SUPPORTING INFORMATION

A Map of Minor Groove Shape and Electrostatic Potential from Hydroxyl Radical Cleavage Patterns of DNA

Eric P. Bishop, Remo Rohs, Stephen C. J. Parker, Sean M. West, Peng Liu, Richard S. Mann, Barry Honig, and Thomas D. Tullius

Supplementary Methods

X-ray structures. We obtained coordinates for the following eight X-ray crystal structures of the Drew-Dickerson dodecamer [d(CGCGAATTCGCG)]$_2$ from the Protein Data Bank. The criterion for selecting these 8 structures was to include all crystal structures of this sequence without any chemical modifications.

PDB ID	1BNA	2BNA	355D	428D	455D	1FQ2	1DOU	1JGR
reference	1	2	3	4	5	6	7	8

NMR structures. We obtained coordinates for the following set of five NMR structures of the Drew-Dickerson dodecamer [d(CGCGAATTCGCG)]$_2$ from the Protein Data Bank. The refinement of this NMR structure included dipolar coupling and chemical shift anisotropy measurements that involve the phosphate groups.[9] The data, therefore, provide information on torsion angle configurations in the phosphodiester backbone, which, combined with NOE measurements, leads to a high-accuracy NMR structure.

PDB ID	1NAJ
reference	9

Monte Carlo simulations. The minor groove profile of the Dickerson dodecamer was predicted based on all-atom Monte Carlo simulations. These simulations were started from ideal B-DNA without any input of sequence-dependent structure. The simulation protocol was identical to the one described elsewhere.[10,11] The minor groove width prediction is based on the average values calculated with the program CURVES[12] for every 10th snapshot along the Monte Carlo trajectory following equilibration.[13]

Tetranucleotides. For the plots in Supplementary Figure 3, tetranucleotides were taken from 884 protein-DNA X-ray structures, and 83 free-DNA X-ray structures, in the Protein Data Bank as of 08/10/2011.The criteria for including a crystal structure in this dataset were a length of the DNA duplex of at least one helical turn (ten base pairs), and the absence of any chemical modifications. In addition, since free DNA can transition to A-DNA due to crystal packing effects, we required free DNA to adopt a B-DNA conformation. Minor groove width for each

tetranucleotide was calculated with CURVES[12] between the two central base pairs by averaging all CURVES levels of the central base pair step.

All 136 unique tetranucleotides are represented in the protein-DNA set. In total, there are coordinates for 7675 tetranucleotide conformations in the dataset for protein-DNA structures. In the free-DNA set, 60 (of 136 possible) unique tetranucleotides are represented, and 293 tetranucleotide conformations in total.

Minor groove width variation in the nucleosome. We used CURVES[12] to calculate the width of the minor groove at each nucleotide of the DNA from the X-ray structures of nucleosome core particles from *Xenopus laevis* (PDB ID 1KX5)[14] and *Drosophila melanogaster* (PDB ID 2PYO).[15] The minor groove width values for each nucleotide are the average of all CURVES levels for a nucleotide.

Distribution of correlations for the ORChID2 nucleosome consensus pattern compared to individual ORChID2 nucleosome sequence patterns. We took the central 140 positions (to eliminate edge-effects from the prediction algorithm) from the ORChID2 nucleosome consensus profile (Figure 3, panels a and b) and scanned this profile across the symmetrized ORChID2 pattern calculated for each of the individual nucleosome-bound sequences from yeast and Drosophila (23,076 and 25,654 sequences, respectively). We retained the maximum Pearson correlation between the consensus profile and the symmetrized ORChID2 pattern of each individual sequence. We plotted this distribution, along with a similar distribution obtained from shuffled versions of the individual sequences as a control. This allowed us to measure the similarity of each individual nucleosome sequence ORChID2 pattern to the ORChID2 consensus (Supplementary Figure 7).

Supplementary References

(1) Drew, H., Wing, R., Takano, T., Broka, C., Tanaka, S., Itakura, K., and Dickerson, R. (1981) Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. U.S.A. 78*, 2179-2183.

(2) Drew, H.R., Samson, S., and Dickerson, R.E. (1982) Structure of a B-DNA dodecamer at 16 K. *Proc. Nat. Acad. Sci. U.S.A. 79*, 4040-4044.

(3) Shui, X., McFail-Isom, L., Hu, G.G., and Williams, L.D. (1998) The B-DNA dodecamer at high resolution reveals a spine of water on sodium. *Biochemistry 37*, 8341-8355.

(4) Shui, X., Sines, C.C., McFail-Isom, L., VanDerveer, D., and Williams, L.D. (1998) Structure of the potassium form of CGCGAATTCGCG: DNA deformation by electrostatic collapse around inorganic cations. *Biochemistry 37*, 16877-16887.

(5) Chiu, T.K., Kaczor-Grzeskowiak, M., and Dickerson, R.E. (1999) Absence of minor groove monovalent cations in the crosslinked dodecamer C-G-C-G-A-A-T-T-C-G-C-G. *J. Mol. Biol. 292*, 589-608.

(6) Sines, C.C., McFail-Islom, L., Howerton, S.B., VanDerveer, D., and Williams, L.D. (2000) Cations mediate B-DNA conformational heterogeneity. *J. Am. Chem. Soc. 122*, 11048-11056.

(7) Woods, K.K., McFail-Islom, L., Sines, C.C., Howerton, S.B., Stephens, R.K., and Williams, L.D. (2000) Monovalent cations sequester within the A-tract minor groove of $[d(CGCGAATTCGCG)]_2$. *J. Am. Chem. Soc. 122*, 1546-1547.

(8) Howerton, S.B., Sines, C.C., VanDerveer, D., and Williams, L.D. (2001) Locating monovalent cations in the grooves of B-DNA. *Biochemistry 40*, 10023-10031.

(9) Wu, Z., Delaglio, F., Tjandra, N., Zhurkin, V.B., and Bax, A. (2003) Overall structure and sugar dynamics of a DNA dodecamer from homo- and heteronuclear dipolar couplings and ^{31}P chemical shift anisotropy. *J. Biomol. NMR 26*, 297-315.

(10) Rohs, R., Sklenar, H., and Shakked, Z. (2005) Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure 13*, 1499-1509.

(11) Rohs, R., Bloch, I., Sklenar, H., and Shakked, Z. (2005) Molecular flexibility in ab initio drug docking to DNA: binding-site and binding-mode transitions in all-atom Monte Carlo simulations. *Nucleic Acids Res. 33*, 7048-7057.

(12) Lavery, R., and Sklenar, H. (1989) Defining the structure of irregular nucleic acids: conventions and principles. *J. Biomol. Struct. Dyn. 6*, 655-667.

(13) Joshi, R., Passner, J.M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M.A., Jacob, V., Aggarwal, A.K., Honig, B., and Mann, R.S. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell 131*, 530-543.

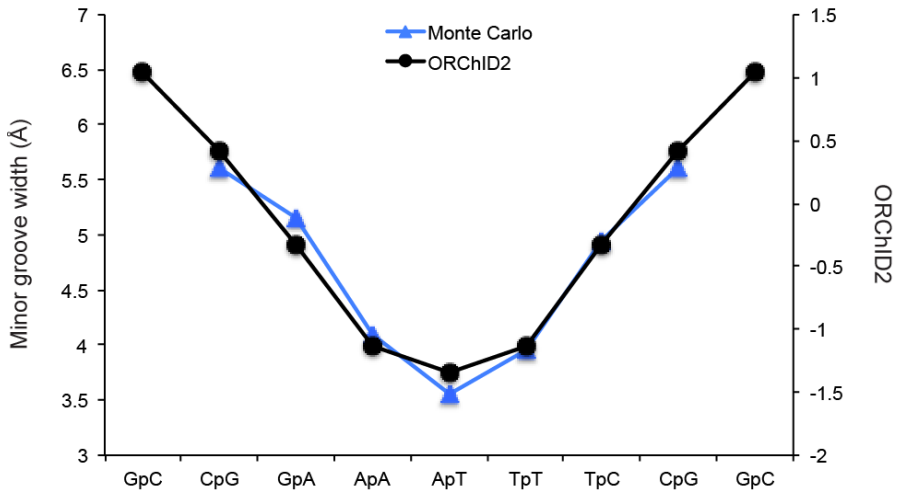
(14) Davey, C.A., Sargent, D.F., Luger, K., Maeder, A.W., and Richmond, T.J. (2002) Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol. 319*, 1097-1113.

(15) Clapier, C.R., Chakravarthy, S., Petosa, C., Fernández-Tornero, C., Luger, K., and Müller, C.W. (2008) Structure of the Drosophila nucleosome core particle highlights evolutionary constraints on the H2A-H2B histone dimer. *Proteins 71*, 1-7.

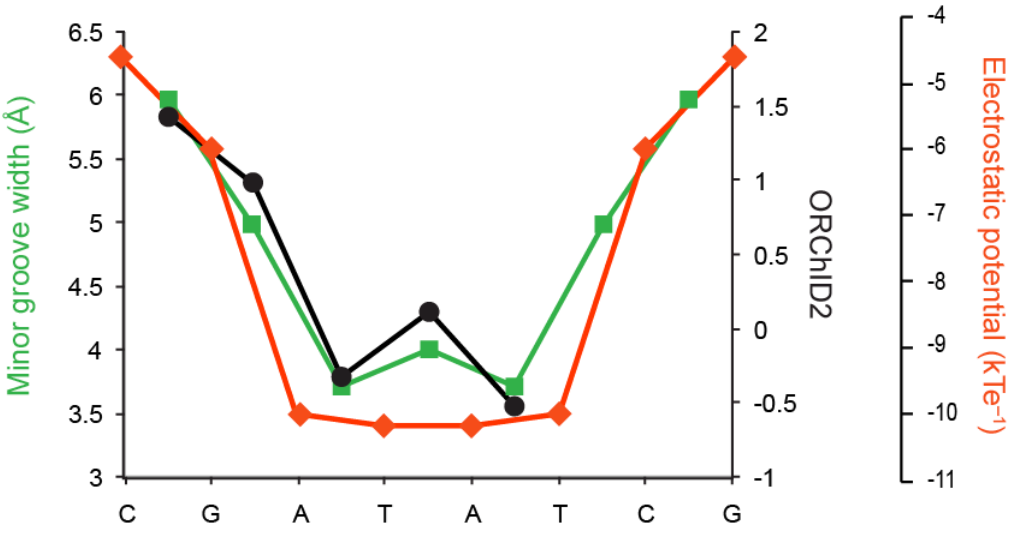
(16) Shatzky-Schwartz, M., Arbuckle, N.D., Eisenstein, M., Rabinovich, D., Bareket-Samish, A., Haran, T.E., Luisi, B.F., and Shakked, Z. (1997) X-ray and solution studies of DNA oligomers and implications for the structural basis of A-tract-dependent curvature. *J. Mol. Biol. 267*, 595-623.

(17) Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009) The role of DNA shape in protein-DNA recognition. *Nature 461*, 1248-1253.

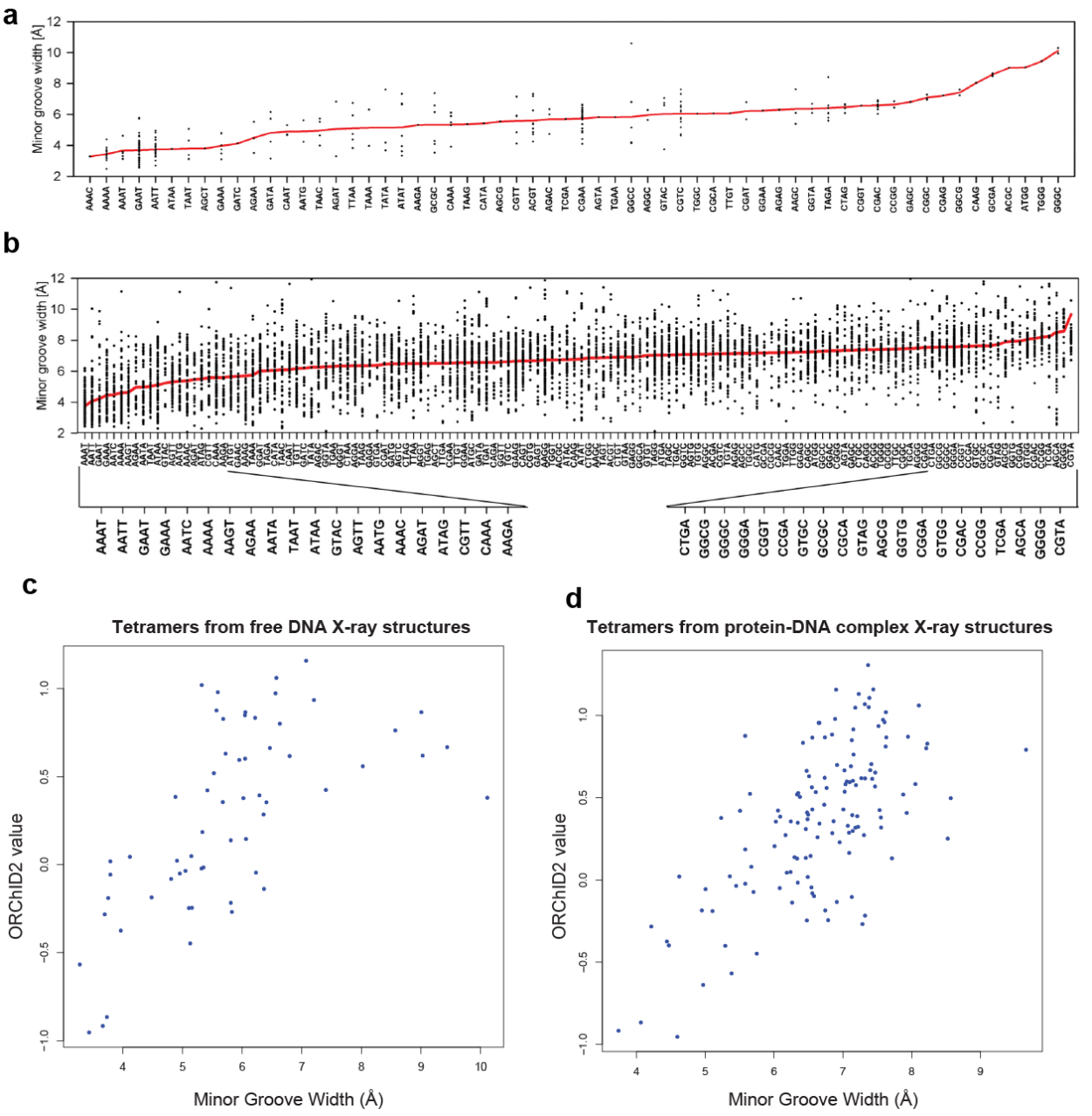
Supplementary Figures



Supplementary Figure 1. Quantitative correlation of the experimental ORChID2 cleavage pattern (black), with minor groove width calculated from a Monte Carlo simulation (blue) of the Drew-Dickerson dodecamer. The Pearson correlation for comparison of the ORChID2 pattern with minor groove width (7 nucleotide positions) is 0.981 (p-value = 9.83×10^{-5}).

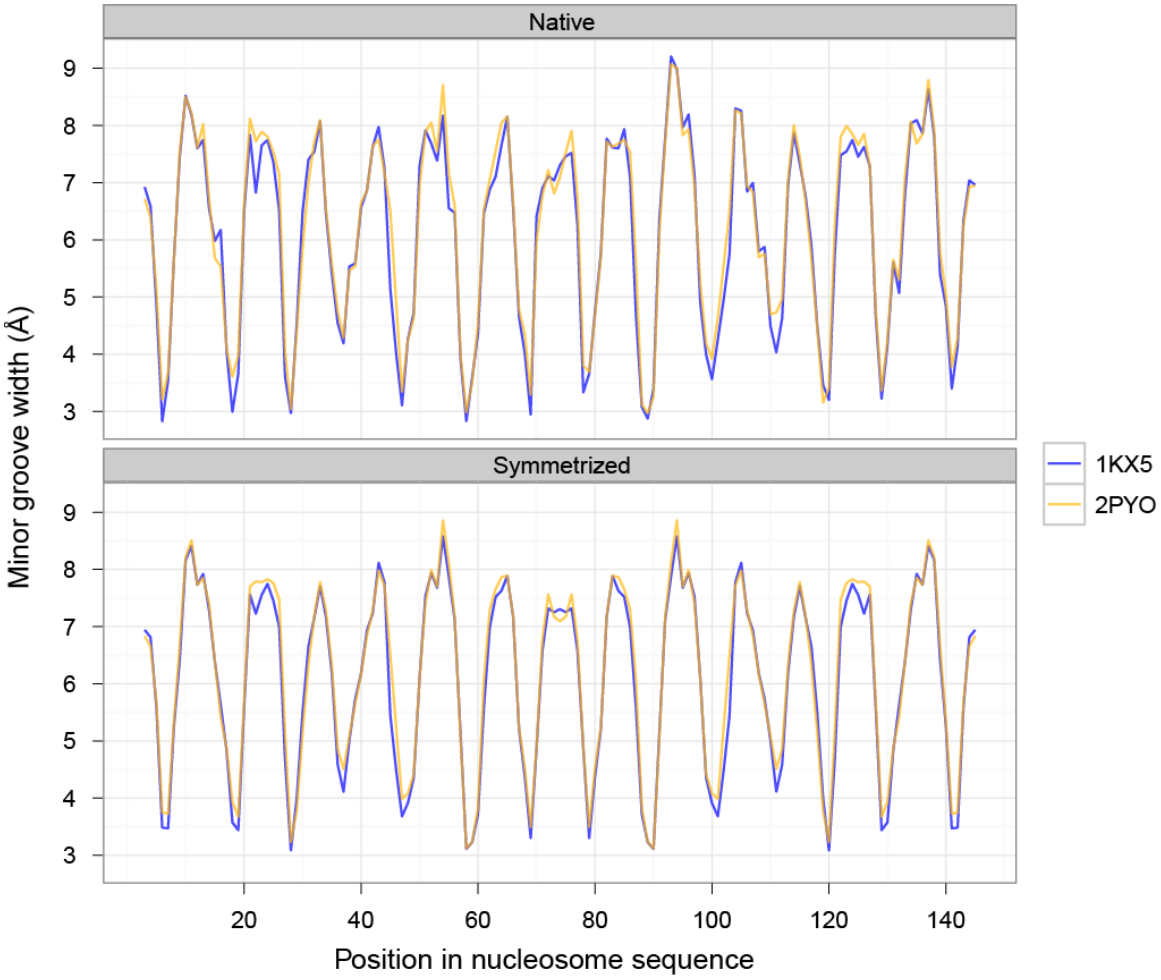


Supplementary Figure 2. Quantitative correlation of the experimental ORChID2 cleavage pattern (black), with electrostatic potential (red) and minor groove width (green) determined from the X-ray structure of $[d(CGCGATATCGCG)]_2$[16] (PDB ID 287D). Minor groove width and electrostatic potential are symmetrized to reflect the symmetry of the nucleotide sequence. The ORChID2 pattern was derived from the experimental cleavage patterns of the 9-mer sequence d(GTATCGCG) and its complement, which are present in the current ORChID database. The Pearson correlation for comparison of the ORChID2 pattern with minor groove width (5 nucleotide positions) is 0.973 (p-value = 5.33×10^{-3}); for comparison of ORChID2 with electrostatic potential (5 nucleotide positions), 0.960 (p-value = 9.44×10^{-3}).

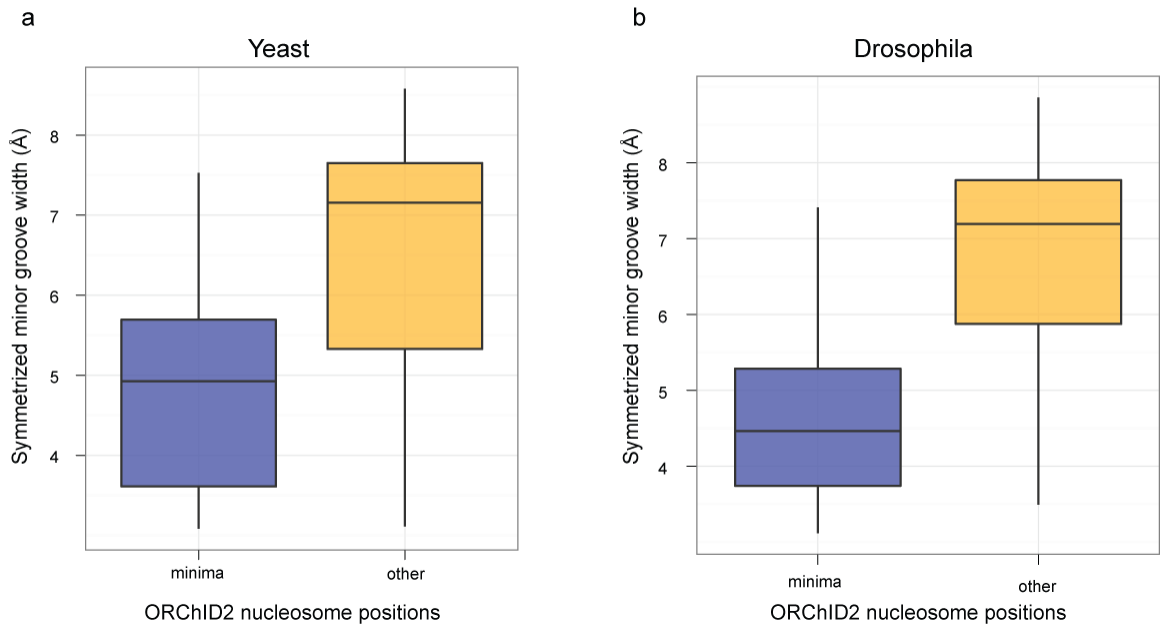


Supplementary Figure 3. Minor groove width between the two central base pairs of each tetramer is plotted for (a) the 60 unique tetranucleotides occurring in free-DNA crystal structures and (b) all 136 unique tetranucleotides from protein-DNA crystal structures. Each individual conformation is shown as black dot, in order to convey structural variability for a given tetramer, and the average width is plotted in red. The sequence, average width, and occurrence in our dataset are given in Supplementary Table 1.

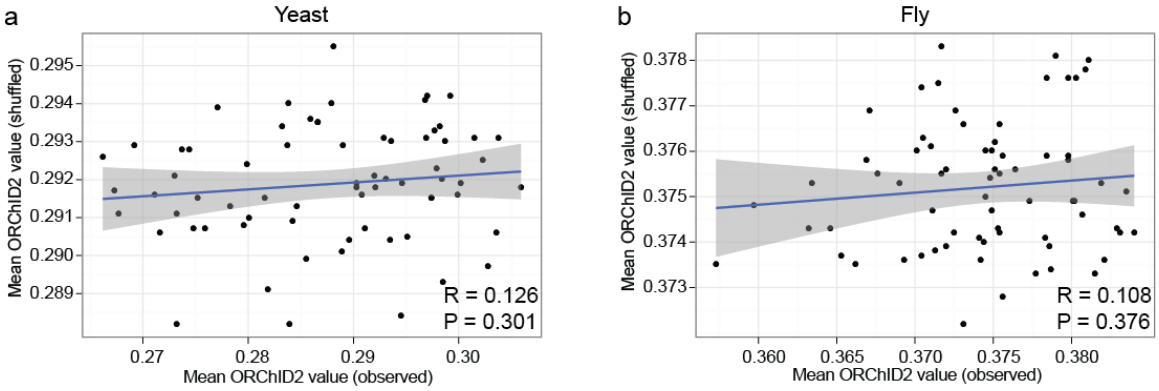
The average minor groove width at the central base pair step is compared with the ORChID2 value, for free DNA crystal structures (c) and for protein-DNA crystal structures (d). The Pearson correlation between minor groove width and ORChID2 values is 0.638 (p-value 4.06 x 10^{-8}) for free DNA structures and 0.653 (p-value < 1.0 x 10^{-16}) for DNA in protein-DNA complexes.



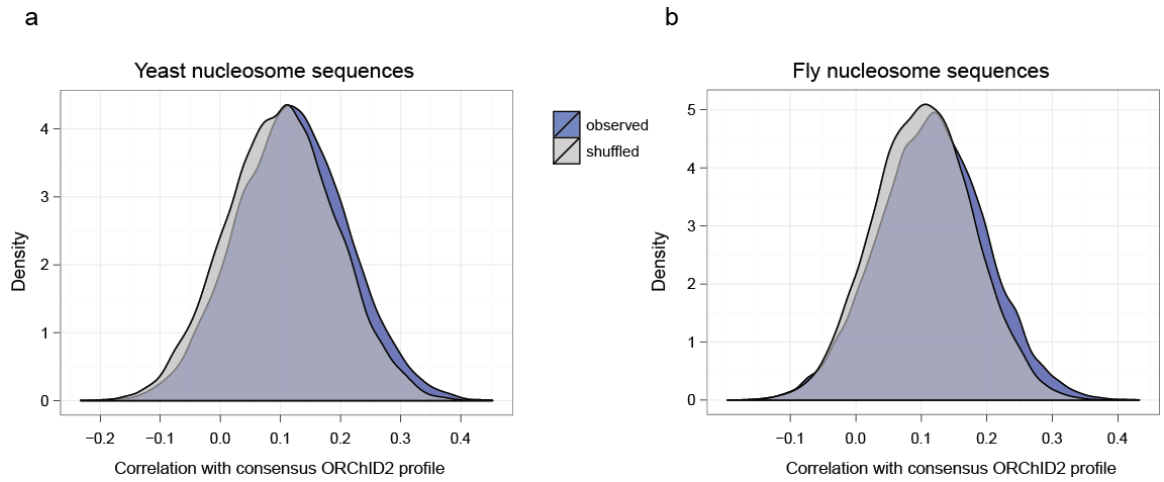
Supplementary Figure 4. Minor groove width variation in the nucleosome. Shown are plots of the minor groove width at each nucleotide position in X-ray crystal structures of two nucleosome core particles (PDB ID 1KX5, blue; PDB ID 2PYO, yellow) (top panel). In the bottom panel, minor groove widths were symmetrized around the nucleosome dyad axis.



Supplementary Figure 5. Minima in the composite ORChID2 pattern of nucleosome-binding sequences occur where the minor groove is narrow in the nucleosome three-dimensional structure. (a) Box plots of the distribution of minor groove widths at the 12 minima, plus one position on either side (a total of 36 positions), in the ORChID2 pattern of the composite set of 23,076 nucleosome-binding sequences from yeast (see Figure 3, panel a) (left, blue box), and at all other positions (right, yellow box). There is a significant difference (p-value = 2.567×10^{-4}; Wilcoxon rank sum test) between these two distributions. (b) Box plots of the distribution of minor groove widths at the 12 minima, plus one position on either side (a total of 36 positions), in the ORChID2 pattern of the composite set of 25,654 nucleosome-binding sequences from Drosophila (see Figure 3, panel b) (left, blue box), and at all other positions (right, yellow box). There is a significant difference (p-value = 4.853×10^{-6}; Wilcoxon rank sum test) between these two distributions.



Supplementary Figure 6. Correlation between mean ORChID2 values for observed and shuffled nucleosome bound sequences in yeast (a) and fly (b). Values plotted on the x and y-axes are derived from the blue and gray lines, respectively, in Figures 3a and 3b. We observe no significant correlation between the observed and shuffled patterns, indicating that the observed consensus nucleosomal ORChID2 pattern is not an artifact of the analysis. Each correlation plot is based on one-half of the nucleosome dyad and is center-aligned by the dyad axis. Gray shading indicates the standard error around the best-fit line.



Supplementary Figure 7. Distribution of correlations for the ORChID2 nucleosome consensus pattern relative to the symmetrized ORChID2 pattern for individual real and shuffled nucleosome-bound sequences. Observed (blue distribution) and shuffled (gray distribution) versions of nucleosome-bound sequences for yeast (a) and fly (b) were compared to their corresponding ORChID2 consensus profiles (blue lines in Figures 3a and 3b). For each species, the distribution of correlations for the observed sequences is significantly greater than the shuffled sequences (p-value < 2.2×10^{-16}; Wilcoxon rank sum test).

Supplementary Table 1: Minor groove width at the center of a tetranucleotide in free DNA and protein-DNA structures. These data are an update (as of 08/10/2011) of a previously published analysis[17] of the PDB.

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

Tetramer	Average minor groove width [Å]	Number of occurrences
AAAC	3.28	1
AAAA	3.43	16
AAAT	3.67	8
GAAT	3.70	45
AATT	3.73	33
ATAA	3.76	1
TAAT	3.79	5
AGCT	3.79	1
GAAA	3.97	6
GATC	4.12	1
AGAA	4.48	3
GATA	4.81	4
CAAT	4.88	3
AATG	4.91	2
TAAC	4.96	4
AGAT	5.05	2
TTAA	5.11	5
TAAA	5.13	2
TATA	5.15	3
ATAT	5.16	7
AAGA	5.32	1
GCGC	5.32	6
CAAA	5.33	5
TAAG	5.35	1
CATA	5.42	1
AGCG	5.53	1
CGTT	5.57	5
ACGT	5.59	9
AGAC	5.68	3
TCGA	5.68	1
CGAA	5.72	23
AGTA	5.81	1
TGAA	5.81	1
GGCC	5.83	6
AGGC	5.95	2
GTAC	6.02	4
CGTC	6.05	13
TGGC	6.05	1
CGCA	6.06	1
TTGT	6.06	1
CGAT	6.22	2
GGAA	6.23	1
AGAG	6.29	1
AAGC	6.36	3
GGTA	6.37	4
TAGA	6.41	7
CTAG	6.46	5
CGGT	6.56	1
CGAC	6.57	9
CCGG	6.63	2
GAGC	6.80	1
CGGC	7.07	3
CGAG	7.20	1
GGCG	7.41	2
CAAG	8.02	1
GCGA	8.57	6
ACGC	9.01	1
ATGG	9.03	1
TGGG	9.44	2
GGGC	10.11	2

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

Tetramer	Average minor groove width [Å]	Number of occurrences
AAAT	3.74	141
AATT	4.06	99
GAAT	4.22	70
GAAA	4.44	97
AATC	4.47	32
AAAA	4.60	153
AAGT	4.62	78
AGAA	4.95	28
AATA	4.97	54
TAAT	5.00	90
ATAA	5.11	98
GTAC	5.23	16
AGTT	5.29	62
AATG	5.36	60
AAAC	5.38	62
AGAT	5.45	30
ATAG	5.51	53
CGTT	5.58	40
CAAA	5.58	111
AAGA	5.58	42
ATGT	5.65	73
GAAC	5.67	42
AAAG	5.70	73
TAAA	5.75	73
GGAT	6.01	95
TAGA	6.03	37
CATA	6.06	50
TAAC	6.08	49
CAAT	6.09	46
TGTT	6.17	86
GATC	6.19	38
TATA	6.24	53
AGAC	6.24	66
GGTA	6.26	44
TGAA	6.30	58
GGGT	6.34	36
CTAA	6.34	54
AGGA	6.34	104
TAAG	6.35	57
GAGA	6.35	28
GTGA	6.38	64
CGAT	6.42	55
GATG	6.46	78
AGTC	6.46	78
CTAG	6.47	36
TTAA	6.48	46
AGGT	6.48	60
TGAG	6.49	60
AGCT	6.49	12
TTGA	6.50	47
CGAA	6.51	73
TTGT	6.53	83
GGAA	6.54	153
ATGC	6.55	60
GATA	6.56	83
TGAT	6.56	61
CAGA	6.56	42
GGTT	6.58	32
TGTC	6.61	97
GAAG	6.63	128
CAGT	6.65	83
CGTG	6.66	51
GAGT	6.66	37
AAGG	6.73	131
TGGT	6.73	50
ACGC	6.74	18
ATAC	6.74	60
CAAG	6.76	35
ATAT	6.79	38
CTGG	6.84	35
AAGC	6.85	36
TAGT	6.86	40
ACGT	6.89	29
CTGT	6.90	101
GTAA	6.91	51
GAGG	6.91	46
GGCA	6.95	43
GTGT	7.01	32

TAGG	7.02	73	CCGA	7.60	30
ATGA	7.03	81	GTGC	7.62	55
TAGC	7.04	57	GCGC	7.63	31
TGAC	7.06	102	CGCA	7.63	26
GGTC	7.07	61	GTAG	7.71	32
AGTG	7.09	65	AGCG	7.88	34
TGTG	7.09	61	GGTG	7.93	40
AGGC	7.10	54	CGGA	7.95	21
ACGA	7.12	70	GTGG	8.05	73
CGTC	7.13	44	CGAC	8.10	43
TGTA	7.13	69	CCGG	8.21	39
AGAG	7.14	21	TCGA	8.23	26
AGCC	7.14	65	AGCA	8.52	79
TGGC	7.14	41	GGGG	8.57	26
CATG	7.15	58	CGTA	9.66	51
GCGA	7.15	15			
CGCG	7.18	14			
CAAC	7.18	30			
TGGA	7.19	89			
TTGG	7.21	29			
GGAG	7.22	36			
CAGC	7.23	64			
ATGG	7.27	97			
GGCC	7.28	30			
GGAC	7.31	53			
CGGG	7.32	35			
AGTA	7.32	26			
GAGC	7.32	39			
CTGC	7.37	55			
CAGG	7.38	36			
ACGG	7.38	37			
TGGG	7.40	50			
GCGG	7.41	33			
TTGC	7.43	57			
CGGC	7.44	52			
TGCA	7.47	24			
AGGG	7.47	57			
CGAG	7.52	20			
CTGA	7.53	53			
GGCG	7.53	38			
GGGC	7.55	57			
GGGA	7.55	115			
CGGT	7.58	64			