Supplementary Information:

Impact of DNA sequences on DNA recognition and 'opening' by Rad4/XPC nucleotide excision repair complex

Debamita Paul[1*], Hong Mu[2*], Amirrasoul Tavakoli[1], Qing Dai[3], Sagnik Chakraborty[4±], Chuan He[3,5], Anjum Ansari[4], Suse Broyde[2†], Jung-Hyun Min[1†]

[1]Department of Chemistry and Biochemistry, Baylor University, Waco, TX 76798, USA
[2]Department of Biology, New York University, New York, NY 10003, USA
[3]Department of Chemistry, The University of Chicago, Chicago, IL 60637, USA
[4]Department of Physics, University of Illinois at Chicago, Chicago, IL 60607, USA
[5]Department of Biochemistry and Molecular Biology, Howard Hughes Medical Institute, The University of Chicago, Chicago, IL 60637, USA

*These authors contributed equally to this work.

†To whom correspondence may be addressed

± Present address: Vector Educational and Consultancy Services, 7/22, Ekdalia, Ballygunge, Kolkata, West Bengal 700019, India

Tel: (254)710-2095 (Min); (212)998-8231 (Broyde)   Fax: (254)710-4272 (Min)

E-mail: JungHyun_Min@baylor.edu; broyde@nyu.edu

Content:
SI Methods, SI Discussion,
SI Tables S1-S5
SI Figures S1-S12
SI Movies S1-S2
SI References

1. SI Methods

1.1. Apparent binding affinities ($K_{d,app}$) determined by competition electrophoretic mobility shift assays (EMSA).

The specified Rad4-Rad23 complexes (0-300 nM) were mixed with 5 nM ^{32}P-labelled DNA (matched or mismatched) in the presence of 1000 nM cold, undamaged DNA (CH7_NX) (**Figures S1**) in a binding assay buffer (5 mM BTP-HCl, 75 mM NaCl, 5 mM DTT, 5% glycerol, 0.74 mM 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate (CHAPS), 500 µg/ml bovine serum albumin (BSA), pH 6.8). Mixed samples were incubated at room temperature for 20 min before being separated on 4.8% non-denaturing polyacrylamide gels. The gels were quantitated by autoradiography using Typhoon FLA 9000 imaging scanner (GE) and Image Lab software (Version 5.2.1 build 11, 2014; Bio-Rad). The apparent dissociation constants ($K_{d,app}$) were calculated as previously described using curve fitting by nonlinear regression using Origin software (OriginPro 9.6.0.172, Academic) (1-7).

1.2. Streak-seeding for crystallization.

To obtain larger crystals, the crystal showers were used for streak-seeding by transferring 5–10 crystals (~20 µm) into a 10 µl drop of harvest buffer (50 mM BTP-HCl, 200 mM NaCl and 12% isopropanol pH 6.8). The crystals were smashed using a fine needle and were transferred into a tube containing 250 µl of the harvest buffer, which was then mixed by gentle vortexing. The freshly prepared seed-stocks always produced better quality crystals. The best diffracting crystal were obtained from drops pre-equilibrated for 1 h (50 mM BTP-HCl, 200 mM NaCl and 0-5% isopropanol, pH 6.8), seeded by passing the tip of a cat's whisker dipped into a fresh seed-stock solution through the drop. The crystals grew to maximum size of ~70-80 µm in 10-12 days.

1.3. Fluorescence lifetime measurements for DNA and DNA protein complexes.

The fluorescence decays of DNA labeled with tC^o and tC_{nitro} were measured using a time-correlated single-photon counting (TCSPC) system (DeltaFlex, HORIBA) equipped with a Ti-sapphire laser source with a tunable range from 690 to 1040 nm (Mai Tai HP, Spectra-Physics), as described previously (5) . The instrument response function (IRF) of the system was measured using a dilute aqueous solution (3% w/w) of LUDOX AM colloidal silica (Sigma-Aldrich). The full-width half maxima (FWHM) was ~425 ps. Fluorescence decay curves were recorded on a 100 ns timescale, resolved into 4,096 channels, to a total of 10,000 counts in the peak channel. The power of the laser delivered to the samples was 0.21 mW/cm^2 as measured by a General Tools Digital UVA/UVB Meter, 280-400 nm (#UV513AB).

1.4. Melting temperatures of DNA duplexes used in this study.

The overall thermal stabilities of the DNA duplexes (**Figures S8**) were measured as follows. The absorbance at 260 nm of each DNA duplex (1.5 µM) was measured in a sample cuvette of path length 1 cm, using a Cary 300 Bio UV-Visible

spectrophotometer equipped with a Varian temperature controller. The absorbance measurements were done from 30 to 85 °C at every 1.0 °C interval. The absorbance data were subsequently averaged (smoothened) over a 2.5 ºC (5 data points) window prior to calculating derivatives using Origin software OriginPro. Melting temperatures were then determined from the melting profiles using methods described previously (2,5).

1.5. Initial models, water box and counterions. The initial models of the two 13-mer DNA sequences (nucleotide steps 10 – 22, **Figure 1B** in the main text) were created using standard B-DNA in Discovery Studio 2.5 (Accelrys Software Inc.). The DNA models were neutralized with Na^+ counterions and solvated with explicit TIP3P water(8) in a cubic periodic box with side length of 59.0 Å using the tLEAP module of the AMBER18 suite of programs (9). The initial models of the Rad4 docked to the two DNA sequences were prepared as in earlier work (10), where the BHD2 and BHD3 domains are close but not bound to the sequences yet. The protein-DNA complexes were neutralized with Na+ counterions and solvated with explicit TIP3P water (8) in a cubic periodic box with side length of 125.0 Å using the tLEAP module of the AMBER18 suite of programs (9).

1.6. MD simulations. We used the ff14SB force field for the MD simulations (11). All MD simulations were carried out using the AMBER18 suite of programs (9). The Particle-Mesh Ewald method (12) with 9.0 Å cutoff for the non-bonded interactions was used in the energy minimizations and MD simulations. Minimizations were carried out in three stages. First, 500 steps of steepest descent minimization followed by 500 cycles of conjugate gradient minimization were conducted for the water molecules and counterions with a restraint force constant of 50 kcal/(mol·$Å^2$) on the solute molecules. Then, 500 steps of steepest descent minimization followed by 500 cycles of conjugate gradient minimization were carried out for the water molecules and counterions with a restraint force constant of 10 kcal/(mol·$Å^2$) on the solute molecules. In the last round, 500 steps of steepest descent minimization followed by 500 cycles of conjugate gradient minimization were carried out on the whole system without restraints. The minimized structures were then subjected to three rounds of equilibration. First, each system was equilibrated at constant temperature of 10º K for 30 ps with the solute molecules fixed with a restraint force constant of 50 kcal/(mol·$Å^2$). Then the system was heated from 10º K to 300º K over 300 ps with the solute molecules fixed with a restraint force constant of 50 kcal/(mol·$Å^2$) at constant volume. In the last round of equilibration, the restraint force constant on the solute was reduced through three steps: at 10 kcal/(mol·$Å^2$) for 200 ps, at 1 kcal/(mol·$Å^2$) for 200 ps, and then at 0.1 kcal/(mol·$Å^2$) for 200 ps with constant pressure at 300º K. Following equilibration, production MD simulations for each system were carried out in a constant-temperature, constant-pressure (NPT)

ensemble at 300º K and constant pressure of 1 Atm for 2 μs. The temperature was controlled with a Langevin thermostat (13) with a 5 ps^{-1} collision frequency. The pressure was maintained with the Berendsen coupling method (14). A 2.0 fs time step and the SHAKE algorithm (15) were applied in all MD simulations. A 1 kcal/mol restraint was applied to the end base pair hydrogen bond donor and acceptor atom pairs during the production MDs.

1.7. Structural analyses

Principal component analysis (PCA) and RMSD analysis. PCA analyses for the ensemble of structures from MD simulations of the DNA duplexes were performed with heavy atoms of the DNA sequences excluding flexible end base pairs, using R (16) with the Bio3D package (17). The structures from each ensemble were superposed to the first frame at the heavy atoms of the base pairs 3' to the run of guanines and serve as our reference for structural changes due to the sequence effect. Due to the fact that each ensemble provides only one cluster, no clustering was performed. The RMSD for the heavy atoms of the DNA sequences excluding flexible end base pairs were calculated for the ensemble of structures in the production MD, using the cpptraj module of AMBER18 (9). PCA analyses for the Rad4 initial binding to the DNA duplexes were carried out as described in earlier work (4). The RMSD for the heavy atoms of the BHD2 domain and potential open site 6-mer DNA sequences (nucleotide steps 14 – 19, **Figure 1B** in the main text) were calculated for the ensemble of structures in the production MD. The best representative structure for each ensemble is defined as the one frame that has the shortest RMSD for the heavy atoms used in the calculation of RMSD values to all other frames.

Base pair step parameters. We measured base pair step parameters (shift, tilt, slide, roll, rise and twist, **Figures 5** and **S10**) between each base pair step of the 13-mer duplexes excluding the flexible end base pairs, using the cpptraj module of AMBER18 (9).

Helix bend and DNA bending direction pseudo-dihedral angle. We measured the total helix bend angle for the duplex excluding flexible end base pairs using Curves+ (18) (**Figure S10**). The DNA bend direction was measured using a pseudo-dihedral angle, defined in **Figure S10**, using the cpptraj module of AMBER18 (9).

Van der Waals stacking interaction energy. The van der Waals stacking interaction energy for the base pairs of the potential open site 6-mer sequences (base pair steps 14 – 19 in **Figure 1B** of the main text) were calculated using the cpptraj module of AMBER18 for the Lennard-Jones potential (9).

Analysis of DNA Untwisting upon Rad4 binding. The untwisting of the lesion-containing DNA along the initial binding trajectory was monitored by the Untwist angle defined as Untwist = Twist $_{initial}$ −Twist.

The twist angle was measured between the end base pairs of the lesion-containing 6-mer (base pair steps 14 and 19 in **Figure 1B** of the main text) using the cpptraj module of AMBER18 (9). Twist $_{initial}$ is the ensemble average twist angle of the lesion-containing 6-mer during the first 1 ns of production MD during which significant untwisting was not observed (**Figure S11**); this ensemble represents the state of the lesion-containing sequence before the engagement of Rad4, especially BHD2. Positive values indicate further untwisting and negative values indicate further twisting.

The AlphaSpace volumes analyses for BHD2 binding to the minor groove. The best representative structure for the initial binding state of each duplex was further analyzed to quantify BHD2's binding into the DNA minor groove around the lesion site. The alpha space volumes of the binding pockets in the DNA and their occupancies by BHD2 were calculated using AlphaSpace v1.0 (19): the DNA was set as the receptor and BHD2 was set as the ligand. The total occupied alpha space volume was used to quantify the extent of BHD2 binding into the DNA minor groove (see **Figure S10** in (4)). The value reflects the curvature and surface area of the DNA minor groove region that is occupied by BHD2.

Analysis of hydrogen bonding between BHD2 and DNA duplexes. Hydrogen bonds between BHD2 and the DNA duplexes were counted for the best representative structures in each ensemble using each pair of donor and acceptor atoms that has a hydrogen bond (heavy-light-heavy atom) angle ≥145° and heavy-to-heavy atom distance ≤ 3.3 Å.

Block average analyses. The DNA structural parameters and van der Waals energies for the stable ensemble of each MD simulation were analyzed using the block averaging method (20,21). In brief, the time series data were divided into "blocks" with a block size that exceeds the longest correlation time, 50 ns in our case. The average for each block was computed and termed "block average". The mean values and the standard deviations of the block averages given in the main text were used to represent the average and the variance of averages.

Molecular structures were rendered using PyMOL 1.3.x (Schrodinger, LLC.) or UCSF Chimera (developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311) (22). All MD simulation data were plotted using MATLAB 7.10.0 (The MathWorks, Inc.).

2. SI Discussion

As described in the main text, the DNA in this 'reverse mode' structure was obtained with a 23-bp matched DNA containing CGC/GCG sequence whereas the 'open' or pseudo-open structures solved so far

contained 24-bp DNA with various lesions and sequences. Here, we argue that the differences in the DNA lengths did not likely cause the different binding modes but rather played a role in capturing existing conformation(s) in solution into crystals by promoting crystal packings for a given conformation.

First of all, it can be structurally demonstrated that adding or removing one nucleotide from one of the DNA ends (L-side) in a 23-bp or 24-bp DNA would not impact any known protein-DNA contacts in either binding modes. For instance, if the protein were to bind in a 'reverse mode' and cap the end of a 24-bp DNA (for instance, containing CCC/GGG), it should have been allowed just the same as was seen in the 23-bp CGC/GCG DNA, as the S-side sequences of the DNA of both DNA are the same. Conversely, if the protein were to 'open' the CGC/GCG DNA, the 23-bp construct had a length sufficient for making all the contacts with the protein in an 'open' conformation, as the contacts on L-side of the DNA does not extend much beyond the potential 'open' site where nucleotides get flipped out which is present in both the 23- and 24-bp DNA (4). However, such alternative structures have not been observed and the crystals of Rad4 tethered to a 23-bp CCC/GGG or with 24-bp CGC/GCG did not diffract well (data not shown). This indicates that these lengths of DNA were not likely optimal for the crystallographic packing of the given complex structure. In other words, the crystals we obtained with 23-bp CCC/GGG were not likely to be in 'reverse mode' (which would have preferred 23-bp for packing) and the crystals we obtained with 24-bp CGC/GCG were not likely to be in an 'open' structure (which would have preferred 24-bp). Thus, the differences in the 'reverse' vs 'open(-like)' mode of binding for CGC/GCG and CCC/GGG sequences are therefore not likely instigated by the 1-bp difference in the lengths but by the other difference in the sample which is the differences in the central DNA sequences.

SI Tables S1-S5

Table S1. Names and sequences of the DNA constructs containing crosslinkable G* (purple) used for the crystallization trials. Yellow highlights indicate the CG/GC repeats. The positions of the CGC/GCG 3-bp sequences after which the constructs are named are indicated in bold.

DNA	Sequences
CGC/GCG (CH9a, 24-bp)	5'–TTGACTCG*ACAT CG**CGC**GC TACAA –3' 3'– ACTGAGCTGTA GC**GCG**CG ATGTTA –5'
CGC/GCG (CH9b, 25-bp)	5'–TTGACTCG*ACAT CG**CGC**GC TACAAA –3' 3'– ACTGAGCTGTA GC**GCG**CG ATGTTTA –5'
CGC/GCG (CH9c, 23-bp)	5'–TTGACTCG*ACAT CG**CGC**GC TACA –3' 3'– ACTGAGCTGTA GC**GCG**CG ATGTA –5'
CGC/GCG (CH9d, 22-bp)	5'–TTGACTCG*ACAT CG**CGC**GC TAC –3' 3'– ACTGAGCTGTA GC**GCG**CG ATGA –5'

Table S2. X-ray diffraction data collection and structure refinement statistics (molecular replacement).

	6UG1 (SC41bxCH9c)
Data collection	
Wavelength (Å)	0.9794
Space group	*P 1*
Cell dimensions	
a, b, c (Å)	53.2 59.6 78.2
α, β, γ (°)	105.5 97.9 107.1
Resolution (Å)	50.0 – 2.9
R_{sym} or R_{merge}	11.6 (51.1)
$I / \sigma I$	9.7 (2.2)
Completeness (%)	90.8% (74.1%)
Redundancy	3.4 (2.9)
Refinement	
Resolution (Å)	37.4-2.8 (2.93-2.83)
No. reflections	17956 (1519)
R_{work} / R_{free}(%)	22.57/27.39
No. atoms	5231
Protein	4300
DNA	931
Water	0
B-factors (Å^2)	76.8
Protein	62.3
DNA (G47)	91.3(91.7)
Water	0
R.m.s. deviations	
Bond lengths (Å)	0.005
Bond angles (°)	0.97

R.m.s., root mean squared.

*Values in parentheses are for highest-resolution shell.

Table S3. (A) Names and sequences of the DNA constructs used for the FLT studies shown in Figures 3 and S6. 'D' indicates tC^o (FRET donor) and 'P' is tC$_{nitro}$ (FRET acceptor). G* is disulfide-modified guanine for tethering. The segments containing consecutive C/G's or alternating CG/GC repeats are in bold. To keep the probes (which are cytosine analogs) in the same positions as before (2,3,5,6), the CG/GC's in the sequences used for crystallization were changed to GC/CG's for FLT. The positions of two flipped-out nucleotides in Rad4-bound 'open' structures are boxed in pink. **(B-E)** Fluorescence lifetimes from MEM analyses: **(B)** Donor lifetimes (τ_D) from samples containing donor-only DNA, CGC/GCG$_F$_D. **(C)** Fluorescence lifetimes (τ_{DA}), fractional populations and FRET efficiencies of samples containing donor/acceptor-labeled DNA, CGC/GCG$_F$. **(D)** Values for the same DNA sequence as CGC/GCG$_F$ except without the crosslinkable G* (CGC/GCG$_F$_noG*). **(E)** Values for samples containing donor/acceptor-labeled CCC/GGG$_F$ or CCC/CCC$_F$. The data are from ref. (5).

[a] Corresponding to the longest lifetime 'zero-FRET' component.

τ_n, A_n, and E_n (n=1,2,3) indicate the fluorescence lifetimes, normalized amplitudes and FRET efficiencies for each Gaussian peak of the samples. The average FRET efficiencies of the samples were calculated as $< E > = 1 - \frac{<\tau_{DA}>}{\tau_D}$, where the donor-only lifetimes (τ_D) for the DNA and DNA with WT Rad4 samples were 4.79 and 4.81 ns respectively. The average FRET efficiencies were calculated using $\tau_D = 4.8$ ns and the uncertainties reported are standard deviations (s.d.) from 3 independent sets of measurements.

A

DNA	Sequences
CGC/GCG$_F$	5'-TTGACTCG*ACATCP**GCGCG**TACAA -3' 3'- ACTGAGCTGTAG**GCGCG**DATGTTA -5'
CGC/GCG$_F$_D	5'-TTGACTCG*ACATCC**GCGCG**TACAA -3' 3'- ACTGAGCTGTAG**GCGCG**DATGTTA -5'
CGC/GCG$_F$_noG*	5'-TTGACTCGACATCP**GCGCG**TACAA -3' 3'- ACTGAGCTGTAG**GCGCG**DATGTTA -5'
CCC/GGG$_F$	5'-TTGACTCG*ACATC**CPCCC**GGTACAA -3' 3'- ACTGAGCTGTA**GGGGG**CDATGTTA -5'
CCC/CCC$_F$	5'-TTGACTCG*ACATC**CPCCC**GGTACAA -3' 3'- ACTGAGCTGTA**GGCCCC**DATGTTA -5'

B

	τ_D
CGC/GCG$_F$_D	4.79 ± 0.026
CGC/GCG$_F$_D + WT	4.81 ± 0.023

9

C

	τ_1	A1	E1	τ_2	A2	E2	τ_3	A3 [a]	E3 [a]	<E>
CGC/GCG$_F$	0.233 ± 0.029	86.102 ± 4.120	0.962 ± 0.009	1.133 ± 0.112	8.133 ± 2.581	0.826 ± 0.0184	4.269 ± 0.596	5.726 ± 2.433	0.119 ± 0.012	0.93 ± 0.004
CGC/GCG$_F$ + WT	0.232 ± 0.059	85.523 ± 0.797	0.951 ± 0.013	1.021 ± 0.228	9.343 ± 1.569	0.797 ± 0.045	3.935 ± 0.773	5.133 ± 0.783	0.176 ± 0.015	0.885 ± 0.006
CGC/GCG$_F$ x WT	0.316 ± 0.012	66.297 ± 0.209	0.941 ± 0.003	1.676 ± 0.124	8.430 ± 0.509	0.650 ± 0.024	4.821 ± 0.249	23.270 ± 0.445	0.004 ± 0.005	0.646 ± 0.002
CGC/GCG$_F$ x Δβ-hairpin3	0.235 ± 0.028	65.076 ± 0.906	0.953 ± 0.004	1.625 ± 0.077	8.036 ± 0.109	0.684 ± 0.016	4.848 ± 0.042	25.69 ± 0.646	0.01 ± 0.008	0.687 ± 0.009

D

	τ_1	A1	E1	τ_2	A2	E2	τ_3	A3 [a]	E3 [a]	<E>
CGC/GCG$_F$ _noG*	0.248 ± 0.026	80.18 ± 6.412	0.944 ± 0.011	1.569 ± 0.023	9.020 ± 2.011	0.644 ± 0.042	4.812 ± 0.242	10.80 ± 1.433	0.01 ± 0.003	0.887 ± 0.004

E

	τ_1	A1 (%)	E1	τ_2	A2 (%)	E2	τ_3	A3[a] (%)	E3 [a]	<E>
CCC/GGG$_F$	0.316 ± 0.018	85.67 ± 1.331	0.935 ± 0.003	1.067 ± 0.177	8.040 ± 0.297	0.783 ± 0.036	4.374 ± 0.301	8.220 ± 1.038	0.110 ± 0.061	0.864 ± 0.000
CCC/GGG$_F$ x WT	0.608 ± 0.070	25.07 ± 1.210	0.876 ± 0.014	1.836 ± 0.009	46.85 ± 1.325	0.628 ± 0.001	4.827 ± 0.135	28.08 ± 0.657	0.188 ± 0.027	0.635 ± 0.006
CCC/GGG$_F$ x Δβ-hairpin 3	0.615 ± 0.011	24.79 ± 1.881	0.874 ± 0.002	1.676 ± 0.018	47.04 ± 4.402	0.659 ± 0.003	4.805 ± 0.181	28.17 ± 2.522	0.024 ± 0.036	0.697 ± 0.116
CCC/CCC$_F$ + WT	0.975 ± 0.019	35.13 ± 1.088	0.802 ± 0.003	1.934 ± 0.002	41.81 ± 1.063	0.607 ± 0.000	4.807 ± 0.019	23.22 ± 0.258	0.024 ± 0.003	0.539 ± 0.000

10

Table S4. Fluorescence lifetimes, fractional populations, and FRET efficiencies for CGC/GCG$_F$ covalently tethered to WT Rad4 (**A**) or to Δβ-hairpin3 (**B**) after being treated with dithiothreitol (DTT) for indicated times (0 min, 5 min, 1 h and 4 h), as shown in **Figure S7**. The average FRET efficiencies were calculated using $\tau_D = 4.8$ ns. [a] Corresponding to the longest lifetime 'zero-FRET' component.

A

	τ_1	A1	E1	τ_2	A2	E2	τ_3	A3 [a]	E3 [a]	\<E\>
CGC/GCG$_F$ x WT	0.291	64.01	0.939	1.626	9.11	0.661	4.474	27.88	0.068	0.616
CGC/GCG$_F$ x WT + DTT_5min	0.326	73.92	0.932	1.826	6.87	0.619	4.431	21.30	0.077	0.706
CGC/GCG$_F$ x WT + DTT_1h	0.348	90.74	0.927	1.818	2.75	0.621	4.373	6.51	0.088	0.861
CGC/GCG$_F$ x WT + DTT_4h	0.362	90.58	0.924	1.808	4.24	0.623	4.586	6.43	0.059	0.865

B

	τ_1	A1	E1	τ_2	A2	E2	τ_3	A3 [a]	E3 [a]	\<E\>
CGC/GCG$_F$ x Δβ-hairpin3	0.231	65.12	0.949	1.526	9.21	0.661	4.574	25.67	0.068	0.606
CGC/GCG$_F$ x Δβ-hairpin3 + DTT_5min	0.317	74.12	0.931	1.726	6.27	0.619	4.431	19.61	0.077	0.712
CGC/GCG$_F$ x Δβ-hairpin3 + DTT_1h	0.328	91.74	0.928	1.806	3.75	0.621	4.373	4.51	0.088	0.851
CGC/GCG$_F$ x Δβ-hairpin3 + DTT_4h	0.342	91.34	0.925	1.812	4.94	0.623	4.453	3.72	0.059	0.869

Table S5. (A) Names and sequences of the DNA constructs used in **Figure 4**. 'D' indicates tC^o (FRET donor) and 'P' is tC$_{nitro}$ (FRET acceptor). **(B)** Fractional populations (%) of the B-DNA conformation (τ_1) as a function of temperature. The fractional populations are computed from the area under the Gaussian peak describing the shortest lifetime component (τ_1) of the MEM lifetime distributions. **(C)** Average FRET versus temperature. The average FRET was calculated as $< E > = 1 - \frac{<\tau_{DA}>}{\tau_D}$, where the donor-only lifetimes (τ_D) for the DNAs were taken as 4.8 ns. The uncertainties reported are standard deviations (s.d.) from 2 independent sets of measurements.

A

CCC/GGG$_F$_noG*	5'-TTGACTC**G**ACAT**CP**CCCGGTACAA -3' 3'- ACTGAGCTGTA**GGGGG**C**D**ATGTTA -5'
CGC/GCG$_F$_noG*	5'-TTGACTC**G**ACATC**P**GCGCGTACAA -3' 3'- ACTGAGCTGTAG**GCGCG****D**ATGTTA -5'

B

Temperature (°C)	10	20	30	40	50	60	70	80
CCC/GGG$_F$_noG*	86.70 ± 0.707	84.45 ± 1.060	80.25 ± 0.071	72.25 ± 0.494	58.85 ± 0.919	35.95 ± 0.353	4.95 ± 0.212	0
CGC/GCG$_F$_noG*	90.01 ± 0.705	89.86 ± 0.709	89.25 ± 0.212	87.15 ± 1.343	82.55 ± 1.909	75.20 ± 0.848	53.35 ± 0.777	0

C

| Temperature (°C)	10	20	30	40	50	60	70	80
CCC/GGG$_F$_noG*	0.796 ± 0.002	0.789 ± 0.002	0.761 ± 0.003	0.716 ± 0.002	0.617 ± 0.003	0.431 ± 0.004	0.318 ± 0.004	0.176 ± 0.001
CGC/GCG$_F$_noG*	0.865 ± 0.001	0.865 ± 0.002	0.859 ± 0.002	0.845 ± 0.003	0.812 ± 0.003	0.733 ± 0.004	0.627 ± 0.003	0.235 ± 0.001

SI Figures and Figure Legends

A

DNA	Sequences	Tm (°C)
CCC/CCC (CH10_NX)	5'-TTGACTC**G**ACATCC**CCC**GCTACAA -3' 3'- ACTGAGCTGTAGG**CCC**CGATGTTA -5'	61.0 ± 0.8
CCC/GGG (CH7_NX)	5'-TTGACTC**G**ACATCC**CCC**GCTACAA -3' 3'- ACTGAGCTGTAGG**GGG**CGATGTTA -5'	76.5 ± 1.0
CGC/GCG (CH9a_NX, 24-bp)	5'-TTGACTC**G**ACATCG**CGC**GCTACAA -3' 3'- ACTGAGCTGTAGC**GCG**CGATGTTA -5'	74.3 ± 1.8
CGC/GCG (CH9c_NX, 23-bp)	5'-TTGACTC**G**ACATCG**CGC**GCTACA -3' 3'- ACTGAGCTGTAGC**GCG**CGATGTA -5'	70.2 ± 1.5

B

		WT Rad4	Δβ-hairpin3
CCC/CCC	$K_{d,app}$ (nM)	59 ± 1	198 ± 10
	R^2	0.99	0.97
CCC/GGG	$K_{d,app}$ (nM)	367 ± 19	512 ± 88
	R^2	0.99	0.95
CGC/GCG 24-bp	$K_{d,app}$ (nM)	386 ± 32	708 ± 28
	R^2	0.90	0.96
CGC/GCG 23-bp	$K_{d,app}$ (nM)	390 ± 33	636 ± 23
	R^2	0.86	0.96

Figure S1. Characterization of the apparent binding affinities between Rad4 and DNA. (A) The construct names and DNA duplex sequences used for the competitive EMSA assays. The positions of the 3-bp sequences after which the constructs are named (e.g., CCC/CCC in the CCC/CCC construct) are indicated in bold. The positions of two flipped-out nucleotides in Rad4-bound 'open' structures are boxed in pink. The deoxyguanosines marked red indicate the positions of the crosslinkable G* used for tethering (e.g., **Figure 1B**). **(B)** Quantification of the bound DNA fractions versus the concentrations of the protein from the EMSA. Symbols of different colors indicate different DNA sequences; filled and empty symbols indicate mismatched and matched DNA, respectively; circles and triangles indicate WT and Δβ-hairpin3, respectively. The error bars indicate ±standard deviations from triplicate gels. Solid and dotted lines indicate the fit curves of the data points for WT and Δβ-hairpin3, respectively. The apparent K_d values ($K_{d,app}$) for various DNAs were calculated as described (1-7). These results showed that Δβ-hairpin3 mutant had weakened specific binding but retained most affinities for nonspecific binding, resulting in a significant loss in the lesion recognition specificity, also as discussed in (5). Neither the WT nor mutant protein showed significant difference in binding to 23 bp versus 24 bp nonspecific CGC/GCG DNA.

C
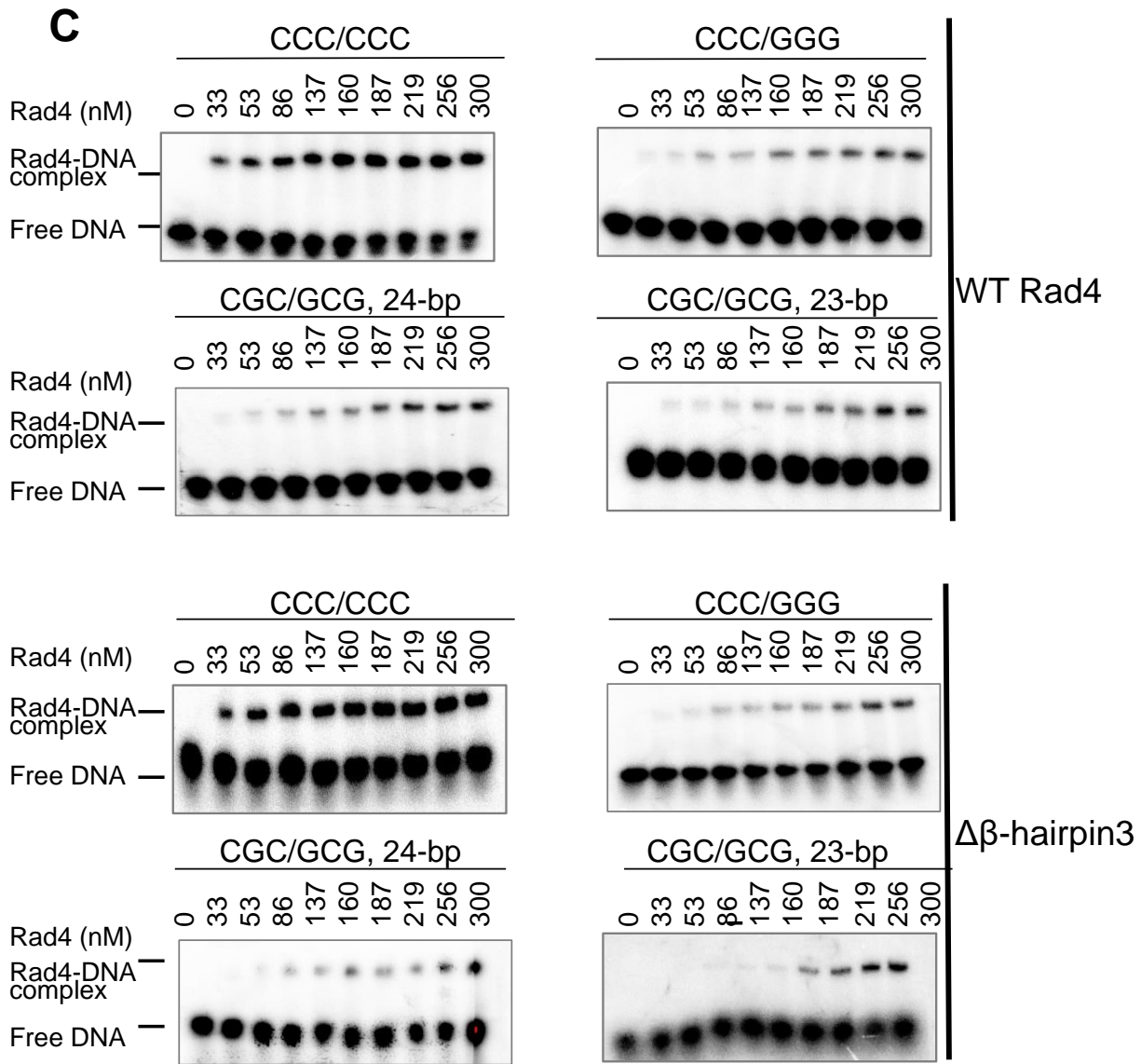
Figure S1. (C) Representative gel images used for (B), showing the binding of different DNA sequences to WT (top) and Δβ-hairpin3 mutant Rad4-Rad23 complexes (bottom).
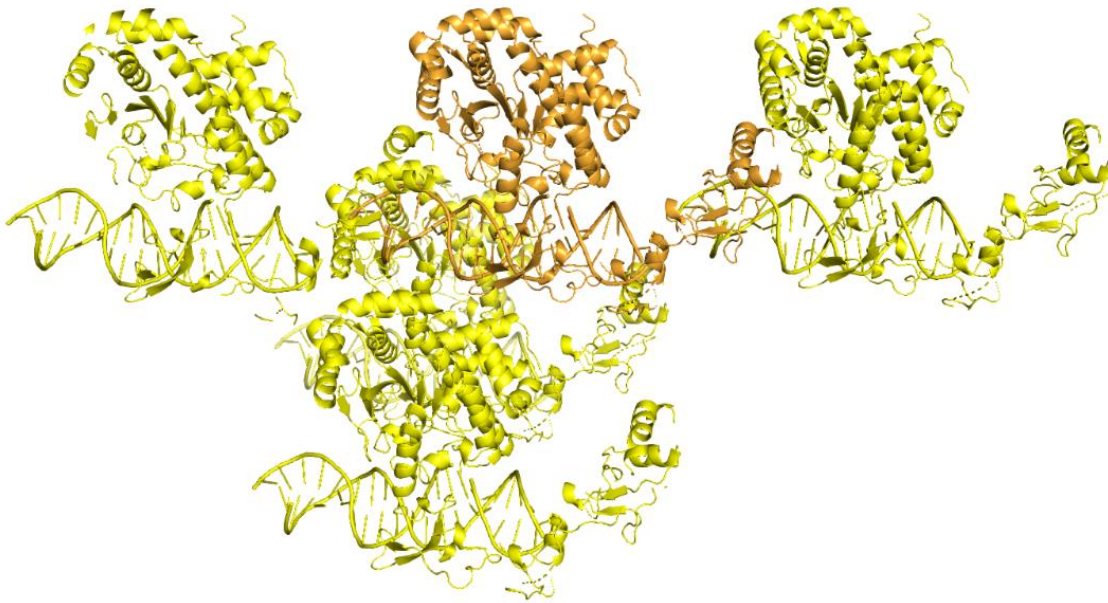
Figure S2. Crystal packing of the Δβ-hairpin3 tethered to CGC/GCG DNA. In this *P1* space group crystal, the BHD2/3 packs against the L-side of the neighboring DNA molecules to make the complexes arrange in a head-to-tail manner.
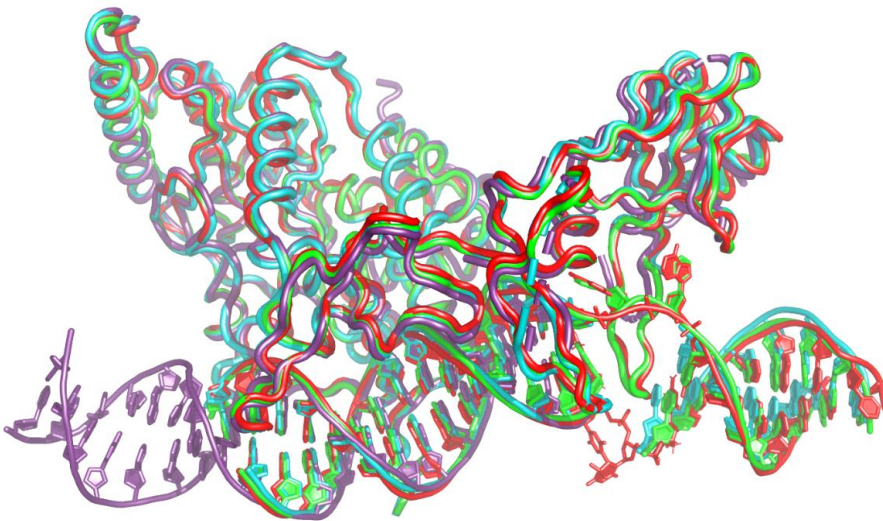
Figure S3. Superposition of the 'reverse mode' and 'open(-like)' Rad4-DNA crystal structures. Δβ-hairpin3 crosslinked to CGC/GCG (PDB ID: 6UGI, purple) shows the 'reverse mode' binding. The WT Rad4 bound to 6-4PP (6CFI, red) and the WT crosslinked to CCC/GGG (4YIR, green) both show the 'open' conformation; Δβ-hairpin3 crosslinked to CCC/GGG (6UBF, cyan) shows the 'open-like' structure.
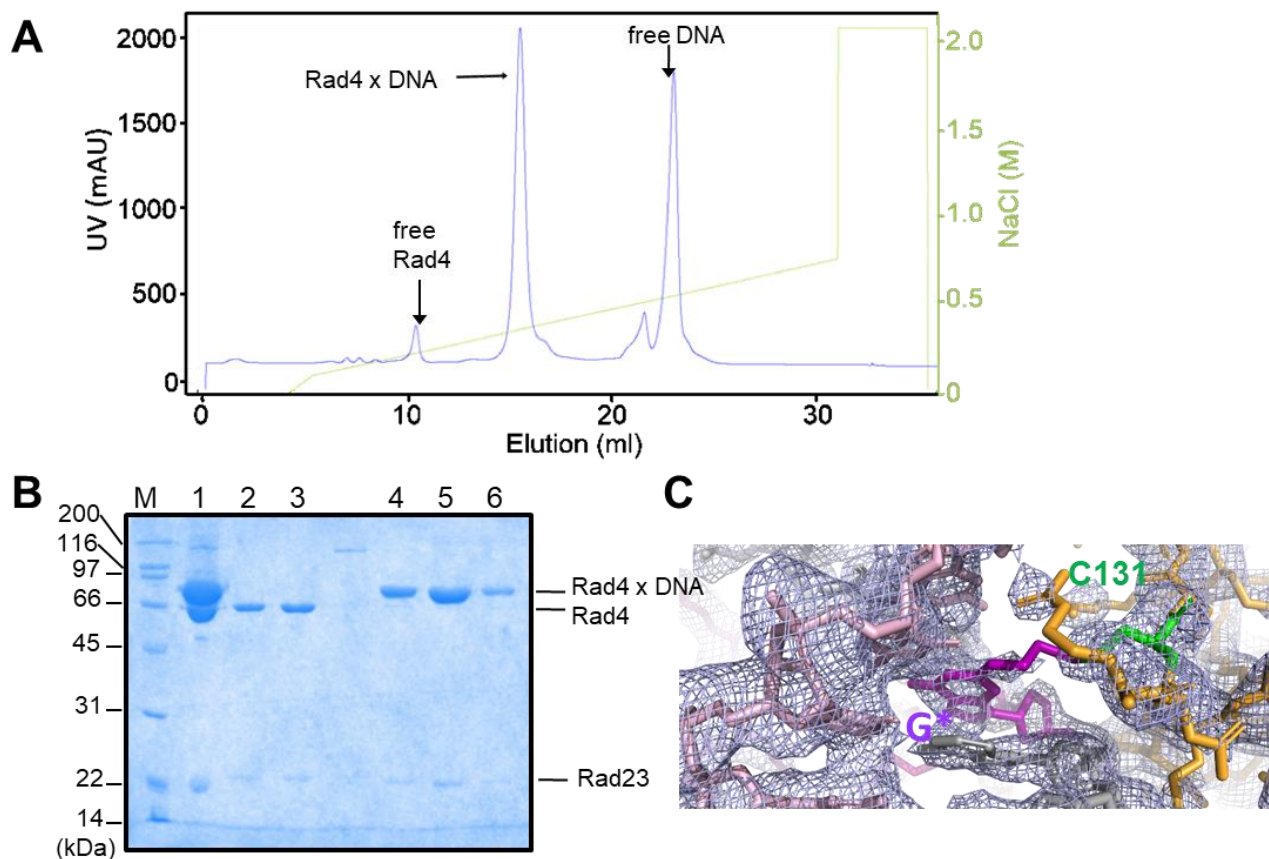
Figure S4. Purification and crystallization of Δβ–hairpin3 Rad4 mutant tethered to CGC/GCG DNA. (A) The mutant Rad4 (Δβ–hairpin3)–Rad23 crosslinked to CGC/GCG (Figure 1B) was purified over a MonoQ column (GE Healthcare) over a 0–2 M NaCl gradient. The peaks corresponding to free Rad4-Rad23 ('free Rad4'), covalently crosslinked protein-DNA complex ('Rad4 x DNA'), and free DNA are indicated with arrows. **(B)** Non-reducing SDS-PAGE gel of the eluted fractions shows that the crosslinked Rad4-Rad23-DNA complex (Lanes 4-6; eluting at 400–480 mM NaCl) was separated from the free Rad4-Rad23 (Lane 2, 3; eluting at 280–320 mM NaCl). Lane 1 shows the input, and Lane M shows the molecular weight marker. **(C)** Composite omit electron density map of the region near the crosslink at a contour level of 1.5 σ for the tethered complex.
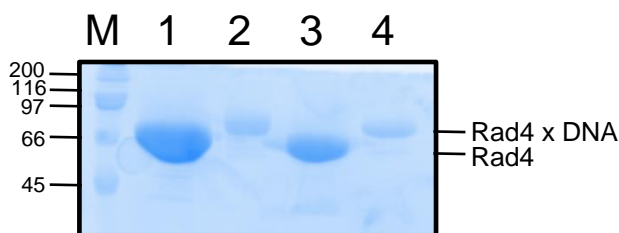

Figure S5. Crosslinked Rad4-Rad23-DNA complexes used in FLT-FRET measurements. Non-reducing SDS-PAGE gel in which lanes 1 & 2 show WT Rad4-Rad23 (SC32) without and with tethered CGC/GCG$_F$ DNA duplex, respectively; lanes 3 & 4 are analogous samples with Δβ–hairpin3 mutant (SC41b). The tethered complexes in lanes 2 and 4 ('CGC/GCG$_F$ x WT' and 'CGC/GCG$_F$ x Δβ–hairpin3', respectively) are used for the FLT-FRET experiments shown in **Figures 3** and **S6**.
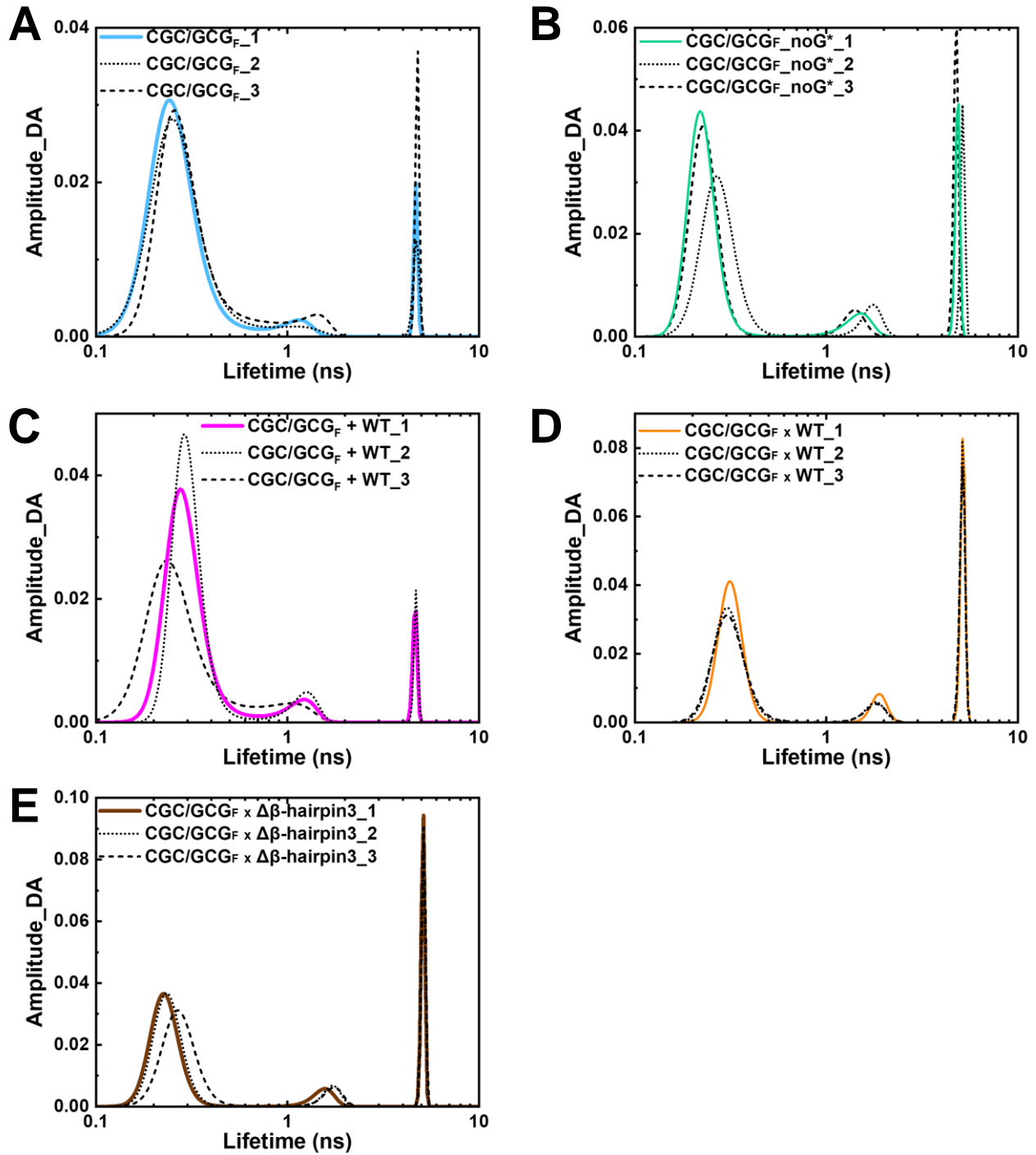
Figure S6. Fluorescence lifetime (FLT) distributions from MEM analyses. Each panel shows 3 independent measurements for a given sample, indicating the reproducibility of the FLT distributions obtained by MEM. The colored lines are those used as representative distributions in Figure 3. **(A)** Unbound CGC/GCG$_F$ (cyan). **(B)** Unbound CGC/GCG$_F$_noG* (green). CGC/GCG$_F$ noncovalently bound to WT Rad4 (**C**, magenta), covalently tethered to WT Rad4 (**D**, orange) or to Δβ−hairpin3 (**E**, brown). All amplitudes indicate the normalized, fractional amplitudes. See Table S3 for the DNA sequences, detailed reports on the lifetimes, fractional amplitudes, FRET efficiencies of each lifetime component as well as the average FRET efficiencies of each sample.
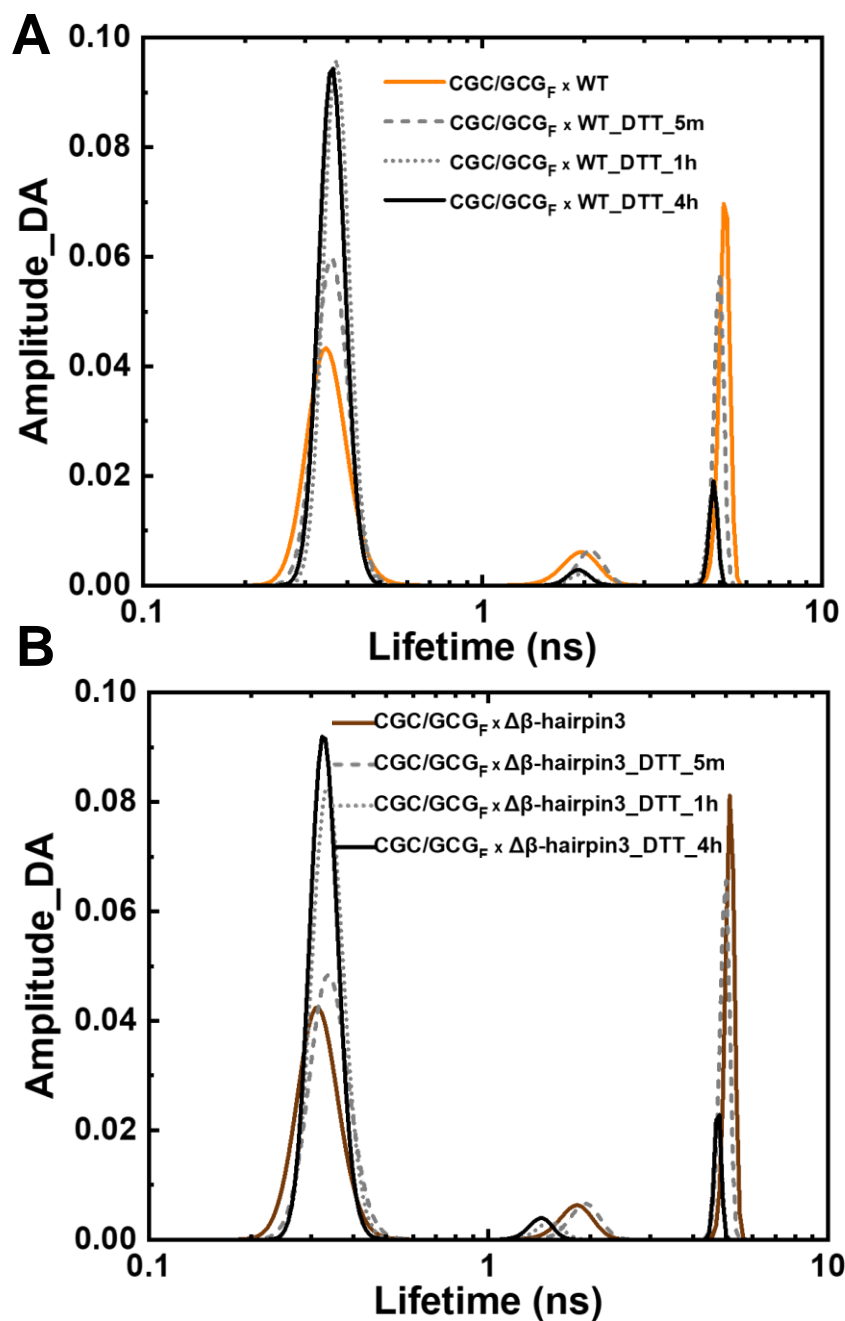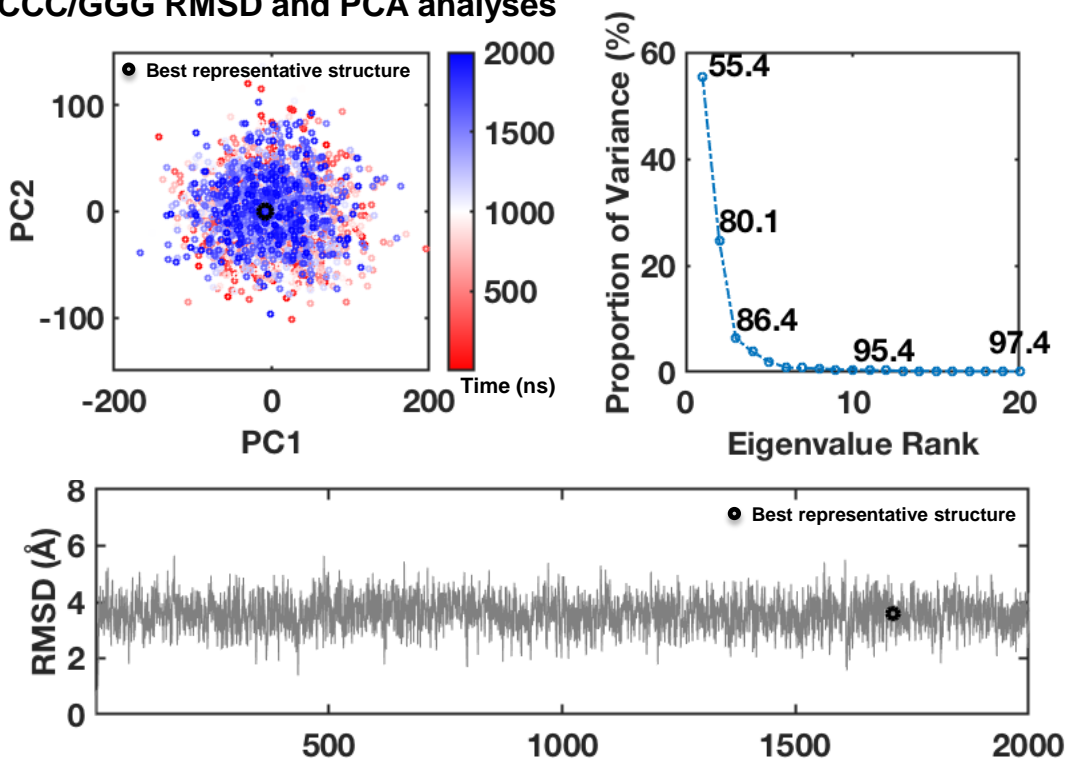
Figure S7. Dithiothreitol (DTT) treatment of the disulfide-tethered DNA x Rad4 complexes.
Samples containing CGC/GCG$_F$ tethered to WT (**A**, orange) or Δβ−hairpin3 (**B**, brown) Rad4-Rad23 complexes were treated with 5 mM DTT, and their FLT distributions were obtained at various time points: 5 min (light grey, long dotted line), 1 h (grey, short dotted line) and 4 h (black, solid line) after the treatment.
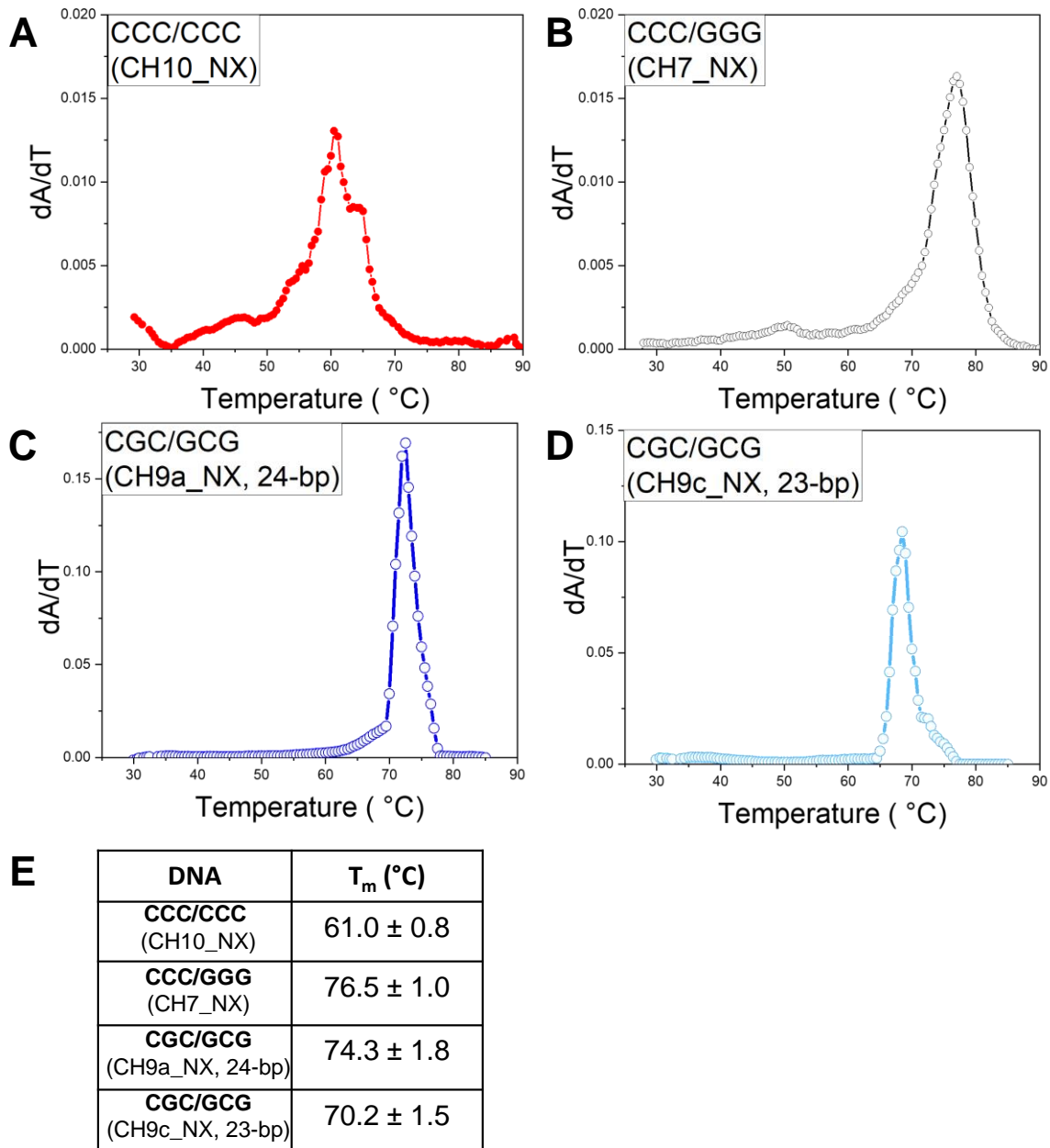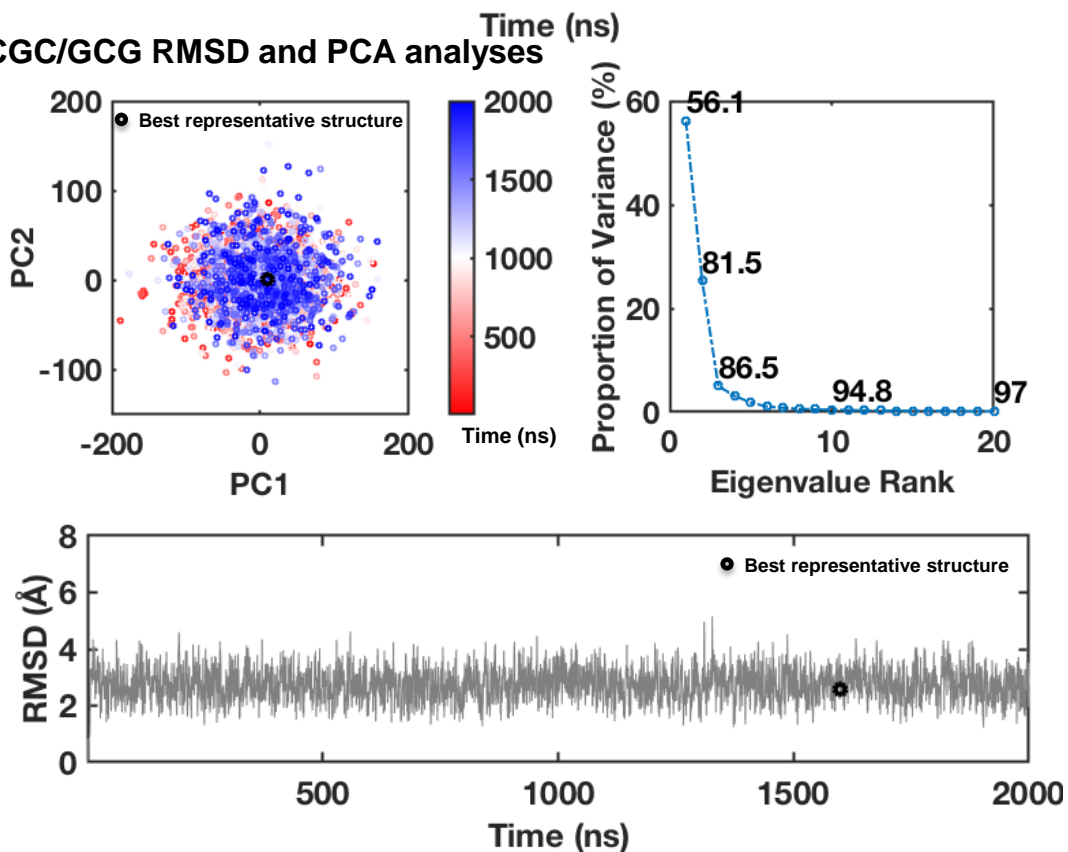
Figure S8. Melting temperature measurements of double-stranded DNA substrates. (A) Melting temperature profiles for the DNA duplexes as measured by the derivative of UV absorbance at 260 nm over temperature. See **Figure S1** for the DNA sequences. **(B)** Average and standard deviations of the T_m's measured for each DNA construct from 2-3 replicates.
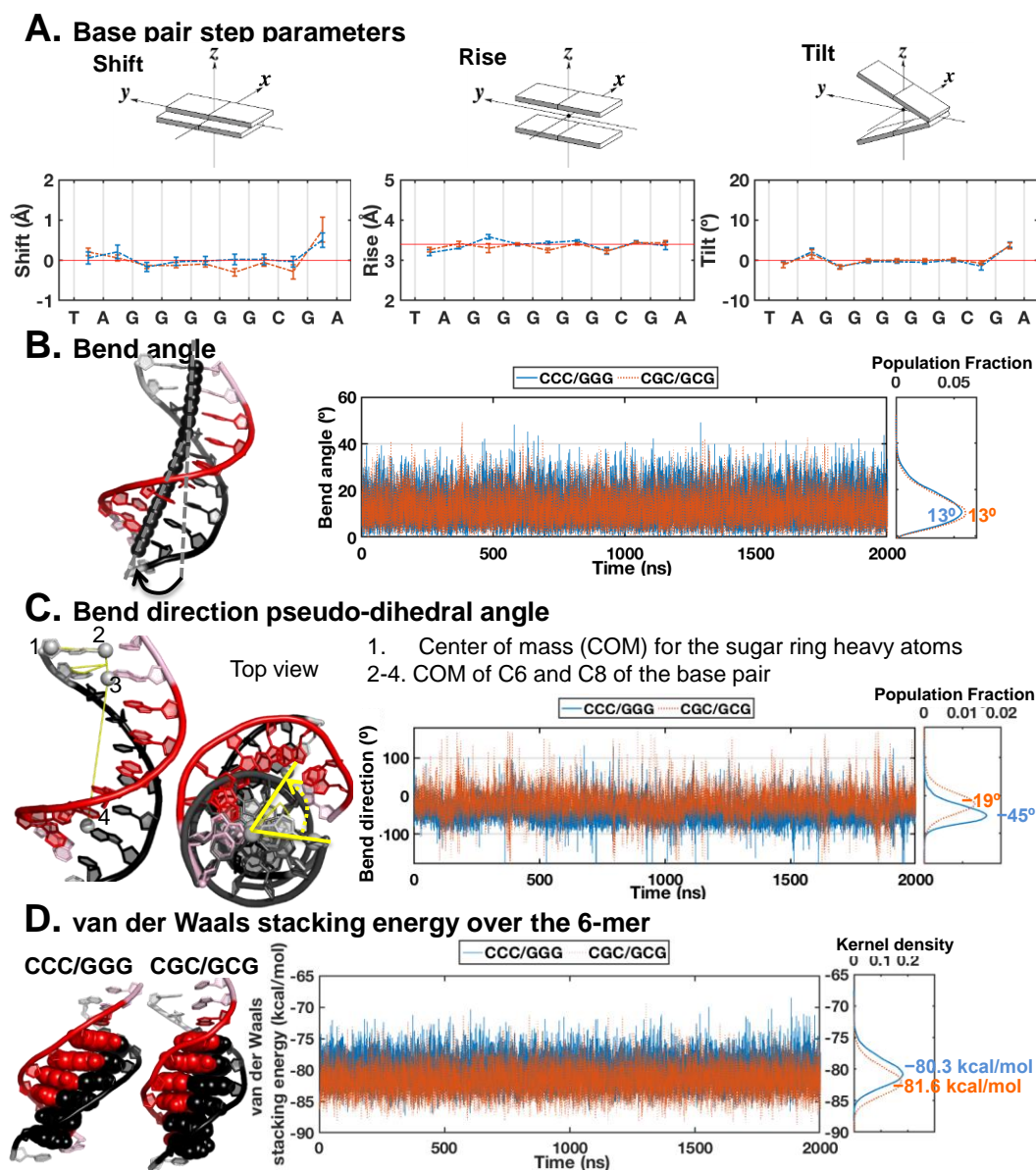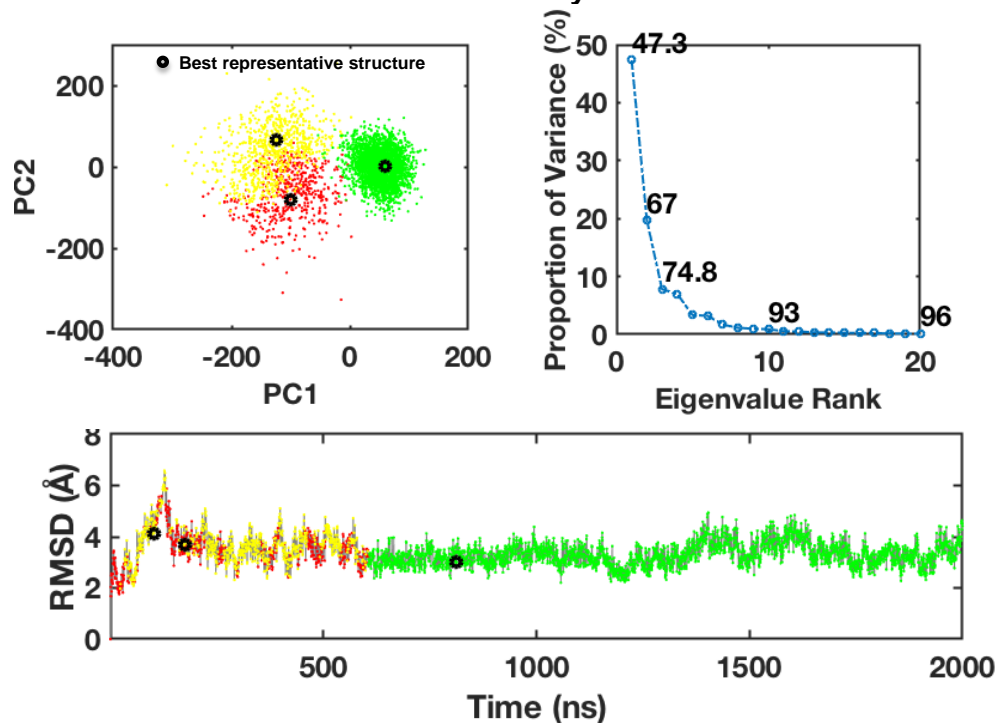
Figure S9. Principal component analyses (PCA) and RMSD for the MD simulations of the (A) CCC/GGG and (B) CGC/GCG duplexes. The 0 to 2 μs ensemble was used for further analyses, since all parameters were stable (single cluster in the PCA plot and stable RMSD values).
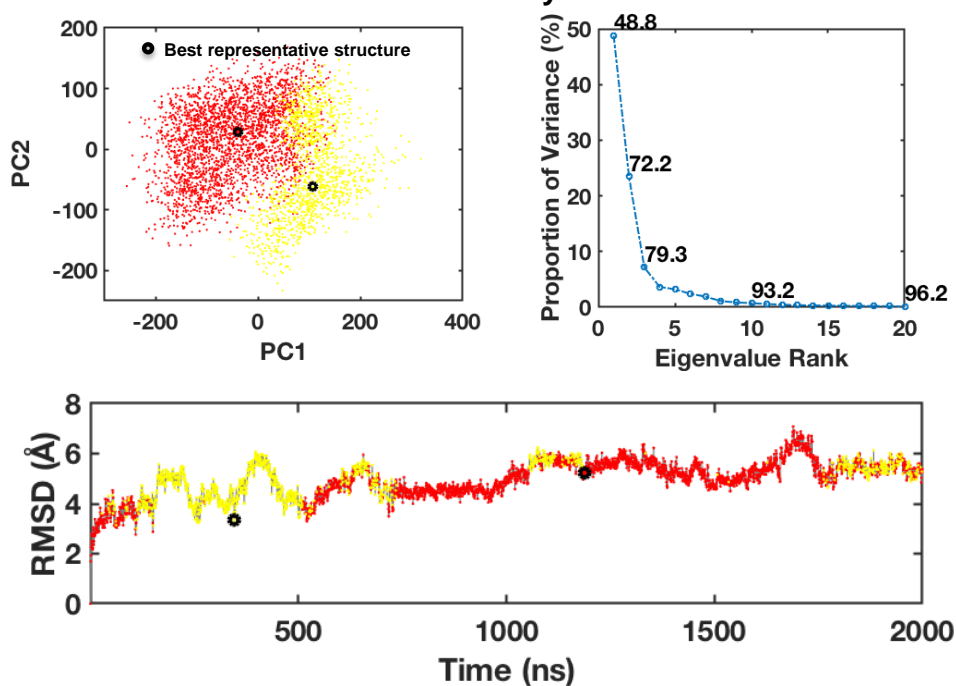
Figure S10. Structural analyses for the MD simulations of the unbound CCC/GGG and CGC/GCG duplexes. (**A**) Shift, rise and tilt base pair step parameters of CCC/GGG (blue) and CGC/GCG (orange). The block averaged means and standard deviations for the parameter values are shown. Illustrations of the base pair step parameters are adapted from 3DNA (23) (**B**) Bend angles. The best representative structure of CCC/GGG is shown on the left to illustrate the bend angle. The helix axis is shown in black spheres. (**C**) DNA bend direction pseudo-dihedral angles. The best representative structure of CCC/GGG is shown on the left to illustrate the definition of the bend direction pseudo-dihedral angle, with the centers of mass (COM) utilized in the computation shown as white spheres. The top view shows the bend direction compared to B-DNA in dark gray. (**D**) van der Waals stacking energy for the potential open site 6-mer (base pair steps 14 – 19 in **Figure 1B**). The best representative structures of CCC/GGG and CGC/GCG are shown on the left with the base pairs of the 6-mer used for the calculation of the van der Waals energy in spheres. Time-dependent values are shown for properties in panels (B), (C) and (D). Kernel densities are calculated for the raw data using the ksdensity function with 200 bins in MATLAB 7.10.0 (The MathWorks, Inc.), and are plotted on the side with mean values labeled. These kernel densities are representative of the population distributions over the range of each property (*e.g.*, bend angle).
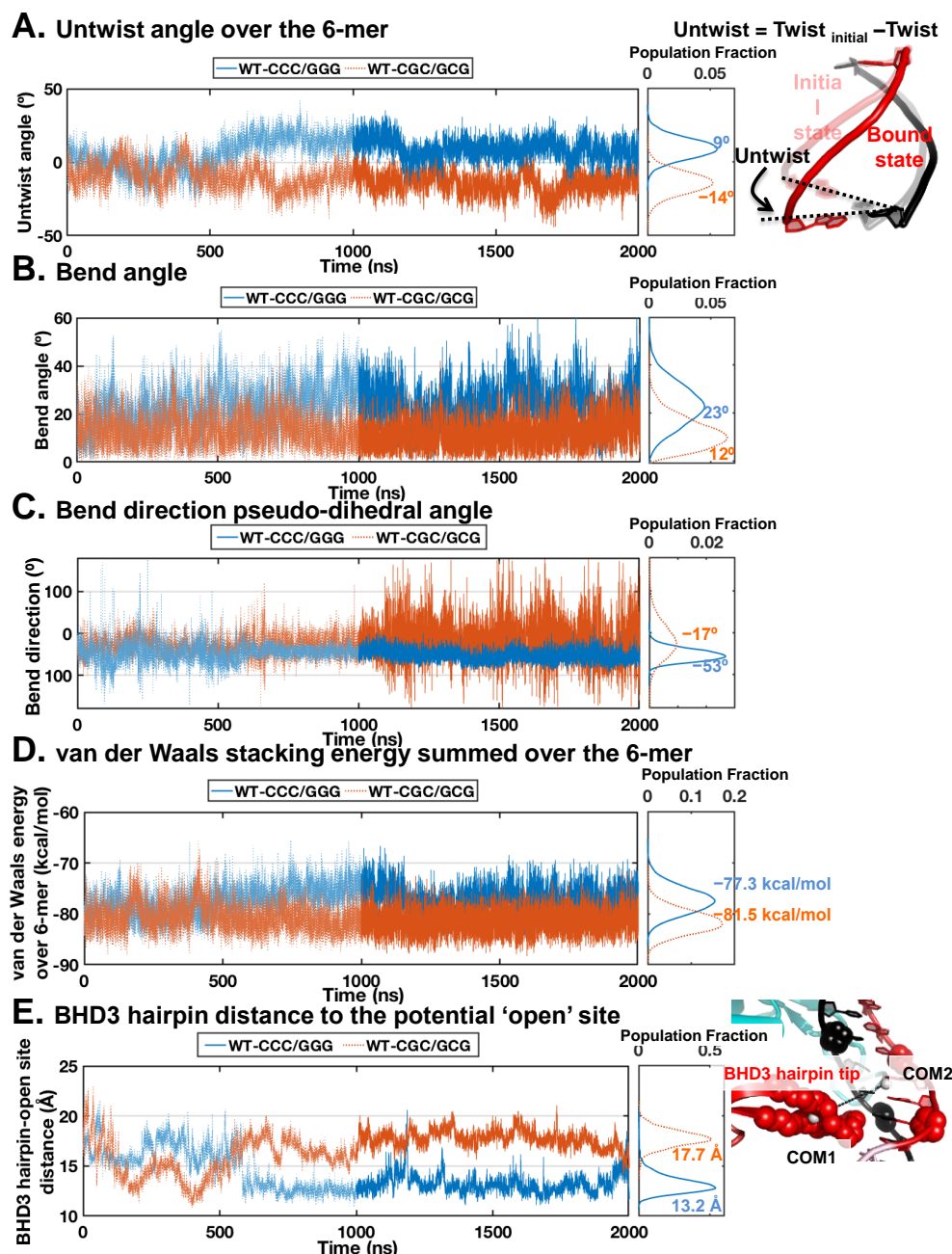
21

A. WT-Rad4-CCC/GGG RMSD and PCA analyses

B. WT-Rad4-CGC/GCG RMSD and PCA analyses

Figure S11. Principal component analyses (PCA) and RMSD for the MD simulations of the Rad4 initial binding to the (A) CCC/GGG and (B) CGC/GCG duplexes. The 1 to 2 μs ensemble was used for further analyses, since one main structure cluster is achieved in ~ 1 μs (with stable RMSD value). Different structure clusters are color-coded. The best representative structure values for each cluster are shown in black circles.

A. Untwist angle over the 6-mer

Untwist = Twist $_{initial}$ − Twist

B. Bend angle

C. Bend direction pseudo-dihedral angle

D. van der Waals stacking energy summed over the 6-mer

E. BHD3 hairpin distance to the potential 'open' site

Figure S12. Structural and energetic analyses for the MD simulations of the Rad4 initial binding to the CCC/GGG and CGC/GCG duplexes. (**A**) Untwist angles illustrated on the right, (**B**) bend angles, (**C**) DNA bend direction pseudo-dihedral angles (as in **Figure S10C**), (**D**) van der Waals stacking energy for the potential 'open' site 6-mer (base pair steps 14–19 in **Figure 1B**), and (**E**) the distance between β–hairpin3 and the center of the potential 'open' site. The distance is calculated between the center of mass (COM) of the β–hairpin3 (residues 597-607) backbone heavy atoms (COM1) and the COM of sugar ring heavy atoms of the nucleotides surrounding the open site (nucleotide steps 15 and 18 in **Figure 1B**) (COM2), illustrated on the right. Time dependent values are shown for all properties, with the values for initial binding states (1 – 2 μs) in darker shade. Kernel densities for the values of the initial binding states are calculated using the ksdensity function with 200 bins in MATLAB 7.10.0 (The MathWorks, Inc.), and are plotted on the side with mean values labeled. These kernel densities are representative of the population distributions over the range of each property (*e.g.* untwist angle).

23

SI Movies S1-S2

Movie S1. Best representative structure for CCC/GGG sequence upon its initial binding with Rad4, showing the engagement of BHD2 with the DNA minor groove.

Movie S2. Best representative structure for CGC/GCG sequence upon its initial binding with Rad4, showing the failure of BHD2 to engage with the DNA minor groove.

SI References

1.   Chen, X., Velmurugu, Y., Zheng, G., Park, B., Shim, Y., Kim, Y., Liu, L., Van Houten, B., He, C., Ansari, A. *et al.* (2015) Kinetic gating mechanism of DNA damage recognition by Rad4/XPC. *Nat. Comm.*, **6**, 5849.

2.   Velmurugu, Y., Chen, X., Slogoff Sevilla, P., Min, J.H. and Ansari, A. (2016) Twist-open mechanism of DNA damage recognition by the Rad4/XPC nucleotide excision repair complex. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, E2296-2305.

3.   Chakraborty, S., Steinbach, P.J., Paul, D., Mu, H., Broyde, S., Min, J.H. and Ansari, A. (2018) Enhanced spontaneous DNA twisting/bending fluctuations unveiled by fluorescence lifetime distributions promote mismatch recognition by the Rad4 nucleotide excision repair complex. *Nucleic Acids Res.*, **46**, 1240-1255.

4.   Paul, D., Mu, H., Zhao, H., Ouerfelli, O., Jeffrey, P.D., Broyde, S. and Min, J.H. (2019) Structure and mechanism of pyrimidine-pyrimidone (6-4) photoproduct recognition by the Rad4/XPC nucleotide excision repair complex. *Nucleic Acids Res.*, **47**, 6015-6028.

5.   Paul, D., Mu, H., Tavakoli, A., Dai, Q., Chen, X., Chakraborty, S., He, C., Ansari, A., Broyde, S. and Min, J.H. (2020) Tethering-facilitated DNA 'opening' and complementary roles of beta-hairpin motifs in the Rad4/XPC DNA damage sensor protein. *Nucleic Acids Res.*, **48**, 12348-12364.

6.   Tavakoli, A., Paul, D., Mu, H., Kuchlyan, J., Baral, S., Ansari, A., Broyde, S. and Min, J.-H. (2021) Light-induced modulation of DNA recognition by the Rad4/XPC damage sensor protein. *RSC Chemical Biology*, **2**, 523-536.

7.   Min, J.H. and Pavletich, N.P. (2007) Recognition of DNA damage by the Rad4 nucleotide excision repair protein. *Nature*, **449**, 570-575.

8.   Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. and Klein, M.L. (1983) Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.*, **79**, 926-935.

9.   Case, D.A., Ben-Shalom, I. Y., Brozell, S. R., Cerutti, D. S., Cheatham III, T. E., Cruzeiro, V. W. D., Darden, T. A., Duke, R. E., Ghoreishi, D., Gilson, M. K., Gohlke, H., Goetz, A. W., Greene, D., Harris, R., Homeyer, N., Izadi, S., Kovalenko, A., Kurtzman, T., Lee, T. S., LeGrand, S., Li, P., Lin, C., Liu, J., Luchko, T., Luo, R., Mermelstein, D. J., Merz, K. M., Miao, Y., Monard, G., Nguyen, C., Nguyen, H., Omelyan, I., Onufriev, A., Pan, F., Qi, R., Roe, D. R., Roitberg, A., Sagui, C., Schott-Verdugo, S., Shen, J., Simmerling, C. L., Smith, J., Salomon-Ferrer, R., Swails, J., Walker, R. C., Wang, J., Wei, H., Wolf, R. M., Wu, X., Xiao, L., York, D. M., Kollman, P. A. (2018). Amber18 ed. University of California, San Francisco, CA.

10.   Mu, H., Zhang, Y., Geacintov, N.E. and Broyde, S. (2018) Lesion Sensing during Initial Binding by Yeast XPC/Rad4: Toward Predicting Resistance to Nucleotide Excision Repair. *Chem. Res. Toxicol.*, **31**, 1260-1268.

11.   Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E. and Simmerling, C. (2015) ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.*, **11**, 3696-3713.

12.   Darden, T., York, D. and Pedersen, L. (1993) Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.*, **98**, 10089-10092.

13.   Loncharich, R.J., Brooks, B.R. and Pastor, R.W. (1992) Langevin Dynamics of Peptides - the Frictional Dependence of Isomerization Rates of N-Acetylalanyl-N'-Methylamide. *Biopolymers*, **32**, 523-535.

14.   Berendsen, H.J.C., Postma, J.P.M., Vangunsteren, W.F., Dinola, A. and Haak, J.R. (1984) Molecular-Dynamics with Coupling to an External Bath. *J. Chem. Phys.*, **81**, 3684-3690.

15. Ryckaert, J.P., Ciccotti, G. and Berendsen, H.J.C. (1977) Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.*, **23**, 327-341.
16. R-Core-Team. (2017), R Foundation for Statistical Computing, Vienna, Austria.
17. Skjaerven, L., Yao, X.Q., Scarabelli, G. and Grant, B.J. (2014) Integrating protein structural dynamics and evolutionary analysis with Bio3D. *BMC Bioinformatics*, **15**, 399.
18. Lavery, R., Moakher, M., Maddocks, J.H., Petkeviciute, D. and Zakrzewska, K. (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.*, **37**, 5917-5929.
19. Rooklin, D., Wang, C., Katigbak, J., Arora, P.S. and Zhang, Y.K. (2015) Alpha Space: Fragment-Centric Topographical Mapping To Target Protein-Protein Interaction Interfaces. *J Chem Inf Model*, **55**, 1585-1599.
20. Flyvbjerg, H. and Petersen, H.G. (1989) Error estimates on averages of correlated data. *The Journal of Chemical Physics*, **91**, 461-466.
21. Yang, W., Bitetti-Putzer, R. and Karplus, M. (2004) Free energy simulations: use of reverse cumulative averaging to determine the equilibrated region and the time required for convergence. *J. Chem. Phys.*, **120**, 2618-2628.
22. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605-1612.
23. Lu, X.J. and Olson, W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108-5121.