CHAPTER 3

Opposites Attract: Shape and Electrostatic Complementarity in Protein-DNA Complexes

ROBERT C. HARRIS,[a,b,†] TRAVIS MACKOY,[b,†] ANA CAROLINA DANTAS MACHADO,[c,†] DARUI XU,[b,d] REMO ROHS*[c] AND MARCIA OLIVEIRA FENLEY*[b]

[a] Department of Physics and [b] Institute of Molecular Biophysics, Florida State University, Tallahassee, FL 32306, USA; [c] Molecular and Computational Biology Program, Departments of Biological Sciences and Chemistry, University of Southern California, 1050 Childs Way, Los Angeles, CA 90089, USA; [d] Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA
*Email: rohs@usc.edu; mfenley@sb.fsu.edu

3.1 Introduction

Protein-DNA binding plays a key role in both gene regulation – through highly specific transcription factor-DNA interactions – and in chromatin organization, through the largely non-specific interactions of histones and other architectural proteins with DNA.[1] This biological importance has inspired the rapid expansion in structural knowledge of protein-DNA binding, derived from X-ray crystallographic and NMR spectroscopic data,[2] ever since

[†] These authors contributed equally.

RSC Biomolecular Sciences No. 24
Innovations in Biomolecular Modeling and Simulations: Volume 2
Edited by Tamar Schlick

Published by the Royal Society of Chemistry, www.rsc.org

the first crystal structures of protein-DNA complexes were solved.[3,4] Several comprehensive reviews have summarized the experimental structural data on protein-DNA complexes.[1,5–10] While many of these reviews concentrated on the structure of the protein in the complex, others have examined the importance of DNA structure.[11–14] For several decades, the complementarity between both the shape and the electrostatic surface potential of the binding partners has been noted,[15,16] but the influence of the structure of DNA on its electrostatic surface potential has only recently been discovered. Regions with a narrow minor groove are associated with enhanced negative electrostatic potential and greater numbers of bound arginines.[17] Therefore, to thoroughly understand protein-DNA binding both partners need to be examined in conjunction, rather than in isolation. This chapter extends the approach presented in an earlier review[1] and examines the roles of shape and electrostatic complementarity in protein-DNA complexes.

3.1.1 Shape and Electrostatic Complementarity

Globular DNA-binding domains of proteins usually have a convex shape complementary to the concave shape of the DNA major groove, and specific protein residues on the protein surface form hydrogen bonds with DNA bases. In addition to those major groove contacts, the protein often 'hugs' the DNA, forming hydrogen bonds and salt bridges with the phosphodiester backbone, and uses flexible linkers and tails to form contacts between protein residues and the minor groove.[17,18] Proteins and DNA adopt complementary shapes to optimize the contacts between both partners. For example, the sequence-dependent shape of a transcription factor-DNA binding site optimizes its contacts with the protein.[17] The ability of DNA to adopt complementary shapes is assisted by the remarkable flexibility of the double helix.[2] DNA can, for instance, wrap around histone octamers and form nucleosomes, which are the basic building blocks of chromatin fibers.[19,20]

The complementary shapes of proteins and DNA are therefore ostensibly important for binding, but this observation must be explained by reference to the underlying forces, such as electrostatic, van der Waals, and hydrophobic forces, that stabilize these complexes. The enrichment of the two basic side chains, arginine and lysine, among the residues contacting DNA[17] causes the binding interfaces of DNA-binding proteins to be predominantly positively charged. On the other hand, DNA is a polyanion with a negative net charge.[21] This negative charge is primarily located on the phosphate groups, which contain the bulk of the charge, while the deoxyribose moieties slightly counter this negative charge and provide an opportunity for hydrophobic interactions in addition to the hydrophobic functional groups of the bases, such as the thymine methyl group.

Although most research has examined shape and electrostatic complementarity in isolation, recent studies have revealed that they are actually closely related.[17] The shape of the dielectric boundary between the low dielectric medium within the protein or DNA molecule and the high dielectric

medium of the solvent enhances differences in electrostatic potential.[22] This effect was previously observed for cavities, first for the enzyme superoxide dismutase[23] and later for transfer RNA,[24] but it was only recently shown to play an important role for the enhancement of negative electrostatic potential in the narrow minor groove.[17] Emphasizing this connection between DNA shape and electrostatic potential and illustrating its role in protein-DNA binding is a goal of this book chapter.

3.1.2 Biological Examples for Shape and Electrostatic Complementarity in Protein-DNA Complexes

A thorough understanding of how shape and electrostatic potential control specificity in protein-DNA binding is expected to provide insights into many biological processes. Transcription factors, for example, differentiate between their *in vivo* DNA binding sites and similar sequences with high precision to activate their target genes. Hox proteins are such transcription factors, which play an important role in the embryonic development across species. These are homeodomain proteins that bind as heterodimers in complex with a cofactor differentiating between very similar target sites. For the *Drosophila* Hox protein Scr (Sex combs reduced), it was shown that nuances in shape and electrostatic potential of its specific DNA target site are recognized through minor groove contacts, enhancing binding specificity.[18] A key arginine of the Scr protein is attracted by the shape and electrostatic potential in a specific region of the DNA, which is slightly altered in minor groove width as a function of base sequence. As a result, the additional arginine-minor groove contact, which is not observed when Scr binds to a Hox consensus site, enhances the interaction of Scr with its *in vivo* specificity site.[18] High-resolution electrostatic potential maps, provided in this book chapter, allow for the exploration of such subtle differences in the electrostatic potential and its effect on protein-DNA binding specificity.

An additional mechanism for DNA to adopt a complementary form to its binding partner is through the deviation from the Watson-Crick double helix through the formation of Hoogsteen base pairs, which effectively extends the four-letter genomic alphabet.[25] In the past, Hoogsteen base pairing has only been observed in complexes with strongly deformed DNA structures, such as complexes with the TATA box binding protein[26] and the integration host factor (IHF).[27] More recently, for the first time, a single Hoogsteen base pair occurred in the context of undistorted B-DNA in a MATα2-DNA complex.[28] Another example is the binding of the tumor suppressor p53 to its various response elements. High-resolution crystal structures have revealed that certain base pairs in the p53 binding site assume Watson-Crick geometry for one sequence[29] and Hoogsteen geometry for a slightly different sequence.[30] The different base-pairing geometry affects the shape of the DNA-binding site, its electrostatic potential and, in turn, the strength of key p53-DNA interactions.[30]

3.1.3 Base and Shape Readout

The terms *direct* and *indirect readout* were coined to describe one of the first crystal structures of a protein-DNA complex, the binding of the tryptophan (trp) repressor to DNA.[4] However, over time and usage, the original meanings of these terms have been expanded to the point where their utility is questionable. To better understand the thousands of available X-ray and NMR structures,[2] we propose to use the more precise terms *base readout* when a protein recognizes a particular base pair, and *shape readout* when a protein recognizes a particular DNA shape.[1] These terms appear to be more useful and can be further subdivided to capture various protein-DNA readout modes.

3.2 Electrostatic Features of DNA

3.2.1 B-DNA

3.2.1.1 Watson-Crick Base Pairs

The electrostatic features of B-DNA (Figure 3.1[b], [e], [f]) are created both by the large negative potential surrounding the phosphate groups of the DNA backbone, which guarantee the overall negative potential on the molecular surface of DNA, and by sequence-dependent effects given by both the local base pair identity and conformation (Figure 3.2).[31] A decomposition of electrostatic potential, based on the chemical groups forming a nucleotide, was previously used to dissect these different contributions.[17] The electrostatic potential contributions from phosphates, bases, and sugar moieties are additive only for solutions of the linear Poisson-Boltzmann equation (LPBE). While the law of superposition does not hold for solutions of the non-linear Poisson-Boltzmann equation (NLPBE), the total potential still appears to be approximately the sum of these contributions. Examining the potential given only by the phosphate groups illustrates some of the electrostatic features of DNA. A further simplification is the model of formal charges where a charge of $-0.5e$ is assigned to each of the two exocyclic phosphate oxygen atoms.

As can be seen from Figures 3.1(a–c), based on formal charges, the phosphate groups on the backbone of B-DNA create a negative potential in both the major and minor grooves of the DNA, but the potential in the minor groove is slightly more negative than that in the major groove because of the closer proximity of the phosphate groups, and, as will be discussed below, due to the shape of the dielectric boundary.[17] Although the minor groove has a more negative potential than the major groove, some debate remains about whether the minor groove has a more negative potential than the phosphate backbone itself. Clearly, the contribution to the potential in the minor groove by the phosphate groups is not sufficient to make the minor groove more negative than the phosphate groups. When the base pairs' contribution is added, some results have indicated that the minor groove is more negative than

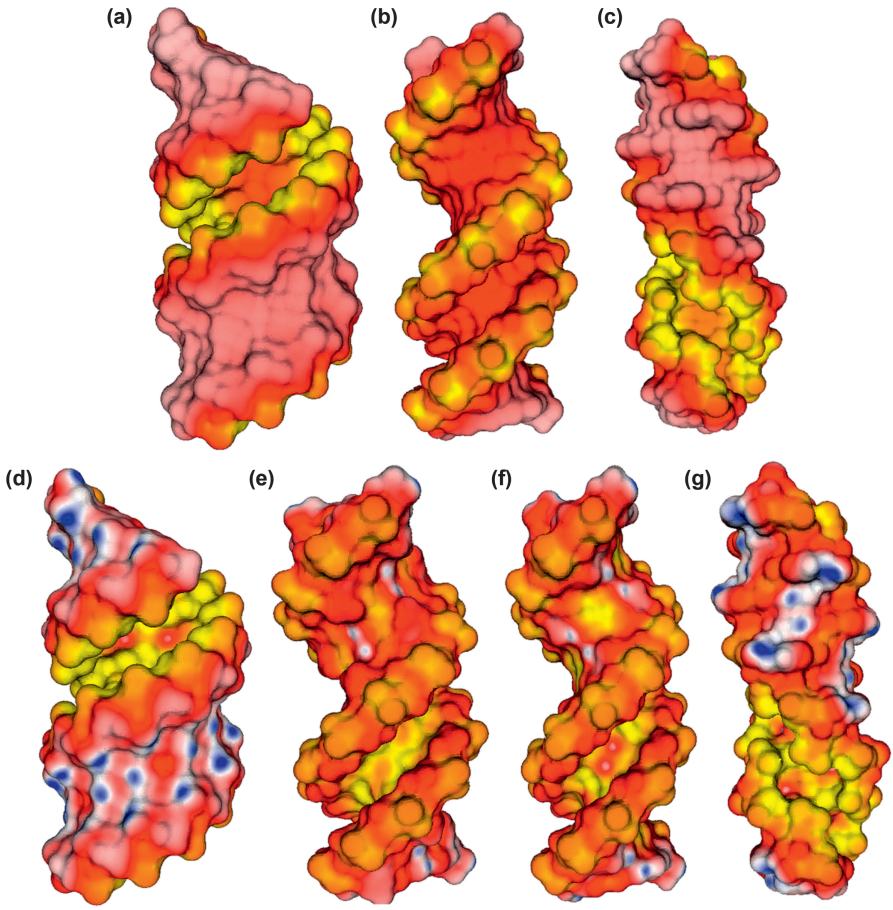

Figure 3.1 Electrostatic potential at the solvent-excluded surface, computed with a probe radius of 1.4 Å, of three ideal DNA double helices (a–c) with a simplified charge distribution of formal charges where all charges on the DNA are set to zero except those on the O1P and O2P atoms, which were set to $-0.5e$, and four additional ideal DNA double helices (d–g) with an all-atom charge distribution based on the Amber94 force field.[136] The electrostatic potential surface maps were computed with the adaptive Cartesian grid (ACG) nonlinear Poisson-Boltzmann equation solver[57] and the ideal DNA structures are based on fiber diffraction data. The electrostatic potential for the DNA modeled with the all-atom charge model is colored from the most negative to the most positive and ranges from -5 kT/e to $+5$ kT/e: *yellow-red*, negative; *white*, neutral; and *blue-green*, positive. On the other hand, the electrostatic potential for the DNA modeled with the 'formal charge' model is colored from the most negative to neutral and ranges from -5 kT/e to 0 kT/e: *yellow-red*, negative; *white*, neutral. The figure shows (a) a canonical A-DNA double-helix, (b) a canonical B-DNA double-helix, (c) a canonical Z-DNA double-helix, (d) a $d(GC)_7$ A-DNA double helix, (e) a $d(AT)_7$ B-DNA double-helix, (f) a $d(GC)_7$ B-DNA double-helix, and (g) a $d(GC)_7$ Z-DNA double-helix.

the phosphate groups,[32] and other results have not.[33] Potentially, this could be due to the use of different molecular mechanics force fields. For example, NLPBE calculations based on the Poltev *et al.* force field indicate that the potential on the phosphate backbone is stronger than the potential in the minor groove,[33] while the results presented here show a minor groove more negative than the phosphate backbone (Figure 3.1). In a more recent study, it was shown that the focusing of field lines in the minor groove, due to the shape of the dielectric boundary, causes enhanced negative electrostatic potential in the center of the groove.[17] By removing the dielectric boundary, it was demonstrated that the position of the phosphate groups alone cannot explain this effect.

When the contributions of the bases and sugar moieties to the potential are added, sequence-dependent effects become apparent. For ideal B-DNA helices shown in Figures 3.1(e) and (f), the potential in the minor groove of the AT-containing helix is about 0.5 kT/e more negative than in the GC helix. This change is caused by the presence of different electronegative and electropositive atoms in A/T and G/C base pairs, as is evident from Figures 3.2(c) and (f) where the potential at the minor groove edge of the G/C base pair is more positive than that at the minor groove edge of the A/T base pair, due to the presence of the electropositive amino N2 group of guanine and because A/T

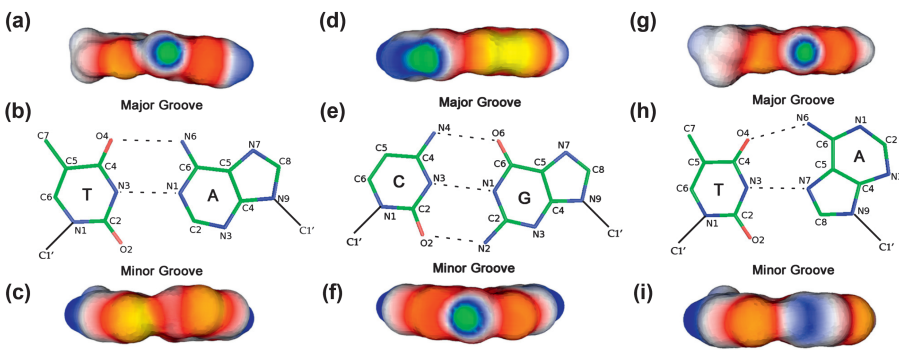

Figure 3.2 Electrostatic potential at the solvent-excluded surfaces, computed with a probe radius of 1.4 Å, of (a–c) a Watson-Crick A/T base pair, (d–f) a Watson-Crick G/C base pair, and (g–i) a Hoogsteen A/T base pair and schematic representations of these base pairs. The electrostatic potential is colored from the most negative to the most positive and ranges from −2 kT/e to +2 kT/e: *yellow-red*, negative; *white*, neutral; and *blue-green*, positive. The glycosidic bonds of the bases are represented by black lines. These electrostatic potential surface maps were generated with an ACG-based nonlinear Poisson-Boltzmann equation solver.[57] The top three images show views of the electrostatic potentials on the major groove edges of the three base pairs, the middle three images are schematic representations of the three base pairs, and the bottom three images show views of the electrostatic potentials on the minor groove edges of the three base pairs.

base pairs typically are located in regions of narrower minor grooves than G/C base pairs. These electrostatic features of B-DNA have been observed in many previous studies.[34–38] In turn, the enhanced negative potential in the AT minor groove is associated with a higher incidence of bound metal ions, cationic protein side chains, and cationic drugs.[17,39–41] The actual difference in potential between the minor groove of AT *vs.* GC-rich sequences varies based on the Poisson-Boltzmann solver used and was reported as 1–1.5 kT/e based on DelPhi calculations.[36] Many reasons can cause this difference in actual value, including the B-DNA model used, the definition of the molecular surface, and the location of the probe where the potential is measured. Importantly, however, the order of magnitude is similar and the electrostatic potential in the minor groove of AT-rich DNA is consistently more negative.

Sequence-dependent effects due to different functional groups at the base edges are also visible in the major groove of B-DNA, as seen in Figures 3.1(e) and(f). Although both base pairs contribute to the negative potential in the major groove due to a larger number of electronegative atoms (Figures 3.2[a] and [d]), both ideal B-DNA helices show periodic positive patches alternating with stronger negative regions. In AT sequences, they are created by the amino N6 and carbonyl O4 groups, and in GC sequences, they are created by the striking asymmetry in the electrostatic potential of the G/C base pair. The guanine of the G/C base pair is strongly negative, while the cytosine contains the electropositive N4 atom, leading to a positive electrostatic potential. This asymmetric potential is also biologically relevant, as studies have shown that metal ions and the cationic side chains of proteins preferentially bind the guanine of the G/C base pair, while small anionic inorganic species[42] (*e.g.* bisulfites) and anionic residues of proteins preferentially bind the cytosine.[43] Additionally, arginines select for guanines in the major groove through the formation of bidentate hydrogen bonds between their guanidinium group and the guanine. Such a bidentate hydrogen bond conveys a high degree of specificity and is, for instance, responsible for the conservation of both G/C base pairs in the CATG core element of the p53 binding site.[29,30,44]

3.2.1.2 Hoogsteen Base Pairs

Interestingly, not all protein-DNA complexes exclusively include canonical Watson-Crick base pairs. A few structures contain Hoogsteen base pairs, which provide an alternative hydrogen bonding geometry (Figure 3.2[h]). Although both A/T and G/C base pairs can assume a Hoogsteen conformation, A/T base pairs have been observed in several structures, like the TATA-box binding protein complex,[26] certain p53-DNA complexes,[30] complexes of DNA with intercalating drugs, and the MATα2-DNA complex,[28] while a Hoogsteen G/C base pair has only been observed in the DNA complex with polymerase iota.[45,46] This difference in frequency probably arises from the necessity of protonating the cytosine base at its N3 position to form the G/C Hoogsteen base pair, while no similar modification is

required to form the A/T Hoogsteen base pair.[47] Regardless, our discussion here will be restricted to the A/T Hoogsteen base pair because of its greater biological relevance.

Figures 3.2(d) and (g) indicate that the major groove of the A/T Hoogsteen base pair presents a similar electrostatic profile to that of the Watson-Crick A/T base pair but the N3 atom generates a negative patch of the adenine base facing the sugar moiety, which is not visible in the figure. As a consequence, the minor groove is significantly less negative (Figure 3.2[i]) because the electronegative N3 atom is no longer present. Since only a few structures have been identified yet with Hoogsteen A/T base pairs, more research is merited so as to determine whether this variation in electrostatic potential is employed in the specific recognition of Hoogsteen A/T base pairs.

One question that has been asked repeatedly in the study of Hoogsteen base pairs is whether these alternative conformations are induced by the presence of their binding partner or whether they co-exist transiently in solution and are identified by their binding partner. The current experimental evidence addressing this important question has provided different answers. For the MATα2-DNA complex, for example, no Hoogsteen base pairs were found in the unbound DNA.[28] A more recent NMR study, however, has found that unbound DNA containing certain base sequences, like CA and TA dinucleotides, can leave its Watson-Crick ground state and adopt Hoogsteen base pairing, at least transiently.[48] If unbound DNA does adopt Hoogsteen base pairing, then perhaps the differences in electrostatic potential noted above could play a role in recognition during the formation of protein-DNA complexes.

3.2.2 A-DNA

Although A-DNA is much less common than its B-form counterpart, A-DNA and its unique electrostatic properties appear to be important for the formation of certain protein-DNA complexes.[49] GC-rich DNA sequences can be induced to form A-DNA helices at low humidity, and GC-rich sequences have also been observed forming A-form or A/B intermediate conformations.[42,50–52] Apparently, the existence of local A-DNA regions is vital for the formation of certain protein-DNA complexes.[1] Since A-DNA typically contains Watson-Crick base pairs, in analogy to B-DNA, the differences in its electrostatic potential are primarily due to the different locations of the phosphate groups and the different shape of both grooves. While in B-DNA the minor groove is narrow and the major groove is wide, in A-DNA the major groove is extremely narrow and the minor groove is very wide and shallow.[1] This larger asymmetry in groove size brings the phosphate groups much closer together and the shape of the dielectric boundary causes a strong electrostatic focusing effect, which in turn generates a strong negative potential in the major groove (Figures 3.1[a] and [d]). On the other hand, the minor groove almost becomes a flattened surface, and parts of the minor groove are so far removed from the phosphates that they even display

regions with neutral or positive electrostatic potential. Apparently, the very negative potential in the major groove of A-DNA is associated with a higher density of bound cations, as X-ray studies have demonstrated that both magnesium and cobalt hexamine often occur in large quantities within the major groove, in some cases bridging the phosphate groups of both DNA strands across the major groove.[53–55] These observations agree with other computational studies,[56] which demonstrate that the density of counterions around A-DNA is highest in the major groove.

Additionally, many complexes require the formation of A-DNA, such as regions in GC-rich stretches of DNA, while mostly preserving the B-form of the remainder of the DNA. When this happens, the DNA bends around the GC region, and the major groove opens up, providing additional hydrogen bonding sites that can contribute to the binding specificity of a protein. One example of such a complex is the N-terminal DNA-binding domain of the Tc3 transposase of *C. elegans* and its transposon DNA. The bound DNA in the Tc3 transposase-DNA complex has a distinct and extensive electronegative potential patch on the G-side of its more A-like major groove, where cationic residues from the alpha helices bind.[57] Apparently, these unique features of the electrostatic potential around A-DNA are very relevant to biological processes. Another example where A-DNA plays a role in protein-DNA recognition is the family of zinc finger proteins, which bind to GC-rich DNA binding sites that exhibit conformations with some characteristics of A-DNA in their unbound state.[49]

3.2.3 Z-DNA

Z-DNA is a left-handed double-helix with a dinucleotide repeat, a smaller helical diameter compared to its B- and A-DNA counterparts, and a peculiar zig-zag arrangement of its sugar-phosphate backbone, which gave Z-DNA its name.[58] As can be seen in Figures 3.1(c) and (g), Z-DNA has a region of very close approach of the phosphate groups and associated narrow minor groove, which in turn generates enhanced negative electrostatic potential in its minor groove compared to the other two forms of DNA. Additionally, while the minor groove of A-DNA has an exceptionally small curvature, the major groove of Z-DNA almost becomes convex, yielding a structure very different from canonical B-DNA.[1]

The formation of Z-DNA is usually unfavorable, but under certain conditions, like high salt concentration, the presence of multivalent ions (*e.g.*, spermine), alternating purine-pyrimidine sequences (usually $d(GC)_2$), and negative DNA supercoiling, it can be stabilized, at least transiently, to perform certain biological functions, like the regulation of gene expression and DNA processing. Although this form of DNA is much less common than B- and A-form DNA, several examples of proteins that bind Z-DNA have been discovered. For example, proteins with the helix-turn-helix Zα domain[59] appear to select for the left-handed helical arrangement of phosphates in Z-DNA, as will

be discussed in the section on the electrostatic complementarity of protein-DNA complexes.

3.3 Structural Complementarity in Protein-DNA Complexes

The requirement that a protein be able to bind DNA places certain constraints upon its shape. These constraints, in turn, lead to the widespread incorporation of common DNA-binding motifs, like zinc-finger binding domains, helix-turn-helix motifs, and leucine zippers in proteins with widely varying functions.[7] Furthermore, during the formation of protein-DNA complexes, both partners can alter their conformation to accommodate the other binding partner. These allosteric transitions allow for the formation of hydrogen bonds and van der Waals contacts between the binding partners, as well as optimize hydrophobic packing and electrostatic interactions at the binding interface. Collectively, these structural adaptations enhance protein-DNA binding affinity and specificity, yielding striking complementarity between the binding partners.

Frequently, DNA must bend to form a complex because the protein requires either a wider major groove into which a DNA-binding domain can be inserted[60] or a narrow minor groove that an unstructured strand can recognize.[1] Additionally, many proteins bind DNA as dimers, and DNA bending is required to enhance interactions at the protein-protein interface. A widely studied example is the papillomavirus E2 protein. DNA is bent in complex with the E2 dimer,[61,62] which is a rare case where high-resolution crystal structures are available for the bound and unbound DNA of both high- and low affinity binding sites.[63,64] As also seen in computational predictions, the high-affinity site already assumes a bent conformation in its free state while the low-affinity site is essentially straight in its unbound form and needs to be bent upon binding.[65] Proteins that bind this way preferentially target AT-rich stretches and A-tracts because these regions of DNA are readily bent. In other complexes, adjacent base pairs of the DNA can become unstacked, which introduces a kink in the helix. Such a kink can allow a hydrophobic side chain of the protein to intercalate into the helix.[66] Still some protein-DNA complexes might require Hoogsteen base pairs while others demand even more radical alterations of the DNA structure.

In addition to the DNA, the protein also undergoes conformational changes upon binding. Many of the cationic side chains of DNA-binding proteins simply reposition themselves to create specific hydrogen bonds and van der Waals contacts and to engage in interactions with the attractive negative electrostatic potential that surrounds DNA.[1] Furthermore, many proteins undergo more global reorganizations upon binding. These changes can include allosteric transitions of loop conformations, as observed for the Runt domain-DNA complex.[67,68] More drastic examples include small peptides that are unstructured in solvent but form alpha helices upon binding,[69] and many larger proteins that contain recognition domains that are unstructured in solution.

3.4 Electrostatic Complementarity at the Interface of Protein-DNA Complexes

3.4.1 Complementarity of Electrostatic Potential Surface Maps

For several decades, researchers have observed that binding interfaces of DNA-binding proteins typically present strong positive electrostatic potential at the interface with their DNA binding partner.[70–72] Even the subset of DNA-binding proteins that possess a negative net charge, like the trp repressor and the Klenow fragment of *E. coli* DNA polymerase I, are typically highly dipolar, with a concentrated patch of positive potential near the DNA-binding interface. In the Klenow fragment of DNA polymerase I, the DNA sits in a deep cleft that has enhanced positive electrostatic potential while other neighboring regions have larger negative potential.[73,74] The idea that the interfaces of DNA-binding proteins should have good electrostatic potential complementarity with the DNA is attractive because of the DNA's high negative charge density, and the introduction of the GRASP software[22,75] has made an analysis of the electrostatic potential complementarity a routine step in the description of a DNA-binding protein. Indeed, analyzing the electrostatic potential complementarity has led to several successful predictions of the location of the protein-DNA interface for complexes whose structures had not yet been solved, and the size of the largest positive potential patch on a protein has been combined with physicochemical and geometric metrics in learning algorithms that predict whether or not a protein is a candidate for binding to DNA.[70–72,76–78]

The DNA-binding proteins discussed in this study also present a high degree of electrostatic potential complementarity (Figure 3.3). Large concentrations of cationic side chains at the binding interface of these complexes create strongly positive electrostatic patches. The electrostatic potential complementarity is particularly evident in the helix-turn-helix motifs that bind Z-DNA,[79] as can be seen in Figure 3.3(i), where the positive electrostatic potential follows the trace of the Z-DNA backbone.

Of course, making predictions of DNA-binding sites from electrostatic potential surface maps requires that the electrostatic potential surface map is numerically accurate. In particular, although using the LPBE to generate such surface potential depictions of electrostatic potential, as has been done for many biological molecules, may offer numerical advantages, the high charge densities of DNA can cause large deviations between the predictions of the LPBE and those of the NLPBE.[57] Additionally, great care must be taken to use a sufficiently fine mesh near the surface because the large charge densities at the protein-DNA interface can lead to rapid changes in potential near the molecular surface. In the present study, the fine mesh spacing near the molecular surface permitted in the ACG PB solver[57] allows for the computation of electrostatic surface potential maps at high resolution.

In addition to numerical approximations, insufficiently detailed knowledge of the binding partners can lead to poor predictions. For example, a protein-DNA complex was reported in the literature that did not appear to have a significant

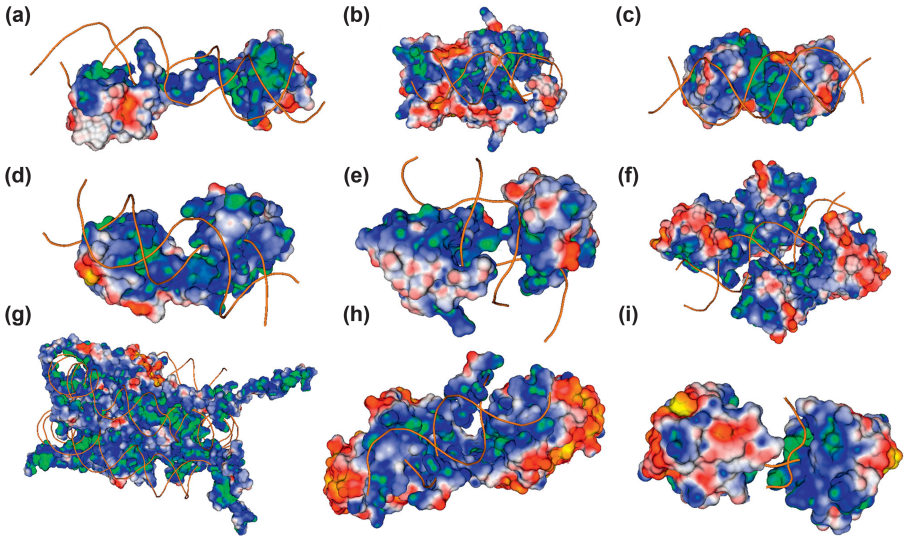

Figure 3.3 Electrostatic potentials at the solvent-excluded surfaces, computed with a probe radius of 1.4 Å, of nine DNA binding proteins or protein complexes: (a) the Tc3 transposase (PDB id: 1U78), (b) the integration host factor (IHF) (PDB id: 2NP2), (c) the phage 434 repressor (PDB id: 2OR1), (d) the MATa1-MATα2 dimer (PDB id: 1AKH), (e) the UBX-EXD dimer (PDB id: 1B8I), (f) the OCT1-PORE complex (PDB id: 1HF0), (g) the histone octamer of the nucleosome (PDB id: 1KX5), (h) the MogR repressor (PDB id: 3FDQ), and (i) the Zα domain of ADAR1 (PDB id: 1QBJ) shown with the traces of the phosphodiester backbones of their DNA partners, computed with the ACG-based nonlinear Poisson-Boltzmann equation solver.[57] These figures show the complementarity between the positive potential at the protein interface and the location of the DNA. The electrostatic potential is colored from the most negative to the most positive and ranges from −5 kT/e to +5 kT/e: *yellow-red*, negative; *white*, neutral; and *blue-green*, positive.

patch of positive potential near the binding interface. However, when the histidines near the interface were protonated, a positive patch appeared.[72] Therefore, obtaining accurate protonation states for titratable residues at protein-DNA interfaces is necessary to obtain a correct electrostatic potential map.

The electrostatic complementarity at protein-DNA interfaces can also be observed at the molecular surface of the DNA. Figure 3.4 illustrates that basic side chains preferentially contact DNA where the electrostatic potential is particularly negative. As was shown for narrow minor groove regions,[17] the negative electrostatic potential can be further enhanced due to the shape of the dielectric boundary through electrostatic focusing.[22] The presence of this effect is clearly visible when the electrostatic potential is compared in the presence of the dielectric boundary *vs*. a situation in which the boundary has been removed.[17] Thus, summarizing our observations, shape complementarity further enhances electrostatic complementarity.

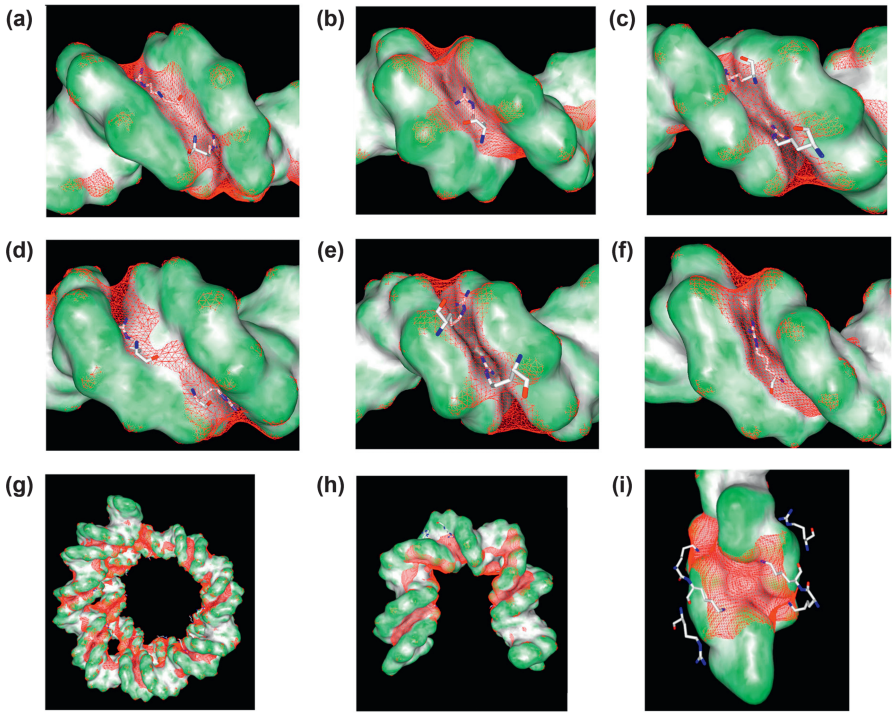

Figure 3.4 Shape and electrostatic potential of the DNA binding sites, computed with the nonlinear Poisson-Boltzmann solver DelPhi[108] and visualized using GRASP2,[75] of nine DNA structures that are bound to the following proteins or protein complexes: (a) the MATa1-MATα2 dimer (PDB id: 1AKH), (b) the UBX-EXD dimer (PDB id: 1B8I), (c) the OCT1-PORE complex (PDB id: 1HF0), (d) the Tc3 transposase (PDB id: 1U78), (e) the phage 434 repressor (PDB id: 2OR1), (f) the MogR repressor (PDB id: 3FDQ), (g) the histone octamer of the nucleosome (PDB id: 1KX5), (h) the integration host factor (IHF) (PDB id: 2NP2), and (i) the Zα domain of ADAR1 (PDB id: 1QBJ). The arginine residues contacting the minor groove within a distance of 6 Å from the base edges are shown in stick representations, and the red mesh represents the -5.0 kT/e isopotential surface calculated with the PB solver DelPhi[108] at physiologic ionic strength. The shape of the solvent accessibility surface is shown in GRASP2[75] surface representations where green represents convex surfaces and dark gray concave surfaces.

3.4.2 Approaches using Electrostatic Complementarity in Predictions of Binding Free Energy Changes

Historically, much of the interest in investigating the electrostatic potential complementarity between DNA, proteins, or ligands has been driven by the desire to predict the electrostatic component of the binding free energy.[80–85] If this energy could be obtained, then it could be combined with other approaches for deriving non-electrostatic components in order to compute the total binding

free energy. If such predictions could be made quickly and easily, then the binding energy of biomolecules and drugs could be optimized computationally, saving substantial time and effort in drug development. Of course, if the electrostatic potential complementarity of a complex is to be used to predict binding free energies, then the degree of electrostatic complementarity in that complex must be quantified. Several studies have therefore attempted to relate the binding free energy to the number of ion pairs, or close ionic contacts in the complex because of the observation that most cationic side chains that interact with DNA bind to the phosphate backbone.[86,87] Based on this model, an arginine or lysine residue is said to form an ion pair if the head group of its side chain lies within a certain cutoff distance to the nearest phosphate group. This model was the basis for the hypothesis that a measure of the salt dependence of the binding free energy ($SK = -dK/d\log[NaCl]$, where K is the binding constant and log[NaCl] is the logarithm of the 1:1 salt concentration) correlates with the number of ion pairs for small cationic organic drugs and small cationic oligopeptides.[88–91] In turn, drawing from theories derived from simple and elegant electrostatic models[92] that neglect dielectric discontinuity and all-atom details of the biopolyelectrolyte charge distribution, some studies have assumed that larger magnitudes of SK are associated with larger electrostatic components of the binding free energy.[87,93]

Early studies on protein-DNA association showed that the NLPBE accurately reproduces the salt dependence of the binding free energies of wild type and mutant proteins but these free energies do not consistently correlate with the number of ion pairs (*e.g.*, short-range phosphate contacts with cationic residues) that are formed upon complex formation.[81] Indeed, it was recently shown that cationic side chains do not preferentially bind to phosphate groups but also intrude the minor groove due to its enhanced negative electrostatic potential.[17] A clustering of cationic side chains near the phosphates, as observed in some complexes,[94] may reflect the smaller desolvation penalty for cationic side chains that contact the phosphate backbone. Additionally, while the NLPBE predicts that SK is correlated with the number of ionic contacts for small proteins and drugs,[90,91] this relationship does not hold for large proteins with either a significant number of anionic residues or a large number of charged residues distant from the binding interface.[95–97] Indeed, the proper way to extend the idea of using ion pairs, which was derived from small proteins and drugs binding to ideal and infinitely long DNA, to predict the electrostatic binding free energy of complexes with large proteins and deformed or finite DNA is unclear. Even the appropriate value of the cutoff distance, which is set at 6 Å in solution to match the Debye length in physiological salt conditions, is ill-defined, and a variety of values for this cutoff distance have been used in an effort to match SK with the structural data.[87] Additionally, many studies have shown that the electrostatic binding free energy is not correlated with SK in the presence of a dielectric discontinuity.[90,91] Although the Coulombic contribution to the electrostatic binding free energy is correlated with SK in the NLPBE for certain complexes,[98] this component is opposed by the desolvation penalty incurred by moving charged hydrophilic groups from the solvent into the low dielectric environment of the interior of the biomolecular complex. Whether a

given electrostatic interaction is favorable or unfavorable to binding is therefore a delicate balance between these two contributions to the energy, and thus the electrostatic binding free energy must be computed from a rigorous physics-based theory, like the NLPBE,[32,99,100] which can calculate these two energy components in a correlated manner. In fact, for most protein-DNA complexes the NLPBE and the generalized Born approach predict that the electrostatic binding free energy is actually unfavorable,[81,90,101,102] as opposed to the predictions of the simpler electrostatic models described above. Unfortunately, the NLPBE contains many free parameters, like the interior dielectric constant and the choice of molecular surface that must be determined by reference to pertinent experimental data. As the relationship between SK and the electrostatic binding free energy does not appear to hold (results not shown), the traditional experimental thermodynamic data for binding free energies, which are used to parameterize these theories, are not useful for this purpose. More theoretical and experimental work is therefore required to solve these difficult problems.

3.4.3 Relationship between Electrostatic Complementarity and Association/Dissociation Constants

If enhanced electrostatic complementarity is not necessarily associated with a more favorable binding free energy and if electrostatic interactions between the two components of the complex are generally unfavorable, the question then arises as to why proteins and DNA exhibit charge and/or electrostatic potential complementarity. One potential explanation is that the electrostatic complementarity between the two binding components provides a steering effect, which enhances their association rate.[103] This explanation receives some support from experimental data,[104] which indicates that the association rate of a protein-protein complex could be enhanced by increasing the electrostatic complementarity of the interface. However, although these changes typically increase the association constant, they also increase the corresponding dissociation constant. This finding also supports the NLPBE's prediction that determining whether or not an electrostatic interaction is favorable requires the correct inclusion of the desolvation penalty. Furthermore, several studies have shown that if the protein is slowly moved away from the DNA, a local electrostatic minimum is located in some distance from the bound conformation.[102,105] This indicates that electrostatics could create a transient encounter complex that would then merge to create the final bound complex.[106] Whether this model does indeed describe how actual protein-DNA complexes form should be further investigated by experimental and computational studies.

3.5 Minor Groove Shape Readout in Protein-DNA Complexes

As discussed in several studies, shape and electrostatics play an important role in protein-DNA binding based on the mechanism of minor groove shape

readout.[1,17,107] It was shown that electrostatic potential in the minor groove is correlated with groove width. Narrower grooves contain stronger negative electrostatic potential and are associated with a larger number of bound arginine residues.[17] These studies, therefore, proposed that arginine binding sites could be predicted by either measuring the minor groove width or computing the electrostatic potential in the minor groove as given by the NLPBE. Here, this work is revisited with an NLPBE solver that contains an adaptive Cartesian grid (ACG), allowing the use of fine mesh spacing near the molecular surface. Additionally, a new method was used for computing the potential at reference points in the minor groove (see Methods section) that eliminates the handful of anomalous potential spikes observed in previous data, which were due to clashes of reference points with the molecular surface in regions of DNA bending.

The data presented here confirm the basic findings of these previous studies,[17,18,107] while the finer mesh spacing employed by the ACG PBE solver reveals additional features in the potential not visible in the analogous figures in Rohs *et al.*[17] (Figures 3.5 and 3.6). In addition to the dependence of the potential on the width of the groove, the potential also appears to contain sequence-dependent effects due to the different chemical composition of the bases. The potential in the minor groove of the ideal B-form DNA with AT sequence was about 0.5 kT/e more negative than that of a GC sequence. This potential difference is smaller than that caused by the width of the minor groove (about 1.5 kT/e), but it is still large enough to be visible in the potential profiles. The minor groove potential presented here appears to display two sequence-dependent effects, the one caused by variations in minor groove width, and the ones due to different functional groups of the bases. The actual absolute values disagree with calculations based on the DelPhi program,[108–110] which was used in studies reporting that minor groove width generated fluctuations in the potential of up to 8 kT/e,[17,107] while the effects due to different functional groups of the bases were much smaller (about 1–1.5 kT/e).[36] This difference might be caused by a number of reasons discussed in earlier sections. The important result, however, is that the effect of DNA shape on electrostatic potential is now shown based on two independent methods. Whereas the effect of different functional groups of the bases does play a role, the impact of shape is larger than the sequence effect, based on DelPhi and ACG calculations. While some experimental measures of the electrostatic potential near the surface of DNA have reported values that are in better accordance with the ones reported here using ACG,[111] we caution to take the actual values of electrostatic potential too literally. Additional experimental measurements of electrostatic potential around nucleic acids in conjunction with different numerical PBE solvers are necessary to determine the origin of the observed discrepancies in absolute values derived from DelPhi and ACG calculations.

Interestingly, the correlation between the potential and the width of the minor groove does not appear to depend on the details of the local charge distribution, as can be seen from Figures 3.7 and 3.8. These figures present plots similar to the ones shown in Figures 3.5 and 3.6 except that formal charges were

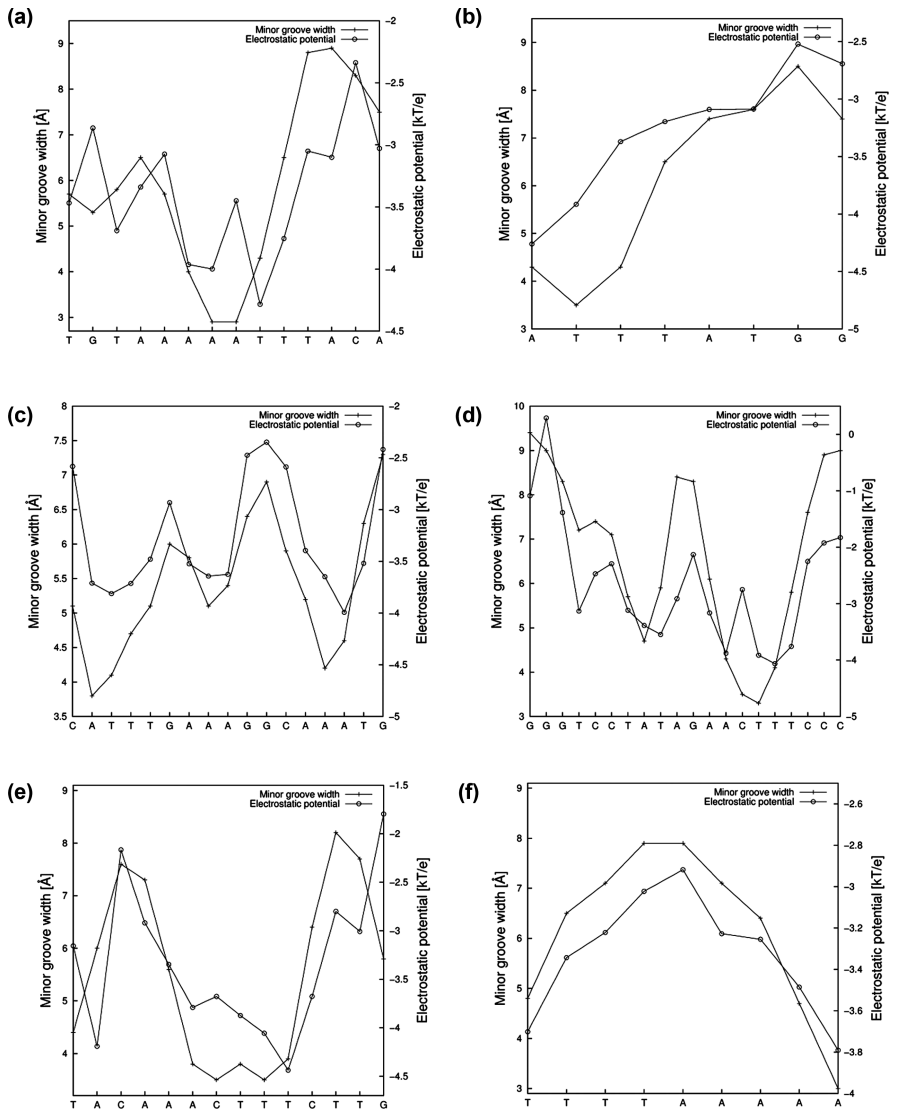

Figure 3.5 Electrostatic potential in the minor groove and minor groove width plotted as a function of sequence for several DNA double helices bound to various specific proteins or protein complexes. The figure includes (a) MATa1-MATα2 (PDB id: 1AKH), (b) UBX-EXD (PDB id: 1B8I), (c) OCT1-PORE (PDB id: 1HF0), (d) Tc3 transposase (PDB id: 1U78), (e) phage 434 repressor (PDB id: 2OR1), and (f) MogR repressor (PDB id: 3FDQ). The electrostatic potential was computed using the ACG nonlinear PBE solver[57] and the parameters described in the Methods section.

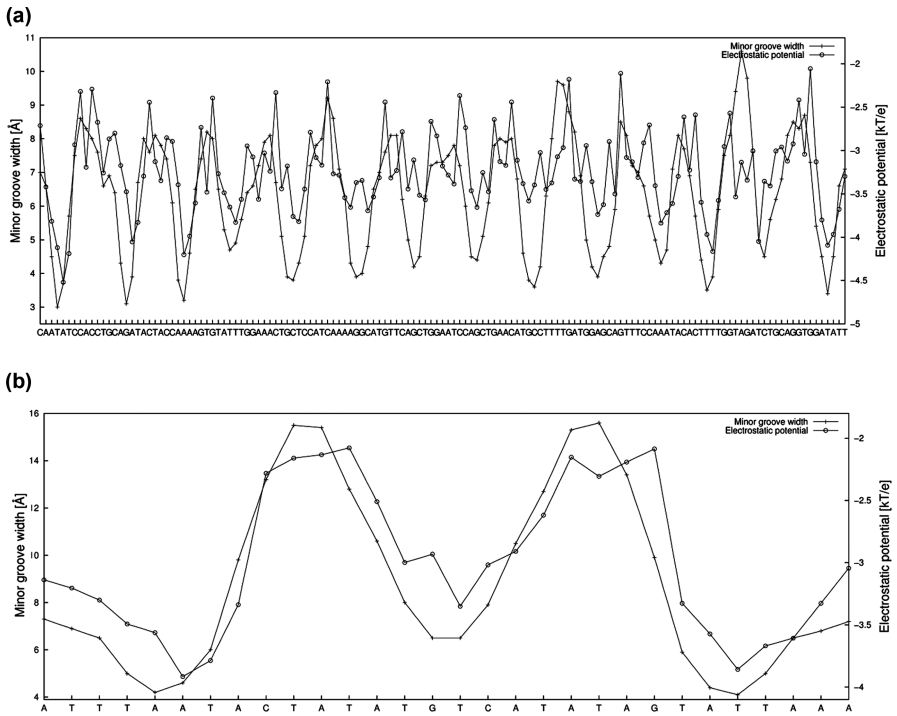

Figure 3.6 Electrostatic potential in the minor groove and minor groove width plotted as a function of sequence for the DNA conformation in two complexes with non-specific binding: (a) the nucleosome core particle (PDB id: 1KX5) and (b) the integration host factor (IHF) (PDB id: 2NP2). While we plot minor groove width and electrostatic potential as a function of base sequence, we are aware of the fact that the electrostatic potential is calculated in approximately the plane of a base pair but minor groove width, especially in 3DNA,[139] as phosphate-phosphate distance, which should actually be plotted between base pairs. Following this more stringent definition, the minor groove profile would shift by one half of a base pair in 3' direction, which would lead to an approximate symmetry of both the electrostatic potential and minor groove profile for the IHF-bound DNA. The electrostatic potential was computed using the ACG nonlinear PBE solver[57] and the parameters described in the Methods section.

used; all of the charges on the DNA were set to zero except for the O1P and O2P atoms, which were assigned a charge of $-0.5e$, where e is the fundamental charge. The sequence-dependent effects are not visible in these figures, indicating that the sequence-dependent effects can be attributed to the dipoles neglected in these calculations.

Additionally, the important observation that the dielectric boundary increases the fluctuations in the potential caused by the minor groove is confirmed by the results presented here. Figure 3.6 shows the potential in the minor groove of a nucleosome (PDB id: 1KX5) with a formal charge distribution

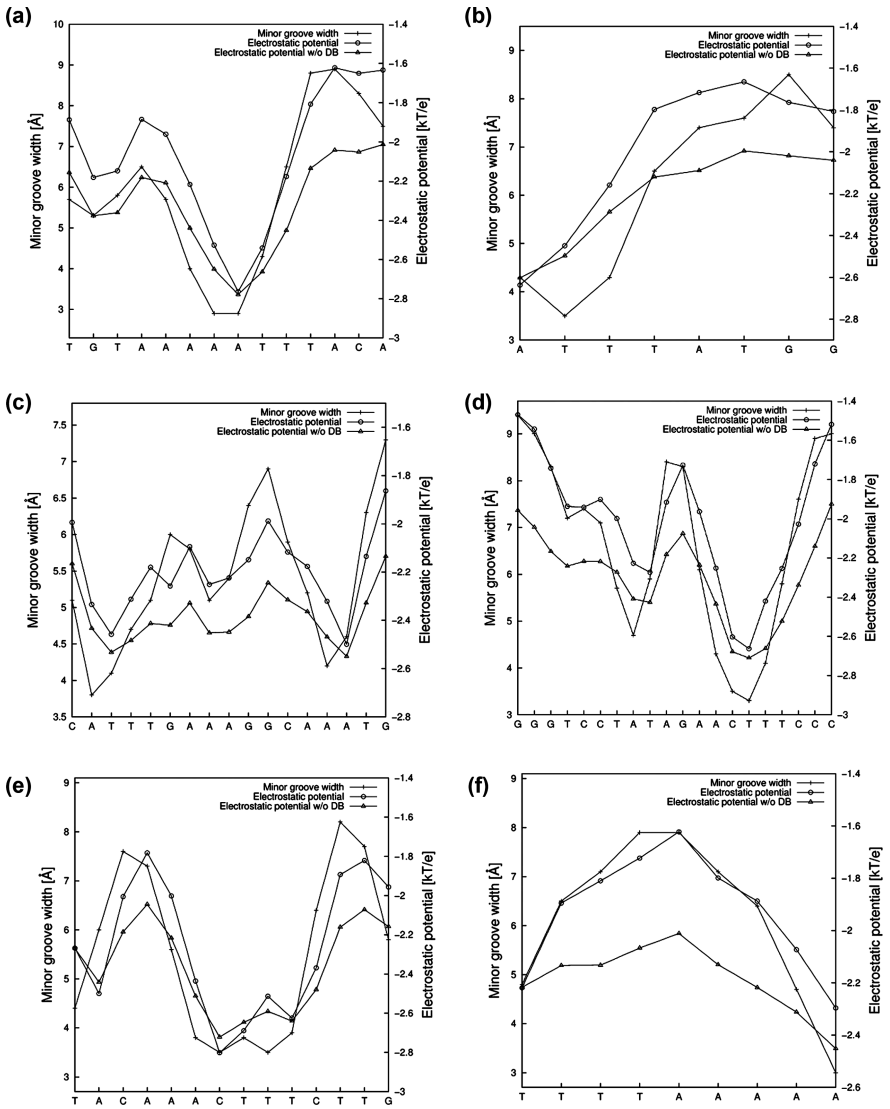

Figure 3.7 Electrostatic potential in the minor groove and minor groove width plotted as a function of sequence for several DNA double helices from protein-DNA complexes computed after setting all partial charges on the DNA to zero except for the O1P and O2P atoms, which were each assigned a charge of $-0.5e$. The figure includes (a) MATa1-MATα2 (PDB id: 1AKH), (b) UBX-EXD (PDB id: 1B8I), (c) OCT1-PORE (PDB id: 1HF0), (d) Tc3 transposase (PDB id: 1U78), (e) phage 434 repressor (PDB id: 2OR1), and (f) MogR repressor (PDB id: 3FDQ). The electrostatic potential was computed using the ACG nonlinear PBE solver[57] and the parameters described in the Methods section.

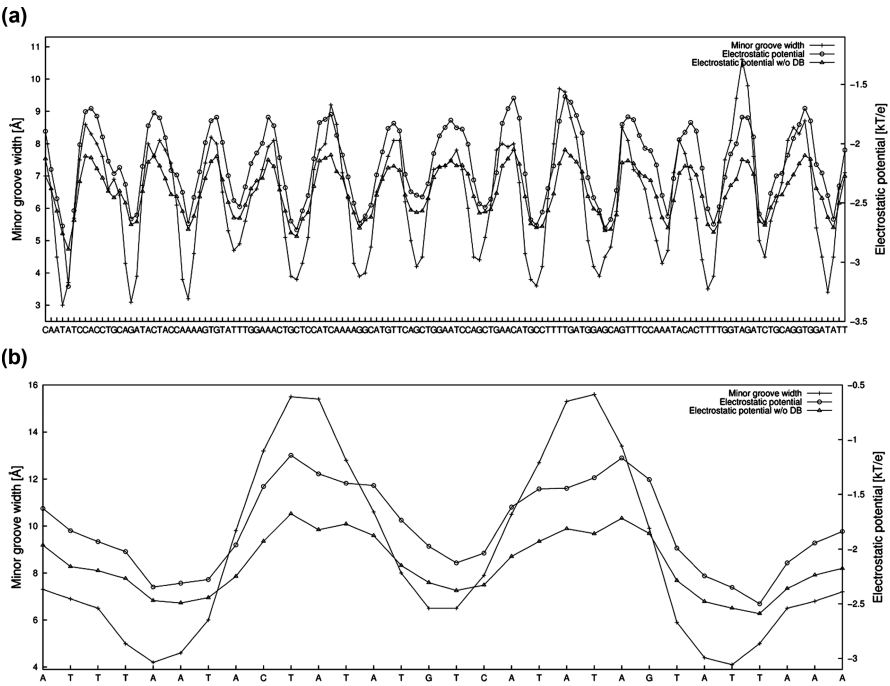

Figure 3.8 Electrostatic potential in the minor groove and minor groove width plotted as a function of sequence for the DNA conformation (a) in the nucleosome core particle (PDB id: 1KX5) and (b) the IHF-DNA complex (PDB id: 2NP2). The electrostatic potential was calculated after setting all partial charges on the DNA to zero except for the O1P and O2P atoms, which were each assigned a charge of $-0.5e$. As for the symmetry of the data for the IHF-DNA complex data see comment in the caption of Figure 3.6 (b). The electrostatic potential was computed using the ACG nonlinear PBE solver[57] and the parameters described in the Methods section.

plotted as a function of base sequence without the dielectric boundary between solute and solvent present. The potentials from the calculations with removed dielectric boundary contained fluctuations of only about 0.5 kT/e compared to about 1.5 kT/e when the dielectric boundary is considered. Again, the actual numbers for electrostatic potential based on DelPhi calculations are different from ACG results, and report a difference of about 8 kT/e between troughs and peaks when the dielectric boundary is present *vs.* only about 2 kT/e when the dielectric boundary is removed. This result demonstrates that the focusing of electrostatic field lines due to the shape of the dielectric boundary plays a major role in nucleosome formation.[17,107]

3.6 Summary and Outlook

This chapter revisits structural and biophysical aspects of protein-DNA recognition. While the presented principles are not novel findings, the effects

have been studied here at higher resolution and with an independent method, which allows a detailed analysis of the shape and electrostatic complementarity in protein-DNA binding.

The approaches presented here are of course tools that are valid for the analysis of much more complex systems. One such complex system is when several transcription factors bind to adjacent sites and cooperative effects influence binding specificity. A prominent example, for which structural information is available, is the enhanceosome where several transcription factors interact with each other while binding to adjacent sites.[112,113] Co-factors that bind DNA, as for instance Extradenticle in the discussed case of Hox proteins,[18] or that do not form contacts with DNA but still affect protein-DNA binding specificity, as observed for CBF-β that enhances the binding of the Runt domain to its DNA target,[67] also contribute through protein-protein interactions to protein-DNA readout. Yet another form of cooperativity is the formation of dimers, as in the case of nuclear receptors such as the well-studied glucocorticoid receptor,[114,115] or tetramers, as discussed for p53.[29,30,44] These dimers and tetramers form biologically relevant assemblies necessary for recognizing DNA *in vivo* sites. In addition, DNA can bind proteins either as a double-helix or in its single-stranded form.[1] In analogy to a recent comparison of protein binding to RNA and DNA,[116] electrostatics is expected to play a much different role in binding of single-stranded *vs*. double-helical DNA to a protein. Double-stranded DNA is characterized by a much higher charge density than single-stranded DNA, while in the latter case bases are much better accessible for stacking interactions with hydrophobic residues.[117] Finally, the interaction of proteins with DNA might interfere or be assisted by the packaging of DNA in nucleosomes.[14,118,119] The structural information on such complexes that combine various levels of interactions is still sparse, but as all these examples demonstrate, there is still much to be learned that goes beyond the scope of this book chapter.

In addition, the genomic era presents a challenge for structural studies of protein-DNA interactions. The amount of available sequence information is ever growing. Whole genomes have been sequenced for several species and transcription factor-DNA binding sequence preferences have been studied with high-throughput methods, such as protein-binding microarrays[120] and the bacterial-one hybrid system.[121] Protein-DNA binding data has been derived at high-resolution from ChIP-seq[122–124] and SELEX-seq[125] experiments, and nucleosome-occupied sequences have been determined based on micrococcal nuclease digestion studies for various organisms.[126–131] The challenge is to connect such sequence information with high-resolution structural data and deduct the general principles for protein-DNA recognition. This has been partially achieved by providing high-throughput tools for the analysis of DNA sequence information for shape.[125,132–134] Forming the connection between high-throughput sequence information on one side and DNA shape and electrostatic potential features on the other side will be an important contribution to be made in the future, which is likely to have a large impact on biological and biomedical research.

3.7 Methods

All structures were either downloaded from the Protein Data Bank or created using the 3DNA package.[135] Unless otherwise noted, charges and radii were taken from the Amber94 force field[136] and assigned using the pdb2pqr server.[137] The minor groove widths were calculated as inter-phosphate distances using the finddist and analyze programs of the 3DNA package, and 5.8 Å was subtracted from the P-P distance to account for the van der Waals radii of the phosphate groups. Midpoints in the minor grooves were calculated as described previously.[17] To avoid reporting anomalously large potentials for midpoints that fall interior to the solvent-excluded molecular surface in regions of deformed DNA, 0.5 Å spheres centered at each midpoint were constructed. The reported potentials are the average over the portions of these spheres exterior to the molecular surface, in addition to the original midpoints if they were also exterior. Electrostatic potentials were computed from the NLPBE with the ACG program developed by Boschitsch and Fenley.[57] The isopotential meshes were generated with the NLPBE using the DelPhi program as described in[17] and visualized using Grasp2.[75] Unless otherwise noted, all electrostatic potential calculations were performed at 0.145 M physiologic salt concentration, with a solvent-excluded molecular surface defined based on a 1.4 Å probe, an interior dielectric constant of 2.0, an exterior dielectric constant of 80.0, a temperature of 298.15 K, a minimum grid spacing of 0.30 Å, a boundary conforming mesh, and a grid that was set to three-times the largest dimension of the molecule. For those calculations that were described as being performed without a dielectric boundary, the interior dielectric constant was 79.9, and the exterior dielectric constant was 80.[138] All other Poisson-Boltzmann parameters remained the same.

Acknowledgments

This work was supported by NIH 5R44GM073391-03 (M.O.F.) and USC start-up funds (R.R.).

References

1. R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig and R. S. Mann, *Annu. Rev. Biochem.*, 2010, **79**, 233–269.
2. R. Rohs, S. M. West, P. Liu and B. Honig, *Curr. Opin. Struct. Biol.*, 2009, **19**, 171–177.
3. A. K. Aggarwal, D. W. Rodgers, M. Drottar, M. Ptashne and S. C. Harrison, *Science*, 1988, **242**, 899–907.
4. Z. Otwinowski, R. W. Schevitz, R. G. Zhang, C. L. Lawson, A. Joachimiak, R. Q. Marmorstein, B. F. Luisi and P. B. Sigler, *Nature*, 1988, **335**, 321–329.
5. S. C. Harrison and A. K. Aggarwal, *Annu. Rev. Biochem.*, 1990, **59**, 933–969.

6. T. A. Steitz, *Q. Rev. Biophys.*, 1990, **23**, 205–280.
7. C. W. Garvie and C. Wolberger, *Mol. Cell*, 2001, **8**, 937–946.
8. C. L. Lawson and H. M. Berman, in *Protein-Nucleic Acid Interactions: Structural Biology*, eds. P. A. Rice and C. C. Correll, Royal Society of Chemistry, 2008.
9. M. Hong and R. Marmorstein, in *Protein-Nucleic Acid Interactions: Structural Biology*, eds. P. A. Rice and C. C. Correll, Royal Society of Chemistry, 2008.
10. N. M. Luscombe, S. E. Austin, H. M. Berman and J. M. Thornton, *Genome Biol.*, 2000, **1**, REVIEWS001.
11. V. B. Zhurkin, M. Y. Tolstorukov, F. Xu, A. V. Colasanti and W. K. Olson, in *DNA conformation and transcription*, ed. T. Ohyama, Georgetown, Tex.: Landes Bioscience; New York, NY.: Springer Science Business Media, 2005.
12. D. M. Crothers and Z. Shakked, in *Oxford Handbook of Nucleic Acid Structures*, ed. S. Neidle, Oxford University Press, London, 1999, pp. 455–470.
13. T. E. Haran and U. Mohanty, *Q. Rev. Biophys.*, 2009, **42**, 41–81.
14. W. K. Olson and V. B. Zhurkin, *Curr. Opin. Struct. Biol.*, 2011, **21**, 348–357.
15. I. T. Weber and T. A. Steitz, *Proc. Natl. Acad. Sci. USA*, 1984, **81**, 3973–3977.
16. J. B. Matthew and D. H. Ohlendorf, *J. Biol. Chem.*, 1985, **260**, 5860–5862.
17. R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann and B. Honig, *Nature*, 2009, **461**, 1248–1253.
18. R. Joshi, J. M. Passner, R. Rohs, R. Jain, A. Sosinsky, M. A. Crickmore, V. Jacob, A. K. Aggarwal, B. Honig and R. S. Mann, *Cell*, 2007, **131**, 530–543.
19. K. Luger, A. W. Mader, R. K. Richmond, D. F. Sargent and T. J. Richmond, *Nature*, 1997, **389**, 251–260.
20. T. J. Richmond and C. A. Davey, *Nature*, 2003, **423**, 145–150.
21. G. S. Manning, *Q. Rev. Biophys.*, 1978, **11**, 179–246.
22. B. Honig and A. Nicholls, *Science*, 1995, **268**, 1144–1149.
23. I. Klapper, R. Hagstrom, R. Fine, K. Sharp and B. Honig, *Proteins*, 1986, **1**, 47–59.
24. K. A. Sharp, B. Honig and S. C. Harvey, *Biochemistry*, 1990, **29**, 340–346.
25. B. Honig and R. Rohs, *Nature*, 2011, **470**, 472–473.
26. G. A. Patikoglou, J. L. Kim, L. Sun, S. H. Yang, T. Kodadek and S. K. Burley, *Genes Dev.*, 1999, **13**, 3217–3230.
27. P. A. Rice, S. Yang, K. Mizuuchi and H. A. Nash, *Cell*, 1996, **87**, 1295–1306.
28. J. Aishima, R. K. Gitti, J. E. Noah, H. H. Gan, T. Schlick and C. Wolberger, *Nucleic Acids Res.*, 2002, **30**, 5244–5252.
29. Y. Chen, R. Dey and L. Chen, *Structure*, 2010, **18**, 246–256.
30. M. Kitayner, H. Rozenberg, R. Rohs, O. Suad, D. Rabinovich, B. Honig and Z. Shakked, *Nat. Struct. Mol. Biol.*, 2010, **17**, 423–429.

31. A. R. Srinivasan, R. R. Sauers, M. O. Fenley, A. H. Boschitsch, A. Matsumoto, A. V. Colasanti and W. K. Olson, *Biophys. Rev.*, 2009, **1**, 13–20.
32. A. H. Boschitsch and M. O. Fenley, *J. Comput. Chem.*, 2004, **25**, 935–955.
33. M. Montrel, V. P. Chuprina, V. I. Poltev, W. Nerdal and E. Sletten, *J. Biomol. Struct. Dyn.*, 1998, **16**, 631–637.
34. P. K. Weiner, R. Langridge, J. M. Blaney, R. Schaefer and P. A. Kollman, *Proc. Natl. Acad. Sci. USA*, 1982, **79**, 3754–3758.
35. A. Pullman and B. Pullman, *Q. Rev. Biophys.*, 1981, **14**, 289–380.
36. B. Jayaram, K. A. Sharp and B. Honig, *Biopolymers*, 1989, **28**, 975–993.
37. K. Chin, K. A. Sharp, B. Honig and A. M. Pyle, *Nat. Struct. Biol.*, 1999, **6**, 1055–1061.
38. N. V. Hud and J. Plavec, *Biopolymers*, 2003, **69**, 144–159.
39. M. L. Kopka, C. Yoon, D. Goodsell, P. Pjura and R. E. Dickerson, *Proc. Natl. Acad. Sci. USA.*, 1985, **82**, 1376–1380.
40. M. A. Young, B. Jayaram and D. L. Beveridge, *J. Am. Chem. Soc.*, 1997, **119**, 59–69.
41. D. Min, H. Li, G. Li, B. A. Berg, M. O. Fenley and W. Yang, *Chem. Phys. Lett.*, 2008, **454**, 391–395.
42. A. G. Tsai, A. E. Engelhart, M. M. Hatmal, S. I. Houston, N. V. Hud, I. S. Haworth and M. R. Lieber, *J. Biol. Chem.*, 2009, **284**, 7157–7164.
43. M. Suzuki, *Structure*, 1994, **2**, 317–326.
44. M. Kitayner, H. Rozenberg, N. Kessler, D. Rabinovich, L. Shaulov, T. E. Haran and Z. Shakked, *Mol. Cell*, 2006, **22**, 741–753.
45. D. T. Nair, R. E. Johnson, S. Prakash, L. Prakash and A. K. Aggarwal, *Nature*, 2004, **430**, 377–380.
46. D. T. Nair, R. E. Johnson, L. Prakash, S. Prakash and A. K. Aggarwal, *Structure*, 2005, **13**, 1569–1577.
47. R. E. Johnson, L. Prakash and S. Prakash, *Proc. Natl. Acad. Sci. USA*, 2005, **102**, 10466–10471.
48. E. N. Nikolova, E. Kim, A. A. Wise, P. J. O'Brien, I. Andricioaei and H. M. Al-Hashimi, *Nature*, 2011.
49. J. W. Locasale, A. A. Napoli, S. Chen, H. M. Berman and C. L. Lawson, *J. Mol. Biol.*, 2009, **386**, 1054–1065.
50. M. Kitayner, H. Rozenberg, D. Rabinovich and Z. Shakked, *Acta Crystallogr. D Biol. Crystallogr.*, 2005, **61**, 236–246.
51. Z. Shakked, D. Rabinovich, O. Kennard, W. B. Cruse, S. A. Salisbury and M. A. Viswamitra, *J. Mol. Biol.*, 1983, **166**, 183–201.
52. M. Eisenstein and Z. Shakked, *J. Mol. Biol.*, 1995, **248**, 662–678.
53. H. Robinson, Y. G. Gao, R. Sanishvili, A. Joachimiak and A. H.-J. Wang, *Nucleic Acids Res.*, 2000, **28**, 1760–1766.
54. Y. G. Gao, J. H. van Boom and A. H. Wang, *Biophys. J.*, 1995, **69**, 559–568.
55. N. V. Hud and M. Polak, *Curr. Opin. Struct. Biol.*, 2001, **11**, 293–301.
56. G. Lamm, L. Wong and G. R. Pack, *Biopolymers*, 1994, **34**, 227–237.

57. A. H. Boschitsch and M. O. Fenley, *J. Chem. Theory Comput.*, 2011, **7**, 1524–2540.
58. A. H. Wang, G. J. Quigley, F. J. Kolpak, J. L. Crawford, J. H. van Boom, G. van der Marel and A. Rich, *Nature*, 1979, **282**, 680–686.
59. T. Schwartz, K. Shafer, K. Lowenhaupt, E. Hanlon, A. Herbert and A. Rich, *Acta Crystallogr. D Biol. Crystallogr.*, 1999, **55**, 1362–1364.
60. B. Konig, J. J. Muller, E. Lanka and U. Heinemann, *Nucleic Acids Res.*, 2009, **37**, 1915–1924.
61. R. S. Hegde, S. R. Grossman, L. A. Laimins and P. B. Sigler, *Nature*, 1992, **359**, 505–512.
62. R. S. Hegde, *Annu. Rev. Biophys. Biomol. Struct.*, 2002, **31**, 343–360.
63. H. Rozenberg, D. Rabinovich, F. Frolow, R. S. Hegde and Z. Shakked, *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 15194–15199.
64. J. Hizver, H. Rozenberg, F. Frolow, D. Rabinovich and Z. Shakked, *Proc. Natl. Acad. Sci. USA*, 2001, **98**, 8490–8495.
65. R. Rohs, H. Sklenar and Z. Shakked, *Structure*, 2005, **13**, 1499–1509.
66. M. H. Werner, A. M. Gronenborn and G. M. Clore, *Science*, 1996, **271**, 778–784.
67. T. H. Tahirov, T. Inoue-Bungo, H. Morii, A. Fujikawa, M. Sasaki, K. Kimura, M. Shiina, K. Sato, T. Kumasaka, M. Yamamoto, S. Ishii and K. Ogata, *Cell*, 2001, **104**, 755–767.
68. D. Bartfeld, L. Shimon, G. C. Couture, D. Rabinovich, F. Frolow, D. Levanon, Y. Groner and Z. Shakked, *Structure*, 2002, **10**, 1395–1407.
69. R. Fairman, R. K. Beran-Steed, S. J. Anthony-Cahill, J. D. Lear, W. F. Stafford II, W. F. DeGrado, P. A. Benfield and S. L. Brenner, *Proc. Natl. Acad. Sci. USA*, 1993, **90**, 10429–10433.
70. S. Jones, P. van Heyningen, H. M. Berman and J. M. Thornton, *J. Mol. Biol.*, 1999, **287**, 877–896.
71. E. W. Stawiski, L. M. Gregoret and Y. Mandel-Gutfreund, *J. Mol. Biol.*, 2003, **326**, 1065–1079.
72. Y. Tsuchiya, K. Kinoshita and H. Nakamura, *Proteins*, 2004, **55**, 885–894.
73. J. Warwicker, D. Ollis, F. M. Richards and T. A. Steitz, *J. Mol. Biol.*, 1985, **186**, 645–649.
74. P. N. S. Yadav, M. J. Modak and J. S. Yadav, *J. Mol. Recognit.*, 1994, **7**, 207–209.
75. D. Petrey and B. Honig, *Methods Enzymol.*, 2003, **374**, 492–509.
76. A. Tomovic and E. J. Oakley, *PLoS ONE*, 2008, **3**.
77. G. Nimrod, A. Szilagyi, C. Leslie and N. Ben-Tal, *J. Mol. Biol.*, 2009, **387**, 1040–1053.
78. M. M. Gromiha and K. Fukui, *J. Chem. Inf. Model*, 2011, **51**, 721–729.
79. T. Schwartz, M. A. Rould, K. Lowenhaupt, A. Herbert and A. Rich, *Science*, 1999, **284**, 1841–1845.
80. D. H. Ohlendorf and J. B. Matthew, *Adv. Biophys.*, 1985, **20**, 137–151.

81. V. K. Misra, J. L. Hecht, K. A. Sharp, R. A. Friedman and B. Honig, *J. Mol. Biol.*, 1994, **238**, 164–280.
82. V. K. Misra, K. A. Sharp, R. A. Friedman and B. Honig, *J. Mol. Biol.*, 1994, **238**, 245–263.
83. M. Kato, S. Braun-Sand and A. Warshel, in *Computational and Structural Approaches to Drug Discovery*, eds. R. M. Stroud and J. Finer-Moore, J. Biomolecular Sciences, Royal Society of Chemistry, London, UK, 2008, pp. 268–291.
84. M. K. Gilson and H. X. Zhou, *Annu. Rev. Biophys. Biomol. Struct.*, 2007, **36**, 21–42.
85. J. A. McCammon, *Curr. Opin. Struct. Biol.*, 1998, **8**, 245–249.
86. M. T. J. Record, T. M. Lohman and P. DeHaseth, *J. Mol. Biol.*, 1976, **107**, 145–158.
87. P. L. Privalov, A. L. Dragan and C. Crane-Robinson, *Nucleic Acids Res.*, 2011, **39**, 2483–2491.
88. D. P. Mascotti and T. M. Lohman, *Biochemistry*, 1997, **36**, 7272–7279.
89. J. B. Chaires, *Biopolymers*, 1998, 201–215.
90. R. C. Harris, J. H. Bredenberg, A. R. Silalahi, A. H. Boschitsch and M. O. Fenley, *Biophys. Chem.*, 2011, **156**, 79–87.
91. M. O. Fenley, R. C. Harris, B. Jayaram and A. H. Boschitsch, *Biophys. J.*, 2010, **99**, 879–886.
92. M. T. J. Record, C. F. Anderson and T. M. Lohman, *Q. Rev. Biophys.*, 1978, **11**, 103–178.
93. N. S. Dosanjh, A. L. West and S. L. J. Michel, *Biochemistry*, 2009, **48**, 527–536.
94. A. G. Cherstvy, *J. Phys. Chem. B*, 2009, **113**, 4242–4247.
95. J. H. Bredenberg, C. Russo and M. O. Fenley, *Biophys. J.*, 2008, **94**, 4634–4645.
96. J. H. Bredenberg, A. H. Boschitsch and M. O. Fenley, *Commun. Comput. Phys.*, 2008, **3**, 1051–1070.
97. R. C. Harris, J. H. Bredenberg, A. R. Silalahi, A. H. Boschitsch and M. O. Fenley, *Biophys. Chem.*, 2011, **156**, 79–87.
98. J. H. Bredenberg and M. O. Fenley, *Commun. Comput. Phys.*, 2008, **3**, 1132–1153.
99. K. A. Sharp and B. Honig, *J. Phys. Chem.*, 1990, **94**, 7684–7692.
100. A. H. Boschitsch and M. O. Fenley, *J. Comput. Chem.*, 2007, **28**, 909–921.
101. B. Jayaram, K. McConnell, S. B. Dixit, A. Das and D. L. Beveridge, *J. Comput. Chem.*, 2002, **23**, 1–14.
102. F. Fogolari, A. H. Elcock, G. Esposito, P. Viglino, J. M. Briggs and J. A. McCammon, *J. Mol. Biol.*, 1997, **267**, 368–381.
103. Y. Levy, J. N. Onuchic and P. G. Wolynes, *J. Am. Chem. Soc.*, 2007, **129**, 738–739.
104. T. Selzer, S. Albeck and G. Schreiber, *Nat. Struct. Biol.*, 2000, **7**, 537–541.
105. L. F. Pineda De Castro and M. Zacharias, *J. Mol. Recognit.*, 2002, **15**, 209–220.
106. G. Schreiber and H.-X. Zhou, *Chem. Rev.*, 2009, **109**, 839–860.

107. S. M. West, R. Rohs, R. S. Mann and B. Honig, *J. Biomol. Struct. Dyn.*, 2010, **27**, 861–866.
108. W. Rocchia, E. Alexov and B. Honig, *J. Phys. Chem. B*, 2001, **105**, 6507–6514.
109. M. K. Gilson, K. Sharp and B. Honig, *J. Comput. Chem.*, 1987, **9**, 327–335.
110. A. Nicholls and B. Honig, *J. Comput. Chem.*, 1991, **12**, 435–445.
111. J. L. Hecht, B. Honig, Y.-K. Shin and W. L. Hubbell, *J. Phys. Chem.*, 1995, **99**, 7782–7786.
112. C. R. Escalante, E. Nistal-Villan, L. Shen, A. Garcia-Sastre and A. K. Aggarwal, *Mol. Cell*, 2007, **26**, 703–716.
113. D. Panne, T. Maniatis and S. C. Harrison, *Cell*, 2007, **129**, 1111–1123.
114. B. F. Luisi, W. X. Xu, Z. Otwinowski, L. P. Freedman, K. R. Yamamoto and P. B. Sigler, *Nature*, 1991, **352**, 497–505.
115. S. H. Meijsing, M. A. Pufall, A. Y. So, D. L. Bates, L. Chen and K. R. Yamamoto, *Science*, 2009, **324**, 407–410.
116. S. Shazman, G. Elber and Y. Mandel-Gutfreund, *Nucleic Acids Res.*, 2011, **39**, 7390–7399.
117. K. E. Max, M. Zeeb, R. Bienert, J. Balbach and U. Heinemann, *J. Mol. Biol.*, 2006, **360**, 702–714.
118. S. Glatt, C. Alfieri and C. W. Muller, *Curr. Opin. Struct. Biol.*, 2011, **21**, 335–341.
119. R. D. Makde, J. R. England, H. P. Yennawar and S. Tan, *Nature*, 2010, **467**, 562–566.
120. M. F. Berger, G. Badis, A. R. Gehrke, S. Talukder, A. A. Philippakis, L. Pena-Castillo, T. M. Alleyne, S. Mnaimneh, O. B. Botvinnik, E. T. Chan, F. Khalid, W. Zhang, D. Newburger, S. A. Jaeger, Q. D. Morris, M. L. Bulyk and T. R. Hughes, *Cell*, 2008, **133**, 1266–1276.
121. M. B. Noyes, R. G. Christensen, A. Wakabayashi, G. D. Stormo, M. H. Brodsky and S. A. Wolfe, *Cell*, 2008, **133**, 1277–1289.
122. A. Valouev, D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers and A. Sidow, *Nat. Methods*, 2008, **5**, 829–834.
123. A. Ozdemir, K. I. Fisher-Aylor, S. Pepke, M. Samanta, L. Dunipace, K. McCue, L. Zeng, N. Ogawa, B. Wold and A. Stathopoulos, *Genome Res.*, 2011, **21**, 566–577.
124. D. S. Johnson, A. Mortazavi, R. M. Myers and B. Wold, *Science*, 2007, **316**, 1497–1502.
125. M. Slattery, T. Riley, P. Liu, N. Abe, P. Gomez-Alcala, I. Dror, T. Zhou, R. Rohs, B. Honig, H. J. Bussemaker and R. S. Mann, *Cell*, 2011, **147**, 1270–1282.
126. E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I. K. Moore, J. P. Wang and J. Widom, *Nature*, 2006, **442**, 772–778.
127. N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. LeProust, T. R. Hughes, J. D. Lieb, J. Widom and E. Segal, *Nature*, 2009, **458**, 362–366.
128. X. Fan, Z. Moqtaderi, Y. Jin, Y. Zhang, X. S. Liu and K. Struhl, *Proc. Natl. Acad. Sci. USA*, 2010, **107**, 17945–17950.

129. N. Ponts, E. Y. Harris, J. Prudhomme, I. Wick, C. Eckhardt-Ludka, G. R. Hicks, G. Hardiman, S. Lonardi and K. G. Le Roch, *Genome Res.*, 2010, **20**, 228–238.
130. Y. Field, N. Kaplan, Y. Fondufe-Mittendorf, I. K. Moore, E. Sharon, Y. Lubling, J. Widom and E. Segal, *PLoS Comput. Biol.*, 2008, **4**, e1000216.
131. T. N. Mavrich, C. Jiang, I. P. Ioshikhes, X. Li, B. J. Venters, S. J. Zanton, L. P. Tomsho, J. Qi, R. L. Glaser, S. C. Schuster, D. S. Gilmour, I. Albert and B. F. Pugh, *Nature*, 2008, **453**, 358–362.
132. S. C. Parker, L. Hansen, H. O. Abaan, T. D. Tullius and E. H. Margulies, *Science*, 2009, **324**, 389–392.
133. S. C. Parker and T. D. Tullius, *Curr. Opin. Struct. Biol.*, 2011, **21**, 342–347.
134. E. P. Bishop, R. Rohs, S. C. J. Parker, S. M. West, P. Liu, R. S. Mann, B. Honig and T. D. Tullius, *ACS Chem. Biol.*, 2011, **6**, 1314–1320.
135. X.-J. Lu and W. K. Olson, *Nucleic Acids Res.*, 2003, **31**, 5108–5121.
136. W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, *J. Am. Chem. Soc.*, 1995, **117**, 5179–5197.
137. T. J. Dolinsky, P. Czodrowski, H. Li, J. E. Nielsen, J. H. Jensen, G. Klebe and N. A. Baker, *Nucleic Acids Res.*, 2007, **35**, W522–W525.
138. M. O. Fenley, C. Russo and G. S. Manning, *J. Phys. Chem. B*, 2011, **115**, 9864–9872.