

Model of full-length HIV-1 integrase complexed with viral DNA as template for anti-HIV drug design

Rajeshri G. Karki^{a,†}, Yun Tang^{a,b,†}, Terrence R. Burke, Jr.^a & Marc C. Nicklaus^{a,*}

^aLaboratory of Medicinal Chemistry, Center for Cancer Research, National Cancer Institute, National Institutes of Health, DHHS, 376 Boyles Street, Frederick, MD 21702, USA; ^bPresent address: Department of Medicinal Chemistry, School of Pharmacy, Fudan University, 138 Yiueyuan Road, Shanghai 200031, P. R. China

Received 24 June 2004; accepted in revised form 7 October 2004
© Springer 2005

Key words: docking, drug-design, human immunodeficiency virus, integrase, molecular modeling

Summary

We report structural models of the full-length integrase enzyme (IN) of the human immunodeficiency virus type 1 (HIV-1) and its complex with viral and human DNA. These were developed by means of molecular modeling techniques using all available experimental evidence, including X-ray crystallographic and NMR structures of portions of the full-length protein. Special emphasis was placed on obtaining a model of the enzyme's active site with the viral DNA apposed to it, based on the hypothesis that such a model would allow structure-based design of inhibitors that retain activity *in vivo*. This was because bound DNA might be present *in vivo* after 3'-processing but before strand transfer. These structural models were used to study the potential binding modes of various diketo-acid HIV-1 IN inhibitors (many of them preferentially inhibiting strand transfer) for which no experimentally derived complexed structures are available. The results indicate that the diketo-acid IN inhibitors probably chelate the metal ion in the catalytic site and also prevent the exposure of the 3'-processed end of the viral DNA to human DNA.

Abbreviations: 3'-P – 3'-Processing; ABNR – adopted basis Newton-Raphson; ASV-IN – Avian sarcoma virus integrase; DKA – Diketo-acid; HIV-1 – human immunodeficiency virus type I; HMG-I (Y) – high mobility group protein isoform I and Y; IN – integrase; LTR – long terminal repeat; MA – matrix protein; MD – molecular dynamics; PR – protease; PDB – Protein Data Bank; RT – reverse transcriptase; SGI – Silicon Graphics, Inc.; SH3 – Src-homology 3; ST – strand transfer; rms – root mean square deviation; Vpr – viral protein R.

Introduction

The *pol* gene of human immunodeficiency virus type 1 (HIV-1) encodes three enzymes that are essential for the virus: protease (PR), reverse transcriptase (RT) and integrase (IN). Agents targeting RT and PR have become successful drugs in the fight against AIDS and these are often

used in combination [1–3]. However, the virus's high mutation rates have given rise to strains that are resistant even to such drug “cocktail” type of therapies. Thus the need for finding new drugs based on inhibition of additional targets is more pressing than ever.

The integrase enzyme of HIV-1 is one such target. It is essential for the viral replication cycle, catalyzing the insertion of the reverse-transcribed viral DNA into the host genome, which is the prerequisite for the formation of the next generation of provirus [4, 5] (Figure 1). It has been

[†]Both authors contributed equally to this work. *To whom correspondence should be addressed. Fax: +1-301-846-6033; E-mail: mn1@helix.nih.gov

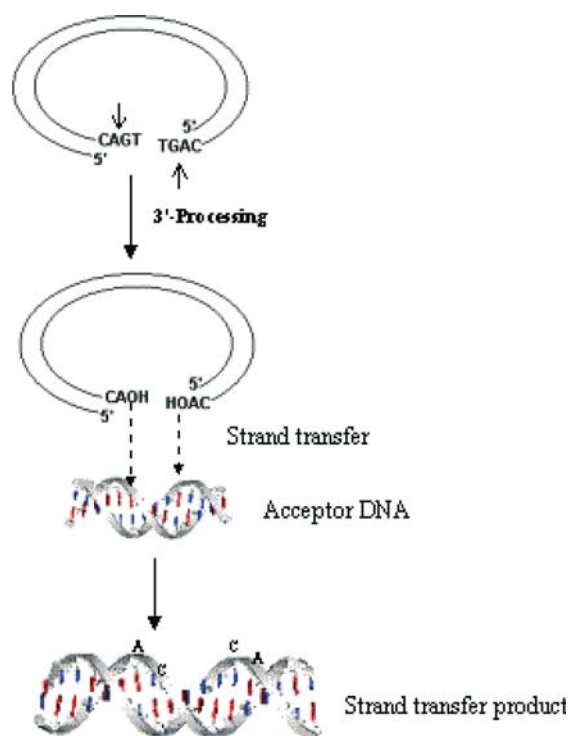


Figure 1. Illustration of the integration process.

shown that viruses encoding catalytically defective integrase do not replicate in cell culture [6]. This lack of replication of integrase defective virus strongly suggests that human host cells contain no endogenous enzymes capable of performing viral integration. The very same absence of close mammalian analogs of IN may result in advantageous therapeutic profiles for agents based on IN inhibition.

In spite of this attractiveness of HIV IN as an anti-AIDS target, the development and delivery to the market of effective IN inhibition-based drugs have lagged far behind drugs directed at RT and PR. Efforts toward development of inhibitors targeted at IN have been hampered, among other factors, by the absence of crystal structures of the full-length enzyme and by uncertainties in the understanding of the biochemical mechanism of proviral integration.

The integration of viral DNA into human DNA proceeds by IN recognizing and binding to attachment sites located at the ends of linear viral DNA, followed by the cleavage of a terminal GT dinucleotide downstream of a conserved CA dinucleotide at both 3'-ends of the viral DNA.

This reaction is known as 3'-processing (3'-P). It occurs in the cytoplasm and results in the formation of a complex of IN and viral DNA, which has the 3'-OH group of the CA unit exposed. This hydroxyl group acts as a nucleophile in the subsequent transesterification reaction, wherein both 3'-processed viral ends are covalently attached to the host cell DNA. These two joining sites on the host DNA are five base pairs apart. In canonical B-DNA, this five base pair spacing would correspond to a distance of approximately 15 Å [7]. The second reaction, referred to as strand transfer (ST) or the integration proper occurs in the nucleus. It has been shown that 3'-P occurs very rapidly while the ST step of viral integration proceeds on a much slower time scale *in vivo* [8]. Both reactions are analyzed in biological assays using the purified enzyme, metal ion cofactor and double stranded oligonucleotides, which function as substitutes for the viral DNA and the host target DNA simultaneously [9]. Variations of this assay have been studied to understand the biochemical reaction and the structure and function of the enzyme [10–15].

IN consists of 288 residues, divided into three distinct domains: the N-terminal domain (residues 1–49), the catalytic core domain (residues 50–212) and the C-terminal domain (residues 213–288) (Figure 2) [7]. The N-terminal domain contains a highly conserved HHCC motif made up of residues H12, H16, C40 and C43 that coordinate a Zn^{2+} ion. The binding of the Zn^{2+} ion to the HHCC motif has been shown to promote the multimerization of IN *in vitro* [16, 17]. The central core domain contains a triad of three highly conserved residues, D64, D116 and E152, which coordinate divalent cations such as Mg^{2+} or Mn^{2+} and are assumed to be key residues of the enzyme's catalytic site. Mutation of any of these three acidic residues abolishes catalytic activity. The core domain is believed to bind viral DNA specifically [18]. The C-terminal domain contains a less conserved sequence, though the overall structure resembles that of the Src-homology 3 (SH3) domain [19, 20]. The C-terminal domain binds viral DNA in a nonspecific manner [10]. All three domains have been shown to be required for the two integration reactions, 3'-P and ST.

While these features of IN are generally supported to a large extent by structural work, much less direct experimental evidence exists for the

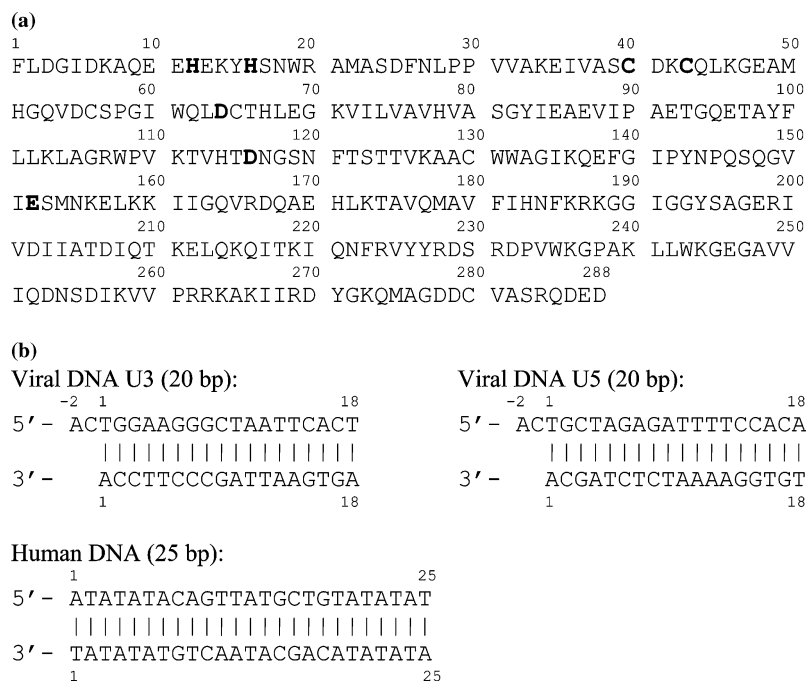


Figure 2. (a) Amino acid sequence of HIV-1 integrase. Residues 1–50: N-terminal domain; 51–212: core domain; 213–288: C-terminal domain (several critical residues are shown in bold font). (b) Viral DNA ends U3 and U5 and human DNA sequences used in this study.

overall configuration of IN *in vivo*. As of today, not even one full-length experimental structure of HIV-1 IN is available. Additional uncertainty pertains to the multimerization state of HIV-1 IN preferred for catalysis. One widely held assumption is that functional IN acts as a multimer [7, 21] at least a tetramer if not an octamer.

The structures of the three domains have been determined separately by means of X-ray crystallography and NMR spectroscopy [20, 22–26]. For each of these domains, an assembly into dimers has been observed. The structures of the combined core and C-terminal domains, and core and N-terminal domains, have also been determined [27, 28]. A number of experiments have each painted a partial picture of how the three domains of IN may organize and interact with viral DNA (Table 1). However, these results by themselves are fragmentary, and it is difficult to directly obtain from them a clear picture of how IN recognizes and binds to viral DNA.

Research efforts spanning more than a decade have recently resulted in identification of clinical candidates against HIV-1 IN [29, 30]. These compounds have validated the approach of basing

anti-HIV drugs on IN inhibition. However, detailed structural data on which the design of these inhibitors may have been based have not been published. In our ongoing efforts to develop IN inhibitors [31–36], we therefore saw the need for deriving a model of the full-length enzyme and its complex with viral DNA on which to base specific IN inhibitor design.

It has become clear in the recent past that ST inhibition correlates more closely with anti-HIV activity in *in vivo* assays and/or clinical application than does inhibition of 3'-P [37, 38]. Likewise, conducting the *in vitro* assay with pre-incubation of the DNA substrate prior to adding the potential inhibitor (versus the inverse sequence of steps that was more commonly used in the past) was found to have a better correlation with *in vivo* activity [37]. These results and other mounting evidence led us to posit that effective inhibition of viral replication based on interfering with the action of IN could be achieved by an inhibitor binding with the IN/viral DNA complex and not with the “empty” enzyme alone, followed by selectively inhibiting the ST reaction. As a consequence, it became necessary to first construct a model of the

Table 1. Summary of features of the structural models of full-length HIV-1 integrase. Models I–III: presented in this paper; A–D: reported in the literature.

Experimental	Model I	Model II	Model III	Model A [39]	Model B [28]	Model C [13]	Model D [40]
Full-length protein-288 residues [43]	Yes	Yes	Yes	No	No	No	No
Viral DNA sequence [47, 48]	Yes	Yes	Yes	Both U5-LTR	Viral DNA sequence different No target DNA	Both U5-LTR	Both U5-LTR
Number of Mg ions in the catalytic site [60, 73]	1	2	2	0	0	0	0
Distance between two 'active' catalytic triad (probably involved in catalytic reaction)	~15 Å	~15 Å	~15 Å	~20 Å	~40 Å	~60 Å ^g	~33 Å
Five base pair separation during strand transfer	Yes	Yes	Yes	Yes	No	Yes	Six base pair
Residues 49–69,	+	+	+	+	+	+	+
139–152,	+	+	+	+	+	+	h
213–246,	+	+	+	+	+	+	+
247–270,	+	+	+	+	+	+	+
271–288	+	+	(See footnote c)	(See footnote d)	(See footnote e)	(See footnote d)	(See footnote d)
cross-linked to different oligonucleotide substrates representing the viral DNA ^a [14, 39]							
Residues 1–12,	–	–	–	+	No target	+	+
49–69,	+	+	+	+	DNA	+	+
139–152,	+	+	+	–		+	+
213–246,	–	–	–	–		–	+
271–288	–	–	–	d		d	d
cross-linked to different oligonucleotide substrates representing the target DNA ^a [14, 39]							
Viral DNA binds with C-terminal domain from one monomer and core domain from another monomer (trans to the active site) [13, 27, 39]	Yes	Yes	Yes	Yes	Yes	Yes	Yes
E152 close to A/T base pair [58]	Yes	Yes	Yes	Yes	Yes ^f	E152 ~17 Å from A ^b	No
K159 cross-linked at Ade1 ^b [12, 15]	Yes	Yes	No	Yes	Yes ^f	K159 ~20 Å from A1	Yes
Ade(–2) ^b from lower strand cross-linked to 139–152 (75% to Y143 & 25% to Q148 [12])	No	No	No	Yes Y143 absent	No	No	No

Cyt(-1) ^b most likely cross-links to W61, Q62, L63 [12]	No	No	No	Possible with some movement	No	No	No
Thy5 ^b from upper strand most likely cross-links to 247–270 [12]	No	No	No	Yes	No	No	Thy4 from lower strand is close to 247–270
A contiguous strip of positive charge extending from the catalytic site along the outside face of the IN52–288 dimer includes residue K159, and continues through residues K186, R187, and K188, and out to residues K211, K215, and K219 of the $\alpha 6$ helix from the paired monomer in the dimer.[27]	Yes	Yes	Yes	Yes	K186 & K188 are far away from the DNA strand in the model.	K159 & K186 are closer to target DNA than to viral DNA	Yes
IN residue E246 binds near position five of lower strand in the viral cDNA [13]	E246 can bind near position 11 of upper strand in viral DNA	E246 can bind near position 11 of upper strand in viral DNA	E246 can bind near position 18 of upper strand in viral DNA	E246 > 20 Å from viral DNA	E246 > 20 Å from viral DNA	E246 > 10 Å from viral DNA	Yes

^aSix integrase subdomains, mapping to amino acids 1–12, 49–69, 139–152, 213–246, 247–270 and 271–288 were found to be cross-linked to different substrates representing both viral and target DNA. A ‘+’ sign indicates such a cross-linking seems possible in the model. A ‘-’ sign indicates the observed peptide is too far away from the DNA strand in the model, so no cross-linking is possible from the model.

^bThe nucleotide numbering corresponds to that depicted in Figure 3.

^cSince the A-form of viral DNA was used, the DNA strand with 20 base pairs was shorter and therefore could not reach the residues 271–288. If longer, cross-linking would be possible from the model.

^dResidues 271–288 missing in the model.

^eResidues G140–Q148 are missing from all four monomers in the structure. Therefore interactions with this portion of the peptide cannot be seen. However, in presence of these residues, cross-linking seems plausible. Also the viral DNA sequence is short (14 base pairs) so interactions with residues 213–246, 247–270 seem plausible but cannot be seen in the modeled structure. As for interactions with the (missing) residues 271–288, it is difficult to make any guess.

^fThe nucleotide sequence for the U5 and U3 LTR being different in this model from that used in *in vitro* studies, the nucleotide that comes close to E152 and K159 is thymine and not adenine. However, this position corresponds to the A/T base pair in the original LTR sequence.

^gThe role of the catalytic triad in the integration reaction is difficult to explain using this model, as the catalytic triad is far away from the site of integration, the distance being ~15 Å and 12 Å from the two LTR 3'-ends, respectively.

^hPossible with some movement.

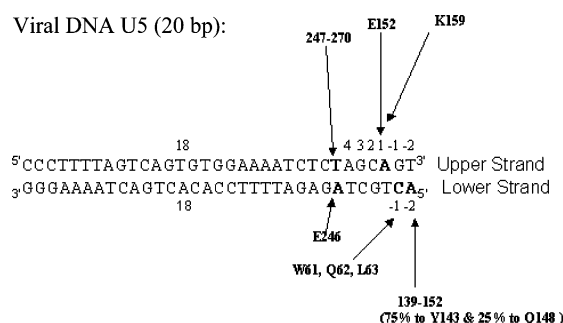


Figure 3. Representation of the experimentally observed cross-linking between U5-LTR and integrase [12–14, 26, 39, 58].

full-length enzyme that would be capable of “holding” a model of the viral DNA oligomer as used in the enzyme-based assays.

Several such models have been proposed for the interaction of viral DNA with a multimer of IN [13, 27, 28, 39–41] (Table 1). However, each model is consistent with only certain experimental results and cannot explain all the biochemical findings (explained in Results and discussion’ and detailed in Table 1). None of these ‘R models contain all the residues 1–288. In addition, metal ions have not been included in most of these models. This makes it more difficult to gain a detailed understanding based on them of how known inhibitors may bind and elicit their activity, given the fact that metal cofactor is essential for the enzyme’s activity. Therefore, we felt the need in our drug-design efforts for truer three-dimensional structural models of the full-length IN complexed with viral DNA.

It is important to note that while we have attempted to build models that are reliable in the global sense of overall structural geometry of the protein–DNA complex, our emphasis was to obtain from this effort a model of the active site that would be more amenable for use in the design of better inhibitors of viral replication. This meant that if modeling decisions had to be taken which would favor either the global aspects of our models or the detailed geometry at the location where the 3’ end of the viral DNA and the enzyme’s active site come together, the latter usually prevailed. These models were developed by making use of, to the extent possible, the various experimental evidence pertaining to HIV-1 (and related) integrase(s), be they of structural nature or of other type such as protein footprinting, cross-linking etc.

To gain additional insight into how the ST process may occur, and thus be inhibitable, structures with simple human DNA models added to the complex were also generated.

These models have already proven useful in initial, limited applications to study the structure-activity relationship and the binding modes of a series of diketo-acid (DKA) HIV-1 IN inhibitors [38, 42] in the context of our structure-based anti-HIV IN inhibitor design. However, the details of the model derivation and the entire range of results and insights gained from these models have not yet been published. They constitute the subject of this paper. Additionally, we present a comparative overview of our model and earlier reported models [13, 27, 28, 39, 40], specifically analyzing how well each agrees with known experimental results and how they facilitate structure-based inhibitor design.

Methods

The amino acid sequence for the full-length protein (288 residues) was taken from literature (Figure 2) [43]. Crystal structure information of the different domains was obtained from the coordinates deposited in the Protein Data Bank (PDB) [44]. The PDB codes of the structural templates used for the study are 1QS4 [45], 1BI4 [24] and 1BL3 [24] (structures of the core domain dimer), 1WJA [23] (solution structure of the N-terminal domain), 1IHV [20] (structure of the C-terminal domain), 1EX4 [27] (the combined core and C-terminal domains) and 1K6Y [28] (the combined core and N-terminal domain). The viral DNA which is known to contain two long terminal repeat (LTR) sequences [46, 47] (hereafter referred to as U3 and U5 LTR), and the human DNA models were built based on reported sequences [47, 48].

Molecular modeling

All the structural models were constructed using the molecular modeling package SYBYL version 6.8 [49], running on a Silicon Graphics (SGI) Octane workstation. They were refined with Molecular Dynamics (MD) simulations using the program CHARMM [50] version c27b3, utilizing the CHARMM all-atom force field [51, 52]. Zinc finger force field parameters were developed using

ab initio calculations with Gaussian 98 [53]. The TIP3P water model was used in the MD simulations [54]. All MD simulations were performed on either an SGI Origin 2000 or Beowulf-type cluster in parallel operation [55]. Three variants of the full-length HIV-1 IN complexed with DNA were modeled using different combinations of the various available experimental results.

Model I: N-terminal domain from NMR structure 1WJA [23], one metal ion, protein bent, viral DNA in canonical B form

The model building of the full-length enzyme was started by extracting the coordinates from the crystal structure of the combined core and C-terminal domains (PDB code 1EX4) [27]. All water molecules were removed. The missing residues in chain A (residues P142–N144) and chain B (residues E138–G149) of 1EX4 were built using the Biopolymer module in Sybyl 6.8 [49]. The conformation of this missing loop was adjusted by Biopolymer's loop search method such that it was consistent with that observed in the crystal structure 1BL3.pdb [24]. The two C-terminal tails, residues Y271–D288, were generated in a random coiled conformation with the SYBYL/Biopolymer/Builder module, as no crystal structure information is available. The C-terminal domain, being the hypothesized DNA binding domain, may possibly bind with the DNA in a dimerized state. Considering this, one of the two long helices linking the C-terminal domain with the core domain was bent at residue T210 to allow dimerization of the C-terminal domains as seen in the crystal structure 1IHV [20]. A monomer of the N-terminal domain of the protein, extracted from the NMR structure 1WJA [23], was then placed above the residues S56 and D55 of the two chains A and B, respectively, in the processed structure of 1EX4 described above. Such a placement was chosen so as to be consistent with the crystal structure of the combined core and N-terminal domain 1K6Y [28]. A loop region was then built by Biopolymer's loop search method to link the N-terminal domains and core domains together. One Mg^{2+} ion was manually placed in each of the active sites of the core domain dimer between the carboxyl groups of residues D64 and D116, as observed in the crystal structure 1QS4.pdb [45]. One Zn^{2+} ion was added to each of the two

N-terminal domains such that they coordinated with the HHCC zinc finger motif. A schematic representation of the model building process is depicted in Figure 4.

This initial structural model of the free IN dimer was minimized *in vacuo* for 1000 steps, using the adopted basis Newton–Raphson (ABNR) method, to remove unfavorable van der Waals contacts, keeping harmonic constraints with a force constant of 20.0 kcal/(mol Å²) on heavy atoms of the dimer. After that, seven chloride ions were added near the long side chains of seven positively charged residues (Lys and Arg) on the surface of the dimer to make the system electroneutral. An MD simulation was then carried out on this model system *in vacuo*, for 500 ps with a time step of 1 fs, using a dielectric constant of 1, keeping harmonic constraints with a force constant of 20.0 kcal/(mol Å²) on atoms N, C, O, and α -C of residues 1–46, 58–136 and 150–270, to obtain the possible conformations for the reconstructed and added parts of the model. The lowest energy conformer from the MD run was extracted and used for further modeling.

Structures of the viral DNA U3 and U5 LTRs and the human DNA were built based on reported sequences (Figure 2) [47, 48] in double stranded B-form using the Biopolymer module in Insight II [56]. Two nucleotides (GT) from the 3'-end of both the U3 and U5 LTRs were removed so that the viral DNA structures represented the post-3'-processed state. The electrostatic potential on the solvent-accessible surface of the full-length protein dimer was calculated using the program GRASP [57] (Figure 5).

Combining available experimental results regarding the possible viral DNA binding sites [14, 26, 39, 58] and the results from the electrostatic potential calculations, the viral DNA was docked into one of the two active sites of the IN dimer, with the conserved CA dinucleotide of the viral DNA located near one of the catalytic residues, E152. The IN-viral DNA U3 complex was built by changing the nucleotide sequence from U5 to U3. The two IN-viral DNA complexes (one with the U3-LTR and the other with the U5-LTR) were composed of 10,261 and 10,262 atoms, respectively. These structural models were solvated separately with a 10 Å aqueous solvent shell using the TIP3P water model [54]. All water molecules within 1.8 Å and beyond 10 Å of any heavy atoms in the model were deleted. Sodium ions were

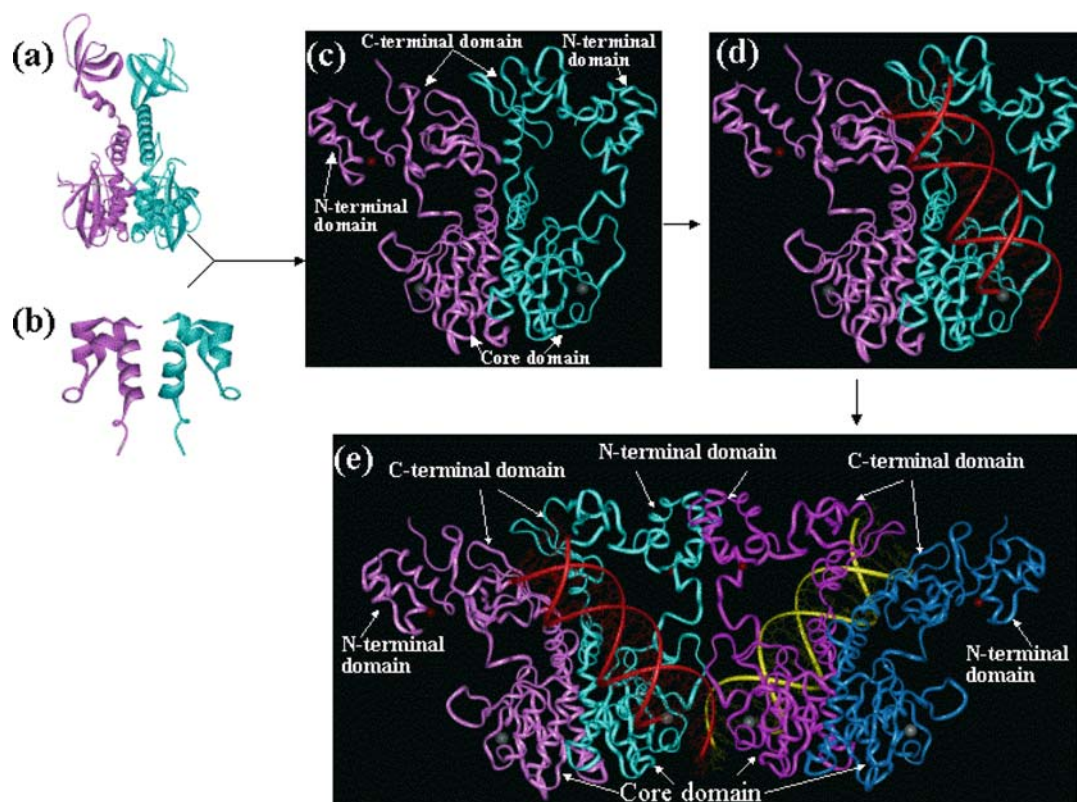


Figure 4. Schematic representation of the model building process. (a) Structure of the combined core and C-terminal domain dimer, PDB code 1EX4 [27]. (b) Structure of the N-terminal domain dimer, PDB code 1WJA [23]. Structural model of (c) full-length integrase dimer; (d) full-length integrase dimer bound with viral DNA U5 LTR; (e) full-length integrase tetramer bound with viral DNA U3 and U5 LTRs.

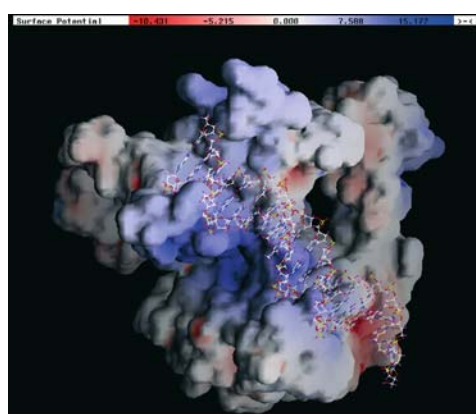


Figure 5. Electrostatic potentials shown on the solvent-accessible surface of the IN dimer, complexed with viral DNA. This figure was produced with the program GRASP [46]. The surface charge distribution was colored as follows. Blue: positive (~ 15.2 kT); red: negative (~ -10.4 kT); white: neutral. (k is the Boltzmann constant and T is the temperature).

placed 3.6 \AA from each phosphorus atom on the O=P=O bisector of DNA to make the systems electroneutral. The resulting solvated systems were very large, with 37,879 and 37,919 atoms for the IN-viral DNA U3 complex and IN-viral DNA U5 complex, respectively. After energy minimization, these systems were subjected to MD simulations without boundary conditions. At first, harmonic forces of $10.0 \text{ kcal}/(\text{mol } \text{\AA}^2)$ were applied to protein atoms N, C, O, α -C of residues 1–46, 58–136, 150–208 and 220–270, to the Zn^{2+} and Mg^{2+} ions, and to the backbone heavy atoms of DNA. The forces were removed gradually. The applied force was $9.0 \text{ kcal}/(\text{mol } \text{\AA}^2)$ starting from 380 ps into the simulation, reduced to $5.0 \text{ kcal}/(\text{mol } \text{\AA}^2)$ after 680 ps and to $1.0 \text{ kcal}/(\text{mol } \text{\AA}^2)$ after 1320 ps. After 1720 ps, the constraints were released completely. The total running time was 3.0 ns with a time step of 2 fs at 310 K. The non-bonded interaction cutoff was set to 14.0 \AA . The minimized average

structures were calculated from the final trajectory corresponding to 2.6–3.0 ns.

Following this, the two IN-viral DNA complexes were combined to form an IN tetramer in a centrally symmetrical fashion, with the two opposite active sites separated by about 15 Å, corresponding to the distance spanned by five base pairs along the major groove of normal B-DNA. This arrangement caused two N-terminal domains, one from each complex, to end up juxtaposed to each other. They were adjusted to form a dimer, using the NMR structure 1WJA [23] as a template, thus functioning as a dual “hook” holding the two IN dimers together.

Finally, a model of IN-viral DNA complexed with human DNA was constructed by placing a 25-base pair DNA model below the gap of the two IN-viral DNA complexes, built using the sequences utilized in *in vitro* assays [48]. Two adenine nucleotides at the 3'-end of the two viral DNA LTRs came close to the backbone of human DNA, such that the number of nucleotides between the points of insertion of the two 3'-processed LTRs on the human DNA was five. Thus the model represents the state after the 3'-P reaction but before the ST reaction.

The whole IN-viral DNA-human DNA complex consisted of 22,044 atoms. It was refined again by MD simulation. Harmonic forces of 10.0 kcal/(mol Å²) were applied on protein atoms N, C, O, α -C of residues 1–46, 58–136, 150–208 and 220–270, to the Zn²⁺ and Mg²⁺ ions, and to the heavy atoms in the backbone of DNA. These constraints were removed gradually, with the applied force being 5.0 kcal/(mol Å²) from 420 ps and 2.0 kcal/(mol Å²) from 800 ps. After 1200 ps, the constraints were released completely. Total running time was 2.6 ns. The minimized average structure was calculated from the final trajectory at 2.2–2.6 ns. These minimized average structures were used for comparison with the initial structure models.

Model II: N-terminal domain from X-ray structure 1K6Y [28], two metal ions, protein bent, viral DNA in canonical B form

A variant of the above model was constructed wherein, instead of the solution structure of the N-terminal domain (1WJA [23]), we used the coordinates of the N-terminal domain from

the crystal structure of the combined core and N-terminal domain (1K6Y [28]). Also, in each of the active sites two Mg²⁺ ions were placed between the carboxyl groups of residues D64, D116 and E152. The location of these metal ions was based on the crystal structures of the core domain of Avian sarcoma virus integrase (ASV-IN) complexed with zinc and cadmium cofactors, respectively (1VSH.pdb and 1VSJ.pdb) [59]. Another reason for our decision to add a second metal ion was the widely held hypothesis that a second metal ion is likely to be carried into the HIV-1 IN active site by the viral DNA [60].

Model III: N-terminal domain from X-ray structure 1K6Y [28], two metal ions, protein not bent, viral DNA in canonical A form

In the third model constructed, coordinates from the crystal structure 1EX4 [27] served as a template for core and C-terminal domains while the coordinates of the N-terminal domain were extracted from the crystal structure of the combined core and N-terminal domain (1K6Y [28]). All missing residues were added in the same way as for Model I. However, unlike Model I, for which the C-terminal domain was dimerized to make it similar to that observed in 1IHV [20], no such modifications were made in this model. All the DNA structures were constructed in A-form unlike for the other models in which the double stranded B-form was used. Two Mg²⁺ ions were placed in the active site as for Models II. Subsequent refinement of both Models II and III was carried out as explained under Model I.

Docking

Docking studies of compounds 1–31 (Table 2) were performed on the structural model of the full-length HIV-1 IN dimer complexed with U5 LTR using the program Glide [61, 62], part of the First Discovery suite (Schrödinger Inc.). Coordinates of the full-length dimer complexed with U5 LTR were extracted from structural Model II and prepared for the Glide calculation by running the scripts *pprep* and *impref*, part of the First Discovery suite. The *pprep* script produces a new receptor file in which all residues are neutralized except those that are relatively close to the ligand

(if the protein is complexed with a ligand) or form salt bridges. The *impref* script runs a series of restrained impact energy minimizations using the Impact utility. Minimization was run until the average root mean square deviation (rms) of the non-hydrogen atoms reached 0.3 Å. In order to study the binding mode of the inhibitors in the presence of metal ions, grid files representing shape and properties of the receptor including Mg^{2+} in the active site were generated. Glide uses two “boxes” to organize the calculation: an enclosing box and a bounding box. The two boxes share a common center. The enclosing box is the larger one of the two and encloses the bounding box. The grid themselves are calculated within the space defined by the enclosing box. The bounding box defines the space through which the center of the defined ligand will be allowed to move during subsequent docking calculation. This box gives a truer measure of the effective size of the search space; the only requirement on the enclosing box is that it be big enough to contain all ligand atoms, even when the ligand center is placed at an edge or vertex of this bounding box. The grid files were generated with the catalytic residues (D64, D116 and E152) as the center of the two boxes, while the size of the bounding box was set to 20 Å so as to explore a larger region of the receptor. The three-dimensional structures of the compounds **1–31** (Table 2), were constructed using the builder tool available through Glide’s Maestro interface. The initial geometry of the structures was optimized using the OPLS-AA force field [63], performing 1000 steps of conjugate gradient minimization. The compounds were subjected to flexible docking using the pre-computed grid files. Glide uses a series of hierarchical filters to search for possible locations of the ligand in the active site region of the receptor. Details about the methodology used by Glide are described elsewhere [64]. Default settings were used for scoring and selection of the best docked poses, with the exception that 1000 (default 400) best docked poses from the second stage of filtering were subjected to minimization by performing 1000 (default 100) steps of conjugate gradient minimization using the OPLS-AA force field [63] and a distance dependent dielectric constant of 2 (a pose means complete specification of the ligand; position and orientation relative to the receptor, core conformation and rotamer-group conformations). For each

compound the 100 top-scored poses were saved and analyzed.

Results and discussion

The goals of our efforts to model full-length IN were to obtain a better understanding of the integration mechanism, to understand the binding modes and mechanism of enzyme inhibition by known inhibitors and to utilize these in a way that would allow us to design new inhibitors. For these reasons, we modeled the complex of the enzyme with viral DNA representing a stage after 3'-P but before ST. We concentrated specifically on obtaining a structure of the area of the protein-DNA complex formed by the active site and the terminal few nucleotides of the viral DNA strands proximal to it. This would be useful for designing active molecules based specifically on ST inhibition. Part of the rationale for this lies in the sequence of steps in the viral life cycle. Following viral entry into the host cell, the virus's genetic material in the form of RNA is released and undergoes reverse transcription into DNA in a reaction catalyzed by RT. Once synthesized, the viral DNA is transported to the nucleus as part of a pre-integration (PI) complex that includes IN, matrix protein (MA), RT, viral protein R (Vpr) proteins and cellular host high mobility group protein isoform I and Y (HMG-I (Y)) [65]. This leads one to conclude that viral DNA is, in effect, always present with IN. IN catalyzes the 3'-P reaction while still in the cytoplasm and then enters the nucleus where it catalyzes the ST reaction [4, 65]. It has been shown that 3'-P occurs very rapidly, with the preprocessed viral DNA subsequently being very tightly bound to IN [8, 66]. In contrast, ST proceeds on a much slower time scale, presumably because the PI complex must locate its target, the host DNA. This is not the case for the viral DNA which is “fed” directly into the IN catalytic site by RT. In agreement with this picture, known inhibitors of IN that are active *in vivo*, especially those belonging to the DKA class, have been found to inhibit the ST reaction much more strongly than the 3'-P reaction, whereas this is not the case for many other compounds that showed good *in vitro* but poor *in vivo* inhibition [37, 38]. It therefore appeared logical to construct a structural model that would specifically allow us to design inhibitors that function in this manner. All the available experimental results were considered for

Table 2. Structures and measured activities of DKA inhibitors used for the docking study [38, 80].

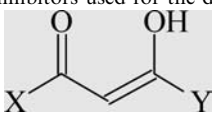





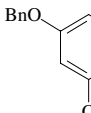
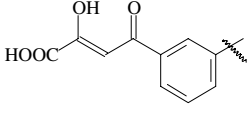
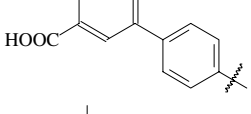

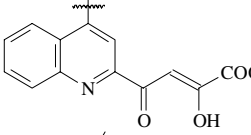

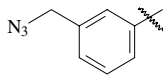
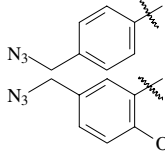
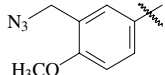
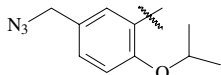
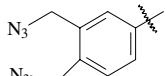
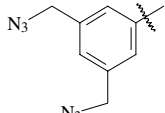
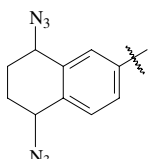
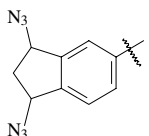
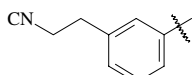
					
Compound	X	Y	Integrase assay (IC ₅₀ μM) ^a	ST	Docking score
3'-P					
1		-COOH	>100	24.2, 25.0	-6.8128
2		-COOH	>100	2.0, 3.0	-7.1007
3		-COOH	65, 100	0.35 ± 0.13	-8.5558
4		-COOH	>100	35	-8.9371
5		-COOH	>100	7.8, 15	-7.4764
6		-COOH	>1000	0.48 ± 0.08	-9.7044
7		-COOH	7.8 ± 2.2	1.83 ± 0.32	-8.3286
8		-COOH	7.2, 7.5	1.28 ± 0.38	-9.1992
9		-COOH	>100	>100	-8.0392
10		-COOH	82	6.5 ± 0.81	-8.2676
11		-COOH	>100	1.43 ± 0.15	-8.1272

Table 2. (Continued)

Compound	X	Y	Integrase assay (IC ₅₀ μM) ^a	ST	Docking score
			3'-P		
12		-COOH	65	0.52 ± 0.10	-7.3173
13		-COOC ₂ H ₅	>100	6.9	-7.4798
14			>100	1.9, 2.0	-8.3564
15			35	0.65 ± 0.19	-7.5102
16			79	1.8	-7.5956
17			40	40	-7.4213
18			>100	>100	-7.8184
19			> 100	34	-7.3540
20			>100	>100	-7.8446
21			>100	>100	-8.9617

Table 2. (Continued)

Compound	X	Y	Integrase assay (IC ₅₀ μM) ^a	ST	Docking score
			3'-P		
22		-COOH	>100	1.53 ± 0.27	-8.6611
23		-COOH	>100	6.1, 7.1	-6.3923
24		-COOH	>100	24, 18	-7.8996
25		-COOH	>100	8.5, 9.0	-7.6564
26		-COOH	>100	25, 23	-7.0654
27		-COOH	>100	0.26, 0.15	-7.9152
28		-COOH	>100	2.0, 2.8	-8.3495
29		-COOH	70	0.32	-7.6265
30		-COOH	>100	0.36	-7.1702
31		-COOH	>100	1.5, 1.8	-7.4388

^aAssays were conducted as reported in [38].

the model building. The generated models were refined by MD simulation first *in vacuo*, followed by MD simulation in explicit solvent as explained in the methods section.

Because of the great variety, incompleteness, and even outright mutual contradiction of the experimental results, three variants of the model were generated using different combinations of the available experimental results. The first two models, termed “I” and “II,” were built with the more common B-form of DNA and a dimerized C-terminal domain. Several alternative possibilities, however, could be explored: a model without dimerization of the C-terminal domain; and use of A-DNA instead of B-DNA, justified by the fact that even less experimental evidence is available for the conformation of the viral DNA than for the protein in the complex. We decided to first explore both of these possibilities using one model, termed “III.” Analysis of the various enzyme-DNA interactions observed in Model III and comparing them with the experimentally observed interactions showed that Model III complied the least with the experimentally observed interactions as compared to Models I and II. This was largely attributed to the short length of the 20 nucleotide long A-form of DNA and to the larger separation between the two C-terminal domains which wasn’t dimerized in Model III. A simple overlay of B-DNA over Model III was done to see if it would fit better than A-DNA. However, because A-DNA is broader than B-DNA (the diameter of A-DNA is approximately 26 Å while that of B-DNA is approximately 20 Å), it fits better in the cleft formed by the C-terminal domain dimer. Also it was realized that A-DNA, because of its shorter length, would not perform better than B-DNA in Models I and II. Therefore the decision was made not to construct additional models with these variations.

Model I of the full-length HIV-1 IN enzyme complexed with viral and human DNA was built using coordinates from the crystal structure of the combined core and C-terminal domains (PDB code 1EX4) [27]. The coordinates for the N-terminal domain were extracted from the NMR structure 1WJA [23]. Model II is different from Model I in two ways. First, in Model II the coordinates for the N-terminal domain were extracted from the crystal structure of the combined core and N-terminal

domain, 1K6Y [28]. Second, two Mg^{2+} ions were placed coordinating with the residues of the catalytic triad. For Model III, the template structures were the same as for Model II. However, unlike Models I and II, in which the C-terminal domain was dimerized to make it similar to the structure observed in 1IHV [20], no such modifications were made in this model. In Models I and II, DNA structures were built in double stranded B-form while in Model III they were constructed in double stranded A-form. There is no experimental evidence as to the conformation of the DNA that is bound to IN. Whereas the naturally occurring form of the DNA is the B-form, at higher salt concentration or low moisture content DNA can change to the A-form. Therefore, in Model III, DNA was built in the A-form to explore this alternative. Figure 4 shows the structures of the individual domains of HIV-1 IN separately, and also after assembly into the model of the full-length enzyme complexed with viral DNA.

All three structural models of the full-length IN tetramer complexed with both viral DNA ends and a human DNA model are centrally symmetric as illustrated in Figure 6. Each viral DNA end is bound to an IN dimer, and two such IN-viral DNA complexes are linked together to form a tetramer through the N-terminal domain dimer bridge, effectively forming a dimer of dimers. In all three structural models, viral DNA binds to the core domain of one monomer and the C-terminal domain of another monomer in a *trans* fashion. Although this binding orientation was modeled we based the decision for such a placement both on results from the calculation of the electrostatic potential on the solvent-accessible surface of the full-length protein dimer using the program GRASP [57] and on experimental results regarding the possible viral DNA binding sites [14, 26, 39, 58]. The electrostatic potential map helped in identifying positively charged residues on the surface of the protein that would favor DNA binding. A separate study has confirmed the *trans* binding of viral DNA to IN [13].

The distance between the two catalytic triads in all three modeled full-length integrase dimers complexed with viral DNA was found to be between 32 and 36 Å. The separation between the points of insertion of two 3'-processed LTRs into the human DNA should be approximately 15 Å, corresponding to the five base pair separation on a

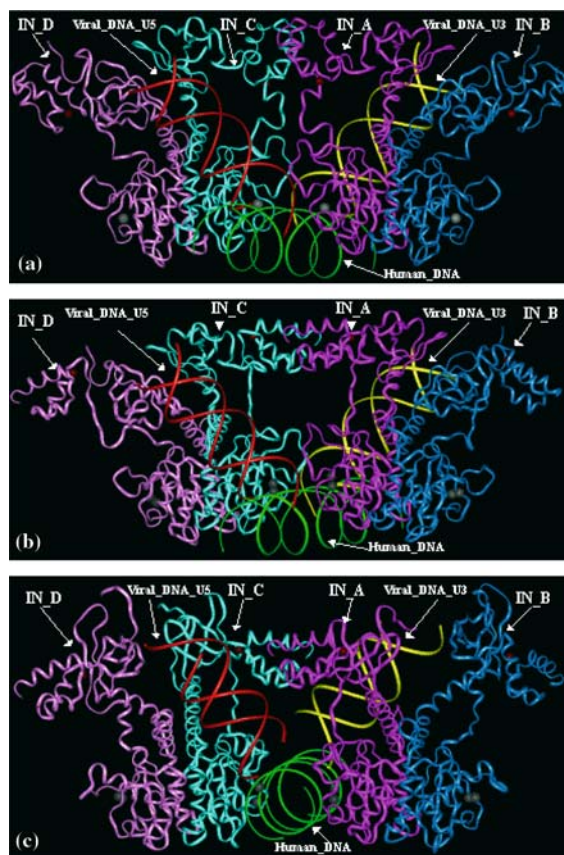


Figure 6. Structural model of the full-length integrase tetramer bound with viral DNA ends and human DNA. (a) Model I; (b) Model II; (c) Model III.

canonical B-DNA [7, 46]. It can also be assumed that the 3'-processed DNA is always complexed with integrase. Based on this, our working hypothesis was that the two catalytic triads involved in the ST reactions come from different dimers. Also, solution structures of the N-terminal domain alone (1WJA [23]) and the combined core and N-terminal domains (1K6Y [28]) show that the N-terminal domain exists as a dimer. However, in our modeled full-length IN dimer, the N-terminal domain was present as a monomer. In order to make this structure consistent with the experimental findings, two IN-viral DNA complexes were combined such that the N-terminal domains, one from each complex, juxtaposed to each other to form a dimer. Depending on the starting coordinates, this N-terminal dimer was consistent with the three-dimensional structure observed in 1WJA [23] (Model I) or 1K6Y [28] (Models II and III),

respectively. This dimerization of the N-terminal domains resulted in the formation of the IN tetramer, which is consistent with the finding that the N-terminal domain promotes tetramerization of HIV-1 IN *in vitro* [7]. Two catalytic triads, one from each dimer, face each other at the interface of the two IN-viral DNA complexes, and are separated by approximately 15 Å. The target human DNA was placed at this interface. In both Models I and II, the conserved adenosine residues of the viral DNA 3'-end are able to make contact with the backbone of the human DNA. The two contact points are five base pairs apart along the major groove. In addition, the two 5'-end unpaired viral nucleotides (CA) are situated in the vicinity of the human DNA, one capable of interacting with the minor groove while the other 5'-end unpaired nucleotides can interact with the major groove. Residues H67, E69, K71, R166, N117 and K159 from the core domain of the full-length IN tetramer lie proximal to the human DNA. Therefore, according to our model the core domain may play an important role in target binding. This is consistent with experimental findings [67].

Comparison of the structural models with experimental results

An important step after constructing each ternary complex structure was to ascertain that the models comply as much as possible with the available experimental data. To this end, each of the modeled structures was evaluated in terms of the interactions made by individual amino acids with the bound DNA. Comparison of these interactions with reported biochemical results are summarized in Table 1.

The electrostatic potential on the solvent-accessible surface of the full-length protein dimer was calculated using the program GRASP [57], which helps to identify residues that are favorable for binding of viral DNA through their positive electrostatic potential. For Model I, these residues were K71, K156, K159, K160, R166, K186, R187, K188, K211, K215, K219, R228, R231, K236, K244, K258, R262, R263, K264 and K266. For Model II the residues were: K156, K159, K160, R166, K186, R187, K188, R228, R231, K236, K244, R262, R263, K264, K266. Finally for Model III the residues were: R20, K34, K42, K46,

K156, K159, K160, K185, K186, K188, R199, K211, K215, K219 and K244. (The 20 nucleotide long A-form of viral DNA is shorter than the B-form and therefore does not reach residues R228, R231, K258, R262, R263, K264, K266 and R284.) The same residues have also been identified as being part of a contiguous strip of positive charge running from the catalytic site along the outside face of the IN^{52–288} dimer [27].

In all three models, the side chain of E152 approached the backbone of the conserved adenosine nucleotide at the 3'-end directly adjacent to the cleaved site in the viral DNA. From studies undertaken to identify amino acids that determine substrate specificity it was concluded that E152 plays a role in the specific recognition of the A/T base pair [58]. In Models I and II, residues K186, K258, R263 and K264 are involved in hydrogen bonding interactions with the backbone phosphate groups of the viral DNA. A protein footprinting approach employed to investigate the accessibility of residues in full-length integrase mutant (F185K and C280S) revealed that the residues K186, K258, R263 and K264 are either involved in, or shielded by, DNA binding [68]. A hydrogen bonding interaction was observed in the models between S230 and the phosphate group of Cyt17 from the lower strand. An earlier study had revealed formation of different cross-linked products between S230C IN mutant and viral DNA. The authors of this latter study concluded that conformational changes accompanying terminal cleavage involve repositioning the C-terminal domain in the vicinity of residue S230, thereby revealing the importance of this residue in IN–DNA interactions [13]. Trans interaction between residue E246 and the upper strand of viral DNA was observed in all the three models. However, the position on the viral DNA where this interaction occurs corresponded to Ade11 in Models I and II, and Thy18 in Model III. Experiments have shown various degrees of cross-linking with a range of nucleotides including the two mentioned above, with the highest level of cross-linking observed with Ade5 from the lower strand [13]. The two distal active sites of the core domains are not used for catalysis according to our models. A somewhat, similar 'crisscrossed' architecture has recently been found in the complex of transposase Tn5 with DNA [69]. We noted a remarkable similarity between the active site region of Tn5 transposase complexed with Tn5

transposon DNA [70] and the active site regions in Models I and II.

Comparison of the structural models with each other

Several models have been proposed recently for the interaction of viral DNA with a multimer of IN [13, 27, 28, 39–41]. A question we therefore wanted to answer was: How do our models agree with, or differ from, the other reported models? For this, the coordinates of the models, henceforth referred to as Model A [39], Model B [28], Model C [13] and Model D [40] were obtained from the originating authors or downloaded from the Protein Data Bank [44]. Each was compared with the others as well as with the reported experimental findings. These comparisons are summarized in Table 1. The salient features of our models are highlighted in the following.

Models I and II are very similar. Differences in the N-terminal domain structure did not result in differences in key interactions between IN residues and the viral DNA observed experimentally. The cross-linking between Ade(–2) from the lower strand and residues 139–152, as well as of Cyt(–1) with W61, Q62, L63 [12] was not directly observed in Models I and II. However, with slight unwinding of the DNA double strand, these interactions would become possible. It is well known that interaction of DNA with a protein often leads to a deformation of the DNA. These could be bending, and/or unwinding of the strands, breaking of the Watson–Crick base pairs or other conformational changes. In fact, it is quite unlikely that the viral DNA interacting with IN would be in pure A- or B-form. Therefore partial unwinding of the strands can be easily hypothesized. (Owing to absence of sufficient experimental evidence for any details of such changes, we elected, however, not to model some more or less arbitrary non-canonical conformation for the viral DNA but to choose the potentially less accurate but less biased canonical A- or B-DNA structures for our models.)

In Model II, two Mg²⁺ ions were placed in the catalytic triad between the carboxyl groups of residues D64 and E152 such that the distance between these two metal ions was approximately 4.0 Å. Although the question regarding the number of metal ions in the active site has not been resolved, both biochemical and structural studies suggest that a model in which the IN active site

binds two metals is plausible, particularly in the presence of DNA complexed with the protein [60, 71–73]. The placement of these metal ions was based on the crystal structures of the core domain of ASV-IN complexed with zinc and cadmium cofactors, respectively (1VSH.pdb and 1VSJ.pdb) [59]. During the first 10 ps of the MD simulation the two metal ions moved apart, with the maximum distance of separation between them being 7.0 Å. Subsequently after 10 ps of MD, the two metal ions reverted back closer to their original positions such that the distance between them was ~ 4.0 Å. In the final refined version of Model II, one of the metal ions coordinated with the carboxyl groups of D64 and D116, while the second metal ion was coordinated with the carboxyl groups of all three residues D64, D116 and E152. The metal was also close enough to the phosphate group of adenosine at the 3'-processed end to possibly interact with it.

Model III is somewhat more different from Models I and II than they are from each other. Many of the interactions observed in Models I and II are not seen in Model III. This is because of the shape and size of the canonical A-form of DNA used. Additionally, in this model no attempt was made to dimerize the C-terminal domain, as had been done for Models I and II. One sees that, if the viral DNA end is docked on the IN surface so as to complement the electrostatic potential maps and to satisfy the experimentally observed IN-DNA cross-linking, it would not be possible to satisfy the constraints imposed by the experimental results (Table I). From this it appears that a bend at residue T210 occurs to facilitate dimerization of the C-terminal domains. Also, critical cross-linking between K159 and Adel at the 3'-processed end is not seen in this model. Thus Models I and II, having the B-form of DNA and dimerized C-terminal domains, were generally found to be more consistent with the experimental results.

In none of the Models A–D are all 288 residues of the full-length enzyme present, nor are there any metal ions. As far as the interactions amongst the domains and monomers are concerned, Model A comes closest to our models. However, Model A consists of an octamer of IN. It remains unresolved whether an IN octamer is required to mediate concerted integration of both viral DNA ends, since the experimental evidence is insufficient at this time. For establishing any of the

well-defined interactions between the eight copies of the enzyme, one would need to construct an octameric model. Although the large-scale multimerization of the enzyme may be crucial for IN functioning *in vivo*, our focus on modeling was to develop inhibitors targeting the catalytically active site. Therefore it was important to model the binding of viral DNA to the catalytic active site as accurately as possible. It was structurally less important whether the model took the form of a single or double tetramer. In Model B, the closest distance between two catalytic triads is ~ 40 Å. This is much larger than the five base pair spacing of the sites of insertion of the two viral DNA ends into target DNA. Model C satisfies the five base pair spacing, however the catalytic triads are situated at a distance of ~ 15 Å from the 3'-end of the LTR, and the role of the catalytic triad in the ST reaction is difficult to understand from this model. In order to compare Model D with our models, we calculated rotation correlation time for each of our modeled dimers and tetramers from Models I to III, in a similar approach as reported by Alexei et al. [40] using the program HYDROPRO version 5a [74]. For our modeled dimers the rotation correlation time was found to be in the range between 54.4 and 66.4 ns, while the time for the tetramers was between 112.8 and 209.2 ns. These values are somewhat higher than those reported experimentally. However, one has to note that in the study reporting Model D, significant uncertainty pertains to the original experimentation [75] on which Alexei et al. based their computational study. Four different rotation correlation times had been obtained for IN, depending on the experimental conditions. These were 20, 40, 80 ns and one between 60 and 100 ns. These values had been *assigned* by the investigators to a monomer, dimer, trimer and tetramer, respectively, based on various factors. The true error limits of the experimentation were considerable, and our results appear to be compatible with the observed rotation correlation times.

From a general point of view it is evident that all models presented are not the only possible solutions. Other inter-domain arrangements forming monomers and inter-monomer arrangements forming dimers and tetramers or octamers are also conceivable. The fact that all models published so far deviate substantially from each other, if they are not outright in conflict with each other, attests to the

degree of uncertainty still inherent in the field. Very recently, a model of the catalytic core domain of IN complexed with viral DNA was reported wherein the IN–DNA interactions were predicted by Fast Molecular docking [76]. The interactions observed in this model are compatible with those observed in our models as far as the core domain interactions with viral DNA LTRs are concerned. This supports the validity of our approach.

We have modeled the full-length IN tetramer paying particular attention to obtaining a structure of the active site complexed with viral DNA that would truly allow structure-based drug design starting from this template. At the same time, our models are allowing us to predict as yet unreported residues from all three IN domains that may be able to cross-link with DNA. These interactions are currently being verified experimentally, helping us explore additional possibilities to inhibit the enzyme.

Docking

The aim of docking studies was to determine possible binding modes of known DKA inhibitors of IN in order to provide information for designing new receptor-based IN inhibitors. Docking was performed on a series of diketo-acids, compounds **1–31** (Table 2), using the program Glide [61], which is part of the First Discovery suite. The rationale for selection of this subset of DKA IN inhibitors was their selective inhibition of the ST reaction [37, 38, 42]. As the enzyme–DNA models generated represent the stage after 3'-P but before ST, we deemed these selective ST inhibitors to be ideal ligands for the docking experiments with our modeled complex. Coordinates of the full-length dimer complexed with U5 LTR extracted from structural Model II served as the receptor for docking studies. Models I and II are generally very similar to each other, particularly in the region of the computationally critical catalytic triad. However, Model II having two metal ions coordinating the three catalytic residues was thought to be more relevant in fulfilling requirements for binding of DKA inhibitors [60]. Model III was not considered for these docking studies because key protein–DNA interactions, including certain experimentally observed ones, could not be seen in it. This was particularly the case for crucial interactions near the catalytic triad between K159 and Ade 1 at the 3'-processed end. The compounds **1–31** were

subjected to flexible docking using pre-computed grid files. For each compound, 100 top-scored poses based on Glide scores [64] were saved and analyzed. The top ranked poses for a subset of compounds are depicted in Figure 7 and the Glide score for these poses is included in Table 2.

Although a large search region was used for the docking study, the ligands docked preferentially close to the active site and in similar geometries. As shown in Figure 7a, the best poses for compounds **1–6** were found to superimpose, with aryl ring and DKA functionalities overlapping across the entire series. The keto-enol functionalites were located in positions capable of chelating one of the two active site Mg^{2+} ions. For all six compounds, one hydrogen-bonding interaction was observed between the enolic hydroxyl group and the backbone carbonyl of D116, a second one between the carboxylic acid carbonyl and the backbone –NH of G118. Compounds **1** and **2** form an additional hydrogen bond with the side-chain of N120 through the carboxylic functionality. In case of compound **4**, a hydrogen bond is found between the inhibitor's benzoyl carbonyl group of **4** and the backbone amide of H67. Compound **6** also forms two additional hydrogen bonds, one between the carboxylic group and N120 and another one between the benzyloxy group and Ade1 at the 3'-processed end of the U5-LTR. The benzoylamino substituent of compound **4**, the phenoxy group of compound **5** and the benzyloxy substituents of compounds **3** and **6** are directed towards the double stranded viral DNA, showing an intercalation-like interaction with the unpaired cytosine residue at the lower strand of U5-LTR.

The binding modes of the bis-diketo acid inhibitors, **7**, **8** and **10**, are slightly different from the previous set (Figure 7b). Though metal chelation seems to be mediated by the diketo functionality for all three compounds, the orientation of these molecules in the active site and the hydrogen bonding interactions are different. In the case of **7**, one diketo acid moiety forms hydrogen bonds with residues D116 and N120, while the second diketo acid functionality extends to the viral DNA end. Hydrogen bonding interactions are observed with H67 as well as with the phosphate backbone of Cyt2 from the upper strand. In compound **8**, the aromatic ring acts as an anchor, with one set of DKA functionality forming interactions with G118 and S119 while a second one is hydrogen-bonded to

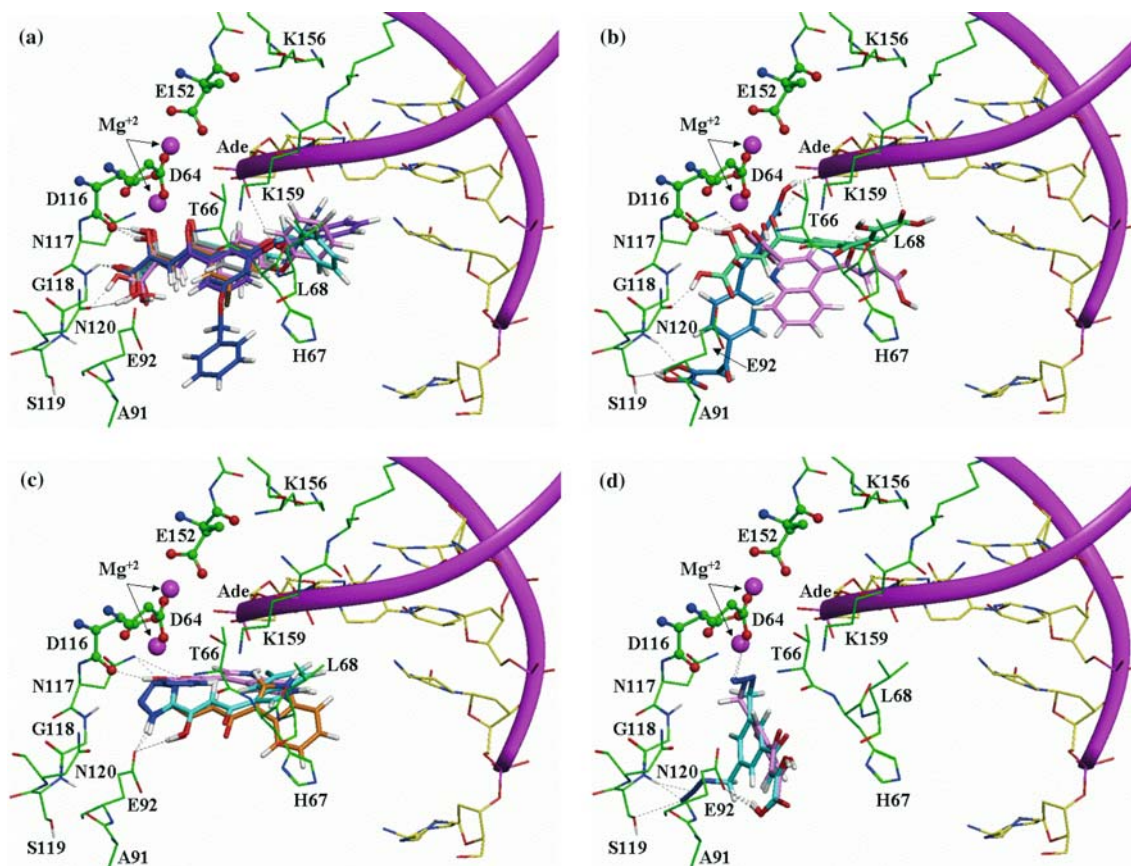


Figure 7. Docking models of diketo-acid HIV-1 integrase inhibitors (Table 2) obtained using Glide. Hydrogen bonds are represented as dashed lines in black. For clarity reasons, only important residues and atoms are shown. Color codes used: nitrogen – blue; oxygen – red; carbon atoms from protein – green; from DNA – yellow; and in (a) Compound 1-gray; 2-orange; 3-aquamarine; 4-purple; 5-pink; 6-blue. (b) Compound 7- aquamarine; 8- blue; 10-pink. (c) Compound 14-orange; 15-pink; 16-cyan. (d) Compound 22-pink and 28-cyan.

the 3'-hydroxy group of Ade1 of the upper DNA strand. Compound 10 shows hydrogen-bonding interactions with D116 and N120. For compounds 14–17, which contain a tetrazole group in place of carboxylic acid functionality, the chelation with the metal ion was mediated by the tetrazole group and not the keto-enol functionality (Figure 7c). Quantum chemical calculations being performed at the *ab initio* level to study and explain these differences in interactions with the metal ion(s) will be published elsewhere.

We also compared the orientation obtained by the docking approach for compound 15 (5-CITEP) with the crystal structure of 5-CITEP in the core domain of IN [45]. Though the binding location is the same, the orientation of 5-CITEP in the two structures is slightly different. Such a difference is to be expected since the structure used for docking contains viral DNA, which is absent in the crystal

structure. Differences of this kind have been attributed previously to crystallographic packing [77, 78].

For compounds 22–30 containing an azido group, two possible orientations were observed. In the top-ranked poses, chelation with the metal ion was mediated by the azido group (Figure 7d). However in about 30 of the 100 best poses, the keto-enol moiety was found to chelate the metal ion.

Because docking scores are known to correlate poorly with enzyme inhibitory activity [79], the lack of strong correlation in this case was not unexpected. However, the binding modes obtained for the 31 DKA inhibitors examined here indicated that compounds capable of chelating the metal ion as well as forming interactions both with the residues of IN and the viral DNA are better IN inhibitors. This is particularly true for selective inhibitors of the ST reaction. Thus we conclude from our docking

studies that the putative roles of substituents on the central aromatic moiety is to position and anchor the entire molecule appropriately in the binding site of the complicated biomacromolecular assembly comprising the IN–DNA complex *in vivo* just prior to strand transfer. Using a similar approach we have addressed the question of metal dependent enzyme inhibitory activity exhibited by certain DKA inhibitors elsewhere [42].

Based on the docking study described above, tentative conclusions can be reached as to the requirements for a compound to inhibit the strand transfer reaction of IN. Inhibitors should possess a central core region, preferably an aromatic moiety that carries substituents capable of hydrogen bonding with one of the residues E92, D116, N117, G118 and N120. Another set of substituents may interact with nucleotides from the viral DNA or chelate the metal ion present in the catalytic site. Further drug design results based on the models presented here will be published in the future. Coordinates of the models can be obtained from the authors upon request.

Conclusions

We have presented structural models of the full-length HIV-1 IN complexed with models of viral and human DNA, that were based on all available experimental information. Using these structural models, docking studies of 31 DKA inhibitors of IN were conducted in an attempt to understand the binding modes of the inhibitors and the structural features necessary for inhibition. Chelation of the metal ion(s), and interaction both with residues in the active site of IN and with the nucleotides at the 3'-end of the LTR were characteristic features exhibited by highly potent DKAs. By binding in this position, the inhibitors may prevent the exposure of the nucleophilic hydroxyl group at the 3'-processed end of the viral LTR to human DNA. This would explain the selective potency in ST inhibition shown by these inhibitors. From the docking analysis of the azido containing compounds, it appeared that the azido group could function as a replacement for the keto-enol functionality present in DKAs. Further studies to test, and possibly exploit, this hypothesis are underway. These models are currently being used for structure-based design of new IN inhibi-

tors and also to improve our understanding of the mechanism of the 3'-P and ST reactions.

Acknowledgements

We thank Robert Craigie, Kui Gao, and Alexei Podtelezchnikov for providing us with the coordinates of their respective models. We thank Peter P. Roller, Yves Pommier, Vinay Pathak, and members of their groups for many useful discussions.

References

1. Autran, B., Carcelain, G., Li, T.S., Blanc, C., Mathez, D., Tubiana, R., Katlama, C., Debré, P. and Leibowitch, J., *Science*, 277 (1997) 112.
2. Palella, F.J., Delaney, K.M., Moorman, A.C., Loveless, M.O., Fuhrer, J., Satten, G.A., Aschman, D.J. and Holmberg, S.D., *New Engl. J. Med.*, 338 (1998) 853.
3. Richman, D.D., *Nature*, 410 (2001) 995.
4. Asante-Appiah, E. and Skalka, A.M., *Antiviral Res.*, 36 (1997) 139.
5. Craigie, R., *J. Biol. Chem.*, 276 (2001) 23213.
6. Cara, A., Guarnaccia, F., Reitz, M.S. Jr., Gallo, R.C. and Lori, F., *Virology*, 208 (1995) 242.
7. Esposito, D. and Craigie, R., *Adv. Virus Res.*, 52 (1999) 319.
8. Roe, T., Chow, S.A. and Brown, P.O., *J. Virol.*, 71 (1997) 1334.
9. Marchand, C., Neamati, N. and Pommier, Y., *Methods Enzymol.*, 340 (2001) 624.
10. Engelman, A., Hickman, A.B. and Craigie, R., *J. Virol.*, 68 (1994) 5911.
11. Engelman, A., Mizuuchi, K. and Craigie, R., *Cell*, 67 (1991) 1211.
12. Esposito, D. and Craigie, R., *EMBO J.*, 17 (1998) 5832.
13. Gao, K., Butler, S.L. and Bushman, F.D., *EMBO J.*, 20 (2001) 3565.
14. Heuer, T.S. and Brown, P.O., *Biochemistry*, 36 (1997) 10655.
15. Jenkins, T.M., Esposito, D., Engelman, A. and Craigie, R., *EMBO J.*, 16 (1997) 6849.
16. Zheng, R., Jenkins, T.M. and Craigie, R., *Proc. Natl. Acad. Sci. U. S. A.*, 93 (1996) 13659.
17. Lee, S.P., Xiao, J.M., Knutson, J.R., Lewis, M.S. and Han, M.K., *Biochemistry*, 36 (1997) 173.
18. Gerton, J.L. and Brown, P.O., *J. Biol. Chem.*, 272 (1997) 25809.
19. Eijkelenboom, A.P., Lutzke, R.A., Boelens, R., Plasterk, R.H., Kaptein, R. and Hard, K., *Nat. Struct. Biol.*, 2 (1995) 807.
20. Lodi, P.J., Ernst, J.A., Kuszewski, J., Hickman, A.B., Engelman, A., Craigie, R., Clore, G.M. and Gronenborn, A.M., *Biochemistry*, 34 (1995) 9826.
21. Hindmarsh, P. and Leis, J., *Microbiol. Mol. Biol. Rev.*, 63 (1999) 836.
22. Dyda, F., Hickman, A.B., Jenkins, T.M., Engelman, A., Craigie, R. and Davies, D.R., *Science*, 266 (1994) 1981.

23. Cai, M., Zheng, R., Caffrey, M., Craigie, R., Clore, G.M. and Gronenborn, A.M., *Nat. Struct. Biol.*, 4 (1997) 567.
24. Maignan, S., Guilloteau, J.-P., Zhou-Liu, Q., Clément-Mella, C. and Mikol, V., *J. Mol. Biol.*, 282 (1998) 359.
25. Goldgur, Y., Dyda, F., Hickman, A.B., Jenkins, T.M., Craigie, R. and Davies, D.R., *Proc. Natl. Acad. Sci. U. S. A.*, 95 (1998) 9150.
26. Eijkelenboom, A.P., Sprangers, R., Hard, K., Puras-Lutzke, R.A., Plasterk, R.H., Boelens, R. and Kaptein, R., *Proteins*, 36 (1999) 556.
27. Chen, J.C., Krucinski, J., Miercke, L.J., Finer-Moore, J.S., Tang, A.H., Leavitt, A.D. and Stroud, R.M., *Proc. Natl. Acad. Sci. U. S. A.*, 97 (2000) 8233.
28. Wang, J.-Y., Ling, H., Yang, W. and Craigie, R., *EMBO J.*, 20 (2001) 7333.
29. Neamati, N., *Exp. Opin. Ther. Patents*, 12 (2002) 709.
30. Hazuda, D.J. and Young, S.D. Inhibitors of human immunodeficiency virus integration. In Clercq, E. (Ed), *Advances in Antiviral Drug Design*, Elsevier Science, Amsterdam, Vol. 4, pp. 63–77.
31. Nicklaus, M.C., Pommier, Y. and Mazumder, A., 210th American Chemical Society National Meeting, Chicago, IL, Aug. 20–25, 1995.
32. Mazumder, A., Wang, S.M., Neamati, N., Nicklaus, M.C., Sunder, S., Chen, J., Milne, G.W.A., Rice, W.G., Burke, T.R. Jr. and Pommier, Y., *J. Med. Chem.*, 39 (1996) 2472.
33. Hong, H.X., Neamati, N., Wang, S.M., Nicklaus, M.C., Mazumder, A., Zhao, H., Burke, T.R. Jr., Pommier, Y. and Milne, G.W.A., *J. Med. Chem.*, 40 (1997) 930.
34. Neamati, N., Hong, H.X., Mazumder, A., Wang, S.M., Sunder, S., Nicklaus, M.C., Milne, G.W.A., Proksa, B. and Pommier, Y., *J. Med. Chem.*, 40 (1997) 942.
35. Nicklaus, M.C., Neamati, N., Hong, H.X., Mazumder, A., Sunder, S., Chen, J., Milne, G.W.A. and Pommier, Y., *J. Med. Chem.*, 40 (1997) 920.
36. Neamati, N., Lin, Z.W., Karki, R.G., Orr, A., Cowansage, M., Strumberg, D., Pais, G.C.G., Voigt, J.H., Nicklaus, M.C., Winslow, H.E., Zhao, H., Turpin, J.A., Yi, J.Z., Skalka, A.M., Burke, T.R. Jr. and Pommier, Y., *J. Med. Chem.*, 45 (2002) 5661.
37. Hazuda, D.J., Felock, P., Witmer, M., Wolfe, A., Stillmock, K., Grobler, J.A., Espeseth, A., Gabryelski, L., Schleif, W., Blau, C. and Miller, M.D., *Science*, 287 (2000) 646.
38. Pais, G.C.G., Zhang, X., Marchand, C., Neamati, N., Cowansage, K., Svarovskaia, E.S., Pathak, V.K., Tang, Y., Nicklaus, M.C., Pommier, Y. and Burke, T.R. Jr., *J. Med. Chem.*, 45 (2002) 3184.
39. Heuer, T.S. and Brown, P.O., *Biochemistry*, 37 (1998) 6667.
40. Podtelezhnikov, A.A., Gao, K., Bushman, F.D. and McCammon, J.A., *Biopolymers*, 68 (2003) 110.
41. De Luca, L., Pedretti, A., Vistoli, G., Barreca, M.L., Villa, L., Monforte, P. and Chimirri, A., *Biochem. Biophys. Res. Commun.*, 310 (2003) 1083.
42. Marchand, C., Johnson, A.A., Karki, R.G., Pais, G.C.G., Zhang, X., Cowansage, K., Patel, T.A., Nicklaus, M.C., Burke, T.R. Jr. and Pommier, Y., *Mol. Pharmacol.*, 64 (2003) 600.
43. Holler, T.P., Foltin, S.K., Ye, Q.Z. and Hupe, D.J., *Gene*, 136 (1993) 323.
44. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., *Nucl. Acids. Res.*, 8 (2000) 235.
45. Goldgur, Y., Craigie, R., Cohen, G.H., Fujiwara, T., Yoshinaga, T., Fujishita, T., Sugimoto, H., Endo, T., Murai, H. and Davies, D.R., *Proc. Natl. Acad. Sci. U. S. A.*, 96 (1999) 13040.
46. Katzman, M. and Katz, R.A., *Adv. Virus Res.*, 52 (1999) 371.
47. Morgan, A.L. and Katzman, M., *J. Gen. Virol.*, 81 (2000) 839.
48. Katz, R.A., Gravuer, K. and Skalka, A.M., *J. Biol. Chem.*, 273 (1998) 24190.
49. SYBYL, Version 6.8, Tripos Inc., St. Louis, MO, 2001 (<http://www.tripos.com>).
50. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., *J. Comput. Chem.*, 4 (1983) 187.
51. MacKerell, A.D. Jr., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kucera, K., Lau, F.T.K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher, W.E., Roux, B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Wiorcikiewicz-Kuczera, J., Yin, D. and Karplus, M., *J. Phys. Chem. B.*, 102 (1998) 3586.
52. Foloppe, N. and MacKerell, A.D. Jr., *J. Comput. Chem.*, 21 (2000) 86.
53. Gaussian 98, Revision A.9, Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Zakrzewski, V.G., Montgomery, J.A. Jr., Stratmann, R.E., Burant, J.C., Dapprich, S., Millam, J.M., Daniels, A.D., Kudin, K.N., Strain, M.C., Farkas, O., Tomasi, J., Barone, V., Cossi, M., Cammi, R., Mennucci, B., Pomelli, C., Adamo, C., Clifford, S., Ochterski, J., Petersson, G.A., Ayala, P.Y., Cui, Q., Morokuma, K., Malick, D.K., Rabuck, A.D., Raghavachari, K., Foresman, J.B., Cioslowski, J., Ortiz, J.V., Baboul, A.G., Stefanov, B.B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Gomperts, R., Martin, L., Fox, D.J., Keith, T., Al-Laham, M.A., Peng, C.Y., Nanayakkara, A., Challacombe, M., Gill, P.M.W., Johnson, B., Chen, W., Wong, M.W., Andres, J.L., Gonzalez, C., Head-Gordon, M., Replogle, E.S., and Pople, J.A. Gaussian, Inc., Pittsburgh PA, 1998 (<http://www.gaussian.com>).
54. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. and Klein, M.L., *J. Chem. Phys.*, 79 (1983) 926.
55. This study utilized the high-performance computational capabilities of the SGI Origin 2000 system and the Biowulf/LoBoS3 cluster at the National Institutes of Health, Bethesda, MD.
56. Computational results obtained using the program Insight II, Accelrys Inc., San Diego, CA, 2000 (<http://www.accelrys.com/>).
57. Nicholls, A., Sharp, K.A. and Honig, B., *Proteins*, 11 (1991) 281.
58. Gerton, J.L., Ohgi, S., Olsen, M., DeRisi, J. and Brown, P.O., *J. Virol.*, 72 (1998) 5046.
59. Bujacz, G., Alexandratos, J., Wlodawer, A., Merkel, G., Andrade, M., Katz, R.A. and Skalka, A.M., *J. Biol. Chem.*, 272 (1997) 18161.
60. Grobler, J.A., Stillmock, K., Hu, B.H., Witmer, M., Felock, P., Espeseth, A.S., Wolfe, A., Egbertson, M., Bourgeois, M., Melamed, J., Wai, J.S., Young, S., Vacca, J. and Hazuda, D.J., *Proc. Natl. Acad. Sci. USA*, 99 (2002) 6661.
61. First Discovery 2.0, Schrödinger, Inc., Portland, OR, 2002 (<http://www.schrodinger.com>).

62. Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., Repasky, M.P., Knoll, E.H., Shelly, M., Perry, J.K., Shaw, D.E., Francis, P. and Shenkin, P.S., *J. Med. Chem.*, 47 (2004) 1739.
63. Jorgensen, W.L., Maxwell, D.S. and Tirado-Rives, J., *J. Am. Chem. Soc.*, 118 (1996) 11225.
64. First Discovery technical notes for version 2.0, Schrodinger, Inc. Portland OR, 2002 (http://www.schrodinger.com/docs/impact2.0/pdf/tech_notes/tech_notes.pdf).
65. Miller, M.D., Farnet, C.M. and Bushman, F.D., *J. Virol.*, 71 (1997) 5382.
66. Roth, M.J., Schwartzberg, P.L. and Goff, S.P., *Cell*, 58 (1989) 47.
67. Katzman, M. and Sudol, M., *J. Virol.*, 69 (1995) 5687.
68. Dirac, A.M.G. and Kjems, J., *Eur. J. Biochem.*, 268 (2001) 743.
69. Rice, P.A. and Baker, T.A., *Nat. Struct. Biol.*, 8 (2001) 302.
70. Davies, D.R., Goryshin, I.Y., Reznikoff, W.S. and Rayment, I., *Science*, 289 (2000) 77.
71. Davies, J.F.I., Hostomska, Z., Hostomsky, Z., Jordan, S.R. and Matthews, D.A., *Science*, 252 (1991) 88.
72. Steitz, T.A. and Steitz, J.A., *Proc. Natl. Acad. Sci. USA*, 90 (1993) 6498.
73. Brautigam, C.A. and Steitz, T.A., *Curr. Opin. Struct. Biol.*, 8 (1998) 54.
74. Garcia de la Torre, J., Huertas, M.L. and Carrasco, B., *Biophys. J.*, 78 (2000) 719.
75. Deprez, E., Tauc, P., Leh, H., Mouscadet, J.-F., Auclair, C. and Brochon, J.-C., *Biochemistry*, 39 (2000) 9275.
76. Adesokan, A.A., Roberts, V.A., Lee, K.W., Lins, R.D. and Briggs, J.M., *J. Med. Chem.*, 47 (2004) 821.
77. Sotriffer, C.A., Ni, H. and McCammon, J.A., *J. Med. Chem.*, 43 (2000) 4109.
78. Sotriffer, C.A., Ni, H. and McCammon, J.A., *J. Am. Chem. Soc.*, 122 (2000) 6136.
79. Tame, J.R.H., *J. Comput.-Aided Mol. Des.*, 13 (1999) 99.
80. Zhang, X., Pais, G.C.G., Svarovskaia, E.S., Marchand, C., Johnson, A.A., Karki, R.G., Nicklaus, M.C., Pathak, V.K., Pommier, Y. and Burke, T.R. Jr., *Bioorg. Med. Chem. Lett.*, 13 (2003) 1215.