



Published in final edited form as:

*FEBS J.* 2009 June ; 276(11): 2926–2946. doi:10.1111/j.1742-4658.2009.07009.x.

## Piecing together the structure of retroviral integrase, an important target in AIDS therapy

Mariusz Jaskolski<sup>1,2</sup>, Jerry N. Alexandratos<sup>3</sup>, Grzegorz Bujacz<sup>2,4</sup>, and Alexander Wlodawer<sup>3</sup>

<sup>1</sup>Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University, Poznan, Poland  
<sup>2</sup>Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland  
<sup>3</sup>Macromolecular Crystallography Laboratory, National Cancer Institute at Frederick, Frederick, MD 21702 USA  
<sup>4</sup>Institute of Technical Biochemistry, Technical University of Lodz, Poland

### Abstract

Integrase (IN) is one of only three enzymes encoded in the genomes of all retroviruses, and the one least characterized in structural terms. IN catalyzes processing of the ends of a DNA copy of the retroviral genome and its concerted insertion into the chromosome of the host cell. The protein consists of three domains, the central catalytic core domain flanked by the N-terminal and C-terminal domains, the latter involved in DNA binding. Whereas the Protein Data Bank contains a number of NMR structures of the N- and C-terminal domains of human immunodeficiency viruses (HIV-1 and HIV-2), simian immunodeficiency virus, and avian sarcoma virus (ASV) IN, as well as X-ray structures of the core domain of HIV-1, ASV, and foamy virus IN, plus several models of two-domain constructs, no structure of the complete molecule of retroviral IN has been solved to date. Although no experimental structures of IN complexed with the DNA substrates are at hand, the catalytic mechanism of IN is well understood by analogy with other nucleotidyl transferases, and a variety of models of the oligomeric integration complexes have been proposed. In this review we present the current state of knowledge resulting from structural studies of IN from several retroviruses. We also attempt to reconcile the differences between the reported structures, and discuss the relationship between the structure and function of this enzyme, which is an important, although so far rather poorly exploited target for designing drugs against HIV-1 infection.

### Keywords

HIV; AIDS; antiretroviral drugs; DNA integration; integrase

### Introduction

Although the existence of retroviruses and their ability to cause diseases have been known for almost a century [1], it was the emergence of the acquired immunodeficiency syndrome (AIDS) in the early 1980s that has provided huge impetus to the structural studies of their protein and nucleic acid components. Retroviruses, most notably human immunodeficiency virus type 1 (HIV-1), are enveloped in a glycoprotein coat and lack the high degree of internal and external symmetry that makes it possible to crystallize many relatively simple

viruses, such as picornaviruses, exemplified by the viruses that cause common cold and polio. It is thus unlikely that high resolution information about the structural organization of intact retroviruses could be obtained with the currently available methods such as crystallography, although significant progress in lower-resolution studies by electron microscopy has given us excellent ideas about global aspects of their structure [2].

A typical retrovirus such as HIV-1 has been described as “Fifteen proteins and an RNA” [3]. Three of these proteins are enzymes that are retrovirus-specific and are encoded by all retroviral genomes [4], although additional enzymes are found in some retroviruses. The structures of two of these enzymes, protease (PR) [5] and reverse transcriptase (RT) [6,7], have been investigated in extensive detail during the last 20 years, using crystallography and NMR spectroscopy. A very large number of such structures, solved for both full-length apoenzymes and for complexes with substrates, products, effectors, and inhibitors, have been published [8-13]. The detailed structural knowledge, based on low-to-medium resolution structures of RT and medium-to-atomic resolution structures of PR, has been of considerable use in the design of clinically-relevant inhibitors of these enzymes [13,14]. At this time, 18 nucleoside and non-nucleoside inhibitors of RT, as well as 10 inhibitors of PR have been approved by the US Food and Drug Administration (FDA) for treatment of AIDS. By contrast, far less is known structurally about the third retroviral enzyme, integrase (IN), and fewer inhibitors of IN have been discovered so far. Only one of them, raltegravir, has recently gained FDA approval as an AIDS drug [15].

Although many anti-HIV drugs are already available, serious side effects and the emergence of drug-resistant mutations necessitate development of novel compounds. The current drugs targeting RT and PR are not without side effects. Significant side effects include myopathy, hepatic steatitis, and lipodystrophy, caused by RT drugs alone, or a combination of RT and PR drugs. RT drugs block several mitochondrial proteins (DNA polymerase  $\gamma$ , uncoupling proteins), while PR drugs such as amprenavir or indinavir block the mechanistically unrelated enzyme, mitochondrial processing protease [16]. Inhibitors of IN appear to be particularly promising [17-19] since, unlike PR and RT, this enzyme does not have direct human homologs. Although such inhibitors might still affect the function of other enzymes, such as RAG1/2 recombinase [20], they have not as yet been shown to cause pathological effects. Drugs against IN might be given in higher, more effective doses with better tolerated side effects. The inhibitors/drugs currently in animal experimental or human clinical trials seem to be keeping this promise, having in the short term fewer side effects as compared to FDA-approved anti-PR or anti-RT drugs. In consequence, drugs targeting IN may be given in sufficiently high doses to fully block the enzyme from integrating viral DNA into the cell genome, thus allowing the host immune system to fight off the infection completely.

Whereas HIV-1 IN is clearly the most medically relevant IN, extensively investigated for over two decades, the enzyme encoded by avian sarcoma virus (ASV) was studied much earlier [21]. In addition, enzymes from other retroviruses, including HIV-2, simian immunodeficiency virus (SIV), human foamy virus (PFV), Mason-Pfizer monkey virus (M-PMV), and feline immunodeficiency virus (FIV) have been investigated as well. Although significant amount of work was done with the latter enzyme [22], it will not be further discussed here since no crystals have been obtained. Similarly, we will not discuss any further M-PMV IN [23], since we are not aware of any advanced structural studies involving this protein.

As will be discussed later, no crystal structure of full-length IN is available at this time. However, many structures of fragments of this enzyme from several different viral sources have been solved by crystallography and NMR in the last 15 years (Table S1 in the Supplementary Material), including several important structures that have appeared since the

last comprehensive review of this subject was published [24]. These data will be discussed below.

## Functional properties of retroviral integrases

In the present review, we focus predominantly on the structural aspects of retroviral integrases and not on the enzymatic mechanism and other functional features of these enzymes, which have been extensively reviewed earlier [24-27]. However, a short introduction about the basics of IN function is necessary to properly interpret the importance of various structural features.

The retroviral genomic RNA is reverse transcribed into a DNA copy by the previously mentioned retroviral enzyme, RT. The function of IN is to insert the resulting viral DNA into the host genome, with the reaction accomplished in two distinct steps (Fig. 1), both catalyzed by a triad of acidic residues in a characteristic D,D(35)E motif (two aspartates and a glutamate, the latter separated from the second aspartate by 35 residues) found in all retroviral INs. In the first, processing step, IN removes the two terminal nucleotides (GT in HIV-1, TT in ASV) from each 3' end of the double-stranded viral DNA. The second step, called "joining" or "strand transfer", involves a nucleophilic attack by the free 3'-hydroxyl of the viral DNA on the target chromosomal DNA, resulting in covalent joining of the two molecules. If the reaction is performed in a concerted manner, the second, coordinated insertion is made into the complementary strand of the target DNA, in a position 5 nucleotides away from the site of the first insertion (in HIV and SIV; 6 nucleotides in ASV). The subsequent removal of the two unpaired nucleotides at each 5' overhanging end of the viral DNA and filling of the gaps are most likely performed by host enzymes.

Although the reactions described above require only the viral and host DNA substrates and divalent metal cofactors used by the integrase during the catalytic mechanism (physiologically  $Mg^{2+}$ , but *in vitro* could also be  $Mn^{2+}$ ), more components are included in the preintegration complex (PIC) that is necessary for the integration to take place in the nucleus [28,29]. Preintegration complexes of HIV-1 have been shown to also contain viral RT and matrix proteins, as well as a number of host proteins. One of the latter proteins, called barrier-to-autointegration factor (BAF), appears to be crucial in preventing autointegration (integration of viral DNA into viral DNA) [30,31]. Whereas the structure of BAF complexed to DNA is known [32], its mode of binding to IN (if any) is not. The only cellular factor that has been shown experimentally to bind directly to IN is lens epithelium-derived growth factor (LEDGF), also known as PC4 and SFRS1 interacting protein 1 (PSIP1) or transcriptional coactivator p75 [33-36]. Structural aspects of its interactions will be discussed below. However, identification of all proteins that participate in creating PICs and assignment of their role is still not complete.

## The amino acid sequence and domain structure of retroviral integrases

A single polypeptide chain of most retroviral INs comprises ~290 residues and consists of three clearly identifiable domains [37], as well as inter-domain linkers. Some important variations are, however, present. For example, PFV IN is significantly longer, comprising 392 residues, and ASV IN is encoded as a 323-amino-acid-long protein that is post-translationally processed to the final polypeptide consisting of 286 residues, which is fully enzymatically active [38]. It must be stressed, however, that definition of the domain boundaries is to a certain extent arbitrary, due to the differences in the lengths of the linking sequences, as well as to difficulties in assignment of the residues at the borders between the domains and the linkers. As shown in Fig. 2, the N-terminal domain (NTD) of HIV-1 IN contains residues 1-46, followed by a linker consisting of residues 47-55. The catalytic core domain (CCD) contains residues 56-202, and is followed by a linking sequence 203-219.

Finally, the C-terminal domain (CTD) contains residues 220-288. The residue numbers at domain boundaries for enzymes from HIV-2 and SIV are approximately the same, whereas they differ for ASV IN (Fig. 2). For PFV IN, a possibility exists that an additional domain consisting of approximately 50 residues might be present at the N terminus, preceding the NTD domain. For practical reasons, slightly different starting and ending points have been utilized for cloning of individual domains and/or two-domain constructs that were used in structural studies. The structures of representative isolated domains of IN are shown in Fig. 3.

The sequence identity/similarity percentages for full-length HIV-1 IN are 58/74% in comparison with SIV IN, and 23/37% with ASV IN, respectively (Fig. 2). These numbers are not fully accurate, since they depend on the correctness of the structure-based alignment of IN from different viral sources. For individual domains, the identity/similarity percentages are for the NTD 55/76% comparing HIV-1 to SIV IN and 26/46% comparing it to ASV IN; for the CCD they are 61/77% and 27/46%, and for the CTD they are 53/68% and 14/25%, respectively. Clearly, sequence conservation is the lowest for the C-terminal domain. It should be stressed that the sequences included in Fig. 2 are shown for enzymes encoded by specific retroviral strains and that quite significant variations between different strains have been observed [39]. In addition, crystallographic studies of some CCDs of IN or of two-domain constructs were only possible after the introduction of mutations (see below).

Until now, no reports of crystallization of isolated NTD or CTD domains have appeared. First crystals of the HIV-1 IN CCD [40] were only obtained after an extensive mutagenesis study, which identified a mutant, F185K, with enhanced solubility [41]. A protein with a substitution F185H, corresponding to the structurally equivalent residue present in ASV IN, was also crystallized [42]. A further mutation W131E was introduced to the HIV-1 IN CCD to enhance solubility even more [43]. The CCD of ASV IN could be crystallized without mutations, although special precautions in protein handling were necessary.

The NTD-CCD construct of HIV-1 IN was crystallized using a soluble variant of the protein with the above-mentioned mutation F185K, as well as with two additional ones, W131D and F139D [44]. The combination of these mutations and use of a specific buffer allowed to increase the protein concentration up to 10 mg/ml and resulted in the growth of diffraction-quality crystals. The same three mutations were also used in crystallization of the CCD-CTD construct of HIV-1 IN, where they were also introduced with the aim of increasing the solubility [45]. Two additional mutations, C56S and C286S, were introduced to prevent non-specific aggregation. However, the structure of the analogous two-domain construct of SIV IN included only a single mutation, F185H, implemented to improve protein solubility [46].

## The catalytic domain of integrase

The central domain of IN (CCD) contains the complete catalytic apparatus and exhibits limited activity even in the absence of the other domains. Although CCD by itself does not carry out the joining reaction, it does support processing, albeit with decreased specificity [47]. CCD also supports a reaction called “disintegration”, in which donor and acceptor DNA molecules are regenerated from a substrate with a Y-letter topology [4]. Due to its importance as the core of the enzyme and because of the failure to crystallize intact INs, CCD was the first target for structural investigation of these proteins.

The structures of the isolated CCDs (Fig. 3B) have been determined in about three dozen crystallographic studies of HIV-1 IN [40,42,43,45,48-51], ASV IN [52-57], and PFV IN [58]. In addition, seven medium-to-low resolution structures of fusion constructs with one of the terminal domains also included CCDs of HIV-2 [59] and SIV [45]. Since crystals of the ASV IN CCD were easier to grow, they were studied more extensively, yielding excellent

structural data, such as the atomic resolution structure with the PDB code 1CXQ [57]. The CCD has been studied in its apo form and in various metal-complexed forms, including the catalytically competent divalent cations  $Mg^{2+}$  and  $Mn^{2+}$ . Again, ASV IN has provided a more exhaustive picture of metal coordination by the catalytic core domain, including occupation of multiple metal sites, or the presence of cations such as  $Zn^{2+}$  that can also act as inhibitors of IN activity. Whereas six structures of small-molecule inhibitor complexes of the HIV-1 and ASV CCD have been published [43,51,56], it has not been possible to elucidate any structure of a DNA complex, although some promising crystallization results were achieved. In variance with the situation concerning the structure of the peripheral IN domains, no solution structure of the CCD is available.

The CCD is built around a five-stranded mixed  $\beta$ -sheet flanked by  $\alpha$ -helices (Fig. 3B). The antiparallel  $\beta 1$ - $\beta 2$ - $\beta 3$  hairpin-type arrangement is extended by two parallel strands  $\beta 4$ ,  $\beta 5$ , which are part of two  $\beta$ - $\alpha$ - $\beta$  crossovers, with the intervening helices  $\alpha 1$ ,  $\alpha 3$  plus a helical turn  $\alpha 2$ , all located on one side of the  $\beta$ -sheet. The other side of the  $\beta$ -sheet is covered by a long helix  $\alpha 4$ , which runs across its face. A helix-turn-helix (HTH) motif leads to a long stretch of nearly 40 residues that has a helical conformation ( $\alpha 5$  and  $\alpha 6$ ) except for a finger-like extrusion that is formed by about 12 residues (Phe185-Ala196 in HIV-1 sequence) in the middle. The finger has a peculiar conformation, extending away from the body of the enzyme (Fig. 3B). Its general conformation is similar in CCDs from different viruses, although it pivots on its points of attachment as a semi-rigid body. Despite its glycine-rich sequence, the finger is stabilized by conserved interactions, for example by a salt bridge (Arg187...Glu198 in HIV-1) anchored at the beginning of helix  $\alpha 6$ . The finger sequence of ASV CCD is the least conserved and, for example, the above salt bridge is not preserved. The amino-acid residues of the finger are hydrophilic, in accord with its solvent exposure in the isolated CCD domain, except for the very tip, which is occupied by a conserved Ile residue. (The presence of Glu203 in an equivalent location in the ASV IN sequence provides again an exception in this regard.) This unusual chemical character of the exposed tip together with lattice contacts formed by the finger loop are most likely responsible for its variations observed in different crystal structures. The C-terminal helix  $\alpha 6$  of the CCD is truncated in PFV IN CCD and is completely absent in the construct of an isolated ASV IN CCD used for crystallographic studies [52,57]. However, the finger structure is clearly seen in the two-domain construct of ASV IN [60], where residues Lys199-Thr207 form an insert between helices  $\alpha 5$  and  $\alpha 6$ . These observations may indicate that selection of Thr207 as the C-terminal boundary of the ASV IN CCD on the basis of extensive studies of many truncation constructs [47] might not correspond to a complete CCD.

The catalytic residues of the D,D(35)E sequence signature found in all INs are presented by the middle of chain  $\beta 1$  (D64), the loop connecting  $\beta 4$ - $\alpha 2$  (the second aspartate), and by the N-terminal segment of  $\alpha 4$  (the glutamate). They are juxtaposed in a row within a patch of negative charge on the surface of the rather flat, slab-like molecule. The active-site face of the slab is opposite to the CCD dimerization face and, therefore, the two active sites of the dimeric enzyme are far apart, nearly as far as the architecture of the dimer allows. Dimerization of the CCD involves a tandem of predominantly hydrophobic  $\alpha 1 \dots \alpha 5'$  interactions, plus hydrophobic contacts between helices  $\alpha 6$  across the dimer two-fold axis, with additional hydrophilic contacts in the middle of the dimer. The latter interactions are interesting because they are connected with the formation of a hydrophilic cavity in the center of the dimer, filled by a few water molecules.

While the  $C\alpha$  traces of the ASV and HIV-1 CCD superpose quite well, the agreement of their dimers is less optimal and reflects a slight but evident difference in the dimer architecture. As a consequence of this difference, the two active sites of the HIV-1 IN CCD dimer are less distant (38.5 vs 42.5 Å, as measured by the separation of the catalytic  $Mg^{2+}$

ions). The distance between the two active sites is incommensurate with a 5-6 bp segment of double-helical B-DNA and suggests that the host DNA must be unwound for coordinated processing of the two strands, or, more likely, that two distinct IN dimers act each on only one insertion point. Until the structure of the complete IN enzyme is solved, it can only be assumed that dimerization of the core domains of the full-length proteins is not different from what has been observed for the isolated CCD domains. This assumption is supported by the consistent picture of CCD dimerization revealed by all structures of two-domain IN constructs and of complexes of IN with LEDGF [35,59].

The CCD of HIV-1 IN used in the first structure determination (1ITG; [40]) contained the F185K mutation introduced to enhance solubility. The cacodylate residue from the crystallization buffer was found attached to the cysteine side chains of the protein, including Cys65 located in the active-site area [40]. The constellation of the catalytic acids (Asp64, Asp116, and Glu152) was found to be in an “inactive”, non-native configuration (Fig. 4A). The distortion of the catalytic apparatus became apparent only later, by comparison with other, unperturbed structures, notably the ASV IN CCD [52,53]. The non-native character of the active site is manifested by the altered conformations of the two aspartic acids, including a major re-orientation of the loop carrying the Asp116 residue, and in complete disorder of the helix fragment with the Glu152 residue and the entire flexible active-site loop in front of it (in total 13 residues, 141-153). It is unlikely that the distortion of the active site was effected by the presence of the unnatural arsenic substituent, as in a related structure of arsenic-free HIV-1 IN (2ITG; [42]), the catalytic aspartic acids are found in exactly the same inactive conformation. Although the structure 1ITG failed to map the functional state of the protein, it provided the first chain tracing and was important in revealing the plasticity of the IN active site and its ability to adopt different conformations.

Perhaps the most significant consequence of the inactive conformation of the catalytic residues is the inability of the two aspartate side chains to bind a catalytic divalent metal cation in a coordinated fashion. Such a cation, revealed by Mg- and Mn-complexes of ASV IN [53,54] and later by Mg complexes of HIV-1 IN [48,49] and PFV IN [58], has an octahedral coordination sphere completed by four water molecules (Fig. 4B). The triad of the catalytic acids can remain in the active conformation even in the absence of metal cations, but then the carboxylate groups are held in place by water-mediated hydrogen-bond bridges (Asp...Wat...Asp64...Wat...Glu). However, as revealed by the atomic-resolution structures of ASV IN, and in agreement with the requirement for basic conditions for IN activity (peak endonuclease activity at pH 8.5 [55]), conformational changes in the active site take place at pH below 6 and consist of protonation and a concomitant swing of the Asp64 carboxylate group out of its metal-coordinating position, and into a dual-hydrogen-bond lock with a neighboring asparagine residue. In addition, changes of pH influence the flexible active-site loop, which in HIV-1 IN comprises the residues 141-147, adjacent to the glutamate-bearing N terminus of helix  $\alpha$ 4, and which in all the crystal structures shows a variable degree of disorder. The flexible active-site loop contains highly conserved residues and appears to be involved directly in substrate contacts [61].

There is little doubt that the metal-coordination site formed between the two aspartate side chains (site I) corresponds to a cation essential for catalysis. The perfect octahedral geometry of this site explains why mutations of the catalytic aspartates cannot be tolerated. However, increasingly larger cations can still be accommodated, from Mg<sup>2+</sup> (mean metal...O distance 2.11 Å), to Mn<sup>2+</sup> (2.23 Å), and even Cd<sup>2+</sup> (2.43 Å) and Ca<sup>2+</sup> (2.46 Å for incomplete coordination sphere). Estimation of metal-binding geometry is more reliable from the ASV IN structures, which are in excellent agreement with expected coordination stereochemistry, for instance with valence parameters [62] of the central ion, which for the structures listed in Table S1 are calculated as 1.95 (1VSD), 1.92 (1A5V) or 1.79 (1VSJ), the

ideal target being 2.00. The corresponding values for the HIV-1 IN data indicate a high level of error, e.g. 1.23/0.91 (1BL3) or even 1.08/0.80/0.79 (1QS4), presumably as a consequence of poor data quality or structure refinement protocols. There is an important difference between ASV and HIV-1 IN in coordinating high-electron metals in site I, connected with the presence of a cysteine residue at position 65 in the latter enzyme. The thiol group of this residue is found in the coordination sphere of the cadmium cations in 1EXQ [45]. Since no such possibility exists in ASV IN, where a phenylalanine residue immediately follows the first catalytic aspartate, high-electron metals may have different impact on the catalytic properties of integrases from these two viruses. With light metals, such as  $Mg^{2+}$ , the thiol group of Cys65 in HIV-1 IN assumes a totally different orientation and, consequently, there is no difference in the coordination chemistry between ASV and HIV-1 IN.

## Structural studies of inhibitor complexes of integrase

Structural data on inhibitor complexes of IN are limited to a few structures of the CCD (Table S1). The structure of an inhibitor 5-CITEP (1-(5-chloroindol-3-yl)-3-hydroxy-3-(2H-tetrazol-5-yl)-propanone) (Fig. 5A) in complex with  $Mg^{2+}$ -containing HIV-1 IN CCD [43] is the only one that includes a compound capable of binding within the active site area of the enzyme. The  $IC_{50}$  value of 5-CITEP, measured in a reaction that monitors 3' processing together with DNA strand transfer, was reported as 2.1  $\mu M$ . This inhibitor was observed in only one of the three independent copies of the enzyme molecule present in the crystal. The molecule of 5-CITEP is located between the coordinated  $Mg^{2+}$  cation and the catalytic Glu152, with which it forms hydrogen bonds (Fig. 5B). The active site of the molecule to which the inhibitor is bound is located close to the crystallographic 2-fold axis, raising the possibility that the exact mode of binding might have been influenced by crystal contacts. The inhibitor makes no direct contacts with either Asp64 or Asp116, and has only an indirect, water-mediated contact with the bound  $Mg^{2+}$  ion. Two symmetry-related molecules of 5-CITEP interact directly with each other. In view of these facts, it is doubtful if this structure represents the true mode of binding that would be present in an IN-DNA complex.

Another IN inhibitor, 4-acetylamino-5-hydroxynaphthalene-2,7-disulfonic acid (Y-3, Fig. 5A), was cocrystallized with the ASV IN CCD in the absence and presence of  $Mn^{2+}$  [56]. This aromatic molecule with several hydrophilic substituents does not bind in the active site of the enzyme but rather on its surface, where it participates in crystallographic contacts, although there is no interference with CCD dimerization. Its presence in the crystals is, however, not a crystallographic artifact since it is observed in the same context at different pH conditions and regardless of metal coordination. Although Y-3 forms no direct interactions with the catalytic residues, it does seem to influence the conformation of the flexible active-site loop by binding to Tyr143 and Lys159 (ASV numbering). Y-3 very likely directly interferes with DNA binding by hydrogen bonding to Lys119, a residue corresponding to His114 in HIV-1 IN, shown to be capable of crosslinking to DNA. It is quite possible that these interactions are the basis of its inhibitory capacity.

The inhibitors discussed above, as well as raltegravir (Fig. 5A), the only IN inhibitor approved for clinical use, are aryl diketo acid derivatives that inhibit strand transfer much more efficiently than 3'-processing [63]. Such compounds are characterized by the presence of  $\alpha$  and  $\gamma$  C=O groups in the vicinity of a carboxylic acid moiety, although the latter group can be replaced by a triazole or tetrazole ring [64]. No structure of raltegravir complexed with IN has been published to date, but it is expected that its mode of binding might involve direct interactions with the divalent cation(s) present in the active site.

A different class of inhibitors for which structural data are available includes arsenic derivatives that were co-crystallized with HIV-1 IN [51]. Crystal structures have been

solved for tetraphenylarsonium chloride and 3,4-dihydroxyphenyl-triphenylarsonium bromide. Both compounds bind in a similar fashion at the interface of the CCD dimer and interact directly with Gln168 that belongs to one of the molecules. Surprisingly, the quality of the electron density maps is much better for the former compound than for the latter one, although only the latter one exhibits measureable inhibitory activity for the disintegration reaction ( $IC_{50}$  of 380  $\mu$ M).

Since IN must form at least a dimer to be catalytically active, prevention of dimerization offers an interesting option for its inhibition [65]. Several studies have reported inhibition of IN activity through the use of peptides derived from amino acid sequences responsible for the dimerization of the CCD domain [66,67], although no structural data are available. In some cases, it was possible to confirm that such peptides disrupted the association-dissociation equilibrium [68] or the crosslinking of the IN dimer [69]. On the other hand, Hayouka et al. [70] have demonstrated that the opposite concept, namely forcing IN to form higher-order oligomers, may be a useful approach to rendering the IN inactive. Specifically, they used peptides (called “shiftides”) derived from the cellular IN-binding protein LEDGF, to inhibit the DNA-binding of IN by shifting the enzyme's oligomerization equilibrium from the active dimer toward the tetramer, which, according to their data, is incapable of catalyzing the first step of integration, i.e. the 3'-end processing.

Development of these and other classes of IN inhibitors is an ongoing process and some very potent inhibitors, with  $IC_{50}$  in the low nanomolar range, are now available [71]. The process that led to the FDA approval of raltegravir, as well as clinical studies of other drug candidates, have been covered in a number of recent reviews [72-74]. In view of the paucity of available structural data on IN inhibitors, the wider subject of IN inhibitors in general cannot be adequately treated within the scope of the current review.

## The N-terminal domain of integrase

NMR structures of the isolated NTD domains were solved for INs from HIV-1 [75] and HIV-2 [76]. Multiple views of the N-terminal domain are also available in medium-resolution crystal structures of a two-domain construct of HIV-1 IN that contains the NTD and CCD domains (1K6Y; [44]) and of the HIV-2 NTD-CCD/LEDGF complex (3F9K; [59]). The solution structure of the HIV-1 IN NTD domain showed the existence of dimers consisting of two interconverting protein forms [75]. The two forms, denoted D (1WJA) and E (1WJC), were observed together in the NMR experiment, with the D form seen mostly above ~300 K, and the E form below that temperature. A form intermediate between these two was reported for an H12C mutant of the N-terminal domain (1WJE; [77]).

The structure of a monomer of the NTD consists principally of four helices (Fig. 3A). Helix 1 comprises residues 2-14 in the E form and 2-8 in the D form, helix 2 residues 19-25, helix 3 residues 30-39 and helix 4 residues 41-45. The segment beyond residue 46 belongs to the interdomain linker and is disordered. A zinc cation is tetrahedrally coordinated by His12, His16, Cys40, and Cys43, although the details of the interactions with the histidine residues differ between forms D and E.

The E form of the NTD is very similar to its counterpart seen in the crystal structure of the two-domain construct (1K6Y; [44]), with an rms deviation of 1.05 Å between molecules A of the models. By comparison, the rms deviations between molecule A and the other three molecules seen in the crystal range from 0.28 to 0.63 Å. Form D of the NTD deviates by almost 2 Å from its crystallographic counterpart. As expected, the interactions of the  $Zn^{2+}$  cation with its ligands in the crystal structure correspond to the structurally closer E form.



The structure of the NTD of HIV-2 IN [78,79] is very similar to its HIV-1 counterpart. A comparison between molecule A of the first model in the assembly in 1E0E (no average structure available) and molecule A of 1K6Y shows an rms deviation of 0.86 Å, although sequence identity between the two proteins is only 55%. The details of the interactions with Zn<sup>2+</sup> are also almost identical in the integrase NTD domains of HIV-1 (E form) and HIV-2. An rms deviation between NTDs belonging to molecules A and B in the structure of the HIV-2 IN NTD-CCD/LEDGF complex (3F9K; [59]) is 0.44 Å, whereas the deviation between NTD A of 3F9K and 1E0E is 1.17 Å.

## The C-terminal domain of integrase

The structure of the isolated CTD of HIV-1 IN (residues 220-270, the carboxyl terminus truncated) was solved independently by two groups using NMR (1IHV; [80] and 1QMC; [78,81]). In addition, the structures of the CCD-CTD constructs were determined by X-ray crystallography for ASV IN (1C0M, 1C1A; [60]), SIV IN (1C6V; [46]), and HIV-1 IN (1EX4; [45]). The structures of the CTD show the presence of dimeric molecules whose subunits were modeled as identical in 1IHV and as very similar in 1QMC (rms deviation 0.34 Å calculated for model 1, since no average structure is available). The rms deviation between these two structures is 1.2 Å. The deviations between the NMR structures of the isolated CTD and the crystallographic models of the two-domain constructs are larger, 1.65 Å between 1IHV and 1EX4 (both HIV-1 IN), 1.87 Å for 1C6V (SIV IN), and 2.05 Å for 1C0M (ASV IN). The four CTD domains present in the crystal structure of ASV IN consist of two very similar pairs (AB and CD, rmsd ~0.15 Å), whereas the rms deviation between molecules A and C is 0.77 Å.

A monomer of the CTD of HIV-1 IN consists of five β-strands (residues 222-229, 232-245, 248-253, 256-262, 266-270), arranged in an antiparallel manner in a β-barrel (Fig. 3C). Eighteen residues that were not included in the constructs used in the NMR experiments are also not seen in the X-ray structures of HIV-1 and SIV IN, and are presumed to be disordered. The topology of the CTD is reminiscent of SH3 domains, found in many proteins that interact with either other proteins or with nucleic acids, although no sequence similarity to SH3 proteins could be detected.

## Two-domain constructs consisting of the NTD and CCD domains

Two structures of the NTD-CCD constructs are available. A 2.4-Å resolution crystal structure of NTD-CCD of HIV-1 IN offers multiple views due to the presence of four molecules in the asymmetric unit (1K6Y; [44]), paired into AB and CD dimers, in which the twofold relationship between the catalytic domains resembles that of the isolated CCDs. Molecules A and D are very similar (rmsd 0.43 Å), whereas molecules B and C are more distant (rmsd 1.85 Å) mostly due to small changes in the interdomain angles. The interdomain linker region (residues 47-55) is disordered in all molecules, but the authors have postulated a pattern of domain connectivity taking into account the presence of NTD... CCD contacts (involving the tip of the finger loop of the CCD and one side of helix 20-24 in the NTD) and of NTD...NTD' interactions in the dimer that would conserve the symmetry of the CCD...CCD' dimer, and arguing that any other NTD-CCD connection would be incompatible with the length of the linker (Fig. 4A). In that interpretation, the distance between the end of the NTD and the beginning of CCD is about 9 Å. However, that view is contradicted by the 3.2-Å resolution crystal structure of the NTD-CCD construct of HIV-2 IN (3F9K), in which 24 IN molecules create 12 crystallographically independent dimers, each interacting with a single molecule of LEDGF [59]. Whereas the connection between NTD and CCD is broken in the electron density map of one of the IN molecules in each assembly, it is unambiguous in the other one, forming an extended chain ~18 Å in length.

Surprisingly, careful analysis of the structure 1K6Y allows re-connection of the separated NTD and CCD domains in all four molecules in exactly the same manner as in the 3F9K structure (Fig. 6C), by the use of symmetry-related domains and of NTD-CCD linkers equivalent to the intact linker from the 3F9K structure. In this model, which differs significantly from the one originally proposed [44], the NTD forms a compact structure with the CCD, using the finger loop of the latter as a docking site with a number of hydrogen-bond and electrostatic points of attachment (Fig. 7). To reconcile the two models of NTD-CCD arrangement, Cherepanov and colleagues [59] have invoked the mechanism of 3D domain swapping. However, while this is certainly a possibility, it may be more prudent to conclude that the arrangement seen in the 3F9K structure is the only model that is currently supported by experiment. The relevance of the observed NTD-CCD interactions to the functional properties of IN is not yet clear.

## Two-domain constructs consisting of the CCD and CTD domains

The structures of two-domain constructs comprising the CCD and CTD were solved independently for HIV-1 IN at 2.8 Å resolution (1EX4; [45]), for SIV IN at 3.0 Å resolution (1C6V; [46]), and for two crystal forms of ASV IN at 2.5 Å (1C0M; [60]) and 3.1 Å (1C1A; [60]) resolution. The crystals of the HIV-1 IN contain two molecules forming a dimer, although the two-fold axis relating the CCD domains differs from the operation connecting the CTD domains. In each molecule, the two domains are connected by a long, well-defined helix comprising residues 195-222. The helix separates the CCD from CTD by as much as 30 Å (Fig. 6D).

The two crystal forms of the ASV IN contain a single dimer, or a pair of dimers. Similarly to what was observed in HIV-1 IN, the symmetry operations between the two domains of each dimer differ for the CCD and CTD. The linker between the CCD and CTD comprises residues 213-223 which assume a completely extended conformation, and not the helical form observed in HIV-1 IN. Thus the number of amino acid residues forming the linker in ASV IN is much smaller than in HIV-1 IN, although the distance between the starting and ending points of these linkers is not very different, at least for one of the two crystallographically independent molecules of ASV IN.

Whereas the crystals of SIV IN also contain two dimers in the asymmetric unit, only a single CTD (denoted X) could be traced unambiguously. The chain connecting it to the CCD domain could not be traced and the authors postulated a connection with chain A of the catalytic domain [46]. If that were the case, the two domains would form a fairly compact molecule, with multiple interdomain contacts. However, an alternative assignment of the visible CTD domain to the D chain of CCD [44] would create an extended two-domain molecule not unlike that of the other two enzymes, although the inter-domain angles would differ in each of the structures. In any case, a comparison of the three structures makes it clear that the arrangement of the domains shows considerable variability and may be influenced by other parts of the molecular complex.

## Inter-domain contacts

One of the measures of the extent of interactions between the domains of IN (dimerization of identical domains, and oligomerization of different domains) is the surface area buried in their interfaces (BSA). Calculations of BSA<sup>a</sup> have been performed for a representative set of IN structures (Table S2). The CCD:CCD interactions extend over a fairly uniform area of about 1000-1650 Å<sup>2</sup>. This area does not depend on the presence of the linkers, at least with

<sup>a</sup>In all buried surface calculations in this article, the reported surface refers to one interacting protein partner, unless stated otherwise.

regard to the NTD-CCD linker (as shown by assigning the linker to either domain, or removing it altogether for the structure 3F9K). The most extensive association (largest BSA) characterizes the CCD:CCD dimer of HIV-1 IN (about 1500 Å<sup>2</sup>), and decreases in the order HIV-1>HIV-2 (~1330 Å<sup>2</sup>)>SIV (~1250 Å<sup>2</sup>)>ASV (~1080 Å<sup>2</sup>)>PFV (~1000 Å<sup>2</sup>).

Homodimeric interactions between the CTDs range between none to negligible (BSA at most ~450 Å<sup>2</sup>). In most structures the CTDs in the dimers of CCD-CTD constructs are far away from each other, possibly due to the influence of crystals contacts. However, even in the solution dimer of isolated CTD, the area of interaction is very limited (~330 Å<sup>2</sup>).

The interaction of isolated NTDs in solution is slightly stronger (~510 Å<sup>2</sup>) but still rather insubstantial. In the dimer of the NTD-CCD construct with the conformation substantiated by the 3F9K structure, direct NTD:NTD interactions are, of course, none because the NTDs fold back on their respective CCDs, and thus are completely isolated from each other. However, even in the model proposed speculatively for the HIV-1 construct, the area of direct NTD:NTD interaction is so small (260 Å<sup>2</sup>) that it can be safely neglected.

Since the NTD in a multi-domain construct folds back on the catalytic core domain, the calculation of the NTD:CCD interaction area will strongly depend on the treatment of the linker peptide (residues 47-55 in the HIV-1 IN sequence). When this sequence, which is anyway disordered in most of the structures, is completely omitted, the value of BSA is ~530 Å<sup>2</sup>. Assigning the linker to the CCD yields a slightly higher apparent buried surface (~670 Å<sup>2</sup>), but the linker certainly should not be treated as part of the NTD domain, since in that case the buried surface would be unreasonably large (~1050 Å<sup>2</sup>).

The interaction between the CCD:CTD domains is very limited, with BSA values falling below 400 Å<sup>2</sup>. In the published 1C6V model (SIV IN), the BSA exceeds 600 Å<sup>2</sup>, but after a more plausible re-interpretation of the assignment of the visible CTD to a CCD, the interaction area drops to ~100 Å<sup>2</sup>, thus becoming insignificant.

As can be gleaned from these calculations, the solvent-excluded buried areas of the homo- and hetero-interactions between the domains of IN are, with the exception of the CCD:CCD contact, not very extensive and their actual values are strongly dependent on the details of the structures used for their calculations, emphasizing once more the flexible nature of this enzyme. It must be also noted that despite the variation of the BSA calculated for the homodimer of the CCDs for IN originating from different viruses, the nature of the interactions is preserved. However, no similar consistency is seen for the homodimers of the other two domains, and the picture of the inter-domain interactions is even less clear.

## Binding of integrase to cellular protein partners

Although a number of proteins have been implicated as putative components of the preintegration complex with integrase [29], the only available structural information is for complexes of the integrase binding domain (IBD) of LEDGF with CCD of HIV-1 IN [35], and with NTD-CCD of HIV-2 IN [59]. The IBD used in these experiments included residues 347-442 of LEDGF. The complex of LEDGF with HIV-1 IN CCD consists of two catalytic domains of IN bound to two IBDs in a fully symmetric fashion. Each IBD interacts with segments of the two CCDs, the latter forming a typical dimer, as observed in all other structures of IN CCDs. The most extensive interactions between IBD and IN are with a segment including residues 166-171 of molecule A (a connecting peptide between helices  $\alpha 4$  and  $\alpha 5$ , described as an unusual helix-turn-helix motif [82]) and bury a surface area of 319 Å<sup>2</sup>. IBD also interacts with residues belonging to helix  $\alpha 3$  (and, to a lesser extent, helix  $\alpha 1$ ) of molecule B (buried surface area 379 Å<sup>2</sup>). Due to the symmetry of the complex, the second IBD molecule interacts with the corresponding areas of molecules B and A of the

CCD. The interactions of IBD with CCD in the complex with HIV-2 NTD-CCD are virtually identical, with additional, mostly electrostatic interactions provided by the N-terminal helix of NTD that belongs to molecule A (buried area 153 Å<sup>2</sup>). It is intriguing, however, that the latter complex, which was prepared by simultaneous co-expression of the interacting proteins, lacks the second IBD, even though the superposition of the common IBD in the two structures is almost exact, and the second binding site is fully formed, including the same positioning of the NTD. The importance of the structurally-derived interactions between IN and IBD was verified by extensive mutational studies of the respective interfaces [59]. It was also reported that areas of full-length LEDGF other than IBD may be involved in interactions with IN [83,84].

## Oligomeric states of full-length integrase and modeling of its structure

The oligomerization state of IN *in vivo* is still not known, but extensive *in vitro* work has shed light on this matter. The isolated NTD, CCD, and CTD domains all remain in solution as dimers, a conclusion uniformly supported by solution chemistry and structural biology studies [1]. However, experiments that found IN...DNA interaction sites by photocrosslinking also suggested that IN acts as an octamer [85]. Comparison of simulation analysis against time resolved fluorescence anisotropy measurements of rotation correlation times could distinguish monomers, dimers, and tetramers, while octamers could not be resolved from higher-order species [86]. At micromolar concentration IN exists as tetramers, octamers, and higher-order aggregates, but such concentration is much higher than cellular. At catalytic (submicromolar) concentration, these experiments showed that IN could exist as a monomer, while addition of Zn<sup>2+</sup> stimulated dimer formation. However, the authors noted that the standard buffer conditions include detergents, which dissociate IN oligomers [86]. Solution small-angle X-ray scattering (SAXS) data for complexes of IN with oligonucleotides also indicated primarily monomeric species [87], with the same caveat regarding possible effects of detergents on the oligomerization state. If detergent is eliminated from the purification and assay experiments, the IN exhibits different assembly and catalytic properties.

It must be pointed out that all of the experiments mentioned above used indirect measurements of the size of IN oligomers. More direct observations involving atomic force microscopy of intact IN complexed to a DNA substrate have shown visually that the size of these complexes is consistent with a tetramer of IN molecules [88]. Similar results were obtained by electron microscopy and single-particle image reconstruction that yielded coarse three-dimensional models at ~27-Å resolution [89]. Finding of a tetramer as the predominant feature agrees with several IN/DNA models, with analysis of IN isolated from nuclear extracts and its complex with LEDGF [35], and with dynamic light scattering experiments.

The assembly of HIV-1 IN into oligomers is different when in complex with Mn<sup>2+</sup> vs Mg<sup>2+</sup> under various *in vitro* conditions [90]. These experiments did not clarify which cation is preferred, but they did show that HIV-1 IN had no active-site cation preference when already in complex with a structural (non-catalytic) Zn<sup>2+</sup> cation. The authors concluded that binding of the catalytic cation and DNA requires a pre-existing specific IN conformation.

A number of models of IN/DNA complexes were proposed, involving either just the CCD [91-95], or the full-length enzymes [96-98]. As no structure of intact IN molecule has been reported to date, the two-domain IN constructs, namely NTD-CCD and CCD-CTD, are being used as starting points for building models of the complete HIV-1 IN protein and IN/DNA complexes [44]. These structures will be informative since they complement each other, and physically fit well together. However, it must be stressed that the IN domains are

connected by flexible linkers allowing significant inter-domain variability, and a three-domain model may not reflect the actual conformation(s) of the intact protein alone or in complex with DNA (Fig. 8).

The starting point for modeling the interactions between full-length IN and DNA was usually based upon experimental structures of recombinases (which bind DNA molecules forming Holliday junctions) [99]. The structure of Tn5 transposase as a synaptic complex transition state intermediate came as a breakthrough for integrase modelers [100]. The prokaryotic Tn5 transposase performs a series of catalytic steps, with distinct processing (endonucleolytic cleavage) and joining reactions, which are very similar to those catalyzed by retroviral IN. Also, its catalytic core domain is structurally very similar to those of retroviral INs. Tn5 functions as a dimer and its DNA binding sites provide a clear template for modeling IN...DNA interactions. These models can be used to predict IN amino acid residues important for DNA binding that can be subsequently tested experimentally.

DNA crosslinking studies implicate certain positively charged or hydrophobic residues for involvement in IN...DNA interactions. Such residues identified in HIV-1 IN include His114, Tyr143, and Lys159 [85]. The DNA-binding CTD domain contains less well-conserved residues which have been identified as important for DNA binding, *viz.* HIV-1 Glu246, Lys258, Pro261, Arg262, Lys264, with some weaker involvement of Ser230 and Arg231 [96]. The somewhat lower degree of sequence conservation in this region may reflect differences in specificity. Finally, the CTD also plays an important role in IN dimerization. When amino acid residues Leu241 and Leu242 along the C-terminal dimer interface are mutated to alanine, they disrupt IN dimerization and strongly reduce catalysis [101]. A comparison of mutants of HIV-1 and ASV INs identified a number of residues (Gln44, Leu68, Glu69, Val72, Ser153, Lys160, Ile161, Gly163, Gln164, Val165, His171, Leu172, Asp229, Ser230, and Asp253) to be responsible for the specificity of binding one of the DNA long terminal repeat ends to IN [97,98]. Further experimental and computational work is needed in order to improve the existing models of the structure of IN and of the interactions with the DNA substrates.

## Structural basis of the enzymatic activity of integrase

The CCD of IN is responsible for the two enzymatic activities of the enzyme, processing and joining. Both reactions are chemically similar, proceeding *via* a nucleophilic attack on a phosphorus atom in the DNA backbone by a donor hydroxyl group (water or the newly formed 3' OH), activated by the catalytic center of the enzyme. *In vitro*, these reactions require  $Mg^{2+}$  or  $Mn^{2+}$  cations, the latter being more efficient. However, because of physiological abundance,  $Mg^{2+}$  is assumed to be the cofactor *in vivo*. Whereas the isolated CCD may exhibit some basal enzymatic activities, the full-length enzyme is necessary for joining to proceed.

The nature and number of divalent metal cations required for catalysis are still under debate. The general composition of the IN active site (a constellation of acid groups) and the similarity of the catalyzed reactions to those carried out by other nucleotidyl transferases would strongly indicate the two-metal-cation mechanism elaborated by Steitz & Steitz [102]. However, despite numerous attempts, it has never been possible to obtain an IN/ $Mg^{2+}$  or  $Mn^{2+}$  complex with two metal cations in the active site (i.e. only site I is filled with  $Mg^{2+}$  or  $Mn^{2+}$ , while site II is empty). On the other hand, it was possible to introduce two cations into the active site by using physiologically irrelevant but stronger binding metals, such as  $Zn^{2+}$ ,  $Cd^{2+}$ , or  $Ca^{2+}$  with ASV IN [54], and  $Cd^{2+}$  with HIV-1 IN [45]. The case of zinc coordination is of special note because, firstly,  $Zn^{2+}$  accepts only four, tetrahedrally arranged ligands, which are a subset of the octahedral sphere of the other cations; secondly,

although it is not a cofactor of IN catalysis *in vivo*, it can support endonucleolytic activity in *in vitro* assays; thirdly, it severely impairs polynucleotidyl transferase activities of IN *in vitro*; and, fourthly, its potential interaction with the CCD is complicated by the fact that it is the major physiological cofactor of the NTD.

The most instructive case is  $\text{Cd}^{2+}$  coordination by IN. One has to clearly distinguish the cases of HIV-1 IN and ASV IN, because the above-mentioned Cys65 residue of HIV-1 IN actually functions as a bridge coordinating both metal centers (sites I and II) simultaneously, replacing in this role the catalytic D64, which is forced away from its active conformation [45]. In this light, the structure of the ASV IN CCD/ $\text{Cd}^{2+}$  complex [54] is more illuminating for providing insights about the possible two-metal-cation functional state of the enzyme.

There is striking similarity between the  $\text{Cd}^{2+}$ -complexed active site of ASV IN and those of other nucleotidyl transferases, most notably of RNase H, which has been described in a ternary complex with  $\text{Mg}^{2+}$  (at sites A and B) and an RNA/DNA hybrid [103], as well as TN5 transposase [104] (Fig. 6). First, the metal...metal distance is nearly identical in ASV IN and RNase H, and compatible with what Yang et al. [105] predict to be required for effective nucleotide-bond hydrolysis (4.0 Å). Additionally, in both cases the two metal centers are connected by two bridging ligands, one of them being a conserved aspartate from the catalytic apparatus (D64 in HIV-1 IN). The other bridge is provided by a water molecule in the ASV IN/ $\text{Cd}^{2+}$  (and also  $\text{Zn}^{2+}$ ) complex but in the RNase H complex structure this water is displaced and its role is assumed by an O atom of the scissile phosphate group of the RNA substrate. This phosphate group is even more essential for the integrity of the functional active site of RNase H because it also fills (albeit with less ideal stereochemistry) an additional site in the coordination sphere of site B of RNase H. (Moreover, the next phosphate of the RNA substrate participates in the activation of the water nucleophile presented in the coordination sphere of site A.) Overall, site B has much less regular stereochemistry, in contrast to the nearly perfect geometry of site A. While the simplicity and approximate mirror symmetry of the two-metal-cation active sites would allow two alternative mappings of the metal centers between RNase H and retroviral IN, there is little doubt that the correct mapping is A-I and B-II. This is because, as site A in the RNase H structure, site I of the IN CCD domain has a nearly perfect coordination sphere, while site II is far less regular, with a missing ligand, large scatter of the  $\text{Cd}^{2+}\dots\text{O}$  distances, and large angular distortions. If this analogy between the active sites of IN and RNase H is correct, then the catalytic metal cation at site I of IN would participate in activating a nucleophilic group (e.g. a water molecule) for attack on a substrate DNA phosphate group. Metal II, on the other hand, would play a role in destabilizing the enzyme-substrate complex, i.e. in driving the reaction forward. At the completion of a reaction cycle, one or both metal cations would probably dissociate as their effective binding (especially at site II) critically depends on the presence of substrate DNA. The parallel between RNase H and retroviral IN has also a chemical aspect because coordination of two  $\text{Mg}^{2+}$  cations by RNase H was easy and occurred at low metal ion concentration only in the presence of the RNA/DNA substrate. With the enzyme alone, the effective  $\text{Mg}^{2+}$  concentration had to be much higher, at non-physiological levels [106]. With ASV IN, it was not possible to introduce a catalytic metal cation at site II, despite a thorough experimental survey, in which elevated metal concentrations were used [54]. This difficulty is correlated with the flexibility of the glutamate element of the active site which participates in the formation of site II. It may be necessary for the enzyme to use external means, such as substrate assistance, to sequester an  $\text{Mg}^{2+}$  cation in site II, with subsequent or simultaneous stabilization of the glutamate side chain.

## Concluding remarks

After an initial spur of activity in the years 1994-2001 that resulted in a wealth of crystal and NMR structures of retroviral INs, only a few new structures have been published in the last eight years. Since many questions, particularly those regarding the structure of the full-length active enzyme and the multiprotein-DNA preintegration complexes, remain to be answered, further structural and biochemical work on this enzyme still needs to be pursued. In addition, IN continues to be an important target for designing anti-HIV drugs, which makes continuation of studies of its structure and function ever more important.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We would like to thank Drs. Alla Gustchina, Anna Marie Skalka, and Mark Andrade for fruitful discussions, and Dr. Peter Cherepanov for providing us with a manuscript and data prior to their publication. This project was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. The research of MJ was supported by a Faculty Scholar fellowship from the National Cancer Institute.

## REFERENCES

1. Coffin, JM.; Hughes, SH.; Varmus, HE. *Retroviruses*. Cold Spring Harbor Laboratory Press; New York: 1997.
2. Grunewald K, Cyrklaff M. Structure of complex viruses and virus-infected cells by electron cryo tomography. *Curr Opin Microbiol*. 2006; 9:437–442. [PubMed: 16829161]
3. Frankel AD, Young JA. HIV-1: fifteen proteins and an RNA. *Annu Rev Biochem*. 1998; 67:1–25. [PubMed: 9759480]
4. Katz RA, Skalka AM. The retroviral enzymes. *Annu Rev Biochem*. 1994; 63:133–173. [PubMed: 7526778]
5. Wlodawer A, Vondrasek J. Inhibitors of HIV-1 protease: a major success of structure-assisted drug design. *Annu Rev Biophys Biomol Struct*. 1998; 27:249–284. [PubMed: 9646869]
6. Sarafianos SG, Das K, Hughes SH, Arnold E. Taking aim at a moving target: designing drugs to inhibit drug-resistant HIV-1 reverse transcriptases. *Curr Opin Struct Biol*. 2004; 14:716–730. [PubMed: 15582396]
7. Ren J, Stammers DK. HIV reverse transcriptase structures: designing new inhibitors and understanding mechanisms of drug resistance. *Trends Pharmacol Sci*. 2005; 26:4–7. [PubMed: 15629197]
8. Wlodawer A, Erickson JW. Structure-based inhibitors of HIV-1 protease. *Annu Rev Biochem*. 1993; 62:543–585. [PubMed: 8352596]
9. Wlodawer A. Rational approach to AIDS drug design through structural biology. *Annu Rev Med*. 2002; 53:595–614. [PubMed: 11818491]
10. Vondrasek J, Wlodawer A. HIVdb: a database of the structures of human immunodeficiency virus protease. *Proteins*. 2002; 49:429–431. [PubMed: 12402352]
11. Louis JM, Ishima R, Torchia DA, Weber IT. HIV-1 protease: structure, dynamics, and inhibition. *Adv Pharmacol*. 2007; 55:261–298. [PubMed: 17586318]
12. Cote ML, Roth MJ. Murine leukemia virus reverse transcriptase: structural comparison with HIV-1 reverse transcriptase. *Virus Res*. 2008; 134:186–202. [PubMed: 18294720]
13. Sarafianos SG, Marchand B, Das K, Himmel DM, Parniak MA, Hughes SH, Arnold E. Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition. *J Mol Biol*. 2009; 385:693–713. [PubMed: 19022262]
14. Mastrolorenzo A, Rusconi S, Scozzafava A, Barbaro G, Supuran CT. Inhibitors of HIV-1 protease: current state of the art 10 years after their introduction. From antiretroviral drugs to antifungal,

- antibacterial and antitumor agents based on aspartic protease inhibitors. *Curr Med Chem.* 2007; 14:2734–2748. [PubMed: 18045120]
15. Summa V, Petrocchi A, Bonelli F, Crescenzi B, Donghi M, Ferrara M, Fiore F, Gardelli C, Gonzalez PO, Hazuda DJ, Jones P, Kinzel O, Laufer R, Monteagudo E, Muraglia E, Nizi E, Orvieto F, Pace P, Pescatore G, Scarpelli R, Stillmock K, Witmer MV, Rowley M. Discovery of raltegravir, a potent, selective orally bioavailable HIV-integrase inhibitor for the treatment of HIV-AIDS infection. *J Med Chem.* 2008; 51:5843–5855. [PubMed: 18763751]
  16. Gerschenson M, Brinkman K. Mitochondrial dysfunction in AIDS and its treatment. *Mitochondrion.* 2004; 4:763–777. [PubMed: 16120431]
  17. Lutzke RA, Plasterk RH. HIV integrase: a target for drug discovery. *Genes Funct.* 1997; 1:289–307. [PubMed: 17366700]
  18. Makhija MT. Designing HIV integrase inhibitors--shooting the last arrow. *Curr Med Chem.* 2006; 13:2429–2441. [PubMed: 16918363]
  19. Nair V, Chi G. HIV integrase inhibitors as therapeutic agents in AIDS. *Rev Med Virol.* 2007; 17:277–295. [PubMed: 17503547]
  20. Melek M, Jones JM, O'Dea MH, Pais G, Burke TR Jr, Pommier Y, Neamati N, Gellert M. Effect of HIV integrase inhibitors on the RAG1/2 recombinase. *Proc Natl Acad Sci U S A.* 2002; 99:134–137. [PubMed: 11756686]
  21. Grandgenett DP, Vora AC, Schiff RD. A 32,000-dalton nucleic acid-binding protein from avian retrovirus cores possesses DNA endonuclease activity. *Virol.* 1978; 89:119–132.
  22. Saenz DT, Poeschla EM. FIV: from lentivirus to lentivector. *J Gene Med.* 2004; 6(Suppl 1):S95–104. [PubMed: 14978754]
  23. Snasel J, Krejcik Z, Jencova V, Rosenberg I, Ruml T, Alexandratos J, Gustchina A, Pichova I. Integrase of Mason-Pfizer monkey virus. *FEBS J.* 2005; 272:203–216. [PubMed: 15634344]
  24. Chiu TK, Davies DR. Structure and function of HIV-1 integrase. *Curr Top Med Chem.* 2004; 4:965–977. [PubMed: 15134551]
  25. Asante-Appiah E, Skalka AM. HIV-1 integrase: structural organization, conformational changes, and catalysis. *Adv Virus Res.* 1999; 52:351–369. [PubMed: 10384242]
  26. Esposito D, Craigie R. HIV integrase structure and function. *Adv Virus Res.* 1999; 52:319–333. [PubMed: 10384240]
  27. Lewinski MK, Bushman FD. Retroviral DNA integration--mechanism and consequences. *Adv Genet.* 2005; 55:147–181. [PubMed: 16291214]
  28. Van Maele B, Busschots K, Vandekerckhove L, Christ F, Debyser Z. Cellular co-factors of HIV-1 integration. *Trends Biochem Sci.* 2006; 31:98–105. [PubMed: 16403635]
  29. Engelman A. Isolation and Analysis of HIV-1 Preintegration Complexes. *Methods Mol Biol.* 2009; 485:135–149. [PubMed: 19020823]
  30. Lin CW, Engelman A. The barrier-to-autointegration factor is a component of functional human immunodeficiency virus type 1 preintegration complexes. *J Virol.* 2003; 77:5030–5036. [PubMed: 12663813]
  31. Hindmarsh P, Leis J. Retroviral DNA integration. *Microbiol Mol Biol Rev.* 1999; 63:836–43. table. [PubMed: 10585967]
  32. Bradley CM, Ronning DR, Ghirlando R, Craigie R, Dyda F. Structural basis for DNA bridging by barrier-to-autointegration factor. *Nat Struct Mol Biol.* 2005; 12:935–936. [PubMed: 16155580]
  33. Cherepanov P, Maertens G, Proost P, Devreese B, Van Beeumen J, Engelborghs Y, De Clercq E, Debyser Z. HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *J Biol Chem.* 2003; 278:372–381. [PubMed: 12407101]
  34. Maertens G, Cherepanov P, Pluymers W, Busschots K, De Clercq E, Debyser Z, Engelborghs Y. LEDGF/p75 is essential for nuclear and chromosomal targeting of HIV-1 integrase in human cells. *J Biol Chem.* 2003; 278:33528–33539. [PubMed: 12796494]
  35. Cherepanov P, Ambrosio AL, Rahman S, Ellenberger T, Engelman A. Structural basis for the recognition between HIV-1 integrase and transcriptional coactivator p75. *Proc Natl Acad Sci U S A.* 2005; 102:17308–17313. [PubMed: 16260736]

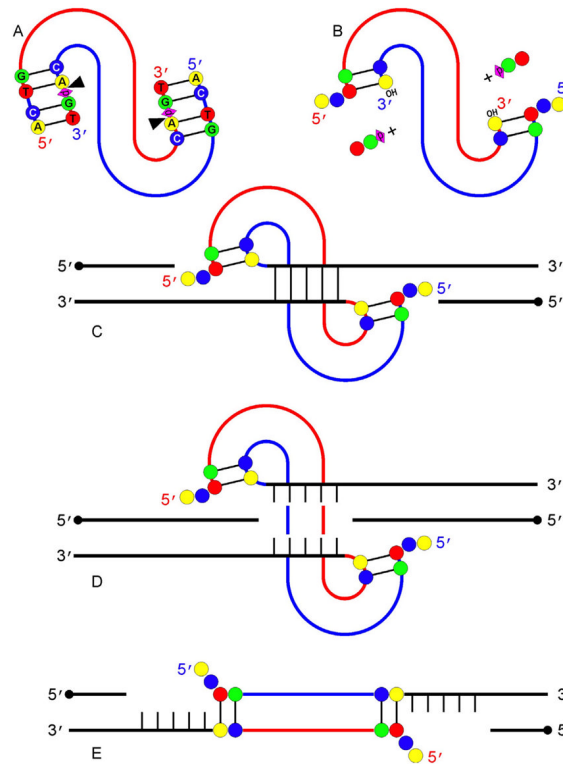


36. Engelman A, Cherepanov P. The lentiviral integrase binding protein LEDGF/p75 and HIV-1 replication. *PLoS Pathog.* 2008; 4:e1000046. [PubMed: 18369482]
37. Bushman FD, Wang B. Rous sarcoma virus integrase protein: mapping functions for catalysis and substrate binding. *J Virol.* 1994; 68:2215–2223. [PubMed: 8139006]
38. Katz RA, Skalka AM. A C-terminal domain in the avian sarcoma-leukosis virus pol gene product is not essential for viral replication. *J Virol.* 1988; 62:528–533. [PubMed: 2447290]
39. Rhee SY, Liu TF, Kiuchi M, Zioni R, Gifford RJ, Holmes SP, Shafer RW. Natural variation of HIV-1 group M integrase: implications for a new class of antiretroviral inhibitors. *Retrovirology.* 2008; 5:74. [PubMed: 18687142]
40. Dyda F, Hickman AB, Jenkins TM, Engelman A, Craigie R, Davies DR. Crystal structure of the catalytic domain of HIV-1 integrase: Similarity to other polynucleotidyl transferases. *Science.* 1994; 266:1981–1986. [PubMed: 7801124]
41. Jenkins TM, Hickman AB, Dyda F, Ghirlando R, Davies DR, Craigie R. Catalytic domain of human immunodeficiency virus type 1 integrase: Identification of a soluble mutant by systematic replacement of hydrophobic residues. *Proc Natl Acad Sci USA.* 1995; 92:6057–6061. [PubMed: 7597080]
42. Bujacz G, Alexandratos J, Zhou-Liu Q, Clement-Mella C, Wlodawer A. The catalytic domain of human immunodeficiency virus integrase: ordered active site in the F185H mutant. *FEBS Lett.* 1996; 398:175–178. [PubMed: 8977101]
43. Goldgur Y, Craigie R, Cohen GH, Fujiwara T, Yoshinaga T, Fujishita T, Sugimoto H, Endo T, Murai H, Davies DR. Structure of the HIV-1 integrase catalytic domain complexed with an inhibitor: A platform for antiviral drug design. *Proc Natl Acad Sci USA.* 1999; 96:13040–13043. [PubMed: 10557269]
44. Wang JY, Ling H, Yang W, Craigie R. Structure of a two-domain fragment of HIV-1 integrase: implications for domain organization in the intact protein. *EMBO J.* 2001; 20:7333–7343. [PubMed: 11743009]
45. Chen JC, Krucinski J, Miercke LJ, Finer-Moore JS, Tang AH, Leavitt AD, Stroud RM. 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; 97:8233–8238. [PubMed: 10890912]
46. Chen Z, Yan Y, Munshi S, Li Y, Zugay-Murphy J, Xu B, Witmer M, Felock P, Wolfe A, Sardana V, Emini EA, Hazuda D, Kuo LC. 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; 296:521–533. [PubMed: 10669606]
47. Kulkosky J, Katz RA, Merkel G, Skalka AM. Activities and substrate specificity of the evolutionarily conserved central domain of retroviral integrase. *Virology.* 1995; 206:448–456.
48. Goldgur Y, Dyda F, Hickman AB, Jenkins TM, Craigie R, Davies DR. Three new structures of the core domain of HIV-1 integrase: an active site that binds magnesium. *Proc Natl Acad Sci USA.* 1998; 95:9150–9154. [PubMed: 9689049]
49. Maignan S, Guilloteau JP, Zhou-Liu Q, Clement-Mella C, Mikol V. 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; 282:359–368. [PubMed: 9735293]
50. Greenwald J, Le V, Butler SL, Bushman FD, Choe S. The mobility of an HIV-1 integrase active site loop is correlated with catalytic activity. *Biochemistry.* 1999; 38:8892–8898. [PubMed: 10413462]
51. Molteni V, Greenwald J, Rhodes D, Hwang Y, Kwiatkowski W, Bushman FD, Siegel JS, Choe S. Identification of a small-molecule binding site at the dimer interface of the HIV integrase catalytic domain. *Acta Crystallogr.* 2001; D57:536–544.
52. Bujacz G, Jaskólski M, Alexandratos J, Wlodawer A, Merkel G, Katz RA, Skalka AM. High resolution structure of the catalytic domain of the avian sarcoma virus integrase. *J Mol Biol.* 1995; 253:333–346. [PubMed: 7563093]
53. Bujacz G, Jaskólski M, Alexandratos J, Wlodawer A, Merkel G, Katz RA, Skalka AM. The catalytic domain of avian sarcoma virus integrase: conformation of the active-site residues in the presence of divalent cations. *Structure.* 1996; 4:89–96. [PubMed: 8805516]

54. Bujacz G, Alexandratos J, Wlodawer A, Merkel G, Andrade M, Katz RA, Skalka AM. Binding of different divalent cations to the active site of avian sarcoma virus integrase and their effects on enzymatic activity. *J Biol Chem.* 1997; 272:18161–18168. [PubMed: 9218451]
55. Lubkowski J, Yang F, Alexandratos J, Merkel G, Katz RA, Gravuer K, Skalka AM, Wlodawer A. Structural basis for inactivating mutations and pH-dependent activity of avian sarcoma virus integrase. *J Biol Chem.* 1998; 273:32685–32689. [PubMed: 9830010]
56. Lubkowski J, Yang F, Alexandratos J, Wlodawer A, Zhao H, Burke TR Jr, Neamati N, Pommier Y, Merkel G, Skalka AM. Structure of the catalytic domain of avian sarcoma virus integrase with a bound HIV-1 integrase-targeted inhibitor. *Proc Natl Acad Sci USA.* 1998; 95:4831–4836. [PubMed: 9560188]
57. Lubkowski J, Dauter Z, Yang F, Alexandratos J, Merkel G, Skalka AM, Wlodawer A. Atomic resolution structures of the core domain of avian sarcoma virus integrase and its D64N mutant. *Biochemistry.* 1999; 38:13512–13522. [PubMed: 10521258]
58. Valkov E, Gupta SS, Hare S, Helander A, Roversi P, McClure M, Cherepanov P. Functional and structural characterization of the integrase from the prototype foamy virus. *Nucleic Acids Res.* 2009; 37:243–255. [PubMed: 19036793]
59. Hare S, Shun M-C, Gupta SS, Valkov E, Engelman A, Cherepanov P. A novel co-crystal structure affords the design of gain-of-function lentiviral integrase mutants in the presence of modified PSIP1/LEDGF/p75. *PLOS Pathogens.* 2009 In press.
60. Yang ZN, Mueser TC, Bushman FD, Hyde CC. Crystal structure of an active two-domain derivative of Rous sarcoma virus integrase. *J Mol Biol.* 2000; 296:535–548. [PubMed: 10669607]
61. Heuer TS, Brown PO. 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; 36:10655–10665. [PubMed: 9271496]
62. Brese NE, O'Keeffe M. Bond-valence parameters for solids. *Acta Crystallogr.* 1991; B47:192–197.
63. Marinello J, Marchand C, Mott BT, Bain A, Thomas CJ, Pommier Y. Comparison of raltegravir and elvitegravir on HIV-1 integrase catalytic reactions and on a series of drug-resistant integrase mutants. *Biochemistry.* 2008; 47:9345–9354. [PubMed: 18702518]
64. Di Santo R, Costi R, Roux A, Miele G, Crucitti GC, Iacovo A, Rosi F, Lavecchia A, Marinelli L, Di Giovanni C, Novellino E, Palmisano L, Andreotti M, Amici R, Galluzzo CM, Nencioni L, Palamara AT, Pommier Y, Marchand C. Novel quinolinonyl diketo acid derivatives as HIV-1 integrase inhibitors: design, synthesis, and biological activities. *J Med Chem.* 2008; 51:4744–4750. [PubMed: 18646746]
65. Camarasa MJ, Velazquez S, San Felix A, Perez-Perez MJ, Gago F. Dimerization inhibitors of HIV-1 reverse transcriptase, protease and integrase: a single mode of inhibition for the three HIV enzymes? *Antiviral Res.* 2006; 71:260–267. [PubMed: 16872687]
66. Lutzke RA, Eppens NA, Weber PA, Houghten RA, Plasterk RH. Identification of a Hexapeptide Inhibitor of the Human Immunodeficiency Virus Integrase Protein by Using a Combinatorial Chemical Library. *Proc Natl Acad Sci USA.* 1995; 92:11456–11460. [PubMed: 8524782]
67. Sourgen F, Maroun RG, Frere V, Bouziane M, Auclair C, Troalen F, Fermandjian S. A synthetic peptide from the human immunodeficiency virus type-1 integrase exhibits coiled-coil properties and interferes with the in vitro integration activity of the enzyme. Correlated biochemical and spectroscopic results. *Eur J Biochem.* 1996; 240:765–773. [PubMed: 8856082]
68. Maroun RG, Gayet S, Benleulmi MS, Porumb H, Zargarian L, Merad H, Leh H, Mouscadet JF, Troalen F, Fermandjian S. Peptide inhibitors of HIV-1 integrase dissociate the enzyme oligomers. *Biochemistry.* 2001; 40:13840–13848. [PubMed: 11705373]
69. Zhao L, O'Reilly MK, Shultz MD, Chmielewski J. Interfacial peptide inhibitors of HIV-1 integrase activity and dimerization. *Bioorg Med Chem Lett.* 2003; 13:1175–1177. [PubMed: 12643937]
70. Hayouka Z, Rosenbluh J, Levin A, Loya S, Lebendiker M, Vepintsev D, Kotler M, Hizi A, Loyter A, Friedler A. Inhibiting HIV-1 integrase by shifting its oligomerization equilibrium. *Proc Natl Acad Sci USA.* 2007; 104:8316–8321. [PubMed: 17488811]
71. Ferro S, De Luca L, Barreca ML, Iraci N, De Grazia S, Christ F, Witvrouw M, Debyser Z, Chimirri A. Docking studies on a new human immunodeficiency virus integrase-Mg-DNA

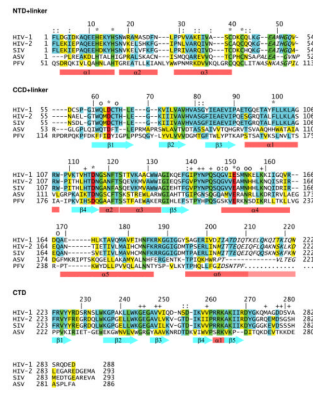
- complex: phenyl ring exploration and synthesis of 1H-benzylindole derivatives through fluorine substitutions. *J Med Chem.* 2009; 52:569–573. [PubMed: 19105658]
72. Jegede O, Babu J, Di Santo R, McColl DJ, Weber J, Quinones-Mateu M. HIV type 1 integrase inhibitors: from basic research to clinical implications. *AIDS Rev.* 2008; 10:172–189. [PubMed: 18820719]
  73. Pace P, Rowley M. Integrase inhibitors for the treatment of HIV infection. *Curr Opin Drug Discov Devel.* 2008; 11:471–479.
  74. Al Mawsawi LQ, Al Safi RI, Neamati N. Anti-infectives: clinical progress of HIV-1 integrase inhibitors. *Expert Opin Emerg Drugs.* 2008; 13:213–225. [PubMed: 18537517]
  75. Cai M, Zheng R, Caffrey M, Craigie R, Clore GM, Gronenborn AM. Solution structure of the N-terminal zinc binding domain of HIV-1 integrase. *Nature Struct Biol.* 1997; 4:567–577. [PubMed: 9228950]
  76. Eijkelenboom AP, van den Ent FM, Wechselberger R, Plasterk RH, Kaptein R, Boelens R. Refined solution structure of the dimeric N-terminal HHCC domain of HIV-2 integrase. *J Biomol NMR.* 2000; 18:119–128. [PubMed: 11101216]
  77. Cai M, Huang Y, Caffrey M, Zheng R, Craigie R, Clore GM, Gronenborn AM. Solution structure of the His12 --> Cys mutant of the N-terminal zinc binding domain of HIV-1 integrase complexed to cadmium. *Protein Sci.* 1998; 7:2669–2674. [PubMed: 9865962]
  78. Eijkelenboom AP, Lutzke RA, Boelens R, Plasterk RH, Kaptein R, Hård K. The DNA-binding domain of HIV-1 integrase has an SH3-like fold. *Nature Struct Biol.* 1995; 2:807–810. [PubMed: 7552753]
  79. Eijkelenboom AP, van den Ent FM, Vos A, Doreleijers JF, Hård K, Tullius TD, Plasterk RH, Kaptein R, Boelens R. The solution structure of the amino-terminal HHCC domain of HIV-2 integrase: a three-helix bundle stabilized by zinc. *Curr Biol.* 1997; 7:739–746. [PubMed: 9368756]
  80. Lodi PJ, Ernst J, Kuszewski J, Hickman AB, Engelman A, Craigie R, Clore GM, Gronenborn AM. Solution structure of the DNA binding domain of HIV-1 integrase. *Biochemistry.* 1995; 34:9826–9833. [PubMed: 7632683]
  81. Eijkelenboom AP, Sprangers R, Hard K, Puras Lutzke RA, Plasterk RH, Boelens R, Kaptein R. Refined solution structure of the C-terminal DNA-binding domain of human immunovirus-1 integrase. *Proteins.* 1999; 36:556–564. [PubMed: 10450096]
  82. Merad H, Porumb H, Zargarian L, Rene B, Hobaika Z, Maroun RG, Mauffret O, Femandjian S. An unusual helix turn helix motif in the catalytic core of HIV-1 integrase binds viral DNA and LEDGF. *PLoS ONE.* 2009; 4:e4081. [PubMed: 19119323]
  83. Cherepanov P, Devroe E, Silver PA, Engelman A. Identification of an evolutionarily conserved domain in human lens epithelium-derived growth factor/transcriptional co-activator p75 (LEDGF/p75) that binds HIV-1 integrase. *J Biol Chem.* 2004; 279:48883–48892. [PubMed: 15371438]
  84. Turlure F, Maertens G, Rahman S, Cherepanov P, Engelman A. A tripartite DNA-binding element, comprised of the nuclear localization signal and two AT-hook motifs, mediates the association of LEDGF/p75 with chromatin in vivo. *Nucleic Acids Res.* 2006; 34:1653–1675. [PubMed: 16549878]
  85. Heuer TS, Brown PO. Photo-cross-linking studies suggest a model for the architecture of an active human immunodeficiency virus type 1 integrase-DNA complex. *Biochemistry.* 1998; 37:6667–6678. [PubMed: 9578550]
  86. Deprez E, Tauc P, Leh H, Mouscadet JF, Auclair C, Brochon JC. Oligomeric states of the HIV-1 integrase as measured by time-resolved fluorescence anisotropy. *Biochemistry.* 2000; 39:9275–9284. [PubMed: 10924120]
  87. Baranova S, Tuzikov FV, Zakharova OD, Tuzikova NA, Calmels C, Litvak S, Tarrago-Litvak L, Parissi V, Nevinsky GA. Small-angle X-ray characterization of the nucleoprotein complexes resulting from DNA-induced oligomerization of HIV-1 integrase. *Nucleic Acids Res.* 2007; 35:975–987. [PubMed: 17259219]
  88. Bao KK, Wang H, Miller JK, Erie DA, Skalka AM, Wong I. Functional oligomeric state of avian sarcoma virus integrase. *J Biol Chem.* 2003; 278:1323–1327. [PubMed: 12446721]
  89. Ren G, Gao K, Bushman FD, Yeager M. Single-particle image reconstruction of a tetramer of HIV integrase bound to DNA. *J Mol Biol.* 2007; 366:286–294. [PubMed: 17157316]

90. Leh H, Brodin P, Bischerour J, Deprez E, Tauc P, Brochon JC, LeCam E, Coulaud D, Auclair C, Mouscadet JF. Determinants of Mg<sup>2+</sup>-dependent activities of recombinant human immunodeficiency virus type 1 integrase. *Biochemistry*. 2000; 39:9285–9294. [PubMed: 10924121]
91. Keseru GM, Kolossvary I. Fully flexible low-mode docking: application to induced fit in HIV integrase. *J Am Chem Soc*. 2001; 123:12708–12709. [PubMed: 11741448]
92. Dayam R, Neamati N. Active site binding modes of the beta-diketoacids: a multi-active site approach in HIV-1 integrase inhibitor design. *Bioorg Med Chem*. 2004; 12:6371–6381. [PubMed: 15556755]
93. Schemes JR, Henchman RH, Siegel JS, Sotriffer CA, Ni H, McCammon JA. Discovery of a novel binding trench in HIV integrase. *J Med Chem*. 2004; 47:1879–1881. [PubMed: 15055986]
94. Savarino A. In-Silico docking of HIV-1 integrase inhibitors reveals a novel drug type acting on an enzyme/DNA reaction intermediate. *Retrovirology*. 2007; 4:21. [PubMed: 17374162]
95. Chen X, Tsiang M, Yu F, Hung M, Jones GS, Zeynalzadegan A, Qi X, Jin H, Kim CU, Swaminathan S, Chen JM. Modeling, analysis, and validation of a novel HIV integrase structure provide insights into the binding modes of potent integrase inhibitors. *J Mol Biol*. 2008; 380:504–519. [PubMed: 18565342]
96. Gao K, Butler SL, Bushman F. Human immunodeficiency virus type 1 integrase: arrangement of protein domains in active cDNA complexes. *EMBO J*. 2001; 20:3565–3576. [PubMed: 11432843]
97. Chen A, Weber IT, Harrison RW, Leis J. 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; 281:4173–4182. [PubMed: 16298997]
98. Dolan J, Chen A, Weber IT, Harrison RW, Leis J. Defining the DNA Substrate Binding Sites on HIV-1 Integrase. *J Mol Biol*. 2008
99. Jayaram M. The cis-trans paradox of integrase. *Science*. 1997; 276:49–51. [PubMed: 9122709]
100. Davies DR, Goryshin IY, Reznikoff WS, Rayment I. Three-dimensional structure of the Tn5 synaptic complex transposition intermediate. *Science*. 2000; 289:77–85. [PubMed: 10884228]
101. Lutzke RA, Plasterk RH. 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; 72:4841–4848. [PubMed: 9573250]
102. Steitz TA, Steitz JA. A general two-metal-ion mechanism for catalytic RNA. *Proc Natl Acad Sci USA*. 1993; 90:6498–6502. [PubMed: 8341661]
103. Nowotny M, Gaidamakov SA, Crouch RJ, Yang W. Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. *Cell*. 2005; 121:1005–1016. [PubMed: 15989951]
104. Lovell S, Goryshin IY, Reznikoff WR, Rayment I. Two-metal active site binding of a Tn5 transposase synaptic complex. *Nat Struct Biol*. 2002; 9:278–281. [PubMed: 11896402]
105. Yang W, Lee JY, Nowotny M. Making and breaking nucleic acids: two-Mg<sup>2+</sup>-ion catalysis and substrate specificity. *Mol Cell*. 2006; 22:5–13. [PubMed: 16600865]
106. Goedken ER, Marqusee S. Native-state energetics of a thermostabilized variant of ribonuclease HI. *J Mol Biol*. 2001; 314:863–871. [PubMed: 11734003]
107. DeLano, WL. The PyMOL Molecular Graphics System. DeLano Scientific; San Carlos, CA: 2002.
108. Nowotny M, Yang W. Stepwise analyses of metal ions in RNase H catalysis from substrate destabilization to product release. *EMBO J*. 2006; 25:1924–1933. [PubMed: 16601679]



**Figure 1. A schematic representation of the reaction catalyzed by retroviral IN during an infection cycle**

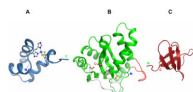
This example shows the activity of HIV-1 IN. The reaction catalyzed by enzymes from other retroviruses may differ in some details, but the general scheme is the same. In the **processing** step (A→B), the 3' ends of viral DNA (colored molecule) are nicked (arrowheads) before the phosphate group (diamond) of the conserved terminal GT dinucleotide (colored beads, A=yellow, C=blue, G=green, T=red), leading to a DNA molecule with a 5' overhang and a free 3' OH group on each strand. In the **joining** step (B→C), host DNA (black) is nicked with a five-nucleotide stagger (vertical bars) on the two strands, and the free 3' ends of the viral substrate are joined to both host strands, preserving DNA polarity. Panels D and E are equivalent to C and are presented to illustrate the topology of the final DNA product (not shown), which is created from molecule E by cellular DNA repair enzymes, which remove the overhanging viral 5' dinucleotides and seal the gaps on both sides of the integrated viral DNA. In the final product, the viral insert is flanked by the repeated stagger sequence, and begins with the conserved TG sequence at each 5' end.



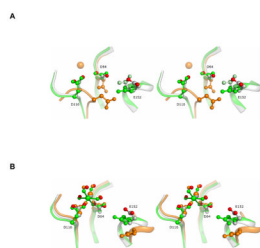
**Figure 2. Amino-acid sequence alignment of retroviral integrases**

The secondary structure of HIV-1 IN is shown below the sequences ( $\alpha$  helices marked as cylinders,  $\beta$  strands as arrows). Symbols:

- Green** All residues identical
- Blue** At least 3 residues identical
- Yellow** Similar residues
- Red** Active site residues
- \*** Metal cation binding
- :** Structurally important
- +** DNA binding
- o** Inhibitor binding



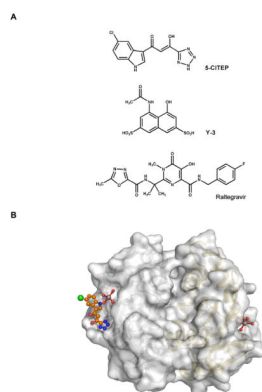
**Figure 3. The structures of the monomers of individual domains of HIV-1 IN**  
(A) The NTD domain (blue) with a  $Zn^{2+}$  cation (large sphere) coordinated (thin lines) by an HHCC motif (ball-and-stick) of a helix-turn-helix (HTH) fold, is represented by the NMR structure 1WJC [75]. (B) The CCD domain (green), shown with the D,D(35)E catalytic residues (ball-and-stick), a magnesium cation (large sphere) coordinated in site I and the flexible active-site loop highlighted in gray, is represented by the crystal structure 1BL3 [49]. The finger loop (red) extrudes from the body of the protein on the right, between helices  $\alpha 5$  and  $\alpha 6$  (C terminus). (C) The CTD domain (red) is represented by the NMR structure 1IHV [80]. This and all subsequent figures were prepared with PyMOL [107].



**Figure 4. The active site of retroviral integrases**

The figures show, in stereoview, the three essential acids of the D,D(35)E motif in selected, least-squares-superposed crystallographic structures of the catalytic core domain in (A) unliganded and (B) Mg-complexed form. The catalytic residues are shown in the context of the protein secondary structure by which they are contributed, namely an extended  $\beta$ -ribbon (the first aspartate, middle of figure), a loop (the second aspartate, left), and an  $\alpha$ -helix (the glutamate, right). The residue numbering D64/D116/E152 is for the HIV-1 IN sequence, and corresponds to D64/D121/E157 in ASV IN. The three divalent-metal-cation-free active sites shown in (A) correspond to the first HIV-1 IN structure (1ITG, orange) [40], solved in the presence of arsenic (part of cacodylate buffer), which reacted with cysteine residues, including one within the active site area (orange sphere), to another medium-resolution structure of HIV-1 IN (1BI4, molecule C, gray with red O atoms) [49], and to atomic-resolution structure of ASV IN (1CXQ, green) [57]. Note that the aspartates in 1ITG have a completely different orientation than in the remaining structures and the entire Asp116 loop has a different, non-native conformation. Another symptom of active-site disruption in the 1ITG structure is the absence in the model of the Glu152 residue, a consequence of disorder in this helical segment. The active sites complexed with the catalytic cofactor  $Mg^{2+}$  (large sphere) are shown (B) for HIV-1 IN, 1BL3 (molecule C, gray with red O atoms) [49], ASV IN, 1VSD (green) [53], and for PFV IN, molecule A of 3DLR (orange) [58]. The structure of the ASV IN has the highest resolution, and its quality is reflected in the nearly ideal octahedral geometry (thin green lines) of the  $Mg^{2+}$  coordination sphere, which in addition to interactions with the carboxylate groups of both active-site aspartates includes four precisely defined water molecules. The coordination geometry of the HIV-1 IN complex 1BL3 is significantly distorted. The view direction in both figures is similar, with a small rotation around the horizontal axis.

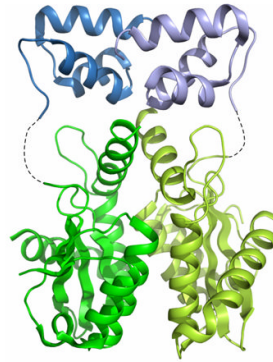




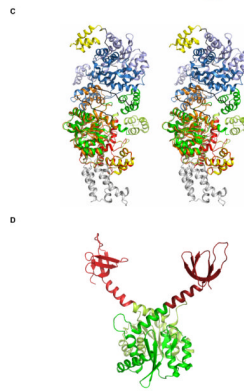
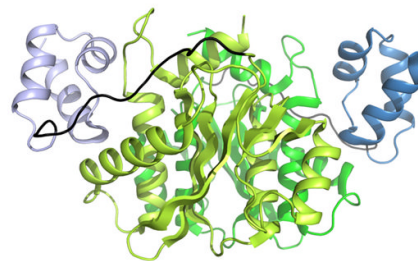
**Figure 5. Small-molecule inhibitors of the CCD domain of retroviral IN**

(A) Chemical diagrams of selected inhibitors discussed in this review. (B) A dimer of the CCD domains (colored silver and gold) of HIV-1 IN shown in surface representation roughly down its two-fold axis. The two active sites are marked by the Mg<sup>2+</sup> cations (gray spheres), with their octahedral coordination spheres formed by the carboxylates of Asp64 and Asp116, and by four water molecules (red spheres). Note that the active sites are located in shallow depressions on the surface of the protein, with the magnesium cations completely exposed to solvent. Next to the active site, a long groove is running on the surface of the protein. In this structure with the PDB code 1QS4 [43], one of the active site grooves is occupied by the 5-CITEP inhibitor, depicted here in ball-and-stick representations with C/N/O/Cl atoms shown in orange/blue/red/green. The two active sites are separated by 40.4 Å, as measured by the distance between the Mg<sup>2+</sup> centers.

A



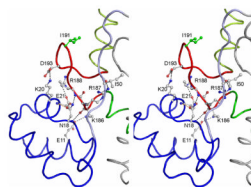
B



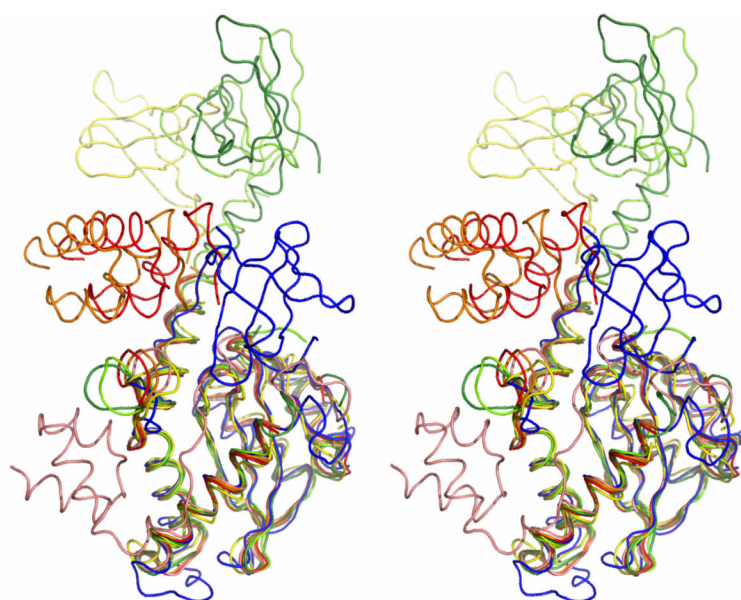
**Figure 6. Three-dimensional structures of dimeric two-domain constructs of HIV integrase determined by X-ray crystallography**

(A) In the NTD-CCD structure (1K6Y, [44]), the linker between the domains is disordered and the speculative NTD (blue) - CCD (green) pairing (dashed line) has been proposed from indirect reasoning, such as the existence of contacts between the NTD domains. (B) Mutual orientation of the NTD (blue) and CCD (green) domains as found experimentally in the structure 3F9K of HIV-2 IN [59]. The linker, visible in molecule A, is shown in black. (C) This stereoview has been constructed by least-squares superposition of the CCD domains of molecules A (red) and B (orange) from the 2:1 complex of HIV-2 IN (3F9K) with the Integrase Binding Domain (IBD) of the LEDGF protein (molecule C, gray), onto the CCDs of molecules A (limon) and B (green) of HIV-1 IN (1K6Y). Note that the smaller, all-helical NTDs of the HIV-1 protein are lifted above (in this view, shooting to the right) the CCD domains, while in the model of HIV-2 IN they “fold back” and adhere to the sides of the CCD dimer. The linkers connecting the NTD and CCD are not present in any of the experimental models shown in this figure, except in molecule A (red) of 3F9K, for which clear electron density allowed to unambiguously connect the domains. The asymmetric unit

of the 1K6Y structure contains another NTD-CCD dimer, here represented in shades of blue. Note that the blue NTD (of molecule D) superposes exactly on the NTD of molecule B (orange) of the HIV-2 NTD-CCD dimer. This unexpected match is a strong indication that, with missing experimental evidence, the pairing of the NTD and CCD domains in the 1K6Y structure does not correspond to the functional conformation of the protein. In a fashion similar to the “blue” NTD (chain D) – “green” CCD (chain B) pairing, also the “green” NTD (chain B) and “blue” CCD (chain D) domains of HIV-1 integrase can be assembled. By generating a symmetry-related copy of the “limon” (chain A) NTD domain (upper yellow), one can complete the entire NTD-CCD dimer of HIV-1 integrase with the blue catalytic cores. Likewise, a crystallographic copy of the NTD domain attributed to chain C (bottom yellow), will complete the HIV-1 integrase with the green catalytic cores. To guide the eye in this complicated view, the missing connections between the NTD and CCD domains have been generated by copying the linker chain from molecule A of the 3F9K model and grafting it into the remaining molecules (black C $\alpha$  traces). In this way, four functional HIV NTD-CCD molecules, assembled into two dimers, have been generated. (D) In the CCD-CTD dimer (1EX4, [45]), the interdomain linker forms a long helix. Because of different degree of deformation of this helix, the relative orientation between the CCD (green) and CTD (red) domains in the two monomers is different. All the NTD-CCD dimers considered in this figure have essentially a common two-fold axis for both domains. This is not true for the CCD-CTD construct (D), where the two-fold axes relating the CCD and CTD domains are not identical.

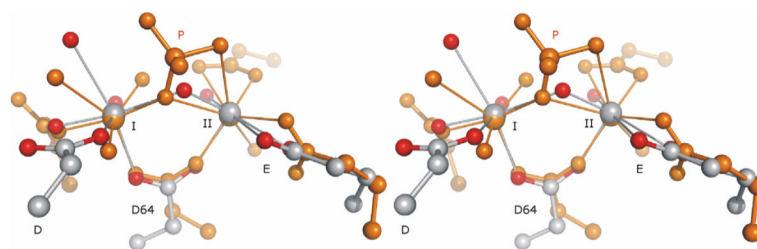


**Figure 7. Stereoview of the NTD docking site at the “finger” structure of the CCD domain**  
 The NTD (blue) is shown with its CCD domain, as found in the LEDGF complex of the HIV-2 protein (PDB code 3F9K, chain A) [59]. The “finger” loop (red), which is located between helices  $\alpha 5$  (green) and  $\alpha 6$  (light green) of the CCD domain, forms five hydrogen bonds/salt bridges (broken line) with the NTD domain, and another one with the linker peptide (light blue) connecting the domains. One of those interactions would require a flip of the side-chain amide group of Asn18, near the entry to an  $\alpha$  helix (upper left) of the NTD domain. However, such a flip would create another impossible NH...HN interaction (orange) at the N-terminus of this helix. The tip of the finger loop is occupied by an isoleucine residue (green ball-and-stick model).



**Figure 8. Stereoview of a structural superposition of several two-domain constructs of retroviral integrases**

The superpositions were calculated using only the C $\alpha$  atoms of the CCD domain (bottom) to show possible mutual orientations of all three domains. Until the structure of intact IN is determined experimentally, this is the best approximation of the three-dimensional model of the enzyme, here shown only for the monomeric molecule. According to available data on dimeric structure of IN domains, a homodimer of IN could be created by rotating the above model by 180° around the vertical line and placing it face-to-face with the original copy, so as to recreate the dimeric interface at the flat face (back of the view) of the CCD domain. The figure uses the following color code: red and orange, molecules A and B of the HIV-1 NTD-CCD protein 1K6Y [44]; salmon, molecule A of the HIV-2 NTD-CCD protein 3F9K [59]; blue, ASV CCD-CTD protein 1C0M [60]; dark/light green, molecules A and B of the HIV-1 CCD-CTD protein 1EX4 [45]; yellow, SIV CCD-CTD protein 1C6V [46]. In the 1C6V structure, the domains that are displayed (D and X) were not interpreted as a single molecule in the original publication. It is evident that the CTD, as represented in this figure (yellow-green-blue colors), covers a wide angular range of its disposition relative to CCD. Note that the NTDs from the 1K6Y structure (red and orange) were linked by the authors to the CCDs without experimental evidence. In a different interpretation of the 1K6Y crystal structure, it is possible to select an NTD partner for the CCD that essentially superposes on the salmon model from the 3F9K structure (see Fig. 6C), which occupies this conformation without any ambiguity as it can be traced *via* an uninterrupted connection to the CCD domain. It is thus very likely that in variance to the CTD domain, the NTD domain has a fixed orientation relative to CCD.



### Figure 9. Divalent metal cation binding sites in integrases

Comparison of the two metal sites found in  $\text{Cd}^{2+}$  complex of ASV IN (PDB code 1VSJ, color code according to element type) [54] and in  $\text{Mg}^{2+}$ /RNA/DNA complex of RNase H (1ZBL, molecule A, orange) [103]. The active sites were superposed by a simple least-squares fit of the metal sites and the carboxylate O atoms of the bridging aspartate residue, which in ASV IN is the first element (Asp64) of the D,D(35)E active site. The metal...metal distance in both structures is nearly identical, 4.10 Å in the RNase H complex and 4.05 Å in the IN complex. Site I, denoted A in [103,108], which in retroviral integrase structures was also seen with the catalytic  $\text{Mg}^{2+}$  or  $\text{Mn}^{2+}$  cations, has a more regular octahedral coordination. The coordination spheres I in the two structures are similar, except that two ligands in the equatorial plane, an aspartate (Asp121 in ASV IN) O $\delta$  atom and a water molecule, have swapped places (in 1ZBL, the aspartate in question has been mutated to asparagine, D192N). Another important difference is that the bridging water molecule in the ASV IN complex is replaced in the RNase H complex by an O atom from the scissile phosphate group (P) of the RNA substrate. The phosphate group has a particularly important role in the formation of site II, as it provides two of the ligands. Site II (denoted B in [103,108]) has a far less regular geometry; the coordination sphere is incomplete in 1VSJ or highly distorted in 1ZBL. Site II has never been seen occupied by  $\text{Mg}^{2+}$  or  $\text{Mn}^{2+}$  in metal complex structures of retroviral INs, and the glutamic acid which takes part in its formation (Glu157 in ASV IN) is the most mobile element of the IN active site. From this comparison, it is very likely that a proper site II of retroviral integrase, occupied by a catalytically-competent metal cation ( $\text{Mg}^{2+}$  or  $\text{Mn}^{2+}$ ), could only be formed with the participation of a DNA substrate.