Domain structure and three-dimensional model of a group II intron-encoded reverse transcriptase

FORREST J.H. BLOCKER,[1] GEORG MOHR,[1] LORI H. CONLAN,[2] LI QI,[2,3] MARLENE BELFORT,[2] and ALAN M. LAMBOWITZ[1]

[1]Institute for Cellular and Molecular Biology, Department of Chemistry and Biochemistry, and Section of Molecular Genetics and Microbiology, School of Biological Sciences, University of Texas at Austin, Austin, Texas 78712, USA
[2]Wadsworth Center, New York State Department of Health, Center for Medical Sciences, Albany, New York 12208, USA

ABSTRACT

Group II intron-encoded proteins (IEPs) have both reverse transcriptase (RT) activity, which functions in intron mobility, and maturase activity, which promotes RNA splicing by stabilizing the catalytically active RNA structure. The LtrA protein encoded by the *Lactococcus lactis* Ll.LtrB group II intron contains an N-terminal RT domain, with conserved sequence motifs RT1 to 7 found in the fingers and palm of retroviral RTs; domain X, associated with maturase activity; and C-terminal DNA-binding and DNA endonuclease domains. Here, partial proteolysis of LtrA with trypsin and Arg-C shows major cleavage sites in RT1, and between the RT and X domains. Group II intron and related non-LTR retroelement RTs contain an N-terminal extension and several insertions relative to retroviral RTs, some with conserved features implying functional importance. Sequence alignments, secondary-structure predictions, and hydrophobicity profiles suggest that domain X is related structurally to the thumb of retroviral RTs. Three-dimensional models of LtrA constructed by "threading" the aligned sequence on X-ray crystal structures of HIV-1 RT (1) account for the proteolytic cleavage sites; (2) suggest a template–primer binding track analogous to that of HIV-1 RT; and (3) show that conserved regions in splicing-competent LtrA variants include regions of the RT and X (thumb) domains in and around the template–primer binding track, distal regions of the fingers, and patches on the protein's back surface. These regions potentially comprise an extended RNA-binding surface that interacts with different regions of the intron for RNA splicing and reverse transcription.

Keywords: HIV-1; retrotransposon; retrovirus; ribozyme; RNA–protein interaction; RNA splicing

INTRODUCTION

Mobile group II introns encode proteins with both RT activity for intron mobility and RNA splicing ("maturase") activity (Lambowitz et al. 1999; Belfort et al. 2002; Lambowitz and Zimmerly 2004). The IEP promotes splicing by stabilizing the active structure of the intron RNA, which then catalyzes transesterification reactions resulting in ligated exons and excised intron lariat RNA. After splicing, the protein remains bound to the excised lariat RNA to promote intron mobility by a remarkable mechanism in which the intron RNA reverse splices directly into a DNA target site and is reverse-transcribed by the IEP. The group II IEPs that mediate these processes contain an N-terminal RT domain homologous to retroviral RTs, followed by regions without strong sequence similarity to other RTs. These include domain X, where mutations affecting RNA splicing activity have been found, and C-terminal DNA-binding (D) and DNA endonuclease (En) domains, which function in intron mobility (Mohr et al. 1993; San Filippo and Lambowitz 2002). The structural relationship between group II intron and other RTs and how different regions of group II IEPs interact with the intron RNA and DNA target site to promote RNA splicing and intron mobility have remained unclear.

Studies in our laboratories have focused on the *Lactococcus lactis* Ll.LtrB intron, for which an efficient *Escherichia coli* expression system facilitates biochemical and genetic analysis (Matsuura et al. 1997; Cousineau et al. 1998; Saldanha et al. 1999; Cui et al. 2004). Ll.LtrB is a subgroup IIA intron belonging to the "mitochondrial" lineage, one of

Reprint requests to: Alan M. Lambowitz, Institute for Cellular and Molecular Biology, University of Texas at Austin, 1 University Station A4800, Austin, TX 78712, USA; e-mail: lambowitz@mail.utexas.edu.

[3]**Present address:** Albany Medical College, 47 New Scotland Avenue MC-165, Albany, NY 12208, USA.

Abbreviations: D domain, DNA-binding domain; En domain, DNA endonuclease domain; IEP, intron-encoded protein; LTR, long terminal repeat; RNP, ribonucleoprotein; RT, reverse transcriptase.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.7181105.

eight lineages of mobile group II introns defined by phylogenetic analysis of the IEPs and structural features of the intron RNAs (Toor et al. 2001; Lambowitz and Zimmerly 2004). Figure 1 shows a diagram of the Ll.LtrB IEP, denoted LtrA, with HIV-1 RT shown below for comparison. In LtrA and other group II IEPs, the RT domain contains conserved sequence motifs RT1–RT7 found in the fingers and palm of retroviral RTs, along with an upstream motif RT0 characteristic of the RTs of non-LTR-retroelements (Xiong and Eickbush 1990; Malik et al. 1999; Zimmerly et al. 2001). Domain X is downstream of the RT domain in the position corresponding to the thumb and part of the connection domain of HIV-1 RT, and is followed by the D and En domains. The En domain, which carries out second-strand cleavage to generate the primer for reverse transcription of the intron RNA, contains conserved sequence motifs characteristic of the H-N-H family of DNA endonucleases interspersed with two pairs of conserved cysteine residues, similar to an arrangement found in phage T4 endonuclease VII (Gorbalenya 1994; Shub et al. 1994; San Filippo and Lambowitz 2002). The H-N-H active site in LtrA contains a single catalytically essential Mg^{2+} ion, while the conserved cysteine pairs help maintain the higher-order structure of the domain (San Filippo and Lambowitz 2002). The En domain was likely acquired by a pre-existing IEP to facilitate mobility by target DNA-primed reverse transcription (TPRT) (Martínez-Abarca and Toro 2000; Belfort et al. 2002). RTs have a propensity for acquiring additional domains, including at least two unrelated En domains used for TPRT by different non-LTR-retrotransposon RTs, and the RNase H domain of retroviral RTs (McClure 1991; Eickbush and Malik 2002; Moran and Gilbert 2002).

The HIV-1 RT, for which X-ray crystal structures have been determined, is a heterodimer consisting of a p66 subunit, comprised of fingers, palm, thumb, connection, and RNase H domains, and a p51 subunit, which is derived from p66 by a proteolytic cleavage that removes the RNase H domain (Kohlstaedt et al. 1992). The fingers and thumb of p66 form a cleft with the palm at its base containing the RT active site. The common regions of p51 contain the same domains as p66, but their relative orientation differs (Kohlstaedt et al. 1992; Wang et al. 1994; Ren et al. 1995). The p51 subunit in the dimer has no enzymatic activity, but is required for the correct folding of p66, and also makes some contribution to binding the template–primer (Ding et al. 1997; Huang et al. 1998).

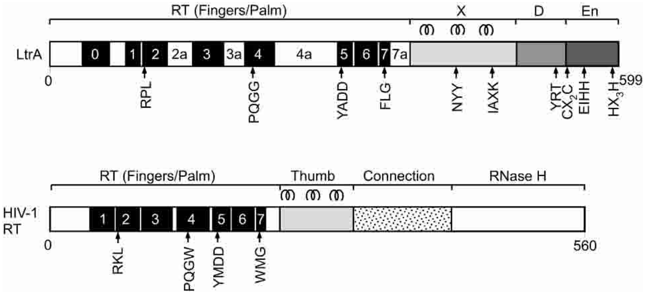

FIGURE 1. Schematics of the LtrA protein and HIV-1 RT. The LtrA protein contains an N-terminal RT domain with conserved sequence blocks RT1–RT7 found in the fingers and palm of HIV-1 RT, an upstream region containing RT0 characteristic of non-LTR-retroelement RTs, and insertions 2a, 3a, 4a, and 7a relative to HIV-1 RT. The RT domain of LtrA is followed by X, DNA-binding (D), and DNA endonuclease (En) domains. HIV-1 RT contains conserved RT sequence blocks RT1–RT7 in the fingers and palm, followed by thumb, connection, and RNase H domains. The three α-helices in the HIV-1 RT thumb and the corresponding predicted α-helices in LtrA's domain X are shown *above* the proteins, and some conserved sequence motifs are shown *below*.

HIV-1 RT is conformationally flexible, with the fingers and thumb in an "open" conformation upon binding template–primer, and the fingers moving down to a "closed" conformation upon binding dNTP (Ding et al. 1998; Huang et al. 1998; Peletskaya et al. 2004). X-ray crystal structures with bound DNA template–primer show that the nucleic acid interacts primarily with the p66 subunit, binding along a groove with positively charged walls that extends from the polymerase active site to the RNase H domain (Bebenek et al. 1997; Ding et al. 1998; Huang et al. 1998). In both the open and closed structures, the template–primer undergoes a transition from A-form to B-form near the RT active site accompanied by a bend of ~40°, which brings the 5′-overhang of the template across the face of the fingers (Ding et al. 1998; Huang et al. 1998). Farther upstream regions of the template strand are thought to interact with distal regions of the fingers (Huang et al. 1998; Peletskaya et al. 2004).

Phylogenetic analyses, based on conserved RT sequences, show that retroelements can be divided into two major classes: LTR-containing retroelements, such as retroviruses and related LTR-containing transposons (e.g., yeast Ty1 elements); and non-LTR-retroelements, a diverse group that includes non-LTR-retrotransposons (e.g., human LINE elements and insect R2 elements), mobile group II introns, bacterial retrons, retroplasmids, and telomerase RTs (Xiong and Eickbush 1990; Malik et al. 1999). The non-LTR-retroelement RTs are thought to be evolutionary ancestors of retroviral RTs (Eickbush and Malik 2002). They use a number of different mechanisms to prime reverse transcription, but a common feature is that the cDNA initiation site is determined primarily by the specific binding of the RT to the template RNA, rather than by base-pairing of a primer as for retroviral RTs (Chen and Lambowitz 1997).

The RTs encoded by non-LTR-retroelements are generally larger than retroviral RTs and have an additional, conserved N-terminal motif, referred to as RT0, proposed to be part of an extended fingers region involved in the specific binding of the template RNA for initiation of reverse transcription (see Fig. 1; Chen and Lambowitz 1997; Bibillo and Eickbush 2002). The non-LTR-element RTs also have longer connecting regions between RT2 and RT3, denoted 2a, and between RT3 and RT4, denoted 3a, and some group II intron RTs also have a longer connecting region between

RT4 and RT5, denoted 4a (Xiong and Eickbush 1990; Malik et al. 1999; Zimmerly et al. 2001). Whether these insertions are nonconserved expansion loops or have specific functions is not known. The region downstream of the RT domain, denoted domain X in group II intron RTs, has no strong sequence homology among RTs from different types of retroelements, and it has remained unclear whether or not domain X is evolutionarily related to the retroviral RT thumb (Mohr et al. 1993).

Here, we investigated the domain structure of group II intron RTs by partial proteolysis and computational analysis. Our results show major proteolytic cleavage sites within RT1 and between the palm and domain X. Domain X and the corresponding region of non-LTR-retrotransposon RTs have three predicted α-helices, whose characteristics and spacing are similar to those of the thumb of retroviral RTs, consistent with a common evolutionary origin. Our model provides an initial structural framework for understanding how group II intron RTs function in RNA splicing and intron mobility.

RESULTS AND DISCUSSION

Controlled proteolysis of the LtrA protein in the presence or absence of Ll.LtrB RNA

Controlled proteolysis can provide insight into structure–function relationships and protein domain structure. Figure 2 shows experiments in which the LtrA protein was treated with either trypsin or Arg-C, which cleave after arginine and lysine, and after arginine residues, respectively. Digestion was allowed to proceed for up to 60 min, and the cleavage products were analyzed by SDS-PAGE. With both proteases, most of the cleavage occurred within the first 10 min and generated bands in the 60-, 43-, 33-, and 27-kDa size ranges (Fig. 2A,B). With both enzymes, some of these bands appeared as doublets, triplets, or quadruplets.

Arg-C routinely generated a faint band of 10 kDa, which was not visible in the trypsin digests (Fig. 2B). Arg-C was therefore used for characterization of the protein fragments, by a combination of HPLC and mass spectrometry, and/or N-terminal sequencing by Edman degradation (Table 1). These analyses allowed identification of the proteolysis products, pointing to two major cleavage sites, one in RT1 and the other between RT7 and domain X. Cleavage at the first major site in RT1 yielded a 10-kDa N-terminal fragment containing RT0 (M1–R85) and a 60-kDa band tentatively identified as RT1/7-X-D-E (R86–K599) (Table 1). Cleavage at the second major site, between RT7 and domain X, produced bands of 43 kDa, corresponding to RT0/7 (M1–R364); 33 kDa, corresponding to RT1/7 (R86–R364); and 27 kDa, corresponding to X-D-E (R365–K599 or S372–K599). Edman degradation sequencing showed that two bands in the X-D-E triplet correspond to digestion products between amino acids 360 and 381 (Fig. 2A). The first fragment is 27.4 kDa, consistent with digestion at R365. The second fragment is 26.7 kDa, starting at S372. The third band in the triplet likely results from digestion at R378.

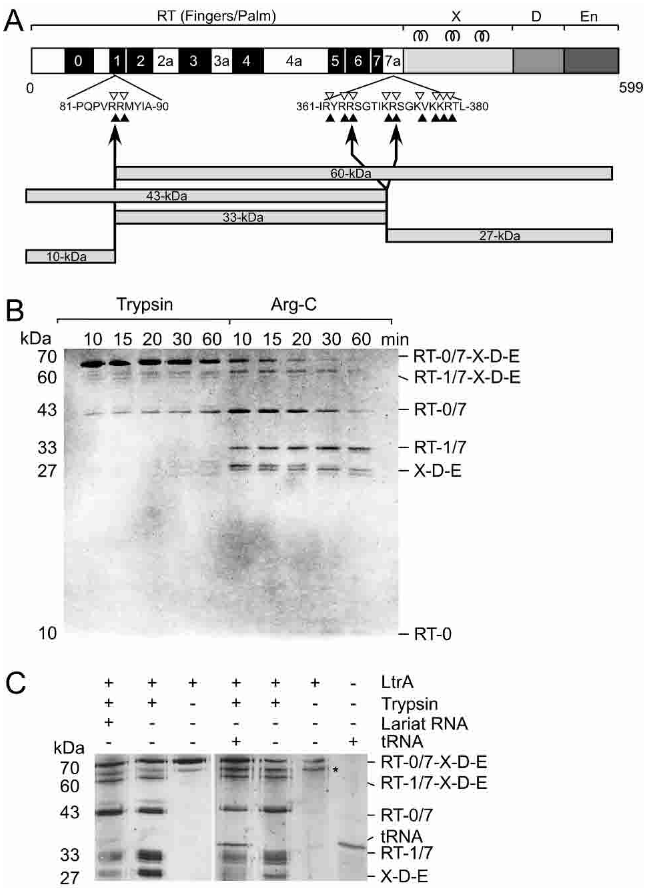

FIGURE 2. Controlled proteolysis of LtrA protein with trypsin and Arg-C. (*A*) Map of LtrA protein with sequences shown at major cleavage sites in RT1 and between RT7 and domain X. Filled arrowheads represent potential Arg-C digestion sites, arrows indicate mapped digestion sites, and open arrowheads represent potential trypsin digestion sites. A schematic representation of the Arg-C digestion products shown in panel *B* is presented *below* the map of LtrA. For details, see Table 1. (*B*) Limited proteolysis of LtrA with trypsin and Arg-C. Proteolysis products were separated by SDS-PAGE, and the gel was stained with Coomassie blue. Fragment assignments at the *right* were determined from Arg-C-digested LtrA, as shown in Table 1. The 10-kDa band is faint in this figure, but clearly and reproducibly present on stained gels, and its identity was verified by mass spectrometry. (*C*) Limited proteolysis of LtrA in the presence and absence of intron lariat and tRNA. LtrA was incubated with intron RNA or tRNA, then digested with trypsin. Proteolysis products were separated by SDS-PAGE, and the gel was silver-stained. Asterisk indicates an unidentified band in LtrA preparations. Nonspecific protection of an exposed region can be seen for the 33- and 27-kDa digestion products.

These results contrast with a previous report, in which complete digestion of LtrA was observed after trypsin addition, leading to the suggestion that the free protein lacks conformational stability (Rambo and Doudna 2004). This difference is likely attributable to trypsin concentration, where these authors used >5-fold higher concentrations (12.5 ng/μL) than we used. We observed partial proteolysis

TABLE 1. Characterization of Arg-C digestion products

		Mass spec[a] MW			
Band	Amino acids	Est.	Meas.	Edman degradation N-term	Domain composition
60 kDa	Arg 86 to Lys 599[b]	60,282	—	—	RT1/7-X-D-E
43 kDa	Met 1 to Arg 364	42,788	42,792	1-MKPTM	RT0/7
33 kDa	Arg 86 to Arg 364	32,890	—	86-RMIYA	RT1/7
27 kDa	Arg 365 to Lys 599	27,392	—	365-RSGTI	X-D-E-1
	Ser 372 to Lys 599	26,762	—	372-SGKVK	X-D-E-2
10 kDa	Met 1 to Arg 85	9915	9914	—	RT0

Described in Figure 2A and text.
[a] Mass spectrometry. (Est.) Estimated from protein sequence; (Meas.) measured by mass spectrometry; (MW) molecular weight.
[b] Assignment tentative, based on molecular weight (MW) estimates from SDS-PAGE.

at 1.9 ng/μL trypsin after 16 min, but complete LtrA degradation at 3.8 ng/μL trypsin by 2 min. Furthermore, our proteolysis results were similar over a range of salt concentrations (100–450 mM NaCl) and temperatures (25°–37°C) (data not shown), suggesting that LtrA has a stable conformation even in the absence of RNA.

The LtrA protein binds tightly to excised intron lariat RNA in vitro (Saldanha et al. 1999). The tryptic digest in Figure 2C shows that addition of intron lariat RNA had relatively little effect on the proteolytic digestion, with only a small decrease in the amounts of the 33-kDa (RT1/7) and 27-kDa (X-D-E) fragments. We confirmed by filter-binding experiments that addition of lariat RNA resulted in RNP formation, and assays measuring reverse splicing, endonuclease cleavage, and RT activity demonstrated that the RNPs were active (data not shown). Again, these proteolysis results were independent of salt concentration or temperature, and whether the protection was performed with purified lariat, or with products of an in vitro splicing reaction (see Materials and Methods). Similar slight protection was also observed in the presence of tRNA (Fig. 2C). Together, these data indicate that proteolytic cleavage is relatively unaffected by intron RNA binding and provide no evidence that binding of the intron RNA produces substantial conformational changes in the protein, although such changes could occur in regions not probed by proteolytic digestion. These conclusions again contrast with those of Rambo and Doudna (2004), who interpret their observed protection against proteolysis of LtrA upon RNA binding as a major protein conformational change. While RNA binding undoubtedly has some protective effect at high trypsin concentrations, caution needs to be exercised in interpreting these results as major conformational changes. However, we agree with these authors on proteolytic cleavage between the RT and X domains, with our data showing the susceptibility of the RT-X junction to multiple proteolytic cleavages, and some protection of this exposed junction by RNA.

Structure-based sequence alignments of LtrA and HIV-1 RT

Figure 3 shows a sequence alignment comparing the predicted secondary structure of LtrA with the X-ray crystallographically determined and predicted structures of the p66 and p51 subunits of HIV-1 RT. The secondary structure of LtrA was predicted by using the JPred server, which reports 76.4% accuracy (Cuff et al. 1998; Cuff and Barton 2000; see Materials and Methods). As shown in Figure 3, JPred accurately predicted the secondary structure of HIV-1 RT without relying on HIV-1 RT X-ray crystal structures. The alignment between the RT domains of LtrA and HIV-1 RT was based on matches to conserved RT sequences, while domains X and D of LtrA were aligned with the thumb and connection domains of HIV-1 RT by using the predicted secondary structure and manual adjustments to maximize sequence similarity.

Aligned as shown, the predicted secondary structure throughout LtrA's RT and X domains agrees well with the crystallographically determined structures of the RT and thumb domains of the p66 and p51 subunits of HIV-1 RT (Ding et al. 1998; 2HMI.pdb). The major differences in LtrA are the previously discussed N-terminal extension, which includes RT0, and the longer connecting regions between the conserved RT sequence blocks (see Introduction). Although it is unknown if these nonhomologous regions were gained by non-LTR-retroelement RTs or lost by retroviral RTs, for convenience, we refer to them as insertions, denoted 2a, 3a, 4a, and 7a according to the RT motif they follow. Domain X, which is located in the position corresponding to the thumb and upstream part of the connection domain of HIV-1 RT, contains three predicted α-helices, which potentially correspond to α-helices αH, αI, and αJ in the HIV-1 RT thumb (see also below). The spacing between αI and αJ is similar to that in HIV-1 RT, while the spacing between αH and αI is ~16 residues longer, an insertion we refer to as "ti."

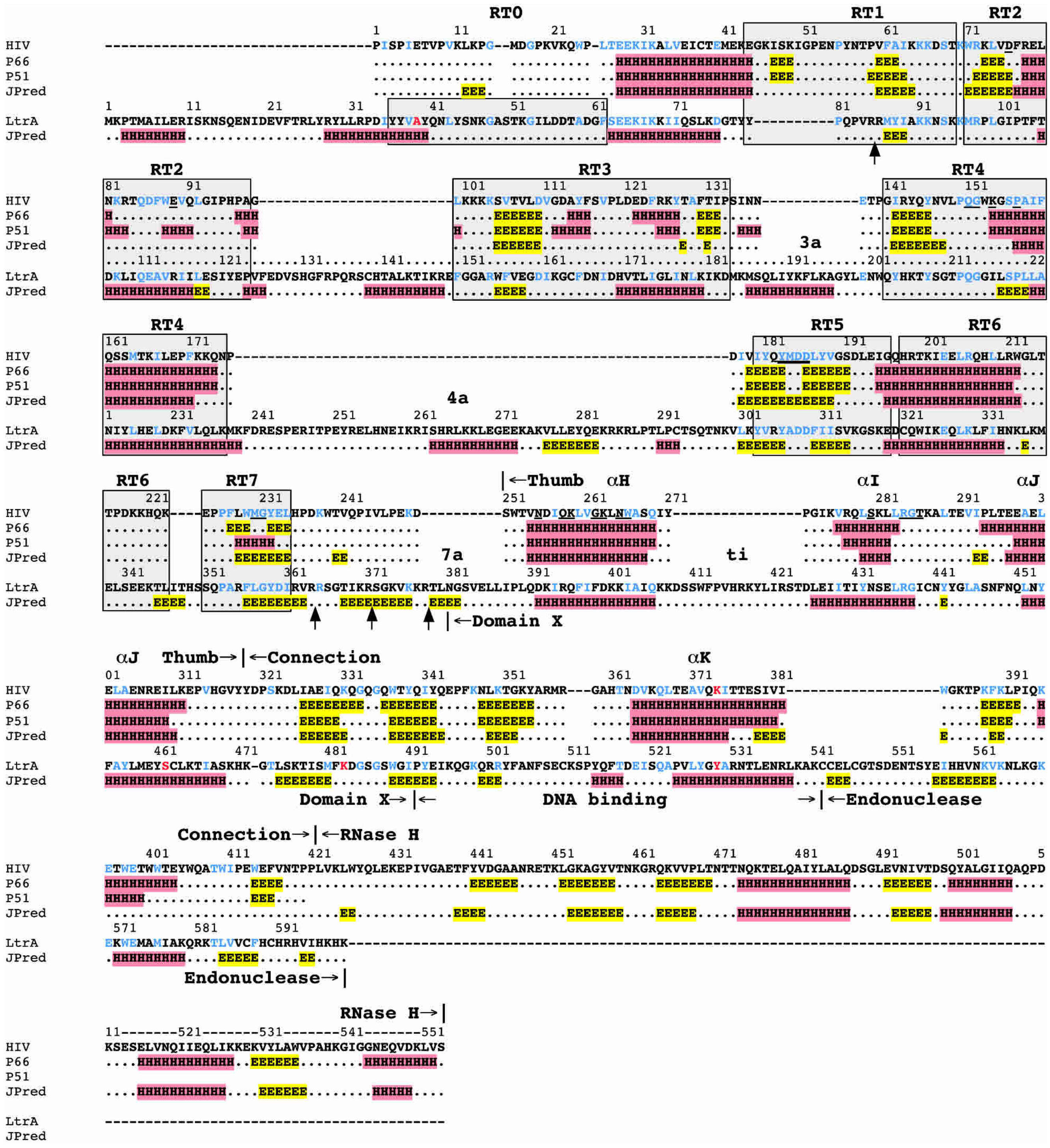

FIGURE 3. Sequence alignment of LtrA and HIV-1 RT. The HIV-1 RT (HIV) sequence is from GenBank accession no. P03366, and the LtrA sequence is from accession no. Q57005. Structural data for the HIV-1 RT p66 and p51 subunits (lines *2* and *3*) are from the X-ray crystal structure of Ding et al. 1998 (2HMI.pdb). The JPred lines show secondary-structure predictions generated by JPred (http://www.compbio.dundee.ac.uk/~www-jpred/submit.html). In both the determined and predicted structures, α-helices (H) are highlighted in red and β-strands (E) in yellow. Conserved sequence blocks RT0–RT7, as defined by Xiong and Eickbush (1990) and Zimmerly et al. (2001), are indicated by gray boxes. Gaps in the sequence alignment are indicated by dashes. Protease-cleavage sites are indicated by arrows pointing to the LtrA sequence. Residues that are identical or similar in the two proteins are shown in blue, with similar amino acids defined as hydrophobic, L, I, V, M, F, W, Y, A; acidic, D, E; basic, R, K; and polar, S, T (Henikoff and Henikoff 1992). Some key amino acid residues discussed in the text, including HIV-1 RT K374 and LtrA A39, S462, K483, and Y529, are shown in red. Amino acid residues in the RT and thumb domains of HIV-1 RT that interact with the template–primer (Ding et al. 1998) are underlined.

The predicted secondary-structure similarities continue between HIV-1 RT's connection domain and LtrA's D domain. The latter was shown previously to consist of an upstream region containing basic amino acid residues (LtrA positions 499–502), followed by a predicted α-helix (LtrA positions 524–538; San Filippo and Lambowitz 2002). Both regions are functionally important, and although not strongly conserved in sequence, can be found in other group II IEPs (San Filippo and Lambowitz 2002). The predicted α-helix in domain D of LtrA aligns with αK in the connection domain of HIV-1 RT. In the latter, αK forms part of the dimerization interface between p66 and p51 and includes some residues that are close to the template–primer, for example, 3.95 Å for K374 in p66 (Ding et al. 1998; 2HMI.pdb). It will be pertinent below that HIV-1 RT K374 aligns with LtrA Y529, the site of a mutation (Y529A) that specifically inhibits second-strand DNA cleavage (these and other key amino acid residues discussed in the text are highlighted in red in Fig. 3; San Filippo and Lambowitz 2002). In the En domain, the program fails to predict a C-terminal α-helix that, based on the structure of other H-N-H endonucleases, is expected to be part of the H-N-H active site (see San Filippo and Lambowitz 2002). In the three-dimensional model below, the En domain was modeled independently based on the X-ray crystal structure of the H-N-H region of phage T4 endonuclease VII (Raaijmakers et al. 2001; 1EN7.pdb).

RT domain insertions

To assess the validity of the predicted secondary structures of LtrA's N-terminal extension and RT domain insertions, we compared them with predicted secondary structures for the same insertions in a collection of non-LTR-retroelement RTs (*Drosophila* jockey, mouse L1, and *Bombyx mori* R2Bm) and representative group II intron RTs (Fig. 4). The latter were selected from different group II intron RT lineages, including the mitochondrial (LtrA and the yeast aI1 and aI2 IEPs), bacterial classes A-E, and chloroplast lineages (Martínez-Abarca and Toro 2000; Zimmerly et al. 2001). Residues conserved in >50% of the aligned group II IEPs are shown in blue.

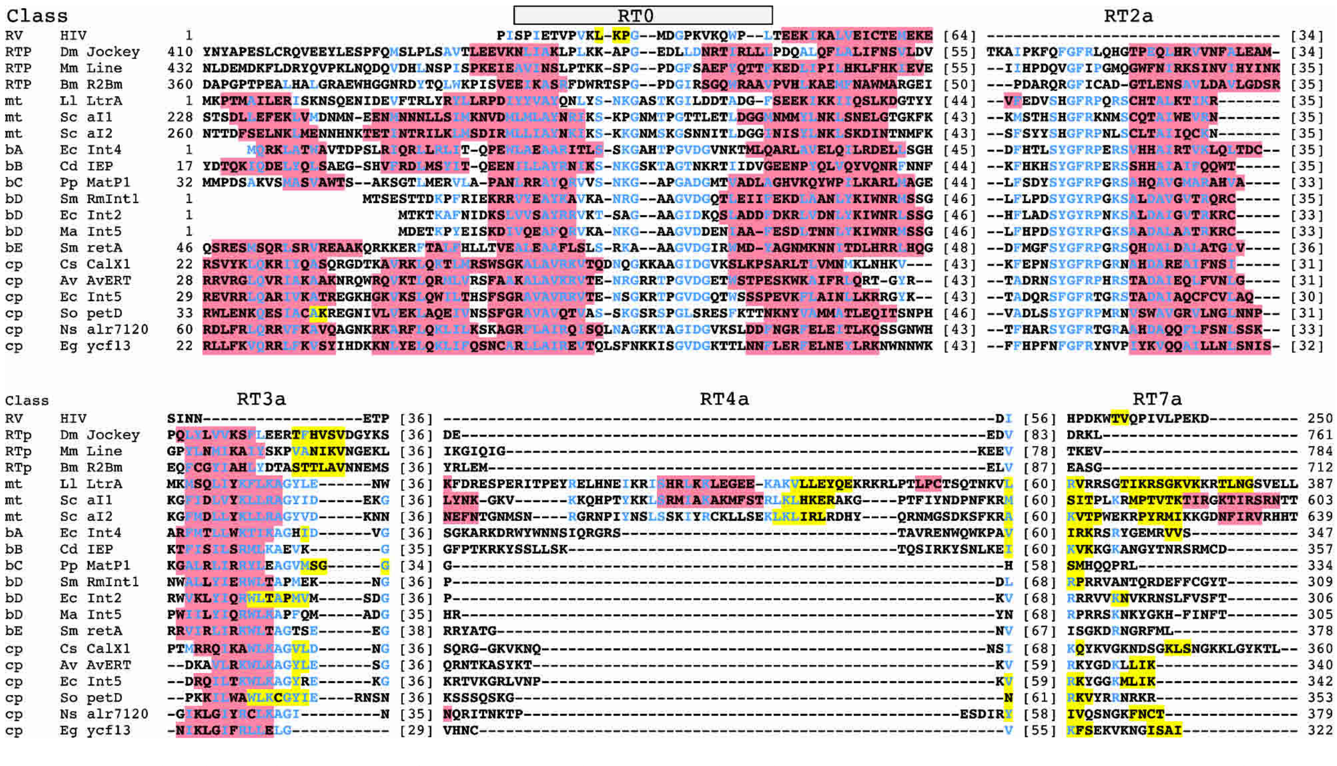

FIGURE 4. Predicted secondary structures for the N-terminal extension and RT domain insertions in non-LTR-retrotransposon and group II intron RTs. Retrotransposons (RTP) are the *Drosophila melanogaster* jockey element (Dm Jockey, accession no. JT0396), *Mus musculus* LINE1 element (Mm Line, GNMSLL), and *Bombyx mori* R2Bm (Bm R2Bm, T18197). Group II IEPs belong to the mitochondrial (mt); bacterial A, B, C, D, and E (bA, bB, bC, bD, bE, respectively); and chloroplast (cp) lineages. Accession numbers are: *Lactococcus lactis* LtrA (Ll LtrA, Q57005); *Saccharomyces cerevisiae* aI1 and aI2 (Sc aI1 and aI2, NP_009310 and NP_009309, respectively); *Escherichia coli* intron 4 (Ec Int4, BAA84894); *Clostridium difficile* IEP (Cd IEP, X98606); *Pseudomonas putida* matP1 (Pp MatP1, AF101076); *Sinorhizobium meliloti* RmInt1 (Sm RmInt1, NP_438012); *E. coli* intron 2 (Ec Int2, S50828); *Methanosarcina aromaticovorans* intron 5 (Ma Int5, AAM07961); *Serratia marcescens* marI1 (Sm marI1, AF027768); *Calothrix* species CalX1 (Cs CalX1, CAA50529); *Azotobacter vinelandii* I-AvERT (Av AvERT, AAL25965); *E. coli* O157:H7 intron 5 (Ec Int5, T00245); *Scenedesmus obliquus* petD (So petD, P19593); *Nostoc* species PC7120 alr7241 (Ns alr7120, BAB78325); *Euglena gracilis* ycf13 (Eg ycf13, NP_041894). α-Helices are highlighted in red, and β-strands in yellow. Conserved residues found in >50% of the group II IEPs are shown in blue.

The alignments show that an N-terminal extension containing RT0 is present in both the non-LTR-retrotransposon and group II intron RTs, but is variable in length, being shortest in the bacterial class D group II intron IEPs (Fig. 4). In all cases, the N-terminal portion of RT0 forms a predicted α-helix, which contains a conserved alanine (LtrA A39) surrounded by predominantly apolar/hydrophobic amino acids, while the C-terminal portion contains additional conserved residues and immediately precedes a predicted α-helix that aligns with an α-helix near the N terminus of HIV-1 RT. In those group II intron RTs with longer N-terminal extensions, RT0 is generally preceded by a region containing two predicted α-helices, which are not predicted for the corresponding region of non-LTR-retrotransposon RTs.

Insertions 2a and 3a also have conserved features in both non-LTR-retrotransposon and group II intron RTs. In all cases, insertion 2a is predicted to consist of an N-terminal loop and a C-terminal α-helix. In the group II intron RTs, the loop contains a strongly conserved sequence motif SYGFRPX(K/R)S (LtrA residues 130–138) (Figs. 3, 4), which is present in shorter form in the non-LTR-retrotransposon RTs. Insertion 3a in all the proteins consists of a predicted N-terminal α-helix followed by a variable length loop region, which contains small predicted β-strands in several cases. The α-helix in 3a contains well-conserved arginine and lysine residues and is followed by conserved sequence AG(hydrophobic)$_2$(acidic)X$_n$G in the group II intron RTs (LtrA residues 184–202; Figs. 3, 4).

By comparison to the above, insertions 4a and 7a are more variable. Insertion 4a is longest in the mitochondrial lineage group II intron RTs, and within that lineage has sequence and predicted secondary-structure similarities not shared by the other proteins. Insertion 7a, the linker between the RT and X/thumb domains, is of variable length, being longest in the three mitochondrial lineage and one chloroplast lineage IEPs (*Calothrix*). It is generally rich in basic amino acids, and its N-terminal part is predicted to form β-strands in a number of proteins.

The finding that the N-terminal extension and other insertions have conserved sequence and predicted secondary features in diverse non-LTR-retroelement and group II intron RTs suggests that they are not simply expansion loops and that some or all may be functionally important.

Predicted secondary structure of domain X

Figure 5 shows the predicted secondary structures of the domain X region for the same collection of non-LTR-retrotransposon and group II intron RTs as in Figure 4. A sequence logo for domain X based on the group II IEP alignment is shown below. Strikingly, in each protein, the predicted secondary structure for domain X has three predicted α-helices, which potentially correspond to α-helices

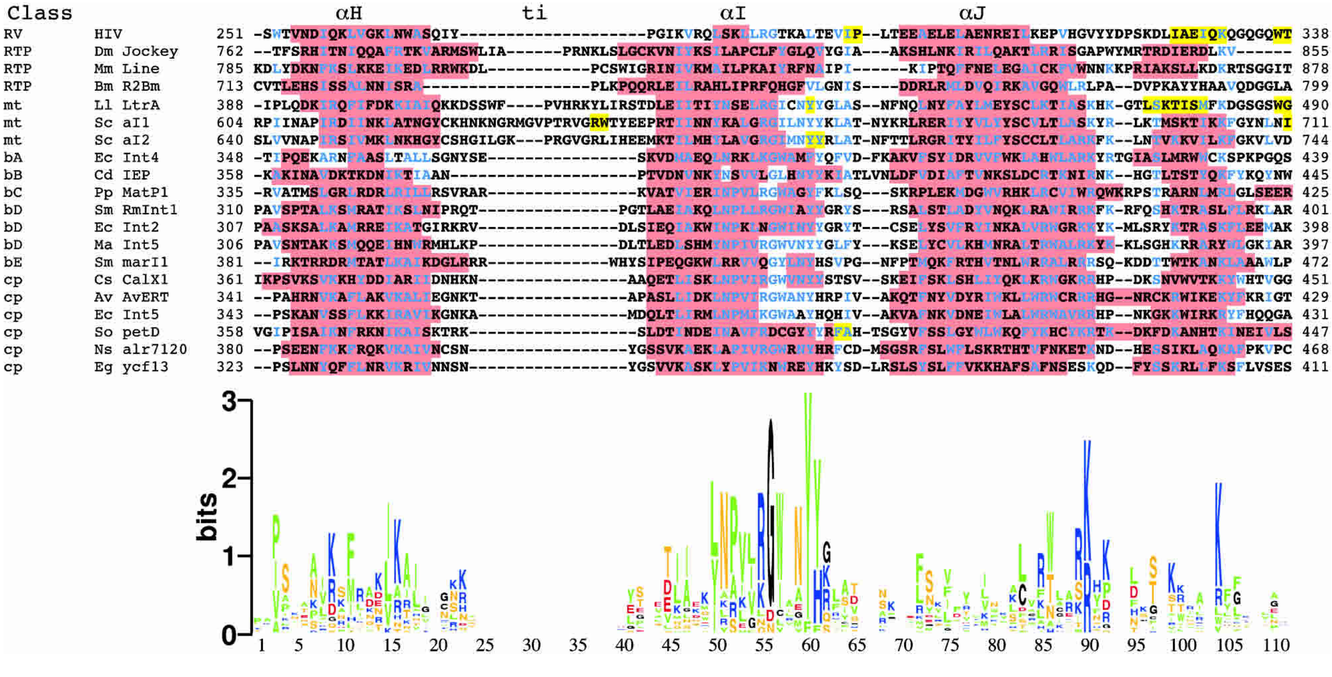
Class
αH
ti
αI
αJ
RV HIV 251 -SWTVNDIQKLVGKLNWASQIY--------------------PGIKVRQLSKLLRGTKALTEVIP--LTEEAELELAENREILKEPVHGVYYDPSKDLIAEIQKQGQGQWT 338
RTP Dm Jockey 762 --TFSRHITNIQQAFRTKVARMSWLIA------PRNKLSLGCKVNIYKSILAPCLFYGLQVYGIA---AKSHLNKIRILQAKTLRRISGAPWYMRTRDIERDLKV------ 855
RTP Mm Line 785 KDLYDKNFKSLKKEIKEDLRRWKDL-----------PCSWIGRINIVKMAILPKAIYRFNAIPI------KIPTQFFNELEGAICKFVWNNKKPRIAKSLLKDKRTSGGIT 878
RTP Bm R2Bm 713 CVTLEHSISSALNNISRA------------------PLKPQQRLEILRAHLIPRFQHGFVLGNIS----DDRLRMLDVQIRKAVGQWLRLPA--DVPKAYYHAAVQDGGLA 799
mt Ll LtrA 388 -IPLQDKIRQFIFDKKIAIQKKDSSWF----PVHRKYLIRSTDLEIITIYNSELRGICNYYGLAS--NFNQLNYFAYLMEYSCLKTIASKHK-GTLSKTISMFKDGSGSWG 490
mt Sc aI1 604 RPIINAPIRDIINKLATNGYCKHNKNGRMGVPTRVGRWTYEEPRTIINNYKALGRGILNYYKLAT-NYKRLRERIYYVLYYSCVLTLASKYR--LKTMSKTIKKFGYNLNI 711
mt Sc aI2 640 SLVVNAPIRSIVMKLNKHGYCSHGILGK---PRGVGRLIHEEMKTILMHYLAVGRGIMNYYRLAT-NFTTLRGRITYILFYSCCLTLARKFK--LNTVKKVILKFGKVLVD 744
bA Ec Int4 348 -TIPQEKARNFAASLTALLSGNYSE----------------SKVDMAEQLNRKLKGWAMFYQFVD-FKAKVFSYIDRVVFWKLAHWLARKYRTGIASLMRWWCKSPKPGQS 439
bB Cd IEP 358 -KAKINAVDKTKDNIKTIAAN--------------------PTVDNVNKYNSVVLGLHNYYKIATLVNLDFVDIAFTVNKSLDCRTKNIRNK--HGTLTSTYQKFYKQYNW 445
bC Pp MatP1 335 -RVATMSLGRLRDRLRILLRSVRAR----------------KVATVIERINPVLRGWAGYFKLSQ---SKRPLEKMDGWVRHKLRCVIWRQWKRPSTRARNLMRLGLSEER 425
bD Sm RmInt1 310 PAVSPTALKSMRATIKSLNIPRQT---------------PGTLAEIAKQLNPLLRGWIAYYGRYS---RSALSTLADYVNQKLRAWIRRKFK-RFQSHKTRASLFLRKLAR 401
bD Ec Int2 307 PAASKSALKAMRREIKATGIRKRV---------------DLSIEQIAKWINPKLNGWINYYGRYT---CSELYSVFRYINKALVRWGRKKYK-MLSRYKTRASKFLEEMAK 398
bD Ma Int5 306 PAVSNTAKKSMQQEIHNWRMHLKP---------------DLTLEDLSHMYNPIVRGWVNYYGLFY---KSELYCVLKHMNRALTRWALRKYK-KLSGHKRRARYWLGKIAR 397
bE Sm marI1 381 --IRKTRRDRMTATLKAIKDGLRRR-------------WHYSIPEQGKWLRRVVQGYLNYHSVPG--NFPTMQKFRTHVTNLWRRALRRRSQ-KDDTTWTKANKLAAAWLP 472
cp Cs CalX1 361 IKPSVKSVKKHYDDIARIIDNHKN----------------AAQETLISKLNPVIRGWVNYYSTSV--SKEIFSKLSHLIYQKLKRWGKRRHP--DKSNVWVTKKYWHTVGG 451
cp Av AvERT 341 --PAHRNVKAFLAKVKALIEGNKT----------------APASLLIDKLNPVIRGWANYHRPIV--AKQTFNYVDYRIWKLLWRWCRRRHG--NRCKRWIKEKYFKRIGT 429
cp Ec Int5 343 --PSKANVSSFLKKIRAVIKGNKA----------------MDQLTLIRMLNPMIKGWAAYHQHIV--AKVAFNKVDNEIWLALWRWAVRRHP--NKGKKWIRKRYFHQQGA 431
cp So petD 358 VGIPISAIKNFRKNIKAISKTRK------------------SLDTINDEINAVFRDCGYYYRFAH-TSGYVFSSLGYWLWKQFYKHCYKRTK--DKFDKANHTKINEIVLS 447
cp Ns alr7120 380 --PSEENFKKFRQKVKAIVNCSN-----------------YGSSVKAEKLAPIVRGWRNYHRFCD-MSGSRFSLWFLSKRTHTVFNKETKND--HESSIKLAQKAFPKVPC 468
cp Eg ycf13 323 --PSLNNYQFFLNRVKRIVNNSN-----------------YGSVVKASKLYPVIKNWREYHKYSD-LRSLSYSLFFVKKHAFSAFNSESKQD--FYSSKRLLFKSFLVSES 411
3
2
1
0
bits
1 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110

FIGURE 5. Predicted secondary structure of the domain X region in non-LTR-retrotransposon and group II intron RTs. The corresponding regions of the thumb and connection domain of HIV-1 RT are shown *above*. The alignment was made by positioning the JPred predicted secondary-structure elements and then adjusted manually to maximize homology. Abbreviations, accession numbers, and symbols are as in Figures 3 and 4. The sequence logo (Schneider and Stephens 1990; Crooks et al. 2004) shows the information content (3 bits = no degeneracy) for each position in domain X of the group II IEPs. Amino acids are colored according to properties: hydrophobic, green (P, L, I, V, M, F, W, Y, and A); basic, blue (R, K, and H); acidic, red (D and E); polar, yellow (N, Q, S, and T); and black (G and C).

αH, αI, and αJ in the thumb of HIV-1 RT. The sizes of these helices and the spacing between them are similar to that in HIV-1 RT, except that the somewhat longer spacing previously noted between αH and αI in LtrA (insertion ti) is also found in the other mitochondrial lineage IEPs. Furthermore, in all the proteins, the predicted αH and αI helices have conserved hydrophobic residues at a spacing of ~3.5 residues, placing them on one side of the helix. The two most strongly conserved motifs in domain X, RGWXNYY (LtrA residues 437–443) and R(K/R)XK (LtrA residues 469–472), are found at or near the C termini of αI and αJ, respectively.

The domain X region of group II intron RTs extends downstream of αJ into the region corresponding to the connection domain of HIV-1 RT (Mohr et al. 1993). This downstream region contains a conserved lysine residue at domain X position 103 (K483 in LtrA) (Fig. 5), along with a number of moderately conserved residues and is predicted to form an α-helix in most proteins, LtrA being one of the exceptions. In HIV-1 RT, this region of the connection domain begins a three-stranded β-sheet, which is part of the dimerization interface.

Together, these findings suggest that there is conserved structural similarity between the domain X regions of group II intron RTs, the corresponding regions of non-LTR-retrotransposon RTs, and the thumb of HIV-1 RT, consistent with a common evolutionary origin of the thumb in all these RTs.

Three-dimensional model of LtrA

The sequence homology between LtrA and HIV-1 RT enabled us to construct a three-dimensional model of LtrA by threading the aligned amino acid sequence of the RT, X, and D domains onto X-ray crystal structures of HIV-1 RT (Fig. 6). LtrA is a homodimer (Saldanha et al. 1999), but there are indications that it is monomeric in solution with dimers assembling on the intron RNA (Rambo and Doudna 2004). Thus, it is possible that the two subunits adopt different tertiary structures dictated by asymmetric interactions with the intron RNA and/or the other subunit. For the modeling, we made the likely assumption that one subunit of LtrA (subunit A) has a structure analogous to that of the active p66 subunit of HIV-1 RT. The second subunit (sub-

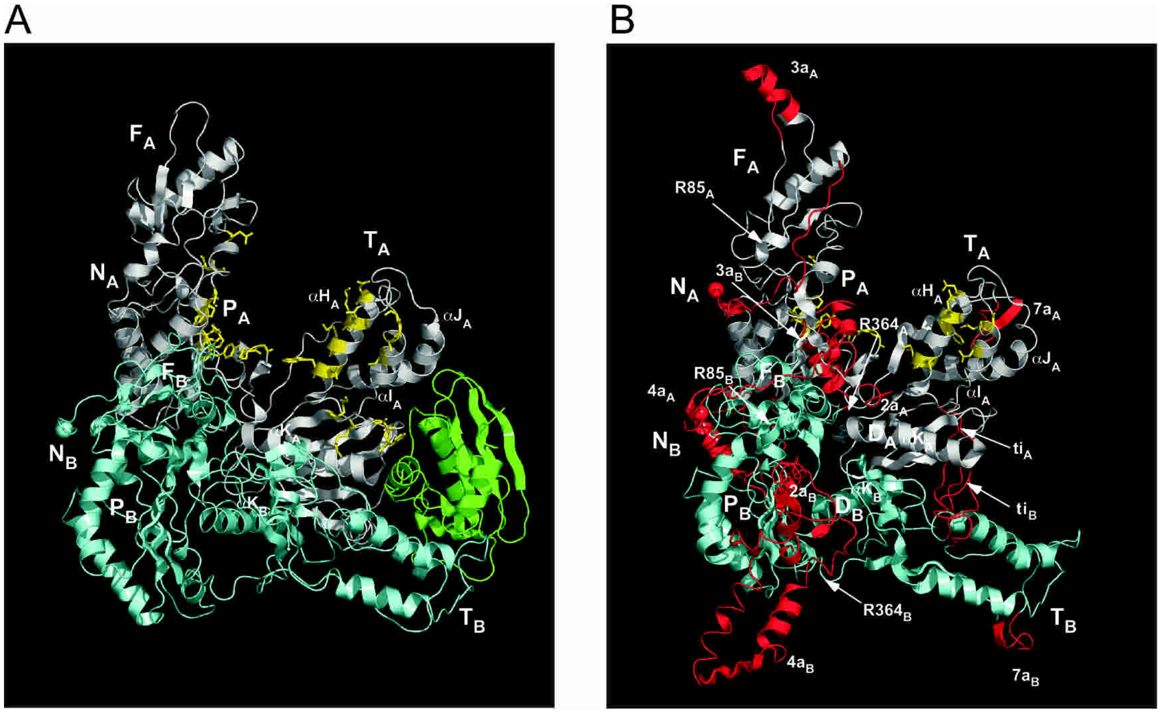

FIGURE 6. Three-dimensional model of LtrA and comparison with HIV-1 RT. (*A*) X-Ray crystal structure of HIV-1 RT (Ding et al. 1998; 2HMI.pdb). The p66 subunit is white with the RNase H domain in green, and the p51 subunit is light cyan. Amino acid residues involved in binding template–primer DNA (Ding et al. 1998; 2HMI.pdb) are shown in yellow with side chains. (*B*) LtrA model. Subunits A and B corresponding to HIV-1 RT p66 and p51 are white and light cyan, respectively. Insertions in LtrA relative to HIV-1 RT are red and named according to the insertion number, followed by a subscript A or B denoting the subunit. Cognate amino acids in the RT and X domain that align with those involved in template–primer binding in HIV-1 RT are shown in yellow with side chains. The matrix used was that of Henikoff and Henikoff (1992; see legend Fig. 3), plus D or K ≡ N, R ≡ Q, Q ≡ K, and N ≡ S (Yeerassamy et al. 2003). Palm (P), fingers (F), thumb (T) in both proteins, and DNA-binding domain (D) in LtrA are indicated, with subscripts A or B denoting subunits A (p66) or B (p51), respectively. The N termini of HIV-RT p66 and p51 and the modeled regions of LtrA subunits A and B (N_A and N_B, respectively) are indicated with spheres. R85 and R364 are amino acid residues at the major proteolytic cleavage sites.

unit B) could be structurally similar to HIV-1 RT's p66 or p51 subunits, or it could have some variation of these structures. To enable us to model dimers based on HIV-1 RT X-ray crystal structures, subunit B of LtrA was threaded onto the structure of the p51 subunit. This is an equivocal assumption, but necessary since the HIV-1 RT is at present the only available RT dimer structure. For threading, we used both the HIV-1 RT "open" structure in the presence of bound template–primer (Ding et al. 1998; 2HMI.pdb) and the "closed" structure with the fingers shifted down to bind the incoming dNTP (Huang et al. 1998; 1RTD.pdb). LtrA threads equally well on both structures, and only the model based on the open structure is shown.

Figure 6 shows ribbon diagrams comparing the LtrA model with the HIV-1 RT structure (Ding et al. 1998; 2HMI.pdb). Insertions in LtrA relative to HIV-1 RT are red; shown in yellow with side chains are amino acid residues that interact with the template–primer in HIV-1 RT and similarly situated cognate residues in LtrA's RT and X domains (see below). Figure 7 shows three views of the LtrA model rendered as surface diagrams, with panels A–C displaying electrostatic surface potential, and panels D–F highlighting the most highly conserved regions in unigenic evolution analysis of splicing-competent LtrA variants (Cui et al. 2004).

LtrA fits very well onto the core tertiary structure of HIV-1 RT, with subunit A folding into fingers, palm, and thumb (Fig. 6). The dimerization interface between the p66 palm and the p51 fingers of HIV-1 RT is maintained in the LtrA model. However, the dimerization interface between the connection domains of p66 and p51 differs significantly, reflecting that most of the HIV-1 RT connection domain is replaced in LtrA by the D and En domains, whose contribution to dimerization is unknown. In the model, the predicted α-helices in the D domains of the two subunits of LtrA are placed at this interface in the same orientation as the αK helices in the two subunits of HIV-1 RT. In both HIV-1 RT and the LtrA model, these helices present long

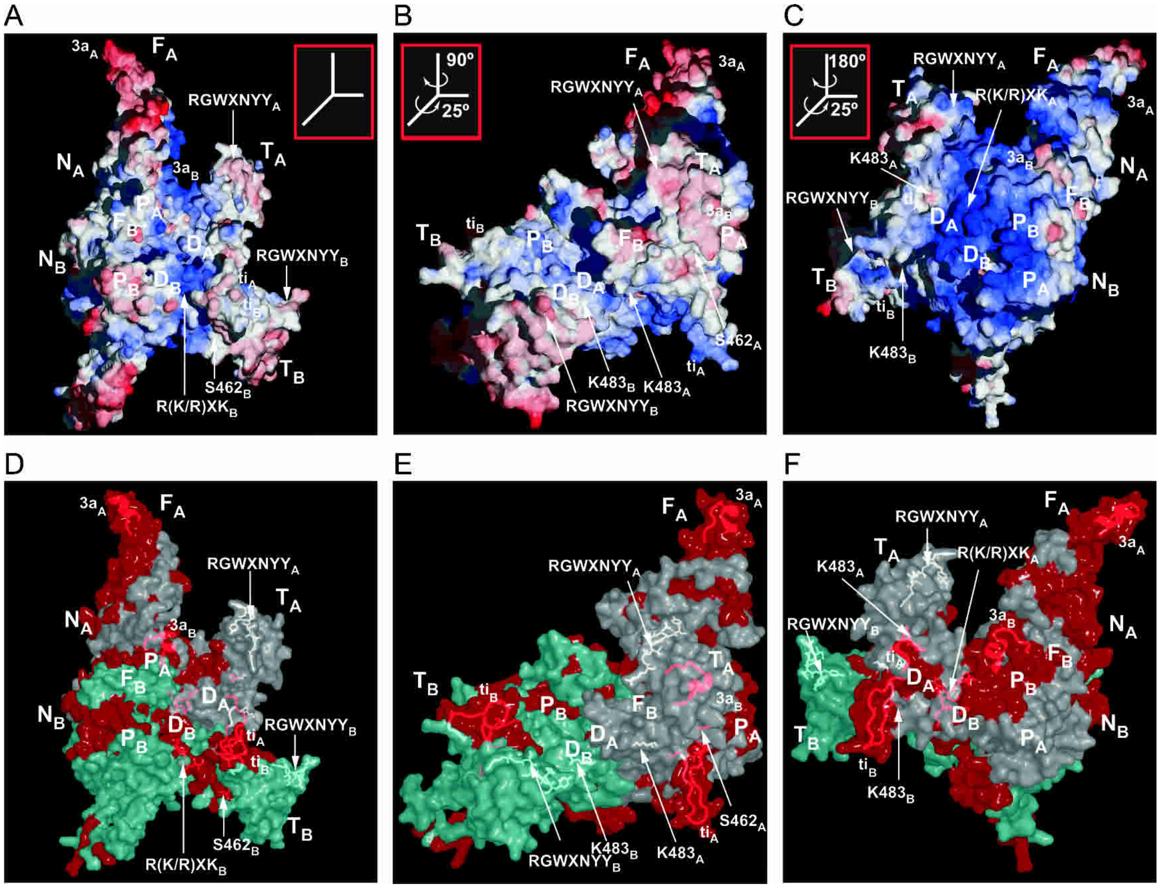

FIGURE 7. LtrA model showing electrostatic surface potential and regions conserved in splicing-competent LtrA variants in unigenic evolution analysis. (*A–C*) Surface diagram colored according to electrostatic potential. The computationally set cutoffs are negative charge (red) -1, neutral (white) $+5$, and positive (blue) $+11$. (*D–F*) Translucent surface diagrams in the same orientations as *A–C*, highlighting regions that were hypomutable in unigenic evolution analysis of splicing-competent LtrA variant selected from a library with random mutations. Residues that fall within a 25-amino-acid sliding window with mutability values ≤ -0.25 are red, and those with scores > -0.25 are white or light cyan for subunits A and B, respectively (see Cui et al. 2004 for an explanation of mutability values). The side chains of key amino acids S462, K483, R(K/R)XKA, and RGWXNYY can be seen at or beneath the surface, while insertions ti and 3a are shown as ribbons beneath the surface. The views in panels *A* and *D* are similar to those in the ribbon diagram of Figure 6B.

hydrophobic surfaces that interact with distal regions of the other subunit. The third dimerization interface between the p51 thumb and the p66 RNase H domain in HIV-1 RT is not present in LtrA, but also appears to be the least important, since p51/p51 homodimers lacking this interface are partially functional (Wang et al. 1994). Weaker dimerization interactions could account for the finding that LtrA appears to dimerize only on RNA binding (Rambo and Doudna 2004).

The model based on HIV-1 RT places the N termini of LtrA subunits A and B in proximity to each other on the same side of the protein. The N-terminal 36 residues of LtrA were not modeled, while the remainder of the N-terminal extension, including RT0, was modeled on the HIV-1 RT structure based on the alignment of the α-helix upstream of RT1 (see Fig. 3). The model places most of the insertions in LtrA's RT domain on the back of the hand and behind the thumb, oriented away from the palm and finger domains, so as not to interfere with the formation of the RT active site (see Materials and Methods for details of how the insertions were modeled).

Domain X models well on the tertiary structure of the HIV-1 RT thumb, with the three predicted α-helices, corresponding to αH, αI, and αJ, packing together in the same relative orientation as in HIV-1 RT. The small insertion, ti, between αH and αI protrudes on the side farthest away from the fingers and the DNA-interaction sites (Fig. 6B). The domain X sequence RGWXNYY (residues 437–443), which is conserved in group II IEPs (see Fig. 5), is situated near the tip of the thumb in both subunits (Fig. 7), while the conserved sequence R(K/R)XK (residues 469–472), which is near the end of αJ, is located on the surface in subunit B, but is buried in a hydrophobic pocket at the base of the thumb in subunit A (Fig. 7). The region of domain X downstream of αJ, which includes the conserved residue K483 (see Fig. 5), lies on the exterior of the protein in both subunits (Figs. 6, 7). K483 is completely solvent-accessible on the surface of the protein, while S462, a site of mutations affecting maturase activity (Moran et al. 1994; Cui et al. 2004), is located in a small hydrophobic pocket (Fig. 7).

Domain D is modeled on the aligned region of the connection domain of HIV-1 RT, with the predicted α-helix in domain D taking the place of αK (Fig. 6). In subunit A, this α-helix lies almost parallel to the wall of the nucleic acid-binding track, while in subunit B, the domain D α-helix is perpendicular to that in subunit A, with its N terminus near the nucleic acid-binding track. As mentioned previously, αK of HIV-1 RT forms part of the dimerization interface between the connection domains and includes some residues that could potentially interact with the template–primer (Ding et al. 1998; Huang et al. 1998). Although there is no experimental evidence that the predicted domain D α-helix is functionally equivalent to αK of HIV-1 RT, the position of the domain D helix in the model raises the possibility that it functions in both DNA-binding and dimerization.

The model readily accounts for the major proteolytic cleavage sites (see positions R85 and R364 in Fig. 6). In both subunits, the Arg-C cleavage site in RT1 (R85) is located in a solvent-exposed loop immediately preceding a small β-strand, and the Arg-C and trypsin cleavage sites at R364, R371, and R378 are located in the β-sheet between the palm and the thumb (see also Fig. 3). The latter cleavage sites are analogous to two protease-sensitive sites found in the linker between the palm and thumb of MMLV RT (Georgiadis et al. 1995). In HIV-1 RT, a protease cleavage site was found in the same β-sheet in RT7 after partial denaturation with SDS (Lowe et al. 1988). With the exception of R371, which is in the middle of a predicted β-strand, all the cleavages in LtrA are predicted to be in loops preceding β-strands (see also Fig. 3).

Potential nucleic acid-binding regions

The template–primer binding track of HIV-1 RT is a positively charged groove that extends from the polymerase active site to the RNase H domain (Bebenek et al. 1997; Ding et al. 1998; Huang et al. 1998). In both the open and closed structures, contacts with the template–primer are made by the p66 template grip (RT2 and RT4), YMDD in RT5, primer grip (RT7), αH and αI in the thumb, parts of the p66 and p51 connection domains, and the RNase H domain (Ding et al. 1998; Huang et al. 1998; Morris et al. 1999).

In the LtrA model, the fingers, palm, and thumb form a putative template–primer binding track analogous to that of HIV-1 RT. In the electrostatic surface potential diagram, this track is seen as a ribbon of positive charge (blue) that extends along the groove formed by the fingers and thumb of subunit A, passes over D_A and into subunit B (Fig. 7A). In subunit A, the track includes regions corresponding to the HIV-1 RT template-grip (RT2 and RT4), RT-5, and primer-grip (RT7), as well as α-helices αH and αI of the thumb/domain X, which interact with the minor groove of the template–primer duplex (Ding et al. 1998; Huang et al. 1998). Furthermore, many of the residues that contact the template-primer in the HIV-1 RT and thumb domains have similarly situated identical or cognate residues in LtrA—12 of 23 using the similarity matrix of Henikoff and Henikoff (1992) and 17 of 23 using a more relaxed matrix with D or K ≡ N, R ≡ Q, Q ≡ K, and N ≡ S (Veerassamy et al. 2003; residues are underlined for HIV-1 RT in Fig. 3 and shown in yellow with side chains in Fig. 6). Interestingly, parts of two conserved insertions in LtrA, $2a_A$ and $3a_B$, lie near residues potentially involved in template–primer binding (Fig. 6B). We note that LtrA is considerably more basic than HIV-1 RT (pIs 9.60 and 8.69, respectively, as calculated by the algorithm ProtParam; Wilkins et al. 1999), and the scale of the electrostatic potential diagram has been appropriately adjusted to highlight the track (see Fig. 7 legend).

Recently, we used unigenic evolution to identify regions of LtrA that are highly constrained ("hypomutable") in splicing-competent LtrA variants isolated from a library containing random mutations (Cui et al. 2004). These studies together with additional biochemical analysis of mutant proteins showed that the N terminus of the RT domain is required for interaction with the high-affinity binding site in intron subdomain DIVa, while other regions of the RT and X domains may interact with conserved regions of the intron's catalytic core. In Figure 7D–F, portions of LtrA that were highly constrained regions in the unigenic evolution analysis (mutability values ≤ −0.25 calculated across a 25-amino-acid residue sliding window) are red. These regions potentially comprise an extended RNA-binding surface consisting of parts of the RT and X domains in and around the template–primer binding track, distal regions of the fingers, including all of insertion 3a, and patches on the back of the molecule, including all of thumb insertion ti. Indeed, insertions 3a and ti were among the most highly constrained sequences in the unigenic evolution analysis, arguing for their functional importance (Cui et al. 2004). S462 and the sequence R(K/R)XK in the thumb fall within the highly constrained regions (Fig. 7D–F), while K483 and RGWXNYY were also constrained in the unigenic evolution analysis (Cui et al. 2004), but are not part of larger regions that meet the statistical or length criteria to be colored red in Figure 7D–F. Additionally, the constrained regions in the nucleic acid-binding track and the back of the hand overlap with the most positively charged regions in the electrostatic potential diagram (Fig. 7D–F). Nucleic acid binding on the back of the hand has been found in T7 RNA polymerase (Tahirov et al. 2002; Yin and Steitz 2002, 2004) and suggested in a model for HIV-1 RT binding of the $\mathrm{tRNA_3}^{\mathrm{Lys}}$ primer (Isel et al. 1999). The highly constrained regions in and around the nucleic acid-binding track in the LtrA model suggest that there could be some overlap in protein regions that bind the intron RNA for RNA splicing and those that later bind the cleaved target DNA for initiation of reverse transcription. If so, the high-affinity binding of DIVa to the N-terminal fingers region and/or other sites on the back side of LtrA may be critical for maintaining contact with the intron RNA, when weaker RNA contacts in the template–primer binding track are displaced by DNA for initiation of reverse transcription.

Model with double-stranded DNA target site

The RNP complex initiates mobility by recognizing DNA target sites, using a mechanism in which both the IEP and base-pairing of the intron RNA contribute to the recognition of DNA target sequences (Singh and Lambowitz 2001). The first step in DNA target site recognition is thought to involve major groove interactions between the IEP and nucleotide residues in the distal 5′-exon region, including T−23, G−21, and A−20 (Singh and Lambowitz 2001). These base interactions bolstered by phosphate-backbone interactions lead to local DNA unwinding, enabling the intron RNA to base pair to target site positions between positions −12 and +2 (EBS/IBS and δ–δ′ interactions). Bottom-strand cleavage between positions +9 and +10 occurs after a lag and requires additional interactions between the IEP and 3′-exon, the most critical being recognition of T+5 (Mohr et al. 2000; Singh and Lambowitz 2001).

We wished to use the LtrA model to visualize how group II IEPs might interact with the DNA target site during intron mobility. It seemed most instructive to consider the bottom-strand cleavage step. This step requires both the initial IEP contacts with the 5′-exon and additional contacts with the 3′-exon (see Mohr et al. 2000; Singh and Lambowitz 2001), but it is not known whether the contacts with the 5′-exon precede those with the 3′-exon or whether the two sets of contacts occur simultaneously.

To incorporate the bottom-strand cleavage step, a modeled En domain (red) and a docked Ll.LtrB target DNA (green and yellow for top and bottom strands, respectively) were added to the LtrA dimer model (Fig. 8). The En domain was modeled on the phage T4 endonuclease VII structure (Raaijmakers et al. 2001; 1EN7.pdb), with the H-N-H active site positioned to cleave the scissile phosphate in the target DNA, as in the X-ray crystal structure of colicin E7 with bound DNA substrate (Hsia et al. 2004; 1PT3.pdb). The N terminus of the En domain is 35 Å from the C terminus of monomer A and 54 Å from that of monomer B, but is left unattached to either subunit. Attachment would require unfolding ~5 amino acids at the C terminus of domain D and ~15 amino acids at the N terminus of the En domain at little cost to the integrity of the model, since these 20 residues are already in an unstructured conformation. The attachment would be easier for monomer B, since there is no intervening protein or DNA, but would require restructuring of intervening protein regions for monomer A.

The trajectory of the DNA backbone was modeled based on the HIV-1 RT model structure described in Peletskaya et al. (2004; 1R0A.pdb), which adds a 5′-extension to the DNA template strand in the crystal structure of Huang et al. (1998; 1RTD.pdb). This structure includes a 40° bend with a change from B-form to A-form DNA 5 bp to the primer side of the RT active site. Such a bend is commonly found for template–primer DNA bound to RTs, RNA polymerases, and some DNA polymerases (Ding et al. 1997). The template (top) strand (green in Fig. 8) continues around the back of the fingers (Peletskaya et al. 2004) and is extended as B-form duplex DNA from position −14 to −30, a region known to be double-stranded in complex with Ll.LtrB RNPs (Singh and Lambowitz 2001). Although the modeled DNA is clearly a simplification of the actual structure in which the intron RNA is base-paired to the top strand between positions −12 and +2 (Singh and Lambowitz 2001), it provides insight into distance considerations.

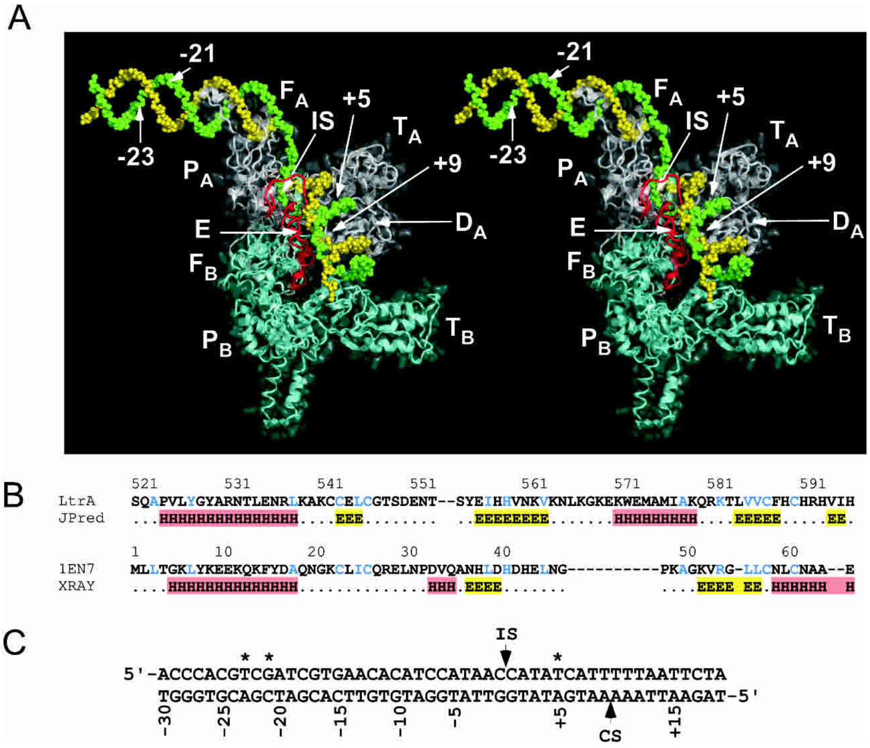
A
-21
F_A
+5
IS
T_A
-23
P_A
+9
D_A
E
F_B
P_B
T_B
B
521 531 541 551 561 571 581 591
LtrA SQAPVLYGYARNTLENRLKAKCCELCGTSDENT--SYEIHHVNKVKNLKGKEKWEMAMIAKQRKTLVVCFHCHRHVIH
JPred ...HHHHHHHHHHHHHHHH....EEE........ ..EEEEEEEE.......HHHHHHHHH....EEEEE.....EE.
1 10 20 30 40 50 60
1EN7 MLLTGKLYKEEKQKFYDAQNGKCLICQRELNPDVQANHLDHDHELNG----------PKAGKVRG-LLCNLCNAA--E
XRAY ....HHHHHHHHHHHHHHH..............HHH.EEEE....... ....EEEE EE.HHHHHH H
C
IS
* * *
5'-ACCCACGTCGATCGTGAACACATCCATAACCATATCATTTTTAATTCTA
TGGGTGCAGCTAGCACTTGTGTAGGTATTGGTATAGTAAAAATTAAGAT-5'
-30 -25 -20 -15 -10 -5 +5 +15
CS

FIGURE 8. LtrA dimer model with the En domain docked to the DNA target site for the bottom-strand cleavage step. (*A*) Stereoviews of the LtrA model with docked DNA. The protein is shown as a ribbon diagram, with the DNA backbone in space-filling representation. The protein model has a translucent surface representation superimposed on the ribbon and space-filling representation to make it easier to see where the protein boundaries fall relative to the DNA. The En domain (red) is modeled on the structure of phage T4 endonuclease VII (Raaijmakers et al. 2001; 1EN7.pdb), with its active site positioned to interact with the scissile phosphate between bottom-strand positions +9 and +10 of the Ll.LtrB DNA target site, as in the structure of colicin E7 with bound DNA substrate (Hsia et al. 2004; 1PT3.pdb). The backbone structure of the DNA is based on that in the HIV-RT model structure of Peletskaya et al. (2004; 1R0A.pdb) extended on both ends as canonical B-form helix and with its sequence changed to match the Ll.LtrB DNA target sequence (panel *C*). The top and bottom strands of the DNA are green and yellow, respectively with bottom-strand position +9, top-strand positions −23, −21 and +5, and the intron-insertion site (IS) between top-strand positions −1 and +1 labeled. Monomers A and B are white and light cyan, respectively. (*B*) Amino acid sequence alignment of LtrA's En domain with endonuclease VII (accession no. P13340). JPred-predicted and X-ray crystallography determined secondary structures are shown below. α-Helices and β-strands are highlighted in red and yellow, respectively. Identical and similar amino acid residues, defined as in Figure 3, are shown in blue. (*C*) Ll.LtrB target sequence. The intron-insertion site (IS) and bottom-strand cleavage sites (CS) are indicated. Asterisks indicate the T−23, G−21, and T+5 nucleotide residues recognized by the IEP.

Despite the considerable uncertainties in the model, it is remarkable that with the En active site placed at bottom-strand position +9, the top strand T+5, which is critical for bottom strand cleavage, is positioned within 4 Å of domain D_A residue Y529, a site of mutations that specifically inhibit second-strand cleavage (San Filippo and Lambowitz 2002). Further, the intron-insertion site (IS) between top-strand positions −1 and +1 falls near the subunit A fingers facing out. From the latter position, the intron RNA, which is not shown in the model, could continue to interact with potential RNA-binding sites in distal regions of the fingers and the back of the hand. Insertions $2a_A$ and $3a_B$ are close to the 3′-end of the priming (bottom) strand.

As indicated above, the initial contacts with T−23, G−21, and A−20 in the distal 5′-exon region of the DNA target site are required for both reverse splicing and bottom-strand cleavage. In the model, top-strand position −21, the most critical residue in the distal 5′-exon region, is approximately one half-turn on the back side of subunit A near the back of the fingers. If the DNA were modeled as straight B-form helix throughout, position −21 would be even farther out of the range of the protein. Thus, in order to contact G−21 and T+5 simultaneously, the DNA would have to be bent backward. An audacious possibility is that the bend could be sharp enough for G−21 to contact domain D on the back side of the protein. Alternatively, positions −21 and +5 may be contacted at different times (e.g., if the protein binds first to the distal 5′-exon region for the initial reverse splicing step, then binds to +5 to catalyze second-strand cleavage), or the two subunits of the dimer may be oriented differently than in HIV-1 RT, so that the D domains of different subunits are on opposite sides of the molecule.

Finally, we note that mutant LtrA proteins deleted for the En domain or with mutations in its conserved sequence motifs retain maturase activity, but lack RT activity (San Filippo and Lambowitz 2002). These findings suggest that the En domain is not required for the formation of the core structure or dimerization, both of which are presumably required for maturase activity, but might at some stage of the reaction interact with and activate the RT domain. Such an interaction could well occur after the bottom-strand cleavage step modeled in Figure 8, concomitant with the repositioning of position +10 to the RT active site to initiate reverse transcription. Recent studies have provided evidence that a potentially analogous interaction with HIV-1 integrase is required for initiation of reverse transcription by HIV-1 RT (Zhu et al. 2004).

Summary

Together, our results suggest that the RT and X domains of group II IEPs are structurally homologous to the RT and thumb domains of HIV-1 RT, except for an N-terminal extension and several insertions that are present in group II

IEPs, but not in retroviral RTs. Furthermore, partial proteolysis is consistent with the modeled structure, and a major protease-sensitive stretch is coincident with prominent cleavage sites in retroviral RT. An important observation from our studies is that the N-terminal extension and insertions 2a and 3a appear conserved structurally in other group II intron and non-LTR-retrotransposon RTs, while insertions 4a, 7a, and ti are conserved structurally only in some lineages of group II intron RTs (Fig. 4). The conservation of these insertions in different RTs as well as unigenic evolution analysis, which shows that insertions 3a and ti are highly constrained in splicing-competent LtrA variants (Cui et al. 2004), suggests that they are functionally important. One possibility is that the N-terminal extension and some of the insertions contribute to the specific binding of the RNA template, which determines the cDNA initiation site in group II intron and non-LTR-element RTs (see Introduction). If so, these regions may have been lost during the evolution of retroviral RTs, concomitant with the streamlining of the cDNA initiation mechanism to use a base-paired RNA primer.

Another major conclusion from our results is that domain X of group II IEPs and the corresponding region of non-LTR-retroelement RTs appear structurally homologous to the thumb of retroviral RT. In all cases, this region is predicted to contain three α-helices, whose size and spacing are similar to those of the three α-helices in the HIV-1 and MMLV RT thumbs (Fig. 5). Furthermore, as in retroviral RTs, we find that a major site of proteolytic cleavage is between RT7 and the thumb/domain X (see also Rambo and Doudna 2004), presumably reflecting similar protein folds that leave the junction exposed. Together, these findings are consistent with a common evolutionary origin for the thumb of all these RTs, with divergence and acquisition of a role in RNA splicing in the case of group II IEPs. Domain X in group II IEPs contains an additional conserved region that is located downstream of the three predicted α-helices and may also contribute to RNA splicing. Additionally, domain X of LtrA and other mitochondrial lineage IEPs contains a small insertion, ti, which is highly constrained in splicing-competent LtrA variants (Cui et al. 2004) and could be a recent structural adaptation for RNA splicing in these IEPs.

The three-dimensional model of LtrA suggests that at least one LtrA subunit likely has a structure analogous to that of the active p66 subunit of HIV-1 RT, with a template–primer binding track that contains appropriately positioned cognates of many of the amino acid residues involved in template–primer binding in HIV-1 RT (Fig. 6). By using the model to display the results of unigenic evolution analysis, we found that regions that are highly constrained in splicing competent LtrA variants are located in and around the template–primer binding track, in the extended fingers region, and on the back of the hand, with the sites in the template–primer binding track and the back of the hand overlapping the most basic regions of the protein. These findings suggest an extended nucleic acid-binding surface that could interact with different regions of the intron RNA to stabilize the active RNA structure. Finally, the docking of target DNA to the LtrA model indicates that LtrA is likely too small to simultaneously contact nucleotide residues in the distal 5′-exon and 3′-exon regions of the DNA target site, unless the target DNA is bent. Thus, the model frames specific questions about DNA and RNA binding that can now be addressed experimentally.

MATERIALS AND METHODS

Protein purification and RNP formation

LtrA was expressed in *E. coli* BL21(DE3) from pImp-1P, and unspliced precursor and lariat RNA were made from pGMΔORF, as described (Saldanha et al. 1999). RNPs were formed with 2 μM LtrA (0.4 mg/mL) and 5 μM gel-purified intron lariat (2 mg/mL), so that all protein molecules are potentially bound. In addition to standard conditions, RNP formation was also done at 450 mM NaCl. Prior to the reaction, 14–28 μM intron RNA (5.6–11 mg/mL) was renatured in 100 mM NaCl, 5 mM $MgCl_2$, 10 mM Tris-HCl (pH 7.5). The mixture was heated to 65°C and slowly cooled to room temperature. Then, 2.8 μM LtrA (0.5 mg/mL) was added, and the buffer conditions were adjusted to either 100 or 450 mM NaCl, 5 mM $MgCl_2$, 20 mM Tris-HCl (pH 7.5), and 5 mM β-mercaptoethanol. The RNA and LtrA were incubated at 25°C for 45 min to allow for complex formation. LtrA was incubated with 2 mg/mL yeast tRNA (Invitrogen) in the same protocol used for RNP formation with purified lariat RNA.

Protease digestion and identification of proteolysis products

LtrA, in the absence of RNA, or in the presence of lariat or precursor RNA, was digested with trypsin (Promega) or Arg-C (Roche) at a final concentration of 1.9–3.8 ng/μL in 100-μL reactions containing 0.4 mg/mL LtrA in Tris-HCl (pH 8). Reaction mixtures were incubated at 37°C or at room temperature for 60 min. Aliquots (20 μL) were removed at different times and quenched by adding SDS-PAGE loading dye. Proteolysis products were separated by 15% SDS-PAGE.

LtrA proteolysis products were analyzed to determine fragment sequences by mass spectrometry or N-terminal sequencing. HPLC and mass spectrometry were performed as described by Derbyshire et al. (1997). For microsequencing, the protein was blotted from the gel, and Edman degradation was performed as described by Yao et al. (1996).

Sequence alignments and secondary-structure predictions

Protein sequences were aligned using conserved sequences in RT0, 1, 2a, 4, 5, and 7, and alignments were refined manually based on predicted secondary structures and multiple sequence alignments of HIV-1, group II intron, and non-LTR-retroelement RTs. Secondary-structure predictions were made by the JPred server (http://www.compbio.dundee.ac.uk/~www-jpred/submit.html; Cuff et al. 1998; Cuff and Barton 2000). This server first uses PSI-BLAST to scan a filtered SWISS-PROT/TRMBL database for related sequences and uses them to generate multiple sequence alignments. PSI-BLAST and HMMPSSM profiles extracted from the alignment are then used as input for Jnet, two connected neural networks trained on a set of 480 known protein structures. The first predicts the propensity for coil, α-helix, or β-sheet at each position using a 17-residue sliding window, and the second uses the output of the first to refine the prediction at each position.

Three-dimensional structural modeling

A three-dimensional model of LtrA was constructed by threading the aligned primary sequence onto X-ray crystal structures of HIV-1 RT (Ding et al. 1998; 2HMI.pdb; Huang et al. 1998; 1RTD.pdb). Threading was done using the alignment interface tool of Swiss Model (Schwede et al. 2003). The threading program automatically models insertions relative to HIV-1 RT by searching the database for similar sequences of known tertiary structure and places them in situ by reducing steric and electrostatic clashes. The secondary structures of the insertions as modeled by Swiss Model were altered manually when necessary to fit the JPred prediction (see above) by adjusting the ϕ/ψ backbone angles, while maintaining the original placement and orientation of the insertions. The model was energy minimized by 400 steps of steepest descent using the GROMOS96 force field with a 9 Å nonbonded cutoff (Scott et al. 1999). The two subunits of LtrA dimers were docked by aligning the modeled monomers to the α-carbons of the HIV-1 RT dimer structure.

LtrA's En domain was modeled on T4 phage endonuclease VII (Raaijmakers et al. 2001; 1EN7.pdb). Swiss Model failed to thread the En domain directly on the T4 endonuclease VII structure because of insufficient sequence homology. Instead, the model was constructed by using the amino acid backbone coordinates of endonuclease VII, and changing the side chains to match the En domain of LtrA. Two small insertions in the En domain (Fig. 8B) were added manually as loops and then energy-minimized. Then the whole structure was energy-minimized using the GROMOS force field. The En domain was positioned on the docked target-site DNA so that the active site interacts with the scissile phosphate between bottom-strand positions +9 and +10, based on the X-ray crystal structure of colicin E7 with bound DNA substrate (Hsia et al. 2004; 1PT3.pdb). Finally, the model was energy minimized with 400 steps of steepest descent minimization using the GROMOS96 force field (Scott et al. 1999).

The target-site DNA was docked to the LtrA model based on the HIV-1 RT model structure described in Peletskaya et al. (2004; 1R0A.pdb), which adds a 5′-extension of the template strand to the template–primer DNA in the crystal structure of Huang et al. (1998). The DNA was extended farther along the same trajectory by adding straight B-form DNA created in InsightII (Accelrys Inc.), and its sequence was changed to match that of the LtrA target site.

ACKNOWLEDGMENTS

We thank Janice Pata and Whitney Yin for insightful suggestions on the manuscript; Noel Miller and Matt Stanger for technical help; Benoit Cousineau and Vicky Derbyshire for advice to the Belfort lab on the proteolysis experiments; and Wadsworth Center core facilities for mass spectrometry and DNA sequencing services. This work was supported by NIH grants GM37949 and GM37951 to A.M.L., and GM39422 and GM44844 to M.B. L.H.C. is the recipient of NRSA fellowship GM06837.

Received September 16, 2004; accepted October 27, 2004.

REFERENCES

Bebenek, K., Beard, W.A., Darden, T.A., Li, L., Prasad, R., Luxon, B.A., Gorenstein, D.G., Wilson, S.H., and Kunkel, T.A. 1997. A minor groove binding track in reverse transcriptase. *Nat. Struct. Biol.* **4:** 194–197.

Belfort, M., Derbyshire, V., Parker, M.M., Cousineau, B., and Lambowitz, A.M. 2002. Mobile introns: Pathways and proteins. In *Mobile DNA II* (eds. N.L. Craig et al.), pp. 761–783. ASM Press, Washington, DC.

Bibillo, A. and Eickbush, T.H. 2002. The reverse transcriptase of the R2 non-LTR retrotransposon: Continuous synthesis of cDNA on non-continuous RNA templates. *J. Mol. Biol.* **316:** 459–473.

Chen, B. and Lambowitz, A.M. 1997. De novo and DNA primer-mediated initiation of cDNA synthesis by the Mauriceville retroplasmid reverse transcriptase involve recognition of a 3′ CCA sequence. *J. Mol. Biol.* **271:** 311–332.

Cousineau, B., Smith, D., Lawrence-Cavanagh, S., Mueller, J.E., Yang, J., Mills, D., Manias, D., Dunny, G., Lambowitz, A.M., and Belfort, M. 1998. Retrohoming of a bacterial group II intron: Mobility via complete reverse splicing, independent of homologous DNA recombination. *Cell* **94:** 451–462.

Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. 2004. WebLogo: A sequence logo generator. *Genome Res.* **14:** 1188–1190.

Cuff, J.A. and Barton, G.J. 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* **40:** 502–511.

Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M., and Barton, G.J. 1998. JPred: A consensus secondary structure prediction server. *Bioinformatics* **14:** 892–893.

Cui, X., Matsuura, M., Wang, Q., Ma, H., and Lambowitz, A.M. 2004. A group II intron-encoded maturase functions preferentially in *cis* and requires both the reverse transcriptase and X domains to promote RNA splicing. *J. Mol. Biol.* **340:** 211–231.

Derbyshire, V., Kowalski, J.C., Dansereau, J.T., Hauer, C.R., and Belfort, M. 1997. Two-domain structure of the *td* intron-encoded endonuclease I-TevI correlates with the two-domain configuration of the homing site. *J. Mol. Biol.* **265:** 494–506.

Ding, J., Hughes, S.H., and Arnold, E. 1997. Protein–nucleic acid interactions and DNA conformation in a complex of human immunodeficiency virus type 1 reverse transcriptase with a double-stranded DNA template-primer. *Biopolymers* **44:** 125–138.

Ding, J., Das, K., Hsiou, Y., Sarafianos, S.G., Clark Jr., A.D., Jacobo-Molina, A., Tantillo, C., Hughes, S.H., and Arnold, E. 1998. Structure and functional implications of the polymerase active site region in a complex of HIV-1 RT with a double-stranded DNA template-primer and an antibody Fab fragment at 2.8 Å resolution. *J. Mol. Biol.* **284:** 1095–1111.

Eickbush, T.H. and Malik, H.S. 2002. Origins and evolution of retrotransposons. In *Mobile DNA II* (eds. N.L. Craig et al.), pp. 1111–1144. ASM Press, Washington, DC.

Georgiadis, M.M., Jessen, S.M., Ogata, C.M., Telesnitsky, A., Goff, S.P., and Hendrickson, W.A. 1995. Mechanistic implications from the structure of a catalytic fragment of Moloney murine leukemia virus reverse transcriptase. *Structure* **3:** 879–892.

Gorbalenya, A.E. 1994. Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family. *Protein Sci.* **3:** 1117–1120.

Henikoff, S. and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89:** 10915–10919.

Hsia, K.-C., Chak, K.-F., Liang, P.-H., Cheng, Y.-S., Ku, W.-Y., and Yuan, H.S. 2004. DNA binding and degradation by the HNH protein ColE7. *Structure* **12:** 205–214.

Huang, H., Chopra, R., Verdine, G.L., and Harrison, S.C. 1998. Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: Implications for drug resistance. *Science* **282:** 1669–1675.

Isel, C., Westhof, E., Massire, C., Le Grice, S.F.J., Ehresmann, B., Ehresmann, C., and Marquet, R. 1999. Structural basis for the specificity of the initiation of HIV-1 reverse transcription. *EMBO J.* **18:** 1038–1048.

Kohlstaedt, L.A., Wang, J., Friedman, J.M., Rice, P.A., and Steitz, T.A. 1992. Crystal structure at 3.5 Å resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science* **256:** 1783–1790.

Lambowitz, A.M. and Zimmerly, S. 2004. Mobile group II introns. *Annu. Rev. Genet.* **38:** 1–35.

Lambowitz, A.M., Caprara, M.G., Zimmerly, S., and Perlman, P.S. 1999. Group I and group II ribozymes as RNPs: Clues to the past and guides to the future. In *The RNA world*, 2d ed. (eds. R.F. Gesteland et al.), pp. 451–485. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Lowe, D.M., Aitken, A., Bradly, C., Darby, G.K., Larder, B.A., Powell, K.L., Purifoy, D.J.M., Tisdale, M., and Stammers, D.K. 1988. HIV-1 reverse transcriptase: Crystallization and analysis of domain structure by limited proteolysis. *Biochemistry* **27:** 8884–8889.

Malik, H.S., Burke, W.D., and Eickbush, T.H. 1999. The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.* **16:** 793–805.

Martínez-Abarca, F. and Toro, N. 2000. Group II introns in the bacterial world. *Mol. Microbiol.* **38:** 917–926.

Matsuura, M., Saldanha, R., Ma, H., Wank, H., Yang, J., Mohr, G., Cavanagh, S., Dunny, G.M., Belfort, M., and Lambowitz, A.M. 1997. A bacterial group II intron encoding reverse transcriptase, maturase, and DNA endonuclease activities: Biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes & Dev.* **11:** 2910–2924.

McClure, M.A. 1991. Evolution of retroposons by acquisition or deletion of retrovirus-like genes. *Mol. Biol. Evol.* **8:** 835–856.

Mohr, G., Perlman, P.S., and Lambowitz, A.M. 1993. Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acids Res.* **21:** 4991–4997.

Mohr, G., Smith, D., Belfort, M., and Lambowitz, A.M. 2000. Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. *Genes & Dev.* **14:** 559–573.

Moran, J.V. and Gilbert, N. 2002. Mammalian LINE-1 retrotransposons and related elements. In *Mobile DNA II* (eds. N.L. Craig et al.), pp. 836–869. ASM Press, Washington, DC.

Moran, J.V., Mecklenburg, K.L., Sass, P., Belcher, S.M., Mahnke, D., Lewin, A., and Perlman, P. 1994. Splicing defective mutants of the *COXI* gene of yeast mitochondrial DNA: Initial definition of the maturase domain of the group II intron AI2. *Nucleic Acids Res.* **22:** 2057–2064.

Morris, M.C., Berducou, C., Mery, J., Heitz, F., and Divita, G. 1999. The thumb domain of the p51 subunit is essential for activation of HIV reverse transcriptase. *Biochemistry* **38:** 15097–15103.

Peletskaya, E.N., Kogon, A.A., Tuske, S., Arnold, E., and Hughes, S.H. 2004. Nonnucleoside inhibitor binding affects the interactions of the fingers subdomain of human immunodeficiency virus type 1 reverse transcriptase with DNA. *J. Virol.* **78:** 3387–3397.

Raaijmakers, H., Törö, I., Birkenbihl, R., Kemper, B., and Suck, D. 2001. Conformational flexibility in T4 endonuclease VII revealed by crystallography: Implications for substrate binding and cleavage. *J. Mol. Biol.* **308:** 311–323.

Rambo, R.P. and Doudna, J.A. 2004. Assembly of an active group II intron–maturase complex by protein dimerization. *Biochemistry* **43:** 6486–6497.

Ren, J., Esnouf, R., Garman, E., Somers, D., Ross, C., Kirby, I., Keeling, J., Darby, G., Jones, Y., Stuart, D., et al. 1995. High resolution structures of HIV-1 RT from four RT inhibitor complexes. *Nat. Struct. Biol.* **2:** 293–302.

Saldanha, R., Chen, B., Wank, H., Matsuura, M., Edwards, J., and Lambowitz, A.M. 1999. RNA and protein catalysis in group II intron splicing and mobility reactions using purified components. *Biochemistry* **38:** 9069–9083.

San Filippo, J. and Lambowitz, A.M. 2002. Characterization of the C-terminal DNA-binding/DNA endonuclease region of a group II intron-encoded protein. *J. Mol. Biol.* **324:** 933–951.

Schneider, T.D. and Stephens, R.M. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* **18:** 6097–6100.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M.C. 2003. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.* **31:** 3381–3385.

Scott, W.R.P., Hünenberger, P.H., Tironi, I.G., Mark, A.E., Billeter, S.R., Fennen, J., Torda, A.E., Huber, T., Krüger, P., and van Gunsteren, W.F. 1999. The GROMOS biomolecular simulation program package. *J. Phys. Chem. A* **103:** 3596–3607.

Shub, D.A., Goodrich-Blair, H., and Eddy, S.R. 1994. Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns. *Trends Biochem. Sci.* **19:** 402–404.

Singh, N.N. and Lambowitz, A.M. 2001. Interaction of a group II intron ribonucleoprotein endonuclease with its DNA target site investigated by DNA footprinting and modification interference. *J. Mol. Biol.* **309:** 361–386.

Tahirov, T.H., Temiakov, D., Anikin, M., Parlan, V., McAllister, W.T., Vassylyev, D.G., and Yokoyama, S. 2002. Structure of a T7 RNA polymerase elongation complex at 2.9 Å resolution. *Nature* **420:** 43–50.

Toor, N., Hausner, G., and Zimmerly, S. 2001. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA* **7:** 1142–1152.

Veerassamy, S., Smith, A., and Tillier, E.R.M. 2003. A transition probability model for amino acid substitutions from blocks. *J. Comp. Biol.* **10:** 997–1010.

Wang, J., Smerdon, S.J., Jäger, J., Kohlstaedt, L.A., Rice, P.A., Friedman, J.M., and Steitz, T.A. 1994. Structural basis of asymmetry in the human immunodeficiency virus type 1 reverse transcriptase heterodimer. *Proc. Natl. Acad. Sci.* **91:** 7242–7246.

Wilkins, M.R., Gasteiger, E., Bairoch, A., Sanchez, J.C., Williams, K.L., Appel, R.D., and Hochstrasser, D.F. 1999. Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.* **112:** 531–552.

Xiong, Y. and Eickbush, T.H. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J.* **9:** 3353–3362.

Yao, R., Nimec, Z., Ryan, T.J., and Galivan, J. 1996. Identification, cloning, and sequencing of a cDNA coding for rat γ-glutamyl hydrolase. *J. Biol. Chem.* **271:** 8525–8528.

Yin, Y.W. and Steitz, T.A. 2002. Structural basis for the transition from initiation to elongation transcription in T7 RNA polymerase. *Science* **298:** 1387–1395.

———. 2004. The structural mechanism of translocation and helicase activity in T7 RNA polymerase. *Cell* **116:** 393–404.

Zhu, K., Dobard, C., and Chow, S.A. 2004. Requirement for integrase during reverse transcription of human immunodeficiency virus type 1 and the effect of cysteine mutations of integrase on its interactions with reverse transcriptase. *J. Virol.* **78:** 5045–5055.

Zimmerly, S., Hausner, G., and Wu, X.-C. 2001. Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Res.* **29:** 1238–1250.