

Supporting Information

Table of Contents

1.	Molecular Dynamics (MD) Simulations	1
2.	System Set Up	2
3.	Sequence Alignment ERGp55-ETV1	9
4.	ERGp55 200 ns RMSD and DNA-Binding plots	10
5.	RMSD Plots ETS Domain DNA-Complex Simulations	11
6.	Binding Distances ETS Domain DNA-Complexes	12
7.	Arg385 & DNA Mutants Binding Distance Plots including WT Repeats	12
8.	Arg367 and Arg370 in the ERG ETS Domain DNA-Complex Crystal Structure and the ERGp55 DNA-Binding Simulation	13
9.	Bridged Hydrogen Bonding Occupancies	14
10.	MM-GBSA Calculation ERGu, mERGu	14
11.	References	15

1. Molecular Dynamics (MD) Simulations

All simulations were performed using the AMBER ff99SB¹ force field parameters for proteins and nucleotides, the GPU accelerated AMBER PME-MD code^{2, 3} and the SPFP precision model⁴ implemented in version 14 of AMBER.⁵ The random seed for initial velocities prior to every simulation was based on the date and time of the simulation.

Implicit Solvent Simulations. Simulations in implicit solvent used a generalized Born solvation model (version igb=5 in AMBER)⁶ implemented in the AMBER software suite⁵ with the corresponding set of atomic radii (mbondi2). The internal dielectric constant was set to $\epsilon_{\text{int}} = 1.0$, the external dielectric constant was set to $\epsilon_{\text{ext}} = 78.5$. An offset of 0.09 Å was chosen in the calculation of effective Born radii. Since implicit solvent simulations were only used for a crude equilibration, the nonpolar solvent term was omitted. Minimization was performed for 1000 steps using 500 steps of steepest descent followed by 500 steps of conjugate gradient. The system was then heated to 325 K in 50 ps using a time step of 0.5 fs, where the temperature increased linearly to 300 K in 30 ps and then again linearly to 325 K in 20 ps. Equilibration was carried out at 325 K with a 1 fs time step. In the heating and equilibration phase bonds to hydrogen were restrained using the SHAKE algorithm.⁷ Temperature regulation was achieved using the weak-coupling algorithm⁸ with a heat coupling constant of 0.5 ps (equilibration) and 1 ps (heating) respectively.

Explicit Solvent Simulations. Minimization in explicit solvent was performed in two steps. First the protein or protein DNA-complex was restrained (harmonic restraints, force constant 10 kcal mol⁻¹ Å⁻²) to relax the solvent molecules for 1000 steps (100 steps of steepest descent followed by 900 steps of conjugate gradient). In the second minimization step the entire system (including solute molecules) was minimized for 5000 steps (500 steps of steepest descent followed by 4500 steps of conjugate gradient). The non-bonded cutoff was set to 12 Å. Subsequently the system was heated to 310 K in a two-step process. In the first step the system was heated to 100 K with a linear increase of temperature in 62.5 ps in the constant volume ensemble (NVE ensemble). In the second step the system was heated to 310 K in the constant pressure ensemble, where during the first 300 ps the temperature increased linearly to 310 K, which was followed by 200 ps for equilibration in the NPT ensemble with isotropic pressure scaling. During the heating process weak restraints (force constant 10 kcal mol⁻¹ Å⁻²) were imposed on the solute molecules. Finally equilibration was run at 310 K using constant pressure and temperature MD (NPT ensemble, with isotropic pressure scaling). The time step during heating and equilibration was 2.5 fs, hydrogen bonds were restrained using the SHAKE algorithm.⁷ Periodic boundary conditions were imposed using the particle mesh Ewald (PME)² method for long-range electrostatics using a cubic spline approximation with a cutoff for the calculation of nonbonded interactions of 8.0 Å. Temperature regulation was achieved using Langevin dynamics with a collision frequency of 1 ps⁻¹. The pressure was regulated using the Monte Carlo barostat (reference pressure 1 bar, frequency of volume change attempts 80 or 100 steps).

2. System Set Up

Since there is no complete experimental structure for ERG, all models were built using the available crystal structures for the ERG ETS domain (PDB-IDs: 4irg, 4iri, 4irh)⁹ and the NMR structure for the PNT domain (PDB-ID: 1sxe,¹⁰ first structure of the NMR-ensemble).

For the systems containing regions without known structure, missing sequences were inserted using a peptide chain generated with LEaP, as implemented in AMBER Tools 14⁵ and combined using the model build function of UCSF chimera.¹¹

ERGp55 Folding Simulation and Generation of the ERGp55 Model

Protein Sequence (Uniprot identifier: P11308-4):^{12, 13}

10	20	30	40	50
MASTIKEALS	VVSEDQSLFE	CAYGTPHLAK	TEMTASSSSD	YGQTSKMSPR
60	70	80	90	100
VPQQDWLSQP	PARVTIKMEC	NPSQVNGSRN	SPDECSVAKG	GKMVGSPDTV
110	120	130	140	150
GMNYGSYMEE	KHMPPNMTT	NERRVIVPAD	PTLWSTDHVR	QWLEWAVKEY
160	170	180	190	200
GLPDVNILLF	QNIDGKELCK	MTKDDFQRLT	PSYNADILLS	HLHYLRETPL
210	220	230	240	250
PHLTSDDVDK	ALQNSPRLMH	ARNTGGAAFI	FPNTSVYPEA	TQRITTRPDL
260	270	280	290	300
PYEPPRRSAW	TGHGHPTPOS	KAAQPSPSTV	PKTEDQRPQL	DPYQILGPTS
310	320	330	340	350
SRLANPGSGQ	IQLWQFLEL	LSDSSNSSCI	TWEGTNGEFK	MTDPDEVARR
360	370	380	390	400
WGERKSKPNM	NYDKLSRALR	YYYDKNIMTK	VHGKRYAYKF	DFHGIAQALQ
410	420	430	440	450
PHPPESLYK	YPSDLPYMG	YHAHPQKMN	VAPHPALPV	TSSSFFAAPN
460	470			
PYWNSPTGGI	YPNTRLPTSH	MPSHLGTY		

The starting structure for the apo ERGp55 full-length simulation in explicit solvent was generated by equilibrating the model built structure (built from the ETS domain and PNT structures, PDB-IDs: 4irh,⁹ 1sxe,¹⁰ plus inserted peptide chain for the remaining protein sequence, see Figure S1) in implicit solvent with harmonic restraints on parts of the structure to maintain the shape of the ETS and PNT domain. The idea behind that was to minimize the size of the structure in a physical manner to then be able to run equilibration in explicit solvent at a considerable speed.

Table S1 - Simulation log for the ERG3+7b full-length system ^a

	Starting Structure	Solvent Type	Restrains [residue no.]	Box size [Å]	Simulation Time [ns]
1	see Figure S1 (a)	Implicit	105-201 PNT domain 294-400 ETS domain	N. A.	100
2	end of 1	Implicit	105-201 PNT domain	N. A.	15
3	end of 2	Explicit	None	8	150
4	end of 3	Explicit	None	12	150
5	end of 4	Explicit	None	20	67
6	end of 5	Explicit	None	25	433

^a Solvent type: 'Implicit' refers to the use of a generalized Born solvent model (see Methods section in main text), 'explicit' implies simulation in TIP3P water. Restrains: Use of harmonic restraints (force constant during heating simulation 10000 kcal mol⁻¹ Å⁻², during equilibration simulation 10 kcal mol⁻¹ Å⁻²) to Cartesian input coordinates to restrain certain parts of the system. Box size: Minimum distance of a solute molecule to the surface of the box of TIP3P water.

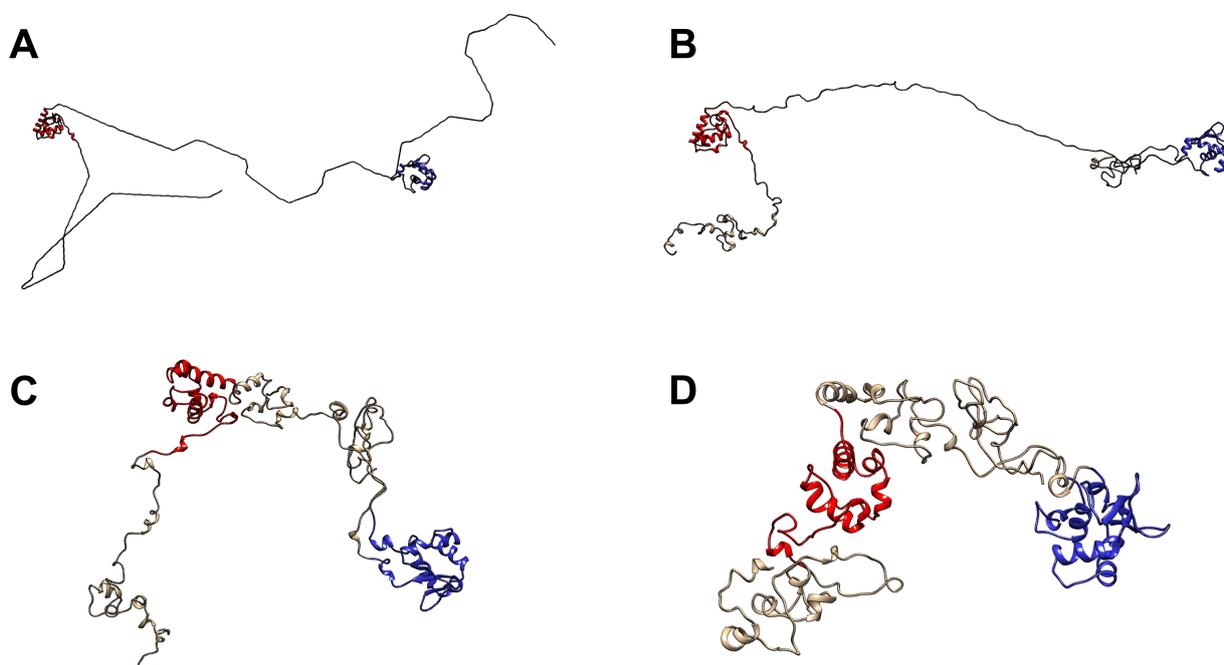


Figure S1 Picture (a) shows the starting structure of the ERGp55 simulation, with the ETS and PNT domain taken from the available structures^{9, 10} and the remaining parts inserted as a linear peptide chain. Picture (b) shows the structure after 100 ns of equilibration in implicit solvent with restraints on the ETS and PNT domain. Picture (c) shows the structure from (b) further equilibrated in implicit solvent with restraints only on the PNT-domain. Picture (d) shows the representative frame of the most populated cluster, obtained from analyzing the equilibration trajectory of structure (c) in explicit solvent (simulation length: 800 ns, snap shot extracted after 555 ns of simulation in explicit solvent). The ETS domain is shown in red, the PNT domain is shown in blue.

The equilibration in explicit solvent was restarted with different box dimensions to ensure decent simulation speed and at the same time avoid self interactions of the protein due to periodic boundary conditions. During the molecular dynamics (MD) run with restraints on the ETS and PNT domain (run 1 in Table S1) the N- and C-terminus folded up in a mostly random coil structure (see Figure S1 (b)). The sequence linking the ETS and PNT domain did not relax due to the applied restraints. During MD run 2 the link between ETS and PNT domain also folded up in a mostly random coil structure, due to the fact that only the PNT domain was left restrained. The dimensions of the system were now feasible to carry out equilibration in explicit solvent. A representative structure (referred to as topERGp55) of the most populated cluster of 800 ns of equilibration in explicit solvent was obtained by using a cluster algorithm implemented in UCSF chimera (step size 128,

yielding 500 structures for analysis).^{11, 14} The snapshot is shown in Figure S1 (d). ERGp55 adopted an overall bent shape. Regions outside the ETS and PNT domain adopted mainly random coil structures with elements of secondary structure (most prominent the α -helix C-terminal to the PNT domain). The exact details of the folding simulation and the native state of ERGp55 however are not subject of this study. Indicated by the fact that no crystal structure was solved for the full-length protein and a circular dichroism study,¹⁵ ERG probably consists of big parts with high flexibility and no secondary structure. The main purpose of the equilibration of ERGp55 was to come up with model structures that deliver more information than the ETS crystal structure alone and maybe describe the disordered state of ERG to some extent. The model structure ERGp55 was then used for the full-length DNA-binding simulations.

Clustering and Analysis:

In order to get a better understanding of the dynamics of the system, we performed clustering analysis as implemented in UCSF Chimera^{11, 14} on the trajectory in order to identify whether the system occupies certain metastable states. We initially performed a clustering analysis on the trajectory after 400 ns (step size 128, yielding 250 structures for analysis) of simulation in explicit solvent earlier in the project (see Figure S2) to be able to extract a structure for the full-length ERG DNA-binding simulations. We decided to use the representative frame of the second most populated cluster (frame 26401 corresponding to 330 ns of simulation in explicit solvent) from the clustering analysis of the 400 ns trajectory, because Tyr371 occupies a special configuration in this frame and we were initially also interested in the behavior of this residue during the DNA binding process, which was proposed to be related to the autoinhibitory regulation of ERG.⁹

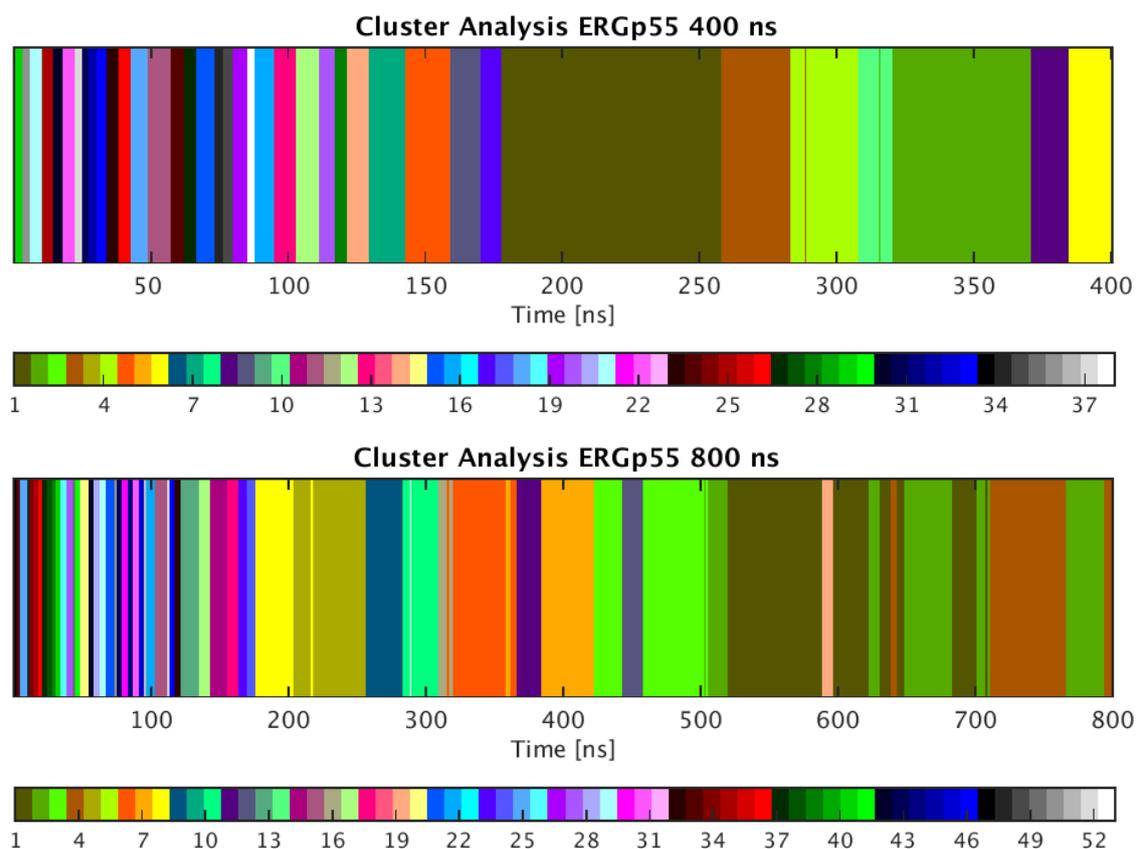


Figure S2 Cluster Analysis plots showing the correspondence of frames throughout the simulations to a cluster. Each color indicates a distinct cluster. The clusters are numbered with decreasing size.

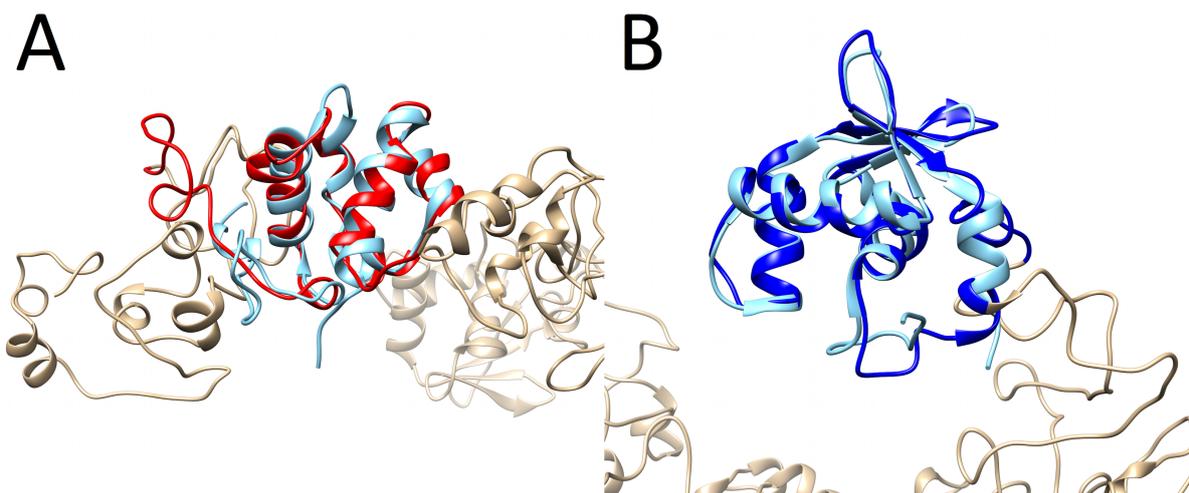


Figure S3 (A) shows the overlap of the experimental PNT domain (PDB-ID: 1sxe¹⁰, first structure of the NMR ensemble) to the PNT region of ERGp55. (B) shows the overlap of the experimental Ets domain (PDB-ID: 4irh⁹) to the Ets region of the ERGp55 structure. The PNT domain is labeled in red and the Ets domain is labeled in blue. The experimental structures are shown in cyan.

The structure is simply named ERGp55 in the main text. The overlap of the ETS and PNT-domain of ERGp55 to the initial model can be seen in Figure S3. The backbone RMSD of the ETS region (amino acids 300-401) to the initial crystal structure (PDB-ID: 4irh⁹) is 3.7 Å, truncating the calculation frame to amino acids 311-391 reduces the RMSD to 1.5 Å. The backbone RMSD of the PNT region (amino acids 105-201) to the PNT NMR-structure (PDB-ID: 1sxe,¹⁰ first structure of the ensemble) is 7.3 Å, truncating the calculation frame to amino acids 129-199 also reduces the RMSD to 2.5 Å. This shows that mainly the terminal regions of the input structures changed their conformation during the equilibration simulation.

We continued the simulation to a length of 800 ns in explicit solvent, to see whether we could obtain any meaningful results from the ERGp55 apo simulation. However as shown in Figure S2 the clustering profile is very diverse. RMSD plots to the first frame and the average structure are shown in Figure S4. The RMSD starts to equilibrate at around 300 ns, we do however not claim that the ERGp55 model nor the simulation is fully equilibrated. Other means of analysis including Markov-State-Model analysis, Cross-Correlation analysis etc. also did not yield any meaningful insights. Indicated by the fact that no crystal structure exist for full length ERG and a circular dichroism study¹⁵ ERG consists of large regions exhibiting random coil structures. This and the lack of suited experimental data makes validation of just the apo ERGp55 model very difficult. Therefore we decided to focus fully on the DNA binding simulations using the model generated as mentioned above for full-length ERGp55 with the rationale that including modelled parts of the protein additionally to ETS and PNT domain without known structures should add information that would be lost by simply using the ETS domain crystal structures.

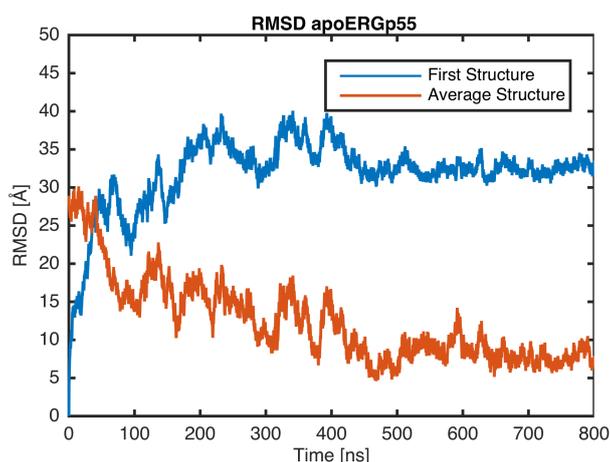


Figure S4 Backbone RMSD analysis of the simulation of ERGp55 in explicit solvent. As a reference both the first frame and the average structure of the trajectory were used. Reference atoms were the CA, C, N and O atoms.

To make sure that the DNA-binding simulations are representative of the real system, we performed extensive validation of the unbiased DNA-binding simulation to experimental data and other simulations using only available structural data, which can be found in the main text. While the procedure used to generate the ERGp55 model, was developed specifically for this purpose, a full length model of ERG was generated and used in simulations before¹⁵. Moreover including regions without known structures into a computational model has also been performed for other system, e.g. to study the dynamics of troponin, which lacks a complete crystal structure as well.¹⁶

Full-Length ERGp55 DNA-Binding Simulation. For the binding simulation of full-length ERGp55, B-DNA nucleic acid structures were generated using NAB¹⁷ as implemented in AMBER Tools 14.⁵ DNA sequences were taken from the -181 ETS binding site (EBS) in the ICAM-1 promoter.^{18, 19} The DNA molecules were 49 base pairs (bp) in length to avoid interaction of the protein with the ends of the DNA-molecule and to be able to show the whole interaction surface between protein and DNA. ERGp55 full-length DNA binding simulations were set up by placing a model built B-DNA structure at a distance (defined as the center of mass of the α 3-helix of the ETS domain to the center of mass of the GGAA recognition sequence) of ~ 23 Å to the full-length ERGp55 protein model (constructed as described above). The NaCl concentration was set to 150 mM by replacing random water molecules with the corresponding amount of ions. The systems were built in UCSF chimera¹¹ and system preparation was performed in LEaP.⁵

ETS Domain DNA-Complex Simulation. The simulation of the ERGdna + DNA system was set up simply using the PDB file of the ERG ETS-domain DNA-complex crystal structure (PDB-ID: 4iri)⁹ as input coordinates. The simulations starting with a different ETS domain construct bound to DNA used the DNA-fragment of the ERG ETS-domain DNA-complex crystal structure (DNA-sequence shown below) and either another ERG ETS domain crystal structure (ERGu and ERGi, with PDB-IDs: 4irg, 4irh⁹ respectively) or a model generated from a modified crystal structure (ERGumax and ERGimax) superimposed onto the ETS domain in the original ETS domain DNA-complex (PDB-ID: 4iri).⁹ ERGumax was generated by deleting the relevant residues in ERGu (PDB-ID: 4irg).⁹ To see whether the crystal structures (ERGi and ERGu, PDB-IDs 4irh and 4irg⁹) and deletion structures (ERGumax)

are stable, 500 ns simulation were performed to analyse the RMSD to the first structure and the average structure during the 500 ns trajectory (see Figure S5). The simulation setup was the same as for the ETS domain DNA-complex simulations (see below), but lacking the DNA structure. Since the average RMSD to the first frame during the 500 ns trajectory did not exceed 3 Å for any of the three systems and the plots of the RMSD to the first frame compared to the RMSD to the average structure were highly similar, it was decided to use the input structures of the apo simulations for the ETS domain DNA-complex simulations. The simulation of apo ERGu was elongated to 700 ns, because of the RMSD peak at the end of the 500 ns period. Analyzing the trajectory revealed that the peak was caused by movements of the N- and C-terminal regions of the ERGu construct, which then remained stable again for 100 ns at the end of the simulation. This movement might be interesting in the context of the autoinhibition of ERG⁹, which was proposed to be related to movements in this region. This is however not subject of this study, which is why no further investigations were carried out in that direction. For details about ERGimax and sequence information of the different ETS constructs see Figure S6 and the next paragraph.

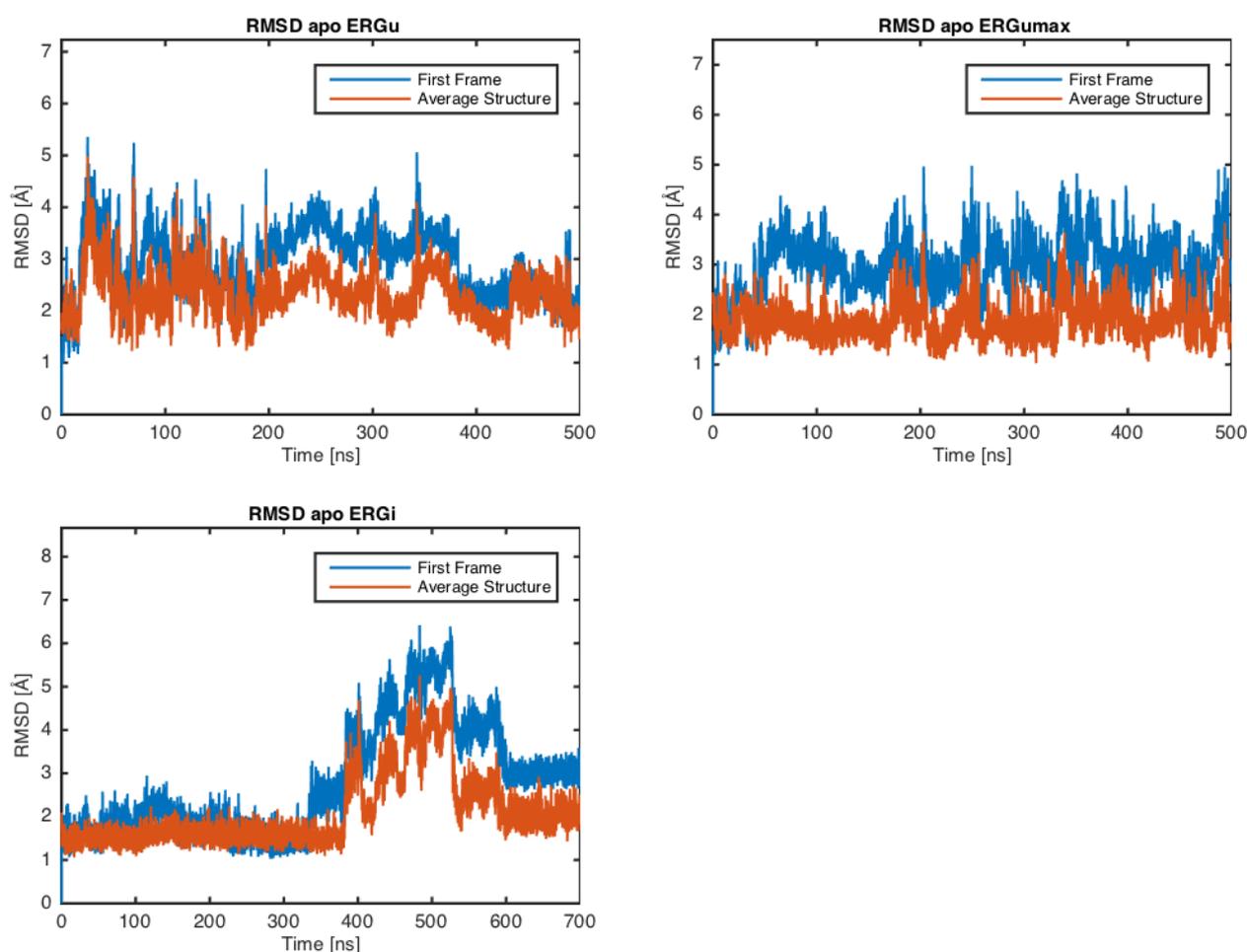


Figure S5 Backbone RMSD plots of the apo Ets domain simulations using the first frame and the average structure of the simulation as a reference. Reference atoms were the CA, C, N and O atoms.

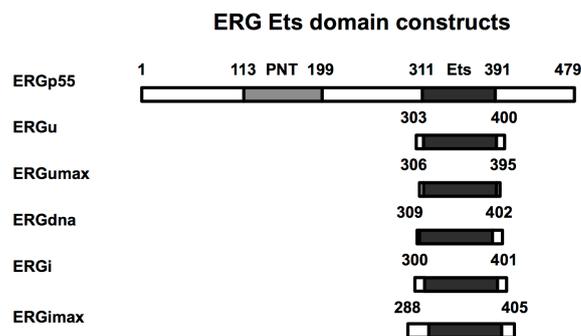


Figure S6 Scheme showing the different ETS domain constructs and the ERGp55 full length protein as used in molecular dynamics simulations.

All systems were solvated in a rectangular box of TIP3P water molecules including counter ions to neutralize the charge with a minimum distance of 10.0 Å of solute molecules to the box boundaries. The NaCl concentration was adjusted to 150 mM by placing the according amount of ions randomly in the box. The ETS domains were aligned in the original complex in UCSF chimera¹¹ and system preparation was performed in LEaP.⁵

DNA-sequence:

5' -GACCGGAAGTGG-3'⁹

Generation of ERGimax Model

ERGimax was generated by adding the relevant amino acids as a linear peptide chain to ERGi (PDB-ID: 4irh)⁹ according to the protein sequence in Regan et al.⁹ The ERGimax model contains one amino acid more than the construct with maximal inhibition in the Regan paper,⁹ which was added by accident, however we assume that the amino acid at the N-terminus will not interfere drastically with the function of ERGimax. The ERGimax simulation was run for 1 μs in explicit solvent. The ERGimax model used in the ETS domain DNA-complex simulation was generated using the cluster analysis tool implemented in UCSF chimera.^{11, 14} Clustering was based on 606 structures extracted throughout the simulation and the representative frame of the most populated cluster (structure obtained after 584 ns) was then used for the ERGimax DNA-complex simulation.

3. Sequence Alignment ERGp55-ETV1

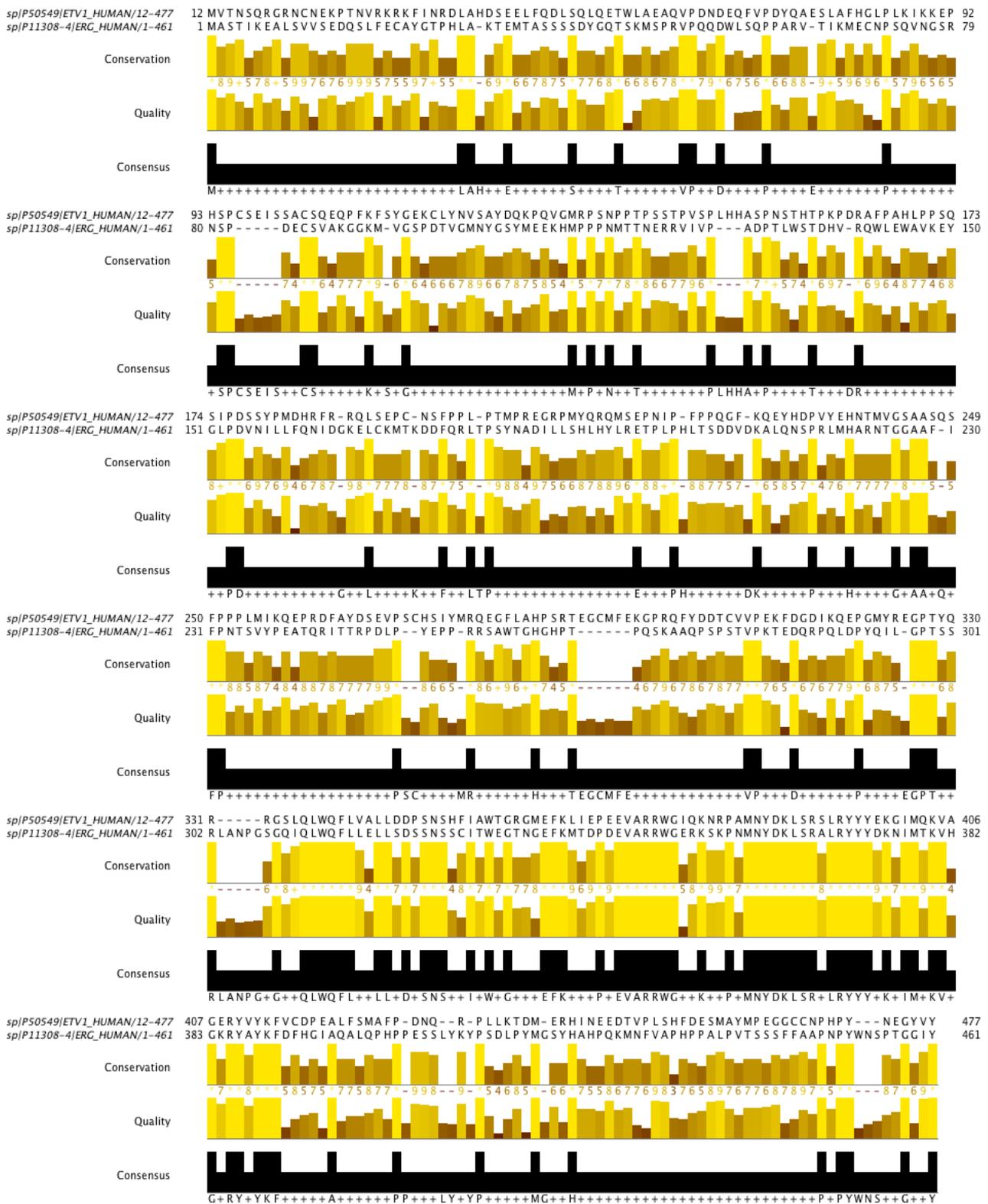


Figure S7 Sequence alignment of ERGp55 (Uniprot identifier: P11308-4)^{12, 13} and ETV1 (Uniprot identifier: P50549-1).^{13, 20} The high sequence similarity in the ETS region (294-400 of ERGp55) can be readily seen. The sequence alignment was performed using Jalview.²¹

4. ERGp55 200 ns RMSD and DNA-Binding plots

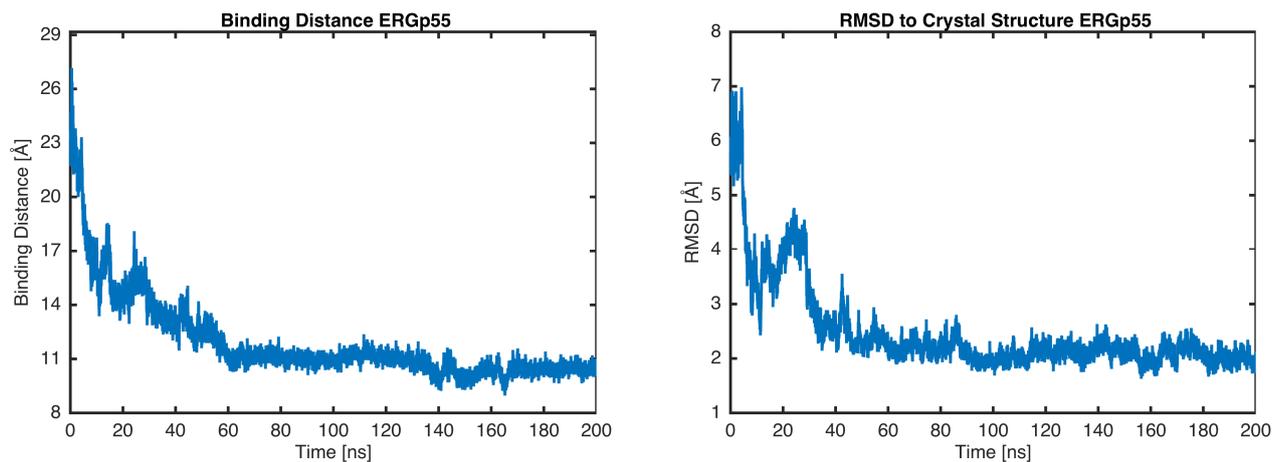


Figure S8 Binding distance plot of the distance between the center of mass of the α 3-helix of the ETS domain and the center of mass of the GGAA recognition sequence. Backbone RMSD plot of the ETS domain DNA-complex (amino acid residues 309-391) to the corresponding parts in the ERG ETS domain DNA crystal structure⁹ (terminal base pairs excluded from RMSD calculation, i.e. use of 10 of 12 bp).

5. RMSD Plots ETS Domain DNA-Complex Simulations

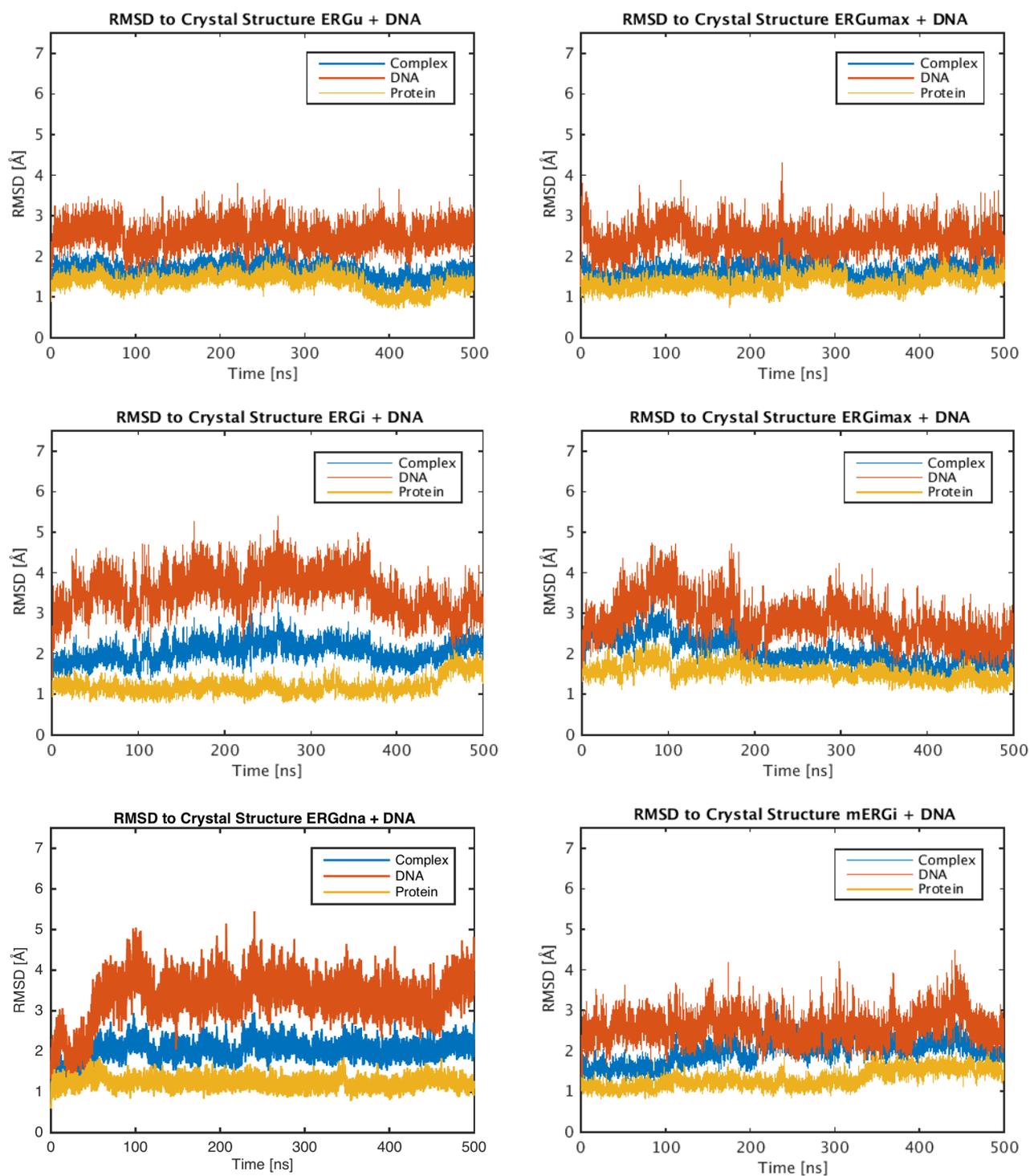


Figure S9 RMSD plots of the ERG ETS domain DNA-complex simulations using the core ETS domain structure (amino acid residues 309-391) and the DNA-structure excluding the terminal base pairs from the ERG ETS domain DNA-complex crystal structure (PDB-ID: 4iri)⁹ as a reference. The RMSD was calculated for the protein and phosphate backbone. Shown are the RMSD plots for the entire protein-DNA complex and the ETS domain and the DNA-molecules alone.

6. Binding Distances ETS Domain DNA-Complexes

Table S2 Binding Distances for the ETS domain DNA-complexes

	Binding Distance [Å]
ERGu	10.5
mERGu	11.2
ERGumax	10.5
ERGi	11.5
ERGimax	10.8
mERGi	11.3
ERGdna	11.6

Average binding distances over the course of 500 ns of simulation of each ETS domain DNA-complex. The DNA binding distance was defined as the distance between the center of mass of the α 3-helix and the GGAA core motif. The DNA fragment was extracted from the ERG ETS domain DNA-crystal structure (PDB-ID: 4iri)⁹.

7. Arg385 & DNA Mutants Binding Distance Plots including WT Repeats

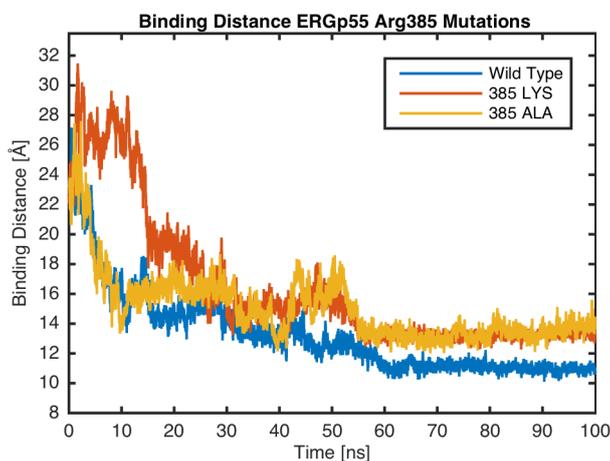


Figure S10 DNA-binding distance plots during 100 ns of the DNA-binding simulation of the WT ERGp55 and the Arg385 lysine and alanine mutants. The binding distance was defined as the distance between the center of mass of the α 3-helix and the GGAA core motif.

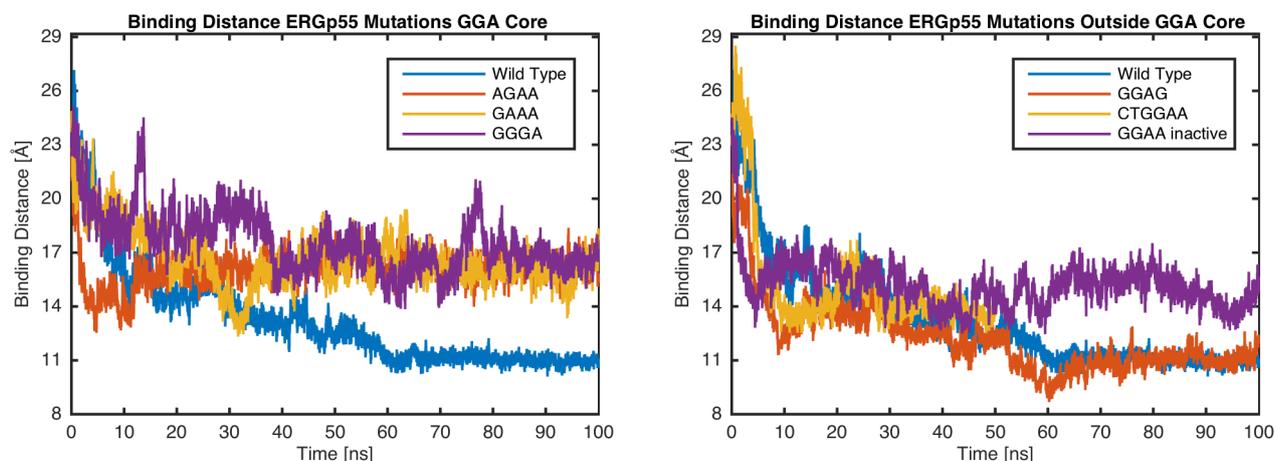


Figure S11 DNA-binding distance plots during 100 ns of the DNA-binding simulation of the WT ERGp55 and DNA mutants in and outside the GGA core sequence as well as an inactive EBS sequence. The binding distance was defined as the distance between the center of mass of the α 3-helix and the GGAA core motif.

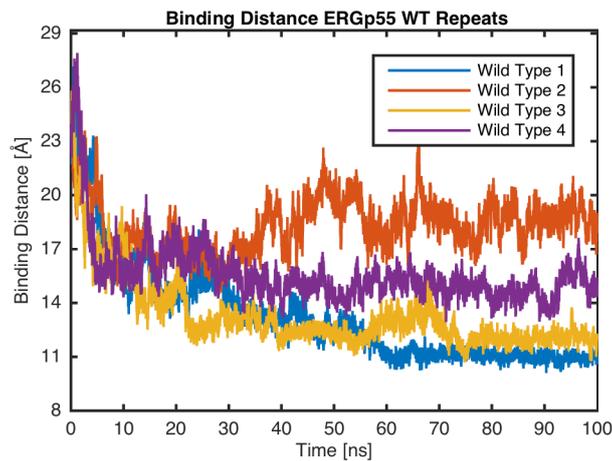


Figure S12 DNA-Binding distance plots during 100 ns of the DNA-binding simulation of the WT ERGp55 simulations. The binding distance was defined as the distance between the center of mass of the α 3-helix and the GGAA core motif.

8. Arg367 and Arg370 in the ERG ETS Domain DNA-Complex Crystal Structure and the ERGp55 DNA-Binding Simulation

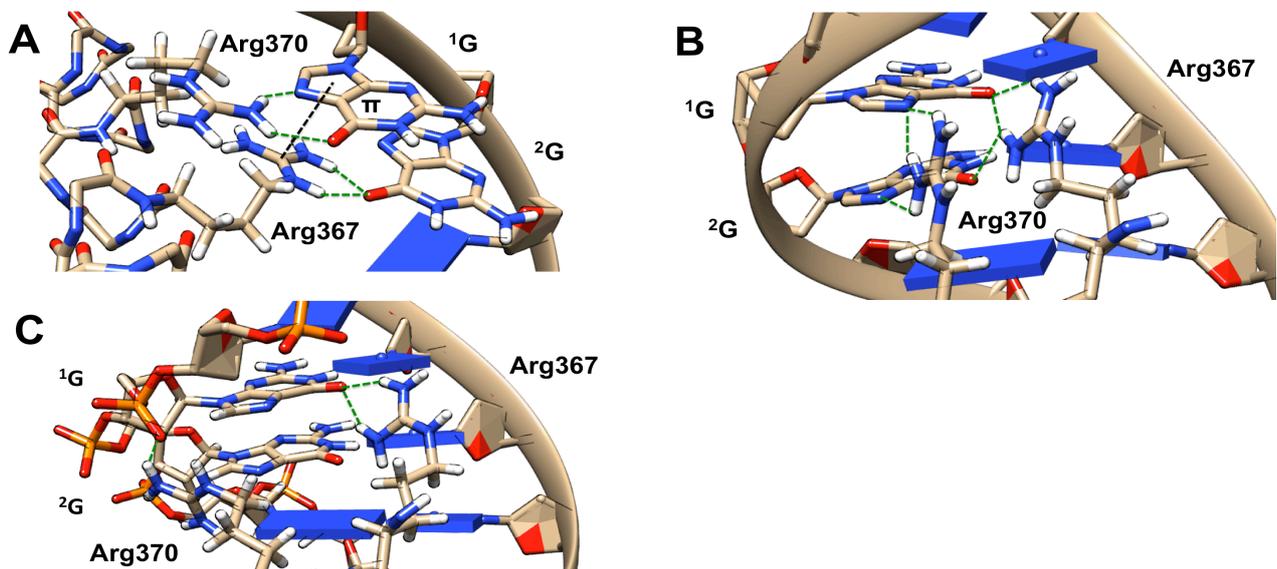


Figure S13 Orientation of Arg367 and Arg370 in the ETS domain DNA-complex crystal structure (PDB-ID: 4iri⁹) (**a**) and during the ERGp55 simulation (state 1 and 2 of Arg370 (**b**) and (**c**) respectively). Hydrogen bonding and salt bridge interactions are indicated by a green-dashed line and the cation- π -stacking interaction is indicated by a black dashed line.

9. Bridged Hydrogen Bonding Occupancies

Table S3 Occupancies of the bridged hydrogen bonds formed between the ERG ETS domain and DNA.^a

	-2 C ^s /G ^a	-1 C ^s /G ^a	1 G ^s /C ^a	2 G ^s /C ^a	3 A ^s /T ^a	4 A ^s /T ^a
ERGP55 D363		18.9^s	26.9^a	27.6^a		
ERGP55 R367		18.9 ^s				
ERGP55 R370				5.2 ^s		
ERGi D363	5.7^s	18.1^s, 19.6^a	34.9^a	51.7^a		
ERGi R367				10.7 ^s	23.6 ^a	5.6 ^a
ERGi R370			23.8^s			
ERGu D363	11.4^s	31.2^s	68.0^a	79.5^a		
ERGu R367		23.4 ^s , 6.1 ^a				
ERGu R370			34.9^s	8.4 ^s		
ERGumax D363	12.8^s	28.3^s	69.4^a	78.4^a		
ERGumax R367		24.4 ^s , 6.1 ^a				
ERGumax R370			43.5^s	7.1 ^s		
ERGDna D363	8.2^s	7.3^s, 20.9^a	5.6 ^s , 29.7^a	59.8^a		
ERGDna R367				6.4 ^s	23.3 ^a	10.1 ^a
ERGDna R370			21.7^s			

^a Occupancy of the bridged hydrogen interaction formed by the residue in the respective system as listed in %. ^s indicates an interaction with the DNA-base on the sense strand and ^a with the antisense strand. Occupancy is defined as the count of the detection of that particular interaction divided by the entire number of frames of the simulation. Interactions highlighted in bold are present in at least 4 of the 5 systems.

10. MM-GBSA Calculation ERGu, mERGu

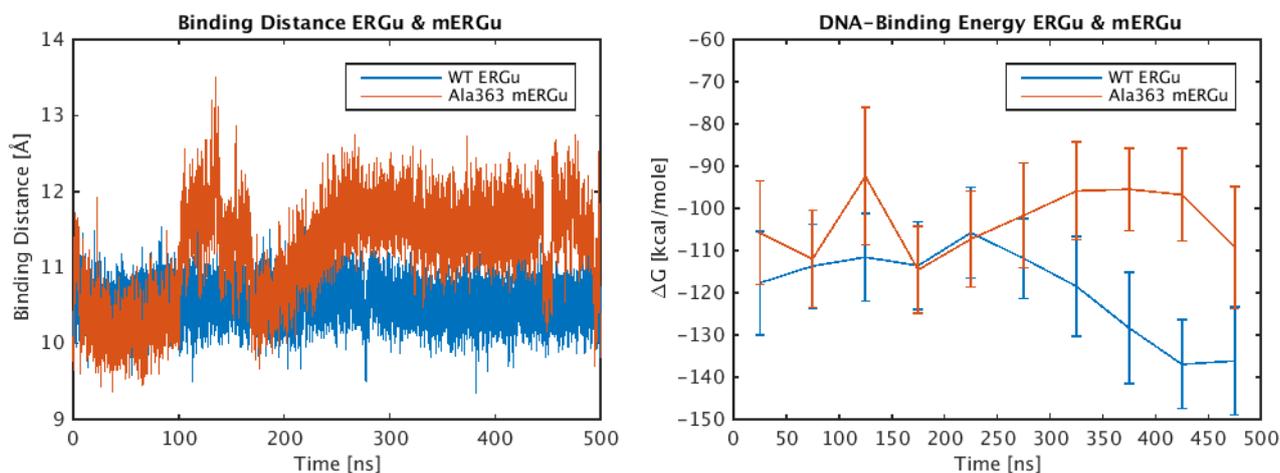


Figure S14 - Binding Distance plot for ERGu and mERGu during 500 ns of simulation in complex with DNA. The binding distance was defined as the distance between the center of mass of the α 3-helix and the GGAA core DNA-motif. The binding free energy plot shows MM-GBSA energies for the binding energy of ERGu and mERGu to DNA during 500 ns of simulation. The binding energies were calculated for time windows of 50 ns i.e. 0-50 ns, 50-100 ns etc. The error bars show the standard deviation of each calculation. The values between two X-axis ticks corresponds to the binding energy during that period.

While the DNA-binding energy of ERGu is in general lower than the binding energy of mERGu, it is hard to make a definite conclusion due to the huge variation in the values for ΔG . The variation in binding energy in the case of mERGu might origin from the fluctuations in the DNA-binding distance (see Figure S14). For the fluctuation of the MM-GBSA DNA-binding energies of ERGu there is no apparent reason, since neither the DNA-binding distance nor the RMSD plots (see Figure S9) show significant fluctuations.

11. References

- 1 V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins*, 2006, **65**, 712-725.
- 2 R. Salomon-Ferrer, A. W. Gotz, D. Poole, S. Le Grand and R. C. Walker, *J Chem Theory Comput*, 2013, **9**, 3878-3888.
- 3 A. W. Gotz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand and R. C. Walker, *J Chem Theory Comput*, 2012, **8**, 1542-1555.
- 4 S. Le Grand, A. W. Götz and R. C. Walker, *Computer Physics Communications*, 2013, **184**, 374-380.
- 5 D. A. Case, J. T. Berryman, R. M. Betz, D. S. Cerutti, T. E. I. Cheatham, T. A. Darden, R. E. Duke, T. J. Giese, H. Gohlke, A. W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, A. Kovalenko, T. S. Lee, S. LeGrand, P. Li, T. Luchko, R. Luo, B. Madej, K. M. Merz, G. Monard, P. Needham, H. Nguyen, H. T. Nguyen, I. Omelyan, A. Onufriev, D. R. Roe, A. Roitberg, R. Salomon-Ferrer, C. L. Simmerling, W. Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu, D. M. York and P. A. Kollman, *AMBER 2014*, University of California, San Francisco, 2014.
- 6 A. Onufriev, D. Bashford and D. A. Case, *Proteins-Structure Function and Bioinformatics*, 2004, **55**, 383-394.
- 7 S. Miyamoto and P. A. Kollman, *J Comput Chem*, 1992, **13**, 952-962.
- 8 H. J. C. Berendsen, J. P. M. Postma, W. F. Vangunsteren, A. Dinola and J. R. Haak, *J Chem Phys*, 1984, **81**, 3684-3690.
- 9 M. C. Regan, P. S. Horanyi, E. E. Pryor, Jr., J. L. Sarver, D. S. Cafiso and J. H. Bushweller, *Proceedings of the National Academy of Sciences of the United States of America*, 2013, **110**, 13374-13379.
- 10 C. D. Mackereth, M. Scharpf, L. N. Gentile, S. E. MacIntosh, C. M. Slupsky and L. P. McIntosh, *Journal of molecular biology*, 2004, **342**, 1249-1264.
- 11 E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng and T. E. Ferrin, *J Comput Chem*, 2004, **25**, 1605-1612.
- 12 D. S. Gerhard, L. Wagner, E. A. Feingold, C. M. Shenmen, L. H. Grouse, G. Schuler, S. L. Klein, S. Old, R. Rasooly, P. Good, M. Guyer, A. M. Peck, J. G. Derge, D. Lipman, F. S. Collins, W. Jang, S. Sherry, M. Feolo, L. Misquitta, E. Lee, K. Rotmistrovsky, S. F. Greenhut, C. F. Schaefer, K. Buetow, T. I. Bonner, D. Hausler, J. Kent, M. Kiekhaus, T. Furey, M. Brent, C. Prange, K. Schreiber, N. Shapiro, N. K. Bhat, R. F. Hopkins, F. Hsie, T. Driscoll, M. B. Soares, T. L. Casavant, T. E. Scheetz, M. J. Brownstein, T. B. Usdin, S. Toshiyuki, P. Carninci, Y. Piao, D. B. Dudekula, M. S. Ko, K. Kawakami, Y. Suzuki, S. Sugano, C. E. Gruber, M. R. Smith, B. Simmons, T. Moore, R. Waterman, S. L. Johnson, Y. Ruan, C. L. Wei, S. Mathavan, P. H. Gunaratne, J. Wu, A. M. Garcia, S. W. Hulyk, E. Fuh, Y. Yuan, A. Sneed, C. Kowis, A. Hodgson, D. M. Muzny, J. McPherson, R. A. Gibbs, J. Fahey, E. Helton, M. Kettelman, A. Madan, S. Rodrigues, A. Sanchez, M. Whiting, A. Madari, A. C. Young, K. D. Wetherby, S. J. Granite, P. N. Kwong, C. P. Brinkley, R. L. Pearson, G. G. Bouffard, R. W. Blakesly, E. D. Green, M. C. Dickson, A. C. Rodriguez, J. Grimwood, J. Schmutz, R. M. Myers, Y. S. Butterfield, M. Griffith, O. L. Griffith, M. I. Krzywinski, N. Liao, R. Morin, D. Palmquist, A. S. Petrescu, U. Skalska, D. E. Smailus, J. M. Stott, A. Schnerch, J. E. Schein, S. J. Jones, R. A. Holt, A. Baross, M. A. Marra, S. Clifton, K. A. Makowski, S. Bosak, J. Malek and M. G. C. P. Team, *Genome research*, 2004, **14**, 2121-2127.
- 13 C. UniProt, *Nucleic acids research*, 2015, **43**, D204-212.
- 14 L. A. Kelley, S. P. Gardner and M. J. Sutcliffe, *Protein Eng*, 1996, **9**, 1063-1065.
- 15 S. P. Gangwar, S. Dey and A. K. Saxena, *PLoS one*, 2012, **7**, e39850.
- 16 J. E. Zamora, M. Papadaki, A. E. Messer, S. B. Marston and I. R. Gould, *Physical chemistry chemical physics : PCCP*, 2016, **18**, 20691-20707.
- 17 T. J. Macke and D. A. Case, in *Molecular Modeling of Nucleic Acids*, American Chemical Society, 1997, vol. 682, ch. 24, pp. 379-393.
- 18 N. H. Dryden, A. Sperone, S. Martin-Almedina, R. L. Hannah, G. M. Birdsey, S. T. Khan, J. A. Layhadi, J. C. Mason, D. O. Haskard, B. Gottgens and A. M. Randi, *The Journal of biological chemistry*, 2012, **287**, 12331-12342.
- 19 K. Degitz, L. Lianjie and S. W. Caughman, *Journal of Biological Chemistry*, 1991, **266**, 14024-14030.
- 20 I. S. Jeon, J. N. Davis, B. S. Braun, J. E. Sublett, M. F. Roussel, C. T. Denny and D. N. Shapiro, *Oncogene*, 1995, **10**, 1229-1234.
- 21 A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. I. Clamp and G. J. Barton, *Bioinformatics*, 2009, **25**, 1189-1191.