Published in final edited form as:

J Chem Theory Comput. 2017 April 11; 13(4): 1794–1811. doi:10.1021/acs.jctc.6b00883.

Direct comparison of amino acid and salt interactions with double-stranded and single-stranded DNA from explicit-solvent molecular dynamics simulations

Casey T. Andrews[†], Brady A. Campbell[†], and Adrian H. Elcock

Department of Biochemistry, University of Iowa, Iowa City, IA 52242

Abstract

Given the ubiquitous nature of protein-DNA interactions, it is important to understand the interaction thermodynamics of individual amino acid sidechains for DNA. One way to assess these preferences is to perform molecular dynamics (MD) simulations. Here we report MD simulations of twenty amino acid sidechain analogs interacting simultaneously with both a 70-base pair double-stranded DNA and with a 70-nucleotide single-stranded DNA. The relative preferences of the amino acid sidechains for dsDNA and ssDNA match well with values deduced from crystallographic analyses of protein-DNA complexes. The estimated apparent free energies of interaction for ssDNA, on the other hand, correlate well with previous simulation values reported for interactions with isolated nucleobases, and with experimental values reported for interactions with guanosine. Comparisons of the interactions with dsDNA and ssDNA indicate that, with the exception of the positively charged sidechains, all types of amino acid sidechain interact more favorably with ssDNA, with intercalation of aromatic and aliphatic sidechains being especially notable. Analysis of the data on a base-by-base basis indicates that positively charged sidechains, as well as sodium ions, preferentially bind to cytosine in ssDNA, and that negatively charged sidechains, and chloride ions, preferentially bind to guanine in ssDNA. These latter observations provide a novel explanation for the lower salt-dependence of DNA duplex stability in GC-rich sequences relative to AT-rich sequences.

Introduction

Direct physical interactions of amino acid sidechains with the bases, sugars and phosphate groups of DNA are one of the principal determinants of the thermodynamics of protein-DNA associations. Following the first identification of key protein-DNA interactions by Seeman *et al.*,¹ comprehensive statistical analyses of contacts in protein-DNA complexes

Correspondence to: Adrian H. Elcock.

[†]these authors contributed equally

Supporting Information

Apparent interaction free energies versus distance from four independent replicate simulations for dsDNA and ssDNA; apparent interaction free energies versus distance for each of the six groups in dsDNA and ssDNA; preservation of stacking interactions in ssDNA; comparison of apparent free energies for interactions with dsDNA with the data reported by de Ruiter and Zagrovic;²¹ comparison of the G_{int} values for amino acid sidechains with ssDNA with the “polar requirement” scale values of Woese; comparison of the G_{int} values for amino acid sidechains with ssDNA with their hydration free energies; comparison of previously reported experimental G_{int} values with guanosine, with the number of heavy atoms in the sidechain; convergence of the interaction preferences with the simulation period. This material is available free of charge via the Internet at <http://pubs.acs.org>.

have been carried out by a number of groups,^{2–5} with more specialized studies of the prevalence of cation- π ⁶ or aromatic interactions also having been reported.^{7, 8} In a number of cases, statistical studies have been used to derive scoring functions for specific transcription factors,⁹ or more general amino acid-nucleic acid interactions,^{10–12} for use in rapidly predicting preferred genomic binding sites for proteins. Others have sought to develop or use more physically motivated energy functions to directly calculate^{13–17} or analyze¹⁸ protein-DNA binding affinities.

Complementing these studies, a number of groups have used free energy simulation methods in an attempt to measure directly the interaction thermodynamics of amino acids with nucleic acid bases. The Sarai group performed early free energy calculations of the interaction of Asn with AT and GC base-pairs using a distance-dependent dielectric model of the solvent;¹⁹ in subsequent work *ab initio* QM calculations were used, although again without explicit modeling of solvation effects.²⁰ Much more recently, the Zagrovic group has reported the first truly comprehensive computational study of the free energies of association of all amino acid sidechains with all of the nucleobases in explicit solvent.²¹ Finally, the Papoian group has used umbrella sampling techniques with a creative helical coordinate system to obtain 3D free energy profiles for Na⁺ and methylguanidinium ions interacting with double-stranded DNA.²²

What we have yet to find in the literature, however, is any study that: (a) uses explicit-solvent MD to comprehensively examine the affinities of all amino acids for the nucleic acid bases in the context of double-stranded DNA, or that (b) directly compares the relative affinities of amino acids for double- and single-stranded DNAs. To address both issues, we have performed explicit-solvent MD simulations of an infinitely replicated 70 base-pair DNA, in both its double-stranded (dsDNA) and single-stranded (ssDNA) states, in an aqueous NaCl solution that also contains 10 copies of 20 amino acid sidechain analogs (i.e. 200 non-salt solute molecules). During “brute force” MD simulations of these systems, the sidechain analogs repeatedly associate with and dissociate from the DNA, thereby providing a “one pot” approach to measuring and comparing the preferences of all types of amino acid sidechains for sites on dsDNA and ssDNA. The simulations show significant differences between the preferences of amino acids (and salt ions) for dsDNA and ssDNA, produce results that are in perhaps surprisingly good qualitative agreement with data obtained from statistical analyses of protein-DNA complexes, and provide what we think is a novel explanation for the lower salt dependences of DNA duplex stability in GC-rich sequences.

Computational Methods

Initial structures of both nucleic acid structures simulated here were generated using the Stroud group's Make-NA server (<http://structure.usc.edu/make-na/server.html>) and formatted to be recognizable by the MD simulation program GROMACS version 4.6.1.^{23, 24} We selected the following sequence containing 70 base-pairs:
TGACGTAATTCATCGAACTTTGCGCTATAC
AAAGGCACCAGTTAGCCCGGGTCTCCTGGAGATGTGACGT. This sequence contains all possible triplet sequences with only four repeating instances (ACG, GAC, TGA, and CGT). Our initial hope in selecting such a sequence was that it might allow us to observe

differences in the affinity of amino acid sidechains for DNA bases depending on the latter's neighboring bases; ultimately, however, we concluded that sampling was insufficient to enable us to draw meaningful conclusions at such a fine level of detail.

We simulated both dsDNA and ssDNA in periodic form, with appropriate bonds, angles and dihedrals replicated in order to build effectively infinitely long DNAs. The dsDNA and ssDNA systems were each simulated in a $238 \times 70 \times 70$ Å box to which periodic boundary conditions were applied in each direction. The decision to perform the simulations in such a way that the DNA is effectively of infinite length was made in order to eliminate any possible end-effects, whereby excessive association of amino acid sidechains (especially aromatic sidechains) might occur at the terminal bases. But while the imposition of such artificial periodicity simplifies the analysis of the interactions (since all of the DNA bases can be considered to be 'internal' rather than 'terminal'), it was pointed out by the reviewers of this paper that it also constrains the DNA in ways that may not be realistic. For the dsDNA, the imposition of periodicity – coupled with the use of constant-volume (NVT) simulation conditions – places limits on the extent to which helical parameters such as twist, and rise can deviate from their initial (standard B-DNA) values. For the ssDNA, on the other hand, the imposition of periodicity and the use of NVT conditions both act to prevent it from forming the very wide range of conformations that are to be expected given its known flexibility.²⁵ In principle, as noted by a reviewer, it might be possible to impose anisotropic pressure coupling in the simulations and thereby allow the ssDNA to become more compact while remaining effectively infinite (and thereby remaining free of end-effects). In practice, however, this would likely require much longer simulation times for proper sampling. In the absence of using such pressure coupling, therefore, it seems reasonable to suggest that the conformations sampled during the simulations, especially those of the ssDNA, are likely to correspond more to those of an idealized, extended state than to those likely to be adopted by a completely unrestrained DNA. It is therefore also quite possible to imagine, as indicated by the reviewers, that the conformational restrictions placed on the DNAs might have noticeable (entropic) consequences for the binding thermodynamics of the amino acid sidechains.

A total of twenty different types of solutes (plus salt) were added to the dsDNA and ssDNA systems. These included sidechain-only models of all common amino acids (except glycine and proline), with histidine also modeled in its fully protonated state, and with N-methylacetamide, included as the twentieth type of solute, acting as a mimic of the protein backbone. We note that we used sidechain-only models instead of capped amino acids because, in exploratory simulations using the latter models, we found that interactions with the DNA were often unduly influenced by the capped termini. Partial charges for the sidechain models were derived from those of the complete amino acids using a similar approach to that outlined by others:^{26, 27} briefly, the C α atom was replaced by a hydrogen atom, the backbone atoms were deleted, and the sum of their partial charges was distributed evenly over the three hydrogen atoms now attached to the C β atom. Ten copies of each type of solute were added at random (non-clashing) locations in the simulation box, giving a total solute concentration of around 30 mg/mL. We also note that, in previous simulation studies that have explored small-molecule interactions with proteins, problems with phase separation have occasionally been encountered at high solute concentrations;^{28–30} that kind

of behavior has been shown, however, to be parameter dependent³¹ and was not observed here.

The nucleic acids were modeled using the AMBER parm99³² force field supplemented with the bsc0 parameters³³ that improve modeling of α and γ torsions, and with improved parameters for the glycosidic torsions for DNA (χ_{OL4}).³⁴ This combination of parameter sets has been shown by the Otyepka and Šponer groups to perform quite well in applications to a variety of challenging systems.^{35, 36} Amino acid sidechains were modeled using the AMBER ff99SB-ILDN force field.^{37, 38} Water was modeled explicitly using the TIP4P-EW model.³⁹ As in our recent work,⁴⁰ this water model was selected in preference to the more commonly used TIP3P model⁴¹ since one of our future goals is to model protein-nucleic acid interactions: when used together with AMBER protein parameters, TIP4P-Ew has been shown to perform well at describing the conformational behavior of small peptides.⁴² Na^+ and Cl^- ions – which were added to 150 mM concentrations in order to crudely mimic physiological concentrations – were modeled using the parameters derived by Joung & Cheatham;⁴³ additional Na^+ ions were added to the ssDNA and dsDNA systems, respectively, to ensure overall system electroneutrality. Previous work by the Cheatham group has indicated that – at least for the r(GACC) system – there is little difference in conformational behavior between simulations that include only enough ions to ensure electroneutrality and those that incorporate additional salt.⁴⁴ We note that improved parameters for DNA-ion interactions in TIP3P water have been derived by the Aksimentiev group⁴⁵ and that, in work reported as the present work was nearing completion, improved parameters have also been reported for amine-phosphate interactions by the same group.⁴⁶ While these new parameters have been shown to dramatically improve the description of dsDNA-dsDNA interactions – and so would be attractive candidates to use in future simulations of protein-DNA systems – they have not yet been optimized for use with the TIP4P-Ew water model that is of interest here (see above).

The dsDNA and ssDNA systems contained 149,818 and 151,098 atoms, respectively. All systems were first equilibrated for 1.35 ns, with the temperature being raised incrementally from 50 to 298 K over the course of the first 350 ps; following this equilibration period, all MD simulations were carried out for a production period of 500 ns. During MD, pressure and temperature were maintained at their equilibrium values using the Parrinello-Rahman⁴⁷ barostat and Nosé⁴⁸-Hoover⁴⁹ thermostat, respectively. All covalent bonds were constrained to their equilibrium lengths with LINCS,⁵⁰ allowing a 2.5 fs timestep to be used. Short-range van der Waals and electrostatic interactions were truncated at 10 Å and longer-range electrostatic interactions were computed using the smooth Particle Mesh Ewald method.⁵¹ During the production period of the simulations, all solute coordinates were saved at intervals of 0.1 ps to give a total of 5 million snapshots for each simulation for subsequent analysis. For both dsDNA and ssDNA, four independent MD simulations were performed, each differing in the initial (randomized) placement of the amino acid sidechains. With a 500 ns production time for each simulation, the combined production time of the simulations reported here is 4 μs .

Computation of G_{int} from MD simulations

As in our previous work,^{52–54} we compute apparent free energies of interaction, G_{int} , between solutes and the DNA from histograms of the minimum distance between any pair of heavy atoms on the solute and the DNA: we have shown previously that this approach produces much more easily interpreted G_{int} profiles than are obtained when interactions are expressed in terms of the distance between the centers of mass of the solutes.⁵³ Also as in our previous work,^{52–54} we obtain properly normalized G_{int} values by comparing the minimum-distance distributions observed in the MD simulations with corresponding minimum-distance distributions obtained when the same solutes are randomly repositioned within the same simulation cell. In the present work, these random placements were performed by resampling each of the 5 million MD snapshots, randomly rotating and translating each of the solute molecules – without altering their internal degrees of freedom – and recomputing the minimum distances. For each type of solute, therefore, we obtain two histograms of the minimum distance between the solute and the DNA: one histogram obtained directly from MD, and one histogram obtained by resampling with random placements. In order to ensure that the computed interaction free energies reach zero at long distances, we uniformly scale the MD-derived histogram so that it matches the randomly-resampled histogram at distances from 30 to 50 Å. For any given separation distance, then, we obtain the effective interaction free energy, G_{int} , using $G_{\text{int}} = -RT \ln (P_{\text{MD}} / P_{\text{random}})$ where P_{MD} is the scaled value of the MD-derived histogram, and P_{random} is the value of the randomly-resampled histogram.

The above type of analysis is relatively straightforward to conduct, and provides (apparent)

G_{int} values that allow the relative preferences of the amino acids for DNA to be measured and compared with experiment (see below). But, as pointed out by a reviewer, it should be remembered that the G_{int} values obtained from applying the above type of analysis to the present situation, in which many different types of solute are simultaneously competing for binding to the DNA, are not necessarily identical to true free energies of interaction that would be obtained from two-component simulations that consider only the binding of a single type of solute to DNA. A more rigorous analysis of the multicomponent systems modeled here might, in principle, be possible using Kirkwood-Buff theory⁵⁵ but, given that the present simulations contain 23 different components, this would be a formidable undertaking. It should be noted, therefore, that the G_{int} values reported throughout this manuscript represent only *apparent* free energies of interaction.

In an attempt to decompose each amino acid sidechain's total apparent free energy of interaction with the DNA into apparent free energies of interaction with each of the different chemical groups of the DNA we follow a similar procedure. We first partition the DNA atoms into six groups: the four bases (adenine, cytosine, guanine, and thymine), the deoxyribose sugars, and the phosphate groups. For each solute molecule in each snapshot we determine the minimum distance between the solute and each of these six groups. We then identify which of these six values is the shortest and increment only the histogram of the group with the shortest distance: the other histograms of the other five groups are not updated. Again, resampling is used to obtain corresponding randomized histograms for each

of the six groups and these are then used to normalize each MD-derived histogram between 30 and 50 Å.

It is to be noted that while the decision to increment only the histogram of the group with the shortest of the six minimum-distances has the disadvantage that it under-emphasizes cases where a single sidechain interacts simultaneously with multiple groups⁴ it has the very significant advantage of ensuring that indirect interactions are not scored as being unduly favorable. If, instead, one allows all six distances to contribute to their respective histograms then physically unreasonable results can be obtained. For example, intercalation of an aromatic amino acid sidechain between the middle two bases in the ssDNA sequence – AGTTAG – would produce a favorable G_{int} value at a separation distance of ~3.4 Å for the residue's interaction with thymine – which would be entirely reasonable – but also (spurious) favorable G_{int} values at separation distances of ~6.8 Å and ~10.2 Å for the residue's interactions with adenine and guanine. The latter two interactions are indirect and would only be computed as being energetically favorable because of the proximity of the adenines and guanines to the thymine that are responsible for the direct interaction; keeping only the shortest distance eliminates this artifact.

Computation of G_{int} from crystallographic data

The Wang group has very recently reported a comparative statistical analysis of amino acid frequencies in the interfaces of protein-ssDNA and protein-dsDNA complexes.⁵⁶ For both types of complex, these authors report amino acid frequencies for three different types of interface environments ('peak', 'flat', and 'valley'; see Figure 4 of ref. ⁵⁶) as well as the total frequencies of each type of interface environment (Figure 3 of ref. ⁵⁶). The data from these two figures can be combined to obtain the relative frequency with which each type of amino acid is found in the interface of protein-ssDNA and protein-dsDNA complexes. The data were extracted by digitizing both figures using WebPlotDigitizer (<http://arohatgi.info/WebPlotDigitizer/>). The effective free energy change representing each amino acid's preference for being found in the interface of a protein-ssDNA complex relative to that of a protein-dsDNA complex was then obtained using $G_{\text{int}} = -RT \ln (P_{\text{SSB}} / P_{\text{DSB}})$ where P_{SSB} and P_{DSB} are the frequencies with which the amino acid is found in the interface of protein-ssDNA and protein-dsDNA complexes, respectively.

Results

Figure 1 shows example configurations of the ssDNA and dsDNA systems at the beginning and end of the production period of the simulations: the amino acid sidechains are initially randomly distributed but during the course of the simulations they repeatedly associate with, and dissociate from, the DNA. The DNA itself is periodically replicated (such that its 'tail' is covalently connected to its 'head' with the appropriate bond, angle, and dihedral terms), but is otherwise free to move within the simulation box according to the influence of the AMBER force field. As noted in the Methods, the imposition of periodicity simplifies analysis by removing end-effects (e.g. clumping of aromatic amino acids that might occur at the terminal bases in the absence of periodicity), but means that our results are probably

more relevant to interactions with long DNAs than to interactions with DNA oligonucleotides.

Using the 5 million simulation snapshots sampled from each simulation, histograms were constructed recording the minimum distance between the heavy atoms of each sidechain and the heavy atoms of the DNA. These histograms were then converted into apparent free energies by comparing them with corresponding histograms obtained from entirely random placement of the same solute molecules within the same simulation box (see Computational Methods for a discussion of the limitations of this approach). The resulting apparent free energies of interaction, G_{int} , calculated for each type of amino acid sidechain with DNA, and averaged over the four independent replicate simulations, are plotted as a function of the separation distance from the nearest heavy atom of the DNA in Figure 2. For one indication of the likely sampling errors, G_{int} plots calculated separately for the four independent 500 ns replicate simulations are shown in Figures S1 and S2 for dsDNA and ssDNA respectively.

In the case of dsDNA (blue lines in Figure 2), the least and most favorable interactions occur, as might be expected, with the negatively charged and positively charged amino acid sidechains, respectively. Interestingly, while the G_{int} profiles for the negatively charged sidechains become repulsive only at comparatively short range (top lines in Figure 2A), the G_{int} profiles for the positively charged sidechains are favorable over a substantially longer range and are non-zero even at a separation distance of 10 Å (bottom lines in Figure 2A). The interactions of all other types of amino acid sidechains are non-zero only at short range. Polar sidechains (Figure 2B) are predicted to have net negative G_{int} values at distances <4 Å, with those of Gln and Asn being the most favorable. Aromatic sidechains (Figure 2C) are also predicted to have net negative G_{int} values at distances <4 Å, with two energetic minima observable for His, Trp and Tyr: the shorter-range minimum reflects hydrogen bonding interactions (which are absent in the G_{int} profile for Phe), while the longer-range minimum is characteristic of hydrophobic contacts between nonpolar groups. Finally, the aliphatic sidechains (Figure 2D) all produce qualitatively identical G_{int} profiles, giving a local minimum with a G_{int} value of ~0 kcal/mol, again at a distance characteristic of hydrophobic contacts.

For interactions with ssDNA (red lines in Figure 2), the behaviors of several of the amino acid sidechains change substantially. For the negatively charged amino acid sidechains (Figure 2A), the short-range repulsion that is apparent with dsDNA is abolished and the net G_{int} values of both D and E with ssDNA become slightly favorable. For each of the positively charged sidechains, the long-range attraction remains noticeable, but the very favorable short-range interaction is weakened. For all other amino acid sidechains, however, the G_{int} values become more favorable with ssDNA than with dsDNA. This is especially true for the aromatic sidechains (Figure 2C) and for the aliphatic sidechains (Figure 2D), for which the interactions with DNA now become net favorable. These changes are largely attributable to the intercalation of aromatic and aliphatic sidechains between bases adjacent in sequence in the single-stranded state: examples of both types of interaction, which occur more frequently between pyrimidine bases (see below), are shown in Figure 3.

The computed differences in the preferences of amino acid sidechains for dsDNA and ssDNA can be directly compared with the results of a very recent analysis of crystallographic data on protein-DNA complexes. The Wang group⁵⁶ has compared the amino acid compositions in the binding interfaces of complexes of dsDNA- and ssDNA-binding proteins; their reported compositions can be converted into effective free energies reflecting the preference of each amino acid for binding to ssDNA relative to dsDNA (see Computational Methods). Figure 4 compares these experimentally-derived free energies with the difference between the G_{int} values for ssDNA and dsDNA computed from the simulations. The agreement is surprisingly good (Pearson correlation coefficient, R_{corr} , of 0.85). In both data-sets, positively charged sidechains are disfavored in ssDNA (relative to dsDNA) while aliphatic, aromatic, and negatively charged sidechains are favored. The good correlation indicates that the MD simulations successfully mirror the differences in the relative affinities of the amino acid sidechains for dsDNA and ssDNA (see Discussion).

Analysis of the MD data can be taken a stage further to examine preferential interactions of the sidechains with each of the four different DNA bases, as well as the sugar and phosphate groups. Again, we obtain apparent free energies of interaction, G_{int} , by comparing the observed frequencies of interaction with corresponding frequencies of interaction obtained by random placement (see Computational Methods). Complete sets of the G_{int} profiles for all amino acid sidechains with dsDNA and ssDNA are shown in Figures S3 and S4 respectively. One straightforward way to compare relative affinities is to plot the minimum value of G_{int} found for each type of amino acid sidechain with each type of DNA base for interaction distances below 5 Å; these values are tabulated in Tables 1 and 2 for dsDNA and ssDNA, respectively, and in Table 3 as the difference between the two (ssDNA-dsDNA). Figure 5A plots these values for dsDNA, with the amino acids arranged from left to right in order of increasingly favorable G_{int} . As was the case when the preferences for dsDNA as a whole were plotted (Figure 2), the apparent interactions with the bases are most favorable for the positively charged amino acids (K, R, and positively-charged histidine, H_p) and least favorable for the negatively charged amino acids (D, E).

The relative preferences that each type of amino acid sidechain has for each of the four different bases can be compared with a variety of prior estimates from the literature. In particular, for each type of sidechain, we can compare the relative G_{int} values for A, C, G and T bases in dsDNA with the corresponding scores in each of the following studies: (a) the early statistical potential functions derived by Mandel-Gutfreund and Margalit,¹¹ (b), the much more recent statistical potential functions reported by Gao & Skolnick,⁵⁷ and (c) the physically-based HINT (Hydropathic INteraction) scores reported by Marabotti et al.¹⁷ Table 4 records the R_{corr} values obtained when the base-preferences of each type of amino acid sidechain are compared between all four data-sets. Shown at the bottom of each column is the average R_{corr} value obtained from each set of comparisons. From this it can be seen that the highest average R_{corr} values are obtained from comparisons of the two sets of statistical potential functions (column 'Mandel:Gao'; $R_{\text{corr}} = 0.45$) and from comparisons of the two sets of physically-based potential functions (column 'dsMD:HINT'; $R_{\text{corr}} = 0.46$). The average degree of correlation between the statistical potential results and the physical potential results, however, is very poor. In an attempt to determine the significance of the reasonably good correlation obtained between the two physically-based studies (i.e. this

work and that of Marabotti et al.), we also include in Table 4 a comparison of our ssDNA results with those reported by Marabotti et al.¹⁷ (column ‘ssMD:HINT’). As should be expected, the correlations in this case are far worse, with the average R_{corr} value changing from 0.46 (using the dsDNA simulation data) to -0.19 (using the ssDNA simulation data).

The penultimate right-hand column in Table 4 (column ‘Ave.’) records the average R_{corr} value obtained from each of the comparisons for each type of amino acid (excluding our ssDNA results); the right-most column re-expresses these values in rank order, with the amino acid producing the highest average degree of correlation between each of the four data-sets being awarded a rank of 1. From this it can be seen that the amino acid sidechain with the best rank is Arg, for which all four data-sets agree that its interaction is most favorable with guanine in dsDNA. Agreement between the four data-sets is also quite good for both Lys and Asp, for which all agree that the most favorable interactions are with guanine and cytosine, respectively. Beyond this, however, the degree of correspondence between the different data-sets is poor, even for amino acids such as Asn and Gln for which a bidentate hydrogen bonding interaction with adenine is commonly cited (see Discussion).

One final point to note regarding the interactions with the various groups of dsDNA concerns the relatively high affinity of sidechains for binding to the ribose sugar. Table 1 records the G_{int} value for each amino acid sidechain interacting not only with each of the four bases but also with the phosphate and sugar groups. When these six values for each type of amino acid sidechain are expressed in rank order – with the most favorable value being given a rank of 1 – it is found that the group exhibiting the most favorable interaction with aliphatic and aromatic sidechains is the deoxyribose sugar. For aliphatic sidechains (Ala, Ile, Leu, Val) the sugar group’s average rank is 1.3 ± 0.5 , with the methyl-containing thymine base being the second most highly ranked group with an average rank of 2.0 ± 0.8 . For aromatic sidechains (Phe, Trp and Tyr) – for which extensive interactions with the deoxyribose group in protein-DNA complex structures have been noted recently⁸ – the average rank is 1.7 ± 0.6 .

Figure 5B shows the minimum G_{int} for the amino acid sidechains with the four DNA bases in ssDNA. The amino acids are arranged in the same order used for the dsDNA results (Figure 5A), so a visual comparison of the two panels indicates that the preferences differ considerably. As noted earlier, in general, the G_{int} values with ssDNA are shifted to more favorable values, with the most consistently favorable interactions being those of the aromatic sidechains. A plot of the difference between the ssDNA and dsDNA G_{int} values is shown in Figure 5C, reinforcing the result noted earlier that the interactions of aromatic, aliphatic, and negatively charged sidechains with the DNA bases are all more favorable in ssDNA than in dsDNA, while the interactions of positive sidechains are all less favorable.

Figure 5C, however, also reveals two interesting preferences of the charged sidechains for interactions with the different bases in dsDNA and ssDNA. First, relative to dsDNA, the positively charged sidechains show a clear preference for interacting with cytosine in ssDNA over the other three bases (compare the green bars with the other bars for K, H_p, and R in Figure 5C). Second, again relative to dsDNA, the negatively charged sidechains show a preference for interacting with guanine in ssDNA over the other three bases (compare the

yellow bars with the other bars for D and E in Figure 5C). Both of these results can be understood in terms of the functional groups that become accessible when Watson-Crick base-pairing is abolished to form the single-stranded state. An un-base-paired cytosine, for example, exposes its N3 and O2 atoms, both of which bear substantial partial negative charges and therefore represent potential bidentate sites of interaction for positively charged amino acid sidechains (see example simulation snapshot in Figure 6A). An un-base-paired guanine, on the other hand, fully exposes the N1-H and the exocyclic NH₂ atoms at the C2 position, again allowing for a bidentate interaction with the carboxylic acid groups of D and E sidechains (Figure 6B). As will be seen later, similar interactions with these bases in the single-stranded state are apparent for Na⁺ and Cl⁻.

One additional feature of Figure 5C concerns the behavior of the uncharged amino acid sidechains. In general, the differences between the computed G_{int} values for ssDNA and dsDNA are most negative (i.e. favorable) for cytosine and thymine (compare green and red bars with the blue and yellow bars in Figure 5C): when the G_{int} values for the four bases are ranked separately for each type of uncharged amino acid sidechain, the average ranks are 1.8 ± 0.9 and 1.9 ± 1.1 for cytosine and thymine respectively, compared with 2.7 ± 0.8 and 3.6 ± 0.5 for adenine and guanine, respectively. This relatively greater affinity for cytosine and thymine in the single-stranded state probably reflects the greater ease of unstacking and exposing the pyrimidines (Py) relative to the purines (Pu). This is suggested by visual analysis of the structures obtained at the end of the four ssDNA simulations: of the 17 Py-Py steps that are present in the initial structure, only 51 ± 6 % are still stacked at the end of the simulation; in contrast, of the 17 Pu-Pu steps, 90 ± 6 % remain stacked (Figure S5). These trends broadly follow the trends that we have seen previously in simulations of dinucleoside monophosphates⁴⁰ and tetranucleotides⁵⁸ performed with the same iteration of the Amber force field.

Figure 5D compares our computed G_{int} values for ssDNA with those reported recently by the Zagrovic group²¹ for interactions of amino acid sidechains with isolated nucleobases. Given that the latter study considers interactions with neutral DNA bases while our study considers interactions with DNA bases in the context of negatively-charged ssDNA, we omit charged amino acid sidechains from the comparison. Importantly, for interactions with all four DNA bases, the qualitative agreement between the two data-sets is excellent, despite the use of quite different simulation force fields and water models, and despite the structural differences between single-stranded DNAs and isolated nucleobases; the only major difference is that our numbers are shifted to significantly more favorable values. Again, to show that such a degree of correlation is not trivial, we can also compare our G_{int} values for dsDNA with those reported by the Zagrovic group. As should be expected, the correlations in this case are much worse than those obtained with our ssDNA results (Figure S6).

As pointed out by a reviewer, we can also compare our computed G_{int} values for interactions with ssDNA, with experimental G_{int} values derived from equilibrium constants by Thomas & Podder,⁵⁹ for interactions of (zwitterionic) amino acids with nucleosides. Figure 5E shows that our computed G_{int} values for interactions of seven neutral sidechains with guanine bases in ssDNA correlate surprisingly well with the corresponding

experimental G_{int} values for interactions with guanosine ($r^2 = 0.98$). No such degree of correlation is obtained when we compare with Woese's more indirectly related "polar requirement" scale⁶⁰ (Figure S7).

As also pointed out by a reviewer, the generally high degree of similarity between our computed G_{int} values for the four different bases of ssDNA (see Figure 5B), suggests that these interactions might be determined primarily by some physicochemical feature of the amino acid sidechain. Perhaps surprisingly, we find essentially no correlation ($r^2 = 0.07$) between the average G_{int} value of each type of sidechain with the four ssDNA bases and the calculated hydration free energy, $G_{\text{hydration}}$, of the same sidechain reported by the Pande group²⁶ (Figure S8). Instead, we find an extremely strong correlation ($r^2 = 0.94$) between our computed average G_{int} values and the number of heavy atoms in the sidechain (Figure 5F). One possible explanation for this finding is that the primary determinants of the simulated interactions are the van der Waals interactions between the amino acid sidechains and the DNA bases since these are the only type of direct interactions that are expected to scale simply with the number, rather than the chemical identity, of the heavy atoms. Despite the lack of correlation with the hydration free energies noted above, however, it is also possible that the finding reflects the entropically favorable release of bound water molecules. Interestingly, while it seems reasonable to be skeptical that 'real life' interactions would scale in such a simple manner, we nevertheless find that the experimental G_{int} values obtained from the work of Thomas and Podder⁵⁹ exhibit exactly the same linear scaling with the number of heavy atoms in the amino acid sidechain (Figure S9).

Finally, we have also analyzed the interactions between the Na^+ and Cl^- ions and the DNA and have computed G_{int} values in the same way that we computed values for the amino acid sidechains. Figure 7A compares the G_{int} profiles of Na^+ with dsDNA (blue) and ssDNA (red). Perhaps surprisingly, the G_{int} value at the distance corresponding to direct contact with the DNA (2.1 Å) is effectively identical in both dsDNA and ssDNA: it is only at longer distances, e.g. at the distance corresponding to a solvent-separated contact (4.1 Å), that the affinity of Na^+ for dsDNA is clearly greater than that for ssDNA. Figure 7B compares the minimum values of G_{int} for Na^+ interacting with each of the six different groups in dsDNA (blue), and ssDNA (red), and also plots the difference between each of these values (green). Of greatest interest is the fact that in going from dsDNA to ssDNA the effective G_{int} value of Na^+ with cytosine becomes more favorable by 0.7 kcal/mol, while its value for adenine becomes less favorable by 0.5 kcal/mol and its value for both guanine and thymine remains unchanged. As is outlined in detail in the Discussion, this increased affinity of Na^+ for cytosine in ssDNA provides a novel potential explanation for the quite different salt dependences of DNA duplex stability in AT-rich and GC-rich regions.

Figure 7C compares the G_{int} profiles of Cl^- with dsDNA (blue) and ssDNA (red); in this case, the greater affinity of Cl^- for ssDNA (which has a significantly lower negative charge density relative to dsDNA) is apparent throughout the full range of interaction distances. Figure 7D compares the minimum values of G_{int} for Cl^- interacting with the six different groups of DNA in the same way as shown for Na^+ in Figure 7B. In going from dsDNA to ssDNA the effective G_{int} values of Cl^- for all four bases becomes more favorable, with that for Cl^- interacting with guanine changing the most. Compounding the trend seen above with

Na⁺, it is apparent that relative to dsDNA, Cl⁻ interacts with guanine more favorably than adenine in ssDNA and interacts with cytosine (slightly) more favorably than thymine.

The impact of these different apparent affinities of the salt ions for the DNA bases in dsDNA and ssDNA can be seen in Figure 8. Figure 8A plots the total number of Na⁺ ions bound per AT (blue) or GC (red) base-pair as a function of the distance from the base-pair in dsDNA. Figure 8B shows the same but for the base-pairs when in the single-stranded state; for this case, the values obtained for the constituent nucleotides of each type of base-pair were simply summed: the total number of Na⁺ ions bound per AT 'base-pair' in the single-stranded state, for example, was obtained by adding the values computed for the A and T nucleotides in ssDNA. In dsDNA (Figure 8A) it can be seen that the apparent affinity of Na⁺ for the two different types of base-pairs is effectively identical; in ssDNA, on the other hand (Figure 8B), it is clear that Na⁺ has higher affinity for a separated GC base-pair than for a separated AT base-pair. Qualitatively identical behavior is observed for the binding of Cl⁻ ions: in dsDNA (Figure 8C), the affinity of Cl⁻ for the two types of base-pairs is identical, whereas in ssDNA (Figure 8D), the affinity for the GC base-pair is clearly greater than that for the AT base-pair. The potential consequences of these differences for understanding the salt dependent stability of duplex DNA is described below.

Discussion

In the present study we have used a "one pot" strategy in an attempt to simultaneously compare the interactions of all types of amino acid sidechains with dsDNA and ssDNA. While similar in spirit to previous simulation studies that have used mixtures of small organic molecules to identify potential binding sites on protein receptors^{28, 30, 61} this is, to our knowledge, the first attempt to simultaneously model the interactions of all types of amino acid sidechain and the first application of such an approach to DNA. As such, while there are many factors that critically affect the specificity of protein-DNA interactions,⁶² the simulations provide a means of isolating one such factor, namely, the intrinsic preferences of the different types of amino acid sidechain for interacting with DNA. While it should be remembered that the G_{int} values that we report are only apparent values (see Computational Methods), we think that this limitation is more than compensated for by the ability to simultaneously compare many different types of sidechain allowed by the "one pot" approach.

Before discussing the principal results of the simulations further, it is important – as always with simulation-based work – to consider whether the sampling achieved by the simulations is sufficient to enable us to draw clear conclusions. With regard to the Na⁺ and Cl⁻ behavior, adequate sampling is suggested by the fact that G_{int} profiles calculated from the four independent replicate simulations for Na⁺ and Cl⁻ are essentially identical to each other (see standard deviations of the minimum- G_{int} values in Tables 1 and 2). A reasonable degree of convergence is in any case to be expected since the 500 ns duration of the simulations is consistent with current estimates of the timescale required for sampling, at least in an averaged sense, the behavior of ions binding to dsDNA.⁶³

For the amino acid sidechains, on the other hand, there is evidence of variability between the independent replicate simulations (see, e.g. Figure S1), and these discrepancies are amplified when the analysis is extended to compare amino acid preferences for the four different types of DNA bases (Figures S3 and S4). Nevertheless, the relative preferences of the amino acid sidechains for binding to the DNA appear to be reasonably well established during the course of the 500 ns simulation periods. Figure S10, for example, compares the ranks of the sidechains when ordered according to their G_{int} values calculated over the last 166 ns of one of the four replicate dsDNA simulations, with the rank ordering obtained from analyzing the first 166 ns (Figure S10A), and the second 166 ns (Figure S10B) of the same simulation. In both cases, the rank orderings are strongly correlated (Spearman rank correlations of 0.85 and 0.93, respectively), with the poorer agreements being obtained for the more weakly binding sidechains (i.e. those with the highest ranks). Reinforcing this result is the finding that the computed G_{int} values computed from the three 166 ns blocks of the same replicate simulation show no consistent trend (Figure S10C): while some values become progressively more favorable (e.g. Arg, Lys, Trp), others become less so (e.g. Cys, Phe, Thr). Coupling these results with the close correspondence obtained between the behaviors of chemically similar sidechains (see, e.g. Figure 2), and the excellent agreement of our ssDNA results with the Zagrovic group's results for nucleobases²¹ (Figure 5D), we think that most of the predictions of the simulations reported here are likely to be robust.

That said, in addition to the obvious issue that the conformational behavior of the DNA – especially that of the ssDNA – will not be fully sampled in simulations of 500 ns duration, one other aspect of behavior that is unlikely to have been fully sampled is the potential intercalation of aromatic and aliphatic sidechains into dsDNA. Intercalation of sidechains is a known feature of many protein-DNA complexes,⁶⁴ and while intercalation events are repeatedly observed in the ssDNA simulations (Figure 3), they are not seen in the dsDNA simulations. Observation of such events would likely require substantially longer simulation times, or enhanced sampling methods. For this reason, we suspect that our estimates of G_{int} for aromatic and aliphatic sidechains with dsDNA may be somewhat too positive (i.e. insufficiently favorable).

Support for this idea can be found in Figure 4, which compares the relative preferences for ssDNA and dsDNA computed from the simulations with the experimental values derived from the recent crystallographic analysis of Wang et al.⁵⁶ If, as proposed above, the G_{int} values for aromatic and aliphatic sidechains with dsDNA are too positive, we should expect the corresponding $G_{\text{int}}(\text{ssDNA} - \text{dsDNA})$ values to be too negative, i.e. shifted to the left in Figure 4. This is indeed the case: Figure 9 redraws the data from Figure 4 but with the data-points for the aromatic and aliphatic sidechains presented as blue circles and all other sidechains presented as red triangles. It can be seen that the aromatic and aliphatic data-points all lie well to the left of a regression line fitted through the data-points for the other sidechains. In addition, it should be noted that the Pearson correlation coefficient for the non-aromatic, non-aliphatic data-points, R_{corr} , is 0.97, indicating that for those sidechains for which intercalation in dsDNA is not expected, the simulations exhibit a surprisingly good ability to reproduce their relative affinities for dsDNA and ssDNA.

While the simulations appear to capture the relative preferences of the amino acid sidechains for DNA as a whole, it was also hoped at the outset of this work that the MD simulations of dsDNA would faithfully capture the relative affinities of each of the amino acid sidechains for the four different DNA bases. There have been many previous attempts to determine these relative affinities using crystallographic analyses (see, for example, Sousa *et al.*⁶⁵ for a review), so the three data-sets that we have chosen (Table 4) only provide a representative comparison rather than a comprehensive one. It is clear, however, that in terms of the preferences exhibited by amino acid sidechains for the four DNA bases, there is little agreement with the existing data-sets (Table 4). In considering this generally low level of agreement, it should be remembered that there are large structural differences between the present MD simulations, which consider only isolated amino acid sidechains, and the experimentally-derived data-sets, which consider interactions within the context of complete protein-DNA complexes. In principle, the use here of sidechain-only models has the advantage of allowing us to assess their intrinsic relative affinities for the DNA bases in a way that is not influenced by the rest of the protein; as such, the simulations are perhaps more likely to reflect the behavior of residues in unstructured protein tails than in structured domains. But it should also be remembered that the C β methyl group, which would normally be obstructed by backbone atoms in real proteins, is fully solvent-exposed in the sidechain models and so becomes a possible new site for interaction, especially with methyl or methylene groups on the DNA; this may contribute in part to the generally high affinity observed for the deoxyribose group (see Results). The presence of the C β methyl group therefore serves as a second potentially confounding factor when comparing the MD-simulated preferences with those derived from crystallographic analyses of protein-DNA complexes (see below).

That said, there is at least agreement between all of the data-sets that, in dsDNA, the positively charged sidechains Arg and Lys interact most favorably with guanine relative to the other bases; this interaction is preferred owing to the possibility of a bidentate interaction with the N7 and O6 atoms exposed in the major groove. For the negatively charged sidechains, Asp and Glu, the three literature data-sets are in agreement that interactions are most favorable with cytosine. This preference is especially strong in the two statistically derived data-sets, but in the HINT-based analysis of Marabotti *et al.*, interactions with adenine are also found to be favorable. In the MD simulations, Asp and Glu exhibit little preference for any of the bases, and although their most favorable G_{int} values are with cytosine and adenine, respectively, this is almost certainly a lucky coincidence. Another commonly cited preference in the literature is that of the amide-containing sidechains Asn and Gln for adenine,^{2, 4, 11} although it should be noted that this is not a consistent result of all statistical analyses.^{5, 57} In contrast, the MD simulations show a preference for guanine, especially so in the case of Gln, for which the preference amounts to 0.82 kcal/mol. The crystallographically determined preference of Asn and Gln for adenine is due to the ability to form bidentate hydrogen bonding interactions with adenine's N7 and exocyclic NH₂ group in the *major* groove – and such interactions are indeed observed during the MD simulations. But, as noted by Seeman and colleagues forty years ago,¹ a very similar set of bidentate hydrogen bonding interactions can also be formed with guanine's N3 and exocyclic NH₂ group in the *minor* groove, and it is this interaction that the simulations

appear to find more favorable. The apparent preference for adenine in crystallographic analyses, therefore, may be partly a reflection of the fact that in protein-DNA complex structures, most amino acid–base contacts (80% in the analysis of Marabotti et al.¹⁷) occur in the major groove. This in turn is likely a function of the greater opportunities for specific recognition of the bases that the major groove offers relative to the minor groove. Obviously, features that reflect aspects of the biological function of DNA-binding proteins (i.e. their need to specifically recognize certain DNA sequences over others) will not be captured by the current simulations.⁶²

A final but unexpected result of the simulations reported here concerns the different relative preferences of Na⁺ and Cl[−] ions for AT and GC base-pairs in dsDNA and ssDNA. We are, of course, not the first to use MD simulations to examine the behavior of ions around dsDNA: detailed simulation studies examining dsDNA-ion interactions have been reported over many years^{66–71} and, with progressively longer simulation times becoming accessible, important new studies continue to be reported.^{63, 72} Simulations focusing on dsDNA-ion interactions have been used to test integral equation theories⁷³ and to model the competitive binding of different types of ions to dsDNA;^{74–76} they can also serve as a point of comparison with experiments that provide important tests of simulation parameters.^{77, 78}

In terms of measuring the sequence dependence of ion binding to dsDNA, the most detailed analyses available to date are those reported recently by the Lavery and Orozco groups. In the first of these studies, a novel helical coordinate system⁷⁹ was used to analyze separately the binding of K⁺ ions to all 136 possible tetranucleotide steps⁶³ in the microsecond-long dsDNA simulation trajectories carried out by the μ ABC consortium.⁸⁰ In the second of these studies,⁸¹ a total of 43 μ s of simulation time of the Dickerson-Drew dodecamer was used to investigate the dependence of DNA conformational properties on the water model and the parameters assigned to the dissolved ions. While the overall behavior of the DNA in this latter study was found to be little affected by changes to the water model or the ion parameters, more specific details of the DNA-ion interactions, as well as their rates of convergence with increasing simulation time, were shown to be very sensitive to the choice of parameters. The present simulations do not, in our view, permit a similar level of base-by-base analysis to be carried out with any confidence, in part because of the complicating presence of a wide variety of other solutes all capable of competing for access to binding sites in the major and minor grooves, but also because the Orozco group's study indicates that for individual base-pair steps, convergence of the ion behavior can remain incomplete even after 3–5 μ s of simulation time.⁸¹ For these reasons, we have limited our analysis here to the averaged behavior of the full population of AT and GC base-pairs in dsDNA and ssDNA, independent of the identities of the neighboring base-pairs.

Just as we are not the first to study the binding of ions to dsDNA, we are not the first to use explicit-solvent MD simulations to model ssDNA. A number of MD studies, in particular, have focused on the effects of salts on the conformational collapse of short ssDNA oligonucleotides. A number of years ago, for example, the Kloo group⁸² reported nanosecond-timescale simulations of the effects of Na⁺ on DNA oligonucleotides ranging from (dT)₂ to (dT)₈. More recently, the Bandyopadhyay group^{83, 84} reported two MD simulation studies exploring the effects of Na⁺ on the conformational behavior of the single-

stranded version of the Dickerson-Drew dodecamer sequence. The Chakrabarti group,⁸⁵ on the other hand, recently reported MD simulations examining the effects of added Mg^{2+} on the behavior of the same sequence. Interestingly, in both sets of studies, collapse of the DNA led to stacking interactions between bases that were not adjacent in sequence; the formation of similar, “non-native” interactions in our ssDNA simulations is likely to be limited by the imposition of artificial periodicity and constant-volume conditions, which together act to prevent collapse of the DNA. Finally, using a very similar force field to that used here, the Pyshnyi group recently reported MD simulations of a very large number of ssDNA and dsDNA oligonucleotides, and showed that they could be used to calculate duplex hybridization enthalpies that were in very good agreement with experimental values.⁸⁶

While there have been a number of prior MD studies of ssDNA, therefore, to our knowledge, this is the first study to directly compare the relative preferences of salt ions for dsDNA and ssDNA and to find a key difference between the average preferences for the nucleotides of AT and GC base-pairs in the double- and single-stranded states. The finding that both Na^+ and Cl^- bind with greater affinity to separated GC base-pairs than to separated AT base-pairs in ssDNA, while their relative affinities are approximately equal in dsDNA, has a straightforward but potentially important implication: it suggests that the salt-dependent increase in duplex stability of DNAs that are rich in GC base pairs should be lower than the salt-dependent increase in duplex stability of DNAs that are rich in AT base pairs. As outlined below, this is consistent with a number of experimental studies.

Frank-Kaminetski first showed⁸⁷ that the melting temperature, T_m , of large DNAs could be expressed in a way that reflects its dependence on NaCl concentration and on its GC content. In terms of the salt dependence of T_m , Frank-Kaminetski's equation can be expressed⁸⁸ as $dT_m / d\log[\text{Na}^+] = 18.30 - 7.04 f(\text{GC})$, where $f(\text{GC})$ denotes the fractional GC content of the DNA; this indicates that the salt dependence of duplex stability decreases with increasing GC content. Blake and Haydock subsequently reported a similar result for Lambda DNA;⁸⁸ their equation reads $dT_m / d\log[\text{Na}^+] = 19.96 - 6.65 f(\text{GC})$. Those same authors also indicated that the salt dependence of T_m was only dependent upon the GC content at salt concentrations above 60 mM. This suggests, consistent with what is seen in this study, that it is a relatively weak binding event that is responsible for the lower salt dependence of GC base-pairs. Interestingly, they also explicitly suggested that the lower salt dependence for GC-rich DNA was due to greater Na^+ binding to the GC single-stranded state than to the AT single-stranded state. While this is also consistent with the behavior reported here, the mechanism proposed by Blake and Haydock to explain greater binding was quite different: they suggested that GC-rich regions are more likely to remain stacked in the single-stranded state and therefore be in a conformation with a higher negative charge density and a concomitant higher affinity for Na^+ . In contrast, in the present study, the greater binding to separated GC base-pairs in the single-stranded state is proposed to be due to binding of Na^+ ions to functional groups in cytosine that become exposed when Watson-Crick base-pairing is disrupted; this greater binding of Na^+ is reinforced by a greater binding of Cl^- ions to guanine.

A later analysis by Blake's group⁸⁹ of the melting behavior of plasmids extended their results to arrive at $dT_m / d\log[\text{Na}^+]$ values for each of the 10 possible nearest neighbor steps

(i.e. ApA, ApC, ... GpG); the average for the ApA, ApT and TpA steps was 20.4 ± 1.5 K, while that for the GpG, GpC and CpG steps was 13.5 ± 3.0 K. A fuller exploration of the effects of salt and sequence composition on the melting temperatures of oligonucleotides (as opposed to much larger plasmids) was reported by Owczarzy et al.⁹⁰ Interestingly, while those authors considered a possible model in which different salt dependences were allowed for each of the 10 possible base-pair steps, they found that results were not significantly improved over a simpler model in which only the GC content determined the salt dependence. More recently, the Ritort group derived salt dependent terms for the 10 nearest neighbor steps using single-molecule techniques, again finding that those containing GC base-pairs are generally lower than those containing AT base-pairs.⁹¹

We note that the ten values reported by the Blake⁸⁹ and Ritort⁹¹ groups correlate only very modestly with each other ($R_{\text{corr}} = 0.30$), but that the correlation coefficient rises to 0.999 when values are instead grouped and averaged according to their GC content (i.e. (a) ApA, ApT, TpA; (b) ApC, ApG, GpA, CpA; (c) CpC, CpG, GpC). The fact that there does not yet appear to be an especially compelling difference between the salt dependences of GpC, CpG and CpC steps in the literature is consistent with the model proposed here since the ability of Na^+ and Cl^- ions to directly coordinate with functional groups on cytosine and guanine, respectively, in the single-stranded state would be expected to be largely independent of the presence or identity of neighboring bases. That said, it should be also be noted that the Lavery group's recent analysis⁶³ of K^+ binding in the large-scale simulations performed by the μABC consortium⁸⁰ has indicated that binding to dsDNA can depend sensitively on not only the nature of the base-pair step but also on the identities of the flanking base-pairs. It is quite possible, therefore, that there might be subtleties to the NaCl-dependence of DNA duplex stability that are not captured by the model proposed here. This is especially likely to be true given the potentially significant parameter dependence of the DNA-ion interactions noted in the Orozco group's recent work.⁸¹

In closing, we note that the simulation data reported here can also, in principle, be used to derive coarse-grained (CG) potential functions for use in simplified but large-scale simulations of protein-nucleic acid systems. We have previously used explicit-solvent MD simulations of all possible pairs of amino acids as a basis for deriving multiple-bead-per-residue CG potential functions^{92, 93} for protein-protein interactions using the Iterative Boltzmann Inversion approach;^{94, 95} similar work has also been reported by the Betancourt group.⁹⁶ The present data should allow a similar approach to be used to derive potential functions for amino acid interactions with the bases, sugar and phosphate groups of DNA in both the single- and double-stranded states. CG simulation models have already been used in a number of very interesting studies of protein-DNA interactions; see, for example, the works of the Levy^{97, 98} and Takada⁹⁹ groups, with at least one involving modeling of DNA in its single-stranded state.¹⁰⁰ The use of potential functions derived from atomistic simulations that have been shown to reproduce relative affinities of amino acids for dsDNA and ssDNA, could enable such simulations to achieve even higher levels of realism, especially in modeling processes such as DNA replication, in which both single- and double-stranded forms of DNA play important roles.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by NIH R01 GM099865 and R01 GM087290 awarded to A.H.E. The authors would like to thank the anonymous reviewers for many insightful comments.

References

1. Seeman NC, Rosenberg JM, Rich A. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci USA*. 1976; 73:804–808. [PubMed: 1062791]
2. Suzuki M. A framework for the DNA–protein recognition code of the probe helix in transcription factors: the chemical and stereochemical rules. *Structure*. 1994; 2:317–326. [PubMed: 8087558]
3. Mandel-Gutfreund Y, Schueler O, Margalit H. Comprehensive Analysis of Hydrogen Bonds in Regulatory Protein DNA-Complexes: In Search of Common Principles. *J Mol Biol*. 1995; 253:370–382. [PubMed: 7563096]
4. Luscombe NM, Laskowski RA, Thornton JM. Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Res*. 2001; 29:2860–2874. [PubMed: 11433033]
5. Lejeune D, Delsaux N, Charlotiaux B, Thomas A, Brasseur R. Protein–nucleic acid recognition: Statistical analysis of atomic interactions and influence of DNA structure. *Proteins: Struct, Funct Bioinf*. 2005; 61:258–271.
6. Michael Gromiha M, Santhosh C, Suwa M. Influence of cation– π interactions in protein–DNA complexes. *Polymer*. 2004; 45:633–639.
7. Baker CM, Grant GH. Role of aromatic amino acids in protein–nucleic acid recognition. *Biopolymers*. 2007; 85:456–470.
8. Wilson KA, Kellie JL, Wetmore SD. DNA–protein π -interactions in nature: abundance, structure, composition and strength of contacts between aromatic amino acids and DNA nucleobases or deoxyribose sugar. *Nucleic Acids Res*. 2014; 42:6726–6741. [PubMed: 24744240]
9. Suzuki M, Yagi N. DNA recognition code of transcription factors in the helix–turn–helix, probe helix, hormone receptor, and zinc finger families. *Proc Natl Acad Sci USA*. 1994; 91:12357–12361. [PubMed: 7809040]
10. Lustig B, Jernigan RL. Consistencies of individual DNA base–amino acid interactions in structures and sequences. *Nucleic Acids Res*. 1995; 23:4707–4711. [PubMed: 8524664]
11. Mandel-Gutfreund Y, Margalit H. Quantitative parameters for amino acid–base interaction: Implications for prediction of protein–DNA binding sites. *Nucleic Acids Res*. 1998; 26:2306–2312. [PubMed: 9580679]
12. Kono H, Sarai A. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins: Struct, Funct Bioinf*. 1999; 35:114–131.
13. Paillard G, Lavery R. Analyzing Protein–DNA Recognition Mechanisms. *Structure*. 2004; 12:113–122. [PubMed: 14725771]
14. Endres RG, Schulthess TC, Wingreen NS. Toward an atomistic model for predicting transcription-factor binding sites. *Proteins: Struct, Funct Bioinf*. 2004; 57:262–268.
15. Morozov AV, Havranek JJ, Baker D, Siggia ED. Protein–DNA binding specificity predictions with structural models. *Nucleic Acids Res*. 2005; 33:5781–5798. [PubMed: 16246914]
16. Spyraakis F, Cozzini P, Bertoli C, Marabotti A, Kellogg GE, Mozzarelli A. Energetics of the protein–DNA–water interaction. *BMC structural biology*. 2007; 7:4. [PubMed: 17214883]
17. Marabotti A, Spyraakis F, Facchiano A, Cozzini P, Alberti S, Kellogg GE, Mozzarelli A. Energy-based prediction of amino acid–nucleotide base recognition. *J Comput Chem*. 2008; 29:1955–1969. [PubMed: 18366021]

18. Jayaram B, McConnell K, Dixit SB, Das A, Beveridge DL. Free-energy component analysis of 40 protein–DNA complexes: A consensus view on the thermodynamics of binding at the molecular level. *J Comput Chem*. 2002; 23:1–14. [PubMed: 11913374]
19. Pichierri F, Aida M, Gromiha MM, Sarai A. Free-Energy Maps of Base–Amino Acid Interactions for DNA–Protein Recognition. *J Am Chem Soc*. 1999; 121:6152–6157.
20. Yoshida T, Nishimura T, Aida M, Pichierri F, Gromiha MM, Sarai A. Evaluation of free energy landscape for base–amino acid interactions using ab initio force field and extensive sampling. *Biopolymers*. 2001; 61:84–95.
21. de Ruiter A, Zagrovic B. Absolute binding-free energies between standard RNA/DNA nucleobases and amino-acid sidechain analogs in different environments. *Nucleic Acids Res*. 2015; 43:708–718. [PubMed: 25550435]
22. Echeverria I, Papoian GA. DNA Exit Ramps Are Revealed in the Binding Landscapes Obtained from Simulations in Helical Coordinates. *PLoS Comput Biol*. 2015; 11:e1003980. [PubMed: 25675216]
23. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. *J Comput Chem*. 2005; 26:1701–1718. [PubMed: 16211538]
24. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J Chem Theory Comput*. 2008; 4:435–447. [PubMed: 26620784]
25. Plumridge A, Meisburger SP, Pollack L. Visualizing single-stranded nucleic acids in solution. *Nucleic Acids Res*. 2016
26. Shirts MR, Pitera JW, Swope WC, Pande VS. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J Chem Phys*. 2003; 119:5740–5761.
27. Deng Y, Roux B. Hydration of Amino Acid Side Chains: Nonpolar and Electrostatic Contributions Calculated from Staged Molecular Dynamics Free Energy Simulations with Explicit Water Molecules. *J Phys Chem B*. 2004; 108:16567–16576.
28. Guvench O, MacKerell AD Jr. Computational Fragment-Based Binding Site Identification by Ligand Competitive Saturation. *PLoS Comput Biol*. 2009; 5:e1000435. [PubMed: 19593374]
29. Seco J, Luque FJ, Barril X. Binding Site Detection and Druggability Index from First Principles. *J Med Chem*. 2009; 52:2363–2371. [PubMed: 19296650]
30. Raman EP, Yu W, Lakkaraju SK, MacKerell AD. Inclusion of Multiple Fragment Types in the Site Identification by Ligand Competitive Saturation (SILCS) Approach. *J Chem Inf Model*. 2013; 53:3384–3398. [PubMed: 24245913]
31. Lexa KW, Goh GB, Carlson HA. Parameter Choice Matters: Validating Probe Parameters for Use in Mixed-Solvent Simulations. *J Chem Inf Model*. 2014; 54:2190–2199. [PubMed: 25058662]
32. Cheatham TE, Cieplak P, Kollman PA. A Modified Version of the Cornell et al. Force Field with Improved Sugar Pucker Phases and Helical Repeat. *J Biomol Struct Dyn*. 1999; 16:845–862. [PubMed: 10217454]
33. Pérez A, Marchán I, Svozil D, Sponer J, Cheatham TE III, Laughton CA, Orozco M. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers. *Biophys J*. 2007; 92:3817–3829. [PubMed: 17351000]
34. Krepl M, Zgarbova M, Stadlbauer P, Otyepka M, Banas P, Koca J, Cheatham TE 3rd, Jurecka P, Sponer J. Reference simulations of noncanonical nucleic acids with different chi variants of the AMBER force field: quadruplex DNA, quadruplex RNA and Z-DNA. *Journal of chemical theory and computation*. 2012; 8:2506–2520. [PubMed: 23197943]
35. Kührová P, Banáš P, Best RB, Šponer J, Otyepka M. Computer Folding of RNA Tetraloops? Are We There Yet? *J Chem Theory Comput*. 2013; 9:2115–2125. [PubMed: 26583558]
36. Stadlbauer P, Krepl M, Cheatham TE 3rd, Koca J, Sponer J. Structural dynamics of possible late-stage intermediates in folding of quadruplex DNA studied by molecular simulations. *Nucleic Acids Res*. 2013; 41:7128–43. [PubMed: 23700306]
37. Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Struct, Funct Bioinf*. 2006; 65:712–725.

38. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct, Funct Bioinf.* 2010; 78:1950–1958.
39. Horn HW, Swope WC, Pitera JW, Madura JD, Dick TJ, Hura GL, Head-Gordon T. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. *J Chem Phys.* 2004; 120:9665–9678. [PubMed: 15267980]
40. Brown RF, Andrews CT, Elcock AH. Stacking Free Energies of All DNA and RNA Nucleoside Pairs and Dinucleoside-Monophosphates Computed Using Recently Revised AMBER Parameters and Compared with Experiment. *J Chem Theory Comput.* 2015; 11:2315–2328. [PubMed: 26574427]
41. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. *J Chem Phys.* 1983; 79:926–935.
42. Beauchamp KA, Lin YS, Das R, Pande VS. Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *J Chem Theory Comput.* 2012; 8:1409–1414. [PubMed: 22754404]
43. Joung IS, Cheatham TE. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J Phys Chem B.* 2008; 112:9020–9041. [PubMed: 18593145]
44. Bergonzo C, Henriksen NM, Roe DR, Swails JM, Roitberg AE, Cheatham TE 3rd. Multidimensional Replica Exchange Molecular Dynamics Yields a Converged Ensemble of an RNA Tetranucleotide. *J Chem Theory Comput.* 2014; 10:492–499. [PubMed: 24453949]
45. Yoo J, Aksimentiev A. Improved Parametrization of Li⁺, Na⁺, K⁺, and Mg²⁺ Ions for All-Atom Molecular Dynamics Simulations of Nucleic Acid Systems. *J Phys Chem Lett.* 2012; 3:45–50.
46. Yoo J, Aksimentiev A. Improved Parameterization of Amine–Carboxylate and Amine–Phosphate Interactions for Molecular Dynamics Simulations Using the CHARMM and AMBER Force Fields. *J Chem Theory Comput.* 2016; 12:430–443. [PubMed: 26632962]
47. Parrinello M, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys.* 1981; 52:7182–7190.
48. Nosé S. A unified formulation of the constant temperature molecular dynamics methods. *The Journal of Chemical Physics.* 1984; 81:511–519.
49. Hoover WG. Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A.* 1985; 31:1695–1697.
50. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A linear constraint solver for molecular simulations. *J Comput Chem.* 1997; 18:1463–1472.
51. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. *J Chem Phys.* 1995; 103:8577.
52. Zhu S, Elcock AH. A Complete Thermodynamic Characterization of Electrostatic and Hydrophobic Associations in the Temperature Range 0 to 100 °C from Explicit-Solvent Molecular Dynamics Simulations. *J Chem Theory Comput.* 2010; 6:1293–1306.
53. Thomas AS, Elcock AH. Direct Measurement of the Kinetics and Thermodynamics of Association of Hydrophobic Molecules from Molecular Dynamics Simulations. *J Phys Chem Lett.* 2011; 2:19–24. [PubMed: 26295208]
54. Andrews CT, Elcock AH. Molecular Dynamics Simulations of Highly Crowded Amino Acid Solutions: Comparisons of Eight Different Force Field Combinations with Experiment and with Each Other. *J Chem Theory Comput.* 2013; 9:4585–4602.
55. Kang M, Smith PE. Kirkwood–Buff theory of four and higher component mixtures. *J Chem Phys.* 2008; 128:244511. [PubMed: 18601352]
56. Wang W, Liu J, Sun L. Surface shapes and surrounding environment analysis of single- and double-stranded DNA-binding proteins in protein-DNA interface. *Proteins: Struct, Funct Bioinf.* 2016; 84:979–989.
57. Gao M, Skolnick J. DBD-Hunter: a knowledge-based method for the prediction of DNA–protein interactions. *Nucleic Acids Res.* 2008; 36:3978–3992. [PubMed: 18515839]

58. Schrodt MV, Andrews CT, Elcock AH. Large-Scale Analysis of 48 DNA and 48 RNA Tetranucleotides Studied by 1 μ s Explicit-Solvent Molecular Dynamics Simulations. *J Chem Theory Comput.* 2015; 11:5906–5917. [PubMed: 26580891]
59. Thomas PD, Podder SK. Specificity in protein–nucleic acid interaction: Solubility study on amino acid–nucleoside interaction. *FEBS Lett.* 1978; 96:90–94. [PubMed: 729797]
60. Woese CR. Evolution of the genetic code. *The Science of Nature.* 1973; 60:447–459.
61. Bakan A, Nevins N, Lakdawala AS, Bahar I. Druggability Assessment of Allosteric Proteins by Dynamics Simulations in the Presence of Probe Molecules. *J Chem Theory Comput.* 2012; 8:2435–2447. [PubMed: 22798729]
62. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. Origins of Specificity in Protein-DNA Recognition. *Annu Rev Biochem.* 2010; 79:233–269. [PubMed: 20334529]
63. Pasi M, Maddocks JH, Lavery R. Analyzing ion distributions around DNA: sequence-dependence of potassium ion distributions from microsecond molecular dynamics. *Nucleic Acids Res.* 2015; 43:2412–2423. [PubMed: 25662221]
64. Werner MH, Gronenborn AM, Clore GM. Intercalation, DNA Kinking, and the Control of Transcription. *Science.* 1996; 271:778–784. [PubMed: 8628992]
65. Sousa F, Cruz C, Queiroz JA. Amino acids–nucleotides biomolecular recognition: from biological occurrence to affinity chromatography. *Journal of Molecular Recognition.* 2010; 23:505–518. [PubMed: 21038352]
66. Young MA, Jayaram B, Beveridge DL. Intrusion of Counterions into the Spine of Hydration in the Minor Groove of B-DNA: Fractional Occupancy of Electronegative Pockets. *J Am Chem Soc.* 1997; 119:59–69.
67. Jayaram B, Beveridge DL. Modeling DNA in Aqueous Solutions: Theoretical and Computer Simulation Studies on the Ion Atmosphere of DNA. *Annu Rev Biophys Biomol Struct.* 1996; 25:367–394. [PubMed: 8800475]
68. Feig M, Pettitt BM. Sodium and Chlorine Ions as Part of the DNA Solvation Shell. *Biophys J.* 1999; 77:1769–1781. [PubMed: 10512802]
69. Auffinger P, Westhof E. Water and ion binding around RNA and DNA (C,G) oligomers1. *J Mol Biol.* 2000; 300:1113–1131. [PubMed: 10903858]
70. Várnai P, Zakrzewska K. DNA and its counterions: a molecular dynamics study. *Nucleic Acids Res.* 2004; 32:4269–4280. [PubMed: 15304564]
71. Savelyev A, Papoian GA. Electrostatic, Steric, and Hydration Interactions Favor Na⁺ Condensation around DNA Compared with K⁺ *J Am Chem Soc.* 2006; 128:14506–14518. [PubMed: 17090034]
72. Dixit SB, Mezei M, Beveridge DL. Studies of base pair sequence effects on DNA solvation based on all-atom molecular dynamics simulations. *J Biosci.* 2012; 37:399–421. [PubMed: 22750979]
73. Howard JJ, Lynch GC, Pettitt BM. Ion and Solvent Density Distributions around Canonical B-DNA from Integral Equations. *J Phys Chem B.* 2011; 115:547–556. [PubMed: 21190358]
74. Giamba u GM, Gebala MK, Panteva MT, Luchko T, Case DA, York DM. Competitive interaction of monovalent cations with DNA from 3D-RISM. *Nucleic Acids Res.* 2015; 43:8405–8415. [PubMed: 26304542]
75. Yoo J, Aksimentiev A. Competitive Binding of Cations to Duplex DNA Revealed through Molecular Dynamics Simulations. *J Phys Chem B.* 2012; 116:12946–12954. [PubMed: 23016894]
76. Savelyev A, MacKerell AD. Competition among Li⁺, Na⁺, K⁺, and Rb⁺ Monovalent Ions for DNA in Molecular Dynamics Simulations Using the Additive CHARMM36 and Drude Polarizable Force Fields. *J Phys Chem B.* 2015; 119:4428–4440. [PubMed: 25751286]
77. Gebala M, Giamba u GM, Lipfert J, Bisaria N, Bonilla S, Li G, York DM, Herschlag D. Cation–Anion Interactions within the Nucleic Acid Ion Atmosphere Revealed by Ion Counting. *J Am Chem Soc.* 2015; 137:14705–14715. [PubMed: 26517731]
78. Gebala M, Bonilla S, Bisaria N, Herschlag D. Does cation size affect occupancy and electrostatic screening of the nucleic acid ion atmosphere? *J Am Chem Soc.* 2016
79. Lavery R, Maddocks JH, Pasi M, Zakrzewska K. Analyzing ion distributions around DNA. *Nucleic Acids Res.* 2014; 42:8138–8149. [PubMed: 24906882]

80. Pasi M, Maddocks JH, Beveridge D, Bishop TC, Case DA, Cheatham T, Dans PD, Jayaram B, Lankas F, Laughton C, Mitchell J, Osman R, Orozco M, Pérez A, Petkevicius D, Spackova N, Spomer J, Zakrzewska K, Lavery R. μ ABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.* 2014; 42:12272–12283. [PubMed: 25260586]
81. Dans PD, Danilne L, Ivani I, Dršata T, Lankaš F, Hospital A, Walther J, Pujagut RI, Battistini F, Gelpí JL, Lavery R, Orozco M. Long-timescale dynamics of the Drew–Dickerson dodecamer. *Nucleic Acids Res.* 2016; 44:4052–4066. [PubMed: 27084952]
82. Martínez JM, Elmroth SKC, Kloo L. Influence of Sodium Ions on the Dynamics and Structure of Single-Stranded DNA Oligomers: A Molecular Dynamics Study. *J Am Chem Soc.* 2001; 123:12279–12289. [PubMed: 11734028]
83. Chakraborty K, Mantha S, Bandyopadhyay S. Molecular dynamics simulation of a single-stranded DNA with heterogeneous distribution of nucleobases in aqueous medium. *J Chem Phys.* 2013; 139:075103. [PubMed: 23968115]
84. Chakraborty K, Khatua P, Bandyopadhyay S. Exploring ion induced folding of a single-stranded DNA oligomer from molecular simulation studies. *PCCP.* 2016; 18:15899–15910. [PubMed: 27241311]
85. Ghosh S, Dixit H, Chakrabarti R. Ion assisted structural collapse of a single stranded DNA: A molecular dynamics approach. *Chemical Physics.* 2015; 459:137–147.
86. Lomzov AA, Vorobjev YN, Pyshnyi DV. Evaluation of the Gibbs Free Energy Changes and Melting Temperatures of DNA/DNA Duplexes Using Hybridization Enthalpy Calculated by Molecular Dynamics Simulation. *J Phys Chem B.* 2015; 119:15221–15234. [PubMed: 26569147]
87. Frank-Kamenetskii MD. Simplification of the empirical relationship between melting temperature of DNA, its GC content and concentration of sodium ions in solution. *Biopolymers.* 1971; 10:2623–2624. [PubMed: 5126533]
88. Blake RD, Haydock PV. Effect of sodium ion on the high-resolution melting of lambda DNA. *Biopolymers.* 1979; 18:3089–3109. [PubMed: 160252]
89. Blake RD, Delcourt SG. Thermal stability of DNA. *Nucleic Acids Res.* 1998; 26:3323–3332. [PubMed: 9649614]
90. Owczarzy R, You Y, Moreira BG, Manthey JA, Huang L, Behlke MA, Walder JA. Effects of Sodium Ions on DNA Duplex Oligomers: Improved Predictions of Melting Temperatures. *Biochemistry.* 2004; 43:3537–3554. [PubMed: 15035624]
91. Huguet JM, Bizarro CV, Forn N, Smith SB, Bustamante C, Ritort F. Single-molecule derivation of salt dependent base-pair free energies in DNA. *Proc Natl Acad Sci USA.* 2010; 107:15431–15436. [PubMed: 20716688]
92. Andrews CT, Elcock AH. COFFDROP: A Coarse-Grained Nonbonded Force Field for Proteins Derived from All-Atom Explicit-Solvent Molecular Dynamics Simulations of Amino Acids. *J Chem Theory Comput.* 2014; 10:5178–5194. [PubMed: 25400526]
93. Frembgen-Kesner T, Andrews CT, Li S, Ngo NA, Shubert SA, Jain A, Olayiwola OJ, Weishaar MR, Elcock AH. Parametrization of Backbone Flexibility in a Coarse-Grained Force Field for Proteins (COFFDROP) Derived from All-Atom Explicit-Solvent Molecular Dynamics Simulations of All Possible Two-Residue Peptides. *J Chem Theory Comput.* 2015; 11:2341–2354. [PubMed: 26574429]
94. Schommers W. A pair potential for liquid rubidium from the pair correlation function. *Phys Lett A.* 1973; 43:157–158.
95. Reith D, Pütz M, Müller-Plathe F. Deriving effective mesoscale potentials from atomistic simulations. *J Comput Chem.* 2003; 24:1624–1636. [PubMed: 12926006]
96. Betancourt MR, Omovie SJ. Pairwise energies for polypeptide coarse-grained models derived from atomic force fields. *J Chem Phys.* 2009; 130:195103. [PubMed: 19466867]
97. Vuzman D, Levy Y. The “Monkey-Bar” Mechanism for Searching for the DNA Target Site: The Molecular Determinants. *Isr J Chem.* 2014; 54:1374–1381.
98. Bhattacharjee A, Krepel D, Levy Y. Coarse-grained models for studying protein diffusion along DNA. *Wiley Interdisciplinary Reviews: Computational Molecular Science.* 2016; 6:515–531.

99. Tan C, Terakawa T, Takada S. Dynamic Coupling among Protein Binding, Sliding, and DNA Bending Revealed by Molecular Dynamics. *J Am Chem Soc.* 2016; 138:8512–8522. [PubMed: 27309278]
100. Mishra G, Levy Y. Molecular determinants of the interactions between proteins and ssDNA. *Proc Natl Acad Sci USA.* 2015; 112:5033–5038. [PubMed: 25855635]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1. Initial and final configurations of typical ssDNA and dsDNA systems

A. Image showing ssDNA surrounded by randomly positioned solutes at the beginning of one of the four replicate MD simulations; water molecules are omitted for clarity. **B.** Same as **A** but showing the configuration at the end of the 500 ns production period of the simulation. **C.** Same as **A** but showing one of the dsDNA systems. **D.** Same as **B** but showing one of the dsDNA systems.

Figure 2. Apparent free energies of interaction of all amino acid sidechains with dsDNA and ssDNA

A. Apparent free energy of interaction, G_{int} , for all charged amino acid sidechains as a function of the minimum distance between heavy atoms of the sidechain and DNA; H_p denotes positively charged histidine. Each line represents the average of four independent replicate MD simulations; blue and red lines represent results for dsDNA and ssDNA, respectively. **B.** Same as **A** but showing results for polar amino acids (plus cysteine). **C.** Same as **A** but showing results for aromatic amino acids (plus the backbone analog, N-methylacetamide, denoted Ace). **D.** Same as **A** but showing results for aliphatic amino acids.

Figure 3. Simulation snapshots illustrating interactions of aromatic and aliphatic sidechains with ssDNA

A. Snapshot showing intercalation of a tryptophan sidechain between two bases with displacement of an intervening base. **B.** Snapshot showing intercalation of a leucine sidechain between bases.

Figure 4. Comparison of computed and experimental estimates of the apparent free energies of interaction of all amino sidechains with ssDNA relative to dsDNA

The x-axis plots, for each amino acid sidechain, the difference between the computed minimum G_{int} value for interaction with ssDNA and the computed minimum G_{int} value for interaction with dsDNA. The y-axis plots the same but using an experimentally derived estimate of the difference based on a recently reported analysis of crystallographic structures⁵⁶ (see text). Error bars for the simulation results represent the standard deviation of the minimum G_{int} values obtained in the four replicate simulations.

Figure 5. Computed apparent free energies of interaction of all amino sidechains with the four DNA bases

A. Minimum G_{int} values for each amino acid sidechain with each type of base in dsDNA; amino acid sidechains are arranged in order of increasingly favorable G_{int} values. **B.** Same as **A** but showing results for ssDNA. **C.** Same as **A** but showing the difference between the ssDNA and dsDNA minimum G_{int} values. **D.** Comparison of the minimum G_{int} values for all uncharged amino acid sidechains with ssDNA, with the values recently reported by de Ruiter & Zagrovic²¹ for interactions with nucleobases. **E.** Comparison of the minimum G_{int} values for seven uncharged amino acid sidechains with guanine bases in ssDNA, with the experimental values reported by Thomas & Podder for interactions with guanosine.⁵⁹ **F.** Comparison of the minimum G_{int} values for all uncharged amino acid sidechains with ssDNA, with the number of heavy atoms in the sidechain.

Figure 6. Simulation snapshots illustrating interactions of charged sidechains with ssDNA
A. Snapshot showing bidentate hydrogen bonding interaction of an Arg sidechain with a cytosine base; numbers indicate the distance between selected heavy atoms in Ångstroms. **B.** Snapshot showing bidentate hydrogen bonding interaction of a Glu sidechain with a guanine base.

Figure 7. Interactions of Na^+ and Cl^- ions with dsDNA and ssDNA

A. Apparent free energy of interaction, G_{int} , for Na^+ as a function of the minimum distance to heavy atoms of dsDNA (blue) and ssDNA (red). **B.** Minimum G_{int} values for Na^+ interacting with each of the four bases (A, C, G, T), the phosphate (P), and the sugar groups (S) in dsDNA (blue) and ssDNA (red), and their difference (green). **C.** Same as **A** but showing results for Cl^- . **D.** Same as **B** but showing results for Cl^- .

Figure 8. Number of bound Na^+ and Cl^- ions per AT and GC base-pairs

A. Cumulative number of Na^+ ions bound per AT base-pair (blue) and GC base-pair (red) in dsDNA as a function of distance from the nearest heavy atom of DNA. **B.** Same as **A** but showing results for ssDNA. **C.** Same as **A** but showing results for Cl^- . **D.** Same as **B** but showing results for Cl^- .

Figure 9. Intercalation of aliphatic and aromatic sidechains in dsDNA is likely underestimated Same as Figure 4 but separating the results for aliphatic and aromatic sidechains (blue circles) from those for all other sidechains (red triangles). The black line is the linear regression line for the red data-points.

Table 1
Minimum values of G_{int} for all amino sidechains and ions with the six groups of dsDNA and with dsDNA as a whole

Column headed “Ade Min”, for example, indicates the minimum value of G_{int} found within 5 Å for each type of sidechain with adenine; Column headed “Ade Std” indicates the standard deviation of the G_{int} values found in the four replicate simulations. Note that the exceptionally high standard deviation for Glu with adenine is due to a very long-lived interaction seen in a single replicate simulation.

	Ade Min	Ade Std	Cyt Min	Cyt Std	Gua Min	Gua Std	Thy Min	Thy Std	Pho Min	Pho Std	Sug Min	Sug Std	DNA Min	DNA Std
Ace	-1.39	0.61	-0.49	0.35	-1.12	0.32	-0.63	0.12	-0.22	0.14	-1.13	0.76	-0.54	0.22
Ala	-0.14	0.08	-0.18	0.12	-0.12	0.17	-0.36	0.01	0.33	0.06	-0.33	0.06	0.06	0.03
Cys	-0.65	0.07	-0.54	0.08	-0.98	0.23	-0.57	0.12	-0.07	0.04	-0.80	0.06	-0.32	0.05
Asp	0.12	0.31	-0.17	0.24	0.38	0.28	0.04	0.14	0.10	0.03	-0.14	0.07	0.15	0.03
Glu	-0.17	2.59	0.22	0.23	0.12	0.32	0.20	0.22	0.11	0.05	-0.26	0.12	0.14	0.05
Phe	-0.73	0.25	-0.83	0.08	-0.89	0.27	-0.60	0.11	0.09	0.05	-1.08	0.26	-0.34	0.14
His	-1.76	0.21	-1.45	0.27	-1.98	0.34	-1.23	0.30	-1.08	0.04	-1.46	0.08	-1.13	0.14
Hip	-3.26	0.34	-3.05	0.28	-4.00	0.10	-3.83	0.35	-2.55	0.08	-2.73	0.28	-2.91	0.17
Ile	-0.44	0.26	-0.27	0.19	-0.67	0.11	-0.63	0.15	0.33	0.07	-0.90	0.15	-0.14	0.10
Lys	-2.11	0.62	-2.20	0.82	-3.15	0.09	-2.69	0.31	-2.00	0.05	-1.13	0.08	-2.09	0.09
Leu	-0.45	0.21	-0.40	0.13	-0.44	0.46	-0.58	0.12	0.45	0.13	-0.65	0.24	0.01	0.14
Met	-0.57	0.13	-0.59	0.22	-0.82	0.27	-0.56	0.08	0.21	0.04	-0.83	0.07	-0.17	0.04
Asn	-1.79	0.22	-1.25	0.22	-1.92	0.15	-1.00	0.27	-0.99	0.05	-1.34	0.25	-1.06	0.05
Gln	-1.71	0.25	-1.59	0.37	-2.53	0.46	-1.26	0.27	-0.90	0.08	-1.58	0.19	-1.21	0.17
Arg	-3.58	0.22	-3.49	0.33	-4.34	0.31	-3.99	0.34	-3.05	0.09	-2.69	0.39	-3.26	0.18
Ser	-0.69	0.06	-0.46	0.08	-1.02	0.18	-0.55	0.06	-0.79	0.04	-0.72	0.03	-0.52	0.03
Thr	-0.85	0.26	-0.57	0.09	-1.01	0.10	-0.57	0.06	-0.57	0.04	-0.90	0.09	-0.43	0.06
Val	-0.38	0.24	-0.25	0.15	-0.47	0.16	-0.48	0.14	0.35	0.08	-0.70	0.11	-0.06	0.08
Trp	-2.68	0.38	-1.77	0.33	-2.07	0.32	-2.02	0.30	-1.00	0.19	-2.15	0.22	-1.36	0.23
Tyr	-1.79	0.61	-1.52	0.72	-2.33	0.18	-1.75	0.74	-0.99	0.08	-2.24	0.36	-1.20	0.25
Na⁺	-2.17	0.01	-1.33	0.03	-2.07	0.02	-1.83	0.08	-2.35	0.01	-1.41	0.02	-1.95	0.01
Cl⁻	1.40	0.15	1.21	0.10	1.53	0.08	1.24	0.04	0.58	0.01	1.19	0.03	0.70	0.01

Table 2
Minimum values of G_{int} for all amino sidechains and ions with the six groups of ssDNA and with ssDNA as a whole

Same as Table 1 but showing data for interactions with ssDNA.

	Ade Min	Ade Std	Cyt Min	Cyt Std	Gua Min	Gua Std	Thy Min	Thy Std	Pho Min	Pho Std	Sug Min	Sug Std	DNA Min	DNA Std
Ace	-1.12	0.12	-1.21	0.08	-1.08	0.13	-1.07	0.06	-0.27	0.10	-1.24	0.05	-0.85	0.04
Ala	-0.50	0.07	-0.52	0.08	-0.46	0.03	-0.56	0.07	0.18	0.04	-0.64	0.03	-0.30	0.04
Cys	-0.99	0.13	-1.02	0.04	-0.99	0.06	-0.93	0.07	-0.23	0.06	-1.06	0.08	-0.73	0.04
Asp	-0.45	0.26	-0.76	0.11	-1.24	0.40	-0.30	0.05	-0.04	0.05	-0.47	0.07	-0.34	0.15
Glu	-0.48	0.19	-0.56	0.10	-1.24	0.29	-0.45	0.18	0.02	0.08	-0.56	0.10	-0.28	0.24
Phe	-2.09	0.09	-1.81	0.28	-2.14	0.09	-2.31	0.10	-0.27	0.04	-1.96	0.10	-1.69	0.05
His	-1.47	0.12	-1.49	0.09	-1.67	0.19	-1.70	0.17	-1.17	0.06	-1.54	0.04	-1.32	0.06
Hip	-1.56	0.09	-2.60	0.14	-2.15	0.11	-1.75	0.07	-2.53	0.06	-1.67	0.08	-2.23	0.04
Ile	-1.36	0.10	-1.46	0.14	-1.44	0.04	-1.50	0.15	0.02	0.11	-1.43	0.09	-1.09	0.09
Lys	-1.10	0.09	-2.26	0.07	-1.93	0.10	-1.64	0.11	-2.04	0.07	-0.99	0.10	-1.84	0.06
Leu	-1.32	0.18	-1.36	0.14	-1.30	0.10	-1.39	0.07	-0.01	0.03	-1.42	0.03	-1.01	0.02
Met	-1.48	0.06	-1.53	0.20	-1.55	0.11	-1.76	0.10	-0.21	0.06	-1.55	0.04	-1.24	0.08
Asn	-1.27	0.04	-1.23	0.07	-1.62	0.13	-1.48	0.17	-1.06	0.04	-1.11	0.03	-1.18	0.05
Gln	-1.34	0.15	-1.28	0.15	-1.74	0.08	-1.40	0.14	-1.01	0.03	-1.19	0.06	-1.23	0.04
Arg	-2.02	0.15	-3.26	0.24	-2.49	0.14	-2.49	0.14	-3.35	0.03	-1.93	0.08	-3.02	0.05
Ser	-0.60	0.02	-0.56	0.15	-0.66	0.04	-0.60	0.06	-0.86	0.05	-0.76	0.07	-0.55	0.05
Thr	-0.74	0.08	-0.79	0.08	-0.73	0.10	-0.72	0.06	-0.65	0.10	-0.94	0.04	-0.54	0.04
Val	-1.00	0.13	-1.04	0.08	-1.03	0.05	-1.07	0.17	0.09	0.08	-1.12	0.10	-0.73	0.07
Trp	-2.74	0.26	-2.82	0.24	-2.87	0.16	-3.20	0.13	-1.57	0.20	-3.01	0.14	-2.59	0.11
Tyr	-2.43	0.23	-2.56	0.24	-2.61	0.17	-2.87	0.17	-1.58	0.19	-2.62	0.09	-2.24	0.14
Na⁺	-1.77	0.02	-2.16	0.04	-2.20	0.02	-2.06	0.05	-2.41	0.01	-0.95	0.01	-2.08	0.02
Cl⁻	0.55	0.03	0.58	0.05	0.38	0.04	0.63	0.02	0.54	0.02	0.91	0.05	0.60	0.02

Table 3
Difference between the minimum values of G_{int} for all amino sidechains and ions with each type of base, ribose, and phosphate for ssDNA and dsDNA

Negative values indicate that interaction with ssDNA is more favorable than with dsDNA.

	Ade Min	Ade Std	Cyt Min	Cyt Std	Gua Min	Gua Std	Thy Min	Thy Std	Pho Min	Pho Std	Sug Min	Sug Std	DNA Min	DNA Std
Ace	0.27	0.62	-0.72	0.36	0.04	0.35	-0.44	0.13	-0.05	0.17	-0.11	0.76	-0.31	0.22
Ala	-0.36	0.11	-0.34	0.14	-0.34	0.17	-0.20	0.07	-0.15	0.07	-0.31	0.07	-0.36	0.05
Cys	-0.34	0.15	-0.48	0.09	-0.01	0.24	-0.36	0.14	-0.16	0.07	-0.26	0.10	-0.41	0.06
Asp	-0.57	0.40	-0.59	0.26	-1.62	0.49	-0.34	0.15	-0.14	0.06	-0.33	0.10	-0.49	0.15
Glu	-0.31	2.60	-0.78	0.25	-1.36	0.43	-0.65	0.28	-0.09	0.09	-0.30	0.16	-0.42	0.25
Phe	-1.36	0.27	-0.98	0.29	-1.25	0.28	-1.71	0.15	-0.36	0.06	-0.88	0.28	-1.35	0.15
His	0.29	0.24	-0.04	0.28	0.31	0.39	-0.47	0.34	-0.09	0.07	-0.08	0.09	-0.19	0.15
Hip	1.70	0.35	0.45	0.31	1.85	0.15	2.08	0.36	0.02	0.10	1.06	0.29	0.58	0.17
Ile	-0.92	0.28	-1.19	0.24	-0.77	0.12	-0.87	0.21	-0.31	0.13	-0.53	0.17	-0.95	0.13
Lys	1.01	0.63	-0.06	0.82	1.22	0.13	1.05	0.33	-0.04	0.09	0.14	0.13	0.25	0.11
Leu	-0.87	0.28	-0.96	0.19	-0.86	0.47	-0.81	0.14	-0.46	0.13	-0.77	0.24	-1.02	0.14
Met	-0.91	0.14	-0.94	0.30	-0.73	0.29	-1.20	0.13	-0.42	0.07	-0.72	0.08	-1.07	0.09
Asn	0.52	0.22	0.02	0.23	0.30	0.20	-0.48	0.32	-0.07	0.06	0.23	0.25	-0.12	0.07
Gln	0.37	0.29	0.31	0.40	0.79	0.47	-0.14	0.30	-0.11	0.09	0.39	0.20	-0.02	0.17
Arg	1.56	0.27	0.23	0.41	1.85	0.34	1.50	0.37	-0.30	0.09	0.76	0.40	0.24	0.19
Ser	0.09	0.06	-0.10	0.17	0.36	0.18	-0.05	0.08	-0.07	0.06	-0.04	0.08	-0.03	0.06
Thr	0.11	0.27	-0.22	0.12	0.28	0.14	-0.15	0.08	-0.08	0.11	-0.04	0.10	-0.11	0.07
Val	-0.62	0.27	-0.79	0.17	-0.56	0.17	-0.59	0.22	-0.26	0.11	-0.42	0.15	-0.67	0.11
Trp	-0.06	0.46	-1.05	0.41	-0.80	0.36	-1.18	0.33	-0.57	0.28	-0.86	0.26	-1.23	0.25
Tyr	-0.64	0.65	-1.04	0.76	-0.28	0.25	-1.12	0.76	-0.59	0.21	-0.38	0.37	-1.04	0.29
Na ⁺	0.40	0.02	-0.83	0.05	-0.13	0.03	-0.23	0.09	-0.06	0.01	0.46	0.02	0.40	0.02
Cl ⁻	-0.85	0.15	-0.63	0.11	-1.15	0.09	-0.61	0.04	-0.04	0.02	-0.28	0.06	-0.85	0.15

Table 4
Correlation coefficients of base preferences for each amino acid sidechain

Each column records the Pearson correlation coefficient, R_{corr} , for a different pair of data-sets. Column ssMD:HINT records R_{corr} for the four G_{int} values computed here for bases in ssDNA with the negative of the Mean HINT-based scores reported for the same bases in Table 2 of Marabotti et al.¹⁷ Column dsMD:HINT records the same but for the four G_{int} values computed here for bases in dsDNA. Column dsMD:HINT records R_{corr} for the four G_{int} values computed here for dsDNA with the negative of the scoring matrix entries reported in Table 2 of Mandel-Gutfreund & Margalit.¹¹ Column dsMD: Gao records R_{corr} for the four G_{int} values computed here for dsDNA with scores reported by Gao & Skolnick⁵⁷ (downloaded from <http://cssb.biology.gatech.edu/skolnick/files/DBD-Hunter/dppp.com>); for purines, scores were computed as the sum of the two reported values for the imidazole and pyrimidine rings (results were slightly worse when the two values were averaged instead). Column "Ave." records the average of the six R_{corr} values shown in each row (excluding that of ssMD:HINT). Column "Rank" expresses these average values in rank order with the highest R_{corr} value being assigned rank 1.

	ssMD:HINT	dsMD:HINT	dsMD:Mandel	dsMD: Gao	HINT:Mandel	HINT:Gao	Mandel:Gao	Ave.	Rank
Ala	0.53	0.14	0.98	0.86	-0.05	0.43	0.77	0.52	4
Cys	-0.25	0.90	-0.51	-0.88	-0.83	-0.61	0.09	-0.31	17
Asp	0.10	0.63	0.79	0.62	0.90	0.25	0.64	0.64	3
Glu	-0.25	0.16	0.11	-0.76	0.98	0.20	0.35	0.17	9
Phe	0.74	0.03	-0.32	-0.29	-0.81	-0.37	0.84	-0.16	13
His	-0.49	0.92	0.49	-0.06	0.10	-0.19	0.32	0.26	5
Ile	-0.23	0.07	0.38	-0.34	0.49	0.00	0.63	0.21	7
Lys	0.19	0.86	0.74	0.79	0.74	0.68	0.18	0.66	2
Leu	-0.66	0.00	0.96	0.54	-0.27	-0.80	0.75	0.20	8
Met	-0.83	0.32	-0.98	-0.61	-0.47	-0.90	0.70	-0.32	18
Asn	-0.64	0.53	0.31	-0.71	0.49	-0.23	0.45	0.14	11
Gln	-0.20	0.26	0.09	0.09	0.20	-0.80	-0.65	-0.14	12
Arg	-0.20	0.93	0.84	0.85	0.74	0.99	0.63	0.83	1
Ser	0.74	0.89	0.84	-0.50	0.52	-0.83	0.03	0.16	10
Thr	-0.62	0.96	-0.82	-0.89	-0.69	-0.83	0.58	-0.28	16
Val	-0.32	0.77	0.46	-0.24	0.12	-0.44	0.75	0.24	6
Trp	-0.16	-0.52	-0.27	-0.91	-0.64	0.69	0.11	-0.26	15
Tyr	-0.86	0.35	-0.69	-0.31	-0.72	-0.51	0.87	-0.17	14

	ssMD:HINT	dsMD:HINT	dsMD:Mandel	dsMD:Gao	HINT:Mandel	HINT:Gao	Mandel:Gao	Ave.	Rank
Ave.	-0.19	0.46	0.19	-0.15	0.05	-0.18	0.45	0.13	