Crystal Structure of the Ribonuclease H Domain of HIV-1 Reverse Transcriptase

JAY F. DAVIES, II, ZUZANA HOSTOMSKA, ZDENEK HOSTOMSKY, STEVEN R. JORDAN, DAVID A. MATTHEWS*

The crystal structure of the ribonuclease (RNase) H domain of HIV-1 reverse transcriptase (RT) has been determined at a resolution of 2.4 Å and refined to a crystallographic *R* factor of 0.20. The protein folds into a five-stranded mixed beta sheet flanked by an asymmetric distribution of four alpha helices. Two divalent metal cations bind in the active site surrounded by a cluster of four conserved acidic amino acid residues. The overall structure is similar in most respects to the RNase H from *Escherichia coli*. Structural features characteristic of the retroviral protein suggest how it may interface with the DNA polymerase domain of p66 in the mature RT heterodimer. These features also offer insights into why the isolated RNase H domain is catalytically inactive but when combined in vitro with the isolated p51 domain of RT RNase H activity can be reconstituted. Surprisingly, the peptide bond cleaved by HIV-1 protease near the polymerase–RNase H junction of p66 is completely inaccessible to solvent in the structure reported here. This suggests that the homodimeric p66-p66 precursor of mature RT is asymmetric with one of the two RNase H domains at least partially unfolded.

AN ESSENTIAL STEP IN THE LIFE CYCLE OF THE HUMAN immunodeficiency virus (HIV), as in other retroviruses, is the reverse transcription of viral genomic RNA. All reactions of this process are catalyzed by a multifunctional viral enzyme reverse transcriptase (RT)(*1*). The conversion of the single-stranded RNA genome into the double-stranded DNA of the provirus requires coordination of the following RT activities: (i) RNA-dependent DNA polymerase, (ii) ribonuclease (RNase) H, and (iii) DNA-dependent DNA polymerase (*2*). Owing to its ability to selectively cleave phosphodiester bonds in the RNA moiety of the RNA-DNA heteroduplex intermediate (*3*), RNase H activity is indispensable at several stages of this complex process. It degrades RNA template during synthesis of the plus strand DNA, and specifically removes both primers by an endonucleolytic mechanism (*4*). Because of its crucial role in the life cycle of retroviruses, RT is a prime target for antiretroviral therapy, especially in connection with HIV infections and AIDS (*5*).

RNase H activity not associated with reverse transcription can be found in various prokaryotic and eukaryotic organisms. The physiological role of RNase H has to some extent been characterized in the bacterium *Escherichia coli*. In contrast to retroviruses, the bacterial enzyme appears nonessential, as mutations in the RNase H gene are not generally lethal (*6*). Analysis of the amino acid sequences suggests homology between *E. coli* RNase H and the carboxyl-terminal portion of RT from HIV and other retroviruses (*7*). Indeed, the predicted localization of the RNase H activity in a COOH-terminal portion of RT was confirmed by Moloney murine leukemia virus (MoMuLV), when it was demonstrated that this COOH-terminal segment of RT expressed separately retained RNase H activity (*8*).

Unlike MoMuLV RT, which is monomeric, HIV-1 RT is a heterodimer composed of two subunits, p66 and p51, which have identical amino termini (*9*) (Fig. 1). The heterodimer is presumably the result of asymmetric processing of p66-p66 homodimer by HIV-1 protease. The COOH-terminal domain of the p66 subunit of HIV-1 RT was expressed separately but, in contrast to MoMuLV, it is not sufficient for RNase H activity. However, the RNase H activity can be reconstituted in vitro by combining this domain with the isolated DNA polymerase domain, p51 (*10*). These results indicate that domains of HIV-1 RT, although structurally distinct, are functionally interdependent, a conclusion also supported by mutagenesis studies (*11*).

In an effort to better understand the structural reasons for the functional organization of the RT-associated RNase H, as well as to establish a structural basis for the design of specific inhibitors directed against this activity, we have determined the 2.4 Å crystal structure of the COOH-terminal domain of HIV-1 RT. Comparison of this structure with the recently determined three-dimensional

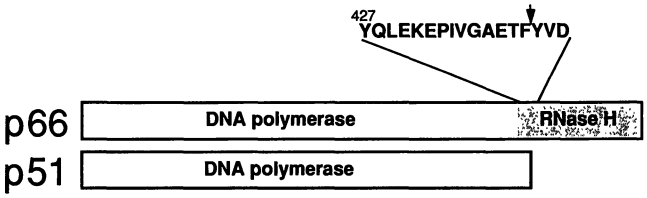


Fig. 1. Schematic representation of the subunits in heterodimeric HIV-1 RT. Relative positions of the DNA polymerase and RNase H domains are indicated. The shaded area highlights the COOH-terminal portion of the p66 subunit, whose three-dimensional structure is reported in this communication. This region, starting from Tyr^{427}, forms a stable domain. It contains the Phe^{440}–Tyr^{441} site that is cleaved by HIV-1 protease (indicated by arrow) to generate the COOH-terminus of p51 during maturation of the RT heterodimer.

The authors are on the staff of Agouron Pharmaceuticals, Inc., 11025 North Torrey Pines Road, La Jolla, CA 92037.

*To whom correspondence should be addressed.

structure of *E. coli* RNase H (*12, 13*) confirms that the COOH-terminal portion of HIV-1 RT indeed represents an RNase H domain, although this domain requires additional factors for RNase H activity.

Structure determination. Purification and crystallization of the RNase H domain of HIV-1 RT have been described (*14*). Briefly, the COOH-terminal portion of the RT gene starting from codon Tyr^{427} and ending at codon Leu^{560} was fused to the *E. coli* dihydrofolate reductase gene by a synthetic linker and overexpressed in *E. coli*. The RNase H portion of this fusion protein forms a tightly folded domain that can be released by the proteolytic action of several enzymes that cleave in the interdomain linker region. For crystallization, the isolated fusion protein was processed in vitro with HIV-1 protease, and the released RNase H domain was purified to homogeneity. In addition to comprising residues 427 to 560 of RT, the protein contains at the NH_2-terminus four additional amino acid residues, Tyr-Ala-Ser-Arg, unrelated to the native RT sequence.

Crystals were grown at 4°C in hanging drops equilibrated against a reservoir solution containing 0.15 M sodium potassium tartrate, 20 percent PEG8000, and 0.1 M sodium citrate, pH 5.2. The starting drops were composed of equal volumes of stock protein solution (protein at 10 mg/ml, 25 mM potassium phosphate, pH 7.0) and reservoir solution. Crystals grew as trigonal prisms belonging to space group $P3_1$ with $a = b = 51.9$ Å and $c = 114.9$ Å and two molecules per asymmetric unit. A typical crystal measures 0.7 by 0.3 by 0.3 mm and diffracts x-rays to a minimum Bragg spacing of about 2.2 Å.

An initial electron density map was calculated at 2.8 Å resolution with phases derived from diffraction data obtained from a crystal soaked for 9 days in a solution containing the crystallization buffer and 9 mM $K_3UO_2F_5$ (Table 1). The map showed the presence of two crystallographically independent molecules in the asymmetric unit, each one characterized by well-defined density suggesting a high percentage of β-sheet and α-helical secondary structure. The map was further improved by the intensity modification approach of Wang (*15*) (Fig. 2).

A model for each independent molecule was built into the solvent flattened map with the use of a computer graphics terminal and the program FRODO (*16*). This process was expedited by the availabil-

Table 1. Diffraction data were collected at 4°C on a Rigaku AFC-6 diffractometer operating at 9 kilowatts and equipped with a graphite monochromator and dual area detectors of the Xuong-Hamlin design (San Diego Multiwire Systems). For the uranium derivative, Bijvoet mates were collected by the inverse beam method such that intensity measurements on an *hkl* reflection at (ω, χ, φ) and the corresponding *hkl* at (ω, −χ, 180° + φ) were separated in time by approximately 6 hours. For phasing and initial modeling, a 2.8 Å native data set was used (84 percent complete, $R_{sym} = 4.4$ percent). For phasing, initial heavy atom positions were obtained from difference Patterson maps and confirmed in anomalous difference Patterson maps. The space group was determined to be $P3_1$ rather than $P3_2$ based on the heavy atom refinement statistics calculated in these enantiomorphic space groups. Parameter and SIRAS phase refinement for two sites (relative occupancies 0.35 and 0.44) were performed with the program HEAVY (*34*). Anisotropic temperature factors were included in the final cycles. Solvent flattening (five cycles, 50 percent solvent content) was performed in accord with the approach of Wang (*15*). The crystallographic *R* factor between observed structure factor magnitudes and those calculated from the solvent flattened map was 0.27 (20 to 2.8 Å data). The average phase shift between SIRAS phases and those calculated from the solvent flattened map was 25.0°. For refinement, X-PLOR (*18*) was used; the initial model was subjected to simulated annealing and atomic positional refinement (overall B factor), followed by iterative refinement of atomic *B* factors and positions, and manual rebuilding. At convergence of this process, refinement was concluded with PROLSQ (*19*). The geometry of this model is typified by an rms deviation of bond lengths from their dictionary values of 0.022 Å and an rms deviation from planarity of 0.018 Å for atoms restrained to lie in a plane.

Data sets	Resolution shells (Å)						
	Overall	20–5.1	5.1–4.1	4.1–3.2	3.2–2.8	2.8–2.6	2.6–2.4
Native							
Reflections measured	32084	5049	6211	9548	4307	3610	3359
Unique reflections	13047	1381	1375	2687	2569	2514	2521
Completeness (%)	93.1	94.5	96.0	96.1	92.5	91.0	90.4
Average I/σ	26.9	71.3	45.8	18.2	6.0	2.6	1.6
R_{sym} (%)*	4.5	3.0	3.7	6.1	9.6	14.1	19.8
$K_3UO_2F_5$							
Reflections measured	59109	12874	13429	21448	11358		
Completeness (%)	99.0	100.0	100.0	99.6	97.5		
R_{sym} (%)*	7.9	5.6	6.7	10.9	24.9		
Figure of merit	0.61	0.74	0.70	0.61	0.49		
$<F_H>$†	49.1	70.3	52.5	40.8	32.7		
$<F_H''>$‡	8.1	10.8	8.7	7.0	5.9		
$<E>$§	11.6	13.7	15.9	8.7	6.1		
$<E''>$‖	8.6	6.1	4.7	8.1	13.2		
$MnCl_2$							
Reflections measured	47364	7574	8786	19397	6377	5230	
Completeness (%)	95.5	98.8	97.3	96.3	94.3	91.0	
R_{sym} (%)*	4.9	3.2	3.7	6.9	13.0	25.8	

Refinement	Resolution shells (Å)						
	Overall	20–5.1	5.1–4.1	4.1–3.2	3.2–2.8	2.8–2.6	2.6–2.4
PROLSQ							
R (%)¶	20.0	22.5	15.5	18.3	22.8	22.9	23.2

*$R_{sym} = \sum_{h}\sum_{i=1}^{N} |\bar{I}(h) - I(h)_i| / \sum_{h}\sum_{i=1}^{N} I(h)_i$ *where* $I(h)_i$ is the *i*th measurement of reflection *h* and $\bar{I}(h)$ is the mean value of the *N* equivalent reflections. †$<F_H>$ is the mean structure amplitude of heavy atoms on an arbitrary scale. ‡$<F_H''>$ is the mean anomalous dispersion structure amplitude of heavy atoms. §$<E>$ is the rms closure error. ‖$<E''>$ is the rms difference between observed and calculated anomalous dispersion contributions. ¶$R = \Sigma|F_{obs} - F_{calc}|/\Sigma F_{obs}$. Only data with $F_{obs}/\sigma(F_{obs}) > 2.0$ were used in the refinement.

ity of refined 1.7 Å coordinates for the *E. coli* RNase H structure (*17*). Since the structure of these two RNase H molecules is quite similar, stretches of secondary structure could often be fit to the electron density by rigid body movements. After further adjustments of main-chain and side-chain torsion angles, the nearly complete model was refined at 2.4 Å resolution with the use of alternating rounds of XPLOR (*18*) simulated annealing refinement and manual rebuilding. During this process omit maps and difference maps were used to adjust atom positions and add residues that could not be reliably modeled in the initial maps. Further refinement with the program PROLSQ (*19*) improved the geometry of the model and corrected for x-ray scattering owing to bulk solvent (*20*).

The final model has an *R* factor of 0.20 and consists of 1803 protein atoms with individual isotropic temperature factors. This model does not include the four NH_2-terminal linker residues, four COOH-terminal residues (six in the second molecule), and residues 538 to 542, which connect β5 to αE (residues 538 to 541 in the second molecule). Also absent from this model are side chains for ten surface residues having weak or ill-defined density in both molecules and ten additional surface side chains with weak density in one or the other molecule. Solvent molecules have not yet been included. A least-squares superposition (*21*) of alpha-carbon (Cα) coordinates for the two crystallographically independent molecules indicates only minor conformational differences between them. The root-mean-square (rms) deviation for 117 Cα positions is 0.31 Å. The largest conformational differences occur for residues at the COOH-terminus, for residues near the unmodeled loop connecting β5 and αE, and for residues 434 to 437 which are involved in a crystal packing interaction.

Overall polypeptide folding. The RNase H domain of HIV-1 RT is an α-β protein folded into a five-stranded mixed β sheet flanked by an asymmetric distribution of four α helices (Fig. 3). Three helices (αA, αB, and αD) cluster together proximate to one face of the central sheet where they cover the adjacent parallel beta strands β1, β4, and β5. The COOH-terminal helix (αE) is situated on the opposite side of the sheet running approximately diagonally across strands β1, β2, and β3. Although each of the five strands is seven amino acids in length or longer, there is a marked splaying apart of the two innermost strands as they propagate toward one edge of the sheet. One consequence of this geometrical distortion of the β sheet is that the number of residues actually involved in hydrogen bonding between these two interior strands is limited to five and three for β1 and β4, respectively. This twisting apart of strands at one edge of the sheet provides a structural framework for the placement and positioning of individual amino acids known to be crucial for enzyme catalysis.

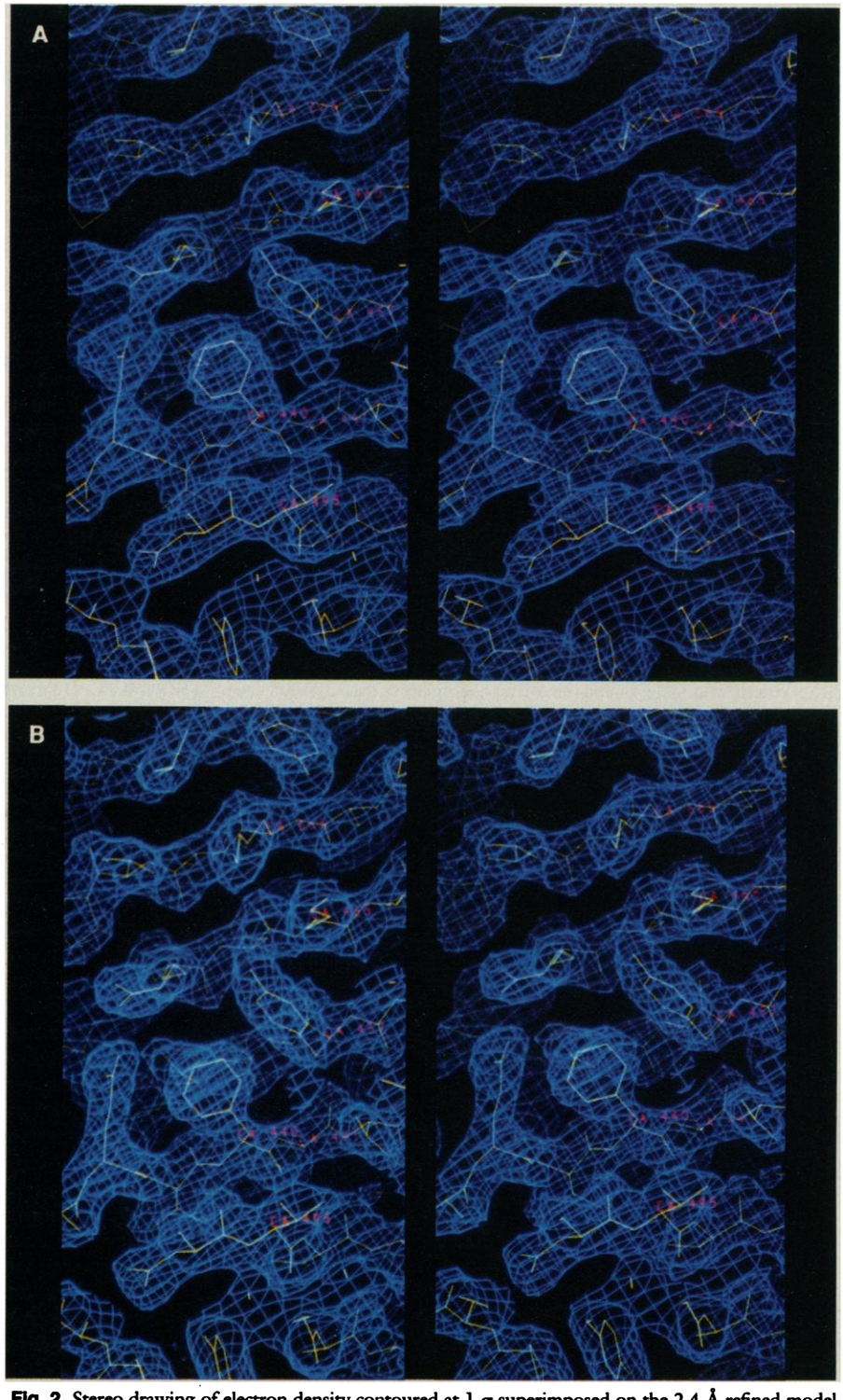


Fig. 2. Stereo drawing of electron density contoured at 1 σ superimposed on the 2.4 Å refined model of the HIV-1 RNase H domain. (**A**) The initial map calculated with observed structure factors (20 to 2.8 Å data) and phases modified by Wang's procedure (*15*). (**B**) Refined ($2F_o - F_c$) map (20 to 2.4 Å data). The view is approximately parallel to the noncrystallographic twofold rotation axis relating the two molecules in the asymmetric unit. Evident is the solvent inaccessible Phe^{440}–Tyr^{441} peptide bond.

The overall folding of the HIV-1 RNase H domain from RT bears a striking resemblance to the folding for *E. coli* RNase H (*12*,

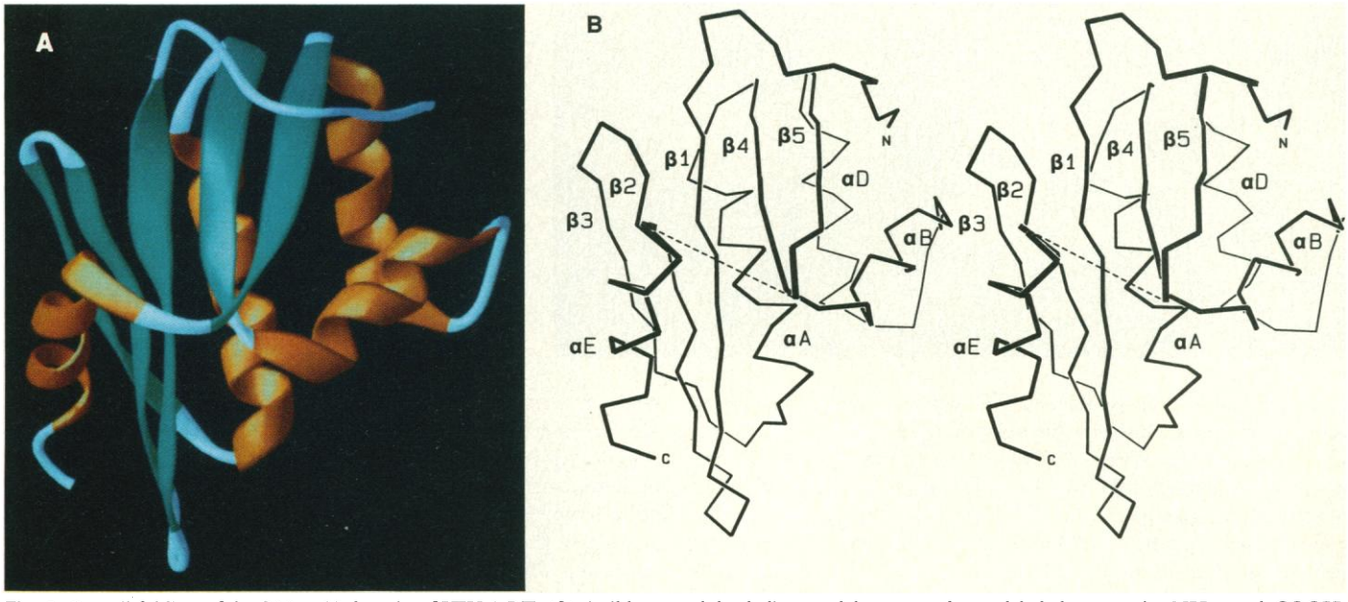


Fig. 3. Overall folding of the RNase H domain of HIV-1 RT. (**A**) A ribbon representation generated with the program RIBBON, version 2.0 (*33*). (**B**) A stereo drawing of the alpha carbon backbone in a similar orientation. The alpha helices and beta strands are labeled as are the NH_2- and COOH-termini. The dashed line represents residues 538 to 542 that are disordered in this structure.

13) even though only 32 amino acids are identical when the two sequences are aligned according to geometrical equivalence (Fig. 4). There are however, structural differences between the two proteins that in some instances are attributable to the unusual relation between retroviral RNase H and its associated DNA polymerase domain. A detailed comparison of the retroviral and bacterial RNase H structures is presented (see below) as a means of documenting features common to both and of identifying particular aspects of the HIV protein that may relate to its role as part of the intact RT heterodimer.

Structure of the RNase H domain of HIV-1 RT and comparison with *E. coli* RNase H. Polypeptide backbone chains for the RNase H domain of HIV-1 RT and for the *E. coli* enzyme were superimposed to establish the geometrical correspondence between residues in the two molecules necessary to align the amino acid sequences. Based on least-squares fitting of selected Cα positions, a high geometrical similarity (rms deviation, 0.98 Å) between these two structures occurs for a 55-residue core comprising three adjacent parallel β strands (β1, β4, and β5) and the flanking helix cluster (αA, αB, and αD). When this superposition is expanded to include residues from all elements of secondary structure (89 Cα coordinates), the rms deviation is 1.23 Å (Fig. 5). The resulting sequence alignment is shown in Fig. 4.

The HIV-1 RNase H domain structure begins with an extended coil for which there is no counterpart in the *E. coli* enzyme. While the four NH_2-terminal residues in this construct are disordered, the remaining 11 residues preceding β1 adopt a stable conformation. The first ordered residue, Tyr^{427}, is buried in a hydrophobic pocket formed by portions of β5, αB, and αD (see Fig. 3). Residues 427 through 433 loop over strands β4 and β5. Residues 434 to 437 form a type II tight turn stabilized by hydrogen bonds between the side-chain carboxamide of Asn^{494} and the backbone carbonyl oxygens of Ile^{434} and Ala^{437}. In all 14 retroviral RT sequences, Asn^{494} is conserved.

Although the interior segments of strand β1, β2, and β3 are similar in both HIV-1 and *E. coli* RNase H, the connections between them differ in the two structures. The NH_2-terminal β strand is ten amino acids in length in both enzymes, but in the HIV-1 structure the β1-β2 connection is formed by four residues in α-helical conformation. In the HIV-1 enzyme, these residues are on

Fig. 4. Geometrical alignment of secondary structure elements in the RNase H domain of HIV-1 RT and in *E. coli* RNase H. In several instances our secondary structure assignments for *E. coli* RNase H differ slightly from those reported by Yang *et al.* (*13*). The criteria we used to assign α helix and β strand secondary structure are main-chain hydrogen bonding patterns. The same criteria were applied uniformly to both structures to avoid introducing artifactual differences resulting from the use of different scoring schemes. For ease of cross-reference, our nomenclature for β strands and α helices conforms to that introduced for the *E. coli* enzyme (*13*). Every fifth residue is marked with an asterisk (*) or the appropriate residue number. Numbering of HIV-1 RT is used for the RNase H domain. Identical residues in the geometrically aligned structures are boxed. Because there is no geometrical similarity in the polypeptide chain between αB and αD in these two proteins, five residues beginning at Pro^{510} in HIV-1 RNase H are arbitrarily centered over the region which serves a similar connecting purpose in the *E. coli* enzyme.

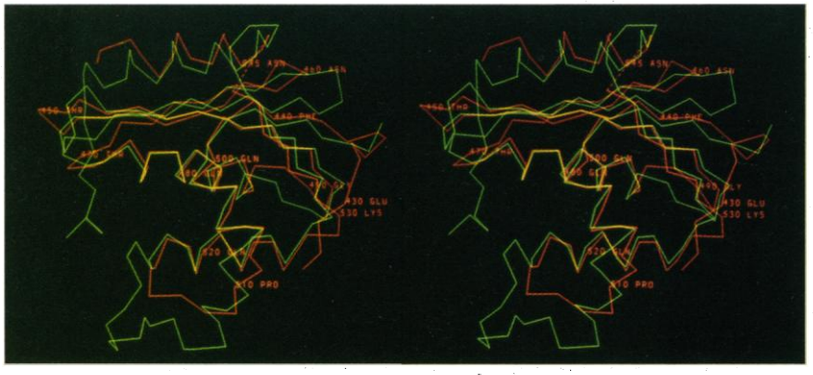


Fig. 5. Superposition of *E. coli* RNase H, in green, on the HIV-1 RNase H domain, in red. The superposition is calculated by minimizing the distance between corresponding Cα positions for 55 residues in the central core. The red dashed line indicates the disordered loop in the HIV-1 structure. Every tenth residue of the HIV-1 RNase H domain is labeled.

the same side of the β sheet as αE, whereas the corresponding residues in the *E. coli* enzyme form a type VIb β turn on the opposite side of the sheet. The next two β strands (β2 and β3) are each two residues shorter than the corresponding strands in *E. coli* RNase H, in part the result of a three-residue deletion in the connection joining β2 and β3 in the HIV-1 enzyme. These residues (459 to 462) form a type I β turn in the HIV-1 enzyme that is stabilized by a hydrogen bond between the side-chain hydroxyl of Thr^{459} and the backbone amide nitrogens of Lys^{461} and Arg^{463}. A twisting of the COOH-terminal end of β2 in the HIV-1 enzyme, relative to its position in the *E. coli* structure, is the result of amino acid sequence differences. In *E. coli* RNase H, Ala^{24} in β2 is packed against Val^{54} in αA. The structurally equivalent residues in the retroviral protein, Tyr^{457} and Leu^{484}, respectively, are much bulkier.

The crossover connection from β3 to β4 is made by αA, a five-turn helix (as in *E. coli* RNase H) that runs antiparallel to both beta strands. The NH_2-terminal residue (Thr^{473}) provides a backbone carbonyl oxygen for hydrogen bonding within the helix and accepts a hydrogen bond, through its side-chain hydroxyl, from the backbone amide nitrogen of Lys^{476}. The importance of these interactions in stabilizing the exposed NH_2-terminus of αA is emphasized by the observation that this residue is either a serine or threonine in 14 retroviral RNase H sequences (*22*). The COOH-terminus of αA is joined to β4 by a three-residue segment that contains a type I tight turn. The side-chain hydroxyl of Ser^{489} stabilizes this turn by donating bifurcated hydrogen bonds to the backbone carbonyl oxygens of residues 485 and 486 and accepting a hydrogen bond from the side-chain nitrogen of Lys^{528}. In the *E. coli* RNase H structure, the αA-β4 connection contains two more residues than does the HIV-1 enzyme, none of which are in tight turn conformation.

Although the NH_2-terminus of αB is in a similar position in both HIV-1 and *E. coli* RNase H, their COOH-termini and the residues immediately following this helix differ significantly. Helix αB begins with a conserved triad of amino acids in both structures (Ser-Gln-Tyr), but the COOH-terminus of this helix is two residues longer in the retroviral enzyme. In the *E. coli* enzyme, the connection between αB and αD is made by an 8-residue helix (αC) followed by a 12-residue loop. In HIV-1 RNase H, this domain is absent. The connection between αB and αD is instead made by a five-residue loop in extended conformation. The position of the helix αD differs slightly in the two enzymes. This difference is due in part to the insertion of a single amino acid between αD and β5 in the retroviral RNase H. Strand β5 is similar in both structures.

Residues connecting β5 to αE are disordered in the HIV-1 protein but well defined in the *E. coli* enzyme. In *E. coli* RNase H, this loop positions a conserved histidine residue adjacent to a cluster of carboxylate side chains implicated in catalysis (*12, 13*). These catalytic site residues lie at the edge of a concave surface thought to be the RNA-DNA heteroduplex binding cleft. Helix αE follows this loop and is four residues shorter in the retroviral RNase H domain. A hydrogen bond between the side-chain hydroxyl of Ser^{553} and the backbone carbonyl oxygen of invariant Asp^{549} four residues further back in the helix causes this shortening of αE. Residues 553 to 556 are in type II tight turn conformation while electron density for the remaining four COOH-terminal residues is too weak to model accurately. This disorder is not unexpected in that these residues represent the NH_2-terminal portion of the RT-integrase cleavage site (*23*) that must be accessible to HIV-1 protease during processing of the gag-pol polyprotein. In *E. coli* RNase H, there are 13 additional residues following αE in extended conformation.

Metal binding at the catalytic site. Divalent metal ions are essential for RNase H activity (*3, 24*). Metal binding to the RNase H domain of HIV-1 RT was studied with x-ray data from a crystal that had been soaked in 45 mM $MnCl_2$. A difference map calculated with diffraction data from this crystal and data from the metal-free protein crystal show two tightly bound Mn^{2+} ions in close proximity to four acidic residues Asp^{443}, Glu^{478}, Asp^{498} and Asp^{549} (Fig. 6). These four residues are among seven amino acids conserved in all known bacterial and retroviral RNase H sequences (*22*). Site-directed mutagenesis studies of *E. coli* RNase H indicate that three of these carboxyls are crucial for enzyme catalysis (*25*). While the conserved acidic residues may serve more than one function in RNase H, our structural results suggest they have an important role in forming these two metal binding sites.

Our finding that Mn^{2+} binds at two sites in the RNase H domain of HIV-1 RT can be contrasted with a report that crystals of *E. coli* RNase H have a single binding site for various divalent cations (*12*). This apparent discrepancy is probably the result of a crystal packing effect in which the active site carboxyl cluster of *E. coli* RNase H interacts with a lysine side chain from a symmetry related molecule in the crystallographic unit cell, thus preventing binding of a second metal ion (*12, 13*).

Yang *et al.* (*13*) have suggested that RNase H may have a catalytic mechanism similar to that for the 3′,5′-exonuclease of DNA polymerase I in which a metal activated hydroxyl ion attacks a pentacoordinate phosphorus intermediate (*26*). The resulting cleavage is accompanied by an inversion of configuration at the phosphorus (*27*). Both enzymes hydrolyze the P-O3′ bond of polynucleotides in a reaction requiring divalent metals bound to an active site having four carboxyl groups. In the absence of substrate or product, the exonuclease domain has one bound divalent metal ion (site A), whereas complexes with deoxynucleoside monophosphates show a second divalent metal (site B) 4 Å from site A (*26*). The two Mn^{2+} ions bound to the RNase H domain are also approximately 4 Å apart, and in both proteins the geometrical arrangement of the corresponding metal-carboxyl clusters is similar. Two carboxyls interact exclusively with the metal at site A and one with the metal at site B while the fourth carboxyl coordinates both metal ions.

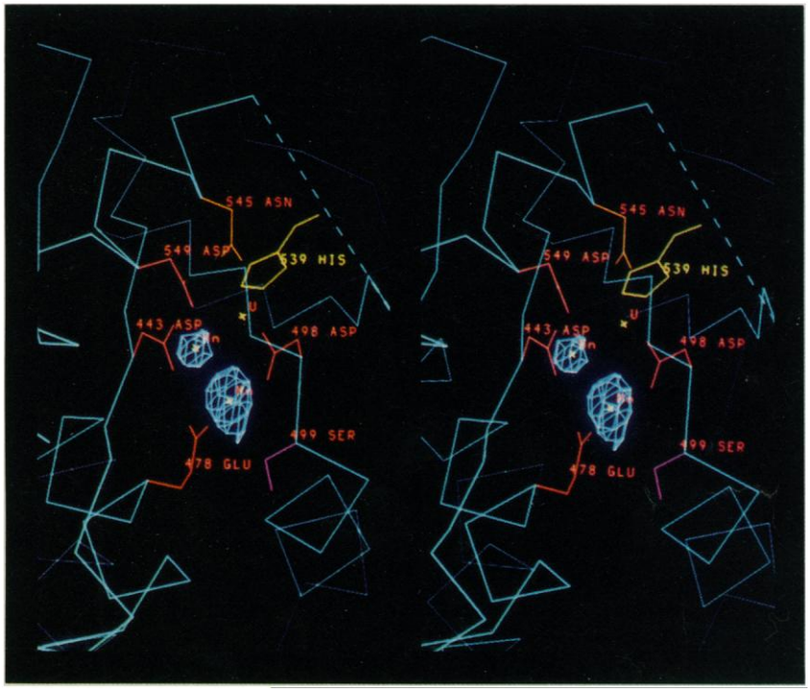
In the complex between the exonuclease of DNA polymerase I and dTMP (5′ deoxythymidylic acid) both metals are coordinated to one oxygen of the 5′ phosphate of the nucleotide (*26*). There are no nucleotides in our RNase H structure; however, the binding site of the $UO_2F_5^{3-}$ anion, used for isomorphous replacement phasing, is only 3 Å from both Mn^{2+} ion binding sites. Moreover, the uranium anion binding site has the same orientation relative to the Mn^{2+} ions as the phosphate of dTMP does to the metal sites in the 3′,5′-exonuclease complex. By analogy, the $UO_2F_5^{3-}$ may be bound at the scissile phosphate binding site of RNase H. These observations are consistent with the proposal that RNase H catalyzed hydrolysis of the RNA of heteroduplex substrates occurs by a mechanism similar to that proposed for the 3′,5′-exonuclease domain of DNA polymerase I.

Implications for positioning the covalently linked RNase H and DNA polymerase domains. The structure reported on here suggests where the covalently linked polymerase domain may be positioned relative to the RNase H domain in RT. The NH_2-terminus occupies a position that in the *E. coli* RNase H structure corresponds to helix αC, which is absent in the retroviral RNase H domain. We noted above that a significant difference between the retroviral and bacterial RNase H proteins is the absence in the former of helix αC and an extended loop that together connect αB to αD in the *E. coli* enzyme. The corresponding connection in the RNase H domain of HIV-1 is accomplished by five amino acids in extended conformation. It appears that the 13-residue deletion in the retroviral RNase H sequence allows portions of the polymerase domain to interact with the RNase H domain in a manner that could not occur if the retroviral enzyme had tertiary structure corresponding to helix αC in the bacterial enzyme. Stated differently, we can now suggest with considerable confidence that some portion of HIV-1 RT polymerase domain immediately preceding the RNase H domain in linear amino acid sequence is in contact with the flat right side of the nuclease domain (as seen in Fig. 3) in direct contact with β5, αB, αD, and the two short loops connecting these structural elements.

Analysis of the solvent accessible surface of crystalline HIV-1 RNase H lends support to this theory. The most extensive packing interaction between individual protein molecules in the crystal lattice occurs between this flat face (β5, αB, αD, and connecting loops) and the corresponding face of a second molecule in the crystallographic asymmetric unit. A survey of subunit interfaces in high resolution crystal structures of 23 oligomeric proteins indicates that at least 600 $Å^2$ of surface area per subunit must be buried to overcome translational and rotational freedom lost on association (*28*). The accessible surface area covered as a result of the observed crystal contact in our structure is 610 $Å^2$ per molecule. If this surface were buried to a similar extent in RT, it would be sufficient to account for the observation that the isolated RNase H and polymerase domains can assemble in vitro to reconstitute RNase H activity (*10*).

In the *E. coli* RNase H structure an extensive concavity implicated in binding the RNA-DNA hybrid substrate contains a prominent cluster of six tryptophan residues some of which are partially exposed (*12*). These tryptophan residues are present in a 40–amino acid stretch of polypeptide chain encompassing residues 81 to 120. Only Trp^{120} has a structural counterpart in the RNase H domain of HIV-1 RT. However, there are six tryptophans in HIV-RT within a short stretch of amino acids (residues 398 to 426) just NH_2-terminal to the RNase H domain. According to the model presented above, these tryptophans would be proximate to the right-hand edge of the retroviral RNase H domain (Fig. 3), in which case they may serve a functional role in hybrid binding.

The protease cleavage site. To generate the mature p66-p51 heterodimer of HIV-1 RT, proteolysis must occur in only one of two p66 subunits of the homodimer. In the structure presented, the HIV-1 protease cleavage site between Phe^{440} and Tyr^{441} (*29*) is inaccessible, consistent with the remarkable stability of this domain in the presence of proteolytic enzymes, including HIV-1 protease (*10, 14*). This can be appreciated by noting that Phe^{440} and Tyr^{441} are the third and fourth residues in β1, the long central strand of the beta sheet (Fig. 2). Since the homodimer and heterodimer have similar RNase H activities (*30*), at least one nuclease domain must be correctly folded. Taken together, these observations imply that the p66 homodimer is asymmetric, with one functionally folded RNase H domain and the other domain unfolded to an extent that



	534											545
HIV-1	A	W	V	P	A	H	K	G	I	G	G	N
HIV-2	A	W	V	P	A	H	K	G	I	G	G	N
SIV_{mac}	A	W	V	P	A	H	K	G	I	G	G	N
EIAV	A	W	V	P	G	H	K	G	I	Y	G	N
VISNA	H	W	V	P	G	H	K	G	I	P	P	N

Fig. 6 (left). Mn^{2+} binding sites with respect to the seven invariant residues in retroviral and bacterial RNases H. The conserved residues (Asp^{443}, Glu^{478}, Asp^{498}, Ser^{499}, His^{539}, Asn^{545}, and Asp^{549}) cluster near the catalytic site. The loop containing His^{539} is disordered in the structure of the HIV-1 RNase H domain. The histidine (yellow) is positioned by analogy with its location in the *E. coli* RNase structure (*17*). Positions of the other side chains are from refined coordinates of the native structure. The $UO_2F_5^{3-}$ position (U) is from the heavy atom refinement process. The Mn^{2+} cations were fit to difference electron density maps calculated with the coefficients $(|F_{Mn}| - |F_{native}|)\alpha_{calc}$. The calculated phases, α_{calc}, are from the refined native model. The map is contoured at 9 σ.

Fig. 7 (top). Comparison of the 534 to 545 region in the HIV-1 RNase H domain with the corresponding region in other lentiviruses. HIV-2, human immunodeficiency virus type 2; SIV_{mac}, simian immunodeficiency virus from macaque monkey; EIAV, equine infectious anemia virus; VISNA, visna lentivirus.

allows protease cleavage.

Although one RNase H domain could be completely unfolded, thereby allowing protease access to an otherwise buried cleavage site, partial unfolding seems a more likely alternative. Some of us have proposed a model for asymmetric processing of HIV-1 RT in which the RNase H domain plays a structural role in stabilizing the dimer via interactions with the polymerase domains (*14*). According to this model the p66 homodimer is asymmetric. Favorable interaction between the two polymerase domains and one folded RNase H domain induces partial unfolding of the other. Protease cleavage of the partially unfolded RNase H domain then stabilizes the resulting heterodimer. A possible advantage of an asymmetric heterodimer may be an increased processivity of RT. This asymmetry would permit DNA synthesis at one end of the RT complex and simultaneous hydrolysis of the reverse-transcribed template in a single RNase H catalytic site at the opposite end.

Does the three-dimensional structure of the HIV-1 RNase H domain described here offer insight into the nature of the putative partially unfolded RNase H domain? At least three points are relevant to the following discussion. (i) Beta strands 1 and 4 separate as they move together toward one edge of the central β sheet creating a prominent cleft that forms part of the catalytic site. The RNase H domain can therefore be considered as two subdomains separated by the β1-β4 split. The left subdomain consists of β1, β2, β3, αE, and connecting loops, while the right subdomain comprises β4, β5, αA, αB, αD, and connecting loops (Fig. 3). (ii) At the COOH-terminus of the RNase H domain, helix αE lies on one side of the β sheet forming an interface with the three NH_2-terminal β strands β1, β2, and β3. Except for Asn^{545} this helix does not interact with any other part of the folded RNase H domain. (iii) Evidence summarized above suggests that in p66 the polymerase domain of HIV-1 RT interacts with the RNase H domain in a contact area consisting of at least β5, αB, αD, and the two loops connecting these elements of secondary structure. Taken together, the above observations allow us to envision a possible structure of the partially unfolded RNase H domain in which these two subdomains are separated from one another along the β1-β4 cleft. The right subdomain is stabilized by interactions with the polymerase domain while the left subdomain is exposed.

A consequence of this unusual processing strategy is the proteolytic release of the COOH-terminal fragment of 120 residues. It seems probable that this fragment can be further degraded by various proteases since as an isolated polypeptide chain it almost certainly cannot assume a native-like fold, having lost the first three residues needed to form the middle β strand in the central β sheet. This may explain unsuccessful attempts by several groups to detect the COOH-terminal fragment released during processing of RT (*31*).

Inactivity of the isolated HIV-1 RNase H domain. The question arises why the isolated RNase H domain reported on here has no detectable activity, yet when combined with p51 in vitro results in reconstitution of RNase H activity on a defined substrate. We now know, from the structural similarity between *E. coli* RNase H and the RNase H domain of HIV-1 RT studied here, that the overall folding of the isolated retroviral protein is substantially correct. Isolated *E. coli* RNase H is enzymatically active and therefore able to bind RNA-DNA hybrid substrate and catalyze its cleavage. Since the catalytic site of the isolated RNase H domain of HIV-1 RT appears to be correctly assembled judged by its geometrical similarity to the corresponding site in *E. coli* RNase H and its ability to bind divalent metal ions, the retroviral domain may be deficient in substrate binding. The lack of productive substrate binding could be correlated with the absence in the retroviral RNase H domain of the lysine and tryptophan rich region incorporating helix αC that was suggested above as a possible structural determinant in binding the RNA-DNA hybrid. Perhaps the DNA polymerase domain of HIV-1 RT is required for productive binding and alignment of the hybrid RNA-DNA substrate. An attractive candidate for this proposed function is the tryptophan rich sequence of 40 to 50 residues immediately upstream of the polymerase-nuclease junction.

Another possibility is that the inactivity of the RNase H domain is a consequence of structural alterations that occur on dissociation from the polymerase domain. There is, in fact, one particular portion of our structure that is highly suggestive in this regard. We noted earlier that for neither of the two crystallographically independent RNase H domains is there interpretable electron density in the final 2.4 Å difference maps for residues 538 to 542 that connect β5 to αE. The corresponding polypeptide chain in *E. coli* RNase H forms a well-defined protruding loop that has been implicated in substrate binding (*13*). The loop contains His^{539}, one of seven invariant residues (*22*). Mutations at the histidine position in *E. coli* RNase H reduced both substrate binding affinity and catalytic rate (*25*). For HIV-1, mutation of His^{539} in RT leads to greatly decreased RNase H and DNA polymerase activities (*30, 32*). The importance of this loop for retroviral activity is further underscored by data presented in Fig. 7. Residues 534 to 545 in HIV-1 are the most highly conserved stretch of 12 amino acids in five lentivirus RNase H sequences. For instance, HIV-1 and HIV-2 have identical primary sequences in these two regions while the next most highly conserved 12 amino acid stretch has three substitutions. The disorder of this loop in our structure provides a possible explanation for the inactivity of the isolated RNase H domain and suggests that its correct positioning may be mediated by an interaction with one or both of the polymerase domains of the p66-p51 heterodimer.

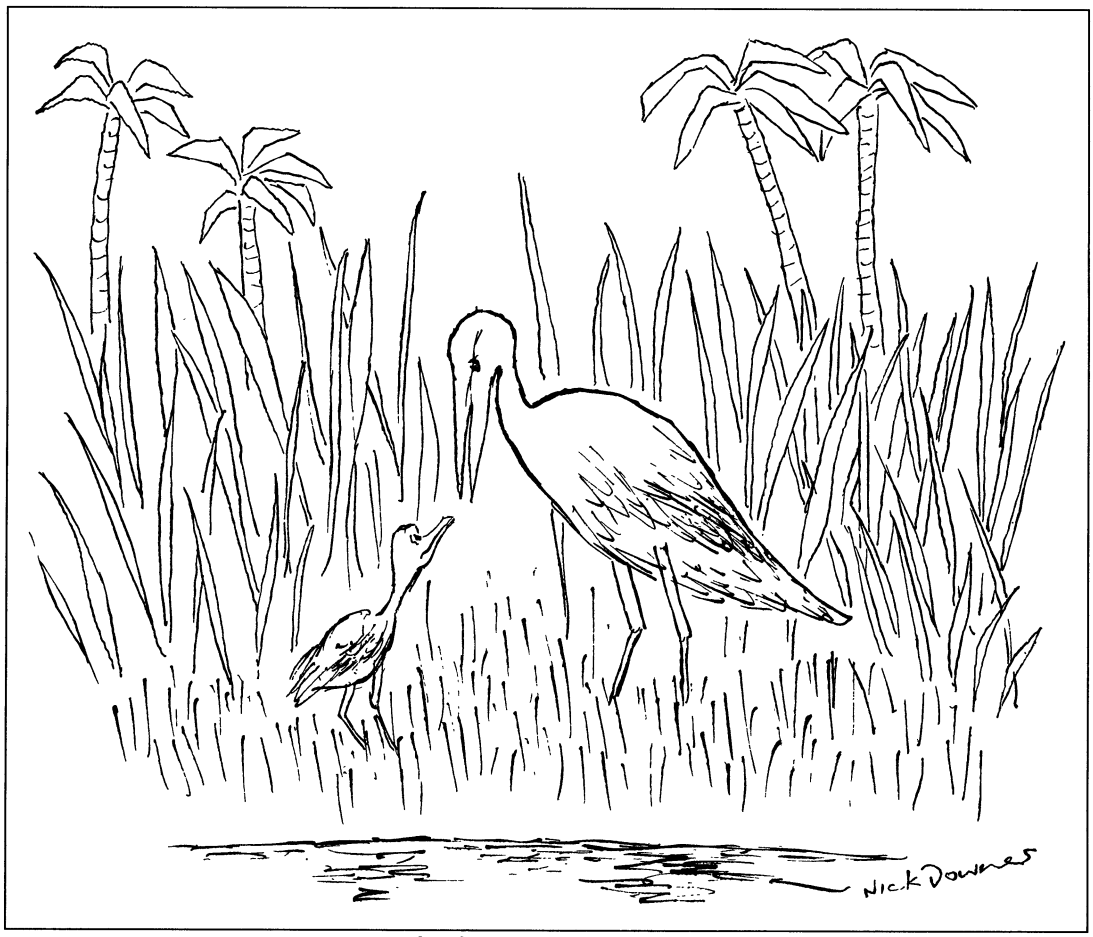
Either or both of these possibilities could be related to the inability of the isolated HIV-1 RNase H domain to catalyze hydrolysis of an RNA-DNA heteroduplex. It remains to be seen whether mutations in the loop connecting β5 to αE in the retroviral domain that mimic the primary amino acid sequence in the corresponding loop in the *E. coli* enzyme can alone or in concert with other modifications (such as in the loop connecting αB to αD) confer nuclease activity to the isolated RNase H domain of HIV-1 RT.

REFERENCES AND NOTES

1. S. Goff, *J. AIDS* **3**, 817 (1990).
2. E. Gilboa, S. W. Mitra, S. Goff, D. Baltimore, *Cell* **18**, 91 (1979).
3. R. J. Crouch and M.-L. Dirksen, in *Nucleases*, S. M. Linn and R. J. Roberts, Eds. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1982), p. 211; R. J. Crouch, *New Biol.* **2**, 771 (1990).
4. C. A. Omer and A. J. Faras, *Cell* **30**, 797 (1982); A. J. Rattray and Champoux, *J. Virol.* **61**, 2843 (1987); A. T. Panganiban and D. Fiore *Science* **241**, 1064 (1988); W.-S. Hu and H. M. Temin, *ibid.* **250**, 1227 (1990).
5. H. Mitsuya, R. Yarchoan, S. Broder, *Science* **249**, 1533 (1990).
6. T. Horiuchi, H. Maki, M. Sekiguchi, *Mol. Gen. Genet.* **195**, 17 (1984); T. A. Torrey, T. Atlug, T. Kogoma, *ibid.* **196**, 350 (1984).
7. M. S. Johnson, M. A. McClure, D.-F. Feng, J. Gray, R. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 7648 (1986).
8. N. Tanase and S. P. Goff, *ibid.* **85**, 1777 (1988).
9. F. di Marzo Veronese *et al.*, *Science* **231**, 1289 (1986); M. Lightfoote *et al.*, *J. Virol.* **60**, 771 (1986).
10. Z. Hostomsky, Z. Hostomska, G. O. Hudson, E. W. Moomaw, B. R. Nodes, *Proc. Natl. Acad. Sci. U.S.A.* **88**, 1148 (1991).
11. V. R. Prasad and S. P. Goff, *ibid.* **86**, 3104 (1989); A. Hizi, S. H. Hughes, M. Shaharabany, *Virology* **175**, 575 (1990).
12. K. Katayanagi *et al.*, *Nature* **347**, 306 (1990).
13. W. Yang, W. A. Hendrickson, R. J. Crouch, Y. Satow, *Science* **249**, 1398 (1990).
14. Z. Hostomska, D. A. Matthews, J. F. Davies, II, B. R. Nodes, Z. Hostomsky, in preparation.
15. B.-C. Wang, *Methods Enzymol.* **115**, 90 (1985).
16. T. A. Jones, *J. Appl. Cryst.* **11**, 268 (1978).
17. W. A. Hendrickson, personal communication.
18. A. T. Brünger, J. Kuriyan, M. Karplus, *Science* **235**, 458 (1987).

19. W. A. Hendrickson, *Methods Enzymol* **115**, 252 (1985).
20. J. T. Bolin, D. J. Filman, D. A. Matthews, R. C. Hamlin, J. Kraut, *J. Biol. Chem.* **257**, 13650 (1982).
21. M. G. Rossmann and P. Argos, *ibid.* **250**, 7525 (1975).
22. R. F. Doolittle, D.-F. Feng, M. S. Johnson, M. A. McClure, *Q. Rev. Biol.* **64**, 1 (1989).
23. L. E. Henderson *et al.*, *J. Virol.* **62**, 2587 (1988).
24. M. C. Starnes and Y. Cheng, *J. Biol. Chem.* **264**, 7073 (1989).
25. S. Kanaya *et al.*, *ibid.* **265**, 4615 (1990).
26. L. S. Beese and T. A. Steitz, *EMBO J.* **10**, 25 (1991).
27. A. P. Gupta and S. J. Benkovic, *Biochemistry* **23**, 5874 (1984).
28. J. Janin, S. Miller, C. Chothia, *J. Mol. Biol.* **204**, 155 (1988).
29. S. F. J. Le Grice, R. Ette, J. Mills, J. Mous, *J. Biol. Chem.* **264**, 14902 (1989); V. Mizrahi, G. M. Lazarus, L. M. Miles, C. A. Meyers, C. Debouck, *Arch. Biochem. Biophys.* **273**, 347 (1989); M. Graves *et al.*, *Biochem. Biophys. Res. Commun.* **168**, 30 (1990).
30. O. Schatz *et al.*, *FEBS. Lett.* **257**, 311 (1989).
31. D. M. Lowe *et al.*, *Biochemistry* **27**, 8884 (1988); A. L. Ferris *et al.*, *Virology* **175**, 456 (1990).
32. M. Tisdale, T. Schulze, B. A. Larder, K. Moelling, *J. Cell. Biochem.*, (Suppl.) **14D**, 179 (1990).
33. M. Carson, *J. Mol. Graphics* **5**, 103 (1987).
34. T. C. Terwilliger and D. Eisenberg, *Acta Crystallogr.* **A 39**, 813 (1983).
35. We thank B. Nodes for technical assistance, W. Yang and W. Hendrickson for providing the refined 1.7 Å coordinates of *E. coli* RNase H, and C. Janson for helping in crystallizations. Coordinates for the model described have been submitted to the Protein Data Bank. Supported in part by NIH grant GM 39599.

7 February 1991; accepted 7 March 1991



"The obstetrician delivered you"