ARTICLES

The role of DNA shape in protein–DNA recognition

Remo Rohs[1]*, Sean M. West[1]*, Alona Sosinsky[1]†, Peng Liu[1], Richard S. Mann[2] & Barry Honig[1]

The recognition of specific DNA sequences by proteins is thought to depend on two types of mechanism: one that involves the formation of hydrogen bonds with specific bases, primarily in the major groove, and one involving sequence-dependent deformations of the DNA helix. By comprehensively analysing the three-dimensional structures of protein–DNA complexes, here we show that the binding of arginine residues to narrow minor grooves is a widely used mode for protein–DNA recognition. This readout mechanism exploits the phenomenon that narrow minor grooves strongly enhance the negative electrostatic potential of the DNA. The nucleosome core particle offers a prominent example of this effect. Minor-groove narrowing is often associated with the presence of A-tracts, AT-rich sequences that exclude the flexible TpA step. These findings indicate that the ability to detect local variations in DNA shape and electrostatic potential is a general mechanism that enables proteins to use information in the minor groove, which otherwise offers few opportunities for the formation of base-specific hydrogen bonds, to achieve DNA-binding specificity.

The ability of proteins to recognize specific DNA sequences is a hallmark of biological regulatory processes. The determination of the three-dimensional structures of numerous protein–DNA complexes has provided a detailed picture of binding, revealing a structurally diverse set of protein families that exploit a wide repertoire of interactions to recognize the double-helix[1]. Nucleotide sequence-specific interactions often involve the formation of hydrogen bonds between amino-acid side chains and hydrogen-bond donors and acceptors of individual base pairs. It has long been recognized that every base pair has a unique hydrogen-bonding signature in the major groove, but that this is not the case in the minor groove[2]. Thus, the expectation has been that the recognition of specific DNA sequences would take place primarily in the major groove by the formation of a series of amino-acid- and base-specific hydrogen bonds[1]. This 'direct readout' mechanism is consistent with observations derived from three-dimensional structures of protein–DNA complexes, but it is far from the entire story.

In many complexes, the DNA assumes conformations that deviate from the structure of an ideal B-form double helix[3–5], sometimes bending in such a way to optimize the protein–DNA interface[6], and in some cases undergoing large conformational changes as in the opening of the minor groove in the complex formed between TBP and the TATA box[7,8]. The term 'indirect readout' was coined[9] to describe such recognition mechanisms that depend on the propensity of a given sequence to assume a conformation that facilitates its binding to a particular protein. The bases involved in such mechanisms need not be in contact with the protein and, for example, can be found in linker sequences that connect two half-sites that are themselves bound by individual protein subunits[10,11].

We recently described an example of a new readout mechanism, the recognition of local sequence-dependent minor-groove shape[12] that is distinct from previously described indirect readout mechanisms. In this case, the sequence-dependence of minor-groove width and corresponding variations in electrostatic potential are used by the *Drosophila* Hox protein Sex combs reduced (SCR) to distinguish small differences in nucleotide sequence[12]. Here we report that this mechanism is a widely used mode of protein–DNA recognition that involves the creation of specific binding sites for positively charged amino acids, primarily arginine, within the minor groove. Minor-groove narrowing is found to be correlated with A-tracts[13,14], usually defined as stretches of four or more As or Ts that do not contain the flexible TpA step[15], but extended here to include as few as three base pairs (see later). Our results offer fundamentally new insights into the structural and energetic origins of protein–DNA binding specificity, and thus have important implications for the prediction of transcription-factor-binding sites in genomes.

Arginine is enriched in narrow minor grooves

The percentage of minor-groove contacts associated with each amino acid, classified according to the width of the minor groove, was determined (Fig. 1a). Arginine constitutes 28% of all amino-acid residues that contact the minor groove and is notably enriched in narrow minor grooves, defined here by a groove width of <5.0 Å (compared to 5.8 Å in ideal B-DNA). Remarkably, 60% of the residues in narrow minor grooves are arginines, compared to 22% in minor grooves that are defined as not narrow—that is, width ≥5.0 Å. A smaller enrichment is also observed for lysine but the overall population of lysines within the minor groove is much less than for arginine.

Binding to the minor groove is a characteristic of many, but not all, protein superfamilies and a large subset of these contact a narrow minor groove (Table 1). Moreover, if the minor groove is contacted, arginines are likely to be involved, and the likelihood that an arginine will be present becomes even greater for narrow minor grooves (Supplementary Table 1).

We compiled the DNA sequence preferences for protein–DNA complexes in which an arginine contacts a narrow minor groove (Fig. 1b). The figure shows that the base pair that has the shortest contact distance with the arginine guanidinium group has a 78%

[1] Howard Hughes Medical Institute, Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biophysics, Columbia University, 1130 Saint Nicholas Avenue, New York, New York 10032, USA. [2] Department of Biochemistry and Molecular Biophysics, Columbia University, 701 West 168th Street, HHSC 1104, New York, New York 10032, USA. †Present address: Institute of Structural and Molecular Biology, School of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, UK.
*These authors contributed equally to this work.

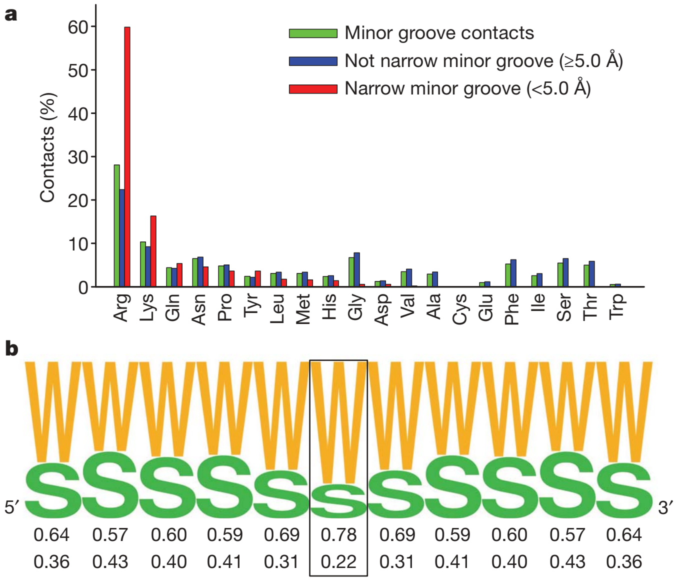

Figure 1 | Amino acid frequencies in minor grooves. a, Histograms for each amino acid illustrate the frequency with which they are observed in any minor groove (green), in minor grooves with a width of ≥5.0 Å (blue), and in narrow minor grooves of width <5.0 Å (red). **b**, Frequency of AT (W) and GC (S) base pairs in sequences of 229 sites contacted by arginines in narrow minor grooves. The central base pair (boxed) is contacted by arginine. Frequencies are symmetrized by using both complementary strands.

probability of being an AT and 22% of being a GC. Neighbouring base pairs in both the 5′ and 3′ directions surrounding the closest contacting base pair also have a strong tendency to be AT. Taken together, these data demonstrate that arginines tend to bind narrow minor grooves in AT-rich DNA.

AT-rich sequences tend to narrow minor grooves

We calculated minor-groove widths for all tetranucleotides contained in Protein Data Bank (PDB) structures for both free DNA (Fig. 2a) and DNA in complexes with proteins (Fig. 2b). There is a large spread of values due in part to end effects and to the effects of crystal packing, but some trends are evident nevertheless. For example, for free DNA structures most of the tetranucleotides with narrow minor grooves (width <5.0 Å) are AT-rich (Fig. 2a and Supplementary Table 2a). Similar behaviour is observed in protein–DNA complexes (Fig. 2b and Supplementary Table 2b). In contrast, tetranucleotides with wide minor grooves have a strong tendency to be GC-rich.

Table 1 | Protein superfamilies with minor-groove contacts

Narrow minor grove	Not narrow minor grove
SRF-like	DNA repair protein MutS, domain I
IHF-like DNA-binding proteins	Origin of replication-binding domain, RBD-like
Histone-fold	DNA/RNA polymerases
DNA breaking-rejoining enzymes	Eukaryotic DNA topoisomerase I, amino-terminal DNA-binding fragment
Zn2/Cys6 DNA-binding domain	Ribonuclease H-like
Homeodomain-like	TATA-box binding protein-like
p53-like transcription factors	
Lambda repressor-like DNA-binding domains	
Winged helix DNA-binding domain	
Leucine zipper domain	
C-terminal effector domain of the bipartite response regulators	
Restriction endonuclease-like	
Glucocorticoid receptor-like (DNA-binding domain)	

Listed are SCOP superfamilies[46] that have an arginine minor-groove contact within a distance of <6.0 Å from the base. Superfamilies that use arginine to contact a narrow minor groove (<5.0 Å) and those that use arginine to contact a not narrow minor groove (≥5.0 Å) are shown. Only superfamilies with a minimum of ten protein chains in PDB structures bound to DNA at least one helical turn long are included. The percentages of chains with minor-groove contacts vary considerably among SCOP superfamilies and are provided in Supplementary Table 1.

The correlation between AT content and groove width is not unexpected given the fact that A-tracts are known to produce narrow minor grooves. However, TpA steps have a tendency to widen the minor groove[15], so it was of interest to determine whether the distinct properties of A-tracts and TpA steps are reflected in our tetranucleotide data set. We find that 67% of tetranucleotides composed only of AT base pairs have a narrow minor groove, but that this number increases to 82% if we exclude TpA steps so as to consider only A-tracts. Even A-tracts of length three have a strong tendency to narrow the minor groove. Forty-three per cent of the tetranucleotides with a minor groove width of <5.0 Å have an A-tract length of three, a percentage that decreases to 11% of tetranucleotides with canonical minor-groove widths (between 5.0 and 7.0 Å) and to 4% of tetranucleotides with minor grooves wider than 7.0 Å (Supplementary Fig. 1). Furthermore, compared to other AT-rich sequences, A-tracts are specifically enriched in DNAs with narrow minor grooves (Supplementary Fig. 1). Thus, although A-tracts are usually thought of as requiring four or more base pairs, in part because a minimum of four is required to rigidify the DNA[14], this analysis shows that A-tracts as short as length three are positively correlated with narrow minor grooves.

Arginines recognize enhanced electrostatic potentials

The minor-groove width and electrostatic potential versus binding-site sequence for several complexes whose binding interface includes an arginine inserted into the minor groove is plotted in Fig. 3 and Supplementary Fig. 2. The correlation of width and potential as well as the tendency of arginines to be located close to minima in width and potential is evident. In this section we highlight a few specific examples of how arginine–minor-groove interactions are used in DNA recognition.

Figure 3a represents the ternary complex of the *Drosophila* Hox protein Ultrabithorax (UBX) and its cofactor Extradenticle (EXD) bound to DNA[16]. In this complex, Arg 5 of UBX, which is a conserved residue across all homeodomains, inserts into a narrow region formed by a 4-base-pair (bp) A-tract. An example of a long and very narrow A-tract that binds α2-Arg 7 from the MATa1–MATα2 complex with DNA is shown (Fig. 3b)[17]. In contrast, α2-Arg 4 inserts into a shallower region at one end of the A-tract, where there are local minima in width and potential that are smaller than at the Arg 7 site in the centre of the A-tract. The two POU domains of the mammalian OCT1 (also known as POU2F1)–PORE complex bind to two A-tracts (Fig. 3c) in which the minima are positioned in such a way as to provide binding sites for four arginines, two from each POU domain[18].

The location of these A-tracts with respect to other nucleotide sequence features can be used to generate specificity, as previously discussed for the Hox protein SCR[12]. In the case of SCR binding, the position of a TpA step within an AT-rich region has a critical role in binding specificity. A similar strategy is used by the motility gene repressor (MogR) in which two long A-tracts separated by a TpA step produce two arginine-binding sites[19] (Fig. 3d). The unique shape recognized by these two arginines probably contributes to the position of the MogR-binding site along the DNA sequence. The overall tendency of TpA steps to widen the minor groove is most apparent when they are positioned between two A-tracts (as in SCR[12] and MogR[19]) where the TpA step acts as a ‘hinge’ between more rigid elements[15,20]. In other contexts, owing to their flexibility, TpA steps can also be accommodated in narrow minor grooves[21]. An example is provided by the bipartite DNA-binding domain of Tc3 transposase in which the arginines bind to a narrow region containing a TATA box[22] that displays enhanced negative electrostatic potential (Fig. 3e).

Although less frequent, arginines also bind narrow grooves associated with non-A-tract sequences. Figure 3f summarizes features of the binding of the 434 repressor to its operator[23] that contains 7 bp

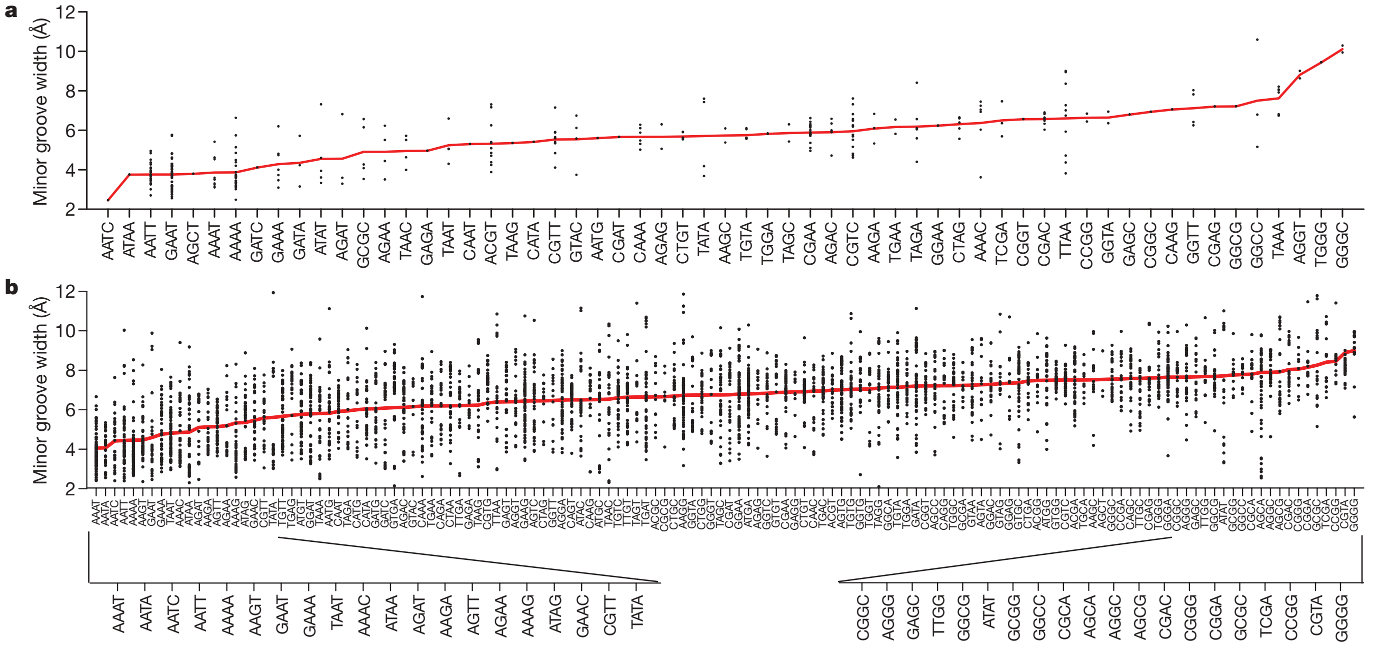

Figure 2 | Distribution of tetranucleotide sequences according to average minor-groove width. Tetranucleotides from structures with a minimum length of one helical turn for which minor-groove width can be defined are ordered by average minor-groove width (red). The widths of all tetranucleotides are shown (black) and the sequence, average width, and occurrence in our data set are given in Supplementary Table 2. **a**, The 59 unique tetranucleotides from free DNA structures. **b**, The set of all 136 unique tetranucleotides derived from protein–DNA complexes.

that are all AT with the exception of a central CG. (The guanine amino group tends to widen narrow grooves, but a single GC base pair can be accommodated with only little disruption).

Arginine minor-groove interactions in the nucleosome

Figure 4a plots minor-groove width and electrostatic potential along the DNA sequence of the nucleosome core particle containing recombinant histones and a 147-bp DNA fragment (PDB code 1kx5)[24]. There are 14 minima in minor-groove width corresponding to regions where the DNA bends so as to wrap around the histone core. As earlier, there is a marked correlation between width and potential. The variation in width between the narrowest and widest regions is about 5 Å, and the difference between the maxima and minima in electrostatic potential is about 6 kT e^{-1} (Fig. 4a). As a consequence, there should be a strong driving force for basic amino acids to bind to narrow regions and indeed arginines are found in 9 of the 14 minima. These arginines are shown in Fig. 4b where the nucleosomal DNA has been colour-coded by minor-groove width. (Although all 14 narrow minor-groove regions are contacted by arginines[24] only 9 satisfy our criteria of $<$6.0 Å between arginine atoms and base atoms in the groove). A similar repeating pattern of narrow minor grooves that are contacted by arginines is seen in all 35 available nucleosome crystal structures (Supplementary Fig. 3a, b).

Because short A-tracts narrow the minor groove and facilitate the bending of DNA, we would expect to see a periodicity of A-tracts in DNA sequences bound by nucleosomes *in vivo*. Previous analyses have focused on dinucleotide statistics[25,26], although it has been known for some time that there is a periodic pattern of AAA and AAT trinucleotides in nucleosome core DNA[27,28]. An analysis of DNA sequences bound *in vivo* by yeast nucleosomes[29] reveals a clear periodicity for A-tracts of at least length three (denoted 3+, Fig. 4c). Moreover, nucleosomal DNAs contain, on average, 10.0 A-tracts of length 3+ (Fig. 4d). Periodicity is also detected for A-tracts of length 4+ and even 5+, although the number per nucleosome decreases to 4.1 and 1.6, respectively (Supplementary Fig. 3). Thus, even though long A-tracts tend to be excluded from the nucleosome[30], A-tracts of ≤5 bp, when present, are used to facilitate bending of the DNA around the histone core.

To evaluate the effect of TpA steps, we compared the periodicities of A-tracts of length three to those of other trinucleotides composed only of AT base pairs. Trinucleotides that contain TpA steps have a much weaker periodic signal than A-tracts of length three, which exclude the TpA step (Supplementary Fig. 4). Together, this analysis suggests that many of the sequence periodicities in nucleosomal DNA reflect the presence of short A-tracts that lead to narrow regions in the minor groove, which are in turn recognized by a complementary set of arginines present on the surface of the nucleosome core particle.

Effects of groove width on electrostatic potential

The remarkable correlation between minor-groove width and electrostatic potential (Figs 3 and 4) is primarily due to the properties of the Poisson–Boltzmann equation that have been extensively discussed in the literature[31]. Biological macromolecules are less polarizable than the aqueous solvent and, in the language of classical physics, can be thought of as a low dielectric region embedded in a high dielectric solvent. Solutions of the Poisson–Boltzmann equation for DNA showed that contours of electrostatic potential owing to backbone phosphates follow the shape of the DNA and that the potentials are the most negative within the grooves[32]. This effect is due to electrostatic focusing, first observed for the protein superoxide dismutase[31], where the narrow active site focuses electric field lines away from the protein and into the high dielectric solvent. The same physical phenomenon produces enhanced potentials in grooves, accounting for the strong correlation described earlier.

To establish the source of the effect in quantitative terms, we calculated the potentials for the MogR-binding site[19] when the dielectric constant is set to 80 both inside the macromolecule and in the solvent (Fig. 5, dashed line) and for the case where the two dielectric constants are different (Fig. 5, solid line). Notably, a large enhancement of electrostatic potentials is only observed when the dielectric constant of the macromolecule and solvent are different, reflecting the focusing of electric field lines described qualitatively earlier. The small effect seen when the dielectric constant is the same results from the phosphates being closer to the centre of the groove when it is narrow (see Supplementary Fig. 5 for a breakdown of the contributions to the net electrostatic potential). Both sets of calculations were

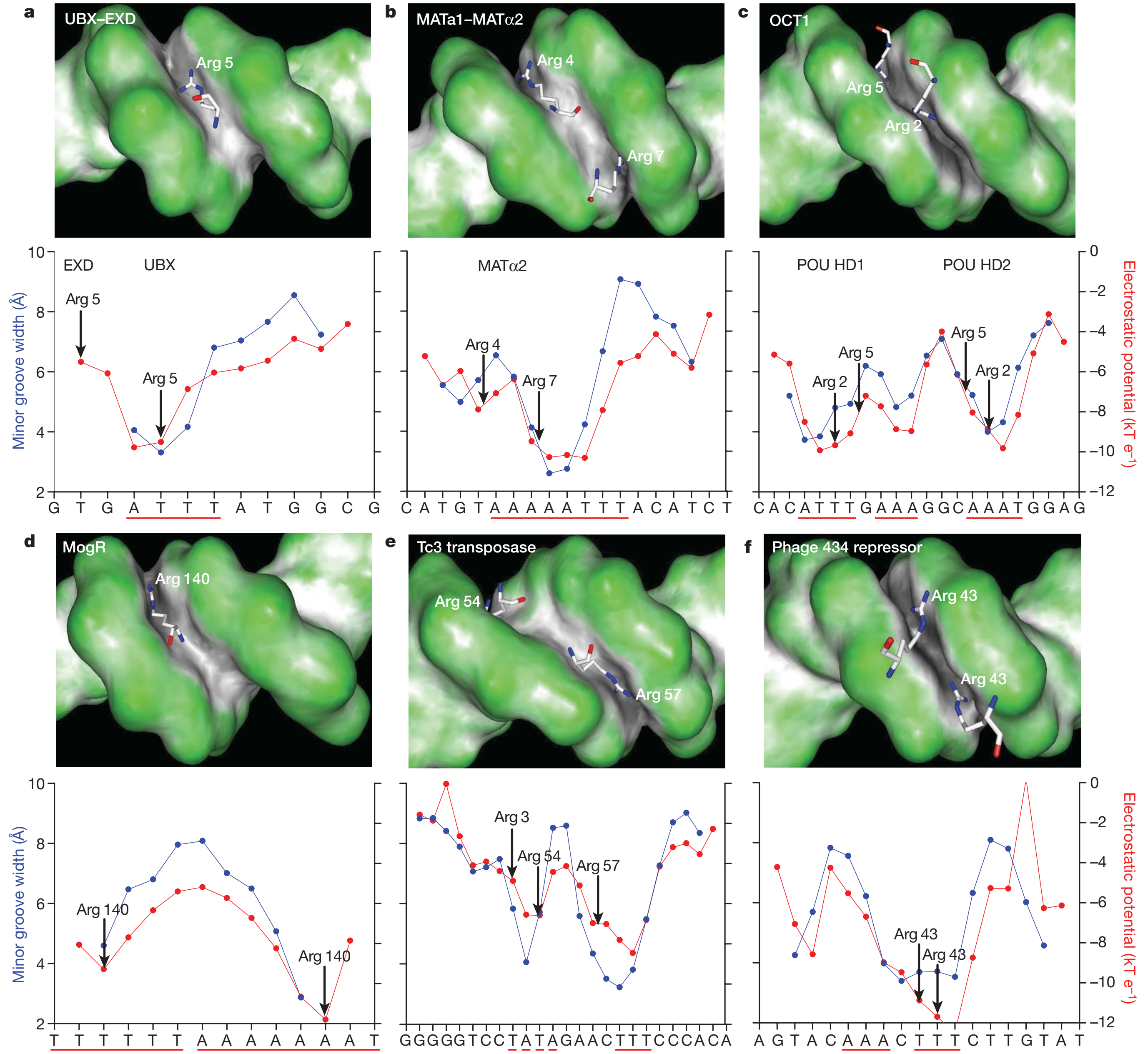

Figure 3 | Specific examples of minor-groove shape recognition by arginines. a–f, DNA shapes of the binding sites of UBX–EXD (PDB code 1b8i)[16] (**a**), MATa1–MATα2 (PDB code 1akh)[17] (**b**), and OCT1–PORE (PDB code 1hf0)[18] (**c**), the MogR repressor (PDB code 3fdq)[19] (**d**), the Tc3 transposase (PDB code 1u78)[22] (**e**) and the phage 434 repressor (PDB code 2or1)[23] (**f**) are shown in GRASP surface representations[31,47], with convex surfaces colour-coded in green and concave surfaces in grey/black. Plots of minor-groove width (blue) and electrostatic potential in the centre of the minor groove (red) are shown below. Arginine contacts (defined by the closest distance between the guanidinium groups and the bases) are indicated. A-tract sequences are highlighted by a solid red line, the TATA box in **e** by a dashed line.

carried out at physiological salt concentrations. Although ionic strength influences the absolute values of the potentials, the dielectric boundary effect remains essentially the same (Supplementary Fig. 6).

Why are arginines preferred over lysines?

It is surprising that there is a substantial population of arginines in the minor groove and a large enrichment when the groove is narrow, whereas the effects for lysines are more modest (Fig. 1a). Arginines have been known for some time to be enriched relative to lysines in protein–protein[33] and protein–DNA[34] interfaces, and the difference has generally been attributed to the ability of the guanidinium group to engage in more hydrogen bonds than the amino group of lysine[35]. To evaluate this idea we determined the number of hydrogen bonds formed by all the arginines and lysines in our data set that penetrate the minor groove. Surprisingly, on average, less than one hydrogen bond is formed by either amino-acid side chain to DNA (0.9 for arginine and 0.6 for lysine), and the standard deviations are such that this difference is insignificant (Supplementary Table 3).

An alternative explanation derives from the difference in the size of the cationic moieties of the two residues. According to the classical Born model, the solvation free energies of ions are proportional to the inverse of their radii[31], suggesting that it is energetically less costly to remove a charged guanidinium group from water than it is to remove the smaller amino group of a lysine. To test this quantitatively, we calculated the change in free energy in transferring arginine and lysine from water to a medium of dielectric constant 2 (see Methods for details). The difference in the transfer free energies between the two residues ranges from 2.3 to 6.6 kcal mol^{-1}, depending on the force field that was used, with lysine consistently having the higher value (Supplementary Table 4). These results indicate that

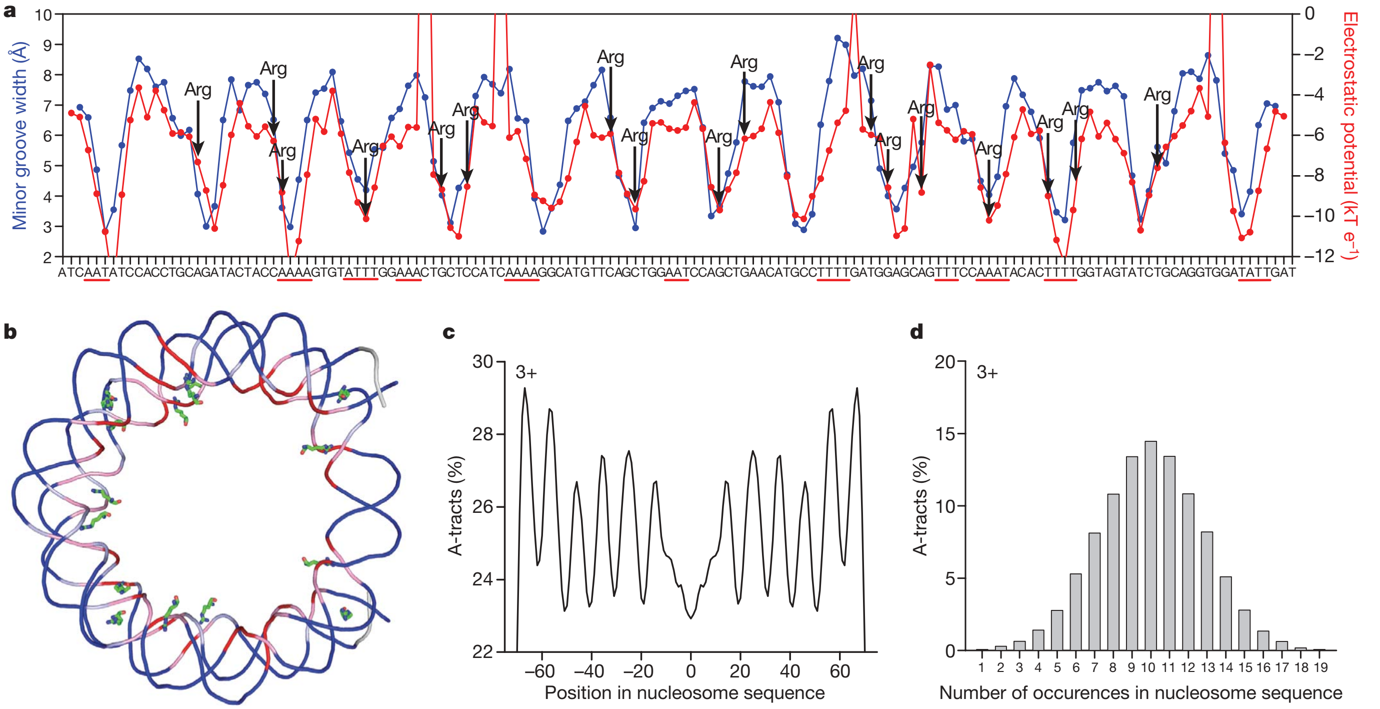

Figure 4 | Minor-groove shape recognition in the nucleosome. **a**, Correlation of minor-groove width of the nucleosome core particle (PDB code 1kx5)[24] (blue) and electrostatic potential (red). Arginine contacts (defined by the closest distance between the guanidinium groups and the bases) are indicated. A-tract sequences are highlighted by solid red lines. **b**, Schematic representation of the DNA backbone in the nucleosome colour-coded by minor-groove width (red ≤4.0 Å, pink >4.0 Å and ≤5.0 Å, light blue >5.0 Å and ≤6.0 Å, dark blue >6.0 Å), including all arginines that contact the minor groove. **c**, The distribution of A-tracts of length 3 bp or longer in 23,076 yeast nucleosome-bound DNA sequences[29]. **d**, Histogram of the occurrence of A-tracts of length three or longer in the same data set[29].

the higher prevalence of arginines compared to lysines in minor grooves is due, at least in part, to the greater energetic cost of removing a charged lysine from water than removing a charged arginine.

Concluding remarks

We have shown that there is a marked enrichment of arginines in narrow regions of the DNA minor groove that provides the basis for a new DNA recognition mechanism that is used by many families of DNA-binding proteins. A readout mechanism on the basis of groove width requires a connection between sequence and shape. This connection seems to be provided in part by A-tracts, which have a strong tendency to narrow the groove, producing binding sites for arginines that, when spaced appropriately on the protein surface, offer a complementary set of positive charges that can recognize local variations in shape. Arginines often insert into the minor groove as part of short sequence motifs (for example, Arg-Gln-Arg in the Hox protein SCR[12], Arg-Lys-Lys-Arg in POU homeodomains[18], Arg-Pro-Arg in Engrailed[36], Arg-Gly-His-Arg in MATa1–MATα2 (ref. 17), Arg-Arg-Gly-Arg in the nuclear orphan receptor[37] and Arg-Gly-Gly-Arg in the human orphan receptor[38]), thus offering a variety of presentation modes that can contribute to the specificity of DNA shape recognition.

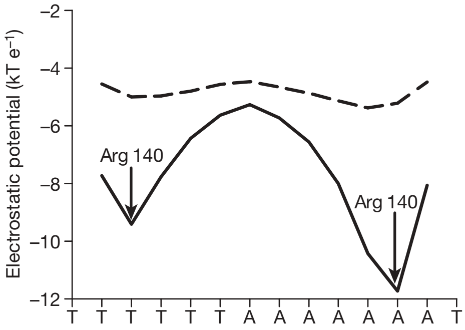

Figure 5 | The biophysical origins of the negative potential of narrow minor grooves. Electrostatic potential in the minor groove of the MogR-binding site (PDB code 3fdq)[19], calculated in the presence of a dielectric boundary ($\varepsilon = 2$ in solute and $\varepsilon = 80$ in solvent, solid line) and in the absence of a boundary ($\varepsilon = 80$ in both solute and solvent, dashed line).

The tendency of A-tracts to narrow the minor groove is primarily due to their ability to assume conformations, by propeller twisting, that lead to the formation of inter-base-pair hydrogen bonds in the major groove[15]. This network is disrupted by TpA steps as notably seen in the MogR-binding site[19]. GC base pairs also have a tendency to widen the minor groove. The combination of these and other factors, such as the effects induced by flanking bases that are not directly located within the binding site[39], can produce a complex minor-groove landscape that offers numerous possibilities for specific interactions with proteins. Indeed, minor-groove geometry is no doubt the result of the interaction of intrinsic and protein-induced structural effects.

The physical mechanisms described here are markedly evident in the nucleosome. The energetic cost of narrowing and bending the DNA in regions where the backbone faces inward will be reduced by the presence of short A-tracts that have an intrinsic propensity to assume such conformations and hence to bend the DNA[28]. Furthermore, the penetration of arginines into the minor groove at sites where the DNA bends and the groove is narrow[21,40] provides an important stabilizing interaction.

The variations in DNA shape observed in protein–DNA complexes often reflect conformational preferences of free DNA[4,10,41]. Sequence-dependent conformational preferences have also been observed in computational studies[11,21,42] and, most recently, analysis of hydroxyl radical cleavage patterns shows that DNA shape is under evolutionary selection[43]. Such observations indicate that the role of DNA shape must be taken into consideration when annotating entire genomes and predicting transcription-factor-binding sites. The biophysical insights described here, together with the increased availability of high-throughput binding data, offer the hope of major progress in understanding how proteins recognize specific DNA sequences and in the development of improved predictive algorithms.

METHODS SUMMARY

Minor-groove geometry was analysed with Curves[44] for all 1,031 crystal structures of protein–DNA complexes in the PDB that have any amino acid contacting base atoms. Protein side chains contact the minor groove in 69% of those structures that have at least one helical turn of DNA. The probabilities for each amino acid to contact the minor groove were calculated for three groups of DNAs: total, narrow and not narrow. Proteins were grouped on the basis of 40% sequence identity. The properties of free DNAs and DNAs bound to proteins were analysed on the basis of the minor-groove widths of tetranucleotides, defined at the central base-pair step.

All 35 crystal structures of the nucleosome available in the PDB were analysed. The analysis of nucleosomal DNA is based on 23,076 sequences in an *in vivo* yeast data set[29]. The signal for a sequence motif in nucleosomal DNA is positive for a base pair when the base pair comprises any part of the sequence motif. Frequencies were symmetrized by analysing both complementary DNA strands.

Electrostatic potentials were obtained from solutions to the non-linear Poisson–Boltzman equation at physiological ionic strength using the DelPhi program[31,45]. Regions inside the molecular surface of the DNA were assigned a dielectric constant of 2, whereas the solvent was assigned a value of 80. The potential is reported at a reference point at the centre of the minor groove. The reference point is located close to the bottom of the groove in approximately the plane of a base pair. This definition provides a measure of electrostatic potential as a function of base sequence. Solvation free energies of amino acids were calculated for extended conformations of arginine and lysine side chains and compared for four different force fields.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 24 July; accepted 2 September 2009.

1. Garvie, C. W. & Wolberger, C. Recognition of specific DNA sequences. *Mol. Cell* **8,** 937–946 (2001).
2. Seeman, N. C., Rosenberg, J. M. & Rich, A. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA* **73,** 804–808 (1976).
3. Travers, A. A. DNA conformation and protein binding. *Annu. Rev. Biochem.* **58,** 427–452 (1989).
4. Shakked, Z. *et al.* Determinants of repressor/operator recognition from the structure of the *trp* operator binding site. *Nature* **368,** 469–473 (1994).
5. Lu, X. J., Shakked, Z. & Olson, W. K. A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.* **300,** 819–840 (2000).
6. Hegde, R. S., Grossman, S. R., Laimins, L. A. & Sigler, P. B. Crystal structure at 1.7 Å of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target. *Nature* **359,** 505–512 (1992).
7. Kim, Y., Geiger, J. H., Hahn, S. & Sigler, P. B. Crystal structure of a yeast TBP/TATA-box complex. *Nature* **365,** 512–520 (1993).
8. Kim, J. L., Nikolov, D. B. & Burley, S. K. Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature* **365,** 520–527 (1993).
9. Otwinowski, Z. *et al.* Crystal structure of *trp* repressor/operator complex at atomic resolution. *Nature* **335,** 321–329 (1988).
10. Hizver, J., Rozenberg, H., Frolow, F., Rabinovich, D. & Shakked, Z. DNA bending by an adenine–thymine tract and its role in gene regulation. *Proc. Natl Acad. Sci. USA* **98,** 8490–8495 (2001).
11. Rohs, R., Sklenar, H. & Shakked, Z. Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure* **13,** 1499–1509 (2005).
12. Joshi, R. *et al.* Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* **131,** 530–543 (2007).
13. Burkhoff, A. M. & Tullius, T. D. Structural details of an adenine tract that does not cause DNA to bend. *Nature* **331,** 455–457 (1988).
14. Haran, T. E. & Mohanty, U. The unique structure of A-tracts and intrinsic DNA bending. *Q. Rev. Biophys.* **42,** 41–81 (2009).
15. Crothers, D. M. & Shakked, Z. in *Oxford Handbook of Nucleic Acid Structures* (ed. Neidle, S.) 455–470 (Oxford Univ. Press, 1999).
16. Passner, J. M., Ryoo, H. D., Shen, L., Mann, R. S. & Aggarwal, A. K. Structure of a DNA-bound Ultrabithorax–Extradenticle homeodomain complex. *Nature* **397,** 714–719 (1999).
17. Li, T., Jin, Y., Vershon, A. K. & Wolberger, C. Crystal structure of the MATa1/MATα2 homeodomain heterodimer in complex with DNA containing an A-tract. *Nucleic Acids Res.* **26,** 5707–5718 (1998).
18. Reményi, A. *et al.* Differential dimer activities of the transcription factor Oct-1 by DNA-induced interface swapping. *Mol. Cell* **8,** 569–580 (2001).
19. Shen, A., Higgins, D. E. & Panne, D. Recognition of AT-Rich DNA binding sites by the MogR repressor. *Structure* **17,** 769–777 (2009).
20. Stefl, R., Wu, H., Ravindranathan, S., Sklenar, V. & Feigon, J. DNA A-tract bending in three dimensions: solving the dA_4T_4 vs. dT_4A_4 conundrum. *Proc. Natl Acad. Sci. USA* **101,** 1177–1182 (2004).
21. Tolstorukov, M. Y., Colasanti, A. V., McCandlish, D. M., Olson, W. K. & Zhurkin, V. B. A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *J. Mol. Biol.* **371,** 725–738 (2007).
22. Watkins, S., van Pouderoyen, G. & Sixma, T. K. Structural analysis of the bipartite DNA-binding domain of Tc3 transposase bound to transposon DNA. *Nucleic Acids Res.* **32,** 4306–4312 (2004).
23. Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M. & Harrison, S. C. Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science* **242,** 899–907 (1988).
24. Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. Solvent mediated interactions in the structure of the nucleosome core particle at 1.9 Å resolution. *J. Mol. Biol.* **319,** 1097–1113 (2002).
25. Trifonov, E. N. & Sussman, J. L. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc. Natl Acad. Sci. USA* **77,** 3816–3820 (1980).
26. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442,** 772–778 (2006).
27. Satchwell, S. C., Drew, H. R. & Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* **191,** 659–675 (1986).
28. Travers, A. A. & Klug, A. in *DNA Topology and its Biological Effects* (eds Cozzarelli, N. R. & Wang, J. C.) 57–106 (Cold Spring Harbor Press, 1990).
29. Field, Y. *et al.* Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLOS Comput. Biol.* **4,** e1000216 (2008).
30. Segal, E. & Widom, J. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.* **19,** 65–71 (2009).
31. Honig, B. & Nicholls, A. Classical electrostatics in biology and chemistry. *Science* **268,** 1144–1149 (1995).
32. Jayaram, B., Sharp, K. A. & Honig, B. The electrostatic potential of B-DNA. *Biopolymers* **28,** 975–993 (1989).
33. Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. Studies of protein-protein interfaces: a statistical analysis of the hydrophobic effect. *Protein Sci.* **6,** 53–64 (1997).
34. Nadassy, K., Wodak, S. J. & Janin, J. Structural features of protein-nucleic acid recognition sites. *Biochemistry* **38,** 1999–2017 (1999).
35. Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.* **29,** 2860–2874 (2001).
36. Kissinger, C. R., Liu, B. S., Martin-Blanco, E., Kornberg, T. B. & Pabo, C. O. Crystal structure of an engrailed homeodomain-DNA complex at 2.8 Å resolution: a framework for understanding homeodomain-DNA interactions. *Cell* **63,** 579–590 (1990).
37. Meinke, G. & Sigler, P. B. DNA-binding mechanism of the monomeric orphan nuclear receptor NGFI-B. *Nature Struct. Biol.* **6,** 471–477 (1999).
38. Gearhart, M. D., Holmbeck, S. M., Evans, R. M., Dyson, H. J. & Wright, P. E. Monomeric complex of human orphan estrogen related receptor-2 with DNA: a pseudo-dimer interface mediates extended half-site recognition. *J. Mol. Biol.* **327,** 819–832 (2003).
39. Rohs, R., West, S. M., Liu, P. & Honig, B. Nuance in the double-helix and its role in protein-DNA recognition. *Curr. Opin. Struct. Biol.* **19,** 171–177 (2009).
40. Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423,** 145–150 (2003).
41. Locasale, J. W., Napoli, A. A., Chen, S., Berman, H. M. & Lawson, C. L. Signatures of protein-DNA recognition in free DNA binding sites. *J. Mol. Biol.* **386,** 1054–1065 (2009).
42. Tolstorukov, M. Y., Virnik, K. M., Adhya, S. & Zhurkin, V. B. A-tract clusters may facilitate DNA packaging in bacterial nucleoid. *Nucleic Acids Res.* **33,** 3907–3918 (2005).
43. Parker, S. C., Hansen, L., Abaan, H. O., Tullius, T. D. & Margulies, E. H. Local DNA topography correlates with functional noncoding regions of the human genome. *Science* **324,** 389–392 (2009).
44. Lavery, R. & Sklenar, H. Defining the structure of irregular nucleic acids: conventions and principles. *J. Biomol. Struct. Dyn.* **6,** 655–667 (1989).
45. Rocchia, W. *et al.* Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J. Comput. Chem.* **23,** 128–137 (2002).
46. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540 (1995).
47. Petrey, D. & Honig, B. GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol.* **374,** 492–509 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by National Institutes of Health (NIH) grants GM54510 (R.S.M.) and U54 CA121852 (B.H. and R.S.M.). The authors thank A. Califano for many helpful conversations.

Author Contributions R.R., A.S., R.S.M. and B.H. contributed to the original conception of the project; S.M.W. and R.R. generated and analysed the data assisted by P.L.; and R.R., S.M.W., R.S.M. and B.H. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to B.H. (bh6@columbia.edu) or R.S.M. (rsm10@columbia.edu).

METHODS

Calculation of minor-groove width. There were in total 1,031 crystal structures of protein–DNA complexes in the PDB as of 1 June 2008, in which the DNA was contacted by any amino-acid side chain at a distance <6.0 Å from base atoms. Of these structures, 567 contained at least one helical turn, and no chemical modifications or deformations that prevent the calculation of minor-groove width. Groove geometry was analysed using Curves[44] and minor-groove width was calculated as a function of base sequence by averaging all the Curves levels given for each nucleotide.

Statistical analysis of protein–DNA contacts. Of the 567 protein–DNA structures in our data set, 392 have at least one minor-groove contact defined by a distance of <6.0 Å between any base and side-chain atoms. To avoid an oversampling bias, proteins in this data set that shared $\geq 40\%$ sequence identity were grouped to create 109 groups. The average number of contacts within each group was subsequently averaged over all 109 groups. These averages were divided by the sum of the average number of contacts for all amino acids to calculate the total minor-groove contacts, contacts in not narrow minor grooves (≥ 5.0 Å), and contacts in narrow minor grooves (<5.0 Å), for each amino acid.

Hydrogen-bond contacts between amino-acid side chains and the DNA bases and phosphates, water molecules and other protein atoms were identified with the HBplus program[48].

Structural annotation of DNA-binding proteins. The proteins in our data set of protein–DNA complexes were classified in SCOP[46] superfamilies. Proteins for which SCOP annotations were not available were annotated manually or using the ASTRAL database[49].

Correlation of tetranucleotide structure and sequence. Tetranucleotides in free DNA and protein–DNA complexes were used to analyse the base sequence propensity of minor-groove regions as a function of minor-groove width. The minor-groove width of a tetranucleotide was defined by the average of all Curves[44] levels for groove width of the second nucleotide and the first level of the third nucleotide, which describes groove width at the central base-pair step. End regions and irregular tetranucleotides were excluded by requiring groove width definitions for at least one Curves level of each of the four nucleotides. Tetranucleotides from nucleosomal DNA were excluded from this analysis because the DNA is strongly deformed and the spacing between narrow regions is fixed at about one helical turn, thus adding a bias to the results. When applied to the 521 protein–DNA complexes in our data set, these criteria allowed the analysis of all 136 possible unique tetranucleotides. When applied to the 88 free DNA structures in our data set, the same criteria resulted in the analysis of 59 unique tetranucleotides. To increase coverage for the free DNA data set, NMR structures were included if dipolar coupling data were used in the refinement.

Propensity of sequence motifs in nucleosomes. The structural analysis of nucleosomes includes all 35 crystal structures in the PDB as of 1 May 2009. The sequence analysis was based on 23,076 nucleosome sequences of length 146–148 bp in a yeast *in vivo* data set[29]. These nucleosome sites were scanned for sequence motifs such as A-tracts of different lengths, TpA steps, or other AT-rich regions. A given motif contributed to a positive signal for any base pair that overlapped that motif, thus longer motifs contributed signals to more nucleotide positions. The frequencies of all motifs were symmetrized by analysing both complementary strands.

Calculations of electrostatic potentials. Electrostatic potentials were obtained from solutions to the non-linear Poisson-Boltzman equation at 0.145 M salt using the DelPhi program[31,45]. Partial charges and atomic radii were taken from the Amber force field[50]. The interior of the molecular surface of the solute molecule (calculated with a 1.4 Å probe sphere) was assigned a dielectric constant of $\varepsilon = 2$, whereas the exterior aqueous phase was assigned a value of $\varepsilon = 80$. Debye–Hückel boundary conditions and five focusing steps were used with a cubic grid size of 165 (a grid size of 185 was used for the nucleosome).

The electrostatic potential is reported at a reference point close to the bottom of the minor groove approximately in the plane of base pair *i*. The reference point *i* is defined as the geometric midpoint between the O4′ atoms of nucleotide $i + 1$ in the 5′–3′ strand, and nucleotide $i - 1$ in the 3′–5′ strand[12]. Where the DNA strongly bends into the major groove the reference point can clash with the guanine amino group and cause large positive potentials (as seen in Fig. 4a for three regions of the nucleosome).

Desolvation free energies were calculated with the DelPhi program[31,45] for the transfer of arginine and lysine side chains in extended conformations from water to a medium of dielectric constant $\varepsilon = 2$. Transfer free energies were calculated for each of the two side chains based on charge distributions and atomic radii taken from Amber[50] and three other force fields (see Supplementary Table 4).

48. McDonald, I. K. & Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793 (1994).
49. Brenner, S. E., Koehl, P. & Levitt, M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* **28**, 254–256 (2000).
50. Cornell, W. D. *et al.* A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).