The 'helix clamp' in HIV-1 reverse transcriptase: a new nucleic acid binding motif common in nucleic acid polymerases

Thomas Hermann, Thomas Meier, Matthias Götte and Hermann Heumann*
Max-Planck-Institut für Biochemie, D-82152 Martinsried, Germany

Received August 1, 1994; Revised and Accepted October 5, 1994

ABSTRACT

Amino acid sequences homologous to ^{259}*KLVGKL (X)$_{16}$ KLLR*^{284} of human immunodeficiency virus type 1 reverse transcriptase (HIV-1 RT) are conserved in several nucleotide polymerizing enzymes. This amino acid motif has been identified in the crystal structure model as an element of the enzyme's nucleic acid binding apparatus. It is part of the helix–turn–helix structure, αH–turn–αI, within the 'thumb' region of HIV-1 RT. The motif grasps the complexed nucleic acid at one side. Molecular modeling studies on HIV-1 RT in complex with a nucleic acid fragment suggest that the motif has binding function in the p66 subunit as well as in the p51 subunit, acting as a kind of 'helix clamp'. Given its wide distribution within the nucleic acid polymerases, the helix clamp motif is assumed to be a structure of general significance for nucleic acid binding.

INTRODUCTION

There have been several attempts in the past to develop a unifying concept for template-dependent nucleic acid polymerases. Most of these investigations were aimed at finding similarities or homologies in structure and function of polymerases. Sequence comparisons were performed in order to detect primary structure homologies. With the availability of crystal structure data of polymerases, such as Klenow fragment of *E.coli* DNA polymerase I (KF), bacteriophage T7 RNA polymerase (T7 RNAP) and HIV-1 RT, three-dimensional structure comparisons also became possible (1–5).

Relying on available information from both methods, amino acid sequence homology searching including secondary structure prediction as well as crystal structure analysis, we intended to derive benefit from combining the strengths of those methods.

Numerous groups have performed sequence alignments of RNA- and DNA-dependent DNA and RNA polymerases in order to detect common conserved motifs (6–9). These studies have shown that a conserved motif exists among all RNA-dependent DNA and RNA polymerases, comprising two aspartic acid residues surrounded by a set of unpolar amino acids (10–12). This 'Asp–Asp' motif was assumed to reside within the polymerization active site. However, due to lack of structural data, these sequence-based studies were only of limited value in attributing functional significance to conserved sequence regions. Primary structure comparisons confined to smaller subgroups of closely related nucleic acid polymerases are more fruitful, as shown for example by analysis of retroviral RNA-dependent DNA polymerases in which six major homology regions, termed A, B, C, D, E and F were identified (12). Site-directed mutagenesis in HIV-1 RT (reviewed in 13) proved that the Asp–Asp motif, located in region E (12), forms a portion of the RT's polymerization active site, in line with previous suggestions. This conclusion was confirmed by crystal structure analysis of Kohlstaedt *et al.* (4).

Recently, availability of three dimensional structural data from crystals of KF, T7 RNAP, and HIV-1 RT has provided a new basis for comparison between polymerases. Three-dimensional structure comparisons have revealed that the so-called 'finger–palm–thumb' elements common to all three enzymes (4,5,2,14) form a groove which is thought to accommodate the nucleic acid substrate. However, since there is only slight similarity between the primary structures of these three polymerases, extensive structure-sequence correlation with the aim of defining motifs was difficult.

In our search for more homologies possibly existing within the group of nucleotide polymerizing enzymes, we tried a different approach. Instead of comparing large primary structure segments, we focused our motif search on small substructure domains to which a function could be attributed. Possible candidates for such motifs were segments which are assumed to play a role for function of nucleic acid polymerases, such as nucleic acid binding. The basis for the selection of possible motifs was the crystal structure model of HIV-1 RT. This enzyme is a protein heterodimer consisting of subunits of 66kD (p66) and 51kD (p51) respectively (reviewed in 13). Although both subunits contain the same 51 kD N-terminal amino acid portion, they differ in their overall conformation, as revealed by X-ray structure analysis. Site-directed mutagenesis along with crystal structure analysis has shown that the catalytic center for nucleotide polymerization is located in the N-terminal portion of p66 (3).

*To whom correspondence should be addressed

The p51 subunit is folded differently from p66, such that this part of the sequence is buried in the interior of the protein. RT contains an additional enzymatic activity, RNase H which is located in the 15kD C-terminal part of the p66 subunit. Previous biochemical investigations and chemical probing experiments as well as crystal structure analyses showed that the distance between the active sites of polymerization and RNase H activity is approximately 18 nucleotides of A-form nucleic acid (15,16,5). Although the heterodimer of p66 and p51 is necessary for RT activity *in vitro*, the role of p51 in enzymatic function is still largely unknown.

It is known from the X-ray structure of HIV-1 RT that interactions of nucleic acid and the protein occur mostly via the sugar–phosphate backbone of the nucleic acid (5). Kohlstaedt *et al.* have already identified some structural elements of RT probably involved in nucleic acid binding, such as the 'fingers', 'palm' and 'thumb' domains within the p66 subunit (4). However, these structural domains were not found to contain any previously described DNA or RNA binding motif.

It was the aim of our work to identify motifs engaged in nucleic acid binding, which HIV-1 RT shares with other polymerases. We present evidence that a sequence known to act as a nucleic acid guiding and binding structure in the α-helix-H/turn/α-helix-I segment of the p66 thumb (5) is a conserved motif in many other nucleic acid polymerases. Molecular modeling studies on RT in complex with a nucleic acid fragment comprising 27 nucleotides reveal that the same motif located in the p51 subunit is probably also involved in nucleic acid binding in HIV-1 RT.

MATERIALS AND METHODS

Protein sequence database searching and alignment

All protein sequences were extracted from the Swiss-Prot Sequence Database. The FASTA algorithm of Pearson and Lipman (17), part of the UWGCG package (Genetics Computer Group Madison), was employed for searching the Swiss-Prot database for protein sequences containing the nucleic acid binding motif from HIV-1 RT. Different runs of FASTA were performed in order to scan motifs with a varying gap size between the two conserved sequence patterns. Gap sizes taken into consideration ranged from 5 to 24 amino acids. Single comparisons of HIV-1 RT sequences against other protein sequences were performed using COMPARE together with DOTPLOT as implemented in the UWGCG program package. Alignments of protein sequences were done using the UWGCG BESTFIT algorithm.

Secondary structure prediction and selection criteria

The secondary structure of protein segments containing the motif plus additional forty amino acids flanking both sides was predicted using two different methods: the PHD neural network system of Rost and Sander (18) and a modified Bayes statistical approach based on the methodology of Maxfield and Scheraga (19) as implemented in the SYBYL molecular modeling package (TRIPOS Associates St Louis). Amino acid sequences were checked for amphiphilic helices using a program for flexible continuous helical wheel mapping (T.Hermann, unpublished).

The selection criteria were the following: Sequences were selected whenever one or both of the two secondary structure prediction methods predicted helical structure for at least half of the residues within a segment comprising the motif plus additional eight amino acids flanking each side; at least half of the residues for which helical structure was predicted had to be located within the bipartite motif itself. A minimum of four basic residues (Arg, Lys) had to occur in the selected segment, at least one of which had to be located within each of the two conserved regions of the motif.

Molecular modeling

Model building of the HIV-1 RT and nucleic acid was done using algorithms implemented in the SYBYL molecular modeling package. The starting C_α coordinates of HIV-1 RT and coordinates for phosphate atoms of a complexed 19/18 nt DNA came from the crystal structure analysis of Jacobo-Molina *et al.* (5), deposited in the file 1HMI of the Brookhaven Protein Data Bank (PDB). A knowledge based algorithm was employed to construct the RT peptide backbone to the C_α atoms (20). The geometry of Pro residues was corrected individually. Side chains were added to amino acids Lys^{238} throughout Val^{317} in both subunits of HIV-1 RT using a procedure from SYBYL. The conformation of the added side chains was refined by energy minimization with a conjugate gradient minimizer under the AMBER force field (21). During this procedure the geometry of the peptide backbone was kept fixed. For the RNase H domain of HIV-1 RT both, backbone and side chain coordinates were taken from the crystal structure analysis of isolated RNase H domain as deposited in the PDB file 1HRH (22).

Modeling of an 18 nt double stranded DNA fragment complexed to HIV-1 RT was done by positioning single base pairs governed by phosphate atom coordinates provided in the PDB file 1HMI. The sequence of the DNA fragment was as described by Jacobo-Molina *et al.* (5). The base pairs were finally linked and the conformation of the DNA backbone was refined by energy minimization under AMBER while the geometry of the bases and the positions of the phosphate atoms were kept fixed.

A 27 nt double stranded RNA fragment in complex with HIV-1 RT was built from standard A-form base pairs using the protein model as a guide but without allowing for any interactions between nucleic acid and protein during calculations on the RNA. The RNA backbone was energy refined while constraining its geometry in the region of the polymerization active site to the corresponding portion of the DNA model built as previously described. The position of the phosphodiester linkage between the nucleotides 18 and 19 where the RNase H cuts was kept within a range of 4 Å to the active site of the RNase H domain of the HIV-1 RT.

Molecular dynamics calculation

A model of HIV-1 RT constructed as previously described was subjected to a molecular dynamics calculation in order to scan the conformational space available to the side chains of the p51 helix clamp. Therefore the geometry of the peptide backbone was constrained to the crystal structure while the modeled side chains of the p51 helix clamp were allowed to move. A dynamics calculation of 120 ps under the AMBER force field was performed at 300 K in steps of 1 fs with a coupling of $\tau = 0.1$ ps to an external constant temperature bath using an algorithm from the SYBYL package. A distant dependent dielectric constant of $\epsilon = 1.0$ was used along with a cutoff for nonbonded interactions of 12 Å. Data was collected at intervals of 200 fs and the root mean square deviation was calculated at these steps for each amino acid side chain within the p51 helix clamp neglecting hydrogen atoms. The side chain mobility of the residues was calculated as the average root mean square deviation of non-hydrogen atoms of each side chain over the whole period of 120 ps.

RESULTS

Selection of a nucleic acid binding motif in the p66 subunit of HIV-1 RT

Template dependent nucleic acid polymerases require nucleic acid binding structures that function to hold the template in place. It is expected that this protein/nucleic acid interaction should be largely sequence independent, occurring via the sugar–phosphate backbone which is uniform for all nucleotides. Because all template directed polymerases must bind nucleic acids more or less nonspecifically, we looked for related structural motifs across several classes of nucleic acid polymerases.

As a basis for our search for nucleic acid binding motifs we used information from X-ray structure analysis of HIV-1 RT. Biochemical studies (23–25) and crystal structure analysis show that RT binds nucleic acid via the 'fingers' and 'thumb' domains (4) of the p66 subunit, acting as a device to position the template/primer relative to the polymerase active site (5).

The thumb consists essentially of a bundle of four alpha helices. Two of them, namely helix αH (Asn^{255} to Ser^{268}) and helix αI (Gln^{278} to Thr^{286}) (4) are part of a helix–turn–helix segment (Val^{254} to Ala^{288}, Fig. 1) which participates in nucleic acid binding as suggested by the crystal structure model. This helix–turn–helix segment, for which we propose the term 'helix clamp' contains several Arg and Lys residues, which may participate in binding the negatively charged nucleic acid backbone. The two helices of the helix clamp contain a distribution of basic and hydrophobic amino acids characteristic of amphiphilic helices, in which 'functional' residues, such as Arg and Lys reside on those parts of the helices exposed to solvent and alternate respectively with two or three 'helix building' amino acids like Gly, Ala, Val, Leu and Ile.

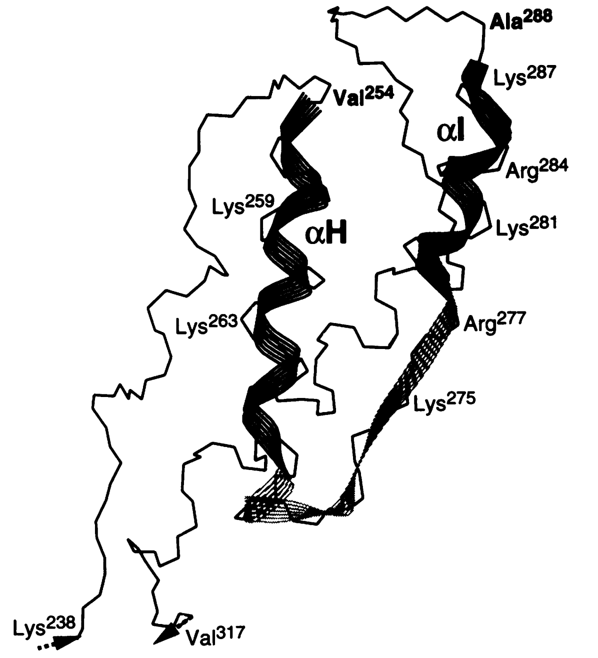


Figure 1. Peptide backbone structure of the thumb subdomain (Lys^{238}–Val^{317}) in the p66 subunit of HIV-1 RT. The helix clamp consists of the helix–coil–helix segment of Val^{254} to Ala^{288} displayed in ribbon-representation. Helices αH and αI are in black, the turn between them in grey. Residues with basic side chains within the helix clamp are labeled. The peptide backbone was modeled from published crystal structure data as described in the text.

Detection of the helix clamp in other nucleic acid polymerases

In order to analyze whether the helix clamp is a motif of general significance for nucleic acid polymerases we evaluated patterns from the known primary, secondary and tertiary structure of this motif which are suitable for pattern search among polymerases. These patterns were partial sequences of the αH–turn–αI thumb region, secondary structure elements having helical character, and the abundance of basic amino acids like Arg and Lys. These comparisons revealed that the sequence $K^{259}LVGKL\ (X)_{16}$-$KLLR^{284}$ of HIV-1 RT consisting of two stretches of six and four amino acids is a conserved sequence homologous to that found in a number of other polymerase enzymes (Fig. 2). A portion of this motif, namely $K^{259}LVGKL$, has been reported as part of region F, one of the six conserved regions (A–F) occurring in several reverse transcriptases (12).

Statistical evaluations of the amino acid distribution in the region of the helix clamp motif throughout the different nucleic acid polymerases containing this motif yielded the consensus sequence $U\ (X)_4\ BLUGBU\ (X)_{9-20}\ BUUB\ (X)_4\ U$, where X is any amino acid, U is one having a hydrophobic side chain (i.e. Ala, Ile, Leu, Val, Met) and B is one having a basic side chain (i.e. Arg or Lys).

Surprisingly the helix clamp motif is found in nucleic acid polymerases of organisms throughout the eucaryotes, bacteria and their viruses, as Table I shows. However, it was not possible to relate these motifs evolutionary, suggesting that it is the structure rather than the specific sequence that is conserved. The helix clamp motif is not limited to reverse transcriptases, but is found scattered throughout every subgrouping of polymerases, including eucaryotic and procaryotic DNA polymerases, and both RNA and DNA dependent RNA polymerases. However, nucleic acid polymerases in which the helix clamp motif is not found include the Klenow fragment of *E.coli* DNA polymerase I and bacteriophage T7 RNA polymerase, two nucleic acid polymerases with known crystal structures. These polymerases may bind and guide nucleic acid via structures other than the helix clamp.

In HIV-1 RT the two conserved segments of the motif are located within the helices αH and αI. They are separated by a region of nonconserved amino acids containing portions of the helices and a turn structure. The nonconserved sequence comprises 16 amino acids in HIV-1 RT. Its size varies in other polymerases and transcriptases from 9 to 20 amino acids. The two helices of the helix clamp motif contain clusters of amino acids having basic side chains which are arranged in such a way that they can act as a guide for the nucleic acid backbone (Fig. 1).

Secondary structure predictions were performed for each polymerase selected in the previous step. The employed sequences comprised the motif itself extended at both sides by additional forty amino acids. These analyses revealed that the two conserved parts of the bipartite helix clamp motif and the immediately flanking sequences are located in regions predicted to have alpha-helical structure. The hydrophobic and hydrophilic amino acids

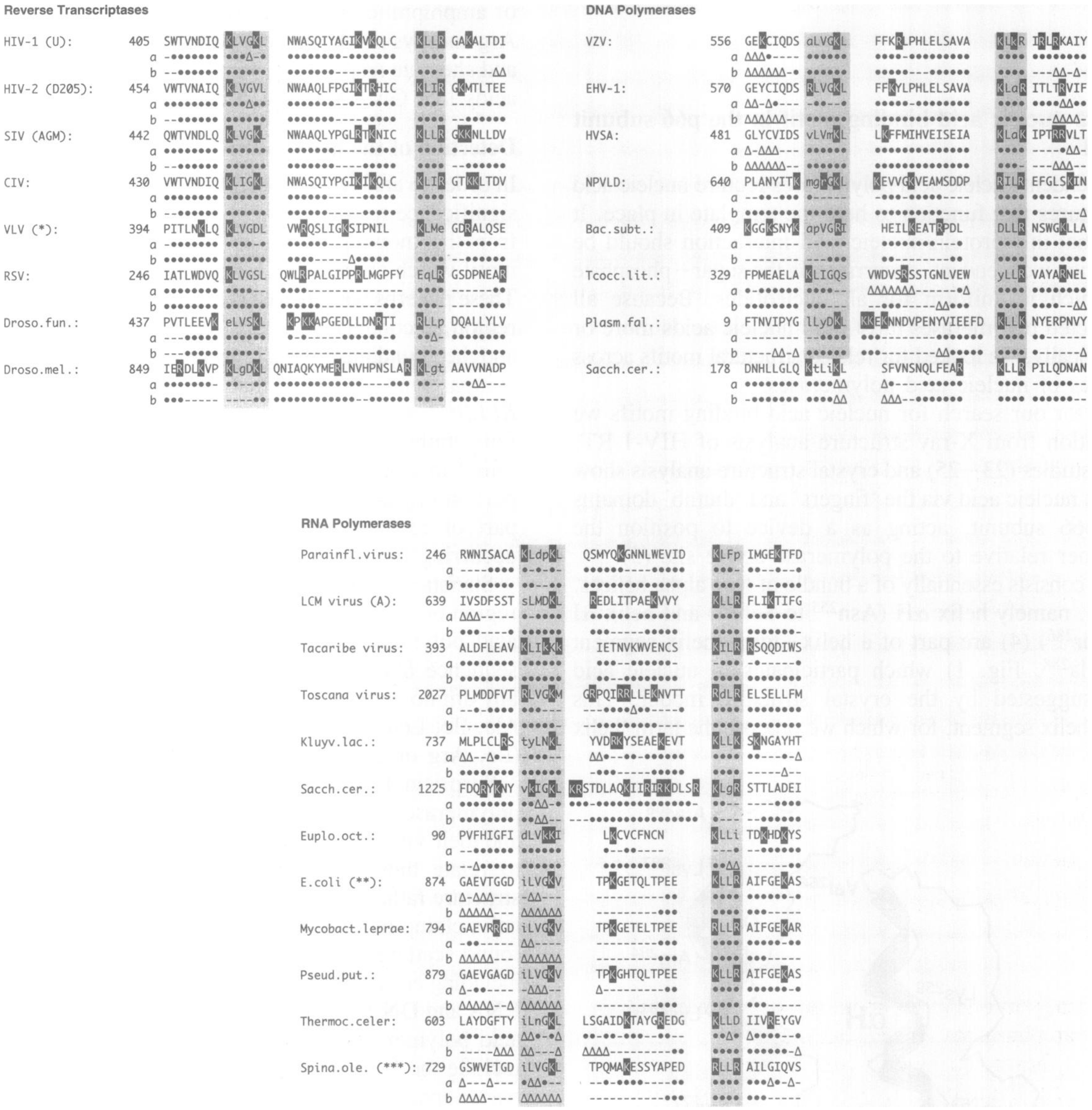
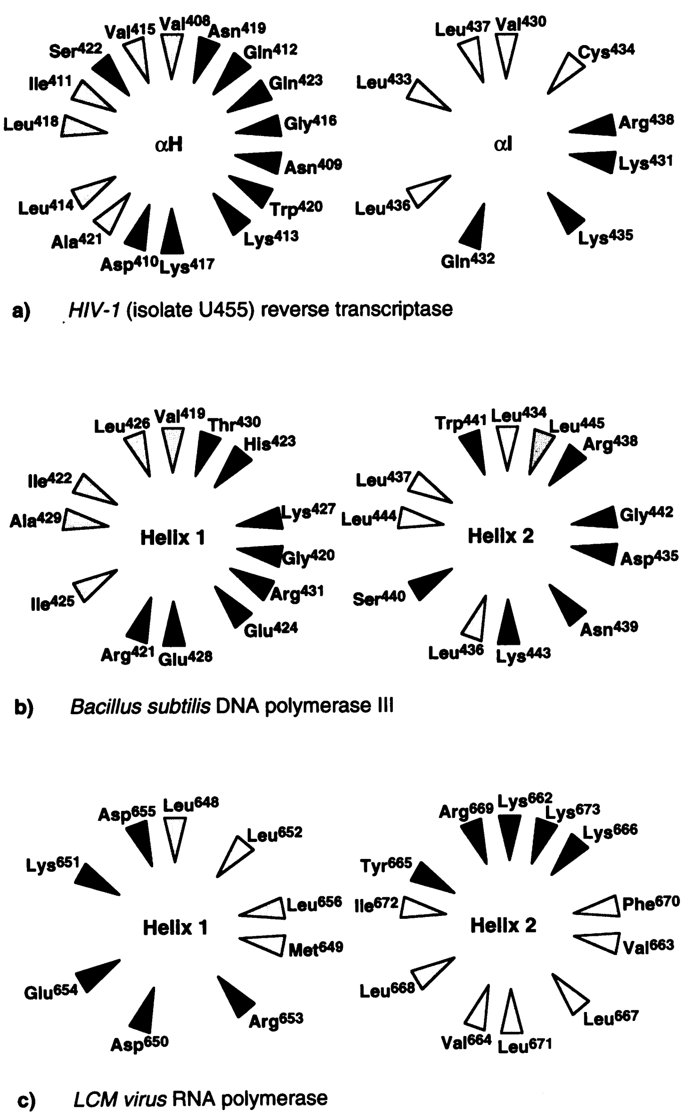


Figure 2. Compilation of nucleic acid polymerases containing the helix clamp motif (gray boxes). The sequences shown each include eight residues flanking the motif. Conserved residues with respect to the motif in HIV-1 according to the matrix of Dayhoff are in capital letters within the helix clamp motif sequence. Basic amino acids (K = Lys and R = Arg) are emphasized in dark boxes. The sequences between the two parts of the motif are not aligned but centered. Below each sequence the predicted secondary structure is shown [calculated with the methods of a) Maxfield and Scheraga (19), and b) Rost and Sander (18), details see Methods]: dots indicate alpha helices, triangles indicate beta sheets and dashes indicate coil structure. The abbreviations are: HIV-1, 2 = *Human Immunodeficiency Virus Type 1*, 2; SIV = *Simian Immunodeficiency Virus*; CIV = *Chimpanzee Immunodeficiency Virus*; VLV = *Visna Lentivirus*; RSV = *Rous Sarcoma Virus*; Droso. fun. = *Drosophila funebris*; Droso. mel. = *Drosophila melanogaster*; VZV = *Varicella Zoster Virus*; EHV-1 = *Equine Herpesvirus Type 1*; HVSA = *Herpesvirus saimiri*; NPVLD = *Lymantria Dispar Multicapsid Nuclear Polyhedrosis Virus*; Bac. subt. = *Bacillus subtilis*; Tcocc. lit. = *Thermococcus litoralis*; Plasm. fal. = *Plasmodium falciparum*; Sacch. cerv. = *Saccharomyces cerevisiae*; Parainfl. virus = *Parainfluenza virus*, LCM virus = *Lymphocytic Choriomeningitis Virus*; Kluyv. lac. = *Kluyveromyces lactis*; Euplo. oct. = *Euplotes octocarinatus*; E.coli. = *Escherichia coli*; Mycobact. leprae = *Mycobaterium leprae*; Pseud. put. = *Pseudomonas putida*; Thermoc. celer = *Thermococcus celer*, Spina. ole. = *Spinacia oleracea*. Where multiple virus strains are available from the database the listed isolate is given as abbreviation in brackets. A further specification of the listed nucleic acid polymerases is given in Table I. Interestingly, one transposable element coding for a reverse transcriptase in *Drosophila* shows up in the list. Furthermore all nucleotide polymerases from plants which do contain the motif are RNA-directed RNA polymerases of chloroplasts. The listed DNA polymerases throughout belong to the B family of DNA polymerases. Among the RNA polymerases a group exists, comprising the enzymes from *E.coli*, *Mycobacterium leprae*, *Pseudomonas putida*, *Thermococcus celer* and the plant chloroplasts, in which all criteria for selection of the polymerases are fulfilled with the exception that the first portion of the bipartite motif is predicted to be in β-sheet rather than in helical conformation. (*): almost identical in sequence and predicted secondary structure are the respective regions in *Ovine Lentivirus* (OMVV) where the helix clamp motif starts at K^{383}, and in *Caprine Arthritis-Encephalitis Virus* (CAEV) where the helix clamp motif starts at K^{407}. (**): identical in sequence and position with *Salmonella* typhimurium. (***): almost identical in sequence and predicted secondary structure are the respective regions in the RNA polymerases of chloroplasts in *Marchantia polymorpha* where the helix clamp starts at V^{732}, in *Zea mays* (I^{746}), in *Oryza sativa* (I^{746}), in *Saponaria officinalis* (I^{179}), and in *Nicotiana tabacum* (I^{737}).

Table I. Specification of the nucleic acid polymerases of Fig. 2 containing the helix clamp motif

Reverse Transcriptases:		
HIV-1, HIV-2	RT from *pol* polyprotein	
SIV	↓	
CIV	↓	
VLV, OMVV, CAEV	↓	
RSV	↓	
Droso. fun.	RT from transposable element	
Droso. mel.	↓	
DNA Polymerases (all DNA-directed):		
VZV	DNAPol of family B	
EHV-1	↓	
HVSA	↓	
NPVLD	↓	
Tcocc. lit.	↓	
Plasm. fal.	↓	(δ subunit)
Sacch. cer.	↓	(α subunit)
Bac. subt.	DNAPol III (γ subunit)	
RNA Polymerases:		
Parainfl.virus	RNA-directed RNAPol	(β subunit)
LCM virus	↓	↓
Tacaribe virus	↓	
Toscana virus	↓	
Kluyv. lac.	DNA-directed RNAPol from plasmid	
Sacch. cer.	DNA-directed RNAPol from mitochondria	
Euplo. oct.	DNA-directed RNAPol II	
E. coli.	DNA-directed RNAPol	(β subunit)
Salm. typhim.	↓	↓
Mycobact.leprae	↓	↓
Pseud. put.	↓	↓
Plant chloroplasts	↓	↓
Thermoc.celer	DNA-directed RNAPol	(A' subunit)



a) *HIV-1* (isolate U455) reverse transcriptase

b) *Bacillus subtilis* DNA polymerase III

c) *LCM virus* RNA polymerase

Figure 3. Helical wheel representation of the helices within the helix clamp regions of three selected nucleotide polymerases from *Human immunodeficiency virus type 1* (**a**), *Bacillus subtilis* (**b**), and *Lymphocytic choriomeningitis virus* (**c**). Decreasing hydrophobicity of amino acid side chains is indicated by darker shading on a three step scale (white → grey → black). Hydrophobicity was assigned to residues according to Kyte and Doolittle (32).

within the helical portions of the motif are distributed such that the helices are amphiphilic. This becomes clear when the amino acid sequence is shown in a helical wheel representation (Fig. 3).

In order to judge the importance of residues for structure and function of the helix clamp motif the sequences of closely related isolates of HIV-1 were compared. There are two amino acid positions in the region of the helix clamp which deviate in RTs of different isolates, namely the residues at positions 272 and 275. The first one, located in the turn between the two helices of the motif, can be either Ala or Pro suggesting that Pro is not required for turn formation. Arg^{275} is replaced by Lys in some HIV-1 isolates, suggesting that an amino acid having a basic side chain is required at this position. Comparison with other retroviral RTs revealed that sequences surrounding the helix clamp motif are also highly conserved. Amino acid variations are restricted to exchanges of hydrophobic residues by other unpolar amino acids and of amino acids with basic side chains by other basic residues (Lys to Arg and vice versa).

Modeling of contacts between the helix clamp of the p66 subunit and DNA

Primary and secondary structure comparisons of polymerases containing the helix clamp motif with HIV-1 RT, suggest that the helix clamp has nucleic acid binding function. Assessment of this hypothesis requires information about the arrangement of side chains of the amino acids in HIV-1 RT forming the helix clamp, with respect to the nucleotides of the nucleic acid substrate. Since this information is not available from the crystal structure, molecular modeling studies were performed in order to analyze whether or not contacts between the basic residues of the helix clamp and the nucleic acid substrate are possible. The data forming the basis for our modeling studies came from X-ray structure analysis of HIV-1 RT. Jacobo-Molina *et al.* have published the coordinates of the C_α atoms of HIV-1 RT in complex with a DNA fragment of 18/19 nucleotides (nt) (5) and also the coordinates of the phosphate atoms of the DNA fragment. Davies *et al.* have previously published the structure of the isolated RT-associated RNase H domain including amino acid side chains (22). This information was used in order to develop

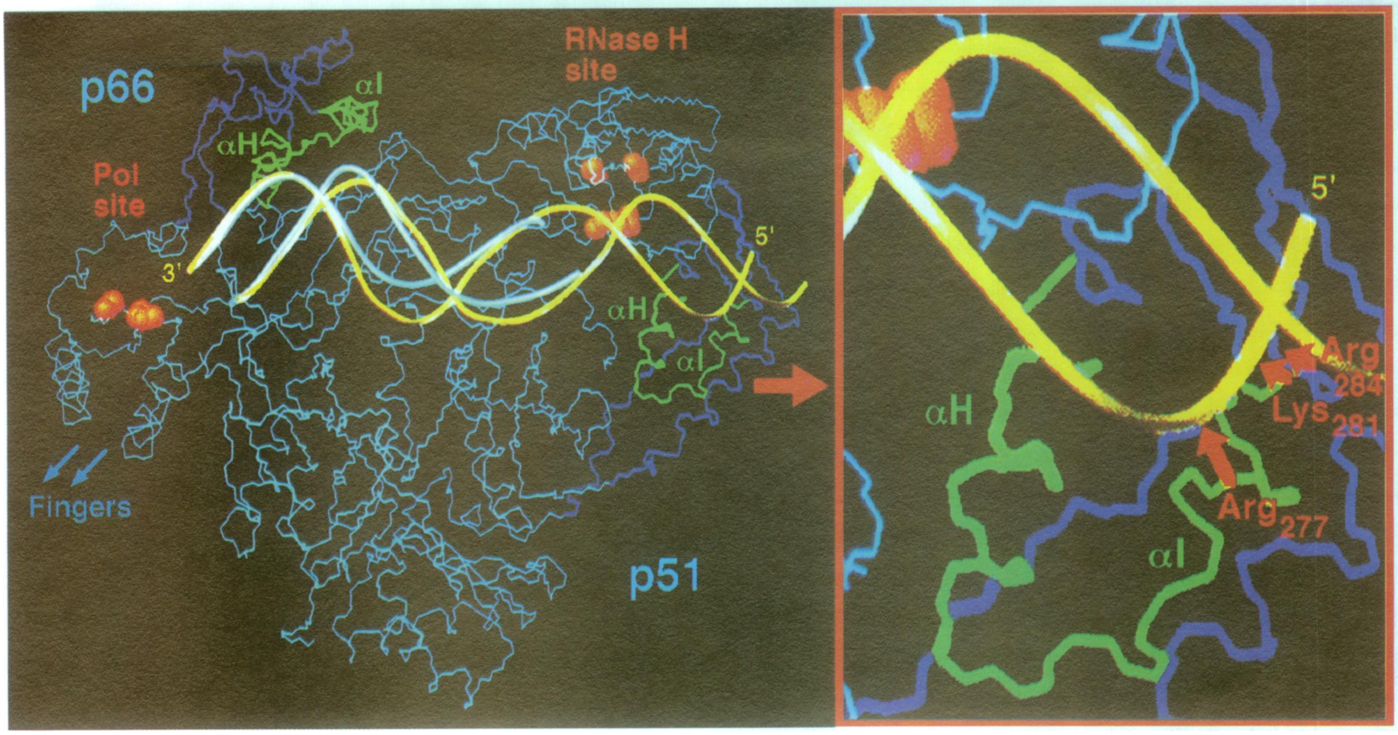


Figure 4. Model of HIV-1 RT in complex with template/primer nucleic acid summarizing the results from crystal structure data (5) and our molecular modeling studies. For the sake of clarity parts of the fingers and palm subdomains of p66 were omitted (residues Pro^1 to Asp^{86} and Asp^{113} to Pro^{157}). Residues of the active sites of polymerization and RNase H are shown as red spheres. The thumbs of both subunits are colored in dark blue and the helix clamps therein (αH−turn−αI) are emphasized in green. The 27 nt template/primer RNA from our modeling studies is shown in backbone tube representation (yellow). The termini of the primer strand are marked by 3′ and 5′, respectively. For means of comparison the 18/19 nt template/primer DNA from crystal structure analysis is superimposed (light blue tube). The right part of the figure shows a detail of possible interactions of basic side chains within the p51 helix clamp with the backbone of the modeled primer strand RNA in the region of 23 to 25 nucleotides upstream its 3′ terminus (see also Fig. 6). The peptide backbone and the template/primer were modeled from published crystal structure data as described in the text.

a more detailed structural model of the RT/DNA complex by means of molecular modeling techniques. For that purpose the peptide backbone of RT heterodimer was reconstructed employing a knowledge based algorithm (20). Construction of the side chains was confined to those amino acids participating in the formation of the helix clamp, namely Lys^{238} to Val^{317}. An all-atom model of the DNA was built using the published coordinates of the phosphorous atoms of the 18/19 nt DNA fragment complexed to RT. The base pair geometry was constrained using standard DNA parameters as described in Methods. The transition of nucleic acid conformation from A- to B-form observed in the crystal structure was taken into account by successively changing constraints on the DNA sugar backbone during model building, going from standard A- to B-form parameters. The resulting model is shown in Figure 4. This model is not intended to represent the actual conformation of the helix clamp subdomain, but rather to aid in exploration of the conformational space accessible to the amino acid side chains in this region.

Our analysis indicates that interactions between the functional amino acids (Arg and Lys) of the helix clamp and the phosphate groups of the complexed DNA are possible, as summarized in Figure 5. The side chains of two Lys in helix αH could interact with the backbone of the primer strand close to its 3′-terminus. The template strand of the DNA possibly has contacts with several Arg and Lys residues in helix αI. A maximum of six basic side chains altogether could participate in binding nucleic acid via the p66 thumb subdomain.

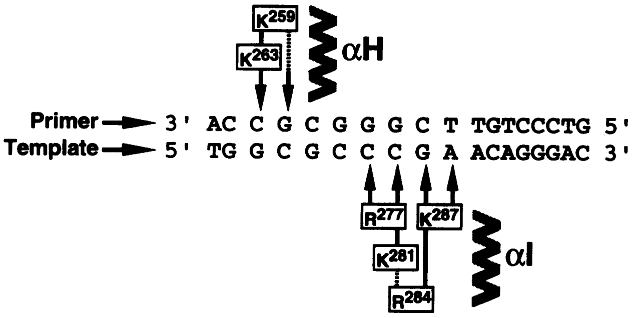


Figure 5. Possible interactions of basic residues Arg (R) or Lys (K) in the HIV-1 RT p66 helix clamp with backbone phosphates of a complexed 18 nt dsDNA derived from molecular modeling studies on a RT/DNA model based on crystal structure data (5). Broken lines indicate interactions in which the protein side chains are further away from template/primer residues, yet still within the range of possible interaction. The residues contacting the primer strand are located within helix αH while those interacting with the template reside in helix αI.

Modeling studies on possible interactions between the p51 helix clamp and a 27 nt RNA fragment

Given that both p66 and p51 contain the helix clamp motif, we constructed a model to ask whether the p51 helix clamp could

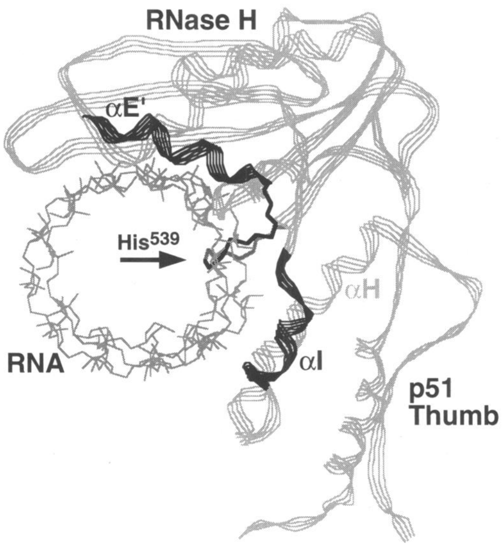


Figure 6. Structural elements of the RNase H domain and the p51 thumb contacting nucleic acid in the HIV-1 RT/27 nt primer/template model. Only the 12 terminal base pairs of RNA are displayed with the base side chains omitted. The peptide backbone of the RNase H domain in p66 and the thumb of p51 is shown in ribbon representation (grey). Alpha helices that are thought to contact bound nucleic acid are αE′ (Gly^{543} to Ile^{556}) at the C-terminus of the RNase H domain and αI in the p51 thumb (black emphasized ribbons). The N-terminus of helix αE′ is connected to the loop containing His^{539} (peptide backbone displayed in black) protruding into the major groove of the template/primer.

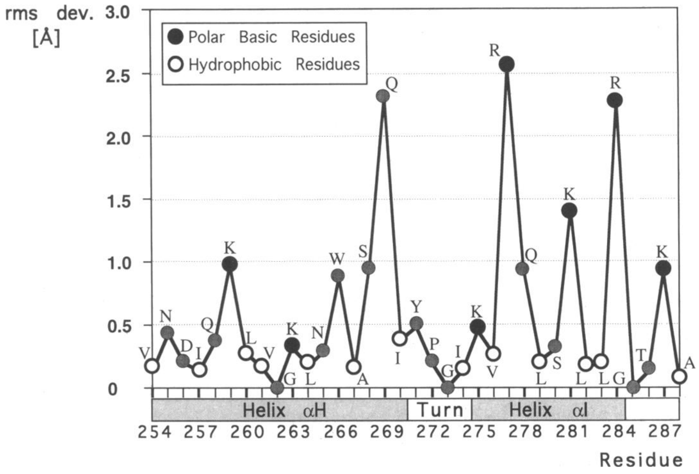


Figure 7. Plot of the relative side chain mobility of residues Val^{254}–Ala^{288} within the thumb of the HIV-1 RT p51 subunit during a 120 ps molecular dynamics calculation at 300 K. The side chain mobility is given as rms deviation (details described in Methods). The secondary structure of the sequence observed in the crystal structure is displayed below (α-helical regions in grey, coil in white). Arg and Lys residues are marked with filled circles, amino acids with hydrophobic side chains with open circles.

also be involved in template binding. Crosslinking experiments by Peliska and Benkovic suggested that contacts exist between the p51 subunit and template/primer (26). In order to test the hypothesis that these contacts may occur via residues in the p51 helix clamp, we performed molecular modeling studies similar to those described for the p66/DNA interaction. However, the structural information available from published crystallographic studies is not suitable for modeling the interaction of residues in the helix clamp in p51 with nucleic acid, since the DNA fragment of 18/19 nucleotides used in the crystal structure analysis does not extend to p51. In addition, this short DNA fragment shows a 45° bend due to transition from canonical A to B form. Simply elongating the short primer/template in the primer 5′ direction would lead to collision of nucleic acid with the protein. For a larger nucleic acid fragment bound to RT a less bent conformation is thus expected. Therefore, we believe that a correction of the DNA structure is necessary in the region of the primer 5′-terminus when considering a longer nucleic acid fragment. With these considerations in mind we built a double stranded template/primer RNA of 27 nucleotides in A-conformation which resembles the crystallographic data at the primer 3′-terminal region close to the polymerization active site. The conformation of the template/primer RNA was constrained by the requirement that the primer 3′-terminus should be located close to the polymerization active site of the RT and the center of RNase H activity should be positioned 18 nucleotides upstream, close to the phosphodiester bond of the template strand which is hydrolyzed during RNase H activity. The distance of 18 nucleotides is in accordance with results from biochemical experiments (15,16). These constraints led to bending of the template/primer towards the protein. Such bending of nucleic acid upon binding to RT was previously suggested by Kohlstaedt *et al.* (4). The template/primer spanning the space between the active sites of polymerization and RNase H activity was elongated by additional 9 base pairs in A-form. Figure 4 shows RT containing the modeled primer/template RNA of 27 nucleotides. For comparison the crystallographically determined 18/19 nt DNA is also depicted.

Analysis of our RT/RNA model shows clearly that interactions between the helix αI within the p51 helix clamp and the sugar–phosphate backbone of the primer strand RNA are possible (Figs. 4 and 6). Analogous to the interaction of nucleic acid with the p66 helix clamp, side chains of basic residues in the *KLLR*^{284} motif could contact RNA phosphate groups. Since in the case of the p51 helix clamp we are dealing with an extrapolated model, the interactions in this model must be judged more cautiously than the data describing the situation at the p66 thumb where information from crystal structure analysis guides the modeling. Highly probable candidates for specific p51/nucleic acid interactions are the phosphates of the RNA primer strand 23 to 25 nucleotides upstream the 3′ terminus and the side chains of Arg^{277}, Lys^{281} and Arg^{284} within the p51 thumb (Fig. 4). Distinct from the p66 thumb where both helices αH and αI of the helix clamp may interact with the nucleic acid, a participation of the p51 helix αH in nucleic acid binding is unlikely. This is due to the fact that the p51 thumb makes important contributions to the interface of both RT subunits and helix αH of the p51 subunit is involved in this contact. Another close contact between protein and nucleic acid observed in our model is that of the sequence at the C-terminus of the p66 subunit. The helical region of Gly^{543} to Ile^{556} (helix αE') (5) may interact with both strands of the nucleic acid (Fig. 6). Furthermore, a loop of the p66 RNase H domain containing His^{539} at its top projects into the major groove of the final 27 nt RNA model (Fig. 6). As a consequence His^{539} is likely to make contacts with bases of the nucleic acid, as previously observed by Jacobo-Molina *et al.* (5). This is supported by experimental data indicating an important role of His^{539} for RT activity (27).

In addition to our work on static models of the RT/nucleic acid complexes we carried out molecular dynamics calculations in order to obtain information about the conformational space accessible to the protein side chains. We confined these calculations to the p51 subunit without considering nucleic acid. The coordinates of the protein backbone were constrained to the crystal structure while the side chains of the amino acids were permitted to move according to Newton's equations of motion. Analysis of the amino acids in the p51 helix clamp revealed high conformational mobility of the side chains of basic residues while amino acids that are considered to function as helical structure building elements show a low side chain mobility (Fig. 7). The side chains of basic amino acids, located at helix positions on the solvent accessible surface of the helix clamp, protrude into the space where template/primer contacts RT according to our static models. A high conformational mobility of the side chains of basic amino acids would thus facilitate interaction with template/primer.

DISCUSSION

Comparison of sequence, structure and function of HIV-1 RT with other template dependent polymerases revealed that a common motif, the helix clamp exists, conferring nucleic acid binding and guiding function. In HIV-1 RT this motif has the amino acid sequence $V\ (X)_4\ K^{259}LVGKL\ (X)_{16}\ KLLR^{284}\ (X)_4\ L$. Comparison with other polymerases reveals a consensus sequence of $U\ (X)_4\ BLUGBU\ (X)_{9-20}\ BUUB\ (X)_4\ U$, where X is any amino acid, U is one having an unpolar side chain and B is one having a basic side chain (Lys or Arg). The conclusion that the helix clamp motif (αH−turn−αI) in the thumb of HIV-1 RT confers nucleic acid binding function rests on information from crystallographic studies, which have shown that the helix clamp motif is oriented towards the nucleic acid binding cleft. Using the RT-C_α and nucleic acid phosphate coordinates from Jacobo-Molina *et al.* (5) as a basis, we show that contacts are possible between side chains of basic amino acids Lys of helix αH and the backbone of the primer strand, and also between basic Lys and Arg in Helix αI and the template strand. Relying on the sequence and structure of the helix clamp motif in HIV-1 RT, criteria were developed for searching similar motifs in other polymerases. These criteria were sequence homology, secondary structure similarity (namely arrangement of residues in α-helical regions), and clustering of basic amino acids at one side of amphiphilic helices. Applying these criteria in a search of protein sequences in the sequence data bank turned up exclusively nucleic acid polymerases, indicating that the helix clamp motif is an amino acid sequence specific for polymerases. Search by sequence homology alone was not sufficient to select only polymerases. Including the criteria for structural homology enhanced the stringency of the search with the result that the selection of proteins was restricted to polymerases having the helix clamp motif as common sequence. However, not all nucleic acid polymerases in the sequence database were picked out by applying our search criteria, e.g. KF and T7 RNAP polymerase did not show up. Grouping the organisms into those containing polymerases with and without the helix clamp motif shows no larger phylogenetic distance between the groups than within the groups. Also within the group of organisms having polymerases containing the helix clamp motif no correlation exists between sequence homology within the helix clamp motif and phylogenetic distance. Moreover the motif appears at variable locations in the polymerase genomes. Given that this sequence motif is quite simple, consisting of only a few degenerately specified amino acid positions, the observed heterogeneity in its occurrence throughout the polymerases suggests that it may have emerged in several polymerase families independently as a consequence of the common need to bind nucleic acid nonspecifically.

While the residues of the polymerase active site of p66 are buried in p51 (28), the thumb regions of both subunits are similarly folded, as can be seen in the crystal structure model (5). The observation that the p51 subunit of HIV-1 RT heterodimer contains the helix clamp with an almost identical local geometry as compared to the p66 subunit raised the question whether the helix clamp also has nucleic acid binding capacity in p51. The published crystal structure analysis of HIV-1 RT in complex with a 18/19 nt template/primer DNA cannot provide an answer to this question, since this DNA fragment was too short to illuminate potential interactions of the p51 thumb with a long template/primer.

We modeled a template/primer of 27 nucleotides RNA in order to study the interaction of nucleic acid with RT. The following constraints were applied: 1) We assumed that the template/primer adopts A-form. Evidence for that is provided by hydroxyl radical footprinting experiments and by crystallographic studies. 2) The template/primer was docked to RT using the active sites of polymerization and RNase H as fix points for the 3′-terminus of the primer and the hydrolyzed phosphodiester bond of the template, respectively. 3) The distance on the template between the two active sites was fixed at 18 nucleotides of A-form RNA, in line with crystallographic and enzymatic analyses.

These modeling studies suggest that RT interacts with the template/primer up to 25 base pairs upstream of the primer 3′-terminus. This does not necessarily contradict hydroxyl radical footprinting studies of Metzger *et al.* (29) which show protection of only 18 nucleotides. Rather, we consider the results from hydroxyl radical footprinting as a lower limit of the estimated region tightly interacting with nucleic acid. Loosely bound nucleic acid appears as accessible to hydroxyl radicals if the nucleic acid associates and dissociates during the exposure time of the probe. Crosslinking studies (26) and very recently DNase I footprinting studies (30) provide evidence that the region interacting with RT is extended at least 25 nucleotides upstream the 3′-terminus of the primer.

Our studies show that the template/primer extended upstream of the RNase H active site passes closely over the p51 thumb. The shortest distance between amino acids within p51 and the terminal residues of the template/primer model was observed after extrapolating 7−8 nucleotides from the upstream end of the 18 nt RNA spanning the active sites, in agreement with the results of the crosslinking experiments. The amino acids nearest to the template/primer are residues within the helix αI of the p51 helix clamp. This is in line with a speculation of Nanni *et al.* (31) who suggested that both helices αH and αI of the p51 thumb could form contacts with nucleic acid.

Basic residues in the p51 thumb may contribute to the template/primer binding capacity of HIV-1 RT apart from nucleic acid binding elements around the polymerase active site in p66, and apart from another probable interaction of a loop in the RNase H domain containing His^{539}. The contact of His^{539} to the bases of bound nucleic acid was already predicted by Jacobo-Molina *et al.* (5) and is confirmed by our model. Site-directed mutagenesis studies to prove the suggested role of amino acids in helix αI in the p51 thumb for nucleic acid binding are currently underway in our lab.

The proximity of the p51 helix clamp to the active site of the RT's RNase H domain led us to speculate that any binding interaction of the p51 helix clamp with nucleic acid may be related to the endonucleolytic RNA digesting function inherent to HIV-1 RT. The p51-mediated nucleic acid binding capacity probably plays a role for guiding the nucleic acid strand lagging behind the RNase H domain during nuclease activity of RT. Moreover, nucleic acid binding via p51 may be important for the process of strand transfer. During reverse transcription initially synthesized short DNA products must be removed from the template viral RNA genome and rehybridized after translocation to the subsequent DNA elongation site (summarized in 26). It has been suggested that this so-called strand transfer process requires an intermediate in which three strands of nucleic acid meet closely, in preparation for formation of a three stranded intermediate (26). We speculate that the helix clamp within the p51 subunit may play a supportive role in assembling this intermediate.

ACKNOWLEDGEMENTS

The authors thank Andrew Wedel for valuable discussion during preparation of the manuscript. T.H. thanks the Studienstiftung des deutschen Volkes for support. This work was supported by the Bayer AG.

REFERENCES

1. Ollis, D. L., Brick, P., Hamlin, R., Xuong, N. G. and Steitz, T. A. (1985) *Nature*, **313**, 762–766.
2. Sousa, R., Chung, Y. J., Rose, J. P. and Wang, B. C. (1993) *Nature*, **364**, 593–599.
3. Arnold, E., Jacobo-Molina, A., Nanni, R. G., Williams, R. L., Lu, X., Ding, J., Clark jr., A. D., Zhang, A., Ferris, A. L., Clark, P., Hizi, A. and Hughes, S. H. (1992) *Nature*, **357**, 85–89.
4. Kohlstaedt, L. A., Wang, J., Friedman, J. M., Rice, P. A. and Steitz, T. A. (1992) *Science*, **256**, 1783–1790.
5. Jacobo-Molina, A., Ding, J., Nanni, R. G., Clark jr., A. D., Lu, X., Tantillo, C., Williams, R. L., Kamer, G., Ferris, A. L., Clark, P., Hizi, A., Hughes, S. H. and Arnold, E. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 6320–6324.
6. Johnson, M. S., McClure, M. A., Feng, D. F., Gray, J. and Doolittle, R. F. (1986) *Proc. Natl Acad. Sci. USA*, **83**, 7648–7652.
7. Poch, O., Sauvaget, I., Delarue, M. and Tordo, N. (1989) *EMBO J.*, **8**, 3867–3874.
8. Delarue, M., Poch, O., Tordo, N., Moras, D. and Argos, P. (1990) *Protein Eng.*, **3**, 461–467.
9. Blanco, L., Bernard, A., Blasco, M. A. and Salas, M. (1991) *Gene*, **100**, 27–38.
10. Kamer, G. and Argos, P. (1984) *Nucleic Acids Res.*, **12**, 7269–7282.
11. Argos, P. (1988) *Nucleic Acids Res.*, **16**, 9909–9916.
12. Larder, B. A., Purifoy, D. J. M., Powell, K. L. and Darby, G. (1987) *Nature*, **327**, 716–717.
13. Jacobo-Molina, A. and Arnold, E. (1991) *Biochemistry*, **30**, 6351–6361.
14. Yadav, P. N. S., Yadav, J. S., Arnold, E. and Modak, M. J. (1994) *J. Biol. Chem.*, **269**, 716–720.
15. Gopalakrishnan, V., Peliska, J. A. and Benkovic, S. J. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 10763–10767.
16. Götte, M., Fackler, S., Hermann, T., Gross, H. J., Cellai, L., LeGrice, S., Heumann, H., *in preparation*.
17. Pearson, W. R. and Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.
18. Rost, B. and Sander, C. (1993) *J. Mol. Biol.*, **232**, 584–599.
19. Maxfield, F. R. and Scheraga, H. A. (1976) *Biochemistry*, **15**, 5138–5153.
20. Claessens, M., VanCutsem, E., Lasters, I. and Wodak, S. (1989) *Protein Eng.*, **2**, 335–345.
21. Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. and Weiner, P. (1984) *J. Am. Chem. Soc.*, **106**, 765–784.
22. Davies, J. F., Hostomska, Z., Hostomsky, Z., Jordan, S. R. and Matthews, D. A. (1991) *Science*, **252**, 88–95.
23. Sobol, R. W., Suhadolnik, R. J., Kumar, A., Byeong-Jae, L., Hatfield, D. L. and Wilson, S. H. (1991) *Biochemistry*, **30**, 10623–10631.
24. Basu, A., Ahluwalia, K. K., Basu, S. and Modak, M. J. (1992) *Biochemistry*, **31**, 616–623.
25. Kumar, A., Hyeung-Rak, K., Sobol, R. W., Becerra, S. P., Byeong-Jae, L., Hatfield, D. L., Suhadolnik, R. J. and Wilson, S. H. (1993) *Biochemistry*, **32**, 7466–7474.
26. Peliska, J. A. and Benkovic, S. J. (1992) *Science*, **258**, 1112–1118.
27. Schatz, O., Cromme, F., Grüninger-Leitch, F. and LeGrice, S. F. J. (1989) *FEBS Lett.*, **257**, 311–314.
28. Hostomsky, Z., Hostomska, Z., Fu, T. and Taylor, J. (1992) *J. Virol.*, **66**, 3179–3182.
29. Metzger, W., Hermann, T., Schatz, O., LeGrice, S.F.J. and Heumann, H. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 5909–5913.
30. S.F.J. LeGrice, *personal communication*.
31. Nanni, R.G., Ding, J., Jacobo-Molina, A., Hughes, S.H. and Arnold, E. (1993) *Perspectives in Drug Discovery and Design*, **1**, 129–150.
32. Kyte, J. and Doolittle, R.T. (1982) *J. Mol. Biol.*, **157**, 105–132.