Supplementary material

Methods

The model system we use is a family A high fidelity polymerase from *Bacillus stearothermophilus*: Bacillus Fragment (BF). This enzyme is structurally homologous to the Klenow fragment from *E. coli* (50 % sequence homology) and exhibits a high efficiency (200 base pairs/sec) and processivity (111 nucleotide bases) for accurate (fidelity of 10^{-8}) DNA replication. Crystals of BF are catalytically active (the enzyme can synthesize base pairs in crystal) and this property has been exploited by Beese et. al. to obtain high resolution crystal structures of the enzyme at a number of points along the replication cycle [1,2].

System preparation

We prepared four model systems **G:C**, **G:A**, 8oxo**G:C** and 8oxo**G:A** (see Figure S1) using the insight II modeling software [3] , starting from the crystal structure of a closed ternary BF-DNA-dCTP complex (PDB id: 1LV5 [1]). These correspond to cases of correct/incorrect nucleotide incorporation opposite an undamaged/oxidatively damaged **G** template base respectively. The Mn^{2+} ion at the catalytic site in the crystal structure was replaced with a Mg^{2+} ion. Crystallographic waters were discarded. Missing atoms in the crystal structure were added including the terminal primer **A** O3′. For the mispair the incoming dCTP in 1LV5 was replaced with a dATP. For the oxidative damage cases the **G** base in the **G:C** and **G:A** models were modified to 8oxo**G** by adding oxygen and hydrogen atoms at C8 and N7 respectively and by modifying the double bond between C8 and N7 to a single bond. Hydrogen atoms were added to the models using the HBUILD [4] utility in CHARMM with HIS protonation states chosen according recommendations from the WHATIF web interface (http://swift.cmbi.kun.nl/WIWWWI). Protonation states for all other charged groups were chosen according to their pKa values in aqueous solution [6] at a pH of 7.0 (ASP→ -1, GLU→-1, LYS→ +1, ARG→+1). The models were then solvated using SOLVATE

1.0 [7] which also neutralizes the system by placing Na^+ and Cl^- ions at isotonic concentrations (0.154 mol/l), with a Debye-Huckel distribution at 300 K. A total of 98 Na^+ and 66 Cl^- ions were added to neutralize the systems.

The **G:C** system is directly derived from the 1LV5 crystal structure[1]. The other three models **G:A**, 8oxo**G:C** and 8oxo**G:A** were constructed by replacing the **G** with 8oxo**G** and/or replacing the incoming dCTP with a dATP. An important issue to consider during the modeling was the conformation of the template **G**/8oxo**G** opposite the incoming nucleotide for the latter three systems for which no BF/DNA/dNTP ternary complexes have been crystallized. For correct nucleotide incorporation opposite undamaged DNA substrates, it is known that in the open (inactive) state of the enzyme prior to nucleotide insertion that the template base is in a syn conformation (characterized by a glycosidic torsion angle χ= 0 degrees between the sugar and base groups). Sometime during the nucleotide insertion stage the template base switches to an anti (χ= 0 degrees) conformation which is preserved and observed in post-insertion structures. However the template base can adopt the syn conformation in the event of a mispair and/or in the event of damage[2,8,9]. While there are no crystal structures for a closed ternary complex with an 8oxo**G:**dCTP pair at the active site, a ternary complex of T7 pol I/8oxo**G**DNA/dCTP (prior to catalysis) shows the lesion carrying template base at the active site to be in an anti conformation [9]. However, there are no crystal structures for a BF/DNA/dNTP ternary complex with either **G:A** or 8oxo**G:A** mispair at the active site. Crystal structures of oligonucleotide sequences show that **G** can adopt either a syn or anti conformation opposite an incoming dATP[2] while structures of post-insertion complexes in BF [8], pol β [10] , and T7 DNA pol [9] find the template base carrying the lesion in a syn conformation in 8oxo**G:A** systems. We thus carried out simulations for a **G:A** system with the template **G** in a syn and anti conformations and modeled the 8oxo**G:C** and 8oxo**G:A** systems in anti:anti and syn:anti conformations respectively. For the **G:A** system the anti:anti simulations were stable but for the syn:anti simulations we observed an syn→ anti template flip with a fast timescale of 800 ps (see Figure S2). This result shows that an anti conformation is reached pre-chemistry and that the base flipping reaction is not rate limiting even in misincorporation reactions for BF. As proposed on basis of indirect kinetic evidence[11] the rate limiting step for mismatch reactions is most-likely the catalysis step owing to a distorted pairing geometry between the templating **G** (in an anti conformation) and the incoming dATP. In contrast simulations for 8oxo**G** lesion with an incoming dATP were stable with a syn conformation for the template base which reduces the distortion of the catalytic site thereby significantly enhancing the rate of the misincorporation. We also initiated unconstrained molecular dynamics trajectories for the 8oxo**G:C** and 8oxo**G:A** systems with the glycosidic torsion restricted to χ=90°, (i.e., close to the presumed transition state between *syn* and *anti* conformations) during the equilibration phase. These simulations showed that the 8oxo**G** template base adopts an *anti* conformation opposite an incoming dCTP (Fig S1 bottom left insert) and a *syn* conformation opposite an incoming dATP (Fig S1 bottom right insert); see also Fig S2. Taken together with the existing structural evidence, these imply that a *syn* conformation of the lesion opposite a dATP and an anti conformation of the lesion opposite dCTP are most-likely to be the stable ground states, thereby validating our model systems.

Forcefield parameterization

The CHARMM27 [12] forcefield was used to perform MD simulations. Parameters (partial charges for nonbonded interactions and force constants for bonded interactions) compatible with CHARMM27 for the 8oxo**G** residue were constructed as described by Foloppe et. al. [13]. Partial charges were assigned and refined to reproduce ab-initio 8oxo**G** dipole moments, base-water dimer interaction energies and distances. These values were then used in a genetic algorithm based optimization scheme developed in our lab (Y. Liu, R. Radhakrishnan, unpublished) to construct and refine the CHARMM force field parameters for bond, angle and dihedrals of the 8oxo**G** residue to reproduce ab-initio vibrational frequencies. The new parameters thus obtained, were then used to refine the partial charges further, and the entire procedure was repeated until convergence was reached. The resulting root-mean-squared deviation (RMSD) of σ = 79.78 cm^{-1} between the newly parameterized CHARMM normal mode frequencies and ab-initio vibrational frequencies is within acceptable limits for small molecules [13]

Simulation protocols

The NAMD simulation package [14,15] with the CHARMM27 force field was used to minimize and equilibrate each model system and for subsequent production runs. The model systems were enclosed in a solvent box (dimensions 111 Å x 91 Å x 95 Å) of 27068 water molecules and periodic boundary conditions were applied. A 12.0 Å cutoff was applied for non-bonded interactions wherein a switching potential was turned on at 10.0 A. The particle mesh Ewald method [16] was used for the treatment of long range electrostatics. The rigidbonds option (i.e., the rattle algorithm) was used to constrain all bonds involving hydrogen atoms to their values in the CHARMM parameter file. The equilibration protocol for each system was as follows: systems were subjected to two initial rounds of minimization (10000 steps), heating from 0-300K (50000 steps) and NVT equilibration (50000 steps) with 1fs timesteps. The protein and DNA fragment was held fixed in the first round while the second round was unconstrained. Subsequently an NPT equilibration (with a 2 fs timestep) was carried out to obtain the correct density/box size for each system. Finally a 100 ps NVT equilibration run was carried out to arrive at the equilibrated configuration. Following the equilibration 10 ns NVT production runs were carried out. The RMSD of the protein backbone was monitored and data from the last 5ns during which the rmsd was found to be stable (Fig S3) was used for subsequent analysis.

Free energy simulations for pre-organization of catalytic sites

We consider the key coordinates for catalysis to be the O3′-P_α (d_a) and O3′-Mg (d_b) distances and perform a two dimensional umbrella sampling wherein the two distances are constrained at different values in the vicinity of the simulation average. Umbrella sampling simulations were performed in CHARMM (c32a1) for a reduced system comprising of all protein, DNA, and dNTP atoms , MG^{2+} ions and water molecules within a 3.0 Å shell of the BF-DNA-dNTP ternary complex. For each of the four model systems starting structures were obtained from the last 5ns of the 10 ns production runs with average (over last 5 ns) d_a and d_b values. We vary d_a and d_b over a range of values $d_a^{max} – d_a^{min}$ and $d_b^{max} – d_b^{min}$, appropriately chosen to sample and include deviations from the ideal catalytic geometry in steps of 0.5 A. The value of (d_a^{max},d_b^{max}) in units of Å was set to be (4.0, 4.0), (5.0,5.0), (5.0,4.0), (4.0,4.0) for the **G:C**, **G:A**, 8oxo**G:C** and 8oxo**G:A** systems, respectively, while the value of d_a^{min} =d_b^{min} was set to be 2.0 Å for all four systems. At each grid point we perform two rounds of steepest-descent minimization followed by Langevin dynamics (γ=10 ps^{-1}) at 300 K with a 1 fs timestep. The first round is performed with a high value of forcing restraints (2000 kcal/mol/Å^2) applied to d_a and d_b and consists of 1000 step minimization and dynamics. The second round is performed with a lower value of the forcing restraint (20 kcal/mol/Å^2) and consists of 1000 step minimization and 10000 step Langevin dynamics. The first round of minimization/dynamics run brings the system to the desired grid point and the second round performs umbrella sampling around it. We thus obtained 25, 49, 35 and 25 umbrella sampling windows for the **G:C**, **G:A**, 8oxo**G:C** and 8oxo**G:A** systems, respectively. Data from these windows were then used to construct unbiased probability distributions and free energy surfaces according to the WHAM [17,18] algorithm. The error in free energies is estimated to be ± 0.9 k_BT (0.54 kcal/mol) from the standard deviations obtained from three completely different umbrella sampling simulations performed for the **G:C** system. Since identical conditions are used to collect data we expect the errors to be of the same order for the **G:A**, 8oxo**G:C** and 8oxo**G:A** systems.

Principal Component Analysis

Principal component analysis (PCA) [19,20] of MD simulations provides us with a framework to project out independent motions in an MD trajectory and sort them in the order of their dominance (the strongest motions first). This is achieved by diagonalizing the variance-covariance matrix of atomic fluctuations along the trajectory. The resulting eigenvectors are the uncoupled principal components (PCs), (modes orthogonal to each other) and the eigenvalues reflect their magnitude (strength) in the trajectory. Since the formalism requires a well-defined average geometry as a reference around which the variance-covariance matrix of atomic fluctuations will be constructed, we chose the average geometry of the ternary complex with bound waters as the reference. The PCA calculation was performed for a small region around the catalytic geometry which included all heavy atoms of the incoming dNTP, six residues of the DNA template strand (including the template **G**/8oxo**G** of the nascent base pair), four residues of the DNA primer strand (including the terminal **A**), the two Mg^{2+} ions, two polymerase aspartate residues **D830 and D653** which coordinate the Mg^{2+} ions, residues from two helices forming the polymerases fingers and four residues R615, Y714, Q797 and H829 crucial for polymerase fidelity. The software program CARMA [21] was used to perform PCA on our system. CARMA also enables us to visualize principal modes by projecting out the atomic fluctuations due to the modes along the MD trajectory. The top 10 principal component modes contained most of the atomic fluctuations in the MD trajectory for all systems studied (70% for G:C, 72 % for 8oxoG:C and 80% for 8oxoG:A).

1. Johnson SJ, Taylor JS, Beese LS. Processive DNA synthesis observed in a polymerase crystal suggests a mechanism for the prevention of frameshift mutations. P Natl Acad Sci USA 2003;100(7):3895-3900.
2. Johnson SJ, Beese LS. Structures of mismatch replication errors observed in a DNA polymerase. Cell 2004;116(6):803-816.
3. Insight II molecular modelling software. San Diego: Molecular Simulations Inc.; 2000.
4. Brünger AT, Karplus M. Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. Proteins Struc Func Genet 1988;4:148-156.
5. Doublie S, Tabor S, Long AM, Richardson CC, Ellenberger T. Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 angstrom resolution. Nature 1998;391(6664):251-258.
6. Radhakrishnan R, Schlick T. Fidelity discrimination in DNA polymerase beta: differing closing profiles for a mismatched (G:A) versus matched (G:C) base pair. . J Am Chem Soc 2005;127:13245-13253.
7. Grubmuller H, Heymann B, Tavan P. Ligand binding: Molecular mechanics calculation of the streptavidin biotin rupture force. Science 1996;271(5251):997-999.
8. Hsu GW, Ober M, Carell T, Beese LS. Error-prone replication of oxidatively damaged DNA by a high-fidelity DNA polymerase. Nature 2004;431(7005):217-221.

9. Brieba LG, Eichman BF, Kokoska RJ, Doublie S, Kunkel TA, Ellenberger T. Structural basis for the dual coding potential of 8-oxoguanosine by a high-fidelity DNA polymerase. Embo Journal 2004;23(17):3452-3461.

10. Krahn JM, Beard WA, Miller H, Grollman AP, Wilson SH. Structure of DNA polymerase beta with the mutagenic DNA lesion 8-oxodeoxyguanine reveals structural insights into its coding potential. Structure 2003;11(1):121-127.

11. Joyce CM, Benkovic SJ. DNA polymerase fidelity: Kinetics, structure, and checkpoints. Biochemistry-Us 2004;43(45):14317-14324.

12. Jr ADM, Bashford D, Bellott M, Dunbrack RL, Jr, Evanseck MJFJD, Fischer S, Gao J, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, B. Prodhom andW. E. Reiher I, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. Journal of Physical Chemistry B 1998;102:3586--3616.

13. All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. J Comp Chem 2000;21:105-120.

14. Kale L, Skeel R, Bhandarkar M, Brunner R, Gursoy A, Krawetz N, Phillips J, Shinozaki A, Varadarajan K, Schulten K. NAMD2: Greater scalability for parallel molecular dynamics. Journal of Computational Physics 1999;151(1):283-312.

15. Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kale L, Schulten K. Scalable molecular dynamics with NAMD. Journal of Computational Chemistry 2005;26:1781-1802.

16. Molecular Dynamics Simulations of Solvated Biomolecular Systems: The Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, RNA, and Proteins. J Amer Chem Soc 1995;117:4193-4194.

17. Roux B. The Calculation of the Potential of Mean Force Using Computer-Simulations. Computer Physics Communications 1995;91(1-3):275-282.

18. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. Multidimensional Free-Energy Calculations Using the Weighted Histogram Analysis Method. Journal of Computational Chemistry 1995;16(11):1339-1350.

19. Amadei A, Linssen ABM, Berendsen HJC. Essential Dynamics of Proteins. Proteins-Structure Function and Genetics 1993;17(4):412-425.

20. Amadei A, Linssen ABM, deGroot BL, Aalten DMFv, Berendsen HJC. An Efficient Method for Sampling the Essential Subspace of Proteins. J Biomol Struct Dynam 1996;13:615-625.

21. Glykos NM, Kokkinidis M. Structural polymorphism of a marginally stable 4-alpha-helical bundle. Images of a trapped molten globule? Proteins 2004;56(3):420-425.

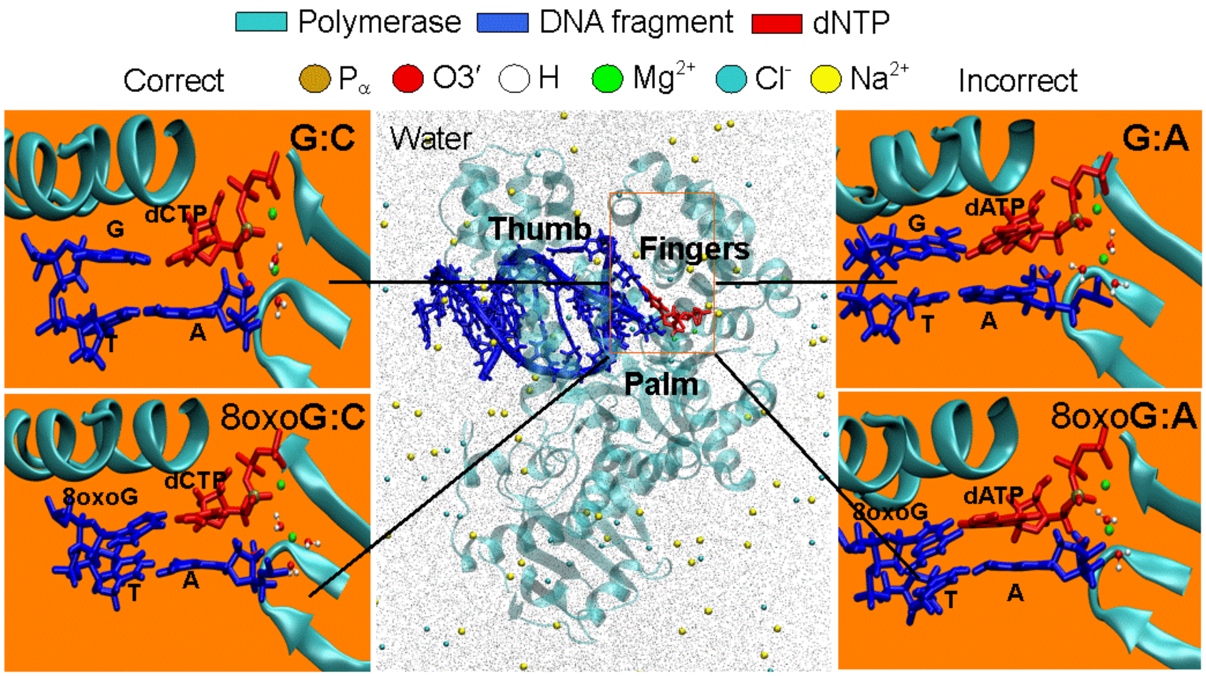

Figure S1: Simulations are carried out on fully solvated and neutralized ternary complexes (center) for the Bacillus fragment (BF). The four insets show the average active site geometry (DNA, incoming dNTP, Mg^{2+}, bound waters, parts of the polymerase fingers and palm domains) from 5 ns classical simulations for the four model systems for correct/incorrect nucleotide incorporation opposite an undamaged/oxidatively damaged **G** template base as indicated.

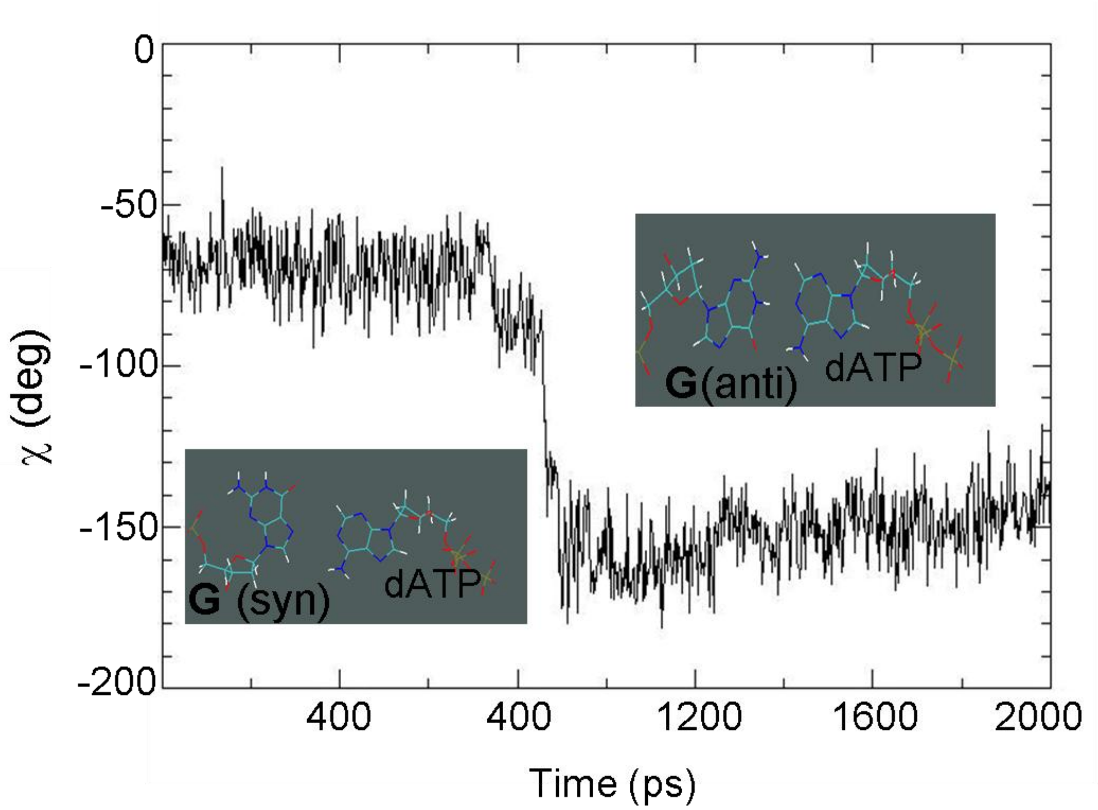

Figure S2: Values of the Glycosidic angle χ during a 2ns production run for the **G**(syn)**:A** case . The templating base **G** starts out in a syn conformation as depicted on the LHS (top left), flipping over to the anti conformation (bottom left).

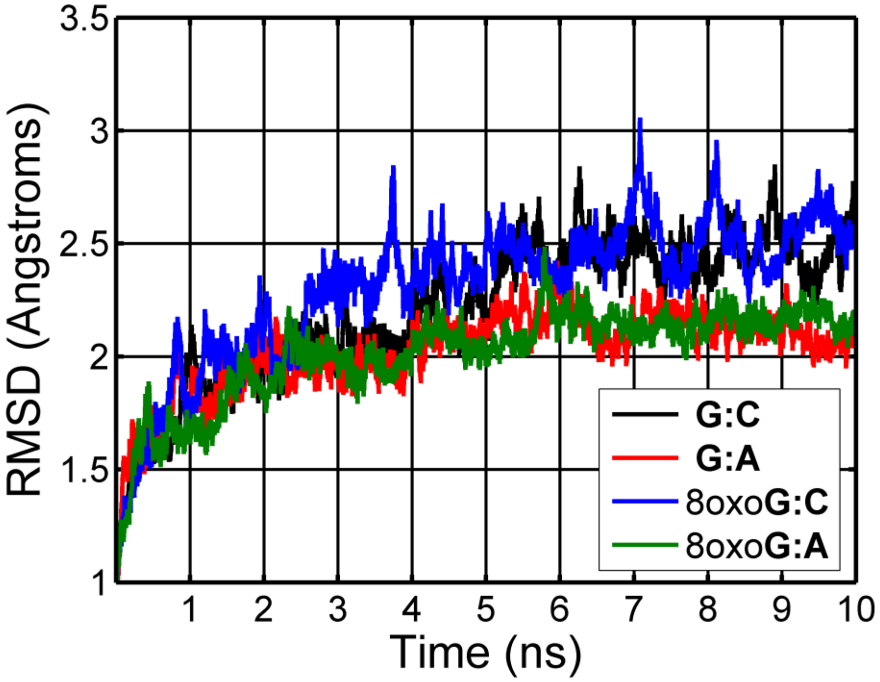

Figure S3: All atom root-mean-squared-deviation (RMSD) along a 10 ns, unconstrained, classical NVT production run. The plateau in the last 5 ns is interpreted as the equilibrium phase; data from the last 5 ns was used in further analysis.

G:C (Control)

Structural parameters Distances Atom1—Atom2	**MM** Å	**QM-MM** Å	**Crystal structure** Å
G:C1'—dCTP:C1' $^{(IS)}$	10.57 (0.15)	10.53 (0.12)	[a]10.56 [b]10.3 (0.2)
T:C1'—**A**:C1' $^{(PIS)}$	11.11 (0.23)	11.10 (0.20)	[a]10.10 [b]10.3 (0.2)
dCTP:P_{α}— **A**:O3′	3.51 (0.26)	3.00 (0.10)	-
MG2—A:O3′	2.66 (0.56)	2.14 (0.09)	[c] ~2.0*
MG2—dCTP:$O2_{\alpha}$	2.43 (0.35)	2.26 (0.12)	[a]2.66 [c] 2.4
MG2—D831:$O1_{\delta}$	1.81 (0.04)	2.03 (0.07)	[a]2.66 [c] 2.4
MG2—D653:$O2_{\delta}$	1.80 (0.04)	2.09 (0.07)	[a]2.93 [c] 2.4
MG2-MG1	3.62 (0.18)	3.67 (0.10)	[a]3.54** [c] 3.6
MG2-WATER1:O	1.94 (0.06)	2.06 (0.06)	[c] 2.6
MG2-WATER2:O	1.93 (0.05)	2.08 (0.08)	[c] 2.5
D830:$O1_{\delta}$—D653:$O2_{\delta}$	2.70 (0.10)	2.98 (0.13)	[a]3.82
A:H3T—D830:$O1_{\delta}$	3.02 (0.24)	2.83 (0.22)	-
A:H3T—dCTP:$O1_{\alpha}$	3.20 (0.33)	3.16 (0.27)	-

Mispair and Oxidative damage

Structural parameters Distances Atom1—Atom2	**G:A MM** **Å**	8oxo**G:C MM** **Å**	8oxo**G:A MM** **Å**
G:C1'—dCTP:C1' $^{(IS)}$	12.04 (0.25)	10.63 (0.16)	10.68 (0.23)
T:C1'—**A**:C1' $^{(PIS)}$	10.66 (0.21)	10.80 (0.23)	10.87 (0.24)
dCTP:P_{α}— A:O3′	4.74 (0.27)	4.50 (0.42)	3.36 (0.18)
MG2—A:O3′	4.72 (0.22)	3.77 (0.23)	2.47 (0.39)
MG2—dNTP:$O2_{\alpha}$/$O1_{\alpha}$	2.83 (0.19)	3.63 (0.15)	2.38 (0.35)
MG2—D831:$O1_{\delta}$	1.83 (0.04)	1.83 (0.04)	1.81 (0.04)
MG2—D653:$O2_{\delta}$	1.82 (0.04)	1.79 (0.04)	1.81 (0.04)
MG2-MG1	3.80 (0.11)	4.13 (0.11)	3.65 (0.12)
MG2-WATER1:O	2.00 (0.07)	1.94 (0.05)	1.95 (0.06)
MG2-WATER2:O	1.96 (0.06)	1.91 (0.05)	1.94 (0.06)
MG2-WATER3:O	1.95 (0.06)	2.02 (0.08)	-
D830:$O1_{\delta}$—D653:$O2_{\delta}$	2.62 (0.08)	2.60 (0.07)	2.70 (0.10)
A:H3T—D830:$O1_{\delta}$	4.42 (0.24)	2.76 (0.49)	2.94 (0.18)
A:H3T—dCTP:$O1_{\alpha}$/ $O2_{\alpha}$	4.77 (0.34)	4.14 (0.51)	3.07 (0.28)

* Modeled

** Mn replaces MG1 in the crystal structure

IS insertion site (From ref [1]: the site occupied by the incoming nucleotide and its pairing template base n)

PIS post insertion site (from ref [1] : the n -1st base pair)

(a) Crystal structure of BF ternary complex (PDB id: 1LV5)

(b) From Johnson and Beese [2]

(c) From Doublie et. al. [5]

Table S4: Top: Comparison of structural data for the **G:C** case obtained from classical MM trajectories (last 5 ns) and QMMM trajectories (last 5 ps) MD simulations with crystal structure values for BF and T7 DNA pol. Bottom: MM averages of structural data for **G:A**, **8oxoG:C**, and 8oxo**G:A** simulations.

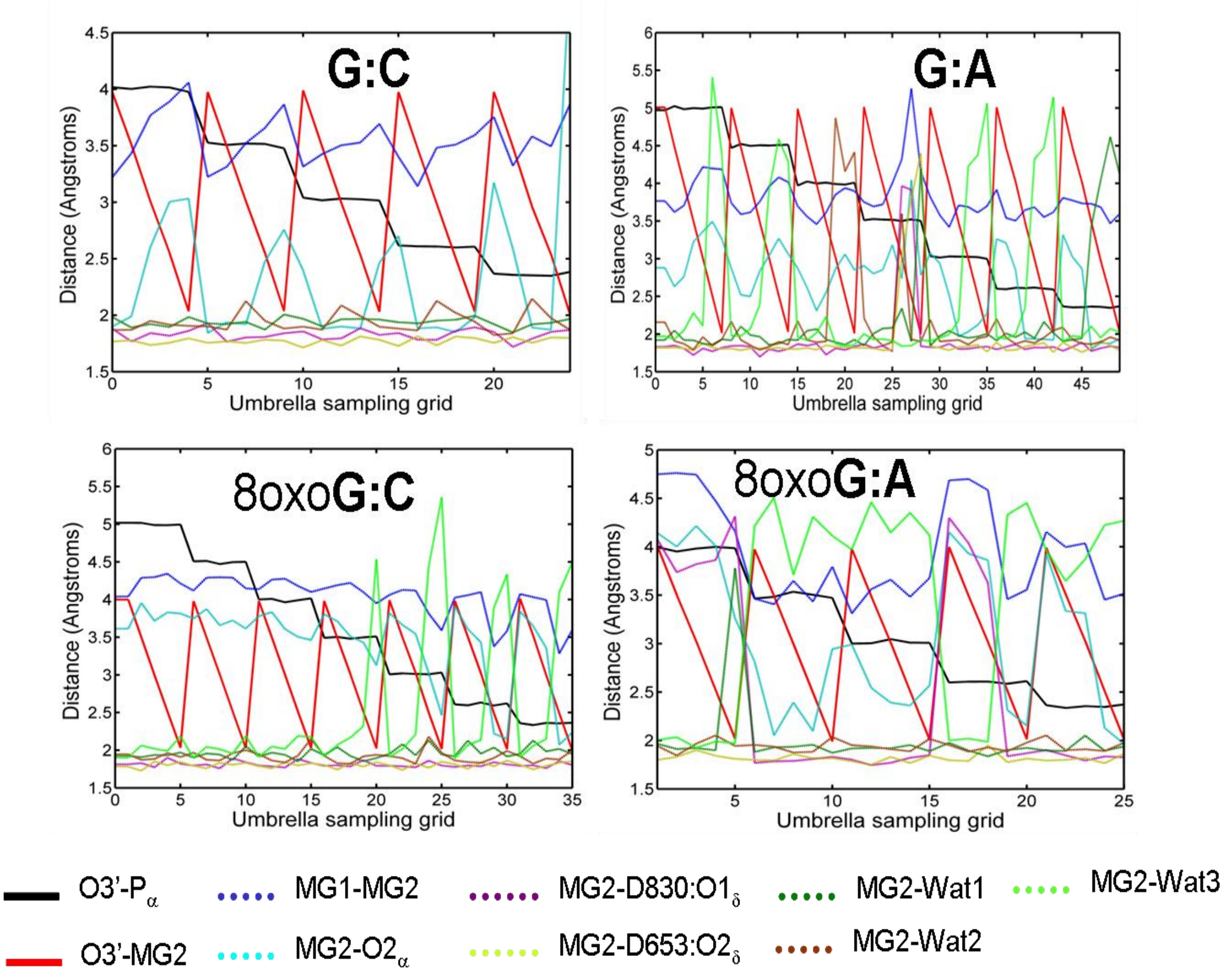

Figure S5: Top left: Variation of catalytic site distances during the constrained umbrella sampling runs for the control **G:C** system where d_a (O3′-P$_\alpha$) and d_b (O3′-MG2) are constrained at 5 different values each(4.0 Å to 2.0 Å in steps of 0.5 Å). The X-axis in the plot runs over the 25 grid points with d_b varying fastest. Similar plots are shown for the **G:A**, 8oxo**G:C** and 8oxo**G:A** model systems as indicated.

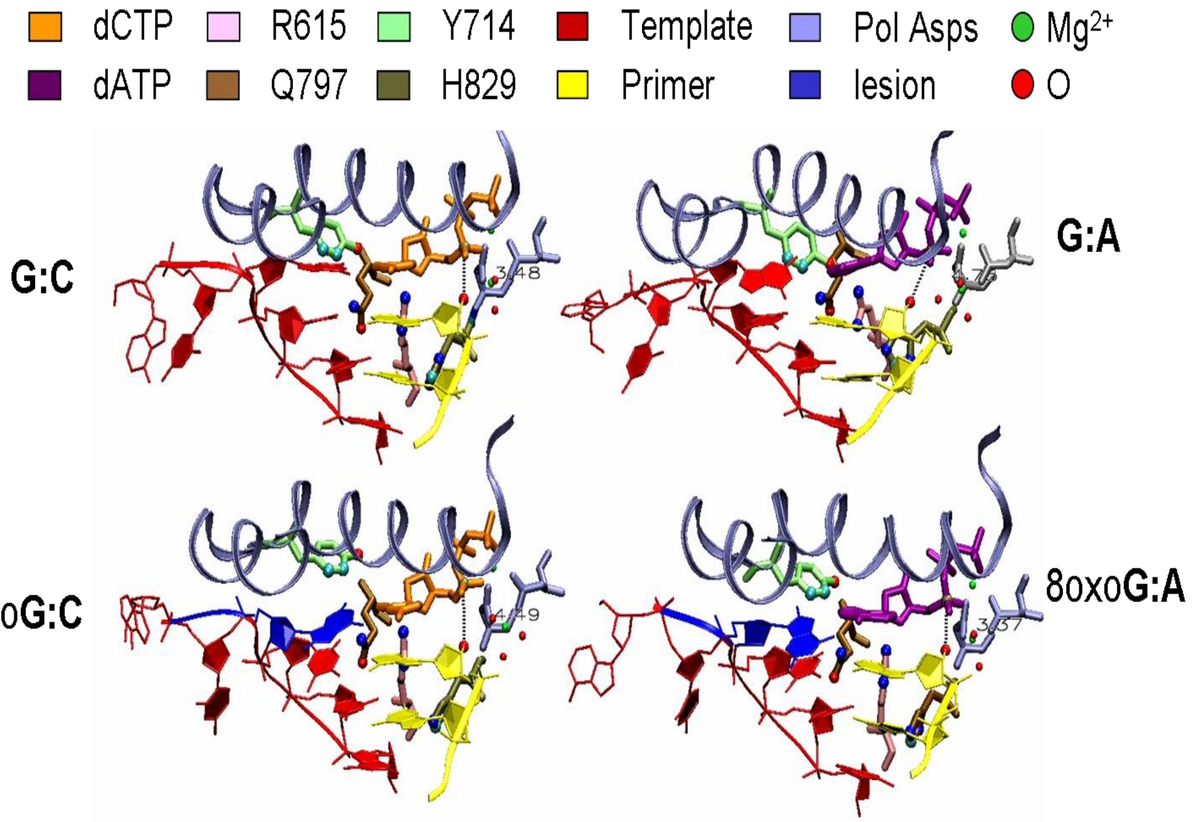

Figure S6: The region considered in principal component analysis. The region comprises of bases of the DNA template (red) and the primer (yellow), the incoming nucleotide, conserved components of the catalytic site (acidic aspartate residues D830 & D653 (iceblue-licorice), Mg^{2+} ions and bound water molecules), two O-helices (iceblue-ribbons), (the O and O1-helix) from BFs fingers domain forming the polymerase finger domain and four polymerase residues identified by experimental mutagenesis studies to be crucial for polymerase fidelity. The catalytic $O3'$-P_{α} distance is shown (dashed line)

This figure also serves as a caption for **Movie S6** which shows the showing the dominant principal component mode (mode 1) of the active site for the four systems. The movie shows that for all four model systems the dominant mode at the active site shows motions of the polymerase fingers, highly correlated with distortions of the DNA fragment. While these distortions are efficiently communicated to the catalytic site in the case of the **G:C** control due to an optimally organized active site it is not so for the other three cases.

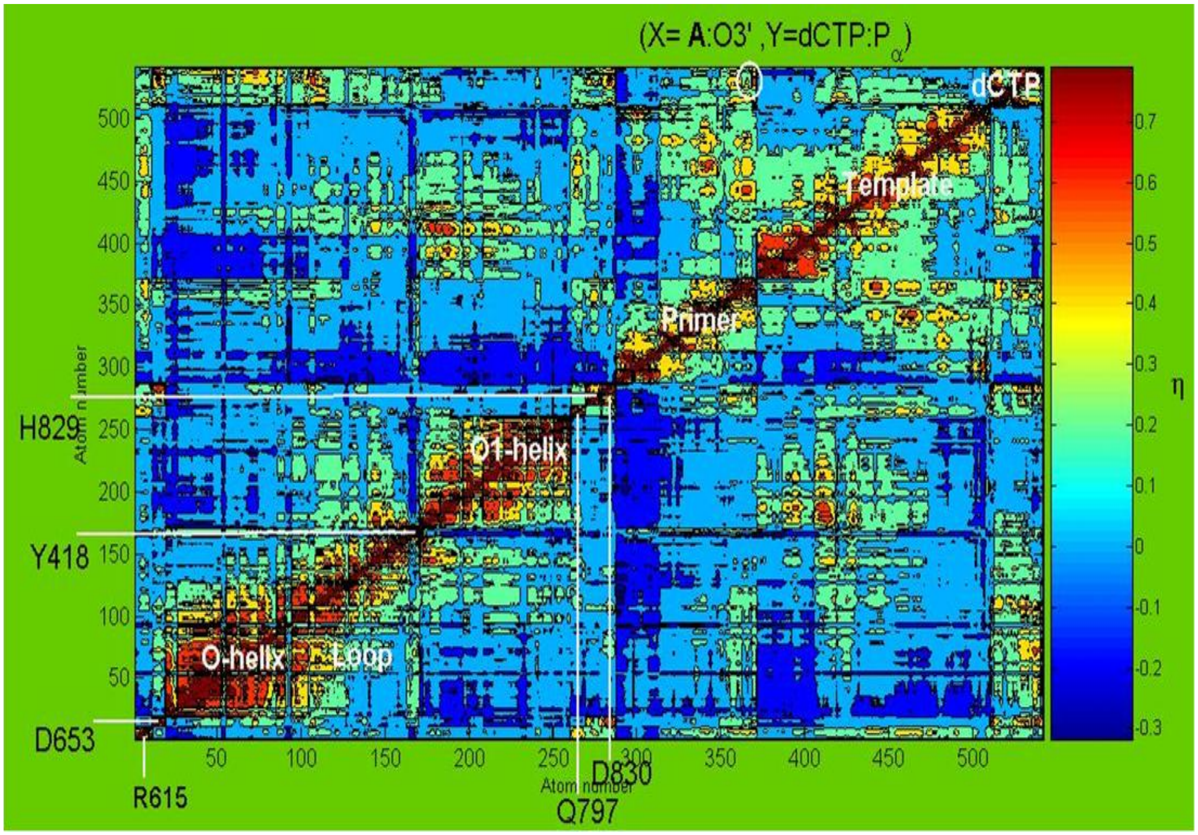

Figure S7: Correlations between vector displacements (**r - <r>**) of atoms in the active site region (Figure S5) for the **G:C** system. Here **r** is a vector drawn from the origin to the atom of interest with average value **<r>.**

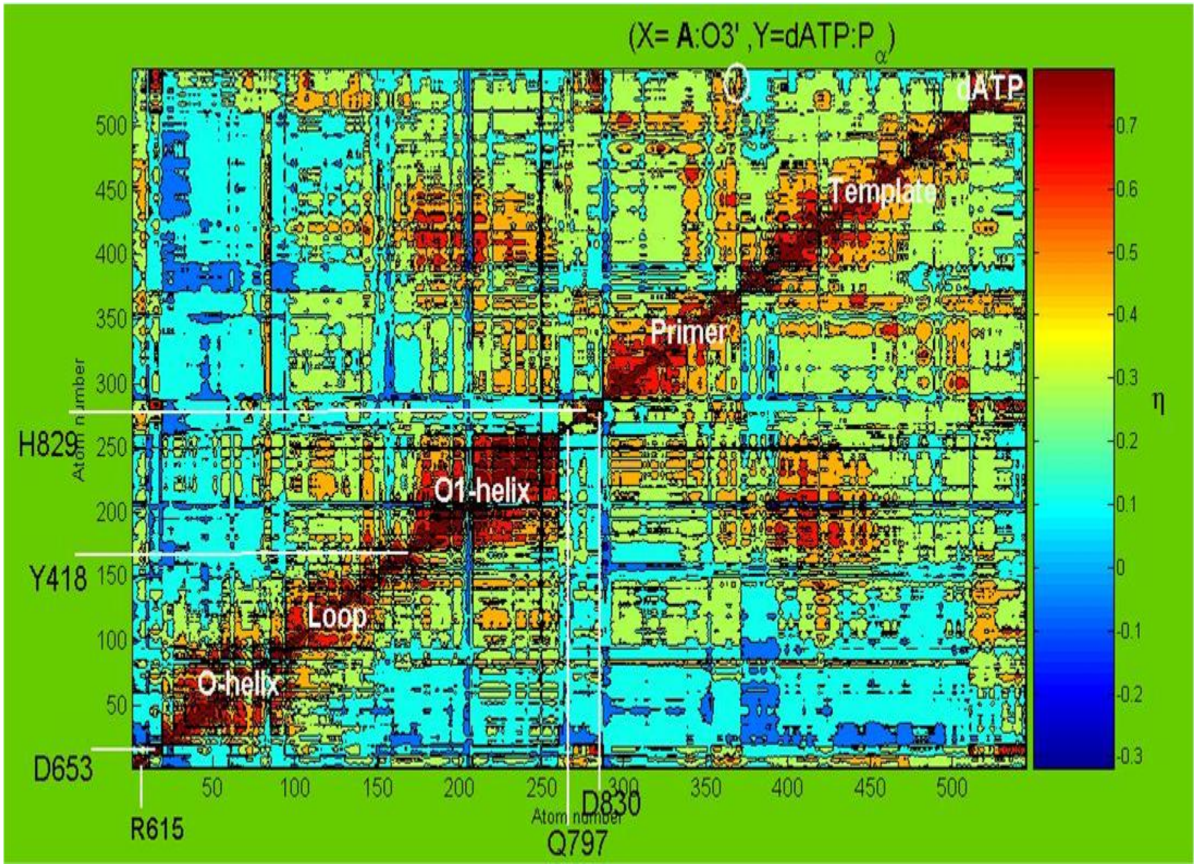

Figure S8: Correlations between vector displacements (**r - <r>**) of atoms in the active site region (Figure S5) for the **G:A** system.

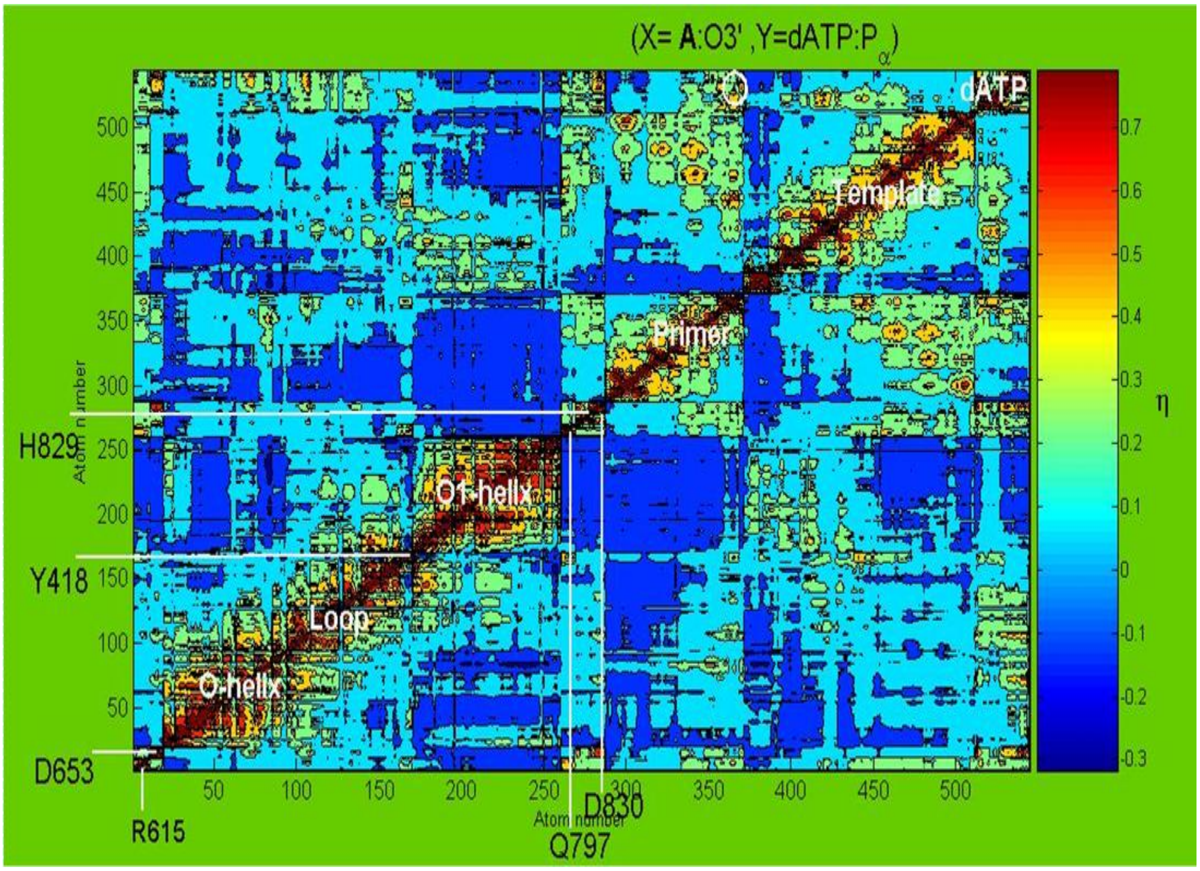

Figure S9: Correlations between vector displacements (**r - <r>**) of atoms in the active site region (Figure S5) for the 8oxo**G:C** system.

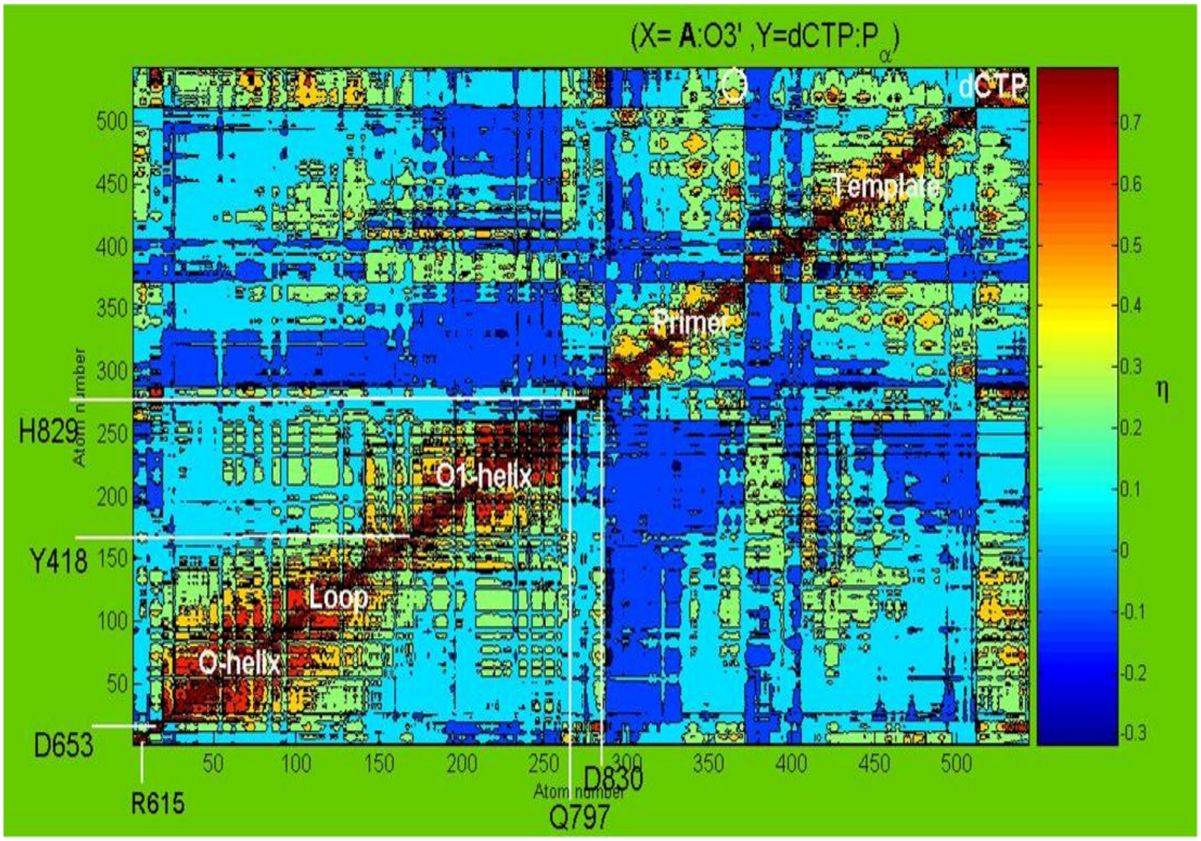

Figure S10: Correlations between vector displacements (**r - <r>**) of atoms in the active site region (Figure S5) for the 8oxo**G:A** system.