Supplementary Table 1: Frequency of arginine-minor groove contacts among protein superfamilies.

SCOP superfamily	Total # of chains	% Contacting minor groove	% Contacting narrow minor groove	% Arginine contacting minor groove	% Arginine contacting narrow minor groove
C-terminal domain of RNA polymerase alpha subunit	2	100.0	100.0	100.0	100.0
ACT-like	2	100.0	100.0	100.0	100.0
TrpR-like	1	100.0	100.0	100.0	100.0
SRF-like	17	94.1	70.6	94.1	70.6
IHF-like DNA-binding proteins	12	100.0	66.7	100.0	66.7
Histone-fold	232	92.7	76.3	92.2	66.4
DNA breaking-rejoining enzymes	58	82.8	81.0	67.2	62.1
Zn2/Cys6 DNA-binding domain	14	42.9	42.9	42.9	42.9
Homeodomain-like	89	70.8	46.1	67.4	39.3
p53-like transcription factors	70	50.0	37.1	44.3	37.1
Putative DNA-binding domain	5	80.0	20.0	40.0	20.0
lambda repressor-like DNA-binding domains	41	46.3	17.1	26.8	17.1
Winged helix DNA-binding domain	94	55.3	21.3	35.1	10.6
Leucine zipper domain	44	9.1	9.1	9.1	9.1
C-terminal effector domain of the bipartite response regulators	22	9.1	9.1	9.1	9.1
Restriction endonuclease-like	83	91.6	9.6	8.4	4.8
Glucocorticoid receptor-like (DNA-binding domain)	42	40.5	7.1	23.8	4.8
N-terminal domain of MutM-like DNA repair proteins	3	100.0	0.0	100.0	0.0
SPOC domain-like	2	100.0	0.0	100.0	0.0
DNA-binding domain of intron-encoded endonucleases	2	100.0	0.0	100.0	0.0
Holliday junction resolvase RusA	1	100.0	0.0	100.0	0.0
P-loop containing nucleoside triphosphate hydrolases	9	77.8	0.0	77.8	0.0
DNA repair protein MutS, domain I	18	66.7	0.0	61.1	0.0
Hedgehog/intein (Hint) domain	2	100.0	100.0	50.0	0.0
HMG-box	2	100.0	0.0	50.0	0.0
Origin of replication-binding domain, RBD-like	17	29.4	23.5	29.4	0.0
DNA/RNA polymerases	84	98.8	2.4	27.4	0.0
Eukaryotic DNA topoisomerase I, N-terminal DNA-binding fragment	11	27.3	0.0	27.3	0.0
His-Me finger endonucleases	8	100.0	0.0	25.0	0.0
Viral DNA-binding domain	8	37.5	0.0	25.0	0.0
Ribonuclease H-like	24	20.8	0.0	12.5	0.0
TATA-box binding protein-like	34	100.0	0.0	2.9	0.0
Homing endonucleases	29	79.3	51.7	0.0	0.0
Group I mobile intron endonuclease	2	100.0	50.0	0.0	0.0
Nucleotidyltransferase	8	100.0	0.0	0.0	0.0

DNA-binding protein LAG-1 (CSL)	3	100.0	0.0	0.0	0.0
Replication terminator protein (Tus)	2	100.0	0.0	0.0	0.0
DNA ligase/mRNA capping enzyme, catalytic domain	1	100.0	0.0	0.0	0.0
DNA repair protein MutS, domain III	13	61.5	0.0	0.0	0.0
S-adenosyl-L-methionine-dependent methyltransferases	8	50.0	0.0	0.0	0.0
Cyclin-like	3	33.3	0.0	0.0	0.0
beta and beta-prime subunits of DNA dependent RNA-polymerase	11	27.3	0.0	0.0	0.0
HLH, helix-loop-helix DNA-binding domain	14	7.1	0.0	0.0	0.0
Lesion bypass DNA polymerase (Y-family), little finger domain	60	0.0	0.0	0.0	0.0
Ribbon-helix-helix	58	0.0	0.0	0.0	0.0
lambda integrase-like, N-terminal domain	34	0.0	0.0	0.0	0.0
C2H2 and C2HC zinc fingers	27	0.0	0.0	0.0	0.0
E set domains	20	0.0	0.0	0.0	0.0
KorB DNA-binding domain-like	4	0.0	0.0	0.0	0.0
Sigma3 and sigma4 domains of RNA polymerase sigma factors	4	0.0	0.0	0.0	0.0
SMAD MH1 domain	4	0.0	0.0	0.0	0.0
Eukaryotic DNA topoisomerase I, dispensable insert domain	3	0.0	0.0	0.0	0.0
S13-like H2TH domain	3	0.0	0.0	0.0	0.0
DNA topoisomerase I domain	2	0.0	0.0	0.0	0.0
Zinc finger design	2	0.0	0.0	0.0	0.0
RuvA domain 2-like	1	0.0	0.0	0.0	0.0
A DNA-binding domain in eukaryotic transcription factors	1	0.0	0.0	0.0	0.0
DNA-binding domain	1	0.0	0.0	0.0	0.0
DNA polymerase beta, N-terminal domain-like	1	0.0	0.0	0.0	0.0
RNA polymerase	1	0.0	0.0	0.0	0.0
GCM domain	1	0.0	0.0	0.0	0.0

For all DNA-binding SCOP superfamilies[46], the table lists the total number of chains in contact with DNA of at least ten base pairs in length, the percentage of these chains that contact the minor groove, the percentage that contact a narrow minor groove with a groove width of <5.0 Å, the percentage that uses arginine to contact the minor groove within a distance of <6.0 Å , and the percentage that uses arginine to contact a narrow minor groove. SCOP superfamilies with arginine-minor groove contacts are highlighted. Arginine contacts in narrow minor grooves are highlighted in dark blue; arginine contacts in minor grooves of width ≥5.0 Å are in light blue; and any other minor groove contacts are in green.

Supplementary Table 2: Minor groove width at the centre of tetranucleotides in free DNA and protein-DNA structures

a. Tetranucleotides from free DNA structures (sorted by average width)

Sequence	Minor groove width (Å)	Number of occurrences			
AATC	2.5	1	TGTA	5.8	6
ATAA	3.8	1	TGGA	5.8	2
AATT	3.8	34	TAGC	5.9	2
GAAT	3.8	48	CGAA	5.9	20
AGCT	3.8	1	AGAC	5.9	5
AAAT	3.9	11	CGTC	6.0	14
AAAA	3.9	24	AAGA	6.1	3
GATC	4.1	1	TGAA	6.2	2
GAAA	4.3	7	TAGA	6.2	6
GATA	4.4	3	GGAA	6.2	2
ATAT	4.6	5	CTAG	6.3	6
AGAT	4.6	3	AAAC	6.4	6
GCGC	4.9	5	TCGA	6.5	3
AGAA	4.9	4	CGGT	6.6	1
TAAC	5.0	4	CGAC	6.6	9
GAGA	5.0	1	TTAA	6.6	10
TAAT	5.2	4	CCGG	6.6	2
CAAT	5.3	1	GGTA	6.6	2
ACGT	5.3	11	GAGC	6.8	1
TAAG	5.4	1	CGGC	6.9	1
CATA	5.4	1	CAAG	7.1	1
CGTT	5.5	9	GGTT	7.1	4
GTAC	5.5	5	CGAG	7.2	1
AATG	5.6	1	GGCG	7.2	1
CGAT	5.7	1	GGCC	7.5	3
CAAA	5.7	6	TAAA	7.6	6
AGAG	5.7	2	AGGT	8.8	2
CTGT	5.7	6	TGGG	9.4	2
TATA	5.7	4	GGGC	10.1	2
AAGC	5.7	2			

b. Tetranucleotides from protein-DNA structures (sorted by average width)

Sequence	Minor groove width (Å)	Number of occurrences			
AAAT	4.1	77	CAGT	6.5	49
AATA	4.1	33	ATAC	6.5	41
AATC	4.4	28	CAAG	6.5	17
AATT	4.4	55	ATGC	6.5	30
AAAA	4.5	94	TAAC	6.5	29
AAGT	4.5	54	TGTC	6.6	53
GAAT	4.6	30	TTGT	6.6	50
GAAA	4.8	47	TAGT	6.6	17
TAAT	4.8	69	TGAT	6.6	49
AAAC	4.8	35	ACGC	6.7	12
ATAA	4.9	52	CGCG	6.7	7
AGAT	5.1	20	CTGC	6.7	35
AAGA	5.1	27	AAGG	6.7	69
AGTT	5.1	50	GGTA	6.7	26
AGAA	5.2	19	CTGG	6.7	29
AAAG	5.3	44	GGGT	6.8	19
ATAG	5.4	40	TAGC	6.8	41
GAAC	5.5	31	CGAT	6.8	32
CGTT	5.6	25	GGAA	6.8	79
TATA	5.6	29	ATGA	6.8	55
TGTT	5.7	45	AGAG	6.8	12
TGAG	5.7	32	GGTC	6.8	28
ATGT	5.8	50	GTGT	6.9	21
GGAT	5.8	49	CGAA	6.9	46
TAAA	5.8	39	GAGG	6.9	11
AATG	5.8	38	CTGT	6.9	53
CAAT	5.9	35	CAAC	6.9	23
TAGA	5.9	27	TGAC	7.0	59
CATG	6.0	19	ACGT	7.0	23
CATA	6.0	31	AGTG	7.0	47
GATG	6.1	33	TGTG	7.0	43
GATC	6.1	30	GGTG	7.0	20
GTGA	6.1	49	TGGT	7.1	21
AGAC	6.1	47	TAGG	7.1	48
GTAC	6.2	9	GGCA	7.1	31
CAAA	6.2	45	TGTA	7.1	43
TGAA	6.2	31	TGGA	7.2	28
CAGA	6.2	27	GATA	7.2	59
CTAA	6.2	42	CGGT	7.2	27
TTGA	6.2	22	AGCC	7.2	31
GAGA	6.2	17	CAGG	7.2	21
TAAG	6.2	31	TGGC	7.2	34
CGTG	6.3	37	GCGA	7.2	12
TTAA	6.4	23	GTAA	7.2	35
GAGT	6.4	27	AGTA	7.3	14
AGGT	6.4	28	GGAC	7.3	23
GAAG	6.4	77	GTAG	7.3	17
AGTC	6.5	37	GGAG	7.3	13
CTAG	6.5	14	GTGC	7.4	53
GGTT	6.5	16	CTGA	7.5	28
AGGA	6.5	56	ACGG	7.5	16

ATGG	7.5	25	GGCG	7.7	24
GTGG	7.5	38	ATAT	7.7	21
CGTC	7.5	26	GCGG	7.8	13
ACGA	7.5	52	GGCC	7.8	12
TGCA	7.5	16	CGCA	7.8	19
AAGC	7.5	15	AGCA	7.9	46
AGCT	7.5	3	AGGC	7.9	22
GGGC	7.6	33	AGCG	7.9	26
CCGA	7.6	21	CGAC	8.0	24
CAGC	7.6	30	CGGG	8.1	10
TTGC	7.6	30	CGGA	8.2	17
CGAG	7.6	17	GCGC	8.3	23
TGGG	7.6	30	TCGA	8.4	17
GGGA	7.7	55	CCGG	8.5	16
CGGC	7.7	22	CGTA	8.9	37
AGGG	7.7	33	GGGG	9.0	13
GAGC	7.7	20			
TTGG	7.7	19			

(a) The 59 tetranucleotides in free DNA structures are ordered by increasing groove width. The table lists the tetranucleotide sequence, minor groove width on average, and number of occurrence of each tetranucleotide in the free DNA dataset.

(b) The 136 unique tetranucleotides in protein-DNA complexes are ordered by the same criterion. The table lists the tetranucleotide sequence, minor groove width on average, and number of occurrence of each tetranucleotide in the protein-DNA dataset.

Supplementary Table 3: Statistics of hydrogen bonds between arginine and lysine side chains with DNA

a. Arginine – Nonredundant dataset

	DEFAULT CRITERIA		RELAXED CRITERIA	
	Average	Std. Dev.	Average	Std. Dev.
protein side-chain to DNA bases:	0.63	0.62	1.37	1.15
protein side-chain to DNA backbone:	0.29	0.39	1.32	0.84
protein side-chain to DNA:	0.92	0.92	2.70	1.67

b. Lysine – Nonredundant dataset

	DEFAULT CRITERIA		RELAXED CRITERIA	
	Average	Std. Dev.	Average	Std. Dev.
protein side-chain to DNA bases:	0.37	0.50	0.71	0.74
protein side-chain to DNA backbone:	0.24	0.38	0.90	0.77
protein side-chain to DNA:	0.61	0.63	1.61	1.11

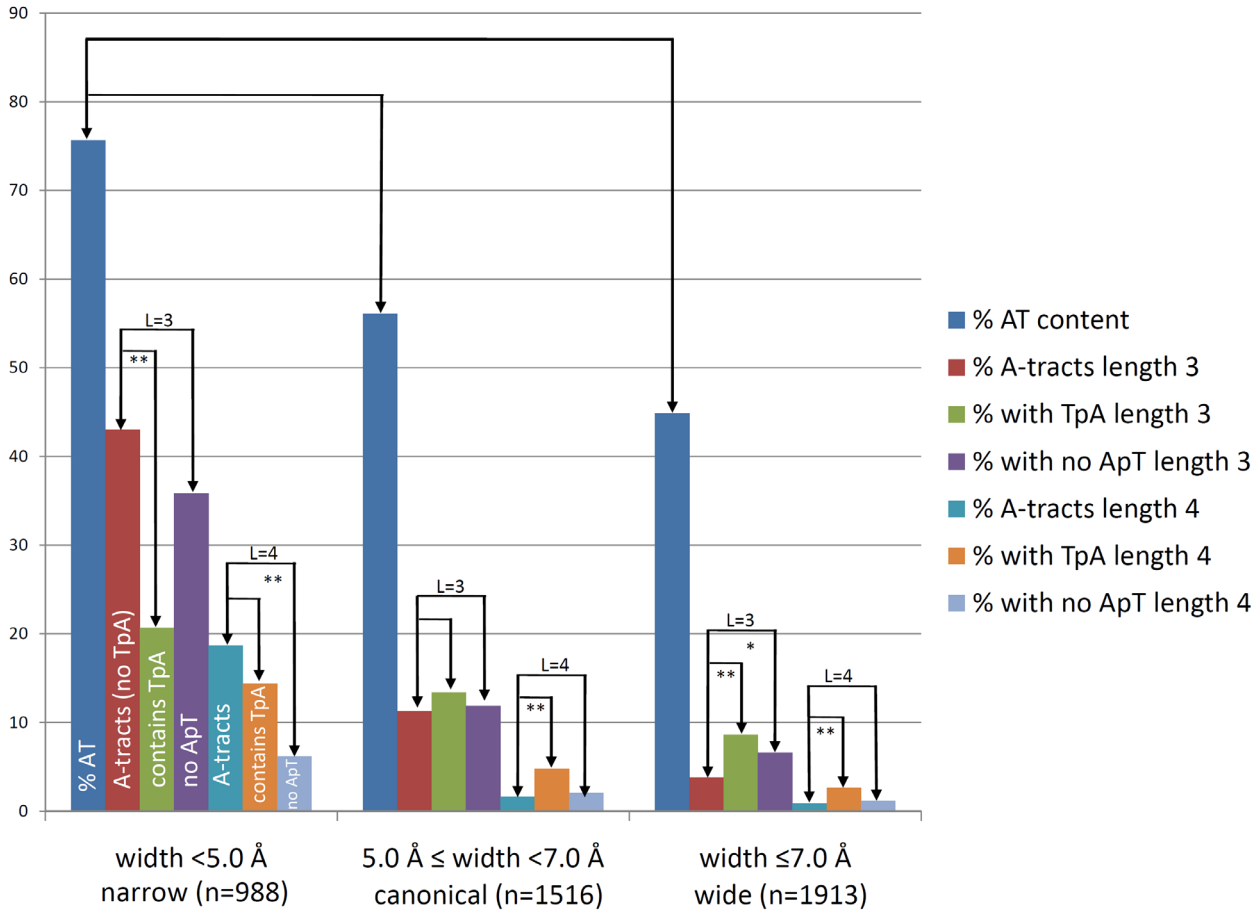
The number of hydrogen bonds for (a) 796 arginines and (b) 276 lysines that contact the minor groove in 392 protein-DNA complexes. The number of hydrogen bonds counted using the HBplus program[48] with default criteria do not show a significant difference between arginine and lysine. Relaxing the criteria by increasing the maximum acceptor-hydrogen donor distance from 2.5 to 3.25 Å and decreasing the minimum acceptor-hydrogen-donor angle for allowed hydrogen bonds from 90° to 80° shows an enrichment of arginine hydrogen bonds in comparison to lysine. However, due to large standard deviations the difference between both side chains is still not significant.

Supplementary Table 4: Transfer free energies of ionized arginine and lysine side chains

RESIDUE (ε=80/ε=2)	AMBER94	CHARMM	OPLS	PARSE
LYSINE SIDE CHAIN (kcal mol^{-1})	36.53	39.27	36.98	41.10
ARGININE SIDE CHAIN (kcal mol^{-1})	34.24	34.58	30.39	35.20
LYSINE vs. ARGININE (kcal mol^{-1})	2.29	4.69	6.59	5.90

Changes in free energy for the transfer of arginine and lysine side chains from water (dielectric constant ε=80) to a low dielectric medium (ε=2) were calculated with DelPhi[31,45] using four different force fields, AMBER94[50], CHARMM[51], OPLS[52], and PARSE[53]. The difference between the transfer free energies of arginine and lysine consistently indicate a higher desolvation cost for lysine in comparison to arginine.

Supplementary Figure 1: Statistical analysis of tetranucleotides present in protein-DNA complexes



The 4,426 tetranucleotides present in protein-DNA complexes in the PDB, for which groove geometry could be analysed based on Curves[44,54] at all three base pair steps, were classified according to their minor groove width. Tetranucleotides were classified as narrow (<5.0 Å, n=988), canonical (≥5.0 Å, <7.0 Å, n=1,514), or wide (≥7.0 Å, n=1,912). The graph represents several parameters that were calculated for these three groups of tetranucleotides, defined as follows:

% AT: The percentage of AT base pairs.

"A-tracts": The percentage of tetranucleotides that contain an A-tract of length three (L=3) (AAA, AAT, ATT, TTT) or of length four (L=4) (AAAA, AAAT, AATT, ATTT, TTTT).

"contains TpA": The percentage of tetranucleotides that contain a three or four nucleotide sequence composed only of AT base pairs that includes a TpA step (thus excluding A-tracts). For length three, these include: ATA, TAA, TAT, TTA. For length four, these include: ATAA, ATAT, ATTA, AATA, TAAA, TAAT, TATA, TTAA, TATT, TTAT, TTTA.

"no ApT": The percentage of tetranucleotides that contain a three or four nucleotide sequence composed only of AT base pairs but do not include an ApT step. This definition is useful because it includes the same number of sequences as the A-tract definition. For length three, these include: AAA, TAA, TTA, TTT. For length four, these include: AAAA, TAAA, TTAA, TTTA, TTTT.

The direct comparison between A-tracts (with no TpA steps) and AT sequences without ApT steps was chosen because the ApT step is in structural and energetic terms the opposite of the TpA 'hinge' step. Base stacking stabilizes ApT steps whereas TpA steps are flexible due to the small overlap between base pairs, a structural difference that was reported to result in TpA step melting 20 K below the ApT step[28].

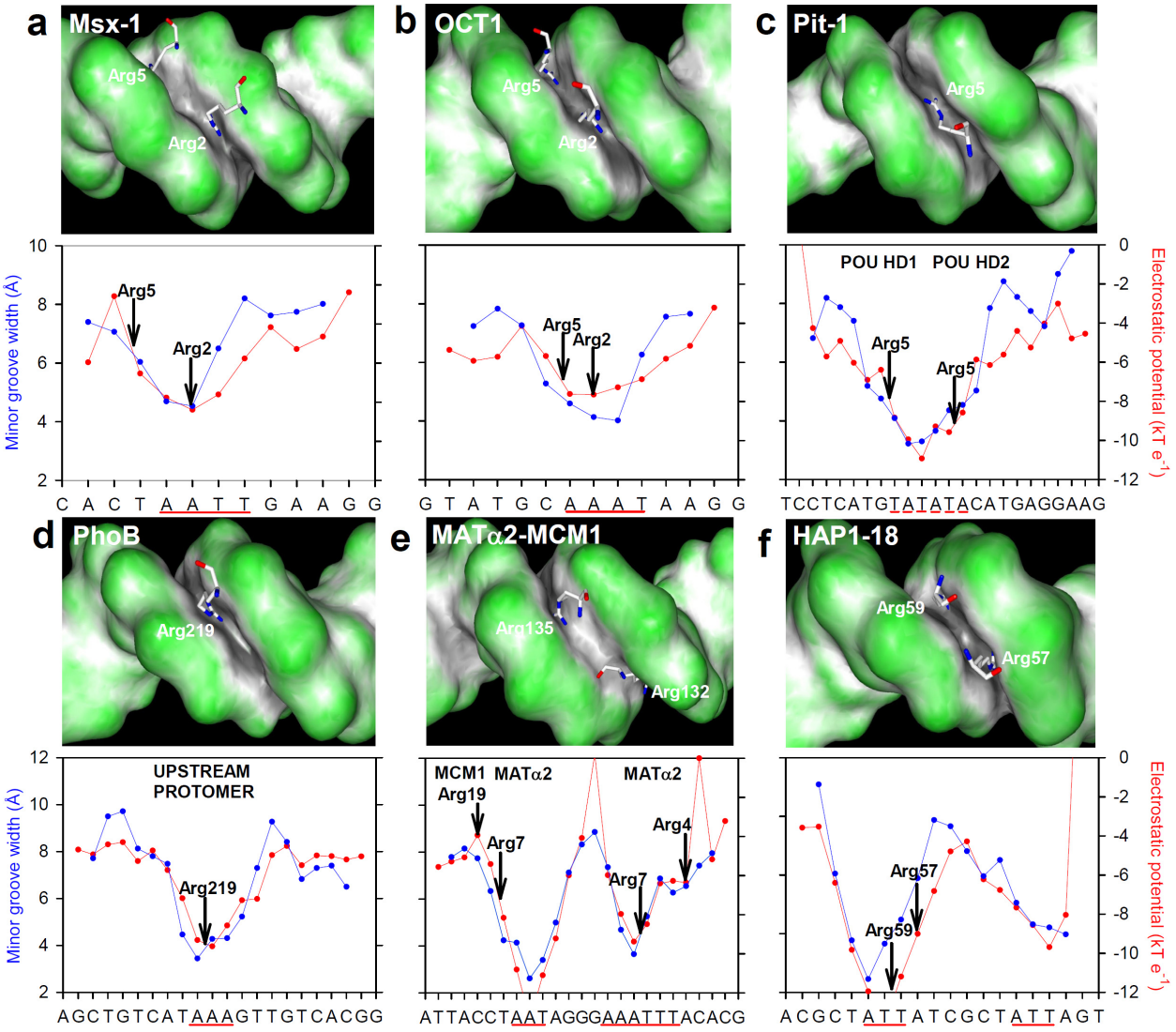
From these data, we conclude:

1) Compared to sequences that do not have narrow minor grooves, sequences with narrow minor grooves tend to be AT-rich.

2) A-tracts are significantly enriched in sequences with narrow minor grooves. This enrichment is seen for both A-tracts of length three and of length four as observed from the following pair wise comparisons: A-tracts vs. "contains TpA" and A-tracts vs. "no ApT" (indicated by the arrows). Statistically significant comparisons are indicated by asterisks (** $p<0.001$; * $p<0.05$). These comparisons show that there is only an enrichment of A-tracts in the narrow group of sequences.

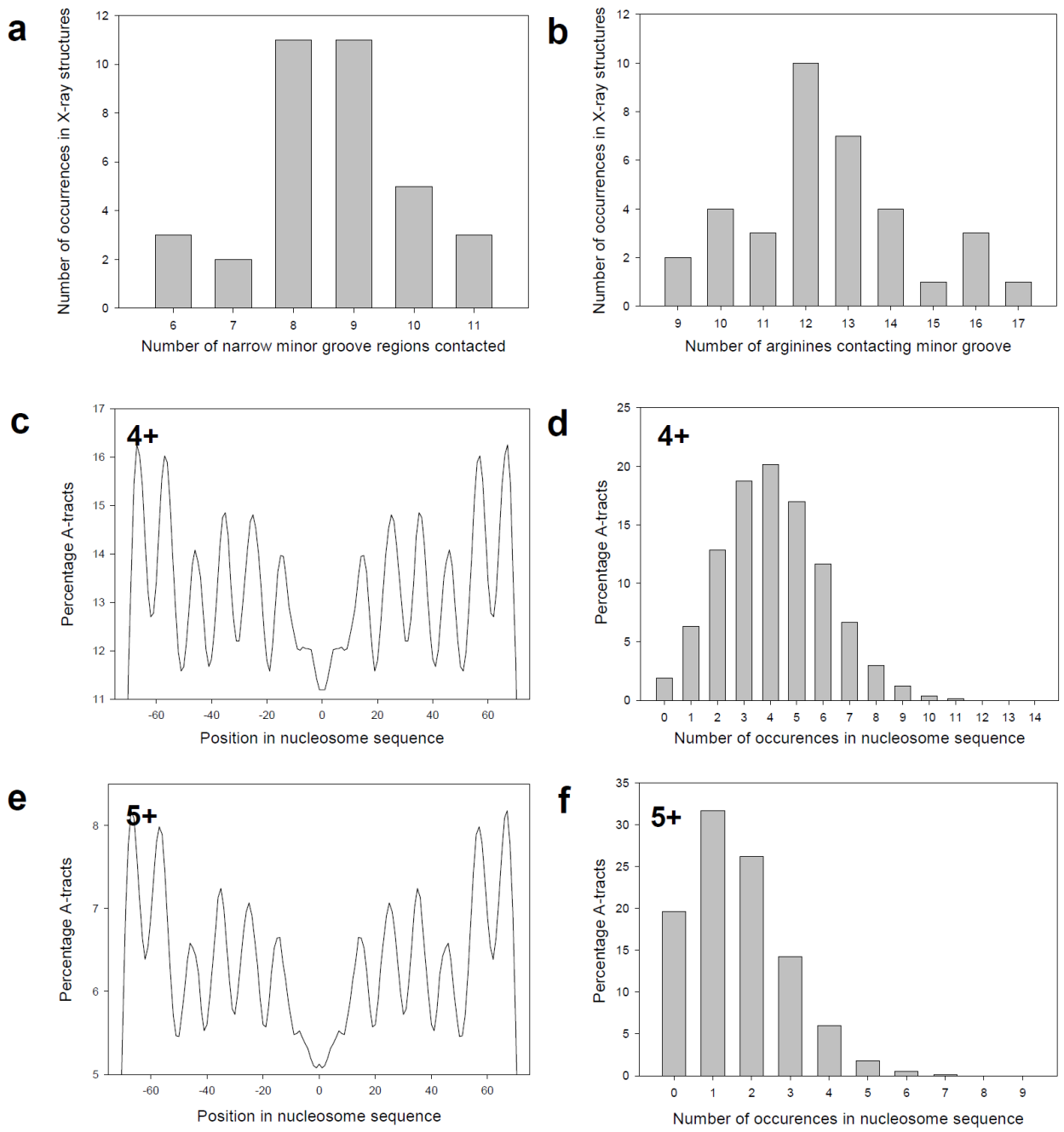
Statistical significance was determined using Fisher's exact test.

Supplementary Figure 2: Additional examples of minor groove shape recognition by arginines



The recognition of arginine residues in narrow minor groove regions is illustrated for representative examples, the three homeodomain proteins (a) Msx-1 1ig7[55], (b) OCT1 POU 1oct[56], and (c) Pit-1 POU 1au7[57], and (d) the PhoB response regulator 1gxp[58], (e) the MATα2–MCM1 complex 1mnm[59], and (f) the Hap1-18 activator 2hap[60]. The upper panel characterizes the shape of the molecular surface of the DNA binding sites with convex surfaces colour-coded in green and concave surfaces in grey/black. The surface representations are generated with the GRASP2 program[47]. The lower panel plots minor groove width (blue) and electrostatic potential in the centre of the minor groove close to the base edges (red) as functions of base sequence. The arginine contacts in the lower panels are defined based on the closest distance between the guanidinium groups and the bases. A-tract sequences are highlighted by a solid red line, the TATA box in (c) by a dashed line.

Supplementary Figure 3: Statistical analyses of nucleosome structures and *in vivo* sequences



(a) Histogram of the number of narrow minor groove regions per structure in 35 nucleosome structures in the PDB derived by X-ray crystallography that bind arginines within <6.0 Å from the bases.

(b) Histogram of the number of arginines per structure bound to narrow minor groove regions in the 35 available crystal structures of the nucleosome. Only arginines with any side

chain atom within a distance of <6.0 Å from a base atom of any nucleotide within a region with a minor groove width of <5.0 Å is considered as intruding the minor groove. In many cases, there are additional arginines in the region outside of the narrow minor groove with distances of >6.0 Å.

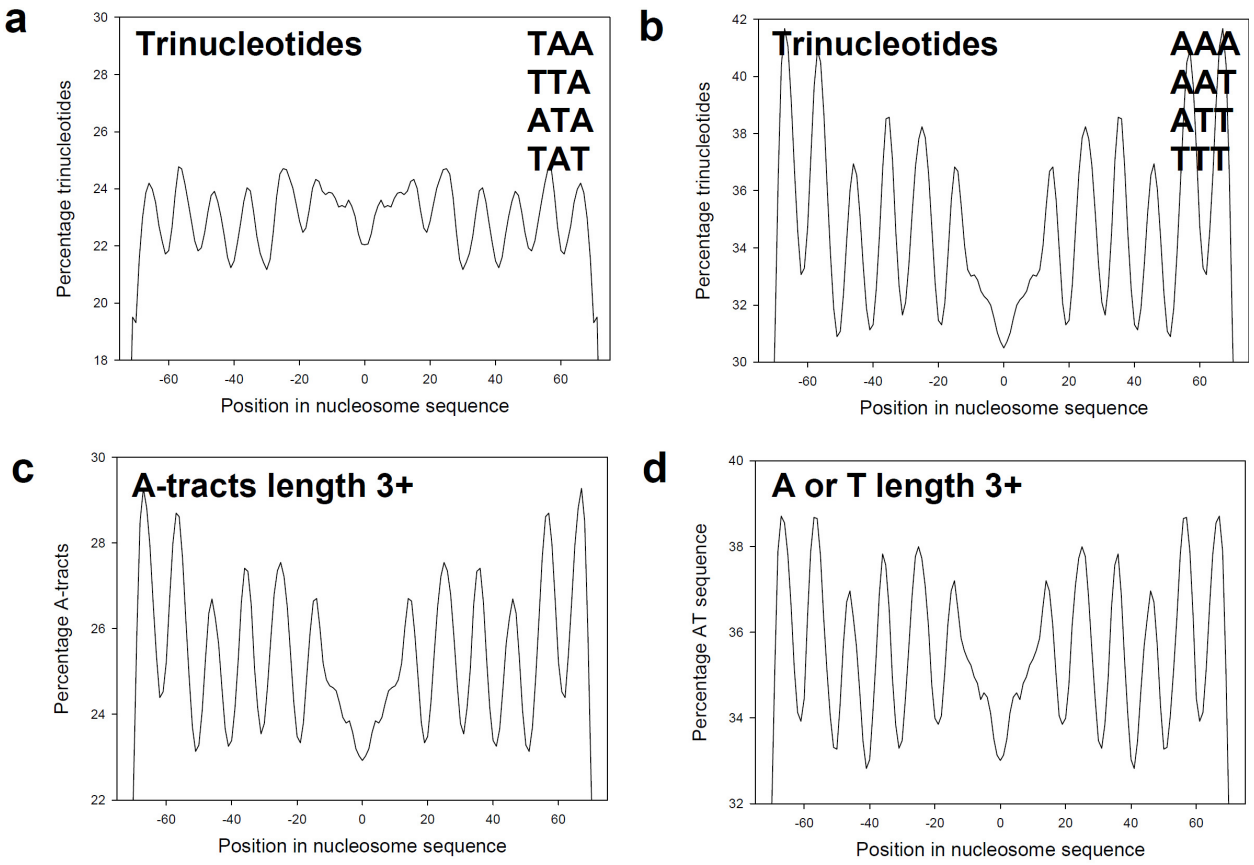
(c) The distribution of A-tracts four base pairs in length or longer in 23,076 *in vivo* yeast nucleosome sequences[29] illustrates that the occurrence of these sequence motifs has a periodicity of a helical turn, similar to the distribution of narrow minor groove regions in known nucleosome structures. The frequency is symmetrized by using both complementary strands.

(d) Histogram of the occurrence of A-tracts of length four or longer in the *in vivo* yeast dataset of 23,076 yeast nucleosome sequences[29].

(e) The distribution of A-tracts five base pairs in length or longer in 23,076 *in vivo* yeast nucleosome sequences[29] illustrates the occurrence of these sequence motifs with a periodicity of a helical turn. The frequency is symmetrized by using both complementary strands.

(f) Histogram of the occurrence of A-tracts of length five base pairs or longer in the *in vivo* yeast dataset of nucleosome sequences[29]. The A-tract occurrence decreasing with A-tract length is in accordance with the depletion of long A-tracts in nucleosomes reported previously[29,30,61].

Supplementary Figure 4: Comparison of periodicity signals of A-tracts vs. AT-rich elements that include TpA steps in *in vivo* sequences



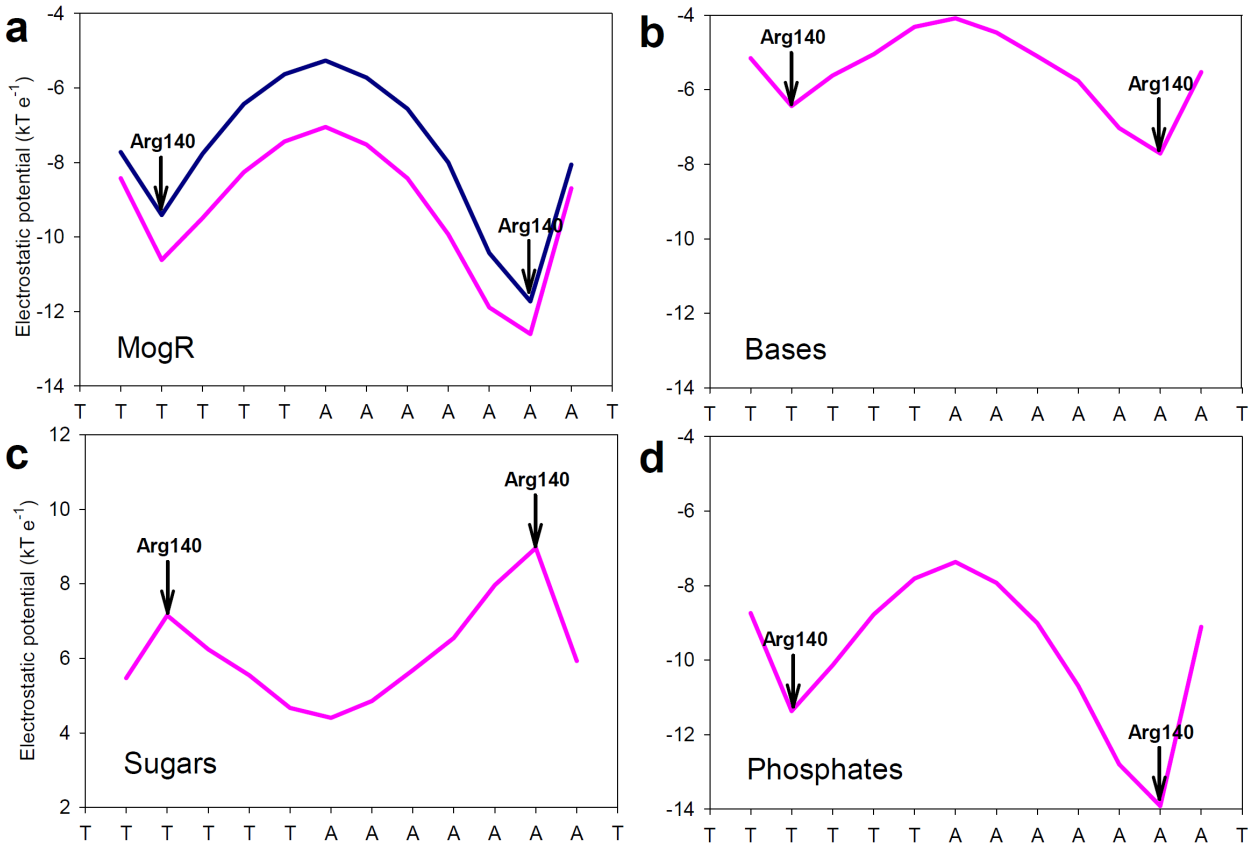
(a) The distribution of the non-A-tract trimers ATA, TAT, TAA, and TTA (containing TpA steps) in 23,076 *in vivo* yeast nucleosome sequences[29] illustrates that the occurrence of these sequence motifs has a periodicity of a helical turn. Note that this signal is significantly weaker than seen for short A-tracts (compare with panel b). Frequencies are symmetrized by using both complementary strands.

(b) The distribution of A-tract-containing trimers AAA, AAT, ATT, and TTT in 23,076 *in vivo* yeast nucleosome sequences[29] illustrates that the occurrence of short A-tracts has a significantly more pronounced periodicity signal than AT-rich trimers containing TpA steps (Supplementary Figure 4a).

(c) The distribution of A-tracts of length three base pairs or longer in 23,076 yeast nucleosome-bound DNA sequences[29] (reproduced from Figure 4c).

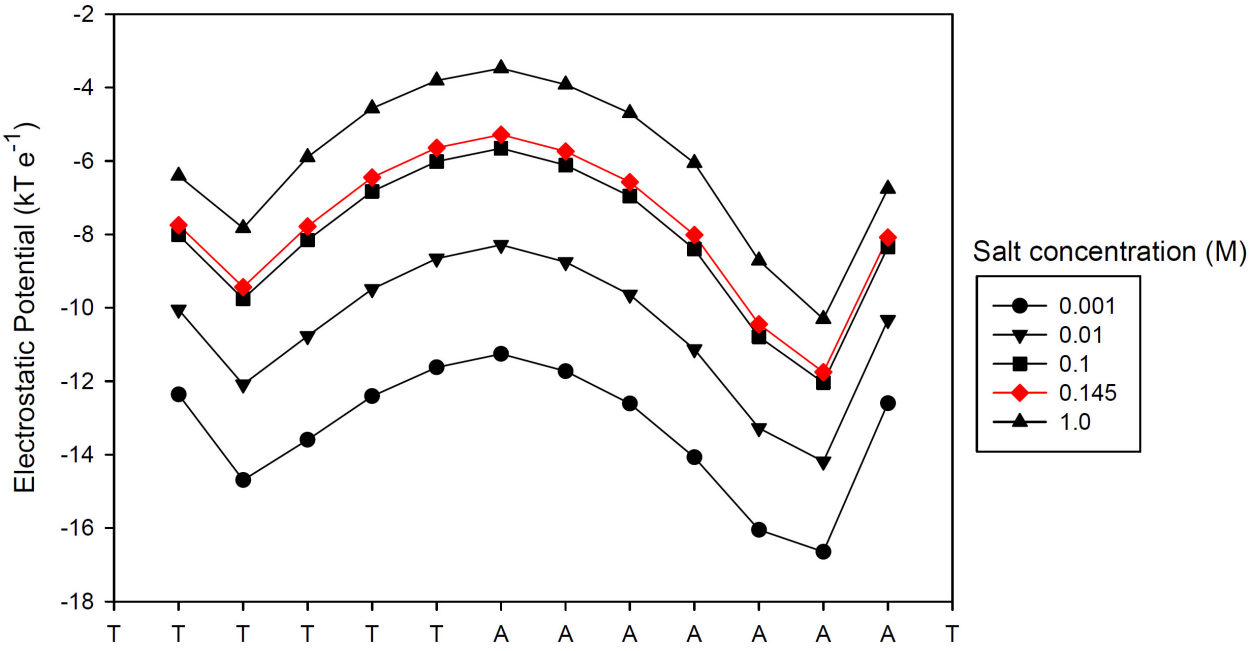
(d) In direct comparison to panel c, this graph shows the distribution of AT-rich regions of length three base pairs or longer in 23,076 yeast nucleosome-bound DNA sequences. Without excluding TpA steps the number of occurrences is greater than for A-tracts, but, in contrast to the comparison shown in panels a and b, the difference in amplitude between these periodicity signals changes very little (from 4.1% to 4.4%, averaged over all peaks and troughs).

Supplementary Figure 5: Decomposition of the electrostatic potentials into individual contributions



Potentials calculated with the linearized PB equation[31] are additive and make it possible to identify the contributions of different components of a nucleotide. (a) Electrostatic potential as a function of nucleotide sequence calculated with the DelPhi program[45] for the MogR binding site 3fdq[19] based on the non-linear and linear PB equation, shown in dark blue and magenta, respectively. Contributions to the potential calculated with the linear PB equation of (b) the bases, (c) the sugar moieties, and (d) the phosphates.

Supplementary Figure 6: Electrostatic potentials of MogR binding site for different salt concentrations



Electrostatic potential in the minor groove of the MogR binding site 3fdq[19], calculated at salt concentrations indicated in the figure. All results represent solutions to the non-linear Poisson-Boltzmann equation obtained from the DelPhi program[31,45]. Although the absolute numbers change, the pattern of the sequence dependence is not sensitive to ionic strength.

Supplementary References:

51 MacKerell, A.D. *et al.*, All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B* 102 (18), 3586-3616 (1998).

52 Jorgensen, W.L., Maxwell, D.S., & TiradoRives, J., Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *Journal of the American Chemical Society* 118 (45), 11225-11236 (1996).

53 Sitkoff, D., Sharp, K.A., & Honig, B., Accurate calculation of hydration free energies using macroscopic solvent models. *J Phys Chem* 98, 1978-1988 (1994).

54 Stofer, E. & Lavery, R., Measuring the geometry of DNA grooves. *Biopolymers* 34 (3), 337-346 (1994).

55 Hovde, S., Abate-Shen, C., & Geiger, J.H., Crystal structure of the Msx-1 homeodomain/DNA complex. *Biochemistry* 40 (40), 12013-12021 (2001).

56 Klemm, J.D., Rould, M.A., Aurora, R., Herr, W., & Pabo, C.O., Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules. *Cell* 77 (1), 21-32 (1994).

57 Jacobson, E.M., Li, P., Leon-del-Rio, A., Rosenfeld, M.G., & Aggarwal, A.K., Structure of Pit-1 POU domain bound to DNA as a dimer: unexpected arrangement and flexibility. *Genes Dev* 11 (2), 198-212 (1997).

58 Blanco, A.G., Sola, M., Gomis-Ruth, F.X., & Coll, M., Tandem DNA recognition by PhoB, a two-component signal transduction transcriptional activator. *Structure* 10 (5), 701-713 (2002).

59 Tan, S. & Richmond, T.J., Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. *Nature* 391 (6668), 660-666 (1998).

60 King, D.A., Zhang, L., Guarente, L., & Marmorstein, R., Structure of HAP1-18-DNA implicates direct allosteric effect of protein-DNA interactions on transcriptional activation. *Nat Struct Biol* 6 (1), 22-27 (1999).

61 Peckham, H.E. *et al.*, Nucleosome positioning signals in genomic DNA. *Genome Res* 17 (8), 1170-1177 (2007).