameric. Model A: composed by the superimposition of the crystal structures of the dimer of HIV-1 IN$^{52\text{-}288}$ (PDB code 1EX4) and the dimer of HIV-1 IN$^{1\text{-}212}$ (PDB code 1K6Y), modelled with viral DNA in canonical form, without metal and target DNA. Model B: composed by assembly of the NMR monomer structure of HIV-1 IN$^{1\text{-}47}$ (PDB code 1WJA), and the monomer X-ray crystallographical structure of HIV-1 IN$^{56\text{-}209}$ (PDB code 1BIS), HIV-1 IN$^{219\text{-}270}$ (PDB code 1IHV), modeled with viral DNA and target DNA in canonical form, without metal.

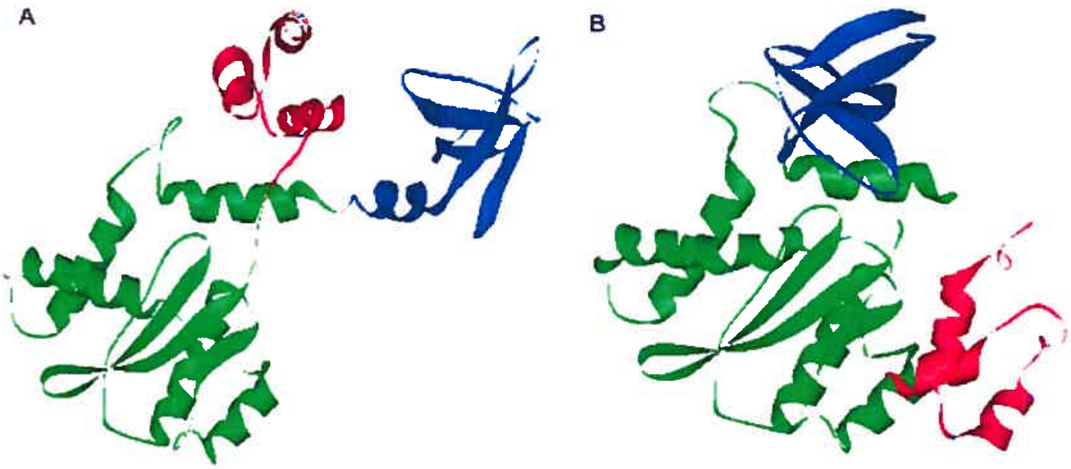

FIGURE. 3.1. Comparison of the integrase monomer structure of model A (panel A) and model B (panel B). Three domains of the integrase monomer are shown in different colors: *N*-terminal domain is in red, core domain is in green, and *C*-terminal domain is in blue.

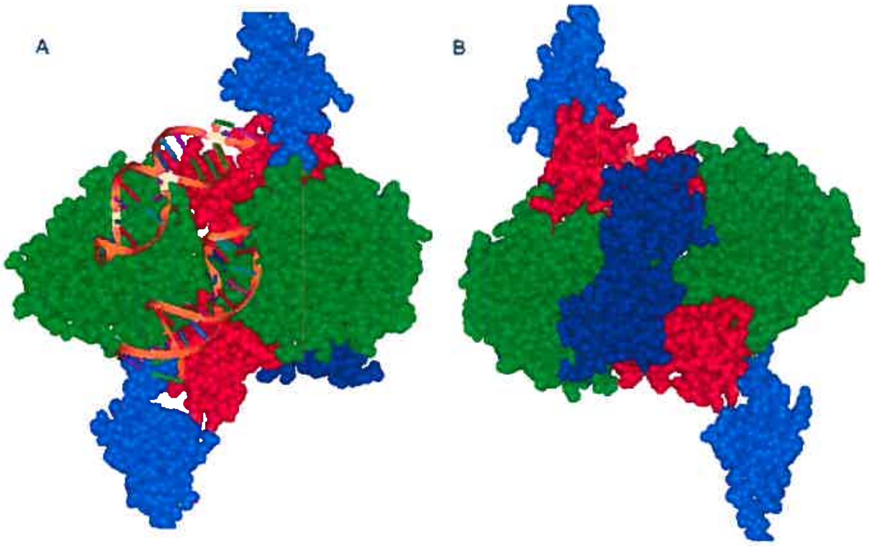

FIGURE. 3.2. Model A of HIV-1 integrase tetramer complexed with viral DNA. The integrase domains are colored as in Fig. 3.1. The viral LTRs are colored as dark orange DNA strands. The host DNA is not present in the model. Two views of the same complex are presented. (A) Front view. (B) Back view.

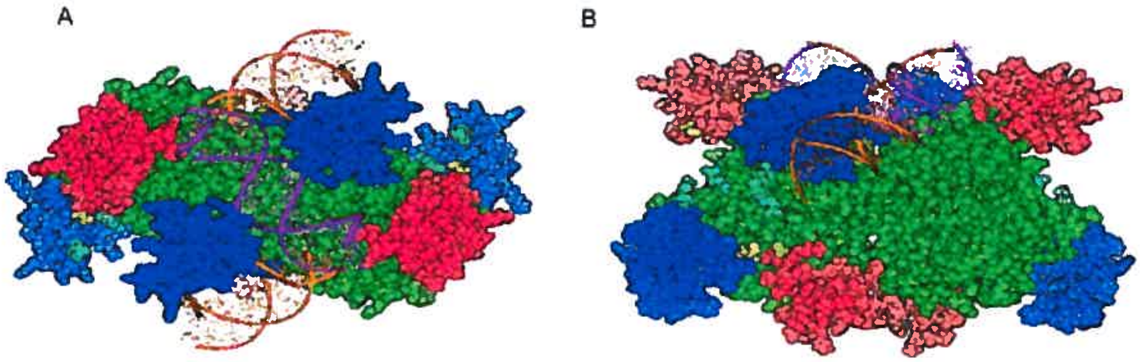

FIGURE. 3.3. Model B of HIV-1 integrase tetramer complexed with DNA. The integrase domains are colored as in Fig. 3.1. The viral LTRs are colored as dark orange DNA strands. The host DNA is in purple. Two views of the same complex are presented. (A) Front view. (B) Back view. Adapted from Figure 4 of A.A.Podtelezhnikov, *et.al.[146]*.

3.3 Detailed comparison of tetramer Models A and B

Model A takes advantage of two resolved multiple-domain structures of HIV-1 integrase, namely the catalytic core with *C*-terminal domain structure, and the *N*-terminal domain and catalytic core structure. This model is based on superposition of the conserved catalytic core of the two two-domain structures, which results in a structurally plausible full-length integrase dimer. In the case of model B, complete integrase structure modeling is based on the structures of three separate domains.

Structural comparison of different integrases indicates considerable flexibility in the linkage between the *N*-terminal domain, the *C*-terminal domain and the catalytic core [12, 30]. We may think this flexibility makes definitive modeling of the complete integrase structure from three separate domains difficult and that two models constructed independently would be substantially different. However, the constructed models result in similar tetramer structures, both of which have similar key interactions between IN residues and the viral DNA observed experimentally. The main difference is that, overall, the *N*-terminal domain and the *C*-terminal domain in Model A do not pack against the core domain as tightly as in Model B. This is particularly noticeable for the *C*-terminal domain in Model A, which has no interaction with the core domain at all, while it is not the case in Model B.

3.3.1 Juxtaposition of the core domain and the C-terminal domain

The spatial arrangement of the *C*-terminal domain relative to the core domain in Model A is based on the resolved crystal structure of the core domain + *C*-terminal domain. On the other hand, Model B takes advantage of measurements using time-resolved fluorescence anisotropy (TFA) [146] to estimate the separation between the centers of mass of the domains. However, TFA does not imply anything about the relative orientation of the integrase domains. Furthermore it is only the range of the rotational correlation times that was determined in the TFA experiments and one can question the validity of modelling the integrase quaternary structure at higher resolution based on results of TFA. However, we can't exclude the close interaction between the *C*-terminal and core domains, since the *C*-terminal domain can contact the core domain, as in the case of RSV integrase [30]

3.3.2 Role of DNA in construction of Models A and B

Model A was constructed using the two-domain structures that were resolved in absence of viral or target DNA. Model A may thus only reflect the conformation and spatial arrangement of the domains in absence of viral and target DNA; as there is evidence that binding of DNA can trigger conformational changes[89]. Model B was carefully examined against the data from cross-linking and footprinting experiments in the presence of viral/target DNA[146], increasing the likelihood of producing a reasonable model of the conformation and the arrangement in the presence of viral/target DNA. But again, it is a complex network of DNA binding rather individual domains that are unilaterally responsible for binding of HIV-1 IN to viral or target DNA; as a result, crosslinking data offer difficulties to interpret for modeling.

Model B was constructed with application of some structural constraints. However, some of the constraints may be arbitrary. For example, when building the tetramer from two dimers, the authors proposed that the interface between the dimers should have extensive contacts between them. While this hypothesis is plausible, there is no clear evidence for this; in addition, in the absence of clear knowledge about the DNA binding sites, it is difficult to conclude on the protein-protein interfaces. Furthermore, the *C*-terminal or / and *N*-terminal domains were placed to be involved in this interface as much as possible, which is not very useful and practicable for higher resolution modeling [146]. It should be noted that, despite the different considerations used for building the two models, the *N*-terminal domain and the *C*-terminal domain are close to each other in three dimensions in both models.

Neither Model A nor Model B is composed of all 288 residues of the full-length enzyme, nor are there any metal ions. The distance between the two active sites in the dimers constituting Models A and B are 40 Å and 33 Å, respectively. This distance is much greater than the five base pair spacing of the insertion sites of the two viral DNA ends into the target DNA, which is in the range of 15 Å to 20 Å (depending on the conformation of the bound DNA). This observation is consistent with the proposal of a tetrameric (or higher order) structure for IN, to allow the required proximity of active site elements via multimerization. Both models place the viral DNA along a contiguous strip of positively charged residues starting from catalytic site of one monomer and extending to the *C*-terminal domain of another monomer, which comply with the "trans to the active site" activity that has been observed [12, 43, 89]. The strips include residues K159, K186, R187, K188, and extends out to K211, K215, and K219 of the α6 helix. For example, both models place K159 close to the same adenosine nucleotide in viral DNA. However, there is a minor but important difference: in model A, the side chain of E152 approaches the backbone of the conserved adenosine nucleotide at the 3'-end directly adjacent to the cleaved site in the viral DNA (the 3'-CA motif). This is consistent with the previous experimental conclusion that E152 plays a role in the specific recognition of that specific A/T base pair from studies undertaken to identify amino acids that determine substrate specificity [84]. On the other hand, E152 is far away from conserved adenosine nucleotide at the 3'-end viral DNA in Model B. Residue E246 in the *C*-terminal domain binds near position five of the lower strand in the viral DNA in Model B, which is confirmed experimentally [89], while there is a 20 Å distance between the viral DNA and E246 in Model A. These points illustrate that neither Model A nor Model B takes into account all the previous experimental data.

3.3.3 Biological relevance of the models

After the models were constructed, the next question is: how biologically relevant are these dimer-of-dimers models? Model B can potentially explain the effects of the F185K substitution in HIV-1 integrase, which makes integrase more soluble and this effect might prevent correct formation of the integrase tetramer [64] and disrupt integration *in vivo* and *in vitro* [97, 138]. Model B places this residue in the interface of two integrase dimers. Additional support for Model B comes from structural similarity of this model with the homologous Tn5 transposase–DNA complex recently resolved by X-ray crystallography [147]. Model A is also reminiscent of the structure of Tn5 transposase dimer in complex with its DNA substrate. Both models are similar to the Tn5 structure.

The major concern about the functional relevance of tetrameric Model A is that its construction is based on the spatial arrangement of four monomers of the *C*-terminally truncated HIV-1 IN$^{1\text{-}212}$, as observed in its crystallographic structure. However, gel filtration shows that this same HIV-1 IN$^{1\text{–}212}$ fragment is exclusively dimeric at physiological salt concentrations, casting doubt on the physiological relevance of the dimer-dimer interface observed in the crystal structure of HIV-1 IN$^{1\text{-}212}$. However, tetramers may exist with the full-length integrase protein since the *C*-terminal two-domain protein (IN$^{50\text{–}288}$) forms both dimers and tetramers in solution [97], pointing to a crucial role for the *C*-terminal domain in tetramerization.

3.4 Integrating the data from the insertional mutations into the models

Most mutants we made were inactive for both disintegration and strand-transfer activities (refer to Chapter 2, particularly Table 1). It is impossible to tell whether the inactivity of our mutants is due to structural disruption or whether the bulky insertions prevent specific interactions such as IN-DNA, monomer-monomer, or dimer-dimer interaction, without further characterization of IN mutants. Our results and existing experimental data indicate important functional conservation even within disordered regions of the crystallographic structures: for example, substitution of Gly140 and 149, which is in the loop between $\beta5$ and $\alpha4$ with more constrained Ala residues, impaired catalysis of HIV-1 IN [17]. For this reason, we will focus on the active mutants and only the active mutants were analyzed against Model A and Model B.

Two insertions obtained at N27_L, which is located at one end of the helix-bundle in the loop connecting the second and third α–helices of the HIV-1 IN *N*-terminus, almost abolished integration activity. These two insertions produce steric hindrance in the tetramer of model A and B. Specifically, these insertions disrupt *N*-terminal domain dimerization in Model B and interfere protein-protein interaction between *N*-terminal and *C*-terminal domain in Model A (Fig. 3.4 and Fig. 3.5). Our results provide indirect evidence that integration activity requires higher order complex formation while disintegration requires only a dimeric or even a monomer state since insertions N27_L retain full disintegration activity. We postulate that they also might interfere with an essential conformational change induced upon zinc binding. Previous studies showed that in the presence of Zn^{+2}, the HIV-1 IN multimerizes from a dimer to a tetramer, which might be essential for the integrase activity *in vitro* [45].

Insertions (D55_C and P58_G) in the hinge between the *N*-terminal domain and the core domain are active in both disintegration and integration activities. Placing insertions in this region thus causes no interference with DNA binding or intramolecular and intermolecular interactions in model B, but not model A, in which the insertions will prevent the docking of target DNA (Fig. 3.6 A and B).

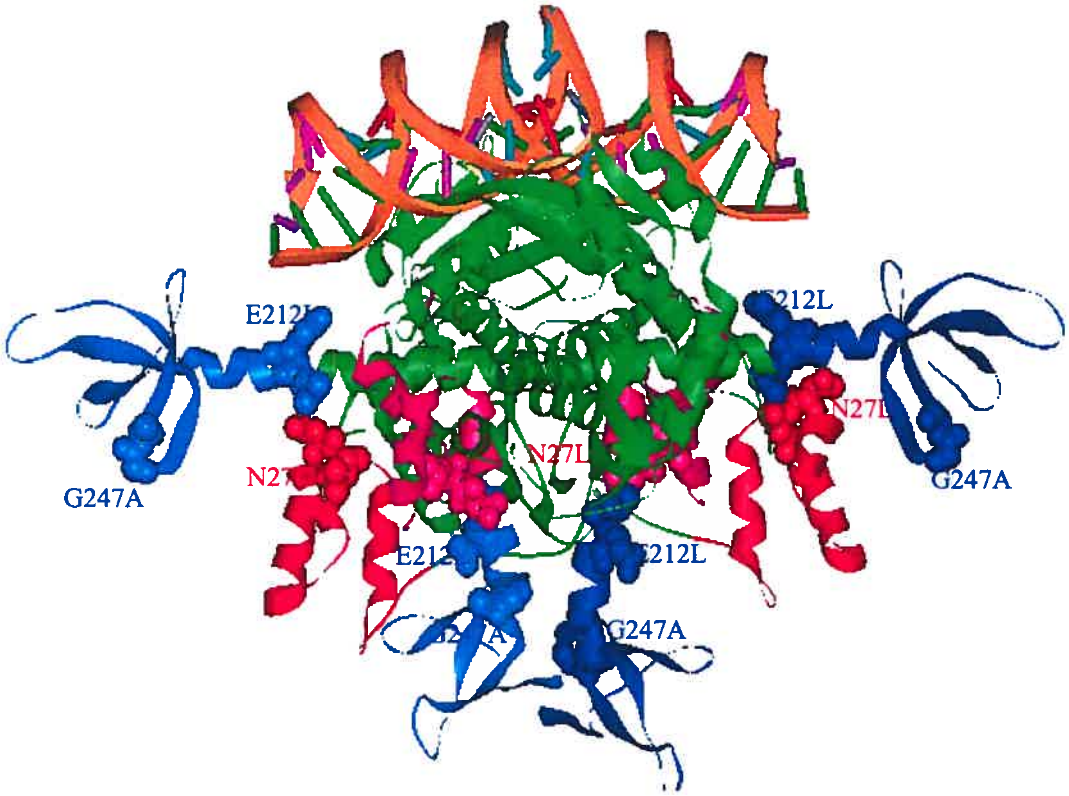

FIGURE. 3.4 Insertions N27_L, E212_L and G247_A in the context of model A of HIV-1 integrase tetramer complexed with viral DNA. The integrase domains are colored as in Fig. 3.1. Amino acids neighboring the insertions are shown as space-filling residues. The viral LTRs are in orange. The host DNA is not present in model A. The imagine was generated with MSIVIEWLAB.

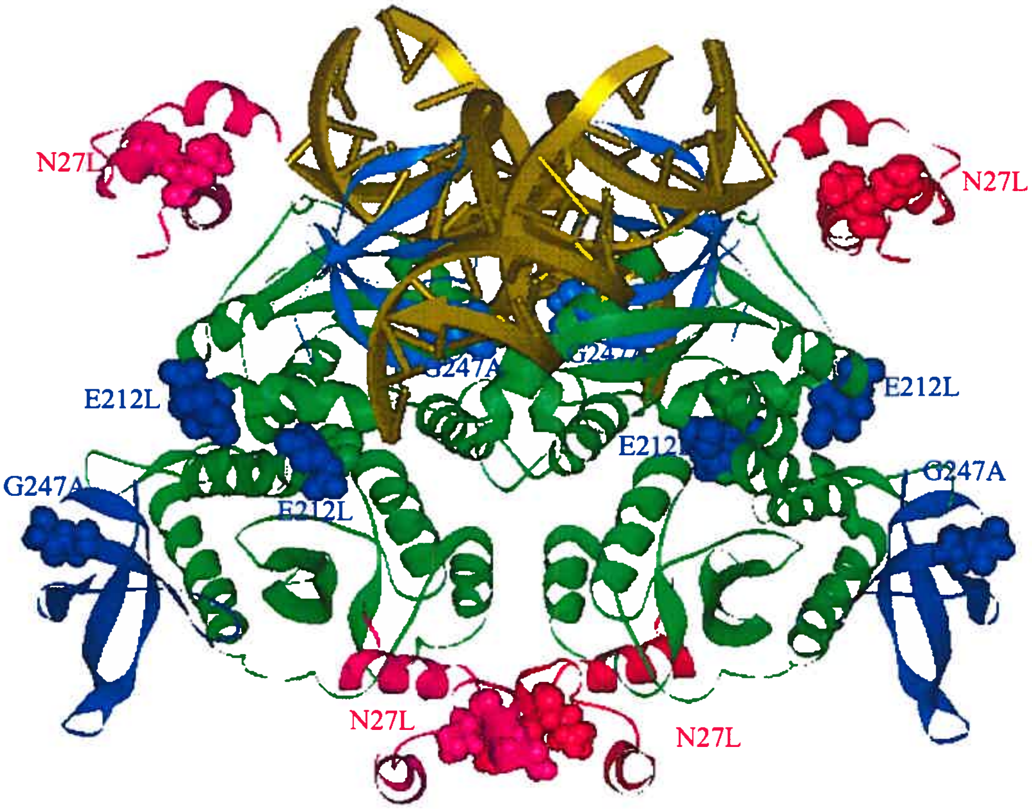

FIGURE 3.5 Insertions N27_L, E212_L and G247_A in the context of model B of HIV-1 integrase tetramer complexed with viral DNA and host DNA. The integrase domains are colored as in Fig 3.1. Amino acids neighboring the insertions are shown as space-filled residues. The viral LTRs are in orange. The imagine was generated with MSIVIEWLAB.

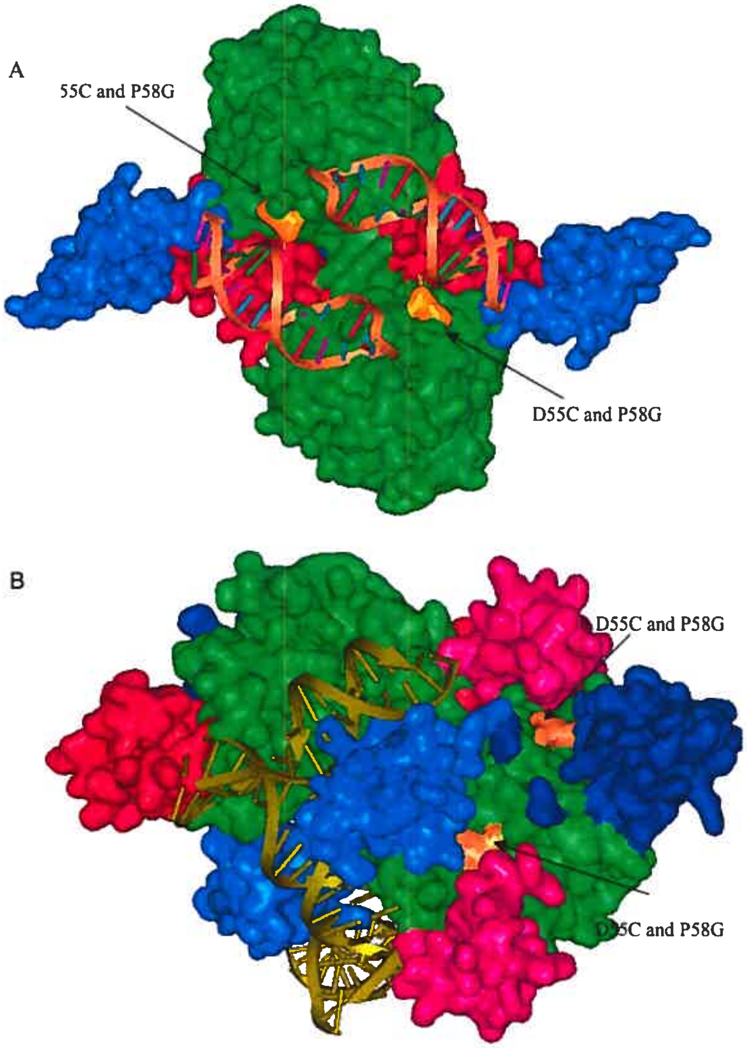

FIGURE. 3.6. Insertions D55_C and P58_G in the context of model A (Panel A) and model B (Panel B). The integrase domains are colored as in Fig. 3.1. Amino acids neighboring the insertions are shown by orange space-filled residues. The viral LTRs are shown as orange DNA strands (panel A) or yellow DNA strands (panel B). The host DNA is not present in the model A. The imagine was generated with MSIVIEWLAB.

The insertional mutant E212_L maintained nearly full disintegration and integration activities. E212_L is within the region connecting the core domain and the *C*-terminal domain, which consists of an extended alpha-helix ($\alpha 6$) with a bend, or kink, at the center. Inserting the amino acids at position 212 in both model A and model B does not interfere with DNA binding or its intramolecular and intermolecular interactions. The presence of a kink in one of the two $\alpha 6$ helices (thus, the dimer is asymmetric) which is near a known proteolytic site, as well as this new data showing tolerance to insertions, suggests that the two $\alpha 6$ helices are flexible during the integration process.

G247_A mutant retained full integration and disintegration activity. The context of G247 differs in Model A and B. Insertions of 19 amino acids in Model B would predict disruption of the interaction between the two *C*-terminal domains with the two core domains, whereas within the tetramer of Model A, this position is modeled away from the dimer interface of two *C*-terminal domains as well as the site of DNA binding. Our results are consistent with this tetramer model.

Since both models have only 270 amino acids, the insertions after D270 can not be positioned into the tetramer models, but by virtual extension of D270, we can predict that the insertions after D270 would not disrupt the interaction between both intermolecular, intramolecular of IN, and viral and host DNA, which is consistent with experimental results in both MuLV IN (M. Roth and coworkers) and HIV-1 IN (this study).

Conclusion: Model A and model B are both tetramer models, model A may be preferred in the spatial arrangement of the *N*-terminal and *C*-terminal domain relative to the core domain since this model is based on superposition of the conserved catalytic core of the two two-domain structures, which results in a structurally plausible full-length integrase dimer. The model B is preferred in the placement of viral and target DNA since their positioning are based on solid biochemical data and target DNA is absent in model A. Our insertion G247_A would favor the arrangement of *C*-terminal domains in model A. From a general point of view, both models we compared might not be the only possible solutions to the tetrameric arrangement of HIV-1 IN. Other inter-domain arrangements forming full-length monomers and inter-monomer arrangements forming dimers and tetramers or octamers are also possible. Nonetheless, our data provides important insights complementing other previously reported work on the assembly of functional HIV-1 IN.

Chapter IV

Discussion

Retroviruses are characterized as simple or complex based on the organization of their genomes, although this is not a taxonomic classification. All retroviruses contain the three major coding domains, *gag*, *pol* and *env*, in their genome. While the simple retroviruses (MuLV, for example) carry only this elementary information, complex retroviruses (HIV-1, for example) carry additional regulatory proteins that are derived from multiply spliced messages. IN proteins from different retroviral species vary in size and amino acid homology [148] [48] [149]. Two regions of strong similarity are shared between retroviral IN proteins, a proposed zinc finger motif or HHCC region in the amino terminus, and a central core region containing the D, D (35)E motif. The *C*-terminal portion of retroviral INs is the least conserved and has been functionally characterized to be the site of nonspecific DNA binding. The M-MuLV IN is approximately 14 kDa larger than HIV-1 and avian sarcoma-leukosis virus INs. This larger size of M-MuLV IN is accounted for by differences in its *N*- and *C*-termini. Approximately 50 amino acids precede the HHCC domain of the M-MuLV IN, and the *C*-terminus contains a unique 36-amino-acid insertion [148].

Several systems have been developed for "genetic footprinting" of a gene based upon the generation of a library of random inserts and screening those pools for selectable phenotypes. The systems are based on bacterial transposons including Tn5, Tn7, and Mu, or viruses [104, 109, 150-153]. These systems have the potential to screen the entire population of insertions before and after a selection process through positional mapping of the inserts by PCR.

Recent studies have demonstrated methods of comprehensive analysis of a gene by constructing an entire set of mutants of the gene [105, 107]. In the process of developing a complete functional map of the M-MuLV genome, a library of M-MuLV proviral vector insertion mutants was generated using a linker scanning system by the research group of Monica Roth at the Robert Wood Johnson Medical School, University of Medicine and Dentistry of New Jersey. After mutagenesis, each proviral vector within the library contained a single 15 base pair insertion at a random location within the target fragment. This mutagenized fragment, the 3' terminal two-thirds of the pol gene, included the last half of the reverse transcriptase (RT) reading frame the IN reading frame. All insertions were identical and code for a unique *Pme* I restriction enzyme site that was used for mapping studies of individual clones selected from the library, in a manner similar to the work presented in Chapter 2 of this thesis. The effects of the insertions were examined *in vivo* (Jennifer Puglia, Tan Wang, Christine Smith-Snyder, Marie Cote, Michael Scher, Joelle Pelletier, Sinu John, Colleen B. Jonsson and Monica J. Roth, manuscript in preparation) and are reported below in order to compare the effects of the M-MuLV insertion mutants with the effects we observed for the HIV-1 insertion mutants.

Regions in the MuLV *pol* gene were identified which functionally tolerate various linker-insertions. These correspond to the RT/IN proteolytic junction the junction between the IN catalytic core domain and the *C*-terminal domain and the C-terminus of IN (Monica Roth, personal communication).

Comparison of two related IN proteins between the in vitro HIV-1 IN study presented in Chapter 2 and the data obtained in the group of Monica Roth for the *in vivo*

MuLV IN study reveals a general agreement with respect to the function of the *C*-terminus of the IN and the viability of insertions within the α6 helix (Fig. 4.1. and Fig. 4.2): the hinge region between the core and *C*-terminal domain and extreme *C*-terminal amino acids can tolerate the insertions in both MuLV IN and HIV-1 IN.

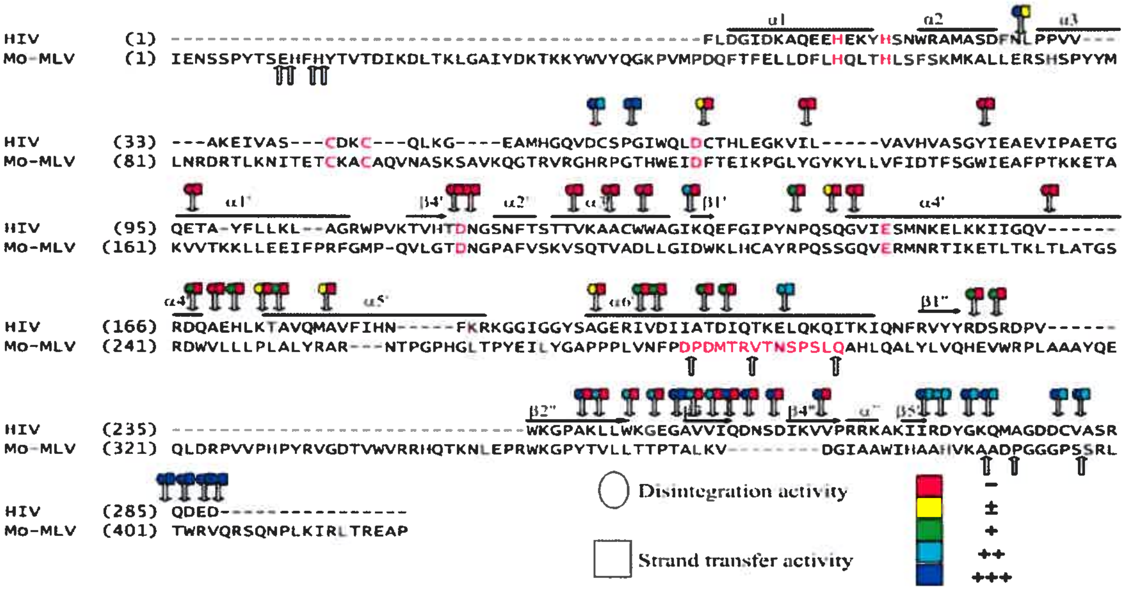
α1
α2
α3
HIV (1) --FLDGIDKAQEEHEKYHSNWRAMASDFNLPPVV---
Mo-MLV (1) IENSSPYTSEHFHYTVTDIKDLTKLGAIYDKTKKYWVYQGKPVMPDQFTFELLDFLHQLTHLSFSKMKALLERSHSPYYM
HIV (33) ---AKEIVAS---CDKC---QLKG----EAMHGQVDCSPGIWQLDCTHLEGKVIL-----VAVHVASGYIEAEVIPAETG
Mo-MLV (81) LNRDRTLKNITETCKACAQVNASKSAVKQGTRVRGHRPGTHWEIDFTEIKPGLYGYKYLLVFIDTFSGWIEAFPTKKETA
α1'
β4'
α2'
α3'
β1'
α4'
HIV (95) QETA-YFLLKL--AGRWPVKTVHTDNGSNFTSTTVKAACWWAGIKQEFGIPYNPQSQGVIESMNKELKKIIGQV------
Mo-MLV (161) KVVTKKLLEEIFPRFGMP-QVLGTDNGPAFVSKVSQTVADLLGIDWKLHCAYRPQSSGQVERMNRTIKETLTKLTLATGS
α4'
α5'
α6'
β1''
HIV (166) RDQAEHLKTAVQMAVFIHN-----FKRKGGIGGYSAGERIVDIIATDIQTKELQKQITKIQNFRVYYRDSRDPV------
Mo-MLV (241) RDWVLLLPLALYRAR---NTPGPHGLTPYEILYGAPPPLVNFPDPDMTRVTNSPSLQAHLQALYLVQHEVWRPLAAAYQE
β2''
β3''
β4''
α''
β5''
HIV (235) --------------------------------WKGPAKLLWKGEGAVVIQDNSDIKVVPRRKAKIIRDYGKQMAGDDCVASR
Mo-MLV (321) QLDRPVVPHPYRVGDTVWVRRHQTKNLEPRWKGPYTVLLTTPTALKV--------DGIAAWIHAAHVKAADPGGGPSSRL
HIV (285) QDED-----------------
Mo-MLV (401) TWRVQRSQNPLKIRLTREAP
○ Disintegration activity
□ Strand transfer activity
🟥 –
🟨 ±
🟩 +
🟦 ++
🟦 +++

Figure 4.1. Positions of each insertion (indicated by arrows) and their activity (using the indicated color scheme) relative to disintegration (circle) and strand transfer activity (square) are shown in the alignment of HIV-1 (HIV) and MuLV (Mo-M-MULV) IN proteins. The information for *in vitro* effects of HIV-1 IN insertions are given above its sequence. The information for MuLV IN insertions is given below its sequence: upward-arrows indicate the sites of insertion resulting in viable viruses in the *in vivo* MuLV IN assay. The amino acid sequence alignment of MuLV and HIV-1 IN was based on Johnson [154]. Dots indicate sequence alignment gap/insertion. Numbering from the *N*-terminus of MuLV IN includes alignment gaps. The GenBank accession number for MuLV IN sequence is NC 001501. Known structural elements of HIV-1 IN, determined by crystallography of recombinant HIV-1 IN [12, 13] are also shown (bold horizontal lines) above the respective homologous segments. Their PDB accession numbers are 1K6Y and 1ex4, respectively. `HHCC' and `DDE' motifs are highlighted by red color. The 14 amino acid region in MuLV IN (DPDMTRVTNSPSLQ), which is highlighted in red, was found to be tolerant of 5 amino acid insertions *in vivo*. This region corresponds to the HIV-1 IN sequence IATDIQFKELQKQI [148].

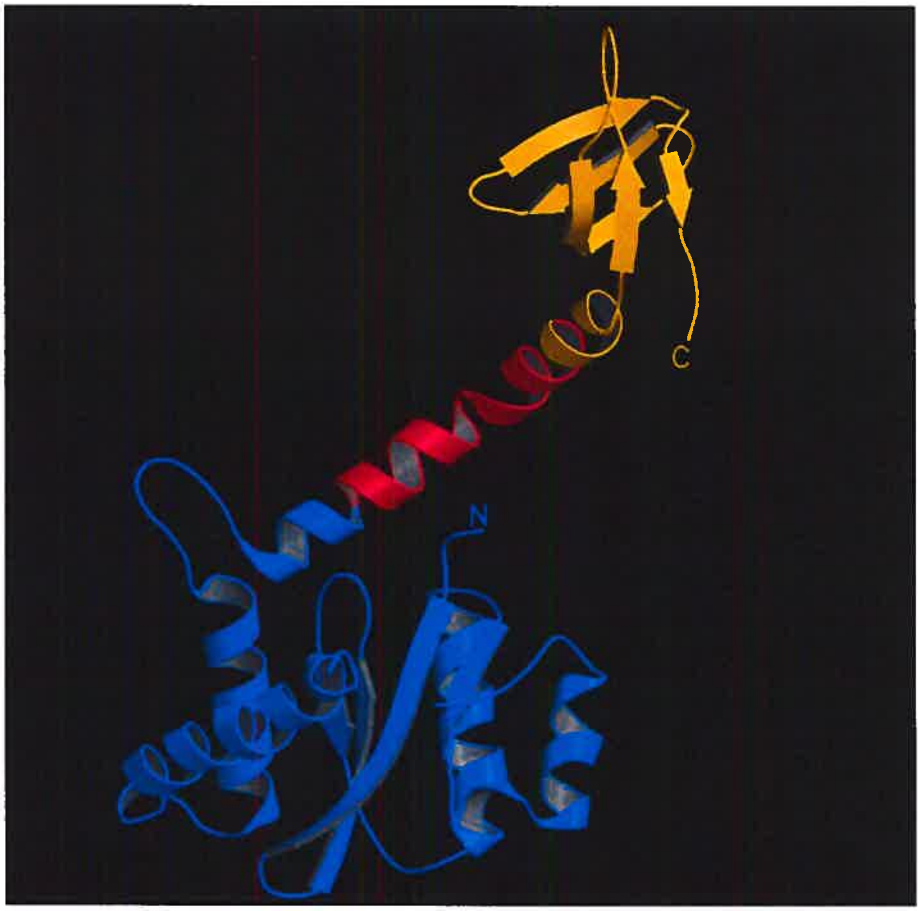

Figure 4.2. MuLV viable domain mapped onto the HIV-1 Core-*C*-terminus structure (1EX4). The 14 amino acid region in MuLV IN (DPDMTRVTNSPSLQ) was found to be tolerant of 5 amino acid insertions *in vivo*. This region corresponds to the HIV-1 IN sequence IATDIQFKELQKQI [148], which is highlighted in red (A204-I217 of the A molecule in 1EX4 is taken from the two domain structure of HIV-1 Core+*C*-terminus [12]). The HIV-1 core domain is colored blue; the *C*-terminus is yellow. The figure was generated in MOLSCRIPT V 2.0 [155].

Due to the uneven distribution of the insertions within the two systems, some questions remain open. The *N*-terminal 14 amino acids of the MuLV IN tolerated insertions; however no insertions were identified within the extreme *N*-terminus of the HIV-1 IN. It should be noted that the *N*-terminus of MuLV IN encodes 45 amino acids not conserved in either HIV or ASV related INs [156]. The region tolerant of insertions at the *N*-terminus of MuLV IN maps within this non-conserved region. As a result, the data obtained for the *N*-terminus of MuLV IN does not provide information with respect to the predicted impact of insertions in the *N*-terminus of HIV-1 IN.

These studies, *in vitro* of HIV-1 IN and *in vivo* of MuLV IN, do not allow the direct comparison of the effects of linker-insertions *in vitro* and *in vivo* within the same virus. These types of studies are of interest, as insertions tolerated in *in vitro* systems may be interfering with host-interacting proteins *in vivo*, where additional interacting partners may be required for full function. It is of interest that although functional complementation of MuLV IN was achieved *in vitro* using constructs that stably expressed the *N*-terminal zinc binding domain (IN1-105) with the core-*C*-terminus two-domain fragment (IN 106-404) [157], no viable linker-insertion was identified *in vivo* at the junction of the HHCC domain and the core domain (Jennifer Puglia, Tan Wang, Christine Smith-Snyder, Marie Cote, Michael Scher, Joelle Pelletier, Sinu John, Colleen B. Jonsson and Monica J. Roth, manuscript in preparation). However, in the case of the *in vitro* HIV-1 IN mutational study, three insertions at two positions (D55_C and P58_G) were identified at the transition between the HHCC and core domain, which retain full activity in both disintegration and strand transfer activity of HIV-1 IN.

The boundary defining the insertion-tolerant region between the core domain and the *C*-terminal domain is not well defined in these two studies since a more saturated library within this region would be required. Nonetheless, the 19 amino-acid insertions obtained within HIV-1 IN in this study and the 5 amino-acid insertions obtained within MuLV IN (Jennifer Puglia, Tan Wang, Christine Smith-Snyder, Marie Cote, Michael Scher, Joelle Pelletier, Sinu John, Colleen B. Jonsson and Monica J. Roth, manuscript in preparation) provide some insights into these boundaries. In MuLV IN, the region encoding DPDMTRVTNSPSLQ corresponds to the HIV-1 IN sequence IATDIQFKELQKQI (Fig. 4.1 and Fig. 4.2). A deletion study of MuLV IN identified a stable expressed *C*-terminal domain, the *N*-terminus of which mapped directly within this region (Monica Roth and coworkers, personal communication), supporting this region as a domain boundary. This suggests that the region should be tolerant of insertions. Indeed, the insertion E212_L in HIV-1, which retains both disintegration and strand transfer activity, maps within the 12 amino acid region homologous to MuLV IN (IATDIQFKELQKQI, where the EL are underlined), in the middle of the long α6 helix, which connects the core and *C*-terminal domains [131]. Insertions *C*-terminal to the observed bend tolerated insertions of both 5 and 19 amino acids, *in vitro* and *in vivo* in the HIV-1 and MuLV IN, respectively. The 19 amino acid insertion D207_I maps within the region homologous to MuLV IN (IATDIQFKELQKQI, DI insertion site underlined), yet is not active for disintegration or strand-transfer activity. Thus differences in the boundaries between HIV-1 and MuLV IN were identified. This may reflect the differences in the size of the insertions, where 5 amino acids are tolerated and 19 amino acids are not, or structural differences in the assembly of IN multimers. It is not known whether the insertions into this region present a favorable condition for the virus. In a related insertional study of the Cre recombinase, insertions into the M-N linker increased DNA binding cooperativity [158]. In that system, it was proposed that extending the length of the linker would lead to a smaller bend angle and thus stabilize partner Cre subunits binding to the loxP. In a similar manner, extending the distance between the core and *C*-terminus may assist in the assembly of the synaptic complex consisting of the two viral termini plus the target DNA.

The ability of retroviral particles to stably integrate into the host genome is a great benefit for gene delivery, but the potential for insertional mutagenesis cannot be overlooked [140, 141, 159, 160. Schemes to target integration into alternative positions within the host chromosome frequently involve generation of fusion proteins with novel targeting domains {Bushman, 1995 #96, 161, 162]. The linker-insertion genetic footprint provides a means to identify non-essential regions within proteins capable of withstanding insertions.

Two models have been compared with the interaction of viral DNA with a multimer of IN. The question that we would like to answer will be: is our experimental data preferred by one model than other? Can the two models agree with our experimental data? The two models both predict that our N-terminal insertion would affect the activities of HIV-1 IN, which agrees with our experimental data. Our C-terminal insertion G247_A would favor the arrangement of C-terminal domains in model A. Through both models, we would also predict that some of the loop regions within the Core might be more amenable to mutation given the solvent accessibility shown in the monomer and dimer structures, such as the loops between the N-terminal $\alpha3$ and the Core

β1', the Core β5' and α4' and between the Core α5' and α6. While we did not expect integration activity per se, we expected disintegration since this activity may not require a higher order complex. However, in our studies, insertions located at the Core loops all lost integration activity and had no or barely detectable disintegration activity. This set of mutants in core demonstrates the compactness of IN and underscores the complexity of intramolecular and intermolecular interactions that IN must maintain during the integration process.

References

1. Moss, A.R. and P. Bacchetti, *Natural history of HIV infection.* Aids, 1989. **3**(2): p. 55-61.

2. Levy, J.A., *Pathogenesis of human immunodeficiency virus infection.* Microbiol Rev, 1993. **57**(1): p. 183-289.

3. Fauci, A.S., et al., *Immunopathogenic mechanisms of HIV infection.* Ann Intern Med, 1996. **124**(7): p. 654-63.

4. Bushman, F.D. and R. Craigie, *Sequence requirements for integration of Moloney murine leukemia virus DNA in vitro.* J Virol, 1990. **64**(11): p. 5645-8.

5. Craigie, R., *HIV integrase, a brief overview from chemistry to therapeutics.* J Biol Chem, 2001. **276**(26): p. 23213-6.

6. Asante-Appiah, E. and A.M. Skalka, *Molecular mechanisms in retrovirus DNA integration.* Antiviral Res, 1997. **36**(3): p. 139-56.

7. Hindmarsh, P. and J. Leis, *Retroviral DNA integration.* Microbiol Mol Biol Rev, 1999. **63**(4): p. 836-43, table of contents.

8. Katz, R.A. and A.M. Skalka, *The retroviral enzymes.* Annu Rev Biochem, 1994. **63**: p. 133-73.

9. Mumm, S.R. and D.P. Grandgenett, *Defining nucleic acid-binding properties of avian retrovirus integrase by deletion analysis.* J Virol, 1991. **65**(3): p. 1160-7.

10. Engelman, A. and R. Craigie, *Identification of conserved amino acid residues critical for human immunodeficiency virus type 1 integrase function in vitro.* J Virol, 1992. **66**(11): p. 6361-9.

11. van Gent, D.C., et al., *Complementation between HIV integrase proteins mutated in different domains.* Embo J, 1993. **12**(8): p. 3261-7.

12. Chen, J.C., et al., *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, 2000. **97**(15): p. 8233-8.

13. Wang, J.Y., et al., *Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein.* Embo J, 2001. **20**(24): p. 7333-43.

14. Asante-Appiah, E. and A.M. Skalka, *HIV-1 integrase: structural organization, conformational changes, and catalysis.* Adv Virus Res, 1999. **52**: p. 351-69.

15. Dyda, F., et al., *Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases.* Science, 1994. **266**(5193): p. 1981-6.

16. Goldgur, Y., et al., *Three new structures of the core domain of HIV-1 integrase: an active site that binds magnesium.* Proc Natl Acad Sci U S A, 1998. **95**(16): p. 9150-4.

17. Greenwald, J., et al., *The mobility of an HIV-1 integrase active site loop is correlated with catalytic activity.* Biochemistry, 1999. **38**(28): p. 8892-8.

18. Maignan, S., et al., *Crystal structures of the catalytic domain of HIV-1 integrase free*

and complexed with its metal cofactor: high level of similarity of the active site with other viral integrases. J Mol Biol, 1998. **282**(2): p. 359-68.

19. Bujacz, G., et al., *High-resolution structure of the catalytic domain of avian sarcoma virus integrase.* J Mol Biol, 1995. **253**(2): p. 333-46.

20. Bujacz, G., et al., *The catalytic domain of avian sarcoma virus integrase: conformation of the active-site residues in the presence of divalent cations.* Structure, 1996. **4**(1): p. 89-96.

21. Bujacz, G., et al., *The catalytic domain of human immunodeficiency virus integrase: ordered active site in the F185H mutant.* FEBS Lett, 1996. **398**(2-3): p. 175-8.

22. Bujacz, G., et al., *Binding of different divalent cations to the active site of avian sarcoma virus integrase and their effects on enzymatic activity.* J Biol Chem, 1997. **272**(29): p. 18161-8.

23. Lubkowski, J., et al., *Atomic Resolution Structures of the Core Domain of Avian Sarcoma Virus Integrase and Its D64N Mutant.* Biochemistry, 1999. **38**(45): p. 15060.

24. Lodi, P.J., et al., *Solution structure of the DNA binding domain of HIV-1 integrase.* Biochemistry, 1995. **34**(31): p. 9826-33.

25. Eijkelenboom, A.P., et al., *Refined solution structure of the C-terminal DNA-binding domain of human immunovirus-1 integrase.* Proteins, 1999. **36**(4): p. 556-64.

26. Cai, M., et al., *Solution structure of the N-terminal zinc binding domain of HIV-1 integrase.* Nat Struct Biol, 1997. **4**(7): p. 567-77.

27. Cai, M., et al., *Solution structure of the His12 --> Cys mutant of the N-terminal zinc binding domain of HIV-1 integrase complexed to cadmium.* Protein Sci, 1998. **7**(12): p. 2669-74.

28. Eijkelenboom, A.P., et al., *The solution structure of the amino-terminal HHCC domain of HIV-2 integrase: a three-helix bundle stabilized by zinc.* Curr Biol, 1997. **7**(10): p. 739-46.

29. Chen, Z., et al., *X-ray structure of simian immunodeficiency virus integrase containing the core and C-terminal domain (residues 50-293)--an initial glance of the viral DNA binding platform.* J Mol Biol, 2000. **296**(2): p. 521-33.

30. Yang, Z.N., et al., *Crystal structure of an active two-domain derivative of Rous sarcoma virus integrase.* J Mol Biol, 2000. **296**(2): p. 535-48.

31. Fayet, O., et al., *Functional similarities between retroviruses and the IS3 family of bacterial insertion sequences?* Mol Microbiol, 1990. **4**(10): p. 1771-7.

32. Kulkosky, J., et al., *Residues critical for retroviral integrative recombination in a region that is highly conserved among retroviral/retrotransposon integrases and bacterial insertion sequence transposases.* Mol Cell Biol, 1992. **12**(5): p. 2331-8.

33. Drelich, M., R. Wilhelm, and J. Mous, *Identification of amino acid residues critical for endonuclease and integration activities of HIV-1 IN protein in vitro.* Virology, 1992. **188**(2): p. 459-68.

34. Bushman, F.D., et al., *Domains of the integrase protein of human immunodeficiency virus type 1 responsible for polynucleotidyl transfer and zinc binding.* Proc Natl Acad

Sci U S A, 1993. **90**(8): p. 3428-32.

35. Mazumder, A., et al., *Intermolecular disintegration and intramolecular strand transfer activities of wild-type and mutant HIV-1 integrase.* Nucleic Acids Res, 1994. **22**(6): p. 1037-43.

36. Jonsson, C.B., et al., *Functional domains of Moloney murine leukemia virus integrase defined by mutation and complementation analysis.* J Virol, 1996. **70**(7): p. 4585-97.

37. Schauer, M. and A. Billich, *The N-terminal region of HIV-1 integrase is required for integration activity, but not for DNA-binding.* Biochem Biophys Res Commun, 1992. **185**(3): p. 874-80.

38. Kulkosky, J., et al., *Activities and substrate specificity of the evolutionarily conserved central domain of retroviral integrase.* Virology, 1995. **206**(1): p. 448-56.

39. Engelman, A., F.D. Bushman, and R. Craigie, *Identification of discrete functional domains of HIV-1 integrase and their organization within an active multimeric complex.* Embo J, 1993. **12**(8): p. 3269-75.

40. Goldgur, Y., et al., *Structure of the HIV-1 integrase catalytic domain complexed with an inhibitor: a platform for antiviral drug design.* Proc Natl Acad Sci U S A, 1999. **96**(23): p. 13040-3.

41. Davies, D.R., et al., *Three-dimensional structure of the Tn5 synaptic complex transposition intermediate.* Science, 2000. **289**(5476): p. 77-85.

42. Esposito, D. and R. Craigie, *HIV integrase structure and function.* Adv Virus Res, 1999. **52**: p. 319-33.

43. Heuer, T.S. and P.O. Brown, *Photo-cross-linking studies suggest a model for the architecture of an active human immunodeficiency virus type 1 integrase-DNA complex.* Biochemistry, 1998. **37**(19): p. 6667-78.

44. Burke, C.J., et al., *Structural implications of spectroscopic characterization of a putative zinc finger peptide from HIV-1 integrase.* J Biol Chem, 1992. **267**(14): p. 9639-44.

45. Zheng, R., T.M. Jenkins, and R. Craigie, *Zinc folds the N-terminal domain of HIV-1 integrase, promotes multimerization, and enhances catalytic activity.* Proc Natl Acad Sci U S A, 1996. **93**(24): p. 13659-64.

46. Bushman, F.D. and B. Wang, *Rous sarcoma virus integrase protein: mapping functions for catalysis and substrate binding.* J Virol, 1994. **68**(4): p. 2215-23.

47. Katzman, M. and M. Sudol, *Mapping domains of retroviral integrase responsible for viral DNA specificity and target site selection by analysis of chimeras between human immunodeficiency virus type 1 and visna virus integrases.* J Virol, 1995. **69**(9): p. 5687-96.

48. Khan, E., et al., *Retroviral integrase domains: DNA binding and the recognition of LTR sequences.* Nucleic Acids Res, 1991. **19**(4): p. 851-60.

49. Jonsson, C.B. and M.J. Roth, *Role of the His-Cys finger of Moloney murine leukemia virus integrase protein in integration and disintegration.* J Virol, 1993. **67**(9): p.

5562-71.

50. Vincent, K.A., et al., *Characterization of human immunodeficiency virus type 1 integrase expressed in Escherichia coli and analysis of variants with amino-terminal mutations.* J Virol, 1993. **67**(1): p. 425-37.

51. Hazuda, D.J., et al., *Viral long terminal repeat substrate binding characteristics of the human immunodeficiency virus type 1 integrase.* J Biol Chem, 1994. **269**(6): p. 3999-4004.

52. Lee, S.P., et al., *Zn2+ promotes the self-association of human immunodeficiency virus type-1 integrase in vitro.* Biochemistry, 1997. **36**(1): p. 173-80.

53. Jenkins, T.M., et al., *Catalytic domain of human immunodeficiency virus type 1 integrase: identification of a soluble mutant by systematic replacement of hydrophobic residues.* Proc Natl Acad Sci U S A, 1995. **92**(13): p. 6057-61.

54. Hickman, A.B., et al., *Biophysical and enzymatic properties of the catalytic domain of HIV-1 integrase.* J Biol Chem, 1994. **269**(46): p. 29279-87.

55. Engelman, A., A.B. Hickman, and R. Craigie, *The core and carboxyl-terminal domains of the integrase protein of human immunodeficiency virus type 1 each contribute to nonspecific DNA binding.* J Virol, 1994. **68**(9): p. 5911-7.

56. Woerner, A.M. and C.J. Marcus-Sekura, *Characterization of a DNA binding domain in the C-terminus of HIV-1 integrase by deletion mutagenesis.* Nucleic Acids Res, 1993. **21**(15): p. 3507-11.

57. Lutzke, R.A., C. Vink, and R.H. Plasterk, *Characterization of the minimal DNA-binding domain of the HIV integrase protein.* Nucleic Acids Res, 1994. **22**(20): p. 4125-31.

58. Shibagaki, Y., et al., *Characterization of feline immunodeficiency virus integrase and analysis of functional domains.* Virology, 1997. **230**(1): p. 1-10.

59. Vink, C., A.M. Oude Groeneger, and R.H. Plasterk, *Identification of the catalytic and DNA-binding region of the human immunodeficiency virus type I integrase protein.* Nucleic Acids Res, 1993. **21**(6): p. 1419-25.

60. Pahl, A. and R.M. Flugel, *Characterization of the human spuma retrovirus integrase by site-directed mutagenesis, by complementation analysis, and by swapping the zinc finger domain of HIV-1.* J Biol Chem, 1995. **270**(7): p. 2957-66.

61. Eijkelenboom, A.P., et al., *The DNA-binding domain of HIV-1 integrase has an SH3-like fold.* Nat Struct Biol, 1995. **2**(9): p. 807-10.

62. Musacchio, A., et al., *Crystal structure of a Src-homology 3 (SH3) domain.* Nature, 1992. **359**(6398): p. 851-5.

63. Yu, H., et al., *Solution structure of the SH3 domain of Src and identification of its ligand-binding site.* Science, 1992. **258**(5088): p. 1665-8.

64. Engelman, A., et al., *Structure-based mutagenesis of the catalytic domain of human immunodeficiency virus type 1 integrase.* J Virol, 1997. **71**(5): p. 3507-14.

65. Bushman, F.D. and R. Craigie, *Activities of human immunodeficiency virus (HIV)*

integration protein in vitro: specific cleavage and integration of HIV DNA. Proc Natl Acad Sci U S A, 1991. **88**(4): p. 1339-43.

66. Vink, C., et al., *Human immunodeficiency virus integrase protein requires a subterminal position of its viral DNA recognition sequence for efficient cleavage.* J Virol, 1991. **65**(9): p. 4636-44.

67. LaFemina, R.L., P.L. Callahan, and M.G. Cordingley, *Substrate specificity of recombinant human immunodeficiency virus integrase protein.* J Virol, 1991. **65**(10): p. 5624-30.

68. Sherman, P.A., M.L. Dickson, and J.A. Fyfe, *Human immunodeficiency virus type 1 integration protein: DNA sequence requirements for cleaving and joining reactions.* J Virol, 1992. **66**(6): p. 3593-601.

69. Roth, M.J., N. Tanese, and S.P. Goff, *Gene product of Moloney murine leukemia virus required for proviral integration is a DNA-binding protein.* J Mol Biol, 1988. **203**(1): p. 131-9.

70. Muller, B., et al., *Monoclonal antibodies against Rous sarcoma virus integrase protein exert differential effects on integrase function in vitro.* J Virol, 1995. **69**(9): p. 5631-9.

71. Jenkins, T.M., et al., *Critical contacts between HIV-1 integrase and viral DNA identified by structure-based analysis and photo-crosslinking.* Embo J, 1997. **16**(22): p. 6849-59.

72. Yoshinaga, T., Y. Kimura-Ohtani, and T. Fujiwara, *Detection and characterization*

of a functional complex of human immunodeficiency virus type 1 integrase and its DNA substrate by UV cross-linking. J Virol, 1994. **68**(9): p. 5690-7.

73. van Gent, D.C., et al., *DNA binding properties of the integrase proteins of human immunodeficiency viruses types 1 and 2.* Nucleic Acids Res, 1991. **19**(14): p. 3821-7.

74. Hong, T., et al., *Human immunodeficiency virus type 1 DNA integration: fine structure target analysis using synthetic oligonucleotides.* J Virol, 1993. **67**(2): p. 1127-31.

75. Katzman, M. and M. Sudol, *Influence of subterminal viral DNA nucleotides on differential susceptibility to cleavage by human immunodeficiency virus type 1 and visna virus integrases.* J Virol, 1996. **70**(12): p. 9069-73.

76. Mazumder, A., et al., *Chemical trapping of ternary complexes of human immunodeficiency virus type 1 integrase, divalent metal, and DNA substrates containing an abasic site. Implications for the role of lysine 136 in DNA binding.* J Biol Chem, 1996. **271**(44): p. 27330-8.

77. Pemberton, I.K., M. Buckle, and H. Buc, *The metal ion-induced cooperative binding of HIV-1 integrase to DNA exhibits a marked preference for Mn(II) rather than Mg(II).* J Biol Chem, 1996. **271**(3): p. 1498-506.

78. Vink, C., R.A. Lutzke, and R.H. Plasterk, *Formation of a stable complex between the human immunodeficiency virus integrase protein and viral DNA.* Nucleic Acids Res, 1994. **22**(20): p. 4103-10.

79. Ellison, V. and P.O. Brown, *A stable complex between integrase and viral DNA ends*

mediates human immunodeficiency virus integration in vitro. Proc Natl Acad Sci U S A, 1994. **91**(15): p. 7316-20.

80. Wolfe, A.L., et al., *The role of manganese in promoting multimerization and assembly of human immunodeficiency virus type 1 integrase as a catalytically active complex on immobilized long terminal repeat substrates.* J Virol, 1996. **70**(3): p. 1424-32.

81. Balakrishnan, M. and C.B. Jonsson, *Functional identification of nucleotides conferring substrate specificity to retroviral integrase reactions.* J Virol, 1997. **71**(2): p. 1025-35.

82. Esposito, D. and R. Craigie, *Sequence specificity of viral end DNA binding by HIV-1 integrase reveals critical regions for protein-DNA interaction.* Embo J, 1998. **17**(19): p. 5832-43.

83. Wang, T., M. Balakrishnan, and C.B. Jonsson, *Major and minor groove contacts in retroviral integrase-LTR interactions.* Biochemistry, 1999. **38**(12): p. 3624-32.

84. Gerton, J.L., et al., *Effects of mutations in residues near the active site of human immunodeficiency virus type 1 integrase on specific enzyme-substrate interactions.* J Virol, 1998. **72**(6): p. 5046-55.

85. Scottoline, B.P., et al., *Disruption of the terminal base pairs of retroviral DNA during integration.* Genes Dev, 1997. **11**(3): p. 371-82.

86. Chow, S.A., et al., *Reversal of integration and DNA splicing mediated by integrase of human immunodeficiency virus.* Science, 1992. **255**(5045): p. 723-6.

87. Ellison, V., et al., *Human immunodeficiency virus integration in a cell-free system.* J Virol, 1990. **64**(6): p. 2711-5.

88. Heuer, T.S. and P.O. Brown, *Mapping features of HIV-1 integrase near selected sites on viral and target DNA molecules in an active enzyme-DNA complex by photo-cross-linking.* Biochemistry, 1997. **36**(35): p. 10655-65.

89. Gao, K., S.L. Butler, and F. Bushman, *Human immunodeficiency virus type 1 integrase: arrangement of protein domains in active cDNA complexes.* Embo J, 2001. **20**(13): p. 3565-76.

90. Bushman, F.D. and R. Craigie, *Integration of human immunodeficiency virus DNA: adduct interference analysis of required DNA sites.* Proc Natl Acad Sci U S A, 1992. **89**(8): p. 3458-62.

91. Lutzke, R.A. and R.H. Plasterk, *Structure-based mutational analysis of the C-terminal DNA-binding domain of human immunodeficiency virus type 1 integrase: critical residues for protein oligomerization and DNA binding.* J Virol, 1998. **72**(6): p. 4841-8.

92. Craigie, R., *Hotspots and warm spots: integration specificity of retroelements.* Trends Genet, 1992. **8**(6): p. 187-90.

93. Withers-Ward, E.S., et al., *Distribution of targets for avian retrovirus DNA integration in vivo.* Genes Dev, 1994. **8**(12): p. 1473-87.

94. Gerton, J.L. and P.O. Brown, *The core domain of HIV-1 integrase recognizes key features of its DNA substrates.* J Biol Chem, 1997. **272**(41): p. 25809-15.

95. Rice, P. and K. Mizuuchi, *Structure of the bacteriophage Mu transposase core: a common structural motif for DNA transposition and retroviral integration.* Cell, 1995. **82**(2): p. 209-20.

96. Jones, K.S., et al., *Retroviral integrase functions as a multimer and can turn over catalytically.* J Biol Chem, 1992. **267**(23): p. 16037-40.

97. Jenkins, T.M., et al., *A soluble active mutant of HIV-1 integrase: involvement of both the core and carboxyl-terminal domains in multimerization.* J Biol Chem, 1996. **271**(13): p. 7712-8.

98. Leavitt, A.D., L. Shiue, and H.E. Varmus, *Site-directed mutagenesis of HIV-1 integrase demonstrates differential effects on integrase functions in vitro.* J Biol Chem, 1993. **268**(3): p. 2113-9.

99. Sherman, P.A. and J.A. Fyfe, *Human immunodeficiency virus integration protein expressed in Escherichia coli possesses selective DNA cleaving activity.* Proc Natl Acad Sci U S A, 1990. **87**(13): p. 5119-23.

100. Grandgenett, D.P. and G. Goodarzi, *Folding of the multidomain human immunodeficiency virus type-I integrase.* Protein Sci, 1994. **3**(6): p. 888-97.

101. Asante-Appiah, E., G. Merkel, and A.M. Skalka, *Purification of untagged retroviral integrases by immobilized metal ion affinity chromatography.* Protein Expr Purif, 1998. **12**(1): p. 105-10.

102. Cherepanov, P., et al., *HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells.* J Biol Chem, 2003. **278**(1): p. 372-81.

103. Biery, M.C., et al., *A simple in vitro Tn7-based transposition system with low target site selectivity for genome and gene analysis.* Nucleic Acids Res, 2000. **28**(5): p. 1067-77.

104. Rothenberg, S.M., et al., *Comprehensive mutational analysis of the Moloney murine leukemia virus envelope protein.* J Virol, 2001. **75**(23): p. 11851-62.

105. Singh, I.R., R.A. Crowley, and P.O. Brown, *High-resolution functional mapping of a cloned gene by genetic footprinting.* Proc Natl Acad Sci U S A, 1997. **94**(4): p. 1304-9.

106. Auerbach, M.R., et al., *Functional characterization of a portion of the Moloney murine leukemia virus gag gene by genetic footprinting.* Proc Natl Acad Sci U S A, 2003. **100**(20): p. 11678-83.

107. Laurent, L.C., et al., *Functional characterization of the human immunodeficiency virus type 1 genome by genetic footprinting.* J Virol, 2000. **74**(6): p. 2760-9.

108. Petyuk, V., et al., *Functional mapping of Cre recombinase by pentapeptide insertional mutagenesis.* J Biol Chem, 2004. **279**(35): p. 37040-8.

109. Quinonez, R., et al., *Genetic footprinting of the HIV co-receptor CCR5: delineation of surface expression and viral entry determinants.* Virology, 2003. **307**(1): p. 98-115.

110.Bowerman, B., et al., *A nucleoprotein complex mediates the integration of retroviral DNA.* Genes Dev, 1989. **3**(4): p. 469-78.

111.Farnet, C.M. and W.A. Haseltine, *Determination of viral proteins present in the human immunodeficiency virus type 1 preintegration complex.* J Virol, 1991. **65**(4): p. 1910-5.

112.Hansen, M.S. and F.D. Bushman, *Human immunodeficiency virus type 2 preintegration complexes: activities in vitro and response to inhibitors.* J Virol, 1997. **71**(4): p. 3351-6.

113.Nermut, M.V. and A. Fassati, *Structural analyses of purified human immunodeficiency virus type 1 intracellular reverse transcription complexes.* J Virol, 2003. **77**(15): p. 8196-206.

114.Miller, M.D., C.M. Farnet, and F.D. Bushman, *Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition.* J Virol, 1997. **71**(7): p. 5382-90.

115.Lin, C.W. and A. Engelman, *The barrier-to-autointegration factor is a component of functional human immunodeficiency virus type 1 preintegration complexes.* J Virol, 2003. **77**(8): p. 5030-6.

116.Liano Martinez, H. and J. Mascias Cadavid, *[A critical commentary on the classification of headaches].* Rev Clin Esp, 1995. **195**(2 Spec No): p. 8-15.

117.Lee, M.S. and R. Craigie, *A previously unidentified host protein protects retroviral DNA from autointegration.* Proc Natl Acad Sci U S A, 1998. **95**(4): p. 1528-33.

118.Li, L., et al., *Modulation of activity of Moloney murine leukemia virus preintegration complexes by host factors in vitro.* J Virol, 1998. **72**(3): p. 2125-31.

119.Bukrinsky, M.I., et al., *Association of integrase, matrix, and reverse transcriptase antigens of human immunodeficiency virus type 1 with viral nucleic acids following acute infection.* Proc Natl Acad Sci U S A, 1993. **90**(13): p. 6125-9.

120. Chen, H. and A. Engelman, *The barrier-to-autointegration protein is a host factor for HIV type 1 integration.* Proc Natl Acad Sci U S A, 1998. **95**(26): p. 15270-4.

121. Fassati, A. and S.P. Goff, *Characterization of intracellular reverse transcription complexes of Moloney murine leukemia virus.* J Virol, 1999. **73**(11): p. 8919-25.

122. Fassati, A. and S.P. Goff, *Characterization of intracellular reverse transcription complexes of human immunodeficiency virus type 1.* J Virol, 2001. **75**(8): p. 3626-35.

123. Engelman, A., *In vivo analysis of retroviral integrase structure and function.* Adv Virus Res, 1999. **52**: p. 411-26.

124. Yang, Z.N., et al., *The crystal structure of the SIV gp41 ectodomain at 1.47 A resolution.* J Struct Biol, 1999. **126**(2): p. 131-44.

125. Wang, J.-Y., et al., *Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein.* EMBO J., 2001. **20**: p. 7333-7343.

126. Chen, A., et al., *Identification of Amino Acids in HIV-1 and Avian Sarcoma Virus Integrase Subsites Required for Specific Recognition of the Long Terminal Repeat Ends.* J. Biol. Chem., 2006. **281**(7): p. 4173-4182.

127. Asante-Appiah, E. and A.M. Skalka, *A metal-induced conformational change*

and activation of HIV-1 integrase. J Biol Chem, 1997. **272**(26): p. 16196-205.

128. Lutzke, R.A., C. Vink, and R.H.A. Plasterk, *Characterization of the minimal DNA-binding domain of the HIV integrase protein.* Nucleic Acids Res., 1994. **22**: p. 4125-4131.

129. Lutzke, R.A.P. and R.H.A. Plasterk, *Structure-based mutational analysis of the C-terminal DNA-binding domain of human immunodeficiency virus type 1 integrase: Critical residues for protein oligomerization and DNA binding.* J. Vifol., 1998. **72**(6): p. 4841-4848.

130. Kim, D.J., et al., *Minimal core domain of HIV-1 integrase for biological activity.* Mol Cells, 2000. **10**(1): p. 96-101.

131. Chen, J.C.-H., et al., *Crystal structure of the HIV-1 integrase catalytic core and C-terminal domains: A model for viral DNA binding.* Proc. Natl. Acad. Sci. USA, 2000. **97**(15): p. 8233-8238.

132. Yi, J., et al., *An inhibitory monoclonal antibody binds at the turn of the helix-turn-helix motif in the N-terminal domain of HIV-1 integrase.* J Biol Chem, 2000. **275**(49): p. 38739-48.

133. Engelman, A., et al., *Multiple effects of mutations in human immunodeficiency virus type 1 integrase on viral replication.* J Virol, 1995. **69**(5): p. 2729-36.

134. Asante-Appiah, E. and A. Skalka, *A metal-induced conformational change and activation of HIV-1 integrase.* J Biol Chem., 1997. **272**((26)): p. 16196-205.

135. Asante-Appiah, E., S.H. Seeholzer, and A.M. Skalka, *Structural determinants of metal-induced conformational changes in HIV-1 integrase.* J. Biol. Chem., 1998. **273**(52): p. 35078-35087.

136. Cherepanov, P., et al., *Structural basis for the recognition between HIV-1 integrase and transcriptional coactivator p75.* Proc Natl Acad Sci U S A, 2005. **102**(48): p. 17308-17313.

137. Chen, Z., et al., *X-ray structure of Simian immunodeficiency virus integrase containing the core and C-terminal domain (residues 50-293)-an initial glance of the viral DNA-binding platform.* J. Mol. Biol., 2000. **296**: p. 521-533.

138. Podtelezhnikov, A., et al., *Modeling HIV-1 integrase complexes based on their hydrodynamic properties.* Biopolymers, 2003. **68**: p. 110-120.

139. Bushman, F.D., *Tethering human immunodeficiency virus 1 integrase to a DNA site directs integration to nearby sequences.* Proc. Natl. Acad. Sci. USA, 1994. **91**: p. 9233-9237.

140. Katz, R.A., G. Merkel, and A.M. Skalka, *Targeting of retroviral integrase by fusion to a heterologous DNA binding domain: in vitro activities and incorporation of a fusion protein into viral particles.* Virology, 1996. **217**: p. 178-190.

141. Tan, W., et al., *Fusion proteins consisting of human immunodeficiency virus type 1 integrase and the designed polydactyl zinc finger protein E2C direct integration of viral DNA into specific sites.* J Virol, 2004. **78**: p. 1301-1313.

142. Goulaouic, H. and S.A. Chow, *Directed integration of viral DNA mediated by*

fusion proteins consisting of human immunodeficiency virus type 1 integrase and Escherichia coli LexA protein. J. Virol., 1996. **70**(1): p. 37-46.

143. Bushman, F.D. and M.D. Miller, *Tethering Human immunodeficiency virus type 1 preintegration complexes to target DNA promotes integration at nearby sites.* J. Virol., 1997. **71**(1): p. 458-464.

144. Seamon, J.A., et al., *Differential effects of C-terminal molecular tagged integrase on replication competent Moloney-murine leukemia virus.* Virology, 2000. **274**: p. 412-419.

145. Seamon, J.A., et al., *Inserting nuclear targeting signals onto a replication-competent M-MuLV affects viral export and is not sufficient for cell cycle independent infection.* J. Virol., 2002. **76**: p. 8475-8484.

146. Podtelezhnikov, A.A., et al., *Modeling HIV-1 integrase complexes based on their hydrodynamic properties.* Biopolymers, 2003. **68**(1): p. 110-20.

147. Rice, P.A. and T.A. Baker, *Comparative architecture of transposase and integrase complexes.* Nat Struct Biol, 2001. **8**(5): p. 302-7.

148. Johnson, M.S., et al., *Computer analysis of retroviral pol genes: assignment of enzymatic functions to specific sequences and homologies with nonviral enzymes.* Proc Natl Acad Sci U S A, 1986. **83**(20): p. 7648-52.

149. Lin, T.H., et al., *Secondary structural analysis of retrovirus integrase: characterization by circular dichroism and empirical prediction methods.* Proteins, 1989. **5**(2): p. 156-65.

150. Hoffman, L., et al., *Transposome insertional mutagenesis and direct sequencing of microbial genomes.* Genetica, 2000. **108**(1): p. 19-24.

151. Rothenberg, S.M., et al., *Comprehensive mutational analysis of the Moloney murine leukemia virus envelope protein.* J. Virol., 2001. **75**(23): p. 11851-11862.

152. Biery, M.C., et al., *A simple in vitro Tn7-based transposition system with low target site selectivity for genome and gene analysis.* Nucleic Acids Res, 2000. **28**(5): p. 1067-1077.

153. Singh, I.R., R.A. Crowley, and P.O. Brown, *High-resolution functional mapping of a cloned gene by genetic footprinting.* Proc. Natl. Acad. Sci. USA, 1997. **94**: p. 1304-1309.

154. Johnson, M.S., et al., *Computer analysis of retroviral pol genes: Assignment of enzymatic functions to specific sequences and homologies with nonviral enzymes.* Proc. Natl. Acad. Sci. USA, 1986. **83**: p. 7648-7652.

155. Kraulis, P.J., *MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures.* Journal of Applied Crystallography, 1991. **24**: p. 946-950.

156. Yang, F., J.A. Seamon, and M.J. Roth, *Mutational analysis of the N-terminus of Moloney murine leukemia virus integrase.* Virology, 2001. **291**: p. 32-45.

157. Yang, F., et al., *Functional interactions of the HHCC domain of Moloney murine leukemia virus integrase revealed by non-overlapping complementation and zinc dependent dimerization.* J. Virol., 1999. **73**: p. 1809-1817.

158. Petyuk, V., et al., *Functional mapping of Cre recombinase by pentapeptide insertional mutagenesis.* J. Biol. Chem., 2004. **279**(35): p. 37040-37048.

159. Calmels, B., et al., *Recurrent retroviral vector integration at the Mds1/Evi1 locus in nonhuman primate hematopoietic cells.* Blood, 2005. **106**(7): p. 2530-3.

160. Dave, U.P., N.A. Jenkins, and N.G. Copeland, *Gene therapy insertional mutagenesis insights.* Science, 2004. **303**(5656): p. 333.

161. Katz, R.A., G. Merkel, and A.M. Skalka, *Targeting of retroviral integrase by fusion to a heterologous DNA binding domain: in vitro activities and incorporation of a fusion protein into viral particles.* Virology, 1996. **217**(1): p. 178-90.

162. Bushman, F., *Targeting retroviral integration.* Science, 1995. **267**: p. 1443-1444.

ACKNOWLEDGEMENTS

I would like to take this opportunity to thank my research director, Dr. Joelle Pelletier, for her encouragement, patience, consideration, and constructive criticism. I especially appreciate her great patience and consideration. When I lost my heart, it is she that always encourages me and I value her great encouragement very much. She is always ready to help me and always has good solution. She taught me not only the science but also the way to solve the problems.

I would like also to thank Dr. Colleen B. Jonsson and Professor Joelle Pelletier , for their financial support. Her enthusiasm and vivid participation are greatly appreciated. I appreciate her kind advice and her good judge on my work.

I would like to express many thanks to Dr. Monica Roth for letting me know her unpublished results and use them in my discussion.

I would also like to thank all the members of Joelle Pelletier's research group, Nicolas Doucet, Roberto Chica, Jordan Volpato for suggestions and supports, which are very important for my work.