PNAS
www.pnas.org

Supplementary Information for

Uncovering universal rules governing the selectivity of the archetypal DNA glycosylase TDG

Thomas Dodd, Chunli Yan, Bradley R. Kossmann†, Kurt Martin and Ivaylo Ivanov*
Ivaylo Ivanov
Email: iivanov@gsu.edu

This PDF file includes:

Supplementary Methods
Figs. S1 to S9
Tables S1
Captions for movies S1
References for SI reference citations

Other supplementary materials for this manuscript include the following:

Movies S1

www.pnas.org/cgi/doi/10.1073/pnas.18003323115

Supplementary Methods

Model construction

Models for the pre- and post-extrusion states were constructed from two TDG/DNA crystal structures (PDB ID: 5HF7(1) and 2RBA(2)). For the base interrogation, we built the system with initially separated TDG and 5caC-DNA. We also built TDG-DNA complexes with a G:T mismatch and normal DNA (A:T pair). For consistency, all systems were built with the same DNA sequence 5'-GTACGTGAG-3'. All systems were then solvated with TIP3P(3) water molecules in a box with a minimum distance of 10.0 Å from the surface atoms of the complex to the edge of the periodic simulation box. Counter-ions were added to neutralize the net charge of the complex and reach 150 mM NaCl concentration to mimic physiological conditions. 5caC force field parameters were determined with the Antechamber module of AMBER16(4).

Molecular dynamics for the base interrogation

Steepest decent minimization was performed for 5000 steps. Each system was then slowly heated to 300 K over 50 ps in the NVT ensemble with positional restraints on all heavy atoms using a force constant of 5 kcal/mol/Å^2. Positional restraints were gradually released over 6 ns in the NPT ensemble to fully equilibrate the systems. Production runs were performed in the isothermal-isobaric ensemble (1 atm and 300 K), employing smooth particle mesh Ewald (SPME) electrostatics, 10 Å cut-off for short-range non-bonded interactions and 2-fs time step. After 100 ns of unrestrained MD, 200 ns of accelerated molecular dynamics was employed by boosting both the total potential and the dihedral potential. Calculated values for boost parameters were 8812 kcal/mol and 155 kcal/mol, respectively. Eight snapshots leading up to the base extrusion event were selected and then replicated 10 times. The replicas were each simulated for 100 ns of unrestrained molecular dynamics, reinitializing velocities for each replica. All simulations were performed using the AMBER16 code with the AMBER Parm14SB parameter set.(5)

Path optimization with the Partial Nudged Elastic Band (PNEB) method

To determine a minimum energy path (MEP) connecting the pre- and post-extrusion TDG configurations we employed the PNEB method (6) - a chain-of-replicas method that involves concurrent optimization of a number of copies of the simulated system (denoted as replicas or beads). We chose to represent the path by a total of 28 replicas - 10 copies of the equilibrated initial states, 8 copies of an intermediate state and 10 copies of the final extruded state, respectively. By gradually spreading the replicas from the initial and final states we allow the PNEB optimization process to discover the MEP in a fully unbiased way. All heavy atoms of the TDG/DNA complex were included in the path optimization. For the first 50ps the system was heated to 100K with a Langevin collision frequency of 1000 ps^{-1} using 5 kcal mol^{-1} $Å^{-2}$ spring forces. Heating of the system from 100K to 300K and subsequent cooling back to 100K was performed stepwise over 25 ns.

Path optimization with the String Method (SM)

We then carried out path optimization with the string method, based on the swarms of trajectories method, requiring definition of a lower dimensional space (7). Twenty-eight images from the PNEB path (including the two fixed end points) were sampled along the initial pathway defined in collective variable space. Two collective variables were defined by using an RMSD collective variable (RMSD computed over a selection of atoms relevant to the extrusion transition) and pseudo-dihedral angle denoted in Figure S6. We refined these structures using a swarm of 20 short (2-ps) simulations launched from each image. Images were updated based on mean drift in each swarm, redistributing between end states and relaxing with 980-steps unconstrained and 20-steps restrained simulations. At least 200 iterations were completed for each string (Figure S7). 200 ns of unconstraint simulations for each image (total 11.2 μs) were then performed by taking the final converged structure for both PNEB (28 images) and string methods (28 images). Graphics of the movie were prepared by Chimera (8).

MSM construction

All TICA calculations, clustering, and MSM construction were performed using the PyEMMA software (9). Backbone torsions of the interrogated base and the distances between base pairs and the guanidinium group of Arg275 were used as descriptive coordinates to define the TICA projections. For base interrogation, all independent components (ICs) were computed using a lag time of 50 ps (25 steps). The combined

trajectories were then projected onto the first two ICs. The trajectory frames were then clustered in the projection space using the k-means algorithm, producing 800 clusters. From the clustering data, implied timescales were generated by estimating the transition probability matrix at different lag times (Figure S8). Using these results, 5 macrostates and a lag time of 50 ps (25 steps) were chosen to construct the MSMs for all three base interrogation systems. For the 5caC-DNA base eversion path, independent components were calculated using a lag time of 100 ps (50 steps). The combined trajectories were then projected onto the first two ICs. Trajectory frames were then clustered using uniform time clustering, a method in which data points are selected uniformly in time and assigned using a Voronoi discretization. This produced 800 clusters, from which the implied timescales were then estimated (Figure S9). Based on these results, 6 macrostates and a lag time of 80 ps (40 steps) was chosen for MSM construction.

Supplementary Figures

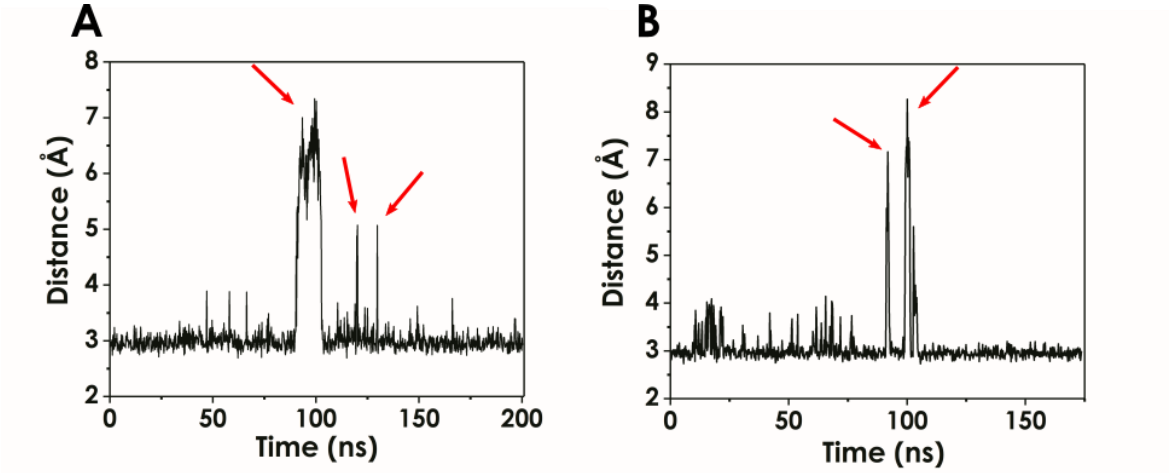


Figure S1. Transient base-opening events observed for regular DNA in the presence of TDG during A) accelerated MD and B) free unbiased MD. Extrusion events (red arrows) were determined by measuring the distance between N1 and N3 atoms of the interrogated base pair along the trajectory. For clarity, only short trajectory segments encompassing such base-opening events are shown.

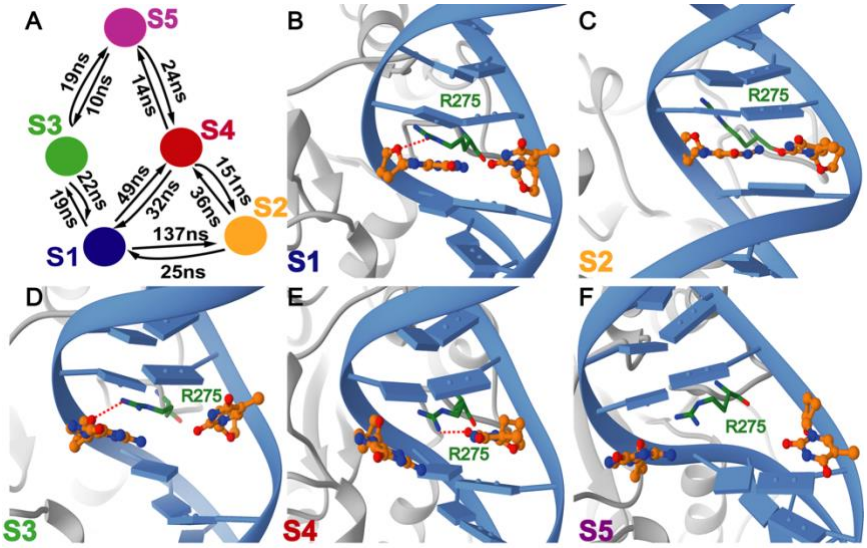


Figure S2. **Representative structures selected from each macrostate of the Markov State Model corresponding to G:T mismatch interrogation by TDG.** Each state is colored according to the color scheme in Figure 2E; panels are labeled by macrostate designation. TDG is shown in grey; DNA is shown in blue. The intercalating Arg275 residue at the tip of the insertion loop is shown in ball and stick representation and colored in green. The G:T mismatch bases are shown in ball and stick representation and colored in orange. The calculated transition timescales determined from transition path theory are included in the first panel.

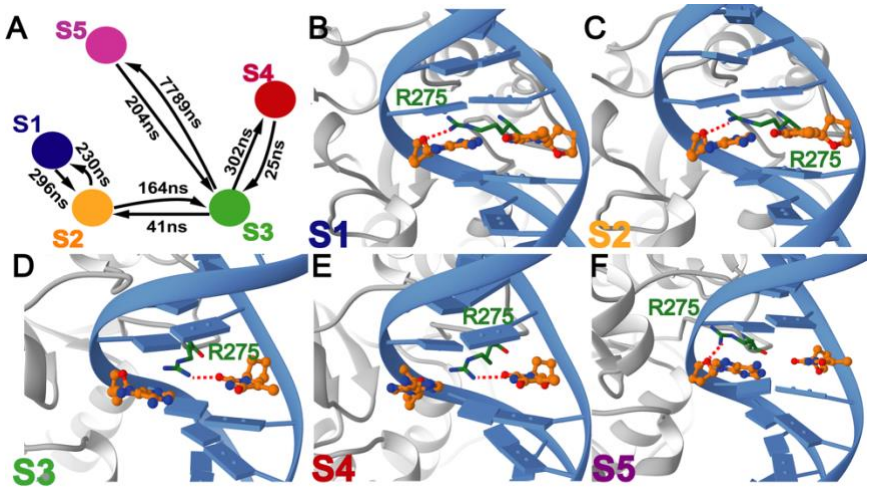


Figure S3. Representative structures selected from each macrostate of the Markov State Model corresponding to A-T base interrogation by TDG. Each state is colored according to the color scheme in Figure 2F; panels are labeled by macrostate designation. TDG is shown in grey; DNA is shown in blue. The intercalating Arg275 residue at the tip of the insertion loop is shown in ball and stick representation and colored in green. The extruded and orphaned bases are shown in ball and stick representation and colored in orange. The calculated transition timescales determined from transition path theory are included in the first panel.

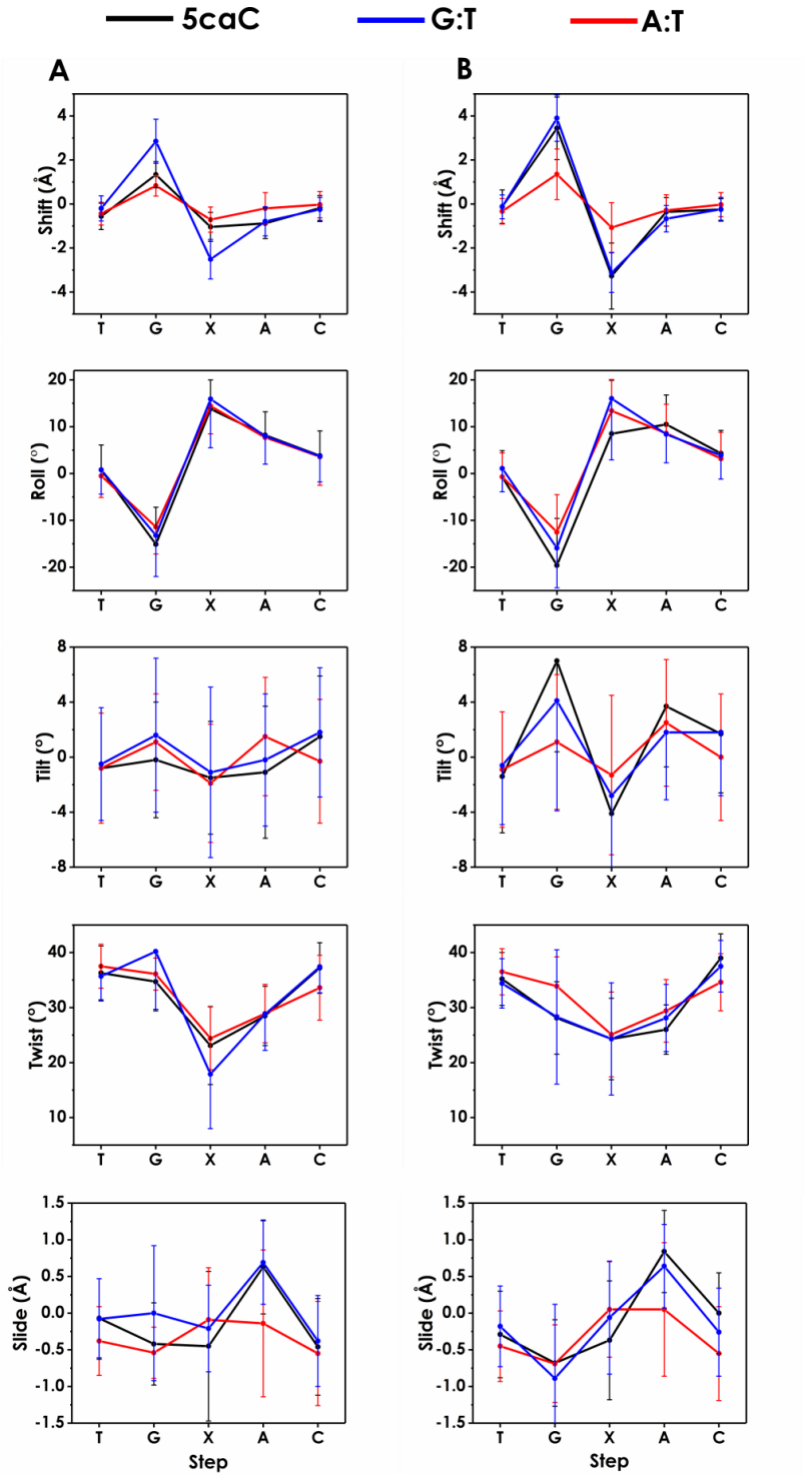


Figure S4. Interbase pair parameters shift, roll, tilt, twist and slide for A) intrahelical state and B) extrahelical state. 5caC is denoted by the black line, while the G:T is denoted by the blue line and the A:T is denoted by the red line. The standard deviation is plotted as error bars.

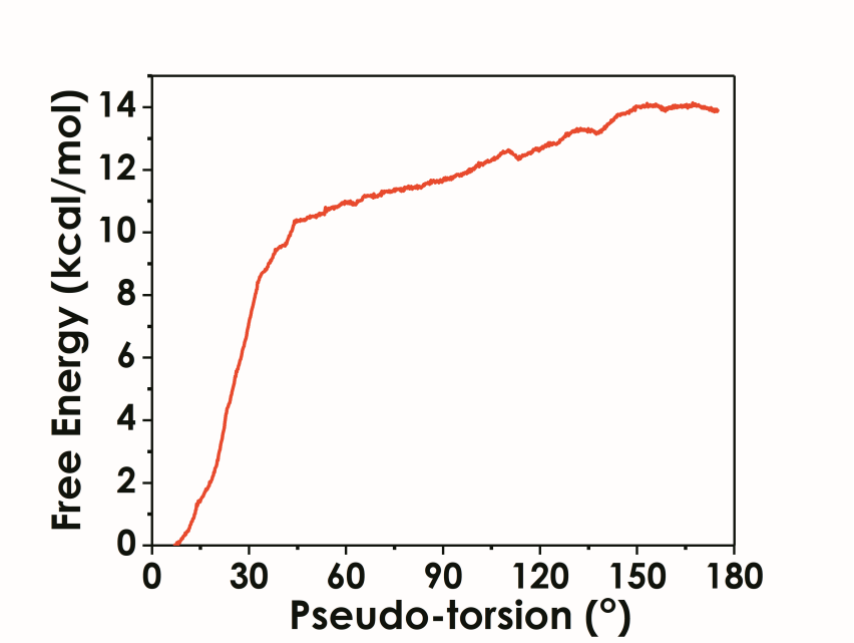


Figure S5. Calculated free energy from umbrella sampling simulations of A:T base pair sequence in the absence of TDG. The pseudo-torsion CV was used to restrain the opposing thymine from its initial intrahelical position to an extrahelical position using 80 bins separated by 2° each

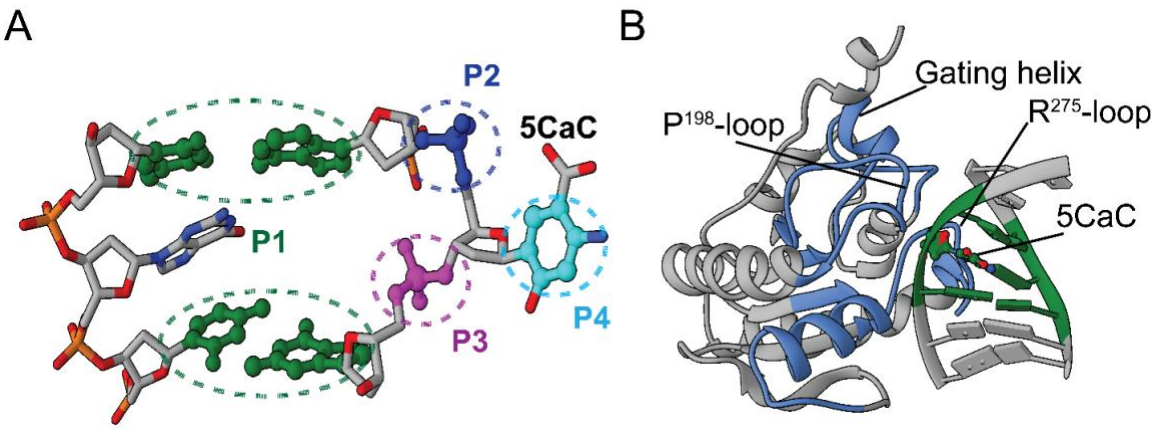


Figure S6. Collective variables used for the finite temperature string method. A) Pseudo-torsion collective variable defined by the centers of mass of four atom groups: P1 (green), P2 (blue), P3 (magenta) and P4 (cyan); B) Root-mean-square deviation (RMSD) collective variable defined over a selection of atoms containing the gating helix and adjacent loops (Cα atoms of residues 136-157), Pro198 loop (Cα atoms of residues 190-200 and 202-205; heavy atoms of residue Lys201), Arg275 loop (Cα atoms of residues 269-274 and 276-283; heavy atoms of residue Arg275), Cα atoms of residues 228-247 and the segment of DNA highlighted in green that includes the 5caC base, orphaned base and the base pairs above and below (all heavy atoms). Protein residues included in the selection are shown in blue; DNA atoms included in the selection are shown in green.

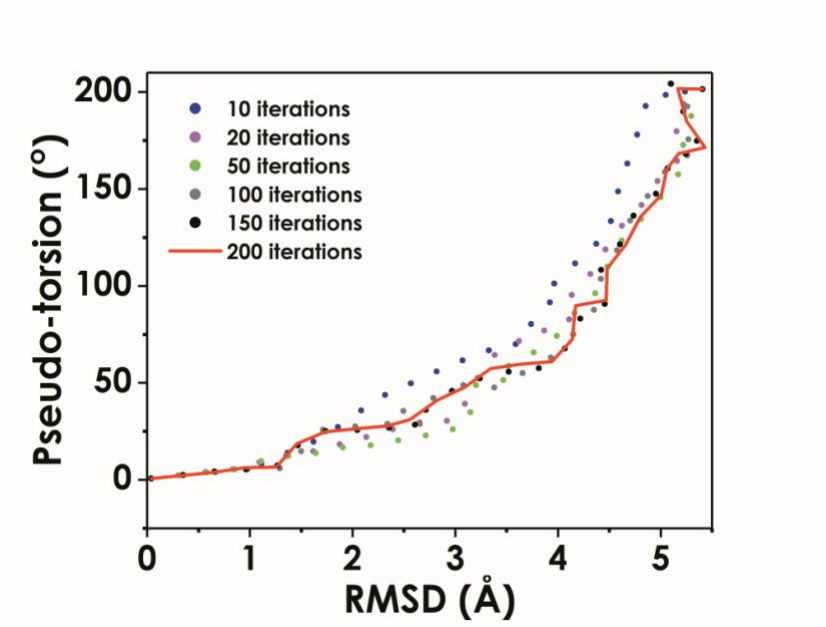


Figure S7. **Convergence of the string method path optimization.** Values of the collective variables (pseudo-dihedral angle and RMSD as shown in Figure S6) are plotted for all the replicas of the simulation system that comprise the string. Same color points represent the positions of the string replicas that have evolved for a pre-specified number of iterations (10 to 200). The final converged positions of the replicas are indicated with a red line.

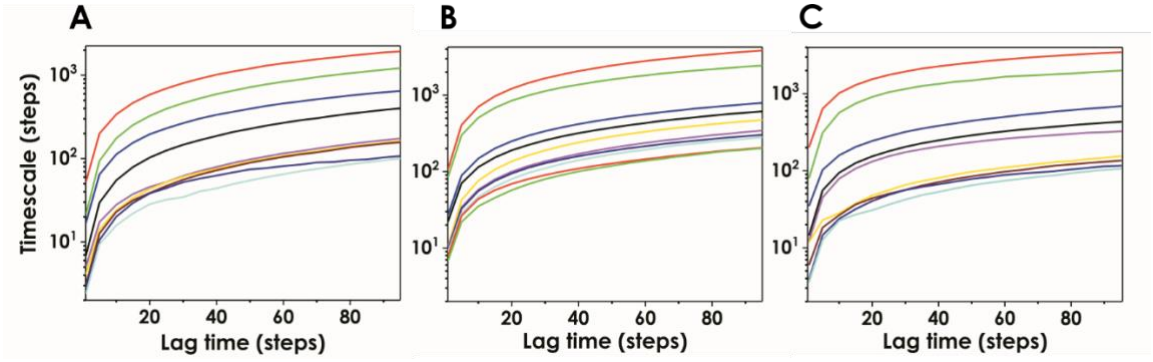


Figure S8. Implied timescales as a function of lag time from analysis of the MD trajectories The following systems were analyzed: (A) TDG in complex with 5caC-DNA (B) TDG with G:T mismatch and (C) TDG with normal DNA (A:T base pair). Colors in the plot are arbitrary and simply aid in visually distinguishing potentially overlapping lines. Gaps separating the slowest implied timescales were used to estimate the number of kinetically distinct macrostates for MSM construction.

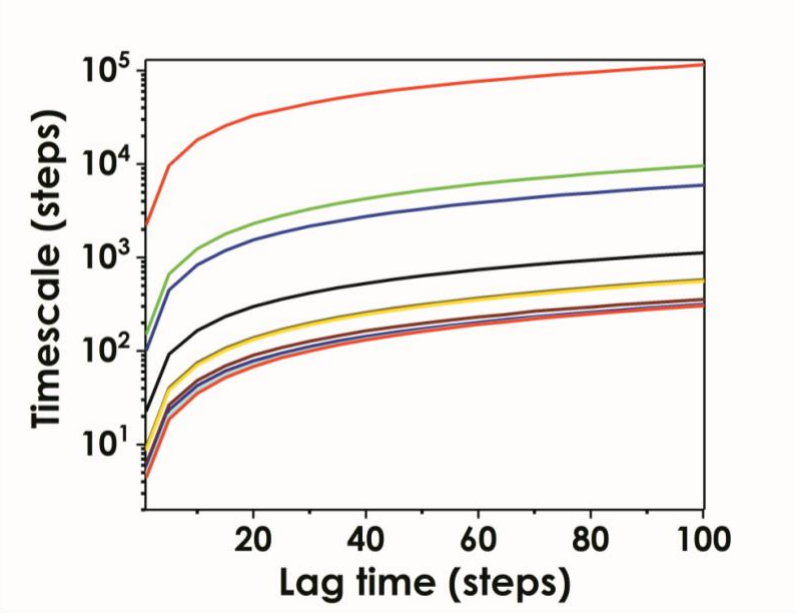
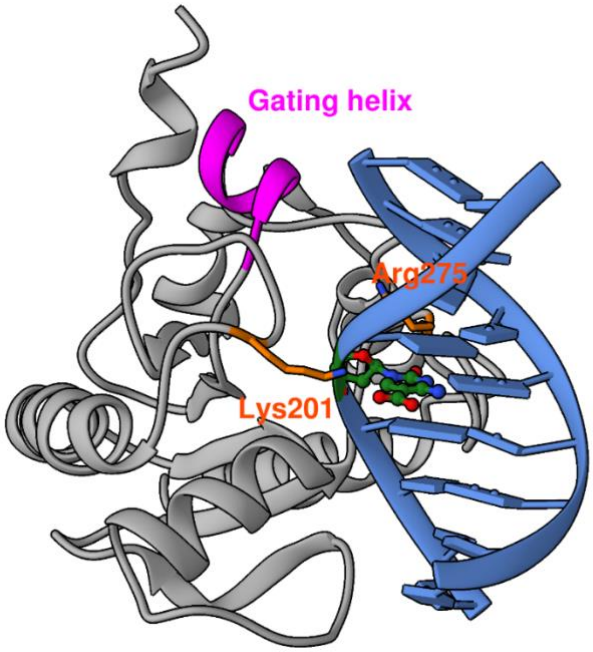


Figure S9. Implied timescales as a function of lag time from analysis of the MD trajectories along the base eversion path in the 5caC-DNA TDG complex. Colors in the plot are arbitrary and simply aid in visually distinguishing potentially overlapping lines. The gaps separating the slowest implied timescales were used to estimate the number of kinetically distinct macrostates for MSM construction. Only the first 10 timescales are shown.

Table S1. Calculated phosphate-phosphate distances for each macrostate identified from the full base eversion path. The flanking phosphate groups on both sides of the 5caC were used for distance measurements.

State	P-P distance (Å)
S1	12.5
S2	11.9
S3	10.9
S4	11.0
S5	9.7
S6	9.5



Movie S1. Base extrusion path optimized with the PNEB method. TDG is shown in grey; DNA is shown in blue. Arg275 in the interrogation loop and Lys201 in the Pro198 loop are shown in orange. The 5caC base is shown in green. The gating helix is colored in magenta.

1. Coey CT*, et al.* (2016) Structural basis of damage recognition by thymine DNA glycosylase: Key roles for N-terminal residues. *Nucleic Acids Res* 44:10248-10258.
2. Maiti A, Morgan MT, Pozharski E, & Drohat AC (2008) Crystal structure of human thymine DNA glycosylase bound to DNA elucidates sequence-specific mismatch recognition. *Proc Natl Acad Sci U S A* 105:8890-8895.
3. Poole PH, Hemmati M, & Angell CA (1997) Comparison of thermodynamic properties of simulated liquid silica and water. *Phys Rev Lett* 79:2281-2284.
4. Wang J, Wolf RM, Caldwell JW, Kollman PA, & Case DA (2004) Development and testing of a general amber force field. *J Comput Chem* 25:1157-1174.
5. Maier JA*, et al.* (2015) ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* 11:3696-3713.
6. Bergonzo C, Campbell AJ, Walker RC, & Simmerling C (2009) A Partial Nudged Elastic Band Implementation for Use with Large or Explicitly Solvated Systems. *Int J Quantum Chem* 109:3781.
7. Pan AC, Sezer D, & Roux B (2008) Finding transition pathways using the string method with swarms of trajectories. *J Phys Chem B* 112:3432-3440.
8. Pettersen EF*, et al.* (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* 25:1605-1612.
9. Scherer MK*, et al.* (2015) PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J Chem Theory Comput* 11:5525-5542.