SI Methods

System Preparation. The parameters for O^6-methylguanine were generated following the protocol in refs. 1 and 2. Hartree-Fock (HF) and Møller-Plesset (MP2) calculations using a 6-31G* basis set were performed using Gaussian98 (3) to estimate stacking interactions with each of the other possible bases, Watson-Crick pairing interactions with both pyrimidines, and hydrogen bonding interactions with water molecules (SI Fig. 9*a*). These energies were scaled as in ref. 1 to account for the lack of explicit polarization in the empirical energy function. The van der Waals parameters for N1, C6, O6, and the methyl group atoms were set by analogy with existing CHARMM atom types, and the charges on those atoms were adjusted by hand to reproduce the target data. The root-mean-square error obtained with the resulting parameters was 0.5 kcal mol^{-1}, which is comparable to figures associated with the existing force field (1). Manually selected internal energy terms yielded good agreement between quantum mechanical and empirical normal mode spectra for the isolated base as well (SI Fig. 9*b*).

The system was prepared from the C145S AGT-dsDNA complex (PDB entry 1T38) (4) as described in detail in ref. 5. In order to limit the computational cost of treating the solvent, we identified atoms that could be constrained without significantly impacting the dynamics of nucleotide flipping based on preliminary steered molecular dynamics simulations for each of the two steps (5). In both cases, the region of mobile atoms was a sphere of 25 Å, but it was centered on the C_α atom of Gly131 for the unstacking step and on the midpoint of the amide hydrogen and C_β atoms of Asn157 for the active-site entry step. The former (latter) resulted in 2117 (1985) atoms in the reaction region, 58 (50) atoms in the buffer region, and 1273 (1413) atoms in the reservoir region.

Classification of Structures. The structures for which we calculated p were drawn from 30 trajectories for each step of the mechanism; because the commitment probability varies rapidly close to the transition state, these structures were taken primarily from points along the paths for which it was observed that there were comparable acceptance rates for shooting forwards and backwards in time. 100 additional trajectories of up to 90 ps were used for each p calculation. In these calculations, the basins used to define the stable states were modified slightly. For the active-site entry step, the active site basin was defined as $3.0 \leq \theta \leq -2.5$ rad (the coordinate is periodic) and the extrahelical intermediate basin was defined as $1.3 \leq \theta \leq 2.5$ rad. For the unstacking step, the intrahelical basin was defined as $-0.6 \leq a_1 \leq 0.4$ rad, $-1.2 \leq a_2 \leq -0.4$ rad and $2.8 \leq s \leq 6.8$ Å and the extrahelical basin was defined as $-0.8 \leq a_1 \leq 0.0$ rad, $-0.6 \leq a_2 \leq 0.0$ rad and $7.0 \leq s \leq 14.0$ Å. The dynamics were terminated when the system spent at least 1.5 ps in a basin. Because a fraction of trajectories were found to become stuck in metastable states that we believe will not be populated significantly on experimentally relevant time scales, we took the transition states to be somewhat closer in p to the reactant basin than traditionally: $0.2 < p < 0.6$ for the unstacking step and $0.3 < p < 0.6$ for the active-site entry step.

Genetic Neural Network (GNN). The neural networks had two input nodes, two hidden layer nodes, and one output node. The genetic algorithm had 200 individuals (combinations of order parameters) and was evolved for 80 generations to optimize the jackknife cross-validated root-mean-square error in p predictions. Typically the simulations converged after 25 generations; results reported were consistently found in five independent trials initialized with different random number generator seeds.

Free Energy Calculations. The free energies were calculated with umbrella sampling (6) and processed with the weighted histogram analysis method (7). The force constants for the harmonic restraints were all 50 kcal mol^{-1} Å^{-2}. For the unstacking step, a total of 640 independent 60-

ps simulations were performed for each base (Gua or mGua). The values of d_1 and d_2 for the restraint minima were varied in increments of 0.1 Å independently from 3.25 to 6.35 Å and 14.05 to 15.95 Å, respectively. For the active-site entry step, a total of 1400 independent 60-ps simulations were performed for each base. The values of d_3 and d_4 for the restraint minima were varied in increments of 0.1 Å independently from 7.05 to 13.95 Å and 8.05 to 9.95 Å, respectively. The starting points for each simulation were selected from the unbiased trajectories with the values of the order parameters closest to the center of each window. The calculations with guanine were performed in the same manner; for the active-site entry step, d_5 was used in place of d_3 for both Gua and mGua.

Estimation of Reaction Rates. To estimate the rates for flipping for mGua and Gua for each of the steps, we used the methods detailed in ref. 8, which builds on ref. 9. Namely, the unimolecular rate is

$$k(\mathbf{e}) = \frac{1}{2\pi}\left(\frac{\det \mathbf{V}_\mathrm{R}}{|\det \mathbf{V}|}\right)^{1/2} \frac{\mathbf{e}^\mathrm{T}\mathbf{D}\mathbf{e}}{|\mathbf{e}^\mathrm{T}\mathbf{V}^{-1}\mathbf{e}|} e^{-\beta\Delta}, \qquad \textbf{[1]}$$

where **D** is the diffusion tensor in the projected space, **V** and $\mathbf{V}_\mathrm{R}$ are the second-derivative matrices (Hessians) for the potential of mean force scaled by the temperature at the saddlepoints and reactant basins, respectively, Δ is the difference in free energy between these points, and **e** is the eigenvector with negative eigenvalue of the product **VD**.

For the Hessians, the appropriate areas of the free energy surfaces in SI Fig. 7 were fit with parabolas; as needed, additional umbrella sampling simulations were performed to extend the free energy surfaces beyond the areas shown to ensure that the basins were sufficiently well defined. The Hessian elements are given in SI Table 1. The diffusion tensors at the saddlepoints were obtained from equation 8 of ref. 8: $D_{ij} = \max\langle \dot{q}_i(0)[q_j(t) - q_j(0)]\rangle$. These elements are also given in SI Table 1 along with the resulting eigenvectors.

Given that the free energies do not include corrections for the charge scaling (SI Tables 2 and

3), the basins and barriers are not perfectly parabolic, and the diffusion tensors were calculated only in the transition state regions, the rates should be viewed as order of magnitude estimates. The effort that would be required to obtain more precise numbers is not justified given the uncertainties inherent in the calculation (e.g., whether the backbone can fully relax to the equilibrated conformation used in the calculations as the protein slides along the DNA). Nevertheless, we feel that the rates are reasonable. The relatively fast pre-exponential factors obtained (in the picosecond range) are consistent with the times observed for commitment to basins during the transition path sampling calculations. As to the free energies, the corrections are expected to be small since the potential based scaling method was employed and the explicit solvent region is very large (10-13).

References

[1] Foloppe N, MacKerell AD. (2000) All-atom empirical force field for nucleic acids: 1. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J Comp Chem* 21:86-104.

[2] MacKerell AD, Banavali NK. (2000) All-atom empirical force field for nucleic acids: 2. application to molecular dynamics simulations of DNA and RNA in solution. *J Comp Chem* 21:105-120.

[3] Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, *et al. Gaussian 98 (Revision A7)*. Gaussian, IncPittsburgh PA (1998) .

[4] Daniels DS, Woo TT, Luu KX, Noll DM, Clarke ND, Pegg AE, Tainer JA. (2004) DNA binding and nucleotide flipping by the human DNA repair protein AGT. *Nat Struct Mol Biol* 11:714-720.

[5] Hu J, Ma A, Dinner AR. (2006) Bias annealing: a method for obtaining transition paths *de novo*. *J Chem Phys* 125:114101.

[6] Torrie JM, Valleau JP. (1977) Non-physical sampling distributions in monte-carlo free-energy estimation-umbrella sampling. *J Comp Phys* 23:187-199.

[7] Kumar S, Bouzida D, Swendsen RH, Kollman PA, Rosenberg JM. (1992) The weighted histogram analysis method for free-energy calculations on biomolecules I The method. *J Comp Chem* 13:1011-1021.

[8] Ma A, Nag A, Dinner AR (2006) Dynamic coupling between coordinates in a model for biomolecular isomerization. *J Chem Phys* 124:144911.

[9] Berezhkovskii A, Szabo A (2005) One-dimensional reaction coordinates for diffusive activated rate processes in many dimensions. *J Chem Phys* 122:014503.

[10] Simonson T, Archontis G, Karplus M (1997) Continuum treatment of long-range interactions in free energy calculations Application to protein-ligand binding. *J Phys Chem B* 101:8349-8362.

[11] Dinner AR, Blackburn GM, Karplus M (2001) Uracil-DNA glycosylase acts by substrate autocatalysis. *Nature* 413:752-755.

[12] Dinner AR, Lopez X, Karplus M (2003) A charge-scaling method to treat solvent in QM/MM simulations. *Theor Chem Acc* 109:118-124.

[13] Ma A, Hu J, Karplus M, Dinner AR (2006) Implications of alternative substrate binding modes for catalysis by uracil-DNA glycosylase: An apparent discrepancy resolved. *Biochemistry* 45:13687-13696.