Orchestration of cooperative events in DNA synthesis and repair mechanism unraveled by transition path sampling of DNA polymerase β's closing

Ravi Radhakrishnan and Tamar Schlick*

Department of Chemistry and Courant Institute of Mathematical Sciences, 251 Mercer Street, New York University, New York, NY 10012

Edited by Charles R. Cantor, Sequenom, Inc., San Diego, CA, and approved February 10, 2004 (received for review December 23, 2003)

Our application of transition path sampling to a complex biomolecular system in explicit solvent, the closing transition of DNA polymerase β, unravels atomic and energetic details of the conformational change that precedes the chemical reaction of nucleotide incorporation. The computed reaction profile offers detailed mechanistic insights into, as well as kinetic information on, the complex process essential for DNA synthesis and repair. The five identified transition states extend available experimental and modeling data by revealing highly cooperative dynamics and critical roles of key residues (Arg-258, Phe-272, Asp-192, and Tyr-271) in the enzyme's function. The collective cascade of these sequential conformational changes brings the DNA/DNA polymerase β system to a state nearly competent for the chemical reaction and suggests how subtle residue motions and conformational rate-limiting steps affect reaction efficiency and fidelity; this complex system of checks and balances directs the system to the chemical reaction and likely helps the enzyme discriminate the correct from the incorrect incoming nucleotide. Together with the chemical reaction, these conformational features may be central to the dual nature of polymerases, requiring specificity (for correct nucleotide selection) as well as versatility (to accommodate different templates at every step) to maintain overall fidelity. Besides leading to these biological findings, our developed protocols open the door to other applications of transition path sampling to long-time, large-scale biomolecular reactions.

Capturing large-scale, long-time conformational rearrangements in biomolecular systems is a well appreciated central objective in structural and computational biophysics. Such motions are involved in drug binding, enzyme catalysis, protein folding, ion permeation through membrane channels, macromolecular assembly, and chromatin condensation. In many cases, experimental data are available on key structural states, kinetic measurements (e.g., rate of catalysis, effect of salt on reaction), and related mutant or variant systems. Modeling and simulation are thus important to complement experimental data by bridging macroscopic kinetic data with all-atom structures through insights into detailed local motions.

Standard approaches for biomolecules (1), molecular dynamics, Monte Carlo, and other specialized techniques,[†] can generate a rich amount of information concerning structural and dynamic properties for complex systems and connect structure and function through a wide range of thermally accessible states. However, sampling the complex configurational space of biomolecules remains a challenge.

Here we describe the application of transition path sampling (TPS; ref. 18) (for an overview, see *Appendix 1*, which is published as supporting information on the PNAS web site) to a long-time, large-scale biomolecular transition, namely "thumb closing" before chemistry in DNA polymerase β (pol β) complexed to primer/template DNA. This pol β conformational change has been inferred from crystallographic structures (19) for pol β (Fig. 1) and is thought to be key in organizing the active site for DNA synthesis (extension of primer strand by one base) and thereby helping the enzyme discriminate a correct rather than incorrect nucleotide (e.g., C rather than A opposite a G). Because mechanistic details are difficult to obtain experimentally, modeling and simulation, subject to the usual approximations and imperfections of force fields, can help describe events at the atomic level and offer insights into rate-limiting conformational and chemical steps.

Mammalian pol β is ideal for studying polymerase mechanisms for efficient and faithful DNA synthesis: it is the smallest eukaryotic cellular DNA polymerase (20), its activity is thought to be governed by an "induced-fit" mechanism in which the correct incoming base triggers a conformational change (19, 21–26), and a wealth of structural and kinetic data are available. Indeed, this induced-fit mechanism (19, 21–26) between the DNA-bound polymerase and the substrate, in which the correct substrate leads to a "closed," tightly bound enzyme/substrate complex where catalytic groups are aligned as required for proper synthesis ("fidelity"), whereas the incorrect unit misaligns components so that repair is hampered (and can lead to "infidelity"), is supported by a large body of structural and kinetic data for polymerases such as the Klenow fragment, Klentaq 1, and HIV-1 reverse transcriptase. After synthesis or repair, which involves a chemical incorporation of the nucleotide in DNA by means of phosphodiester bond formation, the enzyme complex returns to its "open" state, releasing the DNA to allow translocation for the next cycle of polymerization.

Our path sampling of pol β's closing before the chemical reaction employs the state-of-the-art CHARMM force field and is subject to the well documented limitations of modern force fields; nonetheless, the details we unravel, which tie well with a large body of experimental as well as modeling studies, refine polymerase mechanisms by presenting a more detailed and complex view of polymerase kinetic cycles. Specifically, we identify key slow motions and interpret their significance in the context of pol β's reaction pathway. Namely, the calculated reaction profile reveals a cascade of subtle conformational rearrangements rather than subdomain motion *per se* that brings the DNA/protein system to a state nearly competent for the chemical reaction. This complex orchestration of events introduces checks and balances in the enzyme complex and likely helps in fidelity discrimination.

Shaped like a hand, with thumb, palm, and fingers subdomains (Fig. 1), pol β fills single-stranded gaps in DNA with relatively high accuracy (fidelity): it recruits the nucleotide unit (dNTP,

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: TPS, transition path sampling; pol β, DNA polymerase β.

*To whom correspondence should be addressed. E-mail: schlick@nyu.edu.

[†]Activated and long-time processes can be studied by using simplified models [for electrostatics (2), long DNA (3), or proteins (4, 5)], dedicated supercomputing resources (2), Monte Carlo/molecular dynamics combinations (6), aggregate dynamics (7), stochastic models (2, 8, 9), stochastic path approach (10, 11), or multiple or replica dynamics (9, 12). When experimental anchors are available, biomolecular systems can be "steered," "guided," or targeted (13) to study folding/unfolding events and gain valuable insights into common pathways (14–17).

© 2004 by The National Academy of Sciences of the USA

www.pnas.org/cgi/doi/10.1073/pnas.0308585101

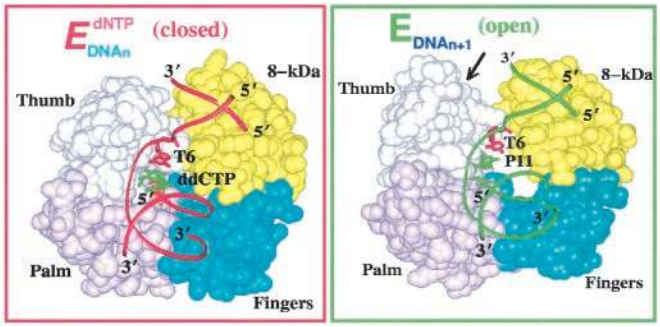

Fig. 1. pol β in open (*Right*) and closed (*Left*) states (19).

2′-deoxyribonucleoside 5′-triphosphate) complementary to the template base (e.g., C opposite G) about 1,000 times more often than the incorrect unit (e.g., T opposite G) (27–33). Closed and open conformations of pol β have been solved by x-ray crystallography by Wilson and coworkers (19, 34) and are related by a large subdomain motion (≈6 Å) of the thumb (Fig. 1, arrow). Kinetic data (28–30, 35–43) suggest slow conformational changes (milliseconds to seconds) before and after the chemical reaction, but their identity is unknown. Prior simulations by molecular and Langevin dynamics simulations, as well as targeted molecular dynamics (TMD), have suggested that key residues in the enzyme active site (e.g., Phe-272 and Arg-258) exhibit subtle conformational rearrangements during the thumb subdomain's motion (14, 15, 44). While these suggestions tie well with, and refine, recent experimental studies (45), details of the reaction pathway (Fig. 6, which is published as supporting information on the PNAS web site) explaining fidelity of the repair or synthesis process remain unknown.

Our goal is to delineate at atomic resolution the sequence of events in the overall catalytic mechanism of DNA polymerases along with energetic estimates and possibly to refine the schematic view of Fig. 6, which focuses on subdomain motions. This refinement is not possible with standard simulation methods that capture short-time dynamics, techniques that employ biasing potentials (13), or approaches that are subject to the approximations of transition-state theory in locating transition-state regions. TPS developed in the Chandler group (18) yields true dynamics at transition states by a clever sampling that conserves equilibrium distributions and the fluctuation-dissipation theorem (46): complete reaction profiles, transition states, barrier free energies, and reaction rates, for the system can ultimately be obtained. However, path sampling has been applied only to relatively small systems [alanine dipeptide isomerization (47), dissociation of water in the condensed phase (48), methanol coupling in zeolites, ice nucleation (49), folding of oligomers modeled as hard spheres (50), and β-hairpin folding (51)]. The challenge of applying TPS to a macromolecular system is certainly practical, to detect and then piece in correctly all the biochemically relevant transition regions, but new procedures are required to assess convergence and estimate free energy barriers. By addressing these computational issues for macromolecules in explicit solvent (our complete protocol is in *Appendix 2*, which is published as supporting information on the PNAS web site), we delineate here key transition-state regions and the associated cooperative dynamics in DNA pol β's closing before the chemical reaction; we also discuss how this collective cascade of subtle conformational changes may play a critical role in pol β's function.

Computational Method

System Preparation. Two models of solvated pol β/DNA/dCTP complexes were prepared from 1BPX (open binary) and 1BPY (closed ternary) crystal structures (19). Hydrogen atoms were added by CHARMM's subroutine HBUILD (52). The open complex was modified by incorporating the incoming unit dCTP (deoxyribocytosine 5′-triphosphate) with nucleotide-binding Mg^{2+}, producing the 1BPX ternary complex.

Also added were the following: a hydroxyl group to the 3′ terminus of the primer DNA strand, missing residues 1–9 of pol β, and specific water molecules coordinated to the catalytic Mg^{2+} (missing in the ternary complex).

Cubic periodic domains for both initial models were constructed by using SIMULAID and PBCAID (53). To neutralize the system at an ionic strength of 150 mM, water molecules with minimal electrostatic potential at the oxygen atoms were replaced by Na^+, and those with maximal electrostatic potential were replaced with Cl^-. All Na^+ and Cl^- ions were placed more than 8 Å away from any protein or DNA atoms and from each other. The electrostatic potential for all bulk oxygen atoms was calculated with DELPHI. The resulting system (see *Lower Center* image in Fig. 4) has 40,238 atoms (including 11,249 water molecules). Consistent with a pH value of 7.0, we assume deprotonated states (i.e., −1 charge each) for Asp-190, Asp-192, and Asp-256, as assumed recently (54). *Appendix 3*, which is published as supporting information on the PNAS web site, provides protonation states of titratable side chains and a brief related discussion. These settings produce a net charge of +7 for pol β, −29 for DNA, and −4 for the dCTP. There are 42 Na^+ ions, 20 Cl^- ions, and 2 Mg^{2+} ions, producing an overall neutral system.

Initial Runs. To generate configurations for initiating path sampling, targeted dynamics were conducted with CHARMM C28a2 (52), implemented with a restraint constant K = 2,000 kcal/(mol·Å^2) on heavy atoms of pol β (see *Appendix 2*). With a decrease from 5 to 0 Å in the offset distance (d_0) over 100 ps, the conformational change is driven from open to closed states. Although unphysical, the targeted trajectory helps suggest transition-state regions. Namely, histograms of geometric variables, dihedral angles associated with key residues 192, 258, and 272 (flip of Asp-192, rotation of Arg-258, and flip of Phe-272), and the thumb's helix N displacement (residues 275–295) that display bimodality lead us to propose candidates for the path sampling order parameters $\{\chi_i\}$, as described in Table 1 and Fig. 2.

Identifying Transition-State Regions. The existence of transition-state regions 2, 3, and 4 was confirmed by calculating the commitment probability distribution (CPD) (47) by using several short (20–100 ps), unrestrained dynamics trajectories. The CPD describes the partitioning of trajectories in proximal free-energy basins near each transition state region (47) (*Appendix 2*) from the transition state ensemble. For example, barrier region 3, partial rotation of Arg-258, corresponds to the dihedral-angle window between the dashed lines for χ_3's evolution in Fig. 7, which is published as supporting information on the PNAS web site (see also Fig. 8 in *Appendix 1*). For each configuration in such a barrier region, four trajectories were initiated with a randomly chosen set of momenta from a Maxwell distribution, and the commitment probability P_B is determined as the fraction of trajectories committing to basin B (i.e., $\chi_3 > 160°$; see Table 1); thus, resolution of the commitment probability is 0.25. From P_B, the probability distribution $P(P_B)$ yields the CPD function of Fig. 7. In general, the initial (unphysical) trajectory (Fig. 7) does not lead to a unimodal function and lacks the peak at 1/2 indicating correct barrier crossing (18) (where $P_B \sim P_A \sim 0.5$). Because the ensemble for Arg-258 committed to an intermediate value of χ_3, the existence of a different barrier region (TS 5) was suggested.

Performing Path Sampling. Our protocol (*Appendix 2*) was performed by using a PERL script that interfaces CHARMM C28a2

Table 1. Order parameters for each transition state (TS) and associated simulation details

TS	Event	χ-order parameter*	χ_{max} state *A*	χ_{min} state *B*	Length, ps TPS/CPD†	No. of trajectories	τ_{mol}, ps‡
1	Partial thumb motion, Watson–Crick pairing of dCTP	rms deviation of residues 275–295 with respect to closed form	3.4 Å	2.7 Å	100/100	150	70 ± 10
2	Asp-192 flip	Dihedral angle C^{γ}–C^{β}–C^{α}–C	100°	140°	10/20	200	4 ± 1
3	Arg-258 rotation, thumb closing	Dihedral angle C^{γ}–C^{δ}–N^{ϵ}–C^{ζ}	120°	160°	10/20	200	4 ± 1
4	Phe-272 flip	Dihedral angle $C^{\delta 1}$–C^{γ}–C^{β}–C^{α}	−25°	15°	10/20	200	3 ± 1
5	Arg-258 rotation, ion motion	Distance: nucleotide-binding Mg^{2+} to $O1^{\alpha}$ of dCTP	—	—	—	—	–

*Our coverage of each TS is given by χ_{min}, χ_{max}, along with lengths τ of TPS simulations and of additional runs for convergence (see Figs. 2 and 7).
†CPD, commitment probability distribution function (see Fig. 7).
‡See Fig. 7 for estimating τ_{mol}.

(52) and handles multiple transition-state regions characterized by sets of arbitrary dihedral angles, distances, and any calculable configurational quantity. CHARMM's Verlet integrator with a time step of 1 fs is used for generating symplectic dynamics trajectories, with electrostatic and van der Waals interactions smoothed to 0 at 12 Å and cubic periodic boundary conditions enforced. The truncation of long-range electrostatic interactions makes the computation of our 40,238-atom system tractable; studies on DNA (e.g., ref. 55) conclude that this approach is satisfactory.

Path segments corresponding to barrier regions (generated as described above, Fig. 7) were used to initiate four different path sampling runs corresponding to TS regions 1–4 (Table 1). Trajectories in each region were harvested by using the shooting algorithm (56) (*Appendix 1* and Fig. 8 *Inset*) to connect by an ensemble two metastable states *A* and *B* by means of a Monte Carlo protocol in trajectory space. States *A* and *B* are defined by our order parameters (Table 1): for TS 1, $\chi_1 > \chi_{max} \in A$ and $\chi_1 < \chi_{min} \in B$; for TS 2–4, $\chi_i < \chi_{max} \in A$ and $\chi_i > \chi_{min} \in B$. In each shooting run, the momentum perturbation size ($dP \approx 0.002$ for TS 1, and 0.005–0.01 for TS 2–4, in units of atomic mass units ×Å/fs) yields an acceptance rate of 30–40%.

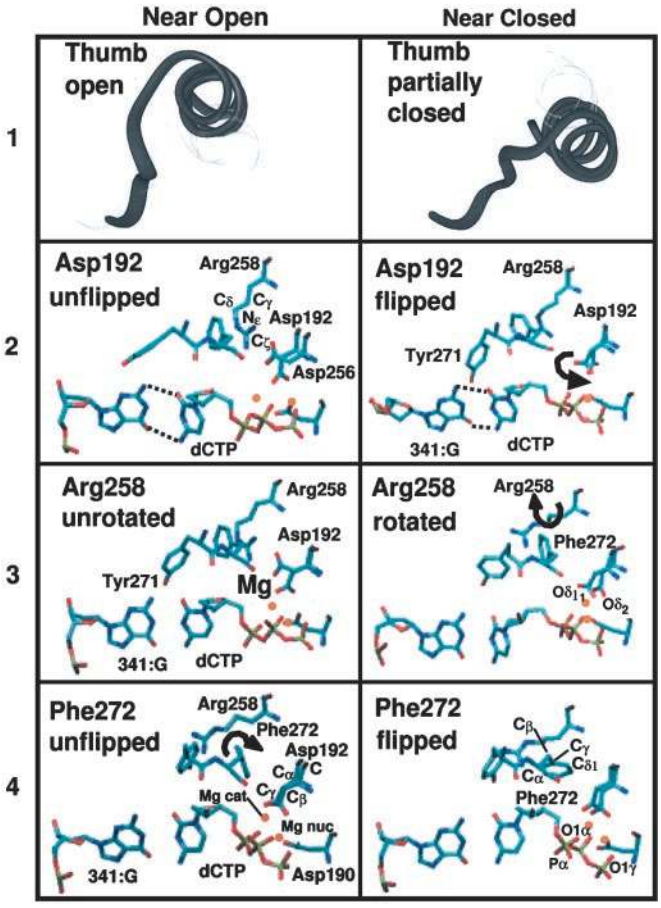

Fig. 2. Molecular snapshots near open (*Left*) and closed (*Right*) states for four transition state regions. (*1*) Partial thumb closing. (*2*) Asp-192 flip. (*3*) Arg-258 partial rotation. (*4*) Phe-272 flip.

Monitoring Adequate Sampling and Convergence. Five sample harvested trajectories (of 150–200 total trajectories in each ensemble; see Table 1) capturing the dynamics for each region 1–4 are shown in Fig. 7. The ergodicity and convergence of each run were monitored by calculating the autocorrelation function of the order parameter $\langle \chi_i(0)\chi_i(t)\rangle$ (Fig. 7, *Top Right*). The gradual decrease for χ_1 and increase for χ_2, χ_3, and χ_4 indicate the decorrelation of the generated trajectories in each TPS run; strongly correlated trajectories would lead to an abrupt change in the correlation function for the chosen values of τ. The characteristic relaxation time τ_{mol} associated with the crossing of each transition-state region is given by the time taken for the gradual transition of the autocorrelation function $\langle \chi_i(0)\chi_i(t)\rangle$ and can be estimated graphically, to yield values in Table 1. The sampling quality was also assessed by calculating order-parameter correlation functions in path space and found to be very satisfactory (see *Appendix 2*).

The convergence of the overall harvested path (collection of 150–200 trajectories in each region) was verified by recalculating commitment probability distribution functions (Fig. 7 *Right*). Unimodal distribution peaks around 1/2 reflect the true saddle nature of the transition state region. The striking contrast between the distributions of the converged path sampling runs and those for the initial trajectory reveals the system's relaxation along coordinates orthogonal to the reaction coordinate, to reach true saddle regions, as discussed in ref. 18.

Piecing Entire Reaction Path. The combined path sampling simulations (750 trajectories of length 10–100 ps) were extended by 20 dynamics simulations of lengths 500 ps to 1 ns to ensure that pol β open and closed structures were visited‡ and trajectories ergodically sampled the phase space between the five barrier regions (see Fig. 3). The decorrelation from the initial path during our path sampling is significant, indicating independence from initial conditions (see Fig. 9 in *Appendix 2*).

TS region 5 was discovered when extending the trajectories

‡Criteria for reaching endpoints are based on rms deviation <2 Å of helix N residues with respect to crystal structures and active-site distances and coordinations (of the two Mg^{2+} ions and P^{α} of dCTP) conforming to two-metal-ion mechanism geometry; see Fig. 5.

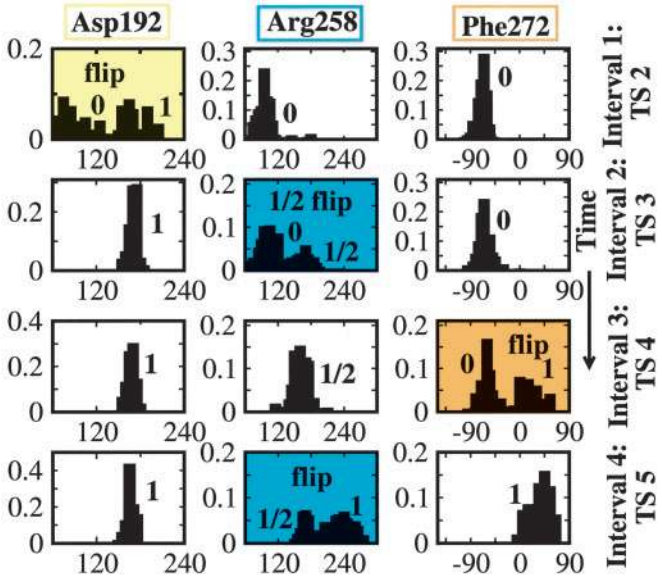

Fig. 3. Normalized probability distributions of dihedral angles (in degrees) characterizing flips of pol β residues. Each column represents a dihedral angle distribution and each row represents a time interval capturing the dynamics in a barrier region. Labels 0, 1/2, and 1 refer to unflipped, partially flipped, and fully flipped states. Highlighted bimodal plots correspond to flipping events.

associated with TS region 4. In overcoming TS region 5, the fully rotated state of Arg-258 is stabilized, and the ligands coordinating the nucleotide binding Mg^{2+} ion undergo subtle rearrangements (see *Results*).

Normalized probability distributions (i.e., unit areas) for the dihedral angles characterizing the flips Asp-192, Arg-258, and Phe-272 are calculated in Fig. 3, from the harvested trajectories. The nonoverlapping nature of the distribution for each residue (i.e., within each column) confirms independence of the motions; the transient time τ_{mol} (Table 1) to cross each transition-state region is much shorter than the time separating the events, justifying performing independent runs for each flip. The combined distribution of dihedral angles for each residue samples the dihedral space between the open and closed conformations of pol β, further ensuring that we have captured intermediate configurations between the barrier regions. The dihedral-angle distributions clearly show that the Asp-192, partial-Arg-258, Phe-272, and full-Arg-258 flip events are local barrier regions in the free energy surface.

The order of the transition states along the pathway (see Fig. 4) was determined by ranking the values of the order parameters in the metastable states. For example, because χ_2 changes from unflipped (in open) to flipped (in closed) values while χ_3, χ_4, and χ_5 remain at values of the open structure (see Fig. 3), we know that TS 2 precedes TS 3–5. The independence of flips associated with barrier regions 2–5 suggests that key residues of the enzyme act discretely to trigger a cascade of subtle changes resulting in systematic assembly of the catalytic region. Near the open and closed states (see Fig. 2) the χ_i values range as follows: χ_1 from $\approx$6 to 1.5 Å, χ_2 from $\approx$90° to 180°, χ_3 from $\approx$100 to 260°, χ_4 from $\approx -50°$ to 50°, and χ_5 from $\approx$3.2 to 1.8 Å.

The combined simulations involved an aggregate time of $\approx$85 ns, requiring 10 months of central processing unit time (approximately half of which was spent on the free energy calculations) on 24 processors of an Origin 3000, 120-MHz processor Silicon Graphics machine at New York University.

Results: Sequence of Conformational Changes

Fig. 4 illustrates the sequence of changes along the pathway: *1*, partial thumb closing; *2*, flip of Asp-192; *3*, partial rotation of Arg-258; *4*, flip of Phe-272; and *5*, rearrangement of catalytic region and stabilization of Arg-258 in the fully rotated state. The approximate free energies shown were calculated by using our method BOLAS (unpublished work and *Appendix 2*), separate from the path sampling protocol.

The conformational landscape that emerges reveals highly cooperative events. Starting from the open ternary complex, the system traverses metastable basin 2 as the thumb subdomain begins to close, during which repuckering of the sugar ring of dCTP occurs: the pseudorotation phase angle changes from about 160° (C2′-*endo*) to 25° (C3′-*endo*) (57) (with mean puckering amplitude, $\bar{\tau}_{max}$, constant at $\approx$44°), facilitating the pairing of dCTP with the guanine template partner. Crossing TS 1 leaves the system in metastable basin 3, in which the thumb is partially closed. The partial thumb closing corresponds to a large subdomain motion during which the rms deviation of helix N residues with respect to the closed structure changes from around 6 to 2.5 Å. The associated dynamics reveal cooperative motion among helix N (thumb) residues, incoming dCTP, guanine template partner, and conserved residue Tyr-271, resulting in the Watson–Crick pairing of dCTP with the template G. Thus, TS 1 selects and helps discriminate the correct incoming dNTP. A similar motion coupling, with biological significance, has been inferred in a recent study involving DNA polymerase I of the A family (58), where the thumb moves in concert with the conserved tyrosine residue (Tyr-714 in DNA polymerase I); this cooperativity has been suggested to block and unblock a preinsertion site that likely shields the template partner against inducing frameshift mutations, thereby enhancing fidelity. Our results for pol β (which lacks such a preinsertion site) suggest that the cooperativity of the thumb motion with Tyr-271, dCTP, and template partner acts similarly as a signaling network for correct nucleotide selection.

TS regions 2–5 define a cascade of global and side-chain motions that (*i*) transition the catalytic region from a disordered arrangement to one nearly ready for the two-metal-ion (Mg^{2+})-catalyzed phosphoryl transfer reaction (59); and (*ii*) stabilize the polymerase in the closed (active) form. With Asp-192's flip (TS 2), the system reaches basin 4. Asp-192's flip occurs concomitantly with the breaking of the salt bridge between Asp-192 and Arg-258 and facilitates the aspartate oxygens $O^{\delta 1}$ and $O^{\delta 2}$ to coordinate the catalytic and nucleotide-binding magnesium ions (see Fig. 4 *4*). Interestingly, the path sampling snapshots in Fig. 4 indicate that the Na^+ ions are mobile in the vicinity of the active site (ions were initially placed 8 Å from each other and from the enzyme/DNA/dCTP complex), as also supported by separate modeling studies (L. Yang, K. Arora, and T.S., unpublished work).

The partial rotation of Arg-258 occurs in crossing TS 3, during which the thumb assumes the fully closed position; the rms deviation (χ_1) decreases from $\approx$2.4 to 1.7 Å between metastable basins 4 and 5. Once in basin 5, Arg-258 interchanges between the partially and fully rotated positions. After crossing TS 4, the flipped state of Phe-272 (χ_4 = 50°) sterically clashes with Arg-258 in its partially rotated state. Thus, metastable basin 6 is a precursor for stabilizing Arg-258 in its fully rotated state. Metastable basin 7 (after crossing TS 5) is characterized by the stabilization of Arg-258 in its fully rotated state, where it participates in a salt bridge with residues Tyr-296 and Glu-295.

Concomitant to the motion in crossing transition states 2–5, the position and coordination of the catalytic Mg^{2+} ion undergo subtle, yet systematic, transformations (Fig. 5) that evolve the system toward the catalytically active configuration. The three conserved aspartate residues (190, 192, and 256 in pol β) coordinate the Mg^{2+} ions, creating a tight binding pocket for the metal ions. The formation of two octahedral complexes (involving the Mg^{2+} ions and their respective ligands), and the transformation to a nearly reaction-

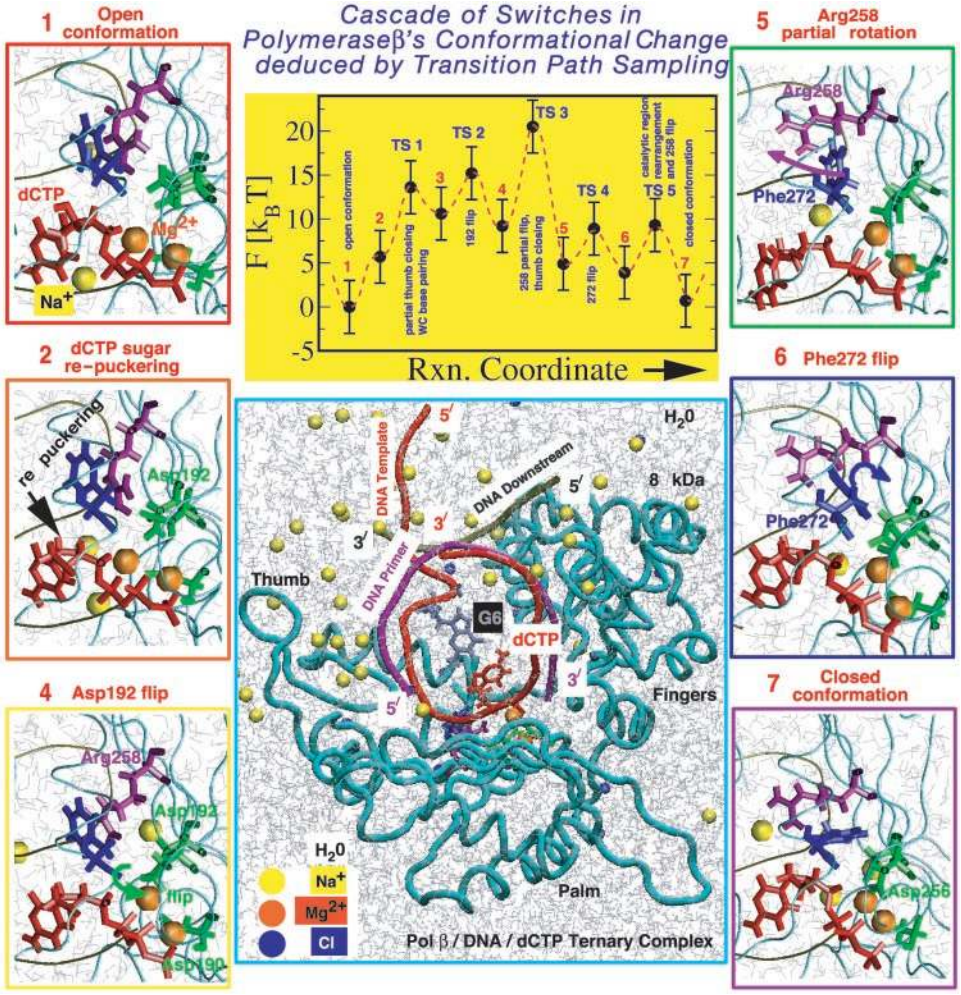

Fig. 4. (*Upper Center*) Overall captured reaction kinetics profile from TPS for pol β's closing transition (for a G·C system) from the open state and estimated energies in units of k_BT (product of Boltzmann constant and absolute temperature). Different barrier regions (TS) and metastable basins along the reaction coordinate are illustrated (images *1*, *2*, *4–6*, and *7*). (*Lower Center*) Structure of pol β/DNA/dCTP ternary complex.

competent state [as envisioned in the two-metal-ion catalytic mechanism (59)], is completed in a cascade of changes involving the aspartate oxygens of Asp-192 and Asp-256 and the Mg^{2+} ions (TS 2–4), and in a subtle rearrangement of the Mg^{2+} ions (TS 5), as evident from Fig. 5.

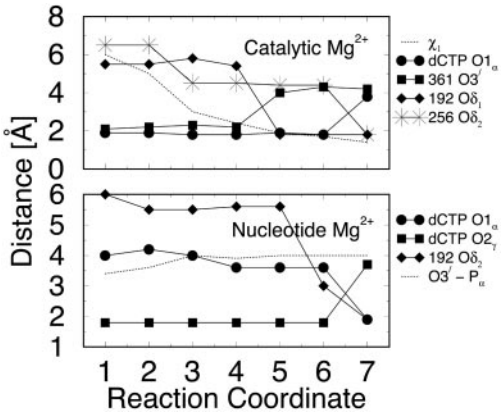

Fig. 5. Evolution of average distances of ligands coordinating the catalytic and nucleotide-binding Mg^{2+} ions along the reaction coordinate. For atom types, see Fig. 2. States 1–7 correspond to Fig. 4. The extent of thumb closing (χ_1 in *Upper*), and a crucial distance for the chemical reaction (O3′ of last primer residue to P^{α} of dCTP in *Lower*) are also provided.

Additional transition state regions, outside of the reaction coordinate considered in Fig. 4, cannot be excluded. For instance, in our open starting conformation, the DNA was already bound to pol β, and the two Mg^{2+} ions along with the dCTP molecule were already in the vicinity of the catalytic region. The binding events of dCTP, Mg^{2+} ions, and DNA could represent additional transition states, although experimental evidence from kinetic studies suggests that they are not rate limiting (27).

Discussion: Rate-Limiting Steps and Implications for Polymerase Efficiency and Fidelity

The developed divide-and-conquer procedure described here for trajectory generation and analysis makes possible the application of TPS to biomolecular systems. Of course, all such simulations are vulnerable to uncertainties in force field parameters, even in the carefully developed and tested CHARMM package, and require careful assessment with respect to available experimental and modeling studies. Although the generated trajectories differ from each other in microscopic details, the collective picture that emerges here suggests a dominant pathway for the closing of the pol β/DNA complex for the correct (G·C) system. This pathway is consistent with the biological relevance of the pol β conformational transition (i.e., selection of incomer dNTP, followed by

assembling of the catalytic region for nucleotidyl transfer reaction). The unraveled complex cascade of subtle conformational rearrangements also introduces "checks and balances" on the nucleotide incorporation process.

Our findings of subtle residue motions rather than subdomain motion *per se* tie well with a large body of experiments that reveal reduced catalytic efficiency of the enzyme (19, 21–26, 45) when some of these key residues (258, 192, and 272) and those associated with the metastable nearby basins (such as 271) are altered. However, our specific description of sequential conformational events and associated energetics (see below) pieces together and extends these kinetic data and findings of prior modeling studies by suggesting a global significance of these collective local changes: a tight orchestration of events at the active site may be essential to the enzyme's function. This view refines the catalytic cycle of Fig. 6 by extending the mechanisms beyond simple subdomain motions. Significantly, a cascade of events guides the enzyme toward the chemistry step; different conditions at the active site may sensitively affect this delicate steering process. Indeed, we expect that the process for a mismatched system (e.g., G·A) at the primer/template terminus will result in multiple pathways rather than a dominant pathway as found for our G·C system here; these alternatives may hamper the closing transition, allowing removal of the incorrect nucleotide unit near the active site.

Significantly, our estimated free energy values along the closing pathway (Fig. 4) suggest that the rate-limiting step is associated with TS 3, namely the partial rotation of Arg-258 (overall barrier of 20.5 $\pm$ 3 k_BT). This value is lower but comparable to the overall rate of the reaction (25–28 k_BT), as obtained from experimental kinetic data [k_{pol}, the overall rate of nucleotide incorporation of 3–90 s^{-1} (27–33)]. This rate-limiting conformational barrier in the same order of the overall (conformational and chemical steps) reaction profile immediately suggests site-specific mutant experiments on Arg-258 to assess Arg-258's effect on reaction efficiency and fidelity. For example, we expect that the substitution of Ala for Arg will reduce the TS 3 barrier and possibly accelerate the conformational closing (i.e., lead to a higher synthesis efficiency and lower k_{pol} value).

Ongoing work in our laboratory combined with a *tour de force* study of the chemical reaction in T7 DNA polymerase (54) also suggest that chemical-reaction barriers define the overall rate-limiting step, consistent with analysis of Showalter and Tsai (60). Still, even in the chemical barrier identified (54), crucial motions of selected residues such as conserved acidic groups in the active site emerge. Thus, while the identification of the single, overall rate-limiting step, as well as events leading up to it or affecting it, have important biological implications for fidelity (60), our results underscore the parallel need to understand at atomic resolution the complex sequence of events in the overall catalytic mechanism of DNA polymerases and to refine the somewhat simplistic kinetic pathway (Fig. 6). These refinements may ultimately lead to design of site-specific therapeutic agents that address diseases caused by DNA polymerase malfunction such as skin cancer and premature aging.

Possible extensions of path sampling to treat the chemical reaction by a combined molecular/quantum mechanical treatment can also be envisioned to further illuminate polymerase mechanisms. Application of other long-time methods that have potential for capturing large-scale motion in biomolecules and associated free energy profiles, such as the stochastic path approach of Elber *et al.* (10, 61), enhanced sampling using Tsallis statistics (62), and the adaptive sampling approach of Vanden-Eijnden and coworkers (E. Weinan, W. Ren, and E. Vanden-Eijnden, personal communication) may also be fruitful.

We thank David Chandler for helpful advice and for clarifying the link between path sampling and reactive flux formalism described in *Appendix 4*, which is published as supporting information on the PNAS web site. We thank Samuel Wilson, William Beard, and Suse Broyde for helpful comments on this work. We are indebted to Linjing Yang and Karunesh Arora for many stimulating discussions throughout this work and for making the data in refs. 14, 15, and 44 available to us for analysis. We thank Martin Karplus for use of the CHARMM program. This research was supported by National Institutes of Health Grant R01 GM55164, National Science Foundation Grant MCB-0239689, and the donors of the American Chemical Society Petroleum Research Fund.

1. Schlick, T. (2002) *Molecular Modeling and Simulation: An Interdisciplinary Guide* (Springer, New York), pp. 345–462.
2. Duan, Y. & Kollman, P. A. (1998) *Science* **282,** 740–744.
3. Schlick, T., Beard, D. A., Huang, J., Strahs, D. & Qian, X. (2000) *IEEE Comput. Sci. Eng.* **2,** 38–51.
4. Brooks, C. L. (2002) *Acc. Chem. Res.* **35,** 447–454.
5. Gan, H. H., Tropsha, A. & Schlick, T. (2000) *J. Chem. Phys.* **113,** 5511–5524.
6. Zhou, R., Berne, B. J. & Germain, R. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 14931–14936.
7. Snow, C. D., Nguyen, H., Pande, V. S. & Gruebele, M. (2002) *Nature* **420,** 102–106.
8. Daggett, V. (2000) *Curr. Opin. Struct. Biol.* **10,** 160–164.
9. Zagrovic, B., Sorin, E. J. & Pande, V. (2001) *J. Mol. Biol.* **313,** 151–169.
10. Elber, R., Cárdenas, A., Ghosh, A. & Stern, H. (2003) *Adv. Chem. Phys.* **126,** 93–129.
11. Zaloj, V. & Elber, R. (2000) *Comput. Phys. Commun.* **128,** 118–127.
12. Daura, X., Jaun, B., Seebach, D., Gunsteren, W. F. V. & Mark, A. (1998) *J. Mol. Biol.* **280,** 925–932.
13. Schlick, T. (2003) *Biophys. J.* **85,** 1–4.
14. Yang, L., Beard, W. A., Wilson, S. H., Broyde, S. & Schlick, T. (2002) *J. Mol. Biol.* **317,** 651–671.
15. Yang, L., Beard, W. A., Wilson, S. H., Roux, B., Broyde, S. & Schlick, T. (2002) *J. Mol. Biol.* **321,** 459–478.
16. Young, M. A., Gonfloni, S., Superti-Furga, G., Roux, B. & Kuriyan, J. (2001) *Cell* **105,** 115–126.
17. Ferrara, P., Apostolakis, J. & Caflisch, A. (2000) *Proteins* **39,** 252–260.
18. Bolhuis, P. G., Chandler, D., Dellago, C. & Geissler, P. L. (2002) *Annu. Rev. Phys. Chem.* **53,** 291–318.
19. Sawaya, M. R., Prasad, R., Wilson, S. H., Kraut, J. & Pelletier, H. (1997) *Biochemistry* **36,** 11205–11215.
20. Wilson, S. H. (1998) *Mutat. Res.* **407,** 203–215.
21. Li, Y., Korolev, S. & Waksman, G. (1998) *EMBO J.* **17,** 7514–7525.
22. Doublié, S. & Ellenberger, T. (1998) *Curr. Opin. Struct. Biol.* **8,** 704–712.
23. Kiefer, J. R., Mao, C., Braman, J. C. & Beese, L. S. (1998) *Nature* **391,** 302–305.
24. Koshland, D. E. (1994) *Angew. Chem. Int. Ed. Engl.* **33,** 2375–2378.
25. Beard, W. A. & Wilson, S. H. (1998) *Chem. Biol.* **5,** R7–R13.
26. Doublié, S., Sawaya, M. R. & Ellenberger, T. (1999) *Structure* **7,** R31–R35.
27. Ahn, J., Werneburg, B. G. & Tsai, M.-D. (1997) *Biochemistry* **36,** 1100–1107.
28. Ahn, J., Kraynov, V. S., Zhong, X., Werneburg, B. G. & Tsai, M.-D. (1998) *Biochem. J.* **331,** 79–87.
29. Vande Berg, B. J., Beard, W. A. & Wilson, S. H. (2001) *J. Biol. Chem.* **276,** 3408–3416.
30. Shah, A. M., Li, S.-X., Anderson, K. S. & Sweasy, J. B. (2001) *J. Biol. Chem.* **276,** 10824–10831.
31. Werneburg, B. G., Ahn, J., Zhong, X., Hondal, R. J., Kraynov, V. S. & Tsai, M.-D. (1996) *Biochemistry* **35,** 7041–7050.
32. Beard, W. A., Osheroff, W. P., Prasad, R., Sawaya, M. R., Jaju, M., Wood, T. G., Kraut, J., Kunkel, T. A. & Wilson, S. H. (1996) *J. Biol. Chem.* **271,** 12141–12144.
33. Menge, K. L., Hostomsky, Z., Nodes, B. R., Hudson, G. O., Rahmati, S., Moomaw, E. W., Almassy, R. J. & Hostomska, Z. (1995) *Biochemistry* **34,** 15934–15942.
34. Pelletier, H., Sawaya, M. R., Kumar, A., Wilson, S. H. & Kraut, J. (1994) *Science* **264,** 1891–1903.
35. Suo, Z. & Johnson, K. A. (1998) *J. Biol. Chem.* **273,** 27250–27258.
36. Kraynov, V. S., Werneburg, B. G., Zhong, X., Lee, H., Ahn, J. & Tsai, M.-D. (1997) *Biochem. J.* **323,** 103–111.
37. Zhong, X., Patel, S. S., Werneburg, B. G. & Tsai, M.-D. (1997) *Biochemistry* **36,** 11891–11900.
38. Dahlberg, M. E. & Benkovic, S. J. (1991) *Biochemistry* **30,** 4835–4843.
39. Kuchta, R. D., Mizrahi, V., Benkovic, P. A., Johnson, K. A. & Benkovic, S. J. (1987) *Biochemistry* **26,** 8410–8417.
40. Wong, I., Patel, S. S. & Johnson, K. A. (1991) *Biochemistry* **30,** 526–537.
41. Patel, S. S., Wong, I. & Johnson, K. A. (1991) *Biochemistry* **30,** 511–525.
42. Frey, M. W., Sowers, L. C., Millar, D. P. & Benkovic, S. J. (1995) *Biochemistry* **34,** 9185–9192.
43. Capson, T. L., Peliska, J. A., Kaboord, B. F., Frey, M. W., Lively, C, Dahlberg, M. & Benkovic, S. J. (1992) *Biochemistry* **31,** 10984–10994.
44. Yang, L., Broyde, S., Beard, W. A., Wilson, S. H. & Schlick, T. (2004) *Biophys. J.* **86,** in press.
45. Beard, W. A., Shock, D. D., Berg, B. J. V. & Wilson, S. H. (2002) *J. Biol. Chem.* **277,** 47393–47398.
46. Crooks, G. E. (1999) Ph.D. Thesis (University of California, Berkeley).
47. Bolhuis, P. G., Dellago, C. & Chandler, D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 5883–5888.
48. Geissler, P. L., Dellago, C. & Chandler, D. (1999) *Phys. Chem. Chem. Phys.* **1,** 1317–1322.
49. Radhakrishnan, R. & Trout, B. L. (2003) *Phys. Rev. Lett.* **90,** 158301.
50. ten Wolde, P. R. & Chandler, D. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 6539–6543.
51. Bolhuis, P. G. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 12129–12134.
52. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983) *J. Comput. Chem.* **4,** 187–217.
53. Qian, X., Strahs, D. & Schlick, T. (2001) *J. Comput. Chem.* **22,** 1843–1850.
54. Florián, J., Goodman, M. F. & Warshel, A. (2003) *J. Am. Chem. Soc.* **125,** 8163–8177.
55. Norberg, J. & Nilsson, L. (2000) *Biophys. J.* **79,** 1537–1553.
56. Bolhuis, P. G., Dellago, C. & Chandler, D. (1998) *Faraday Discuss. Chem. Soc.* **110,** 421–436.
57. Altona, C. & Sundaralingam, M. (1972) *J. Am. Chem. Soc.* **94,** 8205–8212.
58. Johnson, S. J., Taylor, J. S. & Beese, L. S. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 3895–3900.
59. Steitz, T. A. & Steitz, J. A. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 6498–6502.
60. Showalter, A. K & Tsai, M.-D. (2002) *Biochemistry* **41,** 10571–10576.
61. Arora, K. & Schlick, T. (2003) *Chem. Phys. Lett.* **378,** 1–8.
62. Barth, E. J., Laird, B. B. & Leimkuhler, B. J. (2003) *J. Chem. Phys.* **118,** 5759–5768.

Supplementary Information

Figure 6: Pol β's Reaction Pathway

General pathway for nucleotide insertion by DNA polymerases (a), and corresponding crystal closed (b, red) and open (c, green) conformations of a pol β/DNA complex. E: DNA polymerase; dNTP: 2′-deoxyribonucleoside 5′-triphosphate; PP_i: pyrophosphate; DNA_n/DNA_{n+1}: DNA before/after nucleotide incorporation to DNA primer. Binding of correct nucleotides (step 1) induces a conformational change from an open to a closed pol β state (step 2); incorrect units may inhibit or alter this change. The 'thumb' in left-handed pol β corresponds to the 'fingers' in right-handed DNA polymerases. After nucleotide incorporation (step 3), a closed-to-open rearrangement (step 4) occurs prior to the release of product pyrophosphate (step 5). (Figure kindly provided by Linjing Yang).

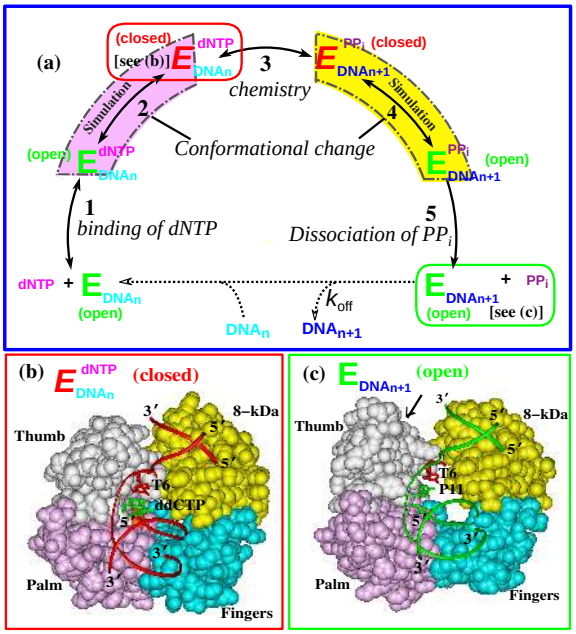

Figure 6: Pol β's reaction pathway.

Figure 7: Convergence Analysis

Left: For four transition state (TS) regions, 5 sample trajectories out of 150–200 harvested in path sampling are shown. The dihedral-angle interval between the dashed lines in TS 3 defines the TS ensemble for Arg258's partial-rotation. **Right (top):** Order parameter autocorrelation functions $\langle \chi_i(0)\chi_i(t)\rangle$ (in units of Å^2 for TS 1, and rad^2 for TS 2–4), where $\langle\cdot\rangle$ denotes the average over the ensemble of generated trajectories. Autocorrelation functions are plotted with initial point $\langle \chi_i(0)\chi_i(0)\rangle \approx \langle \chi_A\rangle^2$ and end point $\langle \chi_i(0)\chi_i(\tau)\rangle \approx \langle \chi_A\rangle\langle \chi_B\rangle$, to indicate crossing the barrier region between A and B over time τ (see Fig. 8 in Appendix A); χ_4 was shifted by 180° before computing $\langle \chi_4(0)\chi_4(t)\rangle$ to include in the same plot. **Right (bottom):** commitment probability distributions (CPDs) for TS regions 1–4 for initial (left) and converged TPS (right) trajectories, each calculated from 150–200 trajectories at a resolution of 0.25.

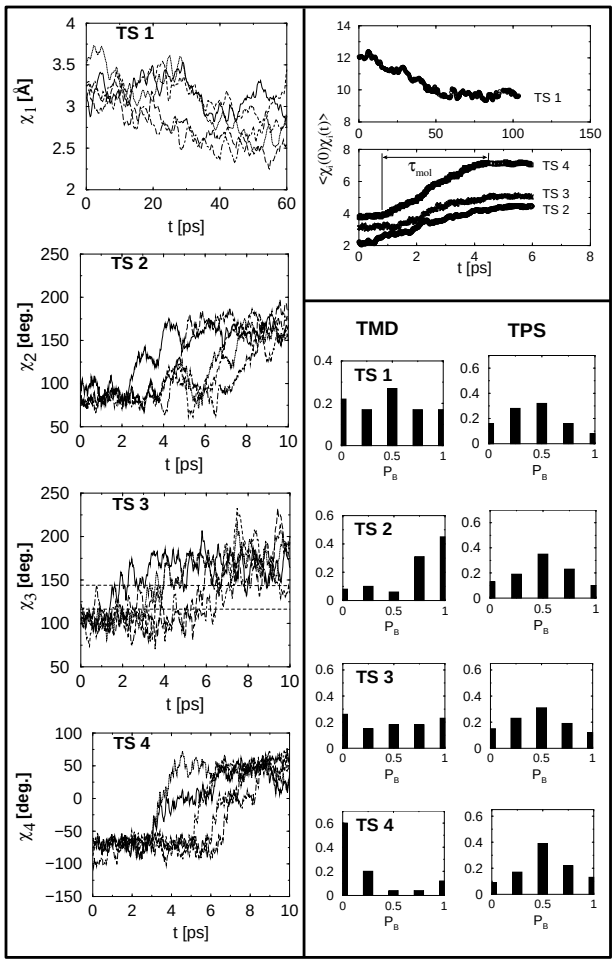

Figure 7: Convergence analysis

Appendix 1: Transition Path Sampling – Background

Transition path sampling (TPS) of Chandler and coworkers [1] aims to capture rare events (excursions or jumps between metastable basins in the free energy landscape) in molecular processes by essentially performing Monte Carlo sampling of symplectic dynamics trajectories, for which acceptance or rejection is determined by selected statistical criteria that characterize the ensemble of trajectories. Method details, beautiful illustrations, and applications to relatively small systems are available [2–5]. Here we only recapitulate essential features.

TPS exploits the separation of timescales in rare-event processes (e.g., long timescale $\tau_{\rm long}$ to bring a system to the top of the free-energy barrier compared to the short timescale $\tau_{\rm short}$ for dynamics within the barrier region) and saddle-like character of the free-energy landscape at transition-state (TS) regions to connect the various free energy basins. Namely, starting from an initial trajectory that captures a barrier crossing — which can be generated by algorithms that use guiding fields — TPS employs Metropolis Monte Carlo (MC) sampling of segments of (reversible and symplectic) molecular dynamics (MD) trajectories longer than $\tau_{\rm short}$ but shorter than $\tau_{\rm long}$. Despite the unphysical nature of the initial sampling trajectories, the protocol converges to yield physically meaningful trajectories passing through the saddle region. TPS samples different molecular dynamics trajectories using the shooting algorithm [6] (which perturbs initial momenta of atoms in a randomly chosen time interval, subject to the conservation of Maxwellian distribution of velocities, total linear and angular momentum, and detailed balance) to perform random walk steps in the space of trajectories, accepted or rejected according to selected statistical criteria (given by a path action as described below, see Eq. **3**) that characterize the ensemble of trajectories [1].

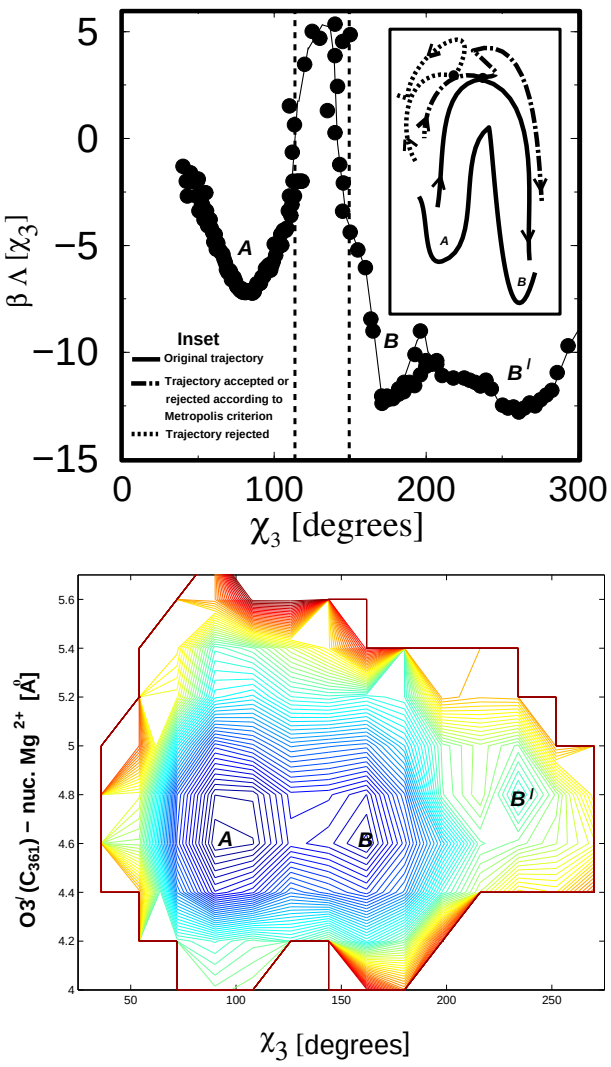

Figure 8: Free energy landscape with basins for Arg-258 rotation: A (unrotated); B (partially rotated); B' (fully rotated). The potential of mean force $\Lambda(\chi_3)$ (Upper) is obtained using umbrella sampling, and the two-dimensional free energy landscape (Lower) is obtained from TPS simulations. Ordinate (Lower) corresponds to distance between the nucleotide binding Mg^{2+} ion and the oxygen O3′ of the last primer (cytosine) residue. The dihedral angle space between the dashed lines (Upper) defines the TS ensemble for Arg-258's partial rotation. *Inset* illustrates how TPS trajectories are harvested (see text).

The shooting algorithm [6] generates an ensemble of molecular dynamics trajectories connecting two local minima (metastable states) A and B (see Fig. 8) in a free energy landscape via Monte Carlo sampling. For a given dynamics trajectory, the state of the system (i.e., basin A or B) is characterized by defining a set of order parameters $\chi = \{\chi_1, \chi_2, \cdots\}$. These order parameters are geometric quantities such as dihedral angles, bond distances, rms deviations of selected residues with respect to a reference structure, and so on. For biomolecules, as we show later, the key to a successful TPS application is identifying these key variables. Here, the groundwork simulations were important [7–9]. To formally identify a basin, *the population operator* h_A indicates if a particular molecular configuration associated with a time t of a molecular dynamics trajectory belongs to basin A:

$$h_A(\chi(t)) = \begin{cases} 1 & \text{if } \chi(t) \in A, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The *trajectory operator* H_B identifies a visit to basin B in a trajectory of length τ:

$$H_B\{\chi\}_\tau = \begin{cases} 1 & \text{if there exists } 0 < t < \tau \text{ such that } h_B(t) = 1, \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The idea in TPS is to generate many trajectories that connect A to B from one such existing pathway (see Figs. 7 and 8 in text). This is accomplished by a Metropolis algorithm that generates an ensemble of trajectories $\{\chi\}_\tau$ of length τ according to a path action $S\{\chi\}_\tau$ given by:

$$S\{\chi\}_\tau = \rho(0)\, h_A(\chi_0)\, H_B\{\chi\}_\tau, \tag{3}$$

where $\rho(0)$ is the probability of observing the configuration at $t = 0$ ($\rho(0) \propto \exp(-\beta E(0))$, in the canonical ensemble). A new trajectory $\{\chi^*\}_\tau$ is generated from an existing trajectory $\{\chi\}_\tau$ using the shooting algorithm [6], by perturbing the momenta of atoms at a randomly chosen time. The perturbation scheme is symmetric, i.e., the probability of generating a new set of momenta from the old set is the same as the reverse probability of generating the old set from the new set. Moreover, the scheme conserves the equilibrium distribution of momenta and the total linear momentum (and, if desired, total angular momentum). The acceptance probability implied by the above procedure is given by

$$P_{\text{acc.}} = \min\left[1, S\{\chi^*\}/S\{\chi\}\right]. \tag{4}$$

Together, these criteria ensure preservation of detailed balance, and thus according to the Metropolis algorithm [10], generate an ensemble of trajectories consistent with the path action S.

The ergodicity and convergence of each TPS run is monitored by calculating the autocorrelation function of the order parameter $\langle \chi_i(0)\chi_i(t)\rangle$ (Fig. 7) associated with each transition state i, where $\langle\cdot\rangle$ denotes the average over the ensemble of generated trajectories. In each

case, the autocorrelation function is plotted from $\langle \chi_i(0)\chi_i(0)\rangle \approx \langle \chi_A \rangle^2$ to the time τ where $\langle \chi_i(0)\chi_i(\tau)\rangle \approx \langle \chi_A \rangle \langle \chi_B \rangle$; this range is spanned during our sampling time τ, indicating that the transition state regions between A and B are crossed during this interval (see Fig. 8). The gradual transition of the autocorrelation functions between these values indicates decorrelation of the generated trajectories in each TPS run; strongly correlated trajectories would lead to an abrupt change in the correlation function for the chosen values of τ. The characteristic relaxation time τ_{mol} associated with the crossing of each TS region is given by the time taken for the gradual transition of the autocorrelation function $\langle \chi_i(0)\chi_i(t)\rangle$, as shown in Fig. 7 (see also Table 1 of main text). The value of τ_{mol} provides an estimate for the timescale associated with barrier crossing at the transition state region (see Appendix 4); the length τ of the MD trajectories in TPS should thus be greater than the transient time to commit to a basin, i.e., $\tau > \tau_{\mathrm{mol}}$.

Conserving the path action as described above both conserves the equilibrium distributions of the individual (metastable) states and ensures that the accepted molecular dynamics trajectories connect the two metastable states in question (Fig. 7). The shooting algorithm based on the Metropolis scheme (e.g., ref. [10]) conserves microscopic reversibility. Taken together, starting from a initial trajectory consistent with the path action, TPS generates an ensemble of MD trajectories guaranteed to converge to the correct ensemble as defined by the path action. The ensemble of trajectories represents configurations that constitute the correct pathway for hopping between the metastable states. Based on work by Crooks [11] and Jarzynski [12], we have shown (unpublished work) that the TPS of symplectic MD trajectories obeying microscopic reversibility satisfies the fluctuation-dissipation theorem [13].

Appendix 2: Transition Path Sampling – Biomolecular Applications

In general, harvesting mechanistic pathways by TPS for large biomolecular systems requires: (i) a robust protocol to generate initial trajectories, (ii) a careful identification of the different transition state regions, (iii) an implementation of the TPS sampling code to work with standard and well established integrators and force fields such as CHARMM [14], (iv) development and application of tests to assess convergence of TPS, and (v) reliable procedures for computing the reaction free energy pathway. We use a divide-and-conquer approach to implement the above steps for biomolecular systems, as we describe in turn.

(i) Generating initial trajectories that capture rare events is accomplished via targeted molecular dynamics (TMD) [15,16], high-temperature MD, and umbrella sampling [17–20] that use guiding (altered) fields. For example, we generate trajectories starting from the crystal open structure that capture pol β's closing using TMD on a pol β/DNA/dCTP complex with explicit water, salt, and magnesium ions [9] and similarly from the closed structure capturing the opening [7]. The TMD simulations were performed by introducing an additional restraint force as implemented in CHARMM based on the rms distance with respect to the closed polymerase conformation (1BPY). The functional form of the rms restraint energy is as follows:

$$E_{\rm rms} \;=\; K\;\left[D_{\rm rms}(X(t), X^{\rm target}) - d_0\right]^2, \qquad (5)$$

where K is a force constant, $D_{\rm rms}$ represents the relative rms distance for a selected set of atoms between the instantaneous conformation $X(t)$ and the reference $X^{\rm target}$, and d_0 is an offset constant (in Å). The restraint force is applied only to the heavy atoms in pol β. With a decrease in d_0 as a function of simulation interval, the conformational change is driven from the initial (open) to final (closed) conformation, and many configurations are generated for inspection.

(ii) Identifying different transition state (TS) regions is based on analysis of histograms of variables characterizing the motions of key residues (i.e., dihedral angles, distances, etc.) in the initial trajectories [free and TMD trajectories of the opening [7] and closing [9] as described above]. Although the initial trajectory is likely far from the physical pathway, a slice of the free energy landscape far from the actual path often displays similar characteristics (e.g., specific slow local motions) as that along the reaction coordinate. As a first approximation, the number of independent variables χ_i that display a bimodal distribution is used to characterize the various TS regions (see Table 1).

The next step is to confirm the existence of the TS regions by initiating a series of short (of order 10–100 ps, see Table 1), unrestrained MD trajectories from the different TS regions and calculating the "commitment probability distribution" (CPD) functions (i.e., probability that a particular trajectory will commit to a particular proximal free energy basin) for each TS region [2]. The CPDs are calculated by first choosing an ensemble of molecular configurations corresponding to a TS region (also called the TS ensemble). For example, in our application, the TS ensemble for the partial rotation of Arg-258 was defined as a window of dihedral-angle space corresponding to the barrier region as shown in Figs. 8 and 7. For each configuration in such a TS ensemble, four trajectories are initiated with a randomly

chosen set of momenta from a Maxwell distribution, and the commitment probability P_B is determined as the fraction of trajectories committing to basin B. The frequency distribution of P_B for the configurations in the TS ensemble yields the CPD. In general, the CPD for each TS region of the initial (unphysical) trajectory is bimodal with a minimum at 1/2 and maxima at 0 and 1 [1] (Fig. 7).

(iii) TPS is implemented using CHARMM version C28a2 by using shooting and shifting [1,21] algorithms and configurational bias to enhance the Metropolis sampling while preserving the detailed balance criteria. The shooting and shifting moves are coded in a PERL script that calls CHARMM for trajectory generation. The code is designed to handle multiple TS regions that are characterized by sets of arbitrary dihedral angles, distances, and any configurational quantity calculable within CHARMM. The Verlet integrator in CHARMM with a time step of 1 fs is used for generating the individual molecular dynamics trajectories in TPS. Electrostatic and van der Waals interactions are smoothed to zero at 12 Å, and periodic boundary conditions are used in three dimensions during the dynamics simulations.

(iv) Testing for convergence of TPS involves (1) calculating the autocorrelation functions associated with order parameters to check for the decorrelation of paths. Paths can be considered decorrelated if the autocorrelation function shows a gradual transition (increase or decrease of the value) between approximately $\langle\chi_A\rangle^2$ and $\langle\chi_A\rangle\langle\chi_B\rangle$.

The sampling quality can also be assessed by calculating order-parameter correlation functions in path space, $\langle\chi_i^*(0)\chi_i^*(n)\rangle_{\rm NS}$, where n represents the harvested trajectories, χ_i^* is the value of the order parameter evaluated at a particular time-slice at the bottleneck of the transition, and NS denotes no shifting with respect to the first trajectory (to remove the trivial decorrelation because of the shifting moves). Fig. 9 shows such a correlation function for the last 30 harvested trajectories for TS 1–4. It is evident from the figure that on an average, every 10th to 20th trajectory is statistically decorrelated; therefore the 150 to 200 trajectories that we generate for each TS ensure sufficiently good sampling.

(2) Recalculating CPD functions for each TS region to ensure that CPDs are unimodal with a peak at 1/2 (Fig. 7). The saddle region associated with each transition state is manifested by a unimodal CPD peaked around 1/2 [2]. (3) Extending a few harvested trajectories and confirming that the open and closed conformations of pol β are sampled. In our case, a global criterion such as the rms deviation of the heavy atoms in the enzyme with respect to the open (1BPX) and the closed (1BPY) crystal structures is used to quantify the proximity to the reference states. A failure to observe the correct bounds of the order parameter autocorrelation functions likely indicates inadequate sampling — MD segments are shorter than $\tau_{\rm short}$ — or use of incorrect variables to characterize the metastable basins. Pathological behavior in 2 above indicates that convergence of TPS has not been achieved, and/or a particular TS region in the initial trajectory is not a saddle in the actual pathway.

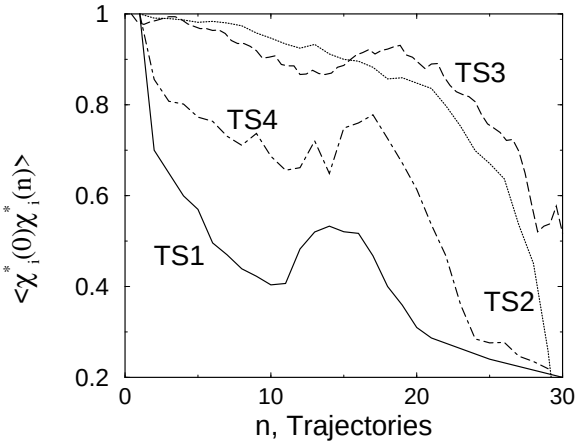

Figure 9: Decorrelation of transition paths for TS 1–4. The plots are normalized to begin at unity by normalizing with $\langle\chi_i^*(0)\chi_i^*(0)\rangle_{\rm NS}$.

Failure to sample the reference states in 3 suggests the presence of additional TS regions and requires additional path sampling near TS regions closest to open or closed conformations of pol β, after which steps **i**–**iv** above are repeated.

(v) Computing free energies along the reaction pathway is the most time consuming and important phase. We have reduced the time and statistical error by applying a novel procedure on the shooting and shifting algorithms rather than a biasing force as described separately (unpublished work). Essentially, the free energy barrier is determined by calculating the potential of mean force using histogram methods and a modified umbrella sampling scheme based on TPS. The order parameter variable χ_i characterizing each transition state i is divided into 6–10 windows. In each window, the probability distribution $P(\chi_i)$ is calculated from a series of modified path sampling runs performed using an appropriate guiding function (action) (unpublished work). The potential of mean force $\Lambda(\chi_i)$ in each window is given by ($\beta = 1/k_BT$):

$$\beta\Lambda(\chi_i) = -\ln[P(\chi_i)] + \text{constant}. \qquad (\mathbf{6})$$

The arbitrary constant associated with each window is adjusted by the method prescribed by Lynden-Bell and coworkers [22] to make the Λ function continuous. The overall free energy is calculated using the equation:

$$\exp(-\beta F) = \int_{\chi_{\min}}^{\chi_{\max}} \exp(-\beta\Lambda(\chi_i))\, d\chi_i, \qquad (\mathbf{7})$$

where the integration domain characterizes the metastable state. The free energies of all metastable states and barriers were calculated by numerically integrating Eq. **7** in the appropriate region.

A coarse-grained potential of mean force along each reaction coordinate is computed for each conformational event, as shown in Figure 10. The order parameter variable χ_i characterizing each transition state i is divided into 6–10 nonoverlapping windows. The probability distribution $P(\chi_i)$ is calculated from a series of path sampling runs (unpublished work) in each order parameter window. The histograms for each window were collected by running 75 trajectories per window.

The potential of mean force was calculated in each discrete window according to Eq. **6**. The slopes of the potential of mean force ($d\beta\Lambda(\chi_i)/d\chi_i$) were evaluated by a linear least squares fit to Λ vs. the χ_i data, in each window, as given in Table 2. The coarse-grained potentials of mean force along the reaction coordinates were calculated from the slopes by using a trapezoidal rule for integration along each χ_i (Fig. 10).

The free energy profiles presented in Fig. 10 involve two main approximations/errors: (i) the statistics obtained from our potential of mean force calculations resolve the free energy to within $\pm 3\, k_BT$. A more accurate resolution of the free energy would involve upwards of 300 trajectories per window. (ii) The slopes $d\beta\Lambda(\chi_i)/d\chi_i$ are assumed to be constant throughout each integration step in the trapezoidal rule. This assumption was deemed reasonable based on analyses of three model systems (unpublished work). This coarse-graining approximation can be circumvented by using overlapping windows as prescribed by Lynden-Bell and coworkers [22].

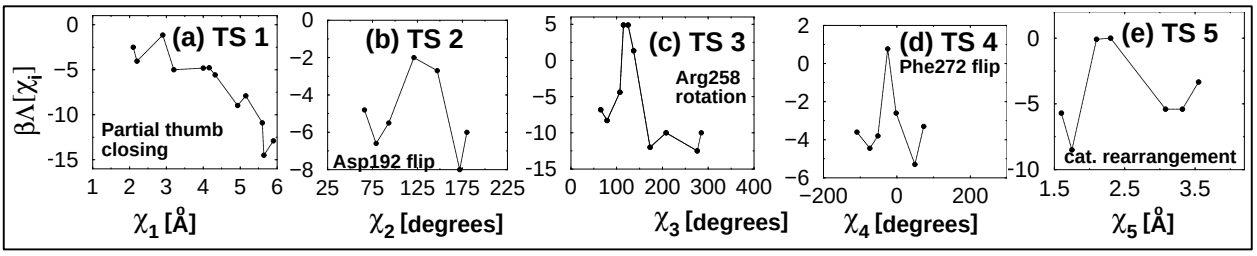

Figure 10: Approximate coarse-grained potential of mean force along the reaction coordinate for different transition state regions. (a) Partial thumb closing (TS 1). The reaction coordinate χ_1 is the rms deviation of the thumb residues (residues 275 to 295) with respect to the closed state. (b) Asp-192 flip (TS 2). The reaction coordinate χ_2 is the dihedral angle characterizing the flip of Asp-192. (c) Arg-258 rotation (TS 3). The reaction coordinate χ_3 is the dihedral angle characterizing the rotation of Arg-258. (d) Phe-272 flip (TS 4). The reaction coordinate χ_4 is the dihedral angle characterizing the flip of Phe-272. (e) Rearrangement of catalytic region and the stabilization of Arg-258 in the fully rotated state (TS 5). The reaction coordinate χ_5 is the distance between the nucleotide binding Mg^{2+} ion and the oxygen atom $O1_\alpha$ of dCTP.

Table 2: Coarse-grained potential of mean force calculations

χ_1 Å	$d\beta\Lambda/d\chi_1$ Å^{-1}	χ_2 deg.	$d\beta\Lambda/d\chi_2$ deg.$^{-1}$	χ_3 deg.	$d\beta\Lambda/d\chi_3$ deg.$^{-1}$	χ_4 deg.	$d\beta\Lambda/d\chi_4$ deg.$^{-1}$	χ_5 Å	$d\beta\Lambda/d\chi_5$ Å^{-1}
2.2	−8.3	72.6	−0.10	59.0	−0.15	−98.0	−0.03	1.7	−20.0
2.8	4.5	85.6	0.08	96.0	0.13	−59.0	−0.03	1.9	22.2
3.1	−15.5	106.0	0.13	112.0	0.92	−45.0	0.18	2.3	0.0
3.6	0.3	134.0	−0.03	119.0	0.00	−15.0	−0.14	2.9	−7.3
4.4	−2.3	159.0	−0.21	142.0	−0.29	28.0	−0.05	3.2	0.0
4.8	−4.2	175.0	0.25	163.0	−0.43	65.0	0.09	3.6	10.1
5.2	5.0			181.0	0.04				
5.5	−4.6			230.0	−0.21				
5.6	−56.2			282.0	0.11				
5.8	4.6								

Steps **i**–**iv** define a self-consistent algorithm for dealing with multiple transition states of the free-energy landscape in complex systems. Together with step **v**, they can yield the overall rate of transition as described in Appendix 4.

Appendix 3: Protonation States

The protonation states of the titratable side chain groups in the enzyme were chosen based on their individual pKa values and consistent with a solution pH of 7.0 as reported in Table 3. Indeed, in the open crystal complex the three conserved Asp groups are well separated from each other and not closely interacting with the dCTP, and therefore this choice of the protonation state based on pKa of the amino acid group and an overall pH of 7.0 is reasonable.

Table 3: Protonation states of amino acids in pol β

Residue	Charge	pKa
Asp	-1	3.9
Glu	-1	4.3
His	0	6.5
Lys	$+1$	10.8
Arg	$+1$	12.5

Still, a body of recent simulation data suggests that the protonation states are far from clear. In ref. [26], the authors show on the basis of a truncated model of the active site in *ab-initio* calculations, density functional theory (DFT) functionals, and specific basis-set used that the geometry could only be optimized if the assumption that Asp-192 was protonated was made. A report by a different group [27] on the same system, truncated pol β active site claims that geometries can be optimized using high-level DFT without assuming that Asp-192 is protonated. These contrasting observations may reflect artifacts of truncating the active site and ignoring the rest of the protein/DNA/solvent environment.

Note also that the protonation state may change as the conformational change occurs. In classical simulations, it is not possible to allow this change in a physically consistent manner, and that is part of the inherent limitations of classical force fields. Further studies are needed to establish the protonation states. These may include quantum mechanics/ molecular mechanics (QM/MM) simulations in which the titratable residue is included in the quantum mechanical (QM) part and then the free energy change associated with protonation/deprotonation reaction is computed. Alternatively, the relative free energies of the protonated and nonprotonated states could be computed using molecular mechanics/ Poisson-Boltzmann surface area (MM/PBSA) or similar methods to determine the pKa.

Appendix 4: Calculating Reaction Rates

Here, we outline how to estimate the rates [based on transition state theory [23]] associated with the transitions between adjacent metastable states in our overall free energy profile (Fig. 4, upper center). We also outline a procedure for obtaining the correction to the transition state theory approximation.

The free energies of the different metastable states and transition-state regions relative to the open and closed states can be obtained from the potential of mean force calculations, see Appendix 2). Using transition state theory [23], the rate of the transition between adjoining metastable states in Fig. 4 is given by

$$k_{\rm TST}^{A\to B} = \frac{1}{\tau_{\rm mol}} \exp(-\beta \Delta F_{AB}^{\rm barrier}), \qquad (\mathbf{8})$$

where $\tau_{\rm mol}$ is the time to cross the transition-state region and commit to basin B, and $\Delta F_{AB}^{\rm barrier}$ is the free energy of the transition-state region between basins A and B relative to basin A. For example, considering the adjacent states A and B as metastable states 3 and 4 (separated by TS 2) in Fig. 4, $\Delta F_{AB}^{\rm barrier} = F(\text{TS 2}) - F(A)$ and $\Delta F_{BA}^{\rm barrier} = F(\text{TS 2}) - F(B)$; Eq. **8** is then used to compute $k^{A\to B}$ and $k^{B\to A}$ associated with TS 2.

In the ideal gas approximation, the pre-factor $1/\tau_{\rm mol} = k_B T/h$, where h is the Plancks constant. In the reactive flux formalism [24], an estimate for $\tau_{\rm mol}$ is given by $w/\langle|\dot{q}|\rangle^*$, where w is the characteristic width to be crossed along the reaction coordinate q, and $\langle|\dot{q}|\rangle^*$ is the rate of change of the reaction coordinate at the transition state surface.

Because we have relevant data in our application from monitoring convergence, we can use the characteristic time to observe the relaxation of the order parameter autocorrelation function (see Fig. 7) as an estimate for $\tau_{\rm mol}$ for TS 1–4. For TS 5, we can use $\tau_{\rm mol} \approx k_B T/h$ because of the relatively small free energy barrier associated with this TS. The rates of transitions between the adjacent metastable states in Fig. 4 can then be calculated using Eq. **8**[§].

Using the individual rates of transitions between adjoining metastable basins, the overall rate of the conformational change can be determined by modeling the overall process as a network of reactions:

$$\text{Open (MS 1)} \underset{k_{21}}{\overset{k_{12}}{\rightleftarrows}} \text{MS 2} \underset{k_{32}}{\overset{k_{23}}{\rightleftarrows}} \text{MS 3} \underset{k_{43}}{\overset{k_{34}}{\rightleftarrows}} \text{MS 4} \underset{k_{54}}{\overset{k_{45}}{\rightleftarrows}} \text{MS 5} \underset{k_{65}}{\overset{k_{56}}{\rightleftarrows}} \text{MS 6} \underset{k_{76}}{\overset{k_{67}}{\rightleftarrows}} \text{Closed (MS 7)}, \qquad (\mathbf{9})$$

where MS 1–7 correspond to the different metastable states in Fig. 4. The network of reactions can be solved using kinetic Monte Carlo simulations [25] to determine the overall rate of transition between the open and closed states. This is a more involved process, which we may pursue in the future.

Still, an order-of-magnitude estimate for the rate (k) of the overall conformational change is available from

$$k \approx (k_B T/h) \times \exp(-\beta \Delta F_{\rm RL}), \qquad (\mathbf{10})$$

where $\Delta F_{\rm RL}$ is the free-energy of the rate-limiting barrier in the reaction profile relative to the open state. In our case, the rate-limiting barrier is TS 3, corresponding to Arg-258's

[§] A correction to the transition state theory approximation may be obtained by computing the transmission coefficient using the Bennett-Chandler method [24].

rotation, for which $\Delta F_{\mathrm{RL}} = 20.5 \pm 3k_BT$; the corresponding range for the rate is 1.5×10^5 to 4×10^2 s^{-1}.

Literature cited

1. Bolhuis, P. G., Chandler, D., Dellago, C. & Geissler, P. L. (2002) *Annu. Rev. Phys. Chem.* **53**, 291–318.
2. Bolhuis, P. G., Dellago, C. & Chandler, D. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 5883–5888.
3. Geissler, P. L., Dellago, C. & Chandler, D. (1999) *Phys. Chem. Chem. Phys.* **1**, 1317–1322.
4. Radhakrishnan, R. & Trout, B. L. (2003) *Phys. Rev. Lett.* **90**, 158301.
5. ten Wolde, P. R. & Chandler, D. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 6539–6543.
6. Bolhuis, P. G., Dellago, C. & Chandler, D. (1998) *Faraday Discuss.* **110**, 421–436.
7. Yang, L., Beard, W. A., Wilson, S. H., Broyde, S. & Schlick, T. (2002) *J. Mol. Biol.* **317**, 651–671.
8. Yang, L., Beard, W. A., Wilson, S. H., Roux, B., Broyde, S. & Schlick, T. (2002) *J. Mol. Biol.* **321**, 459–478.
9. Arora, K. & Schlick, T. (2003) *Chem. Phys. Lett.* **378**, 1–8.
10. Allen, M. P. & Tildesley, D. J. (1990) *Computer Simulation of Liquids* (Oxford Univ. Press, New York).
11. Crooks, G. E. (1999) Ph.D. Thesis, University of California (Berkeley).
12. Jarzynski, C. (1997) *Phys. Rev. Lett.* **78**, 2690–2693.
13. Chandler, D. (1987) *Introduction to Modern Statistical Mechanics* (Oxford Univ. Press, New York).
14. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983) *J. Comp. Chem.* **4**, 187–217.
15. Schlitter, J., Engels, M. & Krüger, P. (1994) *J. Mol. Graphics* **12**, 84–89.
16. Schlitter, J., Engels, M., Krüger, P., Jacoby, E. & Wollmer, A. (1993) *Mol. Simul.* **10**, 291–309.
17. Paci, E. & Karplus, M. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 6521–6526.
18. Bernèche, S. & Roux, B. (2001) *Nature* **414**, 73–77.
19. Bartels, C. & Karplus, M. (1998) *J. Phys. Chem. B* **102**, 865–880.
20. Harvey, S. C. & Gabb, H. A. (1993) *Biopolymers* **33**, 1167–1172.
21. Dellago, C., Bolhuis, P. G. & Geissler, P. L. (2002) *Adv. Chem. Phys* **123**, 1–81.
22. Lynden-Bell, R. M., Van Duijneveldt, J. S. & Frenkel, D. (1993) *Mol. Phys.* **80**, 801–814.

23. Frost, A. A. & Pearson, R. G. (1961) *Kinetics and Mechanism* (Wiley, New York).

24. Chandler, D. (1978) *J. Chem. Phys.* **68**, 2959–2970.

25. Fichthorn, K. A. & Weinberg, W. H. (1991) *J. Chem. Phys.* **95**, 1090–1096.

26. Abashkin, Y. G., Erickson, J. W. & Burt, S. K. (2001) *J. Phys. Chem. B* **105**, 287–292.

27. Rittenhouse, R. C., Apostoluk, W. K., Miller, J. H. & Straatsma, T. P. (2003) *Proteins: Struct. Funct. Genet.* **53**, 667–682.