Article No. jmbi.1999.3463 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2000) 296, 535–548

JMB

Crystal Structure of an Active Two-domain Derivative of Rous Sarcoma Virus Integrase

Zhong-Ning Yang[1]†, Timothy C. Mueser[1]†, Frederic D. Bushman[2] and C. Craig Hyde[1]*

[1]*Laboratory of Structural Biology Research, National Institute of Arthritis Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, MD 20892, USA*

[2]*Infectious Disease Laboratory The Salk Institute, La Jolla, CA USA*

**Corresponding author*

Integration of retroviral cDNA is a necessary step in viral replication. The virally encoded integrase protein and DNA sequences at the ends of the linear viral cDNA are required for this reaction. Previous studies revealed that truncated forms of Rous sarcoma virus integrase containing two of the three protein domains can carry out integration reactions *in vitro*. Here, we describe the crystal structure at 2.5 Å resolution of a fragment of the integrase of Rous sarcoma virus (residues 49-286) containing both the conserved catalytic domain and a modulatory DNA-binding domain (C domain). The catalytic domains form a symmetric dimer, but the C domains associate asymmetrically with each other and together adopt a canted conformation relative to the catalytic domains. A binding path for the viral cDNA is evident spanning both domain surfaces, allowing modeling of the larger integration complexes that are known to be active *in vivo*. The modeling suggests that formation of an integrase tetramer (a dimer of dimers) is necessary and sufficient for joining both viral cDNA ends at neighboring sites in the target DNA. The observed asymmetric arrangement of C domains suggests that they could form a rotationally symmetric tetramer that may be important for bridging integrase complexes at each cDNA end.

Keywords: HIV; integrase; protein structure; Rous sarcoma virus; protein X-ray crystallography

Introduction

Following infection by a retrovirus, the reverse-transcribed cDNA genome must be integrated into a host chromosome for replication to proceed (for reviews, see Coffin *et al*., 1997; Hansen *et al*., 1998). The DNA cutting and joining reactions mediating cDNA integration have been characterized in detail (Figure 1(a) and (b)). Prior to integration, two nucleotides are removed from each cDNA 3′ end (Brown *et al*., 1989; Fujiwara & Mizuuchi, 1988), a reaction that may serve to remove heterogeneous extra nucleotides occasionally added by reverse transcriptase (Miller *et al*., 1997; Patel & Preston, 1994). The recessed 3′ DNA ends are then transferred by integrase to the host target DNA (Brown *et al*., 1987, 1989; Fujiwara & Mizuuchi, 1988). Integration of both viral ends at sites in opposite strands of the target DNA is termed the ''coupled'' or ''concerted'' integration reaction (Figure 1(b)). Each retrovirus displays a characteristic spacing between points of joining, six base-pairs in the case of Rous sarcoma virus (RSV). This intermediate is then processed, presumably by host DNA repair enzymes, to yield a fully integrated provirus.

Purified integrases can remove two nucleotides from the 3′ ends of model substrate DNAs (Bushman & Craigie, 1991; Craigie *et al*., 1990; Katzman *et al*., 1989; Sherman & Fyfe, 1990), and direct the joining of the cDNA ends into target DNA *in vitro* (Bushman *et al*., 1990; Craigie *et al*., 1990; Katz *et al*., 1990). Integrase or its catalytically active fragments can carry out an apparent reversal of the joining step, termed ''disintegration'' (Chow *et al*., 1992) (Figure 1(a)).

Retroviral integrases comprise three protein domains. The amino-terminal domains (N domains) promote DNA binding, enzyme multimerization or both. The larger central domain (amino acid residues 50-217) contains the catalytic

†These authors contributed equally to the work.
E-mail address of the corresponding author: Craig_Hyde@nih.gov

0022-2836/00/020535–14 $35.00/0

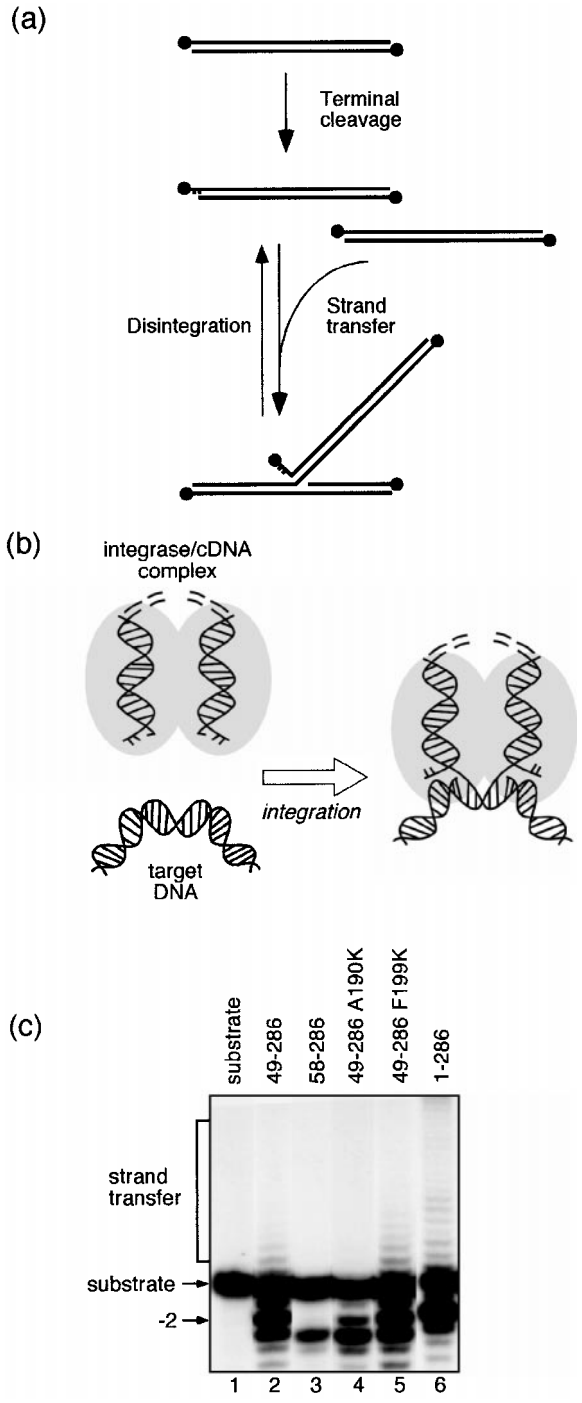

Figure 1. Activities of RSV integrase. (a) Terminal cleavage and strand transfer reactions. DNA 5′ ends are shown as filled circles. The duplex oligonucleotide studied (FB141/FB142) matches the sequence of one end of the unintegrated RSV cDNA. Note that ''disintegration'' reactions employ a Y-shaped substrate resembling the product of integration. (b) Diagram of coupled joining of the two viral cDNA ends to a host target DNA (adapted from Heuer & Brown (1998). A pre-integration complex is shown at the top composed of two cDNA ends and proteins including integrase (gray). Joining of each of the two viral 3′ cDNA ends by integrase produces the coupled integration intermediate shown at the right. The sites of integration on opposing DNA strands are separated by six base-pairs in the case of avian integrases like that from Rous sarcoma virus, and by five base-pairs in the case of HIV integrase. Replication of virus depends critically on both ends of its DNA being integrated correctly in this manner. (c) Autoradiogram of a gel containing products of terminal cleavage and strand transfer reactions directed by the indicated mutant proteins: 1-286 is recombinantly expressed full-length, wild-type protein. The label −2 indicates the terminal cleavage products.

center. A fragment containing only this domain can carry out the permissive disintegration reaction (Bushman *et al.*, 1993; Bushman & Wang, 1994; Kulkosky *et al.*, 1995; Vink *et al.*, 1993). The C-terminal region of retroviral integrases is important for non-specific DNA binding and multimerization (Andrake & Skalka, 1995; Engelman *et al.*, 1994; Mumm & Grandgenett, 1991; Puras Lutzke *et al.*, 1994; van Gent *et al.*, 1991).

The structures of each domain of HIV-1 integrase and the catalytic domain of avian sarcoma virus (ASV) have been solved as isolated fragments by either crystallography or solution NMR methods. The zinc-binding amino-terminal domain (residues 1-49) comprises a three-helix bundle resembling a homeodomain, an unusual conformation for a zinc-binding protein (Cai *et al.*, 1997; Eijkelenboom *et al.*, 1997). The fold of the catalytic domain (solved crystallographically for HIV and ASV) is composed of a five-stranded α/β fold and resembles the structures of several other enzymes involved in polynucleotide phosphotransfer reactions (Yang & Steitz, 1995). The catalytic domain is proposed to dimerize with 2-fold rotational symmetry, as inferred from crystal contacts in both the HIV and ASV structures (Bujacz *et al.*, 1995; Dyda *et al.*, 1994). As a consequence of this dimerization, the two active centers are placed distant from each other, essentially on opposite sides of the complex. The carboxyl-terminal region encompasses a small domain (C domain; residues 220-270) with a five-stranded β-barrel fold resembling an SH3 (Src homology) domain (Lodi *et al.*, 1995; Plasterk, 1995). Solution NMR studies of the isolated HIV C domains propose a dimer model with 2-fold symmetry.

In the case of RSV integrase, two-domain derivatives containing only the catalytic and carboxyl-terminal domains (residues 49-286) are capable of carrying out the full terminal cleavage and strand transfer reactions (Bushman & Wang, 1994). In an effort to obtain structural information on such derivatives, we constructed four mutant forms of RSV IN 49-286, characterized their activities *in vitro*, and tested each for its propensity to crystallize. One of the mutants studied, containing the change F199K, formed diffraction-quality crystals in two space groups, allowing solution of the structure to 2.5 Å resolution.

Results and Discussion

Site-directed mutagenesis to improve crystal growth

A protein fragment containing two of the three domains (residues 49-286) of RSV integrase crystallized readily, but failed to produce diffraction-quality crystals. Therefore, we generated and characterized mutant derivatives with the goal of improving crystallization. Mutant forms were designed using the crystal structures of the closely related ASV integrase catalytic domain (amino acid residues 49-217) (Bujacz *et al.*, 1995, 1997). We mutated residue positions in a hydrophobic surface region of the protein far from the active site to introduce additional charge: mutations A190K, L196K, and F199K. Another mutant form, RSV IN 58-286, removed a region that was apparently disordered in the ASV catalytic domain structure.

Terminal cleavage and strand transfer activities of RSV IN 49-286 and derivatives

RSV IN 49-286 and mutant derivatives were over-expressed, purified and tested for *in vitro* activity using end-labeled oligonucleotide substrates. For terminal cleavage assays, two oligonucleotides were annealed to generate a duplex DNA matching one end of the unintegrated RSV cDNA (Figure 1(a), top). Terminal cleavage generates a labeled DNA strand two bases shorter than the starting substrate. Strand transfer results in the insertion of the recessed 3′ hydroxyl group into another oligonucleotide duplex serving as the integration target (Figure 1(a), bottom), generating labeled DNA strands that are longer than the starting substrate.

Full-length RSV integrase, RSV IN 49-286, and RSV IN 49-286 F199K each display terminal cleavage and strand transfer activities (Figure 1(c), lanes 2, 5, and 6). RSV IN 49-286 A190K displays little correct cleavage (at the −2 position) and no detectable strand transfer (lane 4). RSV IN 58-286 displays no correct cleavage or strand transfer (lane 3). RSV IN 49-286 L196K displayed poor solubility during purification and was not studied further. All proteins tested carry out a non-biological cleavage at the −3 position, an activity that has been seen with other mutant derivatives (Bujacz *et al.*, 1997). Evidently, some of the alterations in RSV IN 49-286 cause a mis-positioning of the substrate, resulting in ectopic cleavage.

Dumbbell disintegration activity of two-domain RSV IN constructs

The observation that some of the derivatives showed little or no correct terminal cleavage and strand transfer activity (RSV IN 58-286 and 49-286 A190K) raised the question of whether the active sites were still intact. To investigate this, RSV IN and derivatives were tested for the ability to carry out disintegration on dumbbell DNA molecules (Chow & Brown, 1994; Chow *et al.*, 1992). The disintegration reaction was previously found to be more permissive than the terminal cleavage and strand transfer reactions, allowing, for example, the demonstration of catalytic activity by isolated catalytic domains of RSV and HIV integrases (Bushman *et al.*, 1993; Bushman & Wang, 1994; Kulkosky *et al.*, 1995; Vink *et al.*, 1993). The dumbbell form of the molecule (Figure 2(a)) is convenient for studies *in vitro*, since it is formed from a single DNA chain and does not require annealing of multiple oligonucleotides. Disintegration, involving the attack of the 3′ hydroxyl group on the branch-point in the Y-shaped substrate, results in the release of the labeled DNA branch (Figure 2(b)). Here, we introduce the use of a dumbbell disintegration substrate for studies of RSV integrase. All the proteins studied were capable of carrying out disintegration on the dumbbell substrate (Figure 2(c)), indicating that the active sites were functionally intact.

Crystallization of RSV IN 49-286 F199K

The various mutant forms crystallized rapidly, generally producing only clusters of thin, stacked plates. Only mutant form F199K produced crystals suitable for diffraction experiments. Interestingly, position F199 in RSV integrase is near a position

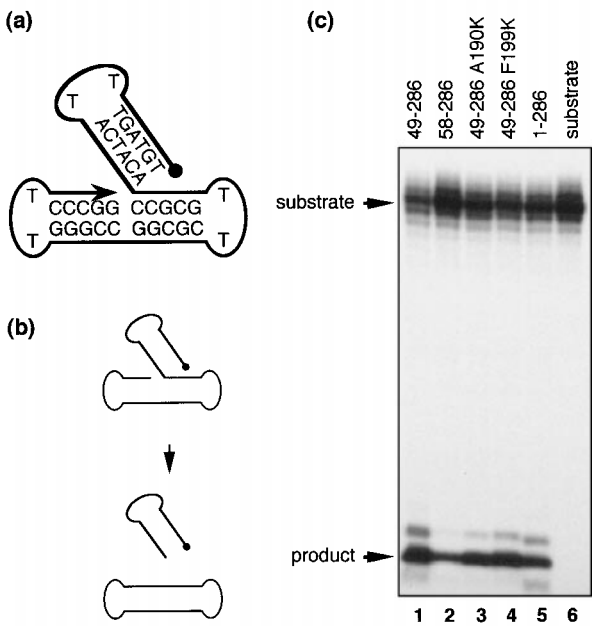

Figure 2. Dumbbell disintegration by RSV IN and mutant derivatives. (a) Sequence of the folded RSV dumbbell disintegration substrate (FB221). The 5′-end (filled circle) is radiolabeled. (b) Diagram of the dumbbell disintegration reaction. (c) Autoradiogram of a gel containing products of dumbbell disintegration with various forms of integrase. The mobilities of the substrates and products are marked.

Table 1. Statistics for data collection and structure refinement

Set	Spacegroup	Wavelength (Å)	Resolution (Å)	Unique reflections	Redundancy	Average I/σ	Completeness (%)	R_{sym} (%)[a]
MAD	$P2_1$	0.9791 (edge)	15.0-3.0	11,250 (1,080)	3.4 (2.9)	15.6 (5.5)	98.6 (96.5)	6.0 (17.7)
	$P2_1$	0.9788 (peak)	15.0-3.0	11,218 (1,085)	4.2 (3.6)	17.8 (5.5)	98.7 (97.2)	6.7 (19.0)
	$P2_1$	0.9711 (remote)	15.0-3.0	11,219 (1,076)	3.4 (2.9)	16.8 (4.2)	98.3 (96.2)	6.1 (21.4)
Native	$P1$	0.9672	15.0-2.53	28,842 (2,583)	1.8 (1.6)	18.3 (5.4)	87.6 (75.7)	3.6 (9.6)

Refinement (15.0-2.53 Å)

				r.m.s.d. from ideality				% in Ramachandran plot	
R (R_{free})[b] (%)	Protein atoms	Solvent molecules	Average B-factor (Å^2)	Bond length (Å)	Bond angle (deg.)	Dihedrals (deg.)	Improper dihedrals (deg.)	Favored	Allowed
21.6 (27.9)	6893	141	29.5	0.0075	1.22	23.0	0.82	86.9	13.0

Monoclinic crystal (spacegroup $P2_1$) $a = 66.24$ Å, $b = 46.34$ Å, $c = 94.31$ Å, $\beta = 101.76°$. Triclinic crystal (spacegroup $P1$) $a = 55.75$ Å, $b = 66.37$ Å, $c = 76.67$ Å, $\alpha = 67.47°$, $\beta = 78.61°$, $\gamma = 90.18°$.

Values in parentheses correspond to the last resolution shell; 3.1-3.0 Å for MAD data sets, or 2.62-2.53 Å for native data set.

[a] $R_{sym} = \Sigma|I_i - \langle I\rangle|/\Sigma|I_i$, where I_i is the intensity of an individual reflection, and $\langle I\rangle$ is the average intensity over symmetry equivalents of that reflection. When Bijvoet pairs were treated as separate reflections, values of R_{sym} for MAD data sets were 4-4.8%.

[b] $R = \Sigma||F_{obs}| - |F_{calc}||/\Sigma|F_{obs}|$, where F_{calc} is the calculated structure factor and summation is over data used in refinement. R_{free} is calculated for a randomly generated 5% of reflections omitted from the refinement process.

on the homologous surface of the HIV catalytic domain where the change F185K was found to improve solubility (Dyda *et al.*, 1994; Jenkins *et al.*, 1996). In an attempt to improve crystal quality, various cryo-protectants were tested. We discovered that the addition of 25% ethylene glycol to initial crystal conditions reduced nucleation, and produced two diffraction-quality crystal forms, the monoclinic ($P2_1$) and the triclinic ($P1$) crystals. Ethylene glycol improved the reproducibility of crystallization and enabled direct freezing of the crystals for cryogenic data collection. Selenomethionine-substituted protein crystallized poorly under these conditions until we discovered that the addition of 20 mM imidazole (pH 6.7) improved the quality of the monoclinic form.

Solution of the two-domain RSV integrase structure

The RSV IN 49-286 F199K structure was solved initially at 3.2 Å resolution using multiple-wavelength anomalous diffraction (MAD) phasing of a monoclinic $P2_1$ crystal containing integrase substituted with selenomethionine (Table 1). A partially refined monoclinic structure was used to solve the 2.5 Å resolution triclinic crystal form by molecular replacement. Prior attempts using molecular replacement methods were successful in locating two and four catalytic domains in the $P2_1$ and the $P1$ cells, respectively, using the crystal structure from ASV (Bujacz *et al.*, 1995). However, no C domain position could be found using either monomer or dimer models of the HIV C domain solved by NMR (Lodi *et al.*, 1995).

Overview of the RSV 49-286 F199K model

The model of the two-domain integrase is presented in Figure 3. Residues 49-213 comprise a compact globular domain similar to the catalytic domains of HIV and ASV integrase solved previously (Bujacz *et al.*, 1995; Dyda *et al.*, 1994). The RSV catalytic domains are tethered by ''linker regions'' (residues 214-219) to the C domains (residues 220-270). The fold of each C domain monomer, resembling an SH3 (Src homology) domain, is similar to that reported previously for the HIV C domain (HIV residues 220-270) determined by solution NMR methods (Eijkelenboom *et al.*, 1995; Lodi *et al.*, 1995).

The C domains associate as a tight dimer but are canted relative to the 2-fold symmetry axis of the catalytic domains (Figure 3(a)). Canted conformations are seen in both crystal forms (Figure 3(b)). The two C domains in each dimer associate in a manner much different from that proposed from NMR studies of the domain from HIV (Eijkelenboom *et al.*, 1995; Lodi *et al.*, 1995). The two polypeptide chains within each dimer can be distinguished as ''proximal'' and ''distal'', depending on whether its C domain is nearer to or more distant from its catalytic domain, respectively.

The catalytic domain

The fold of the catalytic domain and its dimer interface is similar to those observed in the related HIV (Dyda *et al.*, 1994) and closely related ASV integrase domains (Bujacz *et al.*, 1995). The catalytic domains have an α/β structure formed from a five-stranded β-sheet sandwiched between five

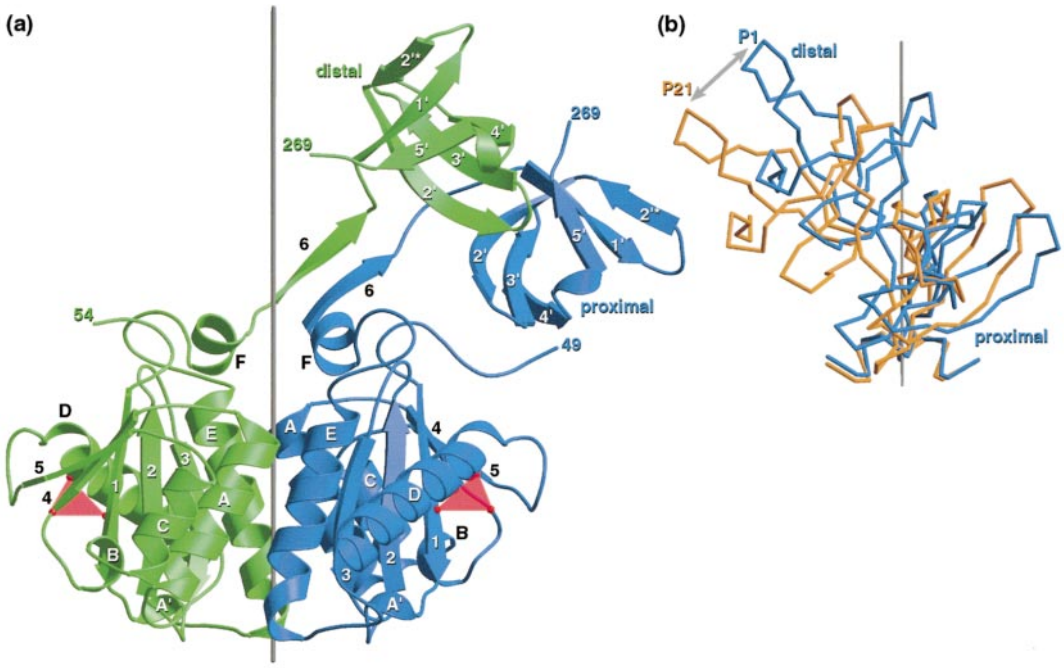

Figure 3. Crystal structure of RSV integrase 49-286. (a) Ribbon diagram of the 49-286 dimer in the triclinic crystal form. Residue numbers mark the range of observed residues at the N or C termini of each fragment. The polypeptide chains are designated proximal (blue) or distal (green). The three acidic residues (D64, D121, and E157) in each chain comprise the active center (red triangles) in the catalytic domains. Linker residues (strand 6) connect the two domains. The C domain dimer is canted by 50° (monoclinic form) and 60° (triclinic form) relative to the local 2-fold axis (gray) of the catalytic domain dimer. C domain strands are labeled 1′ through 5′; strand 2′ from a typical SH3 fold is designated here as two shorter strands, 2′* and 2′. The C-terminal residues (from about 270 to 286) are disordered in all crystal forms. (b) Canted conformations of the two-domain integrase are seen in different crystal forms. Following superposition of the triclinic (blue) and monoclinic (orange) crystal forms on the basis of the catalytic domains (not shown), the C domains differ in tilt by about 20°. The view is at a right-angle to that shown in (a). The conformation of the second dimer of the triclinic unit cell (not shown) is similar to the triclinic (blue) conformation shown. The asymmetric orientation of the C domain dimer positions one C domain much closer to its catalytic domain (the ''proximal''chain).

α-helices (Bujacz *et al.*, 1995, 1997; Dyda *et al.*, 1994; Goldgur *et al.*, 1998; Maignan *et al.*, 1998).

The catalytic centers are formed by a metal-binding cluster of three acidic residues; Asp64, Asp121, and Glu157. A nearby loop (residues 145-154) is well resolved in both distal subunits of the triclinic crystal (Figure 4). These residues are either partially or completely disordered in previous structure reports of isolated catalytic domains (Bujacz *et al.*, 1995; Dyda *et al.*, 1994; Goldgur *et al.*, 1998; Greenwald *et al.*, 1999). Like many other ''flexible loops'' in other enzymes, it may achieve a fixed conformation only when substrate is bound.

The crystal structure of the two-domain RSV protein also provides a view of the C-terminal portion of the catalytic domain of the avian retroviral enzymes. The last ten residues of the ASV construct (encoding residues 54-209) are disordered (Bujacz *et al.*, 1995). Residues 200-207 in RSV integrase form a small hairpin loop projection similar to residues 186-195 in HIV integrase (Dyda *et al.*, 1994). RSV residues 208-213 also form a short helical segment (helix F) similar to HIV residues 196-201. The side-chain of His212 in helix F forms a hydrogen-bonding network with the side-chains of Asp82 and Ser85 (Figure 5).

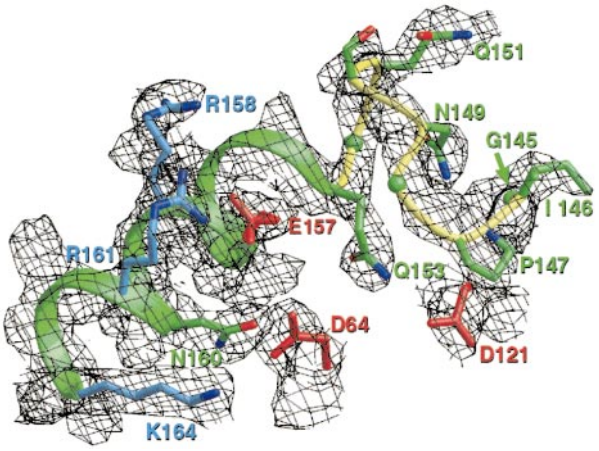

Figure 4. The active-site loop. Electron density ($2F_o - F_c$) map (2.53 Å resolution, contoured at 1.0 σ in the region of the catalytic triad of the triclinic form (''distal'' subunit) showing the ordered conformation of the ''active-site loop'' (residues 145-154, yellow backbone).

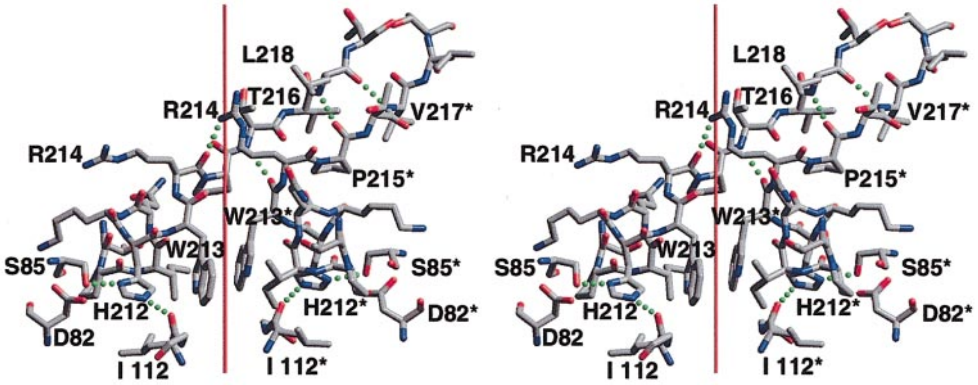

Figure 5. The asymmetric linker. Stereo view of the linker region between the catalytic and C domains, oriented as in Figure 1(a). The canted orientation of the C domain dimer is elicited through asymmetric contacts within the linker region; the parallel β-strands hydrogen bond (green dots) out of register from the sequence. Residues Trp213 and Ile209, located near the dyad of the catalytic dimer, form a hydrophobic core apparently stabilizing this part of the catalytic domain. A hydrogen bond network between His212 and residues Ser85, Asp82 and Ile112 also appear to anchor helix F to the catalytic domain. Residue numbers from the proximal subunit are marked with an asterisk (*).

In previous crystal structures, the catalytic domain interface was formed through contacts between neighboring molecules in the crystal, leaving open the question of whether this interface was merely a consequence of crystal packing. In the two-domain structure, however, we see this dimer occurring in the absence of crystal symmetry, bolstering the idea that it exists in solution. This subunit arrangement places the two active centers on opposite sides of the dimer, evidently too far apart for both to participate in a coupled joining reaction involving both viral cDNA ends spaced by six base-pairs in the target DNA. As discussed below, preservation of this interface forces one to invoke higher-order assemblies, such as tetramers.

The linker region

Residues 214-219 (β-strand 6) link the catalytic domains with the C domain (Figures 3 and 5). The indole side-chains of Trp213 at the end of helix 5 stack with each other across the dyad (Figure 5), marking the point at which the polypeptide chains begin to deviate from the 2-fold symmetry of the catalytic domain dimer.

Strands 6 from both monomers hydrogen-bond to form a parallel two-stranded β-sheet. However, the pairing of residues 214-217 from the proximal chain with residues 216-218 from the distal chain are ''out of register'' (Figure 5), evidently forcing the canted conformation seen in the C domain dimer.

In the HIV domain structure, helices 5 extend further, anti-parallel across the dyad. It is not clear at present whether this represents a real difference in the structures, or if the extended helical conformation of the HIV construct is a consequence of having been truncated just beyond that position. The amino acid sequences in this region of HIV, RSV, and other integrases are not conserved.

The C domains

The C domain monomer comprises five β-strands in a barrel-like SH3 fold (residues 220-270). Structures of the C domain monomers of HIV (Eijkelenboom *et al.*, 1995; Lodi *et al.*, 1995) solved by NMR are very similar to RSV C domains. Figure 6(a) shows a structure-based sequence alignment of RSV and HIV, and the level of sequence conservation among other representative retroviral species. Because many conserved residues are those evidently responsible for stabilizing this fold, it seems likely that this fold is conserved in all species.

The sequence and length of the linkers and the C-terminal residues (those outside of the SH3-like domain fold) are poorly conserved between retroviral strains. Sequences of HIV and SIV integrases suggest that their linkers are longer by about 13 residues.

Structural asymmetry of the RSV integrase C domain dimer

In both our crystal structures of the two-domain RSV integrase, the C domain monomers (Figure 6(b)) dimerize asymmetrically. The C domain dimer interface is formed by strands 1′, 2′ and 5′ (residues 222-224, 239, 242, and 265-268) from the proximal monomer and strands 2′, 3′ and 4′ from the distal monomer (residues 240-244, 246, and 259. Except for minor differences in surface side-chains, the proximal and distal domains are very similar in conformation. In our models, the superposition of the C domains requires a rotation of one domain by just over 90° about an axis passing near the end of the linker residues (Figure 6(b)). The rotations are 93° in the monoclinic form and 97° in the triclinic forms.

The asymmetrical association of C-domains and the ''out of register'' parallel strand interactions in

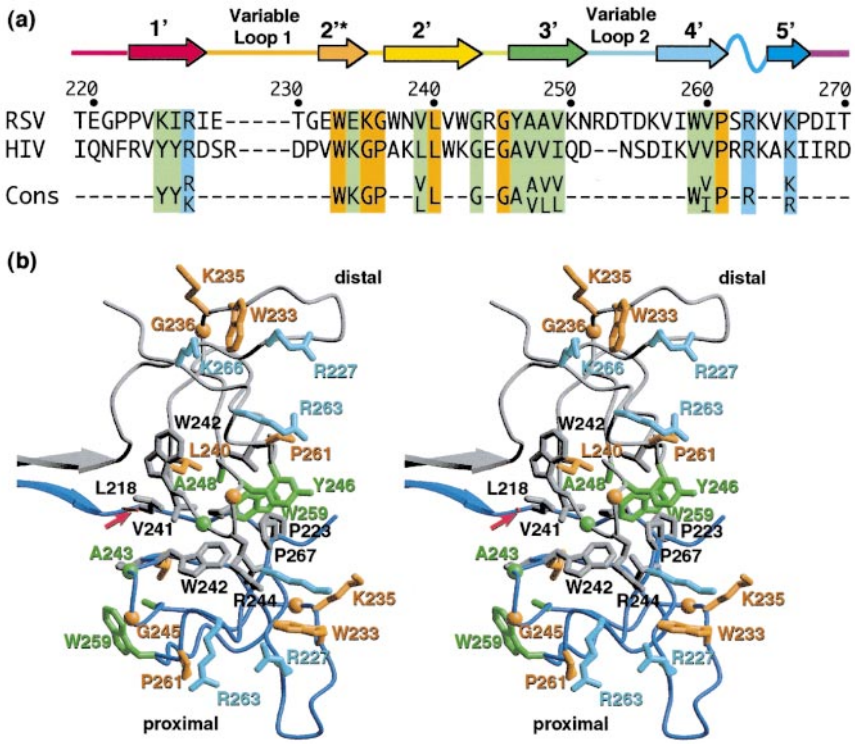

Figure 6. The C domain dimer. (a) Structure-based sequence alignment of retroviral integrase C domains. RSV and HIV residues 220-270 are aligned, along with a consensus sequence derived from 13 retroviruses (after Puras Lutzke *et al.*, 1994); RSV sequence numbering is shown. β-Strands are indicated by colored arrows. Highly conserved residues (found in 11 or more of the 13 sequences) are highlighted orange, highly conserved basic residues are cyan. The remaining consensus residues are green. Loops varying in both length and sequence between retroviral species are labeled. (b) Stereo view of conserved amino acid side-chains and residues comprising the dimer interface. Conserved residues are colored as in (a); non-conserved residues are gray. Note the highly conserved basic residues clustered with Trp233 on the surface of the C domain. The view is shown looking down a rotation axis (arrow) located near the ends of the linker residues that can roughly superimpose the C domains though an approximate 90° rotation.

the linker residues are both likely responsible for the canted conformation of the C-domains (Figure 3(b)). Canted conformations are observed in both crystallographic space groups, despite different crystal packing. In the *P*1 crystal form, C domain dimers form crystal contacts with other C domains through neighboring distal domains, whereas in the $P2_1$ form, contacts are made through the proximal domains.

In view of the number of hydrogen bonds in the linker and intermolecular contacts between, one cannot assume that the asymmetrical packing of C domains is forced by the crystal lattice. There must be, however, some degree of flexibility in the linker region, since the C domains in the two crystal forms differ in tilt (Figure 3(b)) by about 20°. Crystal contacts, therefore, may be responsible for modulating the observed tilt, but not the overall asymmetry in the integrase dimer. The N-terminal residues of the proximal subunits interact with both the proximal C domain and proximal core domain at their interface. Flexibility of the linker region might therefore be due to the absence of the N domain in our construct.

Comparison with HIV C domain structures

The dimer interface observed here differs dramatically from that reported from NMR studies of the isolated domains from HIV integrase (Figure 7). The NMR structures (Eijkelenboom *et al.*, 1995; Lodi *et al.*, 1995) show a symmetrical dimer where the SH3 folds interact through β-strands 2′, 3′, and 4′. In our model, the same 2′, 3′, and 4′ β-strands are used by the distal monomer, but these pack against strands 1′, 2′, and 5′ of the proximal monomer. Surface area calculations show that the buried surface areas in the two viral species are similar (about 850 A^2 for RSV and about 720 A^2 for HIV; PDB code 1IHV).

For either the solution NMR or crystallography studies, the observed structures could, in principle, be an artifact of using fragments at high concentrations. For example, the HIV C domain fragments might associate differently in the presence of the N and core domains. It may be that the RSV C domain dimer is symmetric in the presence of the missing N domain. Because integrase sequences vary most in the C-terminal regions, it is possible that the two proteins assemble differently *in vivo*.

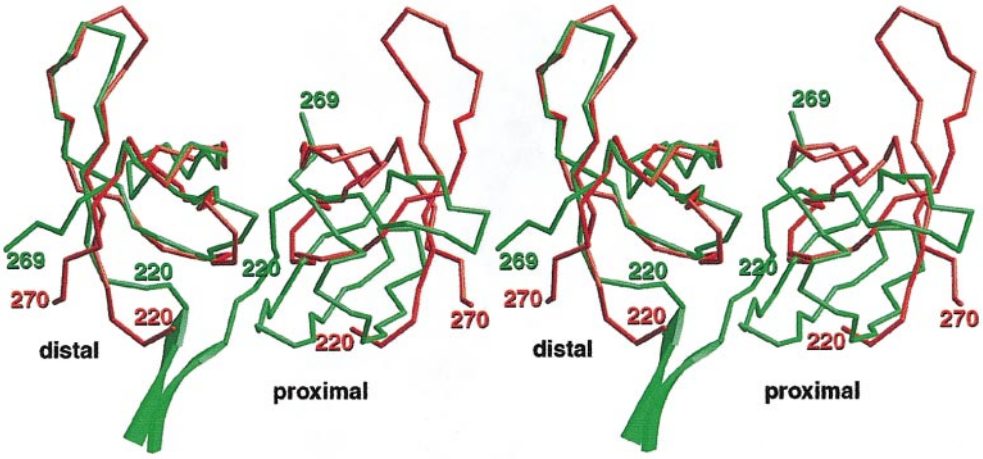

Figure 7. Comparison of RSV and HIV C domain dimerization. The distal domain of the RSV structure superimposes on an HIV monomer (PDB ID: 1IHV) with an r.m.s. deviation in C^{α} positions of about 1.2 Å (excluding insertions). However, the adjoining subunits of RSV and HIV integrase do not superimpose. The HIV dimer is shown with its local 2-fold axis aligned vertically.

As discussed below, we show how the asymmetric C domain interface seen in our structure can be readily extended to provide a possible tetramerization motif. The observed C domain dimer, as we propose, may be one half of a catalytically relevant tetramer. Another possibility is that both dimerization modes are functionally relevant. The protein may perhaps be able to arrange itself for different purposes. For example, integrase may form different multimers in coupled integration complexes at the sites of catalysis, at distant locations within integration complexes, and even other complexes important at other points in the retroviral life-cycle.

Modeling protein-DNA interactions

The observed canted dimers provides the basis for modeling plausible integration complexes (Figure 8). A viral cDNA end is docked with its 3′-OH end at the active center on the observed dimer structure so as to contact the C domains. The best fit (using an idealized linear double-stranded *B*-DNA model) involves simultaneous contacts with both proximal and distal C domains and the catalytic center of the proximal domain. Due to the asymmetrical association of the C domains, the face of the C domain dimer presented to the proximal active site displays two prominent basic surface patches (Figure 8(a)) available for binding with the acidic phosphodiester backbone of the viral DNA end. If the viral DNA 3′-end is first docked into the alternative (distal) subunit active center, it would fail to reach the basic patches on the C domains. Others have reported that bases about 7-11 positions from the 3′-end of the cDNA bind to the C domain (Esposito & Craigie, 1998; Heuer & Brown, 1997). The monoclinic model, with its slightly greater domain tilt (Figure 3(b)), provides better charge complementarity between the DNA and basic residues on the protein. Since this model suggests that each viral end binds to two C domains, it follows that at least a dimer of dimers (or tetramer) must be formed in the coupled joining complex.

Modeling integrase-DNA complexes involved in coupled-joining

Complete integrase-DNA complexes must accommodate three DNA segments: both of the two viral cDNA ends and the target DNA (Figure 1(b)). Although only two catalytic centers are required in principle, one for joining each cDNA end, it does not seem possible to construct a model using both active sites of a single crystallographically observed dimer. Therefore, our models invoke tetramers in which only two of the four active sites are involved in catalysis. The points of joining in the target DNA are known to be spaced six bases apart on opposite strands, corresponding to opposite sides of a single major grove (Pryciak *et al.*, 1992; Pryciak & Varmus, 1992; Pruss *et al.*, 1994a). Cross-linking studies in the HIV system indicate that the cDNA terminal bases are bound by the catalytic domain, while more distal cDNA sequences bind to the C-terminal domain (Esposito & Craigie, 1998; Heuer & Brown, 1997). Points of close approach between HIV integrase and substrate DNA have also been mapped in adduct interference experiments (Bushman & Craigie, 1992; Chow & Brown, 1994). In the following, we assume that the general conclusions from these HIV studies hold for RSV, though further data from the RSV system will be needed to confirm this. In the first of two classes of models (class I) for the full integration complex, the model was constructed by arranging two complexes of an integrase dimer/viral DNA end (presented in Figure 8) onto a model target DNA at the two integration

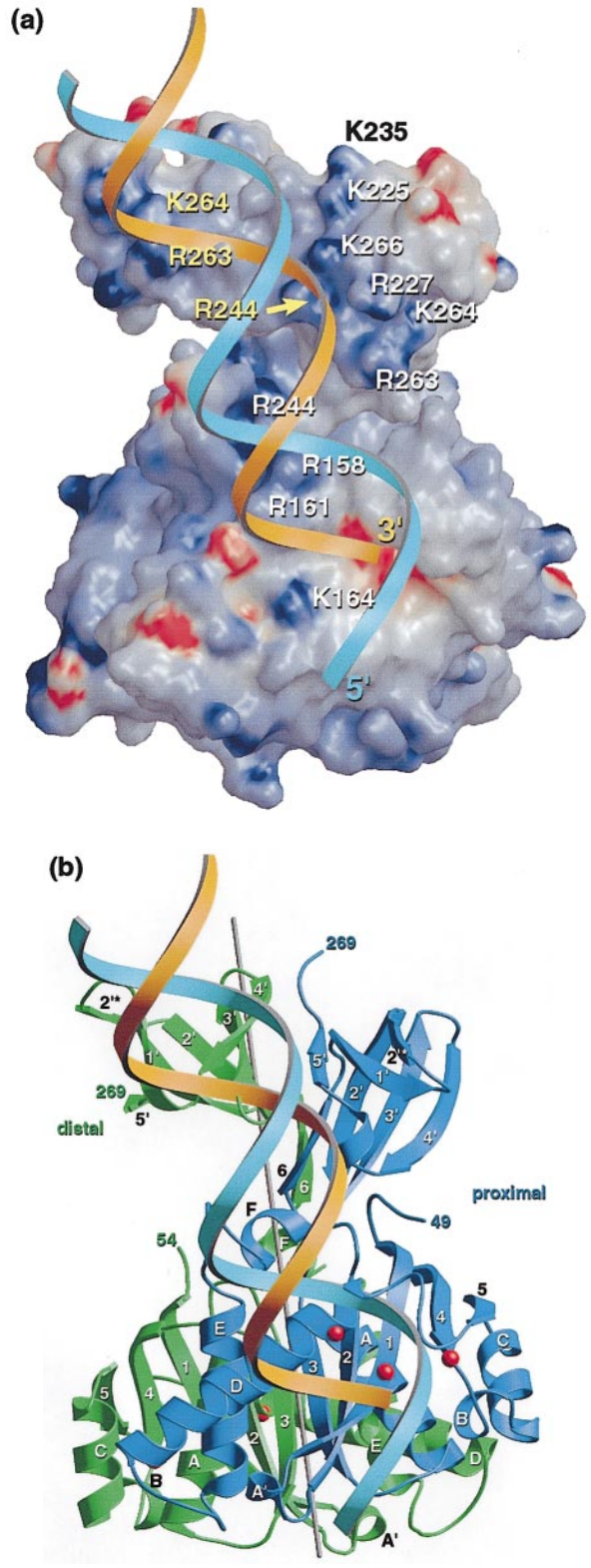

Figure 8. Proposed docking site for the end of the viral cDNA. (a) The two-domain integrase dimer structures in their canted conformations present a basic surface spanning regions of both C domains and the proximal catalytic domain. An idealized model of *B*-DNA (ribbon figure) can be positioned with its 3′ end at the active center (red patch) of the proximal subunit and docked to make contact with a number of basic residues on the protein surface of both domains. The electrostatic surface potential map (GRASP surface) shows basic and acidic regions in blue and red, respectively. The orientation shown is at an approximate right-angle to that of Figure 3(a). Residues with yellow labels belong to the distal subunit. (b) Ribbon figures of the DNA docking model from the same view. Proximal and distal polypeptide chains are colored blue and green, respectively. Active-center residues are marked by red spheres. The local 2-fold axis of the catalytic domains is shown by the nearly vertical gray line. The conformation of the protein in the $P2_1$ crystal form is shown in (a) and (b).

sites (Figure 9(a) and (b)). This model is effectively a dimer of dimers, with an axis of 2-fold symmetry relating the two cDNA/dimer complexes and the target DNA between the sites of integration. The interface between dimers was adjusted to accommodate the three DNAs, while minimizing overlap between the protein domains. The described arrangement leads to a reasonable interdigitation of the carboxyl domain dimers and a tight interface between the catalytic domains. A bent target DNA provides a better fit, in accord with biochemical experiments, indicating that integration is favored at widened sites in major grooves in nucleosomes (Muller & Varmus, 1994; Pruss *et al.*, 1994a,b).

The second model (class II) employs a similar dimer of dimers arrangement with one important exception that makes use of the asymmetrical arrangement of C domain dimers. Since the rotation relating the two C domain monomers in the dimer is roughly 90°, two more monomers can be added to a dimer, each using the same crystallographic interface, to form a closed tetramer.

In this rotationally symmetric C domain tetramer model, each of the four domain interfaces matches the dimer interface observed in the crystals (Figure 9(c)). Both full-length RSV integrase and the C domain alone have been reported to form tetramers in solution (Andrake & Skalka, 1995), consistent with this idea. A key feature of our modeling is that formation of an integrase tetramer can satisfy many requirements for coupled integration of the two cDNA ends. At present, we do not have enough information to distinguish between class I and class II models.

Integrase dimers in their crystallographically observed canted conformations could not be assembled into the proposed class II tetramer unless the protein flexes at the interdomain linkers. If such a movement is made, the viral cDNA can be bound to conserved basic residues on the tetramer, thus also satisfying the constraints described above. In this arrangement, the catalytic domains also fit well on the DNA substrates and form a plausible dimer interface (Figure 9(d)). Because this model involves distortion of the observed conformation, we cannot distinguish between several possibilities for covalent connectivity between the four catalytic domains and the C domains. How-

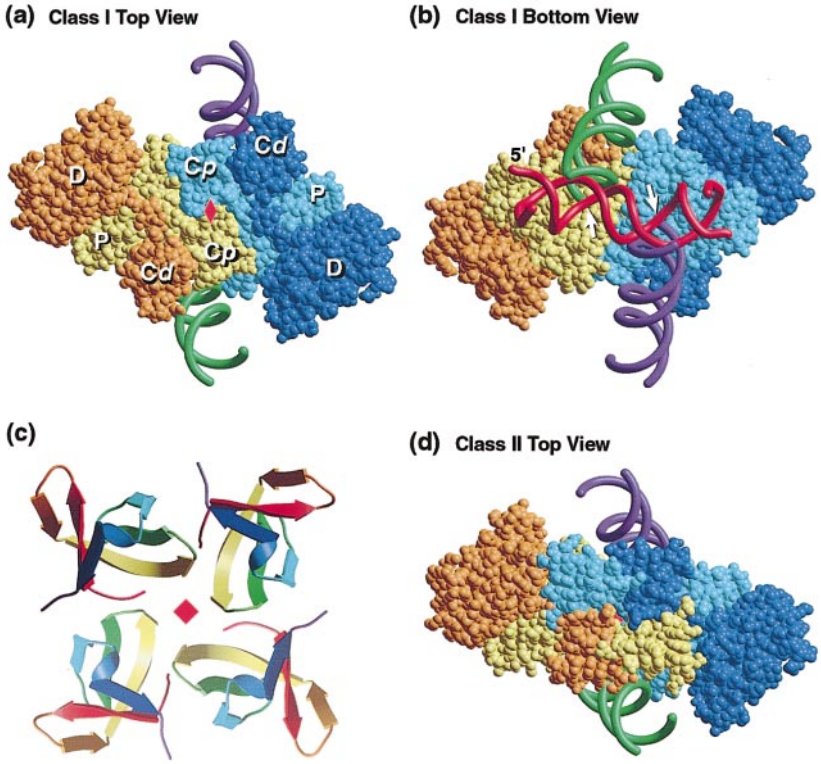

Figure 9. Models for DNA binding and coupled integration by RSV IN 49-286. (a) Class I model for coupled integration based on the crystallographically observed conformation. The overall complex is a dimer of dimers (orange/yellow and blue/cyan for distal/proximal pairs). The active proximal protomers interact tightly near the overall 2-fold axis (red diamond). Distal catalytic domains located at outer edges of the complex are proposed to not participate in catalysis. Proximal C domains (yellow and cyan) cluster near the dyad. Proximal and distal catalytic domains are labeled P and D, and proximal and distal C-domains are label Cp and Cd, respectively. Each viral cDNA end (green and violet) binds primarily to the proximal catalytic domain and both C domains of one integrase dimer. The target DNA is obscured in this view. (b) Class I model, bottom view, showing the target DNA segment (red), two viral cDNA 3′-ends (green and violet) and the two active centers of the proximal subunits. The curved target DNA segment (red) preferred by integrase is taken from the nucleosome core structure (Luger *et al.*, 1997). Active-sites and points of joining in the target DNA are marked with arrows. (c) Model for a C domain tetramer. A symmetric tetramer with 4-fold symmetry follows readily if two additional monomers are added and rotated 90° relative to its neighbors using the same rotation axis that relates proximal and distal monomers (red diamond). Coloring is by secondary structure as in Figure 2(a). (d) Class II model for coupled integration based on the hypothetical tetramer of C domains (top view). The catalytic and C domains in each monomer have been separated to permit formation of the C domain tetramer. Therefore, the connectivity between catalytic and C domains is unknown; the relationship implied by the colors illustrates one possibility. Like the class I model, both catalytic domain dimers are symmetrically disposed around the overall 2-fold axis (red diamond), although the C domains obey a higher 4-fold local symmetry. The essential features of the C domain interactions from the class I model are preserved: a dimer of dimers arrangement in which two of the four active sites participate in catalysis. A bottom view of the class II model (not shown) is similar to that in (b). As in the class I model, each viral cDNA is proposed to bind to two C domains, although the connectivity (a color pattern) cannot be determined. The possibility shown implies that each viral end might bind a dimer productively prior to tetramer formation.

ever, it seems sensible that, like the class I model, each viral cDNA end binds to two C domains. The intriguing possibility exists that binding of a cDNA end alters the conformation of the integrase dimer, making them competent for tetramer formation. Tetramerization in this manner could potentially also help to coordinate the formation of integrase complexes with completion of reverse transcription.

As with our models, Heuer and Brown in cross-linking and modeling studies of HIV integrase also propose the presence of two kinds of protomers in the integration complex; those that participate in catalysis and those that do not. They suggest that integrase monomers not involved in covalent chemistry may nevertheless contribute much of the DNA binding in the complex, leading to a proposal that an octamer may be the active form (Heuer & Brown, 1998).

Integration complexes *in vivo* likely involve further features not yet modeled. Several studies have implicated distortion of both the viral cDNA and target DNA as important for integration, but the conformation of the distorted DNA is unclear (Bor *et al.*, 1995; Bushman & Craigie, 1992; Katz *et al.*, 1998; Muller & Varmus, 1994; Pruss *et al.*, 1994a; Scottoline *et al.*, 1997). The placement of the amino-terminal zinc-binding domains (absent in RSV IN 49-286) remains unknown, although for HIV, the N domain can cross-link to target DNA (Heuer & Brown, 1997). In our models, target DNA makes relatively few contacts with the other two integrase domains, so it is attractive to model the zinc-binding domain as binding to target DNA.

The two-domain RSV integrase structure and inhibitor design

The studies presented here provide new targets for design of potential antiretroviral drugs. Previous studies of isolated catalytic domains have provided views of the integrase active site, but in

pre-integration complexes *in vivo*, the active site is expected to be bound to viral DNA. The new models provide candidate templates for designing inhibitors active against integrase-DNA complexes. Importantly, the new protein-protein interactions observed in the two-domain structure also provide new targets. For example, inhibitors might target the asymmetric interface between the C domains. At present, the degree of similarity between RSV and HIV is unclear, but the phylogenetic conservation of integrase proteins is impressive and inhibitors active against RSV integrase may often inhibit HIV integrase as well.

Materials and Methods

Construction of plasmids encoding derivatives of RSV integrase

Mutant derivatives of RSV integrase were generated by site-directed mutagenesis. Sequences of RSV were derived from the pATV8 clone from the Prague C strain (Katz *et al*., 1982). The starting plasmid pBW5 (Bushman & Wang, 1994), encoding amino acid residues 49-286 of RSV integrase, was fused to a His-Tag sequence supplied by the vector (pET15B, Novagen). To facilitate structural studies, the His-Tag sequence was removed by cloning the *Nde*I to *Pst*I fragment into the pINSD1 (Engelman & Craigie, 1992) expression vector, yielding pBW22. To generate point mutations in the gene for RSV IN 49-286, the Excite Mutagenesis Kit (Stratagene) was used with appropriate mutagenic primers. In this way, point mutations A190K, L196K, and F199K were generated, to yield plasmids pBW31, pBW29, and pBW25, respectively. A deletion mutant encoding RSV IN residues 58-286 was also constructed using the Excite Mutagenesis Kit. Sequences of the mutagenic oligonucleotides used are available upon request. All DNA constructions were analyzed and confirmed by DNA sequencing.

Purification of RSV IN 49-286 and derivatives

Expression plasmids were transformed into *Escherichia coli* BL21(DE3) pLysS (Novagen), and grown in a fermentor (New Brunswick Scientific BioFlo 3000) containing 10 l of Luria broth (LB), 50 μg/ml carbenicillin (Gibco BRL), and 1% (w/v) glucose. The cultures were grown at 37 °C to a cell density of $A_{600} \sim 1.5$, induced with 2 mM IPTG (United States Biochemical) for two hours, then harvested by centrifugation (Beckman JA10, 5000 ***g***) and frozen at −80 °C, typically yielding ~40 g of wet cell paste. Expression levels and identification of the abundant integrase in column fractions were made by visual inspection of Coomassie-stained SDS-PAGE gels.

In a typical protein purification, 40 g of cells was suspended in 250 ml of lysis buffer (buffer B; 25 mM Tris-HCl (pH 7.5), 1.5 M NaCl, 5 mM DTT, 1.7 mM 4-(2-aminoethyl)benzensulfonylfonylfluoride, HCl (AEBSF), 10 mM $MgSO_4$, 1 mM EDTA), sonicated (four to five minutes on ice, 90% duty cycle, output setting 9, Branson Sonifier 250), and centrifuged (100,000 ***g***, Beckman Avanti J-30 I, 20 minutes, 4 °C). The supernatant fluid was poured directly into a 60 ml slurry of SP-Streamline cation-exchange medium (Pharmacia) and the salt concentration lowered about fivefold by the slow addition, with mixing, of about 1 l of dilution buffer (20 mM Tris-HCl (pH 7.5), 5 mM DTT). The diluted supernatant was decanted after allowing the medium to settle. Resuspended medium with bound integrase was packed into a 25 mm diameter × 100 mm glass column (Pharmacia, XK25) and washed with 150 ml of buffer A (buffer B lacking NaCl). Protein was eluted with a sharp linear gradient of 10% - 100% buffer B. Fractions containing integrase (eluting at about 0.5-1.0 M NaCl) were dialyzed against buffer C (50 mM Mes (pH 6.1), 100 mM NaCl, 5 mM DTT, 1.7 mM AEBSF, 10 mM $MgSO_4$, 1 mM EDTA, 4 °C) and loaded onto a 10 mm diameter, 100 mm long, POROS HS 20 column (PerSeptive Biosystems) and eluted with a linear gradient of 10-100% buffer D (buffer C with 1.0 M NaCl). Fractions containing integrase were pooled, and flash frozen for storage at −80 °C after adding glycerol to 10% (v/v).

Preparation of selenomethionine-modified RSV IN 49-286 F199K

Plasmid pBW25 encoding the F199K mutation was transformed into *E. coli* strain DL21 (DE3)pLysS and grown overnight in 50 ml of LeMaster's minimal medium (Yang *et al*., 1990) containing 50 μg/ml methionine, 50 μg/ml carbenicillin, and 35 μg/ml chloramphenicol at 32 °C. The overnight culture was used to inoculate 10 l of fresh LeMaster's medium containing 50 μg/ml L-selenomethionine (Sigma), 50 μg/ml carbenicillin, and 35 μg/ml chloramphenicol. Following growth to $A_{600} = 0.8$ at 32 °C, labeled-integrase was induced with IPTG for three hours at 32 °C. Cells were harvested and protein was purified as above, except that 20 mM DTT was added to all buffers to limit oxidation of selenium.

Activity assays

Assays of the activities of the RSV integrase fragments were carried out essentially as described (Bushman & Wang, 1994). Briefly, the terminal cleavage substrate consisted of two hybridized oligonucleotides, 5′-CTACAAGAGTATTGCATAAGACTACATT-3′ (FB141) and its complement (FB142). The sequence matches that of the U3 end of the unintegrated RSV cDNA. The sequence of the dumbbell disintegration product is shown in Figure 2(a). Substrates (FB141 and the dumbbell disintegration substrate FB221) were ^{32}P end-labeled. Assay mixtures contained 0.5 pmol of DNA substrate, 5 mM $MnCl_2$ or $MgCl_2$, 25 mM Hepes (pH 7.5), 40 mM NaCl, 20 mM 2-mercaptoethanol, 100 μg/ml bovine serum albumin, and 10% glycerol. Enzyme was added last to each assay. Reactions were incubated for one hour at 37 °C, and were stopped by adding sequencing gel loading dye and heating briefly to 95 °C. Reaction products were analyzed by electrophoresis on 15% polyacrylamide DNA-sequencing type gels and visualized by autoradiography.

Crystallization of RSV IN 49-286 F199K

The two-domain integrase mutants, dialyzed (25 mM Bis-Tris-HCl (pH 6.5), 3 mM DTT, three hours, 4 °C) and concentrated (15-20 mg/ml, Centriplus 10 K, Amicon), were tested for crystallization in hanging drop vapor diffusion sparse matrix screens (4 °C and 21 °C, Crystalscreen 1 and 2, and Natrix, Hampton Research, protein and reservoir mixed 1:1 (v/v) to give 2-4 μl drops). Small crystals of the F199 K form initially grew with Natrix condition #23 (0.2 M KCl, 0.01 M $MgCl_2$, 0.05 M sodium cacodylate (pH 6.5), and 10% (w/v) PEG 4000).

Monoclinic ($P2_1$) crystals were produced by hanging drop vapor diffusion (reservoir solution: 0.05 M KCl, 0.01 M $MgCl_2$, 0.05 M sodium cacodylate (pH 7.0), 10 % (w/v) PEG 3350, and 25 % (v/v) ethylene glycol). Triclinic ($P1$) crystals appeared infrequently in the same conditions. Monoclinic ($P2_1$) crystals of selenomethionine-incorporated protein were grown by sitting drop vapor diffusion under modifed native conditions with added imidazole (reservoir solution, 0.050 M KCl, 0.01 M $MgCl_2$, 0.1 M sodium cacodylate (pH 7.2), 0.02 M imidazole, 10 % PEG 3350, and 25 % ethylene glycol). The monoclinic $P2_1$ form contains one dimer in the asymmetric unit, whereas the triclinic $P1$ form has two.

Structure determination of the monoclinic form

Data from monoclinic and triclinic crystal forms were collected at the Brookhaven National Synchrotron Light Source station X9B using either a MAR345 image plate or ADSC Quantum 4 CCD detector. Data were processed using HKL (Denzo/Scalepack) software (Otwinowski & Minor, 1997). The triclinic cell has approximate $C2$ symmetry, but the data fail to index in the monoclinic cell. An interpretable experimentally phased electron density map at 3.2 Å resolution was obtained through MAD methods (SOLVE (Terwilliger & Berendzen, 1999)) using data from selenomethionine-modified monoclinic crystals (Table 1). The catalytic domain was modeled using the structure of ASV integrase catalytic domain (PDB 1VSE) positioned by molecular replacement. The selenium positions superimposed on the methionine positions located in the molecular replacement solution, confirming the solution. Initial MAD phases with a mean figure-of-merit of 0.46 were improved to 0.84 using density modification procedures (program dm). The SH3-like folds of the C domains were located visually in the MAD-phased maps. Monomeric models of C domains of HIV IN (PDB IIHV) were positioned manually, then modified into the RSV sequence. The resulting model containing both catalytic and C domains was partially refined (X-PLOR, Brunger, 1992; SHELX, Sheldrick & Schneider, 1997) to a standard R-factor of 30 % at 3.2 Å resolution.

Structure determination of the triclinic form

The monoclinic dimer model was used in molecular replacement to solve for the high-resolution triclinic crystal form (four monomers per unit cell, R-factor 43 %, correlation coefficient 48 % (AMoRe, Navaza & Saludjian, 1997). The triclinic structure was refined using standard XPLOR and CNS (Brunger, 1992; Brunger *et al.*, 1998) protocols to a final R-factor of 21.6 % (Table 1). The two dimers in the triclinic cell were restrained during refinement by non-crystallographic symmetry corresponding to an approximate 180° rotation between dimers. Manual rebuilding of the model was accomplished using CHAIN (Sack & Quiocho, 1997). The final triclinic structure comprises about 85 % of the sequence: the first five residues at the N termini and C-terminal 17 amino acid residues are partially or completely disordered. In addition, the loop structure near the active site (145-154) was visualized and constructed successfully in two of four independent monomers. Information from the triclinic structure was used to complete the monoclinic structure, which was further refined to an R factor of 25.6 %.

Figures were prepared using programs MOLSCRIPT (Kraulis, 1991), BOBSCRIPT (Esnouf, 1997), Raster3D (Merritt & Bacon, 1997), and GRASP (Nicholls *et al.*, 1991).

RCSB Protein Data Bank accession numbers

Coordinates have been deposited with the RCSB Protein Data Bank (ID: 1C0M and 1C1A).

Note added in proof

A recently published atomic resolution structure of a D64N mutant of the ASV catalytic domain shows an ordered loop conformation similar to that observed in the distal monomer of the triclinic two-domain form (Figure 4) (Lubkowski, J., Dauter, Z., Yang, F., Alexandratos, J., Merkel, G., Skalka, A. M. & Wlodawer, A. (1999). *Biochemistry*, **38**, 13512-13522.). For example, residues 145-154 of RCSB Protein Data Bank model 1c9z (low pH form) superimpose with an r.m.s.d. of 0.39 Å in main-chain atoms.

Acknowledgements

We thank Zbignew Dauter and Beibei Wang for technical help, Nancy Nossal for valuable advice, and Joan Hanley-Hyde for editorial assistance. This work was supported by NIH grants AI34786 and GM56553 (to F.D.B.) and by the NIAMS Intramural Research Program and Intramural AIDS Targeted Antiviral Program, NIH (to C.C.H.). F.D.B. is a scholar of the Leukemia Society of America.

References

Andrake, M. D. & Skalka, A.-M. (1995). Multimerization determinants reside in both the catalytic core and C terminus of avian sarcoma virus integrase. *J. Biol. Chem.* **270**, 29299-29306.

Bor, Y.-C., Bushman, F. & Orgel, L. (1995). *In vitro* integration of human immunodeficiency virus type 1 cDNA into targets containing protein-induced bends. *Proc. Natl Acad. Sci. USA,* **92**, 10334-10338.

Brown, P. O., Bowerman, B., Varmus, H. E. & Bishop, J. M. (1987). Correct integration of retroviral DNA *in vitro*. *Cell,* **49**, 347-356.

Brown, P. O., Bowerman, B., Varmus, H. E. & Bishop, J. M. (1989). Retroviral integration: structure of the initial covalent complex and its precursor, and a role for the viral IN protein. *Proc. Natl Acad. Sci. USA,* **86**, 2525-2529.

Brunger, A. (1992). *X-PLOR A system for Crystallography and NMR. Version 3.1*, Yale University, New Haven, CT.

Brunger, A. T., Adams, P. D., Clore, G. M., DeLane, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallog. sect. D,* **54**, 905-921.

Bujacz, G., Jaskolski, M., Alexandratos, J., Wlodawer, A., Merkel, G., Katz, R. A. & Skalka, A. M. (1995). High-resolution structure of the catalytic domain of avian sarcoma virus integrase. *J. Mol. Biol.* **253**, 333-346.

Bujacz, G., Alexandratos, J., Wlodawer, A., Merkel, G., Andrake, M., Katz, R. A. & Skalka, A. M. (1997). Binding of different divalent cations to the active site of avian sarcoma virus integrase and their effects on enzymatic activity. *J. Biol. Chem.* **272**, 18161-18168.

Bushman, F. D. & Craigie, R. (1991). Activities of human immunodeficiency virus (HIV) integration protein *in vitro*: specific cleavage and integration of HIV DNA. *Proc. Natl Acad. Sci. USA,* **88**, 1339-1343.

Bushman, F. D. & Craigie, R. (1992). Integration of human immunodeficiency virus DNA: Adduct interference analysis of required DNA sites. *Proc. Natl Acad. Sci. USA,* **89**, 3458-3462.

Bushman, F. D. & Wang, B. (1994). Rous sarcoma virus integrase protein: mapping functions for catalysis and substrate binding. *J. Virol.* **68**, 2215-2223.

Bushman, F. D., Fujiwara, T. & Craigie, R. (1990). Retroviral DNA integration directed by HIV integration protein *in vitro*. *Science,* **249**, 1555-1558.

Bushman, F. D., Engelman, A., Palmer, I., Wingfield, P. & Craigie, R. (1993). Domains of the integrase protein of human immunodeficiency virus type 1 responsible for polynucleotidyl transfer and zinc binding. *Proc. Natl Acad. Sci. USA,* **90**, 3428-3432.

Cai, M., Zheng, R., Caffrey, M., Craigie, R., Clore, G. M. & Gronenborn, A. M. (1997). Solution structure of the N-terminal zinc binding domain of HIV-1 integrase. *Nature Struct. Biol.* **4**, 567-577.

Chow, S. A. & Brown, P. O. (1994). Substrate features important for recognition and catalysis by human immunodeficiency virus type 1 integrase identified by using novel DNA substrates. *J. Virol.* **68**, 3896-3907.

Chow, S. A., Vincent, K. A., Ellison, V. & Brown, P. O. (1992). Reversal of integration and DNA splicing mediated by integrase of human immunodeficiency virus. *Science,* **255**, 723-726.

Coffin, J. M., Hughes, S. H. & Varmus, H. E. (1997). *Retroviruses*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Craigie, R., Fujiwara, T. & Bushman, F. (1990). The IN protein of Moloney murine leukemia virus processes the viral DNA ends and accomplishes their integration *in vitro. Cell,* **62**, 829-837.

Dyda, F., Hickman, A. B., Jenkins, T. M., Engelman, A., Craigie, R. & Davies, D. R. (1994). Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science,* **266**, 1981-1986.

Eijkelenboom, A. P. A. M., Puras Lutzke, R. A., Boelens, R., Plasterk, R. H. A., Kaptein, R. & Hard, K. (1995). The DNA binding domain of HIV-1 integrase has an SH3-like fold. *Nature Struct. Biol.* **2**, 807-810.

Eijkelenboom, A. P. A. M., van den Ent, F. M. I., Vos, A., Doreleijers, J. F., Hard, K., Tullius, T., Plasterk, R. H. A., Kaptein, R. & Boelens, R. (1997). The solution structure of the amino-terminal HHCC domain of HIV-2 integrase; a three-helix bundle stabilized by zinc. *Cur. Biol.* **1**, 739-746.

Engelman, A. & Craigie, R. (1992). Identification of conserved amino acid residues critical for human immunodeficiency virus type 1 integrase function *in vitro*. *J. Virol.* **66**, 6361-6369.

Engelman, A., Hickman, A. B. & Craigie, R. (1994). The core and carboxyl-terminal domains of the integrase protein of human immunodeficiency virus type 1 each contribute to nonspecific DNA binding. *J. Virol.* **68**, 5911-5917.

Esnouf, R. M. (1997). An extensively modified version of Molscript that includes greatly enhanced coloring capabilities. *J. Mol. Graph. Model.* **15**, 132-134.

Esposito, D. & Craigie, R. (1998). Sequence specificity of viral end DNA binding by HIV-1 integrase reveals critical regions for protein-DNA interaction. *EMBO J.* **17**, 5832-5843.

Fujiwara, T. & Mizuuchi, K. (1988). Retroviral DNA integration: structure of an integration intermediate. *Cell,* **54**, 497-504.

Goldgur, Y., Dyda, F., Hickman, A. B., Jenkins, T. M., Craigie, R. & Davies, D. R. (1998). Three new structures of the core domain of HIV-1 integrase: an active site that binds magnesium. *Proc. Natl Acad. Sci. USA,* **95**, 9150-9154.

Greenwald, J., Le, V., Butler, S. L., Bushman, F. D. & Choe, S. (1999). The mobility of an HIV-1 integrase active site loop is correlated with catalytic activity. *Biochemistry,* **38**, 8892-8898.

Hansen, M. S. T., Carteau, S., Hoffmann, C., Li, L. & Bushman, F. (1998). Retroviral cDNA integration: mechanism, applications and inhibition. In *Genetic Engineering. Principles and Methods* (Setlow, J. K., ed.), vol. 20, pp. 41-62, Plenum Press, New York and London.

Heuer, T. S. & Brown, P. O. (1997). 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,* **36**, 10655-10665.

Heuer, T. S. & Brown, P. O. (1998). Photo-cross-linking studies suggest a model for the architecture of an active human immunodeficiency virus type-1 integrase-DNA complex. *Biochemistry,* **37**, 6667-6678.

Jenkins, T. M., Engelman, A., Ghirlando, R. & Craigie, R. (1996). A soluble active mutant of HIV-1 integrase: involvement of both the core and carboxyl-terminal domains in multimerization. *J. Biol. Chem.* **271**, 7712-7718.

Katz, R. A., Omer, C. A., Weis, J. H., Mitsialis, A., Faras, A. J. & Guntaka, R. V. (1982). Restriction endonuclease and nucleotide sequence analysis of molecularly cloned unintegrated avian tumor virus DNA: structure of large terminal repeats in circle junctions. *J. Virol.* **42**, 346-351.

Katz, R. A., Merkel, G., Kulkosky, J., Leis, J. & Skalka, A. M. (1990). The avian retroviral IN protein is both necessary and sufficient for integrative recombination *in vitro. Cell,* **63**, 87-95.

Katz, R. A., Gravuer, K. & Skalka, A. M. (1998). A preferred target DNA structure for retroviral integrase *in vitro*. *J. Biol. Chem.* **273**, 24190-24195.

Katzman, M., Katz, R. A., Skalka, A. M. & Leis, J. (1989). The avian retroviral integration protein cleaves the terminal sequences of linear viral DNA at the in vivo sites of integration. *J. Virol.* **63**, 5319-5327.

Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946-950.

Kulkosky, J., Katz, R. A., Merkel, G. & Skalka, A. M. (1995). Activities and substrate specificity of the

evolutionarily conserved central domain of retroviral integrase. *Virology,* **206**, 448-456.

Lodi, P. J., Ernst, J. A., Kuszewski, J., Hickman, A. B., Engelman, A., Craigie, R., Clore, G. M. & Gronenborn, A. M. (1995). Solution structure of the DNA binding domain of HIV-1 integrase. *Biochemistry,* **34**, 9826-9833.

Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature,* **389**, 251-260.

Maignan, S., Guilloteau, J. P., Zhou-Liu, Q., Clement-Mella, C. & Mikol, V. (1998). 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.* **282**, 359-368.

Merritt, E. A. & Bacon, D. J. (1997). Raster3D: photorealistic molecular graphics. *Methods Enzymol.* **277**, 505-524.

Miller, M. D., Farnet, C. M. & Bushman, F. D. (1997). Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. *J. Virol.* **71**, 5382-5390.

Muller, H.-P. & Varmus, H. E. (1994). DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO J.* **13**, 4704-4714.

Mumm, S. R. & Grandgenett, D. P. (1991). Defining nucleic acid-binding properties of avian retrovirus integrase by deletion analysis. *J. Virol.* **65**, 1160-1167.

Navaza, J. & Saludjian, P. (1997). AMoRe: an automated molecular replacement program package. *Methods Enzymol.* **276**, 581-594.

Nicholls, A., Sharp, K. A. & Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Struct. Funct. Genet.* **11**, 281-296.

Otwinowski, Z. & Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307-326.

Patel, P. H. & Preston, B. D. (1994). Marked infidelity of human immunodeficiency virus type 1 reverse transcriptase at RNA and DNA template ends. *Proc. Natl Acad. Sci. USA,* **91**, 549-553.

Plasterk, R. H. A. (1995). The HIV integrase catalytic core. *Struct. Biol.* **2**, 87-90.

Pruss, D., Bushman, F. D. & Wolffe, A. P. (1994a). Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc. Natl Acad. Sci. USA,* **91**, 5913-5917.

Pruss, D., Reeves, R., Bushman, F. D. & Wolffe, A. P. (1994b). The influence of DNA and nucleosome structure on integration events directed by HIV integrase. *J. Biol. Chem.* **269**, 25031-25041.

Pryciak, P. M. & Varmus, H. E. (1992). Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. *Cell,* **69**, 769-780.

Pryciak, P. M., Sil, A. & Varmus, H. E. (1992). Retroviral integration into minichromosomes *in vitro*. *EMBO J.* **11**, 291-303.

Puras, Lutzke R. A., Vink, C. & Plasterk, R. H. A. (1994). Characterization of the minimal DNA-binding domain of the HIV integrase protein. *Nucl. Acids Res.* **22**, 4125-4131.

Sack, J. S. & Quiocho, F. A. (1997). CHAIN: a crystallographic modeling program. *Methods Enzymol.* **277**, 158-173.

Scottoline, B. P., Chow, S., Ellison, V. & Brown, P. O. (1997). Disruption of the terminal base pairs of retroviral DNA during integration. *Genes Dev.* **11**, 371-382.

Sheldrick, G. M. & Schneider, T. R. (1997). SHELXL: high-resolution refinement. *Methods Enzymol.* **277**, 319-343.

Sherman, P. A. & Fyfe, J. A. (1990). Human immunodeficiency virus integration protein expressed in Escherichia coli possesses selective DNA cleaving activity. *Proc. Natl Acad. Sci. USA,* **87**, 5119-5123.

Terwilliger, T. C. & Berendzen, J. (1999). Automated MAD and MIR structure solution. *Acta Crystallog. sect. D,* **55**, 849-861.

van Gent, D. C., Elgersma, Y., Bolk, M. W. J., Vink, C. & Plasterk, R. H. A. (1991). DNA binding properties of the integrase proteins of human immunodeficiency viruses types 1 and 2. *Nucl. Acids Res.* **19**, 3821-3827.

Vink, C., Oude Groeneger, A. M. & Plasterk, R. H. A. (1993). Identification of the catalytic and DNA-binding region of the human immunodeficiency virus type 1 integrase protein. *Nucl. Acids Res.* **21**, 1419-1425.

Yang, W. & Steitz, T. A. (1995). Recombining the structures of HIV integrase, RuvC and RNase H. *Curr. Biol.* **3**, 131-134.

Yang, W., Hendrickson, W. A., Crouch, R. J. & Satow, Y. (1990). Structure of ribonuclease H phased at 2 Å resolution by MAD analysis of the selenomethionyl protein. *Science,* **249**, 1398-1405.

Edited by I. A. Wilson

(Received 8 October 1999; received in revised form 8 December 1999; accepted 8 December 1999)