Molecular BioSystems

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: M. G. Beuerle, N. P. Dufton, A. M. Randi and I. R. Gould, *Mol. BioSyst.*, 2016, DOI: 10.1039/C6MB00506C.

This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the [Information for Authors](#).

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard [Terms & Conditions](#) and the [Ethical guidelines](#) still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

Molecular BioSystems

PAPER

Molecular Dynamics Studies on the DNA-Binding Process of ERG

Matthias G. Beuerle^a, Neil P. Dufton^b, Anna M. Randi^b, Ian R. Gould^{†a}Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

The ETS family of transcription factors regulate gene targets by binding to a core GGAA DNA-sequence. The ETS factor ERG is required for homeostasis and lineage-specific functions in endothelial cells, some subset of haemopoietic cells and chondrocytes; its ectopic expression is linked to oncogenesis in multiple tissues. To date details of the DNA-binding process of ERG including DNA-sequence recognition outside the core GGAA-sequence are largely unknown. We combined available structural and experimental data to perform molecular dynamics simulations to study the DNA-binding process of ERG. In particular we were able to reproduce the ERG DNA-complex with a DNA-binding simulation starting in an unbound configuration with a final root-mean-square-deviation (RMSD) of 2.1 Å to the core ETS domain DNA-complex crystal structure. This allowed us to elucidate the relevance of amino acids involved in the formation of the ERG DNA-complex and to identify Arg385 as a novel key residue in the DNA-binding process. Moreover we were able to show that water-mediated hydrogen bonds are present between ERG and DNA in our simulations and that those interactions have the potential to achieve sequence recognition outside the GGAA core DNA-sequence. The methodology employed in this study shows the promising capabilities of modern molecular dynamics simulations in the field of protein DNA-interactions.

Introduction

Transcription factors are essential components of the regulation of gene expression in eukaryotic cells. Their function is mediated through sequence specific DNA binding and further interactions with other proteins involved in the gene expression process.¹ It is therefore of high biological interest to understand the DNA-binding process at a molecular level. Conventional experimental tools to study protein DNA-interactions mainly include biochemical techniques e.g. electrophoretic mobility shift assays (EMSAs)² and X-ray crystallography. More recently molecular dynamics techniques have proved invaluable in obtaining dynamic information at the atomic level to interpret experimental results in the field of protein DNA-interactions (see ref. 3 and references therein). In this report we study the DNA-interaction of the transcription factor encoded by the ETS-related Gene (ERG), a member of the E26 transformation specific (ETS) protein family.^{4, 5} ERG, the highest expressed ETS transcription factor in mature, quiescent endothelial cells (EC) has been shown to be required for vascular development, during angiogenesis^{6, 7} and in maintaining vascular homeostasis.^{8, 9} ERG drives

expression of genes, which define the endothelial lineage,¹⁰ and controls multiple pathways essential for EC function, partly by regulating EC junctions.^{10, 11} Moreover ERG can promote homeostasis by repressing genes involved in inflammation.^{8, 9, 12}

Whilst constitutively expressed ERG exerts a homeostatic function, its pathological over-expression due to chromosomal translocations is associated with cancer. ERG has been found as a fusion transcript with the transcription factor ELF4 in acute myeloid leukemia¹³ and with the prostate specific gene TMPRSS2 in prostate cancer.¹⁴ The complex lineage-specific function of ERG highlights the need to characterize the mechanisms through which ERG controls gene expression.

The ERG isoform ERGp55 contains two functional domains, namely an ETS and a PNT domain.¹⁵ The ETS domain, which is characteristic for transcription factors of the ETS family, consists of a highly conserved 85 amino acid sequence and shows a characteristic helix-turn-helix fold.⁴ There are numerous ETS domain crystal structures available^{16, 17} including the structure of the ERG ETS domain alone and in complex with DNA.¹⁸ The ETS domain is responsible for sequence specific DNA-binding, recognizing a ¹G²G³A⁴(A/T) cognate sequence.¹⁹ Sequence specificity among different ETS factors is thought to be achieved through the recognition of bases flanking the ¹G²G³A⁴(A/T) core sequence.²⁰

Potential residues involved in DNA-sequence recognition have been analyzed in crystal structures of ETS proteins.¹⁶ A direct recognition mechanism has been identified for the ¹G and ²G base pairs, which is mediated through an interaction with two conserved arginine residues (Arg367 and Arg370 in ERGp55, location on α3-helix shown in Figure 1), which inserts into the DNA major groove.^{4, 16} A conserved tyrosine residue (Tyr371 in

^a Department of Chemistry and Institute of Chemical Biology, Imperial College London, South Kensington SW7 2AZ, United Kingdom.

^b National Heart and Lung Institute (NHLI) Vascular Sciences, Hammersmith Hospital, Imperial College London, London W12 0NN, United Kingdom.

[†] Corresponding Author: Ian R. Gould. Tel: +44 (0)20 7594 5809, Email: i.gould@imperial.ac.uk.

Present Address: Matthias Beuerle, Department of Chemistry, University of Munich (LMU), Butenandtstraße 7, 81377 Munich, Germany.

[†] Electronic Supplementary Information (ESI) available: Additional figures and tables, detailed methodology of molecular dynamics simulations. See DOI: 10.1039/x0xx00000x

PAPER

Molecular BioSystems

ERGp55) on the same helix has been suggested to interact with the ³A base.^{4, 16, 18} An additional indirect recognition mechanism, by which water molecules acting as bridges for hydrogen bonds between amino acids and DNA bases confer additional sequence specificity outside the GGAA cognate sequence has recently been proposed for other ETS factors.¹⁶ In this article we performed molecular dynamics simulations to study the DNA-binding process of ERG. Based on a model built ERGp55 full length structure incorporating available structural information^{18, 21} and a B-DNA structure of a well characterized ERG DNA binding site in the promoter region of ICAM-1,⁸ we were able to simulate DNA-binding reproducing the bound ERG DNA-complex with a RMSD of 2.1 Å compared to the crystal structure. We investigated the importance of salt bridges between ERG and DNA, which revealed Arg385 as a novel key player in the DNA-binding process of ERG. Moreover, we studied the DNA-sequence recognition by ERG and highlighted the role of water-mediated hydrogen bonding and Asp363 in the DNA-sequence recognition process outside the GGAA core motif. Thus, this study identifies new mechanistic aspects of the DNA binding process of ERG, revealed by a novel approach to study protein DNA interactions using molecular dynamics simulations, which is transferable to other open questions in the field.

MATERIAL AND METHODS

Molecular Dynamics (MD) Simulations

All simulations were performed using the AMBER ff99SB²² force field parameters for proteins and nucleotides and the GPU accelerated AMBER PMEMD code.^{23, 24} The simulations were set up using the available structural data for the ETS domain (PDB-IDs: 4iri, 4irh, 4irg)¹⁸ and the PNT domain (PDB-ID: 1sxe)²¹ of ERG. Models containing protein sequences without complete structural data were generated from MD simulations of the respective construct in which these parts were initially inserted as linear peptide chains and equilibrated during the course of the simulation. The DNA-structure used in the ETS domain DNA-complex simulation was extracted from the ERG ETS domain DNA complex crystal structure (PDB-ID: 4iri)¹⁸.

Figure 1 ERG ETS domain DNA-complex X-Ray structure (PDB-ID: 4iri)¹⁸ with the amino acids involved in DNA sequence recognition explicitly shown on the α 3-helix, which inserts into the DNA major groove.

For the DNA-binding simulation of full-length ERGp55, B-DNA nucleic acid structures were generated using NAB²⁵ as implemented in AMBER Tools 14.²⁶ DNA-sequences were taken from the -181 ETS binding site (EBS) in the ICAM-1 promoter.^{8, 27}

A detailed description of the MD simulation protocol and the setup of the individual simulations can be found in the ESI.[†]

Trajectory Analysis

The majority of the trajectory analysis was performed using cpptraj²⁸ as implemented in the AMBER tools software package.²⁶ This includes root-mean-square deviation (RMSD) and B-factor calculations, hydrogen bond analysis, calculations of nucleic acid parameters²⁹ and gridded histogram analysis. For backbone RMSD calculations the CA, C, N and O atoms were used for proteins and the P, OP1 and OP2 atoms were used for nucleotides. For B-factor calculations all atoms were used to provide a mass weighted average for each residue. Estimates for the major groove widths are provided using the average distance between the P atoms of each base pair, while estimates for the minor groove widths are provided using the average distance between the O4' atoms of each base pair. Hydrogen bonds and salt bridges were analyzed using a 3.5 Å distance and a 135° angle cutoff. The gridded histogram analysis used a grid centered on the DNA molecule of the ERGp55 DNA-complex with 0.5 Å spacing in each spatial dimension with a total grid volume of 150 Å x 150 Å x 150 Å (X, Y and Z dimensions). To evaluate, whether the densities obtained from the gridded histogram analysis are in a position to form bridged hydrogen bonds, water molecules were centered in the relevant density positions aligned with a trajectory snapshot after 200 ns and the geometric criteria used for detection of bridged hydrogen bonds in cpptraj²⁸ were measured using UCSF chimera.³⁰ Visualization of the trajectory as well as clustering analysis³¹ was carried out using UCSF chimera.³⁰

MM-GBSA Calculations and Alanine Scanning Mutagenesis

DNA-binding energies were obtained using the molecular mechanics generalized born surface area (MM-GBSA) method from the MMPBSA.py³² script to calculate free energies of binding in combination with the ante-MMPBSA.py script implemented in AMBER Tools.²⁶ The MM-GBSA approach is a well-established trajectory post-processing method to estimate binding energies. A detailed description of the methodology including a review of application studies can be found in the literature.³²⁻³⁴ The MM-GBSA approach was used to estimate the binding energies of ERGp55 binding to DNA using 25 ns (time frame 75-100 ns) of simulation of the DNA binding simulations using an interval of 5 yielding 401 frames for analysis. The input parameters were set according to a MM-GBSA study on nucleic acids³⁵ (Generalized Born Model: Hawkins, Cramer, Truhlar pairwise generalized Born model $igb=1$,³⁵⁻³⁷ radii set *mbondi*, surface tension value 0.0072 N/m², salt concentration 0.15 M, LCPO algorithm to calculate surface area for nonpolar solvation term, internal dielectric constant

$\epsilon_{\text{int}} = 1.0$, external dielectric constant $\epsilon_{\text{ext}} = 78.5$, dielectric radii offset 0.09 Å). The MMPBSA.py³² script was employed to strip solvent and salt molecules from each snapshot and calculate an estimate for the binding energy averaging the results of all frames used in the calculation.

Alanine scanning mutagenesis is an application of the MM-GBSA methodology to estimate the importance of different residues involved in a binding process.³⁸ Here different residues in the WT trajectory are mutated to alanine prior to another MM-GBSA calculation, which provides an estimate in comparison to the WT and the results of other mutants. We have also developed and performed phenylalanine scanning, a variation of the alanine scanning methodology, in which tyrosine is mutated to phenylalanine in the same manner to facilitate a more physically relevant model to compare with the experimental mutations.

Alanine and Phenylalanine Scanning Mutagenesis³⁸ was performed on 25 ns of the ERGp55 DNA-binding simulation. To perform alanine scanning mutagenesis the relevant parameter files were generated manually. Then the trajectory modification and the MM-GBSA energy calculation were performed using the MMPBSA.py³² script. The coordinates of the mutated residue are deleted from the trajectory and replaced by coordinates of an alanine residue in equilibrium geometry.³² Then the MM-GBSA calculation is performed analogous to the conventional MM-GBSA procedure described above. Since phenylalanine scanning is not implemented in AMBER Tools, a script was written that mutates the relevant tyrosine to phenylalanine in the wild type (WT) trajectory. In particular the coordinates of the oxygen (OH, AMBER atom type) of the hydroxyl group of the tyrosine were deleted and the coordinates of the hydrogen (HH) were modified accordingly yielding the mutated trajectory. To obtain reasonable coordinates for HH the vector between the CG and CZ atom of the benzene ring was calculated and scaled to a length of 1.08 Å before it was added to the vector of the coordinates of the CZ atom. This ensured a consistent CZ-HZ bond length of 1.08 Å, which is equivalent to the average CZ-HZ bond length of a random phenylalanine in the ETS domain during 100 ns of the ERGp55 DNA-binding simulation. Additionally the relevant parameter files were generated and the MM-GBSA calculations were performed using the mutated trajectory analogous to the WT calculations.

RESULTS AND DISCUSSION

De Novo DNA-Binding Simulations.

To investigate the DNA-binding process of ERG we used a model-generated, full-length ERGp55 structure and model built B-DNA. Therefore we extracted a representative structure from the equilibration MD simulation of the full-length protein (for details on the generation of the model see ESI†) using clustering analysis.³¹ The DNA sequence used was shown experimentally to be a biologically relevant ERG binding site⁸ and corresponds to the -181 region in the ICAM-1 promoter region.

Figure 2 Setup of the ERGp55 simulation. The full-length ERG model was set up in ~23 Å distance of the GGAA sequence (labeled in purple). The ETS domain is labeled in blue.

The system was set up such that at the beginning of the simulation the center of mass of the $\alpha 3$ -helix of the ETS domain was 23 Å apart from the center of mass of the GGAA cognate sequence (Figure 2). We performed a 100 ns MD simulation, which demonstrated the DNA binding process of ERGp55. The simulation of the full-length ERGp55 binding to DNA resulted in a stable position at the GGAA motif after approximately 60 ns of simulation (for binding distance plot see Figure 3). The average binding distance (defined as the distance between the center of mass of the $\alpha 3$ -helix and the center of mass of the GGAA motif) was 11.0 Å compared to 10.7 Å in the ERG ETS domain DNA-complex crystal structure (PDB-ID: 4iri).¹⁸ During the period of 75-100 ns the average DNA-protein backbone RMSD of the core ETS domain DNA region to the original ERG ETS domain DNA-complex crystal structure was 2.1 Å (Figure 3). For a visualization of the graphical overlap see Figure 3. The minor and major groove widths were calculated to be 12.5 Å and 20.3 Å on average over the last 25 ns of the simulation (crystal structure 12.3 Å and 19.4 Å respectively). In order to compare this simulation to other simulations starting with a bound ERG ETS domain DNA-complex, we performed simulations of the original ETS domain DNA-complex crystal structure¹⁸ and other ETS domains aligned with the experimental structure. Different ETS domain structures were used since it has been shown that ETS domain constructs with different lengths demonstrate different experimental DNA-binding affinities.¹⁸ In particular Regan et al.¹⁸ performed isothermal titration calorimetry experiments to measure DNA-binding affinities of ERG ETS domain constructs of different lengths and showed that they exhibit different DNA-binding affinities. Using crystallographic data and NMR measurements they provided two hypotheses, namely two different conformations of Tyr371 in the DNA-bound and unbound state and conformational dynamics of the ETS domain, to explain this behaviour. Performing simulations with multiple constructs should accommodate those differences in the DNA-binding behaviour and enhance the statistical significance of our simulations. The lengths of the ETS domains used were chosen according to the constructs studied experimentally.¹⁸

Figure 3 Binding distance plot (A) 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 (B) of the ETS domain DNA-complex, the DNA fragment and the ETS domain alone of the core ETS domain (amino acid residues 309-391) to the corresponding parts in the ERG ETS domain DNA crystal structure¹⁸ (terminal base pairs excluded from RMSD calculation, i.e. using 10 of 12 bp). Structural overlap of the ERG ETS domain DNA crystal (PDB-ID: 4iri)¹⁸ structure with a snapshot of the ERGp55 DNA-binding simulation after 100 ns (C).

The constructs were named corresponding to their experimentally obtained DNA-binding strengths ERGinhibited (ERGi), ERGuninhibited (ERGu), maximal inhibited and uninhibited ERG (ERGimax and ERGumax respectively) and ERGdna (the actual experimental crystal structure PDB-ID: 4iri;¹⁸ more details are given in the ESI†). Analysis of those simulations for DNA binding distances and RMSD plots of the ETS domain DNA-complex showed results very similar to the ERGp55 simulation (see Figure S9† & Table S2† for more details). In addition to geometric criteria we analyzed the ERGp55 DNA-binding simulation for salt bridges and hydrogen bonds to the phosphate backbone of DNA that were proposed to mediate the binding process in the ETV1 ETS domain (for a sequence alignment of ETV1 and ERGp55 see Figure S7†).¹⁶ As a reference we also examined these interactions in the MD simulations of the other ERG ETS domain DNA-complex trajectories (Table 1). All eight interactions formed by residues

that are conserved in ERG were observed in our ERGp55 system. An interaction between ERGp55 and DNA outside the ETS region was not detected. In addition to reproducing the experimental results, we observed the interactions of two residues (Lys357 and Arg385), which Cooper et al.¹⁶ proposed to be in position to form interactions with the phosphate backbone (Arg381 and Arg409 in ETV1). Since the DNA molecules used for crystallization did not extend to that region the authors were not able to show these interactions explicitly in any of the experimental structures. The DNA-binding simulation of ERGp55 with a 49 base pair DNA fragment shows that both interactions exist in the equilibrated state of the ERGp55 DNA-system and that the salt bridge of Arg385 is among the most stable interactions in the 25 ns simulation period (Table 1). This suggests a particular importance of this residue that will be examined below.

To further validate the ERGp55 DNA-binding simulation and to show the capabilities of molecular dynamics simulations to relate to experimental data, we examined the fluctuations of Arg367 and Arg370. It has been hypothesized that these residues undergo an order-disorder transition upon DNA binding in the ERG homologue ETV1.¹⁶ To investigate the disorder-order transitions, we calculated B-factors of the ETS domain during the period of 75-100 ns of the ERGp55 DNA-binding simulations and referenced them to B-factors of the ETS domain calculated during the period of 300-800 ns of the apo ERGp55 simulation. Colour mapping the difference between apo and DNA bound B-factor values onto the structure of the ETS domain reveals the change of fluctuations of each residue after the DNA-binding event (Figure 4). The results show that Arg367 and Arg370 indeed show lower B-factors compared to the apo simulation. Moreover 9 of the 10 salt-bridges examined earlier show B-factors with higher than median reductions in the B-factor with the exception of Trp351. One possible explanation why Trp351 is not identified in this analysis is that Trp351 has one of the lowest B-factors in the apo simulation prohibiting further stabilization.

Figure 4 ERGp55 ETS domain colored according to the difference of the B-factors between apo- and DNA-bound form. The color key was created using the median of all values as the white level, therefore red residues are more and blue residues less ordered in the DNA bound state than the median. The labeling unit of the color key is $\text{\AA}^2 \times 3/2 \pi^2$.

Table 1. Duration of hydrogen bonds and salt bridges formed during simulations of the DNA-complexes^a.

	Observed Saltbridges							Potential Saltbridges	
	GLN312	TRP351	LYS355	TYR362	LYS364	TYR372	LYS375	LYS380	ARG385
ERGp55	o	oo	o	o	o	ooo	o	o	ooo
ERGdna	oo	ooo	oo	ooo	oo	ooo	o	oo	NA
ERGi	ooo	ooo	o	x	oo	ooo	o	o	NA
ERGimax	oo	x	o	o	o	ooo	x	x	NA
ERGu	oo	ooo	oo	oo	oo	ooo	oo	o	NA
ERGumax	ooo	ooo	oo	oo	oo	ooo	o	o	NA

^a 'o' indicates that the interaction is present for more than 5%, 'oo' more than 30% and 'ooo' for more than 75%; 'x' indicates the interaction is not present for longer than 5%. 'NA' indicates that the respective interaction cannot be observed in this system due to a too short DNA fragment. The data was collected during the period of 75-100 ns of each simulation.

Another residue showing a significant order-disorder transition is Tyr371, which is speculated to be important in DNA sequence recognition of ETS factors¹⁶ and autoinhibitory regulation of ERG binding to DNA.¹⁸

In summary we were able to reproduce the ERG ETS domain DNA-complex in considerable detail and validated our model by comparing our simulation to experimental findings. It is very interesting to note that the B-factor analysis revealed 9 of the 10 salt bridges including the novel salt bridges Lys357 and Arg385 highlighting the potential of this methodology to identify important residues in the DNA-binding process. We believe that B-factor analysis presents a good method to identify important residues in the binding process as amino acids should in general exhibit a lower flexibility when bound permanently to another molecule rather than when free in solution without any permanent interactions. Residues that are not directly involved in the binding process should not experience a strong change in their flexibility upon DNA-binding. Exceptions might be residues that form strong intramolecular interactions in the unbound state.

Mutations of the ERGp55 Protein- and DNA-Sequence

In order to investigate the specific role of amino acid residues and DNA bases we utilized multiple strategies for mutations in both the ERGp55 and the DNA-sequence. This allowed us to perform MD simulations and MM-GBSA DNA-binding calculations to reveal the importance of distinct amino acid residues and the influence of DNA-sequence variations in the DNA-binding process of ERG.

Alanine Scanning Mutagenesis of Residues Involved in DNA-Binding. To investigate the relevance of Arg385 in DNA-binding and to improve understanding of the importance of residues involved in the DNA binding process of ERG, we calculated MM-GBSA DNA-binding energies for wild type (WT) ERG using the ERGp55 simulation and alanine mutants of residues relevant in the DNA-binding process obtained using alanine scanning mutagenesis.³⁸ In particular we analysed all residues involved in forming a hydrogen bond or salt bridge with the DNA-backbone (Table 1) and residues speculated to be involved in sequence recognition⁴ (Arg367, Arg370 and Tyr371) plus Tyr373, which was examined experimentally.¹⁶ MM-GBSA

energies of the DNA-complex formation were used to rank the contribution to DNA-binding of the residues under study by qualitatively examining the results of alanine scanning mutagenesis. Highly accurate quantitative binding energies cannot reliably be obtained using the MM-GBSA methodology.³⁴ The binding energy of ERGp55 WT to DNA was calculated to be -97.5 kcal/mole. Standard deviations in the range of 9 to 11 kcal/mole, i.e. ~10% of the mean, indicate that the underlying data leading to the results of the MM-GBSA calculations show reasonable scattering. Taking into account the sample size results in standard errors of 0.5 to 0.6 kcal/mole, which under the assumption of a Gaussian error distribution translates into an estimated 95% confidence level of +/- 1.0 - 1.2 kcal/mole. Our results (Figure 5) show that mutation of Arg385 to alanine leads to the largest decrease in DNA-binding energy, again suggesting this residue to be highly instrumental in the binding process of ERG to DNA. When comparing our results to experimental mutation studies of ETS factors we found overlap between experimental and in silico results. It was reported that alanine mutations of the homologues of Lys355 and Lys380 in ETV1 did not show any detectable DNA-binding¹⁶ and Lys355 indeed shows the second weakest binding of the mutations that we examined. Lys380 also shows weaker DNA-binding but not to the same extent as can be seen in Figure 5. In the same study the homologues of Arg367 and Arg370 were mutated to alanine and Tyr371 to phenylalanine. Arg367 and Arg370 showed a significant decrease (50 and 1300 times less respectively) in binding energy while Tyr371 was only marginally affected.¹⁶ Our DNA-binding energy results also show that alanine mutations of Arg367 and Arg370 have a higher impact on the binding energy than the Tyr371 mutant. Conversely mutation of Tyr372 and Tyr373 to phenylalanine showed a significant decrease in DNA-binding affinity experimentally (>1000 times less)¹⁶ and in our study the mutations of Tyr372 and Tyr373 to alanine had a lesser effect than the mutation of Tyr371 to alanine, which showed no significant decrease in binding affinity in experiment.¹⁶ However analysis of the lifetimes of salt bridges performed earlier (Table 1) revealed that the hydrogen bond formed by Tyr372 to the DNA backbone is among the most stable interactions indicating its importance in DNA binding.

View Article Online
DOI: 10.1039/C6MB00506C

Figure 5 Alanine scanning MM-GBSA DNA-binding energies for the alanine and phenylalanine mutants with error bars representing the standard error of the calculation.

An interaction of Tyr373 with the backbone was however not detected. Potential reasons for the deviation of our alanine scanning results from experiment, in addition to the deficiencies of using the MM-GBSA method to calculate binding energies are that alanine scanning mutagenesis only mutates the desired residue to alanine in the calculation of the binding energy but does not perform MD again, which does not allow the environment to adjust to the mutation.

Since experimentally tyrosine residues were mutated to phenylalanine,¹⁶ alanine scanning mutagenesis is not ideal to compare between simulation and experiment. Therefore, we mutated relevant tyrosine residues in the ERGp55 WT simulation to phenylalanine *in silico* to perform those mutations analogous to the mutations carried out in the alanine scanning procedure. The phenylalanine scanning MM-GBSA DNA-binding energy results for Tyr362 and Tyr373 are nearly identical to the MM-GBSA energies obtained from alanine scanning, while the phenylalanine mutants for Tyr371 and Tyr372 showed more favorable DNA-binding energies with a difference of ~5 kcal/mole and ~2 kcal/mole respectively (Figure 5). The DNA-binding energy of the Tyr371 mutant is now significantly closer to the ERGp55 WT energy than the alanine mutant and slightly stronger than the Tyr372 phenylalanine mutant, which is more in line to what has been observed experimentally for ETV1.¹⁶ To summarize, alanine scanning mutagenesis highlights Arg385 as a key residue in the DNA-binding process of ERG and shows reasonable qualitative agreement with experiment, which is remarkable considering that the ERGp55 DNA complex was formed *de novo* by a simple MD simulation without external bias. This includes the energetic importance of Lys355 and the relative binding energies of Tyr371 in comparison to Arg367 and Arg370. Introducing phenylalanine scanning as a variation of alanine scanning increased the overlap to experimental data with regard to Tyr371 and Tyr372. It also has to be taken into account that mismatches between *in silico* alanine scanning mutagenesis and experiment are not necessary a deficiency of the underlying MD trajectory but can also be related to weaknesses of the method itself.

In Silico Mutations of Arg385. To further assess the role of Arg385, we performed simulations with the same setup as the ERGp55 DNA-binding simulation (Figure 2) in which Arg385 was mutated to lysine and alanine. It was hypothesized that in

line with our previous data mutation to alanine should weaken binding drastically. Mutation of Arg385 to lysine should however rescue the binding energy due to the ability of lysine to also form a salt bridge with the DNA-backbone. The plot of the RMSD to the core ERG ETS domain DNA-complex crystal structure throughout 100 ns of simulation in each trajectory shows that the profiles of the simulations of both mutations are very similar and also compare well to the WT profile (Figure 7, Table 2). The binding distance plots are shown in Figure S10.† Due to the highly similar geometries we calculated MM-GBSA DNA-binding energies for the mutant ERGp55 constructs. The alanine mutant showed a binding free energy of $\Delta G = -36.5$ kcal/mole (sd ± 9.9 kcal/mole) and the lysine mutant bound to DNA with $\Delta G = -76.4$ kcal/mole (sd ± 10.7 kcal/mole) compared to the WT DNA-binding energy of -97.5 kcal/mole (sd ± 11.2 kcal/mole). These results underline that Arg385 is essential for DNA-binding. Mutation of Arg385 to lysine rescued the binding energy to some extent, which is to be expected, since lysine is also capable of forming a salt bridge to the phosphate backbone.

Table 2 Average RMSD values and MM-GBSA DNA-binding energies for DNA-binding simulations of ERGp55 and DNA mutants^a.

	RMSD [Å]	ΔG [kcal/mole]	sd $\pm$ [kcal/mole]
¹ G ² G ³ A ⁴ A (WT)	2.1	-97.5	11.2
Ala385	2.8	-36.5	9.9
Lys385	2.8	-76.4	10.7
¹ A ² G ³ A ⁴ A	4.7	-54.9	8.1
¹ G ² A ³ A ⁴ A	5.7	-68.6	10.7
¹ G ² G ³ G ⁴ A	5.4	-75.4	14.2
¹ G ² G ³ A ⁴ G	2.8	-57.1	10.4
⁻² C ⁻¹ T ¹ G ² G ³ A ⁴ A	2.6	-53.8	11.5
Inact. EBS	2.9	-60.9	11.9
Repeat 2	4.7	-51.3	12.6
Repeat 3	2.6	-78.4	19.0
Repeat 4	3.2	-63.0	10.5

^a Average backbone RMSD values to the core ETS domain DNA-Complex crystal structure¹⁸ and MM-GBSA DNA-binding energies evaluated during the period of 75-100 ns of the respective simulation. The superscript denotes the index of the DNA base with the first G from GGAA set to 1. The inactive EBS refers to the GGAA DNA-sequence adjacent to the WT sequence in the -181 EBS of the ICAM-1 promoter. The repeats were performed for the WT ERGp55 system.

Figure 6 DNA sequences of the ICAM-1 promoter^{8,27} and mutations thereof used in the DNA binding simulation of full-length ERGp55. The red box highlights the GGAA sequence referred to as WT in text. The blue box highlights the inactive EBS adjacent to the WT ERG binding site.

Phylogenic analysis of the sequence conservation in the ETS domain throughout the ETS family of transcription factors revealed that the homologue of Arg385 is mutated to lysine in the ELK subfamily of ETS transcription factors,³⁹ which also explains, why mutation of Arg385 to lysine rescues DNA binding energy.

Mutation of the DNA Recognition Sequence. We performed simulations with mutated DNA sequences, to determine whether we would be able to discriminate the WT ERGp55 DNA-binding sequence against mutated versions. Experimentally ERG has been shown to have a strong sequence preference for ¹G, ²G, ³A and a less strict sequence preference for ⁻²C, ⁻¹C and ⁴(A/T).²⁰ Therefore, we mutated the DNA bases in the GGA core and the bases ⁻¹C and ⁴A. Mutations of the relevant base pair were always performed to the other member of the DNA base family, i.e. A to G and T to C. Additionally, we performed a DNA binding simulation to the GGAA sequence adjacent to the binding site studied in this report so far, which was shown experimentally not to be an active ERG binding site,⁸ to see whether an *in silico* discrimination against this sequence is possible (for schematic representation of DNA-sequences see Figure 6). RMSD plots of the simulations to the core ERG ETS domain DNA-complex crystal structure are shown in Figure 7. The simulations of ERGp55 binding to DNA with mutations in the GGA core showed significantly increased RMSD values compared to the simulation of the WT DNA-sequence (Table 2). In comparison, the mutations flanking the GGA core including the inactive GGAA binding sequence show RMSD values more similar to the

WT simulation (Table 2). The binding distance plots also show that mutations in the GGA core lead to a significant increase in the DNA binding distance of around ~5 Å, while mutations outside the GGA core show a maximum deviation of ~3 Å (Figure S11†). Notably the binding distance profile of the GGAG mutation is almost identical to the WT profile. We also calculated MM-GBSA energies for each system during the period of 75-100 ns for qualitative comparison of the different systems. The results are shown in Table 2. Again it has to be noted that MM-GBSA energies are of qualitative and not quantitative nature.³⁴ It is evident that all the mutations resulted in a strong decrease in the DNA-binding energy. Interestingly the mutants flanking the GGA core, where the ETS domain DNA complexes were structurally rather similar to the WT ERGp55 DNA-complex according to RMSD analysis, also show drastically decreased binding energies.

This includes the GGAG mutant, which showed a highly similar binding distance profile compared to the WT simulation. To determine the statistical significance of our results we repeated the WT ERGp55 simulations to a total number of four repeats, each starting with a different random seed for the initial velocities. The RMSD plots to the core ETS domain DNA-complex (Figure 8) show that repeats 3 and 4 show geometries within a RMSD difference of ~1 Å to our tested model system (repeat 1), while repeat 2 exhibits a much higher RMSD value (approximately ~2.6 Å higher). MM-GBSA energies are shown in Table 2. Based on the criteria of binding energy and geometric resemblance to the crystal structure we can discriminate repeats 1 and 3 from all the mutant (both ERG and DNA mutants) simulations. Repeat 4 shows geometric similarity but weaker DNA-binding, while repeat 2 clearly failed to reproduce those criteria.

In summary we were able to discriminate all the mutants (ERG amino acid and DNA sequence) against our validated ERGp55 system (repeat 1) using the criteria of binding distance, RMSD and MM-GBSA binding energy. The simulations of ERG binding to DNA mutants outside the GGA core, which show less strict sequence discrimination experimentally, show higher geometric similarity to our validated ERGp55 system in comparison to mutants in the GGA core with strong experimental sequence preference.²⁰ This is not true for the calculated DNA-binding energies, which do not show a consistent trend.

Figure 7 Backbone RMSD plots to the core ETS domain (amino acid residues 309-391) DNA-complex crystal structure¹⁸ (terminal base pairs excluded, i.e. using 10 of 12 bp) during 100 ns of simulation for the simulations containing mutations of Arg385 and the DNA sequence.

Figure 8 Backbone RMSD plots to the core ETS domain (amino acid residues 309-391) DNA-complex crystal structure¹⁸ (terminal base pairs excluded, i.e. using 10 of 12 bp) during 100 ns of simulations of the repeats of the WT ERGp55 DNA-binding simulation.

Based on the repetitions carried out for the WT simulation, we assume a ~50% confidence level for predictions deduced from this methodology. This coarse estimate originates from the fact that two of the four repeats can be discriminated from all the mutant simulations using the criteria of geometric similarity and binding energy. This confidence level relates to the validity of DNA-docking simulations for protein DNA-complexes without a known experimental DNA-complex structure regarding the criteria of geometric similarity and DNA-binding energy. Obviously significantly more statistics of protein DNA-docking simulations would be necessary to obtain a more rigorous value. We wish to highlight that we expect a significantly higher confidence level for our results from the DNA-binding simulation of the ERGp55 DNA-complex (repeat 1), since we performed various validation studies relating to experimental data. Therefore we attach a high degree of confidence that this simulation represents a highly qualified model of the actual system.

Sequence Recognition by the ERG ETS Domain

One of the major open questions in the field of ETS transcription factors is how DNA-sequence recognition is mediated. ERG has been shown to have a strong preference for the ¹G²G³A core sequence with a less strong preference for ²C¹C upstream and ⁴(A/T) downstream of the core motif.²⁰ The mechanistic details of DNA-recognition especially outside the GGAA core motif are however not completely clear.

Sequence Recognition through Hydrogen Bonding. In order to address this question, we monitored the hydrogen bonding interactions formed by residues Arg367, Arg370 and Tyr371, which are considered to be the main amino acids involved in sequence recognition of ETS transcription factors⁴ during the 500 ns of the ETS domain DNA-complex simulations and the time period of 75-200 ns of our ERGp55 simulation, which we extended to 200 ns for the purpose of more rigorous statistical sampling of the unbiased ERGp55 simulation (Table 3, for 200 ns binding distance and RMSD plots see Figure S8[†]). As can be seen in Figure S8[†] the complex remained perfectly stable during the second 100 ns retaining the high geometric similarity to the crystal structure of the ETS domain DNA-complex. The average binding distance during the 75-200 ns

simulation period is 10.7 Å exactly as in the crystal structure, while the average RMSD remained at 2.1 Å.

In the ERGp55 system we did not detect any hydrogen bonds to the base pairs ³A and ⁴A. Arg367 and Arg370 each formed interactions with both ¹G and ²G in the DNA major groove, which indicates sequence specific interactions. Furthermore, Arg370 forms nonspecific interactions with the DNA-backbone. Tyr371 forms hydrogen bonds to the phosphate backbone two basepairs downstream of the GGAA core during approximately half of the period of observation of the ERGp55 simulation. The sequence recognition of ¹G and ²G by Arg367 and Arg370 corresponds to previous studies on ETS transcription factors.^{4, 16}

The orientation of the amino acid residues involved in DNA sequence recognition (Arg367 and Arg370) in the ERGp55 simulation deviate, however, significantly from the crystal structure. In the crystal structure Arg367 forms a hydrogen bond with ²G and a cation- π interaction to ¹G (Figure S13[†]).^{16, 18, 40} Our simulations show that Arg367 is orientated perpendicular to the plane of the DNA bases, which enables hydrogen bonds with the oxo-groups of both guanine bases simultaneously. The orientation of Arg370 also deviates from the crystal structure. In the full-length ERGp55 simulation we detect hydrogen bonds formed by Arg370 to the backbone at ¹G and groove interactions with ¹G and ²G (Table 3). In the crystal structure Arg370 interacts only with ²G in the DNA major groove. In the ETS domain DNA-complex simulations, Arg367 plays the major role in hydrogen bonding to the DNA-major groove, forming interactions to ¹G, ²G and ¹C (missing in ERGi system) with durations decreasing in that order. The hydrogen bonds to ¹C even though present only for 13-20% of our simulations might enhance the sequence specificity at the -1 position relative to the GGAA core. Arg370 mainly forms interactions to the phosphate backbone, which could explain its energetic importance in the DNA-binding process in comparison to Arg367, since it has been proposed that protein-backbone interactions are the main drivers of the DNA binding affinity for ETS transcription factors.¹⁶ In comparison to the ERGp55 simulation, the interaction of Tyr371 with the phosphate backbone is only observed in the ETS domain DNA-complex simulations of the ERGdna and ERGi system. Additional hydrogen bonds of Tyr371 to DNA were not observed. These results confirm the DNA sequence recognition mechanism as proposed from experiment⁴ of ¹G and ²G by direct hydrogen bonding with Arg367 and in the case of the ERGp55 system, Arg370. Moreover, we detect a hydrogen bonding interaction of Arg367 with ¹C, which might add to the sequence specificity outside the GGAA core. While no data was found explaining the different interactions of Arg370 with DNA, possible explanations might be related to the adjacent regions present in the full-length ERGp55 construct, since it was already shown that the length of the ETS constructs is important for e.g. DNA-binding energies.¹⁸ Another hypothesis is that the unbiased DNA-binding simulation resulted in a slightly different bound state compared to the ETS domain DNA-complex simulations, which all started out in the same position.

Table 3. Duration of hydrogen bonds formed by Arg367, Arg370 and Tyr371 to the GGAA core motif and adjacent bases ^a.

	-1 C ^s /G ^a	1 G ^s /C ^a	2 G ^s /C ^a	6 T ^s /A ^a		-1 C ^s /G ^a	1 G ^s /C ^a	2 G ^s /C ^a	6 T ^s /A ^a
ERGP55 R367		0.86 ^{s,g}	0.51 ^{s,g}		ERGDna R367	0.16 ^{s,g}	0.68 ^{s,g}	0.57 ^{s,g}	
ERGP55 R370	0.19 ^{s,b}	0.27 ^{s,g} , 0.17 ^{s,b}	0.11 ^{s,g}		ERGDna R370	0.42 ^{s,b}	0.30 ^{s,b}		
ERGP55 Y371				0.52 ^{a,b}	ERGDna Y371				0.29 ^{a,b}
ERGi R367		0.77 ^{s,g}	0.36 ^{s,g}		ERGimax R367	0.12 ^{s,g} , 0.12 ^{a,g}	0.76 ^{s,g}	0.13 ^{s,g}	
ERGi R370	0.29 ^{s,b}	0.20 ^{s,b}			ERGimax R370	0.61 ^{s,b}	0.12 ^{s,b}		
ERGi Y371				0.17 ^{a,b}	ERGimax Y371				
ERGu R367	0.20 ^{s,g}	0.98 ^{s,g}	0.50 ^{s,g}		ERGumax R367	0.13 ^{s,g}	0.97 ^{s,g}	0.50 ^{s,g}	
ERGu R370	0.46 ^{s,b}	0.39 ^{s,b}			ERGumax R370	0.58 ^{s,b}			
ERGu Y371					ERGumax Y371				

^a Hydrogen bonds were recorded during the period between 75-200 ns for ERGP55 and 0-500 ns for the other systems shown. The number corresponds to the fraction of time that the longest interaction between two atoms of all the different interactions that have been detected between the two residues has been present. 's' denotes an interaction with the sense strand and 'a' with the antisense strand. 'g' denotes an interaction with the base in the major groove of DNA, while 'b' denotes an interaction with the DNA backbone. DNA bases not listed here did not interact with the ERG ETS domain.

We did not detect any sequence specific interactions for Tyr371 and DNA-base pairs ³A, ⁴A and ⁻²C, ⁻¹C. Therefore, we decided to analyze other mechanisms of DNA-sequence recognition.

Sequence Recognition through Bridging Hydrogen Bonds. To further examine the DNA-sequence recognition by the ERG ETS domain, we assessed bridging hydrogen bonds, i.e. hydrogen bonds that are mediated through a water molecule. Such interactions have been proposed to add to sequence specificity in the field of ETS transcription factors by Cooper et al.,¹⁶ who found a conserved network of water molecules in crystal structures of ETS domain DNA-complexes (Etv5 and Fev). Those interactions were proposed to explain the sequence recognition of the ⁻¹C and potentially ⁻²C base. We performed hydrogen-bonding analysis and searched for interactions, which are 'bridged' via a water molecule. The analysis showed that the amino acids Asp363, Arg367 and Arg370 indeed form those interactions with base pairs in the DNA major groove (Table S3[†]). However, the results were quite heterogeneous and therefore we only report bridged hydrogen bonds that were found in at least 4 of the 5 systems (during the period 75-200 ns of ERGP55 and 0-500 ns of ERGDna, ERGi, ERGu, ERGumax). To classify these interactions, we calculated occupancies, which were defined as the number of bridged hydrogen bonds formed divided by the entire number of frames used for analysis. Arg370 forms a bridged hydrogen bond with ¹G, which is present in 4 out of 5 systems and the average occupancy of that interaction is 31%. Interestingly this interaction was detected in all ETS domain DNA-complex simulations and not in the ERGP55 DNA-binding simulation. Since we observed earlier that Arg370 occupies a position in which it hydrogen bonds mainly to the phosphate backbone (Table 3) in the ETS domain DNA-complex simulations, this bridged interaction might add to its role in the sequence recognition in the DNA major groove of ERG. The two different interactions observed represent another hypothesis as to why the hydrogen bond interactions of Arg370 in the DNA major groove were not observed in the ETS

domain DNA-complex simulations. Depending on the exact position of Arg370 it might be possible that either direct or bridged hydrogen bonds to the DNA major groove are formed. A transition between the two states was however not observed in our simulations.

Asp363 forms bridged hydrogen bonds with ⁻²C (4 of 5 systems, average occupancy 10%) and ⁻¹C (all systems, average occupancy 21%) on the sense strand and with ¹C (all systems, average occupancy 46%) and ²C (all systems, average occupancy 59%) on the antisense strand. Since hydrogen-bonding analysis does not give a detailed geometric representation, we performed a grid analysis to show water density throughout our simulations, to assess whether the results of the hydrogen bonding analysis are sensible for Asp363. Therefore, we analyzed a 500 ns ERGu simulation and a simulation of the same system in which Asp363 was mutated to alanine (mERGu). We chose to use an ETS domain DNA complex in favor of the full-length system, since this enabled more conformational sampling due to the higher simulation speed of a smaller system and therefore a more converged water density. We calculated water densities around the DNA-duplex using gridded histogram analysis and then subtracted the water density of the aspartate mutant from the WT system, to remove the bulk water density, which is generated through unspecific interactions of water with e.g. the phosphate backbone. The results show that we indeed see water densities in positions that are compatible with the geometric criteria that were employed throughout this study to detect bridged hydrogen bonds (Figure 9). This represents another indication for bridged hydrogen bonds between Asp363 and DNA. While there is still other water density present, the relevant densities are among the most dominant ones. It is very interesting that we observe a very similar phenomenon as proposed from crystallographic data for ERG ETS homologues in our simulations, which highlights the potential importance of Asp363 and bridged hydrogen bonding in the sequence recognition of base pairs inside and outside the GGAA core sequence.

Figure 9 The figure shows the structure of the ERGu ETS domain DNA-complex after 200 ns of simulation with the green mesh depicting water densities obtained from gridded histogram analysis as described in the text. The water molecules were placed manually into the density for better visualization and the black dashed lines show potential bridged hydrogen bonds between Asp363 and ^{-2}C , ^{-1}C , ^{1}C and ^{2}C .

A detailed analysis illustrating how bridged hydrogen bonds contribute to sequence recognition for ETS factors has been given by Cooper et al.¹⁶ While performing the same analysis is not possible in our simulation due to the high complexity and the dynamic nature of the solvent shell in our simulations, we performed a bridged hydrogen bond analysis for the ERGp55 DNA-binding simulation to the adjacent GGAA sequence in the ICAM-1 promoter (time period 75–100 ns, termed inactive EBS in Figure 6). The DNA bases at the -1 and -2 position relative to that GGAA core sequence are both thymine bases, which have been shown to have negligible binding specificity compared to both cytosine bases as in the -181 ICAM-1 EBS mainly studied here.²⁰ In this simulation no bridged hydrogen bonds were detected between Asp363 and DNA, which is an indication for the sequence specificity of the bridged hydrogen bonds detected in the -181 EBS system.

To assess whether Asp363 plays an important energetic role in the DNA-binding process we performed alanine-scanning mutagenesis on the full-length ERGp55 system as described earlier. The binding energy of the mutated trajectory is -100.3 kcal/mole (sd +/- 11.1 kcal/mole) in comparison to the wild type energy of -97.5 kcal/mole (sd +/- 11.2 kcal/mole). MM-GBSA calculations of ERGu (Figure S14†) and mERGu indicate a tighter binding of ERGu, the data is however very variant so a clear conclusion, whether Asp363 is energetically relevant is not possible. Experimental studies also showed that mutation of the homologue aspartate residue to alanine in ETV1 did not result in a significant difference in the DNA-binding affinity.¹⁶ Two possible explanations, which save the role of Asp363 in sequence recognition despite its lack of contribution to the DNA-binding energy are that it has been shown that the energetic importance of residues, which confer sequence specificity is small compared to the residues which are essential for DNA-binding in ETS factors¹⁶. Moreover the proximity of the negatively charged aspartate to the phosphate backbone (Figure 9) might contribute unfavorable free energy, which balances with the free energy of the water-mediated hydrogen bonding. The latter hypothesis in combination with the absence of the bridged hydrogen in the ERGp55 inactive EBS DNA-binding simulation could be an

explanation as to why ERGp55 binds to that particular GGAA sequence in the -181 region of the ICAM-1 promoter.

Conclusions

ERG DNA-Binding Simulations. While molecular dynamics has been used frequently to study protein-DNA-interactions (see for example ref. 3 and references therein), the conventional approach is to start simulations in the DNA bound state. Our simulations, which started in an unbound configuration, reproduced the experimental ERG ETS domain DNA-crystal structure and experimental data for the DNA binding process of ERG with remarkable detail. It is worth emphasizing that apart from the structural arrangement no external bias, as exemplified by targeted-MD or umbrella sampling techniques, was used. This approach allowed us to investigate previously unattainable protein-DNA interactions e.g. the salt bridge between Arg385 and DNA, which is not covered by existing experimental structures.¹⁶ We are aware that this interaction can also be studied by performing simulations of the bound crystal structure with an elongated DNA-fragment, we decided, however, to follow the approach described in detail in this article for two main reasons. First, performing unbiased DNA-binding simulations along with extensive validation to different experimental data reduces the model bias to the input data and also allows for further comparison with simulations of the bound protein DNA-complex. Second and more important to us was to showcase the detail which the unbiased DNA-binding simulation, along with the methodology used and developed throughout this study, reproduced experimental data and allowed us to make predictions, which originated from the different analysis strategies. Another noteworthy example of the success of unbiased MD simulation is a study on the self-assembly of lipid bilayers.⁴¹ While no obvious benefit or findings related to using a full length structure were detected so far, we showed where possible that our system represents a highly qualified model and we will be making use of the entire ERGp55 construct in follow up studies. To our knowledge this is the first time that the structure of a protein-DNA complex was obtained *de novo* with high accuracy using this kind of molecular dynamics docking approach. It is possible that extending this methodology to other transcription factors with existing crystal structures for the apo protein and known DNA binding sequence could guide experiments by proposing an initial model for the protein-DNA complex. Furthermore the accuracy of future models will be enhanced empowered through the drastic acceleration of molecular dynamics code throughout the last years,²³ which enables longer simulation timescales. For example increasing the simulation timescales by two orders of magnitude would make it feasible to sample sliding motions of transcription factors along DNA *in silico*⁴², allowing the sampling of a significantly higher amount of phase space. Those considerations highlight that this approach has the potential to be an important application of molecular dynamics simulations in the field of protein-DNA interactions.

DNA Sequence Recognition by ERG. The analysis of hydrogen bonds formed by Arg367 throughout the simulations carried out in this study reproduced their known role in the sequence recognition of ETS transcription factors through interactions with ¹G and ²G, while the interactions formed by Arg370 were only detected in the unbiased ERGp55 full length simulation.⁴ Additionally we examined the possibility of water-bridged hydrogen bonds in sequence recognition, which was recently proposed for other ETS transcription factors.¹⁶ Our simulations show very similar bridged hydrogen bonds formed by Asp363 and Arg370 to base pairs in the DNA major groove. These results enforce the recently proposed role of water-mediated hydrogen bonding in the sequence recognition of ETS transcription factors and encourage experimental studies elucidating the relevance of Asp363.

In conclusion, the ERG DNA-binding simulation represents a proof of concept study that shows, that MD can deliver significant insights for protein-DNA interactions. We report a detailed profile for Arg385 indicating that it plays an essential role in the formation of the ERG DNA-complex. Moreover, we propose a novel role for bridged hydrogen bonding in the DNA-sequence recognition of ERGp55 in and outside the GGA core sequence, similar to what has been proposed recently for other ETS factors.¹⁶

Acknowledgements

M.G.B. would like to acknowledge funding from the German Academic Exchange Service (DAAD).

Notes and References

- 1 B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts and P. Walter, *Molecular Biology of the Cell*, Garland Science, 2014.
- 2 M. M. Garner and A. Revzin, *Nucleic acids research*, 1981, **9**, 3047-3060.
- 3 A. van der Vaart, *Biochimica et biophysica acta*, 2015, **1850**, 1091-1098.
- 4 P. C. Hollenhorst, L. P. McIntosh and B. J. Graves, *Annual review of biochemistry*, 2011, **80**, 437-471.
- 5 B. M. Degnan, S. M. Degnan, T. Naganuma and D. E. Morse, *Nucleic acids research*, 1993, **21**, 3479-3484.
- 6 G. M. Birdsey, A. V. Shah, N. Dufton, L. E. Reynolds, L. Osuna Almagro, Y. Yang, I. M. Aspalter, S. T. Khan, J. C. Mason, E. Dejana, B. Gottgens, K. Hodivala-Dilke, H. Gerhardt, R. H. Adams and A. M. Randi, *Developmental cell*, 2015, **32**, 82-96.
- 7 G. M. Birdsey, N. H. Dryden, A. V. Shah, R. Hannah, M. D. Hall, D. O. Haskard, M. Parsons, J. C. Mason, M. Zvelebil, B. Gottgens, A. J. Ridley and A. M. Randi, *Blood*, 2012, **119**, 894-903.
- 8 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. DOI: 10.1039/C6MB00506C
- 9 A. Sperone, N. H. Dryden, G. M. Birdsey, L. Madden, M. Johns, P. C. Evans, J. C. Mason, D. O. Haskard, J. J. Boyle, E. M. Paleolog and A. M. Randi, *Arteriosclerosis, thrombosis, and vascular biology*, 2011, **31**, 142-150.
- 10 F. McLaughlin, V. J. Ludbrook, J. Cox, I. von Carlowitz, S. Brown and A. M. Randi, *Blood*, 2001, **98**, 3332-3339.
- 11 F. McLaughlin, V. J. Ludbrook, I. Kola, C. J. Campbell and A. M. Randi, *Journal of cell science*, 1999, **112** (Pt 24), 4695-4703.
- 12 L. Yuan, V. Nikolova-Krstevski, Y. Zhan, M. Kondo, M. Bhasin, L. Varghese, K. Yano, C. V. Carman, W. C. Aird and P. Oettgen, *Circulation research*, 2009, **104**, 1049-1057.
- 13 J. H. Martens, *The international journal of biochemistry & cell biology*, 2011, **43**, 1413-1416.
- 14 S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X. W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J. E. Montie, R. B. Shah, K. J. Pienta, M. A. Rubin and A. M. Chinnaiyan, *Science*, 2005, **310**, 644-648.
- 15 S. Carrere, A. Verger, A. Flourens, D. Stehelin and M. Duterrque-Coquillaud, *Oncogene*, 1998, **16**, 3261-3268.
- 16 C. D. Cooper, J. A. Newman, H. Aitkenhead, C. K. Allerston and O. Gileadi, *The Journal of biological chemistry*, 2015, **290**, 13692-13709.
- 17 C. D. Cooper, J. A. Newman and O. Gileadi, *Biochemical Society transactions*, 2014, **42**, 130-138.
- 18 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.
- 19 F. D. Karim, L. D. Urness, C. S. Thummel, M. J. Klemsz, S. R. Mckercher, A. Celada, C. Vanbeveren, R. A. Maki, C. V. Gunther, J. A. Nye and B. J. Graves, *Genes & development*, 1990, **4**, 1451-1453.
- 20 G. H. Wei, G. Badis, M. F. Berger, T. Kivioja, K. Palin, M. Enge, M. Bonke, A. Jolma, M. Varjosalo, A. R. Gehrke, J. Yan, S. Talukder, M. Turunen, M. Taipale, H. G. Stunnenberg, E. Ukkonen, T. R. Hughes, M. L. Bulyk and J. Taipale, *The EMBO journal*, 2010, **29**, 2147-2160.
- 21 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.
- 22 V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins*, 2006, **65**, 712-725.
- 23 R. Salomon-Ferrer, A. W. Gotz, D. Poole, S. Le Grand and R. C. Walker, *J Chem Theory Comput*, 2013, **9**, 3878-3888.
- 24 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.
- 25 T. J. Macke and D. A. Case, in *Molecular Modeling of Nucleic Acids*, American Chemical Society, 1997, vol. 682, ch. 24, pp. 379-393.

PAPER

Molecular BioSystems

View Article Online
DOI: 10.1039/C6MB00506C

- 26 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.
- 27 K. Degitz, L. Lianjie and S. W. Caughman, *Journal of Biological Chemistry*, 1991, **266**, 14024-14030.
- 28 D. R. Roe and T. E. Cheatham, *J Chem Theory Comput*, 2013, **9**, 3084-3095.
- 29 X. J. Lu and W. K. Olson, *Nucleic acids research*, 2003, **31**, 5108-5121.
- 30 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.
- 31 L. A. Kelley, S. P. Gardner and M. J. Sutcliffe, *Protein Eng*, 1996, **9**, 1063-1065.
- 32 B. R. Miller, T. D. McGee, J. M. Swails, N. Homeyer, H. Gohlke and A. E. Roitberg, *J Chem Theory Comput*, 2012, **8**, 3314-3321.
- 33 J. Srinivasan, T. E. Cheatham, P. Cieplak, P. A. Kollman and D. A. Case, *Journal of the American Chemical Society*, 1998, **120**, 9401-9409.
- 34 S. Genheden and U. Ryde, *Expert opinion on drug discovery*, 2015, **10**, 449-461.
- 35 V. Tsui and D. A. Case, *Biopolymers*, 2001, **56**, 275-291.
- 36 G. D. Hawkins, C. J. Cramer and D. G. Truhlar, *J Phys Chem-Us*, 1996, **100**, 19824-19839.
- 37 G. D. Hawkins, C. J. Cramer and D. G. Truhlar, *Chem Phys Lett*, 1995, **246**, 122-129.
- 38 I. Massova and P. A. Kollman, *Journal of the American Chemical Society*, 1999, **121**, 8133-8143.
- 39 V. Laudet, C. Hanni, D. Stehelin and M. Duterrque-Coquillaud, *Oncogene*, 1999, **18**, 1351-1359.
- 40 K. A. Wilson, J. L. Kellie and S. D. Wetmore, *Nucleic acids research*, 2014, **42**, 6726-6741.
- 41 A. A. Skjevik, B. D. Madej, C. J. Dickson, C. Lin, K. Teigen, R. C. Walker and I. R. Gould, *Physical Chemistry Chemical Physics*, 2016, **18**, 10573-10584.
- 42 E. G. Marklund, A. Mahmutovic, O. G. Berg, P. Hammar, D. van der Spoel, D. Fange and J. Elf, *Proceedings of the National Academy of Sciences of the United States of America*, 2013, **110**, 19796-19801.