JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

pubs.acs.org/JACS

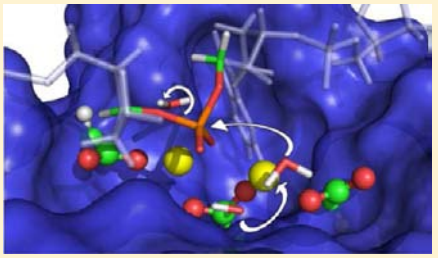
The Catalytic Mechanism of HIV-1 Integrase for DNA 3′-End Processing Established by QM/MM Calculations

António J. M. Ribeiro, Maria J. Ramos, and Pedro A. Fernandes*

REQUIMTE, Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre s/n, 4169-007 Porto, Portugal

Supporting Information

ABSTRACT: The development of HIV-1 integrase (INT) inhibitors has been hampered by incomplete structural and mechanistic information. Despite the efforts made to overcome these limitations, only one compound has been approved for clinical use so far. In this work, we have used all experimental information available for INT and similar enzymes, to build a model of the holo-integrase:DNA complex that includes an entire central core domain, a ssDNA GCAGT substrate, and two magnesium ions. Subsequently, we used a large array of computational techniques, which included molecular dynamics, thermodynamic integration, and high-level quantum mechanics/molecular mechanics (QM/MM) calculations to study the possible pathways for the mechanism of 3′ end processing catalyzed by INT. We found that the only viable mechanism to hydrolyze the DNA substrate is a nucleophilic attack of an active site water molecule to the phosphorus atom of the scissile phosphoester bond, with the attacking water being simultaneously deprotonated by an Mg^{2+}-bound hydroxide ion. The unstable leaving oxoanion is protonated by an Mg^{2+}-bound water molecule within the same elementary reaction step. This reaction has an activation free energy of 15.4 kcal/mol, well within the limits imposed by the experimental turnover. This work significantly improves the fundamental knowledge on the integrase chemistry. It can also contribute to the discovery of leads against HIV-1 infection as it provides, for the first time, accurate transition states structures that can be successfully used as templates for high-throughput screening of new INT inhibitors.



INTRODUCTION

The human immunodeficiency virus (HIV) has three enzymes that are fundamental for its life cycle: protease (PR), reverse transcriptase (RT), and integrase (INT). The current treatment for HIV-1 infected patients consists in a cocktail of PR and RT inhibitors. The cocktail slows down the progression of the disease but the unavoidable emergence of resistance finally renders them completely inefficient. This and their harsh secondary effects are fueling continuous efforts for the discovery of new inhibitors with clinical applicability. As INT inhibition has an effective role on slowing down the progression of AIDS, the last 20 years have witnessed a large global effort for discovery and development of new drugs that target INT. Despite the initial expectations, only one INT inhibitor (Raltegravir) has been developed and approved by the FDA so far,[1] in part due to the lack of structural and mechanistic information. This situation sharply contrasts with the situation of PR and RT, where structural and mechanistic information is abundant and resulted in a total of 11 and 19 drugs already in clinical use.

INT is thought to be active as a homotetramer.[2] Each monomer has three domains, constituted by 288 residues in total. The N-terminal domain includes residues 1−50; the central-core domain includes residues 51−212; and finally, the C-terminal domain includes residues 213−288.[2,3] The three domains are required for INT reactions.[4] INT catalyzes two essential steps in the replication cycle of HIV, named "3′ end processing" and "strand transfer". In the 3′ end processing reaction, INT removes two nucleotides of both 3′ ends of the viral DNA. This reaction happens in the cytoplasm, where INT only has access to the viral DNA. In the second reaction, INT inserts the processed viral DNA into the host DNA.[3,5−7] Additionally, INT (together with several host proteins) is involved in the translocation of viral DNA from the cytoplasm to the nucleus of the host cell, after processing the 3′ end of the viral DNA and before catalyzing the strand transfer reaction.[8] Both reactions happen in the same catalytic center and involve similar breaking and formation of phosphodiester bonds. Mutagenesis studies show that the catalytic residues required for the reactions are Asp64, Asp116, and Glu152,[9] together with, most probably, two magnesium ions. Hypotheses about the catalytic mechanisms of INT may be derived from indirect comparison with enzymes that catalyze similar reactions, such as DNA polymerase I[10] or ribonuclease H.[11] However, an atomic-level catalytic mechanism for HIV-1 INT rooted in solid scientific evidence is still missing. One of the basic problems is that there is no crystallographic structure of INT bound to its DNA substrate. Earlier computational studies on this reaction provided important insights on the problem.[12,13] However, the modeling of the substrate as dimethylphophate, the inclusion of only one magnesium ion in the active site, the very high energy barriers obtained (above the experimental limits imposed by the enzyme turnover), and the inability for

Received: May 22, 2012
Published: July 13, 2012

explaining the role of the fundamental Glu152 residue in the catalysis show that further studies on this system are needed to explain the catalytic mechanism, the mutagenic results, and the kinetics of the enzymatic process.

Further limitations on the available crystallographic data have impaired the development of more complete enzyme models. There is still no crystallographic structure of a complete INT monomer, although models have been constructed based on docking procedures[14] or on the modeling of structures with overlapping domains.[15−19] Furthermore, the precise number of magnesium ions in the active center (either one or two), is still a matter of debate, even though all evidence point out to a two-magnesium chemistry. A tetrameric structure of the related prototype foamy virus (PFV) INT bound to the substrate DNA has been reported very recently.[20] This groundbreaking result allows now for a much more reliable modeling of the substrate binding into HIV-1 INT. Despite the fact that we have built the model of HIV-1 INT:DNA discussed in this work before the PFV INT structure became available, the comparison of the two structures (PFV INT:DNA crystallized and HIV-1 INT:DNA modeled, independently) shows remarkably similar active sites.

In this work, we establish the reaction mechanism of 3′ processing catalyzed by INT, with atomic detail. For that purpose, we had to use an array of techniques that included molecular modeling, docking, molecular dynamics, thermodynamic integration, and high level quantum mechanics/molecular mechanics (QM/MM) calculations with the MPWB1K density functional and large basis sets at the QM level and the Amber force field at the MM level. These computational methods have been widely used in the past to describe the catalytic mechanism of many enzymes.[21−24] We have considered many mechanistic hypotheses and present here the three most viable. All are based on a two-magnesium chemistry. We used the INT catalytic domain as framework; the substrate was modeled as a pentanucleotide and the solvent was also included as a dielectric continuum. The three mechanisms differ mostly in how the nucleophile (a water molecule) is deprotonated during the attack to the phosphodiester bond.

Besides the fundamental advance in the understanding of the chemistry of HIV-1 infection, this work also provides accurate transition states (TS) structures that can be used as templates for high-throughput screening of new INT inhibitors.

METHODS

The overall line of work in this study can be summarized in the following nine tasks:

(Task i) Modeling of the complete active site of INT, i.e., a central core domain with two magnesium ions in the catalytic center using the pdb file 1QS4;[25]

(Task ii) Molecular dynamics simulation of the enzyme model of INT obtained in Task (i) with explicit solvent, to relax the enzyme;

(Task iii) Docking of the DNA substrate, a pentanucleotide, into the active site;

(Task iv) Molecular dynamics simulation of the complex INT:DNA with explicit solvent to relax the system;

(Task v) Free energy calculations, using thermodynamic integration, to estimate the free energy involved in exchanging a water molecule for a hydroxide ion inside the active center and in bulk solution;

(Task vi) QM/MM calculation of the potential energy surface (PES) for each of the possible mechanisms for 3′-end processing, to be able to establish the correct pathway for the INT:DNA system;

(Task vii) Pure QM calculations with implicit water solvation on a subregion of the INT:DNA complex, to calculate the contribution of the solvent;

(Task viii) Molecular dynamics simulations with explicit solvent, starting from the structures of the reactants, to generate a set of different conformations of the reactant state, in order to evaluate the dependence of the energetics of the mechanism on the specific enzyme conformation used in the calculations;

(Task ix) Calculation of the PES of the reaction mechanism starting from several initial enzyme conformations, to measure the effect of the enzyme conformational spread on the activation energies of the catalytic cycle.

The zero point energy of the systems was not included. It gives a small contribution to activation and reaction energies (1−2 kcal/mol). It would have been desirable to include it but it was too much of a computational effort as we have worked with a model of over 2500 atoms. Moreover, the zero point energy of the system never affects the choice between alternative mechanistic hypotheses. As the differences in energy between mechanistic hypotheses are usually quite large, the discrimination between different pathways is much more robust to methodological approximations than the calculation of absolute activation and reaction energies.

The text below gives full detail of all the calculations, as well as of the molecular system used in each one of these steps. The nine tasks will now be detailed, one by one.

Modeling of the INT Central-Core Domain with Two Magnesium Ions (*Task i*). The central core domain of INT was taken from the PDB structure 1QS4[25](chain A), which contains the central core domain of INT, one magnesium ion and the inhibitor 5CITEP, at 2.10 Å resolution. This structure was the only one with an inhibitor bound in the active site. A very easy and basic modeling was done to introduce the nondefined four residues 141−144, through superposition with the 1BIS[26] structure (using the software PYMOL.[27]), which includes the complete central core domain at 1.95 Å resolution. The inhibitor and all water molecules were removed from the crystallographic structure, with the exception of the waters coordinated to Mg^{2+} (four water molecules). To build a two-Mg^{2+} active site, a second Mg^{2+} ion (plus four bound water molecules to complete its octahedral shell) was modeled into the catalytic center. The side chain of Glu152 was rotated in order to be closer to Asp64 and the magnesium was placed between the carboxylates of these two residues (see Figures 1 and 3). Henceforth, we will name this added magnesium as Mg^{2+}_{lg} (leaving group magnesium), and the already existing one as Mg^{2+}_{nuc} (nucleophile magnesium). The pK_a of every INT residue was then calculated using the H++ server.[28] The resulting pK_a's (Supporting Information Table SI-I) show that all residues should be considered in their physiological protonation states. The holoenzyme was relaxed with a 2 ns molecular dynamics simulation (full details on the MD simulations can be found in the section Molecular Dynamics Simulations below). Both structures, before and after the molecular dynamics simulation, were placed in Supporting Information (files holo_before_md.pdb and holo_after_md.pdb).

Molecular Dynamics Simulations (*Tasks ii, iv*, and *viii*). The AMBER 9 package[29] was used in the molecular dynamics simulations, with the ff03 force field. The simulations were carried out before (*task ii*) and after (*task iv*) docking the substrate into the active center, to relax the INT and INT:DNA structures, respectively. Explicit TIP3P[30] water molecules were employed, filling a periodic box with margins of at least 12 Å beyond every protein atom. For the unbound INT, we have added 9933 waters molecules that, together with INT, corresponded to 32178 atoms. The box size was 79 Å × 79 Å × 64 Å at the start of the simulation. For the INT:DNA model, we have added 10 524 water molecules to the complex and that, together with the INT:DNA complex, corresponded to 34 124 atoms. The box size was 77 Å × 74 Å × 75 Å at the start of the simulation. In both runs, the short-range van der Waals interactions were truncated at 10 Å and the Coulombic interactions were calculated with the PME method[31] (also with a cutoff of 10 Å in the real part of the sum). All atoms were free to move in these molecular dynamics simulations. A time step of 1 fs was used. After a warm-up dynamics of 20

dx.doi.org/10.1021/ja304601k | J. Am. Chem. Soc. 2012, 134, 13436−13447

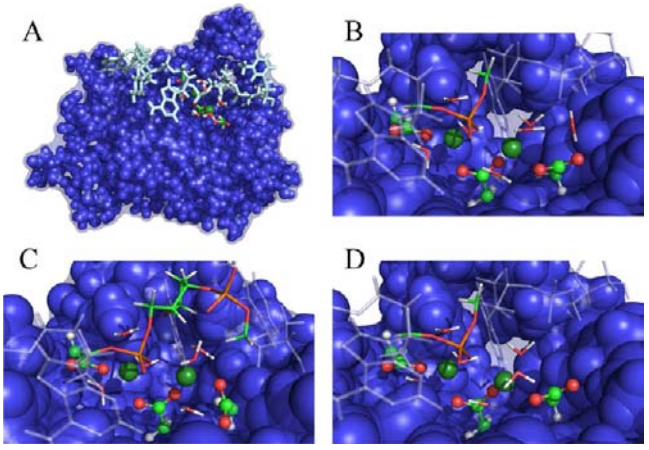


Figure 1. (A) The model used for the QM/MM calculations. The atoms of the complete central core domain are represented as blue spheres. The DNA substrate is represented in tubes and colored in blue. The magnesium ions are represented as dark green spheres. The water molecules bonded to the magnesium ions and the catalytic triad are colored by atom. (B–D) A detail of the model with the QM/MM partition used in the (B) Aspartate-Base, (C) Phosphate-Base, and (D) Hydroxide-Base mechanistic hypotheses. The QM atoms are colored by atom type and the MM atoms are colored blue.

ps (from 0 to 300 K) at constant volume, a production dynamics was run in the NPT ensemble with the Langevin thermostat and isotropic position scaling, at 300 K and 1 bar.

The five molecular dynamics simulations that were run subsequently to generate initial structures for the QM/MM calculation of the PES of the most viable catalytic mechanism (*task vii*) used the same system and options as the INT:DNA simulations (*task iv*). A different seed to generate random initial velocities was chosen for each of the five runs, and a total of 25 atoms (the 2 Mg^{2+} ions, 6 water molecules coordinated to the metals, and the 5 oxygen atoms from the catalytic residues and the substrate that also coordinate the two metals) were constrained, to keep the core reactants as similar as possible in all runs and change only the enzymatic scaffold. The simulations ran for 2 ns for unbound INT (*task ii*), for 5 ns for the INT:DNA complex (*task iv*) and for 1 ns to generate each initial structure for the subsequent QM/MM calculations (*task vii*). This time scale is sufficient to relax the INT:DNA structure, while maintaining it as close as possible to the X-ray structure. The experimental structure represents a thermal average over many different conformations coming from many molecules in the crystal. Therefore, it is more representative of the ensemble than a single structure coming from a MD simulation.

In the MD simulations, Mg^{2+} ions were treated as positive charged ions, with no covalent bonds to the enzyme residues, water molecules or substrate. Furthermore, we did not make any re-parameterization of charges for the residues around the metals. It can be hypothesized that this kind of force field parameters may eventually led to deviations in the structures, such as changes in the coordination sphere of the metals. However, no such conformation changes were observed in our MD simulations. The coordination to the metals remained the same throughout the MD simulation: the RMSD for the Mg^{2+} atoms and respective coordination spheres remained at circa 0.8 Å along all the 5ns dynamics, in relation to the first structure. (See Figure 4 in the SI) We stress that during the QM/MM calculations, the metals and their coordination spheres were included in the QM layer and treated with DFT. Being so, the polarization induced by the metals, and eventual covalent interactions were taken into account in the energetic calculations.

Docking of the Substrate (*Task iii*). The DNA substrate (5-GCAGT-3) was created as a linear chain of nucleotides in Xleap (AMBER 9)[29] and docked into the active site of INT using GOLD.[32,33] The genetic algorithm and Goldscore[34] were used as search and score algorithms. A sphere of 20 Å centered on the Mg^{2+}_{lg} was defined as the search space. The water molecules coordinated to the Mg^{2+} ions were included and allowed to rotate. The substrate was free to move, except for three deliberated knowledge-based constraints derived from experimental data on INT, as well as from similarity with other enzymes that catalyze related reactions.[10,11,35] These constraints can be summarized as follows: First, the 5′ → 3′ direction was imposed to the substrate. The direction of DNA in INT is well-known to be 5′ → 3′ and we can constrain the solutions to obey this restriction. Second, a hydrogen bond between N7 of the third base (an adenine) and K159 is experimentally shown to exist.[36] Adding this second constraint to the first almost "freezes" the translational and rotational freedom of the ligand. Third, the cleaved phosphodiester bond must lie between the two Mg^{2+} ions, in agreement with its position in all other bimetallic enzymes that catalyze similar chemical reactions.

After adding this third constraint to the other two, only very few degrees of freedom remain unknown. The remaining options do not affect in any meaningful manner the position and conformation of the reactive region of the enzyme.

One hundred poses were generated and a structure belonging to the top 10 rated results (according to Goldscore) was chosen to proceed with the work. The choice of the specific structure among the 10 best scored was based on the comparison with the structures of DNA-Polymerase I[10] and Ribonuclease H,[11] which have bound DNA substrates, an extremely similar active site and perform equivalent reactions. A detailed discussion about the similarities between Ribonuclease H, DNA-Polymerase I, and INT can be found in the Results and Discussion sections. Our model of the central core domain with the two magnesium ions and the docked substrate can be seen in Figure 3 in the Results section.

The docking protocol followed here is more a "modeling technique" than a "prediction technique" because the three geometric constraints that the Michaelis complex must obey simultaneously completely restrict the variety of poses that the substrate may take in within the binding site.

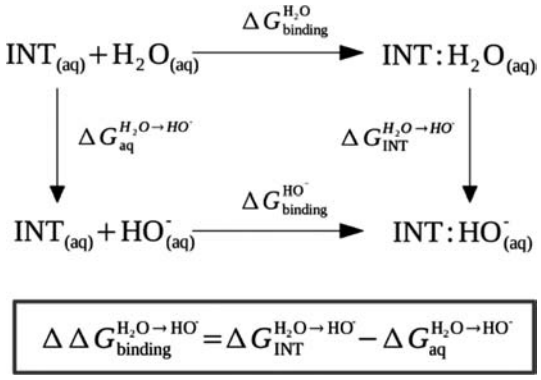
After docking the substrate, we ran MD simulations with the complex (please see section Molecular Dynamics Simulation above for details), and used the last structure of the simulation as a starting point for the QM/MM calculations.

Free Energy Calculations (*Task v*). Thermodynamic integration (TI) is one of the most powerful and accurate computational methods to calculate the difference in Gibbs energy of binding between two similar ligands of a given receptor.[37] Therefore, we have carried out TI calculations (using the software AMBER 10[38]) to calculate the free energy involved in exchanging one water molecule by a hydroxide ion at the active site. The simulations involved two transformations: transforming a hydroxide ion into a water molecule in a periodic box with bulk water and transforming an Mg^{2+}-bound water molecule into an Mg^{2+}-bound hydroxide ion at the active site of INT. Scheme 1 depicts the thermodynamic cycle used to calculate the free energy of exchange. Both transformations were done in the two directions, forward and backward.

Explicit TIP3P[30] water molecules were used, filling up a periodic box with margins of 12 Å beyond the hydroxide ion or beyond the protein, with a starting size of (67, 75, 74) Å for the INT:DNA system and (31, 31, 30) Å for the hydroxide ion in solution. The short-range van der Waals interactions were truncated at 10 Å. The Coulombic interactions were taken into account with the PME method[31] and a cutoff of 10 Å.

Each transformation (water into hydroxide and vice versa) was done in two stages, each stage with the transformation of the force field potential divided in nine steps (with even increments between 0 and 1). We ran a simulation of 400 ps for each of the nine steps. The first stage corresponded to the neutralization of the atomic charge of one hydrogen atom of the water molecule that is to be converted into a hydroxide ion. As TIP3P water hydrogens do not have van der Waals parameters, this is sufficient to annihilate the hydrogen atom. The second stage corresponded to the modification of the remaining atomic charges of the oxygen and hydrogen atoms into the appropriate hydroxide atomic charges.[39] The transformations were always made in both directions (i.e., water into hydroxide and hydroxide into water) to evaluate the hysteresis. The statistical error was calculated through the propagation

Scheme 1. The Gibbs Energy Associated with the Exchange of an Active Site-Magnesium Bound-Water Molecule by a Bulk Solution Hydroxide Is Given by $\Delta\Delta G_{binding}^{H_2O\rightarrow HO^-}$ and Is Calculated by the Difference between the Gibbs Energy of Transforming a Water Molecule into a Hydroxide Ion at the INT Active Site ($\Delta G_{INT}^{H_2O\rightarrow HO^-}$) and in Bulk Solution ($\Delta G_{aq}^{H_2O\rightarrow HO^-}$)



of the standard deviation of the results in each of the nine steps of the two stages, corrected by the correlation time of the data. The last was calculated using a data-block technique.[40] As the statistical uncertainty (±0.2 kcal/mol) is higher than the histeresis (±0.1 kcal/mol), we adopted the first as the uncertainty of $\Delta\Delta G_{binding}^{H_2O\rightarrow HO^-}$.

QM/MM Calculations (*Tasks vi* and *ix*). QM/MM calculations were done with GAUSSIAN03,[41] within the ONIOM scheme.[42−44] The QM/MM starting coordinates to explore the three hypotheses of reaction mechanism considered in this work were taken from the end of the MD simulation of the INT:DNA complex (*task iv*). The QM high level layer was described at the MPWB1K/6-311++G(2d,2p)//B3LYP/6-31G(d) level.[45−51] The choice of the specific functionals and basis sets was based on a previous extensive benchmarking study on the performance of DFT in the description of the hydrolysis of phosphodiester bonds.[52] According to that study, the present theoretical level is expected to provide activation and reaction energies very close to the results of CCSD(T) extrapolated to the CBS limit for this particular reaction.[52]

The number of atoms in the QM layer ranged between 47 and 66, depending on the explored mechanism. It included the acetate portion of the side chain of the three residues of the catalytic triad, the two magnesium ions and respective water coordination spheres (six water molecules), and a portion of the substrate. (See Figure 1 for details on the precise substrate atoms included in the QM layer) As seen on Figure 1, the reactants structures for all three hypotheses are almost the same. Slight variations are due to intrinsic differences in the models (e.g. one has an hydroxide ion while the other two have a water in the same place) or to structural rearrangements during QM/MM optimizations that make a conformation more productive for a certain pathway (e.g., better positioned for the phosphate to act as a base). The valences of the truncated bonds were completed by the addition of hydrogen link atoms. The total number of atoms in the two layers ranged between 2551 and 2555, depending on the explored mechanism. The MM part of the model was treated with the AMBER force field[53] as implemented in GAUSSIAN03. The electrostatic interaction between the QM and the MM layers was treated with the electrostatic embedding method (inclusion of the MM point charges into the QM Hamiltonian). Transition states, intermediates, and products geometries were obtained by scanning the proper coordinates of the system. To obtain the first TS, we started from the reactants and successively shortened the distance between the oxygen of the nucleophile water and the phosphorus of the scissile phosphate, with −0.05 Å increments. To obtain the second TS (in aspartate-base and phosphate-base mechanisms), we started from the intermediate and successively stretched the bond between the same phosphorus and the oxygen of the leaving group, with 0.05 Å increments. The potential energy surface for the three reactions is shown in the Supporting Information. Apart from the scanned coordinates, all atoms in the model were free to move in these calculations.

Solvent Effects (*Task vii*). The solvent was included through high-level single point QM calculations with the IEFPCM polarizable continuum model[54−56] at the MPWB1K/6-311++G(2d,2p) level. For these calculations, smaller models were built, which allowed for a pure DFT approach. These models contained between 159 and 163 atoms (defined by a radius of 5 Å around the scissile phosphoester bond, and then by completing the truncated residues), corresponding to a larger portion of the substrate, the two magnesium ions, the catalytic triad, and the neighboring residues Cys65, Thr66, and Asn155. Figure 2 shows all the atoms included in these calculations. A structure with the atoms included in these calculations is given in Supporting Information (iefpcm.xyz). The incomplete valences originated by the truncation of the residues were completed by the addition of hydrogen atoms. These QM calculations were done with GAUSSIAN03.

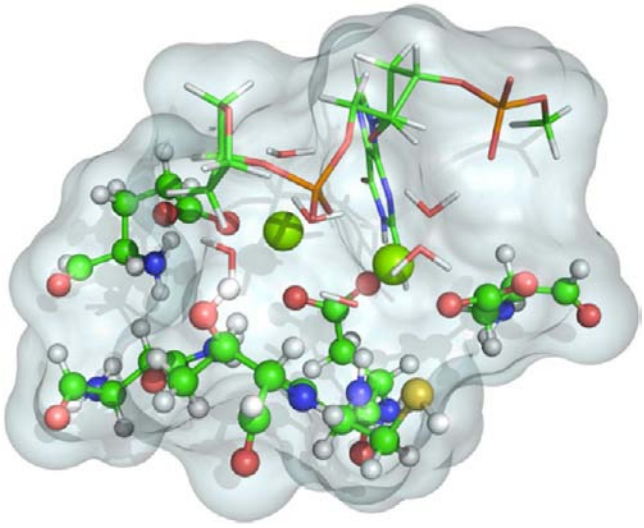


Figure 2. The atoms included in the smaller model to calculate the solvent contribution with the IEFPCM polarizable continuum model. The enzyme atoms are represented in ball and stick. The substrate and water molecules are shown as sticks.

The final energies presented in Table 1, are the sum of the ONIOM(MPWB1K/6-311++G(2d,2p):AMBER)//ONIOM(B3LYP/6-31G(d):AMBER) electronic and force field energies plus the contribution of the solvent calculated with the smaller model.

RESULTS

1. Modeling of the Holoenzyme. We started the modeling by keeping only the central core domain of INT. There is strong evidence pointing to the fact that this domain is the only one required for the chemical transformation. The other domains of INT have important roles (such as DNA binding for example), but they do not participate directly in the chemical transformations.

In fact, it is interesting to notice that other similar retroviral integrases of others viruses such as the Rous sarcoma virus (RSV) and the feline immunodeficiency virus (FIV) do not need other domains, beyond the catalytic, to perform the chemical reaction.[3,57,58] Moreover, the opposite reaction of strand-transfer (disintegration), is catalyzed in vitro by HIV-1 INT only with the central core domain, even though it performs exactly the same chemistry transformations as the ones involved in strand

Table 1. Activation and Reaction Energies for All the Stationary Points of the Three Mechanisms Explored in This Work[a]

	QM/MM energy	ΔE_{Solv}	total energy
Aspartate-Base			
TS1	33.6	2.2	35.8
INT	24.7	6.3	30.9
TS2	28.2	5.8	34.0
P	17.5	−1.0	16.5
Phosphate-Base			
TS1	28.7	1.6	30.3
INT	26.3	6.0	32.3
TS2	33.1	3.2	36.3
P	6.8	−7.0	−0.2
Hydroxide-Base			
TS	10.6	1.4	12.0
P	−17.4	0.2	−17.2

[a]The QM/MM energy refers to the energy of the whole system without the aqueous solvent. ΔE_{solv} corresponds to the contribution of the aqueous solvent for the activation or reaction energy. All values are given in kcal/mol.

transfer.[9] Finally, the most stringent evidence that the chemical reaction takes place only within the central core domain is that all the active site and all catalytic residues are exclusively located in this domain, and at a large distance from the other two domains. These facts show that the N-terminal and C-terminal domains are not directed involved in the chemical reaction, even though they have important accessory roles (probably in the recognition and positioning of the DNA substrate, among other roles). Therefore, they could be omitted from the model for the purpose of our study.

Concerning the substrate, we used the single chain pentanucleotide 5-GCAGT-3. The tetranucleotide 5-CAGT-3 is the optimal motif for the action of INT.[59] As the reaction occurs between the second and the third nucleotides (counting from the 3′ end) we think that the truncation at the fifth nucleotide is far enough from the reaction center to frontier effects to manifest. Furthermore, we used single chain DNA as all evidence show that the terminal nucleotides of the viral DNA are unpaired before the reaction takes place.[60]

Regarding the Mg^{2+} ions, we first tested an INT model with just one magnesium ion. In such model, the activation energy for the reaction amounted to 75 kcal/mol. We have not seen any way to reduce this gigantic barrier. The value of 75 kcal/mol makes sense, as it is close to the gas-phase barrier for the same reaction in model compounds (87.8 kcal/mol).[52] It corresponds to the "intrinsic chemical barrier" for phosphodiester bond cleavage, without stabilization or destabilization by the environment. Furthermore, with just one Mg^{2+}, the fundamental role of Glu152 cannot be justified (see one_mg_ts.pdb in the Supporting Information). Note that mutation of Glu152 renders the enzyme almost inactive.[9] The most plausible option is, therefore, the inclusion of two Mg^{2+} ions coordinated with the three catalytic residues (Asp64, Asp116, and Glu154). In fact, albeit the few available crystallographic structures of HIV-1 INT have only one Mg^{2+}, it is widely accepted, by comparison with other similar enzymes, that a second Mg^{2+} will bind the catalytic core of INT together with substrate binding.[5,10,61,62] A structure with just one Mg^{2+} and the substrate has an excess of negative charge (−2, considering the first shell of coordination of the Mg^{2+} ions) where a nucleophilic attack is expected to occur.

In summary, our model with two magnesium ions can justify the fundamental role of Glu152, is consistent with the structure of all related enzymes and, as it will be shown later on, is the only one that catalyzes the reaction with a turnover rate that is compatible with the experimental turnover.

Very recently, a crystallographic structure of PFV INT with bound DNA was published.[20] Although this structure was still unavailable when we first modeled our INT:DNA complex, we were pleased to verify a very large similarity between both, in particular in what concerns the organization of the catalytic residues and Mg^{2+} ions. We emphasize that this PFV INT:DNA structure also has two Mg^{2+} ions because it has the substrate already bound. Not much can be said about the position of the scissile phosphoester in the two enzymes, as the experimental PFV INT:DNA structure corresponds to a step after the 3′-end processing reaction, and hence, the cleaved dinucleotide is not present in the catalytic center. A superposition of the structure of PFV INT with the structure of our model at the transition state position is shown in Figure 4B and discussed in the following section.

2. Modeling of the INT:DNA Complex. After modeling the central core domain, we performed a guided docking of the substrate and carried out a subsequent 5 ns molecular dynamics simulation to relax the system. The obtained model is depicted in Figure 3. Please note that this docking is not at all affected by the uncertainty that predates typical docking protocols, due to the imposition of three well-known knowledge-based constraints (summarized beforehand in the section Docking of the Substrate (Task iii)) that completely limit the translational and rotational degrees of freedom of the substrate.

The scissile phosphoester is placed between the two Mg^{2+} ions. The coordination sphere of both metals is octahedral, comprising three water molecules, the carboxylate of two residues of the catalytic triad and the scissile phosphoester group. One of the magnesium ions, the *nucleophilic magnesium ion* (Mg^{2+}_{nuc}), is coordinated with Asp64 and Asp116 and the other magnesium ion, the *leaving group magnesium ion* (Mg^{2+}_{lg}), is coordinated with Asp64 and Glu152.

The arrangement of the active site resulting from the modeling is very similar to other enzymes that catalyze similar reactions, such as polymerase I or Ribonuclease H.[10,11,35] A superposition of Rinonuclease H (RNase H) active center bound to DNA (PDB: 1ZBL) with the INT active center shows this similarity (Figure 4A). Eleven "equivalent atoms" of each structure (connected by a black line) were used to superimpose the two structures, yielding the very low rmsd of 0.66 Å. The position of the scissile phosphoester between the Mg^{2+} ions and the position of the nucleophile and of the leaving group in these cases are almost identical in RNase-H and in our model of INT. The coordination of the ions to the enzyme is also similar, apart from specific differences related to the topology of the active site.

The coordination of the Mg^{2+} ions in our model is also in agreement with the recent PFV INT crystallographic structure complexed with DNA and an inhibitor (Raltegravir) (PDB ID: 3OYA).[20] The PFV INT structure was crystallized in a state after the 3′-end processing reaction, with the two processed nucleotides having left the active site. A superposition of the active center of our model with the active center of the crystallographic structure of the PFV INT:DNA complex is shown in Figure 4B. The coordination of the acidic residues to Mg^{2+}_{nuc} differs only very slightly in the two structures. The acidic residues of both enzymes bind in adjacent positions of the octahedral coordination sphere of the metal ions. The

dx.doi.org/10.1021/ja304601k | J. Am. Chem. Soc. 2012, 134, 13436−13447

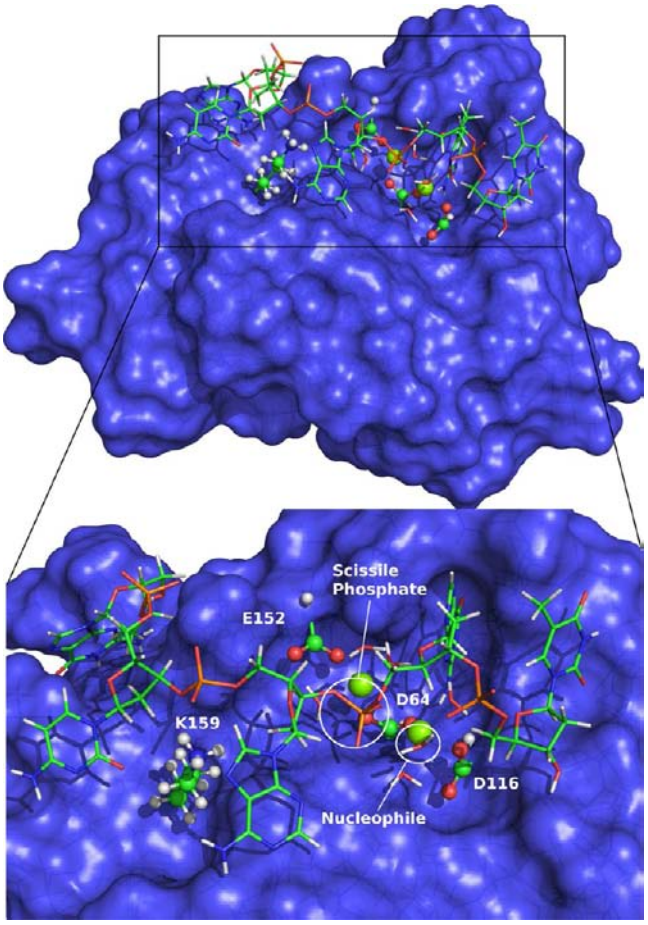


Figure 3. (Top) The INT:DNA model used in this study. The central core domain of INT is shown as a blue surface. The DNA substrate and the water molecules from the magnesium coordination shells are shown in stick representation and colored by atom type. The catalytic triad and lysine 159 are shown in ball and stick representation and are colored by atom type. The magnesium ions are the two green spheres. (Bottom) A detailed view of the active site of the INT:DNA complex.

coordination of the acidic residue in the center of the picture (D128 in PFV: D64 in INT) to the Mg^{2+} ion is exactly the same in the two structures. The PFV INT acidic residue in the left (E221) is coordinated to the metal with both oxygen atoms. In the INT:DNA model, the binding is established by E152 and by a water molecule, but in very similar positions to PFV INT. An equivalent bidentated coordination of E152 in our model is not feasible due to the folding of the main chain. This small difference might be attributed to the fact that, after all, we are superimposing two distinct enzymes.

Direct comparison between the positions of the scissile phosphoester cannot be made as it is not present in the PFV INT structure. Nevertheless, it is extremely encouraging to see that the three atoms of the inhibitor bound to PFV INT coordinated to the Mg^{2+} ions overlap almost flawlessly with the oxygen atoms of the nucleophile, phosphate, and the water molecule that donates a proton to the leaving group.

The overall orientation of the substrate is also in agreement with other enzymes and INT models.[14] An interaction between the N7 of adenine and the side chain of Lys 159 can be seen in the model (see Figure 3). Such interaction was proposed to be important for the integration process and was observed experimentally.[36] Our INT:DNA structure differs from other structures in the position of the nucleotide bases. This was expected, as our structure binds single stranded DNA, whereas other structures bind invariably double stranded DNA (note that experimental evidence shows that DNA extremities are impaired during 3′-end processing).[60]

The overall position of the part of the substrate that is crucial for the reaction is similar to most of the related enzymes; the placement of the other phosphates and bases is important for DNA recognition but should not change the reaction pathway or its energy. An exception is the phosphate of the 3′-terminal nucleotide of the substrate, as it may participate in the reaction, deprotonating the nucleophile (this phosphate is also in a good position in our INT model to perform this role). The 5 ns molecular dynamics of the INT:DNA confirmed the stability of the structure we have modeled, with a stable rmds of around 2.5 Å (see Supporting Information Figure SI-2).

3. The Catalytic Mechanism. The INT:DNA structure and the mutagenic data[9] suggests only three possible mechanisms for the 3′-end processing reaction. Scheme 2 illustrates these hypotheses. All comprise the attack of a nucleophile originated in the Mg^{2+}_{nuc} coordination sphere. The attack of a nucleophile coming from bulk solution (either H_2O or HO^-) was also

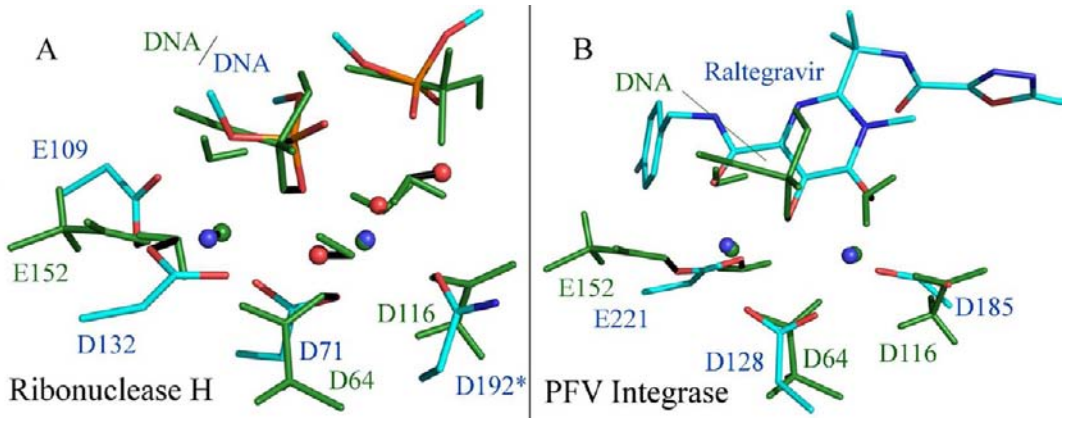
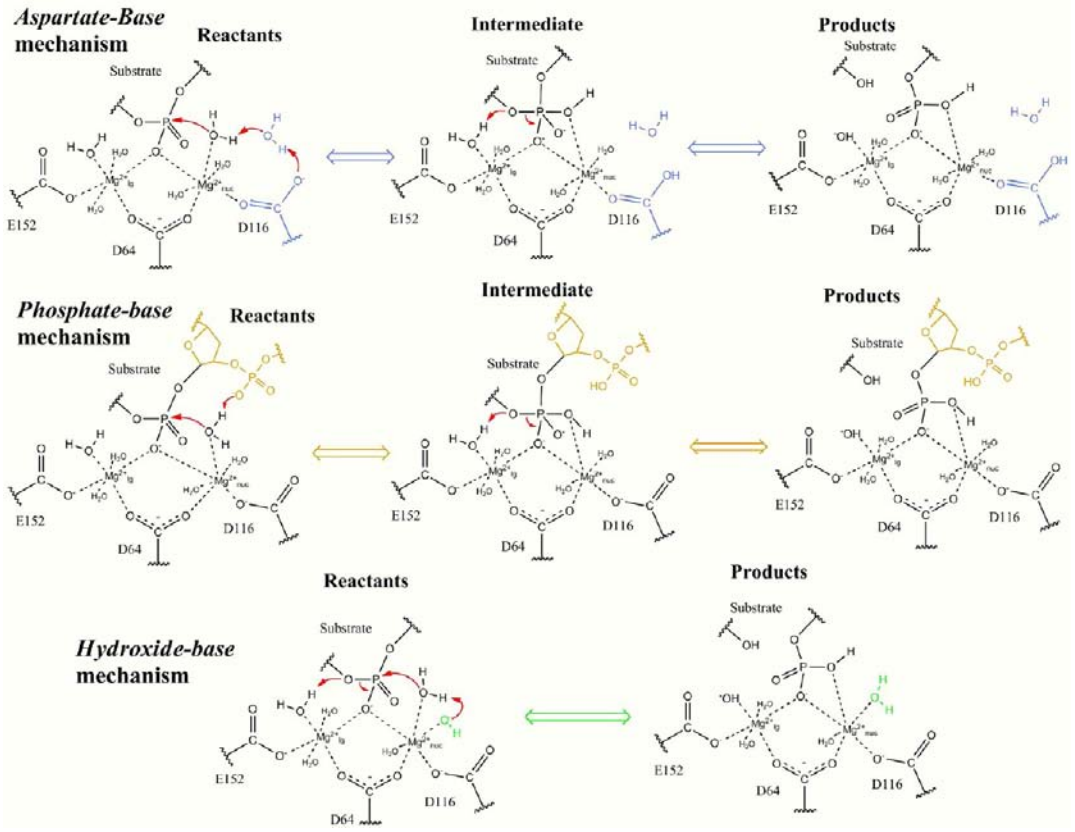


Figure 4. (A) Superposition of the active center of our model of HIV-1 INT and RNase H; (B) Superposition of the active center of our model of HIV-1 INT and PFV INT complexed with DNA and with the inhibitor Raltegravir (right). In both panels, HIV-1 INT atoms are colored green. RNase H and PFV INT carbons and Mg^{2+} ions are colored blue. The red spheres are the oxygen atoms of water molecules present in the crystallographic structure. The black lines connect equivalent atoms in the two structures.

dx.doi.org/10.1021/ja304601k | J. Am. Chem. Soc. 2012, 134, 13436−13447

Scheme 2. Schematic Representation of the Possible Mechanisms for Phosphodiester Bond Hydrolysis That Were Studied in This Work



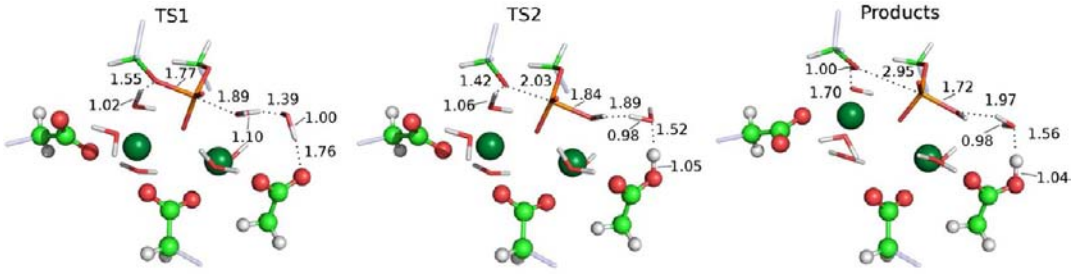


Figure 5. Stationary points for the *Aspartate-Base* pathway (reactants not shown). Relevant interatomic distances (in angstrom) are included. The representations include all the atoms in the high level QM layer. The rest of the enzyme was deleted for clarity. The light blue sticks represent the QM/MM link atoms. Protein atoms are represented in ball and stick, and the substrate and water molecules are represented as sticks. The two dark green spheres correspond to Mg^{2+}_{nuc} (right) and Mg^{2+}_{lg} (left).

considered in preliminary calculations. However, when such external nucleophiles come close to the phosphoester scissile bond, they invariably bind to Mg^{2+}_{nuc} and release a coordinated H_2O, turning the mechanism into one of the three already shown in Scheme 2.

The three hypotheses show some similarities and differ mostly in the species that accepts a proton from the nucleophile (the activating base). We named the mechanisms as *Aspartate-Base*, *Phosphate-Base*, and *Hydroxide-Base*, accordingly. Apart from these, we consider that there are no more plausible hypotheses for the 3′-end processing mechanism. Details of the mechanisms can be better visualized in Figures 5−7 and Scheme 2.

The most important interatomic distances are given in the text. Supporting Information Tables SI-2 to SI-4 show further details of the geometric changes that take place along the reaction pathways. We will discuss subsequently the three mechanisms, one by one, starting with the *Aspartate-Base* hypothesis. In order to simplify the description of the geometric changes along the reaction coordinates, we will adopt the notation (*a*; *b*; *c*), where *a*, *b*, and *c* refer to interatomic distances (in Å) at the reactants, transition state, and products. OW refers to a water oxygen and HW to a water hydrogen.

3.1. The Aspartate-Base Mechanism. In this hypothesis, the nucleophile is a water molecule. The water molecule attacks the phosphoester bond and simultaneously it is deprotonated by Asp116, with the help of a bridging water molecule (see Scheme 2 and Figure 5). Asp116 is quite restricted in terms of conformational rearrangements, due to its coordination to the Mg^{2+} ion, which precludes its approximation to the water molecule to deprotonate it directly. In the molecular dynamics simulations, the distance between the phosphorus of the scissile phosphoester bond and the basic oxygen of Asp116 is always

dx.doi.org/10.1021/ja304601k | J. Am. Chem. Soc. 2012, 134, 13436−13447

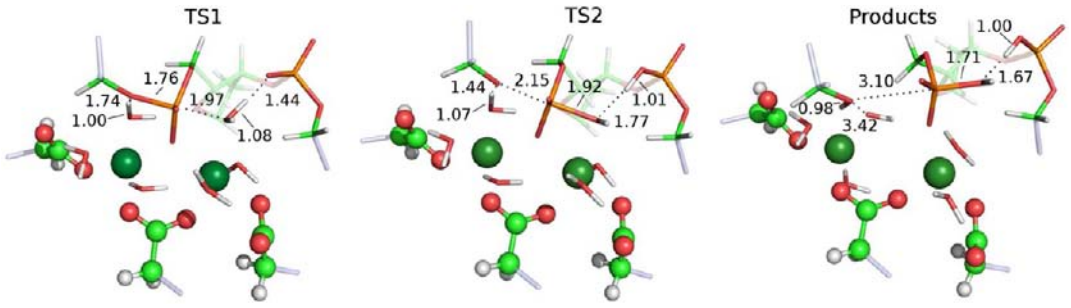


Figure 6. Stationary points for the *Phosphate-Base* pathway (reactants not shown). Relevant interatomic distances (in angstrom) are included. The representations include all the atoms in the high level QM layer. The rest of the enzyme was deleted for clarity. The light blue sticks represent the QM/MM link atoms. Protein atoms are represented in ball and stick, and the substrate and water molecules are represented as sticks. The two dark green spheres correspond to Mg^{2+}_{nuc} (right) and Mg^{2+}_{lg} (left).

larger than 6 Å. Therefore, it is impossible for the nucleophilic water to attack the phosphorus and at the same time give the proton to the asparte, as it needs to move away from the aspartate to attack the phosphate (see Supporting Information Figure SI-3). The assistance of a second water molecule is indispensable. Both water molecules were already present in the positions shown in Figure 5 in the preceding molecular dynamics simulation of the INT:DNA complex. The nucleophilic attack of the water molecule to the phosphorus atom of the scissile bond (with concomitant proton transfer from the water molecule to Asp116) results in the formation of a high energy pentacoordinated intermediate (INTM). At the transition state (TS), the bond between the phosphorus atom and the water oxygen (OW−P) is almost formed (3.04; 1.89; 1.84) but the OW−HW bond is only slightly elongated toward the bridging water (1.02; 1.10; 1.89). Consistently, the length of the hydrogen bond between the proton of the bridging water and the Asp116 oxygen is still large (1.72; 1.76; 1.05). In the product of this step, a metastable pentacoordinate intermediate is formed, with the new bond and the leaving group OW−P bond almost equivalent in length, and Asp116 protonated. The barrier of this step amounts to 35.8 kcal/mol and the reaction energy (the energy of the pentacoordinate intermediate) amounts to 30.9 kcal/mol.

The second step corresponds to the breaking of the bond that connects the pentacoordinated intermediate phosphate to the 3′ oxygen (O3′) of the ribose of the leaving group (see Scheme 2). Upon breaking the O3′−P bond, the O3′ oxoanion deprotonates a water molecule coordinated to the Mg^{2+}_{lg}. The TS for the elimination is an early TS (contrarily to the late first TS), with the O3′−P bond that is being broken still just slightly elongated (1.83; 2.03; 2.95). The Mg^{2}_{lg}-bound acidic water is still protonated, with HW−OW distances of (1.02; 1.06; 1.70) and the transfer only occurs at the products when the pentacoordinated intermediate is fully resolved. The energy involved in each step of this reaction (as well as for the next ones), decomposed in its components, is given in Table 1.

The activation energy for this step is 3.1 kcal/mol (34.0 kcal/mol in relation to the initial reactants) and the global reaction energy is 16.5 kcal/mol. The potential energy profile for this reaction shows two high energy TSs, with a first, rate limiting step of 35.8 and a metastable pentacoordinate intermediate. The solvent raises the energy of the intermediate and second transition state by ca. 6 kcal/mol and stabilizes the products by 1.0 kcal/mol.

3.2. The Phosphate-Base Mechanism. In the second reaction pathway, the nucleophile is also a water molecule and the reaction proceeds in a similar manner as the preceding *Aspartate-Base* mechanism (see Scheme 2 and Figure 6). However, the 3′-terminal phosphate, instead of Asp116, deprotonates the nucleophilic water molecule during its attack on the phosphoester bond. Moreover, the nucleophilic water proton is transferred directly to the phosphate basic oxygen, without the intervention of any bridging water molecule, due to the short distance between them. The attack occurs during the first step of the reaction. At the TS, the OW−P distance shortens so much that the new bond is almost fully formed (3.02; 1.97; 1.92). The nucleophilic water proton is not transferred at this point, with OW−HW distances of (1.00; 1.08; 1.75). The phosphate scissile bond is just very slightly elongated (1.65; 1.76; 1.80). The nucleophilic water proton only migrates to the phosphate in the product of this step. The activation energy is 30.3 kcal/mol and the resulting intermediate (INTM) is again a metastable pentacoordinated intermediate with an energy of 32.3 kcal/mol. TS1 is a true transition state structure at the ONIOM-(B3LYP/6-31G(d):AMBER) and ONIOM(MPWB1K/6-311++G(2d,2p):AMBER). However, the energy of INTM becomes slightly higher than the energy of TS1 upon the *a posteriori* addition of the implicit solvation energy.

The second step corresponds to the breaking of the scissile phosphoester bond. At the TS, the deoxyribose oxygen of the leaving dinucleotide is partially unbound from the phosphate, with O3′−P distances of (1.80; 2.15; 3.10) and partially protonated by a water molecule of the coordination sphere of the Mg^{2}_{lg}, with O3′−HW distances of (1.69; 1.44; 0.98). The energy of this transition state is 36.3 kcal/mol above energy of the reactants (4.0 kcal/mol above the pentacoordinated intermediate) and the overall reaction energy is −0.2 kcal/mol.

3.3. The Hydroxide-Base Mechanism. This mechanism assumes that the nucleophilic water molecule is deprotonated by an HO^- ion coordinated to the Mg^{2}_{nuc} (Scheme 2 and Figure 7). The first step of this mechanism is to exchange an Mg^{2}_{nuc}-bound water molecule for a bulk solvent hydroxide ion.

To evaluate the free energy involved in this process, one has to have two effects in mind: the entropic penalty for finding the hydroxide molecule, a species with a low concentration of 10^{-7} in solution, in a small volume of the active site, around Mg^{2}_{nuc}, and the free energy associated with the chemical exchange of a Mg^{2+}-bound water molecule for an Mg^{2+}-bound hydroxide ion, which includes contributions from metal binding and from changing the environment (bulk sovent/enzyme).

The free energy of reducing the accessible volume of a particle (which comes from the reduction in translational entropy) can be calculated with the simple particle-in-a-box model, which results in eq 1 below,[63]

dx.doi.org/10.1021/ja304601k | *J. Am. Chem. Soc.* 2012, 134, 13436−13447

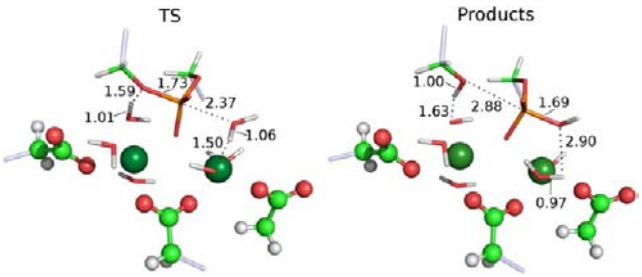


Figure 7. Stationary points for the *Hydroxide-Base* pathway (reactants not shown). Relevant interatomic distances (in angstrom) are included. The representations include all atoms in the high layer. The rest of the enzyme was deleted for clarity. The light blue sticks represent the QM/MM link atoms. Protein atoms are represented in ball and stick, and the substrate and water molecules are represented as sticks. The two dark green spheres correspond to $Mg^{2+}{}_{nuc}$ (right) and $Mg^{2+}{}_{lg}$ (left).

$$\Delta G_{trans} = -k_B T \ln \frac{V_f}{V_i} \quad (1)$$

where V_f and V_i stand for the final and initial accessible volumes, k_B is the Boltzmann constant, and T is the absolute temperature. At neutral pH and 310.15 K, this correction amounts to +11.7 kcal/mol, considering a volume ratio given by the accessible volume of a hydroxide ion at a concentration of 10^{-7} M and a final accessible volume of 3 water molecules. The free energy is quite insensitive to the exact value of the final volume and the value of 3 water molecules was chosen because it corresponds to the number of coordination positions free for the hydroxide ion around $Mg^{2+}{}_{lg}$. The difference between the active site volume and the bulk solution volume is so large that other values will change the free energy only by a few tenths of kcal/mol.

To calculate the second energy contribution, the free energy involved in changing one Mg^{2+}-bound water molecule by a hydroxide ion we ran a set of molecular dynamics simulations and calculated that free energy through thermodynamic integration (TI). We have transformed a hydroxide ion into a water molecule (and vice versa), both in a water box and in the active site of the protein. The result of the TI calculation is −9.0 ± 0.2 kcal/mol (see the Methods section and Supporting Information Table SI-5 for more details on the TI results), meaning that the Mg^{2+}-bound hydroxide and a water molecule in solution are 9.0 kcal/mol more stable than an Mg^{2+}-bound water molecule and a hydroxide ion in solution. The positive charge of the Mg^{2+} ion contributes significantly to this overstabilization.

The sum of these two energies is +2.7 kcal/mol, meaning that the total cost of exchanging a water molecule by a hydroxide ion at the active site is quite small. This value also shows that the predominant coordination is made by water molecules (and not by hydroxide ions), in agreement with the modeling previously done.

This mechanism is different from the previous in the number of elementary steps, as it generates the products through only one reaction step. The transition state consists in the attack of the nucleophilic water to the phosphorus atom of the scissile bond, with concomitant breaking of the opposite O3′−P bond. No pentacoordinated intermediate is formed. At TS1, the OW−P distance is halfway between reactants and products (3.42; 2.37; 1.69). The acidic proton of the nucleophilic water makes a short hydrogen bond with the $Mg^{2+}{}_{nuc}$-bound hydroxide ion but keeps connected to the water molecule, with HW−OW distances of (1.00; 1.06; 2.90). At this stage, the scissile O3′−P is only very slightly elongated (1.68; 1.73; 2.88).

In the products of this step, the O3′−P bond brakes down and O3′ deprotonates a water molecule bound to the $Mg^{2+}{}_{lg}$ ion. This reaction has an activation energy of 12.0 kcal/mol and a reaction energy of −17.2 kcal/mol. The effect of the solvent is smaller than that on the other two mechanisms, raising the TS1 energy by 1.4 kcal/mol and barely affecting the reaction energy (Table 1).

As a whole, the energies for the *Hydroxide-Base* mechanism match much better the expected kinetics than the energies of the other pathways. In fact, this is clearly the only mechanism that is consistent with the experimental kinetics, which sets an upper barrier of 24.6 kcal/mol for the whole process. All the other hypotheses have barriers over 35 kcal/mol, surpassing by far both the kinetics of the *Hydroxide-Base* mechanism determined limit and the rate-limiting barrier experimentally observed.

In summary, all lines of evidence point to the fact that the *Hydroxide-Base* mechanism is not only the one that matches all experimental observations but also that it is the only one which is feasible for this enzyme.

As this mechanism is clearly the only one viable and consistent, we have made a further small approximation, which consisted in transferring the contributions of the zero point energy (ZPE) and vibrational entropy calculated for the same reaction in a small molecular model[52] to the present large enzyme model. The contributions from the ZPE and from entropic effects are always very small in this kind of systems and transformations. They amount to 0.4 or 1.0 kcal/mol for the TS in vacuum and in solvent and contributions of −0.5 and −1.0 for the reaction energy in vacuum and in solvent. We used the values in between (0.7 kcal/mol for the activation energy and −0.8 kcal/mol for the reaction energy) as in the enzyme we have an intermediate dielectric. The result is completely insensitive to these choices but the addition of ZPE and entropic contributions allows us to calculate free energies, instead of enthalpies, and add up the free energy for hydroxide exchange in a solid thermodynamic ground. Adding all these terms (activation energy, ZPE, entropic effects, solvation, and water/hydroxide exchange), we end up with an activation free energy of 15.4 kcal/mol and a reaction energy of −15.3 kcal/mol.

3.4. Exploration of the Conformational Space for the Hydroxide-Base Mechanisms. After showing that the *Hydroxide-Base* mechanism was the only viable hypothesis for the 3′-end processing reaction of INT, we did an additional series of calculations to ensure that the results were not affected by the precise choice for the initial enzyme geometry. This preoccupation arises from single molecule experiments,[64,65] where it can be shown that differences between the folding state of different molecules (static disorder) and conformational fluctuations around a given folding at a time scale comparable with the chemical kinetics (dynamic disorder) affect the rate of the chemical reaction (K_{cat}). These effects are usually minor and will not invert the choice of a given chemical pathway over another, in particular in this case where all hypotheses beyond the chosen one have activation energies higher by 20 kcal/mol and that exceed the experimental activation energy limit by 12 kcal/mol. Theoretical studies leaning over this problem[66,67] showed that maximum difference between barriers coming from different initial configurations may amount to 11 kcal/mol. In general, the differences in activation energies coming from different enzyme structures amount to a modest 1−4 kcal/mol.

To confirm our findings, we also did this test for the *Hydroxide-Base* mechanism. Starting from the QM/MM structure of the reactants of the *Hydroxide-Base* mechanism, we did four 1 ns

molecular dynamics simulations (with different initial velocities). From the last structure of each MD simulation, we explored again the potential energy surface of the mechanism and located the transition state and products. The obtained activation energies (QM/MM energies, without solvent contribution, at the ONIOM(B3LYP/6-31G(d):Amber)) were 7.6, 11.2, 12.6, and 17.4 kcal/mol. The energy in the original scan was 11.7 kcal/mol, at this level of theory. We can see that the activation energies span a range of 9.8 kcal/mol, with a maximum value of 17.4 kcal/mol. We have not weighted and averaged these values because it makes no sense to change from a PES based in geometry optimizations to an ensemble averaged PES with just five states. What the results show instead is that the *Hydroxide-Base* mechanism is always the preferred mechanism, whatever initial enzyme configuration we use (within the limits of the sampling carried out). In conclusion, whatever the value we take from the five potential energy surfaces, this mechanism is always consistent with experimental kinetics and much faster than any other alternative mechanism. However, the differences in activation energy found using different starting conformations of the enzyme point to the need of a careful evaluation of this effect in QM/MM studies of enzymatic reactions.

DISCUSSION

1. The Kinetics of the Chemical Reaction. After 3′-end processing, the INT:DNA complex is transferred to the nucleus, where the strand transfer reaction (also catalyzed by INT) takes place. It is necessary that the DNA does not dissociate from INT before its integration into the host genome. The very small 3′-end processing turnover experimentally measured (0.1 h^{-1}) is due to the very slow rate-limiting process of substrate dissociation, and not to the chemical reaction.[68,69] It has not been possible to determine (experimentally) the kinetics for the chemical step so far. Therefore, we have no experimental value to compare directly with the computed ones. The turnover value of 0.1 h^{-1} provides only an upper limit (24.6 kcal/mol according to transition state theory) for the energy of the rate-limiting barrier of the chemical reaction. Our results are currently the best quantitative answer for the kinetics of the chemistry of 3′-end processing. Moreover, the activation energy found here (15.4 kcal/mol) is consistent with the expected kinetics for an enzyme catalyzed reaction.

2. The Exonuclease Reaction of the *Escherichia coli* DNA Polymerase I as a Guide for Enzymatic Phosphodiester Cleavage. The 3′−5′ exonuclease reaction of *E. coli* DNA polymerase I (DNAP-I) was the first enzymatic phosphodiester hydrolysis proposed to be catalyzed by two metal ions.[70] The mechanistic proposal was based in the analysis of a preceding crystallographic structure of the Klenow fragment of *E. coli* DNA Polymerase I (DNAP-I) bound to a single stranded DNA substrate.[10] The active sites and reactions catalyzed by INT and DNAP-I are very similar. Therefore, it has been assumed that INT might share the chemical mechanism of DNAP-I.[3] It is very interesting to note that the binding structure of DNA resulting from our INT model is also very similar to the one seen in the structure of DNAP-I bound to DNA mentioned above. Both structures have the phosphate group placed between the two Mg^{2+} ions. They also both have the nucleophile coordinated to one Mg^{2+} ion, while the leaving group is stabilized by the other Mg^{2+} ion (through a water molecule in the INT case).

The side chains that coordinate the Mg^{2+} ions are slightly different (but still similar) in the two cases, due to intrinsic differences between the two enzymes. For example, one Glu bound to the Mg^{2+}_{nuc} ion of DNAP-I is replaced by one water molecule in INT, generating a similar coordination shell. A theoretical study by A. Warshel and co-workers[71] describing the cleavage reaction in this system reported activation energies of 25 kcal/mol for a hydroxide nucleophile recruited from solution, and of 35 kcal/mol for a water nucleophile. The hydroxide mechanism has only one transition state while the water mechanism has two. The latter hypothesis is similar to our *Aspartate-Base* mechanism, as a proton transfer occurs from the nucleophile to a glutamate, although without the bridging water. The activation energies obtained in the two enzymes (RNAP-I and INT) for the same *Aspartate-Base* mechanism are remarkably similar (35 and 36 kcal/mol). In the case of the *Hydroxide-Base* mechanism, both studies arrive at similar transition states and similar single step mechanisms, even though the activation energies obtained in INT are lower than in DNAP-I.

3. Structures and Mechanisms of Enzymes with Similar Active Centers. Besides DNAP-I, there are crystallographic structures of a substantial number of enzymes that catalyze phosphoester bond formation or bond hydrolysis with distinct, although similar, active site geometries.[35] Examples include RNase H,[11] MutH endonuclease,[72] Group I intron ribozyme,[73] T7 DNA polymerase,[74] T7 RNA polymerase[75] or the BamHI restriction endonuclease,[76] among others. All of these enzymes have two Mg^{2+} ions in the active center, without any exception. While the T7 RNA polymerase and the T7 DNA polymerase catalyze the polymerization of the respective polynucleotide chains using triphosphate nucleotides as substrates, and the Group I Intron ribozyme has a ribose oxygen as nucleophile, the active centers of the other mentioned enzymes are particularly similar to the INT case and can be very elucidative to us.

In the RNase H and MutH cases, the phosphate of the nucleotide adjacent to the scissile phosphoester is hydrogen-bonded to the water nucleophile, in the same way as in our INT model for the *Phosphate-Base* mechanism. On the other hand, BamHI has an active site geometry that resembles much more our INT model used to study the *Aspartate-Base* mechanism, but without the bridging water used in INT.

Theoretical studies performed on some of these enzymes have been helpful to hypothesize the several possible pathways studied here and to compare the activation energies for each one. In a recent work with Rnase H,[77] a water molecule and a hydroxide ion were tested as nucleophiles. Activation energies of 16.9 and 10.5 kcal/mol were reported. The base was a phosphate group, which is consistent with the geometry of that particular active site. Another study on BamHI[78] also tested a water molecule and a hydroxide ion as nucleophiles. The activation energies calculated were 23.4 kcal/mol for the hydroxide (recruited from bulk solution) and 29.6 kcal/mol for the water molecule (located in the magnesium coordination sphere). In this mechanism, the water molecule is deprotonated by a glutamate, in a similar way as in our *Aspartate-Base* mechanism. Therefore, there is plenty of chemical precedent for the mechanisms tested here and for the energies that we have obtained.

4. The Mechanism Choice for the Integrase 3′-End Processing Reaction. These many examples of chemical diversity concerning the hydrolysis of phosphodiester bonds emphasize the relevance of our study. Detailed theoretical calculations were needed to understand the INT chemistry. Our results revealed a much smaller activation energy when the active site is more negative. The *Hydroxide-Base* was the favored mechanism, as it is by far the most competent kinetically

(activation energy of 15.4 against 36 kcal/mol for the alternative pathways) and is the only one that has a chemical kinetics compatible with a typical enzymatic reaction and compatible with the experimental turnover.

In this mechanism, the nucleophile comes from the coordination sphere of the Mg^{2+}_{nuc} ion. We have not described here other hypotheses in which the nucleophile arises from the solution because after testing them we have seen that there is absolutely no space for a nucleophile coming from outside to fit in a position to attack the phosphodiester bond, due to the proximity to the coordination sphere of the Mg^{2+}_{nuc} ion.

The activation energies of the *Aspartate-Base* and *Phosphate-Base* mechanisms, 35.8 and 36.3 kcal/mol respectively, are too high for an enzyme catalyzed reaction. Even though they are very appealing, they can be safely ruled out in HIV-1 IN based on kinetic grounds. Another poignant evidence that these mechanistic hypotheses are not real is that their activation energy is equivalent to the activation energy of the uncatalyzed hydrolysis of phosphodiester bonds in solution (35 kcal/mol).[44] These reactions are *not* catalyzed by the enzyme.

CONCLUSION

The goal of this study was to get an atomic-level description of the 3′-end processing reaction of HIV-1 INT. For that purpose, we started by building a model of the central core domain of halo-INT with a ssDNA substrate made of five nucleotides in the appropriate sequence. Techniques of molecular modeling, molecular docking, and molecular dynamics were employed, together with all the experimental data available on INT and similar enzymes. The final model has two magnesium ions in the active center coordinated to the catalytic residues, and the critical phosphate group located between them. The overall geometry of the active center correlates very well with the exonuclease active center of *E. coli* polymerase I, Ribonuclease H, the PFV INT, and other nucleases. This model was the framework to the subsequent QM/MM studies performed and it is a result in itself, as it represents the most consistent and robust INT:DNA structure put forward so far.

Subsequently, we used the quantum mechanical/molecular mechanical calculations and an implicit solvation model to explore the possible chemical mechanisms for the catalytic reaction of 3′-end processing. We conceived and tested three mechanistic hypotheses for the cleavage of the phosphodiester bond, which we have named *Aspartate-Base*, *Phosphate-Base*, and *Hydroxide-Base*. These mechanistic hypotheses were put forward based not only in our inspection and understanding of the system under study, but also in the mechanisms available in the literature for closely related enzymatic systems. A water molecule is the nucleophile in all of the three mechanistic hypotheses, donating a proton to the base in question. We found that the most favorable mechanism is the *Hydroxide-Base* mechanism with an activation energy of 15.4 kcal/mol. The reaction proceeds through a single transition state with the nucleophile attacking the scissile phosphoester together with its deprotonation by the base and elimination of the leaving group. Using the catalytic Asp116 as the base, we reach an activation energy of 35.8 kcal/mol, and using the phosphate of the nucleotide adjacent to the scissile phosphoester bond, we get a barrier of 36.3 kcal/mol. These last two values are too high for an enzymatic catalysis reaction; they do not match the experimental turnover and they are equivalent to the activation energy of the uncatalyzed reaction in solution. Therefore, they can be discarded as options for the 3′-end processing reaction of HIV-1 INT. There are no other alternative mechanisms consistent with the enzyme structure and available experimental data.

As a final note, we think that this work is an important and valid contribution to the field of drug discovery toward the HIV-1 virus. Not only does it improve the fundamental description of the chemistry of INT 3′-end processing reaction at an atomic level, but it also provides accurate structures of the holo-INT at the resting state and at the transition state, which might be very helpful in the discovery of new competitive inhibitors that target the INT active site. The utility of the present structures can be confirmed by the superposition shown in Figure 4B, which depicts a snapshot from the transition state of the *Hydroxide-Base* mechanism whose metal-coordinating atoms are flawlessly superimposed to the ones of the inhibitor Raltegravir.

ASSOCIATED CONTENT

Supporting Information

Table with the pK_a results for all the ionizable residues as obtained by the H++ server; tables with relevant interatomic distances for the three mechanisms along the reaction paths; table with the detailed results of the TI calculations; plots of the rmsd of the protein−substrate complex of the substrate (only), and of the metals coordination spheres along the 5 ns MD; plot of the distances between the nucleophile, the scissile phosphate and the Asp116 along the 5 ns MD; structures of our model before the docking, before and after the 2 ns relaxation MD; structure of the model used in the IEFPCM calculations; structure of the model used in the QM/MM calculations; structure of the TS of the model with one magnesium in the active center; structures and energies of all stationary states of the QM/MM calculations; potential energy surface of the three reactions. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

pafernan@fc.up.pt

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work has been financed by the program FEDER/COMPETE and by the Fundação para a Ciência e a Tecnologia (Project PTDC/QUI/68302/2006). A.J.M.R. thanks FCT for a doctoral scholarship (SFRH/BD/61549/2009)

REFERENCES

(1) Grinsztejn, B.; Nguyen, B.-Y.; Katlama, C.; Gatell, J. M.; Lazzarin, A.; Vittecoq, D.; Gonzalez, C. J.; Chen, J.; Harvey, C. M.; Isaacs, R. D.; Protocol 005 Team.. *Lancet* **2007**, *369*, 1261−1269.
(2) Jaskolski, M.; Alexandratos, J. N.; Bujacz, G.; Wlodawer, A. *FEBS J.* **2009**, *276*, 2926−2946.
(3) Chiu, T. K.; Davies, D. R. *Curr. Top. Med. Chem.* **2004**, *4*, 965−977.
(4) Drelich, M.; Wilhelm, R.; Mous, J. *Virology* **1992**, *188*, 459−468.
(5) Asante-Appiah, E.; Skalka, A. M. *Antiviral Res.* **1997**, *36*, 139−156.
(6) Delelis, O.; Carayon, K.; Saib, A.; Deprez, E.; Mouscadet, J.-F. *Retrovirology* **2008**, *5*, 114.
(7) Engelman, A.; Mizuuchi, K.; Craigie, R. *Cell* **1991**, *67*, 1211−1221.
(8) Poeschla, E. M. *Cell. Mol. Life Sci.* **2008**, *65*, 1403−1424.
(9) Engelman, A.; Craigie, R. *J. Virol.* **1992**, *66*, 6361−6369.
(10) Beese, L. S.; Steitz, T. A. *EMBO J.* **1991**, *10*, 25−33.
(11) Nowotny, M.; Gaidamakov, S. A.; Crouch, R. J.; Yang, W. *Cell* **2005**, *121*, 1005−1016.

(12) Bernardi, F.; Bottoni, A.; De Vivo, M.; Garavelli, M.; Keserü, G.; Náray-Szabó, G. *Chem. Phys. Lett.* **2002**, *362*, 1−7.
(13) Ruiz-Pernia, J. J.; Alves, C. N.; Moliner, V.; Silla, E.; Tunon, I. *THEOCHEM* **2009**, *898*, 115−120.
(14) De Luca, L. *Biochem. Biophys. Res. Commun.* **2003**, *310*, 1083−1088.
(15) Wang, L.-D.; Liu, C.-L.; Chen, W.-Z.; Wang, C.-X. *Biochem. Biophys. Res. Commun.* **2005**, *337*, 313−9.
(16) Karki, R. G.; Tang, Y.; Burke, T. R.; Nicklaus, M. C. *J. Comput.−Aided Mol. Des.* **2004**, *18*, 739−60.
(17) Ferro, S.; De Luca, L.; Barreca, M. L.; Iraci, N.; De Grazia, S.; Christ, F.; Witvrouw, M.; Debyser, Z.; Chimirri, A. *J. Med. Chem.* **2009**, *52*, 569−73.
(18) Fenollar-Ferrer, C.; Carnevale, V.; Raugei, S.; Carloni, P. *Comput. Math. Methods Med.* **2009**, *9*, 231−243.
(19) De Luca, L.; Vistoli, G.; Pedretti, A.; Barreca, M. L.; Chimirri, A. *Biochem. Biophys. Res. Commun.* **2005**, *336*, 1010−6.
(20) Hare, S.; Gupta, S. S.; Valkov, E.; Engelman, A.; Cherepanov, P. *Nature* **2010**, *464*, 232−U108.
(21) Alberto, M. E.; Marino, T.; Ramos, M. J.; Russo, N. *J. Chem. Theory Comput.* **2010**, *6*, 2424−2433.
(22) Himo, F. *Theor. Chem. Acc.* **2006**, *116*, 232−240.
(23) Leopoldini, M.; Marino, T.; Michelini, M.; Rivalta, I.; Russo, N.; Sicilia, E.; Toscano, M. *Theor. Chem. Acc.* **2007**, *117*, 765−779.
(24) Ramos, M. J.; Fernandes, P. A. *Acc. Chem. Res.* **2008**, *41*, 689−698.
(25) Goldgur, Y.; Craigie, R.; Cohen, G. H.; Fujiwara, T.; Yoshinaga, T.; Fujishita, T.; Sugimoto, H.; Endo, T.; Murai, H.; Davies, D. R. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 13040−13043.
(26) Goldgur, Y.; Dyda, F.; Hickman, A. B.; Jenkins, T. M.; Craigie, R.; Davies, D. R. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 9150−9154.
(27) Delano, W. L. *The PyMOL Molecular Graphics System*, version 1.2r2; Schrödinger, LLC: New York, 2009.
(28) Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja, V.; Heath, L. S.; Onufriev, A. *Nucleic Acids Res.* **2005**, *33*, W368−71.
(29) Case, D. A. et al. *Amber 9*; University of California: San Francisco, CA, 2006.
(30) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(31) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(32) Jones, G.; Willett, P.; Glen, R. C. *J. Mol. Biol.* **1995**, *245*, 43−53.
(33) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. *J. Mol. Biol.* **1997**, *267*, 727−748.
(34) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. *Proteins* **2003**, *52*, 609−623.
(35) Yang, W.; Lee, J. Y.; Nowotny, M. *Mol. Cell* **2006**, *22*, 5−13.
(36) Jenkins, T. M.; Esposito, D.; Engelman, A.; Craigie, R. *EMBO J.* **1997**, *16*, 6849−6859.
(37) Simonson, T.; Archontis, G.; Karplus, M. *Acc. Chem. Res.* **2002**, *35*, 430−437.
(38) Case, D. A. et al. *Amber 10*; University of California: San Francisco, CA, 2008.
(39) Pang, Y. P. *Proteins* **2001**, *45*, 183−9.
(40) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Allen, M. P., Tildesley, D. J., Eds.; Oxford University Press: New York, 1989; Vol. *57*, p 385.
(41) Frisch, M. J. et al. *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.
(42) Dapprich, S.; Komaromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *THEOCHEM* **1999**, *461*, 1−21.
(43) Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *9*, 1170−1179.
(44) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357−19363.
(45) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
(46) Ditchfield, R.; Hehre, W. J.; Pople, J. A. *J. Chem. Phys.* **1971**, *54*, 724−&.
(47) Petersson, G. A.; Allaham, M. A. *J. Chem. Phys.* **1991**, *94*, 6081−6090.
(48) Petersson, G. A.; Bennett, A.; Tensfeldt, T. G.; Allaham, M. A.; Shirley, W. A.; Mantzaris, J. *J. Chem. Phys.* **1988**, *89*, 2193−2218.
(49) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650−654.
(50) Mclean, A. D.; Chandler, G. S. *J. Chem. Phys.* **1980**, *72*, 5639−5648.
(51) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908−6918.
(52) Ribeiro, A. J. M.; Ramos, M. J.; Fernandes, P. A. *J. Chem. Theory Comput.* **2010**, *6*, 2281−2292.
(53) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(54) Cossi, M.; Scalmani, G.; Rega, N.; Barone, V. *J. Chem. Phys.* **2002**, *117*, 43−54.
(55) Cossi, M.; Barone, V.; Mennucci, B.; Tomasi, J. *Chem. Phys. Lett.* **1998**, *286*, 253−260.
(56) Mennucci, B.; Tomasi, J. *J. Chem. Phys.* **1997**, *106*, 5151−5158.
(57) Bushman, F. D.; Wang, B. B. *J. Virol.* **1994**, *68*, 2215−2223.
(58) Shibagaki, Y.; Holmes, M. L.; Appa, R. S.; Chow, S. A. *Virology* **1997**, *230*, 1−10.
(59) Vink, C.; Vangent, D. C.; Elgersma, Y.; Plasterk, R. H. A. *J. Virol.* **1991**, *65*, 4636−4644.
(60) Scottoline, B. P.; Chow., S.; Ellison, V.; Brown, P. O. *Genes Dev.* **1997**, *11*, 371−382.
(61) Bujacz, G.; Alexandratos, J.; Wlodawer, A. *J. Biol. Chem.* **1997**, *272*, 18161−18168.
(62) Dupureur, C. M. *Curr. Opin. Chem. Biol.* **2008**, *12*, 250−255.
(63) Warshel, A. *Computer Modeling of Chemical Reactions in Enzymes and Solutions*; John Wiley & Sons: New York, 1997.
(64) Smiley, R. D.; Hammes, G. G. *Chem. Rev.* **2006**, *106*, 3080−94.
(65) Polakowski, R.; Craig, D. B.; Skelley, A.; Dovichi, N. J. *J. Am. Chem. Soc.* **2000**, *122*, 4853−4855.
(66) Lodola, A.; Sirirak, J.; Fey, N.; Rivara, S.; Mor, M.; Mulholland, A. J. *J. Chem. Theory Comput.* **2010**, *6*, 2948−2960.
(67) Zhang, Y.; Kua, J.; McCammon, J. A. *J. Phys. Chem. B* **2003**, *107*, 4459−4463.
(68) Lee, S. P.; Kim, H. G.; Censullo, M. L.; Han, M. K. *Biochemistry* **1995**, *34*, 10205−10214.
(69) Tramontano, E.; La Colla, P.; Cheng, Y. C. *Biochemistry* **1998**, *37*, 7237−7243.
(70) Steitz, T. A.; Steitz, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6498−6502.
(71) Fothergill, M.; Goodman, M. F.; Petruska, J.; Warshel, A. *J. Am. Chem. Soc.* **1995**, *117*, 11619−11627.
(72) Lee, J. Y.; Chang, J.; Joseph, N.; Ghirlando, R.; Rao, D. N.; Yang, W. *Mol. Cell* **2005**, *20*, 155−166.
(73) Stahley, M. R.; Strobel, S. A. *Science* **2005**, *309*, 1587−1590.
(74) Doublie, S.; Tabor, S.; Long, A. M.; Richardson, C. C.; Ellenberger, T. *Nature* **1998**, *391*, 251−258.
(75) Yin, Y. W.; Steitz, T. A. *Cell* **2004**, *116*, 393−404.
(76) Viadiu, H.; Aggarwal, A. K. *Nat. Struct. Biol.* **1998**, *5*, 910−916.
(77) De Vivo, M.; Dal Peraro, M.; Klein, M. L. *J. Am. Chem. Soc.* **2008**, *130*, 10955−62.
(78) Mones, L.; Kulhanek, P.; Florian, J.; Simon, I.; Fuxreiter, M. *Biochemistry* **2007**, *46*, 14514−14523.