MOLECULAR AND CELLULAR BIOLOGY, May 1992, p. 2331–2338
0270-7306/92/052331-08$02.00/0
Copyright © 1992, American Society for Microbiology

Vol. 12, No. 5

Residues Critical for Retroviral Integrative Recombination in a Region That Is Highly Conserved among Retroviral/Retrotransposon Integrases and Bacterial Insertion Sequence Transposases

JOSEPH KULKOSKY,[1] KATHRYN S. JONES,[1] RICHARD A. KATZ,[1] JOSEPH P. G. MACK,[2] AND ANNA MARIE SKALKA[1]*

Fox Chase Cancer Center, Institute for Cancer Research, Philadelphia, Pennsylvania 19111,[1] and Crystallography Laboratory, National Cancer Institute-Frederick Cancer Research and Development Center, Frederick, Maryland 21702[2]

Received 25 November 1991/Accepted 21 February 1992

Our comparison of deduced amino acid sequences for retroviral/retrotransposon integrase (IN) proteins of several organisms, including *Drosophila melanogaster* and *Schizosaccharomyces pombe*, reveals strong conservation of a constellation of amino acids characterized by two invariant aspartate (D) residues and a glutamate (E) residue, which we refer to as the D,D(35)E region. The same constellation is found in the transposases of a number of bacterial insertion sequences. The conservation of this region suggests that the component residues are involved in DNA recognition, cutting, and joining, since these properties are shared among these proteins of divergent origin. We introduced amino acid substitutions in invariant residues and selected conserved and nonconserved residues throughout the D,D(35)E region of Rous sarcoma virus IN and in human immunodeficiency virus IN and assessed their effect upon the activities of the purified, mutant proteins in vitro. Changes of the invariant and conserved residues typically produce similar impairment of both viral long terminal repeat (LTR) oligonucleotide cleavage referred to as the processing reaction and the subsequent joining of the processed LTR-based oligonucleotides to DNA targets. The severity of the defects depended upon the site and the nature of the amino acid substitution(s). All substitutions of the invariant acidic D and E residues in both Rous sarcoma virus and human immunodeficiency virus IN dramatically reduced LTR oligonucleotide processing and joining to a few percent or less of wild type, suggesting that they are essential components of the active site for both reactions. On the basis of similarities with enzymes that catalyze analogous reactions, we propose that the invariant D and E residues may participate in coordination of the metal cofactor (Mn^{2+} or Mg^{2+}) required for the catalytic activities of IN. We further speculate that a metal-DNA complex may be necessary to position both LTR and target DNA substrates for nucleophilic attack during the cleavage and joining reactions.

Retroviral integration is dependent upon the interaction of three macromolecular components: (i) a virus-encoded protein, integrase (IN), (ii) specific sequences at the long terminal repeat (LTR) termini of viral "donor" DNA, and (iii) a host DNA "target." Results from the study of several different retroviral systems reveal a stepwise pathway for the interaction of these components during integration in a natural virus infection: (i) removal of a dinucleotide from the 3′ ends of linear viral DNA (4) in the cytoplasm, which we refer to as the processing reaction, (ii) migration of the viral DNA to the cell nucleus within a higher-order protein complex (2), (iii) staggered cleavage of host DNA generating 5′-strand extensions, likely coupled to single-strand joining of the processed 3′ hydroxyl ends of viral DNA to the 5′ phosphate ends of the host target DNA, and (iv) repair and ligation of the gapped unjoined opposite strands. This final step produces a short duplication of host sequences flanking the integrated proviral DNA (for reviews, see references 24, 29, 32, and 33).

The role of IN in processing viral DNA has been established. In vivo, this site-specific endonucleolytic activity of IN was inferred from the observation that the production of recessed 3′ LTR ends is defective in viruses that bear mutations restricted to the IN domain (26). Direct biochemical evidence that the endonucleolytic activity of IN is responsible for processing was first obtained in vitro using a purified preparation of the Rous sarcoma virus (RSV) IN protein and short duplex oligodeoxynucleotide substrates whose sequences represented either of the LTR ends of RSV DNA (20). In this system, IN specifically removed 2 terminal nucleotides from the 3′ ends of these oligodeoxynucleotide substrates, as expected from the processing event observed in vivo (3, 26). Heterogeneous products were also produced in this in vitro reaction, some of which were greater in length than the uncleaved substrate. Such products were shown to arise from the IN-catalyzed joining of the processed strand of one oligonucleotide to various sites in another oligonucleotide which, in this case, is a presumed surrogate for host target DNA (8, 19). This assay and other assays, based upon genetic selection (6, 8, 19), have unequivocally demonstrated that IN alone is sufficient for the major steps of the retroviral recombination reaction: viral DNA processing, host DNA cleavage, and single-strand joining of virus and host DNA.

Although the multifunctional role of IN has been established, the region(s) of the protein that are associated with its various activities has not yet been clearly defined. We previously utilized computer-assisted alignment of the deduced amino acid sequences of 80 retroviral/retrotransposon INs and other proteins with related activities to map conserved domains of the protein. Our alignments and those of

* Corresponding author.

others (13, 16, 27), together with deletion analysis of RSV IN (22), highlighted a functional, highly conserved 50-amino-acid domain, which we referred to as D(35)E, since it is composed of invariant aspartate (D) and glutamate (E) residues consistently separated by 35 amino acids. Further inspection, prompted by sequence comparisons with the bacterial transposon Tn*552* (27), suggested that the D(35)E domain may be part of a larger constellation of conserved residues in retroviral/retrotransposon IN proteins and certain insertion sequence (IS) element transposases. We refer to this larger constellation as the D,D(35)E region, emphasizing the three invariant acidic amino acids. It seems likely that residues in this region contribute to the common functions of DNA recognition, cutting, and joining shared by these groups of proteins. If so, substitution of key amino acids within it should affect one or more of these activities significantly. To test this hypothesis, we introduced several amino acid substitutions in invariant, selected conserved, and selected nonconserved residues within the D,D(35)E regions of RSV IN and human immunodeficiency virus (HIV) IN. The results provide experimental evidence for the importance of the invariant acidic residues, which proved to be exquisitely sensitive to even conservative replacements. In contrast, conserved residues are somewhat less critical, since their replacement leaves significant residual activity. We suggest functions for some of the residues in the D,D(35)E region on the basis of biochemical and structural information available for enzymes with related activities.

MATERIALS AND METHODS

Amino acid sequence alignment. Sequences were obtained from the Genpept (release 64.3), EMBL (release 27.0), and PIR-Protein (release 29.0) data bases (all current to July 1991) using the Genetics Computer Group Sequence Analysis Software Package, version 7.0 (11). A phylogenetically representative subset of the known retroviral IN proteins was used to find sequences similar to that of the region of RSV IN between residues W-61 and K-164, using version 4.40 of the Profile package of programs (14). The most closely related sequences were IN proteins from retrotransposons and transposases of ISs. These sequences were aligned with computer assistance (15). The phylogenetic tree distance data (in Felsenstein format) was converted to postscript format using a program written by J. P. G. Mack. The alignment was converted to postscript format by using a program based on routines written by Michael Gribskov. The alignment of phylogenetically representative members of the set of retroelements and ISs for the residues near D-64 and near the D(35)E domain (D-121 to E-157) of RSV IN is shown in Fig. 1.

The source of retroelement sequences has been described previously (21, 22). Figure 1 shows a subset of the ISs used in the alignment (sequence name, organism, accession number): IS*2*, *Escherichia coli*, JQ0040; IS*3*, *E. coli*, TQECI3; IS26, *Proteus vulgaris*, X00011; IS136, *Agrobacterium tumefaciens*, X04282; IS*240*-A, *Bacillus thuringiensis*, M23740; IS*629*, *Shigella sonnei*, P16942; IS*861*, *Streptococcus agalactiae*, A30868; IS986, *Mycobacterium tuberculosis*, P19774; orf-w2, *Lactococcus lactis*, M37396. Other proteins with D,D(35)E or related motifs, not shown in the alignment, are *pilB* from *Neisseria gonorrhoeae* (30) and a segment of the West Nile virus polyprotein (7) which is C terminal to the VI protein of this flavivirus.

Oligonucleotide-directed mutagenesis. Synthetic DNA oligonucleotides that incorporated a restriction site to verify the change were used to direct specific amino acid substitutions in residues throughout the D,D(35)E region of RSV and HIV IN employing conventional procedures described elsewhere (22). The sequences of the mutagenic oligonucleotides for RSV IN were as follows:

W61L: 5′-GGGACCCCTACAGATACTGCAGACAGACTTTACG-3′
*Pst*I
W61F: 5′-GTTTGGGACCCCTACAATATTCAGACAGACTTTACGC-3′
*Ssp*I
D64E: 5′-AGATATGGCAGACAGAATTCCACGCTTGAGCCTAGAA-3′
*Eco*RI
D121A: 5′-AAGGCCATAAAAACAGCTAATGGATCCTGCTTCACGTCTA-3′
*Bam*HI
D121E: 5′-AAGGCCATAAAAACAGAAAATGGATCCTGCTTCACGTCTA-3′
*Bam*HI
F126A: 5′-CAGATAATGGGTCCTGCGCGACGTCTAAATCCACGC-3′
*Aha*II
E133A: 5′-CAGGTCTAAATCCACTCGAGCGTGGCTCGCGAGATGG-3′
*Xho*I
E157A: 5′-GTCAAGCTATGGTAGCCCGGGCCAACCGGCTC-3′
*Sma*I
E157D: 5′-AGGGTCAAGCTATGGTCGATCGGGCCAACCGGCTCC-3′
*Pvu*I
K164A: 5′-AGCGGGCCAACCGGCTGCTAGCAGATAGGATCCGTGTG-3′
*Nhe*I *Bam*HI

The sequences of the mutagenic oligonucleotides for HIV IN were as follows:

T115S: 5′-CCAGTAAAAACAATACACTCAGATAATGGGAGCAATTTCA–3′
*Dde*I
D116E: 5′-GTAAAAACAATACATACTGAGAATGGCAGCAATTTCA-3′
*Dde*I
F121A: 5′-CAGACAATGGCAGCAATGCAACTAGTGCTACGGTTAAGG-3′
*Spe*I
E152A: 5′-AAGTCAAGGAGTAGTAGCATCGATGAATAAAGAATTAAAG-3′
*Cla*I
K159A: 5′-CAAGGAGTAGTAGAATCGATGAATGCAGAATTAAAGAA-3′
*Cla*I

Construction of RSV IN double mutant proteins D121E E1571 and D121A E157A. The double mutants were constructed by recombining restriction fragments from the plasmids containing the appropriate single mutations. Digestion of pRIT2T IN with *Nru*I generates two IN fragments: a large fragment that includes D-121 and a smaller fragment that includes E-157. To construct double mutant protein D121E, E157D, the small *Nru*I-*Nru*I fragment of D121E was replaced with the analogous fragment from E157D. To construct double mutant protein D121A, E157A, the small *Nru*I-*Nru*I fragment of D121A was replaced by the analogous fragment from E157A.

Vector construction, expression, and purification of RSV and HIV IN fusion proteins. The expression vector and purification of the fusion product of protein A joined to RSV IN were as previously described (22). For HIV IN, the amino terminus was modified to incorporate a *Stu*I site; this resulted in substitution of the N-terminal phenylalanine residue of IN to proline. The *Stu*I site was then used to insert a *Stu*I-*Hpa*I fragment containing the HIV IN sequences into the *Stu*I site in the polylinker region of the maltose-binding protein (MBP) expression vector, pMAL-c, obtained from New England BioLabs. The maltose binding protein fusion with HIV IN is referred to as pMBP-IN. Overnight cultures of pMAL-c (control) and pMBP-IN (or its mutated derivatives) were diluted 1:100 in 150 ml of L broth and grown at 37°C to an optical density at 600 nm of 1. Expression of the protein was induced by addition of isopropyl-β-D-thiogalactopyranoside to a final concentration of 1 mM. The pellet of the 150-ml culture was resuspended in 5 ml of buffer containing 10 mM sodium phosphate, pH 7.2–1 M NaCl–1 mM β-mercaptoethanol–1 mM EDTA and then sonicated. The cleared supernatant fraction from a high-speed centrifugation was diluted with salt-free buffer to 0.5 M NaCl and then passed through a column containing 1.5 ml of amylose resin. Wash and elution of the column was as recommended by New England BioLabs. The purified protein was dialyzed in a solution containing 0.5 M NaCl, 10 mM Tris (pH 7.5), 1 mM EDTA, 1 mM β-mercaptoethanol, and 40% glycerol.

Protein concentrations were determined spectrophotometrically using a colorimetric protein staining assay supplied by Bio-Rad. The purity of each preparation (typically >80%) was then estimated by gel analysis. The amounts of IN fusion preparations added to each reaction were normalized to contain equal amounts of full-length protein.

Construction, expression, and purification of nonfusion RSV IN protein. Construction of RSV IN mutants as nonfusion proteins was accomplished by exchange of a *Bss*HII-*Kpn*I fragment carrying the relevant mutation for the equivalent fragment in the wild-type IN expression vector, pRC23-p32 previously described (22, 31). Expression and purification of the nonfused proteins were as described previously (22, 31) except elution by gradient was replaced with step elution from phosphocellulose using NaCl at 0.2, 0.4, and 1 M. The 1 M NaCl fraction containing the IN protein was diluted fourfold and applied to a 0.5-ml poly(U) Sepharose column, and eluted in steps with 0.2, 0.4, and 1 M NaCl in a 0.01 M Tris (pH 7.5) buffer containing 10% glycerol.

Oligonucleotide processing and joining assays. Standard conditions described by Katzman et al. (20) were used to assay 1- to 2-pmol samples of purified fusion proteins or nonfused RSV IN proteins. Incubation was at 37°C in 2 mM Mn^{2+} for the times indicated in the figures. For HIV MBP-IN, 20 pmol of protein and 1 pmol of annealed 25-mer DNA substrate representing the U5 end of HIV LTR sequences were incubated at 37°C in a 15-μl reaction mixture which contained 10 mM Tris (pH 7.5), 50 mM NaCl, 10 mM $MnCl_2$, and 6% dimethyl sulfoxide. Products of both RSV and HIV reactions were analyzed on 20% polyacrylamide gels containing 7 M urea. The amount of product was quantitated by scanning the radioactivity present in the appropriate bands using the Radioanalytic Imaging System (AMBIS, San Diego, Calif.) after size fractionation on the denaturing polyacrylamide gels.

Circle joining assay. Conditions for the circle joining assay were identical to the oligonucleotide assay, except that the U3 oligonucleotides have 2-nucleotide recessed 3′ OH (preprocessed) ends, and 10 ng of a small (1.3-kb) supercoiled plasmid (pAO3) was added to the reaction. The reaction products were fractionated on a 1.4% Tris-borate-EDTA agarose gel containing 1 μg of ethidium bromide per ml.

RESULTS

Effect of amino acid substitutions in the D,D(35)E region on viral DNA processing. As shown in Fig. 1, the D,D(35)E region is characterized by invariant acidic residues present in analogous locations in the IN proteins of all retroelements and in the transposases of the indicated IS elements. For RSV IN, these invariant residues are amino acids D-64, D-121, and E-157, and in HIV they are D-64, D-116, and E-152. Adjacent residues are also highly conserved (shaded in Fig. 1). These include the hydrophobic tryptophan (W-61), the four amino acids surrounding D-121, the hydrophobic phenylalanine (F) 5 amino acids following D-121, a hydrophobic isoleucine (I) in the middle of the D(35)E domain, and finally a lysine (K), 7 amino acids following the last invariant residue, E-157. To test their importance for the processing and joining activities of IN, we made substitutions in each of the invariant residues and in selected conserved residues in both RSV and HIV IN proteins. As a control, a nonconserved residue in this region of RSV IN was also changed.

For experiments with RSV, we introduced amino acid changes into a fusion protein in which IN sequences are joined to the C terminus of the *Staphylococcus* protein A. All of the fusion proteins, as well as nonfused protein A expressed from the same vector, were then purified from the bacterial extracts on the basis of selective affinity to immunoglobulin G. This procedure was chosen to standardize the purification and to minimize contamination with bacterial nucleases, which are not expected to bind to immunoglobulin G. We previously reported that the wild-type version of this fusion protein has DNA-binding and endonuclease activities characteristic of nonfused RSV IN protein (22). Figure 2 shows a denaturing gel analysis of substrate, and the products formed as a function of time of incubation with the protein A-IN wild-type fusion protein and one of the mutant proteins, F126A. In both cases, the conversion of substrate to processed −2 product continued to increase with time. The rate was linear for F126A, even up to 3.5 h. Similar kinetics were observed with other mutant IN fusion proteins. However, quantitation by radioisotope scanning indicated that the amount of −2 product in the wild-type reaction begins to plateau between 2 and 3 h, as the amount of joined product increases. Therefore, we chose a 2-h time point to compare the processing activities of the mutants with that of the wild-type protein.

As illustrated in Fig. 3 (lanes 10 and 14), substitution of invariant acidic residues D-121 or E-157 with alanine (A), which contains only a methyl group in its side chain, resulted in substantial loss in the production of the specific −2 processed product. We estimate by radioisotope scanning that the amount was reduced to 3% or less of wild type, depending upon assay conditions. Conservative replacements at these positions, which retain the carboxylates but produce single methylene shifts in the position of the carboxyl groups, D-64→E (lane 6), D-121→E (lane 11), or E-157→D (lane 15), also resulted in loss of activity, some even more pronounced than that observed with the alanine replacements (compare E-157→A and E-157→D [lanes 14 and 15]). A double mutation representing an exchange of the acidic residues at positions 121 and 157 (D-121→E, E-157→D [lane 17]) also resulted in severely defective cleavage at the −2 processing site. Thus, it appears that the precise positioning of the carboxyl groups of these side chains within this conserved region and relative to one another is critical for cleavage specificity. As expected, IN with alanine substitutions in both D-121 and E-157 (D-121→A, E-157→A) showed no activity over background (Fig. 3, lane 18). The small amount of −1 product, seen also in the protein A controls, probably reflects the presence of contaminating bacterial nucleases.

Substitution of the conserved residues, phenylalanine (F) at RSV position 126 and lysine (K) at position 164, was less detrimental to processing activity. We observed approximately sixfold reductions with these neutral alanine substitutions compared with the wild type (Fig. 3, compare lanes 9, 12, and 16). For a control to confirm the significance of substitutions in this region, we tested the effect of replacing a nonconserved residue, E-133, in the RSV D(35)E domain.

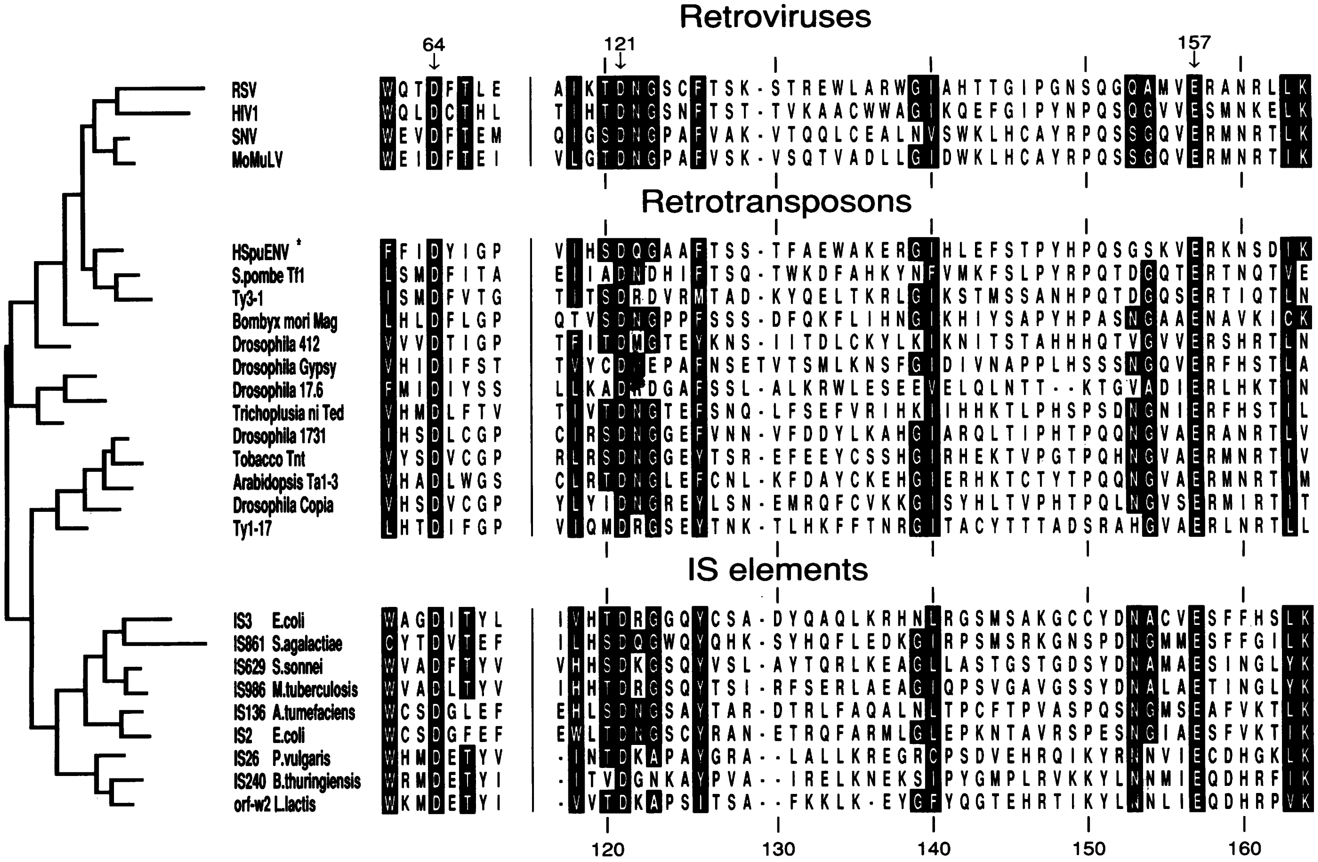
Retroviruses
64
121
157
RSV WQTDFTLE AIKTDNGSCFTSK-STREWLARWGIAHTTGIPGNSQGQAMVERANRLLK
HIV1 WQLDCTHL TIHTDNGSNFTST-TVKAACWWAGIKQEFGIPYNPQSQGVVESMNKELK
SNV WEVDFTEM QIGSDNGPAFVAK-VTQQLCEALNVSWKLHCAYRPQSSGQVERMNRTLK
MoMuLV WEIDFTEI VLGTDNGPAFVSK-VSQTVADLLGIDWKLHCAYRPQSSGQVERMNRTIK
Retrotransposons
HSpuENV * FFIDYIGP VIHSDQGAAFTSS-TFAEWAKERGIHLEFSTPYHPQSGSKVERKNSDIK
S.pombe Tf1 LSMDFITA EIIADNDHIFTSQ-TWKDFAHKYNFVMKFSLPYRPQTDGQTERTNQTVE
Ty3-1 ISMDFVTG TITSDRDVRMTAD-KYQELTKRLGIKSTMSSANHPQTDGQSERTIQTLN
Bombyx mori Mag LHLDFLGP QTVSDNGPPFSSS-DFQKFLIHNGIKHIYSAPYHPASNGAAENAVKICK
Drosophila 412 VVVDTIGP TFITDMGTEYKNS-IITDLCKYLKIKNITSTAHHHQTVGVVERSHRTLN
Drosophila Gypsy VHIDIFST TVYCDNEPAFNSETVTSMLKNSFGIDIVNAPPLHSSSNGQVERFHSTLA
Drosophila 17.6 FMIDIYSS LLKADRDGAFSSL-ALKRWLESEEVELQLNTT--KTGVADIERLHKTIN
Trichoplusia ni Ted VHMDLFTV TIVTDNGTEFSNQ-LFSEFVRIHKIIHHKTLPHSPSDNGNIERFHSTIL
Drosophila 1731 IHSDLCGP CIRSDNGGEFVNN-VFDDYLKAHGIARQLTIPHTPQQNGVAERANRTLV
Tobacco Tnt VYSDVCGP RLRSDNGGEYTSR-EFEEYCSSHGIRHEKTVPGTPQHNGVAERMNRTIV
Arabidopsis Ta1-3 VHADLWGS CLRTDNGLEFCNL-KFDAYCKEHGIERHKTCTYTPQQNGVAERMNRTIM
Drosophila Copia VHSDVCGP YLYIDNGREYLSN-EMRQFCVKKGISYHLTVPHTPQLNGVSERMIRTIT
Ty1-17 LHTDIFGP VIQMDRGSEYTNK-TLHKFFTNRGITACYTTTADSRAHGVAERLNRTLL
IS elements
IS3 E.coli WAGDITYL IVHTDRGGQYCSA-DYQAQLKRHNLRGSMSAKGCCYDNACVESFFHSLK
IS861 S.agalactiae CYTDVTEF ILHSDQGWQYQHK-SYHQFLEDKGIRPSMSRKGNSPDNGMMESFFGILK
IS629 S.sonnei WVADFTYV VHHSDKGSQYVSL-AYTQRLKEAGLLASTGSTGDSYDNAMAESINGLYK
IS986 M.tuberculosis WVADLTYV IHHTDRGSQYTSI-RFSERLAEAGIQPSVGAVGSSYDNALAETINGLYK
IS136 A.tumefaciens WCSDGLEF EHLSDNGSAYTAR-DTRLFAQALNLTPCFTPVASPQSNGMSEAFVKTLK
IS2 E.coli WCSDGFEF EWLTDNGSCYRAN-ETRQFARMLGLEPKNTAVRSPESNGIAESFVKTIK
IS26 P.vulgaris WHMDETYV -INTDKAPAYGRA--LALLKREGRCPSDVEHRQIKYRNNVIECDHGKLK
IS240 B.thuringiensis WRMDETYI -ITVDGNKAYPVA--IRELKNEKSIPYGMPLRVKKYLNNMIEQDHRFIK
orf-w2 L.lactis WKMDETYI -VVTDKAPSITSA--FKKLK-EYGFYQGTEHRTIKYLNNLIEQDHRPVK
120
130
140
150
160

FIG. 1. Alignment of amino acids in the D,D(35)E region of retroviruses, retrotransposons, and bacterial ISs. Sequences are grouped by phylogenetic relatedness as shown in the tree on the left side of the panel. Only representative members of the more than 80 proteins analyzed are included. Amino acids are designated in single-letter code with the highly conserved amino acids indicated by white letters on a black background and less conserved adjacent amino acids indicated by shading. Numbers at top represent amino acid positions of the residues in RSV IN. HSpuENV is a human endogenous spumavirus which, as indicated, appears to be more closely related to retrotransposons than to other retroviruses. SNV, spleen necrosis virus; MoMLV, Moloney murine leukemia virus.

This replacement, E-133→A, had the least detrimental effect on processing (Fig. 3, lanes 9 versus 13). Substitution of tryptophan, W-61 in RSV IN with a small hydrophobic residue, leucine (L), or a more conservative change to phenylalanine, resulted in mutant proteins with no processing activity above background (Fig. 3, lanes 4 and 5). All of the proteins included in Fig. 3 were tested for their ability to bind DNA [oligo(dAT)] in our standard nitrocellulose assay (22, 31). All bound with at least 80% the efficiency of wild type, with exception of the E157 mutants, which were about half as efficient as wild type, and W61F and W61L, whose binding was not above background. This inability to bind DNA may account for the lack of detectable processing activity for the latter two proteins.

A similar strategy was used to test substitutions in HIV IN. In this case, the IN coding sequences were joined to the 3′ end of a gene encoding MBP and amylose resin was used to rapidly purify wild-type IN fusion protein and its mutant versions in a standardized fashion. As shown in Fig. 4A, the fusion protein (MBP-IN) is purified to near homogeneity by a single passage of the soluble bacterial lysate through the amylose resin. Similarly purified MBP was used as a control. Immunoblot analyses (not shown) indicated that the faster-migrating proteins, which are coeluted with MBP-IN from the column, represent degradation products of the full-length

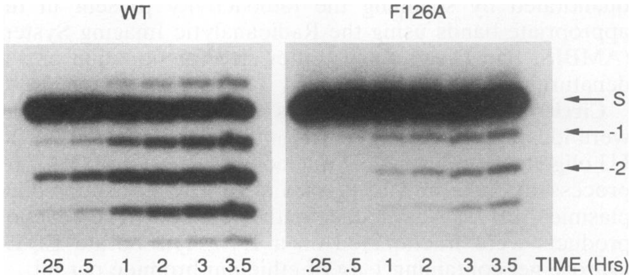

FIG. 2. Time course for processing activity of RSV protein A-IN fusion proteins. Purified protein A-IN fusion proteins were assayed for processing activity using an oligodeoxynucleotide duplex substrate which represents terminal sequences at the U3 end of RSV DNA as described in Materials and Methods. At the times indicated, products of the reactions were separated in a 20% denaturing gel. The −2 band, which contains molecules two nucleotides shorter than the substrate, corresponds to the expected processed product which is subsequently joined to target substrates (18). Protein A-IN wild-type fusion protein (WT) and F126A (protein A-IN with substitution of A for F at position 126 of RSV IN) are shown. To enhance visibility of the products of F126A in the figure, exposure time for autoradiography of the F126A results was twice that of the wild-type protein.

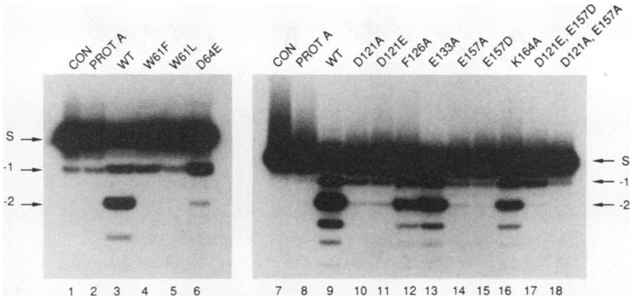

FIG. 3. Viral DNA processing activity of RSV protein A-IN fusion proteins. Purified protein A-IN mutant fusion proteins were assayed for processing activity as described in the legend to Fig. 2. Products of 2 h reactions were separated in a 20% denaturing gel. The −2 band, which contains molecules 2 nucleotides shorter than the substrate(s), corresponds to the expected processed product. CON, control (no protein); PROT A, protein A (control); WT, protein A-IN wild-type fusion protein.

fusion proteins. The estimated lower limit in molecular mass of affinity-purified product is 42 kDa; this protein comigrates with MBP. Generation of small amounts of degradation products is not uncommon with this expression system apparently as a result of protease susceptibility of the fusion polypeptides. Nevertheless, as with RSV, the purified HIV IN wild-type fusion protein exhibited DNA processing activity characteristic of the nonfused version (28), producing a −2 product preferentially (Fig. 4B, lane 2).

As illustrated in Fig. 4B, substitution mutants MBP-IN F121A and K159A exhibit similar levels of processing; both only about 10% or less of wild type. This is somewhat lower than the level of activity observed for the analogous mutants (F126A and K164A) of RSV IN. As with RSV, the HIV D116E mutant is more severely defective (lane 4). In contrast to its RSV counterpart, the HIV E152A substitution mutant (lane 6) seemed to retain some processing activity. Nevertheless, this mutant is also significantly reduced in activity compared with its wild-type counterpart. The most active substitution in the HIV series is T-115→S (Fig. 4B, lane 3). As illustrated in Fig. 1, threonine (T) or serine (S) is conserved at this position throughout the retrovirus family and the analogous substitution in RSV IN (T-120→S) also exhibits near wild-type levels of activity (18).

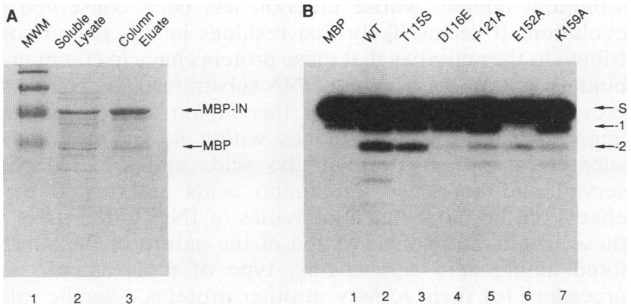

FIG. 4. Purification and activity of HIV IN fusion proteins. (A) Protein gel analysis of crude lysate and purified HIV IN fusion protein. Proteins are detected by Coomassie blue staining. Lanes: 1, molecular weight markers (MWM); 2, proteins in the soluble fraction of *E. coli* lysate; 3, proteins purified by single passage of the crude lysate through amylose affinity resin. (B) Processing activity of HIV MBP-IN substitution mutants. Proteins purified as shown in panel A were assayed for processing activity, using an oligodeoxynucleotide duplex substrate which represents the 25 terminal base pairs at the U5 end of HIV DNA. Because the HIV IN fusion protein is less active than that of RSV (Fig. 2), products of a 3 h incubation were compared.

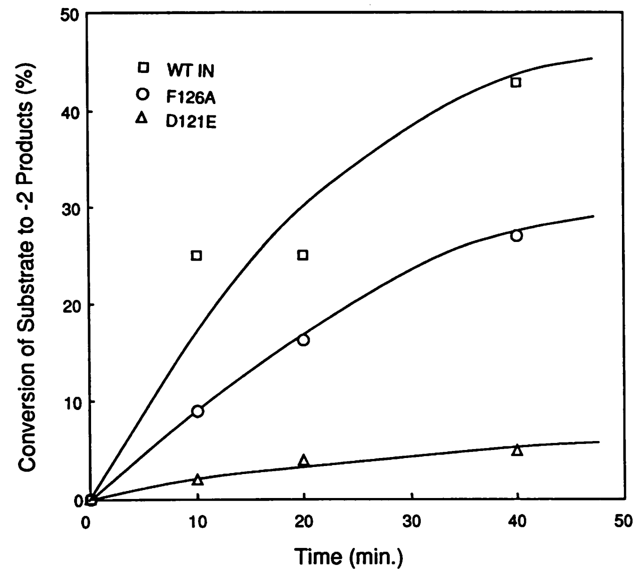

FIG. 5. Processing activities of nonfused RSV IN substitution mutants proteins. Reaction conditions were identical to those described in the legend to Fig. 2. Samples of the reaction were withdrawn and analyzed at the times indicated. Conversion of substrate to −2 product was determined from the amount of radioactivity present in the relevant band, detected by isotope scanning and expressed as a percentage of total counts in the reaction.

Relative processing activities of substituted proteins are retained in nonfused proteins. We have observed that the RSV IN fusion proteins are typically 5- to 10-fold less active than the nonfused forms under equivalent assay conditions (22). In order to determine whether the relative processing activities of the mutants were influenced by linkage to protein A, we analyzed two of the mutant proteins, D121E and F126A, in a nonfused form. With D121E, activity was barely detected at early times in the incubation (Fig. 5). Even after 40 min, less than 5% of the substrate was processed to form the −2 product. The F126A nonfusion protein was more active, producing approximately 30 to 50% of the amount of product observed with wild type at the various times tested. The data in Fig. 5 confirm our earlier observation that the nonfused IN proteins process the substrate more rapidly than their fusion counterparts, requiring only 20 to 30 min rather than 2 to 3 h to convert equivalent amounts of substrate. Nevertheless, the relative amounts of processed product formed by the wild type and by mutants (i.e., wild type IN > F126A > D121E) are similar for both forms of the protein.

Effect of amino acid substitutions in the D,D(35)E region on DNA strand joining. By using LTR oligonucleotide substrates whose termini are identical to the processed viral DNA (preprocessed ends), it is possible to measure the joining activity of IN independently of processing. Joining of the preprocessed 3′ ends to other oligonucleotides in the same reaction produces a population of products that are longer than the substrate. As shown in Fig. 6 (lane 3), joined oligonucleotides are clearly visible in the reaction catalyzed

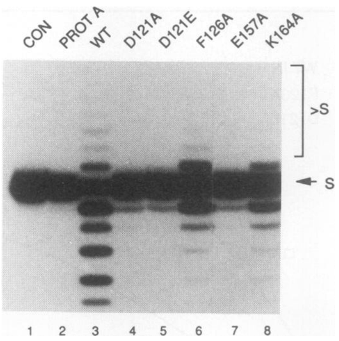

FIG. 6. DNA joining activity of protein A-RSV IN fusion proteins. Purified protein A-IN fusion proteins were assayed for DNA joining activity using an oligodeoxynucleotide duplex substrate with a 3′ recessed terminus that represents the processed U3 end of RSV DNA. The reaction was for 2 h. Joined products (>S) are separated from substrate (S) and smaller cleavage products in a 20% denaturing gel. Lanes: 1, control (CON) (no protein); 2, control (protein A [PROT A]); 3, protein A-IN wild-type fusion protein (WT); 4 to 8, protein A-IN substitution mutants as indicated above the lanes.

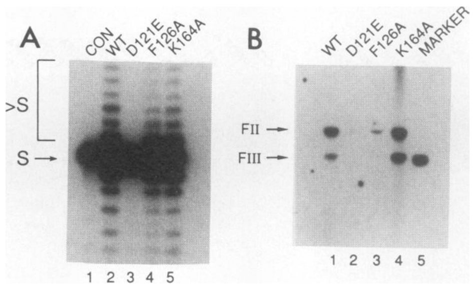

FIG. 7. Joining assay using supercoiled plasmid DNA target. Nonfused RSV IN proteins were tested in these reactions. Conditions were as described in the legend to Fig. 6, except that 10 ng of a 1.3-kb supercoiled plasmid (pAO3) was added to the reaction and incubation was for 30 min. The products were deproteinized by phenol extraction and then fractionated on a 20% polyacrylamide gel (A) or a 1.2% agarose gel (B). FIII; linear products resulting from the presumed concerted joining of two oligonucleotides to the plasmid; FII; relaxed circles, the result of a single oligonucleotide joining. Lane 5, ^{32}P-labelled pAO3 linearized by digestion with *Eco*RI.

by the RSV protein A-IN wild-type fusion protein. No joined products were detected with either of the D-121 substitutions or with the E-157→A substitution. Joining by HIV MBP-IN D116E was impaired to a similar degree relative to HIV MBP-IN wild-type fusion protein in a comparable assay (data not shown). As with processing, the F126A (lane 6) and K164A (lane 8) fusion proteins displayed intermediate joining activities of approximately 50 and 30% of wild type, respectively.

Joining of preprocessed viral oligonucleotides to targets of various sequence and structure can also be tested by addition of other target DNAs to the joining assay. Heterologous plasmid DNA is a particularly useful target, since joined products of the reaction can be detected with great sensitivity. The background is negligible, and instead of a population of different length products, only two products are formed, both easily distinguished: (i) linear DNA, likely the product of concerted joining of two oligonucleotides to plasmid DNA and (ii) relaxed DNA with an oligonucleotide "tail," the product of joining of one oligonucleotide to a single end of the nicked supercoiled DNA. We used the nonfused forms of the proteins characterized in Fig. 5 for these tests so that the maximal activities could be measured. Since the amount of plasmid target DNA in the reaction is very small, oligonucleotide-to-oligonucleotide joining can be measured simultaneously in this assay by electrophoresis of a portion of the reaction in a sequencing gel, as in Fig. 6. With respect to oligonucleotide-to-oligonucleotide joining (Fig. 7A), the nonfused mutant proteins exhibited relative activities similar to the protein A fusions shown in Fig. 6. Both F126A (Fig. 7A, lane 4) and K164A (lane 5) were approximately two- to threefold less active than the wild-type nonfused protein. No joined oligonucleotide products were detected with D121E (Fig. 7A, lane 3). The assay to detect oligonucleotides joined to plasmid DNA (Fig. 7B) again showed D121E to be virtually inactive; only a trace amount of radioactivity was detected at the position of relaxed plasmid DNA (Fig. 7B, lane 2). However, the behavior of the two other mutants tested was not as expected from the oligonucleotide joining results. RSV IN K164A, which showed a decrease in the oligonucleotide joining assay, was slightly more active than the wild-type IN in the plasmid assay (Fig. 7B, lane 4). On the other hand, F126A, which was only slightly less active than wild type in the oligonucleotide assays in either fused form (Fig. 6, lane 6) or nonfused form (Fig. 7A, lane 4), was much less active in joining the preprocessed LTR oligonucleotides to the supercoiled target plasmid (Fig. 7B, lane 3). These results indicate that mutations can influence substrate interactions differentially and suggest that structural features in the DNA may be distinguished by residues of the D,D(35)E region during target DNA selection.

DISCUSSION

We have identified a constellation of amino acids that are conserved in retroviral/retrotransposon integrases and bacterial IS transposases. We refer to it as the D,D(35)E region, for the three invariant, acidic residues, which are D-64, D-121, and E-157 in RSV IN and D-64, D-116 and E-152 in HIV IN. The similarity of sequences in this region of the retroelement/transposase family suggests that it encodes a structural domain whose function has been conserved in evolution. It seems likely that residues in this region contribute to the activities that these proteins have in common—binding, cutting, and joining DNA substrates. To extend our previous deletion analysis of this region and to test the importance of specific residues within it, we have made substitutions of invariant amino acids and selected conserved and nonconserved amino acids and tested their effects on the biochemical activities of IN. On the basis of these results and consideration of the nature of the substituted amino acid side chains, type of replacement, and precedent for their activity in other proteins, specific roles can be proposed for some of the substituted residues.

Conserved W, F, and K residues may be important for DNA recognition. W-61 (of RSV IN and HIV IN) is highly conserved in the retroviral IN proteins and in IS transposases but is replaced by other hydrophobic residues in retrotransposons (Fig. 1). Substitutions of RSV IN W-61 with two such residues, L or F, completely eliminated sequence-independent DNA-binding activity in a nitrocellulose filter binding assay (22, 31), a property not observed with any of the other IN substitution mutants analyzed in this study. These mutants were also defective in both processing and joining activities. The reason for the DNA binding defect is unknown. W-61 might normally interact directly with the DNA substrates, or alternatively, substitution of this residue may alter protein structure or multimerization potential so that DNA binding and other activities are abrogated. Substitution of glutamine (Q) or alanine (A) for the nearby residue T-66, which is also invariant in retroviral IN proteins and conserved in IS elements, has no or little effect on DNA binding, processing, or joining (18). Thus, the results with the W-61 replacements cannot be explained simply by an unusual sensitivity to changes within this general region.

Substitution of alanine for RSV F-126, a residue conserved in retroviruses, or for K-164, a residue conserved in retroviruses and IS elements but not retrotransposons, produced proteins with partial processing activity. Similar results were obtained with the analogous substitutions F-121 and K-159 in HIV IN. Hydrophobic and basic amino acids are often involved in direct interactions with DNA and the partial defects observed are at least consistent with impaired substrate interactions. It may be possible to reveal such interactions by more detailed kinetic analyses and substrate binding studies.

Although RSV F126A and K164A IN exhibited similar joining activities when LTR oligonucleotides were both donor and target, they differed markedly in their abilities to join viral LTR oligonucleotides to heterologous, supercoiled plasmid DNA. The severe defect of F126A in the plasmid joining assay suggests that this hydrophobic residue may participate in recognition of this target. It also suggests that sequence or structural differences between targets may be important in promoting interaction of IN with specific sites during integration.

The invariant D and E residues are critical for catalysis. In contrast to the phenylalanine and lysine substitution mutants discussed above and the conserved histidines, H-9 and H-13 (22), and T-66 of RSV IN (18), replacement of the invariant D or E residues in the D,D(35)E region (RSV D-64, D-121, and E-157 or HIV D-116 and E-152) dramatically reduced both LTR processing and joining activities, even when the substitutions were highly conservative. The dependence of both processing and joining on these invariant residues suggests that they participate in both reactions. As has been documented for other nucleases (5, 10, 17), marked sensitivity to conservative changes is a property that is typical of residues which are components of catalytic centers. In addition, acidic residues are characteristically involved in metal binding (23). Since both the processing and joining reactions of IN require Mn^{2+} or Mg^{2+} one possible role for these invariant residues would be coordination of the metal cofactor(s) required in these reactions. Such a role has recently been proposed for a similar cluster of acidic residues from analysis of the crystal structures of *E. coli* and HIV RNase H (9, 35). Like the IN D(35)E domain, *E. coli* and HIV RNase H contain D and E residues, separated by 37 and 34 amino acids, respectively. The essential role of the relevant residues (D-10 and E-48) in catalysis by *E. coli* RNase H was further demonstrated by the pronounced deleterious effect of their replacement on enzymatic activity (17). Nonconservative substitutions such as D-10→N or E-48→Q eliminated activity, while conservative changes such as D-10→E or E-48→D produced enzymes with 1 to 8% of wild-type activity. Although the nature of our LTR cleavage assays makes rigorous quantitation of IN activity somewhat difficult, we observed similar results with the invariant acidic residues in which conservative substitutions yielded proteins with activities only a few percent of wild-type protein.

The exonucleolytic function of *E. coli* DNA *PolI* may provide the most relevant model for understanding the precise mechanism of DNA cutting involving acidic D and E residues. This protein is believed to utilize a cluster of such acidic residues to stabilize a pentacoordinate phosphorus DNA transition state which facilitates substrate hydrolysis (1, 10). It seems possible that the carboxylate groups of the invariant D and E residues of IN may complex with metal to establish a similar DNA-enzyme intermediate. Cleavage of DNA strands may then be accomplished by nucleophilic attack via a hydroxyl group of an amino acid side chain on the protein (18, 21), the 3′ OH of processed LTR end (12, 25), or a water molecule (12, 34). We hypothesize that the invariant acidic residues in the D,D(35)E region are critical for metal binding and positioning of the DNA substrate for nucleophilic attack. Further enzymatic and structural analysis should allow us to test this hypothesis more directly. Since the D,D(35)E region is conserved in retroviral/retrotransposon integrases and bacterial IS transposases, we expect that the results will elucidate general mechanisms relevant to these and other types of recombinases.

ACKNOWLEDGMENTS

We gratefully acknowledge Jenny Glusker, James Kaiser, Doug Markham, and Christoph Seeger for critically reading this manuscript and for providing helpful suggestions. J.P.G.M. wishes to thank Michael Gribskov for helpful discussions and for providing the routines used in making Fig. 1. We also acknowledge the excellent technical assistance of Patricia Roat.

The research of J.P.G.M. was sponsored in part by the National Cancer Institute, DHHS, under contract N01-CO-74101 with ABL-Basic Research Program. We acknowledge the National Cancer Institute for allocation of computing time and staff support at the Advanced Scientific Computing Laboratory of the Frederick Cancer Research and Development Center. This work was supported by National Institutes of Health grants (CA-47486, CA-06927, and RR-05539), a grant from the Pew Charitable Trust, and also by an appropriation from the Commonwealth of Pennsylvania. J.K. is a Michael Bennett Scholar of the American Foundation for AIDS Research (AmFar).

REFERENCES

1. **Beese, L. S., and T. A. Steitz.** 1991. Structural basis for the 3′-5′ exonuclease activity of *Escherichia coli* DNA polymerase I: a two metal ion mechanism. EMBO J. **10:**25–33.
2. **Bowerman, B., P. Brown, J. M. Bishop, and H. Varmus.** 1989. A nucleoprotein complex mediates the integration of retroviral DNA. Genes Dev. **3:**469–478.
3. **Brown, P., B. Bowerman, H. Varmus, and J. M. Bishop.** 1989. Retroviral integration: structure of the initial covalent product and its precursor, and a role for the viral IN protein. Proc. Natl. Acad. Sci. USA **86:**2525–2529.
4. **Brown, P. O., B. Bowerman, H. E. Varmus, and J. M. Bishop.** 1987. Correct integration of retroviral DNA *in vitro*. Cell **49:**347–356.
5. **Buechler, J., and S. Taylor.** 1988. Identification of aspartate-184 as an essential residue in the catalytic subunit of cAMP-dependent protein kinase. Biochemistry **27:**7356–7361.
6. **Bushman, F. D., T. Fujiwara, and R. Craigie.** 1990. Retroviral DNA integration directed by HIV integration protein *in vitro*. Science **249:**1555–1558.
7. **Castle, E., T. Nowak, U. Leidner, G. Wengler, and G. Wengler.** 1985. Sequence analysis of the viral core protein and the membrane associated proteins V1 and NV2 of the flavivirus West Nile virus and the genome sequence for these proteins. Virology **145:**227–236.
8. **Craigie, R., T. Fujiwara, and F. Bushman.** 1990. The IN protein of Moloney murine leukemia virus processes the viral DNA

ends and accomplishes their integration *in vitro*. Cell **62**:829–837.

9. **Davies, J., Z. Hostomska, Z. Hostomsky, S. Jordan, and D. Matthews.** 1991. Crystal structure of the ribonuclease H domain of HIV-1 reverse transcriptase. Science **252**:88–95.
10. **Derbyshire, V., N. D. F. Grindley, and C. M. Joyce.** 1991. The 3'-5' exonuclease of DNA polymerase I of *Escherichia coli*: contribution of each amino acid at the active site to the reaction. EMBO J. **10**:17–24.
11. **Devereux, J., P. Haeberli, and O. Smithies.** 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12**:387–395.
12. **Engelman, A., K. Mizuuchi, and R. Craigie.** 1991. HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. Cell **67**:1211–1221.
13. **Fayet, O., P. Ramond, P. Polard, M. F. Prére, and M. Chandler.** 1990. Functional similarities between retroviruses and the IS3 family of bacterial insertion sequences? Mol. Microbiol. **4**:1771–1777.
14. **Gribskov, M., R. Luethy, and D. Eisenberg.** 1989. Profile analysis. Methods Enzymol. **182**:146–159.
15. **Hein, J.** 1990. Unified approach to alignment and phylogenies. Methods Enzymol. **183**:626–645.
16. **Johnson, M. S., M. A. McClure, D.-F. Feng, J. Gray, and R. F. Doolittle.** 1986. Computer analysis of retroviral *pol* genes: assignment of enzymatic functions to specific sequences and homologies with nonviral enzymes. Proc. Natl. Acad. Sci. USA **83**:7648–7652.
17. **Kanaya, S., A. Kohara, Y. Miura, A. Sekiguchi, S. Iwai, H. Inoue, E. Ohtsuka, and M. Ikehara.** 1990. Identification of the amino acid residues involved in an active site of *Escherichia coli* ribonuclease H by site-directed mutagenesis. J. Biol. Chem. **265**:4615–4621.
18. **Katz, R., J. P. Mack, G. Merkel, J. Kulkosky, C. Zheng, J. Leis, and A. M. Skalka.** Requirement for a conserved serine in both processing and joining activities of retroviral integrase. Proc. Natl. Acad. Sci. USA, in press.
19. **Katz, R. A., G. Merkel, J. Kulkosky, J. Leis, and A. M. Skalka.** 1990. The avian retroviral IN protein is both necessary and sufficient for integrative recombination *in vitro*. Cell **63**:87–95.
20. **Katzman, M., R. A. Katz, A. M. Skalka, and J. Leis.** 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.
21. **Katzman, M., J. P. G. Mack, A. M. Skalka, and J. Leis.** 1991. A covalent complex between retroviral integrase and nicked substrate DNA. Proc. Natl. Acad. Sci. USA **88**:4695–4699.
22. **Khan, E., J. P. G. Mack, R. A. Katz, J. Kulkosky, and A. M. Skalka.** 1991. Retroviral integrase domains: DNA binding and the recognition of LTR sequences. Nucleic Acids Res. **19**:851–860.
23. **Knighton, D., J. Zheng, L. Eyck, V. Ashford, N. Xuong, S. Taylor, and J. Sowadski.** 1991. Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. Science **253**:407–414.
24. **Kulkosky, J., and A. M. Skalka.** 1990. HIV DNA integration: observations and inferences. J. Acquired Immune Defic. Syndr. **3**:839–851.
25. **Mizuuchi, K., and R. Craigie.** 1986. Mechanism of bacteriophage Mu transposition. Annu. Rev. Genet. **20**:385–429.
26. **Roth, M. J., P. L. Schwartzberg, and S. P. Goff.** 1989. Structure of the termini of DNA intermediates in the integration of retroviral DNA: dependence on IN function and terminal DNA sequence. Cell **58**:47–54.
27. **Rowland, S.-J., and K. G. H. Dyke.** 1990. Tn552, a novel transposable element from *Staphylococcus aureus*. Mol. Microbiol. **4**:961–965.
28. **Sherman, P. A., and J. A. Fyfe.** 1990. Human immunodeficiency virus integration protein expressed in *Escherichia coli* possesses selective DNA cleaving activity. Proc. Natl. Acad. Sci. USA **87**:5119–5123.
29. **Skalka, A. M.** 1988. Integrative recombination of retroviral DNA, p. 701–724. *In* R. Kucherlapati and R. Smith (ed.), Genetic recombination. American Society for Microbiology, Washington, D.C.
30. **Taha, M. K., M. So, H. S. Seifert, E. Billyard, and C. Marchal.** 1988. Pilin expression in Neisseria gonorrhoeae is under both positive and negative transcriptional control. EMBO J. **7**:4367–4368.
31. **Terry, R., D. A. Soltis, M. Katzman, D. Cobrinik, J. Leis, and A. M. Skalka.** 1988. Properties of avian sarcoma-leukosis virus pp32-related *pol*-endonucleases produced in *Escherichia coli*. J. Virol. **62**:2358–2365.
32. **Varmus, H. E., and P. Brown.** 1989. Retroviruses, p. 53–108. *In* D. Berg and M. Howe (ed.), Mobile DNA. American Society for Microbiology, Washington, D.C.
33. **Varmus, H. E., and R. Swanstrom.** 1985. Replication of retroviruses, p. 75–134. *In* R. Weiss, N. Teich, H. Varmus, and J. Coffin (ed.), RNA tumor viruses. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
34. **Vink, C., et al.** 1992. Site-specific hydrolysis and alcoholysis of human immunodeficiency virus DNA termini mediated by the viral integrase protein. Nucleic Acids. Res. **19**:6691–6698.
35. **Yang, W., W. A. Hendrickson, R. J. Crouch, and Y. Satow.** 1990. Structure of ribonuclease H phased at 2 Å resolution by MAD analysis of the selenomethionyl protein. Science **249**:1398–1405.