Title
Revealing Atomic-scale Molecular Diffusion of a Plant Transcription Factor WRKY domain protein along DNA

Authors
Dai, Liqiang
Xu, Yongping
Du, Zhenwei
et al.

Publication Date
2020

DOI
10.1101/2020.02.14.950295

PNAS
www.pnas.org

Main Manuscript for

Revealing Atomic-scale Molecular Diffusion of a Plant Transcription Factor WRKY domain protein along DNA

Liqiang Dai[1, 2†], Yongping Xu[3†], Zhenwei Du[3], Xiao-dong Su[3]*, and Jin Yu[4]*

[1] Shenzhen JL computational science and applied research institute, Shenzhen, Guangdong 518129, China
[2] Beijing Computational Science Research Center, Beijing, 100193, China
[3] State Key Laboratory of Protein and Plant Gene Research, and Biomedical Pioneering Innovation Center (BIOPIC), School of Life Sciences, Peking University, Beijing 100871, China
[4] Department of Physics and Astronomy, Department of Chemistry, NSF-Simons Center for Multiscale Cell Fate Research, University of California, Irvine, CA 92697, USA

† These authors are of equal contributions
* Corresponding authors: Jin Yu, Xiao-dong Su
Email: jin.yu@uci.edu; xdsu@pku.edu.cn

ORCID: 0000-0001-8224-1374

Classification
Biophysics and Computational Biology
Keywords
facilitated diffusion, molecular dynamics, specific recognition, WRKY, transcription factor
Author Contributions
Jin Yu and Xiao-dong Su designed the study. Liqiang Dai conducted all simulations, analyzed data. Jin Yu supervised the simulation conduction and data analyses. Yongping Xu solved the crystal structure of WRKY-DNA complex and conducted ITC experiment. Zhenwei Du conducted the single-molecule fluorescence experiments of WRKY diffusing on DNA. Jin Yu and Liqiang Dai wrote and edited the manuscript.

This PDF file includes:

Abstract

Transcription factor (TF) target search on genome is highly essential for gene expression and regulation. High-resolution determination of TF diffusion along DNA remains technically challenging. Here we constructed a TF model system using the plant WRKY domain protein in complex with DNA from crystallography and demonstrated microsecond diffusion dynamics of WRKY on DNA employing all-atom molecular dynamics (MD) simulations. Notably, we found that WRKY preferentially binds to one strand of DNA with significant energetic bias comparing to the other or non-preferred strand. The preferential DNA strand binding becomes most prominent in the static process, from non-specific to specific DNA binding, but less distinct during diffusive movements of the domain protein on the DNA. Remarkably, without employing acceleration forces or bias, we captured a complete one-base pair stepping cycle of the protein tracking along major groove of DNA with homogenous poly-A sequence, as individual hydrogen bonds break and reform at the protein-DNA binding interface. Further DNA groove tracking of the protein forward or backward, with occasional sliding as well as strand crossing to the minor groove of DNA, have also been captured. The processive diffusion of WRKY along DNA has been further sampled via coarse-grained MD simulations. The study thus provides unprecedented structural dynamics details on a small TF domain protein diffusion, suggests how it approaches to specific recognition site on DNA, and supports further high-precision experimental detection. The stochastic movements revealed in the TF diffusion also provide general clues on how other protein walkers step and slide along DNA.

Significance Statement

In transcription factors search for target genes, 1D diffusion of the protein along DNA is essential. Experimentally, it remains challenging to resolve individual diffusional steps of protein on DNA. Here, we report mainly all-atom equilibrium simulations of a WRKY domain protein in association and diffusion along DNA. We unprecedently demonstrate a complete stepping cycle of the protein for one base pair on DNA within microseconds, along with stochastic motions. Processive protein diffusions on DNA have been further sampled at a coarse-grained level. We have also found preferential DNA strand association of the domain protein, which becomes most prominent toward specific DNA binding, and can be common for small domain proteins to balance movements on the DNA with the sequence recognition.

Introduction

The search and recognition processes of Transcription factors (TFs) on DNA are of fundamental importance in gene expression and regulation. To locate sufficiently fast a target gene site on genome that is wrapped within three dimensional space, the TFs may proceed with a facilitated diffusion process, alternating between one dimensional (1-D) movements along DNA and three dimensional (3-D) intra-cellular diffusion, accompanied by occasional jumping, hopping, and inter-segment transfer (1-6). Experimental detection on protein searching motions or 1-D diffusion along DNA have provided evidence on the facilitated diffusion (7-12). Nevertheless, as TF protein movements of base pair (bp) distances on DNA can take place as fast at microseconds, tracking the 1-D TF diffusion at such a high temporal and spatial resolution remains technically challenging (13-17).

On the other hand, high resolution determinations of protein-DNA complex structures (18) allow one to investigate corresponding conformational dynamics by employing all-atom molecular dynamics (MD) simulations, via high-performance computing (19-21). The protein recognition on specific DNA has been actively examined in recent years using the MD technologies (22-26). In comparison, the protein association with non-specific DNA has been less examined. It is commonly expected that nonspecific association and movements of protein on the DNA happen slowly for the timescale of the simulations and cannot be well sampled via the atomistic MD. Indeed, either comparatively short MD simulations (nano or sub-microseconds) were conducted (22), or external forces were added to accelerate the protein movements or enhance samplings, such as employing targeted MD or umbrella sampling simulations (24, 27-29). In case that comparatively long or extensive MD simulations have been conducted, one recent study concentrates on association processes of a chromatin protein with DNA (30), but not yet the protein movements. For exemplary all-atom simulation studies on the protein movements along DNA, however, the proteins of concerns have been motor proteins such as RNA polymerases (31, 32), or the single-stranded DNA-binding protein (33), and DNA repair proteins (34, 35) . In this work, we focus on a model TF and present mainly unbiased all-atom microseconds equilibrium simulations of the diffusion dynamics of the TF protein along the double stranded (ds) DNA, with simulation samplings accumulated over 100 microseconds. The protein factor under our current investigation is a WRKY domain protein from *Arabidopsis thaliana* WRKY1.

WRKY proteins are a large family of transcription factors (TFs) in plants playing a broad range of important functions for signal response, stress control, and disease resistance (36, 37). The number of WRKY family members in *Arabidopsis* reaches over 70, and all of them include a DNA binding domain about 60 amino acids that is called the WRKY domain. The WRKY domain proteins are featured by a highly conserved ‘WRKYGQK’ region and a zinc finger motif, both of which turn out to be indispensable for maintaining the DNA binding function. Previously, an *apo* C-terminal domain structure of *Arabidopsis* WRKY1 had been made available (38).

Recently, a high-resolution crystal structure of the N-terminal WRKY domain protein in complex with a specific DNA binding sequence is obtained (39). Based on this structure, we performed atomistic MD simulations on the protein-DNA complexes (with a 34-bp dsDNA) in explicit solvent conditions, constructed for both specific and non-specific DNA binding systems. We identified strong or biased association of WRKY with one strand of DNA (the preferred strand; referred as the Crick strand in the crystal structure (39)), comparatively weak association with the other strand (the non-preferred strand), and such preferential strand association demonstrates most prominently into the static and specific protein-DNA binding. Notably, our simulations revealed 1-bp (i.e., base pair) cyclic stepping motions of the domain protein with a full set of hydrogen bonds (HBs) breaking and reforming at the protein-DNA backbone interface, spontaneously, as the protein tracks along the DNA groove and frequently adjusts its orientations to align with the helical groove. Moreover, the simulations also captured events of protein sliding stochastically on the DNA, e.g., attempts at larger step size (> 1 bp), directional reversal, and moving across the DNA strand. The processive diffusion of the WRKY domain protein along DNA have been further sampled by coarse-grained (CG) MD simulations, conducted at various ionic concentrations and along different DNA sequences. Accompanied single-molecule fluorescence measurements confirmed on the WRKY 1-D processive diffusion along DNA.

Results

Specific vs non-specific DNA association of WRKY with varied stabilities

We conducted microseconds equilibrium MD simulations on the WRKY-DNA complexes, with a specific DNA binding sequence (W-box) and a slightly varied but non-specific DNA sequence, respectively. The specific protein-DNA complex had been constructed directly from the crystal structure (39), with DNA extended to 34 bp (see supplementary information **SI Methods**). Since no crystal structure had been made available for WRKY binding on non-specific DNA, the non-specific protein-DNA binding complex was modeled from the crystal structure by converting the specific core sequence of DNA (5'-CTGGTCAAAG-3' on the preferred strand) to the slightly varied nonspecific sequence (5'-CTGATAAAAG-3') and conducting equilibrium simulation. Using the isothermal titration calorimetry (ITC), we also determined the WRKY dissociation constants with DNA on the above specific and non-specific sequences as K_D=0.1 μM and 8 μM, respectively (see **SI Methods** and **Fig S1**).

By conducting and comparing two 10-μs MD simulations of WRKY (modeled at an ionic concentration of 150 mM) on the specific and non-specific sequences (see **Fig 1**), one notices well localization of WRKY on the DNA around the specific sequence, with comparatively small longitudinal ($\Delta X \sim 1.2 \pm 0.8$Å) and rotational movements ($\Delta\Theta \sim 9.8 \pm 7.4°$) of the protein center of mass (COM), after ~ 2 μs pre-equilibration. In comparison, WRKY modeled on the non-specific DNA demonstrated more significant rotation-coupled relaxation on the DNA, with the protein COM shifted both

longitudinally and rotationally (ΔX ~3.3±0.8 Å and ΔΘ ~ 30.0±9.8°). Note however that the protein did not translocate yet along the DNA since the HBs formed at the protein-DNA interface on the non-specific DNA sequence are still well maintained in the full simulation (e.g., see **Fig 2**).

The protein-DNA structural alignments according to the nearby/bound DNA segment (~10 bp; see **SI Fig S2A**) show that the conformational re-arrangements of WRKY on the non-specific DNA are significant (ΔRMSD ~ 8 Å comparing to the initial structure, see also **SI Fig S2B**), which are substantially larger than that on the specific DNA (ΔRMSD ~ 3 Å). Meanwhile, structural alignments according to protein core structure (excluding the peripheral loop region E107-Y114, G162-K168) demonstrate almost no conformational changes of the domain protein on the non-specific DNA vs on the specific DNA (**SI Fig S2C**). In addition, by measuring protein axial or orientational angle (following the beta strands) with respect to the DNA long axis, one can see that the protein orientational angle with respect to DNA rotates from ~62±5° in the specific DNA binding to ~81±9° in the non-specific binding (**SI Fig S2D**). Hence, the positional re-arrangements of the protein on the non-specific DNA mainly come from the orientational changes of the protein on the DNA. Besides, the non-specific DNA in association with the protein also shows slightly larger fluctuations in its major groove size than that the specific DNA (20.2±0.7Å specific and 20.6±1.2Å non-specific), with detectable correlations between the groove size variation and the protein orientation on the DNA (**SI Fig S2E**). Two movies are provided for viewing the WRKY relaxations on the specific and non-specific DNA (see **SI Movie S1** and **S2**), respectively.

Additionally, we constructed a mutant (mt) WRKY-K122A, with a lowered affinity with specific DNA as being measured by the ITC experiments K_D ~ 1 μM (**Fig S1**). Correspondingly, we performed MD simulation for this mt-WRKY, modeled on the original specific DNA sequence. The results show that the mt-WRKY starts re-arranging along DNA similarly as the non-specific wild-type (wt) binding complex (**Fig 1**). The positional relaxation of the mt-WRKY on DNA show intermediate behaviors in between the specific and the non-specific DNA binding complexes.

WRKY association with DNA is strongly biased onto one strand and the bias is strongly maintained in the specific DNA binding complex

By close examinations, we identified detailed interactions at the protein-DNA interface, for both the specific and the non-specific binding systems (see **Fig 2A**). The schematics summarizing the HB and salt-bridge interactions between WRKY and DNA strands are provided (see **SI Fig S3A**). In particular, we found substantial HB interactions between the protein and the preferred DNA strand (~7 to15 HBs), for both specific and non-specific cases. In the specific binding (**Fig 2B** *left* and **SI Fig S3A**), HBs are formed via K125 to G15, K122 & R131& Y133 & Q146 to G16, Y119 & K144 & Q146 (water mediated) to T17, and Y119 & R135 to C18. Among them, arginine and lysine (R131, R135, & K144) can also form salt-bridge interactions with

the phosphate groups on the DNA; Y119 specifically forms a HB with the C18 base, while K122 also specifically forms a HB with the G16 base (in the core GGTC). In contrast, there are much fewer HBs formed between WRKY and the other or non-preferred DNA strand (~ 2 to 5 HBs). It mainly involves HB and salt-bridge interactions between R117/K118 (on the β2 strand) and the backbone DNA, and three water-mediated HBs from W116 & Q121 & Y134 (see **SI Fig S3A)**. Note that W116, R117, K118, Y119, Q121 and K122 are from the WRKYGQK motif. Except for Y119 and K122 associating specifically with the two core DNA bases on the preferred strand, the rest interact with the DNA backbone of the non-preferred strand.

In comparison, one sees that the HB association of WRKY with the non-specific DNA (**Fig 2B** *middle* and **SI Fig S3B**): K125 to G15, R131 & Y133 & Q146 to A16, Y119 & K144 to T17, R135 to A18 on the preferred strand; G153, Q154 & R149 (on the loop connecting β4 and β5 strand) forming HB and salt-bridge interactions to the DNA backbone; Q121 forming HB with the non-preferred strand.

Consistently, the base specific K122 and Y119 association with the altered core sequence (GATA) becomes absent in the non-specific complex. No water mediated HBs are identified at the protein-DNA interface for the non-specific DNA binding nor for the mutant protein case (**SI Fig S3C**). In the mutant, K122A loses contact with the specific core DNA sequence while most other HBs with the preferred DNA strand are preserved.

The preferred strand HB associations also demonstrate larger fluctuations on the non-specific DNA or for the mutant protein than in the original specific DNA binding (**Fig 2C**). Furthermore, WRKY associates with the non-preferred DNA strand via the β2 strand on the specific DNA, while it has β4&5 strands to associate with the non-preferred strand on the non-specific DNA (**Fig 2B**). Such alteration of the protein-DNA binding interface happens together with significant re-orientation of the protein on the non-specific DNA and instabilities. In the mutant, due to loss of the specific HB contact from K122, the protein also re-orients on the DNA, and associates with the non-preferred DNA strand involving β4&5 strands.

Additionally, we calculated average electrostatic (*ele*) and van der Waals (*vdW*) interaction energies at the protein-DNA interface, for respective DNA strands and the core protein (excluding the peripheral loop region as above; see **SI Table S1**). The protein interactions with the *preferred strand* are quite strong in both the specific (*ele & vdW:* -146±21 & -33±5 kcal/mol) and non-specific DNA binding systems (-145±25 & -35±6 kcal/mol). With an updated DNA force field (see **SI Methods**), the sampled electrostatic association can be even stronger for the specific DNA binding (*ele & vdW:* -158±12 & -29±4 kcal/mol) than for the non-specific system (*ele & vdW:* -131±23 & -25±6 kcal/mol). In contrast, the protein interactions appear much weaker with the *non-preferred strand* (*ele & vdW:* -43±21 & -23±5 kcal/mol specific vs -49±21 & -18±6 kcal/mol non-specific); with the updated DNA force field, the

interaction strengths vary somehow (-32±21& -21±5 kcal/mol specific vs -45±25 & -15±6 kcal/mol non-specific) but remain much weaker than that with the preferred DNA strand.

The hydrophobic interactions between the protein and DNA have been monitored as well (see **SI Fig S4**). The involved hydrophobic residues with the preferred stand also appear more than those with the non-preferred strand, in both specific and non-specific (or mutant) cases. In addition, we counted water molecules around the surface of protein or the protein-DNA interface (within 5 Å). For about a same amount of waters (slight above 300 waters) surrounding WRKY in the specific and non-specific DNA system, fewer waters stay close to the protein-DNA interface on the specific DNA (~36±6 excluding the waters mediating HBs) than on the non-specific DNA (~ 47±8). Thus, the hydrophobic interactions at the protein-DNA interface also favor the protein-DNA specific binding and bias toward the preferred DNA strand.

Atomistic simulation of WRKY diffusion along homogenous poly-A DNA with rotation-coupled protein motions sampled

In the above 10-μs simulation of WRKY binding on the non-specific DNA, we have not yet detected diffusion of the protein. In order to probe essential protein translocation or displacements of protein contacts along DNA, we modeled WRKY on homogenous poly-A dsDNA at a length of 34-bp. It was expected that protein contacts made on homogeneous DNA sequences could be synchronized easier to support comparatively fast protein translocation. On such poly-A DNA, we accordingly captured one complete stepping cycle of the WRKY diffusion, i.e., for 1-bp distance on the DNA, via equilibrium atomistic simulation (see **Fig 3)**. We analyze both the COM movements of the protein along DNA and then collective HB dynamics at the protein-DNA interface.

In **Fig 3A**, representative snapshots from two MD trajectories are shown, demonstrating WRKY moving forward (+X direction, or toward *right*) and backward (toward *left*) along DNA, respectively. The longitudinal (along X) and rotational motions (mapped on the Y-Z plane) of the protein COM along DNA are demonstrated in **Fig 3B**&**C**. In the forward direction, mainly four spatial states reveal (labeled 1 to 4; with the initial pre-equilibrated state 0), according to helical motions of the protein COM on the DNA (**Fig 3B**): In the first ~1.86 μs, WRKY tracks slightly forward along the major groove of the DNA, closely following the preferred strand, moving from state 1 to 2 (1→2; Δx~1.1 Å and $\Delta\Theta$ ~ 21.9°) ; during 1.86-3.08 μs, however, it slightly retracks back to state 1 (2→1**);** at ~ 4.96 μs, the protein quickly steps forward, advancing about 1-bp within 0.2 μs (1→3**;** Δx~1.9 Å and $\Delta\Theta$ ~ 27.1°); after that (> 7.5 μs), WRKY slides forward (3→4 ; Δx~ -0.9 Å and $\Delta\Theta$ ~16.9°) but adjusts its spatial orientation on the DNA to better align with the major groove at the next location (see **SI movie S3** for the protein 1-bp stepping). Comparing to the static DNA binding case, there is still no conformational changes of the protein core during

the forward movements. Nevertheless, the protein orientational changes on the DNA are substantial (the orientation angle spans from ~ 77±10° in the first 5 μs to ~ 56±10° in the last 5 μs; see **SI Fig S5**), so that the domain protein can adjust and re-align with the DNA helical track. Meanwhile, the DNA groove size varies (between ~18 to 22Å) and the variation correlates with the protein orientational change (**SI Fig S5**).

The protein diffusion captured in the backward direction (after a pre-equilibrated state 0) also starts with slight forward motions (0→1) within the first ~1.43 μs (Δx~1.6 Å and ΔΘ ~26.0°), similar to that in the forward trajectory; then it is followed by retracking (1→2'; Δx~0.7 Å and ΔΘ ~ -16.9°) at ~1.43 μs and moving backward at ~3.9 μs (2'→1'; Δx~ -2.3 Å and ΔΘ ~10.3. There is a further sliding backward at ~5.18 μs (1'→3'within 1.3 μs; Δx~ -0.8 Å and ΔΘ ~ -36.2°). At ~ 7.02 μs, a strand-crossing event of WRKY on the DNA in the backward direction shows. Right after that, WRKY binds onto the minor groove of the DNA (see **SI Movie S4**). A conformational change of the protein (ΔRMSD~ 3Å, with HBs between β4 and β5 strands broken) starts at ~1.2 μs, right *before* the transition 1→2', and a new protein conformation is reached at ~4.1 μs (right upon moving backward into state 1', with new HBs formed between β5 and β2) and maintained thereafter. The protein orientational changes on the DNA (from 68±12˚ in the first 5 μs to 46±10˚ for the last 5 μs) and accompanied DNA major groove width variations show similarly (between 18 to 22Å) as during the forward trajectory.

To enhance samplings of the protein movements along the poly-A DNA, we additionally launched ten comparatively short simulation runs (2-4μs each; see individual mappings and accompanied protein-DNA interfacial HB dynamics in **SI Fig S6** and **Fig S7**, respectively), starting from various intermediate states along both the forward and backward diffusional paths. The accumulated samplings along the forward and backward paths are shown in **Fig 3C** (*right*). One sees that that the protein COM follows dominantly a rotation-coupled path, tracking along the helical DNA groove. The protein COMs are also mapped into several population states on the X-Θ plane, which accordingly reveal equilibrium distributions and hence estimated free energetics along the diffusion path (~2 k_BT; see logarithmic probability mapping in **SI Fig S6**).

Further, we had conducted two additional 10-μs equilibrium simulations of WRKY on the poly-A DNA, with one repeated run and the other with a varied initial condition (from a 2-μs equilibrated protein conformation from the non-specific DNA binding). Though protein reorientation or relaxation on the DNA persist (via the COM motions; see **SI Fig S8A-C**), no further protein stepping or diffusion along the poly-A DNA was sampled, as HBs at the protein-DNA interface were stably maintained in both simulations. Mapping of all samplings obtained for WRKY on the poly-A DNA (simulation accumulated ~70 μs, see **SI Table S2**), including static and forward/ backward diffusion, are provided (**SI Fig S8D**). In summary, the longitudinal and rotational motions of the protein are largely coupled during the regular helical

tracking motions of the protein; only occasionally, the domain protein skips the groove tracking and slides across the DNA strand.

WRKY stepping along poly-A DNA with cyclic HB breaking and reforming at the protein-DNA interface sampled in the atomic simulations

In the 10-μs all-atom simulations of WRKY along poly-A DNA, by close inspections on how protein individual residues break and reform HB contacts with the DNA backbone during the protein diffusion, we show the representative protein stepping schematics or HB dynamics on the DNA (see **Fig 4**). According to the HB dynamics revealed at the protein-DNA interface from the simulation (**SI Fig S9A**), we define different HB configurations (**I** to **VII**) and connect them to the protein COM states (i.e., state 1 to 4 from **Fig 3**). Among eight key residues frequently forming HBs with the preferred DNA strand, the very front residue R135 (NH1/NH2) that initially bonds with A18 backbone (O2P; 0.4~1.86 μs as configuration or **config I**, protein COM state 1), has the HB broken first and then fluctuates to form a new HB with A19 (**config II** at ~2 μs, state 2), as other contacts almost remain intact. At ~5 μs, as the protein moves forward (state 3), most of HBs break within ~80 ns, while the central K144 shifts its HB with A17 to A18, and K122 forms a new HB with A17 (**config III**); four of the front HBs (but not the one by R135) reform quickly (**config IV**, for ~30 ns), then R131 reforms HB with A17 (**config V**, for ~ 40 ns), the backside K125 reforms HB with A16 (**config VI**, for ~ 60 ns), and finally, R135 reforms HB with A19, which concludes the 1-bp stepping cycle (**config VII or config I** recovered, at ~ 5.2 μs). Note that for **config III** to **VII** transit quickly (within ~0.2 μs), as the protein COM remains at state 3. During this stepping cycle for 1-bp (~7 μs), therefore, the protein COM first oscillates back and forth (with protein orientational changes) and then moves forward (via state transitions $\mathbf{1 \rightarrow 2 \rightarrow 1 \rightarrow 3}$ as in **Fig 3**), i.e., tracking along the DNA major groove, reorienting with the groove (~5 μs), until the majority of HBs suddenly shifted (broken & reformed). Further movements revealed in the simulation (6-10 μs, see **SI Fig S9B**) account for some protein slipping (~2 bp step, incomplete): the HBs break in a way slightly differently, e.g., Y119 breaks contact first and then R135; the middle and rear contacts break and have not yet reformed, while the COM of protein shifts ~ 2 bp. The schematics of protein-DNA interfacial HBs formed on the non-preferred strand is also provided (**SI Fig S9C**): though there are only 2-3 HBs formed occasionally, one finds that R118 breaks and reforms HB with the DNA phosphates from T20' to T23' throughout the 10-μs simulation (across 2 ~ 3 bp).

In the backward movements of WRKY along DNA, the protein also tracks along the major groove initially (< 6.6 μs; see **SI Fig S10**). It starts with R131 squeezing on the neighboring K125 (from **config I′** 0.42-3.9 μs to **II′** 3.9~5.18 μs or the protein COM state $\mathbf{1 \rightarrow 2' \rightarrow 1'}$) to break the back contact K125-A15. After R131 forming stable contact with A15, the COM of protein moves backward ($\mathbf{1' \rightarrow 3'}$), Q146-A16 HB breaks as K144 slides backward to contact A16 (**config III′**). The continuous movements of the COM have most of the HBs broken (**config IV′**). After that, the

middle region re-adjusts with shifted nucleotides (**config V′ & VI′**, 5.22-5.47μs). Finally, the edge residue R135 reforms HB with A17 (**config VII′**, 5.47-5.87μs), the initial set of contacts almost reform, except for the one from K125 to A14. However, WRKY seems to reduce its association then with the DNA, and crosses the preferred strand to move from the major groove to the minor groove (see **SI Fig S11A**): One can find five residues (R131, Y133, K142, K144, and Q146) res-establish contacts with the non-preferred strand after crossing the strand and sliding further ~ 2 bp backward along the DNA, while K142 and K144 keep associating with both strands (see **SI Fig S11B** for 6.6-10 μs with the protein COM state **4′ → 5′**).

The WRKY electrostatic association bias to the preferred DNA strand is less distinct during protein diffusion than in the static DNA binding

Further, we calculated the WRKY interaction energetics with respective strands of DNA during the forward and backward diffusional movements (see **SI Fig S12A** and **Table S1**). During the 1-bp stepping along the forward path (< 5 μs), the interaction energetics with the preferred strand (*ele* -137±27 and *vdW* -23±8 kcal/mol) are weaker than in static binding; the *ele* interactions with the non-preferred strand (-42±21 kcal/mol) are nevertheless similar to that in the static binding. Later (during 5-10 μs, with protein slipping and stochastic motions), the *ele* interactions weakens on the preferred strand further (-109±33 kcal/mol) but is still stronger than that on the non-preferred strand (-65±35 kcal/mol). A similar trend reveals in the backward movements, i.e., the WRKY *ele* association with the preferred strand is weaker than that in the static binding (-130±23 and -119±35 kcal/mol, in early and late simulation stage, respectively), while the association with the non-preferred strand strengthens (-35±25 and -79±47 kcal/mol, in early and late stage, respectively). With stochastic movements become prominent in the late stage of the forward/backward diffusion, the accompanied *ele* fluctuations also increase while the protein energetic distinctions between the two DNA strands decrease, comparing to the static binding systems. The average energetics obtained from various simulation systems (including those performed under the updated DNA force field) are summarized in **Fig 5A.**

Since WRKY demonstrates more or less bias in association with the preferred DNA strand over the non-preferred one, from the static binding to diffusional movements, we then analyzed the energetic disparities and correlations between the two DNA strands in association with the protein (see **SI Fig S12B**). In order to measure how differently the protein interaction energetics are contributed by the two DNA strands, we calculated t-values that characterize average energetic differences over the standard errors or fluctuations (see **SI methods**). The *ele* t-value in static and specific DNA binding is indeed highest (t_s^{ele}=172 for specific and t_{ns}^{ele}= 143 for nonspecific, with both p-values < 10^{-5}; note that for N=2500 samples, t-value needs to be larger than 98 or 128 to be statistically significant, i.e., with p-value <0.05 or 0.01), which then becomes much lower in diffusion (t_f^{ele}=140 and 46 for early and late stage *forward*, and t_b^{ele}=139 and 34 for early and late stage *backward*). Next, to assess

whether the protein association energetics with the two DNA strands are dynamically correlated, the Pearson correlation coefficients r were calculated between the time-dependent energetic data sets. In specific and nonspecific DNA binding r_s^{ele}=-0.072 and r_{ns}^{ele}=0.099 (for N=2500 and p-value < 0.05 or 0.01, one needs |r|> 0.04 or 0.05). Hence, correlations are detectable between the two-strand energetics. In the forward and backward diffusion, the correlation strength can become much larger, occasionally, e.g., from r_f^{ele} = -0.008 to -0.52 *forward* and from r_b^{ele} =-0.048 to 0.19 *backward* (**SI Table S1** for full energetics, t-value and r values for various simulation systems). The analyses thus indicate that WRKY electrostatic association shows less bias and more coordination between the two DNA strands during the protein diffusion than in the static DNA binding.

Coarse-grained simulations of WRKY on processive diffusion along DNA

Under current technology one cannot yet sample processive diffusion of the protein on DNA using unbiased atomic simulations, to do that we conducted coarse-grained (CG) MD simulations to the WRKY-DNA complex, using CafeMol (40) (see **SI Methods**). In the CG presentation, each amino acid is represented by a sphere, while each nucleotide is represented by three spheres (41). Correspondingly, there is no specific or the HB interactions modeled between protein and DNA. Nevertheless, the electrostatic association between the charged amino acids and phosphate groups on DNA have been well taken into account. In **SI Fig S13A-D**, we show that in the CG simulations, WRKY conducts processive diffusion along DNA at variable ionic conditions (from 50 to 150 and 200 mM). In particular, at 150 mM (as modeled in the all-atom MD), WRKY demonstrates mainly the DNA groove tracking (~73%) with occasional strand crossing motions (~1.5%; see **SI Fig S13E-F** and **SI Movie S5**), seemingly consistent with observations in the all-atoms simulations. In comparison, at lower and higher ionic conditions (50 mM and 200 mM), highly regular groove-tracking motions and occasional 'micro-dissociation' or hopping events show, respectively (**SI Fig S13B&D** and **SI Movie S6&S7**), which correspond well to weak and strong charge screening situations. Interestingly, we have also detected variations of stepping size of WRKY along DNA on different DNA sequences (poly-A, poly-AT, and random sequence, all at 150 mM; see **SI Fig S14**), with poly-A showing a high chance of 1-bp stepping (~44%), and poly-AT (with 2-bp periodicity) showing a lowered percentile of the 1-bp steps (~20%) but a notable portion of 2-bp steps (~17%). The observations suggest that DNA sequence periodicity and variations modulate the protein-DNA interactions (even in the absence of HB interactions), which then impact on the protein step size variations.

In addition, we verified the processive diffusion of WRKY on DNA experimentally, obtaining single molecule fluorescence measurements (see **SI Methods** and **Fig S15**), albeit current resolution is not high enough to discern base pair movements of the protein. The diffusion coefficient has been estimated at about 0.05 $bp^2/\mu s$, which is consistent with our computational samplings of the protein stepping on the DNA in

the 10-μs all-atom MD simulations.

Discussion

In this work, based upon high-resolution structures of DNA binding complexes of a representative TF domain protein WRKY, we demonstrated microseconds molecular dynamics of protein diffusion along DNA with unprecedented details. To avoid artifacts from external force or bias, equilibrium all-atom simulations were conducted for both static binding and protein diffusion on DNA, which were accumulated to ~100 μs (under Amber99SB-ILDN force field (42) for protein and Amber94 force field (43) for nucleic acids, using Gromacs (44); see **SI Methods** for details), including several 10-μs long MD simulations and multiple distributed 2-3 μs runs to improve samplings (see list of simulations in **SI Table S2**). In addition, ~ 40 μs all-atom simulations were also conducted with an updated DNA force-field Parmbsc1 or BCS1 (45), which seems to slightly stabilize DNA backbone motions. The simulations conducted under the updated force field reproduced the dominant features of WRKY-DNA binding on specific and non-specific DNA (**SI Fig S16**), with biased association still toward the preferred DNA strand. The 1-bp stepping dynamics of the WRKY domain protein on the poly-A DNA has also been well captured under the updated force field (**SI Fig S17** and **Movie S8**).

The WRKY domain protein preferentially binds one strand of DNA and the protein orientational change on the DNA is significant between specific and non-specific DNA binding. For protein static binding and recognition on DNA, both specific and non-specific DNA (with slightly varied core sequences GGTC and GATA) binding complexes of WRKY were examined, together with a mutant K122A protein complex with the specific DNA. The corresponding protein-DNA binding affinities were determined via accompanied ITC measurements, with dissociation constants measured at 0.1μM, 8μM, and 1μM for the specific, non-specific, and mutant K122A systems, respectively. In all these systems, one DNA strand is always preferentially bound by WRKY. The other strand, the non-preferred one, interacts with the protein much weaker to allow protein to associate with the DNA differently between specific and non-specific binding modes, e.g., via variable protein β-strand regions. Consequently, the domain protein varies its orientation and affinity to different DNA sequences, and recognizes certain bases on the preferred DNA strand upon the specific DNA binding. In the simulations, we have found no essential conformational change of the domain protein from the specific to non-specific DNA binding, and to regular tracking along DNA, though an occasional protein conformation transition was detected prior to directional reversal of the protein on the DNA. Meanwhile, current study shows that relative conformational changes between the domain protein and DNA, i.e., the re-orientation of the protein on the DNA, are highly significant and contribute essentially to the distinction between the non-specific and specific DNA binding (**Fig 1**&**2**, **SI Fig S2**&**16**, or summarized in **Fig 5B**). Such findings provide structural clues to previous work suggesting switch of TF conformational mode in DNA search and recognition (4, 5, 46, 47). Correspondingly, the protein center of the

mass can demonstrate deviations on DNA during the protein orientational relaxation on the DNA, even when there is no real movements or translocation of the protein.

WRKY 1-bp stepping on poly-A DNA is detected from all-atom equilibrium simulations while its processive diffusion statistics is obtained from coarse-grained simulations. With current computing technologies, for TF protein diffusion with an average stepping cycle lasting over tens of microseconds, it is still hard to sample the protein movements at atomic resolution. Our all-atom simulations show that protein-DNA interfacial HB contacts are constantly present and the corresponding dynamics can be rate limiting to hinder the protein diffusion. For homogenous poly-A DNA, the HB dynamics seems to be facilitated along the identical DNA sequences. Consequently, we were able to identify a complete 1-bp protein stepping cycle following the major groove of DNA, which is regulated by collective motions of residues or HB dynamics at the protein-DNA interface, as individual HBs break and reform throughout the cycle. In fact, three protein stepping events were detected from six 10-μs long atomic MD simulations of the WRKY on poly-A DNA (one event captured under the updated DNA force field). Combining these simulations with multiple distributed simulations performed additionally along the forward and backward diffusion paths, the dominant rotation-coupled DNA helical tracking motions of the protein are demonstrated, with ~2 k_BT diffusional free energetics estimated (**SI Fig S6**), which is consistent with previous measurements (13, 14). Furthermore, by conducting CG simulations of WRKY at the residue level on ~200 bp DNA, processive protein diffusion along DNA were sampled at various ionic conditions and sequence patterns (**SI Fig S13**&**14**). Stochastic directional reversal and DNA strand crossing events have been well sampled in the CG simulations, while such events were captured only once or twice in the atomic simulations. The corresponding processive 1D diffusion of WRKY along DNA was also confirmed by accompanied single molecule florescence experiments, albert the measurements were not at high resolution to detect protein stepping motions. Interestingly, in current CG simulations of the processive diffusion of WRKY, high percentiles of 1-bp stepping motions of the domain protein show along homogeneous poly-A DNA, while the percentile of the 1-bp stepping drops (or with more 2-3 bp steps, **SI Fig S14**) for WRKY moving along the random DNA sequences. It thus suggests that the DNA sequence patterns make direct impacts on the protein stepping statistics.

Stochastic variations revealed in the domain protein stepping provide clues for understanding step size variations in other nucleic acid walkers. In the all-atom simulations, both forward and backward movements of the WRKY domain protein along DNA have been revealed. In the forward direction, right after an elementary 1-bp stepping of WRKY, stochasticity is noticeable as WRKY slips further for ~ 2 bp, incompletely, as the related HBs break and part of them reform at the 2-bp distance. In the other case, some protein conformational transition (HBs broken between β-strand 4 & 5) occurs right before the protein moving backward; soon after ~1-bp stepping backward, the protein shows prominent stochastic motions on the DNA as

crossing the preferred strand to move from the major to the minor groove. Such types of diffusional motions of protein along DNA have been captured in current (see **SI Fig S13**) and previous CG simulation studies (48, 49), though no protein side chain motions or protein-DNA HBs can be modeled in the CG system. The WRKY domain protein stepping on DNA is comparable to other nucleic acids walkers, or molecular walkers following a quasi-periodic track. For example, motor proteins such as DNA packaging motors or helicases had been detected with variable stepping sizes from single molecule measurements (50-54). The stepping motions of the motor proteins can be similarly fast as the TF proteins, though substrate binding or chemical catalysis and mechano-chemical coupling that supports directional movements of the motor proteins can be quite slow (e.g., over milliseconds). Although various models were presented to explain diverse stepping behaviors of motor proteins, from current simulation, one would infer that the multiple stepping sizes simply arise because of non-synchronized and rate-limiting motions of individual protein residues forming HB contacts on the DNA backbone. Besides, stochasticity always plays a significant role in the protein stepping or sliding due to thermal fluctuations. The simulated TF protein stepping dynamics, stochastic variations, and DNA sequence effects would await experimental validations at the sufficient high or bp resolution.

The protein electrostatic association bias with one strand of DNA can be marginally maintained to assist the domain protein diffusion and maximally employed to support protein DNA sequence recognition. To further understand how such a WRKY domain protein searches and locates specific target sequence on the DNA, we note that even though WRKY distinguishes the two strands of DNA by almost always associating tightly with the preferred strand, the disparity between protein association with the two strands varies from dynamical search to static binding or recognition stage. During diffusion or stochastic movements of protein on the DNA, the disparity on the protein vdW association with two strands almost vanishes, while the electrostatic differentiation persists but is only marginally maintained, likely due to protein random reorientations on the DNA. For protein regular stepping or groove tracking along DNA, the protein-DNA energy disparity between the two strands increases, as constant re-orientation of the protein happens along the DNA helical track. It appears that some coordination between the two DNA strands in association with the protein supports the protein movements. Quasi-static protein binding on the DNA with lowered fluctuations then enhances the protein association disparities between the two DNA strands. Such enhanced protein-DNA strand bias may contribute to fine-tuning the protein orientation and to support specific DNA sequence recognition on the preferred DNA strand. Due to additional electrostatic stabilization or reduced fluctuations particularly on the preferred DNA strand, the corresponding energetic distinction between the preferred and non-preferred strand becomes maximized for the protein on the specific DNA. The biased protein association with the preferred DNA strand can also perturb base pairing in the duplex DNA, thus assist base readout on the DNA strand for sequence recognition.

Hence, for a small domain TF protein, our studies bring a working scenario in which the biased protein association with one strand of the dsDNA can be marginally sustained during stochastic search process to facilitate fast protein movements, while the bias can be maximally employed into the quasi-static binding and DNA sequence recognition as the protein reorients and stabilizes on the specific DNA. Such a scenario is related to protein geometry on the DNA helical structure, so it seems to apply for monomeric or small TF domain proteins that fit with the DNA groove. Interestingly, for dimeric proteins with two DNA binding domains, such as Myc-Max we recently studied, it is found that the two basic regions or domains bind respectively with the two complementary strands of DNA, i.e., with each domain preferentially bound with one strand (55). The movements of such a dimeric TF protein on DNA then rely largely on coordination between the two domains. Such a perspective is supported by recent structure-based bioinformatic analyses (56), which show statistically that multi-specific TFs intend to form more HBs with one strand than with the other on the DNA, while highly specific DNA binding proteins, typically dimeric type-II restriction endonucleases, associate non-preferentially with both DNA strands. Combining with these findings, the biased DNA strand association scenario appears to be generic for small TF domain proteins to balance target search and recognition on the DNA. For larger or oligomeric TF proteins, however, additional considerations of protein internal degrees of freedom or coordination are needed.

Materials and Methods

Detailed descriptions about obtaining the crystal structure, the setup of atomic and coarse-grained simulations, the Isothermal Titration Calorimetry (ITC) experiments, and the single molecule florescence experiments are provided in ***SI Appendix***.

Acknowledgements

This work has been supported by NSFC Grant #11775016 and #11635002. JY has been supported by the CMCF of UCI via NSF DMS 1763272 and the Simons Foundation grant #594598 and start-up fund from UCI. We acknowledge the computational support from the Special Program for Applied Research on Super Computation of the NSFC Guangdong Joint Fund (the second phase) under Grant No. U1501501 and from the Beijing Computational Science Research Center (CSRC).

References

1. Berg OG & von Hippel PH, Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol* 193(4):723-750 (1987).

2. von Hippel PH & Berg OG, Facilitated target location in biological systems. *J Biol Chem* 264(2):675-678 (1989).

3. Halford SE & Marko JF, How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res* 32(10):3040-3052 (2004).

4. Slusky M & Mirny LA, Kinetics of Protein-DNA Interaction: Facilitated Target Location in Sequence-Dependent Potential. *Biophy J* 87(6):4021-4035 (2004).

5. Bauer M & Metzler R, Generalized Facilitated Diffusion Model for DNA-Binding Proteins with Search and Recognition States. *Biophy J* 102(10):2321-2330 (2012).

6. Shvets AA, Kochugaeva MP, & Kolomeisky AB, Mechanisms of Protein Search for Targets on DNA: Theoretical Insights. *Molecules* 23(9):2106 (2018).

7. Richetti M, Metzger W, & Heuman H, One-dimensional diffusion of Escherichia coli DNA-dependent RNA polymerase: a mechanism to facilitate promoter location. *Proc Natl Acad Sci U S A* 85(13):4610-4614 (1988).

8. Blainey PC, van Oijen AM, Banerjee A, Verdine GL, & Xie XS, A base-excision DNA-repair protein finds intrahelical lesion bases by fast sliding in contact with DNA. *Proc Natl Acad Sci U S A* 103(15):5752-5757 (2006).

9. Tafvizi A, Huang F, Fersht AR, Mirny LA, & van Oijen AM, A single-molecule characterization of p53 search on DNA. *Proc Natl Acad Sci U S A* 108(2):563-568 (2011).

10. Hammer P*, et al.*, The lac Repressor displays facilitated Diffusion in Living Cells. *Science* 336(6088):1595-1598 (2012).

11. Redding S & Greene EC, How do proteins locate specific targets in DNA? *Chem Phys Lett* 570(35):1-11 (2013).

12. Gorman J & Greene EC, Visualizing one-dimensional diffusion of proteins

along DNA. *Nat Struct Mol Biol* 15(8):768 (2008).

13. Blainey PC, *et al.*, Nonspecifically bound proteins spin while diffusing along DNA. *Nat Struct Mol Biol* 16(12):1224-1229 (2009).

14. Marklund EG, *et al.*, Transcription-factor binding and sliding on DNA studied using micro-and macroscopic models. *Proc Natl Acad Sci U S A* 110(49):19796-19801 (2013).

15. Robinson AD & Finkelstein IJ, High-Throughput Single-Molecule Studies of Protein-DNA Interactions. *FEBS Lett* 588(19):3539-3546 (2014).

16. Liu C, Liu Y-L, Perillo EP, Dunn AK, & Yeh H-C, Single-Molecule Tracking and Its Application in Biomolecular Binding Detection. *IEEE Journal of Selected Topics in Quantum Electronics* 22(4):6804013 (2016).

17. Stracy M & Kapanidis AN, Single-molecule and super-resolution imaging of transcription in living bacteria. *Methods* 120:103-114 (2017).

18. Nadassy K, Wodak SJ, & Janin J, Structural Features of Protein-Nucleic Acid Recognition Sites. *Biochemistry* 38(7):1999-2017 (1999).

19. MacKerell Jr AD & Nilsson L, Molecular dynamics simulations of nucleic acid–protein complexes. *Curr Opin Struct Biol* 18(2):194-199 (2008).

20. Klepeis JL, Lindorff-Larsen K, Dror RO, & Shaw DE, Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol* 19(2):120-127 (2009).

21. Perilla JR, *et al.*, Molecular dynamics simulations of large macromolecular complexes. *Curr Opin Struct Biol* 31:64-74 (2015).

22. Furini S, Barbini P, & Domene C, DNA-recognition process described by MD simulations of the lactose repressor protein on a specific and a non-specific DNA sequence. *Nucleic Acids Res* 41(7):3963-3972 (2013).

23. Etheve L, Martin J, & Lavery L, Protein–DNA interfaces: a molecular dynamics analysis of time-dependent recognition processes for three transcription factors. *Nucleic Acids Res* 44(20):9990-10002 (2016).

24. Wieczor M & Czub J, How proteins bind to DNA: target discrimination and dynamic sequence search by the telomeric protein TRF1. *Nucleic Acids Res* 45(13):7643-7654 (2017).

25. Khabiri M & Freddolino PL, Deficiencies in Molecular Dynamics Simulation-Based Prediction of Protein–DNA Binding Free Energy Landscapes. *J Phys Chem B* 121(20):5151-5161 (2017).

26. Pandey B, Grover A, & Sharma P, Molecular dynamics simulations revealed structural differences among WRKY domain-DNA interaction in barley (Hordeum vulgare). *BMC Genomics* 19(1):132 (2018).

27. Furini S, Domene C, & Cavalcanti S, Insights into the Sliding Movement of the Lac Repressor Nonspecifically Bound to DNA. *J Phy Chem B* 114(6):2238-2245 (2010).

28. Qi Y, *et al.*, Strandwise translocation of a DNA glycosylase on undamaged DNA. *Proc Natl Acad Sci U S A* 109(4):1086-1091 (2012).

29. Marklund EG, *et al.*, Transcription-factor binding and sliding on DNA studied using micro- and macroscopic models. *Proc Natl Acad Sci U S A*

110(49):19796–19801 (2013).

30. Zacharias M, Atomic Resolution Insight into Sac7d Protein Binding to DNA and Associated Global Changes by Molecular Dynamics Simulations. *Angew Chem Int Ed Engl* 131(18):6028-6033 (2019).

31. Silva D-A*, et al.*, Millisecond dynamics of RNA polymerase II translocation at atomic resolution. *Proc Natl Acad Sci U S A* 111(21):7665-7670 (2014).

32. Da L-T*, et al.*, T7 RNA polymerase translocation is facilitated by a helix opening on the fingers domain that may also prevent backtracking. *Nucleic Acids Res* 45(13):7909-7921 (2017).

33. Maffeo C & Aksimentiev A, Molecular mechanism of DNA association with single-stranded DNA binding protein. *Nucleic Acids Res* 45(21):12125-12139 (2017).

34. Peng S*, et al.*, Target search and recognition mechanisms of glycosylase AlkD revealed by scanning FRET-FCS and Markov state models. *Proc Natl Acad Sci U S A* 117(36):21889-21895 (2020).

35. Tian J, Wang L, & Da L-T, Atomic resolution of short-range sliding dynamics of thymine DNA glycosylase along DNA minor-groove for lesion recognition. *Nucleic Acids Res* 49(3):1278-1293 (2021).

36. Rushton PJ, Somssich IE, Ringler P, & Shen QJ, WRKY transcription factors. *Trends Plant Sci* 15(5):247-258 (2010).

37. Eulgem T, Rushton PJ, Robatzek S, & Somssich IE, The WRKY superfamily of plant transcription factors. *Trends Plant Sci* 5(5):199-206 (2000).

38. Duan M-R*, et al.*, DNA binding mechanism revealed by high resolution crystal structure of Arabidopsis thaliana WRKY1 protein. *Nucleic Acids Res* 35(4):1145-1154 (2007).

39. Xu Y-p, Xu H, Wang B, & Su X-D, Crystal structures of N-terminal WRKY transcription factors and DNA complexes. *Protein Cell* 11(3): 208-213 (2020).

40. Kenzaki H*, et al.*, CafeMol: A coarse-grained biomolecular simulator for simulating proteins at work. *J Chem Theory Comput* 7(6):1979-1989 (2011).

41. Freeman GS, Hinckley DM, & de Pablo JJ, A coarse-grain three-site-per-nucleotide model for DNA with explicit ions. *J Chem Phys* 135(16):10B625 (2011).

42. Lindorff-Larsen K*, et al.*, Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics* 78(8):1950-1958 (2010).

43. Cornell WD*, et al.*, A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J Am Chem Soc* 117(19):5179-5197 (1995).

44. Berendsen HJ, van der Spoel D, & van Drunen R, GROMACS: a message-passing parallel molecular dynamics implementation. *Comput Phys Commun* 91(1-3):43-56 (1995).

45. Ivani I*, et al.*, Parmbsc1: a refined force field for DNA simulations. *Nat methods* 13(1):55 (2016).

46. Hu L, Grosberg AY, & Bruinsma R, Are DNA transcription factor proteins

maxwellian demons? *Biophys J* 95(3):1151-1156 (2008).

47. Zhou H-X, Rapid search for specific sites on DNA through conformational switch of nonspecifically bound proteins. *Proc Natl Acad Sci U S A* 108(21):8651-8656 (2011).

48. Tan C & Takada S, Dynamic and structural modeling of the specificity in protein–DNA interactions guided by binding assay and structure data. *J Chem Theory Comput* 14(7):3877-3889 (2018).

49. Saito M, Terakawa T, & Takada S, How one-dimensional diffusion of transcription factors are affected by obstacles: coarse-grained molecular dynamics study. *Molecular Simulation* 43(13-16):1315-1321 (2017).

50. Dumont S*, et al.*, RNA translocation and unwinding mechanism of HCV NS3 helicase and its coordination by ATP. *Nature* 439(7072):105 (2006).

51. Myong S, Bruno MM, Pyle AM, & Ha T, Spring-loaded mechanism of DNA unwinding by hepatitis C virus NS3 helicase. *Science* 317(5837):513-516 (2007).

52. Tomko EJ, Fischer CJ, Niedziela-Majka A, & Lohman TM, A nonuniform stepping mechanism for E. coli UvrD monomer translocation along single-stranded DNA. *Mol Cell* 26(3):335-347 (2007).

53. Moffitt JR*, et al.*, Intersubunit coordination in a homomeric ring ATPase. *Nature* 457(7228):446 (2009).

54. Chemla YR, Revealing the base pair stepping dynamics of nucleic acid motor proteins with optical traps. *Phys Chem Chem Phys* 12(13):3080-3095 (2010).

55. Dai L & Yu J, Inchworm stepping of Myc-Max heterodimer protein diffusion along DNA. *Biochem Biophys Res Commun* 533(1):97-103 (2020).

56. Lin M & Guo J-t, New insights into protein–DNA binding specificity from hydrogen bond based comparative study. *Nucleic Acids Res* 47(21):11103-11113 (2019).

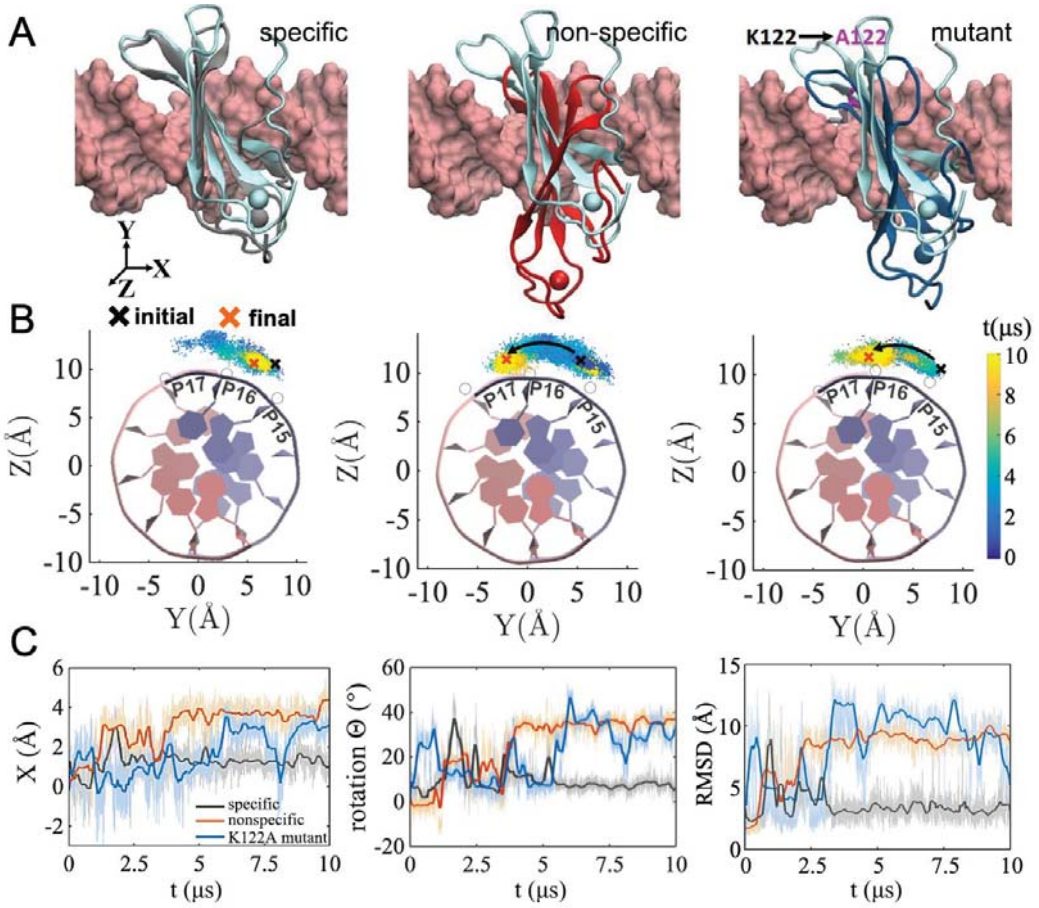


Fig 1 Specific and non-specific DNA association of WRKY. (A) Comparisons of the initial (cyan) and final (gray, red, blue) structures of the simulation of the wild-type (wt) protein binding on the specific DNA *(left)*, the non-specific DNA *(middle)*, and the mutant (mt) protein (K122A) binding on the specific DNA *(right)*. The XYZ-axis is denoted and the longitudinal axis of DNA follows the X direction. (B) The rotational relaxation of the center of mass (COM) of the protein along DNA projected onto the Y-Z plane. The initial and final positioning (due to the protein relaxation on the DNA but not translocation) are denoted. The time evolution is represented by coloring (from blue to yellow). The DNA structure is shown for reference. (C) The relaxation of the protein COM along X & Θ and the protein-DNA RMSDs, showing the protein re-arrangements on the DNA (see **SI Fig S2** with technical details), for respective simulation systems (wt specific, dark curves; non-specific, orange; and K122A mt, blue). Note that the heavy lines are smoothed from the original time series (X(t), Θ(t), and RMSD(t)) over a sliding window ~ 100 ns, and a similar data smooth procedure is conducted for other plots.

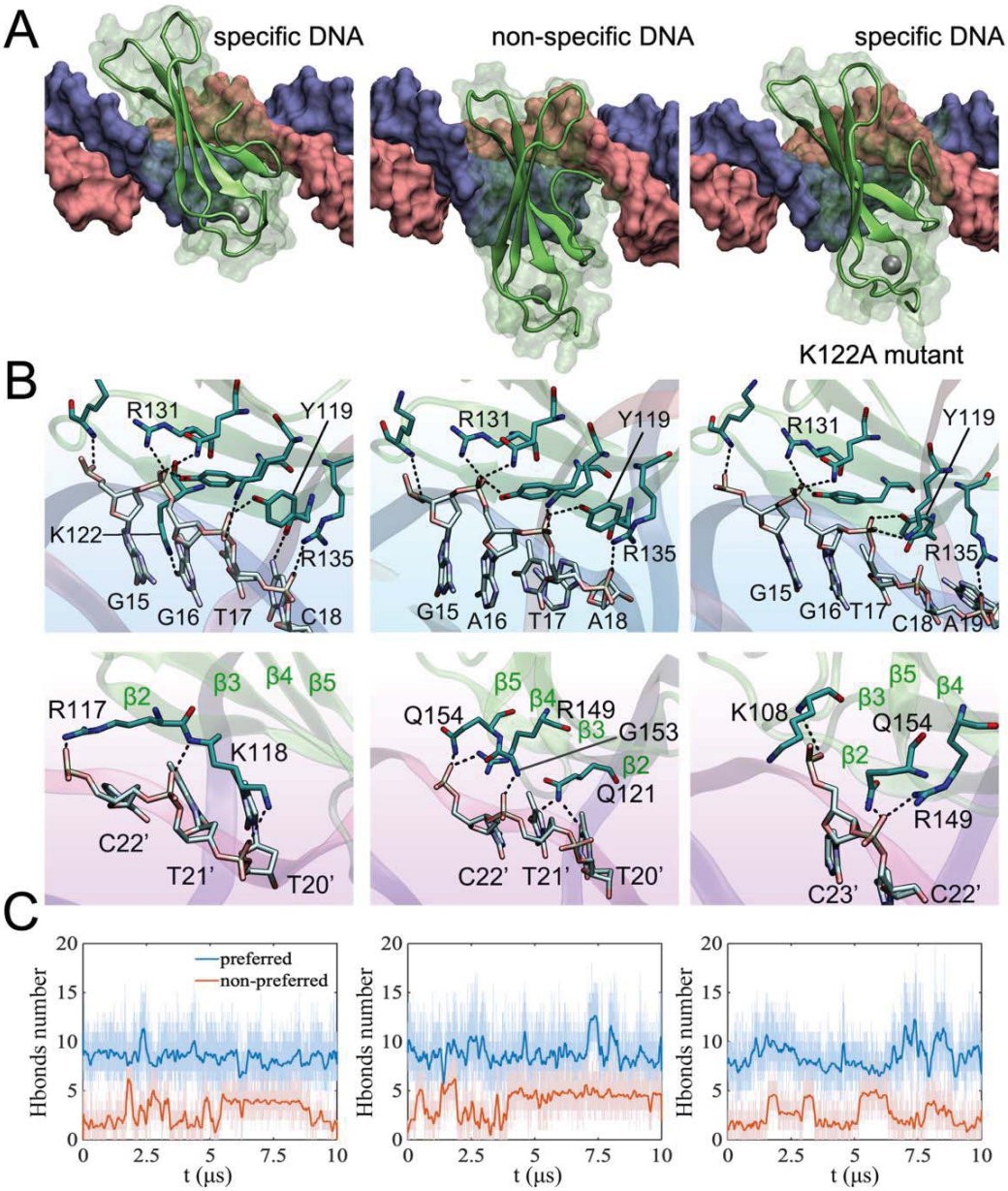


Fig 2. The association between the WRKY domain protein and respective strands of DNA. (A) WRKY association on the specific DNA (core sequence: GGTC; *left*), the non-specific DNA (core: GATA; *middle*), and the mutant K122A on the specific DNA (*right*). The equilibrated protein-DNA complexes are shown in surface representation, with the protein colored in green, and the DNA strands in blue (the preferred) or pink (the non-preferred). (B) The hydrogen bonds (HBs) at the protein-DNA interface are shown on the preferred strand (*top*) and the non-preferred strand (*bottom*). (C) Time-dependent HB statistics at the protein-DNA interface are provided on the respective DNA strands. The HB is defined by a cut-off distance of 3.5 Å between the donor and acceptor atoms and an associated donor-hydrogen-acceptor angle of 140°; the HBs are counted for those formed > 50 ps within a sliding 1-ns simulation window.

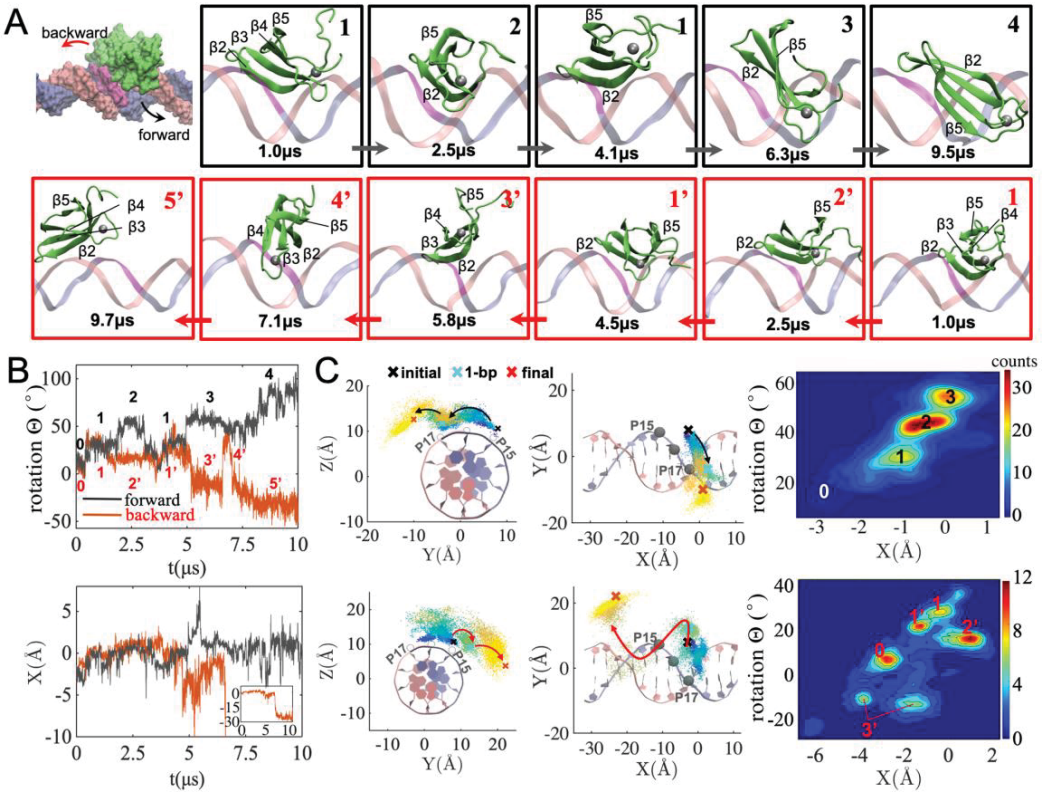


Fig 3. The diffusion of WRKY along poly-A DNA in the forward and backward direction revealed from two 10-μs atomistic MD simulations. (A) The representative structural snapshots taken from the simulation trajectories forward (*top*, from the left to the right, via state 1 → 2 → 1→ 3 → 4 according to the protein COM movements shown in B) and backward (*bottom*, from the right to the left, via state 1 → 2' → 1'→ 3' → 4'→5', with primed labels to differentiate from the forward states, as characterized in B), with the WRKY domain protein shown in green and two DNA strands in blue (the preferred strand) and pink (the non-preferred strand). (B) The helical trajectories of the protein COM along the DNA, shown for the angular Θ(t) (*top*) and the longitudinal movement X(t) (*bottom*) from the simulation. The coordinate system is defined the same as in **Fig 1**. Five (forward, dark line) and Six (backward, orange line) states are identified along the angular coordinates. (C) The protein COM helical motions along DNA are mapped on the Y-Z plane (*left*) and then on the X-Y plane (*middle*), colored by the simulation time (blue to yellow, as in **Fig 1B**). The dsDNA cartoon is also shown for reference. The further sampled protein movements mapped on the X-Θ plane for respective forward (*top right*) and backward (*bottom right*) paths (from an original 10-μs forward/ backward simulation and additional five distributed runs for 2-4μs each; colored according to counts of a total of 28412 snapshots into 200x200 grids on the X-Θ plane; see **SI Fig S6** for individual maps and -ln*P* mapping with normalized counts or probability *P*). Note that the state 3' identified from (B) splits into two populations along -X(t) with a same Θ(t) due to the strand crossing.

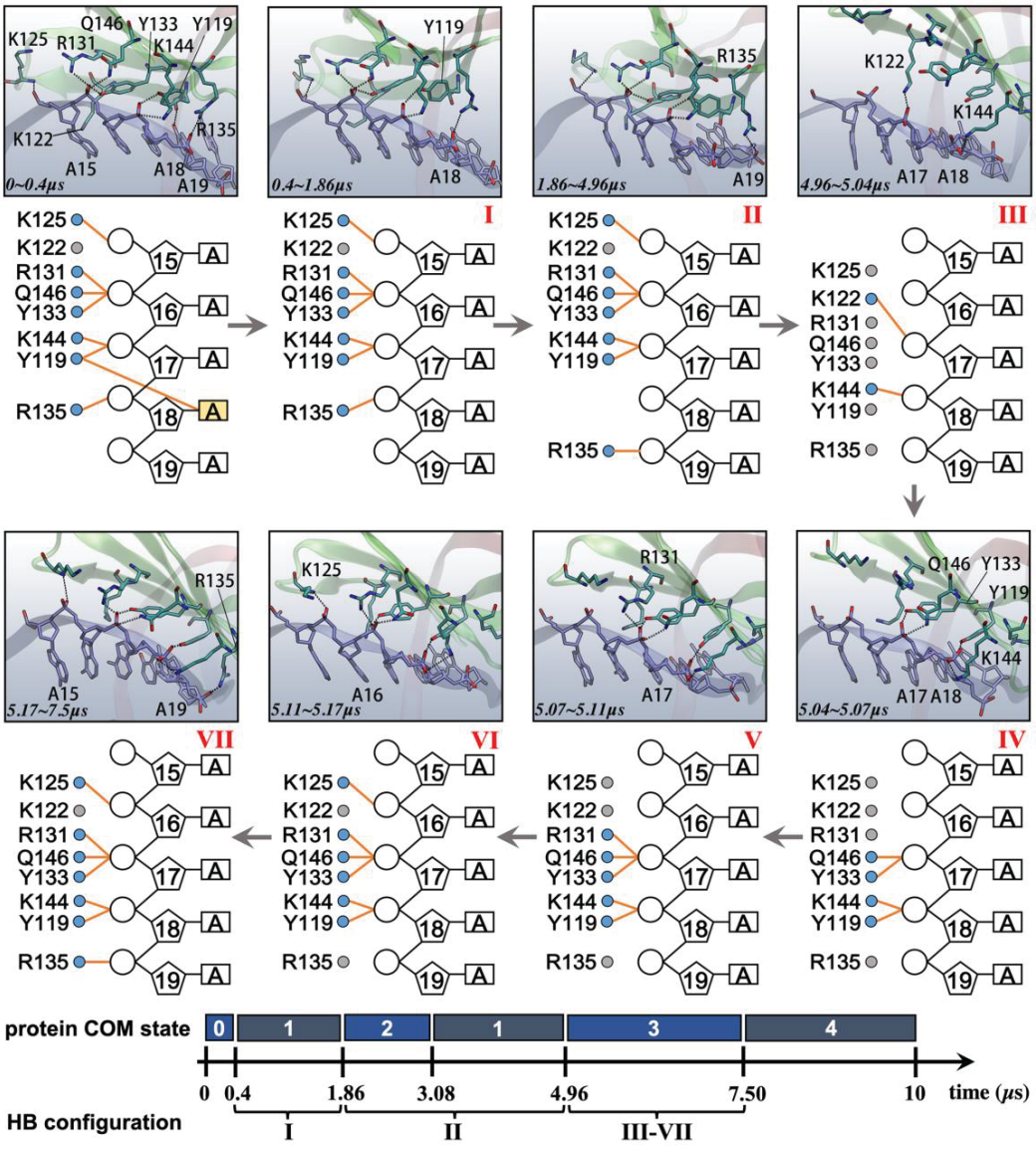


Fig 4. The stepping schematics and structural views of WRKY moving forward along poly-A DNA during diffusion from the 10-μs all-atom equilibrium MD simulation. Since WRKY associates closely with the preferred strand of DNA, we show schematics of eight key residues (filled circle) from WRKY that make HB contacts with the preferred strand (open circle, pentagon and rectangle for the phosphate, sugar and base of a nucleotide, respectively). The HBs in the schematics are depicted in orange lines. The corresponding molecular views at the protein-DNA interface are provided (the preferred and non-preferred strand in blue and pink, respectively; WRKY protein in green). The configurations I to VII are defined according to the HB dynamics at the protein-DNA interface (see text and **SI Fig S9A**).

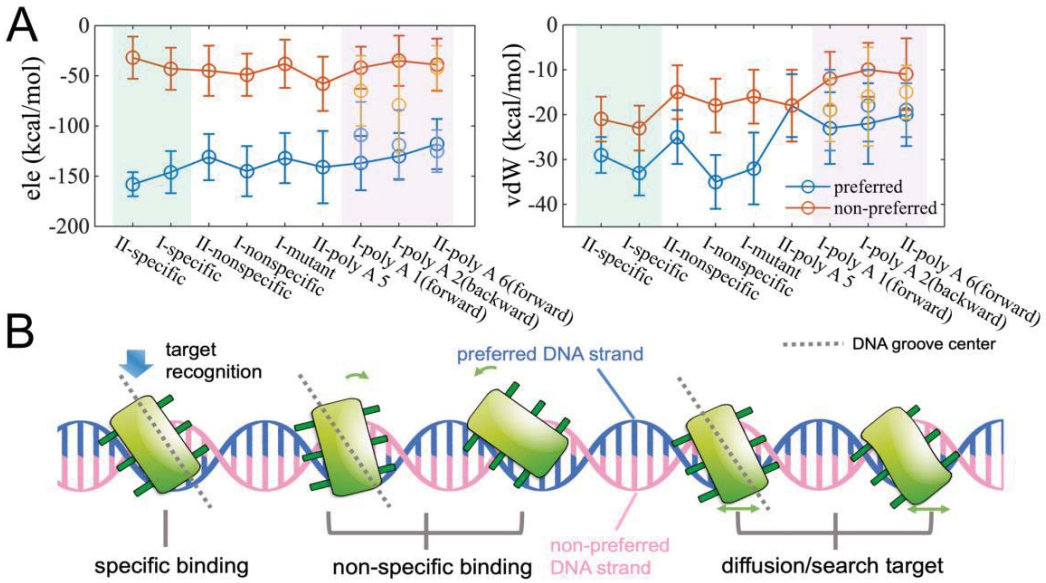
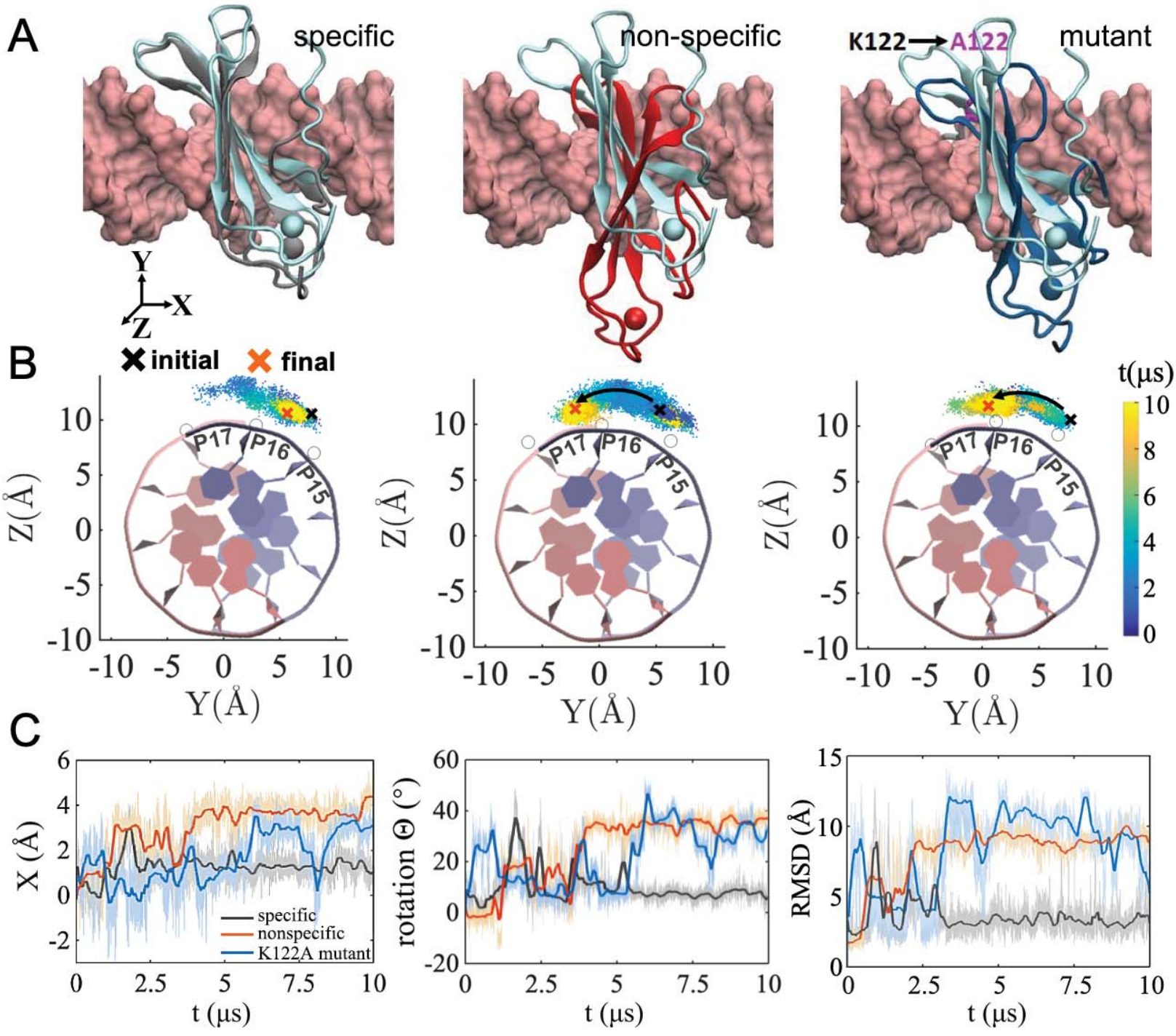
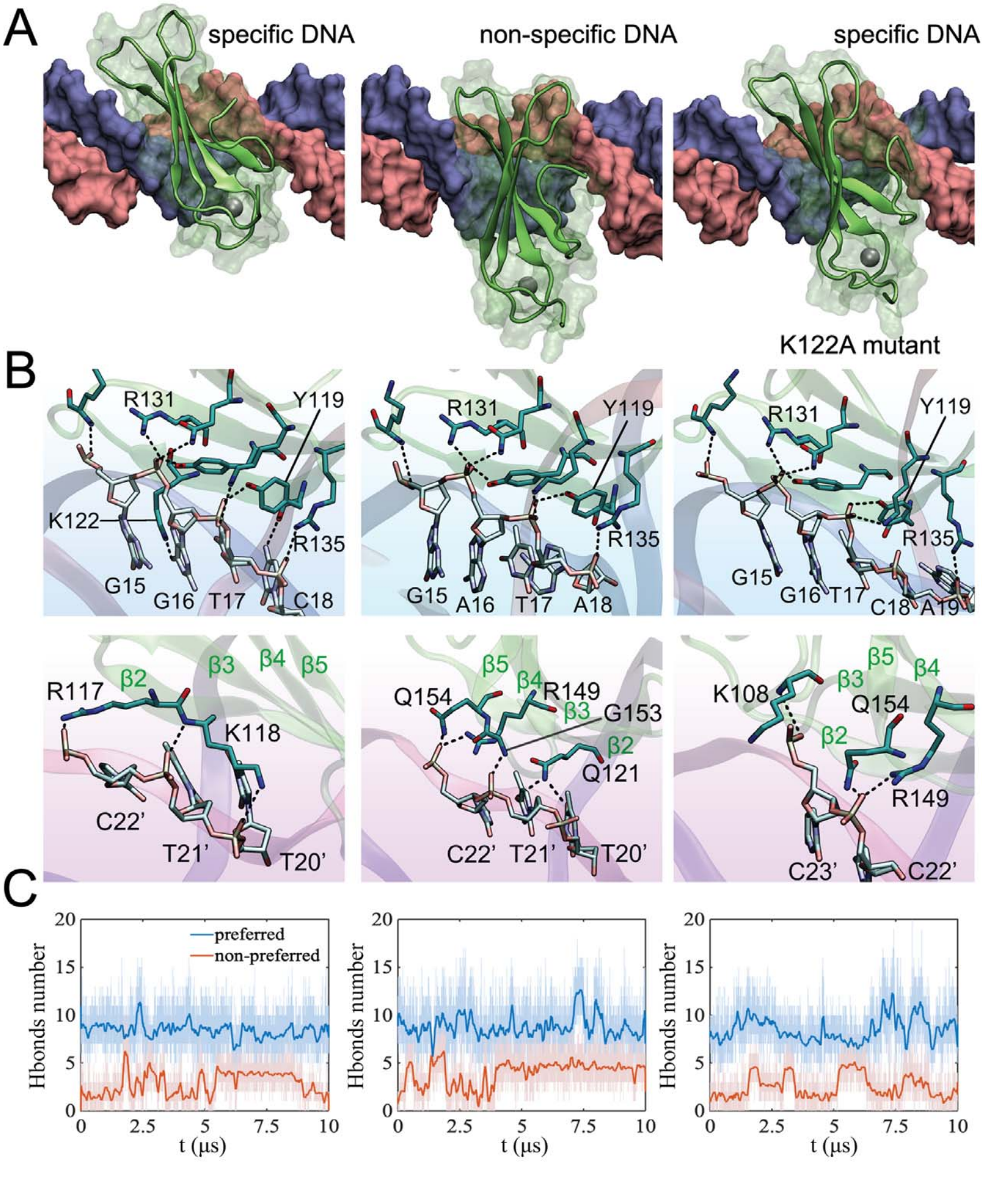
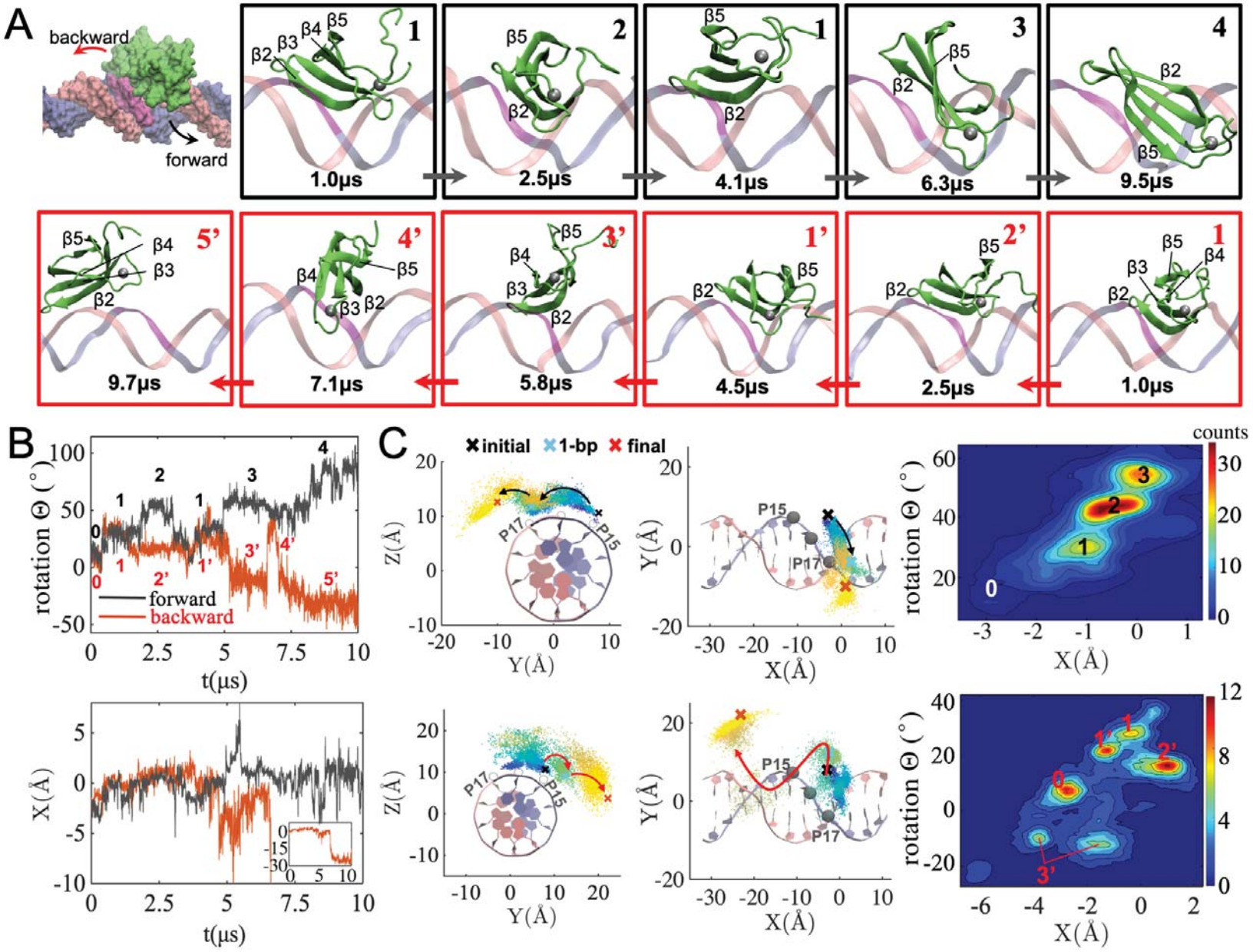
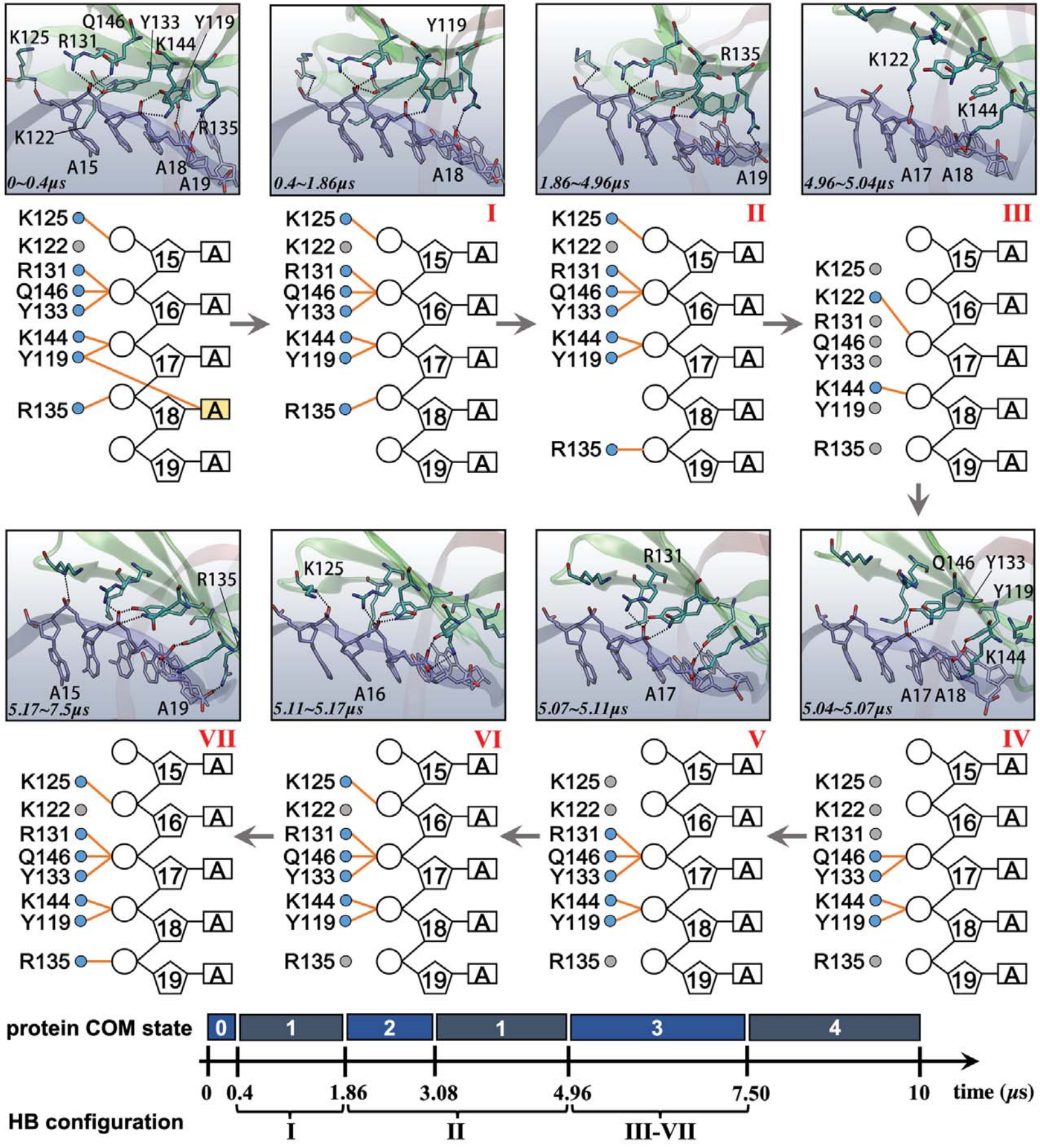


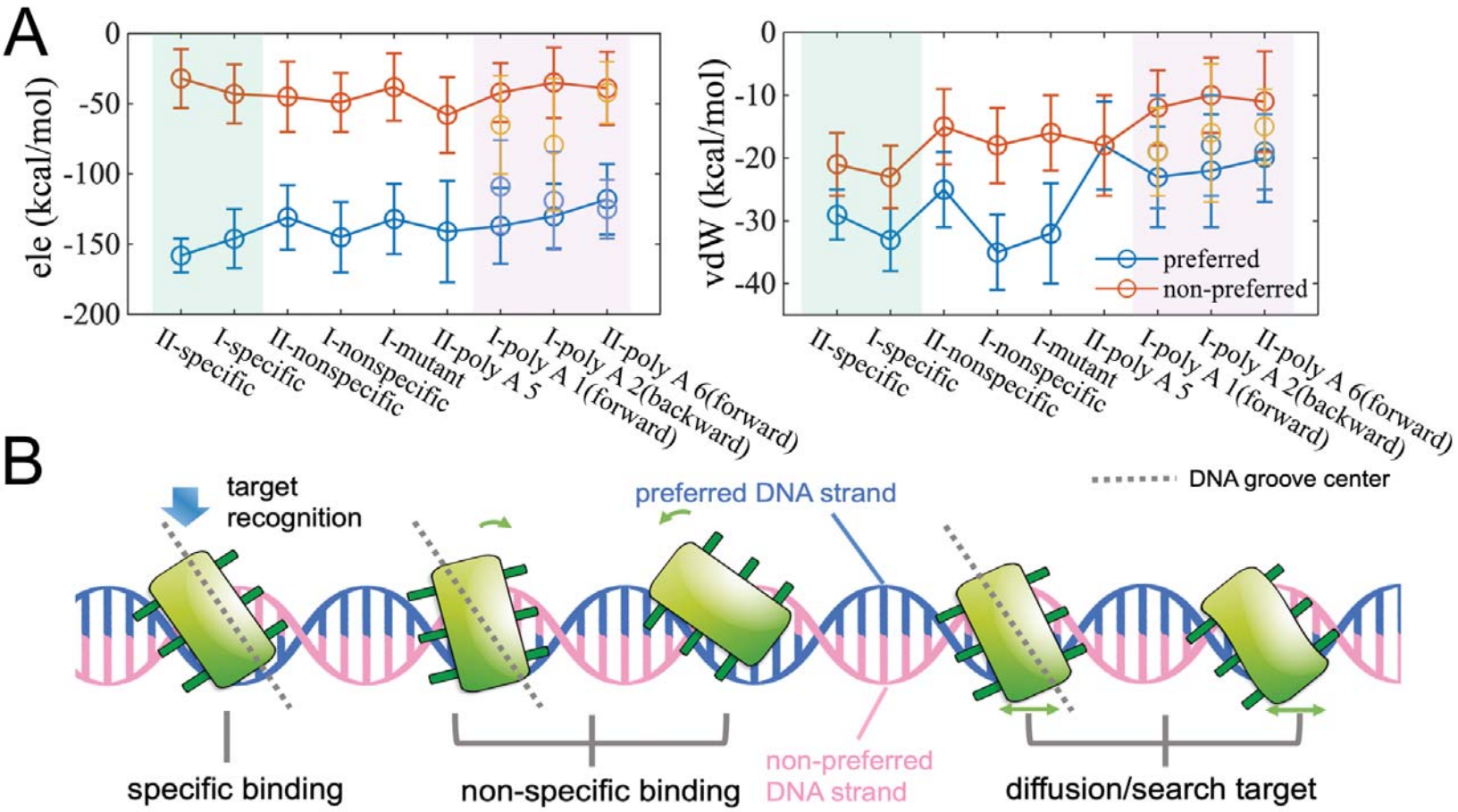
Fig 5. The biased DNA strand association of the WRKY domain protein in diffusion and recognition on DNA. (A) The association energetic between WRKY and two respective DNA strands (preferred in blue, and non-preferred red). The energetics were calculated from all-atom MD simulations from the specific, non-specific DNA binding and K122A systems, and from the forward (poly-A1) and backward (poly-A2) diffusion systems (indicated with prefix I), and additionally also under the updated DNA force field (with prefix II). The specific binding cases are colored in light green background, and diffusive cases in light purple. (B) Schematics on the suggested scenario of the small TF domain protein search and recognition of specific sequence on DNA. The domain protein re-orients constantly during diffusion, following the DNA helical track. The protein diffusional search can be facilitated via modulating the bias and coordination between the two associating DNA strands: with less bias and more coordination between the two strands to assist the protein diffusion, and with more bias and less coordination between the two strands to support DNA sequence recognition of the protein on the preferred DNA strand.


A
specific
non-specific
K122
A122
mutant
Y
X
Z
B
initial
final
t(μs)
10
8
6
4
2
0
P17
P16
P15
Z(Å)
Y(Å)
C
X (Å)
rotation Θ (°)
RMSD (Å)
t (μs)
specific
nonspecific
K122A mutant


A
specific DNA
non-specific DNA
specific DNA
K122A mutant
B
R131
Y119
K122
R135
G15
G16
T17
C18
R131
Y119
R135
G15
A16
T17
A18
R131
Y119
R135
G15
G16
T17
C18
A19
β2
β3
β4
β5
R117
K118
C22'
T21'
T20'
β5
β4
Q154
R149
β3
G153
β2
Q121
C22'
T21'
T20'
β5
β3
β4
K108
Q154
β2
R149
C23'
C22'
C
preferred
non-preferred
Hbonds number
t (μs)
0
2.5
5
7.5
10
15
20


A
backward
forward
β2
β3
β4
β5
1
2
1
3
4
1.0μs
2.5μs
4.1μs
6.3μs
9.5μs
5'
4'
3'
1'
2'
1
9.7μs
7.1μs
5.8μs
4.5μs
2.5μs
1.0μs
B
rotation Θ (°)
t(μs)
forward
backward
X(Å)
C
initial
1-bp
final
Z(Å)
Y(Å)
X(Å)
P15
P17
counts
rotation Θ (°)


Q146
Y133
Y119
K125
R131
K144
K122
R135
A15
A18
A19
0~0.4μs
Y119
A18
0.4~1.86μs
R135
A19
1.86~4.96μs
K122
K144
A17
A18
4.96~5.04μs
I
II
III
K125
K122
R131
Q146
Y133
K144
Y119
R135
15
16
17
18
19
A
R135
A15
5.17~7.5μs
A19
K125
A16
5.11~5.17μs
R131
A17
5.07~5.11μs
Q146
Y133
Y119
K144
A17
A18
5.04~5.07μs
VII
VI
V
IV
protein COM state
0
1
2
1
3
4
0
0.4
1.86
3.08
4.96
7.50
10
time (μs)
HB configuration
I
II
III-VII


A
ele (kcal/mol)
0
-50
-100
-150
-200
vdW (kcal/mol)
0
-10
-20
-30
-40
preferred
non-preferred
II-specific
I-specific
II-nonspecific
I-nonspecific
I-mutant
II-poly A 5
I-poly A 1(forward)
I-poly A 2(backward)
II-poly A 6(forward)
B
target recognition
preferred DNA strand
DNA groove center
non-preferred DNA strand
specific binding
non-specific binding
diffusion/search target