SUPPLEMENTARY DATA

RNA polymerase alpha subunit recognizes the DNA shape of the Upstream Promoter element

Samuel Lara-Gonzalez[1], Ana Carolina Dantas Machado[2], Satyanarayan Rao[2], Andrew A. Napoli[1], Jens Birktoft[1], Rosa Di Felice[2,3,4], Remo Rohs[2,3,5,6,*], Catherine L. Lawson[1,7,*]

[1] Department of Chemistry & Chemical Biology, Rutgers, The State University of New Jersey, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA

[2] Quantitative and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA

[3] Department of Physics and Astronomy, University of Southern California, Los Angeles, CA 90089, USA

[4] CNR-NANO Modena, Via Campi 213/A, 41125 Modena, Italy

[5] Department of Chemistry, University of Southern California, Los Angeles, CA 90089, USA

[6] Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA

[7] Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA

*To whom correspondence should be addressed. Tel: 848-445-5494; Fax: 732-445-4320; Email: cathy.lawson@rutgers.edu

Correspondence may also be addressed. Tel: 213-740-0552; Fax: 213-740-8631; Email: rohs@usc.edu

Present addresses:

Samuel Lara-Gonzalez, IPICYT, Instituto Potosino de Investigación Científica y Tecnológica A.C., División de Biología Molecular, Camino a la Presa San José 2055, Lomas 4a. Sección 78216 San Luis Potosí, S.L.P. México

Ana Carolina Dantas Machado, Department of Medicine, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Satyanarayan Rao, Department of Biochemistry and Molecular Genetics, University of Colorado Denver School of Medicine, Aurora, CO 80045, USA

SUPPLEMENTARY METHODS AND MATERIALS

Molecular Dynamics (MD) simulations

The three crystal structures reported in this study were the starting structures for the MD simulations. Their Protein Data Bank IDs are:

- CAD: PDB ID 3N4M
- ASD: PDB ID 3N97
- CAD_KO: PDB ID 5CIZ

The starting structure of the complex CAD+ASD was obtained by superimposing and merging the CAD and ASD complexes, based on the common structural parts, namely the two αCTDs with the associated DNA sequences. Trajectories for free DNA duplexes taken from these structures are indicated with _DNA *(e.g.* CAD_DNA).

The starting structures for the complexes with mutated αCTDs were obtained from the file with PDB ID 3N4M, by mutating Arg to Ala.

The starting structures for the unbound DNA were taken from the respective complexes.

Using Gromacs, each system was placed at the center of an orthorhombic cell filled with TIP3P water molecules (1), with a water thickness of at least 30 Å between neighboring replicas in order to avoid spurious interactions due to periodic boundary conditions. Sodium ions were added to neutralize the simulation cell.

Each system was then minimized for 5,000 steepest descent steps. Next, it was equilibrated at a temperature of 300 K and a pressure of 1 atmospheric unit with the following four-step protocol:

1. NVT ensemble at 100 K for 200 ps;
2. NVT ensemble at 200 K for 200 ps;
3. NVT ensemble at 300 K for 200 ps;
4. NPT ensemble at 300 K and 1 atm for 2 ns.

The duration of the production run after this preparation was 300 ns, which we used for statistical analysis. The trajectories were saved every 200 ps for the entire system including water molecules and every 1 ps for the ensemble protein+DNA+ions.

Analysis of MD trajectories

We selected one snapshot of the trajectory every 10 ps and determined the shape of the DNA duplex with the software Curves 5.3. Using our own automated procedure packaged in the toolkit Trj2Shape, we efficiently analyzed several snapshots passing the allowed constraints of DNA shape features (2). In particular, for each snapshot Trj2Shape (i) prepares Curves 5.3 acceptable input snapshot by renaming nucleotides; (ii) prepares the input parameter file (crv) for Curves; (iii) runs Curves; (iv) extracts four DNA shape features values (minor groove width (MGW), roll (Roll), helix twist (HelT) and propeller twist (ProT)) from the Curves output; (v) annotates as artifact if shape feature value is beyond the user-defined range (currently MGW range [1.5-12 Å]). Excluding the artifacts, Trj2Shape then calculates the average shape parameter at each base pair (bp) position. Trj2Shape estimates MGW value at each bp by averaging three MGW values at the last level of previous bp, and first two levels of the current bp according to a standard approach

(3,4). We retrieved information only for base pairs at which the MGW could be calculated for more than 90% of the snapshots, which practically excluded 3 bp at each terminus. The Traj2Shape approach is publicly available to users (2).

Clustering of the trajectories was carried out with Gromacs, considering 1 snapshot every 10 ps. The least square fit and root mean square deviation (RMSD) calculations were performed over the DNA's phosphorus atoms. The RMSD cutoff was adjusted to obtain a significant number of structures with significant population. The most populated structures presented later as supporting results accounted for the population percentages reported in Table S1.

Selected representative structures that embody the final time span of the trajectories of CAD and CAD_KO_R265Aα1 were subjected to analysis with DNAproDB (5,6) at https://dnaprodb.usc.edu. DNAproDB is a database and interactive visualization tool that reveals contacts and contact types between protein and DNA in protein–DNA complexes. The database includes data for over 4,500 structures contained in the PDB. The web-visualization tool can upload user structures.

Residue-residue contact maps of complexes were computed using the Gromacs command 'gmx mdmat' that calculates distance matrices consisting of the smallest atom-atom distance between two residues averaged over the input trajectory. The difference contact map between two complexes (Figure 4B) and conversion to an image file was done with the Gromacs command 'gmx xpm2ps'.

Monte Carlo (MC) simulations

We predicted unbound DNA structures of oligonucleotides bound by CAD or CAD_KO using all-atom MC simulations based on a previously described protocol (7,8). Each predicted structure represents the average conformation of 1.5 million MC cycles following an equilibration period of 0.5 million MC cycles. The MC sampling combined collective and internal variables with explicit sodium counter ions and an implicit solvent description (7).

SUPPLEMENTARY TABLES AND FIGURES

Table S1. Clustering analysis for trajectories. Three most representative clusters were retained. A few trajectories are represented by a single cluster, irrespective of the RMSD cutoff. The centroid frame of each cluster was considered the most representative structure in the cluster. In an approximation, the most representative structure of the most populated cluster was also considered the most representative structure of the trajectory. The most representative structures of the first and second most populated clusters were used for the RMSD analysis of bound versus unbound DNA (Table S2). For the contact maps reported in Figure 4, we used the entire cluster identified in the final simulation span (highlighted in yellow). For the DNAproDB analysis shown in Figure 4, we used the centroids of the clusters highlighted in yellow. The RMSD between first and second most representative structures ranges between 1.0 and 2.5 Å.

Structure	**Population (%)**	**Centroid frame (ps)**	**Time span (ps)**
CAD			
r1	43	89,320	7,090–176,820
r2	33	226,480	180,820–300,000
r3	2	291,300	285,160–295,780
ASD			
r1	92	218,420	160–300,000
r2	--	--	--
r3	--	--	--
CAD_KO			
r1	65	127,400	110–268,770
r2	4	287,440	275,170–299,990
r3	1	4,330	2,080–5,540
CAD+ASD			
r1	51	118,450	23,450–175,500
r2	41	236,130	175,510–300,000
r3	6	12,700	6,830–23,430
CAD: DNA only			
r1	30	268,200	171,340–299,940
r2	22	129,670	79,540–183,590
r3	7	51,330	1,550–72,490
ASD: DNA only			
r1	44	231,230	2,400–291,290
r2	--	--	--
r3	--	--	--

CAD KO: DNA only			
r1	93	137,100	0–299,880
r2	--	--	--
r3	--	--	--
CAD+ASD: DNA only			
r1	62	155,660	87,580–298,900
r2	28	83,490	0–-87,500
r3	1	189,950	189,210–190,840
CAD_R265Aα1			
r1	42	82,960	12,440–150,000
r2	39	236,070	181,470–300,000
r3	10	152,380	150,010–180,650
CAD_R265Aα2			
r1	67	203,290	10,020–300,000
r2	0.7	121,340	117,600–123,280
r3	0.4	125,40	10,270–13,030
CAD_KO_R265Aα1			
r1	37	204,800	122,960–243,500
r2	24	94,810	4,520–122,900
r3	17	262,690	243,510–299,990
CAD_KO_R265Aα1_noCAP			
r1	42	215,790	155,860–300,000
r2	35	27,200	980–143,160
r3	3	179,220	151,460–183,230

Table S2. Root mean square deviations (RMSD) between bound and unbound DNA derived from MD simulations. The left two columns indicate the population ranking of the cluster whose most representative structure was considered for RMSD calculation. We calculated the RMSD over all heavy atoms between the bound and unbound DNA, excluding 3 bp at each terminal. The RMSD was computed between the most representative structures of the first and second most populated clusters obtained from the trajectories (Table S1).

Unbound	**Bound**	**RMSD (Å)**
CAD		
1	1	3.5
1	2	3.6
2	2	3.6
2	1	3.6
ASD		
1	1	1.8
CAD_KO		
1	1	3.1
CAD+ASD		
1	1	3.5
1	2	2.9
2	2	2.3
2	1	2.3

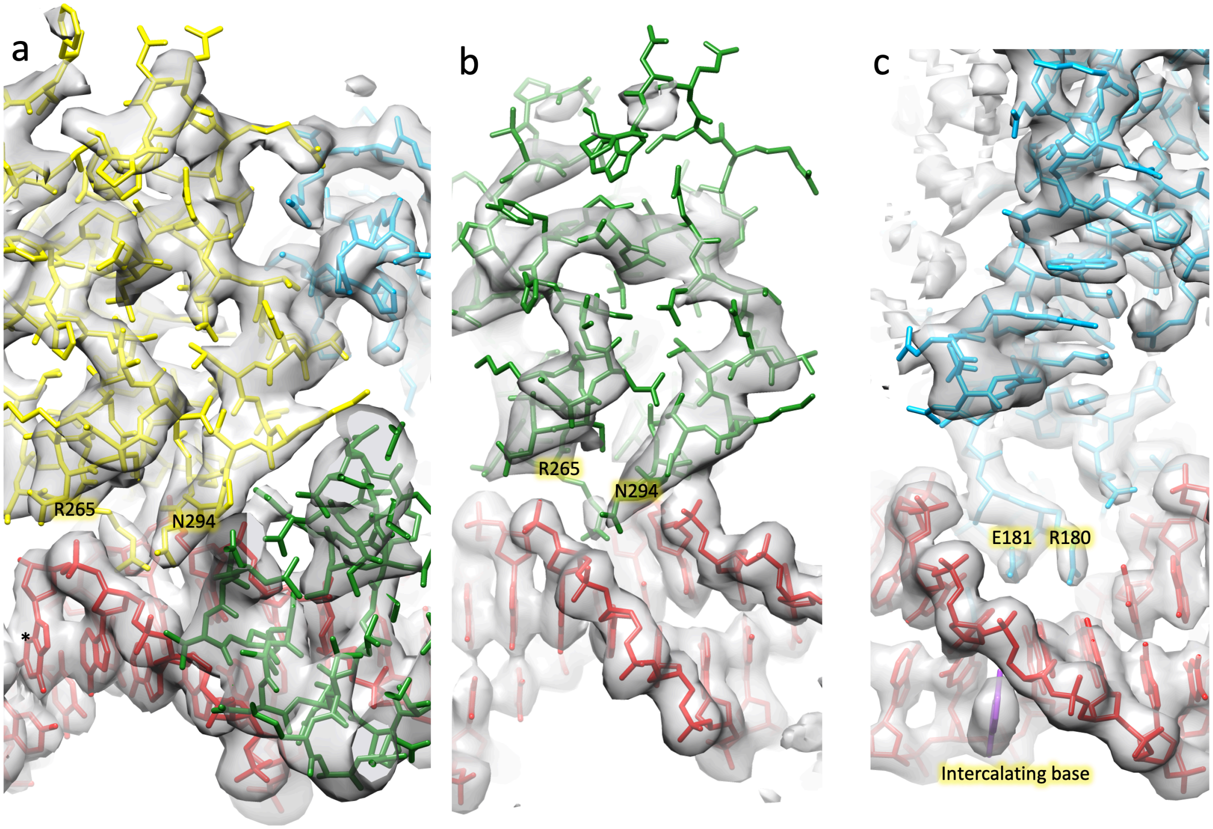

Figure S1. CAD representative electron density. PDB entry 3n4m 2mFo-DFc map is contoured at 1σ; color scheme for atomic coordinates is as described in Figure 1. (a) αCTD[1]-DNA interface. (b) αCTD[2]-DNA interface. (c) CAP-DNA interface; extra density at the position of the DNA major kink, interpreted as a free base (purple).

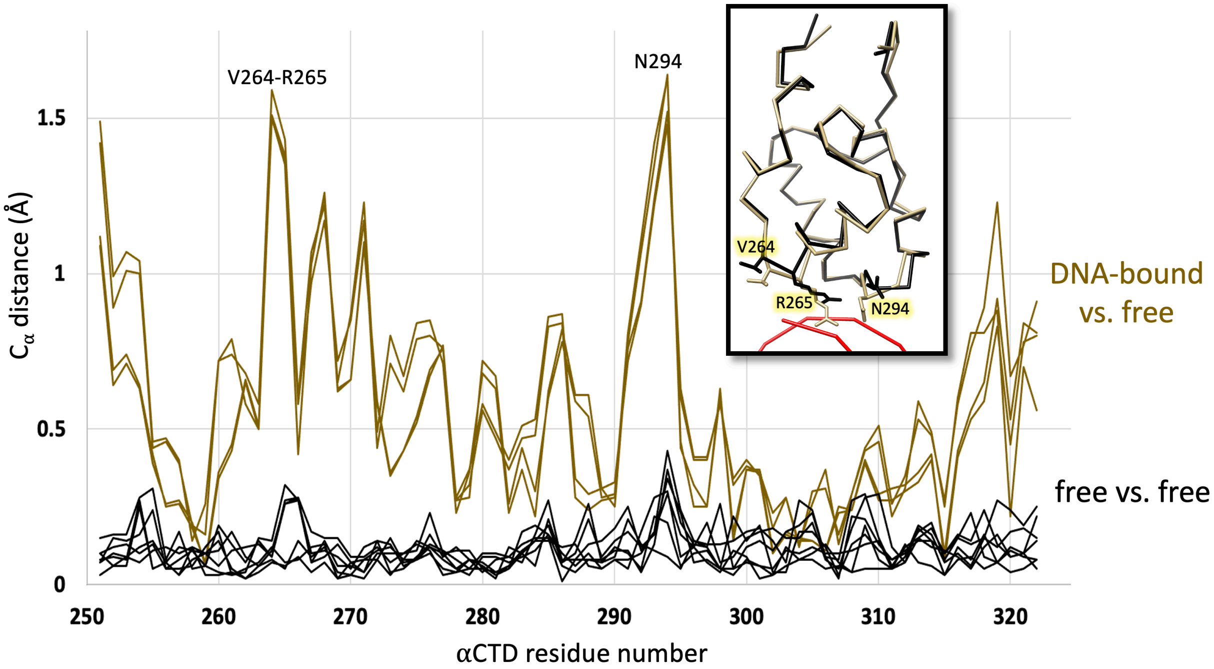

Figure S2. Residual Cα atom distances of DNA-bound vs. free αCTD coordinates following pairwise least-squares fits (free means crystallized without DNA). Gold traces: DNA-bound αCTDs (both αCTDs of the CAD and ASD complexes) vs. free αCTD (PDB ID 3K4G, chain A (9)). Black traces: free αCTDs (3K4G chains B-H) vs. free αCTD (3K4G chain A). Inset: representative Cα trace overlay of bound (gold; CAD αCTD1) vs. free (black; 3K4G chain A). Sidechains are shown for V264, R265, and N294, and a partial phosphate-atom trace is shown for CAD DNA (red).

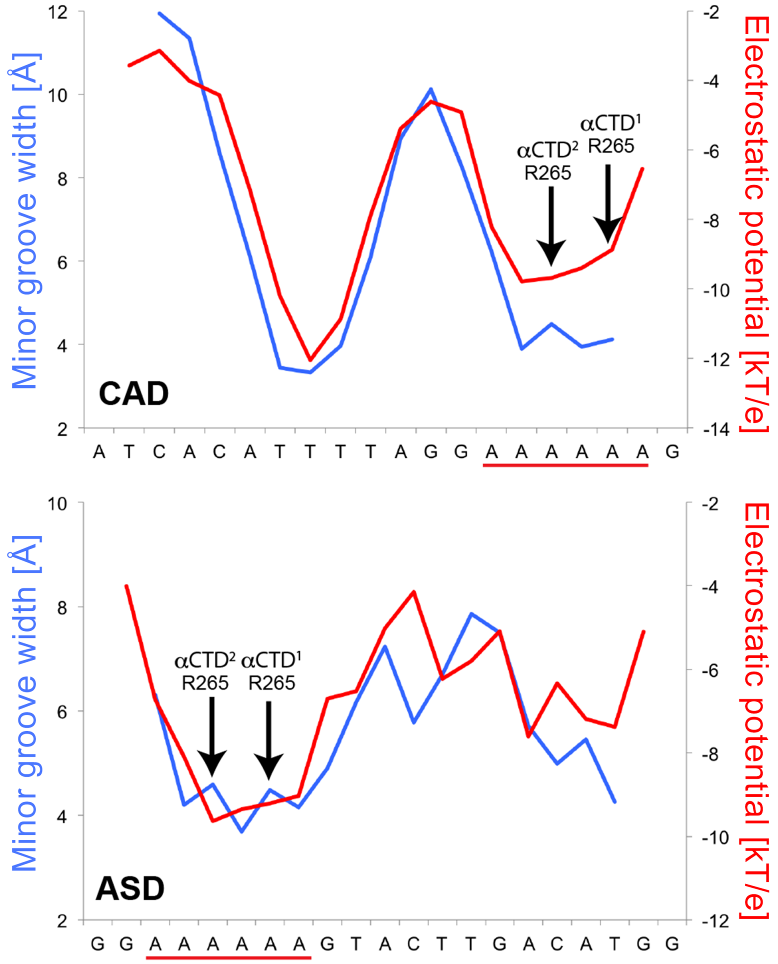

Figure S3. Correlation between minor groove width and electrostatic potential in CAD and ASD co-crystal structures. The A_6-tract (underlined in red) is characterized by a narrow minor groove (blue line plot) and enhanced negative electrostatic potential (red line plot), which strengthens the interactions with the positively charged Arg265 residues of the αCTDs. These plots were evaluated for the DNA duplexes derived from the crystal structures with PDB IDs 3N4M (top panel) and 3N97 (bottom panel). The left-most narrow minor groove region in the top panel is part of the CAP binding site.

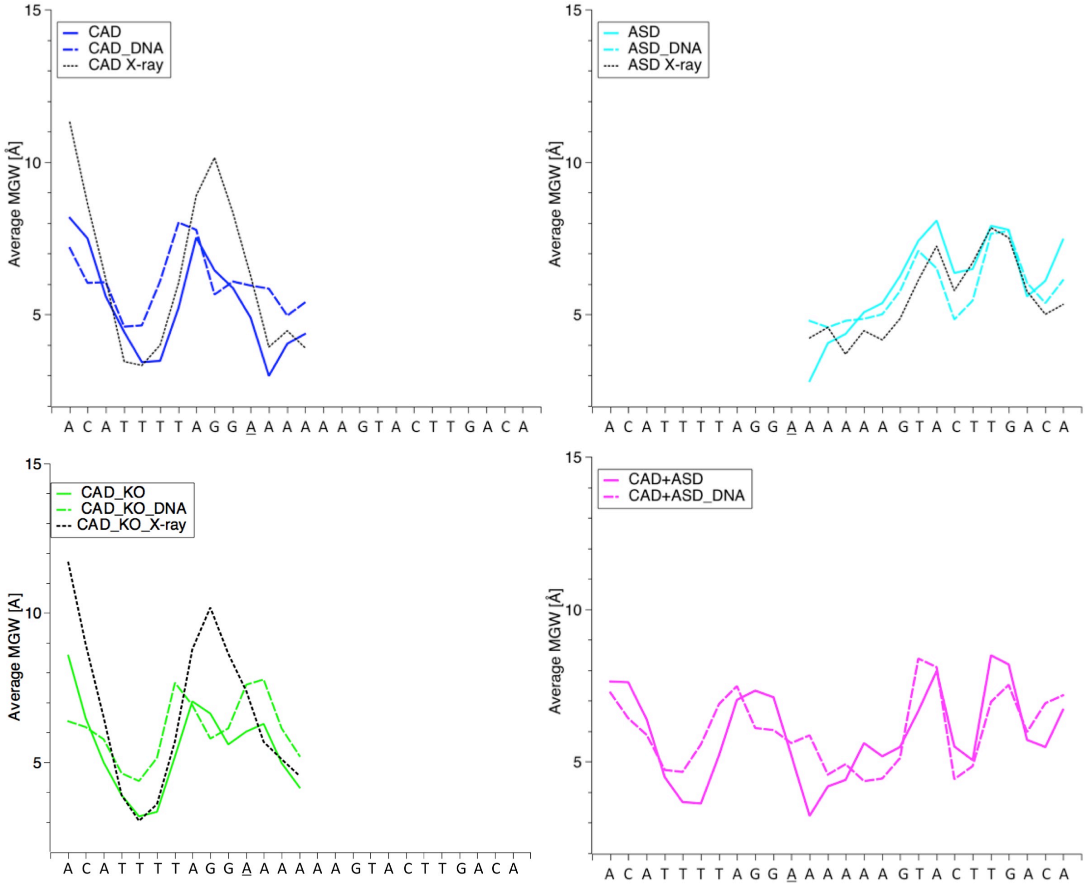

Figure S4. DNA shape of bound versus unbound DNA from MD simulations. The average minor groove width profile of bound and unbound DNA with the same sequence are compared, for bound DNA in the different complexes described in the main text: CAD (top left panel), ASD (top right panel), CAD_KO (bottom left panel), CAD+ASD (bottom right panel). In each panel, minor groove width profiles obtained from MD trajectories of protein-DNA complexes in solution (solid lines) are compared to those obtained for the corresponding unbound DNA duplexes in solution (dashed lines, indicated with '_DNA' in the legends). Although there are differences between the solid and dashed lines in each panel, the trends of minima versus maxima are practically the same, revealing that the DNA itself determines the shape that accommodates the protein. For reference, minor groove width profiles of corresponding X-ray structures are shown as black dotted lines. Our results indicate that the shape of unbound DNA obtained by MD simulation in solution is well conserved in the complexes.

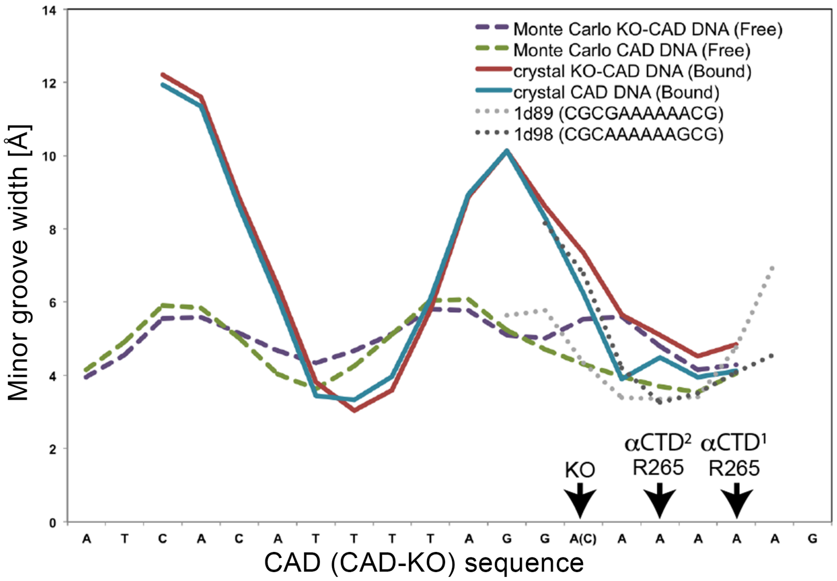

Figure S5. Shortening the A_6-tract widens the minor groove (MC trajectories). Minor groove widths are analyzed based on MC trajectories of unbound DNA in implicit solvent, in comparison to co-crystal structures of protein–DNA complexes and crystal structures of unbound DNA duplexes. The minor groove width of the A-tract region is significantly wider in the CAD-KO complex compared with the CAD complex, both in the simulated structures (purple versus green dashed lines) and in the CAP+αCTD-bound co-crystal structures (red vs. cyan solid lines). Minor groove widths of two free DNA crystal structures are also shown with their A_6-tracts aligned (dotted lines: DNA structures with PDB IDs 1D89 (10) and 1D98 (11)). MC simulations of unbound DNA duplexes taken from the CAD and CAD_KO complexes were performed as described above. The MC results presented here are consistent with MD data shown in Figures 2C and S3. In particular, both MC and MD reveal a smaller variation of the minor groove width across the sequence, as compared to the X-ray trend. Considering that MD and MC simulations are carried out with different force fields, we infer that this underestimation of the variation is not an obvious simulation nor crystal packing artifact. It can be due to statistical smoothing effects, or different behavior of DNA in solution-like simulations vs. the crystalline state (12). The location of the narrowed region of the minor groove width, which is our focus here, is consistently captured by the simulation methods and experimental crystallography.

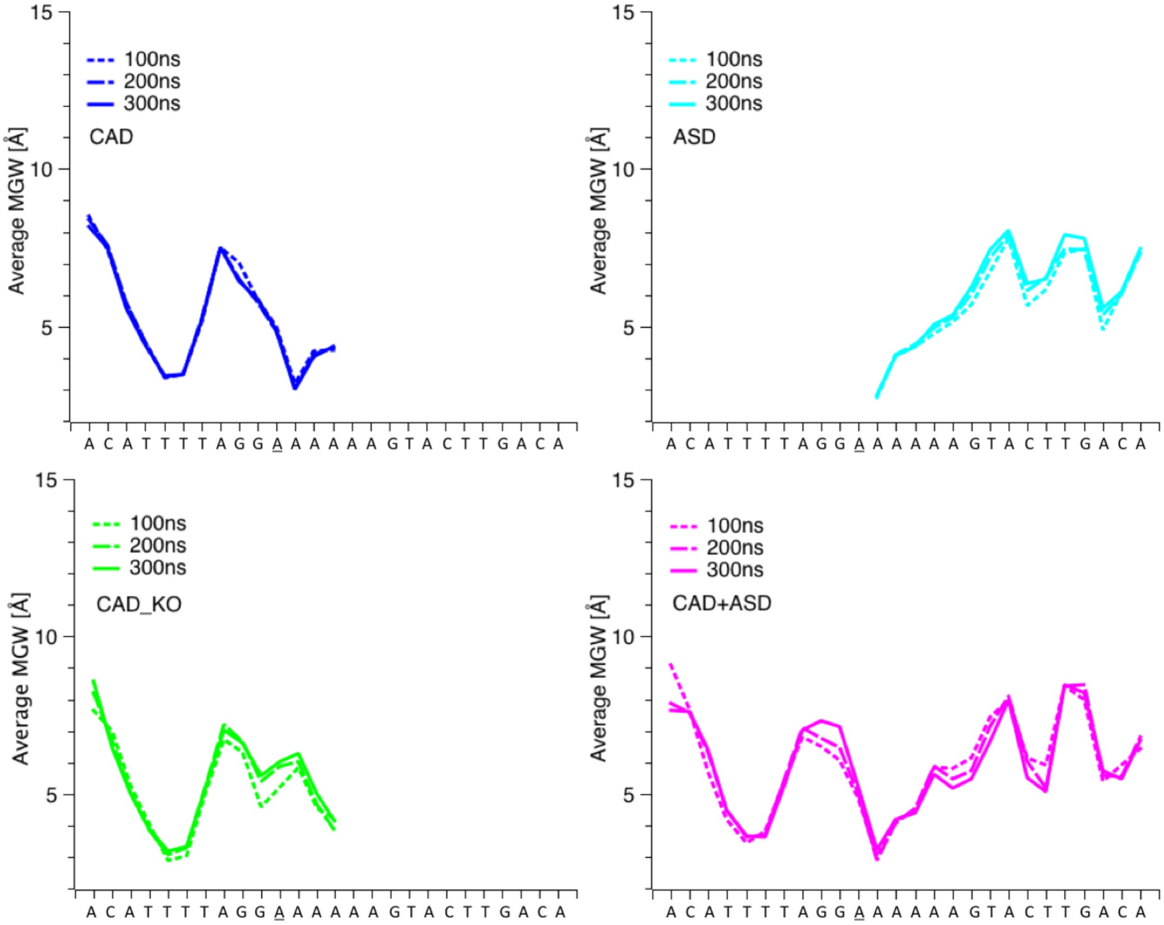

Figure S6. Convergence of MD simulations demonstrates stability of trajectories**.** We focused here on the minor groove width (MGW) as a representative parameter for DNA shape in MD simulations of bound protein–DNA complexes. Figures S6 and S7 plot MGW as a function of nucleotide position averaged over different time intervals within the total simulation time of 300 ns. Minor groove width (MGW) profiles averaged over the first 100 ns (dotted lines), the first 200 ns (dashed lines), and the entire 300 ns duration (solid lines). The underlined position marks the location where the A/C mutation was applied in the CAD_KO system.

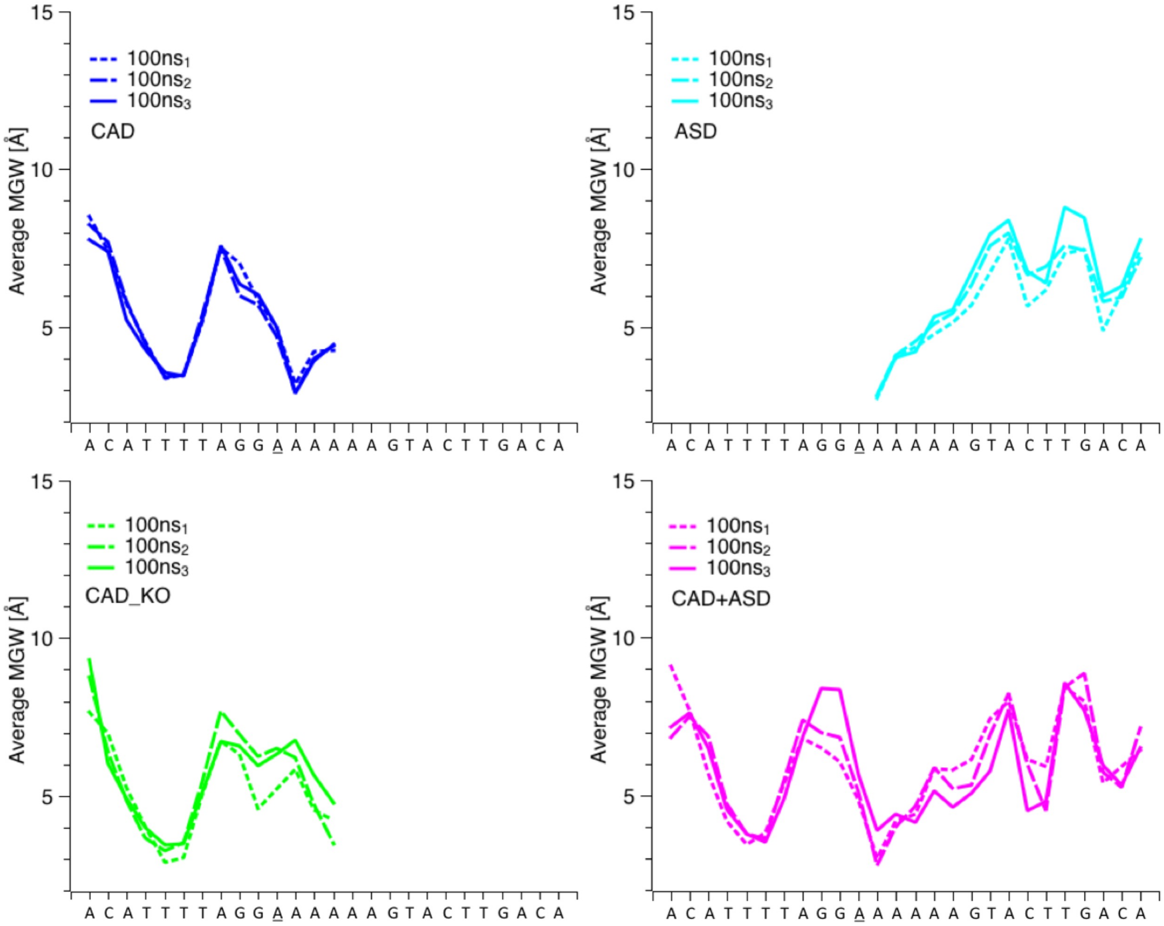

Figure S7. Averaged minor groove width (MGW) profiles. Averaging was performed over 100 ns on the first 1/3 of the simulation (dotted lines), on the second 1/3 of the simulation (dashed lines), and on the third 1/3 of the simulation (solid lines). The underlined position marks the location where the A/C mutation was applied in the CAD_KO system.

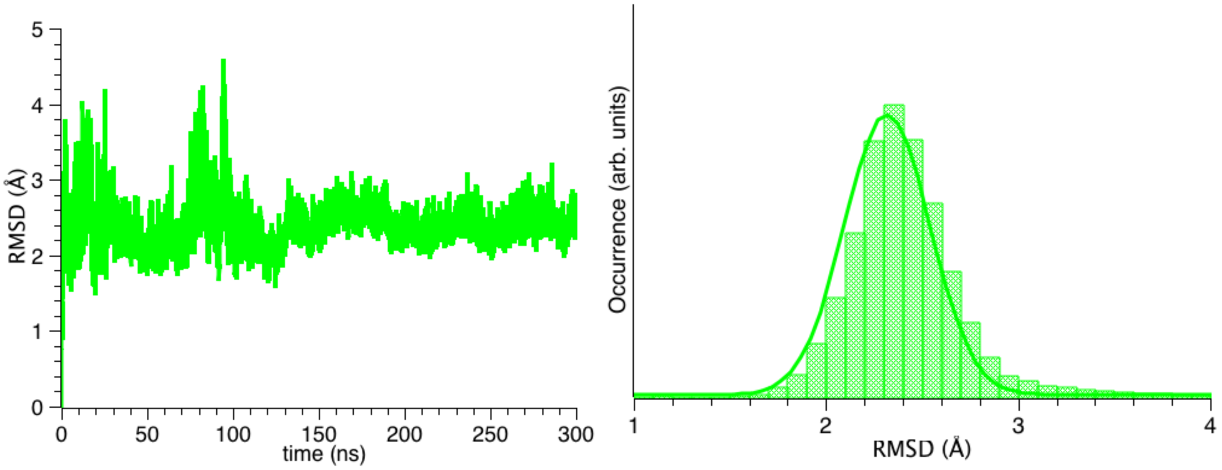

Figure S8. Stability of the CAD_KO complex over 300 ns supports low-resolution crystal structure**.** Time evolution of the RMSD of the CAD_KO complex, computed over the protein and DNA heavy atoms, relative to the equilibrated structure, namely the structure at the end of step 4 of the minimization-equilibration protocol described above (left panel). Histogram of the RMSD and Gaussian fit. (right panel). The fit gives the average RMSD equal to (2.32±0.22) Å, with R^2=0.995. The RMSD data shown here assess a good convergence of this trajectory. Similar trends are obtained for the other trajectories.

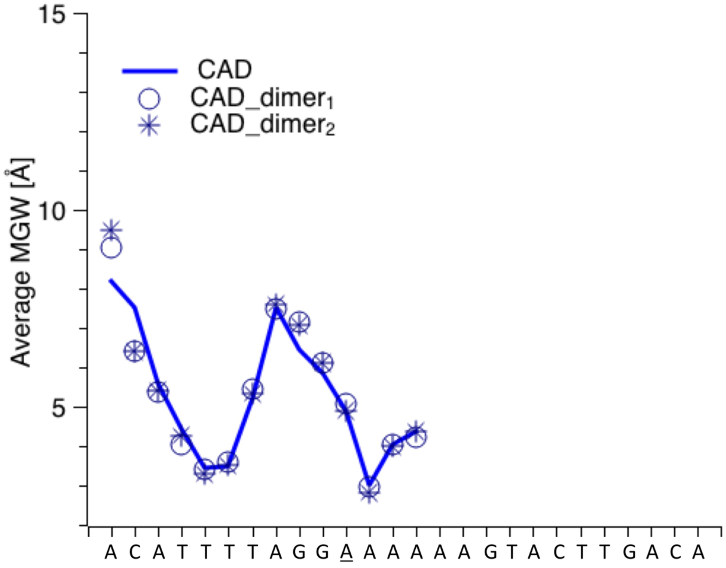

Figure S9. Dimerization does not affect DNA shape in the MD simulation of the CAD complex. Comparison of minor groove width (MGW) for the CAD complex in monomeric form (1/2 of full biological assembly, blue line) and dimeric form (full biological assembly, blue circles and asterisks). Subscripts 1 and 2 identify the two half-assemblies in the CAD dimer. The CAD dimer evolved for 300 ns according to the same simulation protocol as for all other systems in this work. The duplex shape was calculated for 1 snapshot every 20 ps and then averaged over time at each sequence position. The plot clearly indicates that the DNA in both forms has the same shape. The underlined position marks the location where mutation A/C is applied in the CAD_KO system.

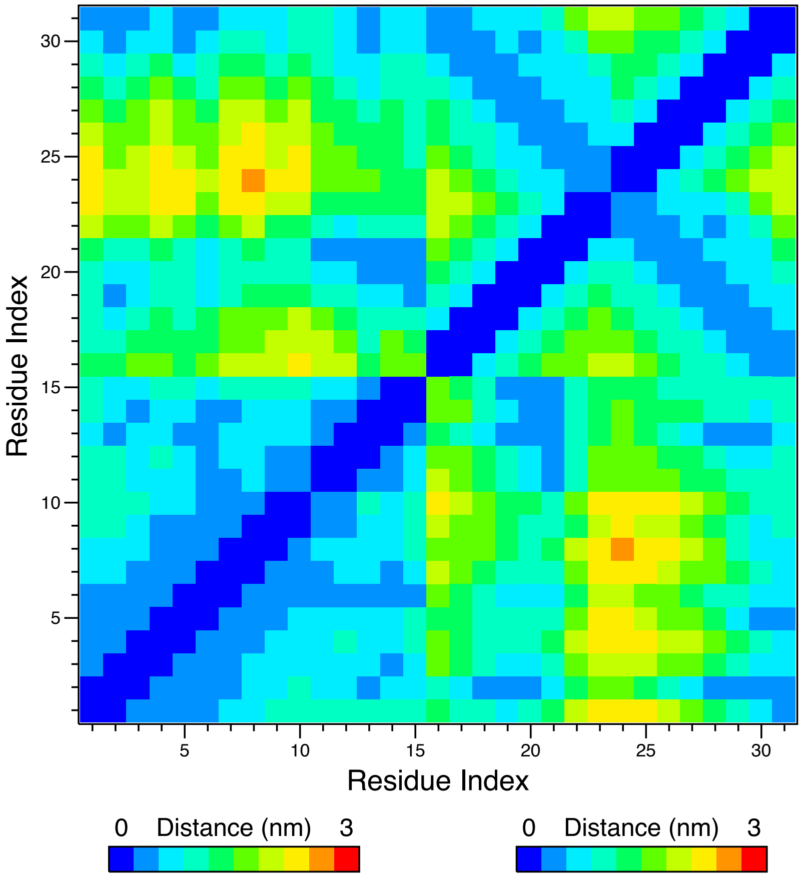

Figure S10. Residue-residue distance contact maps from which the difference contact map of Figure 4B (top panel) was derived. upper left triangle: CAD_R265Aα1, lower right triangle: CAD. The *Distance* in this figure is the effective residue–residue distance (minimum distance over all atom pairs), while the Δ*Distance* in Figure 4B is the relative distance between the mutated protein complex and the native protein complex. Residue indices: binding helix BH (1–10), binding loop BL (11–15), T6+-tract (16–23), A6+-tract (24–31).

SUPPORTING REFERENCES

1. 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.
2. Rao, S., Di Felice, R., and Rohs, R. (2020) Traj2Shape: https://github.com/satyanarayan-rao/Trj2Shape
3. Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Di Felice, R. and Rohs, R. (2013) DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.*, **41**, W56-62.
4. Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R.S., Bussemaker, H.J., Gordan, R. and Rohs, R. (2015) Quantitative Modeling of Transcription Factor Binding Specificities Using DNA Shape. *Proce. Natl. Acad. Sci. USA*, **112**, 4654-4659.
5. Sagendorf, J.M., Berman, H.M. and Rohs, R. (2017) DNAproDB: an interactive tool for structural analysis of DNA-protein complexes. *Nucleic Acids Res.*, **45**, W89-W97.
6. Sagendorf, J.M., Markarian, N., Berman, H.M. and Rohs, R. (2020) DNAproDB: an expanded database and web-based tool for structural analysis of DNA–protein complexes *Nucleic Acids Res.*, **48**, D277-D287.
7. Rohs, R., Sklenar, H. and Shakked, Z. (2005) Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure*, **13**, 1499-1509.
8. Zhang, X.J., Machado, A.C.D., Ding, Y., Chen, Y.H., Lu, Y., Duan, Y.K., Tham, K.W., Chen, L., Rohs, R. and Qin, P.Z. (2014) Conformations of p53 response elements in solution deduced using site-directed spin labeling and Monte Carlo sampling. *Nucleic Acids Res.*, **42**, 2789-2797.
9. Lara-Gonzalez, S., Birktoft, J.J. and Lawson, C.L. (2010) Structure of the Escherichia coli RNA polymerase alpha subunit C-terminal domain. *Acta Cryst. Section D, Biological Crystallography*, **66**, 806-812.
10. Digabriele, A.D. and Steitz, T.A. (1993) A DNA Dodecamer Containing an Adenine Tract Crystallizes in a Unique Lattice and Exhibits a New Bend. *J. Mol. Biol.*, **231**, 1024-1039.
11. Nelson, H.C., Finch, J.T., Luisi, B.F. and Klug, A. (1987) The structure of an oligo(dA).oligo(dT) tract and its biological implications. *Nature*, **330**, 221-226.
12. Azad, R.N., Zafiropoulos, D., Ober, D., Jiang, Y., Chiu, T.P., Sagendorf, J.M., Rohs, R. and Tullius, T.D. (2018) Experimental maps of DNA structure at nucleotide resolution distinguish intrinsic from protein-induced DNA deformations. *Nucleic Acids Res.*, **46**, 2636-2647.