Supplementary Information: Unveiling the pathway to Z-DNA in the protein-induced B-Z transition

Sook Ho Kim[1,2], So-Hee Lim[1,3], Ae-Ree Lee[4], Do Hoon Kwon[1,5], Hyun Kyu Song[1,5], Joon-Hwa Lee[4], Minhaeng Cho[1,3], Albert Johner[6,7], Nam-Kyung Lee[6,7,*] and Seok-Cheol Hong[1,2,*]

[1] *Center for Molecular Spectroscopy and Dynamics, Institute for Basic Science, Seoul 02841, South Korea*

[2] *Department of Physics, Korea University, Seoul 02841, South Korea*

[3] *Department of Chemistry, Korea University, Seoul 02841, South Korea*

[4] *Department of Chemistry and RINS, Gyeongsang National University, Jinju 52828, South Korea*

[5] *Department of Life Sciences, Korea University, Seoul 02841, South Korea*

[6] *Department of Physics, Sejong University,Seoul 05006, South Korea. and*

[7] *Institute Charles Sadron, CNRS 23 rue du Loess 67034, Strasbourg cedex 2, France*

I. PROBABILITY OF A SPONTANEOUS B-Z TRANSITION

The following argument shows that a spontaneous B-Z transition is unlikely for the Core sequence inserted in the stable B-DNA. A ZBP can be found close to the target sequence of size $\sim$1 nm^3 with a probability of $10^{-6} - 10^{-5}$. Let us assume that the complexation kinetics is reaction controlled rather than diffusion controlled. Several factors contribute to the global reaction rate Q: the microscopic reaction rate q, the average number of ZBP in the nanometric reaction volume and the probability to be in the left-handed Z-form, e^{-E_Z/k_BT}, where E_Z includes the domain wall energy $2E_J$ and the B-Z conversion energy per ZBP binding, $6\epsilon_Z$, which is typically $\sim$ 3 - 6 k_BT. We estimate the global reaction rate Q to be $Q \sim 10^{-15}q$. The microscopic reaction rate q cannot exceed the inverse self diffusion time of ZBP of the order of $10^{11}\mathrm{s}^{-1}$, $q \lesssim 10^{11}\mathrm{s}^{-1}$. This leads to the upper bound $Q \lesssim 10^{-4}$ s^{-1}. For a diffusion controlled aggregation, the upper limit for the global rate $Q \sim 10^{-4}\mathrm{s}^{-1}$ is qualitatively recovered. Z-DNA is partially stabilized by hydrogen bonds, so the real barrier against structural fluctuations that may lead to Z-form is likely higher and the rate Q accordingly even lower.

II. PARTITION FUNCTIONS

The grand partition function $\mathcal{Z}_G$ consists of Z_n^B and Z_n^Z, where the superscripts B and Z stand for B-form and Z-form with various number (subscript n) of bound hZα_{ADAR1}. (In the model with 6 binding sites, we take into account n =0, 1, 2, 3, 4, 5, and 6.) Considering the large domain wall energy ($2E_J$), we discard the possibility of Z-form with multiple domains.

$$
\begin{aligned}
Z_n^B &= {}_6C_n e^{nA_B} \qquad \text{(S-1)}\\
Z_n^Z &= \sum_{m=1}^{n} Z_{nm} \quad \text{where} \quad Z_{nm} = \sum c_{nm(x)} e^{\alpha A'_Z + \beta A_Z + \delta A_B} \quad (n = \alpha+\beta+\delta, \quad m = \alpha+\beta, \quad \text{and} \quad x = a, b)
\end{aligned}
$$

$$
\begin{aligned}
Z_0^Z &= 0 \qquad \text{(S-2)}\\
Z_1^Z &= c_{11} e^{A'_Z} = {}_6C_1 e^{A'_Z}\\
Z_2^Z &= c_{21} e^{A'_Z + A_B} + c_{22a} e^{2A'_Z} + c_{22b} e^{A'_Z + A_Z}\\
Z_3^Z &= c_{31} e^{A'_Z + 2A_B} + c_{32a} e^{2A'_Z + A_B} + c_{32b} e^{A'_Z + A_Z + A_B} + c_{33a} e^{3A'_Z} + c_{33b} e^{2A'_Z + A_Z}\\
Z_4^Z &= c_{41} e^{A'_Z + 3A_B} + c_{42a} e^{2A'_Z + 2A_B} + c_{42b} e^{A'_Z + A_Z + 2A_B} + c_{43} e^{2A'_Z + A_Z + A_B} + c_{44a} e^{3A'_Z + A_Z} + c_{44b} e^{2A'_Z + 2A_Z}\\
Z_5^Z &= c_{51} e^{A'_Z + 4A_B} + c_{52} e^{A'_Z + A_Z + 3A_B} + c_{53} e^{2A'_Z + A_Z + 2A_B} + c_{54} e^{2A'_Z + 2A_Z + A_B} + c_{55} e^{3A'_Z + 2A_Z}\\
Z_6^Z &= c_{62} e^{A'_Z + A_Z + 4A_B} + c_{64} e^{2A'_Z + 2A_Z + 2A_B} + c_{66} e^{3A'_Z + 3A_Z}
\end{aligned}
$$

where $c_{nm(x)}$ are constants associated with the entropic contribution of the adsorbed proteins. The first numerical index n indicates the number of bound proteins, the second numerical index m indicates the number of proteins bound to Z-form, and the last (auxiliary), alphabetic index $x = a$ and b is ordered with an increasing number of bound proteins on an existing Z-form. For the case of the 6-binding-site model, $c_{11} = 6, c_{21} = 24, c_{22a} = 8, c_{22b} = 3$,

$c_{31} = 36, c_{32a} = 16, c_{32b} = 12, c_{33a} = 8, c_{33b} = 8, c_{41} = 24, c_{42a} = 8, c_{42b} = 18, c_{43} = 16, c_{44a} = 12, c_{44b} = 2,$ $c_{51} = 6, c_{52} = 12, c_{53} = 8, c_{54} = 4, c_{55} = 6,$ and $c_{62} = 3, c_{64} = 2, c_{66} = 1.$

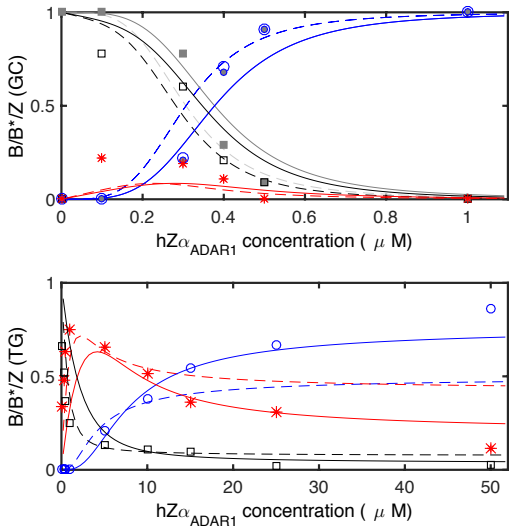

Fig. S 1: 4-binding-site model. Populations of conformational states of DNA (with B* neglected, Z: blue solid circle, B: black solid square; with B* into account, Z: blue open circle, B: black open square, B*: red asterisk) for GC (top) and TG (bottom) repeat Core sequences in the presence of indicated concentrations of hZ$\alpha_{\rm ADAR1}$. Here fits are based on the 4-binding-site model and A_B and A_Z were adjusted for fitting. For GC repeats, A_B= 8.6 k_BT and A_Z = 18.2 k_BT with $a \approx 2.8$ nm (solid line) or 3.0 nm (dashed line). For TG repeats, A_Z = 17.2 k_BT and A_B = 8.7 k_BT with a = 3.8 nm (dashed) or A_B = 8.6 k_BT with a = 3.0 nm (solid). The B* state amounts to 100% and 85% of hZ$\alpha_{\rm ADAR1}$-adsorbed B-DNA state for GC and TG repeats, respectively.

III. 4-BINDING-SITE MODEL

We also considered the case that the number of binding sites available are reduced possibly due to the steric hindrance by dye molecules. The fit based on the 4-binding-site model is shown in Fig. S1. The 4-binding-site model predicts somewhat stronger affinity of hZ$\alpha_{\rm ADAR1}$, A_Z = 18.3 and 17.2 for GC and TG repeats, respectively with A_B = 8.6 and a = 2.8 nm. The difference in Z-DNA affinity is $\sim$ 1.1. However, the fit does not produce the rapid increase of Z-population around the midpoint concentration of hZ$\alpha_{\rm ADAR1}$ for GC repeats. Likewise, the fit by the same model does not produce the experimentally observed increase in Z-population at large protein concentrations for TG (See Fig. S1). The fit works only for the range of hZ$\alpha_{\rm ADAR1}$ from $\sim$ 5 to $\sim$ 25 μM for TG repeats. In particular the best estimated value of the populations of B* (B) state below [hZ$\alpha_{\rm ADAR1}$]$\sim$ 5 μM range is considerably smaller (larger) than the experimental data. Thus, we ruled out this model and conclude that direct hindrance by the dye seems irrelevant.

IV. 6-BINDING-SITE MODEL

A. Fit without the pre-complexation-induced B* state for GC

The populations of B-DNA and Z-DNA obtained by integrating the FRET histogram are fit using the 6-binding-site model assuming that there are B-, B*-, and Z-DNA states for TG repeat sequences and B- and Z-DNA states for GC repeat sequences (Fig. S 2).

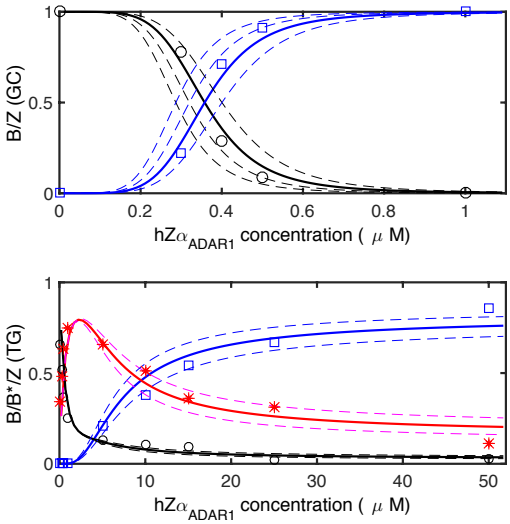

Fig. S 2: 6-binding-site model. Populations of conformational states of DNA (Z: blue square, B: black circle, B*: red asterisk) for GC (top) and TG (bottom) repeat Core sequences in the presence of indicated concentrations of hZ$\alpha_{\rm ADAR1}$, which are fit with various values of A_B and A_Z. Here fits are based on the 6-binding-site model with $a \approx 3.8$ nm, A_B= 9.2 $k_{\rm B}T$, and A_Z = 16.0, 16.1, 16.2, and 16.3 $k_{\rm B}T$ (GC) and 15.7, 15.75, and 15.8 $k_{\rm B}T$ (TG). Assuming the same B-DNA affinity for TG and GC repeats, the best fits are obtained with A_Z = 16.1 $k_{\rm B}T$ (GC) and A_Z = 15.75 $k_{\rm B}T$ (TG) (shown in thick lines). The B* state amounts to 85% of hZ$\alpha_{\rm ADAR1}$-adsorbed B-DNA state. The curves for B, B*, and Z states are shown in black, red and blue, respectively. Considering the presence of B* in GC repeats, we obtain similar affinities A_Z = 16.2 and A_B=9.2 with a = 3.7 nm and $\gamma = 0.85$. See Fig. 3 in the main text.

B. Influence of Boundary Conditions

For the case of flanking B-DNA handles that maintain the free boundary condition with GC or TG repeats, proteins prefer to adsorb on four boundary sites (two from each handle-repeat interface) to nucleate protein binding from each interface. In this case, the number of domain walls costing the junction energy E_J would be $n_J = 1$. For the case of flanking B-DNA handles that maintain the fixed boundary condition with GC or TG repeats, there is no advantage of starting from either interface and nucleation of protein binding can start anywhere on the repeat. Thus, all six potential sites are energetically equivalent and the number of domain walls associated with the junction energy E_J is $n_J = 2$.

It can be also argued that if DNA stacking is partly broken (e.g. due to insertion of dye molecule or unwinding torsion), a smaller junction energy would apply. For free boundary conditions imposed by torsionally flexible handles, a partial transition in the DNA tether involves only one junction. Even if the junction energy is assumed to be halved, according to our analysis, the affinity to Z-DNA does not appear to change by more than 12 % whilst the affinity to B-DNA stays unchanged and so does the difference in Z-DNA affinity between TG and GC repeats. We hence conclude that the effect of the dye molecules on the B/Z interface is negligible.

In Table S 1, the best fit values for affinities are obtained for various junction energy conditions as αE_J with $\alpha = 1 \sim 2$. It turns out that the best fit affinities vary by ~ 10 % upon changing the boundary condition from $\alpha = 2$ to 1.

Table S 1: Junction-energy dependence of hZ$\alpha_{\rm ADAR1}$ affinity to Z-DNA: A_Z is obtained from fits to the model with 6 binding sites on DNA and the fixed boundary condition with varying energy penalties. Here, A_B = 9.2 $k_{\rm B}T$, n_J =2, E_J = 9 $k_{\rm B}T$, and a = 3.8 nm are used.

Domain wall energy	A_Z(GC)	A_Z(TG)
$2E_J$ (fixed B. C.)	16.1	15.8
$1.8E_J$ (90%)	15.8	15.4
$1.6E_J$ (80%)	15.5	15.2
E_J (50%)	14.6	14.3

V. 10-BINDING-SITE MODEL

The populations of B-DNA and Z-DNA obtained by integrating the FRET histogram are fit using the 10-binding-site model (Fig. S 3). It is assumed that the range of DNA flipping by a protein is as large as 5 bps, equivalent to the size of binding site. Accordingly, about 10 proteins can adsorb on Core DNA and over the B-Z junction. The grand partition function $\mathcal{Z}_\mathcal{G}$ now includes combinatorics of Zα adsorption sites with $n = 1$ to 10. We obtained $A_Z = 13.4$ k_BT (GC) and 13.1 k_BT (TG) keeping the same B-DNA affinity as A_B= 8.4 k_BT. Here the best fits are obtained with $a \approx 5.7$ nm. Similar to the 6-binding-site model, the B-, B*-, and Z-DNA states for GC and TG repeat sequences are fit to produce the affinities of hZ$\alpha_{\rm ADAR1}$. The difference of affinities to GC and TG repeat sequences for $n_0 = 5$ is ~ 0.3 k_BT. We conclude that protein affinity difference is not sequence sensitive as long as they are alternating purine-pyrimidine sequences.

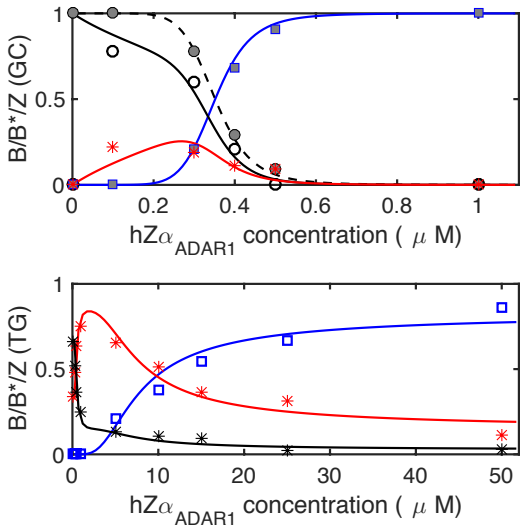

Fig. S 3: 10-binding-site model. Populations of conformational states of DNA (Z: blue square, B: black circle, B*: red asterisk) for GC (top) and TG (bottom) repeat Core sequences in the presence of indicated concentrations of hZ$\alpha_{\rm ADAR1}$, which are fit with various values of A_B and A_Z. Here fits are based on the 10-binding-site model with $a \approx 5.7$ nm. Assuming the same B-DNA affinity A_B= 8.4 k_BT, the best fits are obtained with $A_Z = 13.4$ k_BT (GC) and 13.1 k_BT (TG). The curves for B, B*, and Z states are shown in black, red, and blue, respectively. The filled symbols (B: gray filled circle, Z: gray filled square) in the upper panel (GC) are obtained by hypothesizing no B*-state.

VI. KINETICS

We also acquired kinetic FRET time traces from DNA in the presence of hZα_{ADAR1}. Time trace data exhibited slow kinetics and photobleaching of fluorescent dyes, which make kinetic analysis based on time traces difficult and unreliable. From exemplary and representative times traces, we can still identify the three states discussed in this paper: B-, B*-, and Z-DNA.

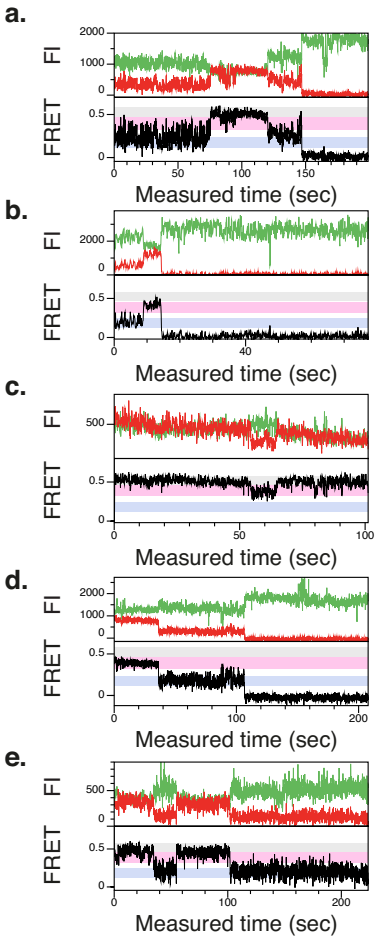

Fig. S 4: Time traces of fluorescence intensities (FI) of donor (green) and acceptor (red) (upper) and FRET efficiency (lower) from $(TG)_{11}$ molecules that exhibit FRET changes in the presence of (a-b) 50 μM and (c-e) 15 μM hZα_{ADAR1}. Even at 50 μM hZα_{ADAR1}, the backward Z-to-B transition was observed. B-, B*-, and Z-DNA states are indicated by gray, pink, and purple bands in FRET time traces, respectively.

To ensure we reach thermal equilibrium, we wait a long time (longer than 0.5 - 2 hr). In Fig. S 5 and S 6, we show that measurements with various incubation times yield the same results as long as the incubation time is longer than 0.5 hr.

For each data point, we collected data from more than 5000 molecules (more than 50 different fields of view in sample and more than 150 molecules in each region). We pooled all the data together to plot histograms. Even if we reduce the data size by choosing subsets of the whole data (seemingly more homogeneous by sharing some of (unimportant) experimental conditions (physical proximity, measurement time, etc)), we still have the same result and conclusion from different sets of data taken at different periods.

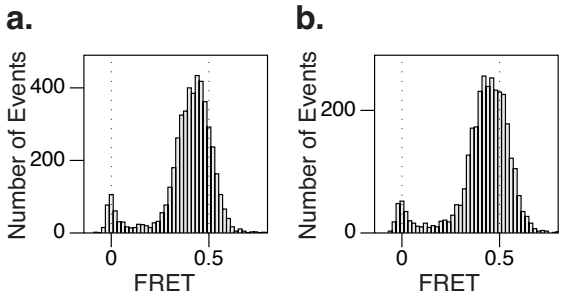

Fig. S 5: FRET histograms from $(TG)_{11}$ in the presence of $[hZ\alpha_{ADAR1}] = 0.5$ μM after incubation for different times: (a) 1 hr and (b) 2 hr. Data with different incubation times are nearly identical.

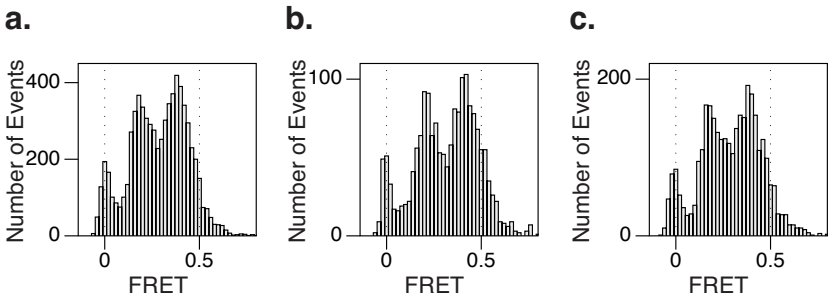

Fig. S 6: FRET histogram from $(TG)_{11}$ in the presence of $[hZ\alpha_{ADAR1}] = 15$ μM constructed from different (sub)sets of data. (a) entire set (file index from 1 - 90) , (b) subset from 1 to 21, (c) subset from 60 to 90. The subset in (c) was collected at least 0.5 hr after the subset in (b). Data from different subsets (different time lapses from the onset of incubation) are nearly identical.

VII. SALT INDUCED B-Z TRANSITION

It is possible to trigger the B-Z transition for $(GC)_{11}$ repeats by increasing salt concentration as reported in earlier studies. The question is whether the B* structure can be formed by salt-induced B-Z transitions. We examined smFRET of TG-repeat DNA by varying salt concentrations. At high salt conditions, electrostatic interactions are screened and hydrophobic residues can rearrange, overcoming weakened electrostatics among charged residues. Naturally these interactions are not specific but rather generic. Although some deformations (leading to a similar FRET value) followed, we expect the structure of deformed DNA to be different from that of the pre-complex with $hZ\alpha_{ADAR1}$. Some exemplary smFRET histogram from TG repeats under rather extreme salt conditions are shown in Fig. S 7. There is no indication of B*.

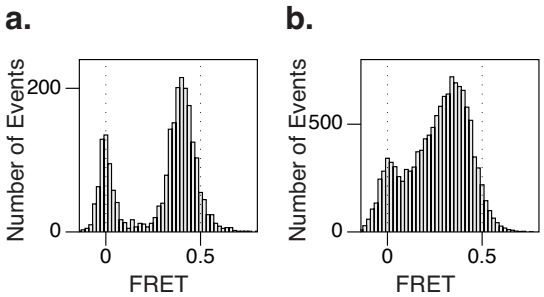

Fig. S 7: FRET histograms for TG repeats (a) 4 M NaCl and (b) 6 M $NaClO_4$.

We reexamined the salt-induced B-Z transition in $(GC)_{11}$ repeats reported in Ref. [10] to check the existence of the B* structure (Fig. 2 therein) as detected in the protein induced B-Z transitions. We also acquired a new set of corresponding FRET efficiency histograms as shown in Fig. S 8. Coincidentally, there are a small number of molecules, the FRET efficiency of which is in the same range (FE $\sim 0.30 - 0.36$) as that of B*-DNA. This was also

noted in Ref. [40]. They are, however, unlikely to be related to (salt-induced) Z-DNA because the third 'peak' already exists under the low salt conditions ($Na^+ \sim 50$ or 100 mM) that disfavor Z-DNA, and the population of the 'peak' remains low and unchanged even in Z-DNA inducing salt conditions.

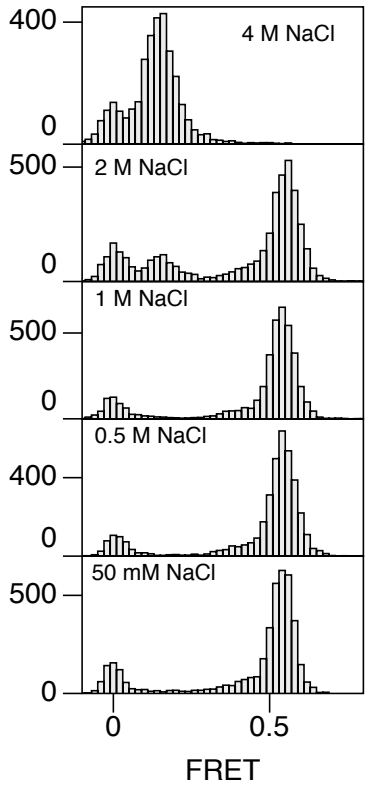

Fig. S 8: FRET histograms for GC repeats obtained at various salt conditions. A small number of molecules with an intermediary FRET efficiency exist in the range of Na^+ from 50 mM to 2 M.

A clear evidence for the entity of the third 'peak' comes from kinetic measurements that follow the B-Z transition in real time under various salt conditions (Fig. S 9). When we examined the FRET time traces that exhibit the interconversion (from B to Z or vice versa), which are rather infrequent (most of FRET traces taken over a few minutes exhibit a single FRET value either B or Z, the relative population of which depends on their thermodynamics), the FRET time traces with a transition event provide no evidence for the existence of an intermediate state on the path of the B-to-Z transition. In other words, one (FRET) state changes to the other (FRET) state directly without going through a third state. Therefore, the third state is likely to come from a misfolded state, as noted in Ref. [40].

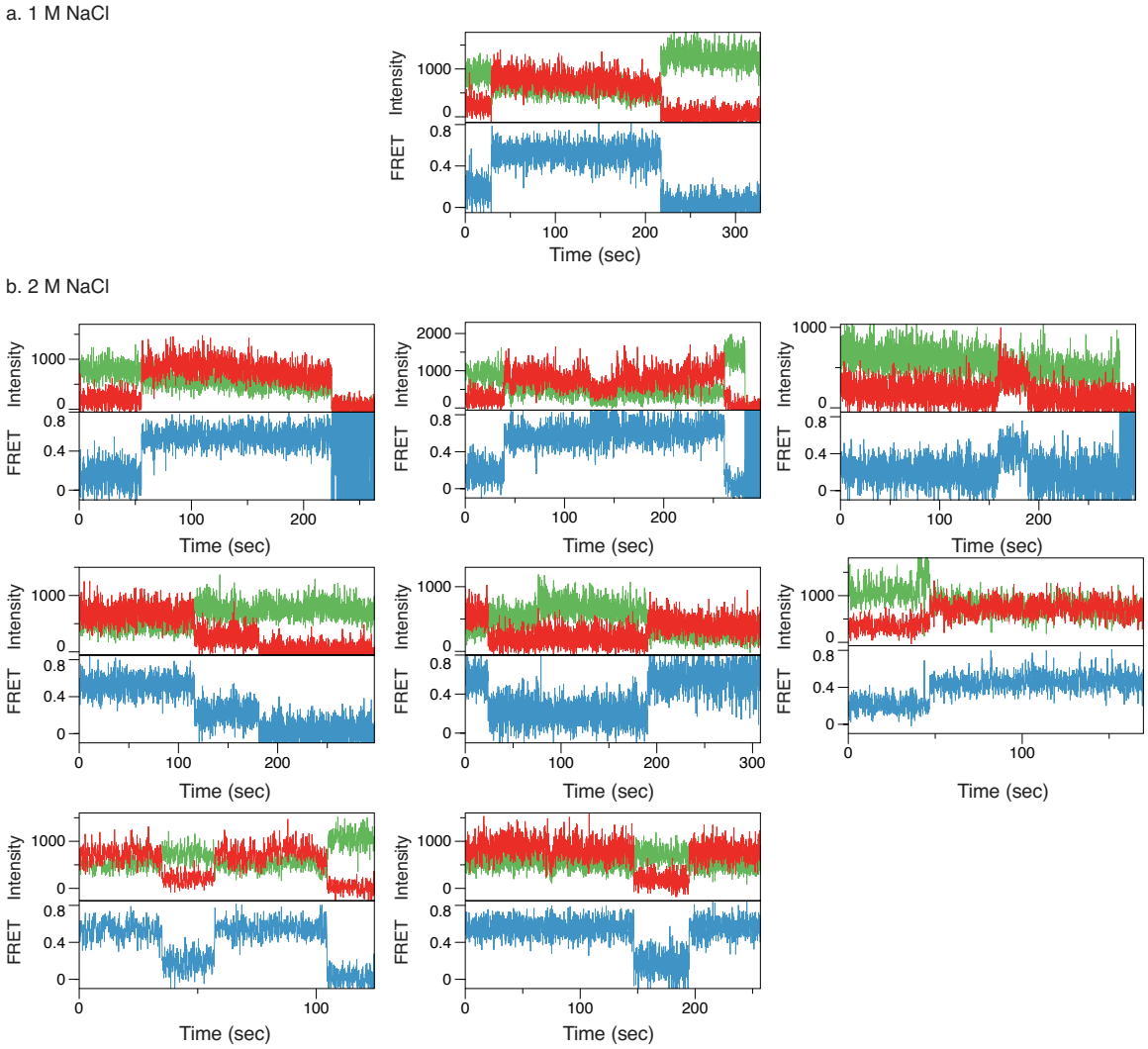

Fig. S 9: Representative FRET time traces for GC repeats that exhibit interconversions between B- and Z-DNA states. (a) 1 M NaCl, (b) 2 M NaCl. Time traces that exhibit the B-Z transitions are as rare as 1 or 2 % of the whole time traces acquired.