JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

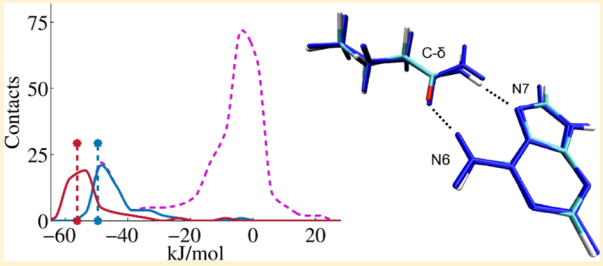
Large-Scale Quantitative Assessment of Binding Preferences in Protein−Nucleic Acid Complexes

Dávid Jakubec,[†] Jiří Hostaš,[†] Roman A. Laskowski,[‡] Pavel Hobza,[†] and Jiří Vondrášek*[,†]

[†]Institute of Organic Chemistry and Biochemistry, Flemingovo náměstí 2, Prague 6, 160 10, Czech Republic

[‡]EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, U.K.

ABSTRACT: The growing number of high-quality experimental (X-ray, NMR) structures of protein−DNA complexes has sufficient enough information to assess whether universal rules governing the DNA sequence recognition process apply. While previous studies have investigated the relative abundance of various modes of amino acid−base contacts (van der Waals contacts, hydrogen bonds), relatively little is known about the energetics of these noncovalent interactions. In the present study, we have performed the first large-scale quantitative assessment of binding preferences in protein−DNA complexes by calculating the interaction energies in all 80 possible amino acid−DNA base combinations. We found that several mutual amino acid−base orientations featuring bidentate hydrogen bonds capable of unambiguous one-to-one recognition correspond to unique minima in the potential energy space of the amino acid−base pairs. A clustering algorithm revealed that these contacts form a spatially well-defined group offering relatively little conformational freedom. Various molecular mechanics force field and DFT-D *ab initio* calculations were performed, yielding similar results.



1. INTRODUCTION

Protein−DNA interactions are critical for important cellular processes such as DNA replication, DNA repair and cell cycle regulation.[1] In eukaryotes, nonspecific protein−DNA association is ubiquitous, as the genetic information is packaged with histone proteins in nucleosomes.[2] For other phenomena, such as the regulation of gene expression, specific DNA sequence recognition with high fidelity is required.

Understanding the rules governing the recognition process would constitute a major accomplishment in the fields of computational biology and bioinformatics. Libraries containing protein DNA-binding motifs and their corresponding recognized DNA sequence patterns have been generated,[3,4] while the continually increasing amount of structural data has opened space for theoretical studies.[5] Despite these ongoing efforts, no unanimous recognition code applicable to all protein families has been described to date (for review, see ref 6). Notable exceptions include zinc finger proteins and transcription activator-like effector (TALE) proteins, whose DNA-binding domains can be engineered to target a specific DNA sequence according to a simple amino acid−nucleotide code.[7]

Two principal modes of specific sequence recognition have been described based on analysis of a large number of experimental structures.[6] Base readout involves direct interactions between a protein DNA-binding domain and the target DNA sequence, typically in the form of hydrogen bonds. The possibility of specific base pair recognition in the major groove by amino acids forming two hydrogen bond contacts was first explored by Seeman based on early experimental data[8] and still constitutes an important tool guiding the prediction of protein−DNA binding sites. A second factor contributing to the specificity of sequence recognition is the DNA shape readout.[6] Local deviations from the B-form have been observed in many protein−DNA complexes, and the propensity to adopt different conformations (e.g., kinks) has been shown to be sequence-dependent.[9] Nonlocal effects, such as an overall bend of the DNA double strand, can enable the formation of interactions that would be impossible in the B-form.[10−12]

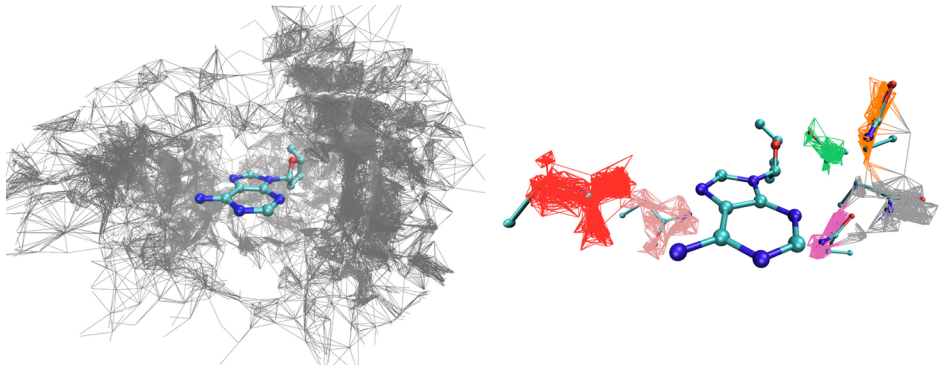
Both modes of sequence recognition are utilized to some extent in the majority of specific protein−DNA interactions.[6] While the nonlocal effects are difficult to quantify, numerous studies have attempted to elucidate the recognition process by investigating small-scale interactions between amino acids and DNA bases. These often draw conclusions from the comparison of experimental structural data with theoretical models.[5,13,14] Few of them, however, involve large-scale investigation of the physicochemical parameters of the protein−DNA interface.[15]

In the present work, we have for the first time quantitatively examined the protein−DNA interactions by calculating the interaction energies for all 20 × 4 amino acid−DNA base pair combinations. Our analysis was performed on all available protein−DNA complexes and therefore draws conclusions that are not specific to any single protein family or DNA binding motif.

Calculations were performed at the molecular mechanics (MM) level utilizing empirical force fields presently used in the

Received: December 22, 2014

DOI: 10.1021/ct501168n
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX



(a) Distribution (b) Clusters

Figure 1. Example distribution and its associated identified clusters for the adenine−asparagine pair. Ball-and-stick representations within clusters in part b are cluster representatives.

realm of protein−DNA interactions. The presented set of amino acid−nucleic acid base contacts offers a unique chance to investigate the performance of the MM methods. We compared the MM results with results from reliable quantum mechanical (QM) methods. The QM methods should adequately describe not only the dominant hydrogen bonding but also electrostatic (in the charged complexes) and dispersion interactions. The dispersion energy in particular plays an important role in any biomolecular complex, and its proper description is of primary importance.[16]

Because we intend to enlarge the complexes investigated (to include interactions with sugar−phosphate backbone elements) in the future, we must use as our reference a method that could be used even for these considerably larger structures. The density functional theory method augmented with empirical dispersion term (DFT-D), which when combined with extended basis sets provides accurate description of virtually any interaction motifs existing in biomolecular complexes, represents a natural choice.[17] On the other hand, a high-level QM technique, such as the coupled cluster method covering single and double excitations iteratively and triple excitations perturbatively (CCSD(T)) using complete basis set (CBS), is needed to describe binding motifs for which the DFT-D method can fail (e.g., charge transfer complexes[18]).

2. COMPUTATIONAL DETAILS

2.1. Data Set. Our data set consisted of 50,205 nucleotide−amino acid pairs extracted from the Protein−DNA interaction atlas (found at http://www.ebi.ac.uk/thornton-srv/databases/sidechains/) generated according to the method described by Luscombe et al.[5] and updated as of March 2014. This atlas contains a total of 1,569 unique structures of protein−DNA complexes from the Protein Data Bank (PDB). Only structures solved by X-ray crystallography to a resolution higher than 2.5 Å and containing nucleic acids consisting of at least 4 base pairs were considered. For each of the four DNA base types, all contacts with each of the 20 amino acids were extracted. A contact was defined where any amino acid side chain atom was within 4.5 Å of any DNA base atom (excluding the DNA sugar−phosphate backbone atoms), as in previous work.[19] This gave a total of 80 distributions, one for each possible nucleotide−amino acid pair. A common frame of reference, centered at the DNA base, was used for each distribution. Figure 1a shows an example distribution for the adenine−asparagine pair.

The amino acids in these distributions tend to cluster in 3D relative to the bases. The sizes and locations of the clusters depend on the interactions the amino acids make with the relevant base and which regions of 3D, relative to the base, are accessible when the base is in the DNA double helix. To identify the largest clusters in each distribution, and pick out a representative amino acid for each, we calculated the RMSD scores between every pair of amino acids in each distribution. Only the three atoms defining the reference frame of the amino acid side chain[5] were used for the RMSD calculations. The amino acid having the largest number of neighbors within RMSD of 1.5 Å was taken to be a cluster representative; its neighbors were taken to be its cluster. After removing the largest cluster, the process was repeated to find the next-largest cluster, and so on until 6 clusters had been identified or the clusters were too small to be significant. Significance was judged by randomly rearranging the amino acids of each distribution and calculating the typical cluster sizes that would arise by chance.

The 80 distributions yielded 272 clusters comprising 13,155 nucleotide−amino acid pairs. The six clusters identified in the example adenine−asparagine distribution are shown in Figure 1b.

Our data set contained a number of complexes of identical proteins interacting with different DNA fragments. To reduce the bias such duplication might introduce, we removed any duplicate nucleotide−amino acid pairs arising from these identical protein complexes. We did not address the bias that might result from nonidentical but homologous protein in our data set (see Discussion).

The removal of duplicate entries brought the final number of contacts within clusters to 4,014 nucleotide−amino acid pairs.

2.2. Interaction Energy Calculations. The presented sets of structures (distributions, clusters, cluster representatives) constitute a hierarchy in which each subsequent set contains an order-of-magnitude fewer contacts than the preceding set. While accurate interaction energy investigations were our primary interest, high-quality QM calculations involving all distributions or even cluster members are currently computationally off limits. We thus had to make a compromise between accuracy and computational demands when choosing the method to use.

The set of cluster representatives involves amino acid−base contacts containing all binding motifs found in biomolecular complexes (van der Waals contacts, hydrogen bonding,

DOI: 10.1021/ct501168n
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

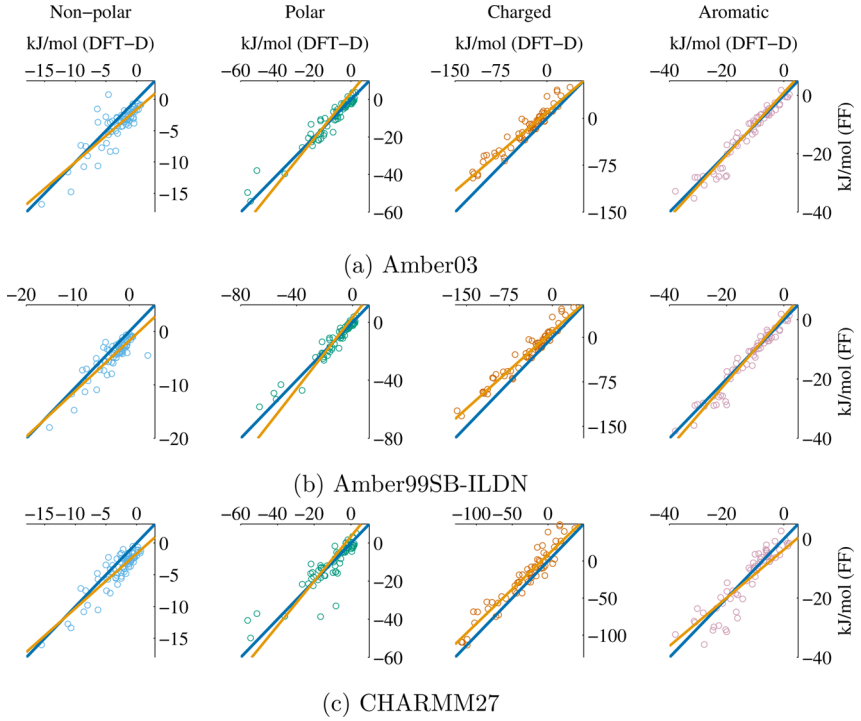


Figure 2. Correlations of DFT-D/B3LYP-D3/def2-TZVPP and force field gas phase interaction energies for cluster representatives featuring various types of amino acids. The blue line has a slope of 1; the orange line corresponds to linear regression.

electrostatic interactions). Rather than sacrifice accuracy, we decided to limit the scope of rigorous electronic structure calculations to contacts involved in this set.

MM methods, on the other hand, allow for very quick interaction energy calculation by the evaluation of a simple potential energy function. As such, the investigation of all contacts at all levels of the aforementioned hierarchy is feasible. Various sets of atomic parameters (force fields) have been developed for MM calculations; the relative performance of three was tested in our work. Nucleic acid base parameters taken from the Amber94 force field (ff94) were combined with amino acid parameters from the Amber99SB-ILDN (ff99SB-ILDN) and Amber03 (ff03) protein force fields.[20] Amber99SB-ILDN improves on the earlier Amber99 and Amber99SB force fields[21,22] by improving isoleucine, leucine, aspartate, and asparagine side chain torsional potentials.[23] Amber03 uses partial charges based on the protein dielectric environment.[24] The CHARMM27 nucleic acid force field uses amino acid parameters taken from protein-oriented CHARMM22.[25,26] A common feature of all these force fields is that they were designed with solvent−solute interactions in mind and parametrized to reflect condensed phase properties. As such, their performance concerning gas phase calculations is not predictable, and the aforementioned electronic structure calculations are required for comparison (see Discussion).

2.2.1. Empirical Force Fields. For each nucleotide−amino acid pair, we prepared a C_α representation of the amino acid by replacing the peptide bond carbonyl and amide groups with hydrogen atoms as described by Berka et al.[19] This procedure capped each standard amino acid side chain with a methyl group and thus eliminated nonspecific interactions between the DNA base and the protein backbone. Proline was modeled as a neutral tetrahydropyrrole and histidine was protonated on ε-N in all contacts. Only the base from each nucleotide was retained with the *N*-glycosidic bond deoxyribose C1′ carbon replaced with a hydrogen atom. All hydrogens were added with a custom Chimera script.[27]

The C_α representation of each amino acid, as well as the modified lone DNA bases, were added to the aforementioned force fields. Atom types of the added C_α hydrogens were HC and HA for the Amber class and CHARMM27 force fields, respectively. Hydrogen atoms added to the proline nitrogen were of type H in all force fields, and the hydrogens replacing the sugar C1′ atom in DNA bases were of type H in Amber and H2N in CHARMM27 force fields. Partial atomic charges were assigned equally to each of the added C_α hydrogen atoms to keep the overall charge of the amino acid an integer: +1 for lysine and arginine, −1 for glutamate and aspartate, and 0 for the rest. By creating the C_α representation, four amino acids obtained symmetry. We did not recalculate the charges of the original hydrogen atoms for alanine, valine, and proline, but we reflected this issue for glycine, in which all four hydrogen atoms were made the same. This was due to the fact that the properties of glycine would differ drastically depending on its orientation with respect to the nucleotide. The partial atomic charge of the added N1 or N9 hydrogen atoms in DNA bases was calculated to retain the null overall charge.

All energy calculations and optimizations were performed using a double-precision version of the GROMACS 4.5.5 package in the gas phase.[28] First, we optimized the hydrogen atom positions in the dimer while keeping the heavy atoms fixed. A single point energy was then calculated on this optimized dimer. The base and the amino acid were then split, and each had its hydrogens optimized, again leaving the heavy atoms intact. The interaction energy was then calculated as the difference between the single point energy of the dimer and the sum of the energies of the optimized monomers.

Unconstrained all-atom optimization of the cluster representatives was performed using MM. The interaction energies

DOI: 10.1021/ct501168n
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

of the fully optimized dimers were then evaluated as described above, with full optimizations replacing the constrained ones.

2.2.2. DFT-D. The DFT-D interaction energy calculations were performed with the TURBOMOLE V6.5 package[29] with the valence triple-ζ plus polarization (def2-TZVPP) basis set, B3-LYP functional, and Grimme's D3 dispersion correction without three body effects.[30] The input geometries of the dimers were the same as those used in the empirical force field calculations. The hydrogen atom optimizations were performed on the DFT-D/B3LYP-D3/def2-TZVPP level for both the interacting dimer, as well as for each individual base and amino acid, prior to the energy calculation. The heavy atoms were constrained to their original positions. The interaction energies were again calculated using the supermolecular approach, by subtracting the sum of the optimized monomer energies from the energy of the interacting dimer. The correction for basis set superposition error (BSSE) was not considered because the term proved to be small (under 4% for the def2-TZVPP basis set, see Discussion), as expected.[30] The interaction energies thus also include part of the relaxation energy due to hydrogen atom optimizations.

Furthermore, selected interaction motifs were investigated at the more accurate DFT-D/B3LYP-D3/def2-QZVP level. To estimate the accuracy of the DFT-D methods used, we calculated the CCSD(T)/CBS interaction energy for one of the guanine−glutamate pairs (see Discussion).

3. RESULTS

3.1. Comparison of Empirical Potentials and DFT-D for Interaction Energy Calculations. The relative correspondence of the interaction energies calculated with the Amber03, Amber99SB-ILDN, and CHARMM27 force fields and the DFT-D/B3LYP-D3/def2-TZVPP results is shown in Figure 2. The 272 cluster representatives were split into groups according to the physicochemical character of each amino acid, similar to previous work on amino acid side chains.[31] Seventy-six dimers comprised the set of nonpolar contacts (G, A, V, I, L, P), 69 were found in the polar set (T, S, N, Q, C, M), 64 were contained in the charged group (K, R, D, E), and 63 were aromatic (F, Y, W, H).

The standard deviations of the differences between DFT-D/B3LYP-D3/def2-TZVPP and MM interaction energies for these respective groups of amino acids are shown in Table 1. These values represent the absolute magnitude of discrepancy between the MM and DFT-D/B3LYP-D3/def2-TZVPP results.

The range of interaction energies observed is very large (about 200 kJ/mol), owing largely to the contacts involving charged and polar amino acid residues. Both the nonpolar and aromatic groups occupy a relatively narrow range of interaction energies (about 20 and 40 kJ/mol, respectively). For this reason, the coefficients of determination (R^2) describing the correspondence between the force field and DFT-D/B3LYP-D3/def2-TZVPP results in relative terms are included after each standard deviation in Table 1. As can be seen, the two Amber force fields yield very similar results, while CHARMM27 deviated from the former when applied to contacts involving polar and aromatic amino acids.

Table 1. Corrected Sample Standard Deviations of the Difference between Force Field and DFT-D/B3LYP-D3/def2-TZVPP Gas Phase Interaction Energies in [kJ/mol] and the R^2 Coefficients of Determination (in Parentheses) between the Force Field and DFT-D/B3LYP-D3/def2-TZVPP Results for Various Sets of Amino Acids

	nonpolar	polar	charged	aromatic
ff03	1.8 (0.69)	9.6 (0.84)	11.9 (0.94)	4.8 (0.84)
ff99SB-ILDN	1.6 (0.76)	9.6 (0.84)	11.2 (0.95)	4.3 (0.88)
CHARMM27	1.5 (0.78)	12.2 (0.73)	12.2 (0.93)	6.9 (0.61)

The blue lines in Figure 2 have a slope of 1 and represent an absolute correspondence between the force field and DFT-D/B3LYP-D3/def2-TZVPP results. Any points found above this line represent structures that were calculated to have a higher stabilization energy using DFT-D/B3LYP-D3/def2-TZVPP than the particular MM force field. The orange regression line shows the drift from the absolute correspondence that is found each for set of contacts. The interaction energies of contacts involving charged amino acids were in general found to be more stabilizing using DFT-D/B3LYP-D3/def2-TZVPP than any of the force fields tested, while the opposite is true for dimers involving nonpolar amino acids. The sets involving polar and aromatic amino acids do not display any such behavior.

The trends of over- or underestimating the interaction energies for structures containing the aforementioned sets of amino acids are shared between all tested force fields. Based on these observations, we conclude that despite the systematic shifting of interaction energies for some groups of amino acids, the correspondence of force field and DFT-D/B3LYP-D3/def2-TZVPP results can be viewed as very good. Tables containing the interaction energies of all cluster representatives in all 20 × 4 amino acid−base pairs can be found at http://pdna-iea.uochb.cas.cz/.

3.2. Interaction Energy Distributions. An analysis of the interaction energy profiles of clusters asociated with a certain amino acid−base pair reveals that the cluster representative indeed represents the most typical interaction energy value found within the members of its associated cluster (Figure 9). This observation is in agreement with the findings of Berka et al. concerning amino acid side chain−side chain interactions.[19]

Most of the fully optimized dimers are significantly more stable than those resulting from either a constrained force field or DFT optimizations, with the stabilization energy increasing occasionally by as much as 120 kJ/mol in pairs involving charged amino acids. This extreme behavior can be expected due to the calculations being performed in the gas phase. An increase of the stabilization energy by at least 5 kJ/mol was observed for all pairs.

The interaction energy profiles for all clusters calculated with all tested force fields, along with the corresponding DFT-D/B3LYP-D3/def2-TZVPP cluster representatives' interaction energy values and the positions of the fully optimized structures, can be found at http://pdna-iea.uochb.cas.cz/.

Next, we turned our attention to comparison of interaction energy profiles for all clusters of the same amino acid−base pair type to the energy profile of the entire corresponding distribution. We observed that some clusters shared the following properties: 1) the cluster consists of contacts with the highest stabilization energy found in the energy profile of the associated distribution, 2) the cluster contains a significant part of all the contacts within that interaction energy range, and 3) the peak corresponding to the cluster-associated part of the distribution is clearly distinguished from the rest of the energy profile. These conditions were met by a single cluster occurring in the adenine−asparagine, adenine−glutamine, adenine−

DOI: 10.1021/ct501168n
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

lysine, cytosine−asparagine, and cytosine−tyrosine energy distribution profiles and by two clusters in the guanine−glutamine distribution. An example of such an energy profile calculated using Amber03 is shown in Figure 3. All but two guanine−glutamine clusters are the largest found for the particular base−amino acid pair.

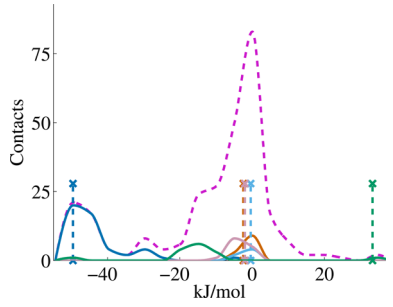


Figure 3. Amber03 interaction energy profile of the adenine−glutamine distribution (dashed purple curve), together with associated clusters (blue, orange, light blue, pink, and green solid lines) and cluster representatives (dashed vertical lines, color corresponds to the color coding of each particular cluster). Areas under curves correspond to the number of contacts in clusters or in the distribution.

The interaction energy profiles of all distributions calculated with all tested force fields can be viewed at http://pdna-iea.uochb.cas.cz/. It is important to note that the positions of cluster energies relative to whole distribution energy profiles remain unchanged when comparing results from the Amber03, Amber99SB-ILDN, and CHARMM27 force fields.

Focusing on the cluster representatives of the these energetically low lying clusters, we compared their structures after constrained optimization to fully optimized dimers (Figures 4 and 5).

Examination of the adenine−asparagine, adenine−glutamine, and the guanine−glutamine contacts revealed common features: the presence of two hydrogen bonds and the proximity of the geometries found in biomolecules to their respective local energy minima. A stabilization energy increase of 5−7 kJ/mol in all these pairs is achieved by parallelization of the planes of the respective hydrogen bond donors and acceptors. The two contacts involving adenine show that the amino acid binds to the major groove of the DNA molecule by two hydrogen bonds involving the amino acid β- or γ-amide group and the amino group and N7 atom of the base (Figures 4a and 4b). In contrast, the glutamine in Figure 4c recognizes guanine in the minor groove, with two hydrogen bonds between the amide group and the amino group and N3 atom of the base. Figure 4d shows an interesting contact in which the glutamine residue disrupts the normal Watson−Crick pairing of guanine with cytosine. The glutamine γ-amide group oxygen acts as an acceptor of two hydrogen bonds involving the N1 atom and the amino group of guanine. A closer look at the PDB structures featuring this contact revealed that it originates from complexes of DNA methyltransferases. These are predominantly mutants of HhaI methyltransferase, which can flip the to-be-methylated cytosine out of the DNA helix, replacing the canonical base pairing with the aforementioned guanine−glutamine contact.[32] This pairing also features the protein main chain NH group as a third hydrogen bond donor.

The remaining cytosine−asparagine, cytosine−tyrosine, and adenine−lysine pairs show more pronounced changes in geometry after full optimization. The cytosine−asparagine contact (Figure 6a) features a single hydrogen bond between the amide group of the amino acid and O2 oxygen of the base. The full optimization of this structure does not lead to a two

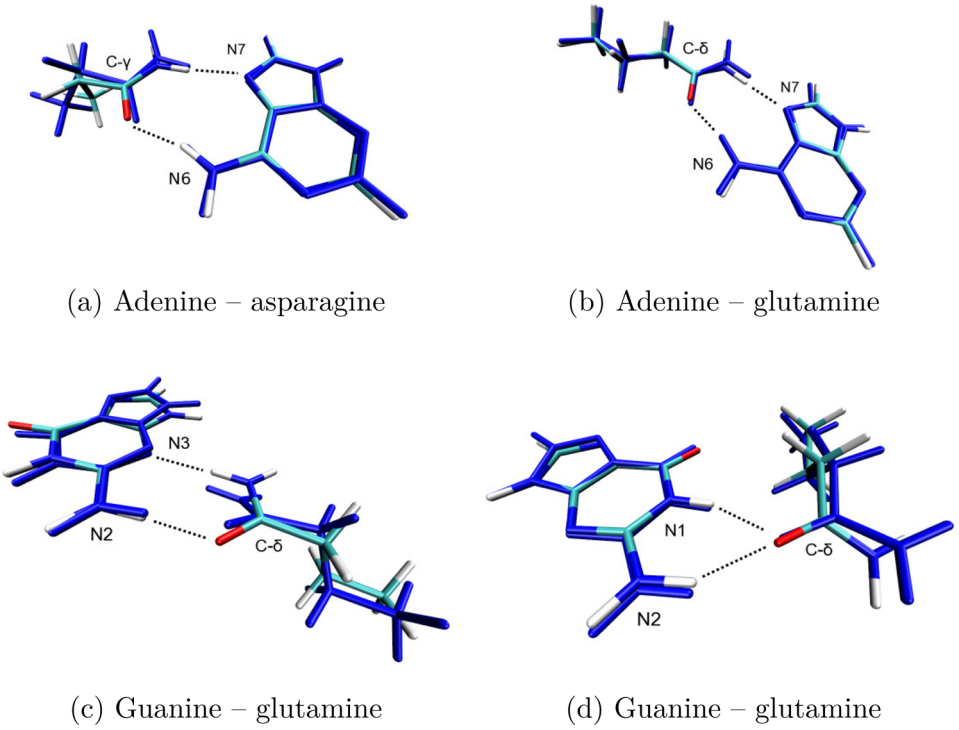
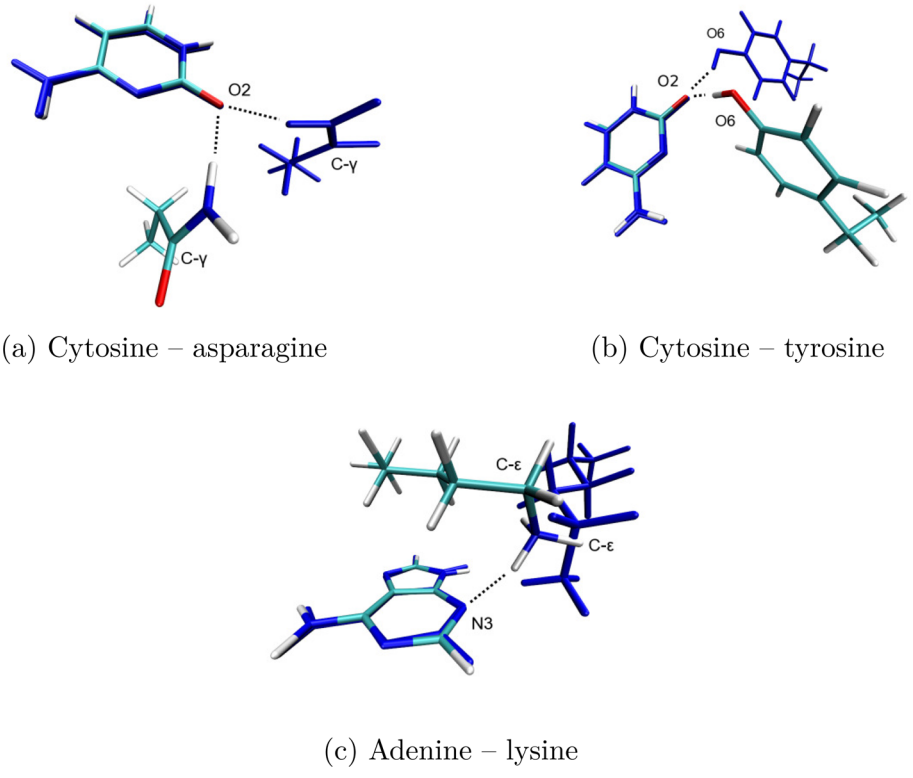


Figure 4. Cluster representatives of distinct low-lying clusters - contacts featuring two hydrogen bonds. Constrained optimizations (blue) and fully optimized structures.

DOI: 10.1021/ct501168n
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX



(a) Cytosine – asparagine (b) Cytosine – tyrosine

(c) Adenine – lysine

Figure 5. Cluster representatives of distinct low-lying clusters: contacts featuring a single hydrogen bond. Constrained optimizations are shown in blue along with the fully optimized structures.

hydrogen bond conformation and instead shifts the asparagine from the minor groove to an unnatural conformation that conflicts with both normal Watson–Crick pairing and the preceding base pairs. It is worth noting that in biomolecules, the cytosine–asparagine contact cannot adopt a two hydrogen bond conformation without interfering with normal base pairing.

The behavior of the cytosine–tyrosine pair is remarkably similar to the previous case (Figure 6b). A single hydrogen bond is formed in the minor groove between the donor phenolic hydroxyl group and the acceptor O2 atom on cytosine. The full optimization of this contact shortens the hydrogen bond by 0.2 Å and shifts tyrosine into a position that conflicts with normal base pairing. This increases the stabilization energy by approximately 11 kJ/mol.

The adenine–lysine contact features a single hydrogen bond between the ε-amino group and N3 atom of the base (Figure 6c). Full optimization of this pair results in a geometry that could not be adopted in biomolecules without heavily distorting the structure of DNA. The aliphatic chain of lysine maximizes the van der Waals contact with the base and positions the C_α deeply within the DNA helix. Because the charged lysine is also likely to interact with the phosphate backbone, we conclude that this contact cannot be treated appropriately using our C_α model and without the context of surrounding residues.

4. DISCUSSION

Our work on elucidating the selectivity of protein–DNA interactions involved several simplifications and without doubt describes only part of the problem. We have so far focused only on the interactions between DNA bases and amino acid side chains. We justified this separation on the basis that the selectivity of binding of the protein to a designated region of DNA should be conveyed by the sequence of the nucleic acid bases. The focus on the pairs consisting of one amino acid and one DNA base immediately removes from the data set the undoubtedly important interaction motifs involving base pairs, multiple base steps, and water-mediated contacts.

While the local topology of the DNA molecule has been known to be responsible for the specificity of some protein–DNA interactions (for example, the *trp* repressor/operator complex[11]), the physicochemical parameters of the sugar–phosphate backbone remain the same regardless of the base sequence and thus play little role in our one side chain–one base correspondence study. It should be noted, however, that to fully describe DNA sequence recognition, the strong electrostatic interactions between charged amino acids and the nucleic acid backbone must be taken into account.[6] These can either aid or disrupt the complex formation, depending on the character of the particular amino acid residues found at the protein–DNA interface.[11] For example, the increased negative electrostatic potential in the minor groove caused by the proximity of the phosphate groups can guide the specific binding of arginine residues.[33] Indeed, we plan to investigate the interactions between a set of amino acids and the sugar–phosphate backbone as a logical next step.

Our approach describes the energetics of binding of proteins to DNA as a sum of pairwise additive terms corresponding to individual amino acid–nucleotide pairs. It has been shown that neglecting the nonadditive contributions resulting from the influence of surrounding DNA bases or amino acid residues is a reasonable first approximation.[34] It should be kept in mind that many-body terms are much more important for polar or charged structures than for nonpolar complexes. An accurate description of many-body terms in biomolecular complexes, in

DOI: 10.1021/ct501168n
J. Chem. Theory Comput. XXXX, XXX, XXX–XXX

which dispersion energy is always important, is tedious as it requires performing calculations at the MP3 level or higher (as the three body dispersion energy is covered only at these levels (e.g., CCSD(T)).

Another important simplification is that all interaction energies were calculated in the gas phase. Contacts involving charged amino acids had the most extreme interaction energies in our models, and these did indeed decrease severely in a simulated ε = 80 implicit solvent environment. The choice of dielectric constant is open to question as the protein−DNA interface constitutes a highly nonhomogeneous environment.[35] Moreover, the entropic terms introduced by the interaction with the solvent are only partially considered, being effectively included in the implicit solvent model and force field parameters. These would certainly play a role in the interactions of DNA bases with nonpolar and aromatic amino acids.[36] Similarly, we ignored the conformational entropy of the binding partners, as this is a nonadditive effect, the calculation of which would have to involve knowledge of the dynamics of the entire protein or DNA molecule.[37]

Our work was burdened by several intrinsic deficiencies of classical force fields. Induction and polarization effects are known to play a major role in biomolecular interactions, providing, of course, that one of the interacting partners is highly polar or charged. These effects are not explicitly described for any of Class I force fields used.[38] Instead, they are implicitly included in the overestimated electrostatic term. Moreover, the force fields have been parametrized to reflect the behavior of biomolecules in a solvent environment.[20,21,24−26]

A high-quality (e.g., CCSD(T)/CBS) study involving cluster representatives is necessary to set a state-of-the-art benchmark against which our current results can be compared.[18] These and similar (MP2.5/CBS) calculations will be part of a subsequent study.

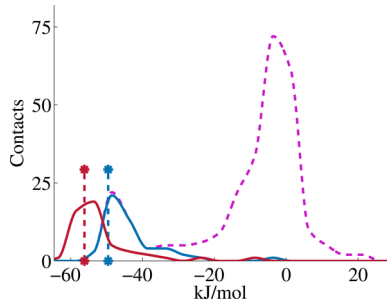


Figure 6. Amber03 adenine−glutamine interaction energy profile (dashed purple curve) and a distinct low-lying cluster (solid blue curve). Crimson solid profile - DFT-D/B3LYP-D3/def2-TZVPP cluster interaction energies; dashed vertical lines - cluster representatives (color coding based on the particular method used). Areas under the curves correspond to the number of contacts found in the clusters or in the distribution.

We verified our observations concerning the distinct low-lying clusters by recalculating all cluster-associated contacts using the DFT-D/B3LYP-D3/def2-TZVPP method. The respective interaction energy profiles reproduced the empirical results very well, despite being shifted by a few kJ/mol toward more negative values (Figure 6). Using the larger def2-QZVP basis set improved the agreement between QM and empirical results (not shown).

To estimate the discrepancy between the presented def2-TZVPP and def2-QZVP results and the benchmark CCSD(T)/CBS values we present the interaction energies for the most stable guanine−glutamate pair calculated by these methods. The CCSD(T)/CBS interaction energy (−169.5 kJ/mol) where BSSE was covered by counterpoise procedure differs only marginally (by 4% and 2%, respectively) from the def2-TZVPP (−175.7 kJ/mol) and def2-QZVP (−171.1 kJ/mol) values which, as mentioned in the Introduction, have not been corrected for BSSE. The BSSE values for this complex were evaluated for both basis sets and were found to be less than 4% of the interaction energy in each case.

The DFT-D/B3LYP-D3/def2-TZVPP and DFT-D/B3LYP-D3/def2-QZVP methods yielded higher stabilization energies for complexes involving charged amino acids compared to the MM results. The only way to decide which of the methods is reasonable we have to perform calculations at a benchmark level — CCSD(T)/CBS, at which charge transfer energies are properly covered.[18]

The comparison of DFT-D/B3LYP-D3/def2-TZVPP and Amber03 interaction energy profiles of all distinct low-lying clusters can be viewed at http://pdna-iea.uochb.cas.cz/.

It is also worth noting that the hydrogen atom optimizations by DFT-D/B3LYP-D3/def2-TZVPP and force fields in general yielded different geometries. The Amber force fields hold the amino groups of DNA bases planar, while the MP2/6-31G**ab initio* calculations correctly show an almost sp^3 hybridization of the amino group nitrogen atom in the gas phase.[39] The penalization for the out-of-plane bending in the force field energy scoring function is enormous. This is illustrated by the presence of hydrogen-bonded complexes in which the heavy atom distance is unrealistically low (below 2.3 Å). We identified three clusters whose representatives meet this criterion. Low-barrier hydrogen bonds characterized by short heavy atom distance have been described for some enzymatic reactions and may be involved in a tight transition state or in substrate binding.[40] More recent work has shown that short hydrogen bonds are ubiquitous within protein structures.[41] The force field interaction energies calculated on these complexes reach extreme values as a result of the steep Lennard-Jones repulsive term and show where the classical force fields fail. To remedy these shortcomings, we propose that either new sets of parameters allowing for greater flexibility of biomolecules should be introduced or the polarization effects must be taken into account. The DFT-D/B3LYP-D3/def2-TZVPP hydrogen atom optimizations, on the other hand, yielded geometries corresponding to the right hybridization states, while providing reasonable interaction energies.

Luscombe et al. investigated a set of 129 protein−DNA complexes and compared the populations of various types of amino acid−base contacts to those obtained by random docking.[5] The pairs were separated into groups featuring hydrogen bonds, van der Waals contacts, and water-mediated interactions. They found that the formation of two hydrogen bonds in guanine−arginine, guanine−lysine, adenine−asparagine, and adenine−glutamine pairs is a universal mechanism of sequence recognition utilized by many protein families. For all of these pairs, the number of observed contacts greatly exceeds their respective expected populations, leading to the conclusion that a one-to-one amino acid−base correspondence is a valid model of sequence recognition for some contacts.[5] We adopted this view and estimated the importance of various amino acid−base contacts for specific pairing on the basis of interaction

DOI: 10.1021/ct501168n
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

energies. We confirm the previous findings concerning adenine−asparagine and adenine−glutamine pairs, adding that the geometries found in biomolecules (Figures 4a and 4a) correspond to unique energy minima.

On the other hand, the distribution energy profile of the guanine−arginine pair shows that there are geometries isoenergetic to that featuring two hydrogen bonds between the arginine guanidino group and guanine O6 and N7 atoms (Figure 7). Given the energetic favorability of this con-

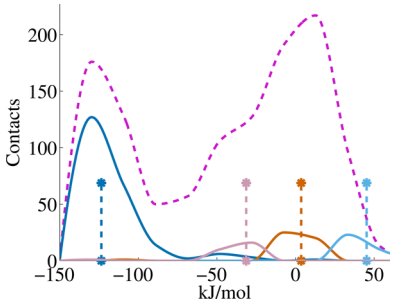


Figure 7. Amber03 guanine−arginine interaction energy profile: an example of the envelope surrounding a low-lying cluster. Color coding: see description in Figure 3.

formation, as well as the absolute number of observed guanine−arginine contacts, we do not challenge the selectivity of arginine toward guanine and instead propose that the energy minimum corresponding to this geometry allows for greater flexibility of the interacting partners compared to the aforementioned adenine pairs. Moreover, our cluster contains only contacts featuring the two η-N atoms of the guanidino group as hydrogen bond donors. A second conformation in which the guanidino ε-N and η-N act as hydrogen bond donors is possible, with interaction energies resembling the former.

Surprisingly, we did not find a cluster of guanine−lysine contacts featuring two hydrogen bonds between the ε-amino group and guanine O6 and N7 atoms. This is likely due to the fact that the positively charged lysine often interacts with the phosphate backbone, and the cluster identification was performed on whole nucleotides. The clusters identified in this distribution are nowhere near the interaction energy minima, as there are plenty of contacts adopting much more favorable conformations (Figure 8). This raises a question regarding whether a common RMSD criterion is applicable to all amino acids. We stated that the cluster interaction energy profile cannot have an envelope of other distribution members to be of functional significance. We consider the geometries of

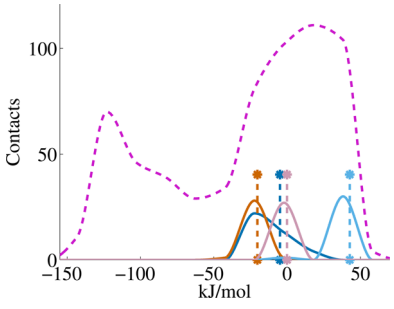


Figure 8. Amber03 guanine−lysine interaction energy profile, in which no clusters were identified at the low end of the energy distribution. Color coding: see description in Figure 3.

these low-lying clusters to be unique, as they are to represent only the mutual orientations of the amino acids and nucleotides that yield significantly more stabilizing interaction energies than the rest of the distributions. The envelopes deny this limiting orientational criterion, as they show that the part of the energy profile covered by the cluster can also be reached by a range of other orientations. If the clusters were not dense enough, however, the cluster definition algorithm would pick up only part of these contacts.

We therefore conclude that the contacts responsible for the specific binding of lysine to guanine have no preferred orientation of the aliphatic lysine side chain and share only the position of the ε-amino group participating in hydrogen bonding. This interaction is dominated by a spherically symmetric potential corresponding to the electrostatic term. There is therefore no reason why a specific orientation of the aliphatic chain relative to the base should be preferred. To describe this set of contacts, we propose the term energy-defined cluster, highlighting the isoenergetic nature of these pairs, while avoiding the usual spatial definition of clusters.

Amino acids capable of being single hydrogen bond donors (S, T, H, Y) can bind to any base in both the minor and major grooves and as such provide no inherent one-to-one amino acid−base selectivity.[8] Despite the lack of any observed preferences compared to a random set, Luscombe et al. dubbed the contacts involving these amino acids context-specific.[5] We identified three pairs (Figure 5: adenine−lysine, cytosine−asparagine, and cytosine−tyrosine) in which a statistically significant fraction of contacts is found in conformations corresponding to distinct energy minima. As these geometries describe the most favorable arrangement the respective pairs can adopt, we suggest that they may help distinguish between generic and context-specific contacts.

We found few discrepancies between the behavior of amino acid−base contacts and the results of Berka et al. concerning side chain−side chain interactions.[31] The interaction energy distribution for most pairs of amino acids consists of one well-defined peak, showing that similar interaction energies are found for many different pair geometries. On the other hand, the interaction energy profiles of many amino acid−base pairs have one or more peaks in addition to the main one, while others have no distinguishable peaks. Some of the amino acid−base interaction energy profiles have shapes which are greatly determined by the peaks corresponding to individual clusters (Figure 9). For pairs of amino acids, the cluster representative of the largest cluster often corresponds to a minimum of the

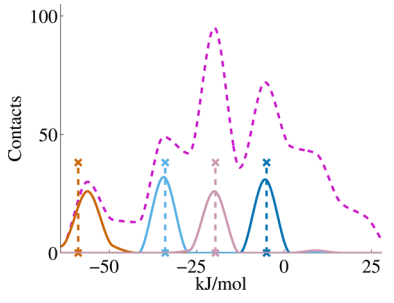


Figure 9. Amber03 adenine−lysine interaction energy profile: an example showing the effect of cluster intection energy profiles on the overall energy distribution. Note how the cluster representatives reflect the average interaction energy of their associated cluster. Color coding: see description in Figure 3.

DOI: 10.1021/ct501168n
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

interaction energy profile of its associated distribution. We observed no such behavior for amino acid−base pairs, in which the minimum of the interaction energy profile can belong to any of the clusters identified or even to no cluster at all. This leads us to conclude that multiple geometries corresponding to distinct energy minima can be adopted for various amino acid−base pairs, depending on the structural or functional needs of each particular protein−DNA complex.

The 1,569 unique PDB structures considered in this work do not address any homology issues other than 100% sequence identity. Removing contacts with 30% and greater sequence identity would drastically reduce the amount of structures. We are aware of the limited size of our data set and are cautious to draw any far-reaching conclusions.

While it remains true that the clusters often make up a significant part of the distributions, their absolute populations are often on the order of 10^1 and therefore quite susceptible to data set bias. This is in contrast to a statistical evaluation of side chain−side chain contacts in proteins,[31] in which the cluster populations were an order of magnitude higher. This is caused by both the relative lack of the protein−DNA crystal structures and the lesser amount of amino acid−nucleotide contacts per structure compared to contacts between amino acids within proteins. On the other hand, the relative populations of amino acid−base clusters merely reflect the current amount of high-quality protein−DNA complexes in the Protein Data Bank. Even if a unique geometry corresponding to a high stabilization energy is, as of this study, found only for a few contacts, it could very well be widespread in nature.

5. ASSOCIATED CONTENT

As previously mentioned, the interaction energies of all amino acid−base pairs as well as their representations (such as in the form of the interaction energy profiles) can be found at http://pdna-iea.uochb.cas.cz/. This site also hosts and provides the geometries of all investigated structures free of charge.

6. CONCLUSIONS

We have for the first time calculated the interaction energies for all amino acid−base pairs found in currently available protein−DNA complexes. We were able to quantitatively examine the underlying aspects of amino acid−base preferences originating from statistical studies.[5,14] We found that amino acid−base geometries capable of one-to-one amino acid−base recognition correspond to unique energy minima with interaction energies distinct from the rest of the distribution. Our observations are in good agreement with DFT-D/B3LYP-D3/def2-TZVPP *ab initio* calculations performed on a smaller set of structures. We plan to extend our data set to structures containing parts of the sugar−phosphate backbone as well as larger contacts spanning several bases. Finally, we plan to make our results more accessible by establishing a Web server providing information about the performance of various computational methods.

AUTHOR INFORMATION

Corresponding Author

*Phone: 420 220 183 267. Fax: 420 220 410 321. E-mail: jiri.vondrasek@uochb.cas.cz.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Education, Youth, and Sports of the Czech Republic (KONTAKT II programme LH11020) and the Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic (research project No. Z40550506). Access to computing and storage facilities provided by ELIXIR CZ and the National Grid Infrastructure MetaCentrum, administered under the programme "Projects of Large Infrastructure for Research, Development, and Innovations".

REFERENCES

(1) Jónsson, Z. O.; Hindges, R.; Hübscher, U. Regulation of DNA replication and repair proteins through interaction with the front side of proliferating cell nuclear antigen. *EMBO J.* **1998**, *17*, 2412−2425.

(2) Luger, K.; Mäder, A. W.; Richmond, R. K.; Sargent, D. F.; Richmond, T. J. Crystal structure of the nucleosome resolution core particle at 2.8 Å resolution. *Nature* **1997**, *389*, 251−260.

(3) Zhu, C.; Byers, K. J. R. P.; McCord, R. P.; Shi, Z.; Berger, M. F.; Newburger, D. E.; Saulrieta, K.; Smith, Z.; Shah, M. V.; Radhakrishnan, M.; Philippakis, A. A.; Hu, Y.; de Masi, F.; Pacek, M.; Rolfs, A.; Murthy, T.; LaBaer, J.; Bulyk, M. L. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.* **2009**, *19*, 556−566.

(4) Badis, G.; Chan, E. T.; van Bakel, H.; Pena-Castillo, L.; Tillo, D.; Tsui, K.; Carlson, C. D.; Gossett, A. J.; Hasinoff, M. J.; Warren, C. L.; Gebbia, M.; Talukder, S.; Yang, A.; Mnaimneh, S.; Terterov, D.; Coburn, D.; Yeo, A. L.; Yeo, Z. X.; Clarke, N. D.; Lieb, J. D.; Ansari, A. Z.; Nislow, C.; Hughes, T. R. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol. Cell* **2008**, *32*, 878−887.

(5) Luscombe, N. M.; Laskowski, R. A.; Thornton, J. M. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.* **2001**, *29*, 2860−2874.

(6) Rohs, R.; Jin, X.; West, S. M.; Joshi, R.; Honig, B.; Mann, R. S. Origins of specificity in protein-DNA recognition. *Annu. Rev. Biochem.* **2010**, *79*, 233−269.

(7) Gaj, T.; Gersbach, C. A.; Barbas, C. F. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* **2013**, *31*, 397−405.

(8) Seeman, N. C.; Rosenberg, J. M.; Rich, A. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. U. S. A.* **1976**, *73*, 804−808.

(9) Rohs, R.; West, S. M.; Liu, P.; Honig, B. Nuance in the double-helix and its role in protein-DNA recognition. *Curr. Opin. Struct. Biol.* **2009**, *19*, 171−177.

(10) Kim, Y.; Geiger, J. H.; Hahn, S.; Sigler, P. B. Crystal structure of a yeast TBP/TATA-box complex. *Nature* **1993**, *365*, 512−520.

(11) Otwinowski, Z.; Schevitz, R. W.; Zhang, R. G.; Lawson, C. L.; Joachimiak, A.; Marmorstein, R. Q.; Luisi, B. F.; Sigler, P. B. Crystal structure of trp repressor/operator complex at atomic resolution. *Nature* **1988**, *335*, 321−329.

(12) Jones, S.; van Heyningen, P.; Berman, H. M.; Thornton, J. M. Protein-DNA interactions: A structural analysis. *J. Mol. Biol.* **1999**, *287*, 877−896.

(13) 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.

(14) Hoffman, M. M.; Khrapov, M. A.; Cox, C. J.; Yao, J.; Tong, L.; Ellington, A. D. AANT: the Amino Acid-Nucleotide Interaction Database. *Nucleic Acids Res.* **2004**, *32*, 174−181.

(15) Jones, S.; Shanahan, H. P.; Berman, H. M.; Thornton, J. M. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.* **2003**, *31*, 7189−7198.

(16) Černý, J.; Hobza, P. Non-covalent interactions in biomacromolecules. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5291−5303.

DOI: 10.1021/ct501168n
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX

(17) Grimme, S. Accurate description of van der Waals complexes by density functional theory including empirical corrections. *J. Comput. Chem.* **2004**, *25*, 1463−1473.
(18) Černý, J.; Pitoňák, M.; Riley, K. E.; Hobza, P. Complete basis set extrapolation and hybrid schemes for geometry gradients of non-covalent complexes. *J. Chem. Theory Comput.* **2011**, *7*, 3924−3934.
(19) Berka, K.; Laskowski, R. A.; Riley, K. E.; Hobza, P.; Vondrášek, J. Representative amino acid side chain interactions in proteins. A comparison of highly accurate correlated ab initio quantum chemical and empirical potential procedures. *J. Chem. Theory Comput.* **2009**, *5*, 982−992.
(20) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(21) Wang, J.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **2000**, *21*, 1049−1074.
(22) 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* **2006**, *65*, 712−725.
(23) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **2010**, *78*, 1950−1958.
(24) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, *24*, 1999−2012.
(25) Mackerell, A. D.; Banavali, N.; Foloppe, N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* **2001**, *56*, 257−265.
(26) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattnos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Marplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Ptudies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(27) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(28) 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.
(29) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. Electronic structure calculations on workstation computers: The program system turbomole. *Chem. Phys. Lett.* **1989**, *162*, 165−169.
(30) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.
(31) Berka, K.; Laskowski, R. A.; Hobza, P.; Vondrášek, J. Energy matrix of structurally important side-chain/side-chain interactions in proteins. *J. Chem. Theory Comput.* **2010**, *6*, 2191−2203.
(32) Klimasauskas, S.; Kumar, S.; Roberts, R. J.; Cheng, X. HhaI methyltransferase flips its target base out of the DNA helix. *Cell* **1994**, *76*, 357−369.
(33) Rohs, R.; West, S. M.; Sosinsky, A.; Liu, P.; Mann, R. S.; Honig, B. The role of DNA shape in protein-DNA recognition. *Nature* **2009**, *461*, 1248−1253.
(34) Benos, P. V.; Bulyk, M. L.; Stormo, G. D. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.* **2002**, *30*, 4442−4451.
(35) Bergqvist, S.; Williams, M. A.; O'Brien, R.; Ladbury, J. E. Heat capacity effects of water molecules and ions at a protein-DNA interface. *J. Mol. Biol.* **2004**, *336*, 829−842.
(36) Guckian, K. M.; Schweitzer, B. A.; Ren, R. X.-F.; Sheils, C. J.; Tahmassebi, D. C.; Kool, E. T. Factors contributing to aromatic stacking in water: evaluation in the context of DNA. *J. Am. Chem. Soc.* **2000**, *122*, 2213−2222.
(37) Foguel, D.; Silva, J. L. Cold denaturation of a repressor-operator complex: the role of entropy in protein-DNA recognition. *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91*, 8244−8247.
(38) Cieplak, P.; Dupradeau, F.-Y.; Duan, Y.; Wang, J. Polarization effects in molecular mechanical force fields. *J. Phys.: Condens. Matter* **2009**, *21*, 333102.
(39) Šponer, J.; Hobza, P. Nonplanar geometries of DNA bases. Ab initio second-order Moeller-Plesset study. *J. Phys. Chem.* **1994**, *98*, 3161−3164.
(40) Cleland, W. W. Low-barrier hydrogen bonds and enzymatic catalysis. *Arch. Biochem. Biophys.* **2000**, *382*, 1−5.
(41) Rajagopal, S.; Vishveshwara, S. Short hydrogen bonds in proteins. *FEBS J.* **2005**, *272*, 1819−1832.

DOI: 10.1021/ct501168n
J. Chem. Theory Comput. XXXX, XXX, XXX−XXX